Visually grounded cross-lingual keyword spotting in speech

SLTU, August 2018

Herman Kamper¹ and Michael Roth²

¹E&E Engineering, Stellenbosch University, South Africa
²Saarland University, Germany

http://www.kamperh.com/
Advances in speech recognition

• Addiction to labels: 2000 hours transcribed speech audio; \( \sim \) 350M/560M words text [Xiong et al., TASLP'17]
• Very different from the "supervision" infants use to learn language
• Sometimes not possible, e.g., for unwritten languages
Advances in speech recognition

- **Addiction to labels**: 2000 hours transcribed speech audio; \( \sim 350\text{M} / 560\text{M} \) 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]

- Very different from the “supervision” infants use to learn language
Advances in speech recognition

- **Addiction to labels**: 2000 hours transcribed speech audio; \( \sim 350M/560M \) words text [Xiong et al., TASLP’17]
- Very different from the “supervision” infants use to learn language
- Sometimes not possible, e.g., for unwritten languages
Images as weak labels for speech

• Maybe we cannot use this type of data for full ASR, but maybe it can be used for other tasks?
• Goal: Use this type of data for cross-lingual keyword spotting
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?

Goal: Use this type of data for cross-lingual keyword spotting
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?
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?

- **Goal**: Use this type of data for cross-lingual keyword spotting
Cross-lingual keyword spotting

Written query:

burning
(English)

Swahili speech corpus
Cross-lingual word prediction from images
Cross-lingual word prediction from images

\[ y_{\text{vis}} \quad \text{hat} \quad \text{man} \quad \text{shirt} \]

\[ \text{VGG} \]

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

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

Swahili speech

\[ 4 / 12 \]
Cross-lingual word prediction from images

$\mathbf{y}_{\text{vis}} \xrightarrow{} \text{hat} \xrightarrow{} \text{man} \xrightarrow{} \text{shirt}$

$0.85 \xrightarrow{} 0.8 \xrightarrow{} 0.9$

VGG

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

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

Swahili speech

4 / 12
Cross-lingual word prediction from images

\( y_{vis} \)  hat  man  shirt

\( f(X) \)

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

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

Swahili speech
Cross-lingual word prediction from images

\[ y_{vis} \quad \text{hat} \quad \text{man} \quad \text{shirt} \]

\[ f(X) \]

Loss \[ \ell \]

\[ \text{VGG} \]

\[ X \]

\[ f(X) \in \mathbb{R}^W \text{ is vector of word probabilities} \]

I.e., a cross-lingual spoken bag-of-words (BoW) classifier.
Cross-lingual word prediction from images

\[ f(X) \]

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

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

Swahili speech

\[ X \]
Cross-lingual word prediction from images

Swahili speech

$f(X)$

Loss

$\ell$

$X$

$\text{max}$

$\text{conv}$

$\text{feedfwd}$

$I.e., a cross-lingual spoken bag-of-words (BoW) classifier$
Cross-lingual word prediction from images

\[ f(X) \in \mathbb{R}^W \text{ is vector of word probabilities} \]

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

Swahili speech
Cross-lingual word prediction from images

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

Swahili speech
Cross-lingual word prediction from images

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

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

Swahili speech
Experimental details

- **Goal:** Use visual grounding for cross-lingual keyword spotting
Experimental details

- **Goal:** Use visual grounding for cross-lingual keyword spotting
- **Proof-of-concept:** Use English speech with German queries

- **Data:** 8,000 images with 5 English spoken captions (∼37 hours)
- **Weak labels:** German visual tagger trained on German Multi30k
Experimental details

- **Goal:** Use visual grounding for cross-lingual keyword spotting
- **Proof-of-concept:** Use English speech with German queries:

- **Data:** 8000 images with 5 English spoken captions (~37 hours)
- **Weak labels:** German visual tagger trained on German Multi30k
Experimental details

- **Goal:** Use visual grounding for cross-lingual keyword spotting
- **Proof-of-concept:** Use English speech with German queries:

  - **Data:** 8000 images with 5 English spoken captions (≈37 hours)
  - **Weak labels:** German visual tagger trained on German Multi30k
Predictions on test data

Given German keyword: ‘Hunde’

\[ f_{w}(X_i) = P_{\theta}(w|X_i) \]: score for whether (English) speech \( X_i \) contains translation of given (German) keyword \( w \)

Evaluation: Does predicted keyword occur in reference translation?
Example predictions (top retrievals)

**Task:** Given written German keyword, find utterances in an unseen English speech collection containing that keyword

- Fahrrad
  - a biker does a trick on a ramp
  - a person is doing tricks on a bicycle in a city

- Straße
  - people on the city street walk past a puppet theater
  - an asian woman rides a bicycle in front of two cars
Example predictions (top retrievals)

**Task:** Given written German keyword, find utterances in an unseen English speech collection containing that keyword

Input: *Fahrrad*
Example predictions (top retrievals)

Task: Given written German keyword, find utterances in an unseen English speech collection containing that keyword

Input: Fahrrad

Output (in top 10):

- Play
  a biker does a trick on a ramp
  a person is doing tricks on a bicycle in a city

Input: Straße (street)

Output (in top 10):

- people on the city street walk past a puppet theater
  an asian woman rides a bicycle in front of two cars
Example predictions (top retrievals)

**Task:** Given written German keyword, find utterances in an unseen English speech collection containing that keyword

Input: *Fahrrad*

Output (in top 10):

- man riding a bicycle on a foggy day
Example predictions (top retrievals)

**Task:** Given written German keyword, find utterances in an unseen English speech collection containing that keyword

**Input:** Fahrrad

**Output (in top 10):**
- man riding a bicycle on a foggy day
- a biker does a trick on a ramp
- a person is doing tricks on a bicycle in a city
Example predictions (top retrievals)

Task: Given written German keyword, find utterances in an unseen English speech collection containing that keyword

Input: Fahrrad

Output (in top 10):
- man riding a bicycle on a foggy day
- a biker does a trick on a ramp
- a person is doing tricks on a bicycle in a city

Input: Straße (street)
Example predictions (top retrievals)

**Task:** Given written German keyword, find utterances in an unseen English speech collection containing that keyword

**Input:** *Fahrrad*

Output (in top 10):
- man riding a bicycle on a foggy day
- a biker does a trick on a ramp
- a person is doing tricks on a bicycle in a city

**Input:** *Straße* (street)

Output (in top 10):
- Play
Example predictions (top retrievals)

Task: Given written German keyword, find utterances in an unseen English speech collection containing that keyword

Input: Fahrrad
Output (in top 10):
- man riding a bicycle on a foggy day
- a biker does a trick on a ramp
- a person is doing tricks on a bicycle in a city

Input: Straße (street)
Output (in top 10):
- a woman in black and red listens to an ipod walks down the street
Example predictions (top retrievals)

Task: Given written German keyword, find utterances in an unseen English speech collection containing that keyword

Input: *Fahrrad*

Output (in top 10):
- man riding a bicycle on a foggy day
- a biker does a trick on a ramp
- a person is doing tricks on a bicycle in a city

Input: *Straße* (street)

Output (in top 10):
- a woman in black and red listens to an ipod walks down the street
- people on the city street walk past a puppet theater
- an asian woman rides a bicycle in front of two cars
Cross-lingual keyword spotting performance

![Bar chart comparing cross-lingual keyword spotting performance for different models.]

- **DETextPrior**: Low performance
- **DEVisionCNN**: Moderate performance
- **XVisionSpeechCNN**: High performance
- **XBoWCNN**: Very high performance

The chart shows a comparison of P@10 performance across different models, with XBoWCNN performing the best and DETextPrior performing the worst.
Example predictions marked as errors

Input: *Feld* (field)
Example predictions marked as errors

Input: *Feld* (field)

Output:

- a brown and black dog running through a grassy field  
  *
Example predictions marked as errors

Input:  *Feld* (field)

Output:

- a brown and black dog running through a grassy field

Input:  *grün(en)* (green)
Example predictions marked as errors

Input: *Feld* (field)

Output:
- a brown and black dog running through a grassy field

Input: *grün(en)* (green)

Output:
- a brown dog is chasing a red frisbee across a grassy field
Example predictions marked as errors

Input: *Feld* (field)

Output:
  • a brown and black dog running through a grassy field

Input: *grün(en)* (green)

Output:
  • a brown dog is chasing a red frisbee across a grassy field

Input: *groß(en)* (big)
Example predictions marked as errors

Input: *Feld* (field)
Output:
- a brown and black dog running through a grassy field *

Input: *grün(en)* (green)
Output:
- a brown dog is chasing a red frisbee across a grassy field *

Input: *groß(en)* (big)
Output:
- a small group of people sitting together outside *
Error analysis by annotator

- DETextPrior
- DEVisionCNN
- XVisionSpeechCNN
- XBoWCNN

$P@10$ (%)
Error analysis by annotator

![Error analysis graph]

- DETextPrior
- DEVisionCNN
- XVisionSpeechCNN
- XBoWCNN

P@10 (%)
Error analysis by annotator

![Error analysis graph](image)
Cross-lingual keyword spotting

Written query:

burning

(English)
Conclusions and future work

• Visual grounding makes it possible to perform cross-lingual keyword spotting without any parallel speech and text or translations
Conclusions and future work

• Visual grounding makes it possible to perform cross-lingual keyword spotting without any parallel speech and text or translations

• Future: Apply approach to a truly low-resource language
Conclusions and future work

- Visual grounding makes it possible to perform cross-lingual keyword spotting without any parallel speech and text or translations
- Future: Apply approach to a truly low-resource language
- Perform error analysis on larger scale
Conclusions and future work

• Visual grounding makes it possible to perform cross-lingual keyword spotting without any parallel speech and text or translations

• Future: Apply approach to a truly low-resource language

• Perform error analysis on larger scale

• Visual tagger improvements: language-agnostic visual recognition
https://github.com/kamperh/
Training: Visually grounded model

German (text) tags

\( \hat{y}_{de} \) Feld Hunde springt

Cross-lingual keyword spotter

Loss \( \ell \)

\( f(X) \)

\( I \)

English speech

VGG-16

\( x \rightarrow f(X) \)

\( \hat{y}_{de} \)

Cross-lingual keyword spotter
Given German keyword: ‘Hunde’

English speech collection (want to search)

\[ f_w(X_i) = P_\theta(w|X_i) \]: score for whether (English) speech \( X_i \) contains translation of given (German) keyword \( w \)