Jenish Maharjan **Week 4**

Paper: *LSTM: A Search Space Odyssey*

Summary:

Greff et al., as the title of the papers suggest, present their exploration of LSTM architecture for recurrent neural networks in the paper "LSTM: A Search Space Odyssey". They specifically present their analysis of eight variants of LSTM architecture on three representative tasks – speech recognition, handwriting recognition and polyphonic music modeling. LSTMs, due to their effectiveness and scalability for modeling learning problems related to sequential data, have been widely researched by many scholars for various tasks. The authors, in a way, bring all the previous studies together with this paper and provide a document that a person new to the field can learn all about LSTMs from. They start by presenting the core idea of the LSTM architecture which include a memory cell capable of maintaining its state over time and non-linear gating units that regulate the information flow into and out of the cells. The authors then move on to describing a vanilla LSTM architecture along with a brief history of LSTMs.

A vanilla LSTM architecture comprises of three gates – input, forget and output gates, block input, a single cell, an output activation function and peephole connections. There is a recurrent connection from the output of the block to the block input and the gates. The architecture uses logistic sigmoid as activation for the gates and hyperbolic tangent as block input and output activation functions. Forget gates and peephole connections were absent in the initial version introduced by Hochreiter and Schmidhuber. For the purpose of this study, the authors analyzed 8 variants of the vanilla architecture – each different from the original by a single change, which helped them isolate the effect of each change. These variants were then evaluated using three different datasets to analyze cross-domain variations. They used the TIMIT Speech corpus dataset for speech recognition, IAM Online Handwriting Database for handwriting recognition and JSB Chorales for polyphonic music modeling.

For the TIMIT and IAM Online tasks, a bidirectional LSTM with two hidden layers – one for processing the input forwards and one for backwards in time – was used. For the JSB Chorales task, a normal LSTM with one hidden layer was used. Both of the networks had a softmax output layer. The loss function for TIMIT and JSB Chorales tasks were set to Cross-entropy Error and for the IAM Online task, to Connectionist Temporal Classification (CTC) error. They employed Stochastic Gradient Descent with updates after each sequence for training. The variants of the included six with one of input gate, forget gate, output gate, input activation function, output activation function or peepholes missing. Another variant had the input and forget gates coupled. The last variant added recurrent connections between all the gates. For hyperparameter search, they used random search motivated by easy implementation and trivial parallelization. The hyperparameters that were tuned were the number of LSTM blocks per hidden layer, learning rate, momentum and standard deviation of Gaussian input noise. In addition to these, for the TIMIT dataset, choice between traditional momentum and Nesterov-style momentum was considered as well.

In the results section, the authors present the comparison of the performance of the variants for the different tasks. While for each task a different variant had the best results, they conclude that improvement in the performance is not significant for any of the eight variants. The authors also present their analysis showing the impact of hyperparameters on the test set performance. They show that learning rate is the most important hyperparameter that larger networks perform better as the size of hidden layer is increased. With a very well-organized paper about their study of LSTM architecture, the authors were able to present new insights on architecture selection and hyperparameter tuning for LSTM networks.

Strengths:

- The authors have used pie charts and box plots which makes understanding their results and analysis more intuitive and easier.
- The contents of the paper are well-organized into meaningful section – specifically the division of content in the results helps emphasize on each of the findings they made in their study. The bulleted conclusion points improved the readability of the section.

Weaknesses:

- Some sections of the paper, specifically the introduction and the history section could be restructured to make it more concise and more readable.
- The functions of each gates could have been described in words in addition to the mathematical equations for a more intuitive understanding.

Confusions:

- I wasn't sure how to interpret the last few graphs in the supplementary material section.
- What does interaction of hyperparameters mean? In my understanding, the degree of interaction would be how changing hyperparameter changes the other.

Discussion Questions:

- How does Connectionist Temporal Classification (CTC) error work? How are loss functions decided for a specific task?
- How does full back propagation through time work?
- What are some other domains/tasks where LSTM architectures are preferred?