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[1,2]

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

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: Cross-modal · Retrieval · 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.

Videos are used in this work for two main reasons. Firstly, they naturally integrate both visual and audio data, providing a weak labeling of one modality with respect to the other. Secondly, the high volume of both visual and audio data allows training machine learning algorithms whose models are governed by a high amount of parameters. The huge scale video archives available online and the increasing number of video cameras that constantly monitor our world, offer more data than computation power available to process them.

Thus we exploit 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.

L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11132, pp. 711–716, 2019.
https://doi.org/10.1007/978-3-030-11018-5_62

2 Related Works

As online music streaming and video sharing websites have become increasingly popular, some research has been done on the relationship between music and album covers [4,5,11,12] and also on music and videos (instead of just images) as the visual modality [2,7,15,17] to explore the multimodal information present in both types of data.

A recent study [10] also explored the cross-modal relations between the two modalities but using images with people talking and speech. It is done through Canonical Correlation Analysis (CCA) and cross-modal factor analysis. Also applying CCA, [18] uses visual and sound features and common subspace features for aiding clustering in image-audio datasets. In a work presented by [13], the key idea was to use greedy layer-wise training with Restricted Boltzmann Machines (RBMs) between vision and sound.

The present work is focused on using the information present in each modality to create a joint embedding space to perform cross-modal retrieval. This idea has been exploited especially using text and image joint embeddings [9,14,16], but also between other kinds of data, for example creating a visual-semantic embedding [6] or using synchronous data to learn discriminative representations shared across vision, sound and text [3].

However, joint representations between the images (frames) of a video and its audio have yet to be fully exploited, being [8] the work that most has explored this option up to the knowledge of the authors. In their paper, they seek for a joint embedding space but only using music videos to obtain the closest and farthest video given a query video, only based on either image or audio.

The main idea of the current work is borrowed from [14], which is the baseline to understand our approach. There, the authors create a joint embedding space for recipes and their images. They can then use it to retrieve recipes from any food image, looking to the recipe that has the closest embedding.

3 Architecture

This research aims to transform two different features representation (image and audio, separately) into a *joint space*.

Our model, depicted in the Fig. 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.

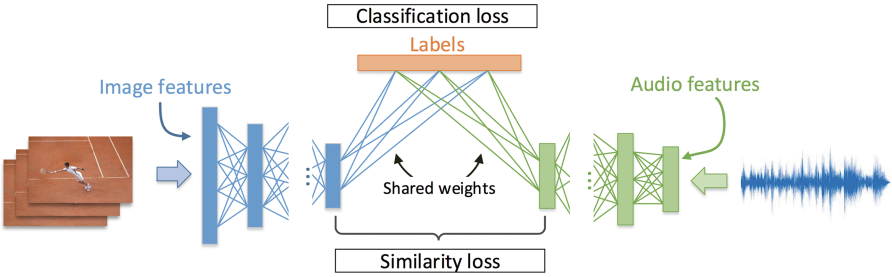

Fig. 1. Schematic of the used architecture.

4 Training

The objective here 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 [14], 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. The key idea here is to have the classification weights from the embeddings to the labels shared between the two modalities. 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 pre-computed and provided by the YouTube-8M dataset [1]. 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. These feature representations are the result of an average pooling of the local audio features computed over windows of one second, and local visual features computed over frames sampled at 1 Hz.

4.1 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.

5 Results

All the experiments presented in this section were developed over a subset of 6,000 video clips from the YouTube-8M dataset [1].

5.1 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.

To have a reference, the random guess result would be k/Number of elements, represented in the first column of each table. The obtained results show a very clear correspondence between the embeddings coming from the audio features and the ones coming from the video features. It is also interesting to notice that the results from audio to video and from video to audio are very similar, because the system has been trained bidirectionally.

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%

5.2 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 Fig. 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

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.

audio are available online[1]. For each result and each query, we also show their YouTube-8M labels.

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

6 Conclusions and Future Work

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.

As future work we would be to make use of the temporal information provided by the individual image and audio features of the YouTube-8M dataset to match audio and images, making use of the implicit synchronization that both modalities have, without needing any supervised control. Thus, the next step in our research is introducing a recurrent neural network, which will allow us to create more accurate representations of the video, and also retrieve different audio samples for each image, creating a fully synchronized system.

The source code and trained model used in this paper is publicly available at https://github.com/surisdi/youtube-8m.

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).

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

References

1. Abu-El-Haija, S., et al.: YouTube-8M: a large-scale video classification Benchmark. CoRR abs/1609.08675 (2016). http://arxiv.org/abs/1609.08675
2. Acar, E., Hopfgartner, F., Albayrak, S.: Understanding affective content of music videos through learned representations. In: Gurrin, C., Hopfgartner, F., Hurst, W., Johansen, H., Lee, H., O'Connor, N. (eds.) MMM 2014. LNCS, vol. 8325, pp. 303–314. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-04114-8_26
3. Aytar, Y., Vondrick, C., Torralba, A.: See, hear, and read: deep aligned representations. arXiv preprint arXiv:1706.00932 (2017)
4. Brochu, E., De Freitas, N., Bao, K.: The sound of an album cover: probabilistic multimedia and information retrieval. In: Artificial Intelligence and Statistics (AISTATS) (2003)
5. Chao, J., Wang, H., Zhou, W., Zhang, W., Yu, Y.: TuneSensor: a semantic-driven music recommendation service for digital photo albums. In: 10th International Semantic Web Conference (2011)
6. Frome, A., et al.: DeViSE: a deep visual-semantic embedding model. In: Neural Information Processing Systems (2013)
7. Gillet, O., Essid, S., Richard, G.: On the correlation of automatic audio and visual segmentations of music videos. IEEE Trans. Circuits Syst. Video Technol. **17**(3), 347–355 (2007)
8. Hong, S., Im, W., Yang, H.S.: Deep learning for content-based, cross-modal retrieval of videos and music. CoRR abs/1704.06761 (2017)
9. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. CoRR abs/1411.2539 (2014)
10. Li, D., Dimitrova, N., Li, M., Sethi, I.K.: Multimedia content processing through cross-modal association. In: Proceedings of the Eleventh ACM International Conference on Multimedia, pp. 604–611. ACM (2003)
11. Libeks, J., Turnbull, D.: You can judge an artist by an album cover: using images for music annotation. IEEE MultiMedia **18**(4), 30–37 (2011)
12. Mayer, R.: Analysing the similarity of album art with self-organising maps. In: Laaksonen, J., Honkela, T. (eds.) WSOM 2011. LNCS, vol. 6731, pp. 357–366. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21566-7_36
13. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the 28th International Conference on Machine Learning, pp. 689–696 (2011)
14. Salvador, A., et al.: Learning cross-modal embeddings for cooking recipes and food images. In: CVPR (2017)
15. Schindler, A., Rauber, A.: An audio-visual approach to music genre classification through affective color features. In: Hanbury, A., Kazai, G., Rauber, A., Fuhr, N. (eds.) ECIR 2015. LNCS, vol. 9022, pp. 61–67. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16354-3_8
16. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. CoRR abs/1511.06078 (2015). http://arxiv.org/abs/1511.06078
17. Wu, X., Qiao, Y., Wang, X., Tang, X.: Bridging music and image via cross-modal ranking analysis. IEEE Trans. Multimedia **18**(7), 1305–1318 (2016)
18. Zhang, H., Zhuang, Y., Wu, F.: Cross-modal correlation learning for clustering on image-audio dataset. In: 15th ACM International Conference on Multimedia, pp. 273–276. ACM (2007)