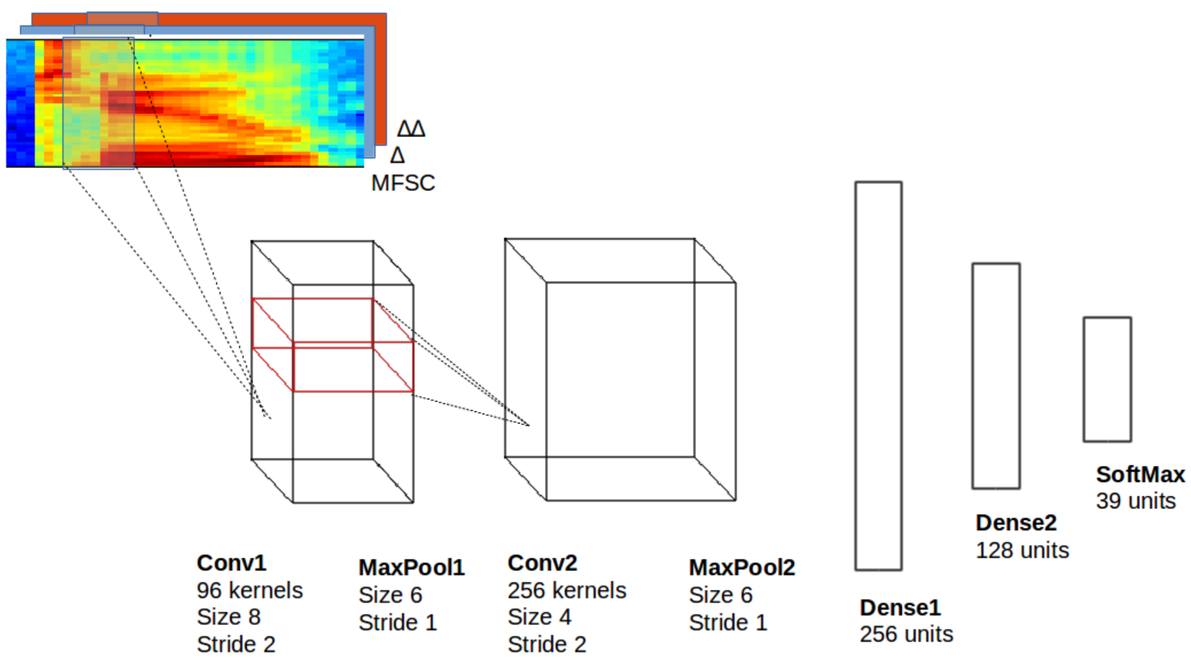
Convolutional Neural Networks for Speech Recognition

HELDER MARTINS & XINYI LIN

Supervised by Giampiero Salvi

Convolutional Neural Network is a variant of Deep Neural Networks where the idea of using fully connected layer is dropped in favor of sharing a smaller set of weights over the input by using the convolution operation

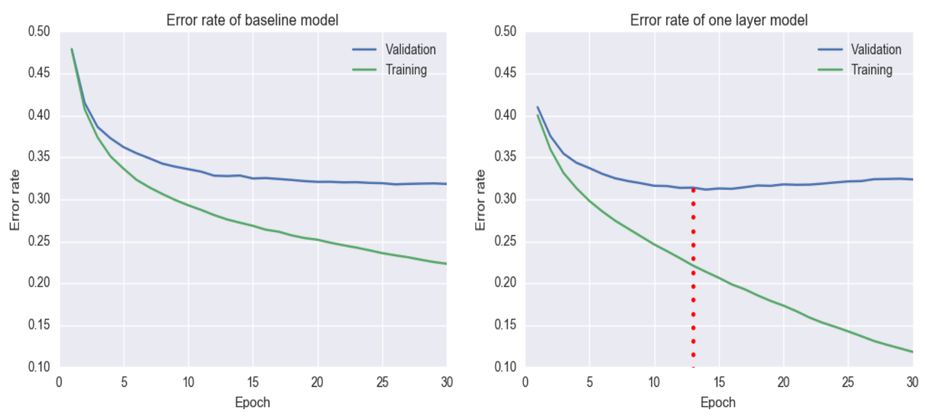


Our reference model can be seen in Figure left, where 2 convolutional layers were used, each one followed by a max-pooling operator. The end of the network contains three dense layers, the last one being the softmax which compute the probabilities for each one the classes.

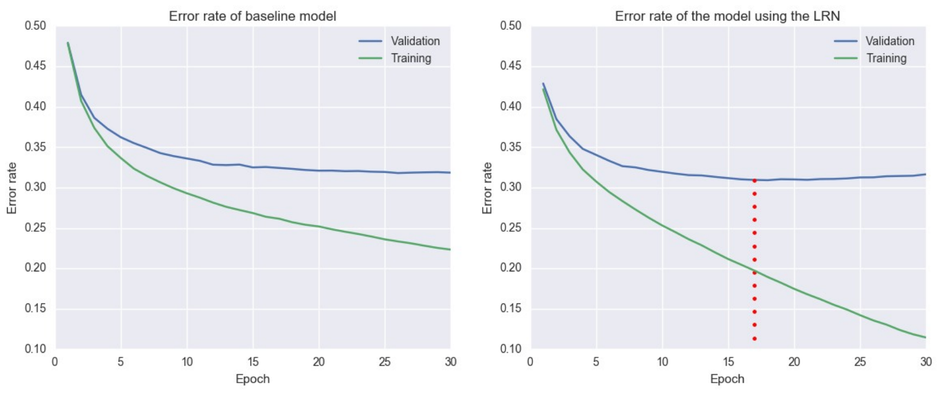
The input can be represented as an "image" of the spectogram static, delta and delta-delta features. The log energy of the mel-frequency spectral coefficients are directly used (instead of the discrete cosine transform) as to maintain the locality of the data over the frequency axis.

The output of the network is 39 correlated phoneme classes, corresponding to a mapping of the original 61 phonemes extracted from the TIMIT dataset.

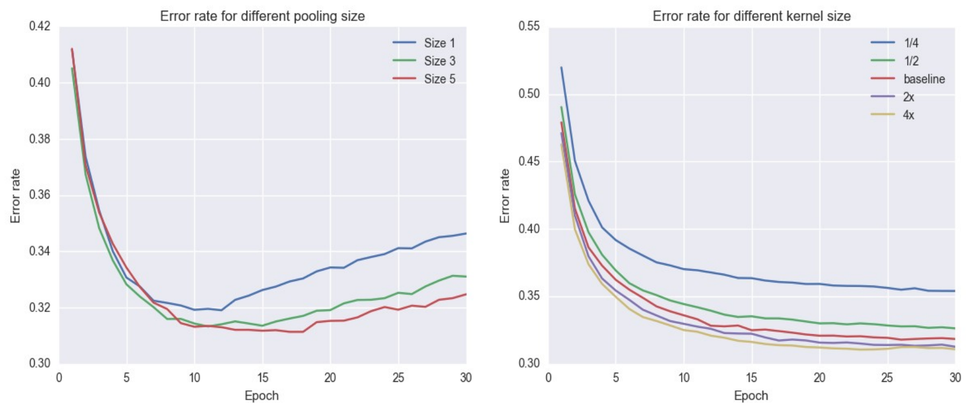
Experiment 1: This graph shows the loss and accuracy computed every 100 iterations during the training of the baseline model. It can be seen that the accuracy increases over time.



Experiment 2: This experiment compared the proposed model with a network composed of one convolutional layer, over 30 epochs. From the right picture below, we can see that one-layer model is overfitting quickly, however their accuracies are still similar.



Experiment 3: Analyzing the effect of adding Local Response Normalization to the baseline model. With it we achieved the best overall result with the least amount of epochs.



Experiment 4: Evaluating the effect of different pooling and kernel sizes over a specific epoch. It can be seen that a too small pooling size causes a faster overfitting to the model with low overall accuracy. Using a larger kernel size we could achieve better result with low phone error rate.