(Outrageously*) Low-Resource Speech Processing

NLP @ Deep Learning Indaba, Kenya, 2019

Herman Kamper

E&E Engineering, Stellenbosch University, South Africa

http://www.kamperh.com/

(Outrageously*) Low-Resource Speech Processing

NLP @ Deep Learning Indaba, Kenya, 2019

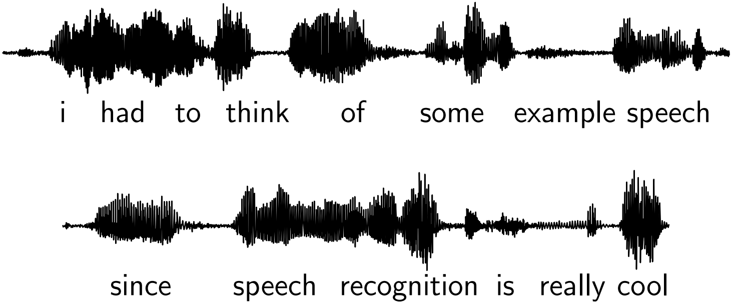
Herman Kamper

E&E Engineering, Stellenbosch University, South Africa

http://www.kamperh.com/

*Title plagiarised from Jade Abbott's DLI talk

Supervised speech recognition

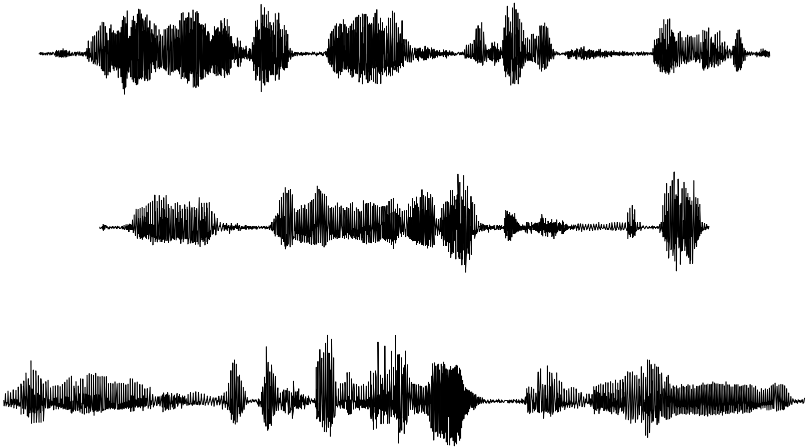


Unsupervised ("zero-resource") speech processing

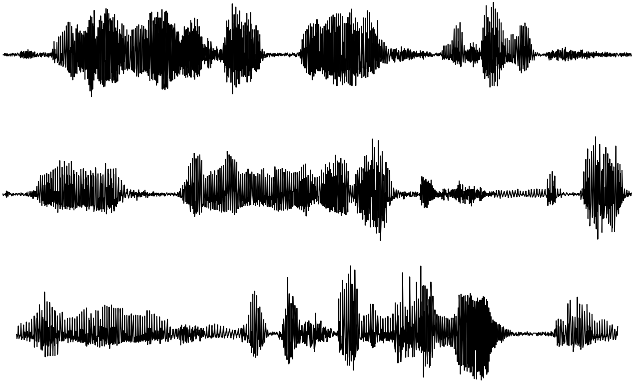
My problem: What can we learn if we do not have any labels?

Unsupervised ("zero-resource") speech processing

My problem: What can we learn if we do not have any labels?



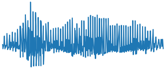
Example: Query-by-example speech search

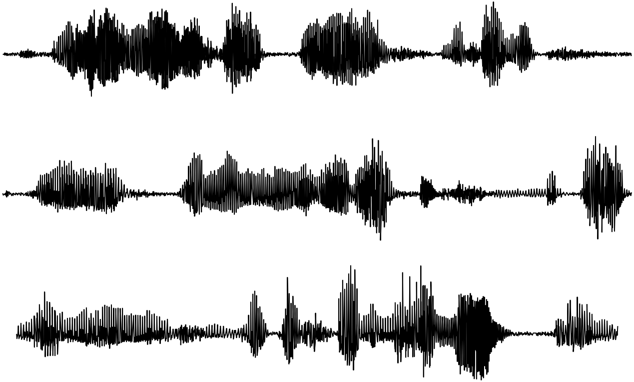


[Jansen and Van Durme, Interspeech'12]

Example: Query-by-example speech search

Spoken query:

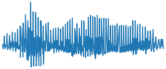


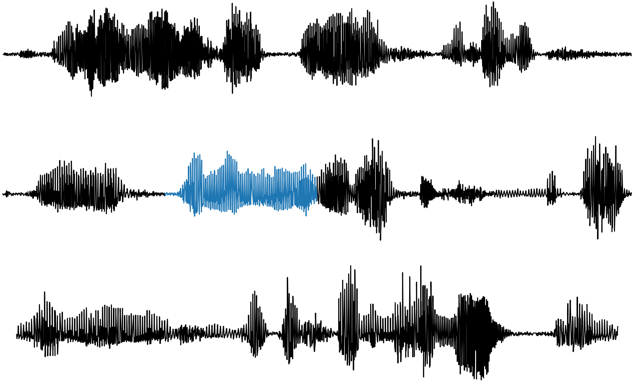


[Jansen and Van Durme, Interspeech'12]

Example: Query-by-example speech search

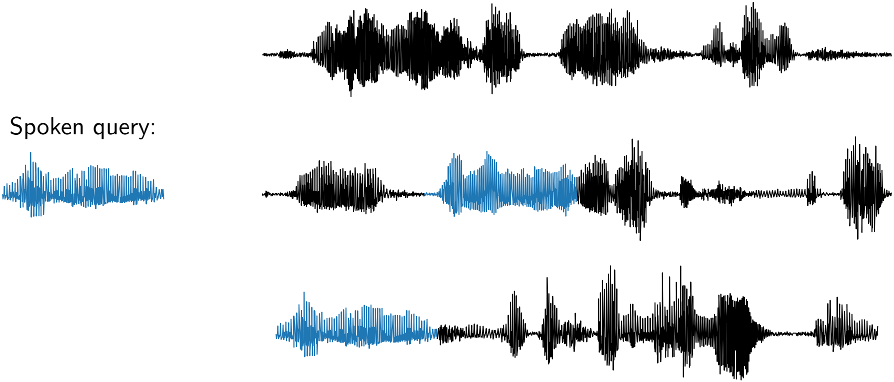
Spoken query:





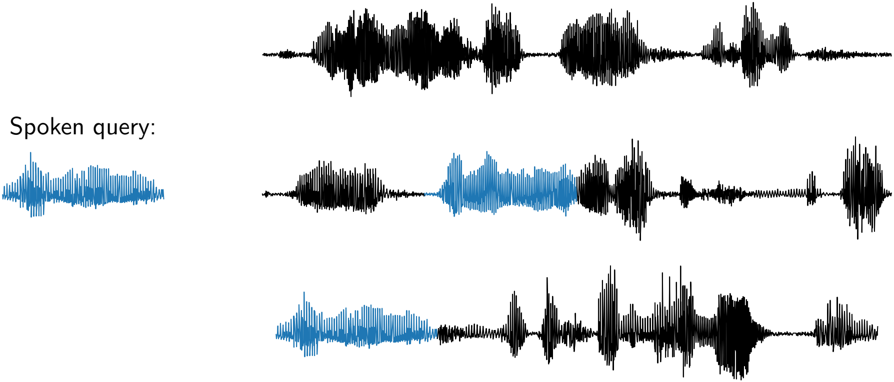
[Jansen and Van Durme, Interspeech'12]

Example: Query-by-example speech search



[Jansen and Van Durme, Interspeech'12]

Example: Query-by-example speech search



Useful speech system, not requiring any transcribed speech

[Jansen and Van Durme, Interspeech'12]

Outrageously low-resource = unsupervised speech processing (outline)

Outrageously low-resource = unsupervised speech processing (outline)

- Why is this problem so important?
  Will try to convince you that this is (one of) the most fundamental machine learning problems, with real impactful applications

Outrageously low-resource = unsupervised speech processing (outline)

- Why is this problem so important?
  Will try to convince you that this is (one of) the most fundamental machine learning problems, with real impactful applications

- What are the key ideas needed to tackle this problem?
  Hopefully you will get some useful tools

Outrageously low-resource = unsupervised speech processing (outline)

- Why is this problem so important?
  Will try to convince you that this is (one of) the most fundamental machine learning problems, with real impactful applications

- What are the key ideas needed to tackle this problem?
  Hopefully you will get some useful tools

- What is still missing?
  What are the open problems and research questions which still need to be solved (according to me)

Why is this problem so important?

1. A fundamental machine learning problem

Problems in unsupervised speech processing:

1. A fundamental machine learning problem

Problems in unsupervised speech processing:

- Learning useful representations from unlabelled speech
- Segmenting, clustering and discovering longer-spanning (word- or phrase-like) patterns

1. A fundamental machine learning problem

Problems in unsupervised speech processing:

- Learning useful representations from unlabelled speech
- Segmenting, clustering and discovering longer-spanning (word- or phrase-like) patterns
- Combined problem of perception, structure, continuous and discrete variables

1. A fundamental machine learning problem

Problems in unsupervised speech processing:

- Learning useful representations from unlabelled speech
- Segmenting, clustering and discovering longer-spanning (word- or phrase-like) patterns
- Combined problem of perception, structure, continuous and discrete variables

"The goal of machine learning is to develop methods that can automatically detect patterns in data ..." — *Murphy*

"Extract important patterns and trends, and understand 'what the data says' ..." — *Hastie, Tibshirani, Friedman*

"The problem of searching for patterns in data is ... fundamental ..." — *Bishop*

1. A fundamental machine learning problem

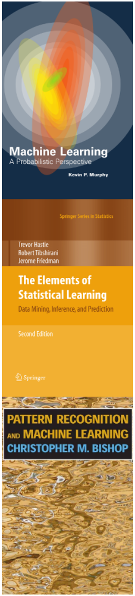
Problems in unsupervised speech processing:

- Learning useful representations from unlabelled speech
- Segmenting, clustering and discovering longer-spanning (word- or phrase-like) patterns
- Combined problem of perception, structure, continuous and discrete variables

"The goal of machine learning is to develop methods that can automatically detect patterns in data . . . " — *Murphy*

"Extract important patterns and trends, and understand 'what the data says' . . . " — *Hastie, Tibshirani, Friedman*

"The problem of searching for patterns in data is . . . fundamental . . . " — *Bishop*



2. Universal speech technology

“Imagine a world in which every single human being can freely share in the sum of all knowledge.”

2. Universal speech technology

"Imagine a world in which every single human being can freely share in the sum of all knowledge."

— *Mission statement stolen from Laura Martinus*

2. Universal speech technology

"Imagine a world in which every single human being can freely share in the sum of all knowledge."

— Mission statement stolen from Laura Martinus

— Who stole it from the Wikimedia Foundation

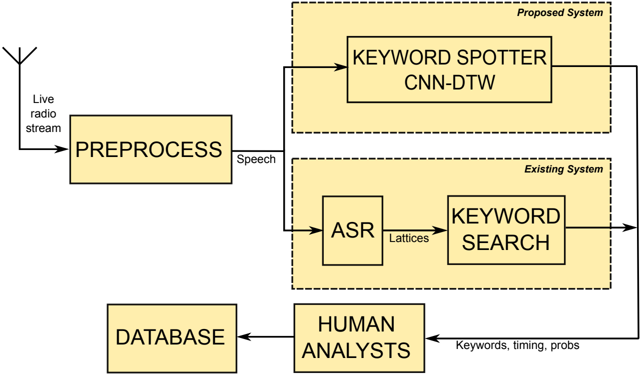
https://15.wikipedia.org/endowment.html

2. Universal speech technology

UN Pulse Lab, Kampala

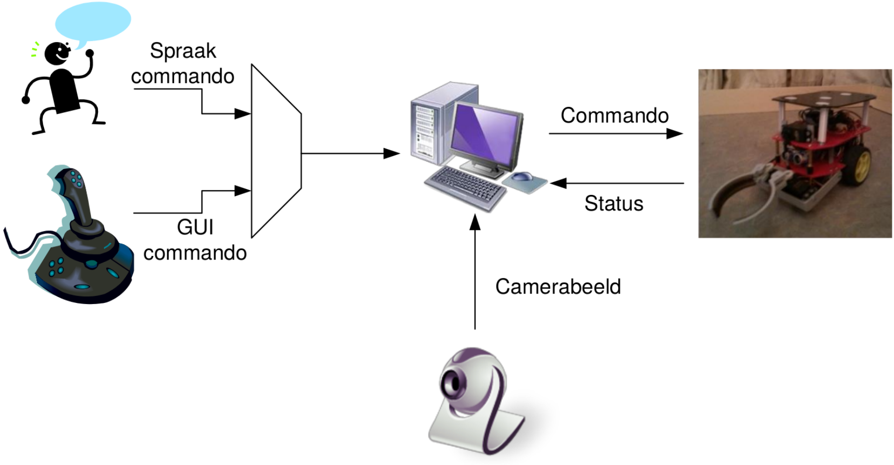
https://www.kpvu.org/post/turn-tune-transcribe-un-develops-radio-listening-tool

2. Universal speech technology



[Saeb et al., 2017; Menon et al., 2018]

2. Universal speech technology



[Renkens, PhD'18]

2. Universal speech technology

Linguistic and cultural documentation and preservation:

http://www.stevenbird.net/

2. Universal speech technology

Academics team up to save dying languages

25/3/2014

A beautifully crafted documentary about Aikuma by Thom Cookes which aired on ABC's program *The World*. This video included a segment about Lauren Gawne and her work on Kagate (Nepal).

Like 0 Tweet 0

http://www.stevenbird.net/

3. Understanding human language acquisition

- Cognitive modelling: Try to uncover learning mechanisms in humans
- A model of human language acquisition: Can probe easily
- Example applications:
  - Identify hearing loss early
  - Predict learning difficulties
  - How much do we need to talk to infants?

https://bergelsonlab.com/seedlings/

Three ideas to tackle these problems

1. Build in the (domain) knowledge we have

1. Build in the (domain) knowledge we have

- Pushing the model in a direction: inductive bias, Bayesian priors, regularisation, data augmentation

1. Build in the (domain) knowledge we have

- Pushing the model in a direction: inductive bias, Bayesian priors, regularisation, data augmentation
- In unsupervised learning this is all we have

1. Build in the (domain) knowledge we have

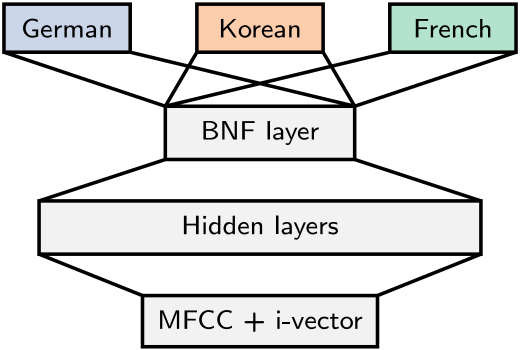
- Pushing the model in a direction: inductive bias, Bayesian priors, regularisation, data augmentation
- In unsupervised learning this is all we have
- We know a lot about languages in general

1. Build in the (domain) knowledge we have

- Pushing the model in a direction: inductive bias, Bayesian priors, regularisation, data augmentation
- In unsupervised learning this is all we have
- We know a lot about languages in general
- Example: Although speech sounds are produced differently in different languages, there are aspects which are shared

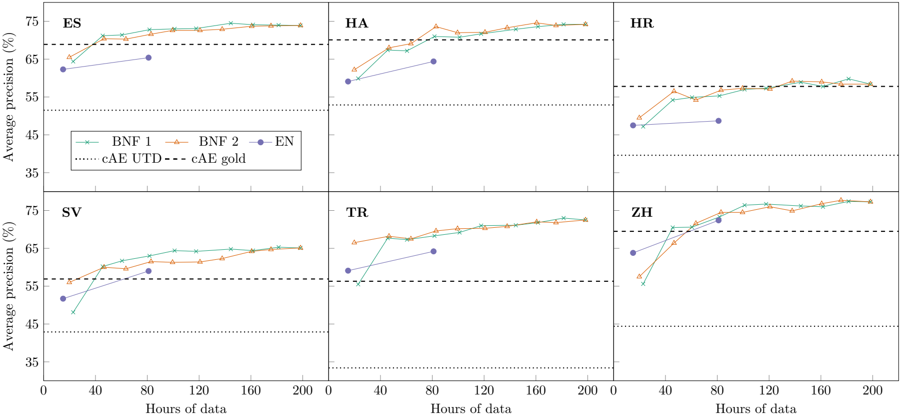
1. Build in the (domain) knowledge we have

Share representations across languages:



[Hermann and Goldwater, 2018; Hermann et al., 2018; `https://arxiv.org/abs/1811.04791`]

1. Build in the (domain) knowledge we have

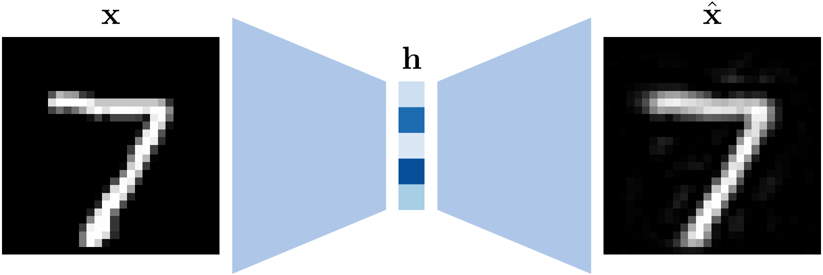


[Hermann and Goldwater, 2018; Hermann et al., 2018; https://arxiv.org/abs/1811.04791]

2. Compression

2. Compression

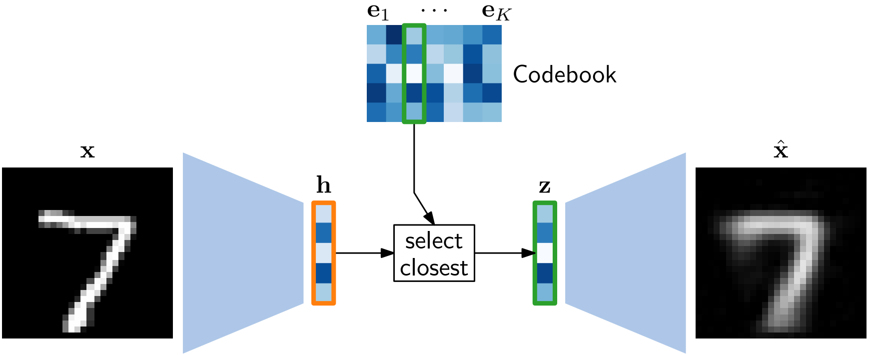
Autoencoder:



Loss for single training example: $J = ||\mathbf{x} - \hat{\mathbf{x}}||^2$

2. Compression

Vector-quantised variational autoencoder (VQ-VAE):



$$\mathbf{z} = \mathbf{e}_k \text{ where } k = \text{argmin}_{j=1}^{K} ||\mathbf{h} - \mathbf{e}_j||^2$$

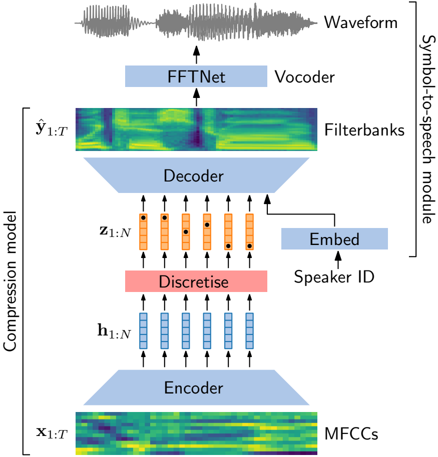
$$J = \alpha||\mathbf{x} - \hat{\mathbf{x}}||^2 + ||\text{sg}(\mathbf{h}) - \mathbf{e}_k||^2 + \beta||\mathbf{h} - \text{sg}(\mathbf{e}_k)||^2$$

2. Compression: An example from our group

Benjamin van Niekerk

André Nortje

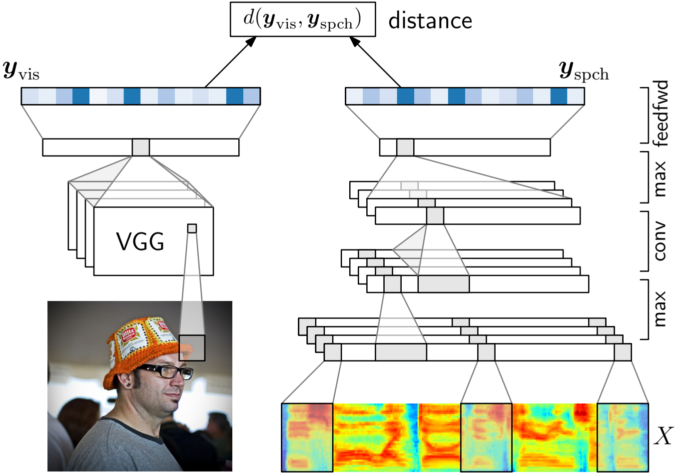
Language	Input	Synthesised output
English	Play	Play
Indonesian	Play	Play



https://arxiv.org/abs/1904.07556

3. Learning from multiple modalities

3. Learning from multiple modalities



[Harwath et al., NeurIPS'16]

3. Learning from multiple modalities

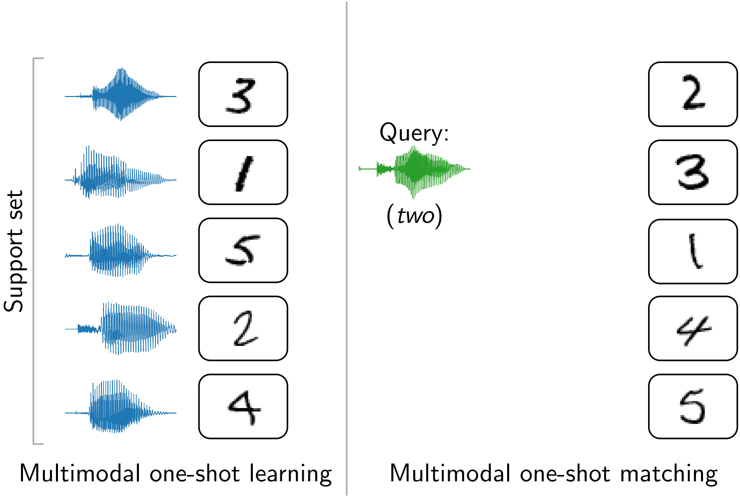
One-shot multimodal learning and matching:

Ryan
Eloff

Leanne
Nortje

3. Learning from multiple modalities

One-shot multimodal learning and matching:



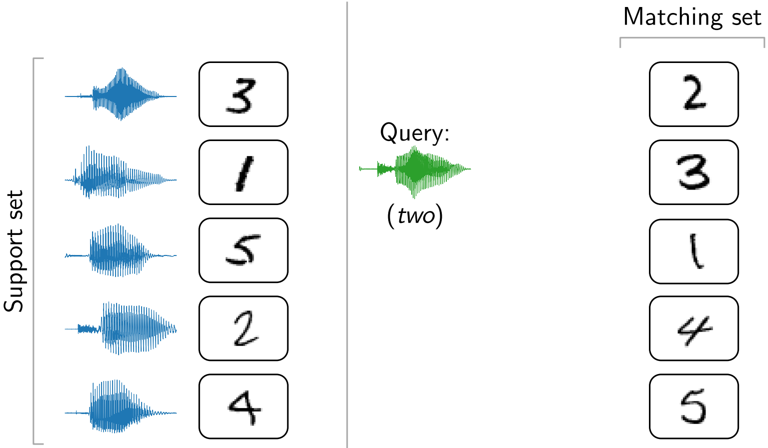
Ryan Eloff

Leanne Nortje

[Eloff et al., ICASSP'19; https://arxiv.org/abs/1811.03875]

3. Learning from multiple modalities

One-shot multimodal learning and matching:



Multimodal one-shot learning | Multimodal one-shot matching

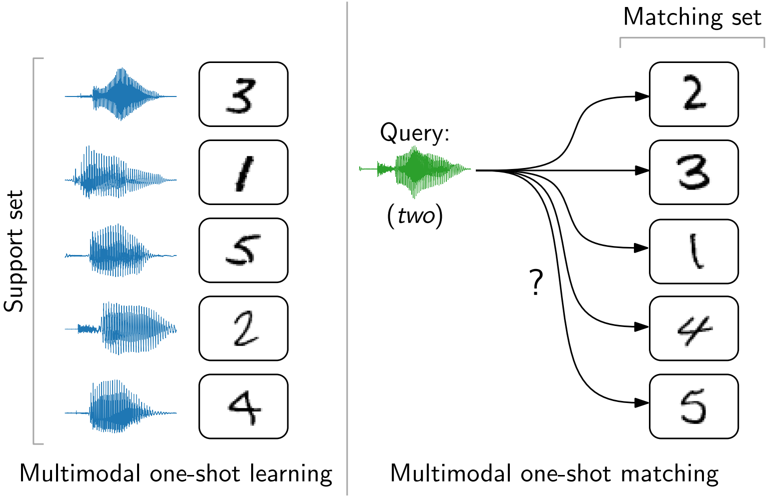
Ryan Eloff

Leanne Nortje

[Eloff et al., ICASSP'19; https://arxiv.org/abs/1811.03875]

3. Learning from multiple modalities

One-shot multimodal learning and matching:



Ryan Eloff

Leanne Nortje

[Eloff et al., ICASSP'19; https://arxiv.org/abs/1811.03875]

The most important missing parts

What I think is still missing

What I think is still missing

- Engineering/technical: Generic ways to incorporate domain knowledge

What I think is still missing

- Engineering/technical: Generic ways to incorporate domain knowledge
- Scientific: What are the mechanisms used for learning language?

What I think is still missing

- Engineering/technical: Generic ways to incorporate domain knowledge
- Scientific: What are the mechanisms used for learning language?
- What are useful, practical applications that we should be working on?

What I think is still missing

- Engineering/technical: Generic ways to incorporate domain knowledge
- Scientific: What are the mechanisms used for learning language?
- What are useful, practical applications that we should be working on? (Instead of just spending time in the shower)

What I think is still missing

- Engineering/technical: Generic ways to incorporate domain knowledge
- Scientific: What are the mechanisms used for learning language?
- What are useful, practical applications that we should be working on? (Instead of just spending time in the shower)
- Real test cases on real low-resource languages

What I think is still missing

- Engineering/technical: Generic ways to incorporate domain knowledge
- Scientific: What are the mechanisms used for learning language?
- What are useful, practical applications that we should be working on? (Instead of just spending time in the shower)
- Real test cases on real low-resource languages
  "...while the authors did make an effort to artificially limit the data availability, I don't think the main claims of the paper ... is generalizable to actual low-resource languages ..." — Reviewer

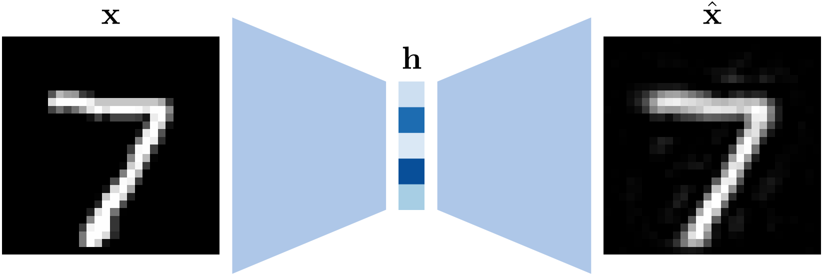
What I think is still missing

- Engineering/technical: Generic ways to incorporate domain knowledge
- Scientific: What are the mechanisms used for learning language?
- What are useful, practical applications that we should be working on? (Instead of just spending time in the shower)
- Real test cases on real low-resource languages
  "...while the authors did make an effort to artificially limit the data availability, I don't think the main claims of the paper ... is generalizable to actual low-resource languages ..." — Reviewer
- Getting data for these test cases

http://www.kamperh.com/

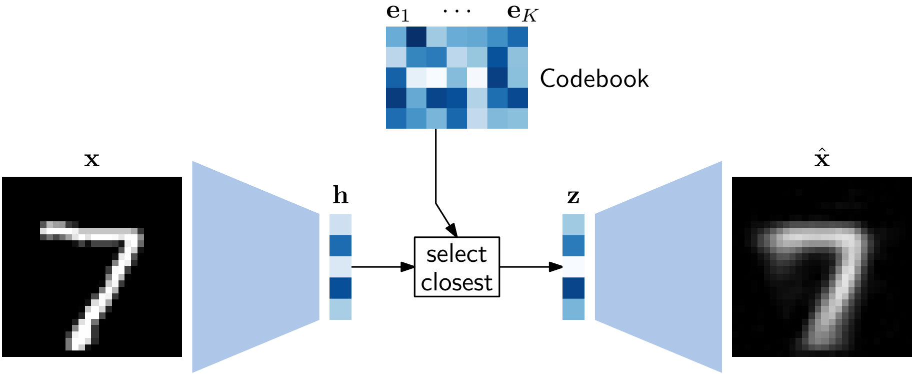
https://github.com/kamperh/

Compression: Autoencoder



Loss for single training example: $J = ||\mathbf{x} - \hat{\mathbf{x}}||^2$

Vector-quantised variational autoencoder (VQ-VAE)



$\mathbf{z} = \mathbf{e}_k$ where $k = \text{argmin}_{j=1}^{K} ||\mathbf{h} - \mathbf{e}_j||^2$

Vector-quantised variational autoencoder (VQ-VAE)

- Loss for single training example:

  $$J = -\log p(\mathbf{x}|\mathbf{z}) + ||\mathrm{sg}(\mathbf{h}) - \mathbf{z}||^2 + \beta||\mathbf{h} - \mathrm{sg}(\mathbf{z})||^2$$

- Assuming spherical Gaussian output:

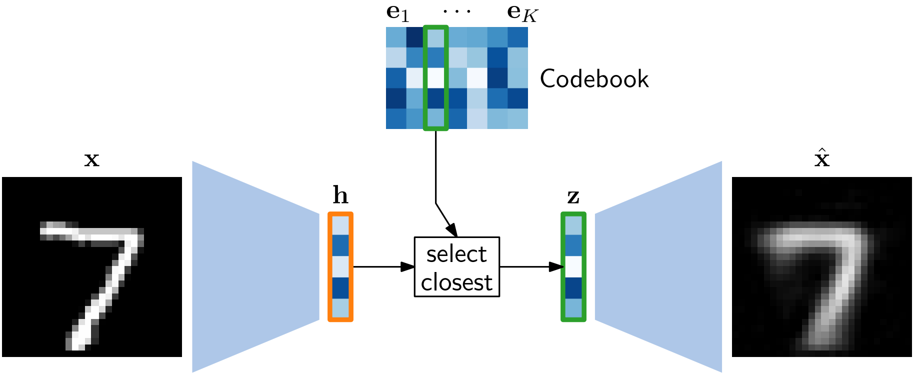
  $$J = \alpha||\mathbf{x} - \hat{\mathbf{x}}||^2 + ||\mathrm{sg}(\mathbf{h}) - \mathbf{z}||^2 + \beta||\mathbf{h} - \mathrm{sg}(\mathbf{z})||^2$$

- Explicitly denoting selected embedding:

  $$J = \alpha||\mathbf{x} - \hat{\mathbf{x}}||^2 + ||\mathrm{sg}(\mathbf{h}) - \mathbf{e}_k||^2 + \beta||\mathbf{h} - \mathrm{sg}(\mathbf{e}_k)||^2$$

- $||\mathbf{x} - \hat{\mathbf{x}}||^2$ is the reconstruction loss
- $||\mathrm{sg}(\mathbf{h}) - \mathbf{e}_k||^2$ updates the embedding codebook, with sg denoting the stop-gradient
- $||\mathbf{h} - \mathrm{sg}(\mathbf{e}_k)||^2$ is the *commitment loss* which encourages the encoder output $\mathbf{h}$ to lie close to the selected codebook embedding $\mathbf{e}_k$

Vector-quantised variational autoencoder (VQ-VAE)



$$\mathbf{z} = \mathbf{e}_k \text{ where } k = \text{argmin}_{j=1}^{K} ||\mathbf{h} - \mathbf{e}_j||^2$$

$$J = \alpha||\mathbf{x} - \hat{\mathbf{x}}||^2 + ||\text{sg}(\mathbf{h}) - \mathbf{e}_k||^2 + \beta||\mathbf{h} - \text{sg}(\mathbf{e}_k)||^2$$

Vector-quantised variational autoencoder (VQ-VAE)

- Quantisation in VQ-VAE:

  $\mathbf{z} = \mathbf{e}_k$ where $k = \text{argmin}_{j=1}^{K} ||\mathbf{h} - \mathbf{e}_j||^2$

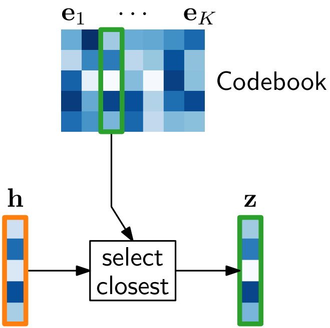
- For backpropagation we need: $\dfrac{\partial J}{\partial \mathbf{h}}$

- Chain rule: $\dfrac{\partial J}{\partial \mathbf{h}} = \dfrac{\partial \mathbf{z}}{\partial \mathbf{h}} \dfrac{\partial J}{\partial \mathbf{z}}$

- What is $\dfrac{\partial \mathbf{z}}{\partial \mathbf{h}}$ with $\mathbf{z} = \text{closest}(\mathbf{e}_1, \ldots, \mathbf{e}_K)$? Cannot solve directly

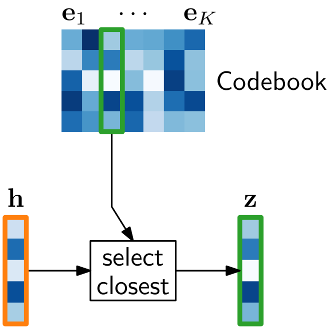
- Idea: If $\mathbf{z} \approx \mathbf{h}$ then we could use $\dfrac{\partial J}{\partial \mathbf{h}} \approx \dfrac{\partial J}{\partial \mathbf{z}}$

- $||\text{sg}(\mathbf{h}) - \mathbf{e}_k||^2 + \beta||\mathbf{h} - \text{sg}(\mathbf{e}_k)||^2$ adds incentive for $\mathbf{z} \approx \mathbf{h}$



Vector-quantised variational autoencoder (VQ-VAE)

- So, why not just use $J = ||\mathbf{x} - \hat{\mathbf{x}}||^2$?
- Then there is no incentive for $\mathbf{z} \approx \mathbf{h}$
- Why not just add $||\mathbf{h} - \mathbf{z}||^2$?
- Might want to update $\mathbf{h}$ and the selected embedding $\mathbf{z} = \mathbf{e}_k$ at different rates



- I.e., might still want $\mathbf{h}$ to sometimes pick different embeddings in the codebook so that these get updated (think about how we add noise in standard STE)
- Answer to both above questions: it works better