Compressing Recurrent Neural Network Model using Vector Quantization

Laboratory for Software Design & Analysis
established in 2005

Kundjanasith Thonglek, Kohei Ichikawa, Keichi Takahashi, Chawanat Nakasan, Hajimu Iida
Email: thonglek.kundjanasith.ti7@is.naist.jp

Introduction

Recurrent Neural Network (RNN) is a neural network model that achieves significant performance on tasks such as machine translation, speech recognition, and language modeling. The goal of language modeling is to predict the next word in a sentence given a history of previous words. However, deploying large language models on mobile devices is challenging due to the limitation in computing resource and storage. In this work, we apply vector quantization techniques to RNN models to reduce the model size without significant accuracy loss. We evaluate the trade-off between model accuracy and model size using the Penn Tree Bank (PTB) dataset.

Compressing recurrent neural network model

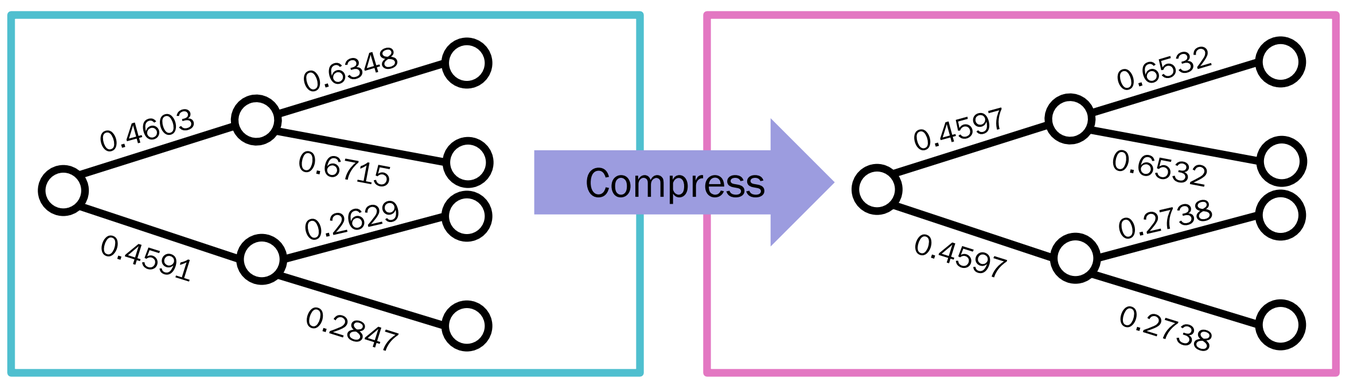
Hyperparameter	SMALL	MEDIUM	LARGE
The initial scale of the weight	0.10	0.05	0.04
The maximum permissible norm of the gradient	5	5	10
The number of epochs trained with the initial learning rate	4	6	14
The total number of epochs for training	13	39	55
The probabilities of keeping weights in the dropout layer	1.00	0.50	0.35
The decay of learning rate for each epoch after initialize	0.50	0.80	0.87

The model processes one word at a time and computes probabilities of the possible values for the next word in the sentence. State of the network is initialized with a vector of zeros and gets updated after reading word.

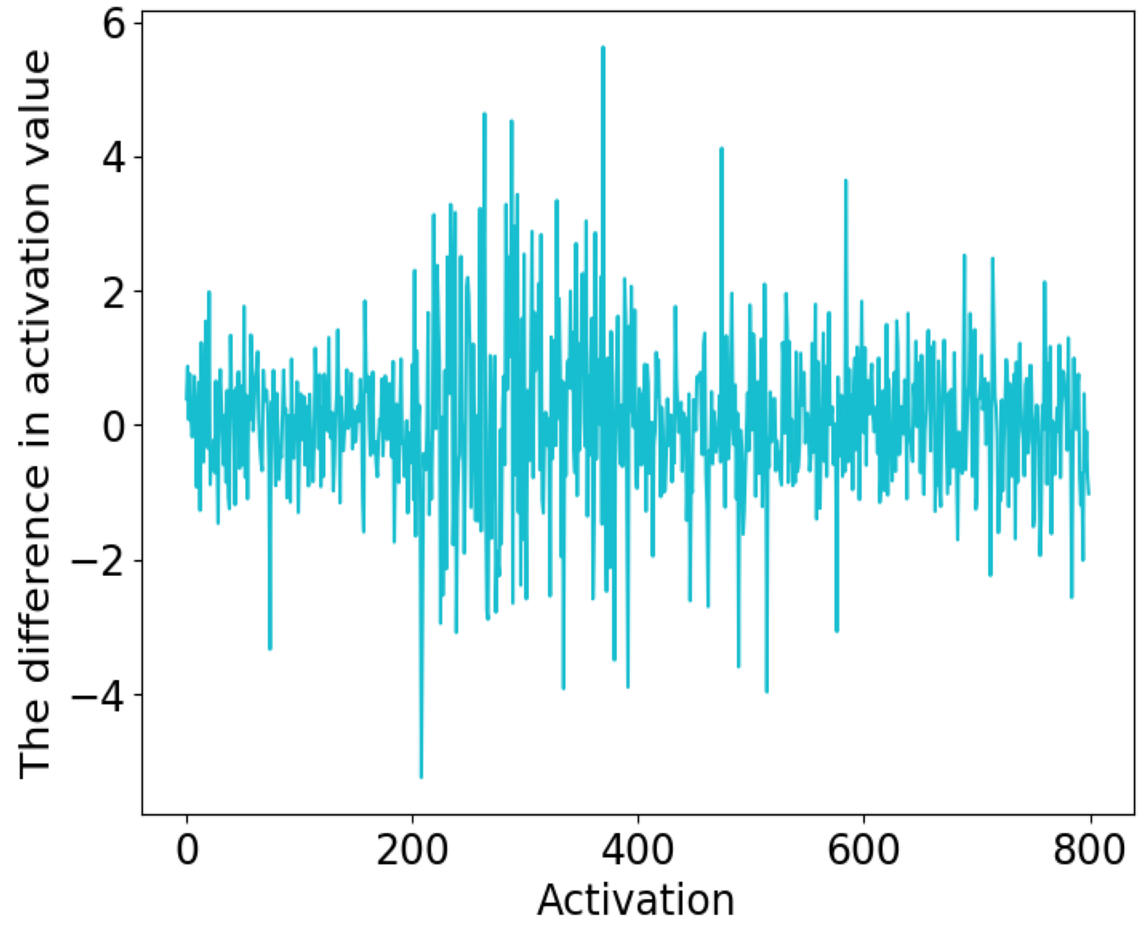
By design, the output of RNN depends on the arbitrarily distant inputs. Every word in a batch correspond to a time.

Perplexity is used by convention in language modeling, is monotonically decreasing in the likelihood of the test data, and is algebraically equivalent to the inverse of the geometric mean per-word likelihood. In the context of natural language, perplexity is an exponentiation of the entropy. A lower perplexity score indicates better generalization performance

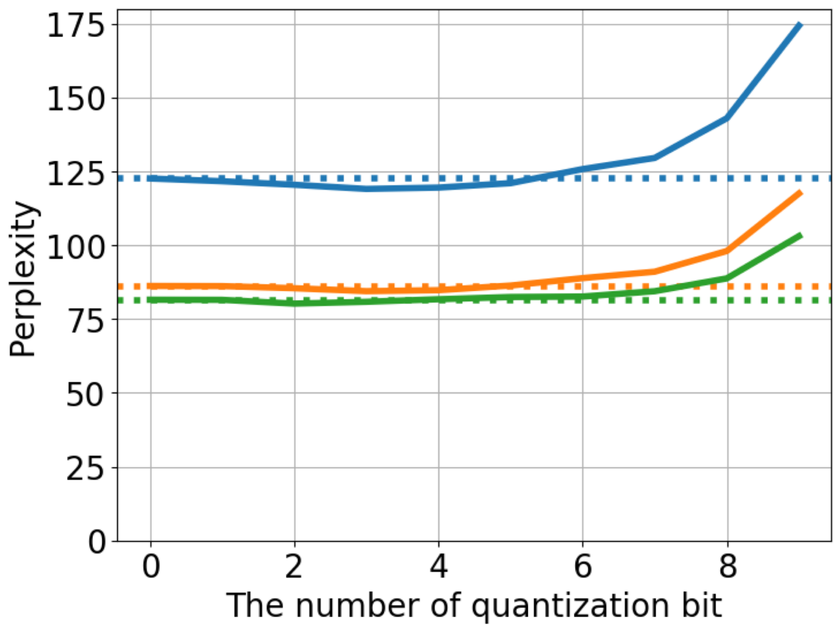
Vector quantization is adapted for lossless data compression when the data exhibit vector structures, such as in a neural network. The key operation in a vector quantization is the quantization of a random vector by encoding it as a binary codeword.

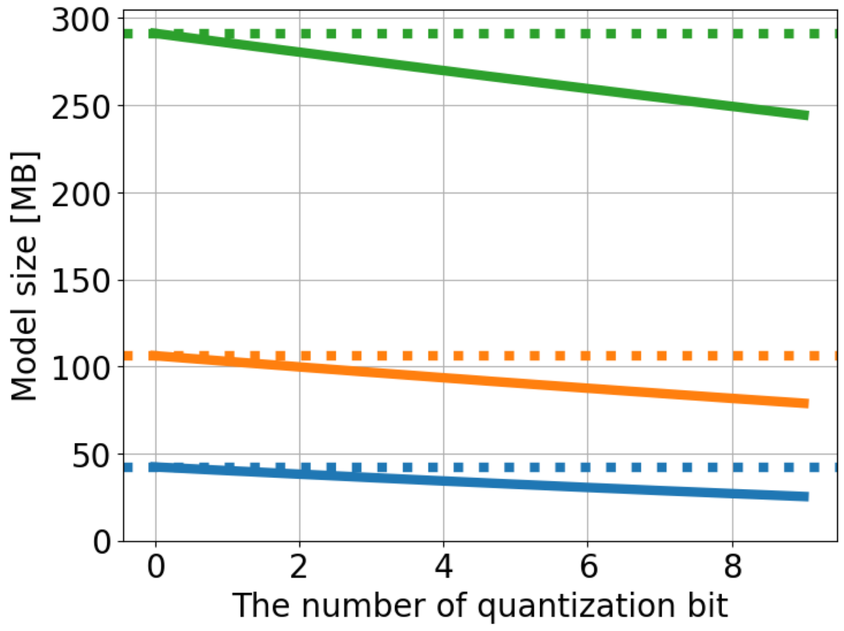


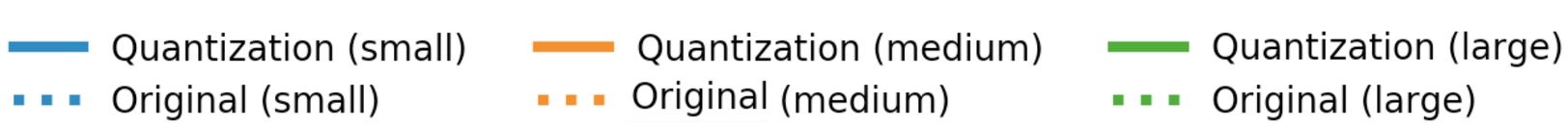
Each input vector can be viewed as a point in an *n*-dimensional space. The vector quantizer is defined by a partition of this space into a set of nonoverlapping *n*-dimensional regions.



Perplexity with the different model size







Increasing the number of quantization bit in vector produces an effect to increase the perplexity exponentially and decrease the model size linearly.

The result indicated that applying vector quantization techniques decreased the model size of small, medium and large model without significant accuracy loss by 23.48%, 14.78% and 10.89% respectively.

Our future work is applying the other lossless compression techniques without significant accuracy loss.