INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Direct multimodal few-shot learning of speech and images

Leanne Nortje, Herman Kamper

Department of E&E Engineering, Stellenbosch University, South Africa
nortjeleanne@gmail.com, kamperh@sun.ac.za

Abstract

We propose direct multimodal few-shot models that learn a shared embedding space of spoken words and images from only a few paired examples. Imagine an agent is shown an image along with a spoken word describing the object in the picture, e.g. *pen*, *book* and *eraser*. After observing a few paired examples of each class, the model is asked to identify the "book" in a set of unseen pictures. Previous work used a two-step indirect approach relying on speech-speech and image-image comparisons across the support set of given speech-image pairs. Instead, we propose two direct models which learn a single multimodal space where inputs from different modalities are directly comparable: a multimodal triplet network (MTriplet) and a multimodal correspondence autoencoder (MCAE). To train these direct models, we *mine* speech-image pairs by using the support set to pair up unlabelled in-domain speech and images. In a speech-to-image digit matching task, direct models outperform indirect models, with the MTriplet achieving the best multimodal five-shot accuracy. We show that the improvements are due to the combination of unsupervised and transfer learning in the direct models, and the absence of two-step compounding errors.

Index Terms: few-shot learning, multimodal modelling, unsupervised learning, transfer learning, speech and images.

1. Introduction

Current audio and vision systems require large amounts of labelled data which is expensive and time-consuming to collect. In contrast, young children are able to learn new words and objects from only a few examples [1–5]. In fact, from only one exposure, infants can learn the word for a shown (visual) object [6]. Can we emulate this in a machine? Imagine an agent is shown images of a *duck*, *horse* and *chicken*, where each image is paired with a spoken word describing the object. After observing a few example pairs from each class, the agent is prompted to identify the visual instance corresponding to the spoken word "horse" from a set of unseen images. This task was formalised in [7]: *multimodal few-shot learning* is the task of learning new concepts from a few paired examples, where each pair consists of two items of the same concept but in different modalities.

Previous work [7, 8] used a two-step *indirect* approach: a spoken query is compared to the spoken examples in the given *support set* of speech-image pairs, and the corresponding image is then used to select the closest item in the unseen *matching set*. The task is therefore reduced to two unimodal comparisons, with the support set acting as a pivot between the modalities. To do the unimodal comparisons, learned speech and image representations are used: [8] compared *transfer learning* to *unsupervised learning*. In transfer learning, representations are used from models trained on labelled background data (not containing any classes seen at test time). The motivation for this is that humans can call on prior knowledge when learning new concepts. In unsupervised modelling, models are trained on unlabelled in-domain data. This is reasonable since, before being shown paired examples, an agent could be exposed to a large amount of unlabelled speech and visual data from its environment.

In this paper we propose combining unsupervised and transfer learning to obtain *direct* models that learn a single multimodal embedding space where observations from either modality can be directly compared, i.e. word and image instances of the same class are mapped to similar representations in a shared space. Training such a multimodal speech-image space has been considered in several studies [9–16]. But they all train models on a large set of paired speech-image examples for a large number of classes. We instead focus on how a multimodal space can be learned from only a few examples for a limited set of classes.

To train direct models, we require matched speech-image pairs. Since the given pairs in the support set is not sufficient to train models directly, we propose to *mine* speech-image training pairs. Spoken/visual items in the support set are compared to unlabelled examples in the respective modalities. New training speech-image pairs are then constructed by using the support set as a pivot. We are therefore making use of unlabelled in-domain data (a form of unsupervised learning). But to do the speech-speech and image-image comparisons in the mining scheme, we actually use representations from transfer-learned speech or vision classifiers (trained on background labelled data). Direct modelling thus combines unsupervised and transfer learning.

We use the mined cross-modal pairs to train two direct models: a multimodal triplet network (MTriplet), which combines two unimodal triplet losses, and a novel multimodal correspondence autoencoder (MCAE), which combines two unimodal correspondence autoencoders (CAEs). A CAE attempts to produce another instance of the same class as the input through a bottleneck layer [17]. The MCAE is similar to other multimodal models, although previous work mostly used autoencoders (AEs) and worked in other modalities [18–23]. The MTriplet is similar to the models of [9–12], but we are first to use these architectures in a multimodal few-shot learning setting. [24, 25] used images with text captions on a similar multimodal few-shot task.

In a speech-to-image digit matching task, we compare our direct models to the best indirect transfer-learned and unsupervised models from [8]. Both direct models outperform the indirect models, with the MTriplet as the top performer. Ablation experiments show that the direct models' superior performance can be attributed to the way in which they combine unsupervised and transfer learning, and because direct matching does not cause a compounding of errors as in the two-step indirect methods.[1]

2. Task: Multimodal few-shot speech-to-image matching

The task of multimodal few-shot matching of speech and images is illustrated in Figure 1 (i). We are given an unseen speech

[1]We release source code at: https://github.com/LeanneNortje/direct_multimodal_few-shot_learning.

http://dx.doi.org/10.21437/Interspeech.2021-49

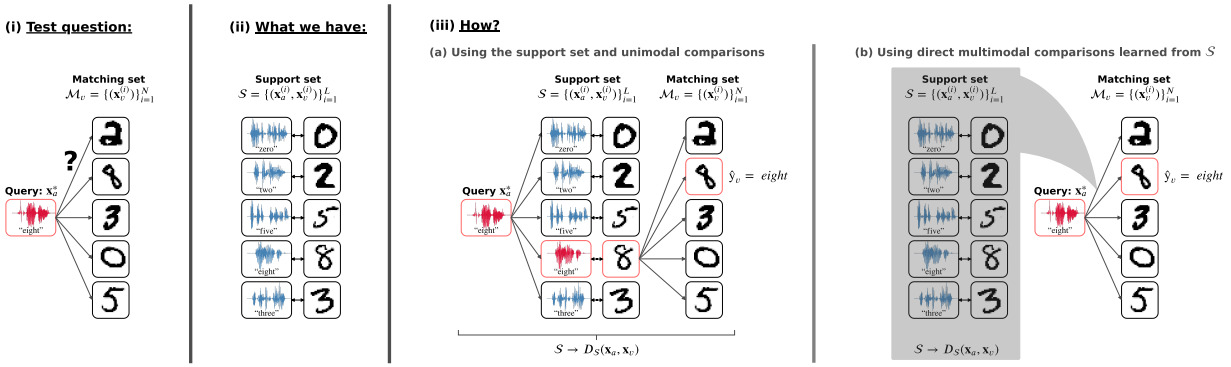

Figure 1: *(i) The multimodal one-shot speech-to-image matching question shown at test time. (ii) The support set given to solve the task. (iii) The support set can be used in two ways: (a) an indirect matching approach [7, 8] which uses two unimodal comparisons across the support set, and (b) the direct matching approach proposed in this paper.*

query $\mathbf{x}_a^*$ and prompted to identify the corresponding image in a matching set $\mathcal{M}_v = \{(\mathbf{x}_v^{(i)})\}_{i=1}^N$ of unseen test images. To do this, we are given a small number of pairs in a *support set* $\mathcal{S}$, where each pair consists of an isolated spoken word $\mathbf{x}_a^{(i)}$ and a corresponding image $\mathbf{x}_v^{(i)}$. Neither the test-time speech query $\mathbf{x}_a^*$ nor the matching set $\mathcal{M}_v$ images occur in the support set. For the *one-shot* case shown in Figure 1 (ii), $\mathcal{S}$ consists of one example pair for each of the L classes. Multimodal one-shot matching can be extended to K-*shot* matching: for L-way K-shot matching, the support set $\mathcal{S} = \{(\mathbf{x}_a^{(i)}, \mathbf{x}_v^{(i)})\}_{i=1}^{L \times K}$ contains K paired speech-image examples for each of the L classes.

To perform this matching task, we need a distance metric $D_\mathcal{S}(\mathbf{x}_a, \mathbf{x}_v)$ between inputs from the two modalities. Previous work, which we use as a baseline (§3), constructed an indirect distance metric while we propose a direct method (§4).

3. Baseline: Indirect multimodal matching

Indirect matching uses two unimodal comparisons across the support set as its metric $D_\mathcal{S}(\mathbf{x}_a, \mathbf{x}_v)$, as shown in Figure 1 (iii) (a). A spoken query $\mathbf{x}_a^*$ is encoded into some representation and then compared to representations for each speech instance $\mathbf{x}_a^{(i)}$ in $\mathcal{S}$ to determine the query's closest spoken match based on the (cosine) distance between representations. The nearest neighbour's paired image is then compared to each image instance $\mathbf{x}_v^{(i)}$ in the matching set $\mathcal{M}_v$ to find its closest image based on the distance between image representations. The closest image is the model's prediction, e.g. in Figure 1 (iii) (a) the *eight* in $\mathcal{M}_v$.

This indirect approach requires unimodal representations that capture similarity within a modality. In [8], two methodologies were considered. The first is *transfer learning*, which is common in unimodal few-shot learning studies [26–32]. This involves training a model on a different but related dataset not containing any of the classes seen at test time [33, 34]. [8] specifically considered unimodal transfer-learned speech and vision classifiers, Siamese [35] triplet networks, and CAEs. All these models were trained on labelled background data not containing any instances of the few-shot classes seen at test time.

The second unimodal representation learning approach is *unsupervised learning*. Before an agent is shown paired examples of new classes, the agent could be exposed to unlabelled speech and visual data from its environment. Some of these unlabelled examples could correspond to the few-shot classes. Therefore, [8] also considered unsupervised unimodal speech and vision AEs and CAEs trained on unlabelled in-domain data.

4. Proposal: Direct multimodal matching

A direct multimodal few-shot learning model aims to learn a single multimodal embedding space which maps inputs of the same class, regardless of modality, to similar representations. As illustrated in Figure 1 (iii) (b), the multimodal space can be used as a direct cross-modal metric $D_\mathcal{S}(\mathbf{x}_a, \mathbf{x}_v)$, allowing a speech query's representation to be directly compared to the image representations in the matching set.

4.1. Cross-modal pair mining

In order to learn a multimodal embedding space, we need pairs of matching images and spoken words. The only pairs we have are those in the given support set $\mathcal{S}$. This small set is not sufficient for training a multimodal model. We therefore use $\mathcal{S}$ to *mine* more in-domain speech-image pairs from a larger set of unlabelled data, using $\mathcal{S}$ as a pivot between unlabelled data in the two modalities. This is similar to ideas used in [36, 37]. Consider the i^{th} speech-image pair in the support set $(\mathbf{x}_a^{(i)}, \mathbf{x}_v^{(i)})$. We find the images in the unlabelled in-domain image data whose closest image in the support set is $\mathbf{x}_v^{(i)}$. Similarly, for spoken words in unlabelled in-domain speech data, we find those for which $\mathbf{x}_a^{(i)}$ is the closest match. From these items matched to the i^{th} pair, we choose a random word and image instance and pair them up, thereby obtaining a novel training pair. Since we use no labels, all the pairs will not be correct; nevertheless, this procedure provides a large number of noisy training examples.

Similar to the indirect matching approach (§3), we need to do unimodal speech-speech and image-image comparisons to perform mining across $\mathcal{S}$. We employ transfer learning, specifically the best unimodal models from [8]: the transfer-learned speech and vision classifiers. For the support set and all the unlabelled speech and image items, we therefore extract features from an intermediate embedding (the layer before the softmax) and then use cosine distance to find the closest matches. Mining thus combines unsupervised learning (from the unlabelled in-domain data) and transfer learning (from background out-of-domain data). Our experiments (§6) show that using unimodal transfer-learned models is essential to the mining process.

We consider two multimodal architectures trained on these mined speech-image pairs: a novel MCAE and an MTriplet.

4.2. Multimodal correspondence autoencoder

The multimodal correspondence autoencoder (MCAE) aims to learn a multimodal space by using a modified correspondence

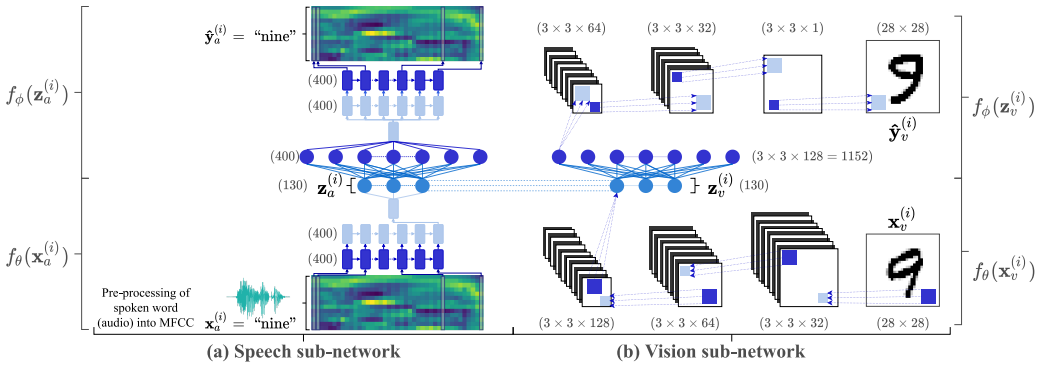

Figure 2: *The MCAE uses (a) an RNN encoder-decoder for processing spoken words, and (b) a CNN encoder-decoder for processing images. The MCAE loss is a combination of CAE losses in each modality and a multimodal loss encouraging the audio and visual latent representations* $\mathbf{z}_a^{(i)}$ *and* $\mathbf{z}_v^{(i)}$ *to be close to each other.*

autoencoder (CAE) loss. A standard CAE is trained to produce another instance $\mathbf{x}_{\text{pair}}^{(i)}$ of the same type as the input $\mathbf{x}^{(i)}$ through a bottleneck layer [17]. The CAE loss is $\ell = ||\mathbf{x}_{\text{pair}}^{(i)} - \hat{\mathbf{y}}^{(i)}||_2^2$, where $\hat{\mathbf{y}}^{(i)}$ is the network output.

The MCAE is a combination of a speech CAE and a vision CAE, as shown in Figure 2. Each CAE consists of an encoder and a decoder. The recurrent neural network (RNN) speech encoder $f_\theta(\mathbf{x}_a^{(i)})$ encodes input $\mathbf{x}_a^{(i)}$ to the representation $\mathbf{z}_a^{(i)}$. The RNN decoder $f_\phi(\mathbf{z}_a^{(i)})$ is conditioned on $\mathbf{z}_a^{(i)}$ to produce output $\hat{\mathbf{y}}_a^{(i)}$. Similar to the acoustic embedding models of [38–42], the encoder produces a fixed-sized representation $\mathbf{z}_a^{(i)}$ for variable duration input $\mathbf{x}_a^{(i)}$. The speech CAE's loss is $\ell_a = ||\mathbf{x}_{a_{\text{pair}}}^{(i)} - \hat{\mathbf{y}}_a^{(i)}||_2^2$. The vision CAE has a similar structure and loss ℓ_v, but uses a convolutional neural network (CNN) encoder and a transposed convolutional vision decoder instead of RNNs.

The MCAE links the speech and vision CAE's with a multimodal loss term that encourages similar latent representations for paired inputs: $\ell_z = ||\mathbf{z}_a^{(i)} - \mathbf{z}_v^{(i)}||_2^2$. The complete MCAE loss for a singular training example is $\ell = \alpha_a \ell_a + \alpha_v \ell_v + \alpha_z \ell_z$, where the α's are loss weights. Each training example therefore consists of $\mathbf{x}_a^{(i)}$, $\mathbf{x}_{a_{\text{pair}}}^{(i)}$, $\mathbf{x}_v^{(i)}$ and $\mathbf{x}_{v_{\text{pair}}}^{(i)}$, where $(\mathbf{x}_a^{(i)}, \mathbf{x}_v^{(i)})$ is a mined speech-image pair. For $\mathbf{x}_a^{(i)}$ we need the paired word $\mathbf{x}_{a_{\text{pair}}}^{(i)}$, and for $\mathbf{x}_v^{(i)}$ we need the paired image $\mathbf{x}_{v_{\text{pair}}}^{(i)}$. Again we do not have class labels for the unlabelled in-domain data, so we mine these unimodal pairs. The procedure is similar to that of the cross-modal mining scheme in §4.1, but here we mine pairs within the modality, so the support set is not used as a pivot. As noted in §1, the overall MCAE model is very similar to earlier AE-based models, such as the one from [22].

4.3. Multimodal triplet network

The multimodal triplet network (MTriplet) is based on the model of [9]. But here we apply it for the first time to multimodal few-shot learning. The MTriplet aims to learn a distance metric between spoken words and images in a single multimodal space by pushing paired cross-modal representations toward each other while pulling non-matching representations away from each other. The network consists of speech and vision encoders which maps inputs $\mathbf{x}^{(i)}$ to representations $\mathbf{z}^{(i)}$. In our case, these encoders have exactly the same architecture as the encoders of the MCAE shown in Figure 2. The model optimises the combination of two triplet losses [43–46] so that the distance between representations of paired cross-modal inputs $(\mathbf{x}_a^{(i)}, \mathbf{x}_v^{(i)})$ are smaller than the distance between representations of non-matched cross-modal pairs $(\mathbf{x}_a^{(i)}, \mathbf{x}_{v_{\text{neg}}}^{(i)})$ and $(\mathbf{x}_{a_{\text{neg}}}^{(i)}, \mathbf{x}_v^{(i)})$, by some margin. Stated mathematically, the loss is:

$$\ell = \max\left\{0, m + d(\mathbf{z}_a^{(i)}, \mathbf{z}_v^{(i)}) - d(\mathbf{z}_a^{(i)}, \mathbf{z}_{v_{\text{neg}}}^{(i)})\right\} + \max\left\{0, m + d(\mathbf{z}_a^{(i)}, \mathbf{z}_v^{(i)}) - d(\mathbf{z}_{a_{\text{neg}}}^{(i)}, \mathbf{z}_v^{(i)})\right\},$$

where m is the margin and $d(\cdot, \cdot)$ is the cosine distance [9]. Each MTriplet training example consists of items $\mathbf{x}_a^{(i)}$, $\mathbf{x}_{a_{\text{neg}}}^{(i)}$, $\mathbf{x}_v^{(i)}$ and $\mathbf{x}_{v_{\text{neg}}}^{(i)}$, where $(\mathbf{x}_a^{(i)}, \mathbf{x}_v^{(i)})$ are mined speech-image pairs and $\mathbf{x}_{a_{\text{neg}}}^{(i)}$ and $\mathbf{x}_{v_{\text{neg}}}^{(i)}$ are negative examples. We mine negatives similarly to the within-modality positives (end of §4.2), but we add constraints inspired by [46–49] to obtain "hard" negatives, i.e. they are non-matching but close to the items in $(\mathbf{x}_a^{(i)}, \mathbf{x}_v^{(i)})$.

5. Experimental setup

Data. As in [7, 8], multimodal few-shot tasks are constructed by pairing isolated spoken digits from the TIDigits corpus [50] with handwritten digit images from MNIST [51]. TIDigits contains speech from 326 speakers. We use the test subsets of TIDigits and MNIST for the multimodal matching task. For training unsupervised indirect models (§3) and for mining pairs in the direct models (§4.1), we use the training subsets as unlabelled in-domain data by removing the class labels. For validating these models, we use the validation subsets (again without labels).

For background speech data we use the Buckeye corpus of English conversational speech from 40 speakers [52]. All utterances (from both TIDigits and Buckeye) are split into isolated words using forced alignments and parametrised as static mel-frequency cepstral coefficients (MFCCs). For background image data, we use the Omniglot dataset of grayscale images from 1623 handwritten character classes [29], which we invert and downsample to 28×28 pixels (to match the MNIST format). All image pixels are normalised to $[0, 1]$. Transfer learning is performed using the labelled training subsets of Buckeye and Omniglot, with validation subsets used for validation. We ensure that the background data does not contain any digit classes.

Models. For the indirect models (§3) we use the same architectures as in [8]. For the direct models (§4), we mine pairs via a five-shot support set, i.e. five word-image pairs for each of the classes. The MCAE architecture is given in Figure 2, with the loss weights set to $\alpha_a = 0.3$, $\alpha_v = 0.3$ and $\alpha_z = 0.4$. The MTriplet architecture is identical to the speech and vision encoders in Figure 2; we use a margin of $m = 0.2$. We use 130-dimensional representations $\mathbf{z}$ throughout to allow for fair

comparisons. Neural networks are implemented in TensorFlow and trained with a learning rate of 10^{-3} using Adam optimisation [53]. The above hyperparameters are based on previous work [8, 9] and weren't explicitly tuned. We did consider the effect of batch size; despite robust performance, we report model stability over five different sizes (from 16 to 256). Early stopping is performed by tracking model loss on mined speech-image validation pairs (so no labels are used for validation either).

Evaluation. As in [8], we evaluate our models on 400 five-shot *episodes* [31]. Each episode's support set contains five word-image pairs for each of the $L = 11$ digit classes ("one" to "nine", as well as "zero" and "oh"), i.e. a five-shot 11-way task. For each episode's matching set, we sample ten digit images not in the support set. If the speech query is either a "zero" or an "oh", it is counted as correct if the model predicts the matching image to be a 0. In each episode, we sample ten different spoken digit queries (not in the support set). The MCAE and MTriplet scores are averaged over five different batch sizes, each trained with five different seeds (i.e. 25 models in total in each case). Scores are reported with 95% confidence intervals.

6. Experimental results

Multimodal five-shot matching results are shown in Table 1. The first row is a naive indirect baseline where matching is performed on the input features: dynamic time warping (DTW) over MFCCs for speech and cosine distance over image pixels. The best overall score of 85.5% is achieved by the direct MTriplet, giving an absolute improvement of more than 25% over the best previous result [8]. This model is followed by the direct MCAE.

Table 1 also shows scores when pair mining (§4.1) is performed using cosine distance over the original features rather than using transfer-learned speech and image classifiers. The latter gives substantially better scores in both the MCAE and MTriplet. The combination of unsupervised learning (mining pairs from unlabelled in-domain data) with transfer learning (using a metric learned from labelled background data) is therefore essential for the superior performance of these direct models. Oracle results (when perfect pairs are used) also show that improvements to the mining approach (potentially by improving transfer learning) could lead to further future improvements.

Both [7] and [8] showed that many of the errors from the indirect models are due to mistakes in speech-speech unimodal comparisons which compound with mistakes in the subsequent image-image matching step. The direct models do not suffer from this, since the two-step approach is replaced with a single cross-modal comparison in the multimodal space. The question then is how much of the direct models' improvements are due to

Table 1: *Multimodal five-shot 11-way speech-to-image matching accuracy (%). Direct models are either trained with pairs mined using cosine distance over the input features or with representations from transfer-learned models. Oracle results with ground truth pairs are given for reference.*

	Model	Accuracy
Indirect multimodal few-shot models	DTW + pixels [7, 8]	41.9
	Transfer-learned classifier [7, 8]	59.7 ± 1.7
	Unsupervised CAE [8]	52.2 ± 0.7
Direct multimodal few-shot models	MCAE (cosine mined pairs)	59.1 ± 3.1
	MCAE (transfer learning pairs)	74.9 ± 1.9
	Oracle MCAE	93.6 ± 1.6
	MTriplet (cosine mined pairs)	67.3 ± 2.0
	MTriplet (transfer learning pairs)	**85.5 ± 1.6**
	Oracle MTriplet	99.1 ± 0.1

Table 2: *Multimodal speech-to-image matching accuracy (%) when direct models are either used directly (as intended) or as though they are indirect models (i.e. two separate unimodal comparisons in the multimodal space via the support set).*

Application mode	Model	
	MCAE	MTriplet
Direct cross-modal comparisons	**74.9 ± 1.9**	**85.5 ± 1.4**
Indirect unimodal comparisons	66.4 ± 2.6	76.2 ± 1.5

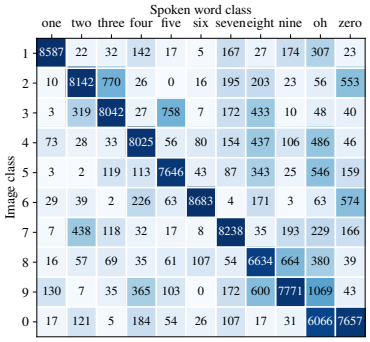

Figure 3: *Confusion matrix counts for the MTriplet from Table 1.*

better comparisons within the multimodal space, and how much are due to not having two comparisons? To quantify this, we use the direct models, but instead of performing direct comparisons, we use them as though they are indirect models, i.e. we use the multimodal space separately to do unimodal speech-speech and image-image comparisons via the support set (§3). Table 2 shows that there is indeed a large gain by not having to do two comparisons (row 1 vs row 2). But even the indirect application of the direct models (row 2) are better than the indirect models (Table 1), showing that the MCAE and MTriplet learn better representations even for within-modality comparisons.

Finally we are interested in the types of mistakes that the direct models make. Figure 3 shows the confusion matrix for the MTriplet from Table 1. We see that the most confusions are for the spoken word "oh" with the image 9. Qualitative analysis shows that this is because many 0 and 9 images are subjectively similar, rather than "oh" and "nine" being acoustically confusable. The fact that confusions are mostly due to image comparisons is supported by using the multimodal models to do *unimodal* few-shot classification: image classification are generally worse than speech classification (scores not shown here).

7. Conclusion

We proposed two novel direct multimodal few-shot learning models which outperformed existing indirect multimodal models by a substantial margin. An MTriplet model performed best. We showed that the improvements over indirect models are due to the combination of unsupervised and transfer learning, which results in more accurate embedding spaces and do not require two comparisons (i.e. no compounding of mistakes). Future work will look into improving the pair mining process and the feasibility of using these models on more realistic datasets.

8. Acknowledgements

This work is supported in part by the National Research Foundation of South Africa (grant no. 120409), CSIR and Saigen Scholarships for LN, and a Google Faculty Award for HK.

9. References

[1] I. Biederman, "Recognition-by-components: A theory of human image understanding." *Psych. Review*, vol. 94, 1987.

[2] G. A. Miller and P. M. Gildea, "How children learn words," *SciAM*, vol. 257, 1987.

[3] R. L. Gómez and L. Gerken, "Infant artificial language learning and language acquisition," *TiCS*, vol. 4, 2000.

[4] B. M. Lake, C. Lee, J. R. Glass, and J. B. Tenenbaum, "One-shot learning of generative speech concepts," in *Proc. CogSci*, 2014.

[5] O. Räsänen and H. Rasilo, "A joint model of word segmentation and meaning acquisition through cross-situational learning." *Psych. Review*, vol. 122, 2015.

[6] A. Borovsky, J. L. Elman, and M. Kutas, "Once is enough: N400 indexes semantic integration of novel word meanings from a single exposure in context," *Lang Learn Dev*, vol. 8, 2012.

[7] R. Eloff, H. A. Engelbrecht, and H. Kamper, "Multimodal one-shot learning of speech and images," in *Proc. ICASSP*, 2019.

[8] L. Nortje and H. Kamper, "Unsupervised vs. transfer learning for multimodal one-shot matching of speech and images," in *Proc. Interspeech*, 2020.

[9] D. Harwath, A. Torralba, and J. Glass, "Unsupervised learning of spoken language with visual context," in *Proc. NIPS*, 2016.

[10] D. Harwath *et al.*, "Jointly discovering visual objects and spoken words from raw sensory input," in *Proc. ECCV*, 2018.

[11] D. Harwath, W. Hsu, and J. Glass, "Learning hierarchical discrete linguistic units from visually-grounded speech," in *Proc. ICLR*, 2020.

[12] K. Leidal, D. Harwath, and J. Glass, "Learning modality-invariant representations for speech and images," in *Proc. ASRU*, 2017.

[13] G. Chrupała, L. Gelderloos, and A. Alishahi, "Representations of language in a model of visually grounded speech signal," in *Proc. ACL*, 2017.

[14] O. Scharenborg *et al.*, "Linguistic unit discovery from multi-modal inputs in unwritten languages: Summary of the "Speaking Rosetta" JSALT 2017 Workshop," in *Proc. ICASSP*, 2018.

[15] D. Merkx, S. L. Frank, and M. Ernestus, "Language learning using speech to image retrieval," *arXiv preprint arXiv:1909.03795*, 2019.

[16] W. N. Havard, J. Chevrot, and L. Besacier, "Catplayinginthesnow: Impact of prior segmentation on a model of visually grounded speech," *arXiv preprint arXiv:2006.08387*, 2020.

[17] H. Kamper, M. Elsner, A. Jansen, and S. J. Goldwater, "Unsupervised neural network based feature extraction using weak top-down constraints," in *Proc. ICASSP*, 2015.

[18] J. Weston, S. Bengio, and N. Usunier, "Large scale image annotation: learning to rank with joint word-image embeddings," *Mach Learn*, vol. 81, 2010.

[19] J. Ngiam *et al.*, "Multimodal deep learning," in *Proc. ICML*, 2011.

[20] A. Frome *et al.*, "DeViSE: A deep visual-semantic embedding model," in *Proc. NIPS*, 2013.

[21] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *Proc. NIPS*, 2013.

[22] C. Silberer and M. Lapata, "Learning grounded meaning representations with autoencoders," in *Proc. ACL*, 2014.

[23] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. ACM*, 2014.

[24] F. Pahde, M. Puscas, T. Klein, and M. Nabi, "Multimodal prototypical networks for few-shot learning," in *Proc. ICCV*, 2021.

[25] Y. Huang and L. Wang, "ACMM: Aligned cross-modal memory for few-shot image and sentence matching," in *Proc. ICCV*, 2019.

[26] L. Fei-Fei, Fergus, and Perona, "A Bayesian approach to unsupervised one-shot learning of object categories," in *Proc. ICCV*, 2003.

[27] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. PAMI*, vol. 28, 2006.

[28] B. M. Lake, R. R. Salakhutdinov, and J. B. Tenenbaum, "One-shot learning by inverting a compositional causal process," in *Proc. NIPS*, 2013.

[29] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, 2015.

[30] G. Koch, "Siamese neural networks for one-shot image recognition," in *Proc. ICML*, 2015.

[31] O. Vinyals *et al.*, "Matching networks for one shot learning," in *Proc. NIPS*, 2016.

[32] A. Parnami and M. Lee, "Few-shot keyword spotting with prototypical networks," *arXiv:2007.14463*, 2020.

[33] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, 2009.

[34] S. Ruder, "Neural transfer learning for natural language processing," Ph.D. dissertation, NUI Galway, Ireland, 2019.

[35] J. Bromley *et al.*, "Signature verification using a "Siamese" time delay neural network," in *Proc. NIPS*, 1994.

[36] X. Cong *et al.*, "Inductive unsupervised domain adaptation for few-shot classification via clustering," *arXiv preprint arXiv:2006.12816*, 2020.

[37] S. Bhosale, R. Chakraborty, and S. K. Kopparapu, "Semi supervised learning for few-shot audio classification by episodic triplet mining," *arXiv preprint arXiv:2102.08074*, 2021.

[38] Y. Chung, C. Wu, C. Shen, and H. Lee, "Unsupervised learning of audio segment representations using sequence-to-sequence recurrent neural networks," in *Proc. Interspeech*, 2016.

[39] Y. Wang, H. Lee, and L. Lee, "Segmental audio Word2Vec: Representing utterances as sequences of vectors with applications in spoken term detection," in *Proc. ICASSP*, 2018.

[40] N. Holzenberger *et al.*, "Learning word embeddings: Unsupervised methods for fixed-size representations of variable-length speech segments," in *Proc. Interspeech*, 2018.

[41] H. Kamper, "Truly unsupervised acoustic word embeddings using weak top-down constraints in encoder-decoder models," in *Proc. ICASSP*, 2019.

[42] P. Peng, H. Kamper, and K. Livescu, "A correspondence variational autoencoder for unsupervised acoustic word embeddings," in *NeurIPS SAS Workshop*, 2020.

[43] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large scale online learning of image similarity through ranking," *Pattern Recognit. Image Anal.*, vol. 5524, 2009.

[44] J. Wang *et al.*, "Learning fine-grained image similarity with deep ranking," in *Proc. CVPR*, 2014.

[45] K. M. Hermann and P. Blunsom, "Multilingual distributed representations without word alignment," in *Proc. ICLR*, 2014.

[46] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Proc. SIMBAD*, 2015.

[47] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, 2015.

[48] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv:1703.07737*, 2017.

[49] A. Jansen, M. Plakal, R. Pandya, D. P. W. Ellis, S. Hershey, J. Liu, R. C. Moore, and R. A. Saurous, "Unsupervised learning of semantic audio representations," in *Proc. ICASSP*, 2018.

[50] R. G. Leonard and G. R. Doddington, "TIDIGITS LDC93S10," Philadelphia: Linguistic Data Consortium, 1993.

[51] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proc. IEEE*, 1998.

[52] M. A. Pitt, K. Johnson, E. Hume, S. Kiesling, and W. Raymond, "The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability," *Speech Commun.*, vol. 45, 2005.

[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.