NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

School of Computer Science and Engineering
College of Engineering

Running CNN efficiently on a FPGA

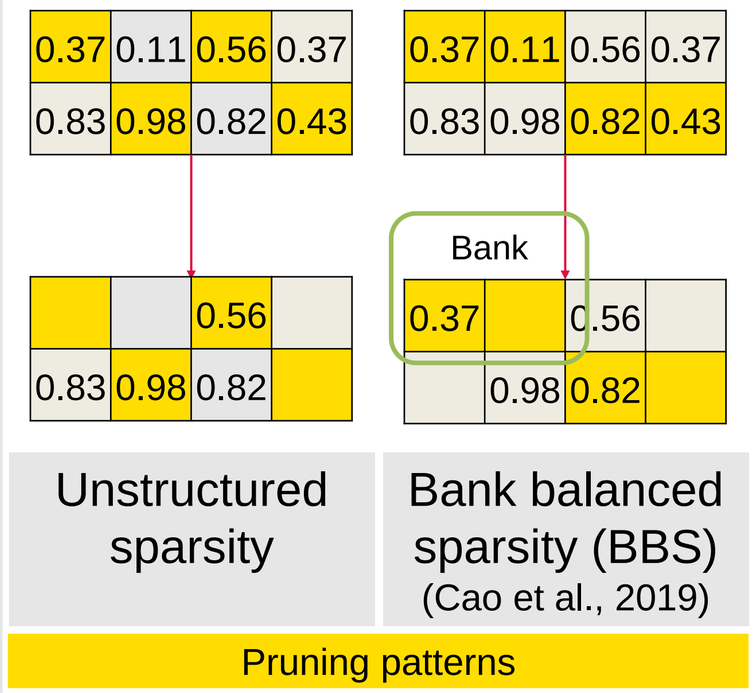
Student: Yang Shenghao

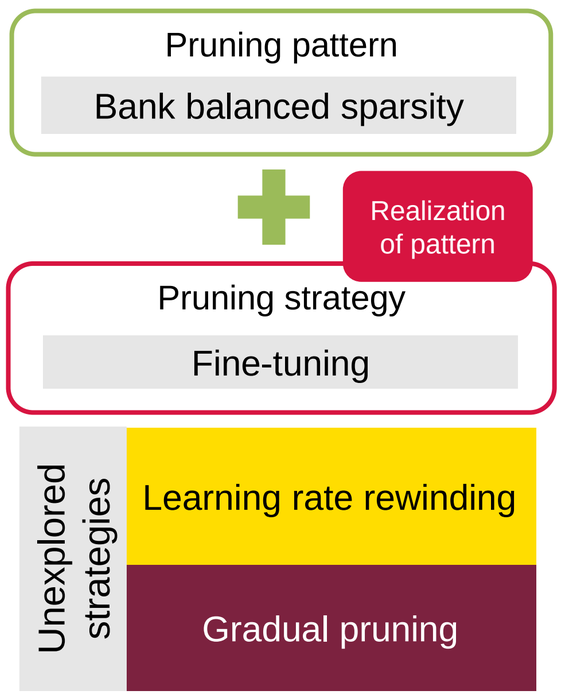
Supervisor: Dr. Weichen Liu

Motivation

Edge devices typically use accelerators to run deep learning models due to their limited computational power. Being both reconfigurable and power efficient, **FPGAs** are a prime candidate.

However, since the at which neural networks are growing is surpassing improvements in accelerator performance, **model compression** methods are being intensely researched.





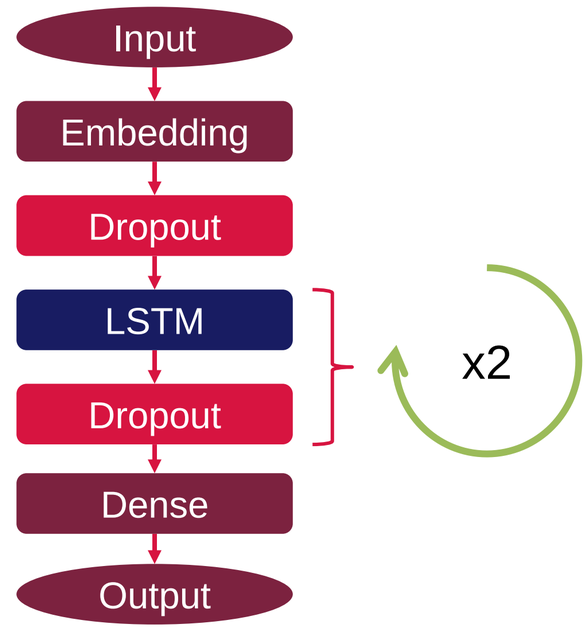
An influential FPGA model compression technique is magnitude pruning using the **BBS structured pruning** pattern – as it allows for more efficient accelerator designs than the unstructured pattern.

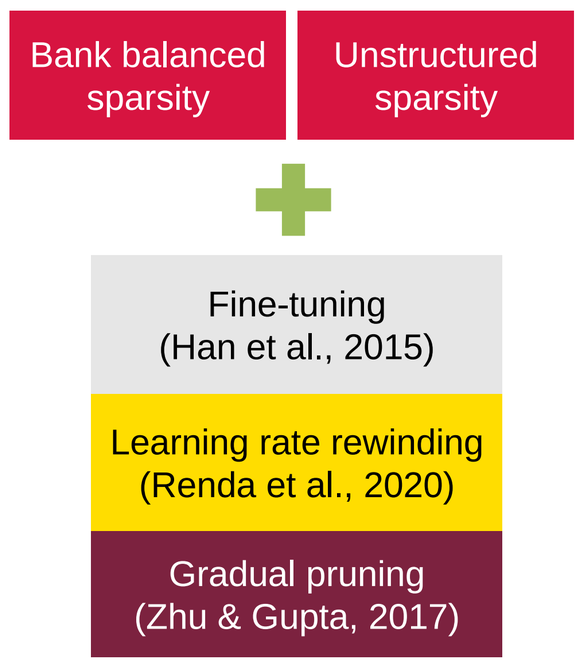
As the original paper only implemented the **fine-tuning** pruning strategy, this project aimed to evaluate BBS' performance on **other prominent strategies**.

Methodology

For evaluation, a LSTM language model equivalent to that used in (Zhu & Gupta, 2017) and (Cao et al. 2019), was used. This model predicts the next words in a sentence given the previous elements.

To evaluate its accuracy, the perplexity ($\exp(loss)$) metric was used. A **lower** perplexity score indicates a model with greater predictive power.



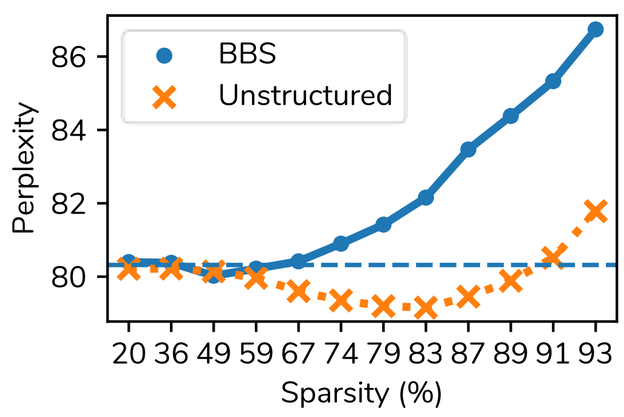


The LSTM model was implemented using Keras and pruning was performed using the TensorFlow Model Optimization Toolkit, with an implementation of BBS added for this project.

Its accuracy was then evaluated on all combinations of two pruning patterns and three pruning strategies, with the perplexity score on the test set used for comparisons.

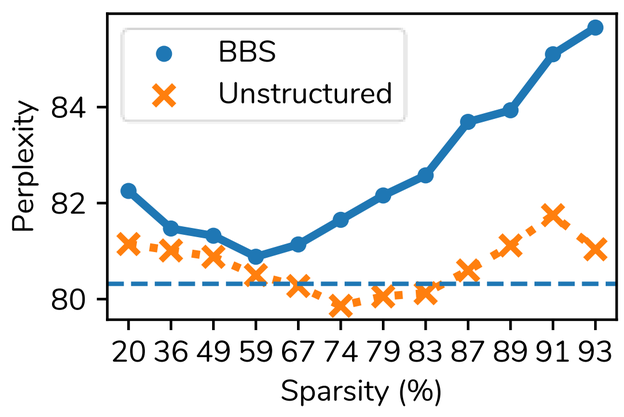
Results

Fine-tuning



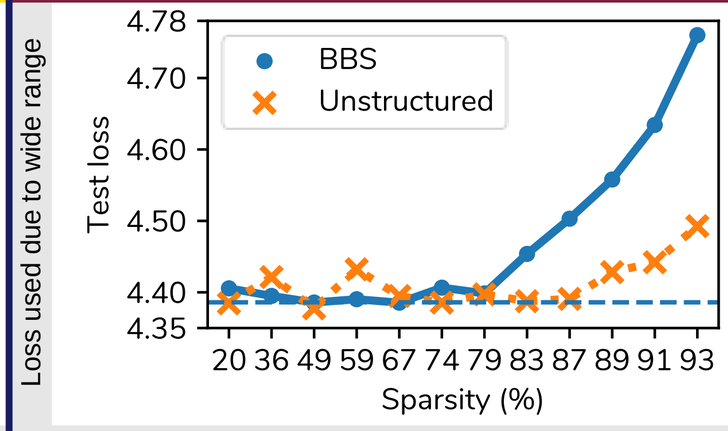
High sparsity: Unstructured pruning yields **greater accuracy**
Medium sparsity: **Equivalent** accuracy

Learning rate rewinding



All sparsity levels: Unstructured pruning **is more accurate** bank balanced sparsity

Gradual pruning



High sparsity: Unstructured pruning yields **greater accuracy**
Medium sparsity: **Equivalent** accuracy

Conclusion

BBS vs Unstructured sparsity

Overall	High sparsity	Unstructured pattern achieves **greater accuracy**
	Low-Medium sparsity	Bank Balanced pattern offers **equivalent accuracy**

Use **BBS** for **low-medium sparsity levels** to exploit hardware efficiency

Use **Unstructured sparsity** for high sparsity levels where storage space is a concern

Comparison between strategies

	Accuracy	Development Speed	Ease Of Use
Fine tuning		Gradual pruning	Learning rate rewinding
Gradual pruning		Learning rate rewinding	Fine tuning Gradual pruning
Learning rate rewinding		Fine tuning (Requires retraining)	(Additional Hyper-parameters)

Recommendations

Given the **differing performance** of BBS **across different strategies**

Pruning patterns should be characterized on a **wider range of pruning strategies** to enable more robust comparisons