UNIVERSITÀ
DI PAVIA

Speech recognition

D. Ligari 518592[1]

[1] *Machine Learning course, University of Pavia, Department of Computer Engineering (Data Science), Pavia, Italy*
Github page: https://github.com/DavideLigari01/speech-recognition.git
Contact: davide.ligari01@universitadipavia.it
Date: May 11, 2023

Abstract—This report presents a lab activity focused on speech recognition. The task is to recognize the pronunciation of a single word from a list of 35 words, using a multilayer perceptron. The dataset used for the task is the Speech Commands Data Set, which includes 105,829 recordings of the 35 words, divided into training, validation, and test sets. Feature extraction has already been performed, and the features are spectrograms that have been made uniform in size. The lab activity encompasses various components, including the visualization of spectrograms, the application of feature normalization techniques, training a multilayer perceptron without hidden layers, and exploring different network architectures. To gain insights into the network's performance, a confusion matrix is constructed to summarize its behavior, and classification errors are thoroughly analyzed. The experiments are replicated using different feature normalization techniques, batch sizes and lambda values, in order to understand how these parameters affect the model's performance.

Keywords— MLP Neural network • Training • Speech recognition • Feature normalization • Confusion matrix

1. MLP Neural network

The Multilayer Perceptron (MLP) neural network is a popular type of artificial neural network used in machine learning. It consists of interconnected layers of nodes or neurons that process data to produce predictions. MLPs employ activation functions, such as sigmoid, tanh, or ReLU, to introduce non-linearity and capture complex patterns. By adjusting the weights of these connections through a process called backpropagation, MLPs can learn from data and make accurate predictions. They are widely used in tasks like image recognition, speech processing, and natural language understanding.

2. Visualize the data

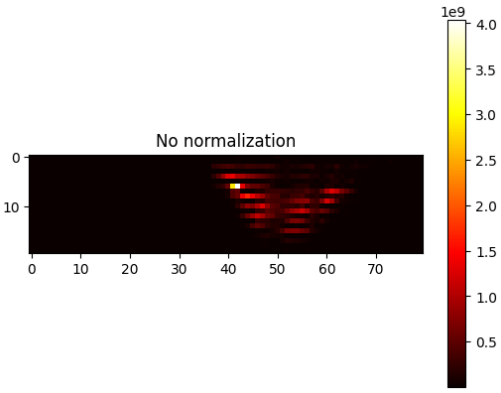

Fig. 1: Spectrogram of a sample of the dataset

3. Batch size selection

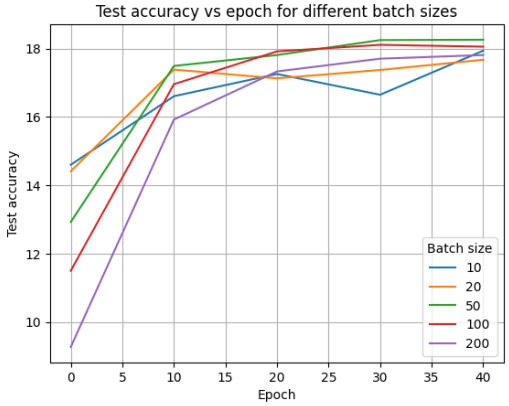

Fig. 2: Test accuracy for different batch sizes

4. Network architecture

a. Choice of the optimal lambda

b. Optimal network

Name server	ip
Google	8.8.4.4
Quad9	149.112.112.112
OpenDNS	208.67.220.220
Comodo Secure DNS	8.20.247.20

Table 1: Best neural network characteristics

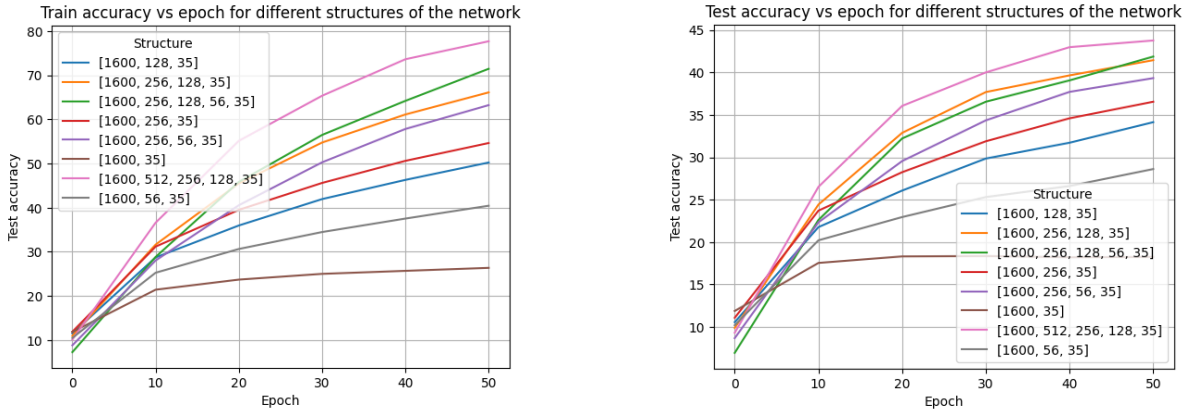

Fig. 3: Train (on left) and test accuracy (on right) for different network architectures

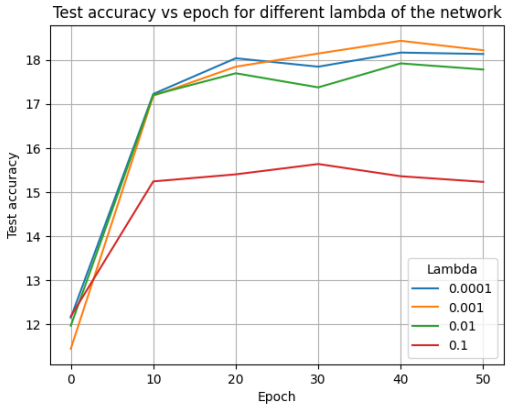

Fig. 4: Test accuracy for different lambda values

5. Model analysis

6. Feature normalization

7. Weights visualization

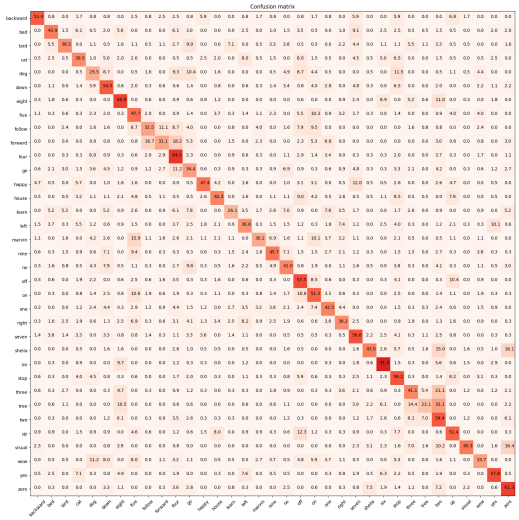

Fig. 5: Confusion matrix of the best model

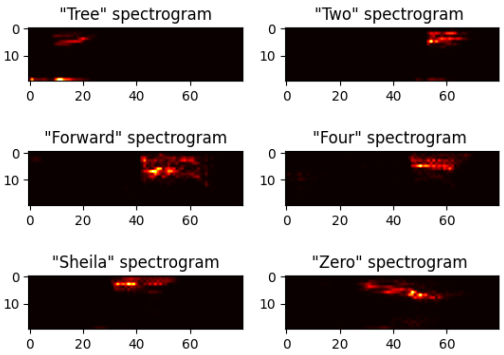

Fig. 6: Spectrogram of the most 3 misclassified words, on left the correct word, on right the confused one

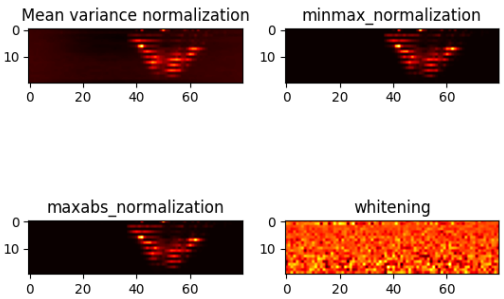

Fig. 7: Spectrogram of a sample of the dataset after different normalizations

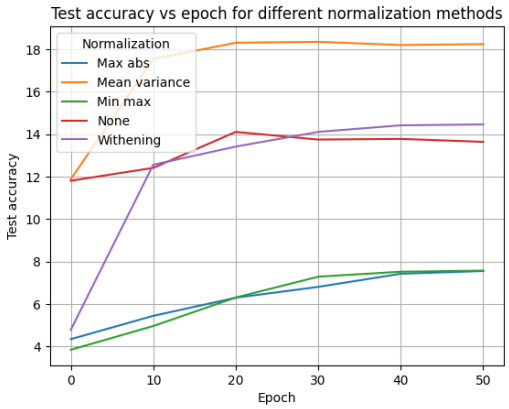

Fig. 8: Test accuracy for different normalizations

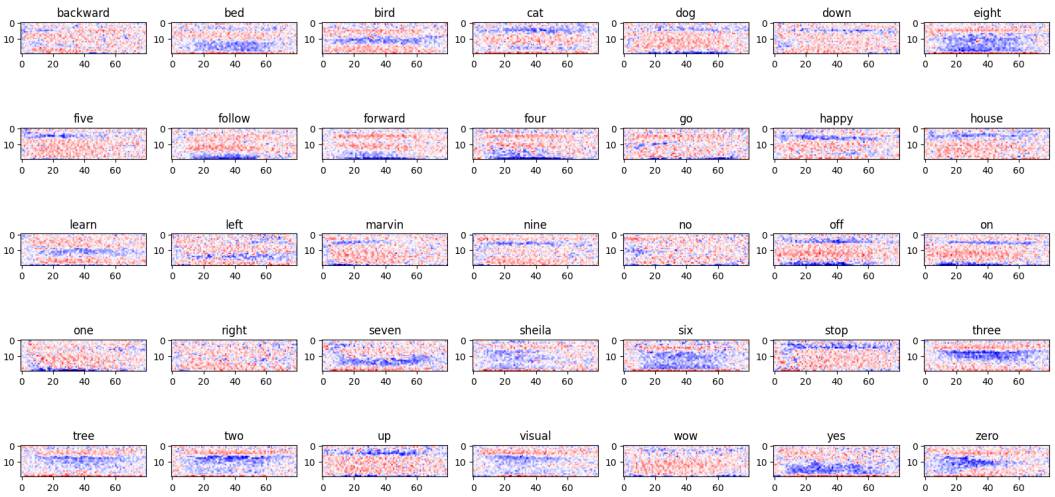

Fig. 9: spectrogram of the weights of the output layer