Towards Localisation of Keywords in Speech using Weak Supervision

Kayode Olaleye
kaykola.olaleye@gmail.com

Supervisor: Dr Herman Kamper

Department of Electrical and Electronic Engineering
Stellenbosch University

Introduction

Detection task

Utterance

A dog running through snow

Transcription

Localisation task

Utterance

| A | dog | running | through | snow |

0 10 40 100 160 200

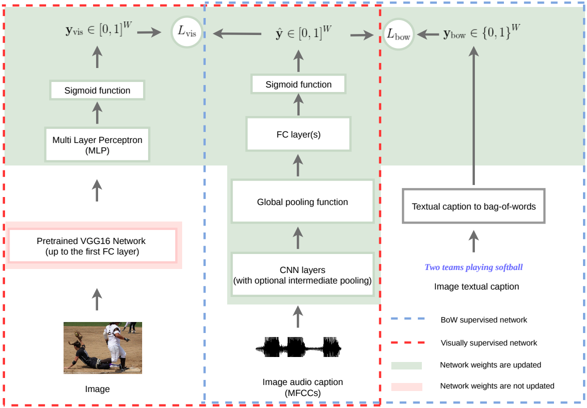
Frame-wise alignment
of transcription

Introduction (cont.)

Image

Utterance

Model structure

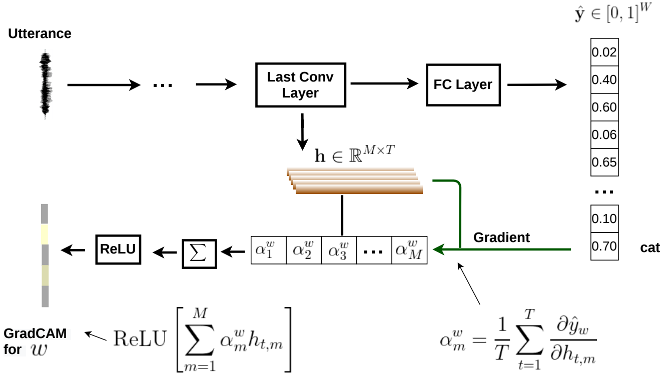

$\mathbf{y}_{\text{vis}} \in [0,1]^W$
L_{vis}
$\hat{\mathbf{y}} \in [0,1]^W$
L_{bow}
$\mathbf{y}_{\text{bow}} \in \{0,1\}^W$
Sigmoid function
Multi Layer Perceptron (MLP)
Pretrained VGG16 Network (up to the first FC layer)
Image
Sigmoid function
FC layer(s)
Global pooling function
CNN layers (with optional intermediate pooling)
Image audio caption (MFCCs)
Textual caption to bag-of-words
Two teams playing softball
Image textual caption
BoW supervised network
Visually supervised network
Network weights are updated
Network weights are not updated

Localisation methods

Two localisation methods:

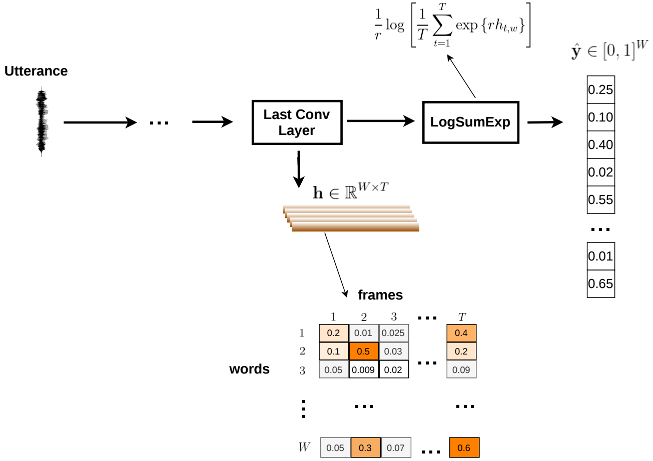
- **GradCAM**
  - Introduced in the vision domain to localise an object in an image.
  - Works with any **trained** CNN architecture.
  - Determines the portion of an input that contributes to a decision of interest using gradient information.
- **PSC**
  - Designed to simulateneously perform detection and localisation of keywords in speech utterance.
  - The CNN architecture is restricted in some ways (*No intermediate max-pooling; no fully-connected layers; LogSumExp function as the global pooling function*).

GradCAM



$$\mathrm{ReLU}\left[\sum_{m=1}^{M} \alpha_m^w h_{t,m}\right]$$

$$\alpha_m^w = \frac{1}{T} \sum_{t=1}^{T} \frac{\partial \hat{y}_w}{\partial h_{t,m}}$$



$$\frac{1}{r}\log\left[\frac{1}{T}\sum_{t=1}^{T}\exp\{rh_{t,w}\}\right]$$

$$\hat{\mathbf{y}} \in [0,1]^W$$

$\hat{\mathbf{y}}$
0.25
0.10
0.40
0.02
0.55
…
0.01
0.65

$$\mathbf{h} \in \mathbb{R}^{W\times T}$$

	1	2	3	…	T
1	0.2	0.01	0.025		0.4
2	0.1	0.5	0.03	…	0.2
3	0.05	0.009	0.02		0.09
⋮		…			…
W	0.05	0.3	0.07	…	0.6

Evaluation and Results

	Supervision method	
Mechanism	BoW	Visual
PSC	63.6	19.1
GradCAM	17.8	16.0

Table 1: Oracle localisation accuracy (%) when assuming perfect detection.

	BoW				Visually-supervised			
Mechanism	P	R	$F1$	Accuracy	P	R	$F1$	Accuracy
PSC	75.2	53.0	62.2	50.4	28.6	8.0	12.5	7.6
GradCAM	17.7	24.5	20.5	13.2	5.0	5.7	5.3	4.4

Table 2: Actual localisation precision, recall, $F1$ and accuracy (%) when taking detection into account with a threshold of $\lambda = 0.4$.

Evaluation and Results (cont.)

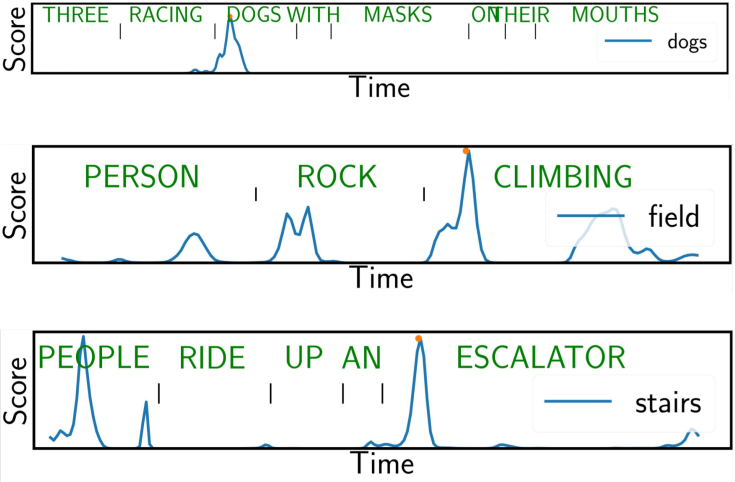


Figure 2: Examples of localisation with the visually supervised PSC mechanism. The keyword being localised is shown on the right of each plot.

Conclusions

- We asked whether keyword localisation in speech is possible with two forms of weak supervision when location information is not provided.
- We attempted to answer the question by comparing two localisation methods (PSC and GradCAM) with two forms of supervision: bag-of-word (BoW) labels and visual context.
- While the GradCAM (a saliency-based method) performed poorly, PSC (a method where localisation is performed as part of the network) performed well with BoW supervision and showed that visual supervision does provide potential for higher precision localisation.
- Our results suggests a mismatch between saliency-based localisation and the multi-label model used here, with a superior detection model performing poorly in localisation. This suggest that better localisation should be possible given a mechanism better aligned to the model and multi-label classification loss.

References

1. Palaz, Dimitri, Gabriel Synnaeve, and Ronan Collobert. "Jointly Learning to Locate and Classify Words Using Convolutional Networks." INTERSPEECH, 2016.
2. Chrupała, Grzegorz, Lieke Gelderloos, and Afra Alishahi. "Representations of language in a model of visually grounded speech signal." arXiv preprint arXiv:1702.01991, 2017.
3. Kamper, Herman, Aristotelis Anastassiou, and Karen Livescu. "Semantic query-by-example speech search using visual grounding." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.
4. Abdel-Hamid, Ossama, et al. "Convolutional neural networks for speech recognition." IEEE/ACM Transactions on audio, speech, and language processing, 2014.
5. Doersch, Carl, and Andrew Zisserman. "Multi-task self-supervised visual learning." Proceedings of the IEEE International Conference on Computer Vision, 2017.

References (cont.)

6. Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." Advances in neural information processing systems, 2016.
7. Harwath, David, and James Glass. "Deep multimodal semantic embeddings for speech and images." 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, 2015.
8. Synnaeve, Gabriel, Maarten Versteegh, and Emmanuel Dupoux. "Learning words from images and speech." NIPS Workshop Learn. Semantics, 2014.
9. Kamper, Herman, and Michael Roth. "Visually grounded cross-lingual keyword spotting in speech." In Proc. SLTU, 2018.
10. Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision, 2017.
11. Harwath, David, et al. "Jointly discovering visual objects and spoken words from raw sensory input." Proceedings of the European conference on computer vision (ECCV), 2018.

References (cont.)

12. Harwath, David, and James R. Glass. "Learning word-like units from joint audio-visual analysis." In Proc. ACL, 2017.
13. Settle, Shane, et al. "Query-by-example search with discriminative neural acoustic word embeddings." In Proc. Interspeech, 2017.
14. Bansal, Sameer, et al. "Towards speech-to-text translation without speech recognition." In Proc. EACL, 2017.
15. Kamper, Herman, Gregory Shakhnarovich, and Karen Livescu. "Semantic speech retrieval with a visually grounded model of untranscribed speech." IEEE/ACM Trans. Acoust., Speech, Signal Process, 2018.
16. Kamper, Herman, Aristotelis Anastassiou, and Karen Livescu. "Semantic query-by-example speech search using visual grounding." In Proc. ICASSP, 2019.
17. Pasad, Ankita, et al. "On the contributions of visual and textual supervision in low-resource semantic speech retrieval." In Proc. Interspeech, 2019.

Thank you for listening!

Supplementary materials

	$\alpha = 0.4$			$\alpha = 0.6$		
Model	P	R	$F1$	P	R	$F1$
Visual supervision:						
PSC	44.5	9.8	16.1	74.7	4.3	8.1
GradCAM	29.3	22.0	25.1	42.7	12.7	19.6
BoW supervision:						
PSC	82.2	49.0	61.4	87.8	46.1	60.4
GradCAM	79.3	52.6	63.2	82.5	50.9	63.0

Table 3: Keyword detection scores (without considering localisation) with threshold α.

Supplementary materials (cont.)

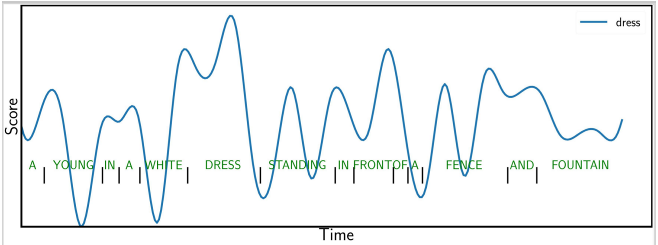


Figure 3: An example localisation with the GradCAM model for the keyword "dress".