AUDIO ATLAS: VISUALIZING AND EXPLORING AUDIO DATASETS

Luca A. Lanzendörfer **Florian Grötschla** **Uzeyir Valizada** **Roger Wattenhofer**
ETH Zurich
{lanzendoerfer, fgroetschla, uvalizada, wattenhofer}@ethz.ch

Late-Breaking / Demo Session Extended Abstract, ISMIR 2024 Conference

ABSTRACT

We introduce Audio Atlas, an interactive web application for visualizing audio data using text-audio embeddings. Audio Atlas is designed to facilitate the exploration and analysis of audio datasets using a contrastive embedding model and a vector database for efficient data management and semantic search. The system maps audio embeddings into a two-dimensional space and leverages DeepScatter for dynamic visualization. Designed for extensibility, Audio Atlas allows easy integration of new datasets, enabling users to better understand their audio data and identify both patterns and outliers. We open-source the codebase of Audio Atlas, and provide an initial implementation containing various audio and music datasets. [1]

1. INTRODUCTION

The increasing size of machine learning datasets presents significant challenges in data visualization and analysis. Traditional tools are often insufficient for effectively managing and interpreting unlabeled audio datasets at scale. Having the ability to visualize large-scale datasets is crucial in helping to understand the structure and patterns within the data. It allows users to quickly grasp relationships between variables, identify trends, and detect outliers or anomalies that may affect the performance of a machine learning model.

Although there have been various projects focusing on providing high-level insight into audio and music datasets, they do not allow users to visualize their own data or were not concerned with large-scale datasets [1, 2, 3, 4, 5]. As such, these tools are mostly unsuitable for machine learning projects. To this end, we present Audio Atlas, an open-source interactive web application that helps users navigate audio and music datasets. Inspired by existing work in the image domain [6], Audio Atlas can visualize any audio dataset, providing a responsive user interface even when displaying tens of millions of samples. To obtain semantically meaningful embeddings, we use CLAP [7], a contrastive neural network trained on audio-text pairs. This enables Audio Atlas to display audio data with meaningful clusters and facilitates effective semantic searches and content exploration without requiring any audio metadata.

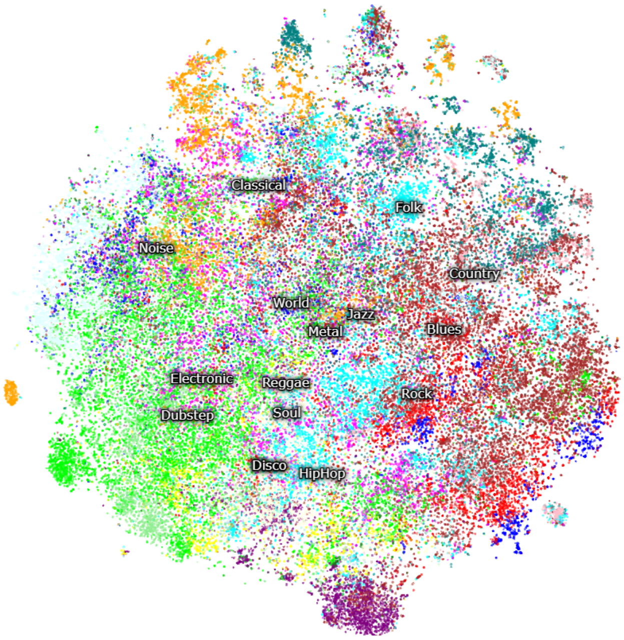


Figure 1: Audio Atlas view of the FMA dataset [8]. A cluster can be highlighted by clicking on the label.

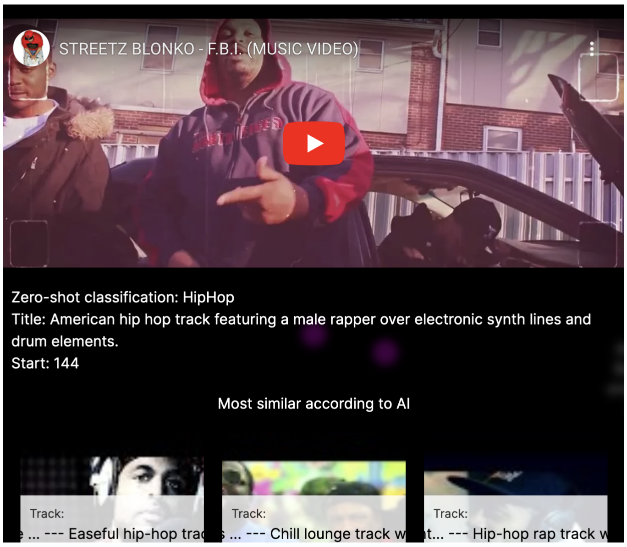
2. AUDIO ATLAS

Audio Atlas is a visualization tool designed to help users interact with audio data through an intuitive and dynamic web interface. The application leverages the Contrastive Language-Audio Pretraining (CLAP) [7] model to generate embeddings that fuse audio and text into a shared vector space. These embeddings are then projected onto a two-dimensional plane using t-SNE [9], and are visualized as a point cloud. We use Milvus [10] to store the CLAP embeddings. Milvus is a high-performance open-source vector database which efficiently manages embeddings and enables nearest-neighbor lookup for semantic searches with both text and audio snippets.

Audio Atlas enables users to perform zero-shot classification on their audio data. Zero-shot classification categorizes datasets without prior explicit training on specific classes, which is particularly useful when labeled data is

[1] https://github.com/ETH-DISCO/audio-atlas

Attribution: L. A. Lanzendörfer, F. Grötschla, U. Valizada, and R. Wattenhofer, “Audio Atlas: Visualizing and Exploring Audio Datasets”, in *Extended Abstracts for the Late-Breaking Demo Session of the 25th Int. Society for Music Information Retrieval Conf.,* San Francisco, United States, 2024.


STREETZ BLONKO - F.B.I. (MUSIC VIDEO)

Zero-shot classification: HipHop
Title: American hip hop track featuring a male rapper over electronic synth lines and drum elements.
Start: 144

Most similar according to AI

Track:
... --- Easeful hip-hop tra

Track:
s ... --- Chill lounge track w

Track:
t... --- Hip-hop rap track w

Figure 2: View when clicking an audio sample or using the search bar. Datasets containing links to YouTube are embedded to allow playback. The grid below shows the most similar results according to the CLAP embedding space.

scarce or missing entirely. The CLAP embeddings are used for zero-shot classification with a user-definable list of classes. The visualization highlights the classes with different colors, and a click on the label highlights only the selected class. The frontend is powered by DeepScatter [11], built on top of WebGL [12] and React [13], and renders the visualizations interactively, allowing users to explore audio datasets by navigating through clusters, performing semantic searches, and listening to the audio snippets. A click on a datapoint opens a detailed view with more information provided in the dataset, such as classes, labels, and descriptions, as well as a list of closest neighbors in the embedding space sorted by similarity (cf. Fig. 2). Our initial implementation provides access to MusicCaps [14], YT8M-MTC [15], VCTK [16], ESC-50 [17], MTG-Jamendo [18], and FMA [8]. Additionally, Audio Atlas remains responsive on large-scale datasets [19].

Furthermore, Audio Atlas can be used for semantic search. Users can search for audio using text or audio, with the provided search bar in Audio Atlas. The modality provided in the search bar (text or audio) is converted into an embedding vector using a contrastive model. The nearest neighbors of this embedding vector are computed using Annoy, an approximate nearest neighbor search framework. This semantic search enables users to find and filter audio using descriptive text, which historically has only been possible if the audio data contained rich annotations. Using contrastive models, we can perform a semantic search on disjoint modalities. While using audio to search for audio has existed previously, using a contrastive approach allows us to search on the basis of semantic meaning instead of the similarity of extracted audio features or waveform similarity.

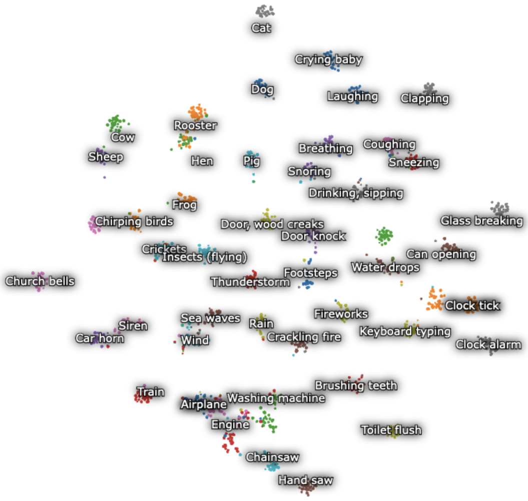


Figure 3: Audio Atlas visualization of ESC-50 dataset. The clusters of the various classes found in the Environmental Sound Classification dataset are clearly visible.

3. USE CASES

We demonstrate a series of practical applications of Audio Atlas, showing its utility in audio analysis through dimensionality reduction and search techniques.

Audio Atlas makes it easy to browse an audio dataset using text and audio queries. We can therefore also qualitatively assess the classification capabilities of the contrastive embedding model. Furthermore, we can explore the dataset with descriptive queries instead of searching for specific tracks using their titles.

Additionally, users can further explore datasets using the upload function with audio queries. By doing so, users are able to find the results that are most similar to their audio file in the dataset according to the similarity of the embedding space, as well as finding the most similar result's label and embedding location. This helps users understand the neighborhood of their query audio file.

In Figure 3, we use the ESC-50 classes to classify the dataset using zero-shot classification. In this way, we can visualize the semantic meanings of the clusters. Moreover, we can see that the t-SNE projection of the CLAP embeddings for the ESC-50 dataset has clustered all classes into local pockets. This clustering helps us explore specific datasets, in addition to understanding what the CLAP model has learned. We can easily apply Audio Atlas to other datasets with various labels to classify and explore the data. Furthermore, we believe Audio Atlas could be used as a novel way to browse music samples. For example, users could explore their unannotated audio libraries simply by describing the sound they are looking for.

In summary, by transforming audio into a visual representation, Audio Atlas gives users a new tool to explore large-scale audio datasets interactively. By making the codebase for Audio Atlas open-source we hope to advance the study of large-scale audio datasets as well as qualitatively assessing the performance of embedding models.

4. REFERENCES

[1] G. McDonald, “Every noise at once,” http://everynoise.com, 2023, accessed: 2024.

[2] Kwinten Crauwels, “Musicmap,” https://www.musicmap.info/, 2022, accessed: 2024.

[3] D. Bogdanov, A. Porter, H. Schreiber, J. Urbano, and S. Oramas, “The acousticbrainz genre dataset: Multi-source, multi-level, multi-label, and large-scale,” in *Proceedings of the 20th Conference of the International Society for Music Information Retrieval (ISMIR 2019): 2019 Nov 4-8; Delft, The Netherlands.[Canada]: ISMIR; 2019.* International Society for Music Information Retrieval (ISMIR), 2019.

[4] Paul Lamere, “The eternal jukebox,” https://eternalbox.dev/, 2012, accessed: 2024.

[5] D. Smilkov, N. Thorat, C. Nicholson, E. Reif, F. B. Viégas, and M. Wattenberg, “Embedding projector: Interactive visualization and interpretation of embeddings,” *arXiv preprint arXiv:1611.05469*, 2016.

[6] F. Grötschla, L. A. Lanzendörfer, M. Calzavara, and R. Wattenhofer, “Aeye: A visualization tool for image datasets,” 2024. [Online]. Available: https://arxiv.org/abs/2408.04072

[7] Y. Wu, K. Chen, T. Zhang, Y. Hui, M. Nezhurina, T. Berg-Kirkpatrick, and S. Dubnov, “Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation,” 2024.

[8] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, “Fma: A dataset for music analysis,” 2017. [Online]. Available: https://arxiv.org/abs/1612.01840

[9] L. van der Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

[10] J. Wang, X. Yi, R. Guo, H. Jin, P. Xu, S. Li, X. Wang, X. Guo, C. Li, X. Xu, K. Yu, Y. Yuan, Y. Zou, J. Long, Y. Cai, Z. Li, Z. Zhang, Y. Mo, J. Gu, R. Jiang, Y. Wei, and C. Xie, “Milvus: A purpose-built vector data management system,” in *Proceedings of the 2021 International Conference on Management of Data*, ser. SIGMOD ’21. New York, NY, USA: Association for Computing Machinery, 2021, p. 2614–2627. [Online]. Available: https://doi.org/10.1145/3448016.3457550

[11] Nomic AI, “Deepscatter,” https://github.com/nomic-ai/deepscatter, 2022, accessed: 2024.

[12] Khronos Group, “Webgl,” http://www.khronos.org/webgl/, 2014, accessed: 2024.

[13] Meta Open Source, “React: A javascript library for building user interfaces,” https://reactjs.org/, 2013, accessed: 2024.

[14] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, M. Sharifi, N. Zeghidour, and C. Frank, “Musiclm: Generating music from text,” 2023.

[15] D. McKee, J. Salamon, J. Sivic, and B. Russell, “Language-guided music recommendation for video via prompt analogies,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 784–14 793.

[16] C. Veaux, J. Yamagishi, and K. MacDonald, “Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit,” 2017.

[17] K. J. Piczak, “ESC: Dataset for Environmental Sound Classification,” in *Proceedings of the 23rd Annual ACM Conference on Multimedia.* ACM Press, pp. 1015–1018. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2733373.2806390

[18] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, “The mtg-jamendo dataset for automatic music tagging,” in *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)*, Long Beach, CA, United States, 2019. [Online]. Available: http://hdl.handle.net/10230/42015

[19] L. A. Lanzendörfer, F. Grötschla, E. Funke, and R. Wattenhofer, “Disco-10m: a large-scale music dataset,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.