Compression of Neural Machine Translation Models via Pruning

Abigail See*, Minh-Thang Luong*, Christopher D. Manning

*equal contribution
Neural Machine Translation
Problem

- Neural Machine Translation models (and neural networks in general) are getting **bigger and bigger**
- **Advantages**: performance improvements!
- **Disadvantages**: over-parameterization leads to large memory requirements and overfitting

This is an obstacle for mobile devices

**How can we reduce over-parameterization?**
Solution

Magnitude-based parameter pruning: delete weights (connections) that are close to zero.

The remaining weights must be retrained to recover performance.

Song Han, Jeff Pool, John Tran, and William Dally. 2015. Learning both weights and connections for efficient neural network. In NIPS.
Main Result

- We can prune **80%** of the weights of a state-of-the-art NMT model, then with some retraining, **surpass** performance of the original model.

- That is, we **compress** the model to **a fifth** of its size with **no performance loss**!
Other Benefits of Pruning

- Pruning acts as a \textit{regularizer}
- Pruning aids the \textit{optimization process}

- The \textit{location} of pruned weights gives insight into the \textit{areas of redundancy} in the NMT architecture.
Come by our poster to learn more!