Project Specification DT2119

Bashar Levin & David Marzban & Johannes Joujo & Nasir Alizade

April 2025

1 Introduction

Classify sound could mean many things, it could be distinguishing between noise and real spoken words, deciding which language is being spoken and what kind of sound it is such as a dog bark or a gunshot. Our group would like to be able to classify speech or audio into these categories.

Group number and member names

proj 4: Bashar Levin & David Marzban & Johannes Joujo & Nasir Alizade

Preliminary project title

Speech/audio classification using LLMs

Type of project (experiment or literature review)

Experimental.

Main problem statement (what is to be done)

Fine-tune GPT-2 to classify speech or audio recordings.

What data to use

The data that will be used to train our model comes from the provided link **different datasets**. The data used could change during the building phase therefore we have not picked a specific dataset now.

What kind of model(s)/algorithm(s) will you use

We are going to use GPT-2 as the model to be fine-tuned. The algorithm we will use are:

What Kind of Model(s)/Algorithm(s) Will You Use

We plan to use the following models and algorithms:

- **Pre-trained GPT-2 model**: Fine-tuned on tokenized versions of audio features.
- **Audio preprocessing**:
  - Convert audio to spectrograms or extract MFCCs (Mel-frequency cepstral coefficients).
  - Flatten or serialize these features into token-like sequences suitable for GPT-2 input.
  - Apply data augmentation techniques such as adding background noise, pitch shifting, and time stretching to improve model robustness and generalization.
- **Fine-tuning procedure**:
  - Apply supervised fine-tuning by adding a classification head (small feedforward network) on top of GPT-2's output embeddings.
  - Training with Cross-Entropy Loss for classification tasks.
- **CNN Encoder**:
  - Before feeding features into GPT-2, a small Convolutional Neural Network (CNN) will be used to automatically extract local time-frequency patterns from spectrogram, enhancing the quality of the tokenized inputs.
- **Baseline**:
  - Use non-fine-tuned GPT-2 with random initialization for classification as a performance baseline.

How will you evaluate the result

The result will be evaluated using the accuracy of the models ability to classify the inputted sound. The original GPT-2 without fine-tuning will be used as a baseline.

A few key references

In this article *Toward the Adoption of Explainable Pre-Trained Large Language Models for Classifying Human-Written and AI-Generated Sentences* [1] Petrillo et al. fine-tuned a model using BERT and tested it by giving it a total of 40 sentences, where half of them are AI-generated and the other half comes from

Wikipedia (written by humans). They wanted to classify if the input was generated by a human or AI.

In the article *Using GPT-2 to Create Synthetic Data toImprove the Prediction Performance of NLP Machine Learning Classification Models* [4], Whitfield shows how to fine-tune a GPT-2 model to generate synthetic pizza reviews.

In the article *Wav2Vec 2.0: A Framework for Self-Supervised Learning of Speech Representations* [2], Baevski et al. show how large transformer models can be adapted to learn from raw audio for downstream tasks such as speech recognition and classification. This article can help us with classification and how to evaluate.

In the article *Scanning dial: the instantaneous audio classification transformer* [3], Jiang et al. propose a modified Transformer model designed to classify extremely short audio clips, with durations too brief for human recognition. They introduce a novel filtering method to create instantaneous audio datasets, as well as a new data augmentation technique to improve performance. Their work shows that Transformers can effectively classify audio signals of only milliseconds in length, which provides valuable insights for our project where efficient and fast audio classification is important.

References

[1] Ahmed Alshehri, Reem Alanazi, Saeed Alqahtani, Ali Alzahrani, and Muhammad Binsawad. Toward the adoption of explainable pre-trained large language models for classifying human-written and ai-generated sentences. *Electronics*, 13(20):4057, 2024.

[2] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.

[3] Huawei Jiang, Husna Mutahira, Unsang Park, and Mannan Saeed Muhammad. Scanning dial: the instantaneous audio classification transformer. *Discover applied sciences*, 6(3):96–, 2024.

[4] Dewayne Whitfield. Using gpt-2 to create synthetic data to improve the prediction performance of nlp machine learning classification models, 2021.