INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Visually grounded few-shot word acquisition with fewer shots

Leanne Nortje, Benjamin van Niekerk, Herman Kamper

MediaLab, E&E Engineering, Stellenbosch University, South Africa
nortjeleanne@gmail.com, benjamin.l.van.niekerk@gmail.com, kamperh@sun.ac.za

Abstract

We propose a visually grounded speech model that acquires new words and their visual depictions from just a few word-image example pairs. Given a set of test images and a spoken query, we ask the model which image depicts the query word. Previous work has simplified this problem by either using an artificial setting with digit word-image pairs or by using a large number of examples per class. We propose an approach that can work on natural word-image pairs but with less examples, i.e. fewer shots. Our approach involves using the given word-image example pairs to mine new unsupervised word-image training pairs from large collections of unlabelled speech and images. Additionally, we use a word-to-image attention mechanism to determine word-image similarity. With this new model, we achieve better performance with fewer shots than any existing approach.
Index Terms: few-shot learning, multimodal modelling, visually grounded speech models, word acquisition.

1. Introduction

Speech recognition for low-resource languages faces a major obstacle: it requires large amounts of transcribed data for development [1]. To overcome this, we can look to how children acquire new words from a few examples without the use of transcriptions [2–6]. E.g. Borovsky et al. [7] shows that children can acquire a word for a visual object after seeing it only once. This has led to recent studies into multimodal few-shot learning [8–10]: the task of learning new concepts from a few examples, where each example consists of instances of the same concept but from different modalities. E.g. imagine a robot seeing a picture of a *zebra*, *kite* and *sheep* while also hearing the spoken word for each concept. After seeing this small set of examples (called a support set) the robot is prompted to identify which image in an unseen set corresponds to the word "zebra".

Building off of a growing number of studies in visually grounded speech modelling [11–15], we consider this multimodal problem of learning the spoken form of a word and its visual depiction from only a few paired word-image examples. Multimodal few-shot speech-image learning was first introduced in [8] and then extended in [9] and [10]. But these studies were performed in an artificial setting where spoken isolated digits were paired with MNIST images of digits. This shortcoming was recently addressed by Miller and Harwath [16], who considered multimodal few-shot learning on isolated words paired with natural images. Their specific focus was on learning a new concept while not forgetting previously learned concepts, i.e. dealing with the problem of catastrophic forgetting. While their methods performed well in a few-shot retrieval task with five classes, it required a relatively large number of samples per class, i.e. many "shots". Our aim is to do visually grounded multimodal few-shot learning on natural images with fewer shots. We do not explicitly focus on the catastrophic forgetting problem (for now), although we do evaluate using the same setup as [16].

There are two core components to our new approach. Firstly, we use the support set to "mine" new noisy word-image pairs from unlabelled speech and image collections. Concretely, each spoken word example in the support set is compared to each utterance in an unlabelled speech corpus; we use a new query-by-example approach (called QbERT) to identify segments in the search utterances that match the word in the support set. We follow a similar approach for mining additional images from the few-shot classes by using AlexNet [17] embeddings and cosine distance for the comparisons between a support set image and unlabelled search images. The mined words and images are then paired up, thereby artificially increasing the size of our support set (in an unsupervised way). This pair mining scheme is very similar to that followed in [10], where it was used on digit image-speech data with simpler within-modality comparisons.

Secondly, our new approach is based on a model with a word-to-image attention mechanism. This multimodal attention network (MATTNET) takes a single word embedding and calculates its correspondence to each pixel embedding to learn how the word is depicted within an image. This is similar to the vision attention part of the model from [18], where the goal was to localise visual keywords in speech (not in a few-shot setting).

We first evaluate our approach on the few-shot retrieval task also used in [16]. We show that MATTNET achieves higher retrieval scores for fewer shots than [16]'s models. We also show that our approach yields more consistent scores with a larger number of few-shot classes. Secondly, we evaluate our approach in a more conventional few-shot classification task where it only needs to correctly distinguish between classes seen in the support set. We consider settings with different numbers of classes and shots, and show that we can achieve accuracies higher than 60% with as little as five shots.

Our core contributions is the new mining scheme operating on natural images and speech, and then the application of a new attention-based model to the task of multimodal few-shot learning. Our experiments show that both these contributions lead to improvements over previous methods.

2. Visually grounded few-shot learning and evaluation

Below we describe the few-shot learning setup as well as the two tasks that we consider in this work.

Visually grounded few-shot learning. Children can learn a new word for a visual object from only a few examples [7]. To attempt to replicate this in a machine learning model, we train

10.21437/Interspeech.2023-668

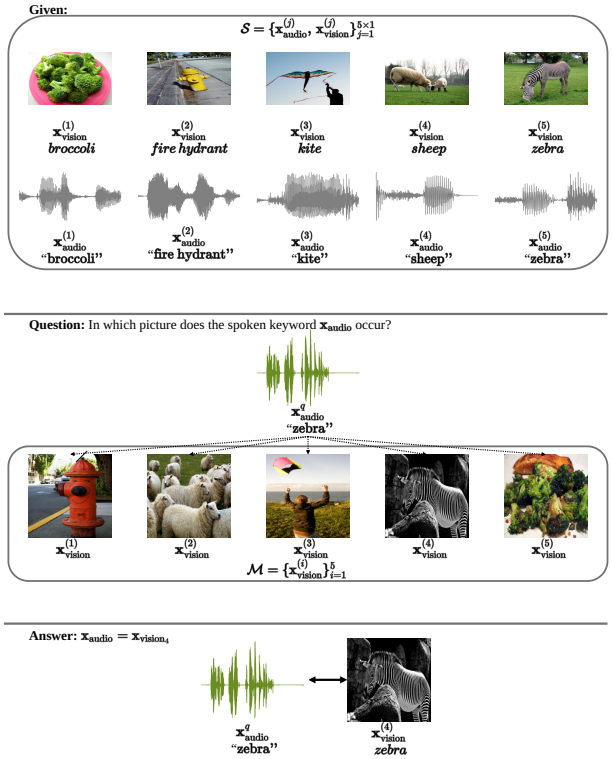

Figure 1: *Given the few examples in the support set $\mathcal{S}$, the multimodal few-shot classification task is to e.g. identify the image depicting the word "zebra" from a set of unseen images.*

a model on a few spoken word-image examples. This set of K examples per class is called the support set $\mathcal{S}$. Each pair in $\mathcal{S}$ consists of an isolated spoken word $\mathbf{x}^{(j)}_{\text{audio}}$ and a corresponding image $\mathbf{x}^{(j)}_{\text{vision}}$. For the *one-shot* case shown in the top part of Figure , $\mathcal{S}$ consists of one word-image example pair for each of the L classes. For the L-way K-shot task, the support set $\mathcal{S} = \{\mathbf{x}^{(j)}_{\text{audio}}, \mathbf{x}^{(j)}_{\text{vision}}\}_{j=1}^{L\times K}$ contains K word-image example pairs for each of the L classes.

Visually grounded few-shot word classification. In this task, illustrated in the middle and lower parts of Figure , we are given an unseen isolated spoken word query $\mathbf{x}^{q}_{\text{audio}}$ and prompted to identify the corresponding image in a matching set $\mathcal{M} = \{(\mathbf{x}^{(i)}_{\text{vision}})\}_{i=1}^{L}$ of unseen test images. $\mathcal{M}$ contains one image depicting each of the L classes. Neither the test-time speech query $\mathbf{x}_{\text{audio}}$ nor any images in $\mathcal{M}$ are duplicated in the support set. This image-speech task was considered in [8–10], but here, for the first time, we use natural images instead of isolated digit images. In contrast to the task that we describe next, this is a conventional few-shot classification task where the model only needs to correctly distinguish between classes seen in the support set, i.e. there are no other background or imposter classes.

Visually grounded few-shot retrieval. In contrast, in this task the goal is to test whether a model can search through a large collection of images and retrieve those that depict a few-shot query, i.e. the matching set $\mathcal{M}$ in this case contains images that depict the L few-shot classes but also images that depict other classes. These additional images might contain completely unseen classes, or background classes potentially seen during pretraining of the few-shot model. The model is penalised if it retrieves one of these imposter images. This few-shot retrieval task was proposed in [16]. Their interest was specifically in measuring catastrophic forgetting. Since their task requires a model to distinguish between few-shot classes and other classes, it can be used to not only determine whether models can be updated to learn new classes from only a few examples, but also how well the model remembers previously learned (background) classes. We do not explicitly focus on the catastrophic forgetting problem, but we want to compare to [16], which is why we also consider this retrieval task.

For both tasks we need a distance metric $D_{\mathcal{S}}(\mathbf{x}^{q}_{\text{audio}}, \mathbf{x}^{(i)}_{\text{vision}})$ between instances from the speech and vision modalities. Below we describe our model that we use as this distance metric.

3. Multimodal few-shot attention

Our approach to determine $D_{\mathcal{S}}(\mathbf{x}^{q}_{\text{audio}}, \mathbf{x}^{(i)}_{\text{vision}})$ relies on two core components: a model with a word-to-image attention mechanism and a method to mine pairs using a few ground truth word-image examples (given in the support set).

3.1. Word-to-image attention mechanism

Our model is shown in Figure and we call it MATTNET (**m**ultimodal **att**ention **net**work). To start off, we adapt the multimodal localising attention model of [18] that consists of an audio and a vision branch. For the vision branch, we replace ResNet50 [19] with an adaptation of AlexNet [17] to encode an image input $\mathbf{x}_{\text{vision}}$ into a sequence of embeddings $\mathbf{y}_{\text{vision}}$. For the audio branch, we use the same audio subnetwork as [18] that consists of an acoustic network f_{acoustic} which extracts speech features from a spoken input $\mathbf{x}_{\text{audio}}$. However, [18] takes an entire spoken utterance as $\mathbf{x}_{\text{audio}}$, whereas we use a single isolated spoken word. We also add a few linear layers to the BiLSTM network f_{BiLSTM} to encode the speech features into a single audio embedding $\mathbf{y}_{\text{audio}}$, similar to acoustic word embeddings [20–23]. We connect the vision and audio branches with a multimodal attention mechanism to compare the word embedding $\mathbf{y}_{\text{audio}}$ to each pixel embedding in $\mathbf{y}_{\text{vision}}$.

To get this word-to-image attention mechanism, we take the keyword localising attention mechanism of [18] which detects whether certain keywords occur in both spoken utterances and images. However, we aim to only detect whether a single isolated spoken word occurs somewhere within an image. More specifically, we calculate attention weights $\mathbf{a}$ over the pixel embeddings by calculating the dot product between $\mathbf{y}_{\text{audio}}$ and each pixel embedding in $\mathbf{y}_{\text{vision}}$. By taking the maximum over $\mathbf{a}$, we get a similarity score $S \in [0, 100]$. The higher S, the more probable it is that the spoken word corresponds to one or more objects in the image. If S is low, it is less probable that any object in the image corresponds to the spoken word.

We train MATTNET by using S in a contrastive loss:

$$
\begin{aligned}
l = \ & \text{MSE}\Big(S(\mathbf{e}_{\text{audio}}, \mathbf{e}_{\text{vision}}), 100\Big) \\
& + \sum_{i=1}^{N_{\text{neg}}} \text{MSE}\Big(\big[S\big(\mathbf{e}^{-}_{\text{audio}_i}, \mathbf{e}_{\text{vision}}\big), S(\mathbf{e}_{\text{audio}}, \mathbf{e}^{-}_{\text{vision}_i}), S(\mathbf{e}_{\text{audio}}, \mathbf{e}^{-}_{\text{vision background}_i})\big], 0\Big) \\
& + \sum_{i=1}^{N_{\text{pos}}} \text{MSE}\Big(\big[S(\mathbf{e}_{\text{audio}}, \mathbf{e}^{+}_{\text{vision}_i}), S(\mathbf{e}^{+}_{\text{audio}_i}, \mathbf{e}_{\text{vision}})\big], 100\Big),
\end{aligned}
\tag{1}
$$

where S is calculated with the word-to-image attention mechanism described above. Intuitively this should push $\mathbf{e}_{\text{audio}}$, $\mathbf{e}_{\text{vision}}$ and the positive examples in $\mathbf{e}^{+}_{\text{audio}_{1:N_{\text{pos}}}}$ and $\mathbf{e}^{+}_{\text{vision}_{1:N_{\text{pos}}}}$ closer. At the same time it should push the negative examples in $\mathbf{e}^{-}_{\text{audio}_{1:N_{\text{neg}}}}$, $\mathbf{e}^{-}_{\text{vision}_{1:N_{\text{neg}}}}$ and $\mathbf{e}^{-}_{\text{vision background}_{1:N_{\text{neg}}}}$ away from these

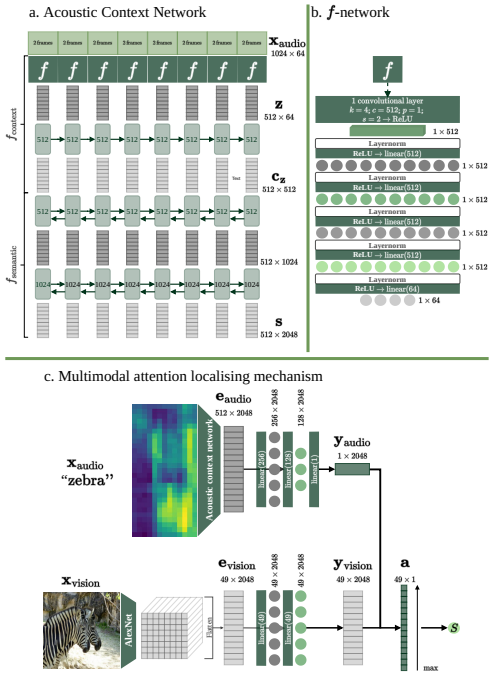

Figure 2: MATTNET *consists of (c) a vision and an audio network. The audio network consists of (a + b) an acoustic context network and an BiLSTM network. These networks are connected with a word-to-image attention mechanism.*

positives. I.e. it should learn the visual depiction of a spoken word class. Therefore, we need positive ($\mathbf{e}^{+}_{\text{audio}_{1:N_{\text{pos}}}}$, $\mathbf{e}^{+}_{\text{vision}_{1:N_{\text{pos}}}}$) and negative ($\mathbf{e}^{-}_{\text{audio}_{1:N_{\text{neg}}}}$, $\mathbf{e}^{-}_{\text{vision}_{1:N_{\text{neg}}}}$, $\mathbf{e}^{-}_{\text{vision background}_{1:N_{\text{neg}}}}$) pairs.

3.2. Few-shot pair mining

For few-shot training, we only have the small number of ground truth examples in the support set $\mathcal{S}$. This would not be sufficient to train the complex model described above. To overcome this, [10] proposed a pair mining scheme. We follow the same high-level idea to mine word-word and image-image pairs: use the audio examples in $\mathcal{S}$ and compare each example to each utterance in a large collection of unlabelled audio utterances, and similarly for the images. The mined items can then be used to construct more word-image pairs for training. While in [10] the unlabelled collection of audio consisted of isolated spoken words (which was artificially segmented), here we consider an unlabelled collection of audio consisting of full spoken utterances (a more realistic scenario).

The simple isolated-word comparison approach used in [10] is not adequate for this setting. We therefore employ another approach. We have a spoken word in our support set that we want to match to unlabelled unsegmented utterances in a large audio collection. This is similar to fuzzy string search, i.e. finding a set of strings that approximately match a given pattern. However, algorithms from string search are not directly applicable to speech since they operate on a discrete alphabet. To bridge this gap, we use QbERT (query-by-example with HuBERT). The idea is to encode speech as a set of discrete units that approximate phones. Then we can apply standard string search algorithms to find examples that match a given query word. We use HuBERT [24] to map input speech into discrete units. Then we divide the units into variable-duration phone-like segments following [25]. Finally, we search the dataset by aligning the query to each utterance using the Needleman-Wunsch algorithm [26]. An alternative to QbERT would have been to use dynamic time warping (DTW), as is done in [10]. However, in a developmental experiment we found that DTW achieves an isolated word retrieval F_1 score of 76.8% while QbERT achieves 98.7%.

Using QbERT, we compare each spoken utterance in an unlabelled collection of audio utterances to each spoken word example in $\mathcal{S}$. For each utterance, we take the highest score across the K word examples per class and rank the utterances from highest to lowest for each class. The first n utterances with the highest scores for a class are predicted to contain the spoken form of the word. Additionally, we use QbERT's predicted word segments to isolate matched words. To mine image pairs, we follow the same steps, but instead we use AlexNet [17] to extract a single embedding for each image and use cosine distance to compare image embeddings to one another. To get word-image pairs, we mine an image from the same predicted class as a segmented word. Negative pairs are taken from the positive pairs of other classes. We also mine an extra negative image from a set known to not contain any of the few-shot classes. Therefore, during the few-shot retrieval task, images containing few-shot classes can be distinguished from images that depicts none of the few-shot classes.

4. Experimental setup

Data: For our experiments, we use the SpokenCOCO Corpus [27] which consists of the MSCOCO [28] images with recorded spoken captions corresponding to the MSCOCO textual captions. We parametrise the utterances as mel-spectograms with a hop length of 10 ms, a window of 25 ms and 40 mel bins. These are truncated or zero-padded to 1024 frames. Images are resized to 224×224 pixels and normalised with means and variances calculated on ImageNet [29] with VGG [30].

We use the SpokenCOCO setup of [16] by splitting it into two main sets: a set which contains only the few-shot classes and a set that does not contain any of the few-shot classes (listed below). We refer to the latter set as background data. The set containing the few-shot classes is split into training and testing sets. We sample the support set $\mathcal{S}$ from this training set (§) and use the Montreal forced aligner [31] to isolate the few-shot words. For our mining approach (§), we need unlabelled audio and image data to mine pairs from; for this we use the remainder of the training data that does not include the support set. From these unlabelled collections, we mine pairs: the $n = 600$ highest ranking examples per class (§). These pairs are split into training and validation pairs.

Models: Figure illustrates our model, MATTNET (§). For the image branch, we use a pretrained adaptation of AlexNet [17] to get a sequence of per pixel embeddings. We use an adaptation of [18]'s audio network for the audio branch pretrained in an self-supervised manner on Libri-Light [32] and multilingual (English and Hindi) Places [33]. The model is initialised by pretraining it on the background data using the contrastive retrieval loss of [34]. We take $N_{\text{pos}} = 5$ and $N_{\text{neg}} = 11$ in Equation after fine-tuning it on the development pairs. For validation, we use the development set to get one positive image $\mathbf{x}^{+}_{\text{vision}}$ and one negative image $\mathbf{x}^{-}_{\text{vision}}$ for each developmental word-image ($\mathbf{x}_{\text{audio}}$, $\mathbf{x}_{\text{vision}}$) pair. The validation task measures whether the model will place $\mathbf{x}_{\text{vision}}$ and $\mathbf{x}^{+}_{\text{vision}}$ closer to $\mathbf{x}_{\text{audio}}$ than it would to $\mathbf{x}^{-}_{\text{vision}}$. We train all models with Adam [35] for

Table 1: *$P@N$ few-shot retrieval scores (%) on the five few-shot classes. K is the number of support-set examples per class.*

Model	K			
	5	10	50	100
Naive fine-tuned [16]	–	–	–	**52.5**
Oracle masking [16]	–	8.4±0.0	24.0±0.1	35.5±0.2
MATTNET	**44.4±0.0**	**43.4±0.1**	**40.2±0.0**	42.5±0.1
MATTNET, no mining	22.0±0.4	24.1±0.8	22.7±0.5	23.2±1.1
MATTNET, $L_{\text{train}} = 40$	39.7±0.6	–	–	–

100 epochs using the validation task for early stopping.

Few-shot tasks: Using the same five classes – *broccoli*, *zebra*, *fire hydrant*, *sheep* and *kite* – as [16], we evaluate our approach on two tasks (as explained in §): a traditional few-shot classification task and a few-shot retrieval task. For both these tasks, the K-shot L-way support set $\mathcal{S}$ contains K ground truth spoken word-image pairs for each of the $L = 5$ classes and is used to mine pairs for training and development. For testing the few-shot classification task, we sample 1000 episodes where each episode contains L spoken word queries $\mathbf{x}^q_{\text{audio}}$, one for each class, and a matching set $\mathcal{M}$ which contains one image $\mathbf{x}^{(i)}_{\text{vision}}$ for each class. However, in the few-shot retrieval task, instead of having one image per class, $\mathcal{M}$ consists of 5000 images $\mathbf{x}^{(i)}_{\text{vision}}$ where some depicts a few-shot class and others do not. Here, 20 query words are taken per class and averaged to get $\mathbf{x}^q_{\text{audio}}$. For each of the L queries $\mathbf{x}^q_{\text{audio}}$, these 5000 images are ranked from highest to lowest similarity. The precision at N ($P@N$) score is the proportion of images in the top N highest ranking images that are from the same class as $\mathbf{x}^q_{\text{audio}}$. N is the actual number of images in $\mathcal{M}$ that depicts the word class.

5. Experimental results

We first want to compare our work directly to that of [16]. Concretely, we compare to two of [16]'s models on the few-shot retrieval task. The first model we compare to is their naive model, which is trained on background classes and fine-tuned on $K = 100$ examples for each of the $L = 5$ classes. The second is their oracle masking model in which the contrastive loss used during fine-tuning ensures that a negative image does not contain any instance of the anchor few-shot class. The results are given in Table ([16] did not report scores for fewer than $K = 10$).

Comparing our full MATTNET model to the oracle masking approach, we see that we outperform [16] consistently across all values of K. Neither MATTNET or oracle masking works as well as the naive fine-tuned approach (line 1), but fine-tuning only works with a large number of shots. We also see that our approach (a bit surprisingly) delivers approximately the same few-shot retrieval scores as K increases, whereas [16]'s scores in line 2, increase. The reason for this is that the models retain contextual information which makes it difficult to disentangle the images containing a few-shot class from background images. However, our approach works particularly well with fewer shots.

To determine the mined pairs' contribution to this performance boost, we do an experiment where we do not update MATTNET on the mined pairs after pretraining it on the background data (not containing any of the few-shot classes). To test this model on the few-shot retrieval task, we use the indirect few-shot method of [8, 9]: each $\mathbf{x}^q_{\text{audio}}$ is compared to each $\mathbf{x}^{(j)}_{\text{audio}}$ in $\mathcal{S}$ to find the audio example closest to the query. The image $\mathbf{x}^{(j)}_{\text{vision}}$ corresponding to the closest $\mathbf{x}^{(j)}_{\text{audio}}$ is then used to calculate the similarity to each image $\mathbf{x}^{(i)}_{\text{vision}}$ in $\mathcal{M}$. Using mined pairs improves the scores with roughly 20% when comparing lines 3 and 4. In the final line of Table we start to investigate how our approach performs when, instead of using just five few-shot classes, we have 40 few-shot classes to learn with $K = 5$ shots. We see that we pay roughly 5% in $P@N$, but even when learning 40 classes, we still outperform the five-class oracle masking approach across all shots considered.

[1]Source code: Support set; MATTNET no mining; MATTNET: 100-shot 5-way; 50-shot 5-way; 10-shot 5-way; 5-shot 5-way; 5-shot 40-way

Table 2: *Few-shot word classification accuracy scores (%). We vary the number of shots per class K. Instead of only considering the five classes from [16], we also look at settings with 40 classes in the support and/or matching sets.*

Model	K	$L_{\mathcal{S}}$	$L_{\mathcal{M}}$	Few-shot accuracy
MATTNET, no mining	5	–	5	50.4
	10	–	5	48.0
	50	–	5	48.5
	100	–	5	47.7
MATTNET, with mining	5	5	5	65.4
	10	5	5	77.5
	50	5	5	86.6
	100	5	5	**90.9**
	5	40	5	63.7
	5	40	40	21.7

To further analyse the performance gains from mining and to also see what happens with more classes, we now consider the conventional few-shot word classification task (§). This task wasn't used in [16]. Table 2 shows that the few-shot classification scores increase as K increases when we use mined pairs. For the no pair mining method, the scores are worse and drops slightly as K increases. Looking at the five-shot case, we see that training on more classes ($L_{\mathcal{S}} = 40$) leads to a slightly lower classification score on the original five classes. The few-shot accuracy when tested on all 40 classes is 21.7%.

All together, our results sets a competitive multimodal baseline for both few-shot retrieval and word classification in settings where the number of shots is small.

6. Conclusion

Our goal was to do multimodal few-shot learning of natural images and spoken words. To do this, we proposed a novel few-shot pair mining method which we use in a new multimodal word-to-image attention model. For the lower-resource scenario where K is small, our model achieves higher few-shot retrieval scores than an existing model. We also set a competitive baseline for natural visually grounded few-shot word classification and present preliminary experiments indicating that our approach can be used on more than five few-shot classes. Future work will look into improving few-shot word classification on more classes.

7. Acknowledgements

We would like to thank DeepMind for funding Leanne Nortje and Google for funding Benjamin van Niekerk. We would also like to thank Tyler Miller and David Harwath for helping with the few-shot retrieval comparisons.

8. References

[1] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Commun.*, 2014.

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

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

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

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

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

[7] 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.*, 2012.

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

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

[10] ——, "Direct multimodal few-shot learning of speech and images," in *Proc. Interspeech*, 2021.

[11] 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.

[12] H. Kamper, A. Anastassiou, and K. Livescu, "Semantic query-by-example speech search using visual grounding," in *Proc. ICASSP*, 2019.

[13] K. Olaleye and H. Kamper, "Attention-based keyword localisation in speech using visual grounding," in *Proc. Interspeech*, 2021.

[14] G. Chrupała, "Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques," *J. Atrif. Intell. Res.*, 2022.

[15] D. Merkx, S. Scholten, S. L. Frank, M. Ernestus, and O. Scharenborg, "Modelling human word learning and recognition using visually grounded speech," *Cogn. Comput.*, 2022.

[16] T. Miller and D. Harwath, Dawid, "Exploring few-shot fine-tuning strategies for models of visually grounded speech," in *Proc. Interspeech*, 2022.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *ACM*, 2017.

[18] L. Nortje and H. Kamper, "Towards visually prompted keyword localisation for zero-resource spoken languages," in *Proc. SLT*, 2022.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016.

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

[21] 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.

[22] 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.

[23] 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.

[24] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *ACM*, 2021.

[25] H. Kamper and B. van Niekerk, "Towards unsupervised phone and word segmentation using self-supervised vector-quantized neural networks," in *Proc. Interspeech*, 2021.

[26] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J. Mol. Biol.*, 1970.

[27] W.-N. Hsu, D. Harwath, C. Song, and J. Glass, "Text-free image-to-speech synthesis using learned segmental units," in *Proc ACL*, 2021.

[28] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common objects in context." in Proc ECCV, 2014.

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009.

[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015.

[31] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using Kaldi," in *Proc. Interspeech*, 2017.

[32] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P. Mazare, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, "Libri-light: A Benchmark for ASR with limited or no supervision," in *Proc. ICASSP*, 2020.

[33] D. Harwath, G. Chuang, and J. Glass, "Vision as an Interlingua: Learning multilingual semantic embeddings of untranscribed speech," in *Proc. ICASSP*, 2018.

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

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