Unsupervised vs. Transfer Learning for Multimodal One-Shot Matching of Speech and Images

Interspeech 2020

authors

Leanne Nortje [nortjeleanne@gmail.com]

Herman Kamper [kamperh@sun.ac.za]

institute

E&E Engineering, Stellenbosch University, South Africa

» Problem

» Problem

- ▷ Vision/speech processing systems require large amounts of labelled data.

» Problem

- Vision/speech processing systems require large amounts of labelled data.
- **Where to find a solution**

» Problem

- Vision/speech processing systems require large amounts of labelled data.
- **Where to find a solution**: Children learn from few examples.

» Problem

- Vision/speech processing systems require large amounts of labelled data.
- **Where to find a solution**: Children learn from few examples.

words $\Longleftrightarrow$ visual objects

» How do Children Learn?

How do Children Learn?

“ice-cream”

» How do Children Learn?

“ice-cream”

“cookie”

» How do Children Learn?

“ice-cream”

“cookie”

“broccoli”

» How do Children Learn?

“ice-cream”

“cookie”

“broccoli”

Which picture is that of an “ice-cream”?

» How do Children Learn?

“ice-cream”

“cookie”

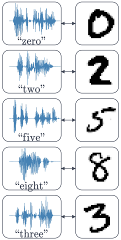
“broccoli”

Which picture is that of an “ice-cream”?

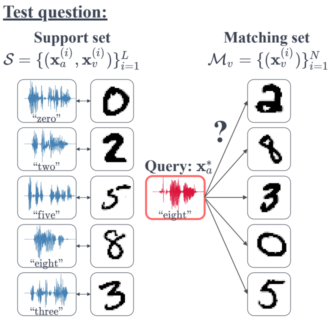
» Multimodal One-Shot Matching

Support set

$\mathcal{S} = \{(\mathbf{x}_a^{(i)}, \mathbf{x}_v^{(i)})\}_{i=1}^{L}$

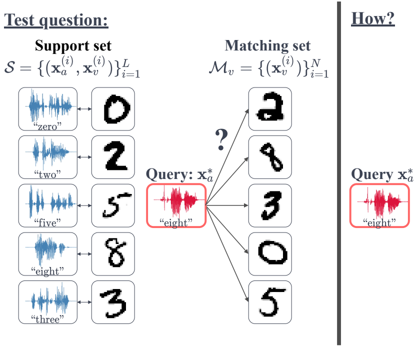


» Multimodal One-Shot Matching



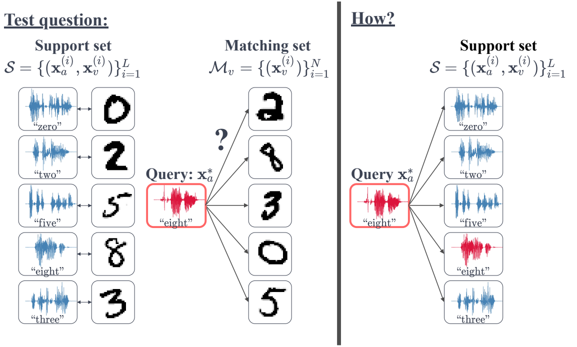
Which image matches the query **"eight"**?

» Multimodal One-Shot Matching



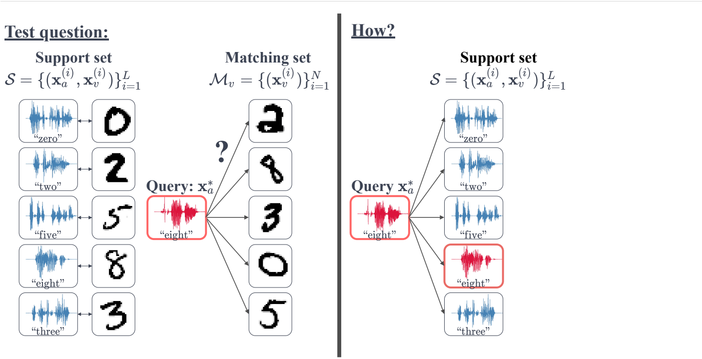
Which image matches the query **"eight"**?

» Multimodal One-Shot Matching



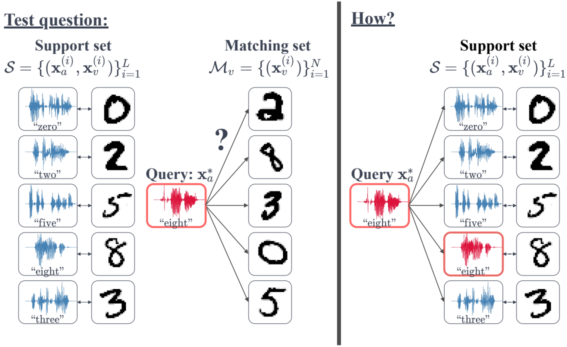
Which image matches the query **"eight"**?

» Multimodal One-Shot Matching



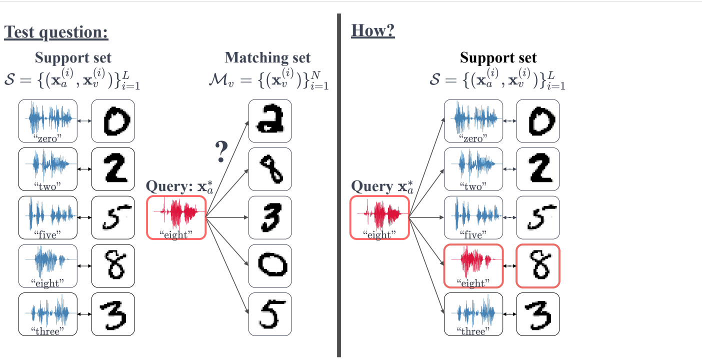
Which image matches the query **"eight"**?

» Multimodal One-Shot Matching



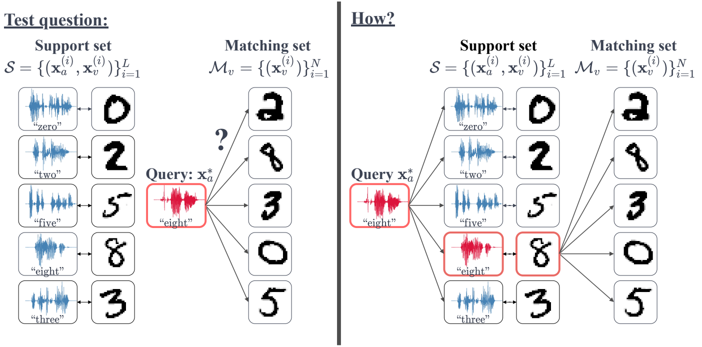
Which image matches the query **"eight"**?

» Multimodal One-Shot Matching



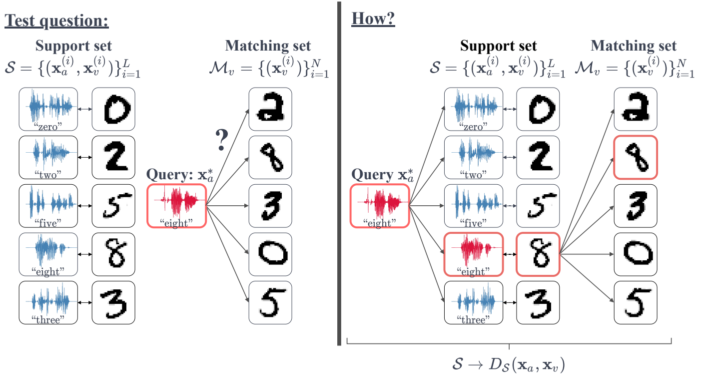
Which image matches the query **"eight"**?

» Multimodal One-Shot Matching



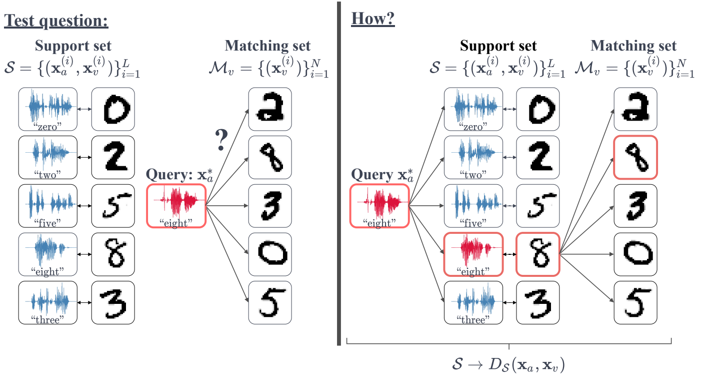
Which image matches the query **"eight"**?

» Multimodal One-Shot Matching



Which image matches the query **"eight"**?

» Multimodal One-Shot Matching



Multimodal one-shot learning: learn from one cross-modal paired example

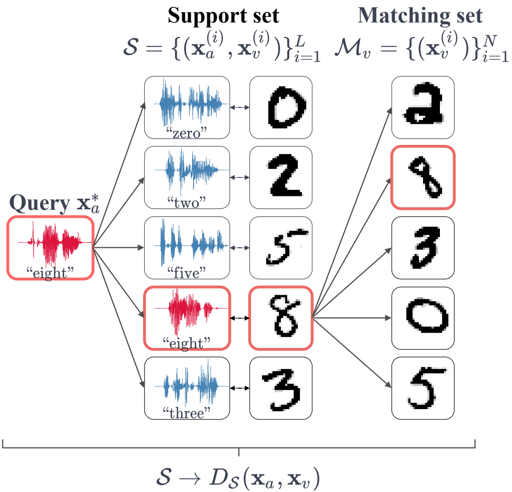
» Multimodal One-Shot Matching

Multimodal one-shot learning: learn from **one** cross-modal paired example.

$$\Downarrow$$

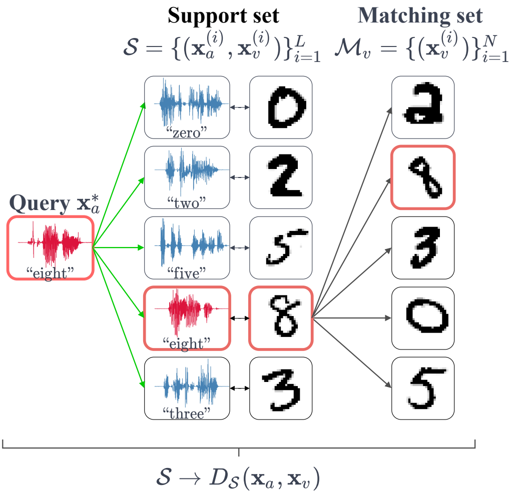
Multimodal few-shot learning: learn from a **few** cross-modal paired examples.

» Multimodal One-Shot Matching



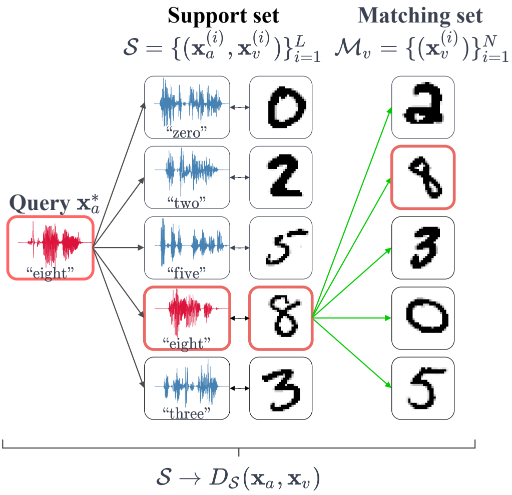
Our approach: a support set + two unimodal comparisons

» Multimodal One-Shot Matching



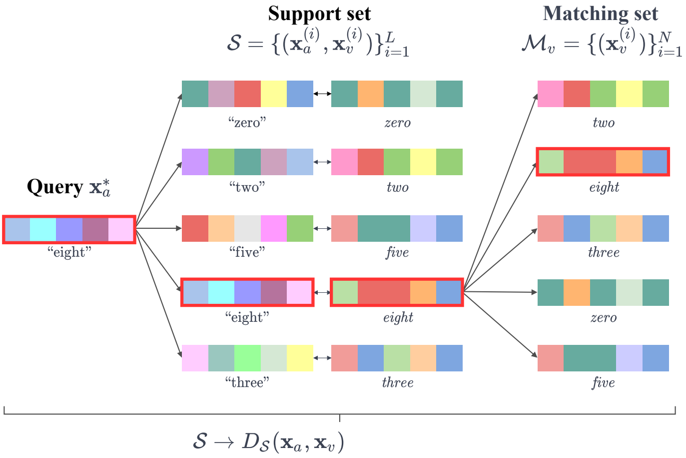
Our approach: a support set + two unimodal comparisons

» Multimodal One-Shot Matching



Our approach: a support set + two unimodal comparisons

» Multimodal One-Shot Matching



Speech comparison: cosine distance between two word representations

Image comparison: cosine distance between two image representations

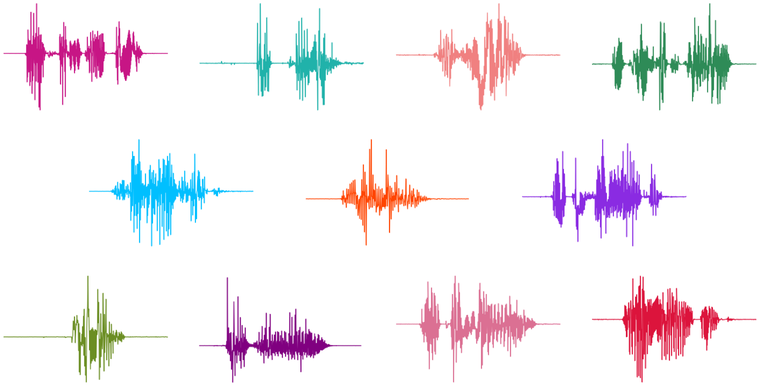
» Unsupervised Learning

» Unsupervised Learning

- Trained on unlabelled in-domain data.

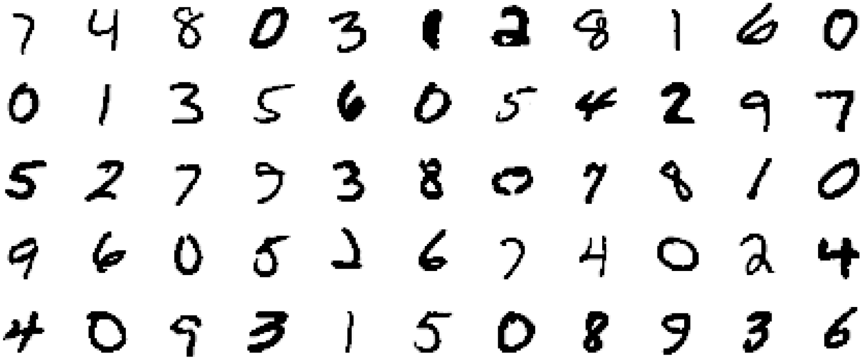
» Unsupervised Learning

- ▷ Trained on unlabelled in-domain data.
- ▷ Unlabelled in-domain data: TIDigits

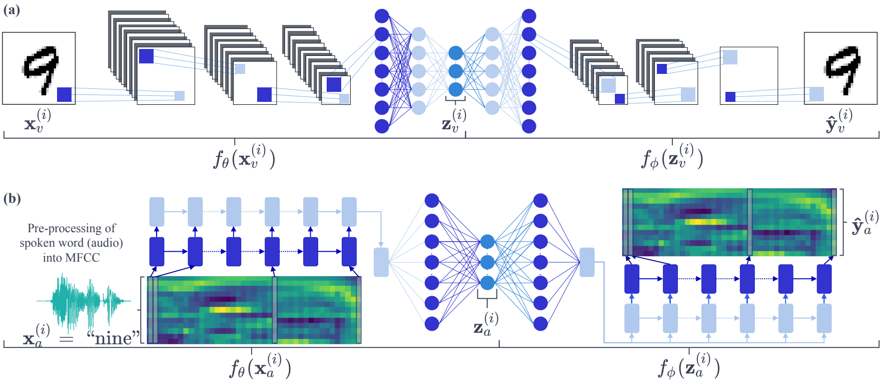


» Unsupervised Learning

- ▷ Trained on unlabelled in-domain data.
- ▷ Unlabelled in-domain data: TIDigits and MNIST

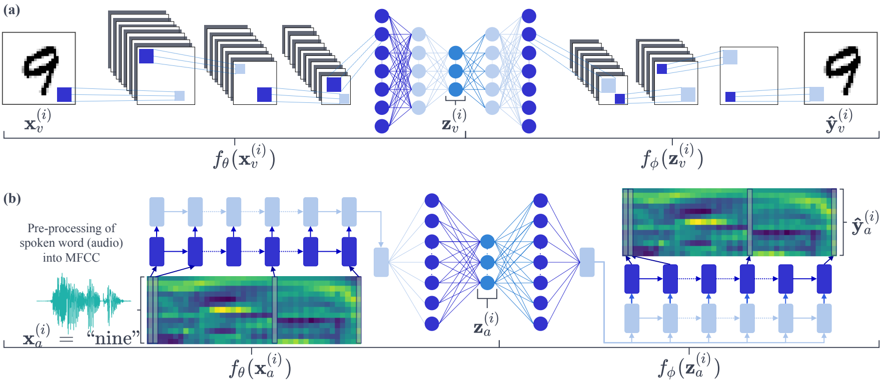


» Unsupervised Learning



» Unsupervised Learning

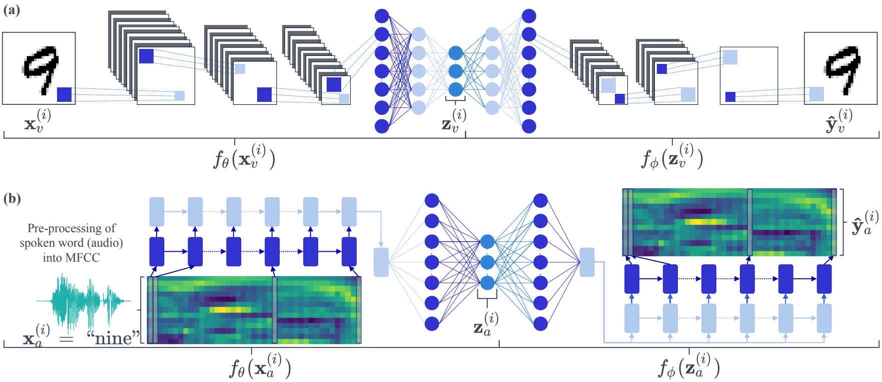
Autoencoder-like model architectures:



» Unsupervised Learning

Autoencoder-like model architectures:

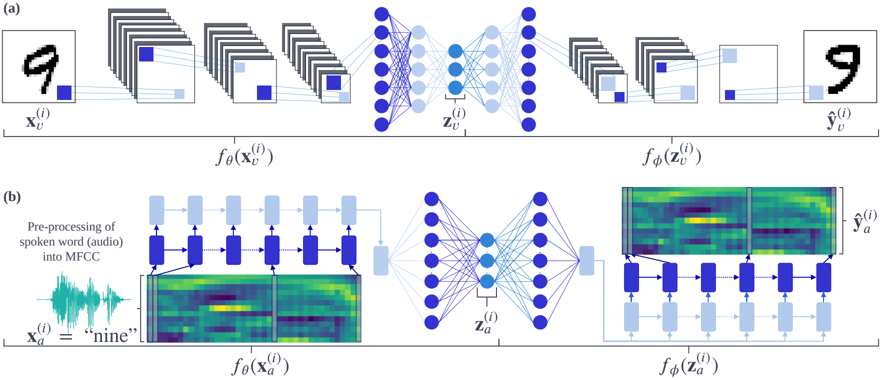
- ▷ Autoencoder (AE)



» Unsupervised Learning

Autoencoder-like model architectures:

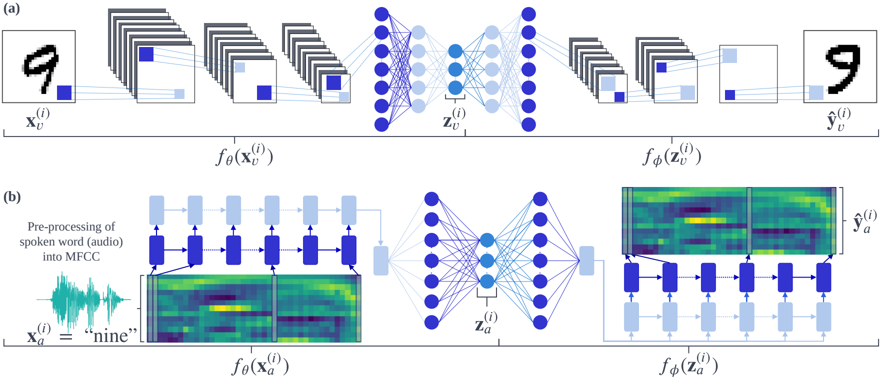
- ▷ Autoencoder (AE)
- ▷ Correspondence autoencoder (CAE)



» Unsupervised Learning

Autoencoder-like model architectures:

- ▷ Autoencoder (AE)
- ▷ Correspondence autoencoder (CAE) (unsupervised within-modality pairs).



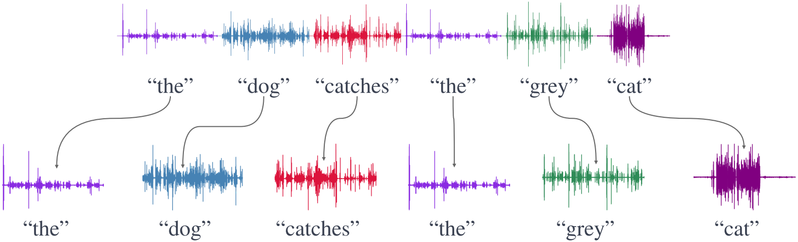
» Transfer Learning

» Transfer Learning

- Trained on labelled background data.

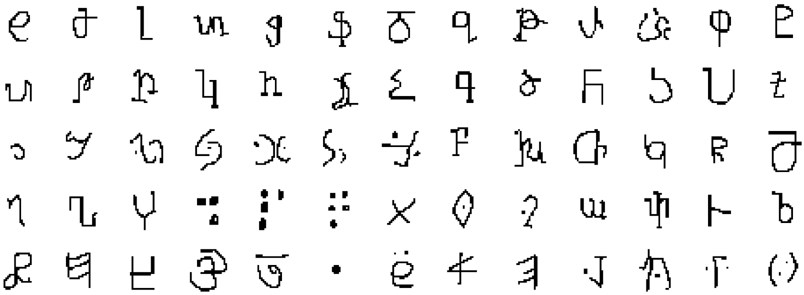
» Transfer Learning

- ▷ Trained on labelled background data.
- ▷ Labelled background data: Buckeye



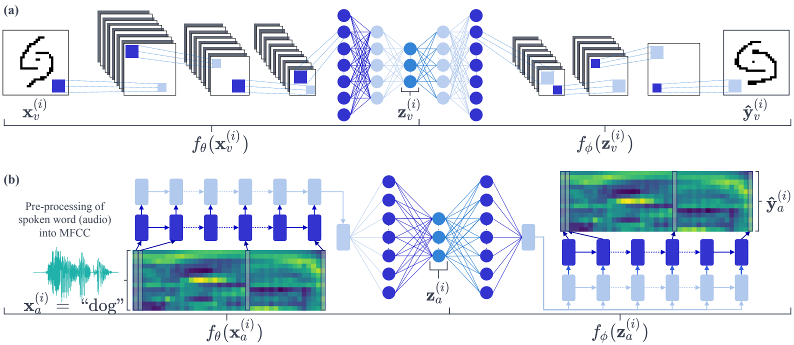
» Transfer Learning

- ▷ Trained on labelled background data.
- ▷ Labelled background data: Buckeye and **Omniglot**.



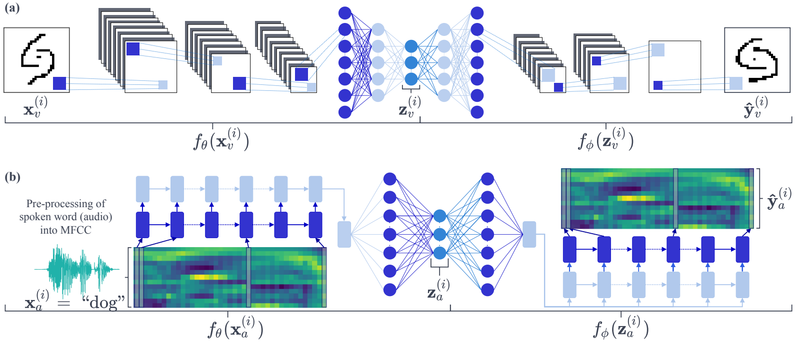
» Transfer Learning

▷ A transfer learned variant of the unsupervised CAE:



» Transfer Learning

- A transfer learned variant of the unsupervised CAE:
- Trained on **ground truth pairs**.



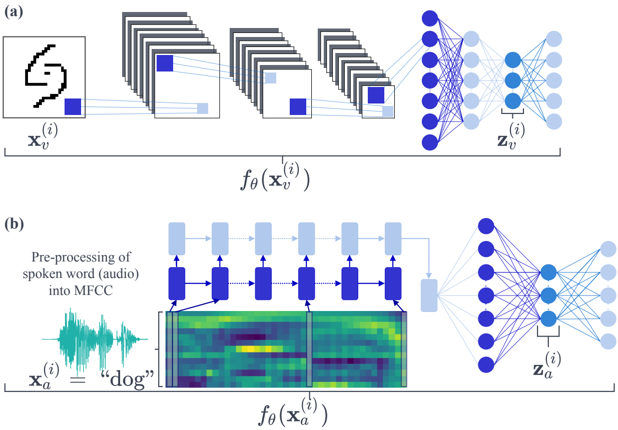
» Transfer Learning

Multimodal few-shot models from Eloff et al. [1]:

» Transfer Learning

Multimodal few-shot models from Eloff et al. [1]:

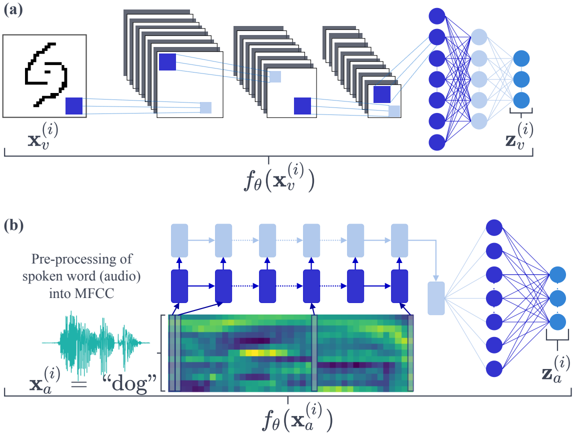
- ▷ Classifiers



» Transfer Learning

Multimodal few-shot models from Eloff et al. [1]:

- Classifiers and
- Siamese Triplet networks.



» K-shot Multimodal Speech and Image Matching

Model		11-way accuracy (%) one-shot	11-way accuracy (%) five-shot
Baseline	DTW + Pixels	31.80	41.88
Transfer learning models	**Classifier**	**56.80 $\pm$ 1.19**	**59.67 $\pm$ 1.73**
	Siamese	54.83 $\pm$ 1.80	59.25 $\pm$ 0.79
	CAE	46.60 $\pm$ 0.69	53.82 $\pm$ 1.07
Unsupervised models	AE	28.99 $\pm$ 0.84	38.68 $\pm$ 1.51
	CAE	42.75 $\pm$ 0.62	52.15 $\pm$ 0.69

» K-shot Multimodal Speech and Image Matching

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

» K-shot Multimodal Speech and Image Matching

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

» Could Unsupervised and Transfer Learning be Complementary?

» Could Unsupervised and Transfer Learning be Complementary?

- **Classifier pairs**: We find unsupervised training pairs using transfer learned classifiers.

» Could Unsupervised and Transfer Learning be Complementary?

- **Classifier pairs**: We find unsupervised training pairs using transfer learned classifiers.

- **CAE with classifier pairs**: Train an unsupervised CAE using these classfier pairs.

» Could Unsupervised and Transfer Learning be Complementary?

- **Classifier pairs**: We find unsupervised training pairs using transfer learned classifiers.

- **CAE with classifier pairs**: Train an unsupervised CAE using these classfier pairs.

- **Transfer learning + CAE fine-tuning**: Pretrain a CAE on ground truth background pairs and then train the CAE on these classifier pairs.

» Towards Combined Transfer and Unsupervised Learning

Model	11-way accuracy (%)	
	one-shot	five-shot
Baseline: DTW + Pixels	31.80	41.88
Transfer learning: Classifier	**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

» Towards Combined Transfer and Unsupervised Learning

Model	11-way accuracy (%) one-shot	five-shot
Baseline: DTW + Pixels	31.80	41.88
Transfer learning: Classifier	**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

» Towards Combined Transfer and Unsupervised Learning

Model	11-way accuracy (%)	
	one-shot	five-shot
Baseline: DTW + Pixels	31.80	41.88
Transfer learning: Classifier	**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

» Towards Combined Transfer and Unsupervised Learning

Model	11-way accuracy (%)	
	one-shot	five-shot
Baseline: DTW + Pixels	31.80	41.88
Transfer learning: Classifier	**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**

» Conclusions

Conclusions

- Transfer learning outperforms unsupervised learning.

Conclusions

- Transfer learning outperforms unsupervised learning.

- Unsupervised learning can be improved by using transfer learning.

Conclusions

- Transfer learning outperforms unsupervised learning.
- Unsupervised learning can be improved by using transfer learning.
- Idealised experiments show the promise of unsupervised learning.

References

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

» Acknowledgements

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