Temporal-aware Cross-modal Embeddings for Video and Audio Retrieval

Amanda Duarte, Didac Surís, Amaia Salvador and Xavier Giró
Universitat Politecnica de Catalunya

The increasing amount of videos online brings several opportunities for training self-supervised neural networks. In this work, we explore cross-modal embeddings between audio and vision by exploiting their alignment on YouTube videos. Joint audio-visual embeddings allow creating links between audio and visual documents by projecting them to a common region of the feature space. They can be applied to enriching radio broadcasts with images, finding sound tracks for user-generated videos or simply enriching a topic search with both audio and video documents.

The idea of creating a joint embedding space across modalities has being exploited by other areas [3, 4]. However, joint representation between the video frames and its audio have yet to be fully exploited. A similar approach to the proposed one was [2], where a soundtrack was retrieved to match a music video. However, this work did not target a synchronization between both modalities.

We aim at training a temporal-aware embedding which can align both audio and visual tracks. Figure 1 presents the basic architecture of our deep neural network, which projects both image and audio features into a joint embedding space.

We use the visual and audio features provided in the YouTube-8M dataset [1]. The dataset includes features at both the clip and frame (temporal window) level. We train embeddings for both scales and assess their quality in a retrieval problem, formulated as using the feature extracted from one modality to retrieve the most similar videos based on the features computed in the other modality.

We aim at not only finding related documents, but synchronize both sequences. The alignment between the two sequences will rely on computing temporal-aware features with recurrent neural networks at different scales. At retrieval time, different scales will be assessed and results evaluated both with ranking metrics and Amazon Mechanical Turk.

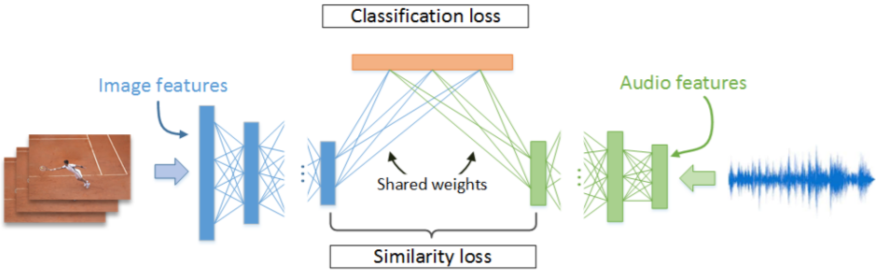

Figure 1: Architecture: Image and audio features are projected into a cross-modal same embedding.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[2] Sungeun Hong, Woobin Im, and Hyun S Yang. Deep learning for content-based, cross-modal retrieval of videos and music. *arXiv preprint arXiv:1704.06761*, 2017.

[3] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning cross-modal embeddings for cooking recipes and food images. In *CVPR*, 2017.

[4] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.