Music Genre Classification using Transformers.

Yuval Hoffman and Roee Hadar

April 2024

Introduction

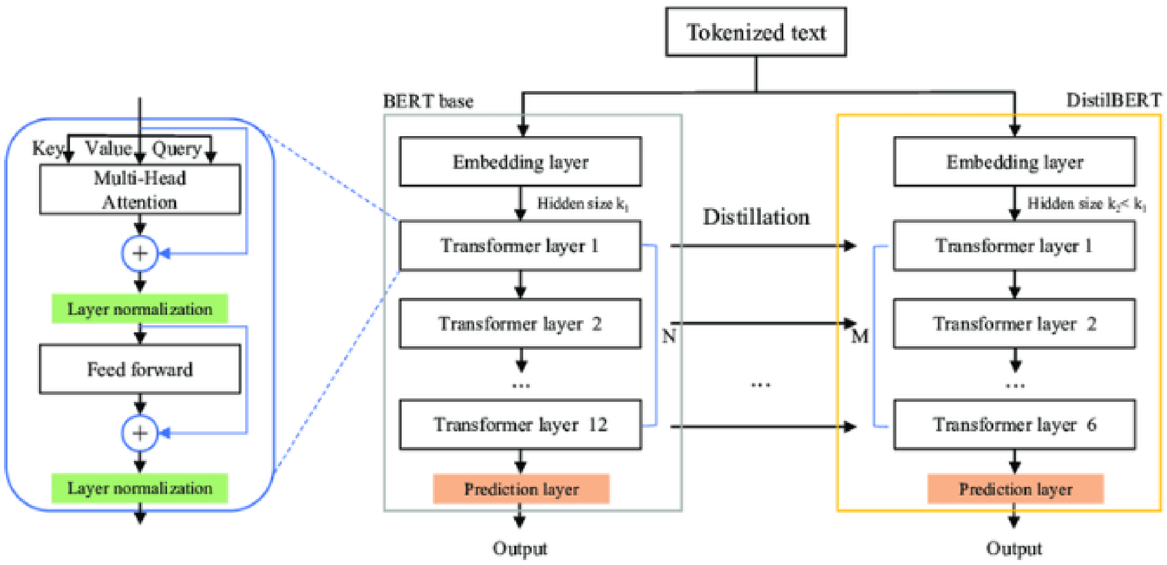
In today's digital era, the proliferation of audio content across various platforms has sparked a growing interest in developing robust methods for automatic audio classification. One prominent application of such techniques is in music genre classification, where algorithms are tasked with identifying the genre of a given audio sample. This task holds significant relevance in music recommendation systems, content organization, and even personalized user experiences.

The emergence of deep learning architectures and pre-trained models has revolutionized the field of audio classification, offering powerful tools for feature extraction, pattern recognition, and classification tasks. Leveraging these advancements, our project seeks to explore the efficacy of fine-tuning pre-trained models on a popular audio dataset, GTZAN. This dataset comprises a diverse collection of audio clips spanning ten distinct music genres, providing an ideal testbed for evaluating the performance of audio classification models.

Through this project, we aim to demonstrate the feasibility and effectiveness of fine-tuning pre-trained models for music genre classification tasks, combined with the usage of different Augmentations. By harnessing the rich features learned by these models from vast amounts of audio data, we endeavour to achieve precise classification accuracy and robustness, thereby contributing to the advancement of automated audio classification techniques.

The project will entail various stages, including data preprocessing, model fine-tuning, training, evaluation, and performance analysis. We will leverage state-of-the-art tools and libraries in machine learning, such as the Hugging Face Transformers library and the `evaluate` library, to streamline these processes and facilitate comprehensive analysis of the model's performance.

Ultimately, our endeavour aims to not only develop a highly accurate and reliable music genre classification model but also to provide insights and methodologies that can be generalized to other audio classification tasks. Through this project, we seek to contribute to the growing body of research in automated audio analysis and pave the way for enhanced audio-based applications and services in diverse domains.



Methodology

Our project delves into several steps, which is described as follows:

First, we lay our hands on the GTZAN dataset, a rich collection of audio clips spanning ten diverse music genres. This dataset serves as our foundation for training and testing our model.

Next, we delve into preprocessing the dataset. We need to refine those raw audio samples into a format our model can digest. To do this, we employ a sophisticated pre-trained feature extractor, which extracts crucial features from the audio data.

With our data primed and ready, we select a pre-trained model that suits our needs. We opt for `distilhubert`, a model known for its efficiency and solid performance in audio tasks. We may note here that, although the model chosen for our task performs well considering audio signals, the model itself has not been trained on songs, but rather on sentences.

Now comes the fine-tuning phase. Here, we tweak our chosen model to adapt its parameters to our music dataset. It's akin to tuning an instrument – we want our model to harmonize perfectly with the nuances of each music genre.

Once our model is honed and ready, it's time for testing. We evaluate its performance by feeding it a variety of songs and observing how accurately it can classify each genre. This stage helps us gauge the model's readiness for real-world applications.

But we're not content with mere accuracy numbers. We dive deeper into performance analysis, exploring metrics like confusion matrices to gain insights into where our model excels and where it struggles.

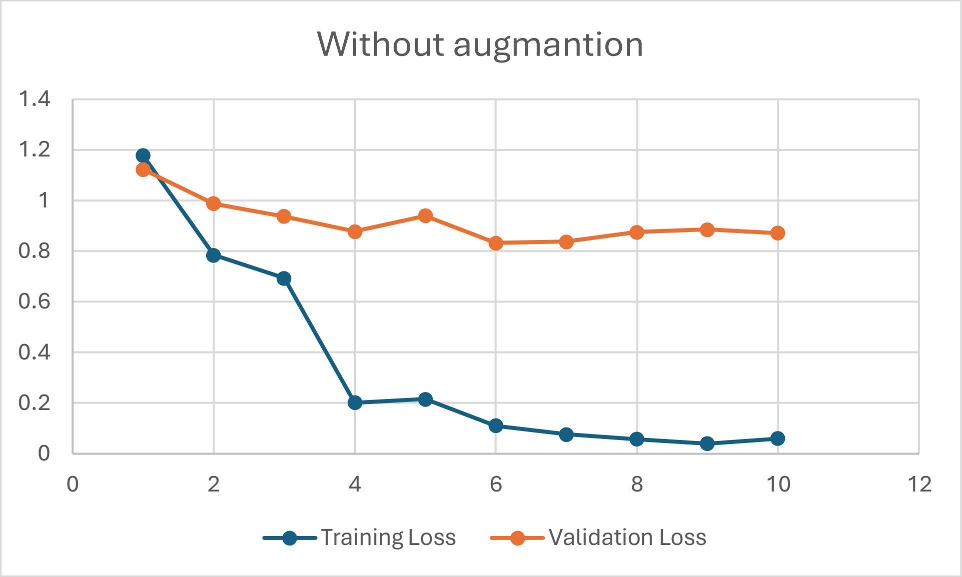
Finally, we inject a dose of creativity with data augmentation techniques. By introducing elements like white noise, pitch variations, and frequency tweaks, we aim to enhance our model's versatility and robustness in handling diverse audio samples.

Results

This following section unveils the results achieved in our project - including the results achieved by each Augmentation employed on the dataset:

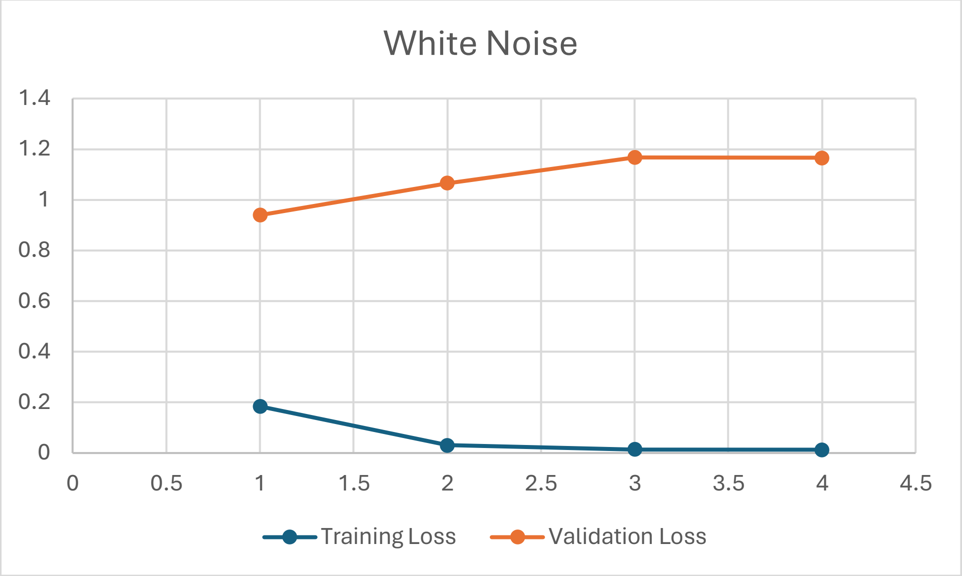
Regular Model Performance:

Our base model, fine-tuned on the GTZAN dataset, demonstrates commendable performance in classifying music genres. With an accuracy score of 78%, it showcases a solid understanding of the subtle differences between different genres. The results showcased in this section are a product of using a small section of the dataset due to GPU limitations, which have affected the success rate of our network. The results are added below:



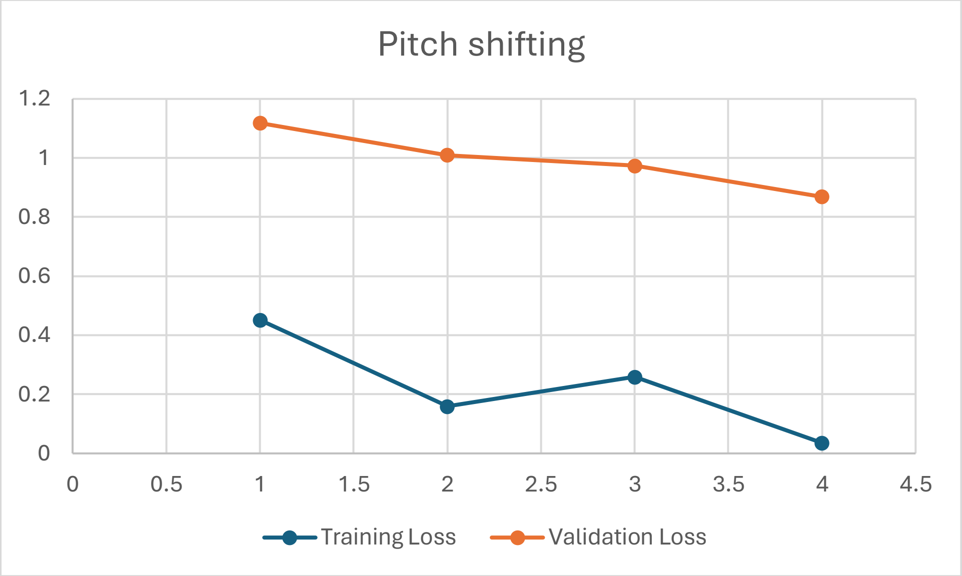
Model with White Noise Augmentation:

Augmenting our model with white noise introduces a layer of complexity, mimicking real-world scenarios where ambient noise can interfere with music recordings. Due to that, we may see that the model performs a bit more poorly, with accuracy rate of about 73%, which is still quite commendable considering the inference of a real-time scenario. The results, as described above, are added below:



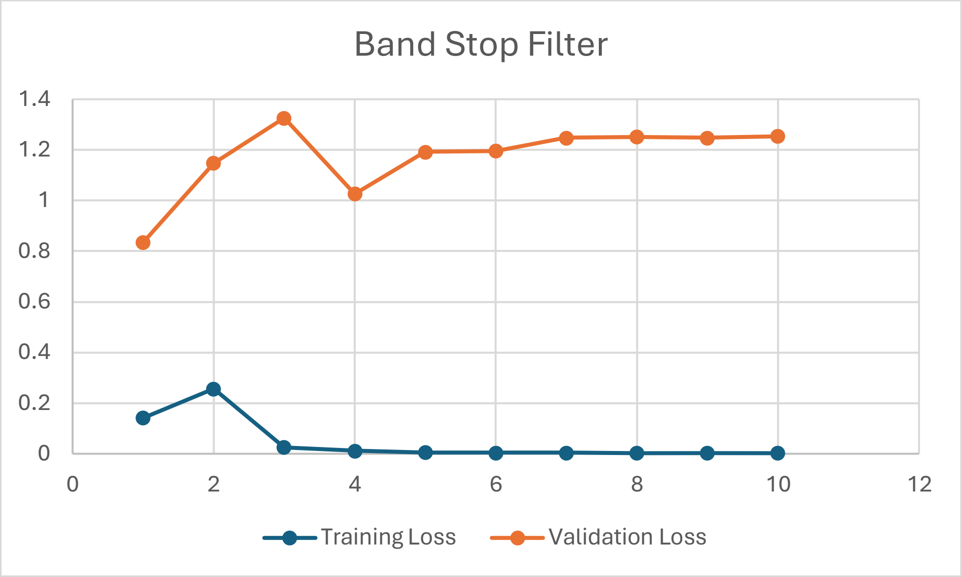
Model with Pitch Shifting Augmentation:

Incorporating pitch shifting into our model's training regimen adds a new dimension to its capabilities. By simulating variations in pitch, our model learns to adapt to different tonalities within music genres. This augmentation strategy yields an accuracy score of 75%, which is less than the results we've achieved without applying the augmentation.



Model with Bandpass Filtering Augmentation:

Implementing bandpass filtering augments our model's ability to discern intricate frequency patterns within audio samples. This augmentation technique enhances the model's sensitivity to specific frequency ranges, resulting in an accuracy score of 78%.



With these nuanced insights into our model's performance across various augmentation scenarios, we gain a deeper understanding of its capabilities and limitations. These results pave the way for future refinements and advancements in music genre classification.

The accuracy for all of the model is:

DATABASE:	GTZN	GTZN+Noise	GTZN+Pitch Shift	GTZN+BSF
Accuracy	78%	73%	75%	78%

Ethical Considerations

Identify Stakeholders:

Stakeholders involved in our project encompass a diverse array of entities, including:

- End-users: Individuals utilizing the music genre classification system for personal enjoyment or professional purposes.
- Music Industry: Record labels, artists, and music streaming platforms impacted by genre classifications.
- Researchers and Developers: Individuals contributing to the development and refinement of the classification model.
- Regulatory Bodies: Government agencies and policymakers responsible for overseeing the use of AI technology in audio classification.
- General Public: Society at large, which may indirectly benefit or be affected by the outcomes of our project.

Analyse Implications:

1. End-users:

- Positive Effects: Enhanced music discovery experience, improved playlist curation, and personalized recommendations.
- Negative Effects: Potential privacy concerns regarding data collection and usage, reliance on algorithmic recommendations leading to reduced diversity in music consumption.

2. Music Industry:

- Positive Effects: Better understanding of consumer preferences, targeted marketing opportunities, and improved content recommendation algorithms.
- Negative Effects: Potential commodification of music genres, reinforcement of existing genre stereotypes, and challenges to artist diversity and creativity.

3. Regulatory Bodies:

- Positive Effects: Opportunity to establish guidelines and regulations governing AI usage in audio classification, ensuring fair and ethical practices.
- Negative Effects: Difficulty in keeping pace with rapid technological advancements, potential for regulatory loopholes leading to misuse of AI systems.

Ethical Considerations:

In addressing the broader ethical implications of our project, we commit to the following goals:

- Privacy: Implementing robust data privacy measures to protect end-users' personal information and ensure consent-based data usage.
- Fairness: Mitigating bias in dataset collection, model training, and evaluation to prevent discriminatory outcomes in genre classification.
- Transparency: Providing clear explanations of how the classification model operates and the factors influencing its predictions, fostering user trust and understanding.
- Safety: Prioritizing the safety and well-being of end-users by ensuring the reliability and accuracy of the classification system and minimizing potential risks associated with misclassification.