Compact Graph Architecture for Speech Emotion Recognition

Hamza Masood (210403)

Implementation Details

All speech signals are brought to the same size using padding and then are converted into a graph of $M = 120$ nodes. Each node represents an overlapping segment of $25ms$. The graph embedding dimension is $Q = 64$. A 5-fold cross validation is applied on the dataset to calculate the weighted and unweighted accuracy on both line and cycle graphs. Weighted accuracy accounts for the class imbalance in the dataset.

The network weights are initialized using Xavier initialization and Adam optimizer is used with a learning rate of 0.01. After every 50 epochs, the learning rate decays by a factor of 0.5 (to avoid overshooting local minima).

The `graphcnn.py` file contains the graph CNN model. The `GraphConv_Ortega` class defines the graph convolution layer. It constructs a normalized Laplacian matrix and performs eigen decomposition to obtain the graph's eigenvectors. These eigenvectors are used to aggregate node features. The `Graph_CNN_ortega` class uses the `GraphConv_Ortega` class to create multiple GCN layers. The output of the GCN layers is pooled using one of sum, mean, or max pooling to get the graph-level embeddings. These embeddings are passed into a classifier that maps them to labels.

The `main.py` script creates an instance of the `Graph_CNN_ortega` class and performs the model training and testing.

Results

Number of Trainable Parameters = 38916

Graph	**WA (%)**	**UA(%)**
Line	59.19	64.81
Cycle	58.53	64.72

Table 1: SER results on the IEMOCAP database in terms of weighted accuracy and unweighted accuracy.

Pooling	Maxpool	Meanpool	Sumpool
WA (%)	55.03	58.53	56.49
UA (%)	62.16	64.72	63.17

Table 2: Comparing different pooling strategies on the IEMOCAP database.

Dataset description

The IEMOCAP dataset has been used for all the experiments. It contains a total of 12 hours of data collected over 5 dyadic sessions with 10 subjects. To be consistent with previous studies, we used four emotion classes: anger, joy, neutral, and sadness. The final dataset contains a total of 4490 utterances - 1103 anger, 595 joy, 1708 neutral and 1084 sad.
The processed dataset can be found at the following link: IEMOCAP