

# MULTIMODAL ONE-SHOT LEARNING OF SPEECH AND IMAGES

by

Ryan Eloff



*Thesis presented in partial fulfilment of the requirements for the  
degree of Master of Engineering (Electrical & Electronic) in the  
Faculty of Engineering at Stellenbosch University*

Supervisor: Dr H. Kamper

Co-supervisor: Prof. H. A. Engelbrecht

March 2020

# PLAGIARISM DECLARATION

1. Plagiarism is the use of ideas, material and other intellectual property of another's work and to present it as my own.  
*Plagiaat is die oorneem en gebruik van die idees, materiaal en ander intellektuele eiendom van ander persone asof dit jou eie werk is.*
2. I agree that plagiarism is a punishable offence because it constitutes theft.  
*Ek erken dat die pleeg van plagiaat 'n strafbare oortreding is aangesien dit 'n vorm van diefstal is.*
3. I also understand that direct translations are plagiarism.  
*Ek verstaan ook dat direkte vertalings plagiaat is.*
4. Accordingly all quotations and contributions from any source whatsoever (including the internet) have been cited fully. I understand that the reproduction of text without quotation marks (even when the source is cited) is plagiarism.  
*Dienooreenkomstig is alle aanhalings en bydraes vanuit enige bron (ingesluit die internet) volledig verwys (erken). Ek erken dat die woordelike aanhaal van teks sonder aanhalingstekens (selfs al word die bron volledig erken) plagiaat is.*
5. I declare that the work contained in this assignment, except where otherwise stated, is my original work and that I have not previously (in its entirety or in part) submitted it for grading in this module/assignment or another module/assignment.  
*Ek verklaar dat die werk in hierdie skryfstuk vervat, behalwe waar anders aangedui, my eie oorspronklike werk is en dat ek dit nie vantevore in die geheel of gedeeltelik ingehandig het vir bepunting in hierdie module/werkstuk of 'n ander module/werkstuk nie.*

Student number / <i>Studentenommer</i>	Signature / <i>Handtekening</i>
Initials and surname / <i>Voorletters en van</i>	Date / <i>Datum</i>

# ABSTRACT

## MULTIMODAL ONE-SHOT LEARNING OF SPEECH AND IMAGES

R. Eloff

*Department of Electrical and Electronic Engineering,  
University of Stellenbosch,  
Private Bag XI, Matieland 7602, South Africa.*

Thesis: MEng (E&E)

March 2020

Humans learn to perform tasks such as language understanding and visual perception, remarkably, without any annotations and from limited amounts of weakly supervised co-occurring sensory information. Meanwhile, state-of-the-art machine learning models—which aim to challenge these human learning abilities—require large amounts of labelled training data to enable successful generalisation. Multimodal one-shot learning is an effort towards closing this gap on human intelligence, whereby we propose benchmark tasks for machine learning systems investigating whether they are capable of performing cross-modal matching from limited weakly supervised data. Specifically, we consider spoken word learning with co-occurring visual context in a one-shot setting, where an agent must learn novel concepts (words and object categories from a single joint audio-visual example. In this thesis, we make the following contributions: (i) we propose and formalise multimodal one-shot learning of speech and images; (ii) we develop two cross-modal matching benchmark datasets for evaluation, the first containing spoken digits paired with handwritten digits, and the second containing complex natural images paired with spoken words; and (iii) we investigate a number of models within two frameworks, one extending unimodal models to the multimodal case, and the other learning joint audio-visual models. Finally, we show that jointly modelling spoken words paired with images enables a novel multimodal gradient update within a meta-learning algorithm for fast adaptation to novel concepts. This model outperforms our other approaches on our most difficult benchmark with a cross-modal matching accuracy of 40.3% for 10-way 5-shot learning. Although we show that there is room for significant improvement, the goal of this work is to encourage further development on this challenging task. We hope to achieve this by defining a standard problem setting with tasks which may be used to benchmark other approaches.

# UITTREKSEL

## **MULTIMODAL ONE-SHOT LEARNING OF SPEECH AND IMAGES**

R. Eloff

*Department of Electrical and Electronic Engineering,  
University of Stellenbosch,  
Private Bag XI, Matieland 7602, South Africa.*

Tesis: MEng (E&E)

Maart 2020

Die mens het die merkwaardige vermoë om taal en visuele konsepte aan te leer sonder geannoteerde afgedata deur gebruik te maak van swak toesig in die vorm van parallelle sensoriese intree. Intussen benodig die beste getoesigde masjienleermodelle massiewe geannoteerde datastelle om te veralgemeen na nuwe intrees. Multimodale eenskootmasjienleer is 'n poging om die gaping tussen die vermoëns van masjienleermodelle te oorbrug. Hier stel ons 'n aantal standaard toetse voor om te bepaal of nuwe masjienleerstelsels die vermoë het om kruismodale passing uit te voer uit slegs 'n paar voorbeelde met beperkte toesig. Meer spesifiek ondersoek ons hoe gesproke woorde wat met ooreenstemmende visuele konsepte voorkom, saam aangeleer kan word in 'n eenskootopstelling waar 'n masjien nuwe konsepte (woord en objekkatgorieë uit 'n enkele gesamentlike oudiovisuele voorbeeld moet aanleer. Ons maak die volgende bydraes: (i ons formaliseer multimodale eenskootmasjienleer uit spraak en beelde; (ii ons ontwikkel twee datastelle wat dien as maatstawwe om kruismodale passing te evalueer: die eerste datastel bestaan uit gesproke syfers met gepaardgaande handgeskrewe syfers en die tweede bestaan uit meer komplekse fotos met geïsoleerde woorde; en (iii ons ondersoek verskeie masjienleermodelle in twee opstellings: een waar enkelmodale modelle uitgebrei word na die multimodale geval en die ander waar oudiovisuele modelle gesamentlik afgerig word. Laastens ondersoek ons die gesamentlike aanleer van gesproke woorde met gepaardgaande visuele konsepte deur gebruik te maak van 'n meta-leer-algoritme. Hierdie model vaar die beste in ons moeilikste toetsomgewing, met 'n kruismodale passingsakkuraatheid van 40.3% vir 10-rigting 5-skoot masjienleer. Ons hoop dat deur hierdie probleem formeel te definieer en standaard toets beskikbaar te stel, ons verdere navorsing in hierdie nuwe en uitdagende veld sal aanmoedig.

## ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to the following people:

- The real MVP, Dr Herman Kamper, who went far above and beyond the expectations of a supervisor. Thank you for your support, guidance and continuous involvement throughout the projects we undertook.
- My co-supervisor, Prof. Herman A. Engelbrecht, for your help, guidance and support throughout the past couple years.
- Kristina Gulordava for your interest and valuable input from the human language acquisition literature.
- The friends that I met in the Media Lab who kept things interesting. Thank you especially to the unofficial “Media Lab Machine Learning” group for the stand-up meetings and many interesting discussions.
- Jeanne—of the House Vindaloo, The Unburnt, Queen of Indian Cuisine, Khaleesi of the Great Spicy Broccoli, Protector of the Naan, Breaker of Eggs and Mother of Avocados. Thank you for your endless support, love, advice and delicious food. Many thanks to your family for their continuous support and amazing rusks.
- My family: Mom, Dad, Kevin, Kaela, Granny, Grandpa. Thank you for your love and support over the years, without which I would not have made it this far. Special shout out to Kaela, an aspiring physiotherapist, who massaged out all stress during the final sprint. Thanks as well to Peppa the family cat who insisted on being part of this thesis, leaving holes in many research papers 🐾.
- My cousins, the Giffard family. This was a tough year for all of us as we fought alongside Crys Giffard in her battle against cancer. Although she left us far too early, I know that her memory will live on in our hearts and that we will continue to support each other through all of life’s ups and downs. Thank you for everything. *Together we are stronger.*

---

The financial assistance of the Stellenbosch University (SU) Wilhelm Frank Scholarship towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to SU.

Thanks to NVIDIA for sponsoring a Titan Xp GPU for this work.

# DEDICATIONS

*This thesis is dedicated to family.*

\* \* \*

*A man should never neglect his family for business.*  
— Walt Disney

# CONTENTS

<b>Plagiarism Declaration</b>	<b>ii</b>
<b>Abstract</b>	<b>ii</b>
<b>Uittreksel</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>iv</b>
<b>Dedications</b>	<b>v</b>
<b>Contents</b>	<b>vi</b>
<b>List of Figures</b>	<b>ix</b>
<b>List of Tables</b>	<b>xii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Problem Statement . . . . .	2
1.3 Approach . . . . .	3
1.4 Thesis Overview . . . . .	4
1.5 Contributions . . . . .	5
<b>I Foundations</b>	<b>7</b>
<b>2 Multimodal One-Shot Learning</b>	<b>8</b>
2.1 The One-Shot Learning Problem . . . . .	8
2.2 Defining Multimodal One-Shot Learning . . . . .	9
2.3 Cross-Modal Matching of Speech to Images . . . . .	10
2.4 Few-Shot Learning of Speech and Images . . . . .	11
<b>3 Related Work</b>	<b>12</b>
3.1 One-Shot Learning Observed in Humans and Cognitive Modelling .	12
3.2 Unimodal Vision One-Shot Learning . . . . .	15
3.3 One-Shot Learning in Other Domains . . . . .	16

3.4	One-Shot Learning with Meta-Learning . . . . .	16
3.5	Zero-Shot Learning . . . . .	17
3.6	Multimodal Representation Learning . . . . .	17
<b>II</b>	<b>One-Shot Learning Handwritten Digits Paired with Spoken Digits</b>	<b>19</b>
<b>4</b>	<b>A Framework for Multimodal One-Shot Learning</b>	<b>20</b>
4.1	Indirect Matching Framework . . . . .	20
4.2	Direct Matching of Raw Features . . . . .	22
4.3	Transfer Learning with Neural Network Classifiers . . . . .	22
4.4	Metric Learning with Siamese Neural Networks . . . . .	25
4.5	Chapter Summary . . . . .	27
<b>5</b>	<b>TIDIGITS and MNIST Experiments</b>	<b>28</b>
5.1	A Multimodal Digits Benchmark Dataset . . . . .	28
5.2	Background Data for Neural Network Models . . . . .	29
5.3	Experimental Setup . . . . .	29
5.3.1	Data Preprocessing . . . . .	30
5.3.2	Model Implementation . . . . .	30
5.3.3	One-Shot Task Evaluation . . . . .	32
5.4	Experimental Evaluation . . . . .	32
5.4.1	One-Shot Speech Classification . . . . .	32
5.4.2	One-Shot Image Classification . . . . .	33
5.4.3	One-Shot Cross-Modal Matching of Speech to Images . . . . .	34
5.4.4	Analysis of Speaker Invariance . . . . .	34
5.5	Chapter Summary . . . . .	36
<b>III</b>	<b>One-Shot Learning Natural Images Paired with Spoken Words</b>	<b>37</b>
<b>6</b>	<b>A Direct Framework for Matching Speech to Images</b>	<b>38</b>
6.1	Direct Matching Framework . . . . .	38
6.2	Deep Audio-Visual Embedding Networks . . . . .	39
6.3	Multimodal Model-Agnostic Meta-Learning . . . . .	45
6.4	Chapter Summary . . . . .	50
<b>7</b>	<b>Flickr 8K and Flickr Audio Experiments</b>	<b>51</b>
7.1	A Multimodal Natural Speech and Images Benchmark Dataset . . . . .	51
7.2	Background Data for Neural Network Models . . . . .	53
7.3	Experimental Setup . . . . .	55
7.3.1	Data Preprocessing . . . . .	55
7.3.2	Model Implementation . . . . .	56
7.3.3	One-Shot Task Evaluation . . . . .	60
7.4	Experimental Evaluation . . . . .	60
7.4.1	One-Shot Speech Classification . . . . .	60
7.4.2	One-Shot Image Classification . . . . .	62

<i>CONTENTS</i>	<b>viii</b>
7.4.3 One-Shot Cross-Modal Matching of Speech to Images . . .	62
7.4.4 Analysis of Speaker Invariance . . . . .	63
7.5 Chapter Summary . . . . .	65
<b>8 Conclusion</b>	<b>66</b>
8.1 Summary and Conclusions . . . . .	66
8.2 Future Work . . . . .	66
<b>IV Appendices</b>	<b>68</b>
<b>A TIDIGITS and MNIST Experiments</b>	<b>69</b>
A.1 Analysis of Speaker Invariance in One-Shot Speech Classification .	69
<b>B Flickr 8K and Flickr Audio Experiments</b>	<b>70</b>
B.1 Natural Speech and Images Benchmark Dataset Image Excerpts . .	70
B.2 Flickr 8K and Flickr Audio Background Dataset Keyword Classes .	73
<b>References</b>	<b>74</b>

## LIST OF FIGURES

2.1	Unimodal one-shot speech learning and classification. . . . .	9
2.2	Multimodal one-shot learning and matching of speech and images. . . . .	10
3.1	Example propose-but-verify experiment (Trueswell <i>et al.</i> , 2013). The word “zud” is presented in two learning trials containing five alternative visual items, with the intended meaning bear. Figure reproduced from Trueswell <i>et al.</i> (2013). . . . .	13
3.2	Example of the zero-shot fast mapping task (Lazaridou <i>et al.</i> , 2014). A potential wampimuk with linguistic context “We found a cute, hairy wampimuk sleeping behind the tree” (a) and the linguistic context together with a projection of the wampimuk image in linguistic space (b). . . . .	14
4.1	Multimodal one-shot learning framework for indirect matching of speech to images. Support set speech items are represented with coloured blocks for simplicity. The speech query “two” is matched to the matching set image of a <i>two</i> using unimodal comparisons and retrieving an auxiliary image query from the multimodal support set which may be compared to the matching set images. . . . .	21
4.2	Optimal non-linear alignment computed with dynamic time warping over the acoustic features for two speech segments, with different lengths and both containing instances of the spoken word “six”. . . . .	23
4.3	Transfer learning illustrated for spoken word classification. The goal is to learn model parameters $\theta$ on a source task (purple circles) which has many labelled examples. The shared knowledge among the source task is used to learn a prior, an optimal configuration of $\theta$ shown at the end of the trajectory, that is useful for related target tasks (blue circles) which have limited data. . . . .	24

4.4	Siamese triplet convolutional neural network that takes in triplet pairs of spoken words for learning speech representations useful for comparison. Speech input features are extracted mel-frequency cepstral coefficients centre zero padded to 120 frames (§5.3.1 and §7.3.1). The anchor and positive examples are instances of spoken word “motorcycle” and the negative example is an instance of spoken word “bicycle”. The parameters of the tied network branches are trained with a triplet loss that optimises the distance between the <i>motorcycle</i> representations to be smaller than the distance between the <i>motorcycle</i> and <i>bicycle</i> representations. . . . .	26
5.1	TIDIGITS isolated spoken digits (a) and MNIST handwritten digit images (b). . . . .	29
5.2	Flickr Audio isolated spoken words (a) and Omniglot handwritten character images (b). . . . .	30
6.1	Multimodal one-shot learning framework for direct matching of speech to images. The support set contains multimodal pairs for the classes. The speech query “bird” is matched directly to the matching set image of a <i>bird</i> in joint audio-visual space. This is accomplished by adapting a joint audio-visual model on the weakly supervised support set during multimodal one-shot learning. In the absence of the support set (i.e. no learning phase) this is equivalent to zero-shot learning. . . . .	40
6.2	Speech network branch of a deep audio-visual embedding network applied to a spoken word “bicycle”. Speech input features are extracted mel-frequency cepstral coefficients centre zero padded to 140 frames (§7.3.1). Convolution layers (1-D) shown in green, batch normalisation layer shown in red, max pooling layers shown in purple, average pooling layer shown in orange and fully-connected linear layer with L2 normalisation shown in yellow. . . . .	42
6.3	Vision network branch of a deep audio-visual embedding network applied to an image of a <i>bicycle</i> . Convolution layers (2-D) shown in green, average pooling layer shown in orange and fully-connected linear layer with L2 normalisation shown in yellow. . . . .	42
6.4	Transfer learning (a) and meta-learning (b) illustrated for spoken word classification. In both cases, the goal is to learn model parameters $\theta$ on a source task (purple circles), where in (a) we consider a task with many labelled examples, while in (b) we consider many tasks with few labelled examples, one of which is shown here. The shared knowledge among the source task (or tasks) is used to learn a prior that is useful for learning related target tasks (blue circles) which have limited data. Here the form of this learned prior is optimal initialisation for model parameters such that fine-tuning may adapt parameters to the target tasks. Meta-learning explicitly optimises this prior such that <i>fine-tuning is effective</i> on few examples. . . . .	44

6.5	Meta-learning demonstrated for spoken word classification tasks. Model parameters $\theta$ are fine-tuned (dashed lines) on a small number of meta-training data from each task and then evaluated on meta-testing data at these updated parameters. The resulting meta-testing error is used to meta-learn $\theta$ (solid line) such that it is fast to adapt to new tasks. . . . .	47
7.1	The image from the natural speech and images benchmark dataset paired with spoken word “bird”. This demonstrates the difficulty of the multi-modal one-shot learning task, where a learner must simultaneously learn a new word, identify which visual object or concept this word refers to and generalise this word-object mapping to unseen instances of the spoken word and visual object. . . . .	52
B.1	Excerpt of images paired with spoken word “bird”. . . . .	70
B.2	Excerpt of images paired with spoken word “surfboard”. . . . .	71
B.3	Excerpt of images paired with spoken word “guitar”. . . . .	72

## LIST OF TABLES

2.1	Multimodal one-shot learning terminology. . . . .	11
5.1	11-way 1-shot and 5-shot speech classification results on isolated spoken digits sampled from the TIDIGITS dataset. . . . .	33
5.2	10-way 1-shot and 5-shot image classification results on handwritten visual digits sampled from the MNIST dataset. . . . .	33
5.3	11-way 1-shot and 5-shot cross-modal matching of isolated spoken digits to handwritten visual digits sampled from the paired TIDIGITS and MNIST benchmark dataset for evaluating MOONSHOT. . . . .	35
5.4	Speaker invariance tests for 11-way 1-shot cross-modal speech-image digit matching. All support set items are from the same speaker as the speech query, except for the support set item actually matching the query. . . . .	35
7.1	10- and 20-way 1- and 5-shot speech classification results on isolated spoken words sampled from the Flickr Audio dataset. . . . .	61
7.2	10- and 20-way 1- and 5-shot image classification results on natural images sampled from the Flickr 8K dataset. . . . .	61
7.3	10- and 20-way 1- and 5-shot cross-modal matching of isolated spoken words to natural images sampled from the paired Flickr Audio and Flickr 8K benchmark dataset for evaluating MOONSHOT. . . . .	64
7.4	Speaker invariance tests for 10-way 1- and 5-shot cross-modal natural speech-image matching. All support set items are from the same speaker as the speech query, except for the support set item actually matching the query. . . . .	64
A.1	Speaker invariance tests for 11-way 1-shot speech classification. All support set items are from the same speaker as the query, except for the support set item actually matching the query. . . . .	69

# 1 | INTRODUCTION

*Ancora Imparo*  
“I am still learning”  
— Michelangelo, age 87

Humans possess the remarkable ability to learn new words and object categories from only one or a few examples (Carey, 1978; Carey and Bartlett, 1978; Markson and Bloom, 1997). This may be observed in children: a child hearing the word “lego” (/ˈlɛɡoʊ/) for the first time in the context of receiving a new toy, can quickly learn to associate the spoken word “lego” to the new (visual) concept *lego*. Current state-of-the-art speech and vision processing algorithms require thousands of labelled examples to complete a similar task (Ngiam *et al.*, 2011), despite the major advances deep learning has made in speech processing (Hinton *et al.*, 2012) and computer vision (Krizhevsky *et al.*, 2012). This has led to research in *one-shot learning* (Fei-Fei *et al.*, 2006; Lake *et al.*, 2011, 2013, 2014, 2015; Koch *et al.*, 2015), where the task is acquisition of novel concepts from only one or a few labelled examples. How can we build intelligent agents that are capable of efficiently learning new spoken words and visual objects from a single experience?

## 1.1 MOTIVATION

One-shot learning studies have primarily focused on problems where novel concepts in a single modality are observed along with class labels. This is different from the example above: the child directly associates the spoken word “lego” to the visual signal of lego without any additional supervision (e.g. class labels), and can generalise this single example to other visual or spoken instances of *lego*. Similarly, a child hearing the word “dog” (/dɒg/) in reference to an animal the child has not seen before, may quickly infer that the concept *dog* generalises to Labrador, German Shepherd and other dog breeds. Humans successfully learn without access to thousands of labelled data samples for each specific task—the focus of unimodal one-shot learning studies. More interestingly, as demonstrated in these examples, they do so without any strong supervisory signals. This motivates *multimodal one-shot learning*, a new problem setting which we formalise in this thesis. Consider an agent such as a household robot that is shown a single visual example of *milk*,

*eggs, butter* and a *mug*, novel objects each paired with a spoken description. During subsequent use, a speech query is given and the agent needs to identify which visual object the query refers to. The specific speech query and visual objects in this testing phase are different instances from those observed before; the agent must effectively generalise from the joint audio-visual examples presented in the learning phase.

Learning to recognise new words and object categories in the one-shot setting is motivated by our observation of humans. In cognitive psychology, research has shown that both children and adults are capable of learning novel words from only one or a few exposures; retaining knowledge for a period of time by a process known as *fast mapping* (Carey, 1978; Carey and Bartlett, 1978; Markson and Bloom, 1997). Importantly, this mechanism of word learning is observed in the presence of some informative context—for example, a visual object. While it is postulated that word learning ensues in co-occurring context, it is difficult to pinpoint how humans manage to do this in complex natural settings which offer many interpretable meanings. Trueswell *et al.* (2013) suggest that humans learn words in this setting by following a one-trial procedure, followed by verification, building on their prior experiences. This form of one-trial learning has also been observed in Bayesian modelling of word learning (Frank *et al.*, 2009). The multimodal one-shot learning problem follows naturally, as we question whether we may develop a machine learning system which exhibits a similar ability to quickly learn novel words and visual objects in such complex settings from only co-occurring context. Such a system is desirable in its ability to learn from limited information, reducing the need for large labelled datasets typically required by data-hungry neural networks, as well as enabling continual learning of new concepts in online settings. Additionally, this could aid further research on human language acquisition. Expanding on Harwath *et al.* (2020), we believe that humans provide living proof of language acquisition from spoken words and visual perception without additional supervision and from limited data—thus it is plausible for a machine learning system.

This setting is relevant in a number of domains: modelling infant language acquisition, where models can be used to test particular cognitive hypotheses (Räsänen and Rasilo, 2015); low-resource speech processing, where new concepts could be taught in an arbitrary language (Besacier *et al.*, 2014); cross-modal retrieval, where a single audio-visual template pair could be used to enable retrieval of images similar to a novel spoken query (Harwath *et al.*, 2016); and robotics, where novel concepts must be acquired online from co-occurring multimodal sensory inputs (Walter *et al.*, 2012; Taniguchi *et al.*, 2016; Thomason and Knepper, 2016; Renkens and Van hamme, 2017, 2018).

## 1.2 PROBLEM STATEMENT

In this thesis, we do not attempt to model human language acquisition nor solve this problem for our artificial agents. Instead, we present multimodal one-shot learning and benchmarks as a step toward the goal of intelligent agents with more human-

like learning abilities. We specifically consider multimodal one-shot learning on datasets of spoken words paired with images.<sup>1</sup> During a multimodal one-shot learning *episode*, a model is shown a set of speech-image pairs, one for each of the novel concepts it should learn. This set, which we refer to as the *support set*, is acquired before the model is evaluated and must be used to learn the new concepts (i.e. spoken words and visual objects). To evaluate the effectiveness of the multimodal one-shot learner, we perform a cross-modal matching task: during testing, the model is shown a spoken word, called the *query*, and a set of test images, called the *matching set*. The test query and matching set contain unseen instances of concepts seen in the support set. The model then needs to predict which test image in the matching set corresponds to the spoken query. We refer to this downstream evaluation task as *cross-modal matching of speech to images*. In assessing this task, we propose two benchmark datasets from which multimodal one-shot learning episodes are sampled. The first contains isolated spoken digits paired with handwritten digit images—an “MNIST” of weakly supervised audio-visual learning. The second contains natural images paired with spoken keywords—a more realistic and complex setting.

### 1.3 APPROACH

We approach this problem within two simple frameworks: (i) indirect matching of speech to images and (ii) direct matching of speech to images. The indirect matching framework extends existing unimodal one-shot learning models to the multimodal case. This is achieved by unimodal comparisons through the support set. Given an input speech query, we find the closest speech segment in the support set. We then take its paired support image as an auxiliary query, and find its closest image in the matching set. This image is predicted as the match. Metrics for speech-speech and image-image comparisons need to be defined, and this is where we take advantage of the large body of work in unimodal one-shot learning to investigate several options. One approach is to use labelled background training data not containing any of the classes which will be seen during one-shot learning. Using such speech and image background data, we specifically investigate Siamese neural networks (Bromley *et al.*, 1994; Chopra *et al.*, 2005) as a way to explicitly train unimodal distance metrics. We also incorporate recently proposed advances (Koch *et al.*, 2015; Chechik *et al.*, 2010; Wang *et al.*, 2014; Hermann and Blunsom, 2014; Hoffer and Ailon, 2015; Schroff *et al.*, 2015; Hermans *et al.*, 2017) for such networks. One disadvantage of this approach is the unimodal distance metrics cannot be fine-tuned on multimodal one-shot support samples. As a result, these models may not take full advantage of the learning samples; in addition, learning may not scale as we observe more learning samples in a few-shot setting.

The direct matching framework considers models that jointly learn audio-visual representations so that we may match speech to images directly. The aim of this

---

<sup>1</sup>Although the focus of this thesis is on speech and images, the multimodal one-shot learning problem is applicable to any source of paired information in multiple modalities.

approach is to improve on the indirect matching framework with models which may be fine-tuned on the weakly supervised multimodal one-shot support set. Similar to before, this approach uses weakly supervised background data. We investigate deep audio-visual embedding networks (DAVEnet) (Harwath *et al.*, 2018) to train a multimodal distance metric. Following this, we investigate a novel extension of the popular model-agnostic meta-learning algorithm (MAML) (Finn *et al.*, 2017a) to allow for fast adaptation in the weakly supervised multimodal case. The MAML algorithm has been shown to generalise better on few-shot and out-of-distribution tasks (Finn and Levine, 2018).

Common to both of these frameworks is the use of background data to train a model that learns to adapt to novel concepts. Once again, this is similar to how humans learn. A baby does not immediately know how to speak, understand language and identify visual objects. Instead, humans must learn by building on their prior experience with such tasks. The type of experience is also important, since knowing how to walk will not enable a child to quickly learn a new language. Then the one-shot tasks that we are interested in must be similar to the prior experience that our models are trained on. In other words, the background data should be sampled from an underlying data-generating distribution similar to that of the one-shot tasks. There is still no free lunch. While the current paradigm in machine learning is to train an expert on one specific task—e.g. recognising a specific set of spoken words or visual objects—we aim to enable agents that continually learn new tasks, building on prior knowledge and from very little data.

## 1.4 THESIS OVERVIEW

This thesis is organised in parts. In Part I we discuss the foundations of multimodal one-shot learning. In Part II we introduce an indirect modelling approach and evaluate models on multimodal digits benchmark tasks. Finally, in Part III we introduce a direct modelling approach and evaluate models on complex multimodal natural setting benchmark tasks. The individual chapters of this thesis proceed as follows:

### *Part I*

- In Chapter 2, we look at the one-shot learning problem, followed by our formal definition of multimodal one-shot learning. We also describe a cross-modal matching evaluation task.
- In Chapter 3, we provide a brief overview of the literature related to multimodal one-shot learning of speech and images.

### *Part II*

- In Chapter 4, we introduce a framework for multimodal one-shot learning relying on indirect matching via unimodal comparisons. Specifically, we inves-

tigate direct feature comparisons and neural network approaches to learning useful features for comparison.

- In Chapter 5, we propose and evaluate multimodal one-shot learning benchmark tasks on handwritten digits paired with spoken digits. We show that our indirect matching models achieve reasonable results, but suffer from compounding errors as a result of successive unimodal comparisons.

### *Part III*

- In Chapter 6, we introduce an improved framework for multimodal one-shot learning which makes use of a joint mapping to directly match speech and images. To achieve this, we investigate deep audio-visual embedding networks and a weakly supervised multimodal variant of model-agnostic meta-learning.
- In Chapter 7, we propose and evaluate multimodal one-shot learning benchmark tasks on natural images paired with spoken words. We present results for both our direct and indirect modelling approaches, showing that the direct matching models improve results on five-shot tasks as they may be fine-tuned directly on multimodal learning samples.
- Finally, we present a summary of our conclusions in Chapter 8.

Chapters 2, 4 and 5 cover work published previously as [Eloff \*et al.\* \(2019\)](#).

## 1.5 CONTRIBUTIONS

Our main contribution is the formal definition of multimodal one-shot learning. We also develop two one-shot cross-modal matching datasets that may be used to benchmark other approaches. As an intermediate evaluation in our work, we consider unimodal one-shot speech and image classification. Apart from [Lake \*et al.\* \(2014\)](#), this thesis is to our knowledge the only work that considers one-shot unimodal learning of spoken language. We present several new models and baselines not considered in [Lake \*et al.\* \(2014\)](#). A subset of the work in this thesis appeared previously as [Eloff \*et al.\* \(2019\)](#) in a paper titled “Multimodal One-shot Learning of Speech and Images”, presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing*. Overall, the goal of this thesis is to motivate others to work on multimodal one-shot learning, which we hope will provide solutions to some of the hardest problems we currently face in machine learning.

To summarise, we make the following core contributions:

- We introduce and formalise the multimodal one-shot learning problem in the context of modelling speech and images.
- We define a cross-modal matching task which may be used to evaluate models for multimodal one-shot learning.

- We propose an indirect matching framework for one-shot cross-modal matching of speech to images, extending unimodal one-shot learning models to the multimodal case.
- We investigate several models within the indirect matching framework, specifically considering direct feature matching, transfer learning with neural network classifiers and metric learning with Siamese neural networks.
- We develop a simple benchmark dataset for multimodal one-shot learning containing spoken digits paired with handwritten digit images, which we use to evaluate our indirect matching models. This results in a number of baselines which may be used to benchmark other approaches.
- We release the splits for the digits benchmark dataset and the code to reproduce the experiments on this dataset as a contribution. The code recipe for the digits experiments is publicly available at:  
<https://github.com/rpeloff/multimodal-one-shot-learning>
- We propose a direct matching framework for one-shot cross-modal matching of speech to images, where we apply joint audio-visual models for comparing speech to images directly.
- We investigate several models within the direct matching framework, specifically considering end-to-end architectures which model speech and images in a joint audio-visual space. We accomplish this by investigating joint speech-image metric learning with deep audio-visual embedding networks. Furthermore we investigate a novel extension of multimodal model-agnostic meta-learning to the multimodal case to explicitly train such networks which learn to learn from small amounts of speech paired with images and no labels.
- We develop a more complex and realistic benchmark dataset for multimodal one-shot learning containing natural images paired with spoken words, which we use to evaluate our indirect and direct matching models. This results in a new baselines which may be used to benchmark other approaches.
- We release the splits for the natural speech-image benchmark dataset and the code to reproduce the experiments on this dataset as a contribution. The code recipe for the natural speech-image experiments is publicly available at:  
<https://github.com/rpeloff/moonshot>
- We consider one-shot unimodal modelling of spoken digits and natural spoken words. To our knowledge, this thesis is the only work that considers one-shot unimodal learning of spoken language, apart from [Lake \*et al.\* \(2014\)](#) who did not consider the neural network models that we investigate.

PART I  
FOUNDATIONS

## 2 | MULTIMODAL ONE-SHOT LEARNING

*Why shoot for the moon? It matters because when you try to do something radically hard, you approach the problem differently than when you try to make something incrementally better.*

— Astro Teller

In this chapter we first describe the unimodal one-shot learning problem as defined in prior work. Thereafter, we provide our interpretation and formal definition of multimodal one-shot learning, which we refer to as *MOONSHOT* (Multimodal One-SHOT learning). Finally, we describe a cross-modal matching task which we use to evaluate our MOONSHOT modelling approaches.

### 2.1 THE ONE-SHOT LEARNING PROBLEM

The goal of unimodal one-shot learning is to build a model that can acquire new concepts after observing only a single labelled example from each class. This model must then successfully generalise to new instances of those concepts in tasks such as classification or regression.

For instance, given one example for each of 10 unique digit images, with labels, the task is to learn from these limited examples such that unseen instances of test digit images may be correctly classified. Formally, in an  $L$ -way one-shot *episode* a model is shown a *support set*  $\mathcal{S} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^L$ , containing one labelled example for each of  $L$  classes. From this set, it must learn a classifier  $C_{\mathcal{S}}$  for unseen test *queries*  $\hat{\mathbf{x}}$ . This is illustrated in Figure 2.1 for five-way one-shot speech classification. In this case the support set  $\mathcal{S}$  contains spoken utterances along with hard textual labels. A model must use this information to classify the spoken test query “two” (/tu/) as the concept label *two*. Note that the test-time query does not occur in the support set itself—it is an unseen instance of a class occurring in the support set.

Evaluating a one-shot learner proceeds by sampling a number of  $L$ -way one-shot episodes (or *tasks*)  $\mathcal{T}$  from a distribution over tasks  $p(\mathcal{T})$  that we are interested in. For each episode, first adapt the model on support set  $\mathcal{S}$ , then test model classification accuracy on the set of queries  $\mathcal{Q} = \{(\hat{\mathbf{x}})\}_{i=1}^N$ . Finally, report the expectation

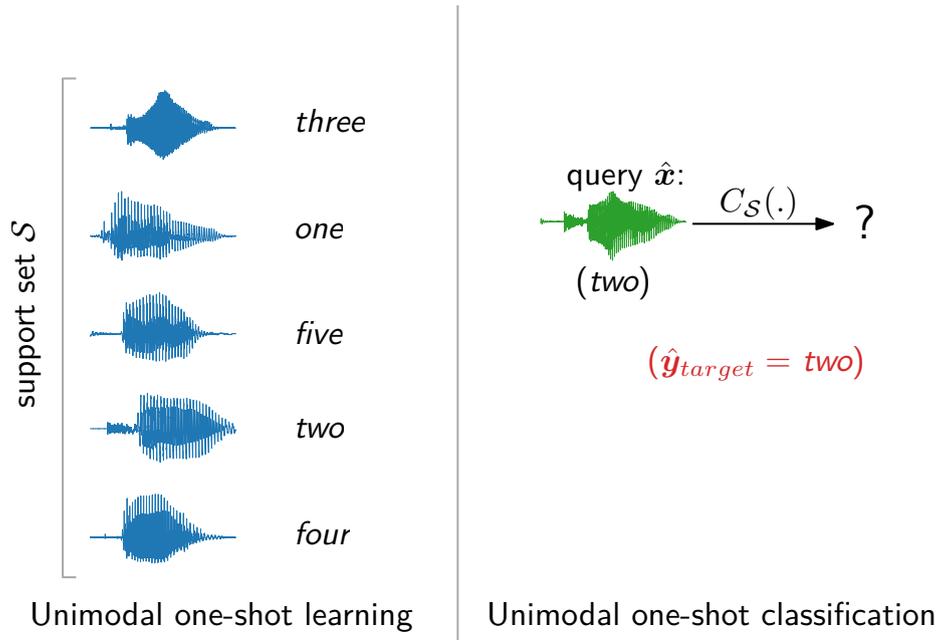


Figure 2.1: Unimodal one-shot speech learning and classification.

across episode scores. It is important to note that a one-shot learner may not accumulate experience across episodes during evaluation—before continuing to the next episode, the model must be reset to its state prior to learning from the current support set.

## 2.2 DEFINING MULTIMODAL ONE-SHOT LEARNING

We now extend unimodal one-shot learning to fit Multimodal One-Shot learning (MOONSHOT). Instead of a labelled unimodal support set, we are now given features in multiple modalities with the only supervisory signal being that these features co-occur. In our case we consider speech and images as the two modalities, although this may be applied to any source of paired multi-sensory information. Consider, for example, a user teaching a robot novel speech-image correspondences by presenting a single paired example per class. The user might provide one example for each of the concepts *milk*, *eggs* and *butter*. The robot must successfully learn these concepts such that it generalises to new instances of the spoken words and visual objects. A new user might utter the query “milk” to which the robot should be able to visually identify *milk* even if this object is different to the instance seen during learning. Formally, in an  $L$ -way one-shot episode we are given a multimodal support set  $\mathcal{S} = \{(\mathbf{x}_a^{(i)}, \mathbf{x}_v^{(i)})\}_{i=1}^L$ , where each spoken caption  $\mathbf{x}_a^{(i)} \in \mathcal{A}$  (audio space) is paired with an image  $\mathbf{x}_v^{(i)} \in \mathcal{V}$  (vision space). The goal is to leverage the paired association between these limited examples such that a model quickly adapts to novel instances of these concepts.

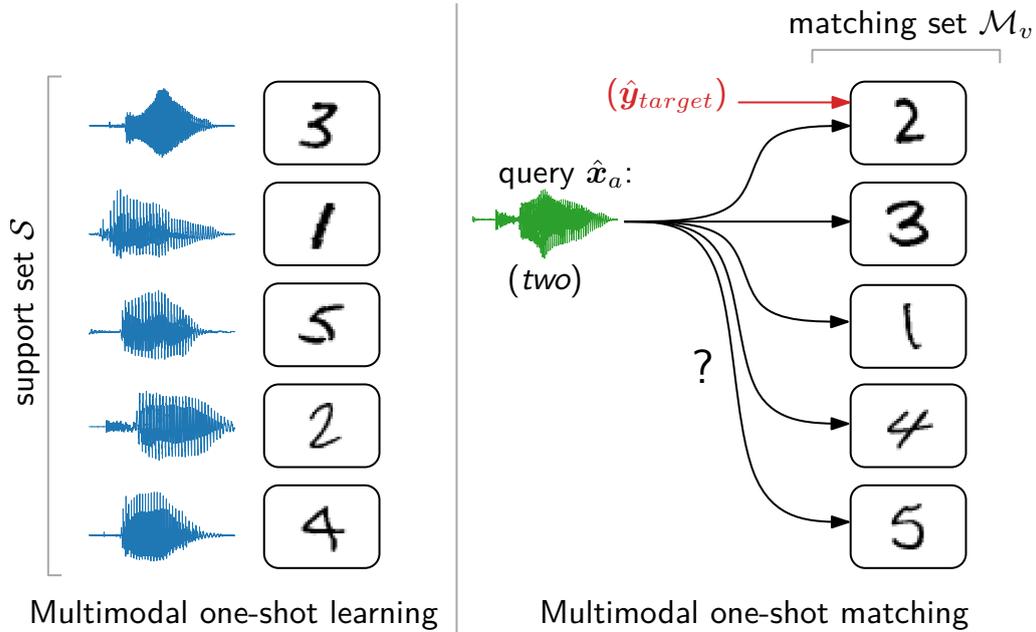


Figure 2.2: Multimodal one-shot learning and matching of speech and images.

### 2.3 CROSS-MODAL MATCHING OF SPEECH TO IMAGES

During test-time, following MOONSHOT, a model is presented with a test query in one modality, and asked to determine the matching item in a test (or matching) set in the other modality. This is related to cross-modal retrieval tasks (Kashyap, 2017; Wang *et al.*, 2018; Eisenschtat and Wolf, 2017) used to evaluate multimodal networks. Formally, as defined by Kashyap (2017) to evaluate their multimodal network, we match query  $\hat{x}_a$  in one modality (speech) to a *matching set*  $\mathcal{M}_v = \{(\hat{x}_v)\}_{i=1}^N$  in the other modality (images) according to some metric  $D_S(\hat{x}_a, \hat{x}_v)$  that is learned from the support set  $\mathcal{S}$  during the MOONSHOT phase. Neither the query  $\hat{x}_a$  nor the items in the matching set  $\mathcal{M}_v$  occur exactly in the support set  $\mathcal{S}$ . We refer to this task as one-shot cross-modal matching. In this thesis we consider one-shot matching where  $N = L$ , such that the matching set contains one instance for each of the  $L$  classes in the support set. This task is illustrated in Figure 2.2, where the support set  $\mathcal{S}$  contains spoken utterances paired with images and no labels. Here a MOONSHOT model must use this paired information to match the spoken query “two” most similar to the image of a *two* in the matching set. Evaluating a MOONSHOT model on this task follows the same procedure as described for unimodal one-shot classification in §2.1, reporting cross-modal matching accuracy.

Table 2.1: Multimodal one-shot learning terminology.

Symbol	Terminology	Definition
$\mathcal{T}$	episode (or task)	one- or few-shot learning episode of novel concepts a model must adapt to
$p(\mathcal{T})$	task distribution	distribution over one-shot learning tasks from which episodes are sampled
$\mathcal{S}$	support set	training data for episode $\mathcal{T}$ containing $L$ classes, $K$ examples per class (unimodal or multimodal)
$\mathcal{Q}$	query set	testing data for episode $\mathcal{T}$ containing unseen instances of the $L$ task classes (speech or images)
$\mathcal{M}$	matching set	testing data for episode $\mathcal{T}$ containing one instance for each of the $L$ task classes, in different modality from the query set, and for cross-modal matching
$C_{\mathcal{S}}$	classifier	one-shot classification model for episode $\mathcal{T}$ adapted on unimodal support set $\mathcal{S}$
$D_{\mathcal{S}}$	metric	one-shot cross-modal metric model for episode $\mathcal{T}$ adapted on multimodal support set $\mathcal{S}$

## 2.4 FEW-SHOT LEARNING OF SPEECH AND IMAGES

One-shot learning can be generalised to  $K$ -shot learning, where, in the unimodal case, a model is shown a support set containing  $L$  novel classes and  $K$  examples per class. This is also referred to as few-shot learning. In multimodal  $L$ -way  $K$ -shot learning, the support set  $\mathcal{S} = \{(\mathbf{x}_a^{(i)}, \mathbf{x}_v^{(i)})\}_{i=1}^{L \times K}$  consists of  $K$  speech-image example pairs for each of the  $L$  classes. Evaluating a multimodal  $K$ -shot learner follows as before, and the matching set contains one instance for each of the  $L$  classes. This would occur, for instance, when a user teaches a robot speech-image correspondences by presenting it with multiple paired examples per class. In a 3-way 5-shot task, a robot matching spoken queries to visual instances could have a speech-image support set with five examples for each of the concepts *milk*, *eggs* and *butter*. A summary of the one-shot terminology and notation used within this thesis is shown in Table 2.1.

## 3 | RELATED WORK

In this chapter we review work related to the Multimodal ONE-SHOT learning (MOONSHOT) problem. We first review one-shot learning observed in humans during language acquisition and cognitive modelling of these learning abilities (§3.1). We then review unimodal one-shot learning literature particularly in the predominant case of computer vision (§3.2). Thereafter, we investigate one-shot learning in other domains (§3.3) and more recent meta-learning approaches to one-shot learning (§3.4). Finally, we briefly review zero-shot learning (§3.5) and multimodal representation learning (§3.6) which are also related to this work.

### 3.1 ONE-SHOT LEARNING OBSERVED IN HUMANS AND COGNITIVE MODELLING

Many studies in the cognitive science community have investigated human language acquisition. We have previously discussed research on fast mapping (§1.1), whereby children and adults may learn novel words from a single exposure (Carey, 1978; Carey and Bartlett, 1978; Markson and Bloom, 1997; Halberda, 2006). For example, Markson and Bloom (1997) demonstrate in a series of experiments on adults and three- and four-year old children that when learning a novel name or fact paired with an unfamiliar object, they were capable of successfully recalling the pair immediately after exposure, as well as after a 1-week and 1-month delay. Notably, these results also suggest that fast mapping is not limited to language and may be the result of human learning and memory abilities. Later work by Trueswell *et al.* (2013) investigates how adults may learn novel words in ambiguous settings. Here the authors propose humans learn in a one-trial procedure which they dub *propose-but-verify*: In an experiment, adult participants are shown a so-called nonce word, a novel word used for this single occasion, in a sequence of learning trials which contain ambiguous visual context and must correctly guess the referent—the item to which the word refers. This is illustrated in Figure 3.1<sup>1</sup> where the word “zud” is presented in two learning trials containing five alternative visual items, with the intended meaning *bear*. Participants who chose the correct referent (the bear) on

---

<sup>1</sup>Reprinted from *Cognitive Psychology*, vol. 66, no. 1, Trueswell, Medina, Hafri and Gleitman (2013), Propose but verify: Fast mapping meets cross-situational word learning, pp. 126–156, Copyright 2013, with permission from Elsevier.



Figure 3.1: Example propose-but-verify experiment (Trueswell *et al.*, 2013). The word “zud” is presented in two learning trials containing five alternative visual items, with the intended meaning bear. Figure reproduced from Trueswell *et al.* (2013).

the first trial were likely to make the same correct choice on the second trial where it appears again. However, if a participant was incorrect on the first trial and selected, for example, the *door*, then they were unlikely to store the alternative word meanings which would allow them to correlate the bear among the two learning trials. Instead humans continue to perform with chance accuracy until they make a correct choice which they may confirm in a subsequent trial. Yet, these results are paradoxical with the observations of fast mapping since it may take many incorrect attempts before learning occurs in this setting of high uncertainty. This is attributed to the type of contextual evidence provided in this experiment which simulates the earliest stages of word learning. Infants must initially undergo a slow learning task whereby they build up a database of words and linguistic knowledge. It is this prior experience that allows children and adults to lock onto the correct referent for novel words encountered in the natural world after one or a few encounters.

Computational work on human language acquisition reinforces the findings presented above. Bayesian word learner models have been shown to acquire novel words from a single exposure by building on the learner’s previous experience (Frank *et al.*, 2009). In a similar vein, Fazly *et al.* (2010) develop a probabilistic model which learns to align words with their referent semantic elements in a paired “scene” of meaning words. This model successfully selects the correct referent when queried with a novel word and manages to retain knowledge after a single trial, consistent with the human fast mapping observations. Work by Lake *et al.* (2019), similar to Paperno *et al.* (2016), stresses that our best machine learning algorithms do not generalise in the way humans do: experiments show that humans successfully learn compositional instructions in a few-shot setting, whereas sequence-to-sequence recurrent neural networks fail to generalise to test data. Most related to the work we present here, Lazaridou *et al.* (2014) investigates cross-modal learning of visual-semantic

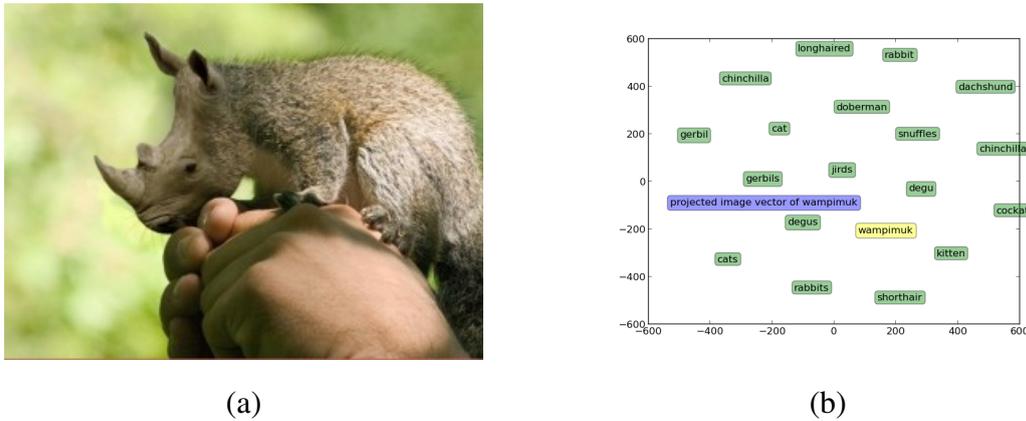


Figure 3.2: Example of the zero-shot fast mapping task (Lazaridou *et al.*, 2014). A potential wampimuk with linguistic context “We found a cute, hairy wampimuk sleeping behind the tree” (a) and the linguistic context together with a projection of the wampimuk image in linguistic space (b).

embeddings for a fast mapping variant of zero-shot learning: a learner is presented an unseen object and must find the correct referent word in limited linguistic context. This is actually the opposite of fast mapping in cognitive psychology where a learner is exposed to a new word and must find the correct referent object. As an example of this task, Figure 3.2 (a) shows a *wampimuk* with the linguistic context “We found a cute, hairy *wampimuk* sleeping behind the tree”. From this statement it is likely that a wampimuk is a small animal. Projecting the visual object onto linguistic space and finding its nearest neighbour we get the zero-shot label “degus”, as shown in Figure 3.2 (b). In the fast mapping zero-shot task, the learner looks at only the limited linguistic context within linguistic space and should find that the visual objects projection vector is closest to the word “wampimuk”; this is shown in Figure 3.2 (b) with the green vectors removed. Lazaridou *et al.* motivate their work with concerns on current successful language models learning entirely from word co-occurrence, arguing that language should be grounded for cognitively plausible models. As a means of communication words are intended to refer to things in the real world (Abbott, 2010). Similar in spirit, the cross-modal matching of speech to images in the MOONSHOT setting may be viewed as investigating an audio-visual fast mapping task. Specifically a learner must learn to match unseen instances of novel spoken words to their correct referent objects following a single exposure. We also note the importance of a learner’s prior experience in the language acquisition literature. This leads us to develop models within the framework of *transfer learning* (Caruana, 1997), where models are trained on some form of source task and encode common knowledge among this task such that it may be applied to some unseen task. We discuss transfer learning in more detail in §4.3. Here we build on prior experience with learning spoken words and visual objects to efficiently learn new concepts in complex natural settings.

## 3.2 UNIMODAL VISION ONE-SHOT LEARNING

Most one-shot learning studies have been primarily interested in image classification. Seminal work focussed on Bayesian modelling approaches (Fei-Fei *et al.*, 2006; Lake *et al.*, 2011, 2013, 2014) which make use of problem specific feature engineering or inference procedures. For example, Lake *et al.* (2015) introduce Bayesian program learning (BPL) which uses specialised stroke information of handwritten characters to perform tasks such as one-shot character classification and generation of new samples for a novel character after a single exposure. Koch *et al.* (2015) demonstrate that *Siamese neural networks* (Bromley *et al.*, 1994; Chopra *et al.*, 2005) trained to learn a suitable metric achieve impressive results on the one-shot classification of handwritten characters. A Siamese model consists of two or three network branches with the same architecture and shared parameters among the branches. These tied networks are trained in the framework of *metric learning* to explicitly learn how to effectively compare inputs. We provide a detailed discussion of Siamese neural networks and metric learning in §4.4. Although BPL was shown to perform better on the same task—3.3% error for BPL versus 8.0% error for Siamese neural networks—it has the advantage of specialised stroke information as a prior for the character drawing process. On the other hand, Siamese neural networks do not make any assumptions on the underlying data distribution and learn to compare novel handwritten character images completely from scratch. This advantage is due to the generalisable composite features learned by neural network layers (Bengio and LeCun, 2007; Donahue *et al.*, 2014), combined with the ability to rapidly acquire new examples in non-parametric models (in this case nearest neighbours). See, for example (Salakhutdinov and Hinton, 2007). Furthermore, simply learning features relevant for nearest neighbour comparison alleviates the problem of catastrophic forgetting inherent to neural networks (French, 1999; Kirkpatrick *et al.*, 2017), where previously learned information is lost as new information is acquired.

Apart from Siamese models (Koch *et al.*, 2015), which is the focus of our indirect matching framework (§4.1), other metric learning based approaches have been proposed (Vinyals *et al.*, 2016; Shyam *et al.*, 2017; Snell *et al.*, 2017) which build on advances in attention and memory mechanisms for neural networks (Bahdanau *et al.*, 2015; Weston *et al.*, 2015). Along with the more recent *meta-learning* (which we discuss later in §3.4) approaches (Santoro *et al.*, 2016; Finn *et al.*, 2017a; Mishra *et al.*, 2018), these have each produced improvements on one-shot image classification tasks. However, only small improvements have been made over Siamese networks: Finn *et al.* (2017a) developed a model-agnostic meta-learning algorithm (MAML) which achieved state-of-the-art results at the time with only 1.4% increase in accuracy over Siamese networks for a 5-way 1-shot learning task. However, MAML is shown to be superior on more difficult tasks. For this reason, we focus first on extending Siamese neural networks to the MOONSHOT problem (see Chapter 2) and investigate recent advances for these networks. In particular, we consider *Siamese triplet networks* (Chechik *et al.*, 2010; Wang *et al.*, 2014; Hoffer and Ailon, 2015;

Schroff *et al.*, 2015; Hermans *et al.*, 2017) (§4.4). Later we investigate the MAML algorithm for the MOONSHOT problem (§6.3).

### 3.3 ONE-SHOT LEARNING IN OTHER DOMAINS

A number of one-shot studies have considered other domains, such as robotics (Walter *et al.*, 2012; Finn *et al.*, 2017b), video (Stafylakis and Tzimiropoulos, 2018), gesture recognition (Wu *et al.*, 2012; Thomason and Knepper, 2016), language modelling (Vinyals *et al.*, 2016) and many more. Lake *et al.* (2014) investigated one-shot speech learning using a generative hierarchical hidden Markov model to recognise novel words from learned primitives. This Bayesian model is based on prior work (Lake *et al.*, 2013) in vision and has displayed strong results, however it relies on prior knowledge of the input data structure (e.g. character strokes and speech phonemes) and may not generalise to other tasks. We consider Siamese neural networks for one-shot unimodal modelling of spoken language to automatically extract speech features that enable fast adaptation from only few examples of novel words.

There has been work on extending one-shot learning to *weakly supervised* (Yu *et al.*, 2018) and *semi-supervised* (Ren *et al.*, 2018) settings. In weakly supervised modelling we attempt to build models which may learn from weaker forms of supervision without direct labels. For example, in MOONSHOT we consider the case of weakly supervised speech and images where there are no labels but the co-occurrence between these modalities serves as a weak supervisory signal. Semi-supervised learning is related, where we have a portion of training data that is labelled as well as a large dataset of unlabelled data. Usually the labelled data is used to learn some mechanism of labelling the unlabelled data so that it may be used for training a model. Both of these studies have focussed on unimodal vision tasks and make use of some form of fully supervised objective to guide the weakly and semi-supervised tasks. The MOONSHOT setting is different in that we consider multiple modalities without any supervision and only co-occurring context.

### 3.4 ONE-SHOT LEARNING WITH META-LEARNING

*Meta-learning*—also referred to as learning to learn—has had a surge of interest in recent years and demonstrated some of the most intriguing results on one-shot learning tasks (Vinyals *et al.*, 2016; Santoro *et al.*, 2016; Finn *et al.*, 2017a; Mishra *et al.*, 2018). Conventional learning algorithms optimise an objective for a single task such as automatic speech recognition. With meta-learning we take a different approach and optimise an objective that learns how to learn new tasks quickly and from few examples. This is different to multi-task learning where a single model must learn a number of related tasks using the shared signal as an inductive bias (e.g. a robot learning to grip a lego block and open a door) (Caruana, 1997). Crucially, multi-task learning makes use of a shared representation among training tasks so that a model

can leverage information learned for each task to support learning of the other tasks. Meta-learning focusses instead on the distillation of training tasks into a prior that enables a model to efficiently learn *new* tasks from small amounts of data (Finn, 2018). One successful approach to meta-learning for one-shot learning is the multi-modal model-agnostic meta-learning algorithm (MAML) (Finn *et al.*, 2017a) which was mentioned earlier. MAML receives its apt name due to the fact that it may be applied to any model architecture and learning problem which is trained with gradient descent. The idea behind MAML is to train a model across a diverse set of training tasks that optimises for initial parameters which generalise well when fine-tuned for one or a few gradient steps on a new task. This simple algorithm has achieved impressive results on one-shot tasks such as sinusoid regression, image classification and reinforcement learning. Other domains have also begun to adopt MAML, for example low-resource neural machine translation (Gu *et al.*, 2018) and robot imitation learning (Finn *et al.*, 2017b). We investigate a novel extension of the MAML algorithm in conjunction with a joint audio-visual model that learns to learn from weakly supervised multimodal speech paired with images (§6.3).

### 3.5 ZERO-SHOT LEARNING

*Zero-shot learning* is a similar topic that investigates agents which may generalise to previously unseen classes without any additional training data (Palatucci *et al.*, 2009; Socher *et al.*, 2013). While closely linked, this is different to one-shot learning since a model does not see any examples from the test classes to learn from. For example, Frome *et al.* (2013) develop a deep visual-semantic embedding model (DeViSE) which learns joint embeddings from images and unannotated text. Using DeVISE, the authors show that it is possible to make zero-shot inferences about thousands of image labels not seen during training. This is similar to our work on joint audio-visual modelling for the MOONSHOT problem and we show that these models are capable of zero-shot inferences (§6.1).

### 3.6 MULTIMODAL REPRESENTATION LEARNING

Our work is also related to learning multimodal representations from paired images and speech (Ngiam *et al.*, 2011; Harwath *et al.*, 2016; Leidal *et al.*, 2017; Kamper *et al.*, 2019; Kashyap, 2017). We are particularly inspired by the use of unlabelled speech, where weaker supervision is obtained in the form of co-occurring images. These studies have shown that this task is possible and that suitable neural network models may acquire language in this setting. We extend this research to the one-shot domain. The cross-modal matching of speech to images (see Chapter 2) is similar to cross-modal retrieval (Rasiwasia *et al.*, 2010; Harwath *et al.*, 2016), where documents (e.g. images) are retrieved from a database in response to a query (e.g. speech). Also related is the use of images as a “pivot” between two languages (Gella

*et al.*, 2017), learning multimodal multilingual representations that can be used for text-to-image retrieval and vice versa. By modelling text and images in a joint space, images most similar to a text query in one language may be retrieved. These images may then be used as an auxiliary query for retrieving text in a different language. This is similar to our indirect matching framework (§6.1).

PART II

ONE-SHOT LEARNING HANDWRITTEN  
DIGITS PAIRED WITH SPOKEN DIGITS

## 4 | A FRAMEWORK FOR MULTIMODAL ONE-SHOT LEARNING

In this chapter, we present Multimodal ONE-SHOT learning (MOONSHOT) models to perform cross-modal matching of speech to images within an indirect matching framework, where matching is reduced to a sequence of two unimodal comparisons. This approach is motivated by extending existing techniques that have been successful in unimodal one-shot learning to the MOONSHOT setting. The models that we present in this chapter are somewhat agnostic to the specific dataset under consideration; we describe the application of these models to different datasets in Chapters 5 and 7. We first describe the general framework with an illustrative example (§4.1) followed by the models we investigate within this framework. Specifically, we describe a direct matching baseline on raw features (§4.2), a neural network transfer learning approach on classification embeddings (§4.3) and a Siamese neural network approach to explicitly learn suitable unimodal metrics (§4.4).

### 4.1 INDIRECT MATCHING FRAMEWORK

We start by describing a framework for indirect matching of speech to images. Assume we have a method or model that can measure similarity within a modality (in the subsequent sections we describe different methods for measuring similarity within a modality). One-shot cross-modal matching is then accomplished by first comparing a query to all of the items in the support set which are in the same modality as the query (e.g. speech). The most similar (speech-image) support-set pair is retrieved and we take the instance of this pair in the matching set modality (e.g. images) as an auxiliary query. Finally, the retrieved auxiliary query instance is used to determine the closest item in the matching set. In the case of matching speech to images, this approach thus defines a metric  $D_S$  as a mapping from audio space  $\mathcal{A}$  to vision space  $\mathcal{V}$ :  $\mathcal{A} \rightarrow \mathcal{V}$ ; thus allowing us to match speech to images by unimodal comparisons through the multimodal support set  $\mathcal{S}$ .

As a concrete example we revisit the problem described in §2.3 which we expand on here in Figure 4.1. A MOONSHOT model is shown a support set  $\mathcal{S}$  containing spoken digit utterances paired with handwritten digit images and no labels. The model must use this paired information to match the spoken query “two” most similar

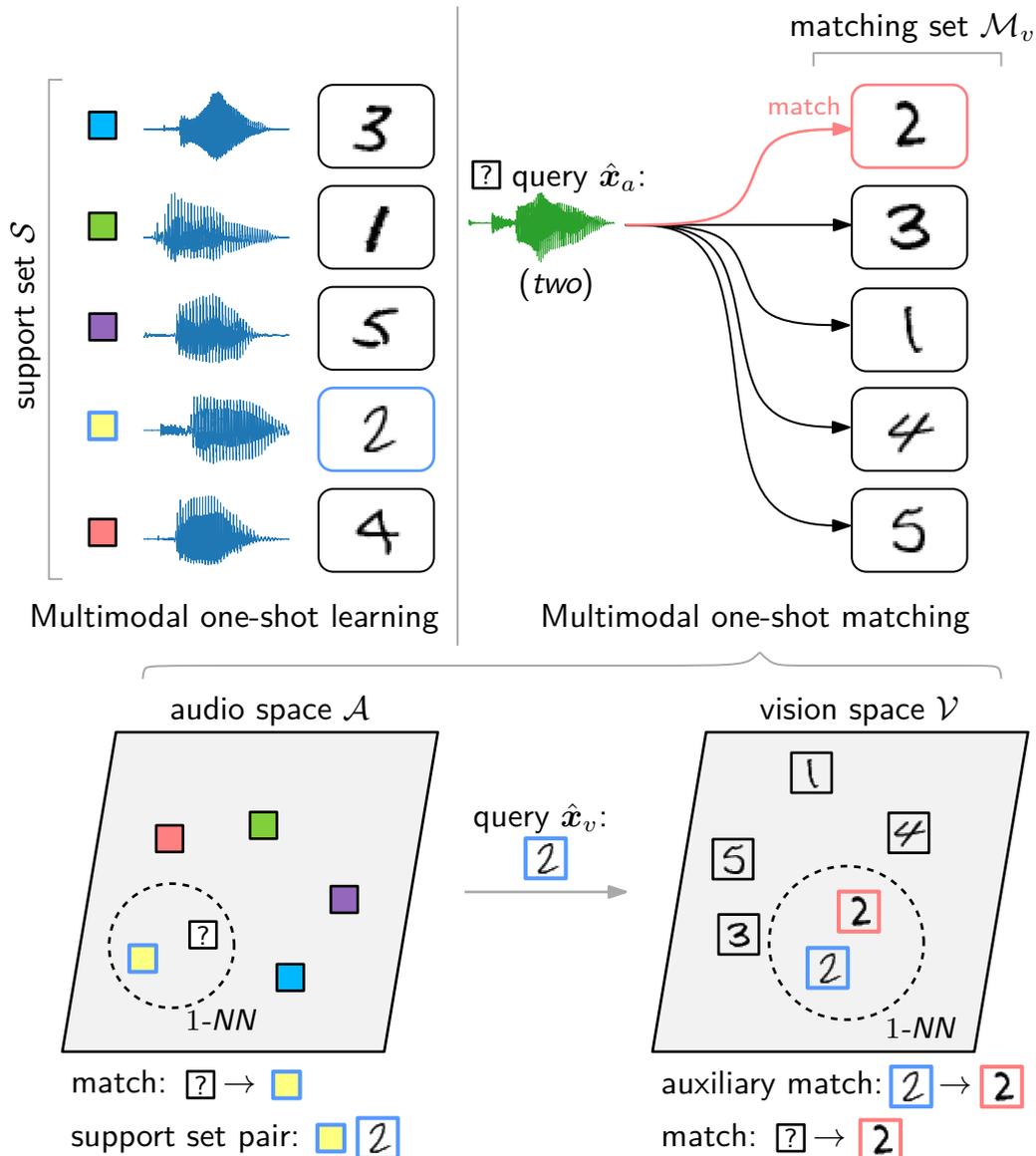


Figure 4.1: Multimodal one-shot learning framework for indirect matching of speech to images. Support set speech items are represented with coloured blocks for simplicity. The speech query “two” is matched to the matching set image of a *two* using unimodal comparisons and retrieving an auxiliary image query from the multimodal support set which may be compared to the matching set images.

to the image of a *two* in the matching set  $\mathcal{M}_v$ . Figure 4.1 demonstrates the indirect matching framework applied to this task, where the support set speech items are represented with coloured blocks for simplicity. The speech query  $\hat{x}_a$ , represented in the figure as  $?$ , is compared to all the support set speech segments in  $\mathcal{A}$  according to the model’s speech-speech metric and 1-nearest neighbour matching (1-NN). We find that the speech query is matched to the speech support set item 2 paired with

an image of a *two*. Taking this paired image as an auxiliary query  $\hat{x}_v$ , we compare to all the matching set images in  $\mathcal{V}$  according to the model’s image-image metric and 1-*NN*, as shown in the bottom part of Figure 4.1. Finally, we find that the auxiliary image query is matched to the matching set image of a *two* and we predict this image as the cross-modal match for the speech query [?].

Several different methods or models can be used to determine within-modality similarity: we compare directly using the raw image pixels and extracted speech features (§4.2), to feature embeddings learned by neural network classifiers (§4.3) and Siamese neural networks (§4.4).

## 4.2 DIRECT MATCHING OF RAW FEATURES

As a first naive approach for measuring similarity within a modality, we consider directly using image pixels and acoustic speech features. Here we first compare the acoustic features of a speech query to the speech segments in the support set, retrieve the image paired with the matching support set speech item, and compare the pixels of this image to the images in the matching set, returning the most similar image. We specifically use cosine similarity between image pixels and dynamic time warping (DTW) (Sakoe and Chiba, 1978) to measure similarity between speech segments. The DTW algorithm calculates an optimal non-linear alignment between two variable length speech sequences which produces a minimum distance when compared. This is demonstrated in Figure 4.2 where DTW is applied over the acoustic features for two speech segments with different lengths, both containing instances of the spoken word “six” (we discuss the acoustic feature later in §5.3.1 and §7.3.1). The non-linear alignment, or optimal warping path, is shown as a dashed line and produces a minimum distance between these two segments when accumulated. We use this approach as our nearest neighbour matching baseline, similar to unimodal one-shot learning studies (Lake *et al.*, 2014; Koch *et al.*, 2015; Vinyals *et al.*, 2016; Finn *et al.*, 2017a).

## 4.3 TRANSFER LEARNING WITH NEURAL NETWORK CLASSIFIERS

Another method, also used in unimodal one-shot learning, is to train a supervised model on a large background dataset. This background dataset should not contain instances of the target one-shot classes. The idea is that features learned by such a model would still be useful for determining similarity on classes which it has not seen (Vinyals *et al.*, 2016). In other words, it follows the *transfer learning* principle: first train a classifier on a large labelled dataset, or source task, and then apply the learned representations to new target tasks which are related but have too few training instances (Donahue *et al.*, 2014). The high-level intuition of transfer learning is demonstrated in Figure 4.3, where we train a model with parameters  $\theta$  on a

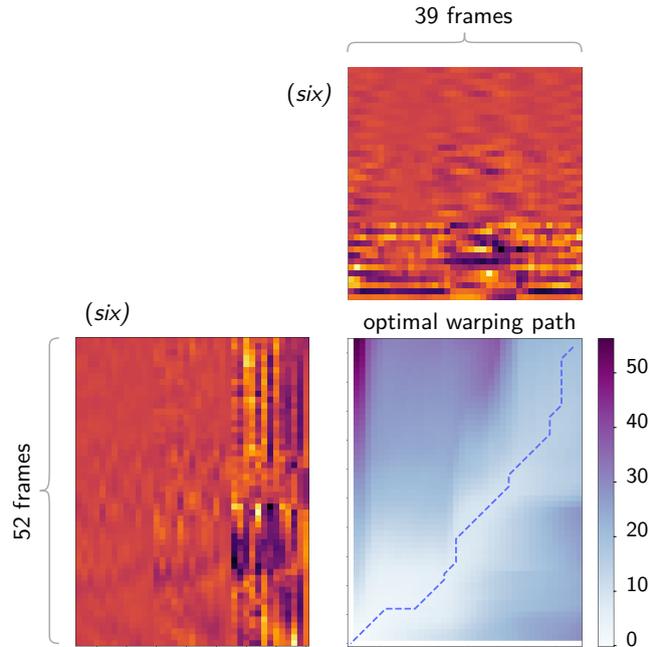


Figure 4.2: Optimal non-linear alignment computed with dynamic time warping over the acoustic features for two speech segments, with different lengths and both containing instances of the spoken word “six”.

source task (i.e. background dataset) of spoken words with many labelled examples for speech classification. Specifically, the figure depicts a space of possible configurations for model parameters  $\theta$  and a hypothetical trajectory (represented as a solid line) through the parameter space during training with gradient descent. Classifying each of the training word classes among the other training word classes may be viewed as a separate source task and the goal is to find common knowledge in the form of  $\theta$  (i.e. learn a prior) which is optimal across these source tasks—also referred to as multi-task learning (Caruana, 1997). This is shown intuitively in the figure, where a set of source tasks, depicted in the figure as purple circles, is classifying spoken words for source classes *bicycle*, *motorcycle*, *dog* and *squirrel*. Note that the coloured circles do not depict the embedding space (i.e. learned representations) determined by the configuration of model parameters  $\theta$ . Instead, the location of a coloured circle in parameter space indicates the optimal configuration of  $\theta$  for the best performance on the task contained within the circle—i.e.  $\theta$  should be in close proximity to the task circle. We learn an optimal configuration of  $\theta$ , depicted in the figure at the end of the trajectory (after the arrowhead), such that the validation error is low among the source tasks—i.e. the optimal  $\theta$  is in close proximity to each of the purple circles. We may then transfer learn a set of target tasks by simply applying the learned configuration of  $\theta$ . For example, classifying novel spoken words *horse* and *bird*—target classes depicted in the figure as blue circles which have limited labelled examples. The source tasks are related enough to the target tasks that the

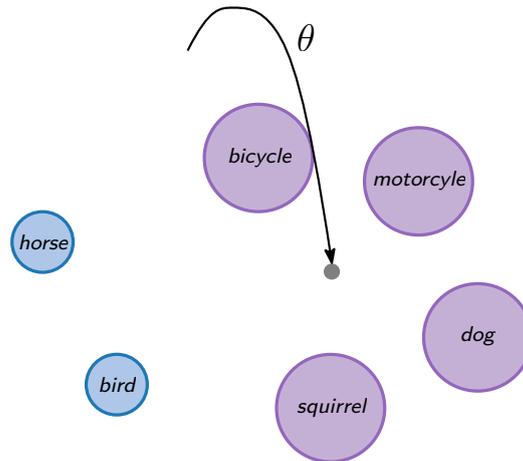


Figure 4.3: Transfer learning illustrated for spoken word classification. The goal is to learn model parameters  $\theta$  on a source task (purple circles) which has many labelled examples. The shared knowledge among the source task is used to learn a prior, an optimal configuration of  $\theta$  shown at the end of the trajectory, that is useful for related target tasks (blue circles) which have limited data.

optimal configuration of  $\theta$  is in close proximity to that of the target tasks—i.e. the learned representations are useful for determining similarity on examples from the target tasks. This approach has been shown to counteract overfitting on such tasks which have limited data.

Here we train such neural network classifiers separately for both the spoken and visual modalities on suitable background data containing *none of the classes* (spoken words and visual objects) tested on during one-shot learning (as described later in §5.2 and §7.1). We specifically build both feedforward neural network and convolutional neural network classifiers and compare their effectiveness transfer learning to the one-shot tasks. We extract representations from the final hidden layer of these networks, before the unnormalised log-probabilities, or *logits* layer, and prior to the rectified linear unit activation. Applying these representations as the learned feature embeddings, we perform nearest neighbour matching using cosine similarity during one-shot learning.

While the representations learned by this transfer learning approach should contain semantic information relevant for comparison, it is largely dependent on the type of features learned by the neural network classifiers. Specifically, these representations are not directly optimised for matching within an embedding space. This leads us to our next approach, where we explicitly optimise representations for nearest neighbour matching.

## 4.4 METRIC LEARNING WITH SIAMESE NEURAL NETWORKS

Our final approach within the indirect matching framework is another form of transfer learning which explicitly learns a metric for comparing features: a supervised model can be trained on a background dataset (the same as in §4.3), not containing any instances of the target one-shot classes, to directly measure similarity between inputs instead of predicting a class label. This is referred to as *metric learning*. Siamese neural networks have been used for this task (Bromley *et al.*, 1994; Chopra *et al.*, 2005; Hadsell *et al.*, 2006) and have been successful in unimodal one-shot learning (Koch *et al.*, 2015). Siamese networks are a natural choice for one-shot learning as they were developed with the purpose of learning in cases where there are potentially many classes, classes are not known in advance or only a few examples are available per class.

A Siamese neural network consists of two identical neural network branches with shared parameters—a set of twin networks, hence the name “Siamese”. These networks are trained to map input features to a target embedding space where the “semantic” relationship between input pairs may be captured based on proximity: inputs of the same type should ideally be mapped to similar embeddings, while inputs that are unrelated should be far apart. Early approaches (Bromley *et al.*, 1994; Chopra *et al.*, 2005; Hadsell *et al.*, 2006; Koch *et al.*, 2015) took in pairs of training examples and either maximised or minimised a distance based on whether the inputs were of *same* or *different* types. Recent studies (Chechik *et al.*, 2010; Wang *et al.*, 2014; Hermann and Blunsom, 2014; Hoffer and Ailon, 2015; Schroff *et al.*, 2015) have argued that the *relative* rather than the *absolute* distance between embeddings are meaningful. For example, images of motorcycles should be more similar to bicycles than they are to dogs. This is not the case for the two branch networks which are optimised to push both bicycles and dogs as far as possible from motorcycles. The more recent approaches overcome this limitation by considering triplet pairs and evaluating their relative similarity. We also follow the triplet approach in this work.

Concretely, let  $x_a$  and  $x_p$  be inputs of the same class, while  $x_a$  and  $x_n$  are of different classes. This triplet pair is input to the respective branch of a Siamese neural network, as illustrated in Figure 4.4 for learning spoken word representations. The intuition is that we want to push the so-called anchor example  $x_a$  and positive example  $x_p$  together such that the distance between them is smaller (by some specified margin) than the distance between the anchor  $x_a$  and negative example  $x_n$ . Models using this approach are sometimes referred to as *Siamese triplet networks*, since there are three shared parameter network branches for inputs  $(x_a, x_p, x_n)$ . We apply this approach where we learn embedding function  $f(\cdot)$  as the final fully-connected linear layer of a convolutional neural network (CNN) followed by L2 normalisation—this is one of the tied network branches, as shown in Figure 4.4. Concretely, we train models with a hinge loss for triplet pairs (Chechik *et al.*, 2010; Wang *et al.*, 2014;

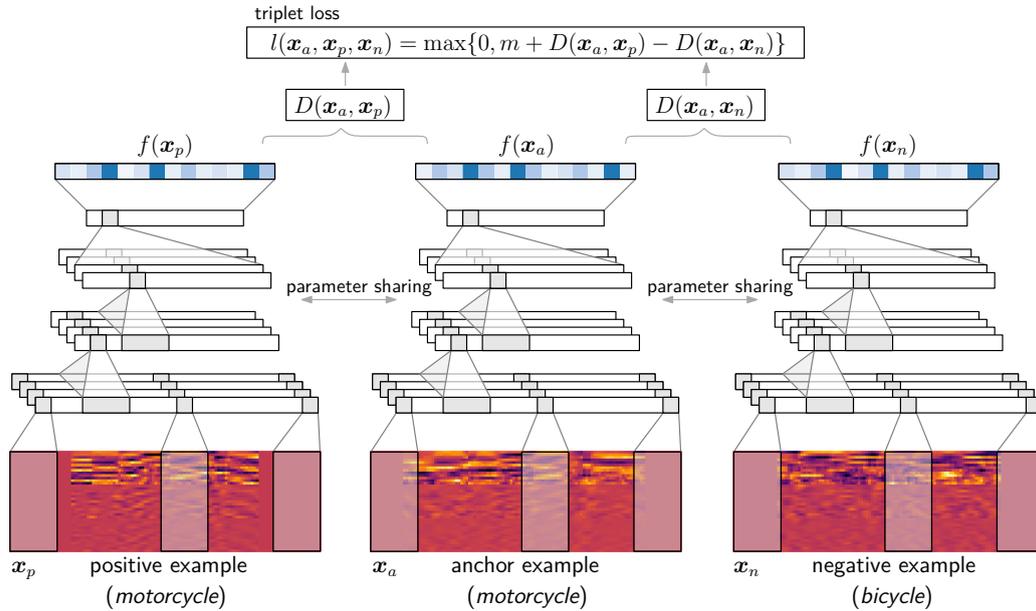


Figure 4.4: Siamese triplet convolutional neural network that takes in triplet pairs of spoken words for learning speech representations useful for comparison. Speech input features are extracted mel-frequency cepstral coefficients centre zero padded to 120 frames (§5.3.1 and §7.3.1). The anchor and positive examples are instances of spoken word “motorcycle” and the negative example is an instance of spoken word “bicycle”. The parameters of the tied network branches are trained with a triplet loss that optimises the distance between the *motorcycle* representations to be smaller than the distance between the *motorcycle* and *bicycle* representations.

(Hermann and Blunsom, 2014; Schroff *et al.*, 2015), defined as:

$$\mathcal{L}(\mathbf{x}_a, \mathbf{x}_p, \mathbf{x}_n) = \max\{0, m + D(\mathbf{x}_a, \mathbf{x}_p) - D(\mathbf{x}_a, \mathbf{x}_n)\} \quad (4.4.1)$$

where  $D(\mathbf{x}_1, \mathbf{x}_2) = \|\mathbf{f}(\mathbf{x}_1) - \mathbf{f}(\mathbf{x}_2)\|_2^2$  is the squared Euclidean distance and  $m$  is the margin between the pairs  $(\mathbf{x}_a, \mathbf{x}_p)$  and  $(\mathbf{x}_a, \mathbf{x}_n)$ .

One problem with this approach is that the number of triplet pairs grows cubically with the dataset, and it may become infeasible to fit all possible triplets in memory. Additionally, naive random sampling of triplet pairs may lead to overfitting on easy triplets, ignoring the few “hard” cases where a negative example is closer to the anchor example than the positive example. This has led to different methods of choosing triplet pairs both for efficiency and to avoid overfitting, such as importance sampling (Wang *et al.*, 2014) or using only the most difficult negative example with the smallest distance  $D(\mathbf{x}_a, \mathbf{x}_n)$  (Settle *et al.*, 2017). We follow the *online semi-hard mining* scheme (Schroff *et al.*, 2015), where all possible anchor-positive pairs in a mini-batch are used. For each positive pair, the most difficult negative example  $\mathbf{x}_n$  with the smallest distance  $D(\mathbf{x}_a, \mathbf{x}_n)$  in the mini-batch satisfying  $D(\mathbf{x}_a, \mathbf{x}_p) < D(\mathbf{x}_a, \mathbf{x}_n)$  is then used, except if there is no such negative example in which case the one with the largest distance is used. These negatives examples

are dubbed *semi-hard* since they lie close to the anchor example but not closer than the positive example. According to [Schroff et al. \(2015\)](#), although it might seem natural to simply choose the hardest negative example in the mini-batch, this constraint is required for stability. For example, there may be cases where false-negatives exist or negative examples are still semantically related to the anchor examples (e.g. motorcycles and bicycles).

The online semi-hard mining approach to triplet sampling also incorporate another recent advance, specifically to improve efficiency: we simplify the three shared-parameter networks with a single neural network that embeds a mini-batch of examples and then samples triplet pairs online from these embeddings. This is done with an efficient implementation of pairwise distances, similar to [Song et al. \(2016\)](#). We build this single network model with online semi-hard mining of triplet pairs, and refer to it as *Siamese CNN (online)*. We also compare to using three shared-parameter networks with the same CNN architecture, where we generate triplets offline at each training step from the current mini-batch. We refer to this model as *Siamese CNN (offline)* in our experiments. Similar to our neural network classifier baselines (§4.3 above), we train separate networks for vision and speech on triplets from large disjoint labelled datasets which do not contain any of the target one-shot classes.

## 4.5 CHAPTER SUMMARY

To summarise, we have introduced a triad of multimodal one-shot learning modelling approaches. These models may be applied within an indirect matching framework for the task of cross-modal matching speech to images. We have described this general framework in detail and shown how it may be used to extend unimodal one-shot learning models to the multimodal setting. Specifically, this framework assumes access to metrics for comparing speech-speech and image-image. We proposed the following models to define these metrics: (i) direct feature matching with image pixels and dynamic time warping over speech segments, (ii) transfer learning feature embeddings with neural network classifiers (separate networks for speech and images) and (iii) explicit metric learning of feature embeddings with Siamese triplet neural networks (separate networks for speech and images). These models have been used in unimodal one-shot learning studies, with the first two delegated as baselines and the third displaying favourable results in comparison. We investigate these models for multimodal one-shot learning, first on a simple benchmark dataset containing spoken digits paired with handwritten digit images (see Chapter 5) and finally on a more complex benchmark dataset containing natural images paired with spoken words (see Chapter 7).

## 5 | TIDIGITS AND MNIST EXPERIMENTS

This chapter investigates Multimodal ONE-SHOT learning (MOONSHOT) on a simple problem using the models and indirect matching framework discussed in Chapter 4. We first create a MOONSHOT benchmark dataset from spoken digit words paired with handwritten digit images (§5.1). We also introduce supervised background datasets (not containing any of the one-shot classes) which we use to train our neural network models for the MOONSHOT task (§5.2). We then discuss the experimental setup that we follow (§5.3), including data preprocessing, the implementation (and training) of our indirect matching models and one-shot task evaluation. Thereafter, we present results for unimodal and multimodal one-shot task experiments on the benchmark dataset and discuss the advantages and limitations of our models (§5.4). The results presented in this chapter have been published previously as [Eloff \*et al.\* \(2019\)](#).<sup>1</sup>

### 5.1 A MULTIMODAL DIGITS BENCHMARK DATASET

For our initial investigation of the MOONSHOT problem we propose a simple benchmark dataset: learning from examples of spoken digits paired with handwritten digit images. For speech we use the TIDIGITS corpus which contains spoken digit sequences from 326 different speakers sampled at 20kHz ([Leonard and Doddington, 1993](#)), and for images we use the MNIST handwritten digits dataset which contains  $28 \times 28$  grayscale images ([LeCun \*et al.\*, 1998](#)). We use the spoken utterances from men, women, and children, and split the digit sequences into isolated digits using forced alignments obtained from the ground truth transcriptions of the speech data.<sup>2</sup>

To form this simple multimodal digits benchmark dataset, we pair each isolated spoken digit with an image of the same type. Unlike previous work which used the same dataset combination for learning multimodal representations ([Kashyap, 2017](#); [Leidal \*et al.\*, 2017](#)), we treat utterances labelled “oh” and “zero” as separate classes, resulting in 11 class labels. We use this combined dataset as our task distribution

---

<sup>1</sup>Full code recipe for the TIDIGITS and MNIST experiments described in this chapter available at: <https://github.com/rpeloff/multimodal-one-shot-learning>.

<sup>2</sup>We make these splits and the forced alignments available in our code recipe.

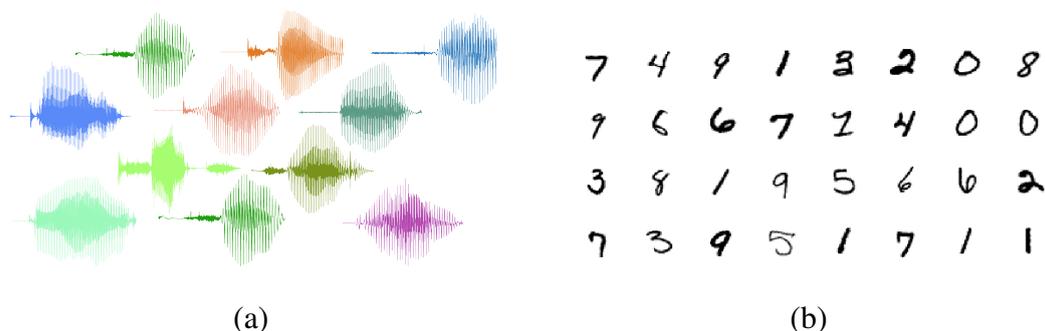


Figure 5.1: TIDIGITS isolated spoken digits (a) and MNIST handwritten digit images (b).

from which we may sample MOONSHOT episodes  $\mathcal{T} \sim p(\mathcal{T})$  (§2.2). Extracts from the TIDIGITS and MNIST one-shot benchmark dataset are shown in Figure 5.1.

## 5.2 BACKGROUND DATA FOR NEURAL NETWORK MODELS

Our neural network models (§4.3 and §4.4) are trained on large labelled background datasets to obtain feature representations which may be applied to the one-shot problem on classes not occurring in the background data. We use the Flickr Audio Caption corpus (Harwath and Glass, 2015) and the Omniglot handwritten character dataset (Lake *et al.*, 2015) as background data for the within-modality speech and vision models, respectively. Utterances in the spoken audio corpus are split into isolated words using forced alignments (provided with the data). We ensure that none of the target digit classes occur in this audio data. Some words in the Flickr corpus overlap with TIDIGITS (e.g. “four”, “seventh”) and we remove these from the background data. None of the Omniglot image classes overlap with digit classes. We use the train and validation data splits from both background datasets. The Flickr Audio splits contain 5 534 isolated spoken word classes, 88 411 speech segments for training and 14 744 speech segments for validation. The Omniglot splits contain 964 handwritten character classes, 19 280 images for training and 13 180 images for validation respectively. Extracts from the Flickr Audio and Omniglot background dataset are shown in Figure 5.2.

## 5.3 EXPERIMENTAL SETUP

In this section we discuss the details of our experiments: preprocessing applied to the speech and image data (§5.3.1), implementation and training procedure of the indirect matching models (§5.3.2) and how we evaluate the unimodal and multimodal one-shot tasks (§5.3.3).

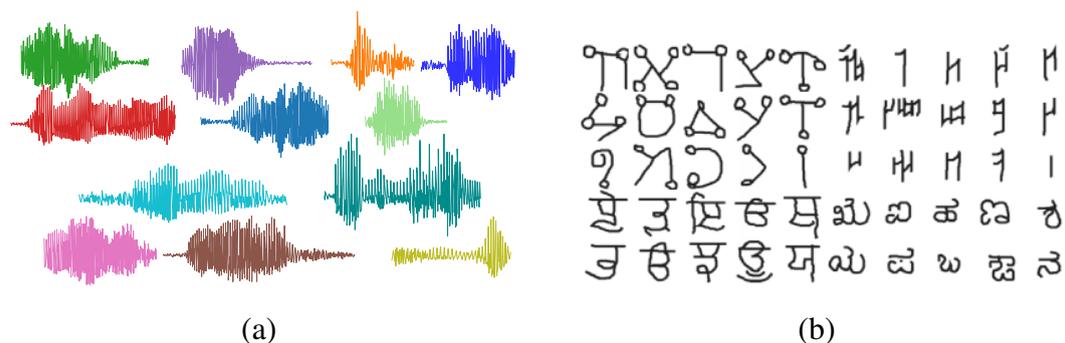


Figure 5.2: Flickr Audio isolated spoken words (a) and Omniglot handwritten character images (b).

### 5.3.1 Data Preprocessing

Speech is parametrised as mel-frequency cepstral coefficients (MFCC) with first and second order derivatives, computed with a 25 millisecond window size and a 10 millisecond frame shift, yielding 39-dimensional feature vectors for both the TIDIGITS corpus and the background Flickr Audio caption corpus. For our neural network models which require fixed length inputs, we centre zero-pad or crop speech segments to 120 frames. Dynamic time warping (DTW) may deal directly with variable length segments given that we normalise out their duration. We simply divide by sum of the lengths of any two segments under DTW comparison. MNIST image pixels are normalised to the range  $[0, 1]$  and no further preprocessing is applied. Images in the background Omniglot dataset are downsampled to  $28 \times 28$  and pixel values are normalised and inverted in order to match MNIST.

### 5.3.2 Model Implementation

Neural network models are implemented in TensorFlow (TF) (Abadi *et al.*, 2015) and we build these networks from the ground up using TF layers—we provide our implementation of these networks in the code recipe as a contribution. The hyperparameters described in this section have been carefully selected such that each of the models under consideration achieve low generalisation error on a held-out validation set. Specifically, we tune all models using unimodal one-shot learning on the validation sets of the background data. Although this results in different model configurations, the idea is that the underlying mechanisms of these models should perform as best as possible during evaluation and provide a fair comparison. We trained the neural network models using the Adam optimiser (Kingma and Ba, 2015) with a learning rate of  $10^{-3}$  which is step decayed by 0.96 at each new epoch. A batch size of 200 is used for the feedforward neural network (FFNN) and convolutional neural network (CNN) classifiers (§4.3), and an epoch consists of a single iteration of the shuffled background dataset.

For the Siamese models (§4.4) we follow an alternate approach where mini-batches are formed using the *batch all* strategy proposed by Hermans *et al.* (2017).

Specifically, we randomly sample  $p$  classes and  $k$  examples per class to produce balanced batches of  $pk$  examples each. This results in  $pk(pk - k)(k - 1)$  valid triplet combinations, maximising the number of triplets within a mini-batch. For our *Siamese CNN (offline)*, trained in the standard way where three networks are explicitly tied (Wang *et al.*, 2014; Hoffer and Ailon, 2015; Schroff *et al.*, 2015), we use  $p = 32$  and  $k = 2$ , giving 3 968 triplets per mini-batch which is the largest batch we could fit on a single NVidia Titan Xp GPU. Our *Siamese CNN (online)* variant additionally makes use of semi-hard online mining of triplet pairs (Schroff *et al.*, 2015), where for each valid anchor-positive pair a semi-hard negative is chosen (see end of §4.4). Using the online mining scheme this results in  $pk(k - 1)$  semi-hard triplet combinations.<sup>3</sup> While this is fewer than in the offline variant, the *Siamese CNN (online)* is capable of large  $pk$  combinations due to the efficient single network implementation and online triplet sampling scheme. We found a large variety of classes to perform best in validation, and choose  $p = 128$  and  $k = 8$  (total of 7 168 triplets per mini-batch). Each epoch consists of 50 or 200 randomly sampled  $pk$  batches for the Siamese CNN (online) and Siamese CNN (offline) models respectively. All models are trained for a maximum of 100 epochs using early stopping based on one-shot validation error on the background data.

Tuning models using one-shot validation error on the background data gave the following architecture for the spoken word Siamese and classification CNNs: an input layer;  $39 \times 9$  convolution with 128 filters; rectified linear unit (ReLU);  $1 \times 3$  max pooling;  $1 \times 10$  convolution with 128 filters; ReLU;  $1 \times 28$  max pooling over remaining units; 2048-unit fully-connected; ReLU. There is no natural intuition for what the translation invariance should be over the frequency axis of the input speech MFCC. We follow the example of Harwath *et al.* (2016) and collapse the entire 39-dimensional frequency axis through the first convolution, an approach which Harwath *et al.* suggest in order to capture complex relationships along this dimension. The consecutive convolutional layers may then be implemented with 1-dimensional convolutions since only the temporal dimension remains, capturing translation invariance along this dimension (e.g. shifted phonemes). Vision Siamese and classification CNNs have the architecture:  $3 \times 3$  convolution with 32 filters; ReLU;  $2 \times 2$  max pooling;  $3 \times 3$  convolution with 64 filters; ReLU;  $2 \times 2$  max pooling;  $3 \times 3$  convolution with 128 filters; ReLU; 2048-unit fully-connected; ReLU; 1024-unit fully-connected. The speech and vision FFNNs have the same structure: 3 fully-connected layers with 512 units each. For the classifier networks (§4.3), the speech networks have an additional 5534-unit softmax output layer (the number of word types in the speech background Flickr Audio dataset), while vision networks have a 964-unit softmax (the number of image classes in the background Omniglot dataset).

<sup>3</sup>There is an error in our prior publication on this work (Eloff *et al.*, 2019) where we describe the *Siamese CNN (online)* as sampling  $pk(pk - k)(k - 1)$  triplet combinations, which is only the case for the offline variant. Experimental results here and in the prior work are however correct.

### 5.3.3 One-Shot Task Evaluation

We first consider one-shot learning in the unimodal case and then the multimodal case. We empirically evaluate our models on the one-shot tasks according to the classification or matching accuracy averaged over 400 test episodes. In the case of MOONSHOT tasks, each episode  $\mathcal{T} \sim p(\mathcal{T})$  randomly samples a multimodal support set  $\mathcal{S}$  of isolated spoken digits paired with digit images from the paired TIDIGITS and MNIST benchmark dataset. In the case of unimodal one-shot learning, episode support sets contain only spoken digits or digit images from the respective benchmark dataset. We consider  $L$ -way one-shot and five-shot (§2.4) tasks for each of the  $L = 11$  classes (the spoken digits “oh” to “nine” and “zero”). For testing in a MOONSHOT task, a matching set is sampled, containing  $N = 10$  digit images (one for each of the image classes 0 to 9) not seen in the support set. Finally, a speech query instance from one of the  $L$  spoken digit classes is sampled, also not seen in the support set. The query then needs to be matched to the correct item in the matching set. The matching set only contains 10 items since there are 10 unique handwritten digit classes.<sup>4</sup> Testing unimodal one-shot learning follows similarly, where, without a matching set, we sample unseen speech or image queries for the classes in the support set. These queries then need to be classified to the correct class. Within a unimodal or multimodal episode, 10 different query instances—and matching sets in the multimodal case—are also sampled while keeping the support set fixed. Results are averaged over 10 models trained with different seeds and we report average accuracies with 95% confidence intervals.

## 5.4 EXPERIMENTAL EVALUATION

We consider both unimodal and multimodal one-shot task experiments on the simple paired speech and images benchmark dataset. First we look at how our proposed models perform on one-shot speech classification (§5.4.1) and one-shot image classification (§5.4.2). Then we move to the MOONSHOT problem, evaluating these models within our indirect framework for cross-modal matching of speech to images (§5.4.3). Finally we investigate a variant of the MOONSHOT problem that tests how invariant these models are to the query speaker (§5.4.4).

### 5.4.1 One-Shot Speech Classification

We first consider unimodal one-shot speech classification, which has (to our knowledge) so far only been considered in [Lake et al. \(2014\)](#). Table 5.1 shows one-shot and five-shot speech classification results. Average training time (in minutes) is also shown; all models trained within a few seconds of the average time. The Siamese models outperform the direct feature matching baseline using DTW, as well as the

<sup>4</sup>We don’t include a second image for digit 0 even though the support set contains “oh” and “zero”.

Table 5.1: 11-way 1-shot and 5-shot speech classification results on isolated spoken digits sampled from the TIDIGITS dataset.

Speech Model	Train time	11-way Accuracy	
		1-shot	5-shot
DTW	–	67.99% $\pm$ 0.29	91.30% $\pm$ 0.20
FFNN Classifier	13.1m	71.39% $\pm$ 0.81	89.49% $\pm$ 0.45
CNN Classifier	60.6m	82.07% $\pm$ 0.92	93.58% $\pm$ 0.98
Siamese CNN (offline)	70.5m	89.40% $\pm$ 0.54	95.12% $\pm$ 0.37
Siamese CNN (online)	15.0m	<b>92.85% <math>\pm</math> 0.38</b>	<b>97.65% <math>\pm</math> 0.22</b>

Table 5.2: 10-way 1-shot and 5-shot image classification results on handwritten visual digits sampled from the MNIST dataset.

Vision Model	Train time	10-way Accuracy	
		1-shot	5-shot
Pixels	–	44.74% $\pm$ 0.47	67.94% $\pm$ 0.32
FFNN Classifier	8.9m	45.28% $\pm$ 0.24	67.29% $\pm$ 0.73
CNN Classifier	9.7m	65.80% $\pm$ 0.57	85.60% $\pm$ 0.53
Siamese CNN (offline)	62.5m	73.48% $\pm$ 0.53	87.66% $\pm$ 0.32
Siamese CNN (online)	14.3m	<b>73.84% <math>\pm</math> 0.56</b>	<b>89.36% <math>\pm</math> 0.32</b>

neural network classifiers. The *Siamese CNN (online)* model achieves best overall performance, outperforming the *Siamese CNN (offline)* variant, while training almost five times faster. The single network with online semi-hard mining is thus more efficient and accurate than the three shared-network approach. None of these Siamese models were considered in [Lake et al. \(2014\)](#).

#### 5.4.2 One-Shot Image Classification

We next consider unimodal one-shot image classification. Table 5.2 shows one-shot and five-shot image classification results. Once again, the Siamese models perform best, outperforming both the neural network classifiers and the direct pixel matching baseline. The *Siamese CNN (online)* model achieves best overall performance with an accuracy of 74%, again outperforming the *Siamese CNN (offline)* variant (while more efficient) and comparing favourably to the best result of 72% reported in [Vinyals et al. \(2016\)](#) which considered the same task. These observations are similar to the case of one-shot speech classification, although the accuracies are overall lower. This influences the results of cross-modal matching in the multimodal case

which makes use of these vision models paired with the speech models seen before.

### 5.4.3 One-Shot Cross-Modal Matching of Speech to Images

We now turn to the MOONSHOT problem. Table 5.3 shows results for one-shot and five-shot cross-modal matching of speech to images on the simple benchmark dataset for MOONSHOT. Here the Siamese models are again stronger overall compared to direct feature matching or transferring features from neural network classifiers. The *Siamese CNN (online)* model achieves our best results, with double the accuracy of direct feature matching (*DTW + Pixels*) using pixel-distance over images and DTW over speech. The *Siamese CNN (offline)* model follows closely, but is again slower to train (§5.4.1 and §5.4.2).

While the Siamese models achieve promising results compared to the baselines here, our best one-shot multimodal accuracy is lower than the accuracy in unimodal one-shot speech classification (see Table 5.1) and unimodal one-shot image classification (see Table 5.2). The multimodal one-shot results here are therefore worse than both the individual unimodal matching results. This is due to compounding errors in our retrieval framework (§4.1): errors in comparisons with the support set affects comparisons in the subsequent matching step. If the model fails to select the correct auxiliary query image from the support set then it is far less likely that the model will match this image to the correct matching set image. This suggests investigating an end-to-end architecture which can directly compare test queries in one modality to the matching set items in the other modality, without doing explicit comparisons to the support set. In addition, it might be useful to fine-tune the parameters of our models on the multimodal support set. The ability to fine-tune models on smaller datasets is a cornerstone of transfer learning. Yet, this is not possible with the models that we have investigated thus far which rely on supervised training procedures (or no training at all in the case of direct feature matching). This motivates investigating an architecture which may be fine-tuned directly on weakly supervised MOONSHOT support sets.

### 5.4.4 Analysis of Speaker Invariance

As a final evaluation of our proposed models and indirect matching framework, we test for speaker invariance in the MOONSHOT setting. In all of the experiments above we chose spoken queries such that the speaker uttering the query does not appear in the support set. This is representative of an extreme case where one user teaches an agent and another user tests the system. An even more extreme case could occur: the matching item in the support set could be the only item not coming from the query speaker. This is problematic since the same word uttered by different speakers might be acoustically more different than different words uttered by the same speaker. We would like the opposite to be true for the representations learned by our models: more similar for the same word from different speakers than different words from the same speaker—i.e. invariant to the specific speaker. We test

Table 5.3: 11-way 1-shot and 5-shot cross-modal matching of isolated spoken digits to handwritten visual digits sampled from the paired TIDIGITS and MNIST benchmark dataset for evaluating MOONSHOT.

Speech & Vision Model	11-way Accuracy	
	1-shot	5-shot
DTW + Pixels	34.92% $\pm$ 0.42	44.46% $\pm$ 0.69
FFNN Classifiers	36.49% $\pm$ 0.41	44.29% $\pm$ 0.56
CNN Classifiers	56.47% $\pm$ 0.76	63.97% $\pm$ 0.91
Siamese CNNs (offline)	67.41% $\pm$ 0.56	70.92% $\pm$ 0.36
Siamese CNNs (online)	<b>70.12% <math>\pm</math> 0.68</b>	<b>73.53% <math>\pm</math> 0.52</b>

Table 5.4: Speaker invariance tests for 11-way 1-shot cross-modal speech-image digit matching. All support set items are from the same speaker as the speech query, except for the support set item actually matching the query.

Speech & Vision Model	11-way Accuracy
	1-shot
DTW + Pixels	28.00% $\pm$ 1.86
FFNN Classifiers	34.95% $\pm$ 2.28
CNN Classifiers	53.71% $\pm$ 2.20
Siamese CNNs (offline)	66.70% $\pm$ 0.92
Siamese CNNs (online)	<b>69.73% <math>\pm</math> 1.04</b>

this worst-case setting in the following experiment: we sample a support set where all spoken digits are from the same speaker as the speech query, except for the one instance matching the query word which is produced by a different speaker. The spoken digits from the same speaker as the speech query distract from the true match and effective models should be invariant to these speakers. Cross-modal matching results for this case are shown in Table 5.4. All of the models experience a drop in accuracy compared to the results in Table 5.3 (first column). This decrease is smallest for the Siamese models, with the *DTW + Pixels* approach dropping most. Similar results are seen for this speaker invariance experiment in the unimodal case; see results in Table A.1 (Appendix A.1). This indicates that the neural models learn features from the background data which are more independent of speaker and can generalise to other speakers, whereas DTW over speech is affected more by speaker mismatch. Additionally, the Siamese neural networks are more independent of speaker information than the classification neural networks as a result of directly optimising for an effective metric space that captures fine grained word similarities. While the one-

shot classes do not occur in the background data, the learning of an effective metric space explicitly normalises out speaker information.

## 5.5 CHAPTER SUMMARY

In this chapter, we developed a simple benchmark dataset for multimodal one-shot learning containing spoken digits paired with handwritten visual digits. We then compared novel Siamese convolutional neural network (CNN) architectures to traditional direct feature matching models and transfer learning with feedforward and convolutional neural networks on this benchmark dataset. We show that a single CNN with a triplet loss (Chechik *et al.*, 2010; Wang *et al.*, 2014) and online semi-hard mining (Schroff *et al.*, 2015) is more efficient and results in higher accuracies on unimodal (§5.4.1 and §5.4.2) and multimodal (§5.4.3) one-shot tasks than the offline variant which uses shared weight networks, both approaches outperforming the direct feature matching and neural network baselines. In addition, we show that the Siamese models are invariant to the specific speaker uttering a speech query (§5.4.4), capturing only semantic word information by directly optimising for an effective metric space.

One disadvantage of the models within the indirect matching framework is the error compounded by making mistakes on the first step of finding an auxiliary query that may be used to compare to the image matching set. We also note that these models cannot directly take advantage of the weakly supervised multimodal support set, for example, by fine-tuning model representations. This is a result of the successive unimodal non-parametric nearest neighbour modelling of the multimodal support set with suitable metrics defined by speech and vision models respectively. These limitations motivate investigating an end-to-end architecture which may directly match a speech query to the image matching set while taking advantage of the support set for one-shot learning. The simple benchmark dataset also does not capture the true complexity of one-shot learning in a natural realistic setting. This leads us to Part III where we investigate end-to-end architectures within a direct matching framework (§6) on natural images paired with spoken words (§7).

PART III

ONE-SHOT LEARNING NATURAL IMAGES  
PAIRED WITH SPOKEN WORDS

## 6 | A DIRECT FRAMEWORK FOR MATCHING SPEECH TO IMAGES

While the indirect matching framework (see Chapter 4) achieved reasonable results on a simple Multimodal ONE-SHOT learning (MOONSHOT) benchmark (see Chapter 5), it displayed two closely linked limitations: compounding errors due to successive comparisons through the multimodal support set and an inability to directly update parametric models on the learning instances of this set. This is attributed to the non-parametric modelling of the support set items in a nearest neighbour framework. Here we introduce a new framework along with two novel MOONSHOT models which aim to improve on these challenges by directly matching speech to images within an end-to-end architecture. As we discuss later, this framework also lends itself to zero-shot learning (§3.5).

We first describe the general direct matching framework with an illustrative example (§6.1). Thereafter, we describe a deep audio-visual embedding network (DAVENet) which models weakly supervised speech and images in a joint audio-visual space (§6.2). Lastly, we introduce a multimodal model-agnostic meta-Learning algorithm (MuMAML) and show how this may be applied to DAVENet within the direct matching framework (§6.3).

### 6.1 DIRECT MATCHING FRAMEWORK

Direct matching of speech to images relies on defining a model that can measure similarity across modalities. This may be as simple as performing a linear projection from the input space of each modality to a joint space where they may be semantically related. See, for example, the linear regression model in (Lazaridou *et al.*, 2014). Using such a model, one-shot cross-modal matching is accomplished by directly comparing a query to the matching set items which are in a different modality to the query. The closest item is then predicted as the match. The multimodal support set is used in order to learn the joint space wherein the one-shot items from different modalities may be compared. Focussing again on speech and images, this defines a metric  $D_S$  as a mapping from both audio space  $\mathcal{A}$  and vision space  $\mathcal{V}$  to joint audio-visual space  $\mathcal{Z}$ :  $\mathcal{A} \rightarrow \mathcal{Z}$  and  $\mathcal{V} \rightarrow \mathcal{Z}$  respectively; audio and visual inputs are separately mapped to a shared space, allowing us to directly compare speech

and images. Importantly, the “one-shot learning” occurs when we adapt a model on the multimodal support set to learn this mapping. In the absence of a support set, assuming that a model can still map inputs to some joint space, this is equivalent to cross-modal zero-shot learning (Socher *et al.*, 2013; Frome *et al.*, 2013; Lazaridou *et al.*, 2014): the query and matching set items are instances of previously unseen classes—there is no explicit learning phase. Our direct matching framework extends itself naturally to this setting by testing cross-modal matching of speech to images prior to training on any support set.

To illustrate the direct matching framework, consider the MOONSHOT problem shown in Figure 6.1: a model is shown a support set  $\mathcal{S}$  containing natural images paired with spoken words for the classes *surfboard*, *bird*, *basketball*, *guitar* and *horse*, and no labels. The model must use this weakly supervised support set to directly match the spoken test query “bird” (/bɜːd/) most similar to the image of a *bird* in the matching set  $\mathcal{M}_v$ . The speech query  $\hat{x}_a$ , represented in the figure as [?], is compared to all the matching set images in  $\mathcal{Z}$  according to the model’s joint audio-visual metric and 1-nearest neighbour matching (1-NN). We find that the speech query [?] is directly matched to the matching set image of a *bird* and we predict this image as the cross-modal match. This task may also be accomplished in a zero-shot setting by applying the model’s joint audio-visual metric and matching *prior* to viewing and learning the MOONSHOT support set.

A number of approaches may be used to learn a model that can compare cross-modal similarity. We specifically investigate feature embeddings learned by deep audio-visual embedding networks (DAVEnet) (§6.2). We then compare this to an extension of DAVEnet that is trained with a novel multimodal model-agnostic meta-learning algorithm (MuMAML) (§6.3). While both of these models may be fine-tuned on the multimodal support set (following the transfer learning principle), the latter directly optimises for representations which are fast to adapt on only a few instances of novel weakly supervised pairs.

## 6.2 DEEP AUDIO-VISUAL EMBEDDING NETWORKS

Our first approach within the direct matching framework is based on the work of Harwath and colleagues (Harwath and Glass, 2015; Harwath *et al.*, 2016; Harwath and Glass, 2017; Leidal *et al.*, 2017; Harwath *et al.*, 2018, 2019, 2020) on modelling spoken language and visual perception from purely unannotated speech and images. Specifically, these studies investigate learning joint audio-visual representations from weakly supervised speech and images which may be used to directly measure similarity across modalities. These representations are then used to solve various tasks such as localisation of visual objects in images given speech queries. This is another form of metric learning (§4.4): an unsupervised model is trained on a background dataset containing paired inputs in different modalities with no labels (i.e. weakly supervised by co-occurring context) in order to directly measure similarity between the inputs. We focus here on the approach proposed in (Harwath

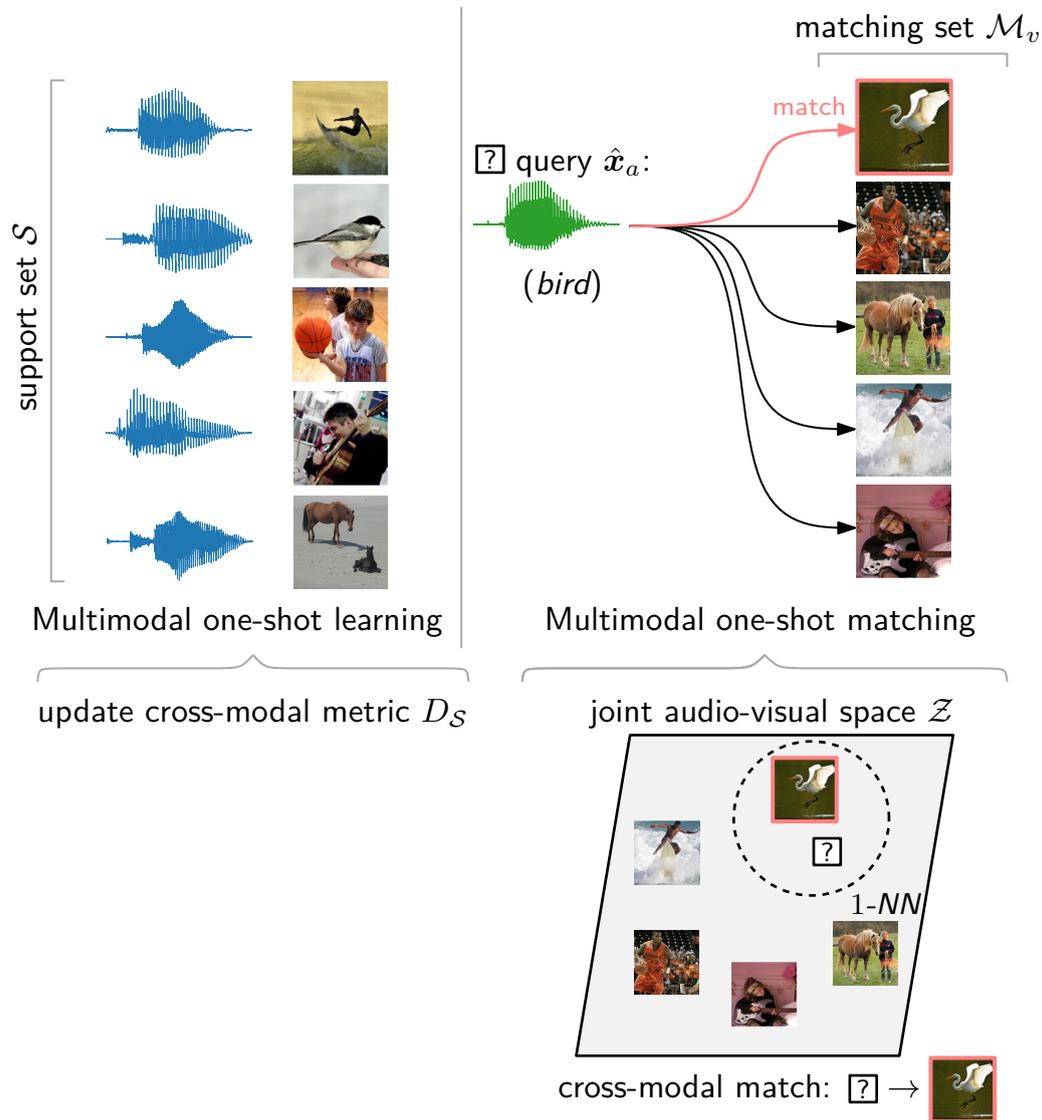


Figure 6.1: Multimodal one-shot learning framework for direct matching of speech to images. The support set contains multimodal pairs for the classes. The speech query “bird” is matched directly to the matching set image of a *bird* in joint audio-visual space. This is accomplished by adapting a joint audio-visual model on the weakly supervised support set during multimodal one-shot learning. In the absence of the support set (i.e. no learning phase) this is equivalent to zero-shot learning.

*et al.*, 2018, 2019) where the authors develop a deep audio-visual embedding network (DAVENet). In our setting, we learn joint audio-visual representations from paired speech and image background data containing *none of the classes* (spoken words and visual objects) tested on during one-shot learning—this is different to the setting considered by Harwath *et al.*, where the goal was cross-modal speech-image retrieval (§3.6).

The DAVENet model consists of two network branches: one for speech and

the other for vision. These networks project speech and images to fixed dimensional embeddings by passing inputs through the respective network branches. The output of these branches have the same fixed size. These embeddings are trained such that semantically related inputs in either modality are mapped close together in a joint audio-visual space. This is achieved by following a metric learning approach, directly influenced by the Siamese neural networks discussed previously (§4.4). Siamese neural networks consist of two (or three in the triplet case) neural network branches with shared parameters. In contrast, the speech and vision network branches of DAVENet are not tied—they may be constructed with different architecture and do not share parameters. Similar to more recent studies on Siamese networks (Chechik *et al.*, 2010; Wang *et al.*, 2014; Hermann and Blunsom, 2014; Hoffer and Ailon, 2015; Schroff *et al.*, 2015), DAVENet employs triplet pairs in order to capture relative similarity between inputs. Both the speech and image instances of a paired input to DAVENet form the anchor of a multimodal triplet pair: the triplet pair with a speech anchor treats the paired image as a positive example and vice versa for the triplet pair with an image anchor. Negatives for these dual triplet pairs are sampled as so-called impostors, where a mismatched image negative is sampled for the speech anchor and a mismatched speech negative is sampled for the image anchor. The two triplet pairs are then used with a straightforward extension of the *triplet loss* (§4.4) to push paired speech and image embeddings to be more similar relative to mismatched speech and image embeddings. DAVENet also follows the efficient Siamese neural network approach and simplifies three shared-parameter networks into a single network (see end of §4.4)—although really DAVENet has two network branches—and embeds a mini-batch of paired speech and image inputs from which it may sample triplet pairs online.

Formally, consider the paired speech-image input  $(\mathbf{x}_a, \mathbf{x}_v)$ . This pair is input respectively to the speech branch (see Figure 6.2) and vision branch (see Figure 6.3) of a DAVENet to produce embeddings in joint audio-visual space. A speech impostor  $\mathbf{x}_a^{\text{imp}}$  and an image impostor  $\mathbf{x}_v^{\text{imp}}$  are also sampled for the speech-image input pair. We discuss later the specific way in which these mismatched instances are sampled. The idea is to push the embeddings of the paired speech and image input  $(\mathbf{x}_a, \mathbf{x}_v)$  close together such that the distance between them is smaller than the distance between embeddings of mismatched pairs  $(\mathbf{x}_a, \mathbf{x}_v^{\text{imp}})$  and  $(\mathbf{x}_a^{\text{imp}}, \mathbf{x}_v)$ —this is equivalent to optimising a Siamese objective for multimodal triplet pairs  $(\mathbf{x}_a, \mathbf{x}_v, \mathbf{x}_v^{\text{imp}})$  and  $(\mathbf{x}_v, \mathbf{x}_a, \mathbf{x}_a^{\text{imp}})$ . We learn embedding functions  $f_a(\cdot)$  and  $f_v(\cdot)$  as the final L2 normalised fully-connected linear layer of the DAVENet speech and vision branches respectively (see yellow speech and vision output layers in Figure 6.2 and Figure 6.3). We train DAVENet with a hinge loss (or triplet loss) for the dual multimodal triplet pairs (Harwath and Glass, 2015; Harwath *et al.*, 2016; Harwath and Glass, 2017; Leidal *et al.*, 2017; Harwath *et al.*, 2018, 2019, 2020), defined as:

$$\begin{aligned} \mathcal{L}(\mathbf{x}_a, \mathbf{x}_v, \mathbf{x}_a^{\text{imp}}, \mathbf{x}_v^{\text{imp}}) = & \max\{0, m + D(\mathbf{x}_a, \mathbf{x}_v) - D(\mathbf{x}_a, \mathbf{x}_v^{\text{imp}})\} \\ & + \max\{0, m + D(\mathbf{x}_a, \mathbf{x}_v) - D(\mathbf{x}_a^{\text{imp}}, \mathbf{x}_v)\} \end{aligned} \quad (6.2.1)$$

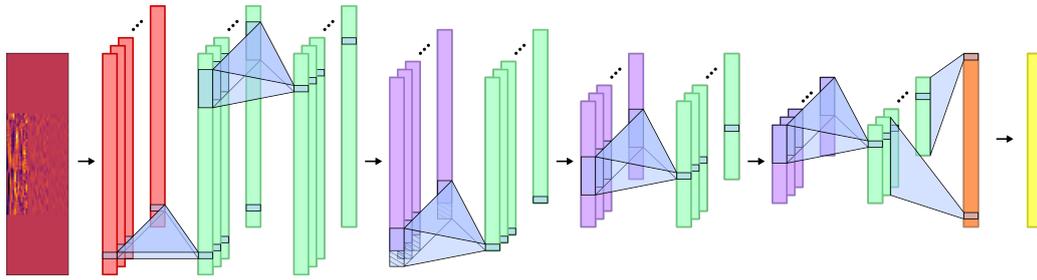


Figure 6.2: Speech network branch of a deep audio-visual embedding network applied to a spoken word “bicycle”. Speech input features are extracted mel-frequency cepstral coefficients centre zero padded to 140 frames (§7.3.1). Convolution layers (1-D) shown in green, batch normalisation layer shown in red, max pooling layers shown in purple, average pooling layer shown in orange and fully-connected linear layer with L2 normalisation shown in yellow.

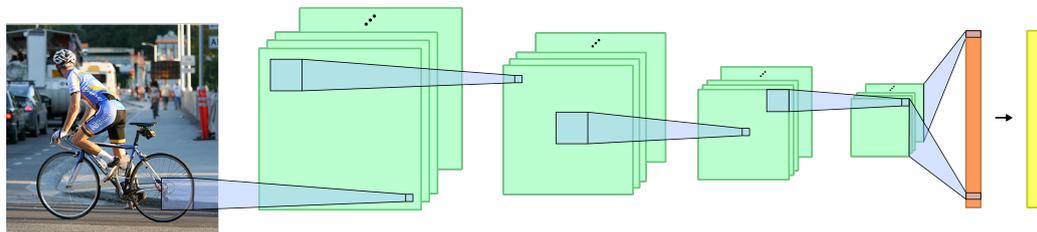


Figure 6.3: Vision network branch of a deep audio-visual embedding network applied to an image of a *bicycle*. Convolution layers (2-D) shown in green, average pooling layer shown in orange and fully-connected linear layer with L2 normalisation shown in yellow.

where  $D(\mathbf{x}_a, \mathbf{x}_v) = \|\mathbf{f}_a(\mathbf{x}_a) - \mathbf{f}_v(\mathbf{x}_v)\|_2^2$  is the squared Euclidean distance and  $m$  is the margin between each of the mismatched pairs.

The success of this model is dependant on the sampling of suitable mismatched pairs against which we may assess relative similarity. We note once again that, in contrast to the original study, we train DAVenet on multimodal background data—disjoint from the one-shot learning classes—containing paired speech and images and no further supervision. For this reason we cannot follow the Siamese neural network approach where negatives (i.e. mismatched pairs) are carefully selected to avoid overfitting (e.g. see online semi-hard mining in §4.4). Instead we follow the approach proposed in [Harwath \*et al.\* \(2020\)](#) which makes use of a blended triplet loss: the sum of two triplet loss terms (see Equation 6.2.1), where the first randomly samples mismatched pairs and the second applies a variant of online semi-hard mining ([Schroff \*et al.\*, 2015](#)) to sample mismatched pairs which are more difficult. Both of these loss terms sample negative examples online from the output embeddings of a paired speech and image mini-batch. Concretely, for each of the paired inputs  $(\mathbf{x}_a, \mathbf{x}_v)$  in a mini-batch, random sampling uniformly selects impostor speech  $\mathbf{x}_a^{\text{imp}}$  and image  $\mathbf{x}_v^{\text{imp}}$  examples, not including the paired input examples themselves. In

the case of semi-hard mining, the most difficult impostor image  $\mathbf{x}_v^{\text{imp}}$  with the smallest distance  $D(\mathbf{x}_a, \mathbf{x}_v^{\text{imp}})$  in the mini-batch satisfying  $D(\mathbf{x}_a, \mathbf{x}_v) < D(\mathbf{x}_a, \mathbf{x}_v^{\text{imp}})$  is selected as the mismatched item for speech input  $\mathbf{x}_a$ , unless there is no such negative image example in which case the one with the largest distance is used. Vice versa for the semi-hard mining of impostor speech  $\mathbf{x}_a^{\text{imp}}$  for image input  $\mathbf{x}_v$ . These mismatched pairs are referred to as *semi-hard* since they are close to the input in the opposite modality but not closer than the distance between the input pair itself. As mentioned in the description of our Siamese neural network approach (§4.4), the purpose of such semi-hard negatives is to combat overfitting as well as ensure stability during training by not naively selecting the hardest impostor examples.

Once trained, a DAVeNet model enables us to compare cross-modal similarity and directly match speech to images. The main advantage of this approach is not only that we may apply it within our direct matching framework (§6.1), but that it also allows us to perform transfer learning by explicitly fine-tuning on a weakly supervised MOONSHOT support set containing novel spoken words paired with visual objects. This is possible since the format of the background data on which DAVeNet is trained is the same as that of the multimodal support set, although the latter contains unseen classes with limited samples. During one-shot cross-modal matching, for each speech-image pair  $(\mathbf{x}_a, \mathbf{x}_v)$  in a multimodal set  $\mathcal{S} = \{(\mathbf{x}_a^{(i)}, \mathbf{x}_v^{(i)})\}_{i=1}^L$ , we therefore sample mismatched pairs as outlined above and apply the blended triplet loss terms (see Equation 6.2.1) to update the parameters of a pre-trained DAVeNet with a small number of gradient descent fine-tuning steps.

The fine-tuning approach to transfer learning is intuitively illustrated in Figure 6.4 (a), which builds on Figure 4.3 (§4.3), where we train a model with parameters  $\theta$  on a background dataset of spoken words with many labelled examples for speech classification—also referred to as a source task.<sup>1</sup> As before, the figure depicts a space of possible configurations for model parameters  $\theta$  and a hypothetical trajectory (represented as a solid line) through the parameter space during training with gradient descent. We specifically consider training on a set of source tasks where a single task is classifying one of the training word classes among the other training word classes. In this example we consider classifying spoken words for source classes *bicycle*, *motorcycle*, *dog* and *squirrel*—depicted in the figure as purple circles. We note again that the coloured circles do not depict the embedding space (i.e. learned representations) determined by the configuration of model parameters  $\theta$ . Instead, the location of a coloured circle in parameter space indicates the optimal configuration of  $\theta$  for the best performance on the task contained within the circle—i.e.  $\theta$  should be in close proximity to the task circle. We learn an optimal configuration of  $\theta$ , depicted in the figure at the end of the trajectory (after the arrowhead), such that the validation error is low among the source tasks—i.e. the optimal  $\theta$  is in close proximity to each of the purple circles. In our previous approach (§4.3), we transfer learn a set of target tasks by simply applying the learned configuration of  $\theta$ . Now, in

<sup>1</sup>We consider the unimodal case in this example for simplicity, although extending to the multimodal case is straightforward.

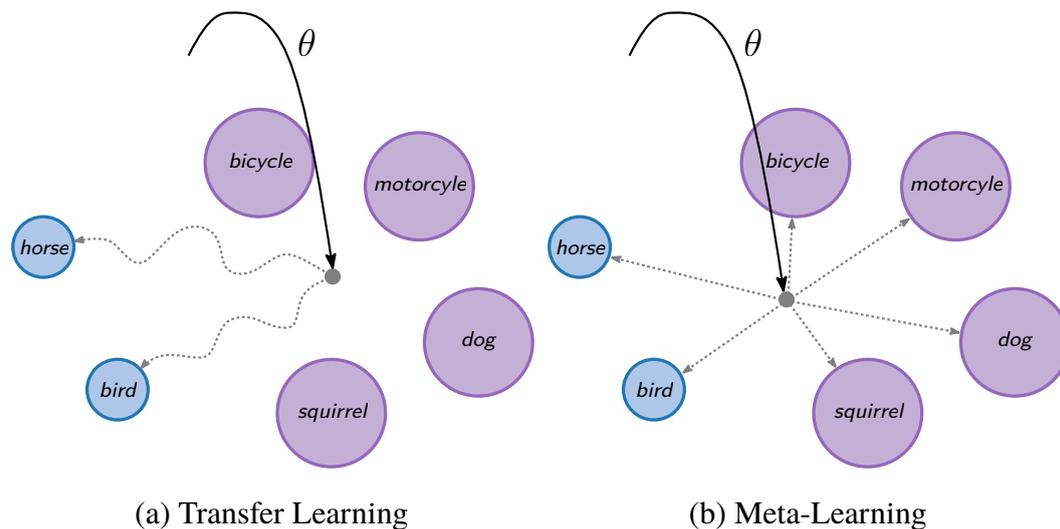


Figure 6.4: Transfer learning (a) and meta-learning (b) illustrated for spoken word classification. In both cases, the goal is to learn model parameters  $\theta$  on a source task (purple circles), where in (a) we consider a task with many labelled examples, while in (b) we consider many tasks with few labelled examples, one of which is shown here. The shared knowledge among the source task (or tasks) is used to learn a prior that is useful for learning related target tasks (blue circles) which have limited data. Here the form of this learned prior is optimal initialisation for model parameters such that fine-tuning may adapt parameters to the target tasks. Meta-learning explicitly optimises this prior such that *fine-tuning is effective* on few examples.

the fine-tuning approach to transfer learning, the learned configuration of  $\theta$  is used as a useful initialisation of the model parameters for target tasks. Specifically, we fine-tune the optimal configuration of  $\theta$  on a small dataset of training examples, or a support set, from a target task using a few steps of gradient descent. For example, classifying novel spoken words *horse* and *bird*—target classes depicted in the figure as blue circles which have limited labelled examples—are each transfer learned by fine-tuning the model parameters which are initialised with the optimal configuration of  $\theta$  among the source tasks. This is illustrated by the hypothetical trajectories (represented as dashed lines) toward adapted  $\theta$  which are optimal for the target tasks. The source tasks are related enough to the target tasks that the optimal configuration of  $\theta$  is in close proximity to that of the target tasks and fine-tuning is efficient with only a few gradient steps on a small support set from the target task. The adapted model parameters result in learned representations which are useful for determining similarity on examples from the target tasks.

There is the possibility of severe overfitting when updating model parameters  $\theta$  with a few gradient updates on small multimodal support sets and we consider this in the experimental chapter (§7.4.3). Once again, we highlight that our previous models within the indirect matching framework (see Chapter 4) are different in that they perform transfer learning only with non-parametric nearest neighbour mod-

elling of a multimodal support set—model parameters are fixed and cannot be fine-tuned. Moreover, these models require supervised background data disjoint from the one-shot learning tasks for transfer learning. DAVenet requires only weakly supervised background data, also disjoint from one-shot learning, containing unannotated speech paired with images.

Our final approach aims to further improve on transfer learning with DAVenet by explicitly training models that are quick and effective to adapt on a small number of weakly supervised multimodal examples. This is accomplished by optimising DAVenet within the framework of meta-learning and we describe this approach next.

### 6.3 MULTIMODAL MODEL-AGNOSTIC META-LEARNING

We now consider a different approach to transfer learning which directly optimises for the ability to learn new tasks quickly and effectively from few examples—this may effectively reduce the possibility of overfitting. To achieve this, we consider meta-learning (§3.4), or learning to learn, which optimises the parameters of a model to learn how to adapt to new tasks. Models within this framework have achieved some of the most successful results on unimodal one-shot learning tasks. These approaches focussed on two types of meta-learning: one that learns a so-called black box meta-policy which effectively updates model parameters (Santoro *et al.*, 2016; Mishra *et al.*, 2018) and another that learns an initialisation for model parameters that is effective for fast adaptation (Vinyals *et al.*, 2016; Finn *et al.*, 2017a). Here we focus on the latter since this simply extends our fine-tuning approach to transfer learning (§6.2) by explicitly rather than implicitly learning a useful initialisation for model parameters. We illustrate this form of meta-learning in Figure 6.4 (b), where we train a model with parameters  $\theta$  on a background dataset of spoken words with many labelled examples for speech classification—we describe later how we extend this to the multimodal case. This model is trained on a variety of many speech classification tasks, each of which contains a number of spoken word classes and only a few labelled examples per class—much like episodes  $\mathcal{T}$  in the one-shot learning problem setting. This may be simulated using a large dataset of spoken words with many labelled examples by repeatedly sampling a number of word classes and only one or a few examples per class: i.e.  $\mathcal{T} \sim p(\mathcal{T})$ . Once again, we consider the example where one such source task is classifying spoken words for classes *bicycle*, *motorcycle*, *dog* and *squirrel* in Figure 6.4 (b). Using the shared knowledge among this and many other such tasks, we meta-learn an optimal  $\theta$ , or prior, which enables us to efficiently transfer learn novel target classes such as the spoken words *horse* and *bird* which have only a few labelled examples. Meta-learning achieves this goal by explicitly learning an optimal initialisation such that *fine-tuning is effective*: model parameters quickly adapt to new tasks with limited data. This is illustrated by the hypothetical trajectories (represented as dashed lines) toward adapted  $\theta$  which are optimal for source tasks and target tasks (depicted in the figure as purple and blue circles respectively); the result of adapted model parameters is learned representa-

tions which are useful for determining similarity on examples from an adapted task. This is different to the fine-tuning approach to transfer learning (§6.2) which only implicitly learns an optimal initialisation for adaptation. The high-level intuition motivating meta-learning is demonstrated by the optimal  $\theta$  in Figure 6.4 (b) which is better positioned among the seen source tasks (each word may be viewed as a separate task) and unseen target tasks and is more direct during fine-tuning in comparison to plain transfer learning, as shown in Figure 6.4 (a).

One successful approach to this type of meta-learning is the popular model-agnostic meta-learning algorithm (MAML) (Finn *et al.*, 2017a,b; Finn and Levine, 2018; Finn, 2018). MAML is a simple approach to meta-learning that is referred to as model-agnostic since it may be applied to any model architecture that is trained with gradient descent, while adding no additional parameters. As previously illustrated, MAML trains a model on many different but related tasks and optimises the model parameters to be easy to fine-tune on small amounts of data from each task. This simple and effective meta-learning algorithm has also achieved some of the most impressive results on unimodal one-shot learning tasks. We consider MAML instead of black-box meta-policy approaches such as memory-augmented neural networks (MANN) (Santoro *et al.*, 2016) and simple neural attentive meta-learner (SNAIL) (Mishra *et al.*, 2018) which perform comparably, but expand the number of learned parameters and place constraints on the model architecture—e.g. MANN requires using a recurrent neural network model architecture and SNAIL requires using temporal convolutions and a causal attention mechanism. We extend MAML to the weakly supervised multimodal setting such that we meta-learn the parameters for a deep audio visual embedding network (DAVENet) (§6.2) for quickly learning new words and object categories. We first describe MAML and then show how we extend this to multimodal model-agnostic meta-learning (MuMAML).

Concretely, in our view of MAML, we consider a *one-shot meta-learning phase*, where at each training step a number of meta-training tasks  $\mathcal{T}$  are sampled from a distribution over related tasks  $\mathcal{T} \sim p(\mathcal{T})$ . Each task contains  $L$  novel classes and we sample one example per class. A model is then trained, or fine-tuned, on the examples from each of these tasks separately based on some objective  $\mathcal{L}$ . A single training step is demonstrated in Figure 6.5 for a set of three speech classification tasks, where each task  $\mathcal{T}$  consists of four word classes—i.e. four-way one-shot learning problem. Importantly, the figure depicts a space of possible configurations for model parameters  $\theta$  and a hypothetical meta-training trajectory (represented as a solid line) through the parameter space during training with gradient descent. Each of the coloured circles represent a single meta-training task; for example, the blue circle depicts the task of classifying spoken words for novel classes  $\{\textit{mountain}, \textit{kayak}, \textit{table}, \textit{pencil}\}$ . Again, these circles do not represent regions in the embedding space. Instead, the location of a coloured circle in parameter space indicates the optimal configuration of  $\theta$  for the best performance on the meta-training task contained within the circle—i.e.  $\theta$  should be in close proximity to the task circle. Fine-tuning the current configuration of  $\theta$  (at the end of meta-training trajectory) with gradient descent on one-shot meta-training data from each task results in the hypothetical trajectories (represented as

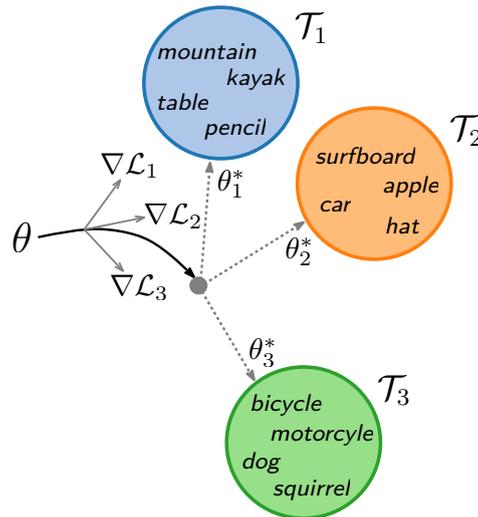


Figure 6.5: Meta-learning demonstrated for spoken word classification tasks. Model parameters  $\theta$  are fine-tuned (dashed lines) on a small number of meta-training data from each task and then evaluated on meta-testing data at these updated parameters. The resulting meta-testing error is used to meta-learn  $\theta$  (solid line) such that it is fast to adapt to new tasks.

dashed lines) toward the adapted model parameters  $\theta_1^*$ ,  $\theta_2^*$  and  $\theta_3^*$  which are optimal for each of the three tasks respectively. Following this fine-tuning step, we sample new unseen instances for each of the meta-training tasks and evaluate the objective  $\mathcal{L}$  on this meta-testing data with the respective adapted model parameters for each task. Finally, we meta-learn the parameters  $\theta$  of the model by computing gradients on the meta-testing errors with respect to  $\theta$  and applying gradient descent with the average across task gradients—this is demonstrated as  $\nabla \mathcal{L}_1$ ,  $\nabla \mathcal{L}_2$  and  $\nabla \mathcal{L}_3$  in the figure for each of the respective tasks and the combined meta-training trajectory (represented as a solid line). This completes a meta-learning step and we repeat this for a large number of steps, each of which samples a new set of tasks and train-test instances. In other words, we consider how well fine-tuning model parameters on limited data from each task performs on some testing data and use this information across tasks to guide model parameters such that it is quick and effective to fine-tune. The goal here is to meta-learn model parameters  $\theta$  such that gradients computed on a few training examples from a new task result in rapid progress. As [Finn et al. \(2017a\)](#) describe it, we are essentially looking to find an initialisation for the model parameters such that small updates to these parameters with gradient descent lead to large improvements in performance on target tasks.

So far we have described a “species” of MAML in the supervised unimodal case. We now turn to the MOONSHOT setting and formally describe a new multimodal model-agnostic meta-learning algorithm (MuMAML) that learns to learn from weakly supervised multimodal paired information. Once again, we focus on learning from paired speech and image background data which does not contain the

target one-shot classes and is weakly supervised by co-occurring context. One disadvantage of this approach is that we require supervised labels for the multimodal background data due to the sampling of meta-training and meta-testing pairs from the same classes during meta-learning. i.e., in contrast to the supervised models in Chapter 4 which require labelled background data separately for the audio and visual modalities, here we require paired labelled data. However, once learned, MuMAML may learn new tasks from weakly supervised paired information with only co-occurring context and no additional supervision.

Specifically, since we are interested in applying MuMAML to DAVeNet, we consider the blended triplet loss described in §6.2. To recap, this weakly supervised objective optimises DAVeNet embedding functions  $f_a(\cdot)$  and  $f_v(\cdot)$  respectively for paired speech and image inputs  $(\mathbf{x}_a, \mathbf{x}_v)$ , defined as:

$$\begin{aligned} \mathcal{L}_{\text{blend}}(f_a, f_v, \mathbf{x}_a, \mathbf{x}_v) &= \mathcal{L}_{\text{random}}(f_a, f_v, \mathbf{x}_a, \mathbf{x}_v) \\ &+ \mathcal{L}_{\text{semi-hard}}(f_a, f_v, \mathbf{x}_a, \mathbf{x}_v) \end{aligned} \quad (6.3.1)$$

where  $\mathcal{L}_{\text{random}}$  and  $\mathcal{L}_{\text{semi-hard}}$  are triplet loss terms (see Equation 6.2.1)—the former randomly samples mismatched pairs and the latter samples mismatched pairs with semi-hard mining. In the context of meta-learning, we consider a task  $\mathcal{T}$  which contains  $L$  paired speech-image classes and generates i.i.d. observations  $(\mathbf{x}_a, \mathbf{x}_v)$  for these classes. During meta-training, we sample one paired input per class from  $\mathcal{T}$  resulting in a batch of  $L$  speech-image pairs  $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_a^{(j)}, \mathbf{x}_v^{(j)})\}_{j=1}^L$ . We then apply the blended triplet loss to optimise parameters  $\theta$  of a DAVeNet model on meta-training data  $\mathcal{D}_{\text{train}}$ , defined as:

$$\mathcal{L}_{\mathcal{T}}(f_{\theta}) = \sum_{(\mathbf{x}_a^{(j)}, \mathbf{x}_v^{(j)}) \in \mathcal{D}_{\text{train}}} \mathcal{L}_{\text{blend}}(f_{\theta}, \mathbf{x}_a^{(j)}, \mathbf{x}_v^{(j)}) \quad (6.3.2)$$

where we have aggregated the parameters of the speech and vision network branches ( $f_a$  and  $f_v$ ) into a single model  $f_{\theta}$  for brevity. This results in a meta-training error which we use to adapt the model parameters  $\theta$  to become  $\theta'$ . Specifically, the updated parameters  $\theta'$  are computed with a step of gradient descent:

$$\theta' = \theta - \alpha \nabla_{\theta} \mathcal{L}_{\mathcal{T}}(f_{\theta}) \quad (6.3.3)$$

where  $\alpha$  is the gradient descent step size (i.e. learning rate). We consider multiple gradient updates in this work and this is achieved by repeated application of Equation 6.3.3. We then sample a set of meta-testing data  $\mathcal{D}_{\text{test}} = \{(\mathbf{x}_a^{(j)}, \mathbf{x}_v^{(j)})\}_{j=1}^N$  from  $\mathcal{T}$ . The adapted model  $f_{\theta'}$  with parameters  $\theta'$  is evaluated with the blended triplet loss on the meta-testing data—i.e. replace  $\mathcal{D}_{\text{train}}$  with  $\mathcal{D}_{\text{test}}$  in Equation 6.3.2 and compute the objective with model  $f_{\theta'}$ . This results in a meta-testing error which is used to optimise initial parameters  $\theta$  for fast adaptation. This is achieved by optimising the adapted model  $f_{\theta'}$  to perform well on the meta-testing data for task  $\mathcal{T}$  with respect to initial model parameters  $\theta$ :

$$\min_{\theta} \mathcal{L}_{\mathcal{T}}(f_{\theta'}) = \min_{\theta} \mathcal{L}_{\mathcal{T}}(f_{\theta - \alpha \nabla_{\theta} \mathcal{L}_{\mathcal{T}}(f_{\theta})}) \quad (6.3.4)$$

**Algorithm 6.1** Multimodal Model-Agnostic Meta-Learning for Speech and Images

**Require:**  $p(\mathcal{T})$ : distribution over multimodal tasks; each task  $\mathcal{T}$  contains  $L$  unique paired speech-image classes

**Require:**  $\alpha, \beta$ : step size hyperparameters

- 1: randomly initialise model parameters  $\theta$
- 2: **while** not done **do**
- 3:   sample batch of tasks  $\mathcal{T}_i \sim p(\mathcal{T}_i)$
- 4:   **for all**  $\mathcal{T}_i$  **do**
- 5:     sample one speech-image example per class  $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_a^{(j)}, \mathbf{x}_v^{(j)})\}_{j=1}^L$
- 6:     evaluate  $\nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(f_{\theta})$  using  $\mathcal{D}_{\text{train}}$  and  $\mathcal{L}_{\mathcal{T}_i}$  in Equation 6.3.2
- 7:     compute adapted parameters with gradient descent  $\theta'_i = \theta - \alpha \nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(f_{\theta})$
- 8:     sample meta-testing speech-image examples  $\mathcal{D}_{\text{test}}^{(i)} = \{(\mathbf{x}_a^{(j)}, \mathbf{x}_v^{(j)})\}_{j=1}^N$
- 9:     **end for**
- 10:   update model parameters  $\theta \leftarrow \theta - \beta \nabla_{\theta} \sum_{\mathcal{T}_i \sim p(\mathcal{T}_i)} \mathcal{L}_{\mathcal{T}_i}(f_{\theta'_i})$  using each  $\mathcal{D}_{\text{test}}^{(i)}$  and  $\mathcal{L}_{\mathcal{T}_i}$  in Equation 6.3.2
- 11: **end while**

where  $\mathcal{L}_{\mathcal{T}}$  is evaluated on  $\mathcal{D}_{\text{test}}$ . Importantly, we compute the weakly supervised objective  $\mathcal{L}_{\mathcal{T}}$  with the adapted model parameters  $\theta'$  while optimising the initial model parameters  $\theta$ . We perform this meta-optimisation across a batch of meta-training tasks  $\mathcal{T}_i$  sampled from a distribution over related multimodal paired information tasks  $\mathcal{T}_i \sim p(\mathcal{T}_i)$ . Specifically, we meta-learn the model parameters  $\theta$  with a stochastic gradient descent method for meta-optimisation across the batch of tasks, resulting in updated model parameters:

$$\theta \leftarrow \theta - \beta \nabla_{\theta} \sum_{\mathcal{T}_i \sim p(\mathcal{T}_i)} \mathcal{L}_{\mathcal{T}_i}(f_{\theta'_i}) \quad (6.3.5)$$

where  $\beta$  is the meta-learning step size. This completes MuMAML and we summarise these steps in Algorithm 6.1.<sup>2</sup>

Essentially MuMAML optimises model parameters such that a few gradient descent steps on limited weakly supervised speech and image data from a new task performs well with respect to test data. Following training of DAVeNet or a similar model with the MuMAML algorithm, learning a new target task is as simple as applying the fine-tuning step described in Equation 6.3.3. For example, we may adapt model parameters in the MOONSHOT setting for target task  $\mathcal{T}$  by evaluating  $\mathcal{L}_{\mathcal{T}}(f_{\theta})$  on a multimodal support set  $\mathcal{S} = \{(\mathbf{x}_a^{(i)}, \mathbf{x}_v^{(i)})\}_{i=1}^L \sim \mathcal{T}$  and performing the gradient descent step. In other words, training in the MOONSHOT setting is exactly the same as in the meta-training step of MuMAML. Therefore MuMAML directly optimises for the MOONSHOT problem.

<sup>2</sup>Note that we show only a single gradient update on meta-training data for brevity, although extending to multiple gradient updates is straightforward.

## 6.4 CHAPTER SUMMARY

In summary, we have introduced a new cross-modal matching framework for the direct matching of speech to images within the multimodal one-shot learning problem setting; the aim of which is to improve on the limitations of the indirect matching framework: compounding errors from successive unimodal comparisons and an inability to fully take advantage of multimodal learning samples. This is achieved with end-to-end architecture that models cross-modal similarity. We proposed a deep audio-visual embedding network (DAVEnet) as an approach to learning a joint audio-visual metric space wherein we may perform cross-modal comparisons. We described how this model builds on Siamese triplet networks to metric learn cross-modal feature embeddings purely from weakly supervised speech paired with images, without any labels and disjoint from the one-shot learning task. One advantage of this model is that it may be fine-tuned directly on a weakly supervised multimodal support set, taking full advantage of these learning samples to update model parameters. Thereafter, we proposed an algorithm for training a DAVEnet model to explicitly learn how to learn quickly and effectively from few examples of novel weakly supervised speech-image pairs. We accomplish this within the framework of meta-learning, and formalise a multimodal model-agnostic meta-learning algorithm. We have shown how this approach directly optimises for the multimodal one-shot learning task. In the following chapter, we compare models within the indirect matching framework to DAVEnet models within the direct matching framework, trained both in the standard way and with our multimodal meta-learning algorithm, for multimodal one-shot learning on a complex benchmark dataset containing natural images paired with spoken words (see Chapter 7).

## 7 | FLICKR 8K AND FLICKR AUDIO EXPERIMENTS

In our previous investigation of Multimodal ONE-SHOT learning (MOONSHOT), we considered a simple benchmark task based on speech and image digits (Chapter 5). We now consider a more naturalistic setting that may capture some of the complexities of true one-shot learning in the wild. To accomplish this goal, we create a new MOONSHOT benchmark dataset from natural images paired with spoken words (§7.1). In addition, we introduce an accompanying natural speech and images background dataset, not containing any of the classes in the MOONSHOT task, which we use to train multimodal neural networks (§7.2). Thereafter, we describe our experimental setup (§7.3), including data preprocessing, the implementation and training of our models, and one-shot task evaluation. We consider both models using an indirect matching framework (Chapter 4), which we have investigated previously, as well as new models within a direct matching framework (Chapter 6). Finally, we present results for unimodal and multimodal one-shot task experiments on the benchmark dataset comparing these models and discuss the advantages and limitations of these approaches (§7.4).<sup>1</sup>

### 7.1 A MULTIMODAL NATURAL SPEECH AND IMAGES BENCHMARK DATASET

Although our previous investigation of a speech-image digit benchmark dataset has shown promising results (see Chapter 5), the task did not capture the complexity of one-shot learning in a naturalistic setting. To be clear, in the case of these speech-image digit pairs, there is no ambiguity as to which visual item or concept a spoken word refers to since only a single visual digit is present. This is unlike more realistic problem settings where an agent may be presented with an ambiguous visual context paired with a spoken description and must learn to identify which item or concept a word refers to. To demonstrate this scenario, consider the example image shown in Figure 7.1, paired with the spoken word “bird”. For an experienced learner who has

---

<sup>1</sup>Full code recipe for the Flickr 8K and Flickr Audio experiments described in this chapter available at: <https://github.com/rpeloff/moonshot>.



Figure 7.1: The image from the natural speech and images benchmark dataset paired with spoken word “bird”. This demonstrates the difficulty of the multimodal one-shot learning task, where a learner must simultaneously learn a new word, identify which visual object or concept this word refers to and generalise this word-object mapping to unseen instances of the spoken word and visual object.

knowledge of what constitutes the concept *bird*, the task of identifying this image as containing a bird may seem trivial. However, consider that the image is paired with the novel spoken word “dax”. Now it is not clear whether this word refers to the bird, the crashing water, the river itself, the rocks or the flora. This is an example of the true MOONSHOT setting where a learner must simultaneously learn to recognise a new spoken word, identify which visual object or concept this word refers to and generalise to new instances of the spoken word or image *dax*. This leads us to introduce a new MOONSHOT benchmark dataset that emulates this setting: learning natural images paired with isolated spoken words. The image in this example is taken from this dataset.

Specifically, we consider the Flickr Audio Caption corpus (Harwath and Glass, 2015) which contains 40 000 spoken descriptions sampled at 16kHz for 8 000 natural images with 5 captions per image in the Flickr 8K corpus (Hodosh *et al.*, 2013). Using the paired spoken captions and images from these datasets, we create a new MOONSHOT benchmark dataset. To construct this dataset, we first split the spoken captions into isolated words using the forced alignments obtained from Harwath and Glass (2015) (computed using the ground truth transcriptions). We apply a filtering process to select text “keywords” from the ground truth transcriptions of the spoken captions. These keywords correspond with the isolated spoken words for each of the images and the goal is to select words which are semantically relevant to their paired images. Keyword filtering involves the following steps applied to the ground truth text transcriptions for each spoken caption:

1. We apply a pre-trained English language model to process the textual sentences from ground truth transcriptions, identifying keyword tokens (i.e. individual words). This involves removing invalid tokens—specifically, commonly used stop words which should be ignored, punctuation and out-of-vocabulary

words—and performing lemmatisation to convert keywords to their dictionary form, or *lemma*, based on the intended meaning within a sentence (e.g. swimming → swim), allowing us to group variations of keywords. We achieve this using the publicly available spaCy large English model (Honnibal and Montani, 2017). This results in a number of keyword-image pairs.

2. We filter the quality of the identified keyword-image pairs by keeping only the keywords that occur in at least 2 of the 5 unique image captions. We found that this requirement significantly reduced the number of highly ambiguous words for a given image.
3. We then filter out keywords that are too infrequent for MOONSHOT. Since we are interested in evaluating a MOONSHOT learner on 1- and 5-shot tasks with up to 15 potential test queries from any given class (i.e. keyword), we remove keywords which are paired with less than 20 unique images.

This process results in a large dataset of filtered keyword-image pairs. We then randomly sample a small set of unique keywords from this data and manually remove a few ambiguous terms. This produced a set of 30 unique *one-shot keyword classes* which we consider for multimodal one-shot learning:

```
asian basketball bench bird blonde boat car cliff climber
dance fire floor ground guitar hair hill horse obstacle
paddle path purple rope sand sled snowboard splash suit
surfboard throw vest
```

We filter the dataset of keyword-image pairs keeping only these one-shot learning keyword classes. Finally, we use this data to form our natural MOONSHOT benchmark dataset, defining a task distribution from which we may sample MOONSHOT episodes  $\mathcal{T} \sim p(\mathcal{T})$ . Specifically, for each keyword-image pair we retrieve the corresponding isolated spoken word (from the same image caption as the keyword) and natural image pair from the split Flickr Audio captions and Flickr 8K images respectively. Our resulting one-shot learning benchmark dataset contains 1 346 unique speech-image pairs with 1 243 unique images. The final splits that we use to form this benchmark dataset are made available as part of our code recipe for these experiments. We show excerpts for a few of the benchmark dataset classes in Appendix B.1.

## 7.2 BACKGROUND DATA FOR NEURAL NETWORK MODELS

Aside from the direct feature matching baseline (§4.2), all of the models that we have proposed rely on some form of large background dataset for building on prior experience to perform one-shot learning. The neural network models within our indirect matching framework (Chapter 4) require supervised unimodal background data for training both speech and vision networks separately. This is different to the models in our direct matching framework (Chapter 6) which require multimodal paired

speech and image background data for joint audio-visual modelling. Specifically, in the case of a deep audio-visual embedding network (DAVEnet), we require only weakly supervised speech and image pairs, without any labels and only co-occurring context. This is advantageous since labelling such large datasets is expensive. In the case of DAVEnet trained with our multimodal model-agnostic meta-learning algorithm (MuMAML) we still require speech and image pairs, however we also require labels for these multimodal examples. As discussed in §6.3, this is due to the requirement for defining a distribution over tasks  $p(\mathcal{T})$  where a task  $\mathcal{T}$  contains a number of classes (i.e. labelled data) from which we may sample paired speech and image meta-training and meta-testing data. In any case, the Flickr Audio Caption corpus (Harwath and Glass, 2015) paired with the corresponding caption images in the Flickr 8K corpus (Hodosh *et al.*, 2013) may be used for training all of these models since this dataset contains natural speech and image pairs as well as textual transcriptions which may be used for labels.

This dataset is the same as the one used to create a natural speech and images benchmark dataset for one-shot learning (§7.1). It is important to ensure that there is no overlap between the background training and validation data classes and the one-shot learning classes. We therefore first split the spoken captions into isolated words using the provided forced alignments and then follow the keyword filtering process (as described in §7.1 above) to the textual transcriptions of the spoken captions, resulting in a large dataset of filtered keyword-image pairs. Thereafter, we remove all images paired with the 30 one-shot keyword classes. Note that we also remove keyword-image pairs in cases where keywords are not in the one-shot keyword classes but have paired images that are in the one-shot learning dataset (i.e. by being paired with one of the one-shot keyword classes). The resulting dataset of keyword-image pairs is used to form our background dataset, disjoint from the one-shot learning benchmark dataset, by again retrieving the corresponding isolated spoken word and natural image pairs from the split Flickr Audio captions and Flickr 8K images respectively. The final background dataset contains 179 keyword classes, 21 094 unique speech-image pairs with 4 622 unique images for training and 2 575 unique speech-image pairs with 769 unique images for validation. For the unimodal models, we train on only the spoken words or images in this paired dataset respectively. When we require labels for supervision we use the lemmatised text keywords. We show the full list of multimodal background keyword classes in Appendix B.2. The final splits are also made available with our code recipe.

One problem with this background dataset is that it only contains 4 622 unique training images which is not sufficient for training a typical deep neural network from scratch, i.e. without any prior pre-training. In the case of our unimodal vision models we therefore make use of an extended background dataset comprised of the Flickr 30K (Young *et al.*, 2014) and MSCOCO (Lin *et al.*, 2014) image datasets. Like Flickr 8K, these datasets contain images paired with text captions. Flickr 8K is in fact included in Flickr 30K and we remove the duplicate instances from the latter. Once again, we need to ensure that these datasets do not contain any of the one-shot classes. We apply the same filtering process as described above to the text captions

in these datasets to identify keyword-image pairs. We remove the one-shot keyword classes and use the resulting dataset of keyword-image pairs to retrieve images from the Flickr 30K and MSCOCO datasets respectively (where corresponding keywords are used as image labels). In combination with the images from our multimodal background dataset, this forms an extended supervised multi-label vision background dataset containing 1 256 keyword classes, 579 085 unique keyword-image pairs with 118 083 unique images for training and 20 313 unique keyword-image pairs with 5 397 unique images for validation.

## 7.3 EXPERIMENTAL SETUP

In this section we discuss the details of our experiments on the natural speech and images benchmark: preprocessing applied to the speech and image data (§7.3.1), implementation and training procedure of our proposed models (§7.3.2) and how we evaluate the unimodal and multimodal one-shot tasks (§7.3.3).

### 7.3.1 Data Preprocessing

We first describe speech preprocessing and then image preprocessing. We parametrise speech as mel-frequency cepstral coefficients (MFCC) with first and second order derivatives, computed with a 25 millisecond window size and a 10 millisecond frame shift, yielding 39-dimensional feature vectors for isolated spoken words in both the background dataset and the one-shot learning dataset. For our neural network models (which require fixed length inputs), we centre zero-pad or crop speech segments to 140 frames. Although dynamic time warping (DTW) may deal directly with variable length segments given that we normalise out their duration, we follow the suggestions of [Mueen and Keogh \(2016\)](#) and re-interpolate all speech segments to have the same length as the longest segment in the batch before applying the DTW algorithm. This simplifies the problem of choosing a suitable normalisation. When comparing MFCC-encoded speech segments using DTW, we standardise the speech features using the mean and variance computed per feature dimension from the background training data.

Image pixels are normalised to the range  $[-1, 1]$ . For direct feature matching baseline with cosine distance over image pixels, we square crop images along their shortest edge and upsample or downsample to  $256 \times 256$ . We take a different approach when training our neural network models and follow the typical procedure in computer vision where image data is augmented to maximise the available training data and improve the generalisation ability of these models. We specifically follow the approach proposed by [He et al. \(2016\)](#) and perform the following steps: (1) an image is randomly resized along its shortest edge to values in the range  $[299, 480]$  for scale augmentation, (2) the image is randomly flipped left or right for horizontal flip augmentation, (3) the image colour is augmented by randomly selecting brightness and saturation adjustments and random ordering of these operations and finally (4)

the image is randomly square cropped to  $299 \times 299$ . This results in an image that is a potentially zoomed in crop of the original image. Testing the neural network models follows the same as direct pixel matching and we simply square crop and resize images along their shortest edge to  $299 \times 299$ . Image inputs to the neural network models are  $299 \times 299$  which is the required input dimension and we discuss this below.

### 7.3.2 Model Implementation

We implement all of our neural network models in TensorFlow (TF) (Abadi *et al.*, 2015) and unless specified otherwise we build these network from the ground up using TF layers—we provide our implementation of these networks in the code recipe as a contribution. We consider convolutional neural networks for modelling both spoken words and visual objects. In all cases, we follow a transfer learning approach to training, where we first train a classification network and apply this network as a pre-trained feature extractor for the data used to train our target neural networks. Essentially we freeze the parameters of a pre-trained classifier network and remove the final layer which computes unnormalised log-probabilities specific to the classification pre-training (i.e. the logits layer). The resulting network, which we refer to as the *base network*, is then used for training all other neural network models by appending additional neural network layers to the top of this network. We note that implementation is simpler as we simply need to extract embeddings for training and testing data using the base network and then apply these embeddings as input to the target neural networks. In each epoch, or iteration of the training data, we sample a new set of 4 random data augmentations for each of the images in the training data (§7.3.1).

We carefully select the hyperparameters described in this section by tuning models using one-shot error on the validation sets of the background data. This results in models which perform as best as possible during evaluation such that their underlying mechanisms may be fairly compared. All of our neural network models are trained for 100 epochs using the Adam optimiser (Kingma and Ba, 2015) with a learning rate of  $3 \cdot 10^{-4}$  which is step decayed by 0.96 after every 2 epochs. We also apply gradient norm clipping to mitigate exploding gradients (Pascanu *et al.*, 2013), where gradients are scaled such that the global L2-norm is not larger than 5. We tuned all our models using unimodal or multimodal one-shot learning respectively on the background validation data. We also perform early stopping based on the one-shot validation task.

**Speech Models** For modelling spoken words we implement a base convolutional neural network with the same architecture as the speech branch of the deep audio-visual embedding network (DAVENet) proposed in Harwath *et al.* (2018, 2019). This network has the following architecture: an input layer; batch normalisation (BN),  $39 \times 1$  convolution with 128 filters; rectified linear unit (ReLU);  $1 \times 11$  convolution with 256 filters; BN; ReLU;  $1 \times 2$  max pooling;  $1 \times 17$  convolution with 512

filters; BN; ReLU;  $1 \times 2$  max pooling;  $1 \times 17$  convolution with 512 filters; BN; ReLU;  $1 \times 2$  max pooling;  $1 \times 17$  convolution with 1024 filters; BN; ReLU; global average pooling; dropout; 2048-unit fully-connected; ReLU; dropout. Here we have changed the embedding dimension of the DAVenet speech branch from 512 in the original architecture to 1024 and added a global average pooling layer (to collapse the temporal dimension) followed by a 2048-unit fully-connected layer and ReLU. We also followed the typical computer vision practice of adding batch normalisation (BN) (Ioffe and Szegedy, 2015) layers and dropout layers (Srivastava *et al.*, 2014). Both of these layers have been shown to regularise neural networks and improve their generalisation ability. We specifically use BN to regularise convolutional layers and dropout to regularise fully-connected layers. We found a dropout rate of 0.2 to work well on the validation task. To train this base network, we add an additional 179-unit fully-connected logits layer with softmax activation for the spoken word classes in the background data (§7.2). We then optimise model parameters to minimise a categorical cross-entropy objective on the training data, sampling batches of 32 spoken words at each step. This results in both a trained speech base network (without the logits layer) and a classifier speech model (§4.3) which we apply within our indirect matching framework and test for one-shot learning. We then trained a Siamese speech model (§4.4) on top of the pre-trained (frozen) base network, adding only a linear 1024-unit fully-connected layer with L2 normalisation, again on the spoken word background data. Like in our previous experiments on the digits benchmark dataset (§7.3.2), we sample balanced batches of  $p = 64$  and  $k = 8$  examples from which we may select triplet pairs. We consider online semi-hard mining of triplet pairs in these experiments and our Siamese speech model corresponds to the *Siamese CNN (online)* described in §4.4.

**Vision Models** To model visual objects in images we use the InceptionV3 network architecture (Szegedy *et al.*, 2016), which has been shown to achieve impressive results compared to the state-of-the-art computer vision models on the ImageNet classification challenge (Russakovsky *et al.*, 2015). Yet, this model has far fewer parameters than most of its competitors such as the VGG network (Simonyan and Zisserman, 2015) that is used by Harwath *et al.* (2018, 2019) as the architecture for the DAVenet vision branch. This smaller architecture is accomplished by incorporating novel factorised convolutions and strong regularisation techniques. To implement this network we use the pre-built module available in TF<sup>2</sup> and discard the final logits layer which is specific to ImageNet classification. Importantly, we do not use the pre-trained ImageNet weights that may be loaded with this module since the visual objects present in the ImageNet images overlap with the one-shot classes that we consider. Instead, we train InceptionV3 from random initialisation. We do not describe the full InceptionV3 architecture here for the sake of brevity,

---

<sup>2</sup>We specifically used the `tf.keras.applications.InceptionV3` module available in TensorFlow 2.0.

but the reader can consult our released code for further details.<sup>3</sup> Here we simply note that InceptionV3 takes as input  $299 \times 299 \times 3$  RGB-colour images, processes images with a sequence of convolutional blocks containing different configurations of convolution layers, and outputs global average pooled (to collapse the spatial dimensions) embeddings with dimension 2048. We encourage the interested reader to investigate (Szegedy *et al.*, 2016) for further details on the InceptionV3 architecture. We also add a dropout layer with rate 0.2 to the top of InceptionV3. We trained this vision base network by adding a final 1256-unit fully-connected logits layer with sigmoid activation for the image classes in the extended multi-label vision background dataset (§7.2). Here we are dealing with a multi-label multi-class classification problem, where all classes present in a given image should be correctly identified. To accomplish this we use each of the network output logits as a Bernoulli variable specifying the probability of the corresponding class being present in an image. Since each image has only 1 or a few labels present (i.e. sparse multiple labels), we found that using a binary cross-entropy objective applied to the logits did not perform well (as measured by precision and recall) and would often converge to predicting no classes as being present in an image (i.e. the majority case). To remedy this we investigated training with focal loss (Lin *et al.*, 2017), an objective that adds a simple factor to the cross-entropy objective to focus in on the more difficult examples which are misclassified. The intuition is that the negative log-likelihood  $-\log(p)$  for easy outputs where  $p \geq 0.5$  and a class is correctly predicted as present is significant enough that the sum over many such outputs outweighs the few difficult cases where  $p < 0.5$  and classes are misclassified. The focal loss mitigates this effect by reducing the magnitude of the loss for easy outputs and is defined as:

$$\mathcal{L}_{\text{focal}} = -y(1-p)^\gamma \log(p) - (1-y)p^\gamma \log(1-p) \quad (7.3.1)$$

where  $\gamma$  is a “focussing” parameter. Larger values for  $\gamma$  decrease the effect of correctly classified outputs at a faster rate. Note that when  $\gamma = 0$  in Equation 7.3.1, it reduces to the standard cross-entropy objective. We optimised the parameters of the InceptionV3 model to minimise the focal loss on the background training data, sampling batches of 32 images at each step. We found  $\gamma = 2$  to work well, resulting in precision of 0.568 and recall of 0.359 on the validation data, improving on binary cross-entropy. Similar to before, this gave a trained vision base network (without the logits layer) and a classifier vision model (§4.3) which we apply within our indirect matching framework and test for one-shot learning. To demonstrate the class leakage that the ImageNet weights introduce, we also consider an *oracle* classifier vision model which loads the InceptionV3 module with pre-trained ImageNet weights and performs no further training. This model has “seen” the one-shot classes which overlap with the ImageNet classes resulting in an unfair advantage over true one-shot learners. Finally, we trained a Siamese vision model (§4.4) on top of the pre-trained (frozen) base network, adding only a linear 1024-unit fully-connected layer with L2 normalisation. As with the Siamese speech model, we sample balanced batches of

<sup>3</sup>Code available at: <https://github.com/rpeloff/moonshot>.

$p = 32$  and  $k = 4$  from which we may select triplet pairs using online semi-hard mining.

**Audio-Visual Models** We train audio-visual neural network models for our direct matching framework (§6.1) building on the base speech and vision networks discussed above. Specifically, we train a DAVeNet audio-visual model (§6.2) by adding small feedforward neural networks on top of the pre-trained frozen speech and vision base networks. We use the same architecture for both the speech and vision networks: 1024-unit fully-connected; ReLU; dropout; 512-unit fully-connected; L2 normalisation. We set the dropout layer to have rate 0.2 as with our prior models. We trained this model on the paired speech and image background dataset (§7.2), without using any labels and considering only weakly supervised pairs, sampling batches of 256 speech-image pairs at each step. During validation we fine-tune DAVeNet on the multimodal one-shot validation data for 10 steps of stochastic gradient descent (SGD) with learning rate  $5 \cdot 10^{-3}$ . Finally, we trained a variant of this model with our MuMAML algorithm (§6.3). Specifically, we train a MuMAML audio-visual model, with the same architecture as the DAVeNet model, on the paired speech and image background data, utilising the labels to sample batches of 4 tasks containing 10 classes at each step for meta-learning. This model was trained for 75 000 steps with a meta-learning rate  $\beta = 3 \cdot 10^{-4}$ , using 5 gradient descent steps on one-shot meta-training data and (inner) learning rate  $\alpha = 5 \cdot 10^{-3}$ . After every 5 000 steps, we decayed the meta-learning rate by 0.96 and validated the model. As with DAVeNet, we fine-tune MuMAML on the multimodal one-shot validation data for 5 steps of SGD with learning rate  $5 \cdot 10^{-3}$  (the same setup as used during meta-training).

**Fine-Tuning Models** We fine-tune both unimodal and multimodal models. In both cases, we found that fine-tuning with the Adam (Kingma and Ba, 2015) optimiser performed better than using vanilla stochastic gradient descent (SGD) and we use this throughout testing. In the unimodal one-shot learning case, speech and vision classifiers are fine-tuned by replacing the logits layer with a new softmax logits layer for the  $L$  one-shot classes. We fine-tune these networks for 15 SGD steps with learning rate  $1 \cdot 10^{-2}$  on shuffled mini-batches (maximum size 32 if possible) sampled from labelled unimodal one-shot support sets. We present two variants of this fine-tuned model, one where we match with cosine similarity on the fine-tuned embedding layer and another where we match with the new softmax classification outputs. For the former we freeze all network layers before the final fully-connected layer (prior to the logits), while for classification we freeze all network parameters except for the new logits layer. Unimodal speech and vision Siamese models are fine-tuned by simply applying the triplet loss on the labelled unimodal one-shot support sets. No extra layers are added and we freeze all network layers besides the final fully-connected layer. We fine-tune these models for 15 SGD steps with learning rate  $1 \cdot 10^{-2}$  for 10-way tasks and  $1 \cdot 10^{-3}$  for 20-way tasks. Note that we cannot fine-tune here in the one-shot case and only consider five-shot fine-tuning since the

triplet loss relies on at least two examples per class to form a minimum of one anchor-positive pair. In the multimodal case, only DAVenet and the MuMAML variant can be fine-tuned. In fact, without fine-tuning these models are evaluated for zero-shot learning. We fine-tune both models by freezing only the base networks and applying the blended triplet loss on the weakly supervised multimodal support set for 15 SGD steps with learning rate  $5 \cdot 10^{-3}$  (the same setup as used during meta-training for MuMAML and background validation for both models).

### 7.3.3 One-Shot Task Evaluation

We evaluate all of our models on one-shot classification or matching tasks, and average accuracies reported over 400 episodes with 95% confidence intervals. Unimodal one-shot learning considers only spoken words or natural images randomly sampled from the respective benchmark dataset modality, while MOONSHOT episodes randomly sample paired natural speech and images from the benchmark dataset. We consider  $L$ -way one-shot and five-shot (§2.4) tasks for both  $L = 10$  and  $L = 20$  classes. We follow the same procedure as in our digits experiments (§5.3.3) when testing on a MOONSHOT task: a matching set is sampled, containing  $N = L$  images (one for each of the  $L$  image classes) not seen in the support set. A speech query instance from one of the  $L$  spoken word classes is sampled, also not seen in the support set. The query then needs to be matched to the correct item in the matching set. Testing unimodal one-shot learning is similar but without a matching set, where we instead sample only unseen speech or image queries for the classes in the support set which need to be classified to the correct class. Within a unimodal or multimodal episode, 15 different query instances—and matching sets in the multimodal case—are sampled while keeping the support set fixed.

## 7.4 EXPERIMENTAL EVALUATION

We now consider unimodal and multimodal one-shot learning experiments on the natural speech and images benchmark dataset. In our initial investigation of this complex dataset we test our proposed unimodal models on one-shot speech classification (§7.4.1) and one-shot image classification (§7.4.2). Thereafter, we evaluate MOONSHOT models within both our indirect matching framework and our new direct framework for cross-modal matching of speech to images (§7.4.3). We then perform one final investigation on the speaker invariance of our MOONSHOT models where we test if they are invariant to the query speaker (§7.4.4).

### 7.4.1 One-Shot Speech Classification

Unimodal one-shot speech classification has so far only been considered in [Lake et al. \(2014\)](#) and our prior work on the digits benchmark dataset ([Eloff et al., 2019](#)) (see Chapter §5). We now consider this task on the isolated spoken words from the Flickr-Audio caption corpus (§7.1) which has not been considered before. Here we

Table 7.1: 10- and 20-way 1- and 5-shot speech classification results on isolated spoken words sampled from the Flickr Audio dataset.

Speech Model	Match	Fine-Tune	10-way Accuracy		20-way Accuracy	
			1-shot (%)	5-shot (%)	1-shot (%)	5-shot (%)
Random	random	–	9.5 ± 0.7	–	4.7 ± 0.6	–
DTW	cosine	–	69.2 ± 1.4	89.6 ± 0.8	60.6 ± 1.3	82.1 ± 1.0
Classifier	cosine	No	88.5 ± 0.9	96.0 ± 0.5	83.9 ± 1.0	93.1 ± 0.7
Classifier	cosine	Yes	<b>89.8 ± 0.9</b>	<b>97.0 ± 0.5</b>	85.1 ± 1.0	94.3 ± 0.6
Classifier	softmax	Yes	89.3 ± 0.9	96.6 ± 0.5	85.1 ± 0.9	93.2 ± 0.7
Siamese	cosine	No	89.6 ± 0.9	96.4 ± 0.5	<b>85.2 ± 1.0</b>	94.2 ± 0.6
Siamese	cosine	Yes	–	96.6 ± 0.5	–	<b>94.4 ± 0.6</b>

Table 7.2: 10- and 20-way 1- and 5-shot image classification results on natural images sampled from the Flickr 8K dataset.

Vision Model	Match	Fine-Tune	10-way Accuracy		20-way Accuracy	
			1-shot (%)	5-shot (%)	1-shot (%)	5-shot (%)
Random	random	–	9.5 ± 0.7	–	4.7 ± 0.6	–
Pixels	cosine	–	15.7 ± 1.0	25.1 ± 1.1	10.1 ± 0.8	20.8 ± 1.0
Oracle	cosine	No	41.3 ± 1.4	57.2 ± 1.4	31.6 ± 1.2	47.5 ± 1.3
Classifier	cosine	No	33.6 ± 1.3	49.9 ± 1.4	24.3 ± 1.1	40.6 ± 1.3
Classifier	cosine	Yes	33.9 ± 1.3	52.4 ± 1.4	24.3 ± 1.2	<b>41.3 ± 1.3</b>
Classifier	softmax	Yes	34.4 ± 1.3	<b>52.6 ± 1.4</b>	25.0 ± 1.1	37.5 ± 1.3
Siamese	cosine	No	<b>35.5 ± 1.4</b>	49.4 ± 1.3	<b>25.7 ± 1.1</b>	40.4 ± 1.3
Siamese	cosine	Yes	–	51.8 ± 1.4	–	40.8 ± 1.3

investigate only the unimodal speech models that we proposed within our indirect matching framework (see Chapter §4) since they may be applied directly to this task. Table 7.1 shows 10- and 20-way 1- and 5-shot speech classification results. The neural network models outperform the DTW direct feature matching baseline. Without fine-tuning, the Siamese speech model performs better compared to the baseline models. However, the classifier model with fine-tuned feature embeddings shows some gains without too much overfitting, achieving the best overall performance on the 10-way tasks. This is less clear for the 20-way tasks where the difference between models is marginal and Siamese models have a slight edge. We note again that we cannot fine-tune the Siamese speech model in the one-shot case.

### 7.4.2 *One-Shot Image Classification*

Next, we consider unimodal one-shot image classification. Once again, we investigate only the models proposed within our indirect matching for one-shot learning complex natural images. Table 7.2 shows 10- and 20-way 1- and 5-shot image classification results. While our direct feature pixel matching baseline performs better than chance, it is outperformed by the neural network models. Here it is not clear which of the neural network models is best overall. The Siamese model performs best on one-shot tasks, beating the classifier model with and without fine-tuning. Meanwhile for five-shot tasks, the classifier model shows increased benefit from fine-tuning. As discussed previously (§7.3.2), we include results for an oracle vision model which was trained for state-of-the-art results ImageNet classification. Without any further training on our background data, this model achieves the best results overall. The ImageNet data contains the classes used here for one-shot learning and this model has an unfair advantage. Yet, the results are still reasonably close to that achieved by our true one-shot learner models. It is clear that one-shot learning of the complex natural images is a challenging task. Accuracies are also overall lower in comparison to one-shot speech classification. To our knowledge, one-shot learning of the Flickr 8K images has not been considered in other studies on unimodal one-shot learning and is clearly a challenging problem. As with our digits experiments, we note the performance of these models influences the results of cross-modal matching within the indirect matching framework for the multimodal case, where these vision models are paired with the speech models seen before.

### 7.4.3 *One-Shot Cross-Modal Matching of Speech to Images*

We now consider the MOONSHOT problem setting. Table 7.3 shows 10- and 20-way 1- and 5-shot cross-modal matching of isolated spoken words to natural images sampled from the paired Flickr Audio and Flickr 8K benchmark dataset. Here the Siamese vision and speech models within the indirect matching framework achieve the best results on one-shot tasks, while the fine-tuned joint audio-visual MuMAML model which directly matches speech to images performs best on five-shot tasks. All of the neural network models perform better than the DTW and pixels direct feature matching baseline, which itself is only marginally better than a random baseline at one-shot matching speech to images. We included results for the oracle vision model paired with the classification speech model discussed previously. Although this model has an unfair advantage, MuMAML narrowly manages to outperform it on the ten-way five-shot task. The results for the DAVENet and MuMAML models without any fine-tuning are zero-shot since these models may compare speech to images without seeing any learning samples. Both zero-shot results are better than the direct feature matching baseline, demonstrating that these models encode a general prior relevant for transfer learning on new unseen tasks. Specifically, when these models are fine-tuned on one-shot learning samples they result in large improvements for one- and five-shot learning compared to zero-shot learning. This also

demonstrates less overfitting than in the case of unimodal one-shot learning models (§7.4.1 and §7.4.2).

Although the Siamese model achieves the best results on one-shot tasks, this model and the classifier model require supervised background data, with the vision branch of these models trained on the large extended background dataset. Meanwhile, the direct matching DAVenet and MuMAML models are trained on a smaller paired speech and image background dataset. These end-to-end architectures could potentially be improved if a larger multimodal background dataset is used for training.<sup>4</sup> We note again that DAVenet has the additional advantage that it requires only weakly supervised background data without any labels. Although, in this case, we train the DAVenet and MuMAML models on top of the pre-trained speech and vision base networks which do make use of supervised data for pre-training. We trained these models in this way since we only have a small training dataset. However, it is still possible to train such networks from scratch with enough data, as shown by [Harwath \*et al.\* \(2018\)](#). The DAVenet and MuMAML model also show large improvements in the case of five-shot tasks, outperforming our other models. This demonstrates that the direct matching models which may be fine-tuned on multimodal support sets take full advantage of the learning samples, scaling better as more examples become available. The best accuracies on each of the MOONSHOT tasks are lower than the accuracies on the respective unimodal one-shot speech classification (see Table 5.1) and unimodal one-shot image classification (see Table 7.2) tasks. We noted previously that this is due to the compounding errors of the models in the indirect matching framework. We observe this effect again in the case of the direct matching models simply due to the difficulty of the MOONSHOT problem in the complex natural setting.

#### 7.4.4 Analysis of Speaker Invariance

In the final investigation of our proposed MOONSHOT models, we consider how invariant these models are to the specific query speaker in the MOONSHOT setting. As in our digits benchmark experiments (§5.4.4), we consider tasks where one user teaches an agent and another user tests the system. Specifically, we consider the extreme case where the matching item in the support set is the only item not coming from the query speaker. We would like our models to be invariant to the query speaker, such that words from the same speaker as the query in the support set (for the incorrect items) are acoustically more different than the correct word from a different speaker. We test this setting following the same procedure as before: we sample a support set where all spoken words are from the same speaker as the speech query, except for the one instance matching the query word which is produced by a different speaker. The spoken words from the same speaker as the speech query distract from the true match and effective models should be invariant to these speakers. Table 7.4

---

<sup>4</sup>We do not perform ablation experiments to test this hypothesis since collecting additional paired speech-image examples to increase the size of the multimodal background dataset is a long and costly process. Instead we motivate this as a potential direction for future work.

Table 7.3: 10- and 20-way 1- and 5-shot cross-modal matching of isolated spoken words to natural images sampled from the paired Flickr Audio and Flickr 8K benchmark dataset for evaluating MOONSHOT.

Cross-Modal Model	Match	Fine-Tune	10-way Accuracy		20-way Accuracy	
			1-shot (%)	5-shot (%)	1-shot (%)	5-shot (%)
Random	random	–	$9.7 \pm 0.8$	–	$5.1 \pm 0.6$	–
DTW + Pixels	cosine	–	$12.6 \pm 0.9$	$12.9 \pm 0.9$	$6.6 \pm 0.6$	$7.3 \pm 0.7$
Oracle	cosine	–	$37.2 \pm 1.3$	$39.7 \pm 1.3$	$27.1 \pm 1.3$	$28.4 \pm 1.2$
Classifiers	cosine	–	$31.6 \pm 1.3$	$32.3 \pm 1.3$	$20.8 \pm 1.1$	$21.7 \pm 1.1$
Siamese	cosine	–	<b><math>32.1 \pm 1.3</math></b>	$33.2 \pm 1.3$	<b><math>21.4 \pm 1.1</math></b>	$23.0 \pm 1.1$
DAVEnet	cosine	No <sup>†</sup>	$14.5 \pm 1.0$	–	$9.5 \pm 0.8$	–
DAVEnet	cosine	Yes	$26.9 \pm 1.3$	$37.5 \pm 1.4$	$16.6 \pm 1.0$	$25.2 \pm 1.2$
MuMAML	cosine	No <sup>†</sup>	$18.1 \pm 1.1$	–	$11.8 \pm 0.9$	–
MuMAML	cosine	Yes	$29.1 \pm 1.3$	<b><math>40.3 \pm 1.3</math></b>	$17.1 \pm 1.0$	<b><math>26.4 \pm 1.2</math></b>

<sup>†</sup> The results shown for DAVEnet and MuMAML without fine-tuning are zero-shot.

Table 7.4: Speaker invariance tests for 10-way 1- and 5-shot cross-modal natural speech-image matching. All support set items are from the same speaker as the speech query, except for the support set item actually matching the query.

Cross-Modal Model	Match	Fine-Tune	10-way Accuracy	
			1-shot (%)	5-shot (%)
Random	random	–	$9.7 \pm 0.8$	–
DTW + Pixels	cosine	–	$12.9 \pm 1.1$	$12.7 \pm 1.0$
Classifiers	cosine	–	$28.6 \pm 2.0$	$30.3 \pm 1.9$
Siamese	cosine	–	<b><math>30.2 \pm 2.2</math></b>	$33.0 \pm 2.1$
DAVEnet	cosine	No <sup>†</sup>	$10.9 \pm 1.1$	–
DAVEnet	cosine	Yes	$25.9 \pm 1.9$	$37.3 \pm 2.1$
MuMAML	cosine	No <sup>†</sup>	$14.9 \pm 1.5$	–
MuMAML	cosine	Yes	$26.0 \pm 2.0$	<b><math>38.8 \pm 2.1</math></b>

shows results for 10-way 1- and 5-shot cross-modal natural speech-image matching under this extreme setting. While most of the models experience a drop in accuracy compared to the results in Table 7.3 (first two columns), the decrease is small and the features learned by the neural network models are reasonably invariant of the

specific speaker, generalising to other speakers. Interestingly, the drop is smallest for the direct matching DAVenet model adapted on the multimodal support set and the indirect matching Siamese vision and speech models. Both of these models directly optimise for an effective metric space and this appears to capture fine-grained word similarities.

## 7.5 CHAPTER SUMMARY

This chapter introduced and investigated a new complex benchmark dataset for multimodal one-shot learning containing natural images paired with spoken words. In addition we developed a related multimodal background dataset not containing the one-shot learning classes which may be used to train neural network models for this setting. We compared several models within two frameworks, one that indirectly matches speech to images using unimodal comparisons and the other that directly matches speech to images by jointly modelling speech and images. Specifically, for indirect matching, we compared transfer learning approaches with both classification and Siamese neural networks for learning separate unimodal networks which may be extended to the multimodal case, where the latter performed better on both unimodal (§7.4.1 and §7.4.2) and multimodal (§7.4.3) one-shot tasks by directly optimising for effective comparisons. We then compared these models to multimodal transfer learning approaches for direct matching that directly optimise for effectively comparing speech and images. Although the indirect matching models performed better on the one-shot tasks, the direct matching models showed best results on five-shot tasks. We discussed that this is due to the advantage that these models may directly update their parameters with a few-steps of gradient descent on the multimodal support set, resulting in better scaling as more examples become available. Specifically, our deep audio-visual embedding network trained with a novel multimodal model-agnostic meta-learning algorithm for fast adaptation on a few weakly supervised examples achieved our best results with a cross-modal matching accuracy of 40.3% for 10-way 5-shot learning and 26.4% for 20-way 5-shot learning. Finally, we demonstrate that all of the neural network models we have proposed are reasonably invariant to the specific speaker.

## 8 | CONCLUSION

### 8.1 SUMMARY AND CONCLUSIONS

This thesis introduced and formalised multimodal one-shot learning, specifically for learning spoken words and visual objects. Observing only one paired speech-image example from each class, a model is asked to pick the correct image for an unseen spoken query. We proposed two benchmark datasets for this task: a simple benchmark comprised of spoken and visual digits and a more complex benchmark comprised of natural images paired with spoken words. To accomplish this task we proposed models within two frameworks, one which indirectly matches speech to images and another that aims to improve on the former by directly matching speech to images. We proposed and evaluated several baseline and more advanced models within these frameworks. We show that a metric learning approach to transfer learning using unimodal Siamese neural networks achieves impressive results on the simple benchmark with a cross-modal matching accuracy of 70.12% for 11-way 1-shot learning. We then show that the unimodal Siamese neural networks and multimodal deep audio-visual embedding networks achieve our most competitive results on the more complex benchmark tasks. Our overall best model learns how to learn from few examples of speech-image pairs within a meta-learning framework for directly matching speech to images. On our most difficult benchmark, this model achieved a cross-modal matching accuracy of 40.3% for 10-way 5-shot learning and 26.4% for 20-way 5-shot learning. We have shown that this problem setting is extremely challenging for a machine learning system—speaking in absolute terms, these scores are low, showing that there is still much to be done. However, these accuracies do indicate that there is promise and that this task is possible.

### 8.2 FUTURE WORK

One speculated limitation that we faced in this work was a lack of sufficient multimodal paired data for training deep audio-visual embedding networks. This is especially true when training this model using our novel multimodal model-agnostic meta-learning algorithm which benefits from a large variety of tasks such that it may learn to learn. These end-to-end architectures for matching speech to images could potentially be improved if a larger multimodal background dataset is used for

training. Future work could investigate obtaining such a background dataset not containing the one-shot classes and evaluating whether this increases the performance of models on multimodal one-shot learning tasks. Another idea that could be investigated in future work is the use of pre-training techniques for the multimodal one-shot learning models, for example, on large unsupervised datasets of spoken words and visual objects without weak supervision in the form of co-occurring context. Even if the unsupervised datasets contain the one-shot classes, there would be no supervisory signal for directly learning these concepts. This may be similar to how humans learn, where young children observe many new visual objects and spoken words in their environment. Only a few such instances are linked by directly co-occurring context—this is where multimodal one-shot learning might occur as children may quickly learn the correspondence between the audio and visual signal by building on prior experience. Extending meta-learning algorithms to the purely weakly supervised case could also be investigated in future work, since our approach requires labelled paired multimodal training data. Finally, by motivating and formalising the multimodal one-shot learning task and creating benchmark datasets with several baseline results, we hope to encourage future work on this challenging problem; the goal of which is to narrow the gap on achieving agents with a more general artificial intelligence, exhibiting the ability to learn new concepts such as spoken words and visual objects from limited data, similar to humans.

\* \* \*

*The real goal of AI is to understand and build devices that can perceive, reason, act, and learn at least as well as we can.*

— Astro Teller

PART IV  
APPENDICES

# A | TIDIGITS AND MNIST EXPERIMENTS

## A.1 ANALYSIS OF SPEAKER INVARIANCE IN ONE-SHOT SPEECH CLASSIFICATION

Table A.1: Speaker invariance tests for 11-way 1-shot speech classification. All support set items are from the same speaker as the query, except for the support set item actually matching the query.

Speech Model	11-way Accuracy
	1-shot
DTW	53.27% $\pm$ 2.87
FFNN Classifier	67.91% $\pm$ 4.05
CNN Classifier	78.09% $\pm$ 3.30
Siamese CNN (offline)	88.77% $\pm$ 1.24
Siamese CNN (online)	<b>92.77% <math>\pm</math> 1.11</b>



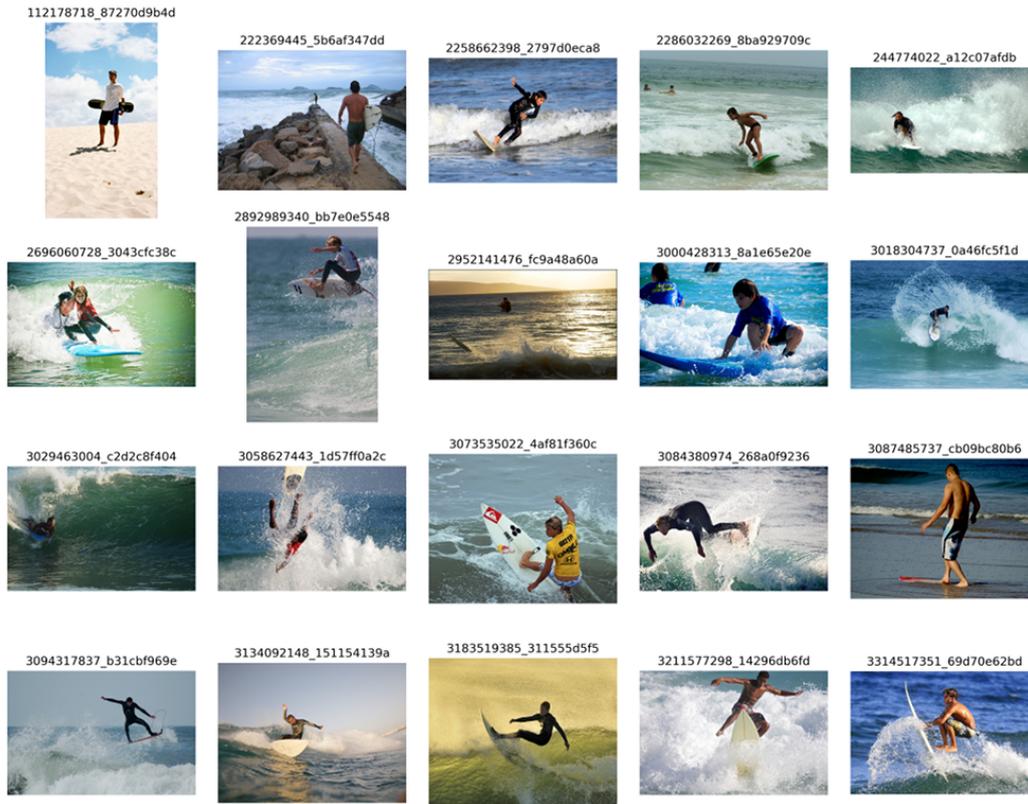


Figure B.2: Excerpt of images paired with spoken word “surfboard”.

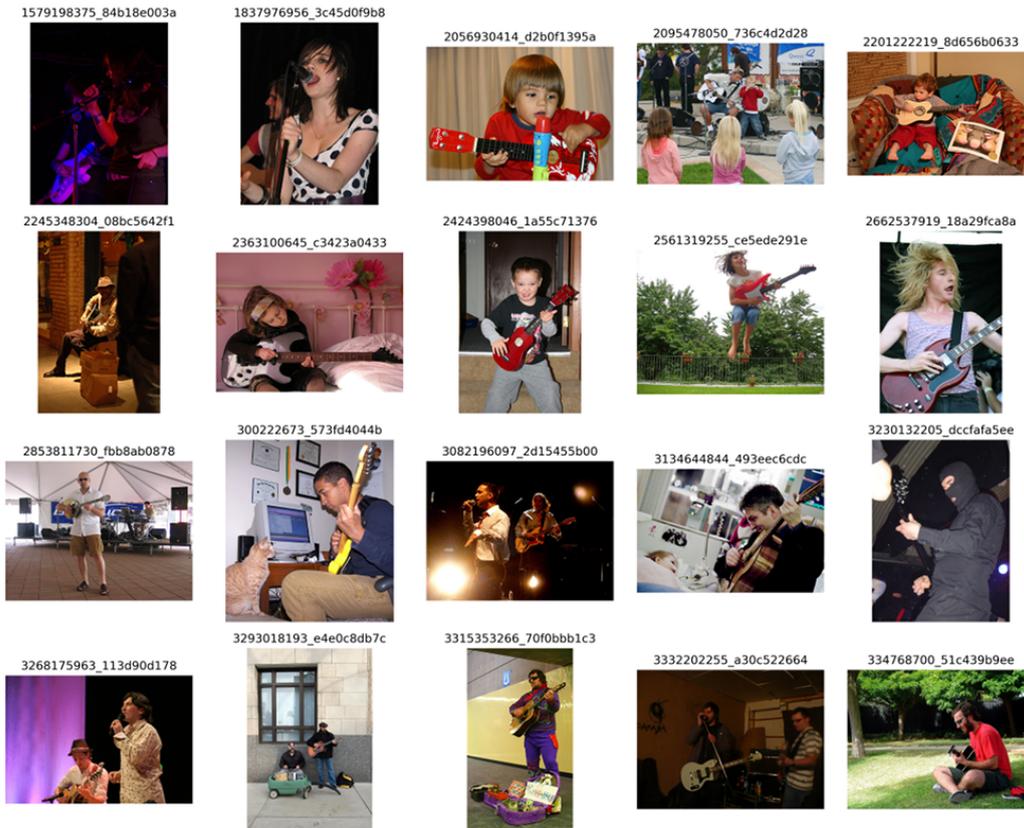


Figure B.3: Excerpt of images paired with spoken word “guitar”.

## B.2 FLICKR 8K AND FLICKR AUDIO BACKGROUND DATASET KEYWORD CLASSES

adult air arm baby background backpack bag ball bar baseball beach  
bed bicycle bike biker bite black blue boy brick brown bubble building  
camera cap carry catch chair chase child city climb coat collar costume  
couple cover cross crowd dirt dog dress drink eat face fall fence  
field fight flag flip fly football fountain frisbee game girl glass  
grass grassy green greyhound group hand hang hat head helmet high hike  
hiker hit hockey hold ice jacket jean jump kick kid lake large laugh  
lay leap leash leg line little look man mountain mouth mud near night  
ocean old orange outside paint pant park people person picture play  
player playground playing point pole pool pose puppy push race rail  
railing ramp read red ride rider river road rock rocky run shirt short  
sidewalk sign sit ski skier slide small smile snow snowboarder snowy  
soccer stair stand step stick street striped sunglass surf surfer swim  
swimming swing table take talk tan team tennis track tree trick trunk  
try walk wall watch wave wear white window woman wood wooden yard yellow  
young

## REFERENCES

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems.  
Available at: <http://tensorflow.org/>
- Abbott, B. (2010). *Reference*. Oxford University Press, Oxford, UK.
- Bahdanau, D., Cho, K. and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In: *Proceedings of International Conference on Learning Representations*.
- Bengio, Y. and LeCun, Y. (2007). Scaling learning algorithms towards AI. In: *Proceedings of Large-Scale Kernel Machines*.
- Besacier, L., Barnard, E., Karpov, A. and Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, vol. 56, pp. 85–100.
- Bromley, J., Guyon, I., LeCun, Y., Säckinger, E. and Shah, R. (1994). Signature verification using a “Siamese” time delay neural network. In: *Proceedings of Advances in Neural Information Processing Systems*, vol. 6, pp. 737–744. Morgan-Kaufmann.
- Carey, S. (1978). The child as word learner. In: Halle, M., Bresnan, J. and Miller, G. (eds.), *Linguistic theory and psychological reality*, pp. 264–293. Cambridge, MA: MIT Press.
- Carey, S. and Bartlett, E. (1978). Acquiring a single new word. In: *Proceedings of the Stanford Child Language Conference*.
- Caruana, R. (1997). Multitask learning. *Machine Learning*, vol. 28, no. 1, pp. 41–75.
- Chechik, G., Sharma, V., Shalit, U. and Bengio, S. (2010). Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, vol. 11, pp. 1109–1135.
- Chopra, S., Hadsell, R. and LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.

- Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E. and Darrell, T. (2014). Decaf: A deep convolutional activation feature for generic visual recognition. In: *Proceedings of International Conference on Machine Learning*.
- Eisenschtat, A. and Wolf, L. (2017). Linking image and text with 2-way nets. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.
- Eloff, R., Engelbrecht, H.A. and Kamper, H. (2019). Multimodal one-shot learning of speech and images. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*.
- Fazly, A., Alishahi, A. and Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, vol. 34, no. 6, pp. 1017–1063.
- Fei-Fei, L., Fergus, R. and Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 594–611.
- Finn, C., Abbeel, P. and Levine, S. (2017a). Model-agnostic meta-learning for fast adaptation of deep networks. In: *Proceedings of International Conference on Machine Learning*.
- Finn, C. and Levine, S. (2018). Meta-learning and universality: Deep representations and gradient descent can approximate any learning algorithm. In: *Proceedings of International Conference on Learning Representations*.
- Finn, C., Yu, T., Zhang, T., Abbeel, P. and Levine, S. (2017b). One-shot visual imitation learning via meta-learning. In: *Proceedings of Annual Conference on Robot Learning*.
- Finn, C.B. (2018). *Learning to Learn with Gradients*. Ph.D. thesis, University of California, Berkeley.
- Frank, M.C., Goodman, N.D. and Tenenbaum, J.B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, vol. 20, no. 5, pp. 578–585.
- French, R.M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, vol. 3, no. 4, pp. 128–135.
- Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M. and Mikolov, T. (2013). DeViSE: A deep visual-semantic embedding model. In: *Proceedings of Advances in Neural Information Processing Systems*, vol. 26, pp. 2121–2129. Curran Associates, Inc.
- Gella, S., Sennrich, R., Keller, F. and Lapata, M. (2017). Image pivoting for learning multilingual multimodal representations. In: *Proceedings of Empirical Methods in Natural Language Processing*.
- Gu, J., Wang, Y., Chen, Y., Li, V.O.K. and Cho, K. (2018). Meta-learning for low-resource neural machine translation. In: *Proceedings of Empirical Methods in Natural Language Processing*.

- Hadsell, R., Chopra, S. and LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.
- Halberda, J. (2006). Is this a dax which I see before me? Use of the logical argument disjunctive syllogism supports word-learning in children and adults. *Cognitive Psychology*, vol. 53, no. 4, pp. 310–344.
- Harwath, D. and Glass, J.R. (2015). Deep multimodal semantic embeddings for speech and images. In: *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*.
- Harwath, D. and Glass, J.R. (2017). Learning word-like units from joint audio-visual analysis. In: *Proceedings of Annual Meeting of the Association for Computational Linguistics*.
- Harwath, D., Hsu, W.-N. and Glass, J. (2020). Learning hierarchical discrete linguistic units from visually-grounded speech. In: *Proceedings of International Conference on Learning Representations*.
- Harwath, D., Recasens, A., Surís, D., Chuang, G., Torralba, A. and Glass, J. (2018). Jointly discovering visual objects and spoken words from raw sensory input. In: *Proceedings of The European Conference on Computer Vision*.
- Harwath, D., Recasens, A., Surís, D., Chuang, G., Torralba, A. and Glass, J. (2019). Jointly discovering visual objects and spoken words from raw sensory input. *International Journal of Computer Vision*, pp. 1–22.
- Harwath, D., Torralba, A. and Glass, J.R. (2016). Unsupervised learning of spoken language with visual context. In: *Proceedings of Advances in Neural Information Processing Systems*, vol. 29, pp. 1858–1866. Curran Associates, Inc.
- He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.
- Hermann, K.M. and Blunsom, P. (2014). Multilingual distributed representations without word alignment. In: *Proceedings of International Conference on Learning Representations*.
- Hermans, A., Beyer, L. and Leibe, B. (2017). In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.
- Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N. and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97.
- Hodosh, M., Young, P. and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, vol. 47, no. 1, pp. 853–899.

- Hoffer, E. and Ailon, N. (2015). Deep metric learning using triplet network. In: *Proceedings of International Workshop on Similarity-Based Pattern Analysis and Recognition*.
- Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.
- Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: *Proceedings of International Conference on Machine Learning*.
- Kamper, H., Shakhnarovich, G. and Livescu, K. (2019). Semantic speech retrieval with a visually grounded model of untranscribed speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 89–98.
- Kashyap, K. (2017). *Learning digits via joint audio-visual representations*. Master's thesis, MIT.
- Kingma, D. and Ba, J. (2015). Adam: A method for stochastic optimization. In: *Proceedings of International Conference on Learning Representations*.
- Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quaa, J., Ramalho, T., Grabska-Barwinski, A., Hassabis, D., Clopath, C., Kumaran, D. and Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of National Academy of Sciences*, vol. 114, no. 3, pp. 3521–3526.
- Koch, G., Zemel, R. and Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. In: *Proceedings of International Conference on Machine Learning*.
- Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. In: *Proceedings of Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105. Curran Associates, Inc.
- Lake, B.M., Lee, C.-Y., Glass, J.R. and Tenenbaum, J.B. (2014). One-shot learning of generative speech concepts. In: *Proceedings of Annual Meeting of the Cognitive Science Society*.
- Lake, B.M., Linzen, T. and Baroni, M. (2019). Human few-shot learning of compositional instructions. In: *Proceedings of Annual Meeting of the Cognitive Science Society*.
- Lake, B.M., Salakhutdinov, R., Gross, J. and Tenenbaum, J.B. (2011). One shot learning of simple visual concepts. In: *Proceedings of Annual Meeting of the Cognitive Science Society*.
- Lake, B.M., Salakhutdinov, R. and Tenenbaum, J.B. (2013). One-shot learning by inverting a compositional causal process. In: *Proceedings of Advances in Neural Information Processing Systems*, vol. 26, pp. 2526–2534. Curran Associates, Inc.
- Lake, B.M., Salakhutdinov, R. and Tenenbaum, J.B. (2015). Human-level concept learning through probabilistic program induction. *Science*, vol. 350, no. 6266, pp. 1332–1338.

- Lazaridou, A., Bruni, E. and Baroni, M. (2014). Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In: *Proceedings of Annual Meeting of the Association for Computational Linguistics*.
- LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324.
- Leidal, K., Harwath, D. and Glass, J.R. (2017). Learning modality-invariant representations for speech and images. In: *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*.
- Leonard, R.G. and Doddington, G. (1993). TIDIGITS LDC93S10. Philadelphia: Linguistic Data Consortium.  
Available at: <https://catalog.ldc.upenn.edu/LDC93S10/>
- Lin, T., Goyal, P., Girshick, R., He, K. and Dollár, P. (2017). Focal loss for dense object detection. In: *Proceedings of IEEE International Conference on Computer Vision*.
- Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L. and Dollár, P. (2014). Microsoft COCO: Common objects in context. [arXiv:1405.0312](https://arxiv.org/abs/1405.0312).
- Markson, L. and Bloom, P. (1997). Evidence against a dedicated system for word learning in children. *Nature*, vol. 385, pp. 813–815.
- Mishra, N., Rohaninejad, M., Chen, X. and Abbeel, P. (2018). A simple neural attentive meta-learner. In: *Proceedings of International Conference on Learning Representations*.
- Mueen, A. and Keogh, E. (2016). Extracting optimal performance from dynamic time warping. In: *Proceedings of the International Conference on Knowledge Discovery and Data Mining*.
- Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H. and Ng, A.Y. (2011). Multimodal deep learning. In: *Proceedings of International Conference on Machine Learning*.
- Palatucci, M., Pomerleau, D., Hinton, G.E. and Mitchell, T.M. (2009). Zero-shot learning with semantic output codes. In: *Proceedings of Advances in Neural Information Processing Systems*, vol. 22, pp. 1410–1418. Curran Associates, Inc.
- Paperno, D., Kruszewski, G., Lazaridou, A., Pham, N.Q., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G. and Fernández, R. (2016). The LAMBADA dataset: Word prediction requiring a broad discourse context. In: *Proceedings of Annual Meeting of the Association for Computational Linguistics*.
- Pascanu, R., Mikolov, T. and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In: *Proceedings of International Conference on Machine Learning*.
- Räsänen, O. and Rasilo, H. (2015). A joint model of word segmentation and meaning acquisition through cross-situational learning. *Psychological Review*, vol. 122, no. 4, pp. 792–829.

- Rasiwasia, N., Pereira, J.C., Coviello, E., Doyle, G., Lanckriet, G.R., Levy, R. and Vasconcelos, N. (2010). A new approach to cross-modal multimedia retrieval. In: *Proceedings of ACM international conference on Multimedia (ACMMM)*.
- Ren, M., Ravi, S., Triantafillou, E., Snell, J., Swersky, K., Tenenbaum, J.B., Larochelle, H. and Zemel, R.S. (2018). Meta-learning for semi-supervised few-shot classification. In: *Proceedings of International Conference on Learning Representations*.
- Renkens, V. and Van hamme, H. (2017). Weakly supervised learning of hidden Markov models for spoken language acquisition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 285–295.
- Renkens, V. and Van hamme, H. (2018). Capsule networks for low resource spoken language understanding. In: *Proceedings of Annual Conference of the International Speech Communication Association*.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C. and Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, pp. 211–252.
- Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49.
- Salakhutdinov, R. and Hinton, G. (2007). Learning a nonlinear embedding by preserving class neighbourhood structure. In: *Proceedings of the International Conference on Artificial Intelligence and Statistics*.
- Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D. and Lillicrap, T. (2016). Meta-learning with memory-augmented neural networks. In: *Proceedings of International Conference on Machine Learning*.
- Schroff, F., Kalenichenko, D. and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.
- Settle, S., Levin, K., Kamper, H. and Livescu, K. (2017). Query-by-example search with discriminative neural acoustic word embeddings. In: *Proceedings of Annual Conference of the International Speech Communication Association*.
- Shyam, P., Gupta, S. and Dukkipati, A. (2017). Attentive recurrent comparators. In: *Proceedings of International Conference on Machine Learning*.
- Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In: *Proceedings of International Conference on Learning Representations*.
- Snell, J., Swersky, K. and Zemel, R. (2017). Prototypical networks for few-shot learning. In: *Proceedings of Advances in Neural Information Processing Systems*, vol. 30, pp. 4077–4087. Curran Associates, Inc.

- Socher, R., Ganjoo, M., Manning, C.D. and Ng, A. (2013). Zero-shot learning through cross-modal transfer. In: *Proceedings of Advances in Neural Information Processing Systems*, vol. 26, pp. 935–943. Curran Associates, Inc.
- Song, H.O., Xiang, Y., Jegelka, S. and Savarese, S. (2016). Deep metric learning via lifted structured feature embedding. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.
- Strivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958.
- Stafylakis, T. and Tzimiropoulos, G. (2018). Zero-shot keyword spotting for visual speech recognition in-the-wild. In: *Proceedings of The European Conference on Computer Vision*.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.
- Taniguchi, T., Nagai, T., Nakamura, T., Iwahashi, N., Ogata, T. and Asoh, H. (2016). Symbol emergence in robotics: A survey. *Advanced Robotics*, vol. 30, no. 11–12, pp. 706–728.
- Thomason, W. and Knepper, R.A. (2016). Recognizing unfamiliar gestures for human-robot interaction through zero-shot learning. In: *Proceedings of International Symposium on Experimental Robotics*.
- Trueswell, J.C., Medina, T.N., Hafri, A. and Gleitman, L.R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, vol. 66, no. 1, pp. 126–156.
- Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K. and Wierstra, D. (2016). Matching networks for one shot learning. In: *Proceedings of Advances in Neural Information Processing Systems*, vol. 29, pp. 3630–3638. Curran Associates, Inc.
- Walter, M.R., Friedman, Y., Anton, M. and Teller, S. (2012). One-shot visual appearance learning for mobile manipulation. *International Journal of Robotics Research*, vol. 31, no. 4, pp. 554–567.
- Wang, J., song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B. and Wu, Y. (2014). Learning fine-grained image similarity with deep ranking. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.
- Wang, L., Li, Y., Huang, J. and Lazebnik, S. (2018). Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 394–407.
- Weston, J., Chopra, S. and Bordes, A. (2015). Memory networks. In: *Proceedings of International Conference on Learning Representations*.

- Wu, D., Zhu, F. and Shao, L. (2012). One shot learning gesture recognition from RGBD images. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.
- Young, P., Lai, A., Hodosh, M. and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In: *Proceedings of Annual Meeting of the Association for Computational Linguistics*.
- Yu, T., Finn, C., Dasari, S., Xie, A., Zhang, T., Abbeel, P. and Levine, S. (2018). One-shot imitation from observing humans via domain-adaptive meta-learning. In: *Proceedings of Robotics: Science and Systems*.