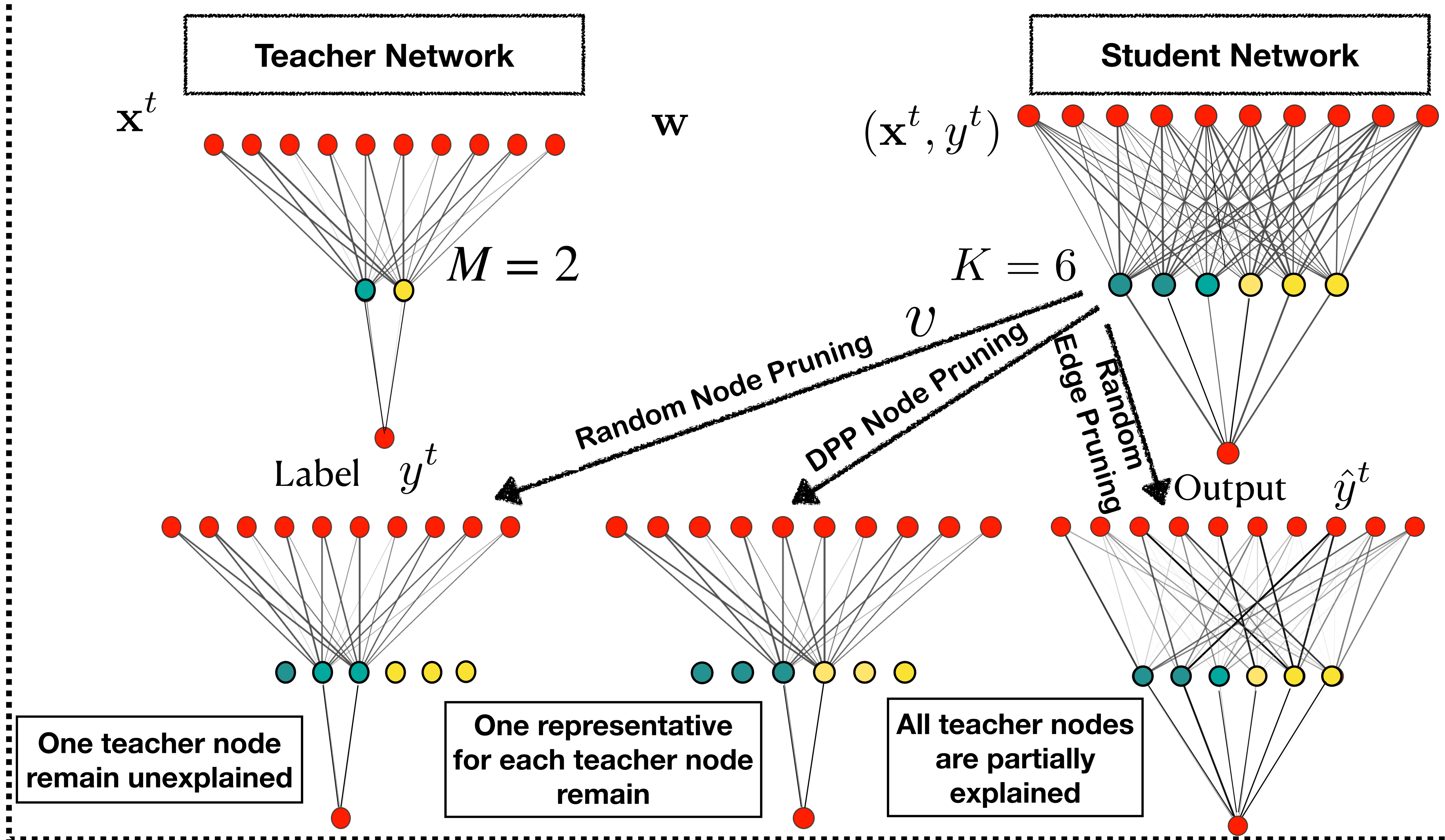
Understanding Diversity Based Neural Network Pruning in Teacher Student Setup

Rupam Acharyya[1], Ankani Chattoraj*[2], Boyu Zhang*[3], Shouman Das[4], Daniel Stefankovic[3]
Mathematics Department, University at Buffalo[1],
Department of Brain & Cognitive Science, University of Rochester[2],
Department of Computer Science, University of Rochester[3],
Department of Mathematics, University of Rochester[4]

Introduction

- **Neural Network Pruning:** Given a large trained neural network, how to reduce the size of the network without degrading its performance much?
- **Motivation:** Currently the pre-trained networks (e.g. BERT) have billions of parameters. Pruning can help reducing the time complexity (of fine tuning) and space complexity.
- **Limitation:** Lots of pruning methods available, but why do they work?
- **This Work:** Takes a step towards explaining pruning performance.

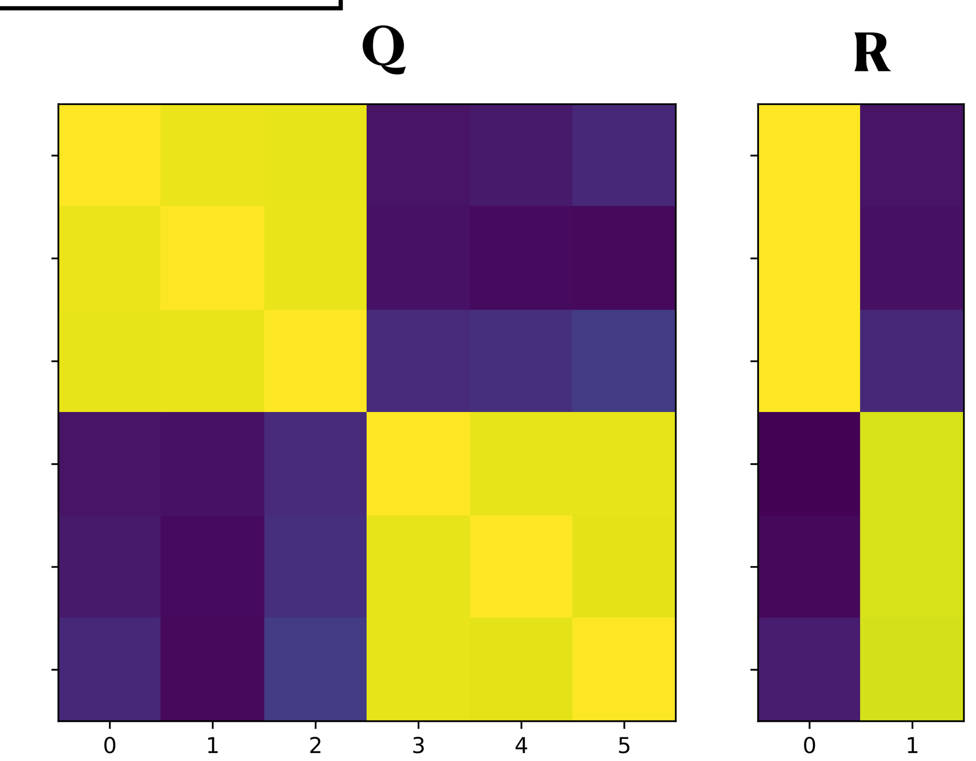
Neural Network pruning in Teacher Student Setup



Previous Work

Generalization Error in Teacher Student Setup:

- Generalization Error (GE): Expected error on the unseen test dataset - a measure of performance of NN.
  - Lower the error better the model.
  - Difficult to compute in general as test data distribution is unknown.



- GE in teacher student setup can be written as function of macroscopic order parameters:
  - Correlation between student hidden nodes (Q).
  - Correlation between student and teacher hidden nodes (R).

Determinantal Point Process (DPP): DPP is a probability distribution to sample diverse subsets of a ground set.

DPP Node Pruning: Sample a subset of nodes for each layer using the DPP defined by the kernel matrix defined as above. Later some *re-weighting* of the edges is needed to compensate for the lost nodes (can be done efficiently).

Result

Comparison between DPP node pruning and Random node pruning:

Theorem: For $k_n \leq M$ we have,
$\mathbb{E}_f\left[\epsilon_{k_n}^{Rand\ Node}(f)\right] \geq \epsilon_{k_n}^{DPP\ Node}(f)$ and $\mathbb{E}_f\left[\hat{\epsilon}_{k_n}^{Rand\ Node}(f)\right] \geq \hat{\epsilon}_{k_n}^{DPP\ Node}(f)$
and, $\epsilon_{k_n}^{Imp\ Node}(f) \geq \hat{\epsilon}_{k_n}^{DPP\ Node}(f)$, i.e., DPP node pruning outperforms random node pruning in the above setup. Here the expectation is taken over the the subsets of hidden nodes of size k_n chosen u.a.r

Comparison between DPP node pruning and Random edge pruning:

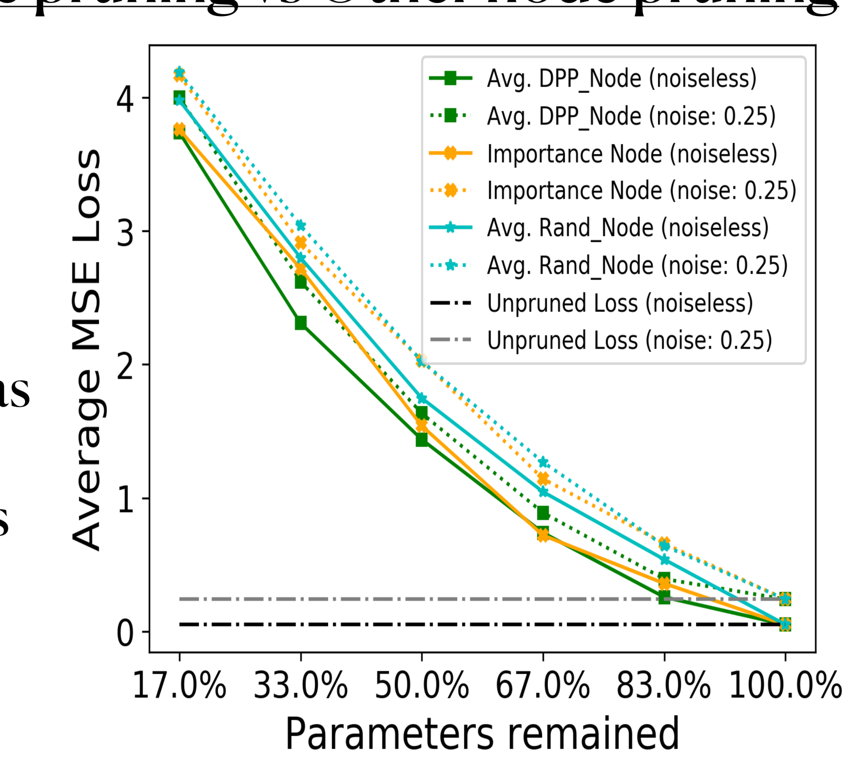
Theorem: Let k_n and c satisfy the equation below, and $0 \leq c \leq \frac{1}{Z}$ and $Z(=\frac{K}{M}) \geq 4$. Then $\epsilon_{k_n}^{DPP\ Node}(f) \geq \epsilon_c^{Rand\ Edge}\left(\mathbb{E}\left[f\right]\right)$,
i.e., Random edge pruning outperforms DPP node pruning in the above setup.

$$\text{Node Edge Ratio}: \frac{k_n}{K} = \lim_{N\to\infty} \frac{k_e}{N} = c$$

Simulations

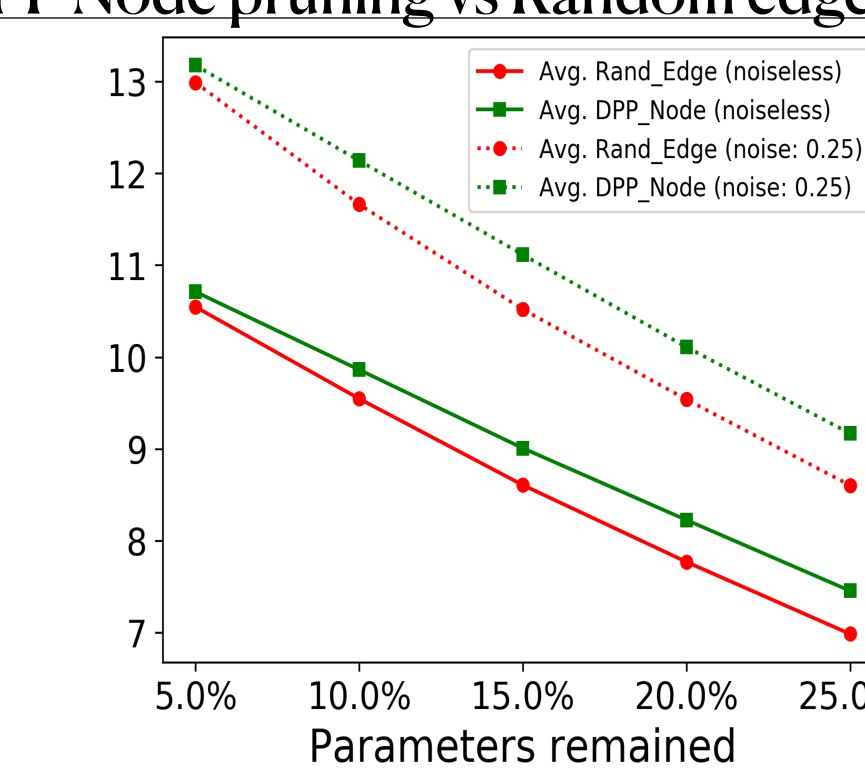
DPP node pruning vs Other node pruning

- Data:
  - Sampled the 800000 i.i.d input samples from $\mathcal{N}(0,1)$ as training data and 80000 as testing data.



- $M = 2, K = 6, N = 500$, and $v^* = 4$

DPP Node pruning vs Random edge pruning



- $K = 5, M = 20, N = 500$, and $v^* = 4$
- Node-to-edge ratio: [1:83, 2:166, 3:250, 4:333, 5:417, 6:500]

Conclusion & Future Work

- Compared different pruning methods in Teacher Student framework - first theoretical comparison.
  - DPP node pruning vs Random and Importance node pruning.
  - Random edge pruning vs DPP node pruning.
- Extend for feed-forward networks with more than two layers and in other neural network architectures.

Reference: 1) Zelda Mariet and Suvrit Sra. *Diversity networks: Neural network compression using determinantal point process.*
2) ebastian Goldt, Madhu Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborov *Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup.*