Compression of Neural Machine Translation Models via Pruning
Abigail See*, Minh-Thang Luong*, Christopher D. Manning

The Problem

- NMT models (and neural networks in general) are getting bigger and bigger.
- Advantages: performance improvements!
- Disadvantages: over-parameterization leads to long running times, large storage size, and overfitting.

How can we reduce over-parameterization?

The Solution

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

Pruning Schemes

The NMT architecture includes several classes of weights (see ‘Our NMT Architecture’).

Should we prune:

- proportionally from each class (class-uniform pruning), or
- in proportion with the standard deviation of each class (class-distribution pruning), or
- without regard to class (class-blind pruning)?

Areas of Redundancy

The location of the pruned weights reveals the areas of redundancy in the network.

Pruning also...

- regularizes the retraining phase.
- aids the optimization process.

Results

- Baseline: state-of-the-art English-German model with 6.1 perplexity and 20.5 BLEU on WMT’14 [1].
- Can prune up to 40% with negligible effect on performance — a sign of redundancy!
- With retraining, can prune 80% and surpass baseline performance!

Conclusion

- Weight pruning is an effective compression method.
- We can make a SOTA model 5 times smaller with slight performance improvement.
- Pruning seems to aid optimization and regularization.
- It also gives insights into areas of redundancy in the NMT architecture.

Citations

[1] Song Han, Jeff Pool, John Tran, and William Dally. 2015b. Learning both weights and connections for efficient neural network. In ICLR.

Acknowledgements

- NSF Award RI-154368
- Bloomberg LP
- NVIDIA Corporation
- DARPA Communicating with Computers (CwC) program under ARO prime contract no. W911NF-15-1-0462

Models via Pruning

How can we reduce over-parameterization?

Pruning Schemes

The NMT architecture includes several classes of weights (see ‘Our NMT Architecture’).

Should we prune:

- proportionally from each class (class-uniform pruning), or
- in proportion with the standard deviation of each class (class-distribution pruning), or
- without regard to class (class-blind pruning)?

Areas of Redundancy

The location of the pruned weights reveals the areas of redundancy in the network.

Pruning also...

- regularizes the retraining phase.
- aids the optimization process.

Results

- Baseline: state-of-the-art English-German model with 6.1 perplexity and 20.5 BLEU on WMT’14 [1].
- Can prune up to 40% with negligible effect on performance — a sign of redundancy!
- With retraining, can prune 80% and surpass baseline performance!

Conclusion

- Weight pruning is an effective compression method.
- We can make a SOTA model 5 times smaller with slight performance improvement.
- Pruning seems to aid optimization and regularization.
- It also gives insights into areas of redundancy in the NMT architecture.

Citations

[1] Song Han, Jeff Pool, John Tran, and William Dally. 2015b. Learning both weights and connections for efficient neural network. In ICLR.

Acknowledgements

- NSF Award RI-154368
- Bloomberg LP
- NVIDIA Corporation
- DARPA Communicating with Computers (CwC) program under ARO prime contract no. W911NF-15-1-0462