[Paper Review]
Sequence to Sequence Learning with Neural Networks

Paul Jason Mello
Department of Computer Science and Engineering
University of Nevada, Reno
pmello@unr.edu

Abstract

"Deep Neural Networks (DNNs) are powerful models that have achieved excellent performance on difficult learning tasks. Although DNNs work well whenever large labeled training sets are available, they cannot be used to map sequences to sequences. In this paper, we present a general end-to-end approach to sequence learning that makes minimal assumptions on the sequence structure. Our method uses a multilayered Long Short-Term Memory (LSTM) to map the input sequence to a vector of a fixed dimensionality, and then another deep LSTM to decode the target sequence from the vector. Our main result is that on an English to French translation task from the WMT'14 dataset, the translations produced by the LSTM achieve a BLEU score of 34.8 on the entire test set, where the LSTM's BLEU score was penalized on out-of-vocabulary words. Additionally, the LSTM did not have difficulty on long sentences. For comparison, a phrase-based SMT system achieves a BLEU score of 33.3 on the same dataset. When we used the LSTM to rerank the 1000 hypotheses produced by the aforementioned SMT system, its BLEU score increases to 36.5, which is close to the previous best result on this task. The LSTM also learned sensible phrase and sentence representations that are sensitive to word order and are relatively invariant to the active and the passive voice. Finally, we found that reversing the order of the words in all source sentences (but not target sentences) improved the LSTM's performance markedly, because doing so introduced many short term dependencies between the source and the target sentence which made the optimization problem easier." [4]

1 Summary

Using prior knowledge of RNN capabilities and limitations, the authors of this work propose utilizing LSTM's to learn large sequences of text with minimal assumptions regarding the structure of the sequence. They prove their methods effectiveness in a translation task between English and French achieving near SOTA BLEU and perplexity scores while only utilizing a simple model.

2 Main Contributions

The main contributions of this work are as follows:

2.1 Key Contributions

- They introduce a variational auto-encoder structure, with encoder and decoder, to LSTM's. The encoder LSTM takes variable length input sequences and maps them to a fixed vector dimension. Then, the decoder LSTM maps the sequence vector to a target output.

- Idea to reverse input sequence to capture short term dependencies.
- Application of LSTM for long sequence generation with no noticeable degradation in quality of generated output over significantly long sequence generations.

2.2 Innovative Aspects

- Utilizing a small network and dataset, thy achieve a BLEU score of 34.81 using five ensembled LSTM's consisting of four layers each, with a beam search on the decoder.
- They demonstrate that reversing the order of the input sequence helps capture the short term dependencies between words and improves the output generation by reducing perplexity by 20% and increasing BLEU scores by 20%.

3 Strengths and Weaknesses

This paper does not aim to achieve SOTA results, but it is a by-product of their well designed system. Below are some of these strengths and some weaknesses.

3.1 Strengths

The authors illustrate the capabilities of LSTM's in long form sequence generation, particularly its ability to provide coherent and consistent translations between English and French on the WMT 14 language dataset with minimal assumptions regarding the structure of the sequence. They achieve a very high BLEU score of 34.81 on this dataset given the size and depth of the model, which are both severely limited. Their impressively coherent generation of long term sequences are overshadowed by their architecture design which encodes and decodes input sequences to target sequences. It is further impressive as the utilization of these dual LSTM's do not incur any significant computation costs. The use of reversed input sequences plays with the concepts of language by helping the model extract the short range dependencies between words better, providing a better generalization of training sequences. Another important strength is the models capability to train on variable length sequences and map them to a fixed size vector dimension. In these strengths they are able to scale their sequence outputs to be better with more data.

The authors utilize many tricks to maximize the accuracy of their model going as far as to use an ensemble of models for prediction, a multi-gpu setup with inter-gpu communication, and utilizing beam search for the decoder.

3.2 Weaknesses

There are little to no weaknesses in this paper considering the date of publication is in late 2014 and the nature of this paper is to probe LSTM's capabilities using simple architectures. Advances in computing and the realization of the importance of data preprocessing have resulted in far better modeling techniques and architectures since this papers release. For that reason I can not say any weaknesses exist, especially given the nature of this model being kept small. However, by today's standards, such as Mamba [1] and modern scaling laws[2], the training data is too small and the gpu optimizations too weak. Additional evaluations on more datasets would also be useful in to assess model quality.

3.3 Areas of Improvements

After reviewing this paper I have little to say about areas of improvement given the date of publication. As described in the weaknesses section 3.2, most of the areas of improvements rely on additional resources like data, computing power. Especially considering that they coded a multi-gpu setup using C++ to communicate calculations between the 4 layers of the LSTMs and the softmax's which were naively implemented.

4 Discussion

Ultimately, the use of LSTMs for long term sequence to sequence generation was a pivotal moment in the AI field. While the idea is relatively simple by today's standards. Sequence to sequence LSTM's started a significant shift in the monetization of AI and subsequently supported the increasing shift to research and development of other more advanced systems in the following few years. It is not far off to say that without sequence to sequence modeling LLMs, in their current state, would not exist today let alone the massive financial investments in the AI industry.

5 Conclusion

The authors propose using an encoder and decoder LSTM architecture for the generation of text to translate between English and French. They demonstrate their results exceed other contemporary models with only a simple toy system consisting of 4-layers. Despite this simple system, the data preprocessing, model type, and architecture all culminated in a particularly strong system for text generation. If nothing else, this paper repeats many of the lessons learned from AlexNet [3], which is to say that doing everything possible to improve model quality and performance will have a profound effect on the models final quality. In this case, reversing the input sequence, distributing gpu loads, using encoder and decoder architecture, using a highly efficient vector representation, and more lead to the SOTA results on text generation.

References

[1] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024.

[2] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012.

[4] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks, 2014.