Graph Neural Networks For Speech Recognition

Shuo Feng sf587, Jingyi Chen jc2498

10/3/2021

1 Motivation and Backgroud

In the last decade or so, the big messy data is well trained by deep learning in many different fields like natural language processing, computer vision, and speech recognition. The innovation speed of deep learning algorithms is very fast while no one network is considered perfect. Most generally, the standard neural network can only capture the dependency information by only regarding the features of nodes, and using Euclidean distance to detect their similarities. However, there are numerous application domains' data are generated from non-Euclidean domains which are represented as graphs with interrelationships and more complex dependencies between various entities. Hence, graph neural networks(GNNs) have emerged in machine learning and demonstrated a superior performance from the message passing between the nodes of graphs by representing information from their neighborhood with arbitrary depth.

2 Problem Statement

Music and audio content have been an increasingly important category of discriminative data, on which classification techniques are intensively explored to correctly address the music genre classification problem. As so far, there are various methods dealing with visual representation as spectrograms, representing spectrum frequencies of the audio signal, for music genre classification. Therefore, developing deep learning techniques and processes for successful music genre classification has a huge space for development. Since Netruel Network is a powerful and comprehensive algorithm in deep learning and Graph Neturel Network is especially efficient in solving image problems, our goal is to explore how GNN is applicable in successful music genre classification and how to improve this algorithm.

3 Dataset Discription

We plan to test our algorithm GTZAN (Tzanetakis and Cook (2002)) in this project. GTZAN consists of genre original, image original and csv file describing features of audio files. The reason of choosing this dataset is that GTZAN provides a clean distribution on a variety of audio sources and conditions.

4 Proposed Methods

We first compress audio data into vector representation of a MEL spectrogram. With prepared transferred image data, we plan to try Siamese neutral Network to extract features from the image since graph neutral network rely on full data to perform well but for complex and high dimensional relation audio data, it is hard for every audio data representation to have full features. Siamese Networks will help to use a new number of features of images to get better predictions. Besides directly applying, we aim to do some modifications on the Siamese process to extract features so that the expected result might be same accuracy level of prediction by less features or higher accuray prediction by more feasible features

Then, we will process modeling using Graph Neutral Network which has strong relational inductive bias to model complex irregular relationship structure among features. Compared with Convolution

Netural Network, Graph Neutral Network convolution solves process non-euclidean data better. We will try different modified aggregator functions F on operation of vertices. In the State of comparing similarity level, a small network is then used to fuse these representations and compute a similarity score. They can be thought of as learning both the representations and the similarity metric. According to similarity metric, generally, using cosine similarity,. In this step, we will conduct different similarity algorithms,