Grounding spoken dialog in cartoon videos

Abstract

1 Introduction

2 Related work

Audiovisual models Aytar et al. (2016); Owens et al. (2016a,b)

Video captioning Krishna et al. (2017); Zhou et al. (2018)

3 Method

4 Results

5 Conclusion

A Pilot

Overfitting a small training sample, with Wav2Letter and Wav2Vec (pretrained).

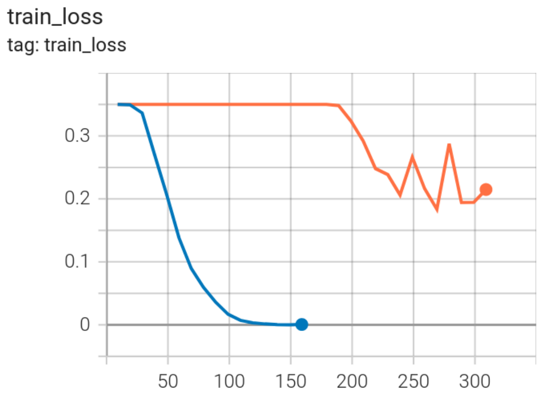


Figure 1: Overfitting on a 1-batch sample. Orange: Wav2Letter; Blue: Wav2Vec. Window 0.

B Cross Modal Contrastive Learning

Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Sound- net: Learning sound representations from unlabeled video. In Adv. Neural Inform. Process. Syst., 2016.

Andrew Owens, Jiajun Wu, Josh McDermott, William Free- man, and Antonio Torralba. Ambient sound provides super- vision for visual learning. In Eur. Conf. Comput. Vis., 2016.

Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Von- drick, Josh McDermott, and Antonio Torralba. The sound of pixels. In Eur. Conf. Comput. Vis., September 2018.

C Modeling video and speech

Here we sketch some ideas for how to model videos with spoken narration or dialog, based on related work.

C.1 Separate encoders

This is the approach taken in Rouditchenko et al. (2020) where the video encoder and the audio encoder both separately embed their respective modality as a vector. The loss function is contrastive, based on max-margin softmax (Ilharco et al., 2019), where the softmax is applied to pairwise dot products between modality-specific vectors:

$$L(\mathbf{x}, \mathbf{y}) = -\operatorname*{mean}_i \left(\log \frac{\exp(\mathbf{x}_i \cdot \mathbf{y}_i - \delta)}{\exp(\mathbf{x}_i \cdot \mathbf{y}_i - \delta) + \sum_{j \neq i} \exp(\mathbf{x}_i \cdot \mathbf{y}_j)} \right) \tag{1}$$

where $\mathbf{x}, \mathbf{y}$ are modality-specific vectors and δ is the margin. The second term in the denominator sums over the negative examples.

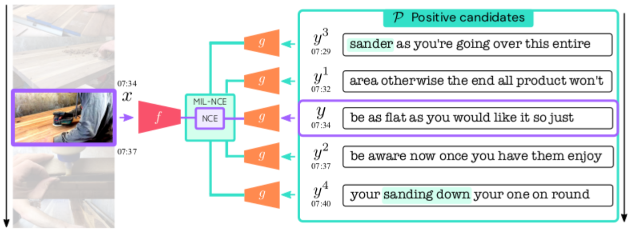
The advantage of this type of approach is its simplicity and the fact that there is a single vector representation for the whole speech fragment.

C.2 Multiple instance learning

In the approach introduced in Miech et al. (2020) the objective is basically eq. (1) without the margin, and multiple positive examples: instead of a single positive example as in eq. (1), a set of potentially positive examples is sampled from closely co-occuring video-text fragments (see fig. 2. This is supposed to address to some extent the issue that the content of the video and the corresponding speech are only weakly and noisily synchronized. There is a lot of verbiage about this in the paper, but it seems like a pretty trivial idea, and I'm not sure why it works better than simply using longer video-text segments (it does seem to work better).

C.3 Matchmap

The matchmap is a concept introduced in Harwath et al. (2018) and which specifies the affinity between each region of the two dimensional visual feature map and each frame in the associated audio, i.e. it is a three-dimensional tensor.


07:34
x
07:37
f
MIL-NCE
NCE
g
$\mathcal{P}$ Positive candidates
y^3 07:29
sander as you're going over this entire
y^1 07:32
area otherwise the end all product won't
y 07:34
be as flat as you would like it so just
y^2 07:37
be aware now once you have them enjoy
y^4 07:40
your sanding down your one on round

(a) **Examples of positive candidates**

Figure 2: Examples of positive examples from Miech et al. (2020).

The similarity between the image and the speech can be computed by a number of scoring functions which aggregate across these dimensions in different ways: the one which tended to work best was defined as:

$$\mathrm{MISA}(M) = \operatorname*{mean}_{t}\left(\max_{r,c}(M_{r,c,t})\right) \tag{2}$$

The extension of this idea to videos associated with speech would simply add an extra dimension to the matchmap, corresponding to time in the visual domain. This would make the matchmap a four-dimensional tensor, specifying the affinity between each region of the each frame of the video with a frame of the audio. Overall similarity between video and speech would involve a scoring function aggregating over the four dimensions: the equivalent to MISA would be:

$$\mathrm{MISA}_4(M) = \operatorname*{mean}_{t}\left(\max_{r,c,\tau}(M_{r,c,\tau,t})\right), \tag{3}$$

where t is time in the audio modality, and τ is time in the video modality. This function has the effect that for each frame of the audio, it finds the best matching frame/region (let us call this *fragment*) of the video, and averages over the scores of these matches. It does this without encouraging any sort of structure to this alignment, for example nothing prevents the same fragment of the video to be the best match for each frame of the audio.

This approach does, *not per*, se produce a single vector representing an utterance and thus if audio-audio retrieval or similarity metric is desired, something more complex than a simple cosine similarity would be needed:

- Mean or max pooling of features across frames in order to obtain a single vector for the utterance;
- Dynamic Time Warping if audio-audio similarity metric is needed.

C.4 Transformer-based models

C.4.1 Fast and Slow

Peng and Harwath (2021) have a model which combines separate encoders with dual cross-modal attention encoder. They work with images split into regions of interest, but we could do something very similar, but with video frames instead of ROIs.

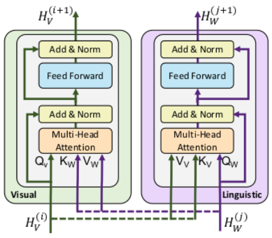


Figure 3: ViLBERT consists of two parallel streams for visual and linguistic data that interact through co-attention transformer layers (image from (Lu et al., 2019)).

C.4.2 Contrastive bi-modal transformer

The model proposed in Sun et al. (2019b) uses the BERT architectire directly to model video-text data. The video is converted to high-level visual tokens and concatenated with transcribed audio. A video-text datapoint is transformed to the the sequence

$$(\,\texttt{[CLS]}\,w_1\ldots w_n\,\texttt{[>]}\,v_1\ldots v_m\,\texttt{[SEP]}\,)$$

and used as input to the standard BERT architecture and trained adapting the BERT training objectives. The `[CLS]` token is used as a joint representation of the video-text datapoint. A similar approach could be used for video-audio data: tokens can correspond to video and/or audio frame features. For training, BERT-like objectives could be used; a interesting alternative would be to adapt this architecture to use with the contrastive loss. We would represent in input as

$$(\,\texttt{[AUD]}\,w_1\ldots w_n\,\texttt{[VID]}\,v_1\ldots v_m)$$

using `[AUD]` and `[VID]` as pooled representations of audio and video respectively, and use these with the loss from eq. (1) (or similar), identifying these representations with x_m and y_n, respectively in the formula.

C.4.3 VilBERT

The ViLBERT paper (Lu et al., 2019) uses a transformer with so-called co-attention to model data which consist of images (represented as a collection of regions) and text. There are not many details in the paper about the actual formulation of these co-attention heads, but there is a picture: see fig. 3. In principle this architecture could also be suitable for modeling video-audio data: instead if image regions we'd be using video frames.

C.4.4 CBT (contrastive bidirectional transformer)

Sun et al. (2019a) is rather convoluted: there are two BERT-like encoders, as well as a separate cross-modal transformer, with a contrastive training objective, which unlike appendix C.4.2 uses a simple joint representation for both modalities.

References

Aytar, Y., Vondrick, C., and Torralba, A. (2016). Soundnet: Learning sound representations from unlabeled video. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 892–900.

Harwath, D., Recasens, A., Surís, D., Chuang, G., Torralba, A., and Glass, J. (2018). Jointly discovering visual objects and spoken words from raw sensory input. In *Proceedings of the European conference on computer vision (ECCV)*, pages 649–665.

Ilharco, G., Zhang, Y., and Baldridge, J. (2019). Large-scale representation learning from visually grounded untranscribed speech. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Krishna, R., Hata, K., Ren, F., Fei-Fei, L., and Niebles, J. C. (2017). Dense-captioning events in videos. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 706–715. IEEE Computer Society.

Lu, J., Batra, D., Parikh, D., and Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.

Miech, A., Alayrac, J., Smaira, L., Laptev, I., Sivic, J., and Zisserman, A. (2020). End-to-end learning of visual representations from uncurated instructional videos. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9876–9886. IEEE.

Owens, A., Isola, P., McDermott, J. H., Torralba, A., Adelson, E. H., and Freeman, W. T. (2016a). Visually indicated sounds. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2405–2413. IEEE Computer Society.

Owens, A., Wu, J., McDermott, J. H., Freeman, W. T., and Torralba, A. (2016b). Ambient sound provides supervision for visual learning. In *European conference on computer vision*, pages 801–816. Springer.

Peng, P. and Harwath, D. (2021). Fast-slow transformer for visually grounding speech.

Rouditchenko, A., Boggust, A., Harwath, D., Joshi, D., Thomas, S., Audhkhasi, K., Feris, R., Kingsbury, B., Picheny, M., Torralba, A., et al. (2020). Avlnet: Learning audio-visual language representations from instructional videos.

Sun, C., Baradel, F., Murphy, K., and Schmid, C. (2019a). Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*.

Sun, C., Myers, A., Vondrick, C., Murphy, K., and Schmid, C. (2019b). Videobert: A joint model for video and language representation learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7463–7472. IEEE.

Zhou, L., Zhou, Y., Corso, J. J., Socher, R., and Xiong, C. (2018). End-to-end dense video captioning with masked transformer. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8739–8748. IEEE Computer Society.