On the Contributions of Visual and Textual Supervision in Low-Resource Semantic Speech Retrieval

Ankita Pasad[1], Bowen Shi[1], Herman Kamper[2], Karen Livescu[1]

[1]Toyota Technological Institute at Chicago, USA [2]Dept. E&E Engineering, Stellenbosch University, South Africa

Overview

- Background: Visual grounding is a commonly used source of weak supervision for tasks involving untranscribed spoken data (e.g. [1,2]).
- Open question: Does visual grounding still help if we have text annotations during training?
- Our setting: A low-resource setting where a fraction of the spoken training corpus is transcribed.
- Our work: Explores how to best combine the two modalities for *semantic speech retrieval*

Query word	Retrieved utterance
kids	a group of young boys playing soccer
beach	a dog retrieves a branch from a beach

Multi-Task Learning (MTL) Approach

What are the multiple tasks?

Visually supervised task (MTL-visSup)

Trained on *image-speech pairs*[1]. An external *image tagger*[2] provides weak labels as ground truth.

Textually supervised task (MTL-textSup)

Trained on *speech-text pairs*[3]. Each transcript provides a multi-hot bag-of-words vector as ground truth.

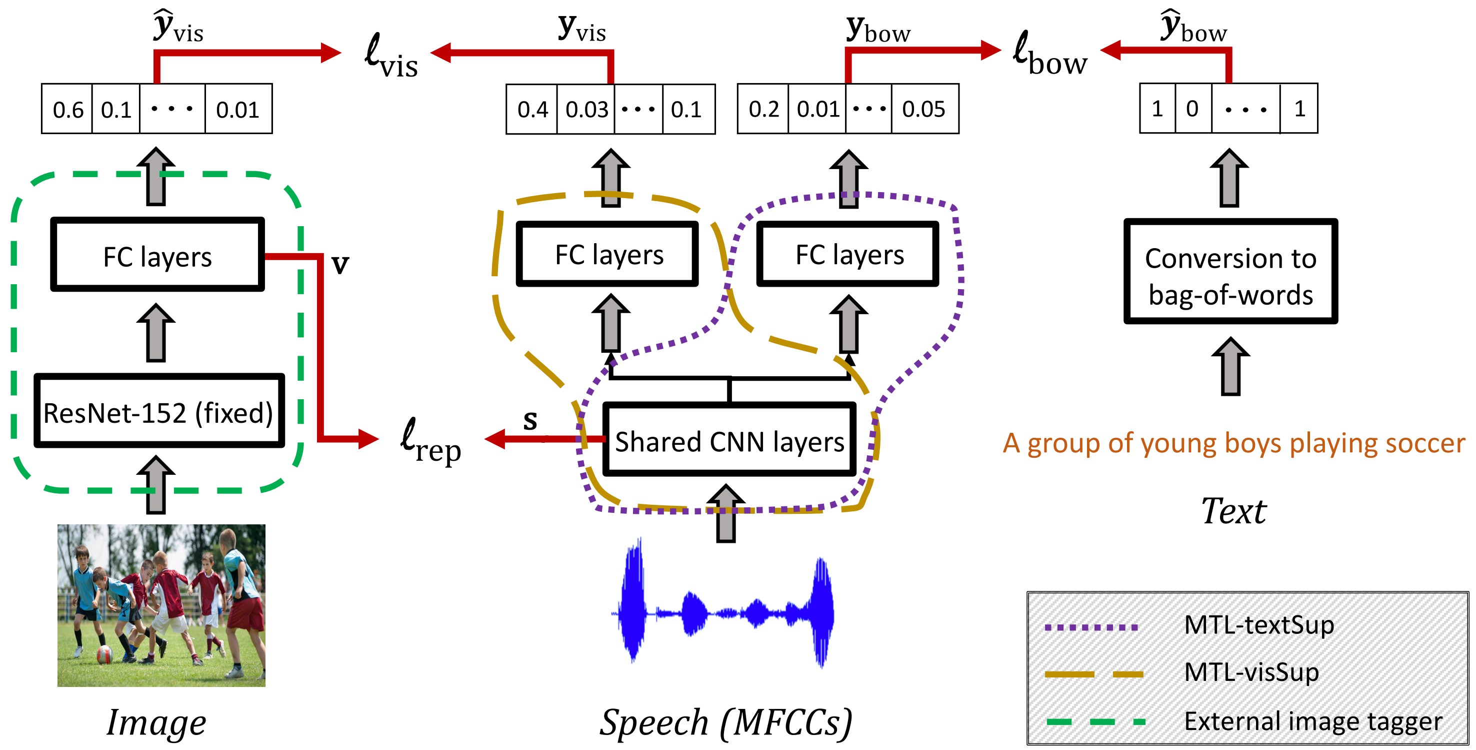
Unsupervised representation learning

Trained using the intermediate visual and speech representations. The former is fixed; the latter is updated during training.

What are the loss functions?

Supervised task losses (sup∈{vis,bow}): summed cross entropy between the predicted and ground truth vectors.

$$\ell_{\text{sup}} = -\sum_{w=1}^{|N_{\text{sup}}|} \{\hat{y}_{\text{sup,w}} \log y_{\text{sup,w}} + (1-\hat{y}_{\text{sup,w}}) \log[1-y_{\text{sup,w}}]\}$$



Representation loss: margin-based contrastive loss with margin m, positive pair $\{\boldsymbol{v}, \boldsymbol{s}\}$, negative pairs $\{\boldsymbol{v}', \boldsymbol{s}\}$ and $\{\boldsymbol{v}, \boldsymbol{s}'\}$, and cosine distance.

$$\ell_{\text{rep}} = \left\{ \frac{1}{|V|} \sum_{\boldsymbol{v}' \in V} \max[0, m + d_{\cos}(\boldsymbol{v}, \boldsymbol{s}) - d_{\cos}(\boldsymbol{v}', \boldsymbol{s})] + \frac{1}{|S|} \sum_{\boldsymbol{s}' \in S} \max[0, m + d_{\cos}(\boldsymbol{v}, \boldsymbol{s}) - d_{\cos}(\boldsymbol{v}, \boldsymbol{s}')] \right\}$$

Total loss: weighted sum of the three losses.

$$\ell = \alpha \cdot \ell_{\text{vis}} + \beta \cdot \ell_{\text{bow}} + (1-\alpha-\beta) \cdot \ell_{\text{rep}}$$

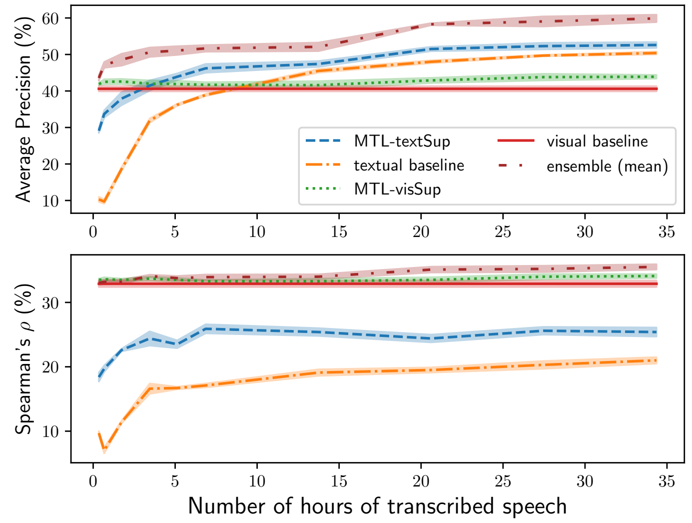
How is the inference done?

Input: Spoken utterances

Output: Scores from either MTL-visSup ($\boldsymbol{y}_{\text{vis}}$) or MTL-textSup ($\boldsymbol{y}_{\text{bow}}$) or a combination of the two

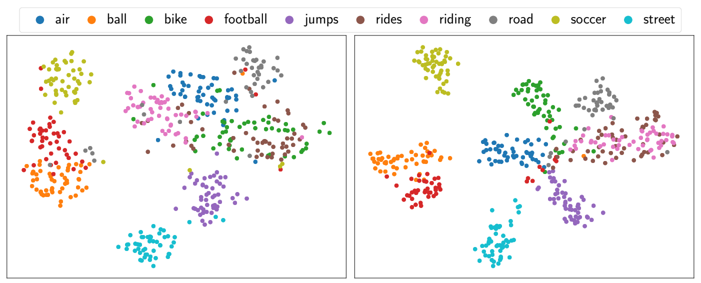
Main Results

The models are evaluated on a corpus of *semantic relevance judgements*[4] (collected by Kamper et al. [1]).



Additional Observations

- Adding representation loss gives a gain of 7-15% on average precision.
- Higher output dimensionality acts as a regularizer in lower supervision conditions.
- The proposed model outperforms pre-training and hierarchical MTL.
- t-SNE visualization of the learned representations in the text baseline (left) and MTL-textSup (right)



Conclusion

- Visual grounding helps even in the presence of textual supervision.
- Proposed MTL approach significantly improves performance at all levels of supervision.
- Joint training with representation loss helps.

Future Work

Domain extension: Does our visually grounded model perform well on speech not describing visual scenes?

Modify text encoder: Can we explicitly encode semantics in the textual supervision?

References

[1] D. Harwath *et al.* Unsupervised Learning of Spoken Language with Visual Context, *NIPS 2016*.

[2] H. Kamper *et al.* Semantic Speech Retrieval with a Visually Grounded Model of Untranscribed Speech, *IEEE TASLP, vol. 27, pp. 89-98, 2019*.

[1]The Flickr8k Audio Captions Corpus consisting of ∼8k images paired with 5 spoken captions each amounting to a total of ∼46 hours of speech data (∼34 hours training, ∼6 hours dev, and ∼6 hours test).

[2]ImageNet pre-trained fixed ResNet followed by fully connected layers trained on the union of MSCOCO and Flickr30k, with ∼149k images (∼107k training, ∼42k dev).

[3]Written transcripts of the Flickr8k Audio Captions. We use subsets of these transcripts with varying sizes: from just ∼21 minutes to the complete ∼34 hours of labelled speech.

[4]∼1k utterances from the Flickr8k Audio Captions Corpus with their semantic relevance for each of 67 query words. Each (utterance, keyword) pair was labeled by 5 annotators. Majority vote of the annotators ("hard labels") and the actual number of votes ("soft labels") for evaluation.