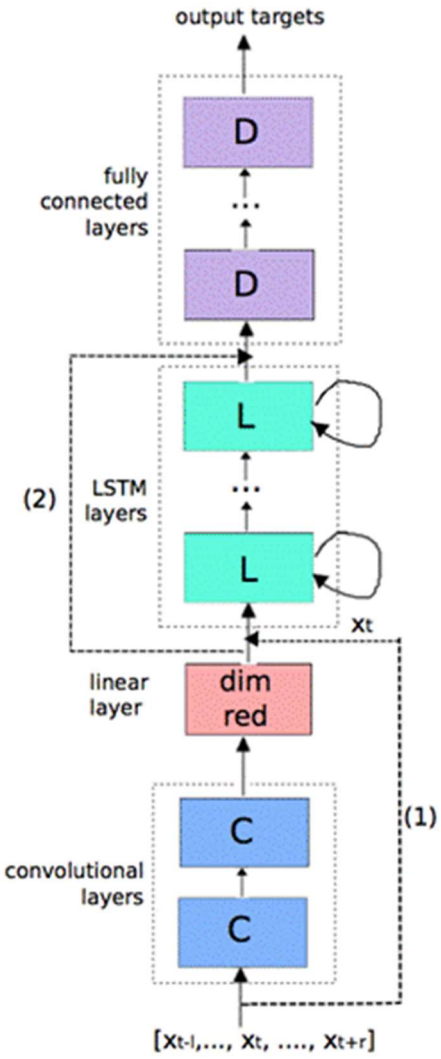
CNN based on digital signal processing

Introduction to the topic

Speech processing is the study of speech signals and the processing methods of signals. The signals are usually processed in a digital representation, so speech processing can be regarded as a special case of digital signal processing, applied to speech signals. Aspects of speech processing includes the acquisition, manipulation, storage, transfer and output of speech signals. The input is called speech recognition and the output is called speech synthesis.[1]

In deep learning, a convolutional neural network (CNN, or ConvNet) is a class of deep neural networks, most commonly applied to analyzing visual imagery. CNNs are regularized versions of multilayer perceptrons. Multilayer perceptrons usually mean fully connected networks, that is, each neuron in one layer is connected to all neurons in the next layer. The "fully-connectedness" of these networks makes them prone to overfitting data. Typical ways of regularization include adding some form of magnitude measurement of weights to the loss function. However, CNNs take a different approach towards regularization: they take advantage of the hierarchical pattern in data and assemble more complex patterns using smaller and simpler patterns. Therefore, on the scale of connectedness and complexity, CNNs are on the lower extreme.[2]

CLDNN (convolutional, long short-term memory, fully connected deep neural networks) is a typical application of CNN. CNN and LSTM can obtain a better improvement than DNN, because CNN is good at reducing the change of frequency domain and LSTM can provide long term memory. In CLDNN, we arrange CNN, LSTM and DNN in to one model, in order to obtain a better efficacy.



Picture 1. Architecture of the CLDNN[3]

The architecture of the CLDNN is shown as above.

Analysis of results including pros and cons.

The experiment shows that when processing 300 hours of speech signal, compared to 4 other types of architectures, CNN+LSTM+DNN has the lowest WER(Word Error Rate).

Method	WER
LSTM	21.6
CNN+2*LSTM	21.8
CNN+3*LSTM	21.5
CNN+LSTM+DNN	20.6
LSTM+DNN	20.8

The experiment also shows that adding one layer of LSTM is helpful but continuing adding LSTM is useless.

The advantages of this CNN-based method is it can extract features from signal and use LSTM or DNN to process the features. The combination of several techniques is very useful. However, it only works when the variation of data is relatively small, otherwise it can only reduce the WER by 2~3%. [4]

Recommendations for a person who want to develop or use such systems

CNN-based method is very useful to extract different voice source from a complicated environment. It can reduce the WER sharply but also the computation rises obviously. The person who want to use this system must consider this trade. If the project needs to be real-time, then the efficacy should be improved before deploying it.

Conclusions

CNN-based method brings a new direction into speech processing area. It can reduce WER sharply on some conditions. But also, it becomes deeper and more complicated, which indicates more computations.

Summary of Group Works

Yerlan Sharipov: Speech processing is the study about extracting meaningful information from signals in digital representation. Over the last two centuries communication method evolved from the Morse code alphabet to direct audio calls.

Olutayo: Regular FST based models declines with increasingly long numeric sequences. There are two corrections, the first one is to including more training data with long numeric sequence. This second one is improve the neural network of the model.

Shen: The neural network algorithm for real-time signal processing and Kalman Filter.

Tan: The front-end processing of the voice signal must be processed to ensure the accuracy of the subsequent processing steps.

Summary of references

[1] https://en.wikipedia.org/wiki/Speech_processing

[2]https://en.wikipedia.org/wiki/Convolutional_neural_network

[3]T N Sainath R J Weiss A Senior et al. "Learning the speech front-end with raw waveform CLDNNs[C]" Sixteenth Annual Conference of the International Speech Communication Association 2015.

[4]Wang X, Zhang P, Zhao Q, et al. Improved End-to-End Speech Recognition Using Adaptive Per-Dimensional Learning Rate Methods[J]. Ieice Transactions on Information & Systems, 2016, 99(10):2550-2553.