Project Name: Recurrent neural network(RNN)
Name: Alec Mabhiza Chirawu
Chinese Name: 亚历克上
Student Number: M202161029

Neural Networks and Deep Learning
Information And Engineering Department
USTB

Professor: 黄旗明(Roland)

22 April 2022

Summary

Long Short-Term Memory (LSTM) is a specific recurrent neural network (RNN) architecture that was designed to model temporal sequences and their long-range dependencies more accurately than conventional RNNs. We show that a two-layer deep LSTM RNN can exceed state-of-the-art speech recognition performance. This architecture makes more effective use of model parameters than the others considered, converges quickly, and outperforms a deep feed forward neural network having an order of magnitude more parameters.

TABLE OF CONTENTS

1 Introduction

Recurrent neural networks (RNNs) are emerging as a powerful tool to model sequence data for speech recognition. RNNs have demonstrated great success in sequence labeling and prediction tasks such as handwriting recognition and language modeling. They can exploit a dynamically changing contextual window over the input sequence history rather than a static one as with feed-forward networks. Long Short-Term Architecture (LSTM) is conceptually attractive for the acoustic modeling of speech. Bidirectional LSTM(BLSTM) networks that operate on the input sequence in both directions to make a decision for the current input have been proposed for phonetic labeling of acoustic frames on the TIMIT speech database. For online and offline handwriting recognition, BLSTM networks used together with a Connectionist Temporal Classification (CTC) layer and trained from unsegmented sequence data, have been shown to outperform state of the art Hidden-Markov-Model (HMM) based system.

1.1 Methodology

Alation is one of the first local explainability methods used in computer vision. Ablation infers image pixel importance by measuring the change in prediction when removing information from a region. The size and shape of the ablation pattern may lead to false positive and false negative in determining importance among features. The permutation method is such a method where the relative importance is determined based on all predictions after single feature values are randomised.

1.2 Scope

The LSTM contains special units called memory blocks in the recurrent hidden layer. Memory blocks contain memory with self-connections storing the temporal state of the network and gates to control the flow of information. The forget gate scales the internal state of a cell before adding it as input to the cell through the self-recurrent connection of the cell, therefore adaptively forgetting or resetting the cell's memory.

2. Findings

2.1 Deep LSTM

Deep LSTM RNNs are similar to deep DNNs in the sense that they are a feedback neural network unrolled in time where each layer shares the same model parameters. The features from a given time instant are only processed by a single nonlinear layer before contributing the output for that time instant. Instead of increasing the memory size of a standard model by a factor of 2, one can have 4 layers with approximately the same number of parameters. This allows for better use of parameters by distributing them over the space through multiple layers.

An LSTM network computes a mapping from an input sequence $x = (x_1, ..., x_T)$ to an output sequence $y = (y_1, ..., y_T)$ by calculating the network unit activations using the following equations iteratively from t = 1 to T:

$$i_t = \sigma(W_{ix}x_t + W_{im}m_{t-1} + W_{ic}c_{t-1} + b_i)$$

$$f_t = \sigma(W_{fx}x_t + W_{fm}m_{t-1} + W_{fc}c_{t-1} + b_f)$$

$$c_t = f_t \Theta c_{t-1} + i_t \Theta g(W_{cx}x_t + W_{cm}m_{m-1} + b_i)$$

$$o_t = \sigma(W_{ox}x_t + W_{om}m_{m-1} + W_{oc}c_{t-1} + b_o)$$

$$m_t = o_t \Theta h(c_t)$$

$$y_t = \phi(W_{ym}m_t + b_y)$$

where the W terms denote weight matrices, Wic , W_{fc} , W_{oc} are diagonal weight matrices for peephole connections, the b terms denote bias vectors (b_i is the input gate bias vector), σ is the logistic sigmoid function, and i, f, o and c are respectively the input gate, forget gate, output gate and cell activation vectors, all of which are the same size as the cell output activation vector m, is the element-wise product of the vectors, g and h are the cell input and cell output activation functions, generally and in this paper tanh, and φ is the network output activation function.

2.2 Deep LSTMP

We know that DNNs generalize better to unseen examples with increasing depth, which makes it harder to overfit to the training data since the inputs to the network need to go through many non-linear functions. With this motivation, we have experimented with deep LSTMP architectures, where the aim is increasing the memory size and eneralization power of the model.
With LSTMP architecture the equations are:

$$r_t = W_{rm} m_t$$
$$y_t = \phi(W_{yr} r_t + b_y)$$

Here the m_{t-1} activation is replaced with r_{t-1}.

2.4 Experiment

We evaluate and compare the performance of LSTM RNN architectures on a large vocabulary speech recognition task – the Google Voice Search task. We use a hybrid approach for acoustic modeling with neural networks, wherein the neural networks estimate hidden Markov model (HMM) state posteriors. We deweight the silence state counts by a factor of 2.7 when estimating the state frequencies. The experiment require more resources and time.

3. Conclusion

RNNs allow us to perform modeling over a sequence or a chain of vectors. These sequences can be either input, output or even both. Therefore, we can conclude that neural networks are related to lists or sequences. So, whenever you have data of sequential nature, you should apply recurrent neural networks.

4. Reference List

Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." *arXiv preprint arXiv:1406.1078* (2014).

Sherstinsky, Alex. "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network." *Physica D: Nonlinear Phenomena* 404 (2020): 132306.

Yin, Wenpeng, et al. "Comparative study of CNN and RNN for natural language processing." *arXiv preprint arXiv:1702.01923* (2017).

Miao, Yajie, Mohammad Gowayyed, and Florian Metze. "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding." *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015.