CORRESPONDENCE BETWEEN AUDIO AND VISUAL DEEP MODELS FOR MUSICAL INSTRUMENT DETECTION IN VIDEO RECORDINGS

Olga Slizovskaia, Emilia Gómez, Gloria Haro
Universitat Pompeu Fabra, Barcelona, Spain
{olga.slizovskaia, emilia.gomez, gloria.haro}@upf.edu

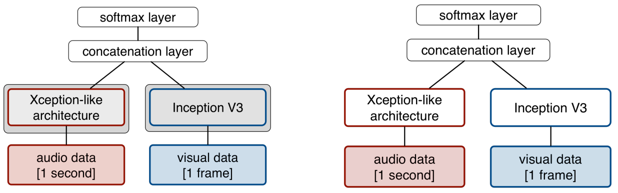
ABSTRACT

This work aims at investigating cross-modal connections between audio and video sources in the task of musical instrument recognition. We also address in this work the understanding of the representations learned by convolutional neural networks (CNNs) and we study feature correspondence between audio and visual components of a multimodal CNN architecture. For each instrument category, we select the most activated neurons and investigate existing cross-correlations between neurons from the audio and video CNN which activate the same instrument category. We analyse two training schemes for multimodal applications and perform a comparative analysis and visualisation of model predictions.

1. INTRODUCTION

The majority of real-world data analysis tasks rely on manifold data. That could be user's history and content features for music recommendation [5], brain activity signals and medical images for diagnosis [4] or a variety of meteorological data. Nevertheless, the most common multimodal information is audio-visual data, as it uses our senses and we experience it every day. We question the task of detecting musical instruments in video recordings and comparing the features learned from a model trained on visual information and a model trained on audio signals.

Previous research has demonstrated that audio-visual analysis can significantly improve the quality of music videos categorisation [3, 7]. However, apart from several papers on audio-visual correspondence [1, 2], there is a lack of detailed studies on data interference for this type of content. The literature mostly refers to the intuition of complementarity of audio and video sources. One approach to study the correspondence of the sources is to ask humans directly which information they may find linked [1]. Another approach is data-driven, such as the work in [2], where the authors train a multimodal CNN in an unsupervised manner to learn joint audio-visual feature representation on common videos. In this study, they proved that the learned representation is an efficient feature set for both modalities.



(a) Feature extraction scheme (b) Joint training scheme

Figure 1: Multimodal CNN architecture and training schemes. Gray boxes refer to pretrained feature extraction subnetworks whose parameters are freezed during the multimodal training.

From previous literature, we can pose the following research question: how do such multimodal features correspond to each other and what is the contribution of each modality to the representation? If we train a multimodal network, do we have an interpretable representation? In this work, we analyse two types of features: audio convolutional deep features and video convolutional deep features, with the goal of answering to this question.

2. METHODOLOGY

We experiment with a subset of FCVID dataset [3] which includes 5154 videos of musical performances with instruments almost equally distributed over 12 categories such as accordion, violin, saxophone or chamber music.

We use a similar CNN architecture as in our previous work [7] schematically presented in Figure 1. For audio analysis we utilise Xception-like architecture operating on the audio signal of 1 second length represented as log-mel spectrograms. We use Inception V3 architecture for video frames analysis and process 1 frame for every audio segment. We examine two training schemes for the multimodal network. As show in Figure 1, in the first case, we train audio and video subnetworks separately and, having them as feature extractors, train the integrated multimodal representation. In the second case, we consider joint training, such as weights in both modalities are adjusted simultaneously.

Then, having a pretrained multimodal network, we compute the features from audio and video subnetworks.

Attribution: Olga Slizovskaia, Emilia Gómez, Gloria Haro. "Correspondence between audio and visual deep models for musical instrument detection in video recordings", 18th International Society for Music Information Retrieval Conference, Suzhou, China, 2017.

For each video recording in the dataset, we extract an audio feature vector of length 1024 and a corresponding visual feature vector of length 2048 as the penultimate layer representations of the feature extraction subnetworks by construction.

2.1 Feature correlation study

Then we compute the importance coefficient of each feature with respect to each specific data category, defined as follows. Let D be a dataset, F a neural network algorithm, $X = (x_1, ..., x_K)$ a feature vector from the last layer of the network, $Y = (y_1, ..., y_C)$ a prediction vector, and $y_c = \sum_{i=1}^{K} w_i^c x_i$ the last layer decision rule, where w_i^c is a weight of the feature x_i for the category y_c.

We compute the importance a_i^c of the feature x_i for the category c as $a_i^c = \frac{\partial y_c}{\partial x_i}$. If there is only one dense/softmax layer beyond the convolutional layers, $a_i^c = w_i^c$. We compute category-wise feature importance as

$$L_{D,i}^c = \frac{1}{N_c} \sum_d \max(w_i^c x_i, 0),$$

where $d \in D$ and $F(d) = c$.

Next, we filter out the least important features for each category and compute pairwise Spearman correlation coefficient for the remaining ones (we take top-10 and top-20 features for audio and video respectively).

2.2 Multimodal Visualisation

We then incorporate the visualisation procedure proposed in [6] called *Gradient-weighted Class Activation Mapping* (Grad-CAM). This visualisation method highlights local regions which influence most the network decision.

3. RESULTS

Figure 2 provides the example correlation matrix for independent training scheme. This matrix shows the cross-correlations between audio and visual features for *SaxophonePerformance* category. Large in absolute values indicate a correlation in the feature pair while zero values mean that two features are independent. We observe that even though audio and visual features have strong intra-modal correspondence, there are no significant inter-modal relations.

As we observe a similar pattern for all other categories, we can conclude that in case of independent representation learning for audio and video, we learn substantive concepts in each modality.

4. CONCLUSION

The cross-modal relations analysis in multi-modal applications could be a perspective research direction in many fields, including MIR. It's especially relevant in connection with increasing impact of neural networks which have a shortage of methods for better understanding and interpretability for the models.

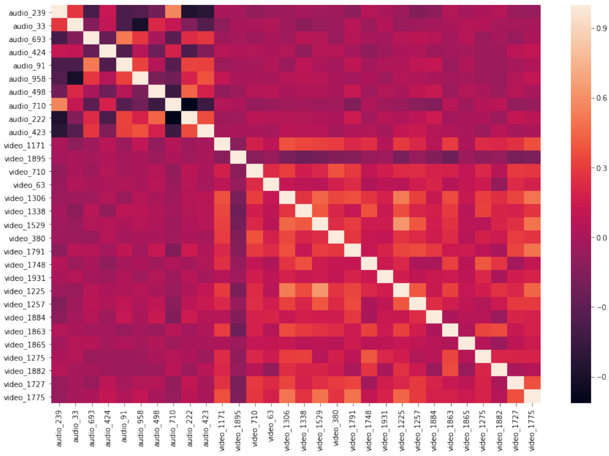


Figure 2: Cross-correlations of the most activated audio and visual features for *SaxophonePerformance* category for the feature extraction training scheme.

5. ACKNOWLEDGEMENTS

This work is partly supported by the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X GPU and WiMIR society for covering the registration expenses.

6. REFERENCES

[1] M. Adeli, J. Rouat, and S. Molotchnikoff. Audiovisual correspondence between musical timbre and visual shapes. *Frontiers in Human Neuroscience*, 2014.

[2] R. Arandjelović and A. Zisserman. Look, listen and learn. In *IEEE International Conference on Computer Vision*, 2017.

[3] Y.-G. Jiang, Z Wu, J. Wang, X. Xue, and S.-F. Chang. Exploiting feature and class relationships in video categorization with regularized deep neural networks. *arXiv preprint arXiv:1502.07209*, 2015.

[4] D. Lahat, T. Adali, and C. Jutten. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proc. of the IEEE*, 103(9):1449–1477, 2015.

[5] S. Oramas, O. Nieto, M. Sordo, and X. Serra. A deep multimodal approach for cold-start music recommendation. In *2nd Workshop on Deep Learning for Recommender Systems, at RecSys 2017*, Como, Italy, 2017.

[6] R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *arXiv preprint arXiv:1610.02391*, 2016.

[7] O. Slizovskaia, E. Gómez, and G. Haro. Musical instrument recognition in user-generated videos using a multimodal convolutional neural network architecture. In *Proc. of the 2017 ACM on International Conference on Multimedia Retrieval*.