ECE Senior Capstone Project 2023 Tech Notes

Apple Orchard

Machine Learning for Audio Applications

By Finn Tekverk, ECE '23

Introduction

Machine learning (ML) is a field within the larger concept of artificial intelligence. It is a type of computing that involves the creation of algorithms and 'models' which make a computer perform one task especially well. It typically involves the use of some statistical analyses in order to identify important trends in data.

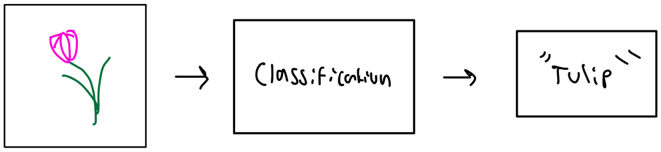
Machine learning is made possible through use of training data. This helps the model to categorize and make decisions for input data from a test case. The quality and quantity of data used for training are a large factor in the success of a model. For example, a model might be given photos of various types of flowers as training data. For training, we will also 'tell' the machine the type of each flower in the image. However, when the model is run for testing, we ask the model to predict the type of flower given only the image. The model will make decisions based on its dataset and predict what kind of flower the new image is.

ML for Audio Applications

Machine learning is used across many disciplines and types of data. For this project, the use of machine learning for audio was especially relevant. There are several applications within audio machine learning, such as sound classification, models which analyze human speech, and noise reduction.

Classification

Classification is an application of machine learning which tries to categorize input data. For example, given photos of animals, a model may classify the types of animals by identifying dogs or cats. In a project by Li and Qin, the classification of traditional Chinese instruments was performed. This project was able to determine if the model was successful by simply comparing the classification output. In other words, the model asked: did the model successfully identify which instrument was played? This creates a binary case for success and failure (successful identification vs. incorrect identification) (Li and Qin).



(Figure 1: Depiction of Classification Model)

Speech

There are a wide variety of machine learning projects concerning human speech. One such project paired visual information, which was given to the model alongside the audio sample. Using video of lip contours and an audio sample, the project evaluated the performance of the combined model. This project is interesting as an audio machine learning project, as it uses a multi-factor approach for data. By combining visual indicators, the model was able to make more informed decisions about the audio sample that it received (Debnath and Roy). This addition in capability often results in a far more complex model.

Noise and Speech

There are also projects that perform speech recognition using data with high levels of noise. A project by Duarte and Colcher asked a similar question, and the two authors created a noisy dataset to evaluate the performance of a few models. Their training set includes vocal audio that cuts in and out, as well as vocal audio with high levels of noise.

After evaluating the performance of the models on noisy data, Duarte and Colcher determined "an average result of 0.4116 in terms of character error rate in the noisy set", for an Signal to Noise Ratio (SNR) of 30 (Duarte, Colcher). This demonstrates the importance of valid and representative training data on the machine learning model produced.

While it would be ideal for a machine learning model to be 100% accurate in all environments, this is simply not realistic. Despite having a large SNR, this project still failed to reliably identify speech.

Ethics in Machine Learning Models

A major concern is the model's ability to recognize voices of different ethnic or racial groups, in order to avoid a model that is unable to assist members of these groups. While our group is ultimately at the mercy of available datasets, we must do our part in making sure to include vocal data from groups that are often underrepresented. This is to ensure that the technology works for as many people as possible, without favoring or exhibiting bias towards one group of individuals.

Conclusion

Machine learning has proven to be a powerful tool for audio applications, including sound classification, speech analysis, and operating within noisy audio environments. The quality and quantity of training data, as well as the type of model used is critical in the performance. Furthermore, using multiple factors such as visual indicators combined with audio data can produce more robust models.

References

1. Debnath, S., and P. Roy. *Audio-Visual Speech Recognition Based on Machine Learning Approach*. International Journal of Advanced Intelligence Paradigms, 2022.

2. Li, Rongfeng, and Zhang Qin. *Audio Recognition of Chinese Traditional Instruments Based on Machine Learning*. Cognitive Computation and Systems, June 2022.

3. Duarte, Julio Cesar, and Sérgio Colcher. *Building a Noisy Audio Dataset to Evaluate Machine Learning Approaches for Automatic Speech Recognition Systems*. ArXiv, 4 Oct. 2021.