Coreset based Data-Independent pruning of Graph Convolutional Networks

TEXAS A&M UNIVERSITY

Presentation by
Anmol Anand
November 29, 2023

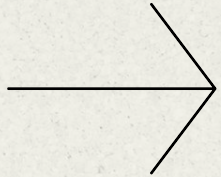
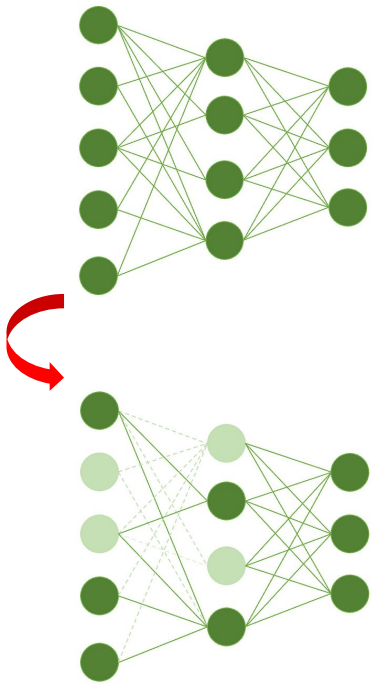


TABLE OF CONTENTS

Background - Coreset based Neural Pruning

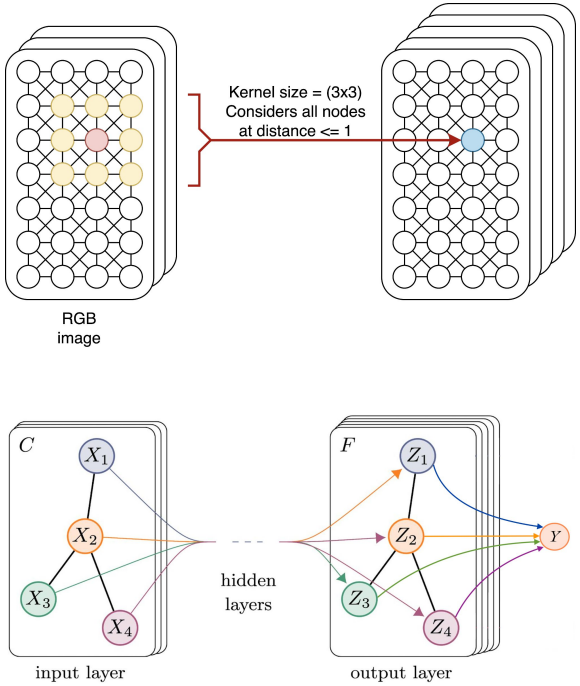
- **Coresets**:
  - A subset of the original dataset that approximately preserves the properties of the full dataset.

- **Coreset based Neural Pruning:**
  - Aimed at retaining a smaller, representative set of neurons that maintain network performance while reducing redundancy.



Background - Graph Convolutional Networks

- Input: Structured graph data (Examples: citation networks, knowledge graphs, etc)
  - NODE FEATURES
  - EDGE FEATURES
- Output:
  - GRAPH CLASSIFICATION (IN THIS PAPER)
  - NODE CLASSIFICATION
  - EDGE CLASSIFICATION



IMDb Graph Datasets

- Samples are ego-network graphs (every node represents a person).
  - Nodes represent actors with features such as:
    - Age
    - Number of movies
    - Awards won
    - Number of interviews
    - Popularity score
  - Edges are between actors who have appeared in same movies.

IMDb-BINARY

Classify IMDb graphs into 2 genres:

- Action
- Romance

IMDb-MULTI

Classify IMDb graphs into 3 genres:

- Action
- Romance
- Sci-Fi

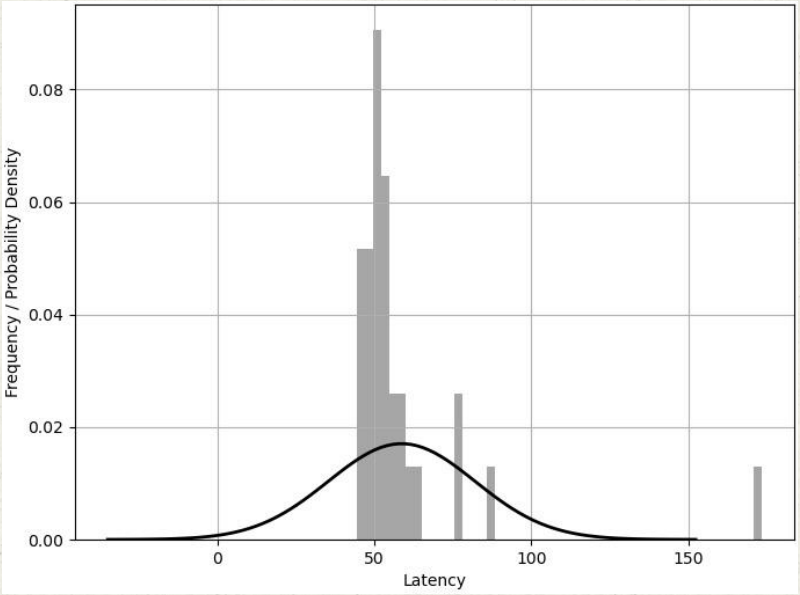
GCN Architecture

Layer	Original Number of Neurons	Number of Neurons after Compression	Compression Factor
Graph Convolution Layer 1	512	512	
Graph Convolution Layer 2	512	256	2
Fully Connected Layer 1	1024	512	2
Fully Connected Layer 2	Number of classes	Number of classes	

- Compression factor 2 on hidden layers.

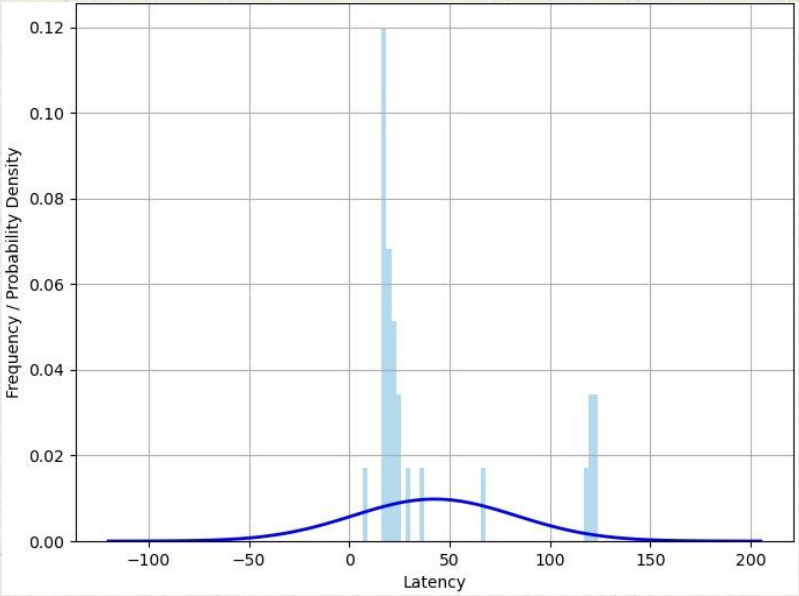
Results: IMDb-BINARY

Average latency to reach the saturation accuracy of 75% while using the
Uncompressed Network: 58.67 seconds



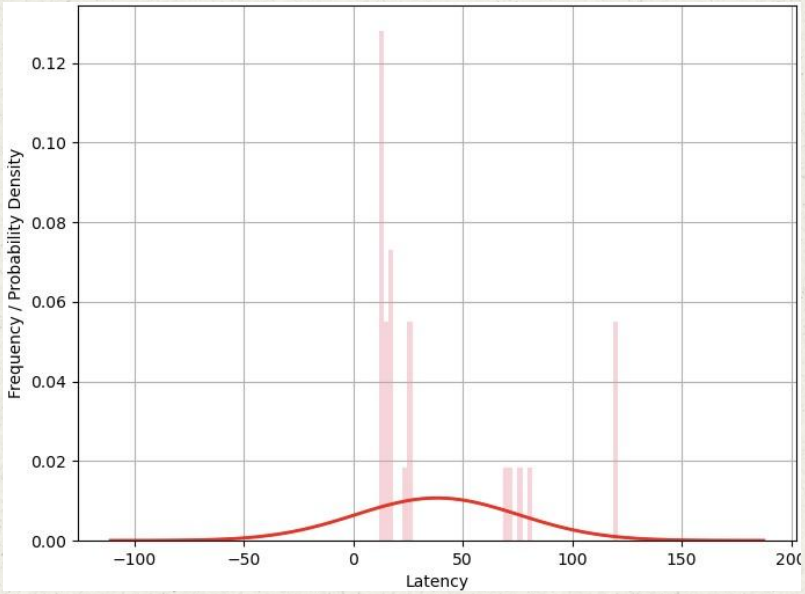
Results: IMDb-BINARY

Average latency to reach the saturation accuracy of 75% while using the
Uniformly Pruned Network: 42.53 seconds



Results: IMDb-BINARY

Average latency to reach the saturation accuracy of 75% while using the **Coreset Pruned Network: 38.21 seconds**



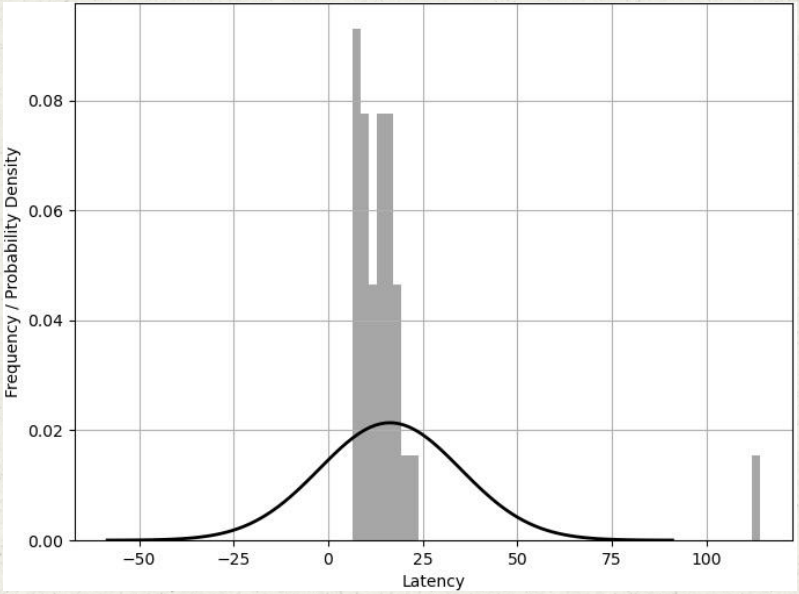
Results: IMDb-BINARY

Compression Type	Expected latency (seconds) to reach saturation accuracy of **75%**
None	58.67
Uniformly Pruning	42.53
Coreset Pruning	38.21

- Coreset Pruning took **35%** less time than the Uncompressed network.
- Whereas, Uniform Pruning took **28%** less time than the Uncompressed network.

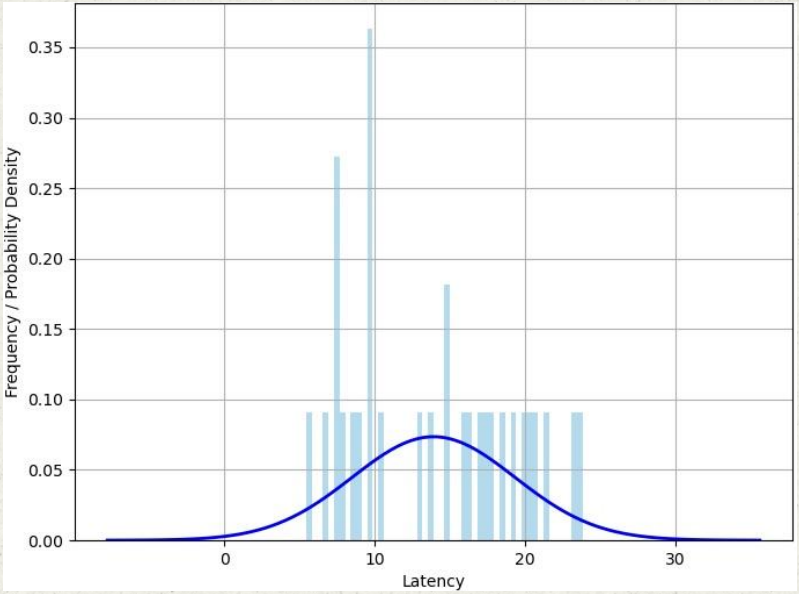
Results: IMDb-MULTI

Average latency to reach the saturation accuracy of 56% while using the
Uncompressed Network: 16.31 seconds



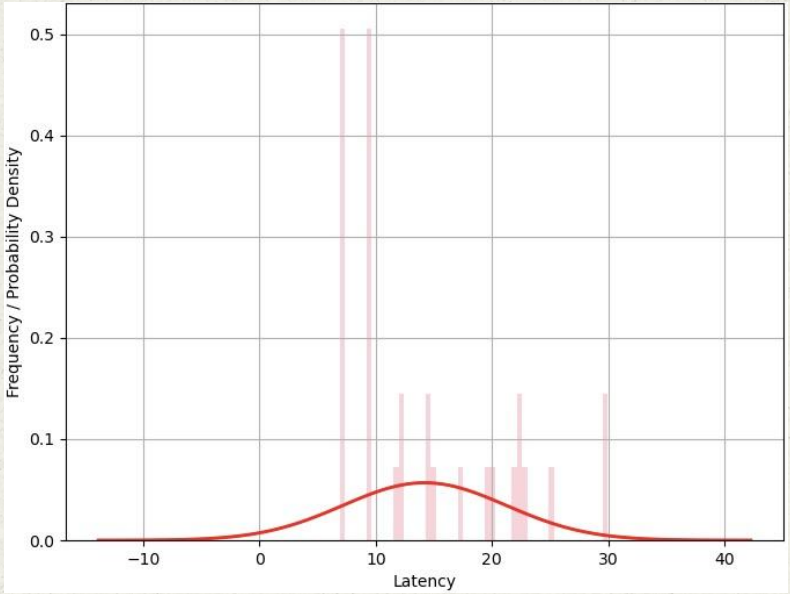
Results: IMDb-MULTI

Average latency to reach the saturation accuracy of 56% while using the **Uniformly Pruned Network: 13.92 seconds**



Results: IMDb-MULTI

Average latency to reach the saturation accuracy of 56% while using the **Coreset Pruned Network: 14.19 seconds**



Results: IMDb-MULTI

Compression Type	Expected latency (seconds) to reach saturation accuracy of **56%**
None	16.31
Uniformly Pruning	13.92
Coreset Pruning	14.19

- Coreset Pruning took **13%** less time than the Uncompressed network.
- Whereas, Uniform Pruning took **15%** less time than the Uncompressed network.

Observations

- **Coreset based pruning** improves performance and reaches saturation accuracy faster than the **unpruned network**.
- **Coreset based pruning** outperformed **Uniform pruning** for IMDb-Binary but not for IMDb-Multi.
- Possible **reasons** why coreset based pruning did not consistently outperform uniform pruning for Graph Convolutional Networks:
    - Using the **same coreset computation algorithm** for GCN layers as FC layers might be a bad idea.
    - **Same network size** for Uniform and Coreset pruned networks. Only differ in the initial network weights.
      Not necessary that initial weights in the coreset based approach will result in faster training.

THANKS!

DO YOU HAVE ANY QUESTIONS?


aanand@tamu.edu


aanand.dev