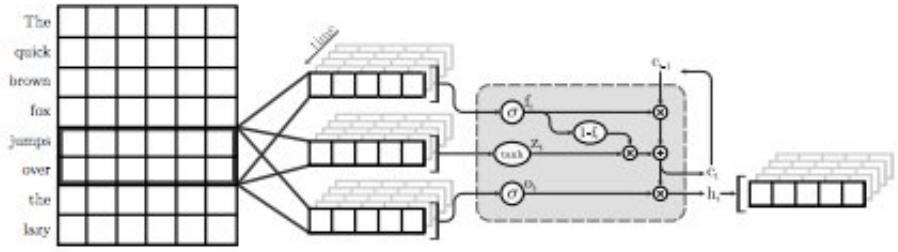
# An evaluation of the accuracy-efficiency tradeoffs of neural language models

October 9 2018, by Ingrid Fadelli




An illustration of the first QRNN layer for language modeling. In this visualization, a QRNN layer with a window size of two convolves and pools using embeddings from the input. Note the absence of recurrent weights. Credit: Tang & Lin.

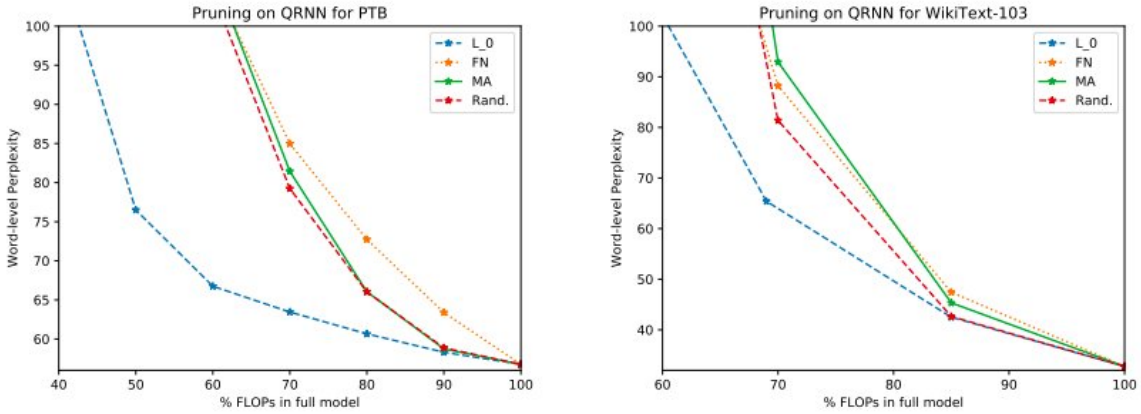
A team of researchers at the University of Waterloo in Canada has recently carried out a study exploring accuracy-efficiency tradeoffs of neural language models (NLMs) specifically applied to mobile devices. In their paper, which was pre-published on arXiv, the researchers also proposed a simple technique to recover some perplexity, a measure of a language model's performance, using a negligible amount of memory.

NLMs are language models based on neural networks through which algorithms can learn the typical distribution of sequences of words and make predictions about the next word in a sentence. These models have a number of useful applications, for instance, enabling smarter software

keyboards for mobile phones or other devices.

"Neural language models (NLMs) exist in an accuracy-efficiency trade-off space where better perplexity typically comes at the cost of greater computation complexity," the researchers wrote in their paper. "In a software keyboard application on mobile devices, this translates into higher power consumption and shorter battery life."

When applied to software keyboards, NLMs can lead to more accurate next-word prediction, allowing users to input the next word in a given sentence with a single tap. Two existing applications that use neural networks to provide this feature are SwiftKey1 and Swype2. However, these applications often require a lot of power to function, rapidly draining the batteries of mobile devices.




Full experimental results on Penn Treebank and WikiText-103. We illustrate the perplexity–efficiency tradeoff space on the test set obtained before applying the single-rank update. Credit: Tang & Lin.

"Based on standard metrics such as perplexity, neural techniques represent an advance in the state of the art language modeling," the researchers explained in their paper. "Better models, however, come at a cost in computational complexity, which translates to higher power consumption. In the context of mobile devices, energy efficiency is, of course, an important optimization objective."

According to the researchers, NLMs have so far primarily been evaluated in the context of image recognition and keyword spotting, while their accuracy-efficiency tradeoff in natural language processing (NLP) applications has not yet been thoroughly investigated. Their study focuses on this unexplored area of research, carrying out an evaluation of NLMs and their accuracy-efficiency tradeoffs on a Raspberry Pi.

"Our empirical evaluations consider both perplexity as well as energy consumption on a Raspberry Pi, where we demonstrate which methods provide the best perplexity-power consumption operating point," the researchers said. "At one operating point, one of the techniques is able to provide energy savings of 40 percent over the state-of-the-art [methods] with only a 17 percent relative increase in perplexity."

In their study, the researchers also evaluated a number of inference-time pruning techniques on quasi-recurrent neural networks (QRNNs). Extending the usability of existing training-time pruning methods to QRNNs at runtime, they attained several operating points within the accuracy-efficiency trade-off space. To improve performance using a small amount of memory, they suggested training and storing single-rank weight updates at desired operating points.

**More information:** Adaptive pruning of neural language models for mobile devices. arXiv: 1809.10282 [cs.CL]. arxiv.org/abs/1809.10282