BacalhauNet: A Tiny CNN for Lightning-Fast Modulation Classification

BacalhauNet Team

José Nuno Grácio Rosa[1], Guilherme Carvalho[2], Daniel Granhão[2], Tiago Filipe Gonçalves[2]

[1] Escola Superior de Tecnologia e Gestão - Politécnico de Leiria

[2] INESC TEC and Faculty of Engineering - University of Porto

Outline

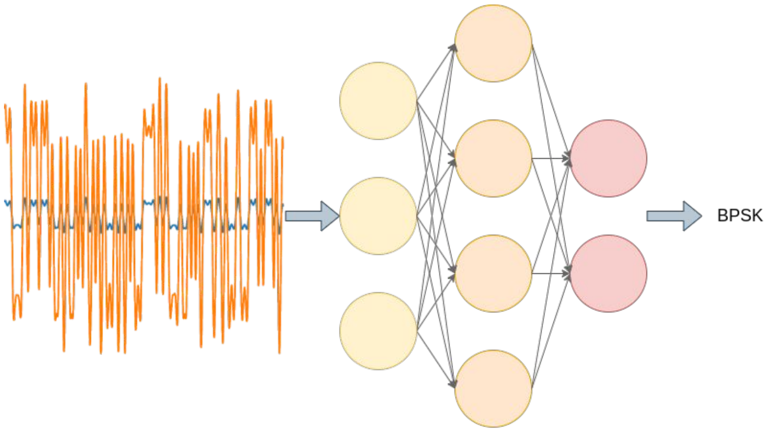
1. Introduction
2. BacalhauNet: A Tiny CNN for Lightning-Fast Modulation Classification
3. Conclusions and Future Work

Introduction

Deep Learning for Improved Radio Efficiency

- There is a **need for improved radio efficiency**
  - Improved spectral allocation can be attained via high quality spectrum sensing and adaptation
  - Solutions such as Dynamic Spectrum Access (DSA) and Cognitive Radio (CR) require Automatic Modulation Classification (AMC)

- Deep Learning has been shown to be competitive for AMC, **but is computationally expensive!**
  - There is a need to **compress** neural networks so that **low latency** and **high throughput** requirements can be met



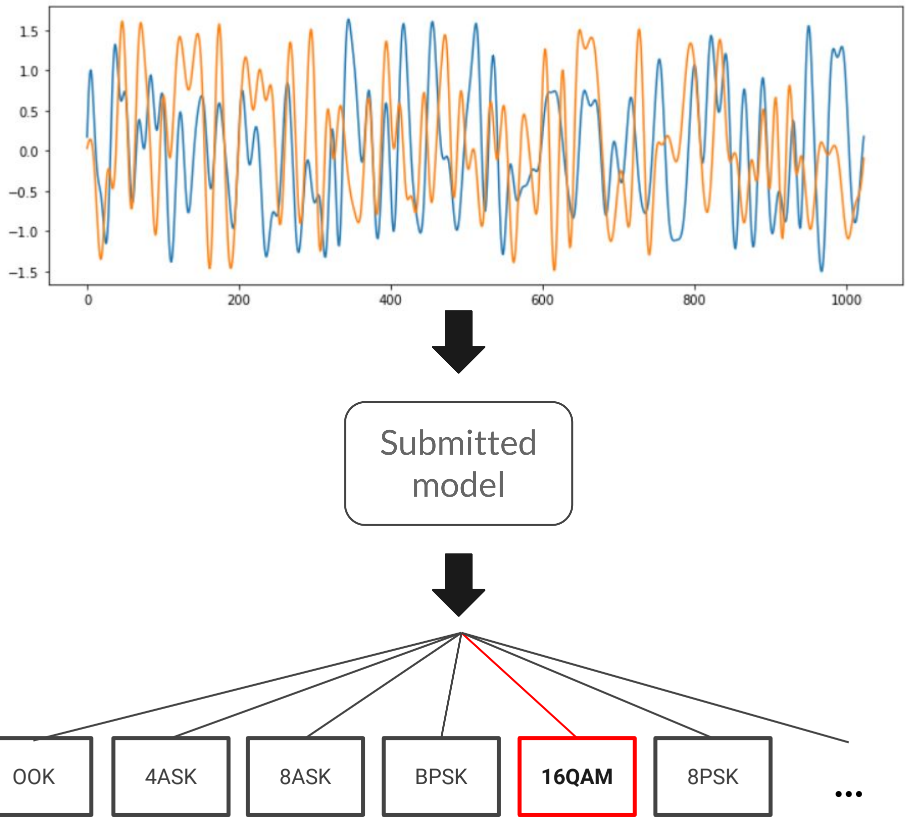
Neural Networks used for Automatic Modulation Classification

The Challenge

- **Main Goal:** design a neural network that achieves **accuracy >= 56%** on dataset **RadioML 2018.01A** while minimizing **inference cost** (the evaluation metric)

- The inference cost is related to the number of parameters of the network as well as the number of required operations

Inference Cost Score = Submission Inference cost / Baseline Inference Cost

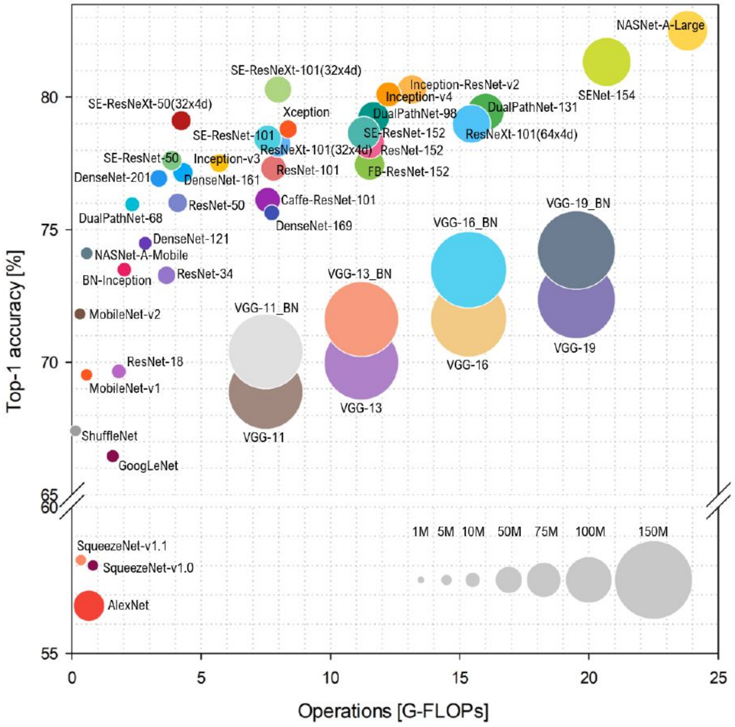


Overview of the Challenge

BacalhauNet: A Tiny CNN for Lightning-Fast Modulation Classification

Starting Point

- Baseline architecture based on a VGG architecture
- Tests using a MobileNetv3-Small based architecture were not satisfying (inference cost score ≈ 7.25)
- **So we built our own network** based on well known blocks

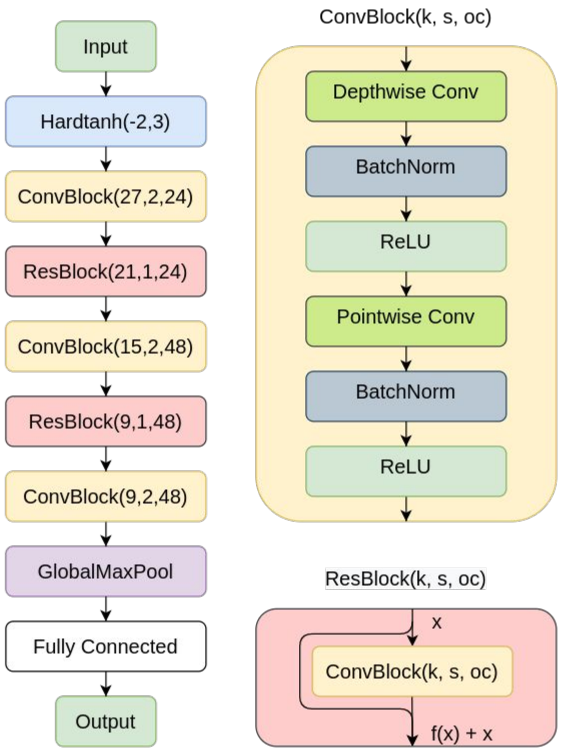


Top-1 Accuracy vs. Computational Complexity of Various Models

BacalhauNet: Proposed Architecture

- BacalhauNet is built on top of commonly used structures:
    - **Depthwise separable convolutions**
    - **Residual connections**

- We used a **design space exploration** approach to find the required number of layers and to optimize the parameters of each layer:
    - **Kernel** size (k)
    - **Stride** length (s)
    - Number of **output channels** (oc)

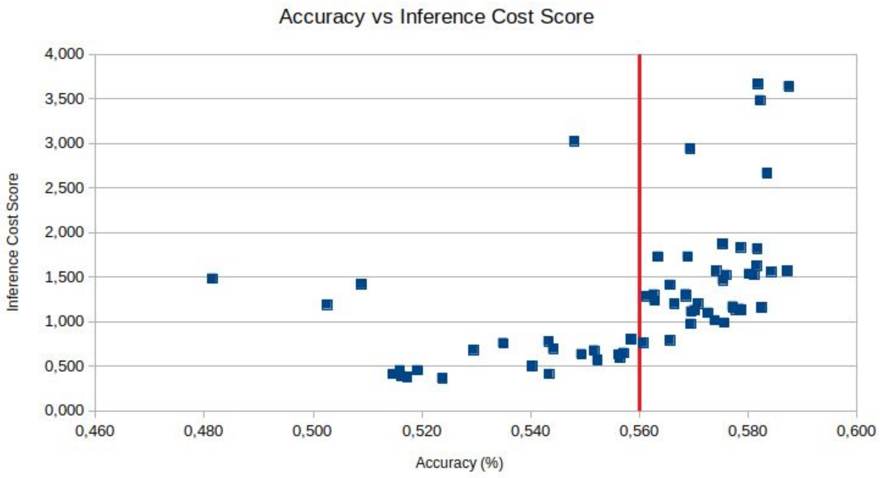


BacalhauNet Architecture

BacalhauNet: Results

- Models with **lower inference cost score were found but discarded** since the accuracy didn't allow us to compress the model as much or didn't reach the 56% threshold
- Selected model:
  - Accuracy ≈ 59%
  - Inference cost score ≈ 1.42
- Selected model compressed **using the baseline method** (quantization to 8 bits):
  - **Accuracy ≈ 59%**
  - **Inference cost score ≈ 0.146**

  ➡ **≈ 6.85x smaller than baseline**



Quantization

- BacalhauNetV1 with **floating-point (FP)** inputs, weights and activations achieved a **good initial inference cost score**

- **Inputs** were quantized to **8 bits**, while **weights and activations** were iteratively quantized down from **8 to 5 bits**

- **Bit-width of 6** selected due to it being a **good compromise between accuracy and inference cost**

Architecture	Bit Width	Test Accuracy Reached	Inference Cost Score
VGG (Baseline)	8 bit	59.47%	1.000
BacalhauNet (Ours)	FP	59.09%	1.416
	8 bit	59.06%	0.146
	7 bit	58.35%	0.100
	6 bit	**58.67%**	**0.078**
	5 bit	55.89%	0.056

Quantization Results

Pruning

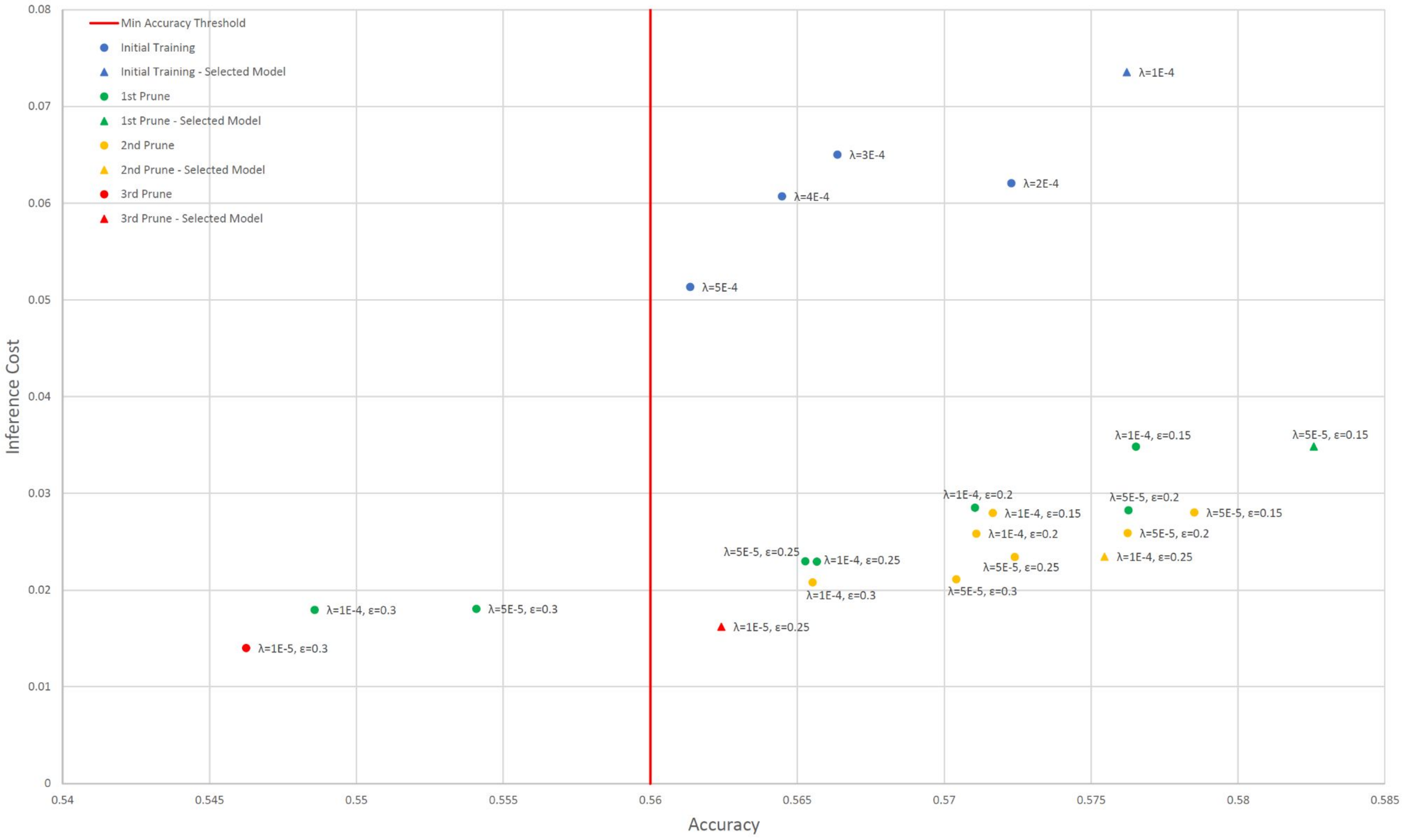
- Both structured and unstructured pruning methods were tested
    - **Unstructured pruning** was selected as that is the method **most favored** by the challenge evaluation method

- **Several iterations of sparsity inducing training and pruning** were performed

- Exploration of 2 variables:
    - **Weight Decay**
    - **Minimum Weight Absolute Value**

Step	Weight Decay	Min. Weight Abs. Value	Accuracy	Inference Cost Score
Original	0.0001	-	57.62 %	0.0735
1st prune	0.00005	0.15	58.26 %	0.0348
2nd prune	0.0001	0.25	57.55 %	0.0235
3rd prune	0.00005	0.25	56.24 %	0.0162

Pruning Results

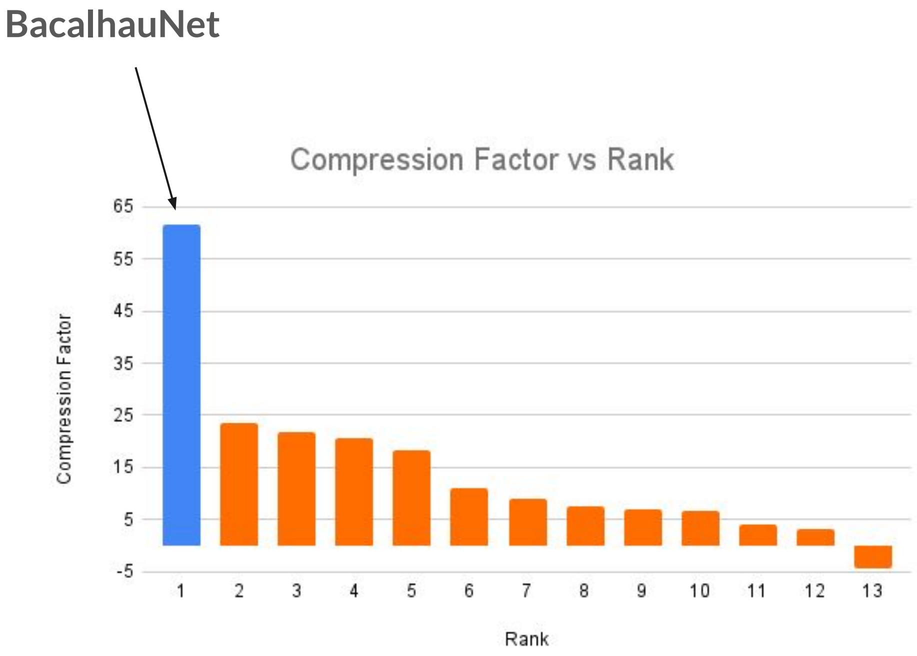


Pruning Design Space Exploration

Conclusions

Conclusions

- Our submission achieves an **inference cost score of 0.0162** (**≈ 61.73x compression**)
- This **enables the implementation** of the proposed neural network in **resource constrained devices**
- **FINN** can be used to **deploy our model onto an FPGA**



PS007 Challenge Results

Thank You!

BacalhauNet Team

POLITÉCNICO DE LEIRIA

INESCTEC

U.PORTO

FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO