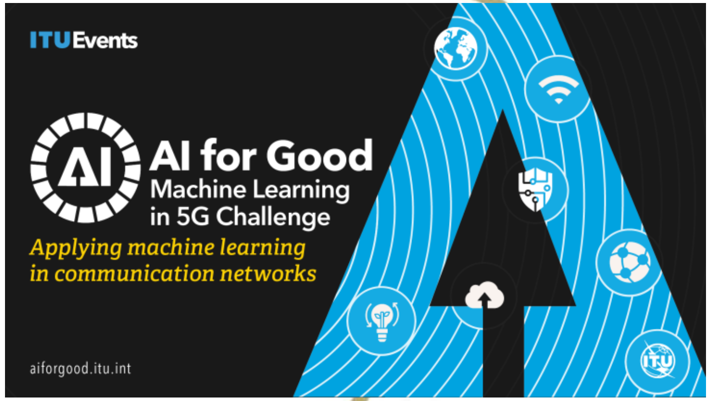


Feedback Magnitude Pruning for Modulation Classification

The A(MC) Team:
Jakob Krzyston
PhD Student @ GT, Research Engineer, GTRI
jakobk@gatech.edu
Dr. Rajib Bhattacharjea
Principal Engineer, DeepSig Inc
raj@deepsig.ai
Dr. Andrew Stark
Senior Research Engineer, GTRI
andy.stark@gtri.gatech.edu

Georgia Tech Research Institute

Background

- Ever increasing demand for wireless data
  - Improved spectrum sensing improves spectral allocation

- Automatic Modulation Classification (AMC)
  - Traditionally methods were statistical/expert-derived

Georgia Tech Research Institute

Background

- Ever increasing demand for wireless data
  - Improved spectrum sensing improves spectral allocation

- Automatic Modulation Classification (AMC)
  - Traditionally methods were statistical/expert-derived
  - Recently, neural networks outperform traditional methods

Background

- Ever increasing demand for wireless data
  - Improved spectrum sensing improves spectral allocation

- Automatic Modulation Classification (AMC)
  - Traditionally methods were statistical/expert-derived
  - Recently, neural networks outperform traditional methods

- Deep learning has shown to yield improved results, but:
  - Memory intensive
  - Computationally expensive
  - Are not developed for specialized hardware (FPGAs)

CLASSIFICATION

Problem Statement

- Develop efficient deep learning approaches for AMC
  - ≥ 56% Accurate on RadioML 2018 dataset (DeepSig)
  - Neural network compression methods:
    - Efficient architectures
    - Quantization
    - Pruning
    - Custom training paradigms

Georgia Tech Research Institute

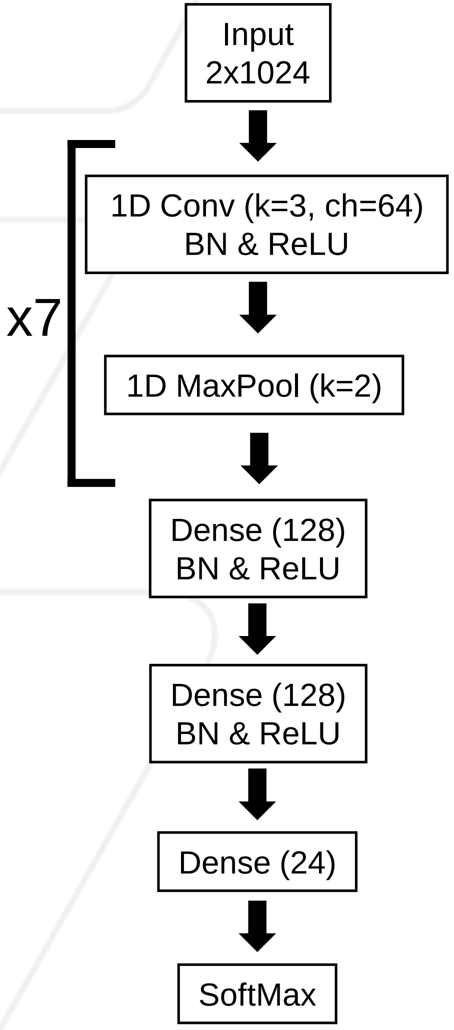
Possible Approaches

- ~~**Modify the provided architecture [Speed]**~~
- **Reduce the quantization with Brevitas [Speed]**
  - Four bits for both weights and activations
- **Prune weights [Speed]**
  - L1 unstructured Iterative Magnitude Pruning (IMP)
  - Prune when accuracy threshold reached
- **Adjust training paradigm [Accuracy]**
  - Learning Rate Scheduler → Reduce LR on Plateau

Georgia Tech Research Institute

Methods

Architecture



IMP (Simplified)

```
for number of pruning epochs do
    for number of training epochs do
        Train model;
        Evaluate model;
        if model accuracy> 56% then
            Save model;
            Prune 20% of the weights;
            Break
        end
    end
end
```

Algorithm 1: IMP with Accuracy Criterion

Compression Summary

Quantity	Original	Final
Bit Ops	807,699,904	24,436,576
Weight Bits	1,244,936	68,072
Compression	1x	9.313x
Sparsity	0%	89.26%

Notes

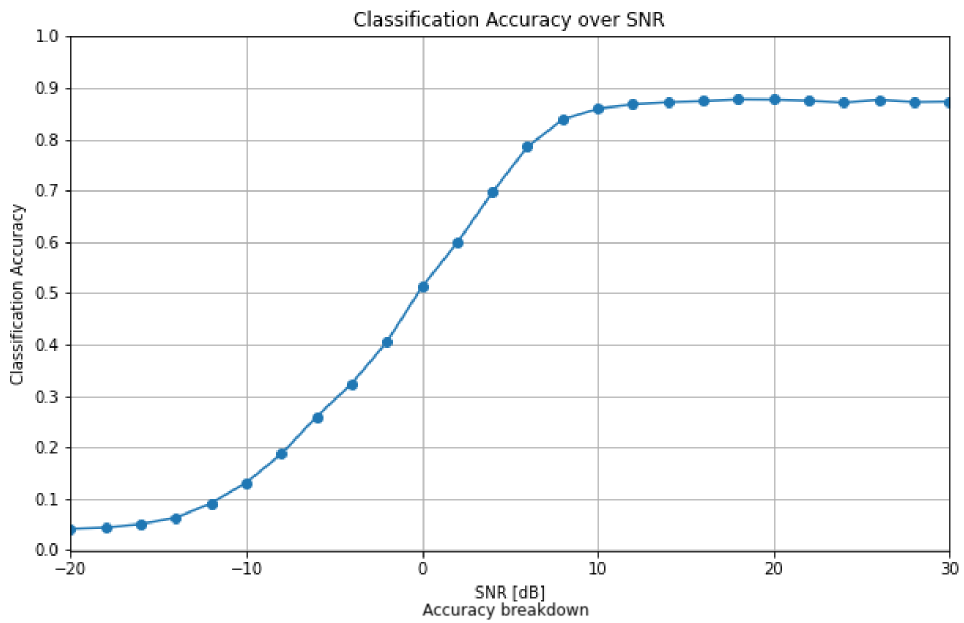
- Sparsity % = 1 – (0.8 ^ 10)
- Compression = 1 / (0.8 ^ 10)

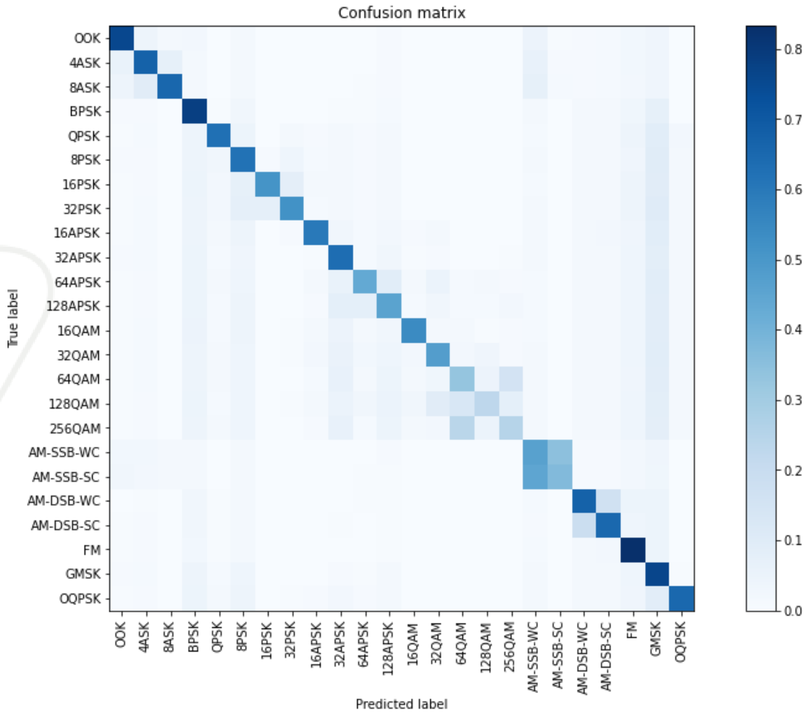
Georgia Tech Research Institute

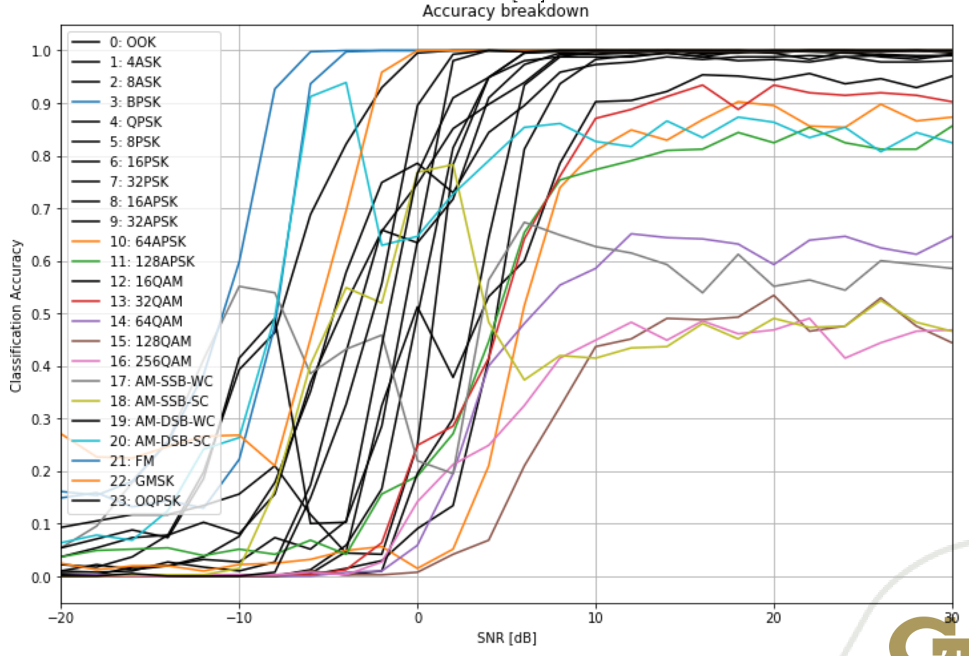
Submission Results

- **Inference Cost Score:**
  - 0.042467

- **Overall Test Accuracy:**
  - 0.5625





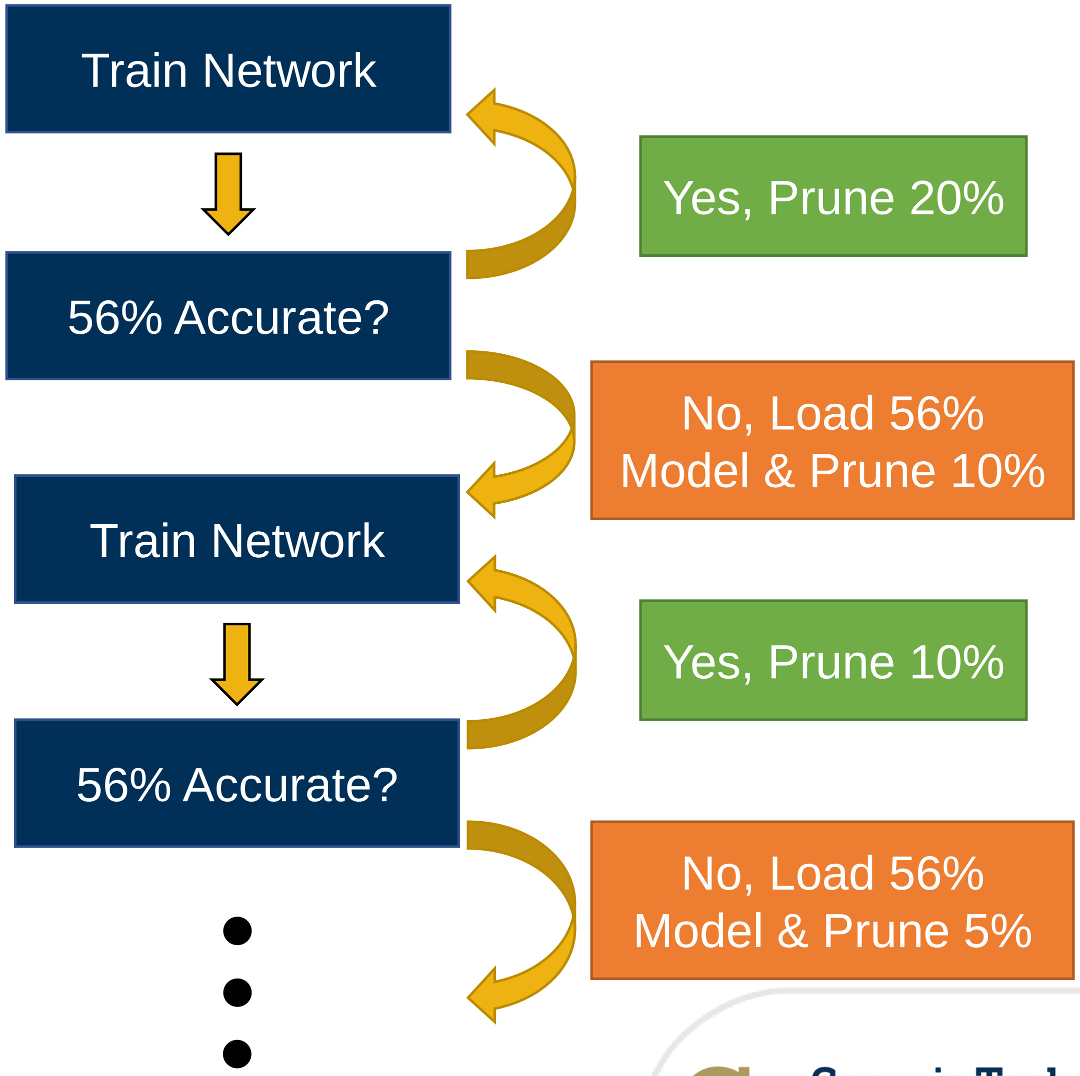


Georgia Tech Research Institute

Feedback Magnitude Pruning

- Feedback Magnitude Pruning (FMP)
  - IMP with pruning rate adjustment
  - Feedback mechanism dictates change
  - Similar to learning rate scheduling

- In this work:
  - L1 Unstructured pruning
  - Pune % = 20
  - Decay factor = 2
  - Stop at 5% pruning threshold



FMP Results

- Variables
  - Pruning method
    - None
    - IMP
    - FMP
  - Bits

Bits	Compression Ratio	Cost	Pruning
8	1	1	None
8	5,821.527	0.0583	FMP
7	5,821.527	0.0482	FMP
6	3,072.602	0.0465	FMP
5	1,621.719	0.0434	FMP
4	9.313	0.0424	IMP
4	813.147	0.0419	FMP

FMP Results

- Variables
  - Pruning method
    - None
    - IMP
    - FMP
  - Bits

Bits	Compression Ratio	Cost	Pruning
8	1	1	None
8	5,821.527	0.0583	FMP
7	5,821.527	0.0482	FMP
6	3,072.602	0.0465	FMP
5	1,621.719	0.0434	FMP
4	9.313	0.0424	IMP
4	813.147	0.0419	FMP

FMP Results

- Variables
  - Pruning method
    - None
    - IMP
    - FMP
  - Bits

Bits	Compression Ratio	Cost	Pruning
8	1	1	None
8	5,821.527	0.0583	FMP
7	5,821.527	0.0482	FMP
6	3,072.602	0.0465	FMP
5	1,621.719	0.0434	FMP
4	9.313	0.0424	IMP
4	813.147	0.0419	FMP

FMP Results

- Variables
  - Pruning method
    - None
    - IMP
    - FMP
  - Bits

Bits	Compression Ratio	Cost	Pruning
8	1	1	None
8	5,821.527	0.0583	FMP
7	5,821.527	0.0482	FMP
6	3,072.602	0.0465	FMP
5	1,621.719	0.0434	FMP
4	9.313	0.0424	IMP
4	813.147	0.0419	FMP

FMP Results

- Variables
  - Pruning method
    - None
    - IMP
    - FMP
  - Bits

Bits	Compression Ratio	Cost	Pruning
8	1	1	None
8	5,821.527	0.0583	FMP
7	5,821.527	0.0482	FMP
6	3,072.602	0.0465	FMP
5	1,621.719	0.0434	FMP
4	9.313	0.0424	IMP
4	813.147	0.0419	FMP

Georgia Tech Research Institute

Conclusion

- We demonstrate the effectiveness of integrating feedback into IMP
  - **Feedback Magnitude Pruning (FMP)**

- FMP compresses networks further than IMP
  - 813x vs 9.313x

- Inference cost of FMP and four bit quantization = **0.0419**

Georgia Tech Research Institute

CLASSIFICATION