Project Scoping Document

Deliverables

- Create a LSTM model for NMT
- Train it, (write a training code and complete training)
- Implement pruning provided in paper, given a trained model
- Experimental report replicating ablation studies
- Implement Optimal Brain Damage(if possible), on trained model. May be not be viable given that computational time of algorithm may not be in line with current models. Size of model may be to large for algorithm, as we need to compute Hessian matrix wrt weights.
- Implement more recent paper (one or two) on pruning out of these:
    - SNIP: Single-shot Network Pruning based on Connection Sensitivity

SNIP: Single-shot Network Pruning based on Connection Sensitivity

Pruning large neural networks while maintaining their performance is often desirable due to the reduced space and time complexity. In existing methods, pruning is done within an iterative...

https://arxiv.org/abs/1810.02340v2

    - Importance of estimation for nn pruning

Importance Estimation for Neural Network Pruning

Structural pruning of neural network parameters reduces computation, energy, and memory transfer costs during inference. We propose a novel method that estimates the contribution of a neuron...

https://arxiv.org/abs/1906.10771v1

    - Picking Winning Tickets Before Training by Preserving Gradient Flow

Picking Winning Tickets Before Training by Preserving Gradient Flow

Overparameterization has been shown to benefit both the optimization and generalization of neural networks, but large networks are resource hungry at both training and test time. Network pruning...

https://arxiv.org/abs/2002.07376

- Pruning filters for efficient ConvNets

Pruning Filters for Efficient ConvNets

The success of CNNs in various applications is accompanied by a significant increase in the computation and parameter storage costs. Recent efforts toward reducing these overheads involve pruning...

https://arxiv.org/abs/1608.08710

-

Neural Pruning via Growing Regularization

Regularization has long been utilized to learn sparsity in deep neural network pruning. However, its role is mainly explored in the small penalty strength regime. In this work, we extend its...

https://arxiv.org/abs/2012.09243

- **All previously mentioned papers are for CNNs, we want to apply similar methods for LSTMs in NMT task.**
- Some older paper on pruning are:
  - G. Augasta, T. Kathirvalavakumar, A Novel Pruning Algorithm for Optimizing Feedforward Neural Network of Classification Problems, Neural Process. Lett. 34 (3), 241–258, 2011
  - H.-J. Xing, B.-G. Hu, Two phase construction of multilayer perceptrons using Information Theory, IEEE T. NeuralNetwor. 20 (4), 715–721, 2009
  - M. Hagiwara, A simple and effective method for removal of hidden units and weights, Neurocomputing, 6, 207–218,1994
- Same ablations as in main paper, but on other 2 additional papers
- Create a website/video with animation to visualize pruning algos (how pruning is performed).

Datasets

We will choose the dataset on the basis of computations available

- WMT'14 English-German data (4.5M sentence pairs)
- IWSLT'15 English-Vietnamese data (133K sentence pairs)

Ablations/Experiments

- Effect on BLUE score by different pruning schemes and percentage of parameters pruned
- Performance of pruned models after pruning, or after pruning and retraining
- Breakdown of perplexity increase by weight class

Timeline

- Till 20th October
  - Read the papers with which we are going to compare
- Till 30th October(checkpoint 2)
  - Create a LSTM model for NMT
  - Train It
  - Implementing pruning provided in the original paper
- Till 15 Nov
  - Experimental report replicating ablation studies in the given paper
  - Start working on comparisons
- Till 30 Nov (checkpoint 3)
  - Perform ablations with other two papers
  - Form a visualisation for pruning