INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Unsupervised vs. transfer learning for multimodal one-shot matching of speech and images

Leanne Nortje, Herman Kamper

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

Abstract

We consider the task of multimodal one-shot speech-image matching. An agent is shown a picture along with a spoken word describing the object in the picture, e.g. *cookie*, *broccoli* and *ice-cream*. After observing *one* paired speech-image example per class, it is shown a new set of unseen pictures, and asked to pick the "ice-cream". Previous work attempted to tackle this problem using transfer learning: supervised models are trained on labelled background data not containing any of the one-shot classes. Here we compare transfer learning to unsupervised models trained on unlabelled in-domain data. On a dataset of paired isolated spoken and visual digits, we specifically compare unsupervised autoencoder-like models to supervised classifier and Siamese neural networks. In both unimodal and multimodal few-shot matching experiments, we find that transfer learning outperforms unsupervised training. We also present experiments towards combining the two methodologies, but find that transfer learning still performs best (despite idealised experiments showing the benefits of unsupervised learning).

Index Terms: one-shot learning, multimodal modelling, unsupervised models, transfer learning, word acquisition

1. Introduction

Young children are able to learn new objects and words from only a few examples [1–4]. In contrast, most conventional vision or speech processing systems require large amounts of labelled data. This has motivated studies into one-shot learning [5–11]: to learn a new concept from one or a few labelled examples. One-shot learning studies have mainly focused on learning new concepts in a single modality. But recently, *multimodal one-shot learning* has also been considered [12]. Instead of observing an item together with a class label, the model observes a pair of items coming from different modalities but representing the same concept. As an example, imagine a household robot is shown examples of *milk*, *eggs*, *butter* and a *mug*, each visual instance being paired with a spoken tag. At test time, the agent is then presented with a spoken query such as "butter", and asked to identify the corresponding visual object.

In [12], this was investigated on a dataset of isolated spoken digits paired with images. To perform multimodal matching at test-time, separate speech-speech and image-image comparisons were combined: a spoken query is compared to all the speech items in a so-called *support set*, the image corresponding to the closest item in the support set is determined, and this image is then compared to all the items in the *matching set* to predict the test image best matching the input speech query. To learn a distance metric within each modality, transfer learning was used by training supervised vision and speech models on background training data not containing any of the one-shot test classes. As in other unimodal one-shot studies in gesture recognition [13, 14], video [15] and robotics [16, 17], this can be motivated by the observation that humans can call on prior knowledge when learning new concepts.

Except for existing knowledge, it is also conceivable that, before being shown paired examples, an agent such as the household robot would be exposed to a large amount of *unlabelled* speech and visual data from its environment. Some of these unlabelled examples could correspond to the classes of interest. Motivated by this observation, we ask how unsupervised models trained on unlabelled in-domain data compares to transfer learning from background data for multimodal one-shot matching.

To learn feature representations for within-modality comparisons, we specifically consider two unsupervised learning strategies. An autoencoder (AE) attempts to reproduce its input at its output through a bottleneck feature layer. The correspondence autoencoder (CAE) tries to reproduce another instance of the input at its output [18]. Since we only have unlabelled data, the CAE samples nearest neighbours to obtain its output targets. We compare these unsupervised models to supervised classifier and Siamese neural networks trained on background data [12]. Each of the models are trained separately on vision and speech data and then used to estimate within-modality similarity.

On the same isolated digit speech-image multimodal one-shot matching task as in [12], we show that transfer learning outperforms unsupervised modelling. We also consider approaches for combining transfer and unsupervised learning. Although this yields improvements over a purely unsupervised model, the best overall performance is still achieved through transfer learning.[1]

2. Multimodal one-shot matching

We first describe unimodal one-shot matching and then extend it to the multimodal case. As an example, we consider one-shot speech classification, illustrated on the left in Figure 1(a). The model is shown a support set $\mathcal{S}$, containing one isolated spoken word with a text label for each of the L word classes. From this set, the model must learn a classifier $C_{\mathcal{S}}$ that can make predictions on an unseen test query $\mathbf{x}_a^*$. One approach is to simply compare the query with each item in the support set and then predict the label of the closest item, as illustrated on the right in Figure 1(a).

Figure 1(b) illustrates *multimodal* one-shot speech-image matching. Instead of labelled examples, the multimodal support set $\mathcal{S} = \{(\mathbf{x}_a^{(i)}, \mathbf{x}_v^{(i)})\}_{i=1}^{L}$ consists of pairs, where each isolated spoken word $\mathbf{x}_a^{(i)}$ has a corresponding image $\mathbf{x}_v^{(i)}$. One pair is given for each of the L classes. At test time, the model is presented with an unseen spoken query $\mathbf{x}_a^*$ and asked to determine the matching image in a test (or matching) set $\mathcal{M}_v = \{(\mathbf{x}_v^{(i)})\}_{i=1}^{N}$ of unseen images, as illustrated on the left in Figure 1(b). Neither the query $\mathbf{x}_a^*$ nor the matching set items

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

http://dx.doi.org/10.21437/Interspeech.2020-87

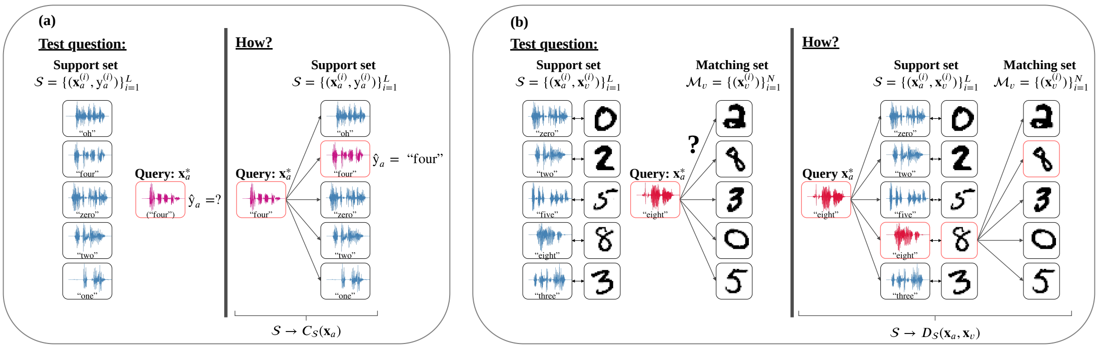
(a)
Test question:
Support set
$\mathcal{S} = \{(\mathbf{x}_a^{(i)}, \mathrm{y}_a^{(i)})\}_{i=1}^{L}$
"oh"
"four"
"zero"
"two"
"one"
Query: $\mathbf{x}_a^*$
("four")
$\hat{\mathrm{y}}_a$ =?
How?
Support set
$\mathcal{S} = \{(\mathbf{x}_a^{(i)}, \mathrm{y}_a^{(i)})\}_{i=1}^{L}$
Query: $\mathbf{x}_a^*$
"four"
"oh"
"four"
$\hat{\mathrm{y}}_a$ = "four"
"zero"
"two"
"one"
$\mathcal{S} \rightarrow C_{\mathcal{S}}(\mathbf{x}_a)$
(b)
Test question:
Support set
$\mathcal{S} = \{(\mathbf{x}_a^{(i)}, \mathbf{x}_v^{(i)})\}_{i=1}^{L}$
Matching set
$\mathcal{M}_v = \{(\mathbf{x}_v^{(i)})\}_{i=1}^{N}$
"zero"
"two"
"five"
"eight"
"three"
Query: $\mathbf{x}_a^*$
"eight"
?
How?
Support set
$\mathcal{S} = \{(\mathbf{x}_a^{(i)}, \mathbf{x}_v^{(i)})\}_{i=1}^{L}$
Matching set
$\mathcal{M}_v = \{(\mathbf{x}_v^{(i)})\}_{i=1}^{N}$
Query $\mathbf{x}_a^*$
"eight"
"zero"
"two"
"five"
"eight"
"three"
$\mathcal{S} \rightarrow D_{\mathcal{S}}(\mathbf{x}_a, \mathbf{x}_v)$

Figure 1: *(a) Unimodal one-shot speech classification and (b) multimodal one-shot speech-image matching. In both cases, the left side illustrates the question shown at test time, and the right side illustrates how the model makes its prediction.*

$\mathcal{M}_v$ occur exactly in the support set $\mathcal{S}$. To perform this task, we need to use $\mathcal{S}$ to construct a distance metric $D_{\mathcal{S}}(\mathbf{x}_a, \mathbf{x}_v)$ between audio queries and test images.

The approach we use (originally proposed in [12]) is to reduce the task to two unimodal comparisons, as shown on the right in Figure 1(b). First, we compare the query $\mathbf{x}_a^*$ to each $\mathbf{x}_a^{(i)}$ in $\mathcal{S}$ to find the query's closest spoken neighbour in the support set. This closest neighbour's paired image is then compared to each image $\mathbf{x}_v^{(i)}$ in the matching set $\mathcal{M}_v$. This closest matching-set image is then selected as the model's prediction. In the figure, this is the image of the rightmost *eight*.

We can also extend one-shot learning to K-shot learning. In unimodal L-way K-shot classification, the support set $\mathcal{S}$ contains L classes and K labelled examples per class. In multimodal L-way K-shot matching, $\mathcal{S} = \{(\mathbf{x}_a^{(i)}, \mathbf{x}_v^{(i)})\}_{i=1}^{L \times K}$ consists of K speech-image pairs for each of the L classes.

3. Feature representations

In the description above we implicitly assume that we have a method or model that can measure similarity within a modality. The aim of this paper is to consider different feature representations for these similarity comparisons, specifically comparing transfer learning (used in [12]) to unsupervised feature learning. To compare the different features, we use the same framework as in [12] where multimodal one-shot learning is performed via two unimodal comparisons (as outlined above, Figure 1(b)-right). Note that this is not an end-to-end approach; future work will explore learning direct cross-modal matching networks.

As a baseline, we use raw speech and image features directly (§3.1). We then consider different neural networks to learn feature representations (§3.2 and §3.3). We use separate networks for learning speech and image features. For both the speech and vision models, we consider two settings: training on unlabelled in-domain data (§3.2) and training on labelled background data (§3.3).

3.1. Raw feature matching

As a nearest neighbour baseline, we use cosine distance over image pixels for image-to-image comparisons, and dynamic time warping (DTW) over MFCCs for speech-to-speech comparisons.

3.2. Unsupervised models on unlabelled in-domain data

We consider two unsupervised models trained on unlabelled in-domain speech and vision data—data which includes unlabelled instances of classes that we will see during one-shot testing.

An autoencoder (AE) is an unsupervised neural network which aims to reconstruct its input through a lower dimensional latent representation that acts as an information bottleneck [19]. As shown in Figure 2, the AE's encoder $f_\theta(\mathbf{x}^{(i)})$ encodes the input $\mathbf{x}^{(i)}$ to the feature representation $\mathbf{z}^{(i)}$. The decoder $f_\phi(\mathbf{z}^{(i)})$ decodes $\mathbf{z}^{(i)}$ to produce the output $\hat{\mathbf{y}}^{(i)}$. We use a squared loss between the network's output $\hat{\mathbf{y}}^{(i)}$ and the desired output $\mathbf{y}^{(i)}$, i.e., $\ell = ||\mathbf{y}^{(i)} - \hat{\mathbf{y}}^{(i)}||_2^2$, with the target set to $\mathbf{y}^{(i)} = \mathbf{x}^{(i)}$.

The correspondence autoencoder (CAE) is identical to the AE but instead of reproducing the input $\mathbf{x}^{(i)}$, it aims to reproduce another instance $\mathbf{x}_{\text{pair}}^{(i)}$ of the same class as the input [18], i.e. we set the target $\mathbf{y}^{(i)} = \mathbf{x}_{\text{pair}}^{(i)}$ in the loss ℓ. The intuition is that the CAE will produce features that are invariant to properties not common to two inputs while capturing aspects that are, such as the class. We consider two variants of the CAE: one trained from scratch and another pretrained as an AE before switching to the CAE loss (denoted as AE-CAE). To train the CAE, we need pairs of items of the same class. Since our in-domain data is unlabelled, we use cosine distance over pixels to find image pairs that are most alike, and DTW to find spoken word pairs predicted to be of the same type. Speaker information is used to ensure that speech pairs are from different speakers.

Using unlabelled in-domain image data, we train unsupervised vision networks with the AE, CAE and AE-CAE losses; we use the architecture shown in Figure 2(a), with a convolutional neural network (CNN) encoder producing the latent feature vector, and a decoder with transposed convolutions. Similarly, we use unlabelled in-domain speech data to train unsupervised speech networks using the AE, CAE and AE-CAE losses; we use an encoder recurrent neural network (RNN) producing the latent feature vector which is then used to condition a decoder RNN, as shown in Figure 2(b). These speech RNNs are similar to the acoustic embedding models of [20–23], since they give a fixed-sized embedding for variable duration input.

3.3. Transfer learning from labelled background data

We next consider training supervised models on labelled background data. These datasets do not contain any instances of the target one-shot classes. The idea is that features learned by such models would still be useful for determining similarity on unseen classes [10]. This is a form of *transfer learning* [24, 25].

We specifically consider supervised classifier and Siamese neural networks, as in [12]. We use identical architectures to the encoder parts of the networks in Figure 2. For the classifiers, we add a softmax layer after the feature embedding layer $\mathbf{z}^{(i)}$ and train the networks with the multiclass log loss.

A Siamese network does not classify an input, but measures similarity between inputs [26]. The network consists of

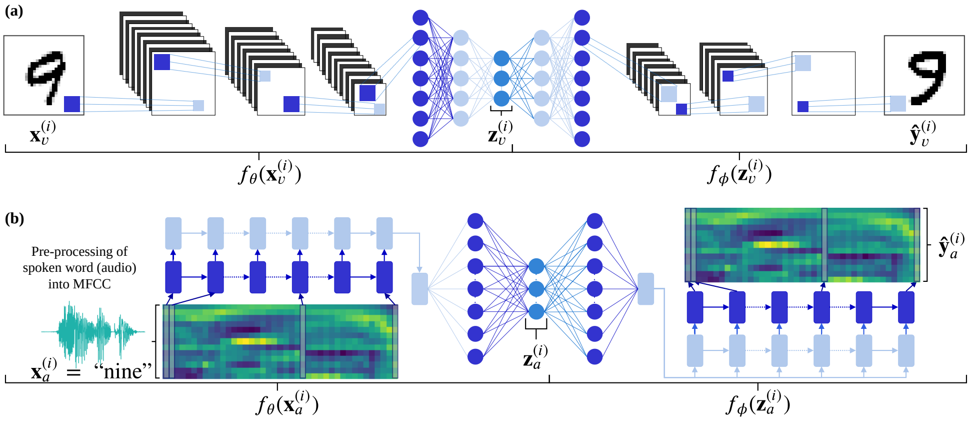
(a)
$\mathbf{x}_v^{(i)}$
$\mathbf{z}_v^{(i)}$
$\hat{\mathbf{y}}_v^{(i)}$
$f_\theta(\mathbf{x}_v^{(i)})$
$f_\phi(\mathbf{z}_v^{(i)})$
(b)
Pre-processing of spoken word (audio) into MFCC
$\mathbf{x}_a^{(i)}$ = "nine"
$\mathbf{z}_a^{(i)}$
$\hat{\mathbf{y}}_a^{(i)}$
$f_\theta(\mathbf{x}_a^{(i)})$
$f_\phi(\mathbf{z}_a^{(i)})$

Figure 2: *(a) Convolutional neural networks (CNNs) are used to learn feature representations for image data and (b) recurrent neural networks (RNNs) are used to learn feature representations for speech data.*

identical sub networks with shared parameters; each network maps its input to an embedding. Ideally, inputs of the same class should have similar embeddings and inputs of different classes should have different embeddings. Say we have inputs $\mathbf{x}$, $\mathbf{x}_{\text{pair}}$ and $\mathbf{x}_{\text{neg}}$, where $\mathbf{x}$ and $\mathbf{x}_{\text{pair}}$ are from the same class and $\mathbf{x}$ and $\mathbf{x}_{\text{neg}}$ are from different classes. We want the distance between the embeddings of $\mathbf{x}$ and $\mathbf{x}_{\text{pair}}$ to be smaller than those of $\mathbf{x}$ and $\mathbf{x}_{\text{neg}}$. We use the triplet hinge loss $l(\mathbf{x}, \mathbf{x}_{\text{pair}}, \mathbf{x}_{\text{neg}}) = \max\{0, m + d(\mathbf{x}, \mathbf{x}_{\text{pair}}) - d(\mathbf{x}, \mathbf{x}_{\text{neg}})\}$, where $d(\mathbf{x}_1, \mathbf{x}_2) = \left\|\mathbf{z}_1 - \mathbf{z}_2\right\|_2^2$ is the squared Euclidean distance between the embeddings $\mathbf{z}_1$ and $\mathbf{z}_2$ of $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively, and m is a margin parameter [27, 28]. To sample negative items, we use the online semi-hard mining scheme, where for each positive pair $(\mathbf{x}, \mathbf{x}_{\text{pair}})$, the most difficult negative pair $(\mathbf{x}, \mathbf{x}_{\text{neg}})$ is sampled (meeting some constraints) [29–31].

Again, separate classifier and Siamese vision CNNs and speech RNNs are trained on labelled background data. We also consider supervised variants of the CAE and AE-CAE approaches, where instead of finding input-output training pairs based on their nearest neighbours (§3.2), we train on ground truth pairs from the background data (these were not considered in [12]). For all of the models, we use the embedding $\mathbf{z}^{(i)}$ as representation for unseen input $\mathbf{x}^{(i)}$.

4. Experimental setup

4.1. Data

We follow the same setup as [12], using a dataset of paired isolated spoken digits and handwritten digit images [32]. Speech data are parametrised as Mel-frequency cepstral coefficients (MFCCs). Image pixels are normalised to $[0, 1]$. We use the TIDigits corpus as our in-domain speech data; the corpus consists of spoken digit sequences from 326 speakers [33]. We split these sequences into isolated digits using forced alignments. As our in-domain image data, we use the MNIST corpus which contains 28 × 28 grayscale handwritten digit images [34]. Although the TIDigits and MNIST datasets are labelled, note that we use it as unlabelled in-domain data for the models in §3.2; we specifically train these unsupervised models on unlabelled isolated examples from the training subsets of these datasets. All one-shot evaluation experiments are then performed on the MNIST and TIDigits test subsets.

For background speech data, we use the Buckeye corpus of English speech from 40 speakers [35]. We use forced alignments to extract a set of labelled isolated words from this set. For background image data, we use Omniglot [36], containing 1623 types of handwritten characters which we invert and downsample to 28 × 28. We ensure that there are no instances of the target digit classes in either the Buckeye or Omniglot background data.

4.2. Models

Neural networks are implemented in TensorFlow and trained using Adam optimisation [37] with a learning rate of 10^{-3}. Model hyperparameters were tuned using unimodal one-shot classification on test subsets of the background data, while early stopping was performed on validation subsets—neither of these background sets have item or class overlap with the final evaluation data. We use a feature embedding dimensionality of 130 in all models to make results comparable. All speech RNNs take static MFCCs as input, but first and second order derivatives are used in the DTW baseline where it is beneficial.

Unsupervised speech RNNs are trained on unlabelled isolated digits from the TIDigits training set using the AE, CAE and AE-CAE losses (§3.2). In all cases, the encoder and decoder each consists of three 400-unit RNN layers. Unsupervised vision CNNs are trained with the AE, CAE and AE-CAE losses on unlabelled images from the MNIST training set. The encoder consists of three convolutional layers with 3×3 kernels and 32, 64 and 128 units; the decoder has the inverse architecture.

For transfer learning (§3.3), we train supervised classifier and Siamese speech RNNs on labelled isolated words from the Buckeye training set. Similarly, we train supervised classifier and Siamese vision CNNs on Omniglot. All these supervised models share the same structure as the encoder components from their unsupervised counterparts. We also train supervised variants of the CAE and AE-CAE speech and vision models on the labelled background data.

4.3. Evaluation

We evaluate models averaged over 400 "episodes" [10]. To construct the support set, each multimodal episode randomly samples a spoken digit and paired image for each of the $L = 11$ classes ("one" to "nine", as well as "zero" and "oh"). A matching set is then sampled for testing, containing ten digit images not in the support set. Finally, a spoken query is sampled, also not in the support set. The speech query then needs to be matched to the correct image in the matching set. The matching set only contains ten digit images since there are only ten unique handwritten digit classes (both "zero" and "oh" are counted as correct if the image is that of a 0). Within an episode, ten different

Table 1: *Unimodal one- and five-shot speech classification.*

Model		11-way accuracy (%) one-shot	 five-shot
Baseline	DTW	65.90	89.45
Transfer learning models	Classifier RNN	**86.87 ± 0.83**	**95.40 ± 0.50**
	Siamese RNN	83.52 ± 2.56	94.34 ± 0.86
	CAE RNN	79.89 ± 1.32	92.16 ± 0.90
	AE-CAE RNN	80.02 ± 1.04	93.91 ± 0.25
Unsupervised models	AE RNN	53.82 ± 1.70	75.58 ± 1.54
	CAE RNN	75.80 ± 1.76	95.14 ± 0.80
	AE-CAE RNN	77.01 ± 1.29	93.30 ± 0.56

Table 2: *Multimodal one- and five-shot speech-image matching.*

Model		11-way accuracy (%) one-shot	 five-shot
Baseline	DTW + Pixels	31.80	41.88
Transfer learning models	Classifier [12]	**56.80 ± 1.19**	**59.67 ± 1.73**
	Siamese [12]	54.83 ± 1.80	59.25 ± 0.79
	CAE	46.60 ± 0.69	53.82 ± 1.07
	AE-CAE	48.15 ± 1.21	56.81 ± 1.21
Unsupervised models	AE	28.99 ± 0.84	38.68 ± 1.51
	CAE	42.75 ± 0.62	52.15 ± 0.69
	AE-CAE	42.81 ± 1.01	50.28 ± 0.29

Table 3: *Multimodal one- and five-shot speech-image matching using models that combine transfer and unsupervised learning.*

Model	11-way accuracy (%) one-shot	 five-shot
Baseline: DTW + Pixels	31.80	41.88
Transfer learning: Classifier [12]	**56.80 ± 1.19**	**59.67 ± 1.73**
CAE with cosine pairs	42.75 ± 0.62	52.15 ± 0.69
CAE with classifier pairs	48.66 ± 1.14	55.59 ± 0.71
Transfer learning + CAE fine-tuning	54.32 ± 2.19	59.37 ± 1.80
CAE with oracle pairs	89.19 ± 0.69	92.81 ± 0.47

query instances are also sampled while keeping the support and matching sets fixed. We report unimodal and multimodal one- and five-shot matching accuracies with 95% confidence intervals averaged over five models trained with different seeds.

5. Experimental Results

5.1. *K*-shot unimodal speech and image classification

We first consider unimodal results in isolation. Table 1 shows one- and five-shot speech classification results. All models except the AE RNN outperform the baseline. The classifier RNN achieves the highest accuracies, followed by the Siamese RNN. In all cases, transfer learning models outperform their unsupervised counterparts, except for the five-shot CAE RNN.

For unimodal image classification (not shown here), the trends are very similar, with the classifier and Siamese CNNs achieving accuracies of around 64% and 84% for the one- and five-shot cases, respectively. Again, these transfer learning models outperform all the unimodal unsupervised image models.

5.2. *K*-shot multimodal speech and image matching

Table 2 shows multimodal one- and five-shot results.[2] In each case, the same model type is used to obtain speech and image features, e.g. the *Classifier* row uses a CNN vision classifier to get image features with an RNN speech classifier for speech features. In both one- and five-shot multimodal matching, the classifier performs best followed closely by the Siamese model. None of the unsupervised models perform as well as these models obtained using transfer learning. For the CAE and AE-CAE losses, the models trained using labelled background data also outperform the unsupervised variants.

5.3. Towards combined transfer and unsupervised learning

It is evident that the transfer learning approach originally followed in [12] outperforms the unsupervised approach developed here. However, the two methodologies might be complementary: transfer learning from background data could capture general properties within a particular modality, while unsupervised learning on unlabelled in-domain data could provide a way to tailor representations to a specific test setting.

As an initial investigation, we propose two combined models here, with results given in Table 3. The *CAE with cosine pairs* (row 3) is repeated from Table 2. Instead of finding nearest neighbours using cosine distance, we use representations from the classifier (trained on background data) to find pairs in the unlabelled in-domain data for training a CAE (as with the standard CAE, speaker information is still used to ensure that pairs are from different speakers). We see that this *CAE with classifier pairs* (row 4) gives a small improvement over the standard CAE. By additionally initialising the CAE by training it on the labelled background data and then fine-tuning it on the in-domain data, we get a further improvement (*Transfer learning + CAE fine-tuning*, row 5). Neither of these approaches, however, outperform the transfer learned classifier (row 2).

In order to see if it is at all possible to achieve better performance with the CAE by using more accurate training pairs, we also give the performance of a CAE trained only using correct pairs in the last row of Table 3. We see that this oracle model outperforms all other approaches, indicating that, if we were able to improve the CAE's training pairs, we might be able to take advantage of an unsupervised learning scheme.

6. Conclusion

We have compared existing and new models for few-shot multimodal speech-image matching. Transfer learning from background data consistently outperformed unsupervised modelling on unlabelled in-domain data on a multimodal one-shot matching benchmark. We also proposed two approaches for combining transfer and unsupervised learning. Although neither improved the best transfer learning approach, performance improved over the standard unsupervised approach. We will therefore also consider other approaches for combining the methodologies in future work. Building on models which directly maps images and unlabelled speech into a joint space [38–41], we will also consider end-to-end solutions for multimodal one-shot learning.

[2] Note that the results here are not directly comparable to that of [12]. We found a small bug in the validation setup of [12]; the scores across models in [12] are comparable, but lower scores are achieved when using the proper validation setup used in this paper. We reran the code of [12] to confirm the scores reported here.

This work is supported in part by the National Research Foundation of South Africa (grant number: 120409), a Google Faculty Award for HK, a DST CSIR scholarship for LN, and funding from Saigen.

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

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

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

[7] B. M. Lake, R. Salakhutdinov, J. Gross, and J. B. Tenenbaum, "One shot learning of simple visual concepts," *CogSci*, vol. 33, 2011.

[8] B. M. Lake, C.-y. Lee, J. R. Glass, and J. B. Tenenbaum, "One-shot learning of generative speech concepts," *CogSci*, vol. 36, 2014.

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

[10] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. NIPS*, 2016.

[11] P. Shyam, S. Gupta, and A. Dukkipati, "Attentive recurrent comparators," in *Proc. ICML*, 2017.

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

[13] W. Thomason and R. A. Knepper, "Recognizing unfamiliar gestures for human-robot interaction through zero-shot learning," in *Proc. ISER*, 2017.

[14] D. Wu, F. Zhu, and L. Shao, "One shot learning gesture recognition from RGBD images," in *Proc IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2012.

[15] T. Stafylakis and G. Tzimiropoulos, "Zero-Shot keyword spotting for visual speech recognition in-the-wild," in *Proc. ECCV*, 2018.

[16] M. R. Walter, Y. Friedman, M. Antone, and S. Teller, "One-shot visual appearance learning for mobile manipulation," *IJRR*, vol. 31, 2012.

[17] C. Finn, T. Yu, T. Zhang, P. Abbeel, and S. Levine, "One-shot visual imitation learning via meta-learning," *arXiv:1709.04905*, 2017.

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

[19] D. Chicco, P. Sadowski, and P. Baldi, "Deep autoencoder neural networks for gene ontology annotation predictions," in *Proc. ACM*, 2014.

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

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

[22] N. Holzenberger, M. Du, J. Karadayi, R. Riad, and E. Dupoux, "Learning word embeddings: Unsupervised methods for fixed-size representations of variable-length speech segments," in *Proc. Interspeech*, 2018.

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

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

[25] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. ICML*, 2014.

[26] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a "Siamese" time delay neural network," in *Proc. NIPS*, 1994.

[27] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proc. CVPR*, 2014.

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

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

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

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

[32] K. Kashyap, "Learning digits via joint audio-visual representations," Thesis, Massachusetts Institute of Technology, 2017.

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

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

[35] 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.

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

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

[38] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. ICML*, 2011.

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

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

[41] D. Harwath, A. Recasens, D. Suris, G. Chuang, A. Torralba, and J. Glass, "Jointly discovering visual objects and spoken words from raw sensory input," in *Proc. ECCV*, 2018.