Multimodal learning from images and speech

KU Leuven & UPF Barcelona, January 2019

Herman Kamper

E&E Engineering, Stellenbosch University, South Africa

http://www.kamperh.com/

Advances in speech recognition

Advances in speech recognition

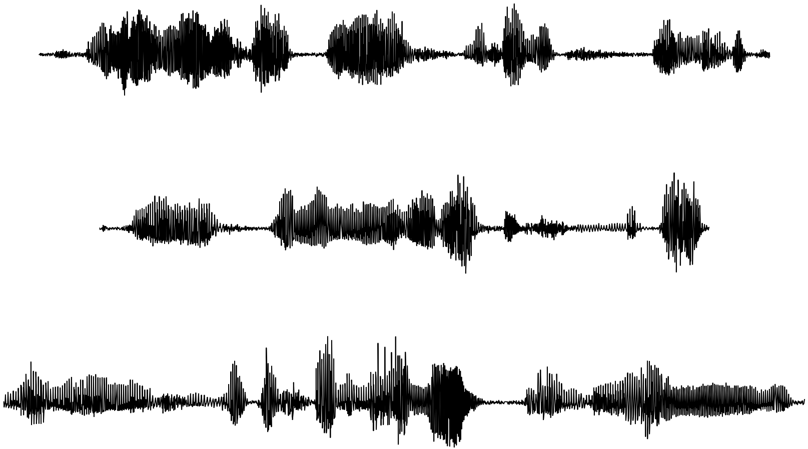
- **Addiction to labels**: 2000 hours transcribed speech audio; $\sim$350M/560M words text [Xiong et al., TASLP'17]

Advances in speech recognition

- **Addiction to labels**: 2000 hours transcribed speech audio; ~350M/560M words text [Xiong et al., TASLP'17]
- Sometimes not possible, e.g., for unwritten languages

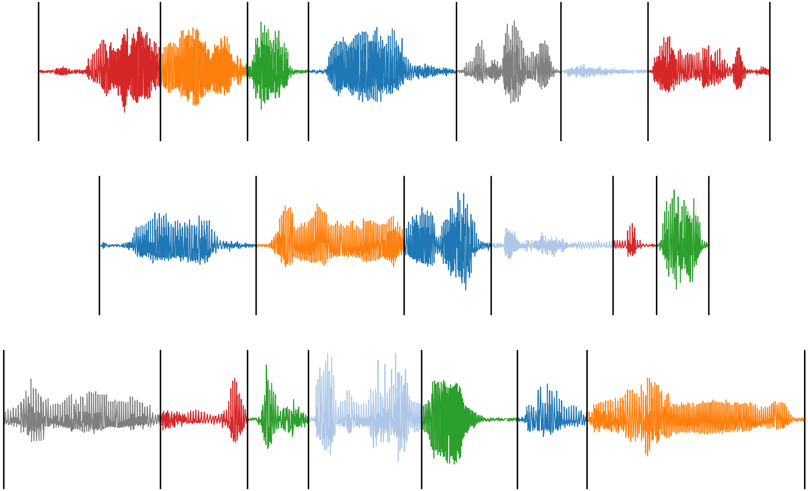
“Zero-resource” speech processing

"Zero-resource" speech processing



[Kamper et al., TASLP'16]

“Zero-resource” speech processing



[Kamper et al., TASLP’16]

Why learn without labels?

Why learn without labels?

- Get insight into human **language acquisition** [Räsänen and Rasilo, '15]

Why learn without labels?

- Get insight into human **language acquisition** [Räsänen and Rasilo, '15]
- Language acquisition in **robots** [Roy, '99]; [Renkens and Van hamme, '15]

Why learn without labels?

- Get insight into human **language acquisition** [Räsänen and Rasilo, '15]
- Language acquisition in **robots** [Roy, '99]; [Renkens and Van hamme, '15]
- Analysis of audio for unwritten languages [Besacier et al., '14]

Why learn without labels?

- Get insight into human **language acquisition** [Räsänen and Rasilo, '15]
- Language acquisition in **robots** [Roy, '99]; [Renkens and Van hamme, '15]
- Analysis of audio for unwritten languages [Besacier et al., '14]
- New **insights** and models for speech processing [Jansen et al., '13]

Why learn without labels?

- Get insight into human **language acquisition** [Räsänen and Rasilo, '15]
- Language acquisition in **robots** [Roy, '99]; [Renkens and Van hamme, '15]
- Analysis of audio for unwritten languages [Besacier et al., '14]
- New **insights** and models for speech processing [Jansen et al., '13]
- but ...

Why learn without labels?

- Get insight into human **language acquisition** [Räsänen and Rasilo, '15]
- Language acquisition in **robots** [Roy, '99]; [Renkens and Van hamme, '15]
- Analysis of audio for unwritten languages [Besacier et al., '14]
- New **insights** and models for speech processing [Jansen et al., '13]
- but ... what about context?

1. Visually Grounded Keyword Spotting

1. Visually Grounded Keyword Spotting

Shane Settle

Michael Roth

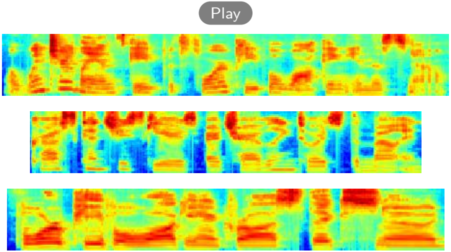
Greg Shakhnarovich

Karen Livescu

Images as weak labels for speech

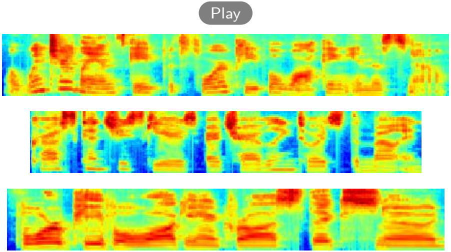
Images as weak labels for speech

Can we use images as weak labels in low-resource settings?



Images as weak labels for speech

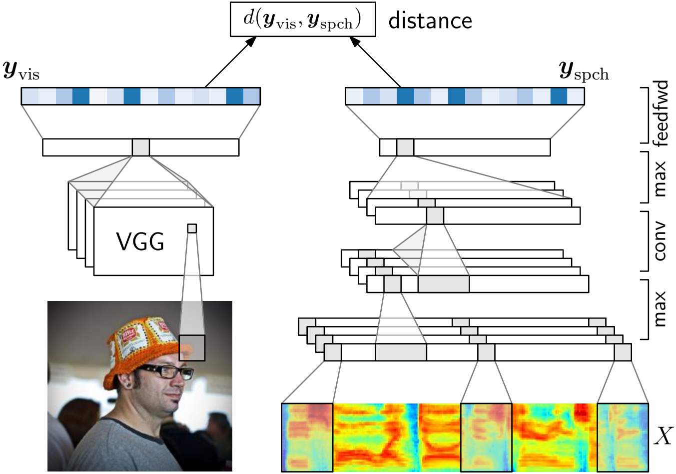
Can we use images as weak labels in low-resource settings?



Maybe we cannot use this type of data for full ASR, but maybe it can be used for other tasks?

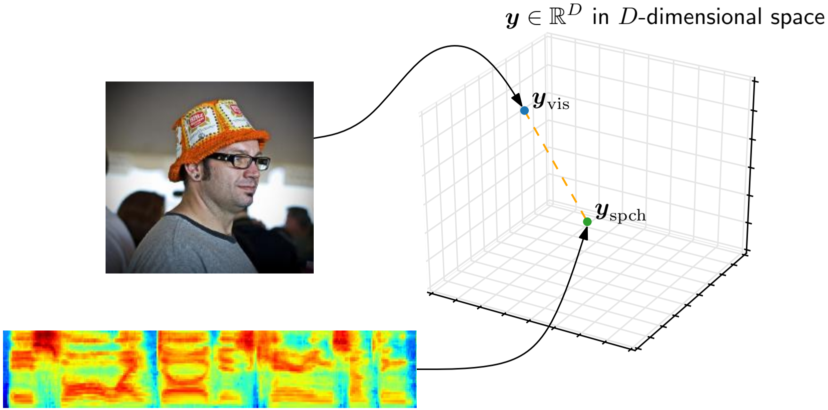
Map images and speech into common space

Map images and speech into common space



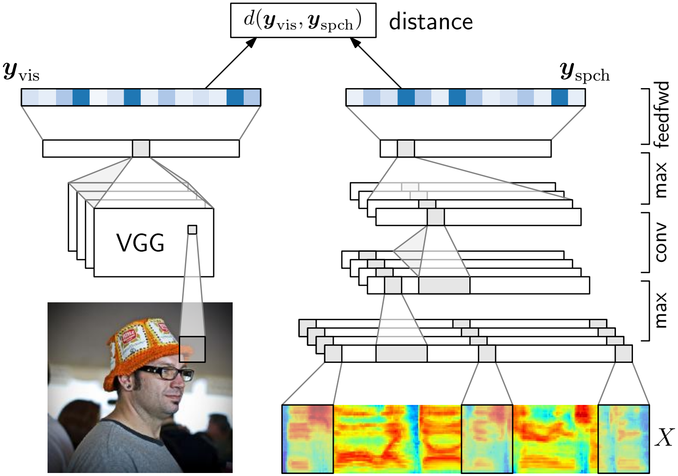
[Harwath et al., NIPS'16]

Retrieval in common (semantic) space



[Harwath et al., NIPS'16]

Can we use (supervised) vision model to get labels?



Cannot obtain textual labels for the speech using this model

Word prediction from images and speech

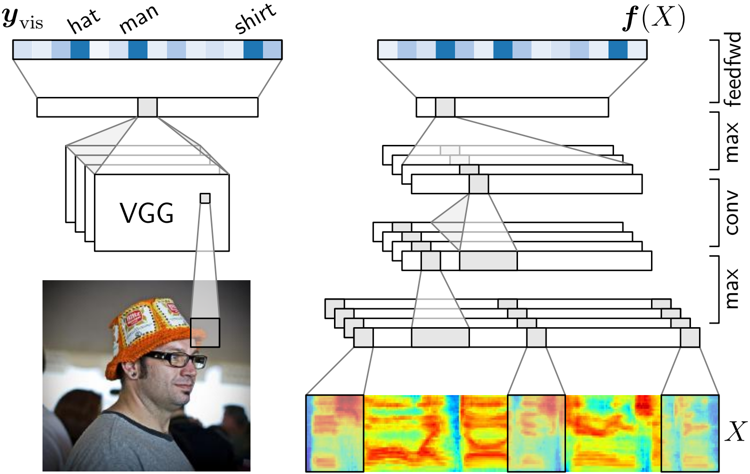
Word prediction from images and speech

[Kamper et al., Interspeech'17]

Word prediction from images and speech

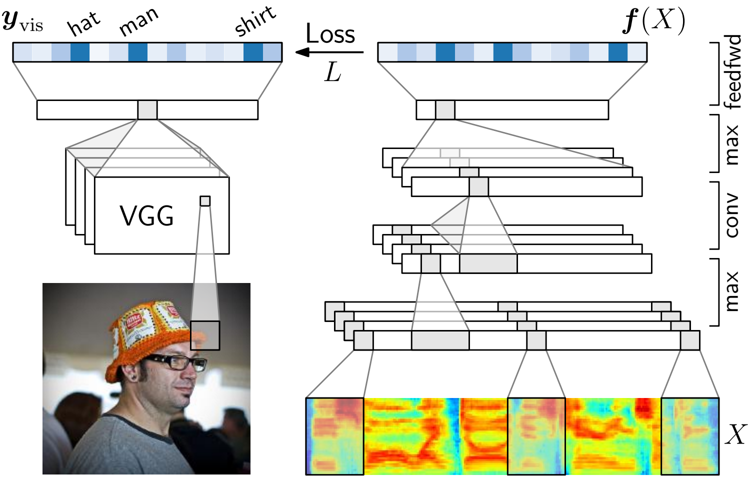
[Kamper et al., Interspeech'17]

Word prediction from images and speech



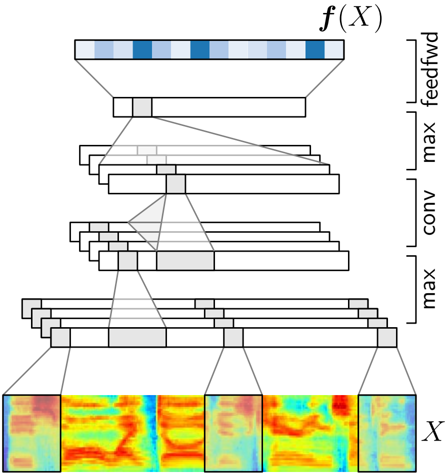
[Kamper et al., Interspeech'17]

Word prediction from images and speech



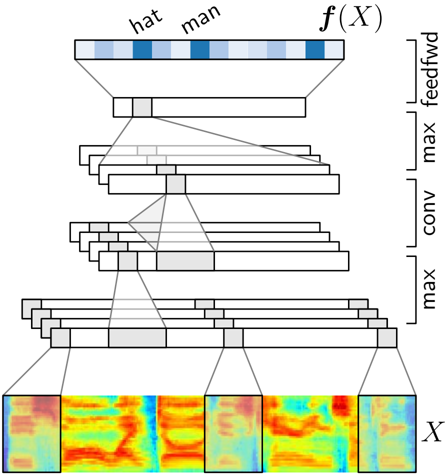
[Kamper et al., Interspeech'17]

Word prediction from images and speech



[Kamper et al., Interspeech'17]

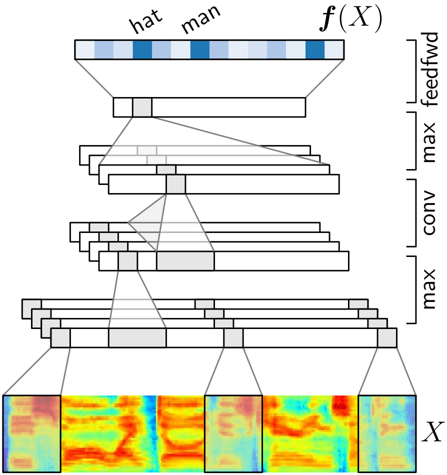
Word prediction from images and speech



[Kamper et al., Interspeech'17]

Word prediction from images and speech

$\boldsymbol{f}(X) \in \mathbb{R}^W$ is vector of word probabilities

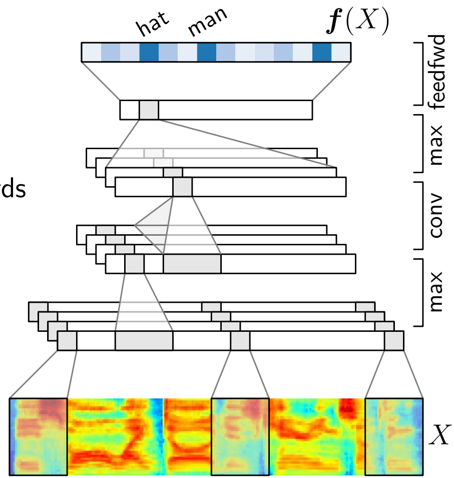


[Kamper et al., Interspeech'17]

Word prediction from images and speech

$\boldsymbol{f}(X) \in \mathbb{R}^W$ is vector of word probabilities

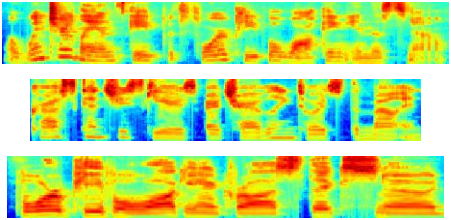
I.e., a spoken bag-of-words (BoW) classifier



[Kamper et al., Interspeech'17]

Images paired with untranscribed speech

We are still in this setting:



- We do not use any of the speech transcriptions during model training (only for evaluation)
- But our resulting model can make bag-of-words (BoW) predictions

Task 1: Spoken bag-of-words prediction

Input utterance	Predicted BoW labels
Play	

Task 1: Spoken bag-of-words prediction

Input utterance	Predicted BoW labels
Play	**bicycle**, bike, **man**, riding, wearing

Task 1: Spoken bag-of-words prediction

Input utterance	**Predicted BoW labels**
man on bicycle is doing tricks in an old building	**bicycle**, bike, **man**, riding, wearing

Task 1: Spoken bag-of-words prediction

Input utterance	**Predicted BoW labels**
man on bicycle is doing tricks in an old building	**bicycle**, bike, **man**, riding, wearing
a little girl is climbing a ladder	child, **girl**, **little**, young
a rock climber standing in a crevasse	climbing, man, **rock**
a dog running in the grass around sheep	**dog**, field, **grass**, **running**
a man in a miami basketball uniform looking to the right	ball, **basketball**, **man**, player, **uniform**, wearing

Task 1: Spoken bag-of-words prediction

Input utterance	**Predicted BoW labels**
man on bicycle is doing tricks in an old building	**bicycle**, bike, **man**, riding, wearing
a little girl is climbing a ladder	child, **girl**, **little**, young
a rock climber standing in a crevasse	climbing, man, **rock**
a dog running in the grass around sheep	**dog**, field, **grass**, **running**
a man in a miami basketball uniform looking to the right	ball, **basketball**, **man**, player, **uniform**, wearing

Task 2: Keyword spotting

Keyword	Example of matched utterance	Type
beach	Play (one of top 10)	
behind		
bike		
boys		
large		
play		
sitting		
yellow		
young		

Task 2: Keyword spotting

Keyword	Example of matched utterance	Type
beach	a boy in a yellow shirt is walking on a beach . . .	
behind		
bike		
boys		
large		
play		
sitting		
yellow		
young		

Task 2: Keyword spotting

Keyword	Example of matched utterance	Type
beach	a boy in a yellow shirt is walking on a beach . . .	correct
behind		
bike		
boys		
large		
play		
sitting		
yellow		
young		

Task 2: Keyword spotting

Keyword	Example of matched utterance	Type
beach	a boy in a yellow shirt is walking on a beach . . .	correct
behind	a surfer does a flip on a wave	
bike		
boys		
large		
play		
sitting		
yellow		
young		

Task 2: Keyword spotting

Keyword	Example of matched utterance	Type
beach	a boy in a yellow shirt is walking on a beach ...	correct
behind	a surfer does a flip on a wave	mistake
bike		
boys		
large		
play		
sitting		
yellow		
young		

Task 2: Keyword spotting

Keyword	Example of matched utterance	Type
beach	a boy in a yellow shirt is walking on a beach ...	correct
behind	a surfer does a flip on a wave	mistake
bike	a dirt biker flies through the air	
boys		
large		
play		
sitting		
yellow		
young		

Task 2: Keyword spotting

Keyword	Example of matched utterance	Type
beach	a boy in a yellow shirt is walking on a beach . . .	correct
behind	a surfer does a flip on a wave	mistake
bike	a dirt biker flies through the air	variant
boys		
large		
play		
sitting		
yellow		
young		

Task 2: Keyword spotting

Keyword	Example of matched utterance	Type
beach	a boy in a yellow shirt is walking on a beach . . .	correct
behind	a surfer does a flip on a wave	mistake
bike	a dirt biker flies through the air	variant
boys	Play	
large		
play		
sitting		
yellow		
young		

Task 2: Keyword spotting

Keyword	Example of matched utterance	Type
beach	a boy in a yellow shirt is walking on a beach ...	correct
behind	a surfer does a flip on a wave	mistake
bike	a dirt biker flies through the air	variant
boys	two children play soccer in the park	
large		
play		
sitting		
yellow		
young		

Task 2: Keyword spotting

Keyword	Example of matched utterance	Type
beach	a boy in a yellow shirt is walking on a beach . . .	correct
behind	a surfer does a flip on a wave	mistake
bike	a dirt biker flies through the air	variant
boys	two children play soccer in the park	semantic
large		
play		
sitting		
yellow		
young		

Task 2: Keyword spotting

Keyword	Example of matched utterance	Type
beach	a boy in a yellow shirt is walking on a beach ...	correct
behind	a surfer does a flip on a wave	mistake
bike	a dirt biker flies through the air	variant
boys	two children play soccer in the park	semantic
large	Play	
play		
sitting		
yellow		
young		

Task 2: Keyword spotting

Keyword	Example of matched utterance	Type
beach	a boy in a yellow shirt is walking on a beach ...	correct
behind	a surfer does a flip on a wave	mistake
bike	a dirt biker flies through the air	variant
boys	two children play soccer in the park	semantic
large	... a rocky cliff overlooking a body of water	
play		
sitting		
yellow		
young		

Task 2: Keyword spotting

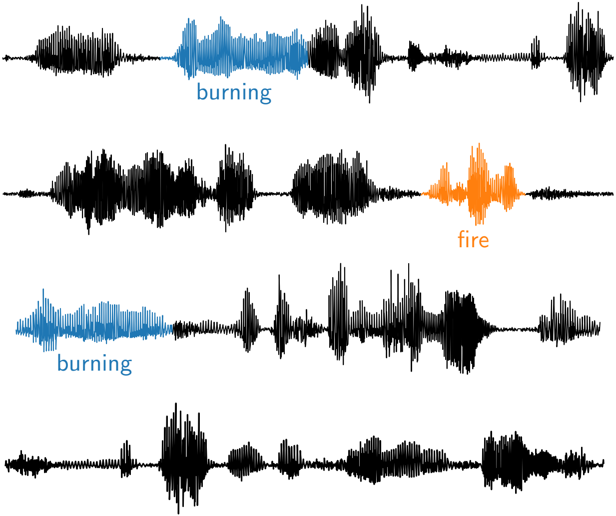
Keyword	Example of matched utterance	Type
beach	a boy in a yellow shirt is walking on a beach ...	correct
behind	a surfer does a flip on a wave	mistake
bike	a dirt biker flies through the air	variant
boys	two children play soccer in the park	semantic
large	... a rocky cliff overlooking a body of water	semantic
play		
sitting		
yellow		
young		

Task 2: Keyword spotting

Keyword	Example of matched utterance	Type
beach	a boy in a yellow shirt is walking on a beach ...	correct
behind	a surfer does a flip on a wave	mistake
bike	a dirt biker flies through the air	variant
boys	two children play soccer in the park	semantic
large	... a rocky cliff overlooking a body of water	semantic
play	children playing in a ball pit	variant
sitting	two people are seated at a table with drinks	semantic
yellow	a tan dog jumping over a red and blue toy	mistake
young	a little girl on a kid swing	semantic

Task 3: Semantic speech retrieval

Written query:
burning



[Kamper et al., TASLP'19]

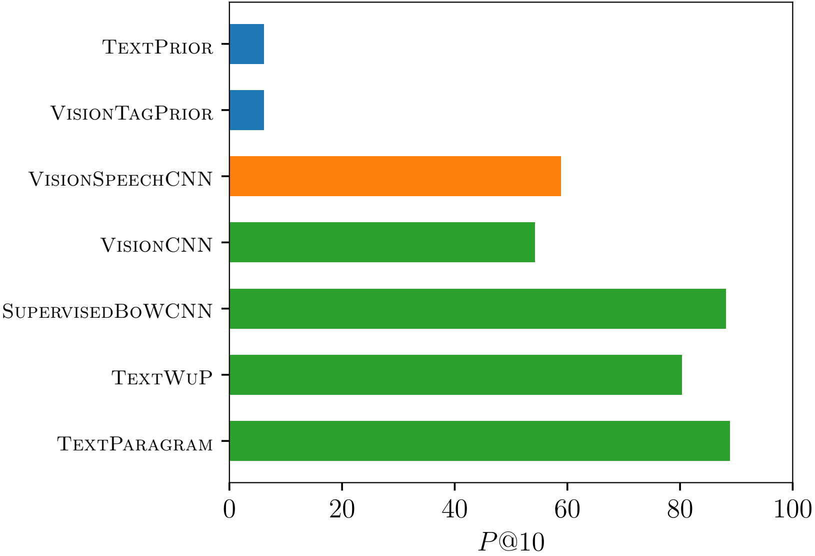
Human (MTurk) evaluation

Human (MTurk) evaluation

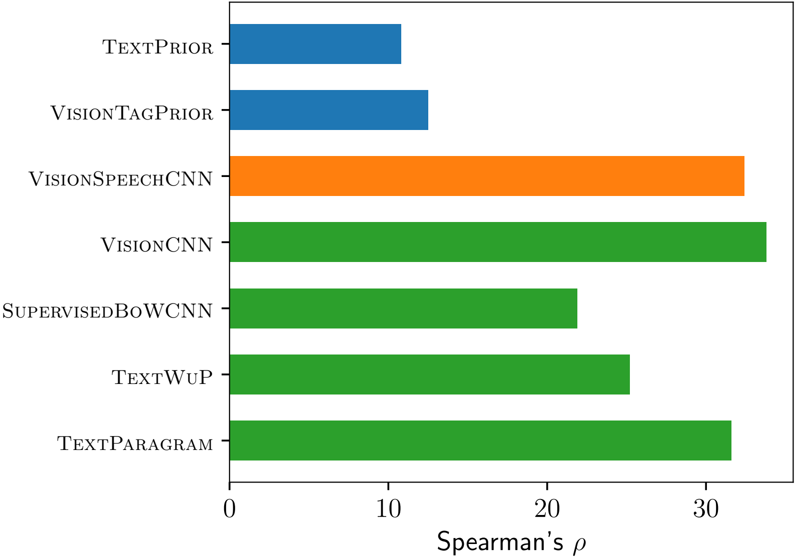
Keyword	Top retrieved utterance	Human label
ocean	man falling off a blue surfboard in the ocean	5 / 5
snowy	a skier catches air over the snow	5 / 5
bike	a dirt biker rides through some trees	4 / 5
children	a group of young boys playing soccer	4 / 5
field	two white dogs running in the grass together	3 / 5
swimming	a woman holding a young boy slide down a water slide into a pool	3 / 5
carrying	small dog running in the grass with a toy in its mouth	2 / 5 $*$
large	a group of people on a zig path through the mountains	1 / 5 $*$
hair	two women and a man smile for the camera	0 / 5 $*$

Task 3: Semantic speech retrieval

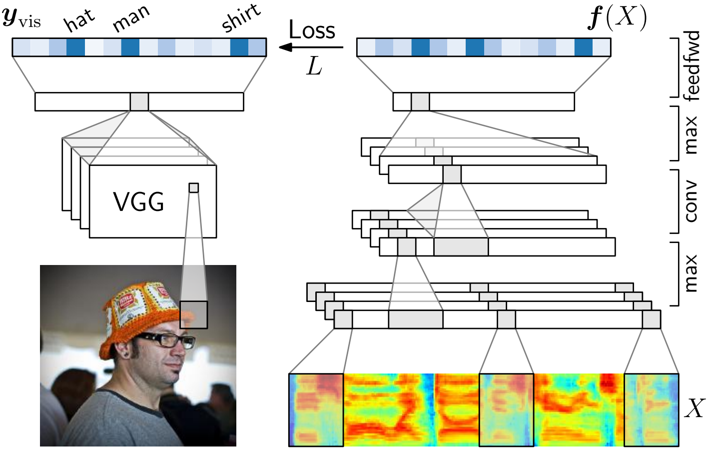
Task 3: Semantic speech retrieval



Task 3: Semantic speech retrieval



But this model is trained for English?

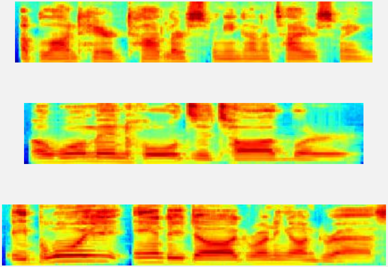


[Kamper et al., Interspeech'17]

Task 4: Cross-lingual keyword spotting

Given English keyword:
'Disease'

Arapaho speech collection
(want to search)

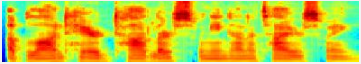


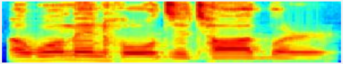
[Kamper and Roth, SLTU'18]

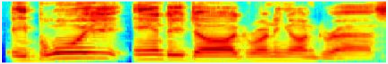
Task 4: Cross-lingual keyword spotting

Given German keyword:
'Hunde'

English speech collection
(want to search)

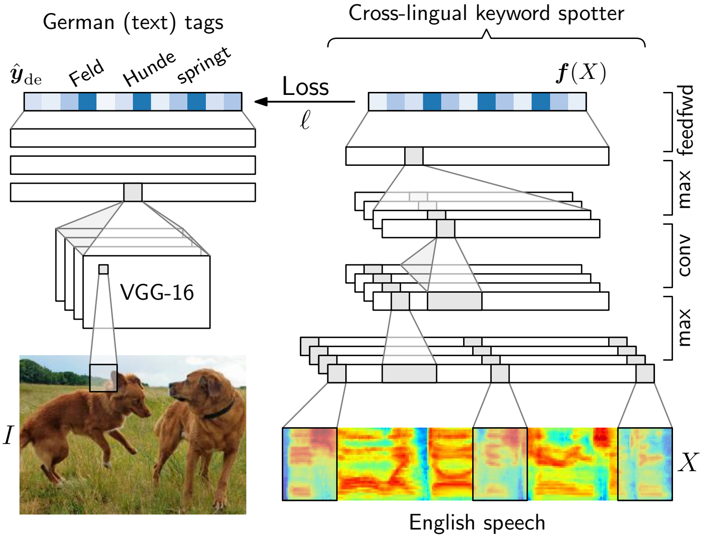






[Kamper and Roth, SLTU'18]

Task 4: Cross-lingual keyword spotting



[Kamper and Roth, SLTU'18]

2. Multimodal One-Shot Learning from Images and Speech

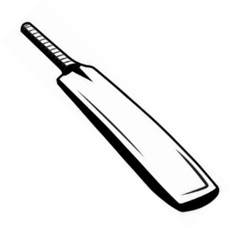
2. Multimodal One-Shot Learning from Images and Speech

Ryan Eloff

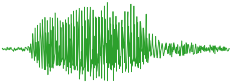
Herman Engelbrecht

You are the robot

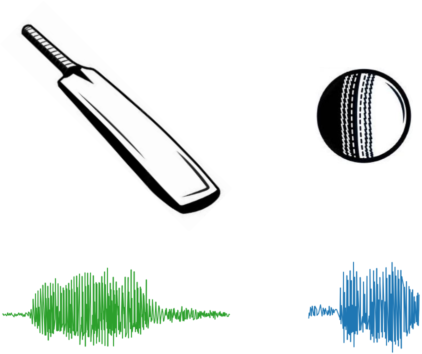
You are the robot



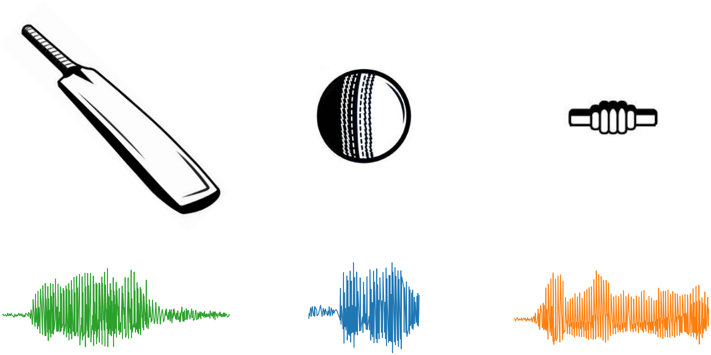
You are the robot



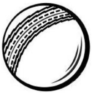
You are the robot



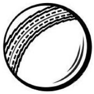
You are the robot

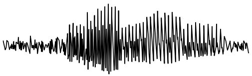


You are the robot



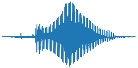
You are the robot

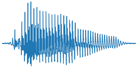


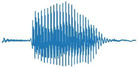


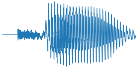
You are the robot

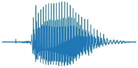
Unimodal one-shot learning and classification


– three


– one

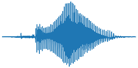

– five

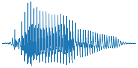

– two

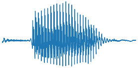

– four

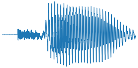
[Fei-Fei et al., PAMI'06]; [Lake et al., CogSci'14]

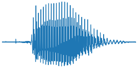
Unimodal one-shot learning and classification


– three

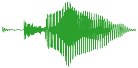

– one


– five


– two


– four

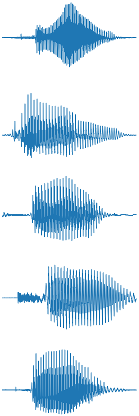
Query:


$\hat{y} = ?$

(*two*)

[Fei-Fei et al., PAMI'06]; [Lake et al., CogSci'14]

Unimodal one-shot learning and classification



– three

– one

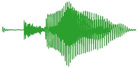
– five

– two

– four

One-shot speech learning

Query:



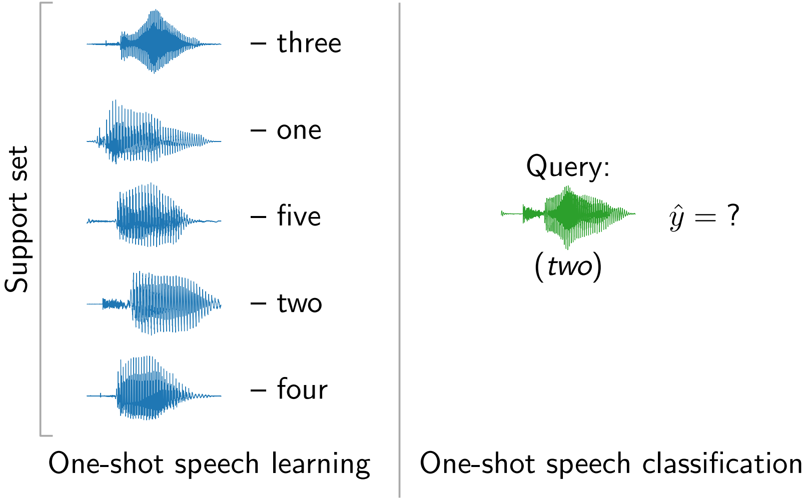
$\hat{y} = ?$

(*two*)

One-shot speech classification

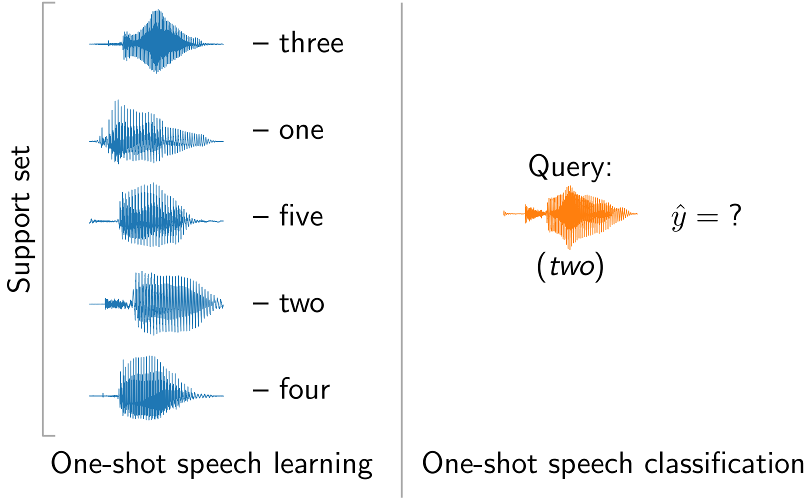
[Fei-Fei et al., PAMI'06]; [Lake et al., CogSci'14]

Unimodal one-shot learning and classification



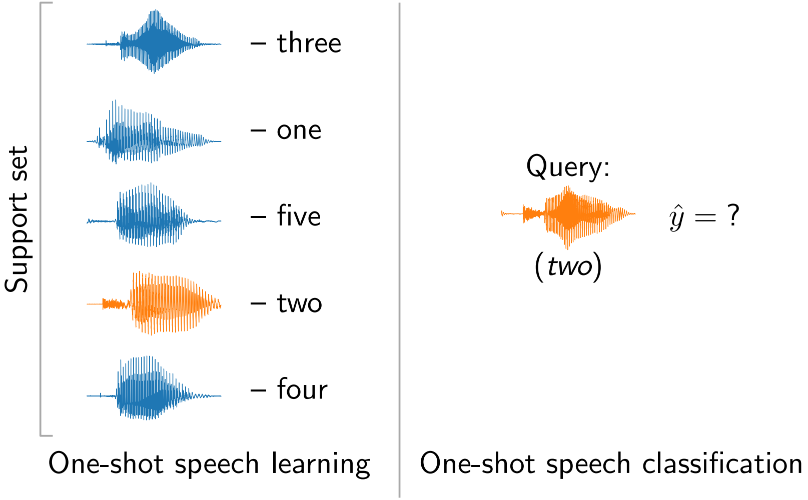
[Fei-Fei et al., PAMI'06]; [Lake et al., CogSci'14]

Unimodal one-shot learning and classification



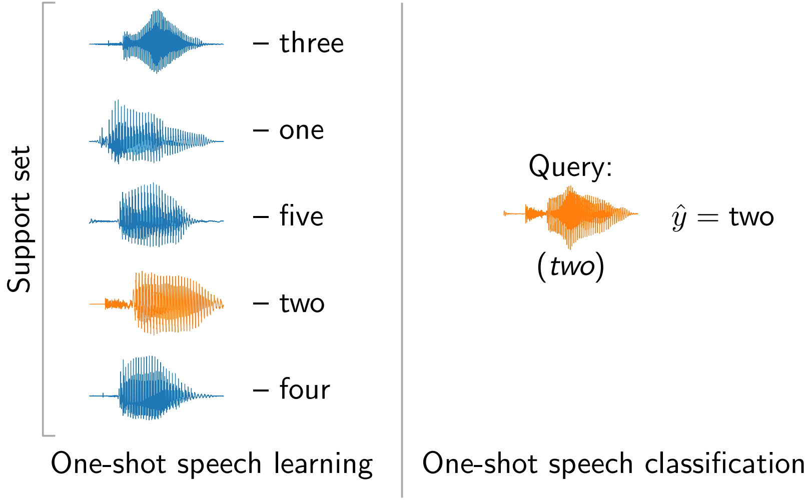
[Fei-Fei et al., PAMI'06]; [Lake et al., CogSci'14]

Unimodal one-shot learning and classification



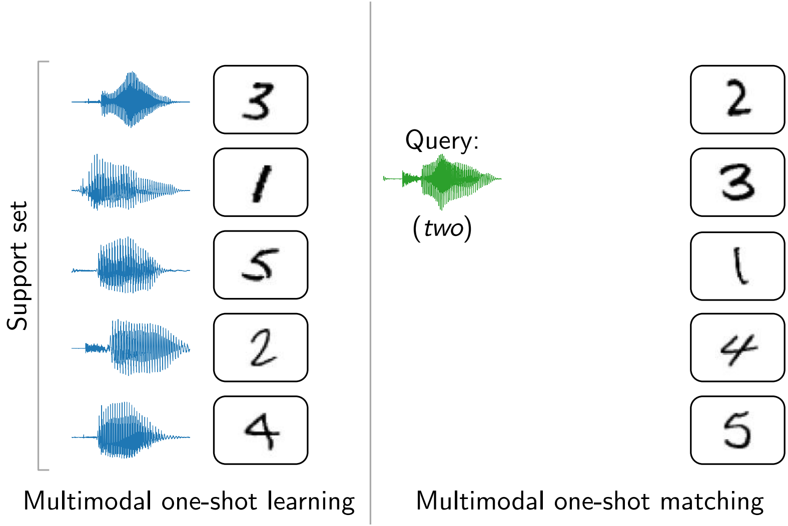
[Fei-Fei et al., PAMI'06]; [Lake et al., CogSci'14]

Unimodal one-shot learning and classification



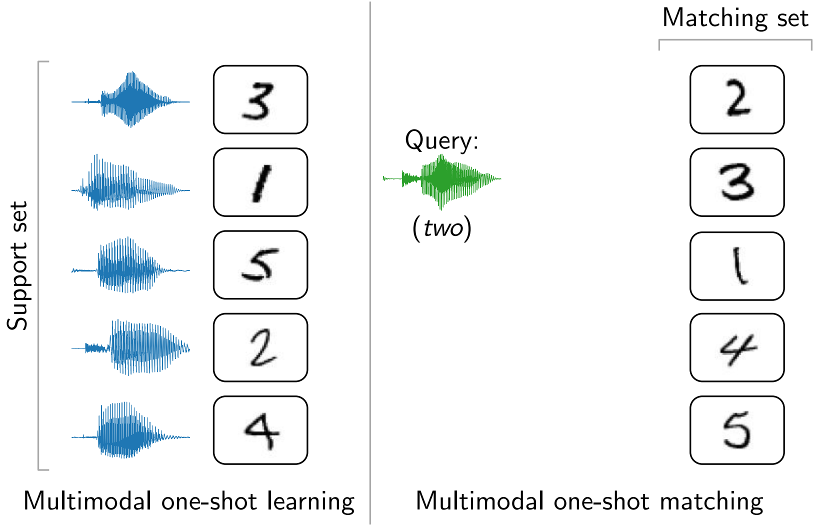
[Fei-Fei et al., PAMI'06]; [Lake et al., CogSci'14]

Multimodal one-shot learning and matching



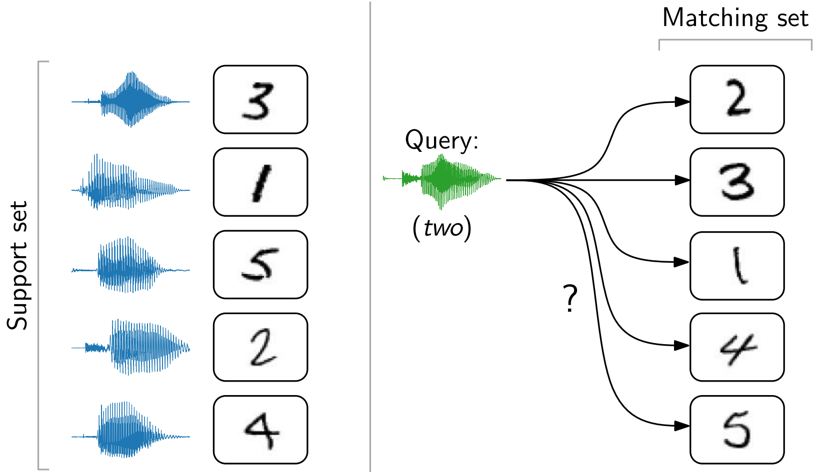
[Eloff et al., arXiv'18]

Multimodal one-shot learning and matching



[Eloff et al., arXiv'18]

Multimodal one-shot learning and matching

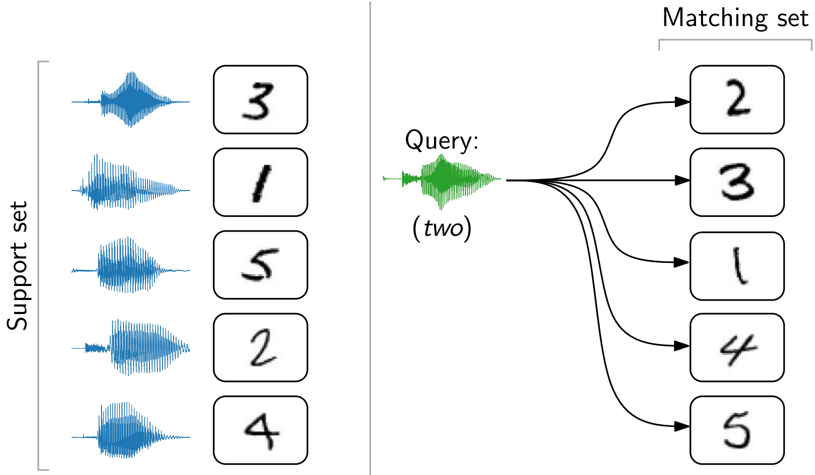


Multimodal one-shot learning

Multimodal one-shot matching

[Eloff et al., arXiv'18]

Our framework

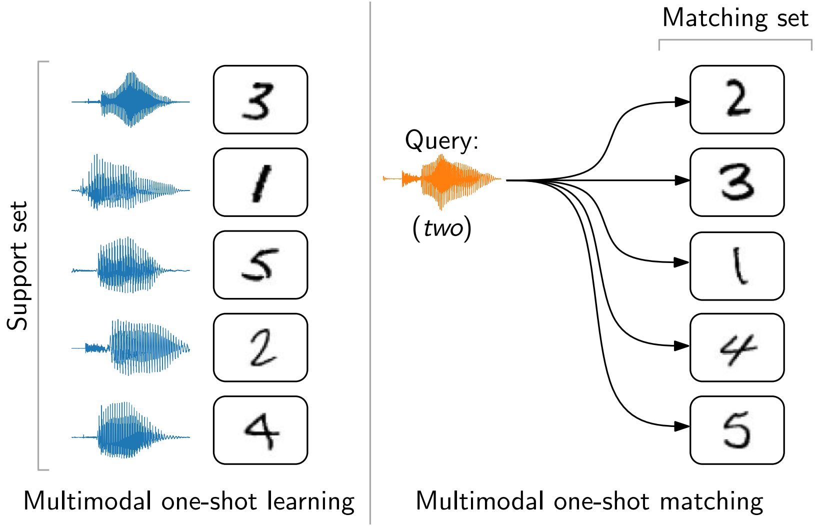


Multimodal one-shot learning

Multimodal one-shot matching

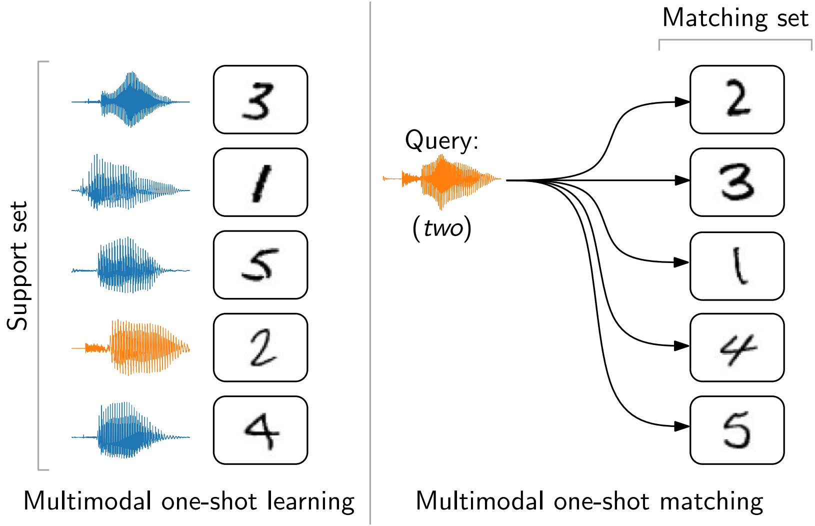
[Eloff et al., arXiv'18]

Our framework



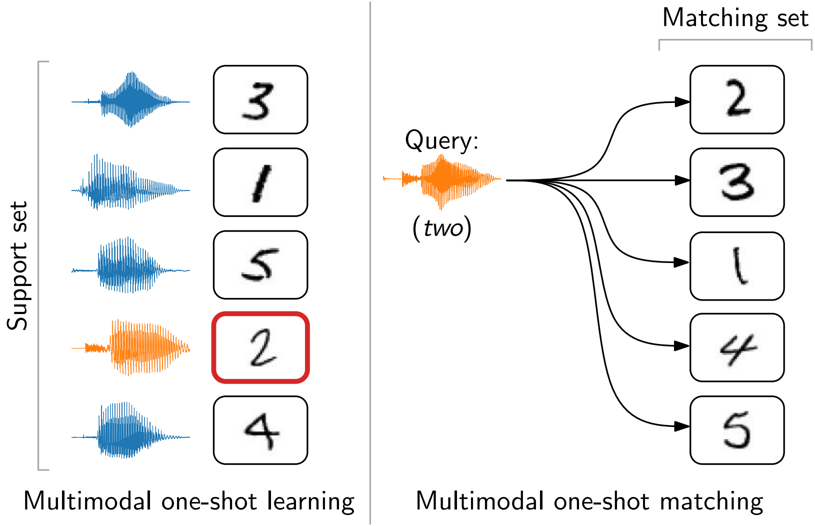
[Eloff et al., arXiv'18]

Our framework



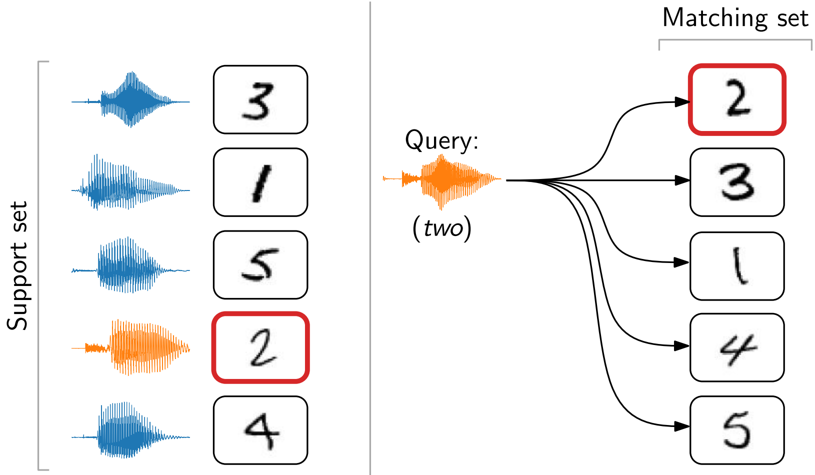
[Eloff et al., arXiv'18]

Our framework



[Eloff et al., arXiv'18]

Our framework

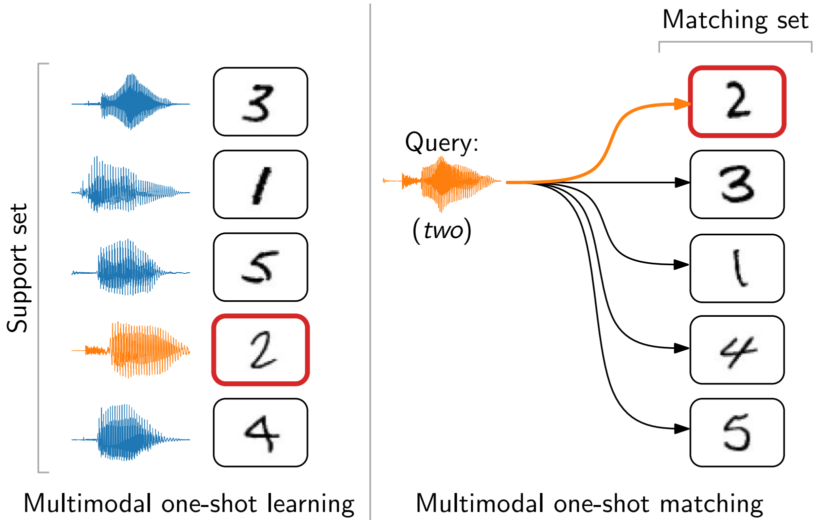


Multimodal one-shot learning

Multimodal one-shot matching

[Eloff et al., arXiv'18]

Our framework



[Eloff et al., arXiv'18]

Our approach to multimodal one-shot learning

Our approach to multimodal one-shot learning

- Requires within-modality distance metrics
- Can be done directly over features: DTW over speech, cosine over image pixels
- Or distance metrics can be learned from background data
- Compare these on TIDigits (speech) paired with MNIST (images)

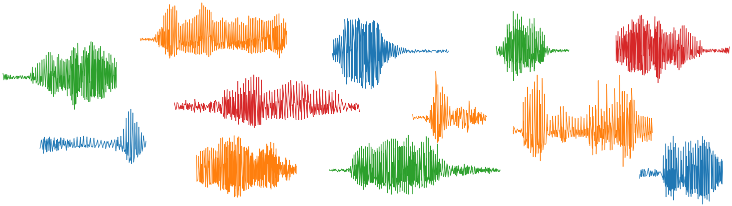
Background data

Omniglot (no digits):

Background data

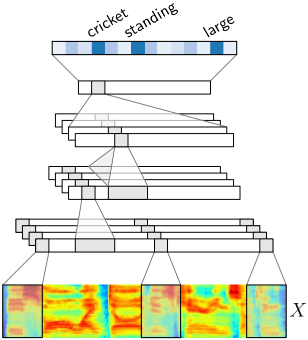
Omniglot (no digits):

Isolated labelled words (no digits):



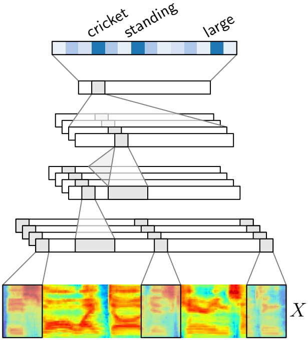
Models for metric learning

Classifier network:

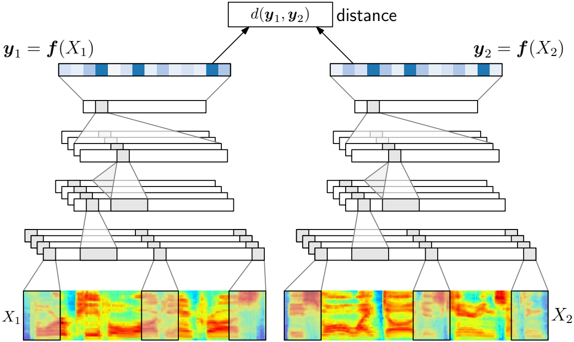


Models for metric learning

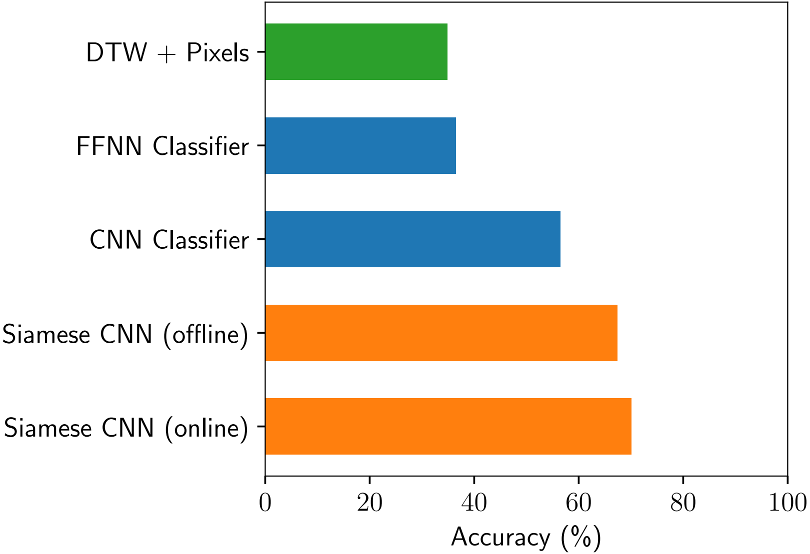
Classifier network:



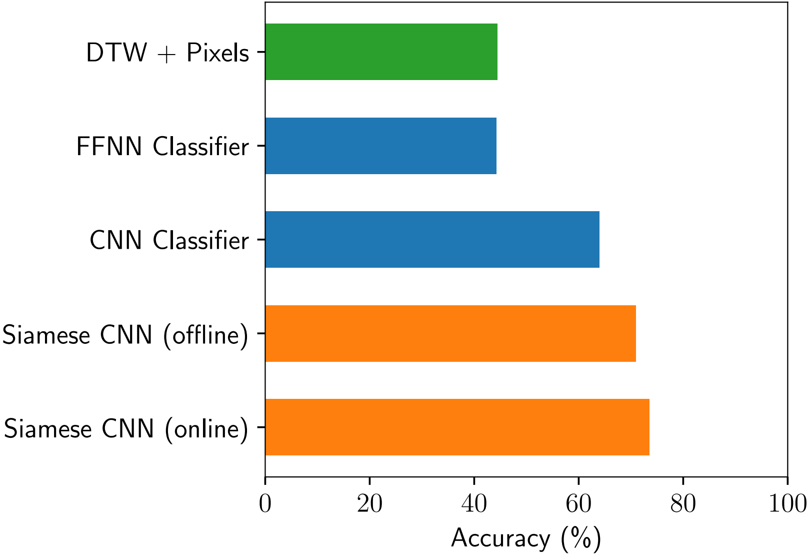
Siamese network:



Multimodal one-shot matching



Multimodal five-shot matching



Takeaways and future work

What to take away from this talk:

Takeaways and future work

What to take away from this talk:

- Visual grounding is useful for dealing with unlabelled speech
- Some things are better when using visual grounding, e.g., one-shot learning, semantic search (?)
- Some things are impossible without it, e.g., keyword prediction from unlabelled speech

Takeaways and future work

What to take away from this talk:

- Visual grounding is useful for dealing with unlabelled speech
- Some things are better when using visual grounding, e.g., one-shot learning, semantic search (?)
- Some things are impossible without it, e.g., keyword prediction from unlabelled speech

Future work:

- Visual grounding of speech paired with videos
- Language universal/agnostic vision systems
- Meta-learning and unsupervised background modelling for one-shot learning
- Developing practical tools for low-resource languages

http://www.kamperh.com/

https://github.com/kamperh/recipe_semantic_flickraudio

https://github.com/rpeloff/multimodal_one_shot_learning

Unimodal one-shot speech classification