Visually grounded cross-lingual keyword spotting in speech

SLTU, August 2018

Herman Kamper[1] and Michael Roth[2]

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

http://www.kamperh.com/

Advances in speech recognition

Advances in speech recognition

- **Addiction to labels**: 2000 hours transcribed speech audio; ∼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]
- Very different from the "supervision" infants use to learn language

Advances in speech recognition

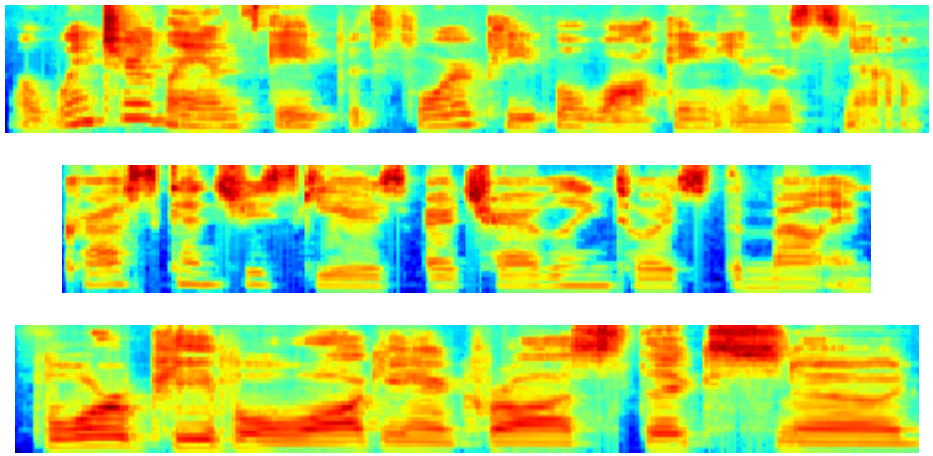
- **Addiction to labels**: 2000 hours transcribed speech audio; ∼350M/560M words text [Xiong et al., TASLP'17]
- Very different from the "supervision" infants use to learn language
- Sometimes not possible, e.g., for unwritten languages

Images as weak labels for speech

Images as weak labels for speech

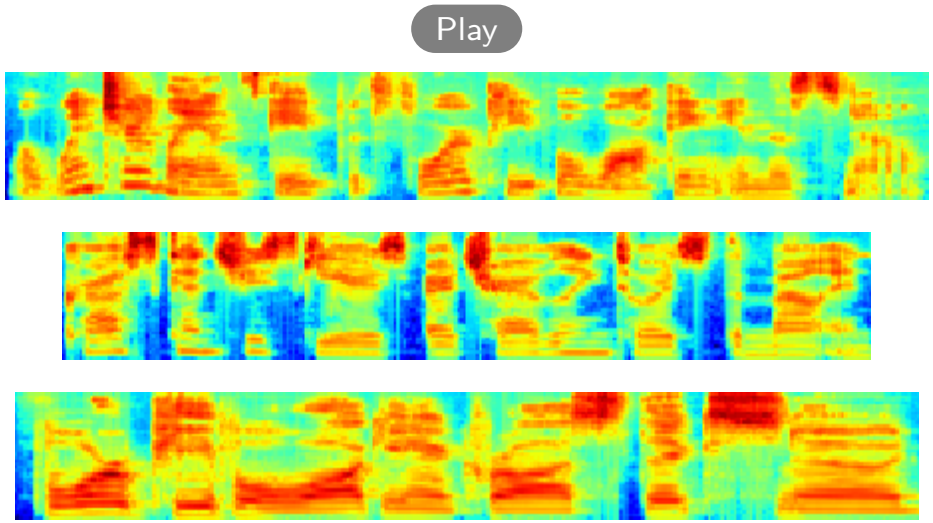
Can we use images as weak labels in low-resource settings?

Play



Images as weak labels for speech

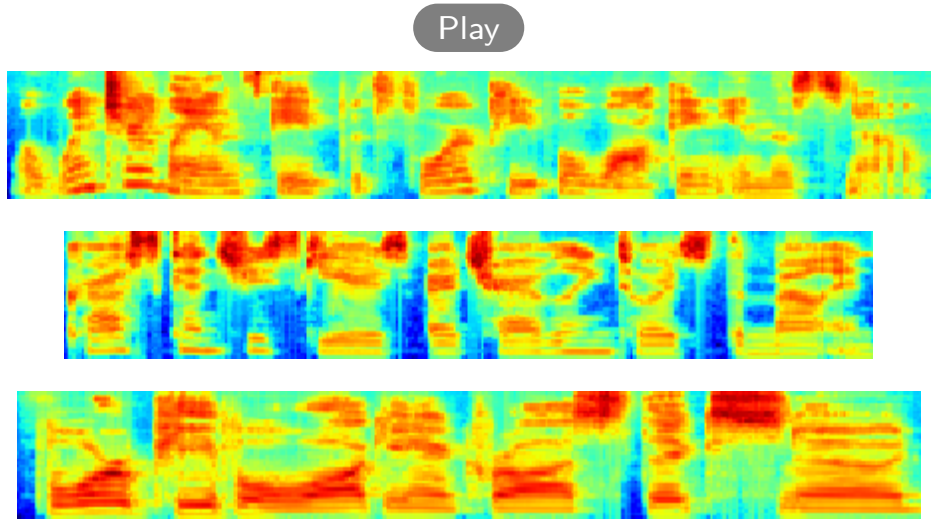
Can we use images as weak labels in low-resource settings?



- 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?



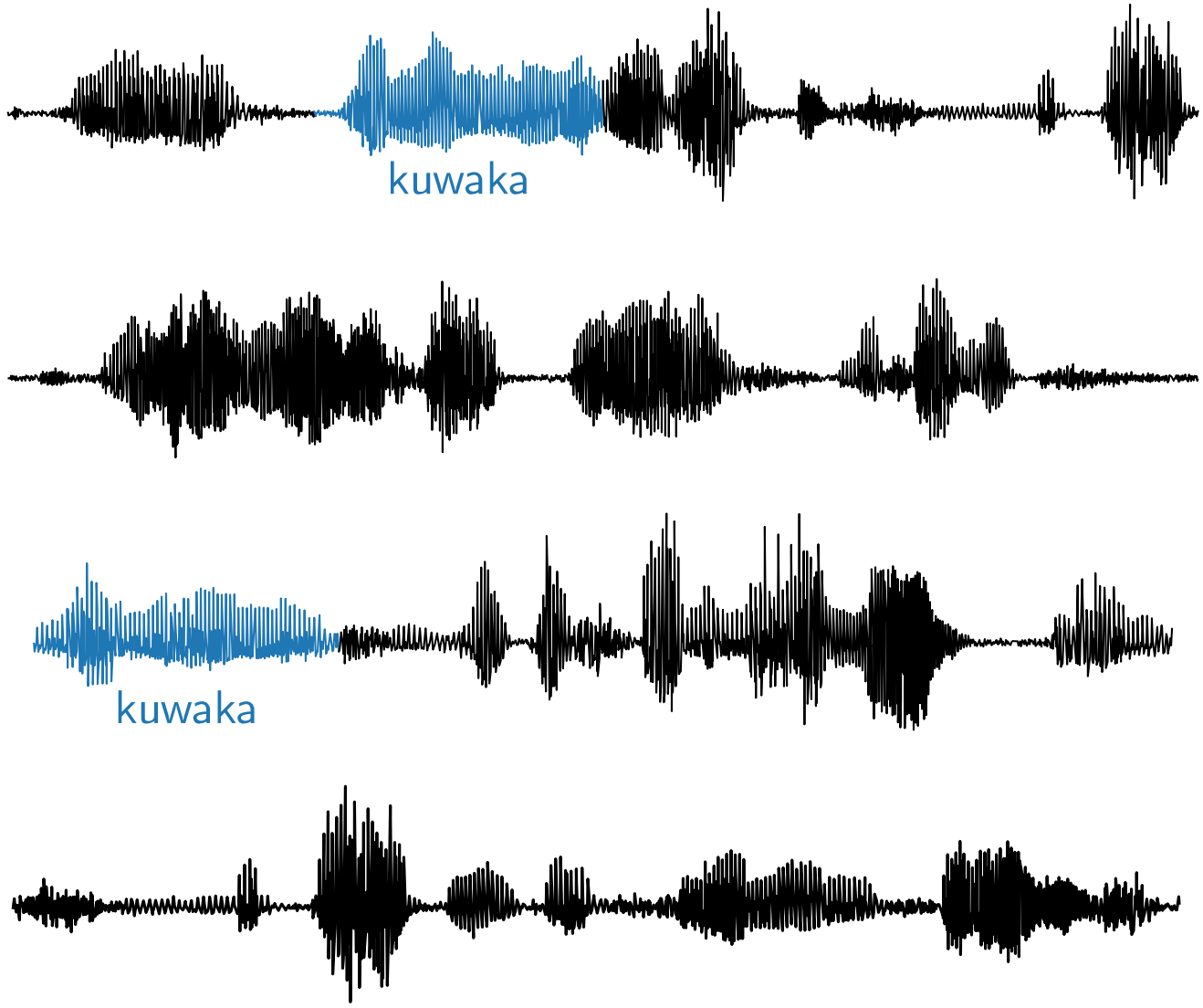
- Maybe we cannot use this type of data for full ASR, but maybe it can be used for other tasks?
- **Goal:** Use this type of data for cross-lingual keyword spotting

Cross-lingual keyword spotting

Written query:

burning

(English)



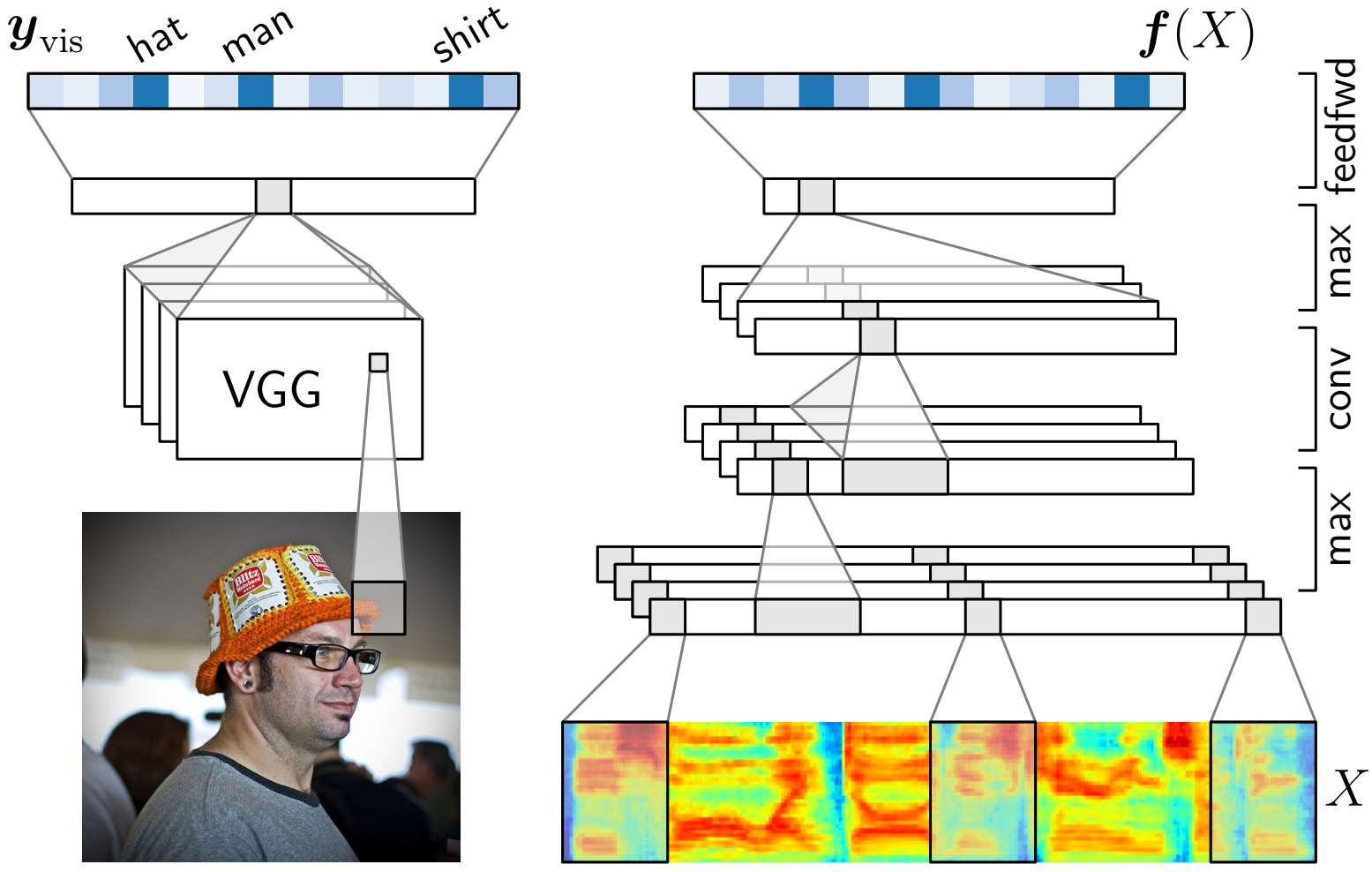
Swahili speech corpus

Cross-lingual word prediction from images

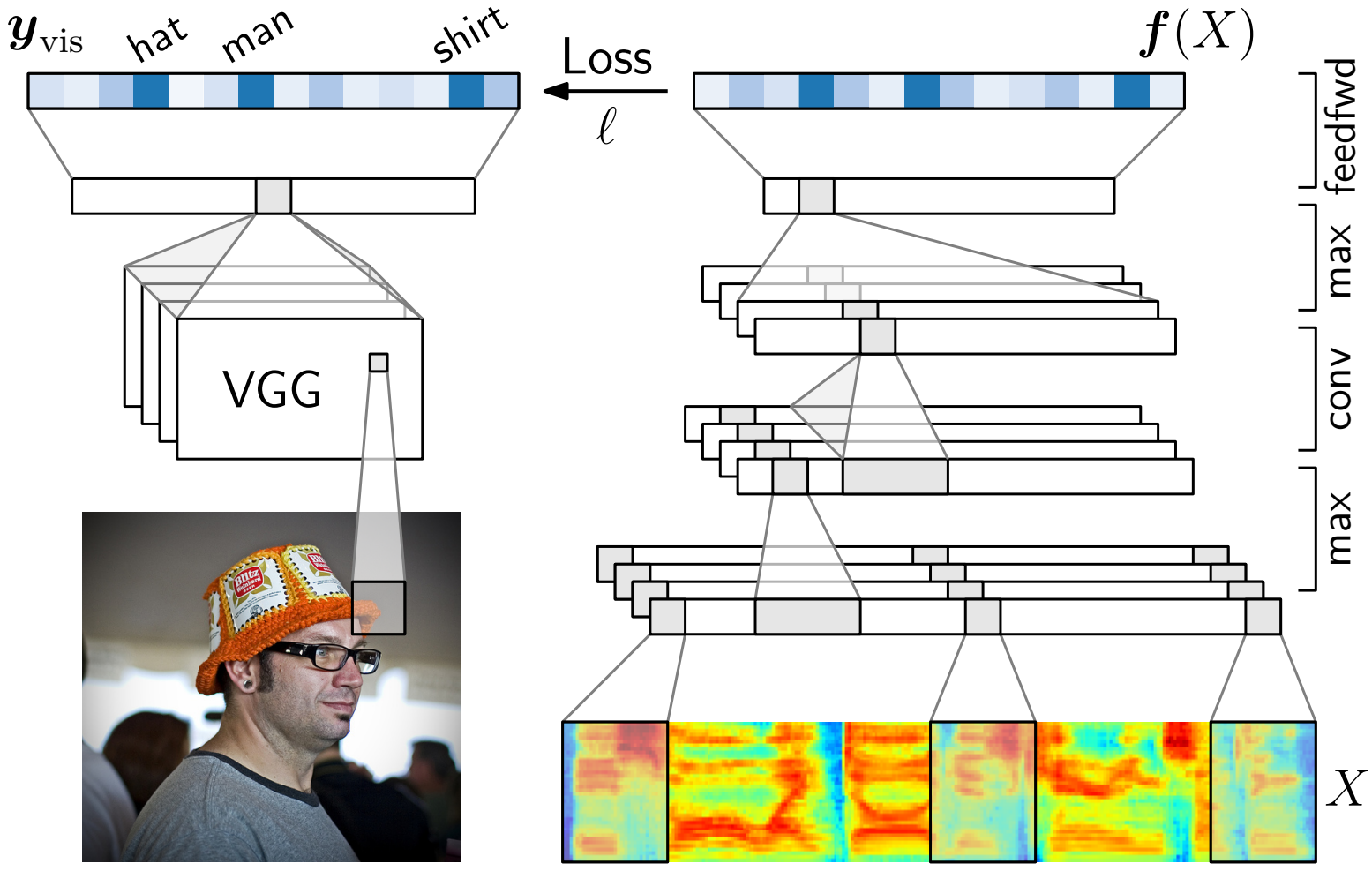
Cross-lingual word prediction from images

Cross-lingual word prediction from images

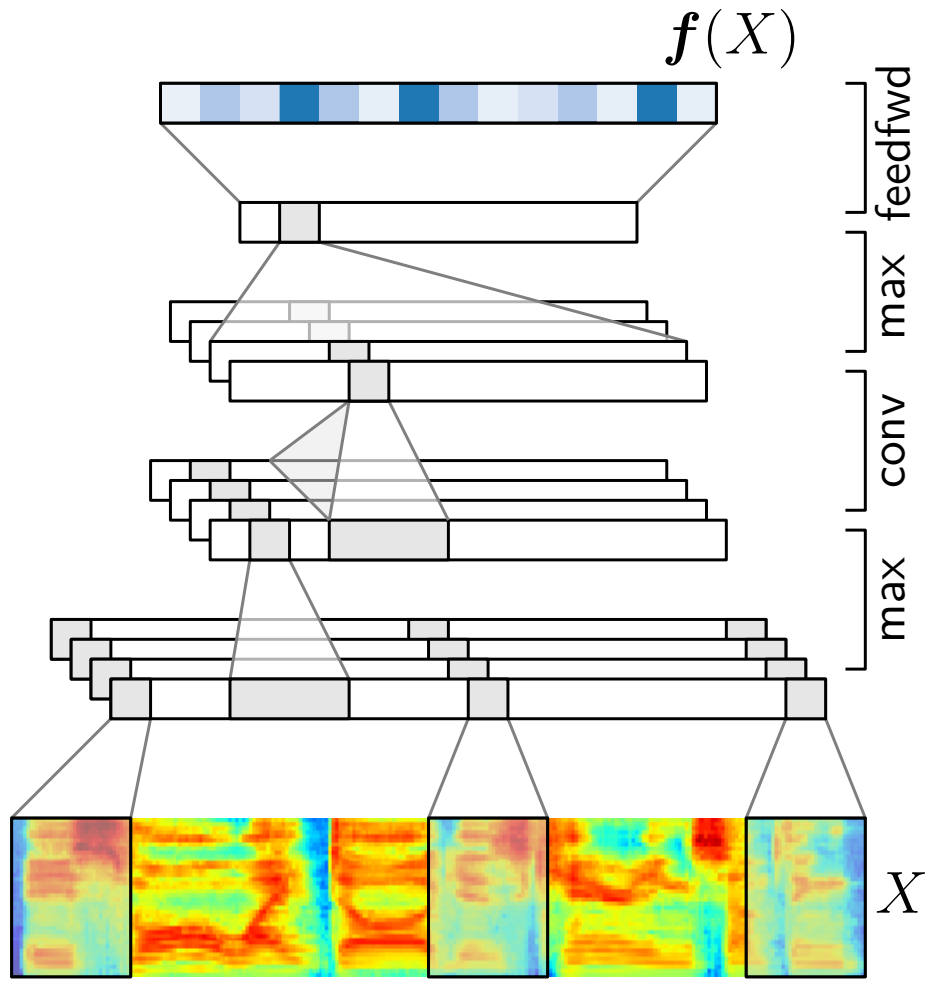
Cross-lingual word prediction from images



Cross-lingual word prediction from images

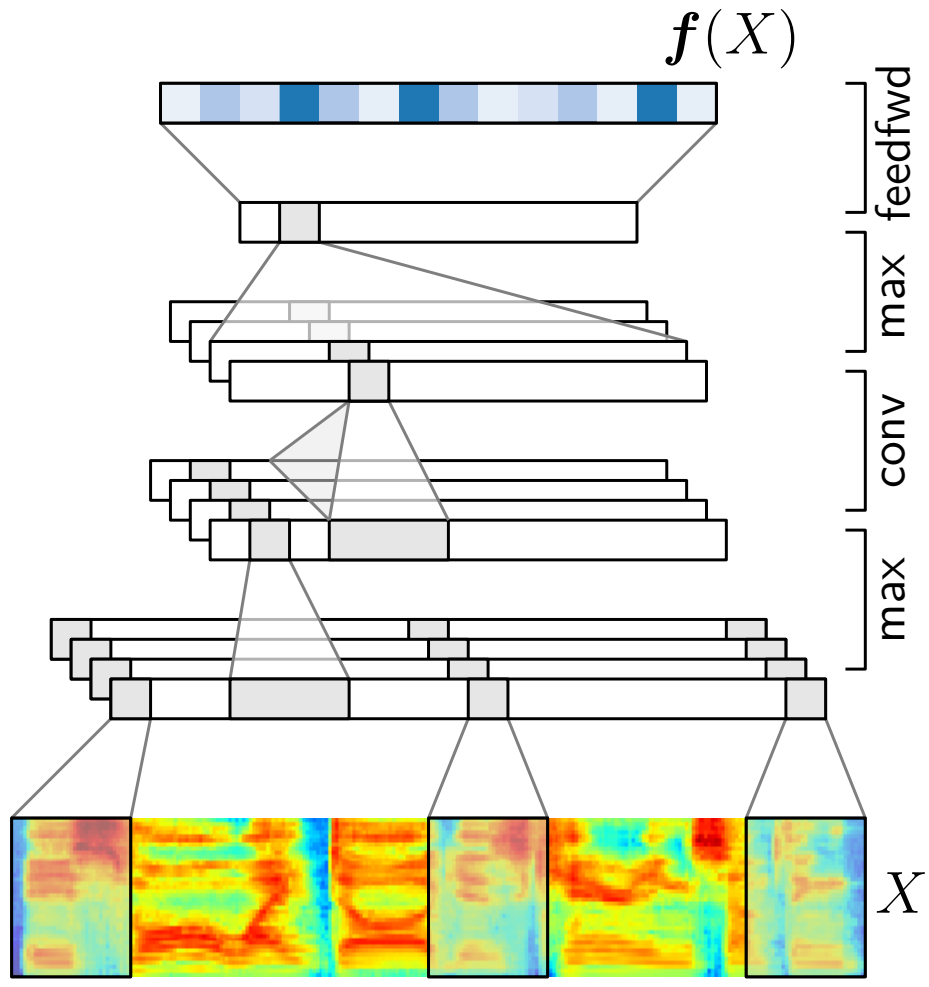


Cross-lingual word prediction from images



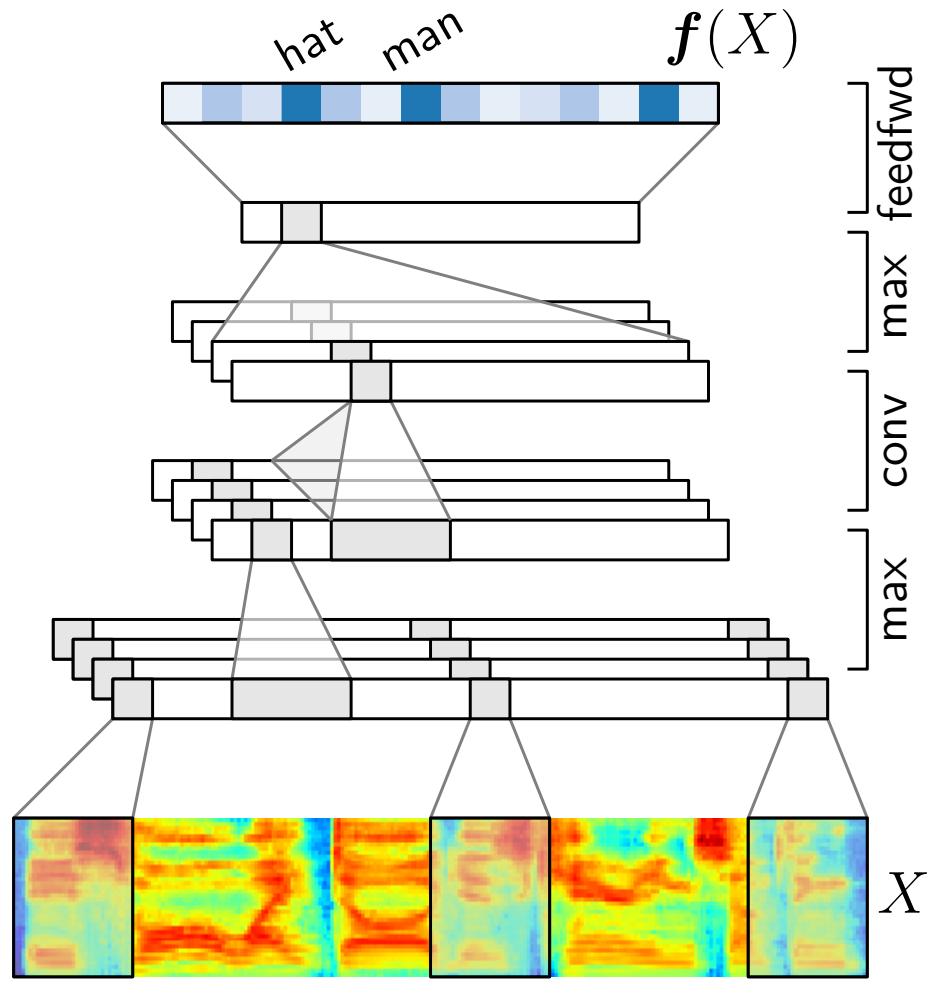
Cross-lingual word prediction from images

Swahili speech



Cross-lingual word prediction from images

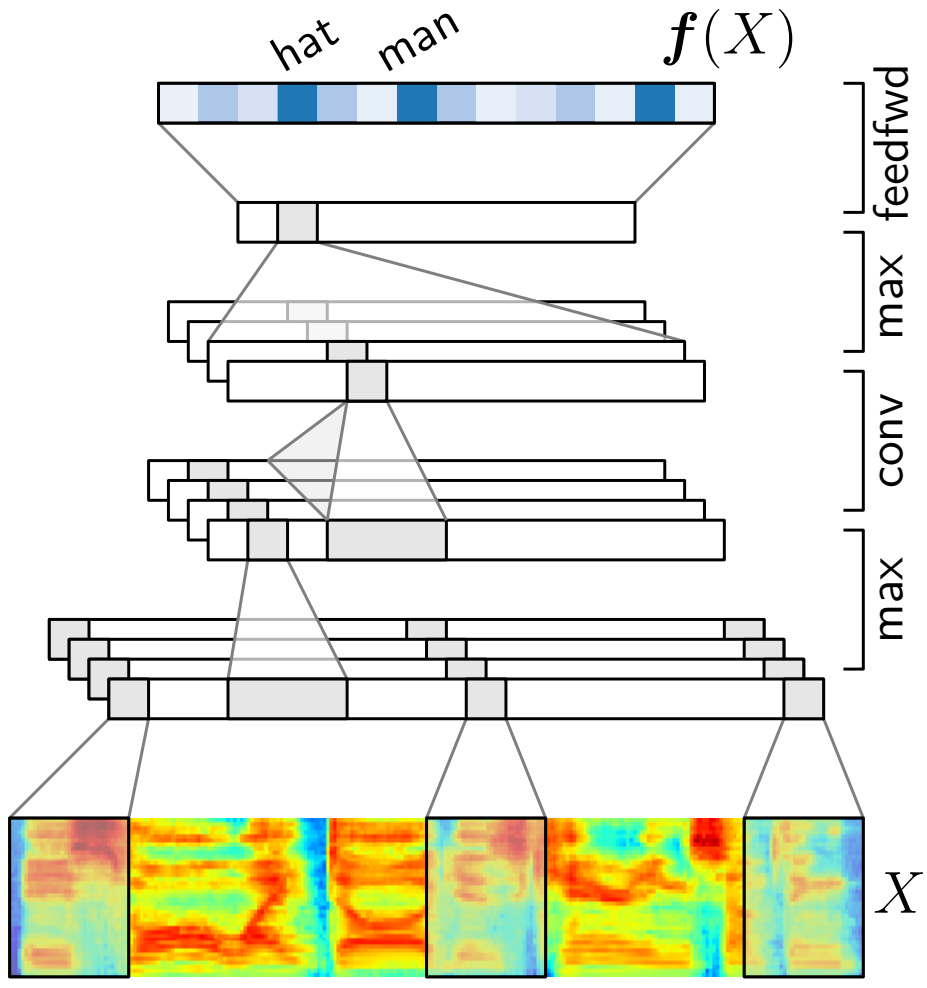
Swahili speech



Cross-lingual word prediction from images

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

Swahili speech

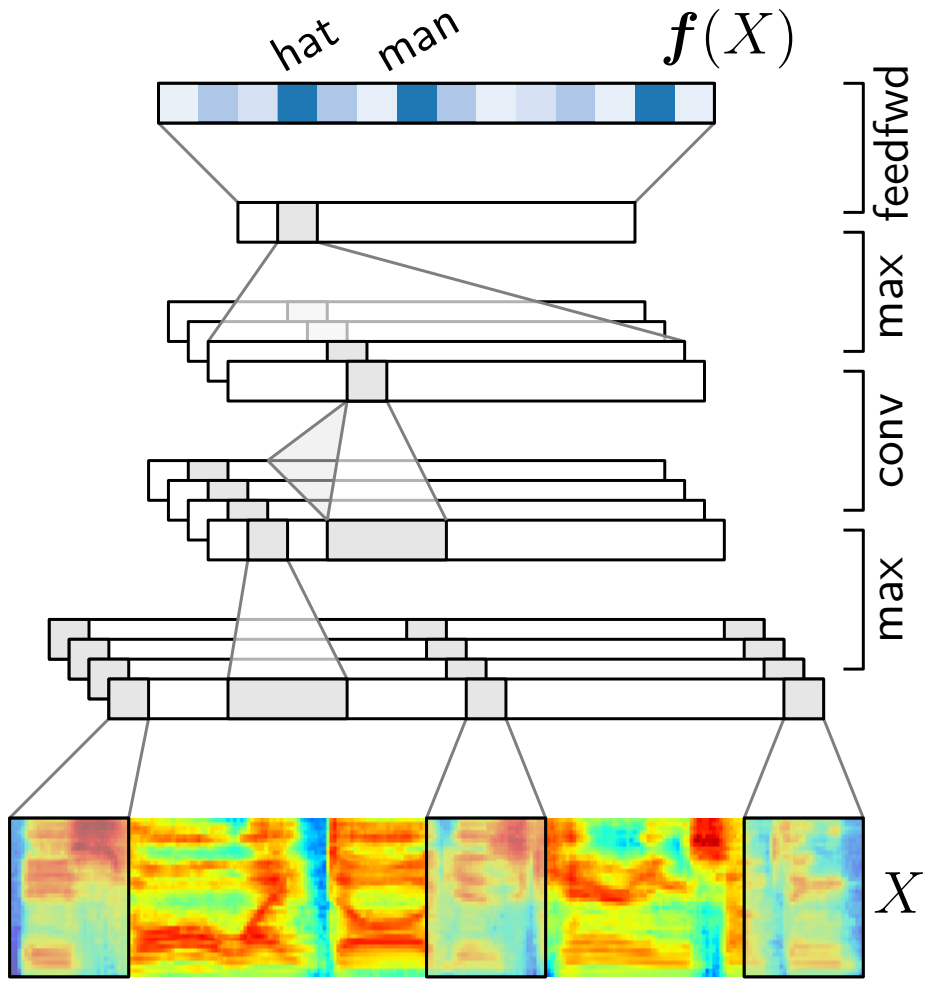


Cross-lingual word prediction from images

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

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

Swahili speech



Experimental details

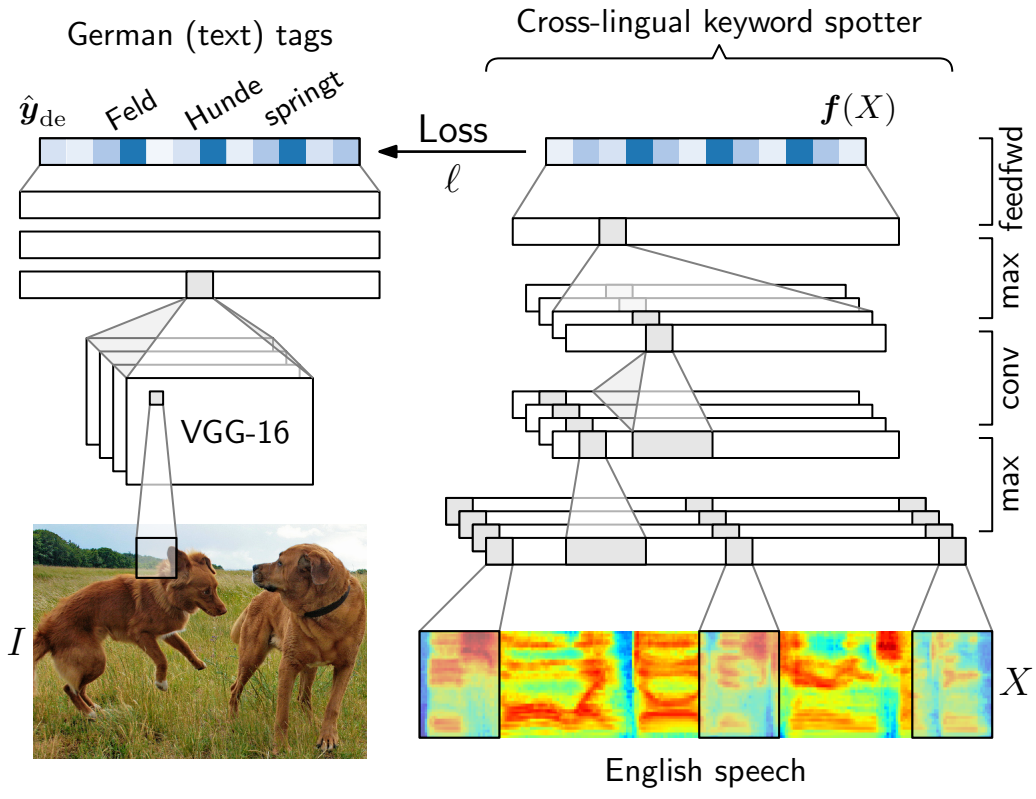
- **Goal:** Use visual grounding for cross-lingual keyword spotting

Experimental details

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

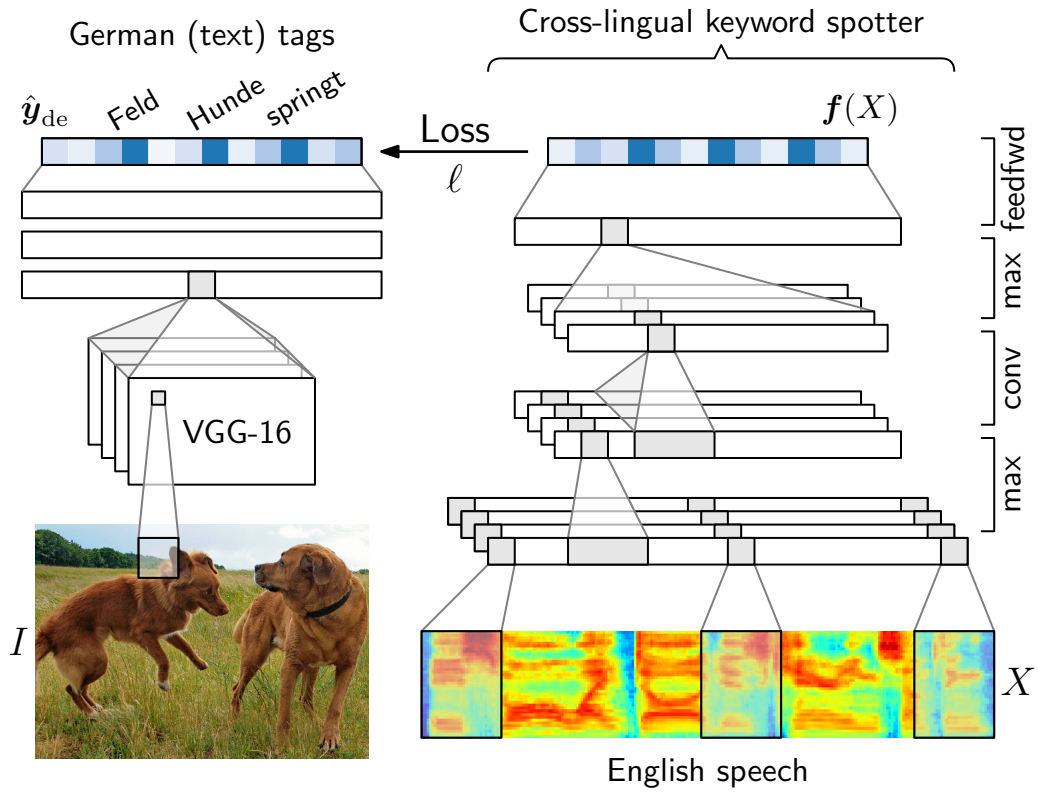
Experimental details

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



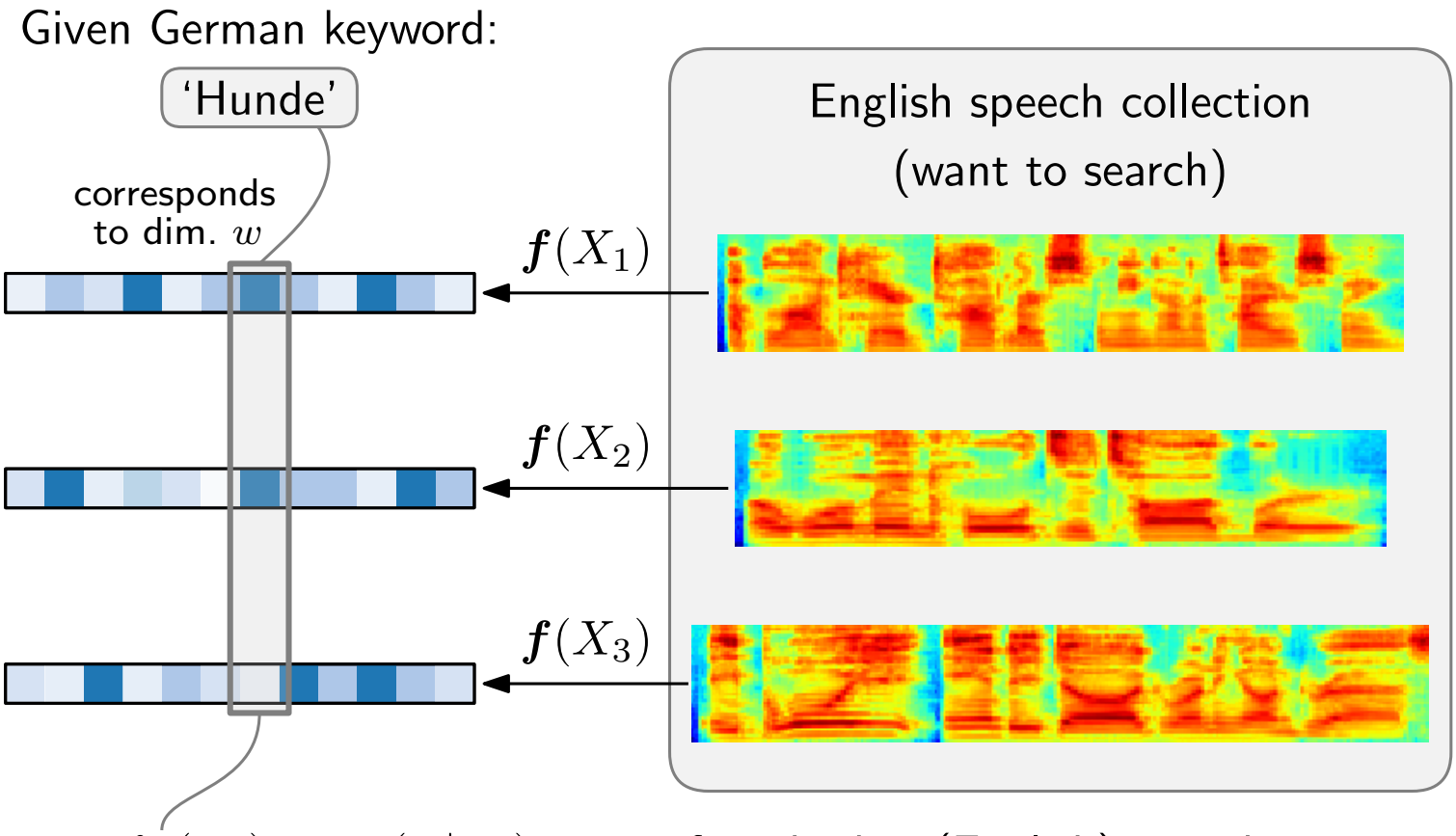
Experimental details

- **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



$f_w(X_i) = P_{\boldsymbol{\theta}}(w|X_i)$: 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

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

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):

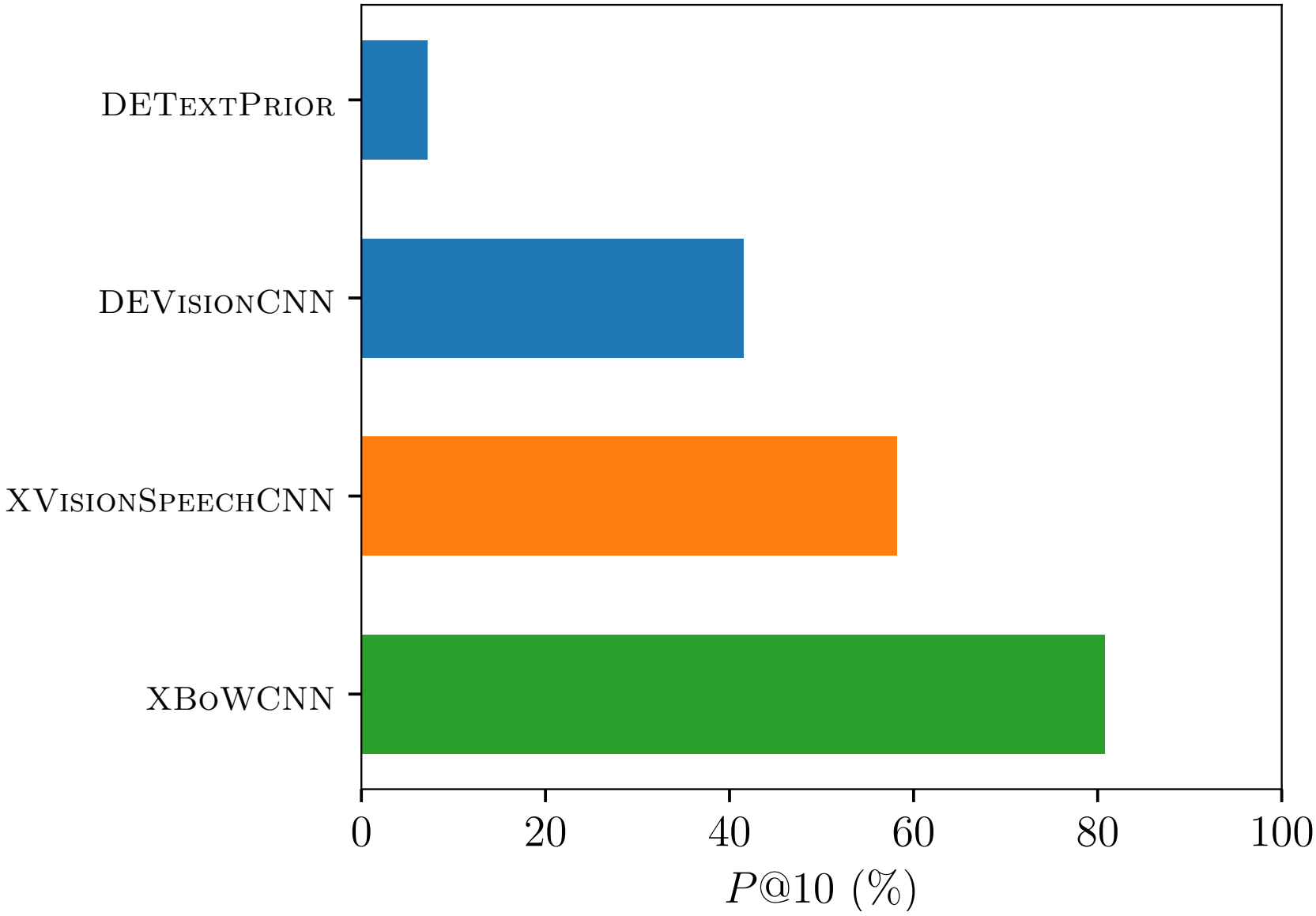
- man riding a bicycle on a foggy day
- a biker does a trick on a ramp
- a person is doing tricks on a bicycle in a city

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



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:

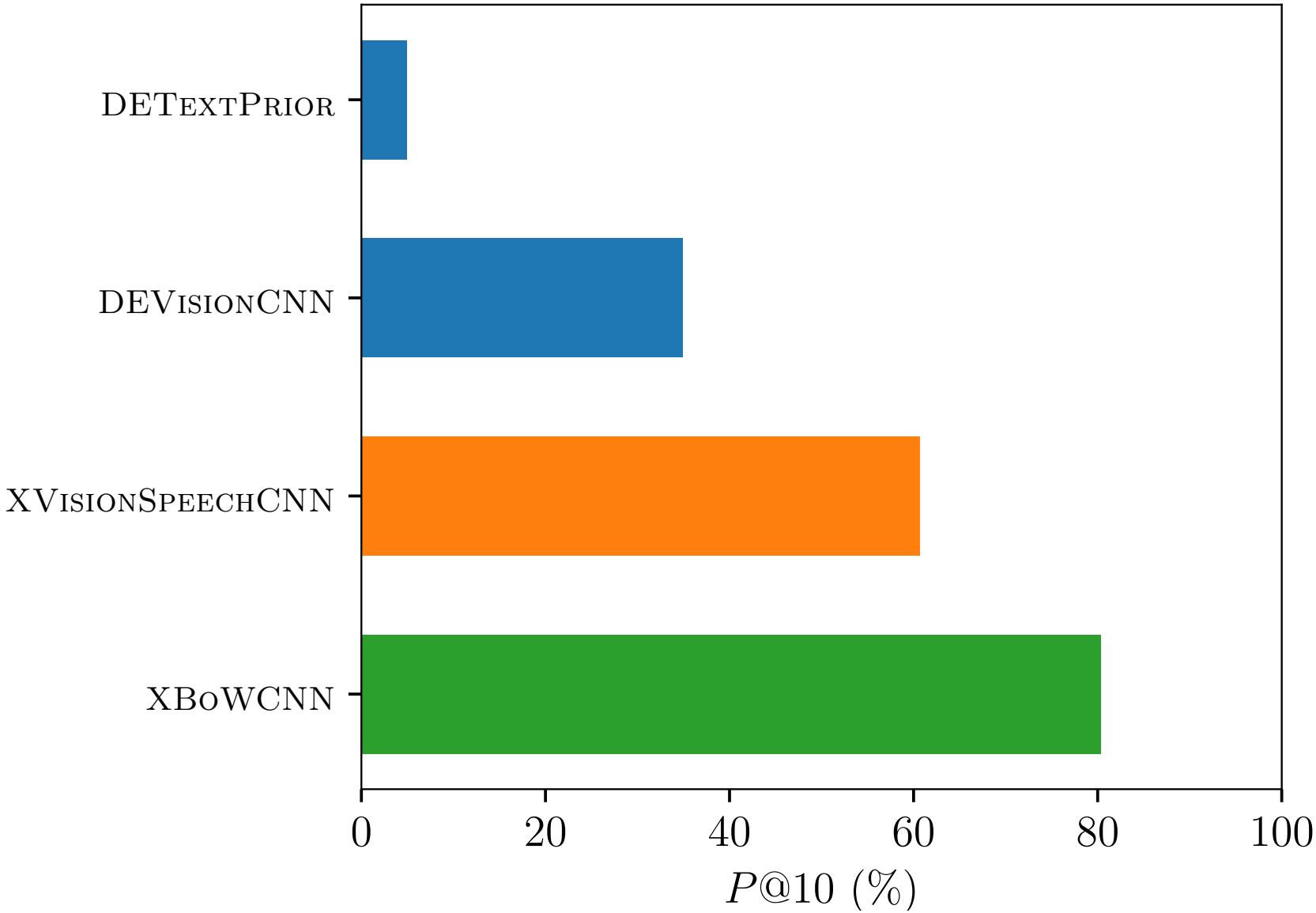
- a brown dog is chasing a red frisbee across a grassy field *

Input: *groß(en)* (big)

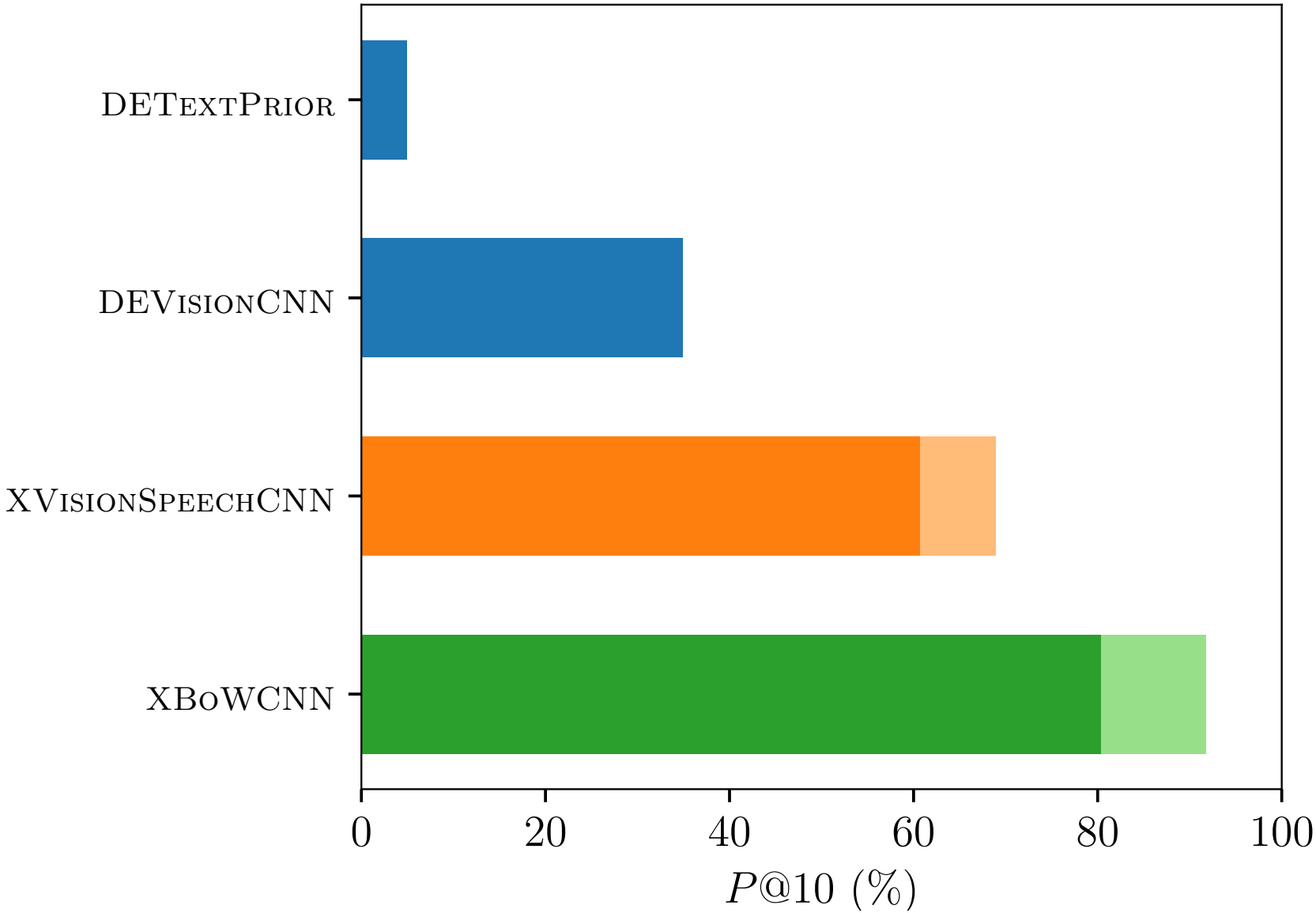
Output:

- a small group of people sitting together outside *

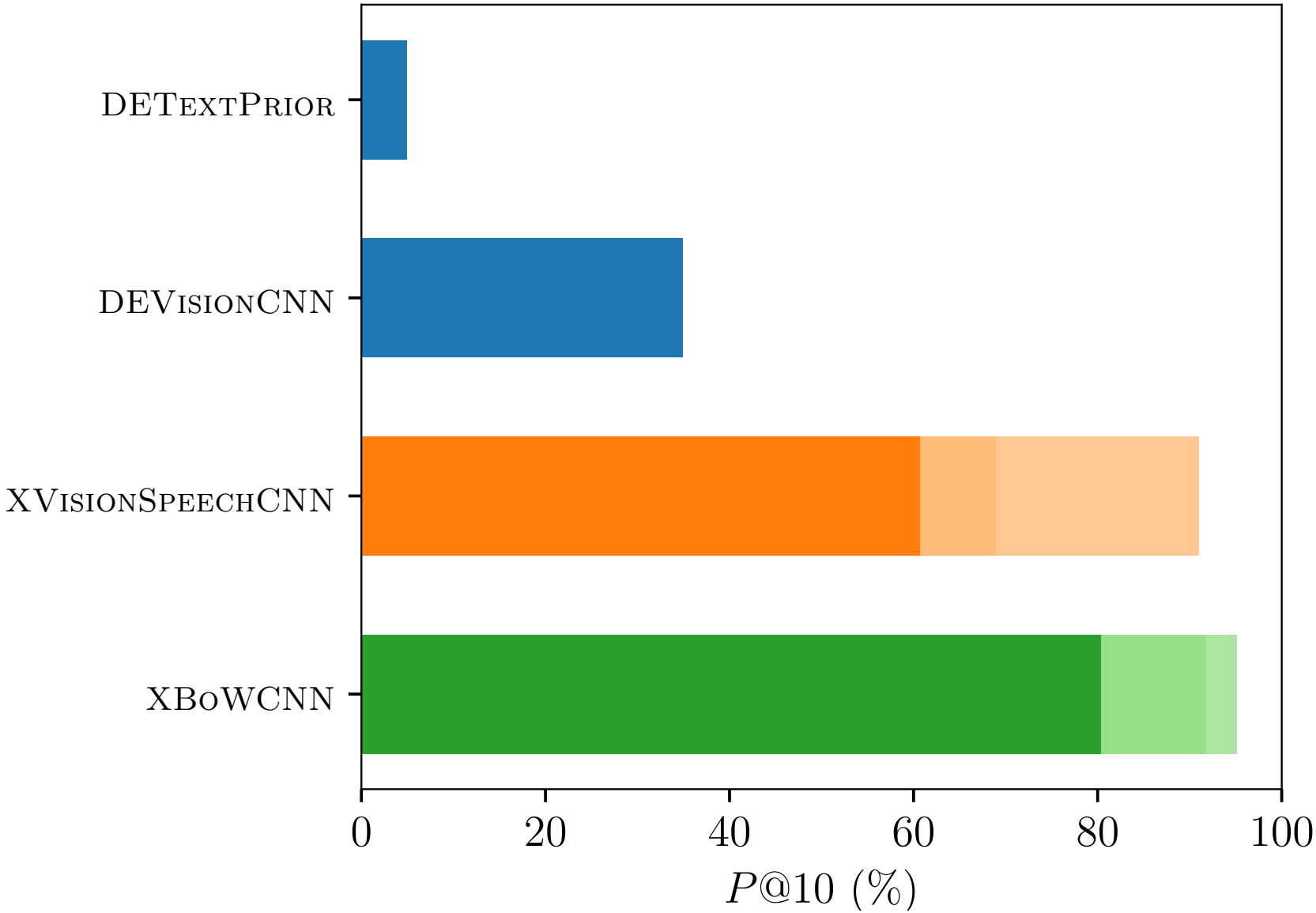
Error analysis by annotator



Error analysis by annotator



Error analysis by annotator

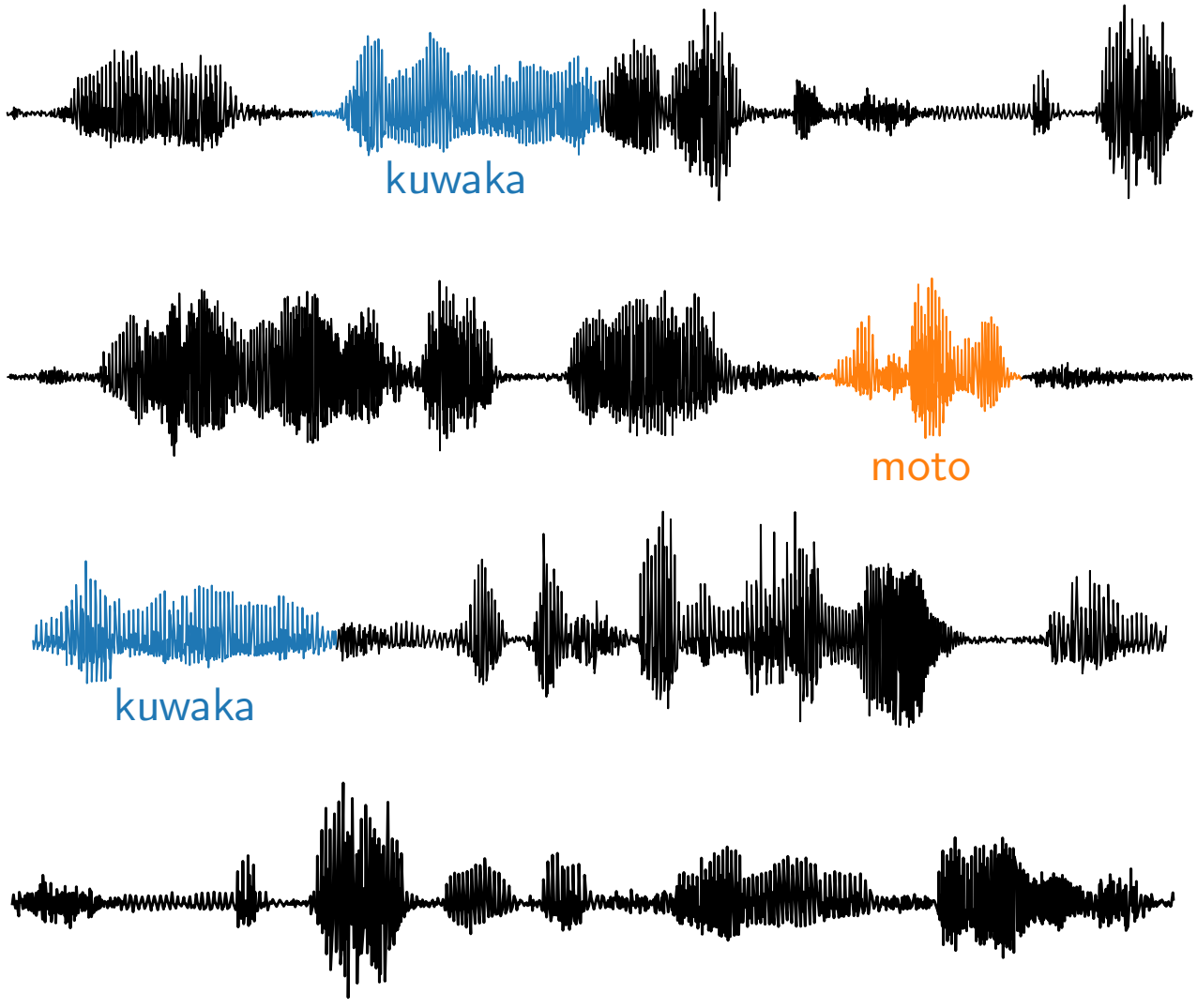


Cross-lingual keyword spotting

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

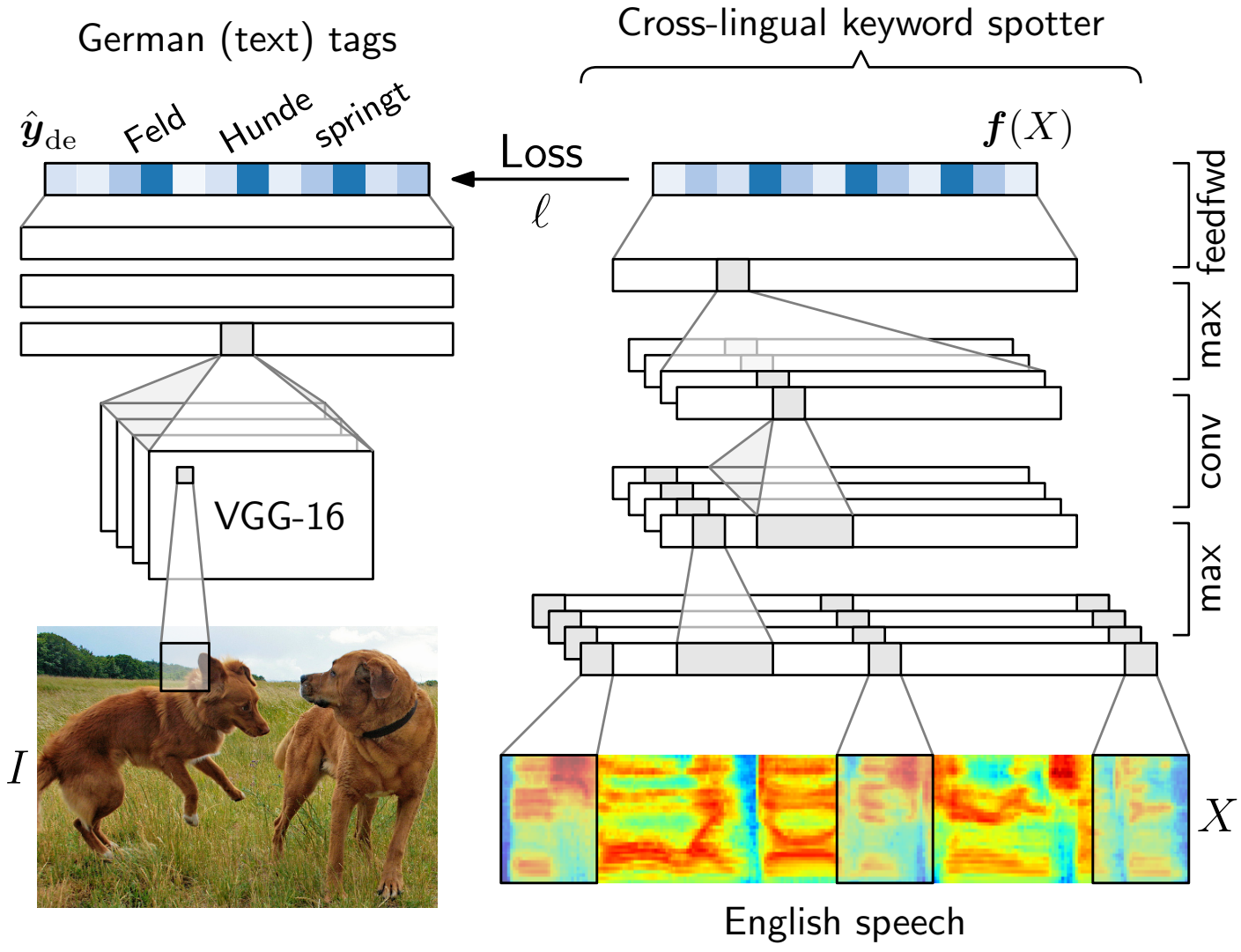
- Visual grounding makes it possible to perform cross-lingual keyword spotting without any parallel speech and text or translations
- Future: Apply approach to a truly low-resource language
- Perform error analysis on larger scale

Conclusions and future work

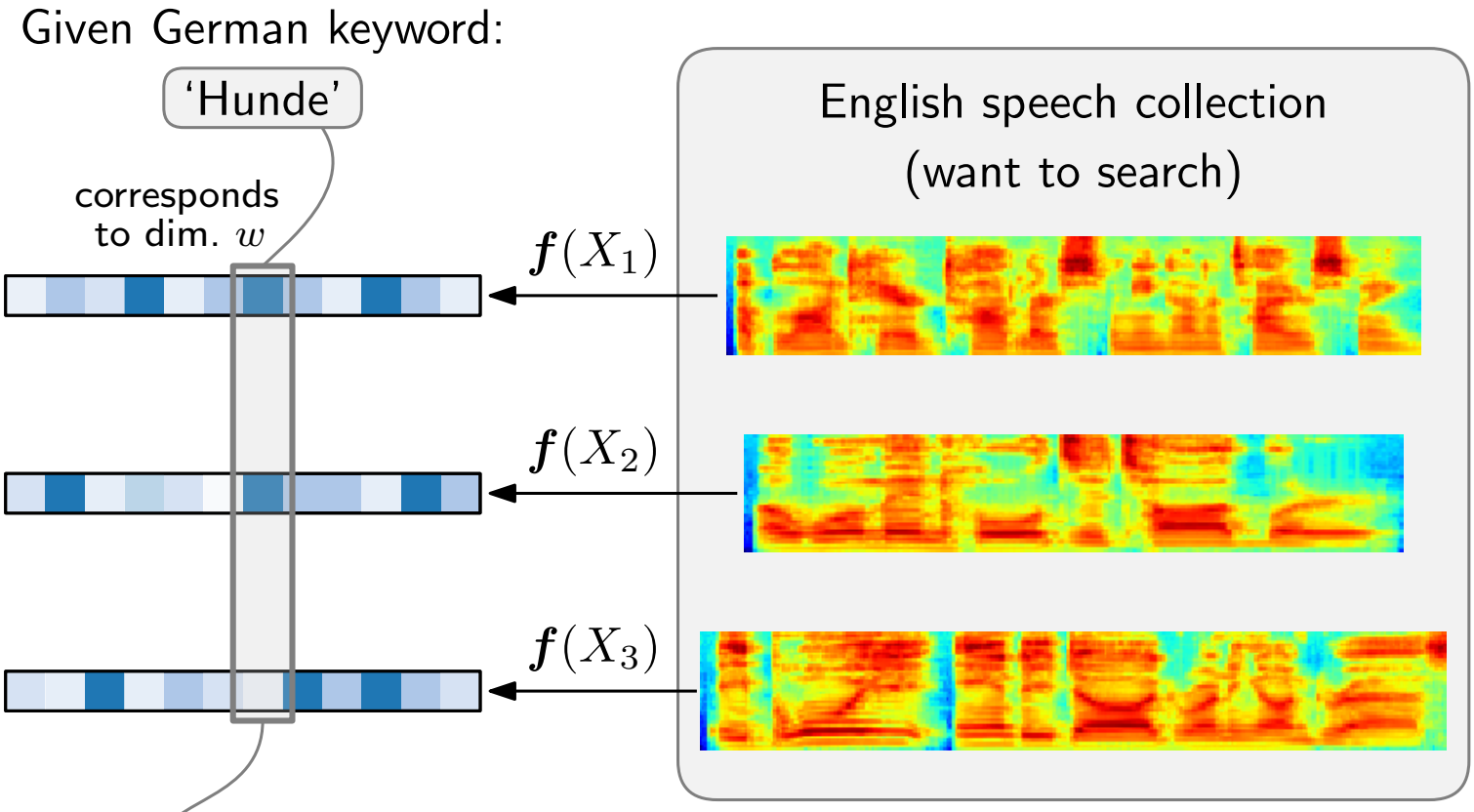
- Visual grounding makes it possible to perform cross-lingual keyword spotting without any parallel speech and text or translations
- Future: Apply approach to a truly low-resource language
- Perform error analysis on larger scale
- Visual tagger improvements: language-agnostic visual recognition

https://github.com/kamperh/

Training: Visually grounded model



Testing: Cross-lingual keyword spotting



$f_w(X_i) = P_{\boldsymbol{\theta}}(w|X_i)$: score for whether (English) speech X_i contains translation of given (German) keyword w