A Review of Google's Light-weighted NLP Model - PRADO and It's Latest Improvements – pQRNN

Huanzhen Hu

University of Illinois at Urbana-Champaign

hh21@illinois.edu

Abstract Recently, there has been a great interest in the development of small and accurate neural networks that run entirely on devices such as mobile phones, smart watches and IoT. This enables user privacy, consistent user experience and low latency. Although a wide range of applications have been targeted from wake word detection to short text classification, yet there are no on-device networks for long text classification. This review will introduce some novel NLP architectures developed by Google, including PRADO and its improvements pQRNN. Both achieved SOTA on most text classification problems and BERT-level performance with only 1/300 parameters of the BERT.

1 Introduction

One of the fundamental tasks in Natural Language Processing is related to long text classification. Given a document, the goal is to assign one or more categories of interest to the text. For a long time, the most successful text classification approaches relied on sparse lexical features. However, with the recent advancements in deep learning, various neural network architectures like CNN (Kim, 2014), LSTM (Zhang et al., 2015), hierarchical attention mechanisms (Yang et al., 2016) showed improvement in performance. In 2018, the Bidirectional Encoder Representations from Transformers (BERT) model, a 24-layer, 1024-hidden, 16-heads, 340M parameter neural network architecture was published by Google and achieved state-of-the-art performance on a number of natural language understanding tasks, including GLUE, SQuAD, and SWAG. Despite the great success, the complex architecture of BERT constrained its operation only on remote servers. The limited memory and computation capacity of personal device raised higher demand on the model. The main challenge is to limit the complexity of the neural network architectures while maintaining high performance.

In 2019, a new neural architecture – PRADO published by Google achieved SOTA on most text classification problems while only used 200K parameters and less. Unlike most models which uses fixed number of parameters on each token, PRADO requires only a few parameters to learn the most relevant or useful tokens for the task.

Recently, Google's researchers announced the latest improvements of PRADO and named it pQRNN. The new model reached the new SOTA of the NLP task with the smallest model size. The novelty of pQRNN is that it combines simple projection operations with the quasi-RNN encoder for fast and parallel processing. The research indicated pQRNN can achieve BERT-level performance on text classification tasks with only 1/300 parameters of the BERT.

2 Basic Idea of PRADO

As shown in Figure 1, PRADO consists of a projected embedding layer, a convolutional and attention encoder mechanism and a final classification layer. Generally speaking, the text input of the NLP model is processed into a form suitable for neural networks by first dividing the text into tokens corresponding to the values in the predefined universal dictionary. The neural network then uses trainable parameter vectors (including embedding tables) to uniquely identify each text segment. However, the way in which text is segmented has a significant impact on the model performance, size, and latency. Since most NLP tasks can be solved by knowing a small subset of these segments and may ignore the subtle differences between segments (Figure 2), allowing the network to determine the most relevant segments for a given task results in better performance. In addition, the network does not need to be able to uniquely identify these segments, but only needs to recognize clusters of text segments. Leveraging these insights, PRADO was designed to learn clusters of text segments from words rather than word pieces or characters, which enabled it to achieve good performance on low-complexity NLP tasks. Since word units are more meaningful, and yet the most relevant words for most tasks are reasonably small, many fewer model parameters are needed to learn such a reduced subset of relevant word clusters.

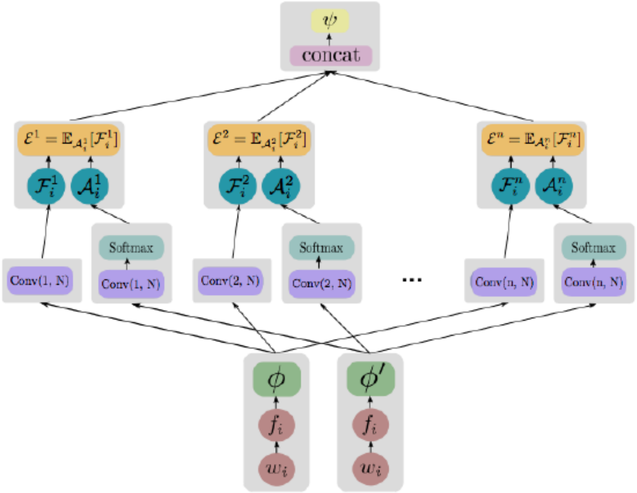


Figure 1: PRADO Model Architecture

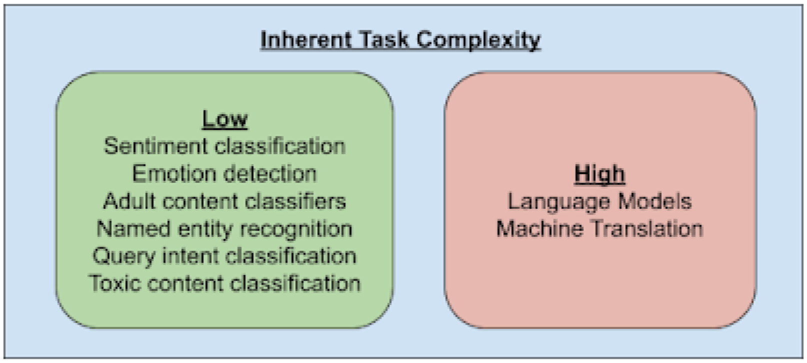
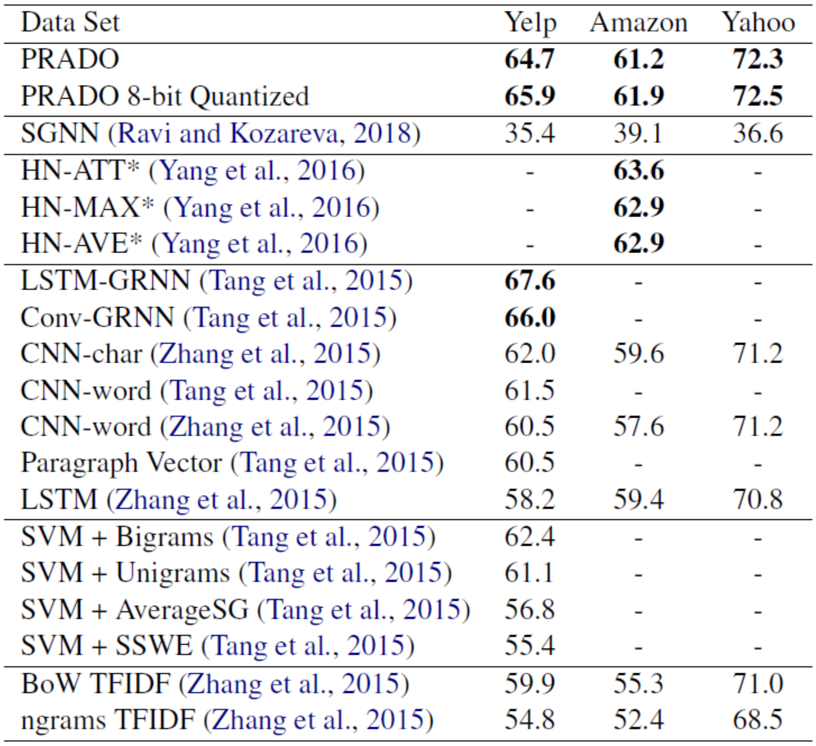


Figure 2: Common NLP Tasks Complexity

3 Performance of PRADO

The model was evaluated on the following large-scale document classification tasks: Yelp reviews, Amazon reviews and Yahoo Answers Table 1 shows the obtained results for each data set and method.

The comparison showed PRADO significantly outperforms existing neural networks approaches like LSTM, CNN-char and CNN-word with +1.1 up to +6.5% depending on the task and data set, and it achieves comparable results to the hierarchical attention models of (Yang et al., 2016). This is very impressive given that PRADO produces magnitudes smaller and compact neural networks.



Data Set	Yelp	Amazon	Yahoo
PRADO	**64.7**	**61.2**	**72.3**
PRADO 8-bit Quantized	**65.9**	**61.9**	**72.5**
SGNN (Ravi and Kozareva, 2018)	35.4	39.1	36.6
HN-ATT* (Yang et al., 2016)	-	**63.6**	-
HN-MAX* (Yang et al., 2016)	-	**62.9**	-
HN-AVE* (Yang et al., 2016)	-	**62.9**	-
LSTM-GRNN (Tang et al., 2015)	**67.6**	-	-
Conv-GRNN (Tang et al., 2015)	**66.0**	-	-
CNN-char (Zhang et al., 2015)	62.0	59.6	71.2
CNN-word (Tang et al., 2015)	61.5	-	-
CNN-word (Zhang et al., 2015)	60.5	57.6	71.2
Paragraph Vector (Tang et al., 2015)	60.5	-	-
LSTM (Zhang et al., 2015)	58.2	59.4	70.8
SVM + Bigrams (Tang et al., 2015)	62.4	-	-
SVM + Unigrams (Tang et al., 2015)	61.1	-	-
SVM + AverageSG (Tang et al., 2015)	56.8	-	-
SVM + SSWE (Tang et al., 2015)	55.4	-	-
BoW TFIDF (Zhang et al., 2015)	59.9	55.3	71.0
ngrams TFIDF (Zhang et al., 2015)	54.8	52.4	68.5

Table 1: Evaluation Results

4 Basic Idea of pQRNN

As shown in Figure 3, the pQRNN model is composed of three building blocks, a projection operator that converts tokens in text to a sequence of ternary vectors, a dense bottleneck layer and a stack of QRNN encoders. The implementation of the projection layer in pQRNN is identical to that used in PRADO and helps the model learn the most relevant tokens without a fixed set of parameters to define them. It first fingerprints the tokens in the text and converts it to a ternary feature vector using a simple mapping function. This results in a ternary vector sequence with a balanced symmetric distribution that uniquely represents the text. This representation is not directly useful since it does not have any information needed to solve the task of interest and the network has no control over this representation. We combine it with a dense bottleneck layer to

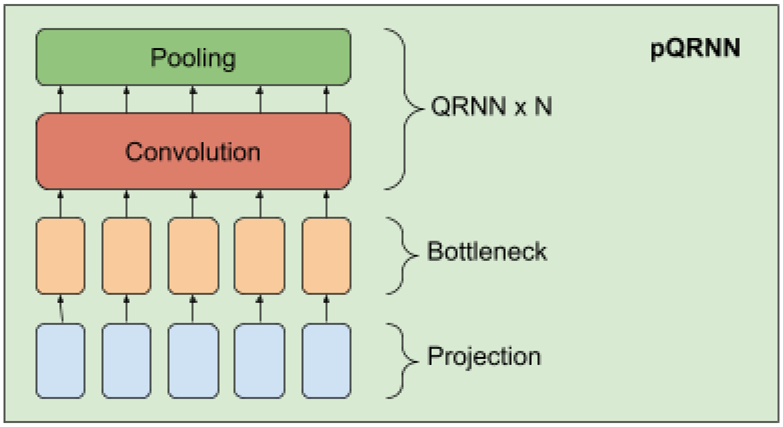


Figure 3: pQRNN Model Architecture

allow the network to learn a per word representation that is relevant for the task at hand. The representation resulting from the bottleneck layer still does not take the context of the word into account. We learn a contextual representation by using a stack of bidirectional QRNN encoders. The result is a network that is capable of learning a contextual representation from just text input without employing any kind of preprocessing.

5 Performance of pQRNN

The model was evaluated on the civil comments dataset and compared it with the BERT model on the same task. As shown in Figure 4. Without any kind of pre-training and just trained on the supervised data, the AUC for pQRNN is 0.963 using 1.3 million quantized (8-bit) parameters. With pre-training on several different data sources and fine-tuning on the supervised data, the BERT model gets 0.976 AUC using 110 million floating point parameters.

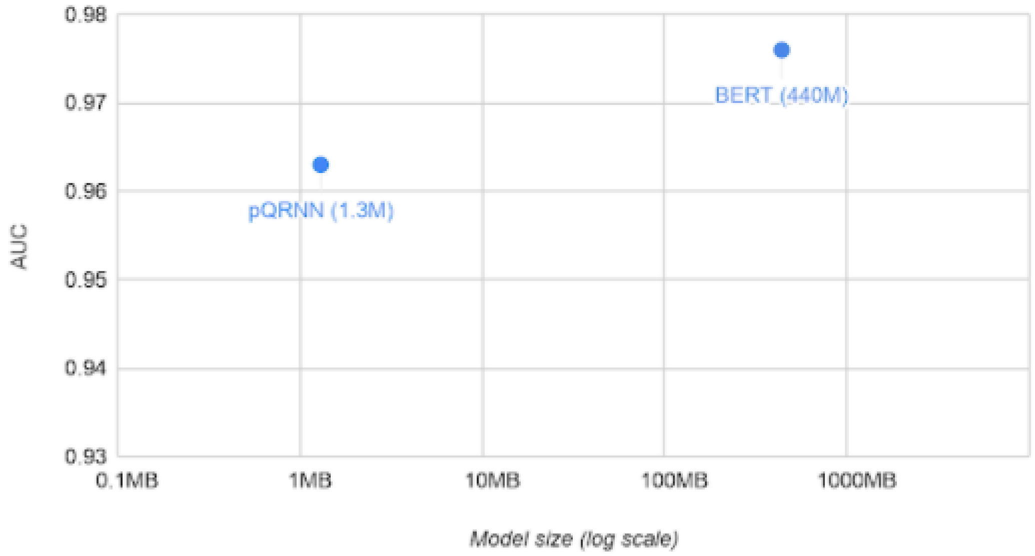


Figure 4: Model Comparison between pQRNN and BERT

6 Conclusion

PRADO is an effective novel trainable projection on-device neural network with attention. It's performance on multiple large-scale document classification tasks is demonstrated. One more step forward, it's latest improvement – pQRNN, can nearly achieve BERT-level performance, despite being 300x smaller. The insights behind PRODA and pQRNN could be the foundation for the next generation of state-of-the-art light-weight text classification models.

Reference

Prabhu Kaliamoorthi, Sujith Ravi, Zornitsa Kozareva. 2019. PRADO: Projection Attention Networks for Document Classification On-Device.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1746–1751. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, Advances in Neural Information Processing Systems 28, pages 649–657. Curran Associates, Inc.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical attention networks for document classification. In NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016, pages 1480–1489.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In EMNLP, pages 1422– 1432. The Association for Computational Linguistics.

https://en.wikipedia.org/wiki/BERT_(language_model)

https://ai.googleblog.com/2020/09/advancing-nlp-with-efficient-projection.html

https://github.com/tensorflow/models/tree/master/research/sequence_projection