A Correspondence Variational Autoencoder for Unsupervised Acoustic Word Embeddings

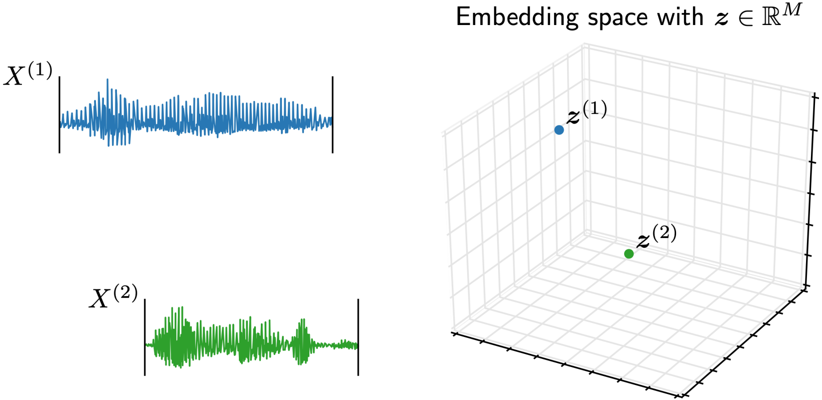
Puyuan Peng[1] Herman Kamper[2] Karen Livescu[3]

[1]University of Chicago, USA

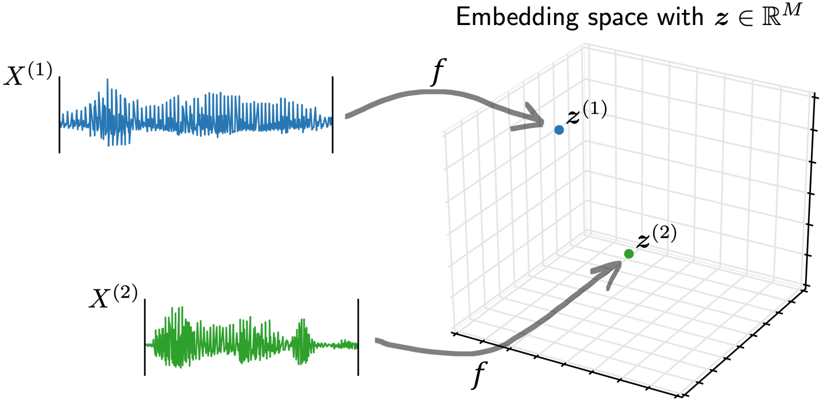
[2]Stellenbosch University, South Africa

[3]Toyota Technological Institute at Chicago, USA

Background: Acoustic word embeddings (AWEs)



Background: Acoustic word embeddings (AWEs)



Background: Why unsupervised acoustic word embeddings

Background: Why unsupervised acoustic word embeddings

- Why unsupervised: Most of the spoken languages in the world are under-resourced

Background: Why unsupervised acoustic word embeddings

- Why unsupervised: Most of the spoken languages in the world are under-resourced
- Why acoustic word embeddings: useful for downstream applications – unsupervised term discovery [Kamper et al., 2016], query-by-example search [Settle et al., 2017]

Approach

Approach

1. Unsupervised term discovery (UTD) system to provide training data [Park and Glass, 2007]

Approach

1. Unsupervised term discovery (UTD) system to provide training data [Park and Glass, 2007]

2. VAE for generative pre-training

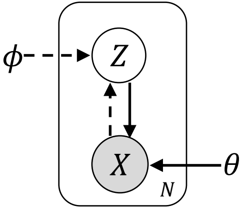
Approach

1. Unsupervised term discovery (UTD) system to provide training data [Park and Glass, 2007]

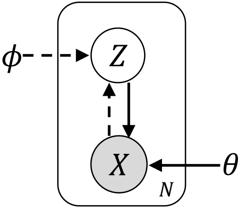
2. VAE for generative pre-training

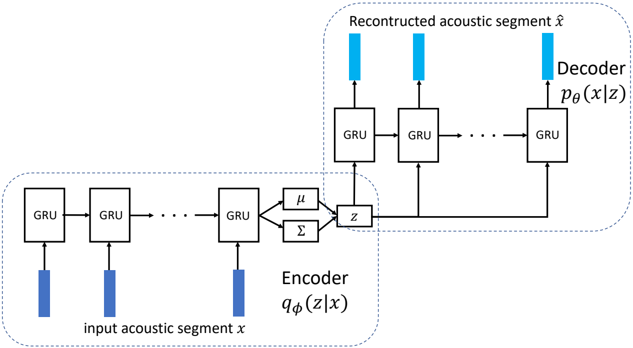
3. Maximal sampling correspondence VAE for training

VAEs for acoustic word embedding



VAEs for acoustic word embedding





Same-different word discrimination task [Carlin et al., 2011]

Same-different word discrimination task [Carlin et al., 2011]

1. Calculate cosine similarity between embeddings

Same-different word discrimination task [Carlin et al., 2011]

1. Calculate cosine similarity between embeddings
2. Classify two acoustic segments as being same or different type based on some threshold

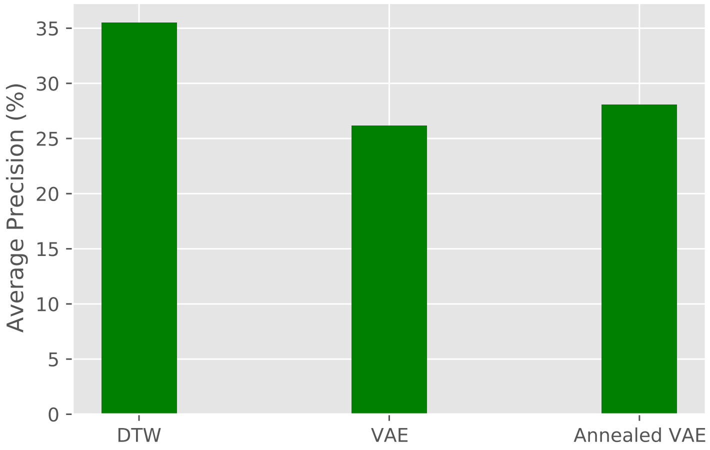
Same-different word discrimination task [Carlin et al., 2011]

1. Calculate cosine similarity between embeddings
2. Classify two acoustic segments as being same or different type based on some threshold
3. Vary the threshold to get precision-recall curve

Same-different word discrimination task [Carlin et al., 2011]

1. Calculate cosine similarity between embeddings
2. Classify two acoustic segments as being same or different type based on some threshold
3. Vary the threshold to get precision-recall curve
4. Report area under the curve i.e. average precision (AP)

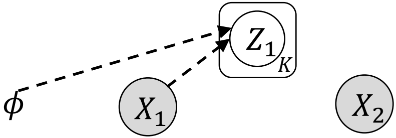
Performance of VAEs



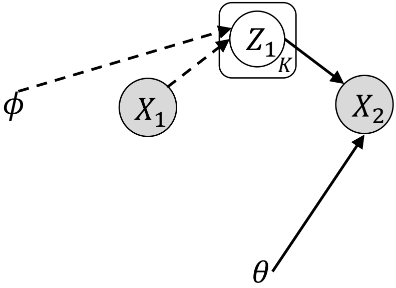
Correspondence VAE (CVAE)

Correspondence VAE (CVAE)

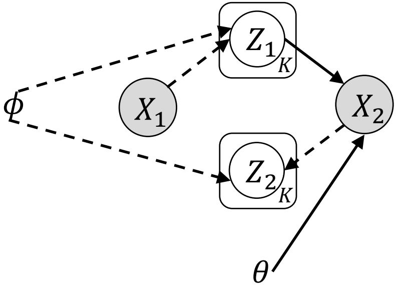
Correspondence VAE (CVAE)



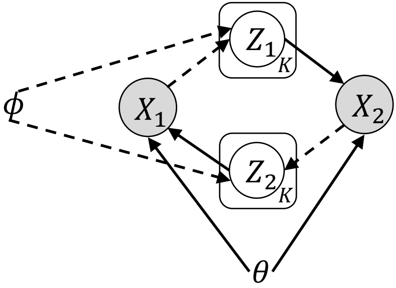
Correspondence VAE (CVAE)



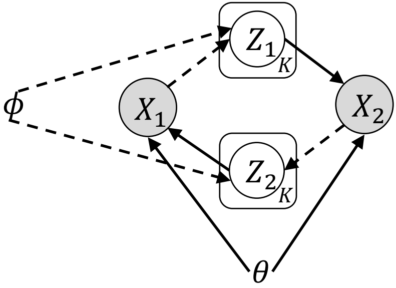
Correspondence VAE (CVAE)



Correspondence VAE (CVAE)

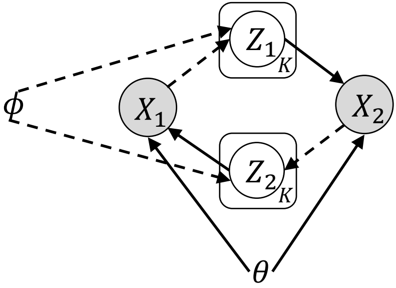


Correspondence VAE (CVAE)



For data pair (x_1, x_2), the objective is

Correspondence VAE (CVAE)



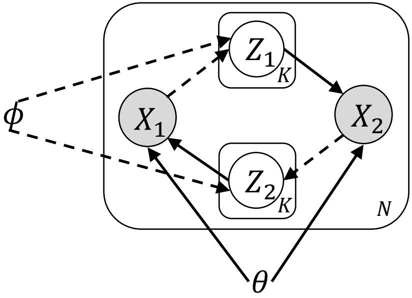
For data pair (x_1, x_2), the objective is

$$J_{\text{CVAE}} = \frac{1}{K} \sum_{k_2=1}^{K} \log p_\theta(x_2|z_1^{(k_1)}) - D_{KL}(q_\phi(Z_1|x_1)||p(Z_1))$$

$$+\frac{1}{K} \sum_{k_1=1}^{K} \log p_\theta(x_1|z_2^{(k_2)}) - D_{KL}(q_\phi(Z_2|x_2)||p(Z_2))$$

$$z_1^{(k_1)} \overset{\text{i.i.d}}{\sim} q_\phi(Z_1|x_1), \quad z_2^{(k_2)} \overset{\text{i.i.d}}{\sim} q_\phi(Z_2|x_2)$$

Correspondence VAE (CVAE)



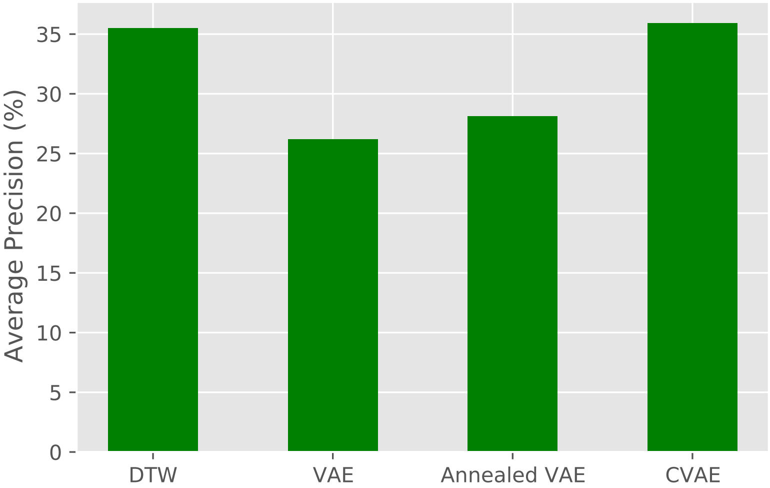
For data pair (x_1, x_2), the objective is

$$J_{\text{CVAE}} = \frac{1}{K}\sum_{k_1=1}^{K} \log p_\theta(x_2|z_1^{(k_1)}) - D_{KL}(q_\phi(Z_1|x_2)||p(Z_1))$$

$$+\frac{1}{K}\sum_{k_2=1}^{K} \log p_\theta(x_1|z_2^{(k_2)}) - D_{KL}(q_\phi(Z_2|x_1)||p(Z_2))$$

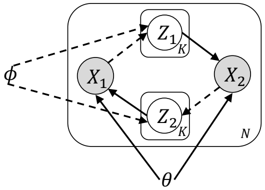
$$z_1^{(k_1)} \overset{\text{i.i.d}}{\sim} q_\phi(Z_1|x_1), \quad z_2^{(k_2)} \overset{\text{i.i.d}}{\sim} q_\phi(Z_2|x_2)$$

Performance of the Correspondence VAE (CVAE)

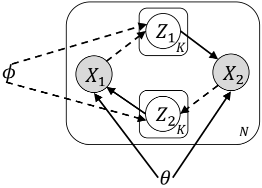


Maximal Sampling Correspondence VAE (MCVAE)

Maximal Sampling Correspondence VAE (MCVAE)

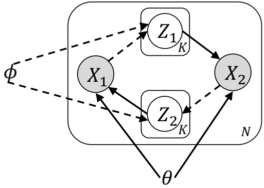


Maximal Sampling Correspondence VAE (MCVAE)



Maximal Sampling

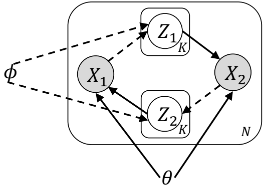
Maximal Sampling Correspondence VAE (MCVAE)



Maximal Sampling

1. Draw samples from posterior:
$z_1^{(1)}, z_1^{(2)}, \cdots, z_1^{(K)} \overset{\text{i.i.d}}{\sim} q_\phi(Z_1|x_1)$

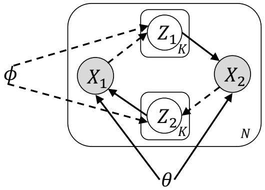
Maximal Sampling Correspondence VAE (MCVAE)



Maximal Sampling

1. Draw samples from posterior:
$z_1^{(1)}, z_1^{(2)}, \cdots, z_1^{(K)} \overset{\text{i.i.d}}{\sim} q_\phi(Z_1|x_1)$
2. Get candidate likelihood models:
$p_\theta(\cdot|z_1^{(1)}), p_\theta(\cdot|z_1^{(2)}), \cdots, p_\theta(\cdot|z_1^{(K)})$

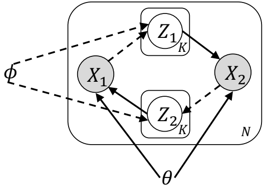
Maximal Sampling Correspondence VAE (MCVAE)



Maximal Sampling

1. Draw samples from posterior: $z_1^{(1)}, z_1^{(2)}, \cdots, z_1^{(K)} \overset{\text{i.i.d}}{\sim} q_\phi(Z_1|x_1)$
2. Get candidate likelihood models: $p_\theta(\cdot|z_1^{(1)}), p_\theta(\cdot|z_1^{(2)}), \cdots, p_\theta(\cdot|z_1^{(K)})$
3. Choose the one that achieves $\max_{k_1} \log p_\theta(x_2|z_1^{(k_1)})$ as the likelihood model

Maximal Sampling Correspondence VAE (MCVAE)

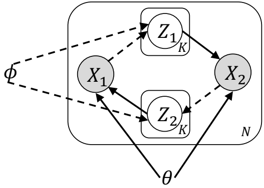


Maximal Sampling

1. Draw samples from posterior: $z_1^{(1)}, z_1^{(2)}, \cdots, z_1^{(K)} \overset{\text{i.i.d}}{\sim} q_\phi(Z_1|x_1)$
2. Get candidate likelihood models: $p_\theta(\cdot|z_1^{(1)}), p_\theta(\cdot|z_1^{(2)}), \cdots, p_\theta(\cdot|z_1^{(K)})$
3. Choose the one that achieves $\max_{k_1} \log p_\theta(x_2|z_1^{(k_1)})$ as the likelihood model

For data pair (x_1, x_2), the objective is

Maximal Sampling Correspondence VAE (MCVAE)



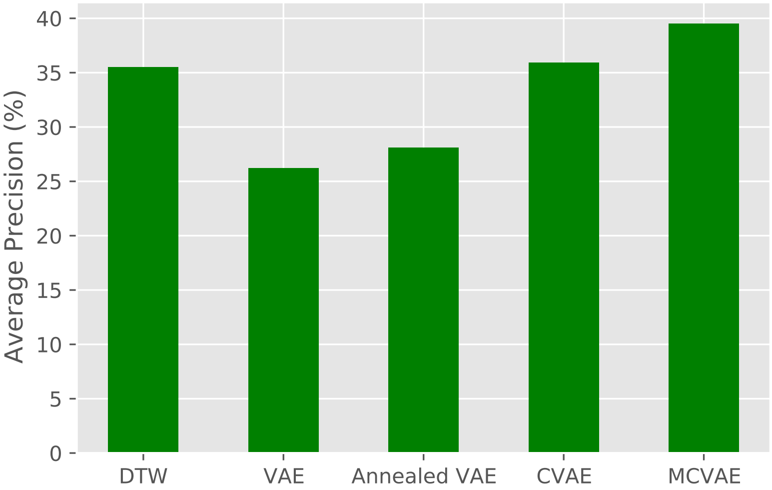
Maximal Sampling

1. Draw samples from posterior: $z_1^{(1)}, z_1^{(2)}, \cdots, z_1^{(K)} \overset{\text{i.i.d}}{\sim} q_\phi(Z_1|x_1)$
2. Get candidate likelihood models: $p_\theta(\cdot|z_1^{(1)}), p_\theta(\cdot|z_1^{(2)}), \cdots, p_\theta(\cdot|z_1^{(K)})$
3. Choose the one that achieves $\max_{k_1} \log p_\theta(x_2|z_1^{(k_1)})$ as the likelihood model

For data pair (x_1, x_2), the objective is

$$J_{\text{MCVAE}} = \max_{k_1} \log p_\theta(x_2|z_1^{(k_1)}) - D_{KL}(q_\phi(Z_2|x_2)||p(Z_2))$$
$$+ \max_{k_2} \log p_\theta(x_1|z_2^{(k_2)}) - D_{KL}(q_\phi(Z_1|x_1)||p(Z_1))$$
$$z_1^{(k_1)} \overset{\text{i.i.d}}{\sim} q_\phi(Z_1|x_1), \quad z_2^{(k_2)} \overset{\text{i.i.d}}{\sim} q_\phi(Z_2|x_2)$$

Performance of the MCVAE



Comparison with prior work

Table 1: Unsupervised word discrimination performance.

	Average Precision (%)	
Model	English	Xitsonga
SiameseRNN [Settle and Livescu, 2016]		
CAE-RNN [Kamper, 2019]		
MCVAE (ours)		
DTW alignment		

Comparison with prior work

Table 2: Unsupervised word discrimination performance.

Model	Average Precision (%)	
	English	Xitsonga
SiameseRNN [Settle and Livescu, 2016]	17.5	25.1
CAE-RNN [Kamper, 2019]	35.5	32.2
MCVAE (ours)	**39.5**	**44.4**
DTW alignment	35.9	28.1

Fine-tunning with labeled data

Fine-tunning with labeled data

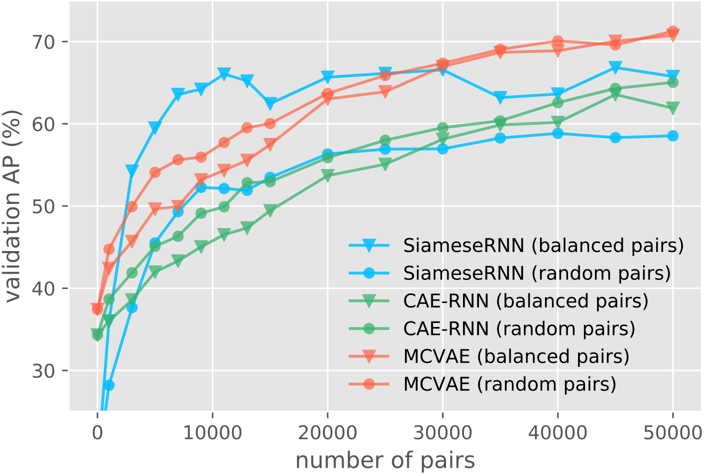
1. Training pair distribution: balanced pairs; random pairs

Fine-tunning with labeled data

1. Training pair distribution: balanced pairs; random pairs
2. Amount of data: from 1k to 50k

Fine-tunning with labeled data

1. Training pair distribution: balanced pairs; random pairs
2. Amount of data: from 1k to 50k



Conclusion

1. Propose the maximal sampling correspondence VAE (MCVAE) – a probabilistic approach for unsupervised acoustic word embeddings
2. MCVAE achieves state-of-the-art performance on unsupervised AWE task
3. MCVAE is robust to the amount and distribution of training data