Web Interface for Exploration of Latent and Tag Spaces in Music Auto-Tagging

Philip Tovstogan [1] **Xavier Serra** [1] **Dmitry Bogdanov** [1]

Abstract

We present an online graphical web interface for the exploration and evaluation of embedding and tag spaces of music auto-tagging systems. It allows for quick and qualitative evaluation of individual and pairwise tag predictions as well as visualization of tag and embedding latent spaces (original and with dimensionality reduction). We provide taggrams and embedding vectors for the MTG-Jamendo dataset from multiple state-of-the-art auto-tagging models that can be explored and compared.

1. Introduction

Music auto-tagging is a common task in music information retrieval. Some of the applications are genre classification and emotion recognition, while there are also works that focus on learning representative embeddings that can be used for transfer learning (Nam et al., 2019). In cases when the ground truth is available objective metrics such as ROC-AUC and PR-AUC give a good assessment of model performance. However, there always remains the question of how well the model generalizes towards other datasets and how it performs in real life.

In large auto-tagging datasets (Bogdanov et al., 2019; Bertin-Mahieux et al., 2011) labels are usually weak and noisy, which makes a qualitative evaluation on the test set slow and costly. Visual interfaces can help during the evaluation allowing people to see if the segment of the audio track with predicted label actually contain this tag. We are unaware of existing tools for evaluation on the segment level, although there are visual systems for exploration and discovery of music (Pampalk, 2001).

In this paper, we introduce an interface that visualizes latent spaces in 2D. We focus on both tag and embedding spaces. It can be useful for researchers who want to do a quick or in-depth qualitative analysis of the auto-tagging models. It is also possible to inspect individual dimensions of the embedding space and get useful insights into the semantics of what the model is learning. The system can also be used by regular users for the exploration of various music collections and/or discovering new music.

2. System Overview

The system is implemented as a Flask web application and is available online,[1] the code is open-source and available on GitHub.[2] We provide documentation on how the code can be used to extract embeddings and taggrams for any private collection and visualize it on a local machine. For demonstration purposes, we use the **MTG-Jamendo Dataset** (Bogdanov et al., 2019) because of the availability of full audio tracks under Creative Commons licenses and because Jamendo API could be used to serve audio.

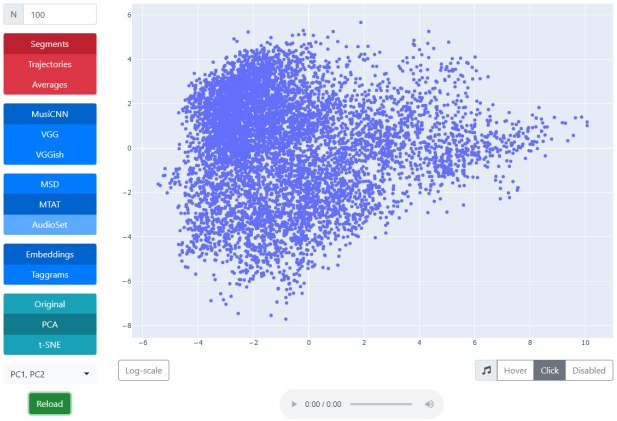


Figure 1: Screenshot of the interface

The interface allows the user to select the number of tracks that are visualized on-screen at the same time. Depending on the technical capabilities of the system that is used to interact with the interface it can vary: 100 tracks can be handled with no problem on most systems (MacBook Air 2017), while more powerful ones could handle 300 – 500.

Each model cuts an input track into multiple **segments** (the length of which varies depending on the model). There are three modes to visualize tracks:

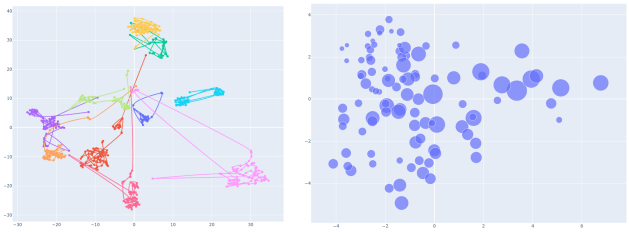
*Equal contribution [1]Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain. Correspondence to: Philip Tovstogan <philip.tovstogan@upf.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 108, 2020.

[1] https://music-explore.upf.edu
[2] https://github.com/MTG/music-explore

- **Segments** – each segment is visualized as a separate point on the graph (Fig. 1).
- **Trajectories** – each track is visualized as a line that connects its consecutive segments (Fig. 2a). It allows for the visualization of separate tracks, but works well only for a small number of tracks (10 or less).
- **Averages** – The track is represented as an arithmetic mean of the values of its segments, visualized as a circle with the diameter proportional to the standard deviation (Fig. 2b).

By hovering or clicking on the point you can listen to that particular segment or track. It is also possible to zoom in and out to take a closer look at clusters of points or areas on the graph that might be of particular interest.



(a) Trajectories (b) Averages

Figure 2: Viewing modes

We use Essentia library with the following pretrained TensorFlow models (Bogdanov et al., 2013; Alonso-Jiménez et al., 2020) to extract tag activation values and embeddings:

- **MusiCNN** – musically-motivated CNN (Pons & Serra, 2019). It uses vertical and horizontal convolutional filters. The model contains 6 layers and 787,000 trainable parameters. Embeddings are extracted from the penultimate layer of size 200.
- **VGG** – architecture from computer vision (Simonyan & Zisserman, 2015) based on a deep stack of 3×3 convolutional filters adapted for audio (Choi et al., 2016). The model contains 5 layers and 605,000 trainable parameters. Embeddings are taken from penultimate layer of size 2×128 and flattened to 256.
- **VGGish** – original implementation of computer vision architecture (Simonyan & Zisserman, 2015) with the number of output units is set to 3087 (Hershey et al., 2017). The number of trainable parameters is 62 million. Embeddings are taken from dedicated embeddings layer of size 128.

The provided MusiCNN and VGG models have been pretrained on top 50 tags of two datasets: **Million Song Dataset (MSD)** (Bertin-Mahieux et al., 2011) and **MagnaTagATune (MTAT)** (Law et al., 2009). The interface makes it easy to switch between training datasets. The VGGish model is pre-trained on **AudioSet** (Gemmeke et al., 2017) and only provides an embeddings layer. The length of segments is 3 sec for MusiCNN and VGG, and 0.96 sec for VGGish.

There are two layers: **taggrams** and **embeddings**. They are available for visualization of all combinations of models and training datasets, except for VGGish, which has only the embeddings layer. The layers can be visualized with the common dimensionality reduction techniques: **PCA** (Wold et al., 1987) (several components available) or **t-SNE** (Maaten & Hinton, 2008). However, we also provide a way to visualize **original** dimensions either by specifying the index of the embeddings layer or choosing two tags in the taggrams layer.

3. Applications

The ability to visualize individual tags in taggrams is very useful for quick qualitative evaluation of auto-tagging system. By listening to segments with high activation values you can immediately hear if the tag is representative (e.g. hearing *guitar* or comparing if it is *slow*er than other segments). Also, in the case of noisy labels it easy to see if the tags with the same semantic meaning have a high correlation (e.g. *vocal* and *vocals* in MTAT), or that semantically mutually exclusive tags have negative correlation (e.g. *vocal* and *no vocals* in MTAT).

The interface is also useful for exploring music. Either by selecting tags to look for new music on the intersection of genres/categories or using dimensionality reduction one can explore the whole latent space from different perspectives. Distance between the points is indicative of similarity relative to selected tags.

Dimensionality reduction also allows for the exploration of the semantics of the learned embedding space. Listening to segments while slowly moving along one of the axes can give insight into the semantics of the largest differences that are learned by auto-tagging systems.

Looking at trajectories with t-SNE, can give insights about the structure of the tracks and its temporal evolution. For example, transitions between voiced and instrumental parts of the track are quite evident (two cyan clusters on Fig. 2a).

4. Future Work

The explainability of deep learning systems is currently an important research topic, and our system can help to understand the semantics of learned latent spaces. However, the question on the subjectivity of perceived semantics remains, thus we want to conduct user experiments to see if there is an agreement in the semantics of principal components or dimensions of embeddings.

Although the system can be used for music exploration, it requires explicit control from the user. The next step would be to evaluate the feasibility of assisted navigation in the latent space as a means to explore and discover new music.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 765068.

References

Alonso-Jiménez, P., Bogdanov, D., Pons, J., and Serra, X. Tensorflow audio models in essentia. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pp. 266–270, 2020.

Bertin-Mahieux, T., Ellis, D. P. W., Whitman, B., and Lamere, P. The million song dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR*, pp. 591–596, 2011.

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J. R., and Serra, X. Essentia: An audio analysis library for music information retrieval. In *Proceedings of the 14th International Society for Music Information Retrieval Conference, ISMIR*, pp. 493–498, 2013.

Bogdanov, D., Won, M., Tovstogan, P., Porter, A., and Serra, X. The MTG-Jamendo dataset for automatic music tagging. In *Machine Learning for Music Discovery Workshop, ML4MD, International Conference on Machine Learning, ICML*, 2019.

Choi, K., Fazekas, G., and Sandler, M. B. Automatic tagging using deep convolutional neural networks. In *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR*, pp. 805–811, 2016.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pp. 776–780, 2017.

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., and Wilson, K. W. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pp. 131–135, 2017.

Law, E., West, K., Mandel, M. I., Bay, M., and Downie, J. S. Evaluation of algorithms using games: The case of music tagging. In *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR*, pp. 387–392, 2009.

Maaten, L. v. d. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov): 2579–2605, 2008.

Nam, J., Choi, K., Lee, J., Chou, S., and Yang, Y. Deep learning for audio-based music classification and tagging: Teaching computers to distinguish rock from bach. *IEEE Signal Processing Magazine*, 36(1):41–51, 2019.

Pampalk, E. Islands of music: Analysis, organization, and visualization of music archives. Master's thesis, Vienna University of Technology, 2001.

Pons, J. and Serra, X. musicnn: Pre-trained convolutional neural networks for music audio tagging, 2019.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR*, 2015.

Wold, S., Esbensen, K., and Geladi, P. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37–52, 1987. ISSN 0169-7439. Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists.