MUSICAL INSTRUMENT RECOGNITION IN *USER-GENERATED VIDEOS* USING A MULTIMODAL CONVOLUTIONAL NEURAL NETWORK ARCHITECTURE

Olga Slizovskaia, Music Technology Group & Image Processing Group, Universitat Pompeu Fabra

Motivation

Musical instrument recognition is a well-known task in music informational retrieval. However, the majority of the existing techniques have been evaluated on good-quality and relatively small-size data and have never been tested on large-scale datasets. On the other hand, modern image and video recognition approaches are able to achieve high performance even for moderate-quality data. We aim to apply multimodal convolutional neural network architecture to take advantages from both audio and visual sources in the context of user-generated videos.

Datasets

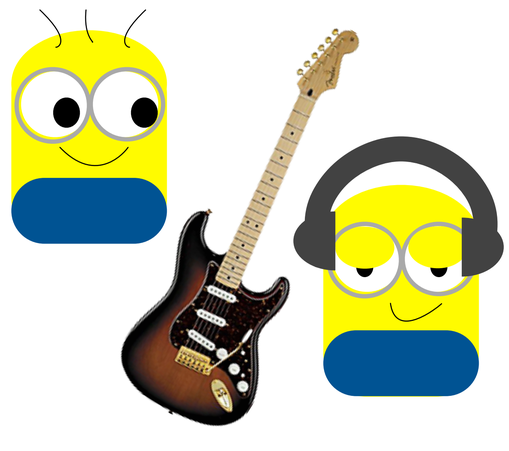
I. FCVID [4] / Music/ Musical Performance With Instruments
II. YouTube8M [5] Dataset / Music / Musical Instruments
✓ at least 1000 videos per class in original dataset
✓ single-label videos
✓ undersampling for top-3 instruments

Property	FCVID	YouTube-8M
Total number of categories	12	13 (46)
Total number of videos	5,154	60,862 (235,260)
Total video duration	259.84 hr	4,152.09 hr
Mean video duration	3.03 min	4.09 min
Videos per category (mean/std)	429 / 101	4,677 / 6,445
Videos used in experiments	5,247	60,802

Architectures

Video

For detecting instruments in static video frames, we experiment with Inception v3 architecture since it's one of the most prominent and successful architectures and it has been showed to provide a notable generalisation ability in various tasks. We explore the inuence of the total number of frames selected from the videos at the training phase. Moreover, we have studied the impact of fine-tuning the model over an independent set of images of musical instruments.



Audio

For audio feature representation learning, we have chosen the model from [2] (Han et al.) as a baseline. We also experiment with a modied model from [1] (Choi et al.) with a final classification somax layer. Both architectures follow the idea of stacking convolutional layers, but the latter has a larger receptive field and exploits more advanced activation function and batch normalization. In addition, we explore a recent Xception [3] architecture for audio-based instrument recognition. We modify the input layer so that the receptive field would be the same as in [1], and employ rectangular filters of size 48×3 at the first layer for better capturing the timbral characteristics of musical instruments. To reflect the changes in the input layer, we set the number of filters for separable convolutions equal to 768.

Multimodality

We individually train audio and video representation models and then exploit learned features from the last layers of the networks to train and evaluate the joint model as shown in Figure 1. Since the specic parameters for the audio and visual networks change for each experiment, we comment on the architecture of the late fusion model. The input layer of the model takes a concatenated feature vector of size (k + 1, n), where k is the number of video frames (plus one vector of the audio features), and n corresponds to the penultimate layer size in the audio and visual networks. The model consists of two fully-connected layers (each layer contains 1024 neurons and ReLU activation function) preceding the batch normalization, and a somax prediction layer.

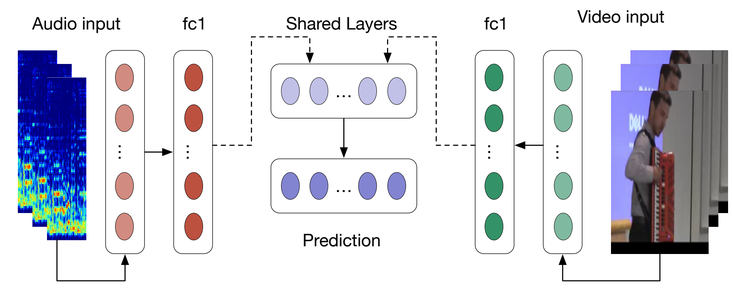
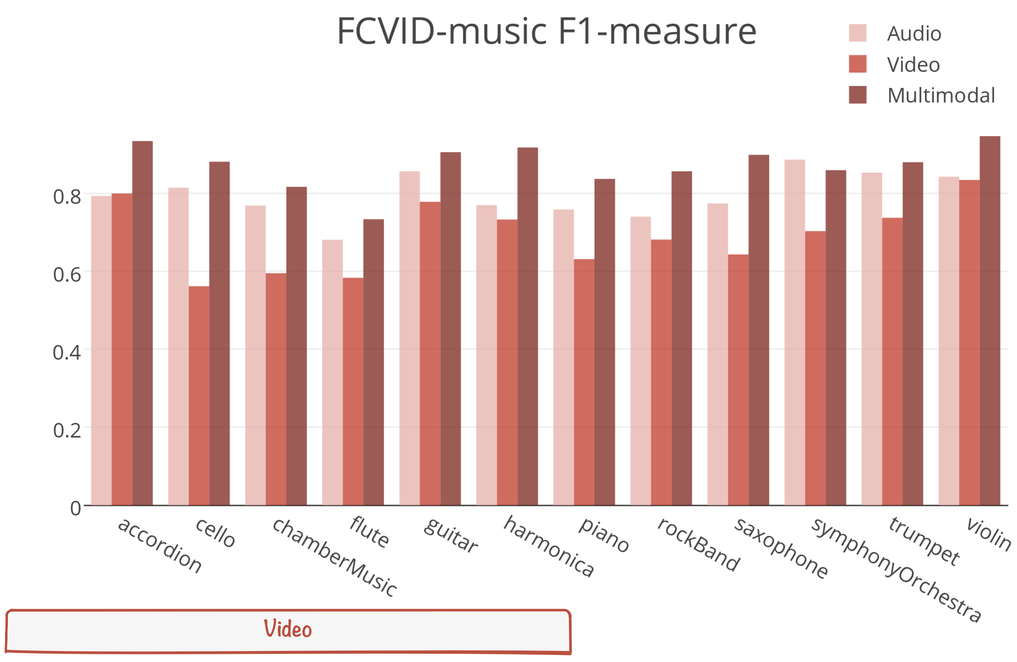
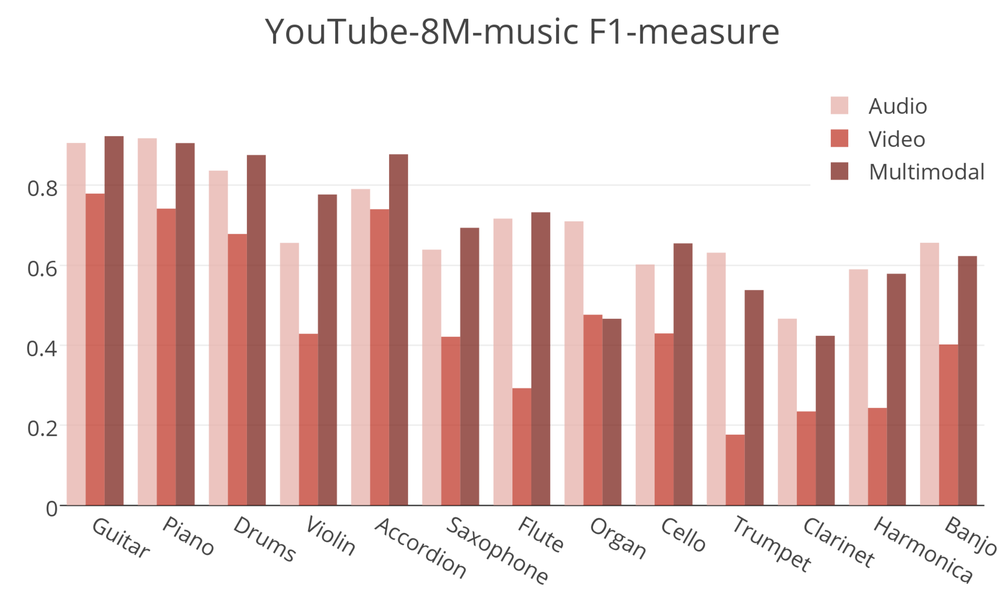


Figure 1: Schematic representation of our multimodal CNN architecture for musical instrument recognition.

Results





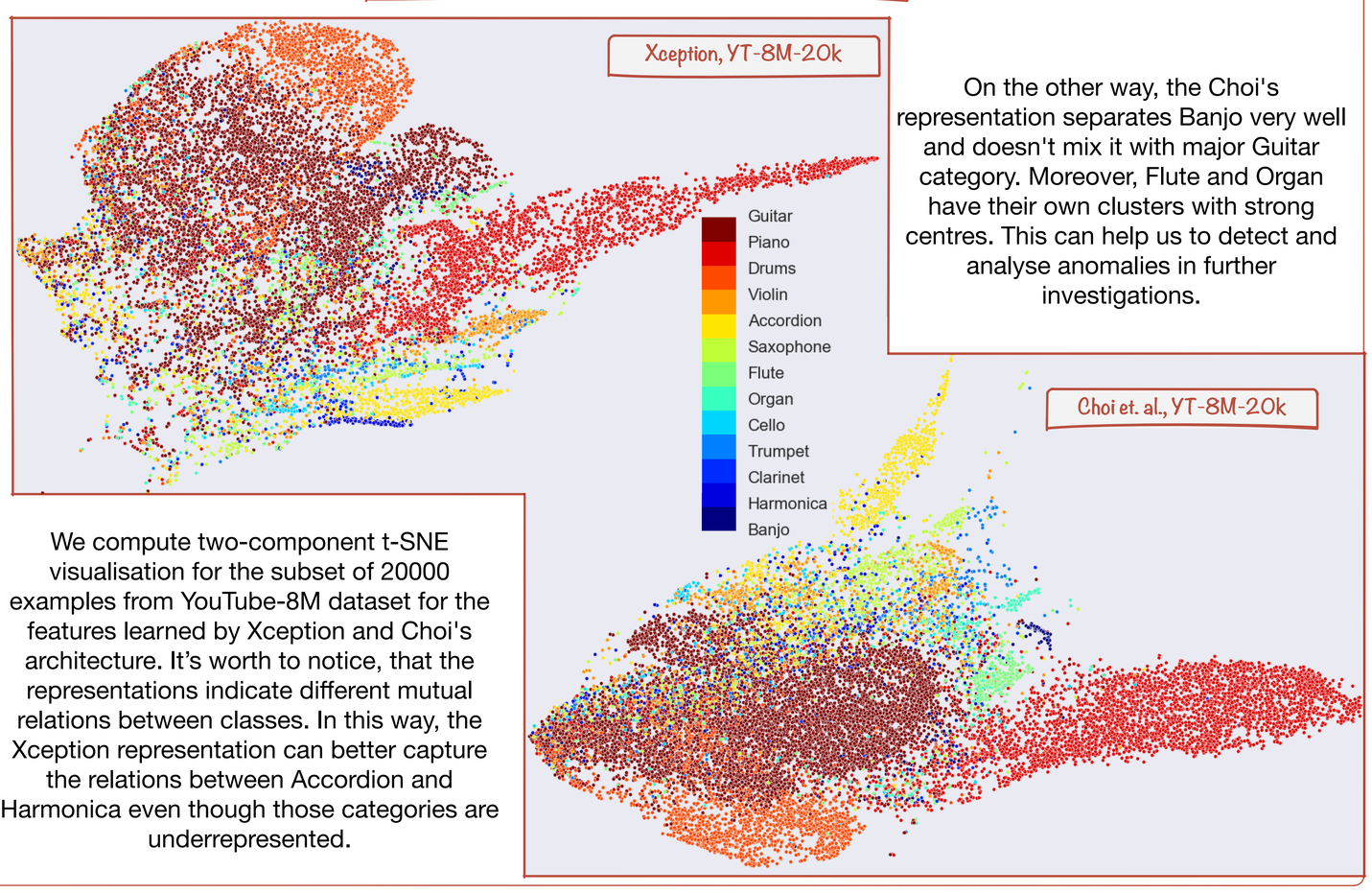
Video

Dataset	FMs	PT	Steps	Time	Hit@1	Hit@3	F1
FCVID	20	No	32K	19h	42.30	64.53	43.16
FCVID	30	No	16K	11h	65.39	81.75	67.29
FCVID	30	Yes	16K	11h	68.77	84.26	70.33
FCVID	50	No	24K	22h	67.47	83.21	69.38
FCVID	50	Yes	21K	19h	**69.39**	**84.32**	**71.23**
FCVID	100	No	43K	98h	68.56	83.97	70.42
FCVID	100	Yes	36K	84h	67.76	83.50	69.16
YT-8M	10	No	58K	82h	61.15	78.45	52.19
YT-8M	20	Yes	57K	92h	70.07	84.20	71.09

Audio

Method		#Params	Dataset	Hit@1	Hit@3	F1
Han et al.		1.5M	FCVID	64.13	76.82	53.64
Choi et al.	+ CC	2.4M	FCVID	77.73	92.05	77.18
Choi et al.	+ USC	2.4M	FCVID	**79.81**	**96.09**	78.71
Xception	+ USC	9.6M	FCVID	78.69	94.44	**79.35**
Han et al.		1.5M	YT-8M	59.37	70.87	56.50
Choi et al.	+ USC	2.4M	YT-8M	**83.58**	94.23	**84.26**
Xception	+ USC	9.6M	YT-8M	83.53	**94.69**	84.16

Feature learning representation study



On the other way, the Choi's representation separates Banjo very well and doesn't mix it with major Guitar category. Moreover, Flute and Organ have their own clusters with strong centres. This can help us to detect and analyse anomalies in further investigations.

We compute two-component t-SNE visualisation for the subset of 20000 examples from YouTube-8M dataset for the features learned by Xception and Choi's architecture. It's worth to notice, that the representations indicate different mutual relations between classes. In this way, the Xception representation can better capture the relations between Accordion and Harmonica even though those categories are underrepresented.

References

[1] Choi, Keunwoo , et al. "Automatic Tagging using Deep Convolutional Neural Networks", 17th International Society for Music Information Retrieval Conference, New York, USA, 2016
[2] Han, Yoonchang, Jaehun Kim, and Kyogu Lee. "Deep convolutional neural networks for predominant instrument recognition in polyphonic music." *arXiv preprint arXiv:1605.09507* (2016).
[3] Chollet, François. "Xception: Deep Learning with Depthwise Separable Convolutions." *arXiv preprint arXiv:1610.02357* (2016).
[4] Jiang, Yu-Gang, et al. "Exploiting feature and class relationships in video categorization with regularized deep neural networks." *arXiv preprint arXiv:1502.07209* (2015).
[5] Abu-El-Haija, Sami, et al. "YouTube-8M: A Large-Scale Video Classification Benchmark." *arXiv preprint arXiv:1609.08675* (2016).
[6] Yue-Hei Ng, Joe, et al. "Beyond short snippets: Deep networks for video classification." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.
[7] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. "Learning Spatiotemporal Features with 3D Convolutional Networks." ICCV 2015.

Acknowledgments

This research is supported by the Spanish Government as a part of Maria de Maeztu Strategic Research Program (MDM-2015-0502). We also thank NVIDIA for hardware donation in the context of NVIDIA Academic Partnership Program.