Cross-modal Embeddings for Video and Audio Retrieval

Didac Surís[1], Amanda Duarte[1,2], Amaia Salvador[1],
Jordi Torres[1,2], and Xavier Giró-i-Nieto[1]

[1] Universitat Politécnica de Catalunya - UPC, Spain
[2] Barcelona Supercomputing Center - BSC, Spain
{amanda.duarte, amaia.salvador, xavier.giro}@upc.edu

Abstract. In this work, we explore the multi-modal information provided by the Youtube-8M dataset by projecting the audio and visual features into a common feature space, to obtain joint audio-visual embeddings. These links are used to retrieve audio samples that fit well to a given silent video, and also to retrieve images that match a given query audio. The results in terms of Recall@K obtained over a subset of YouTube-8M videos show the potential of this unsupervised approach for cross-modal feature learning.

Keywords: Sonorization, embedding, retrieval, cross-modal, YouTube-8M

1 Introduction

Videos have become the next frontier in artificial intelligence. The rich semantics make them a challenging data type posing several challenges in both perceptual, reasoning or even computational level.

In addition to that, the popularization of deep neural networks among the computer vision and audio communities has defined a common framework boosting multi-modal research. Tasks like video sonorization, speaker impersonation or self-supervised feature learning have exploited the opportunities offered by artificial neurons to project images, text and audio in a feature space where bridges across modalities can be built.

This work exploits the relation between the visual and audio contents in a video clip to learn a joint embedding space with deep neural networks. We propose a joint audiovisual space to address a retrieval task formulating a query from any of the two modalities.

2 Architecture

The main objective of this research is to transform the two different features representation (image and audio, separately) into a *joint space*.

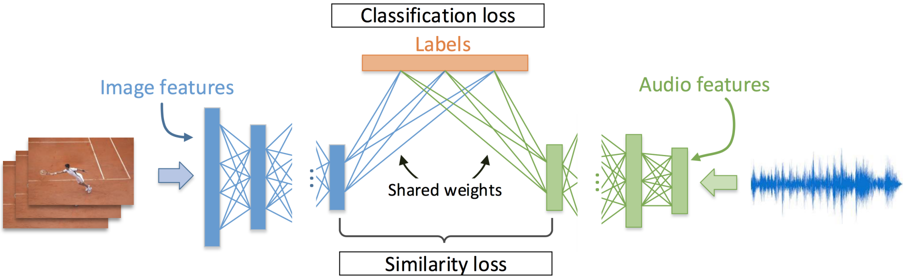


Fig. 1. Schematic of the used architecture.

Our model, depicted in the Figure 1, consists of two separated sets of different sizes of fully connected layers, one for visual features and a second for audio features. Both are trained to be mapped into the same cross-modal representation. We adopt a self-supervised approach, as we exploit the unsupervised correspondence between the audio and visual tracks in any video clip. In the end, a classification from the two embeddings using a sigmoid as activation function is performed, also using a fully connected layer.

Each hidden layer uses ReLu as activation function, and all the weights in each layer are regularized by the L2 norm.

3 Training

The objective of this work is to get the two embeddings of the same video (visual and audio) to be as close as possible (ideally, the same), while keeping embeddings from different videos as far as possible. The notion of ”similarity” or ”closeness” here is mathematically represented by the cosine similarity between the embeddings.

In addition to that, inspired by the work presented in [1], we provide additional information to our system by incorporating the video labels (classes) provided by the YouTube-8M dataset. This information is added as a regularization term that seeks to solve the high-level classification problem, both from the audio and the video embeddings, sharing the weights between the two branches. To that end, the loss function used for the classification is the well known cross entropy loss. This loss is optimized together with the cosine similarity loss, serving as a regularization term. In another words, the system learns to classify the audio and the images of a video (separately) into different classes or labels provided by the dataset. We limit its effect by using a regularization parameter λ.

The features used to train our model are already precomputed and provided by the YouTube-8M dataset [2]. In particular, we use the *video-level* features, which represent the whole video clip with two vectors: one for the audio and another one for the video.

3.1 Parameters and Implementation Details

For our experiments we used the following parameters:

- Batch size of 1024.
- We saw that starting with λ different than zero led to a bad embedding similarity because the classification accuracy was preferred. Thus, we began the training with $\lambda = 0$ and set it to 0.02 at step number 10,000.
- Margin $\alpha = 0.2$.
- Percentage of negative samples $p_{\text{negative}} = 0.6$.
- 4 hidden layers in each network branch, the number of neurons per layer being, from features to embedding, 2000, 2000, 700, 700 in the image branch, and 450, 450, 200, 200 in the audio branch.
- Dimensionality of the feature vector = 250.

4 Results

All the experiments presented in this section were developed over a subset of 6,000 video clips from the YouTube-8M dataset [2].

4.1 Quantitative Performance Evaluation

To obtain the quantitative results we use the Recall@K metric. We define Recall@K as the recall rate at top K for all the retrieval experiments, this is, the percentage of all the queries where the corresponding video is retrieved in the top K, hence higher is better.

The experiments are performed with different dimensions of the feature vector. The Table 1 shows the results of recall from audio to video, while the Table 2 shows the recall from video to audio.

Table 1. Evaluation of Recall from audio to video

k	Recall@1	Recall@5	Recall@10
256	21.5%	52.0%	63.1%
512	15.2%	39.5%	52.0%
1024	9.8%	30.4%	39.6%

Table 2. Evaluation of Recall from video to audio

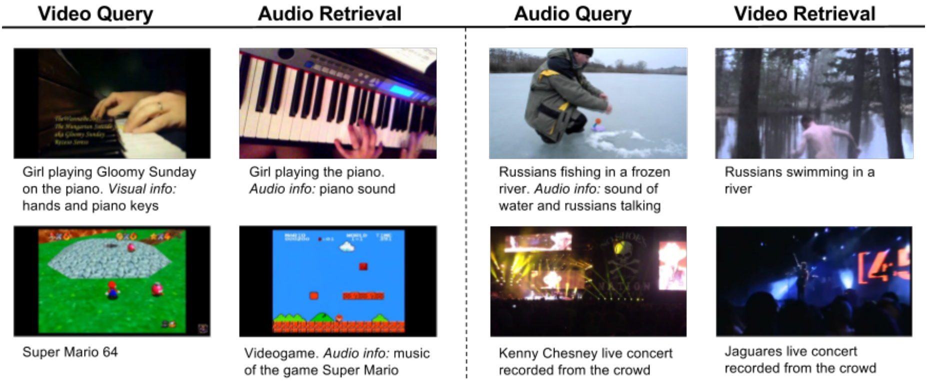
k	Recall@1	Recall@5	Recall@10
256	22.3%	51.7%	64.4%
512	14.7%	38.0%	51.5%
1024	10.2%	29.1%	40.3%

4.2 Qualitative Performance Evaluation

To obtain the qualitative results, a random video was chosen and from its image embedding, we retrieved the video with the closest audio embedding, and the other way around. In case the closest embedding retrieved corresponded to the same video, we took the second one in the ordered list.

On the left side of Figure 2 we can see the results given a video query; and on the right the input query is an audio. Examples depicting the real videos and audio are available online [3]. For each result and each query, we also show their YouTube-8M labels.

[3] https://goo.gl/NAcJah



Girl playing Gloomy Sunday on the piano. *Visual info:* hands and piano keys

Girl playing the piano. *Audio info:* piano sound

Russians fishing in a frozen river. *Audio info:* sound of water and russians talking

Russians swimming in a river

Super Mario 64

Videogame. *Audio info:* music of the game Super Mario

Kenny Chesney live concert recorded from the crowd

Jaguares live concert recorded from the crowd

Fig. 2. Qualitative results. On the left we show the results obtained when we gave a video as a query. On the right, the results are based on an audio as a query.

The results show that when starting from the image features of a video, the retrieved audio represents a very accurate fit for those images.

5 Conclusions

We presented an simple but effective method to retrieve audio samples that fit correctly to a given (muted) video. The qualitative results show that the already existing online videos, due to its variety, represent a very good source of audio for new videos, even in the case of only retrieving from a small subset of this large amount of data. Due to the existing difficulty of creating new audio from scratch, we believe that a retrieval approach is the path to follow in order to give audio to videos. The source code and trained model used in this paper is publicly available at https://github.com/surisdi/youtube-8m.

6 Acknowledgements

This work was partially supported by the Spanish Ministry of Economy and Competitivity and the European Regional Development Fund (ERDF) under contract TEC2016-75976-R. Amanda Duarte was funded by the mobility grant of the Severo Ochoa Program at Barcelona Supercomputing Center (BSC-CNS).

References

1. Salvador, A., Hynes, N., Aytar, Y., Marin, J., Ofli, F., Weber, I., Torralba, A.: Learning cross-modal embeddings for cooking recipes and food images. In: CVPR. (2017)
2. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark. CoRR **abs/1609.08675** (2016)