Introduction

The neural network models are used for large data driven applications. For hardware applications, the neural network models are computationally intensive. The models require large memory space. Different methods are developed in the recent times to compress the models without affecting its performance.

In this project, quantization, pruning and clustering are used as optimization methods to compress a neural network model.

The objective of this project is to explore various optimization methods for the Automatic Modulation Classification (AMC) model and to achieve an optimized model with less memory requirement and less training time without affecting the performance.

Automatic Modulation Classification is the technique used to automatically determine the modulation scheme of a received signal accurately. This has an important role in military applications such as in Cognitive Radios. In this project a deep learning model which is implemented for AMC is taken and optimized using various optimization techniques available in the TensorFlow library package.

Methodology

Base Model

The base model used for AMC is a Deep Learning model using CNN.

The model has two convolutional layers, which is followed by a dense layer of 256 neurons and activation function Relu. There is one more dense layer which has 11 neurons with activation function SoftMax for classifying the modulation schemes.

![Fig: CNN Model used for AMC](image)

1. Quantization

Quantization is the method used to optimize a model by converting the model parameters from float32 to lower bit representation. This process reduces the model size and latency. By reducing the computational complexity

2. Pruning

The connections with small weights are pruned. For this a threshold value is set, and the weights that come under the value of threshold is removed. In this method the number of parameters in a model can be reduced which in turn can reduce the overall size of the model.

![Fig: Pruning a neural network model](image)

3. Clustering

In Clustering methods, the parameters of models are replaced with smaller values. In this process, group the weights into different clusters(N).

Then the centroid value of each cluster is calculated. This value is shared for all the weights in that cluster and thus enabling weight sharing. This method compresses the model.

![Fig: Confusion Matrix of model after INT4 bit quantization(Quantization aware training)](image)

Performance of Base Model is shown below. The base model has an accuracy of 0.56533

![Fig: Confusion Matrix of Base model](image)

Analysis and Results

The Figure below shows, how each methods have reduced the model size and how it affected the accuracy of respective models.

![Fig: Graph showing the comparison of size and accuracy across models](image)

Summary/Conclusions

The model optimization methods have reduced the training time and size of the model by compressing them. Some of these methods are not suitable for the Automatic Modulation Classification problem as they reduce the accuracy of model further.

The quantization aware training and pruning were effective in reducing the size of model without degrading the model performance. The TensorFlow model optimization tool kit is used to implement these methods.

Key References


Acknowledgements

I would like to thank my project advisor Prof. Birsen Sirkeci for her tremendous support throughout the completion of this project. I am grateful to her for giving me an opportunity to do this project under her guidance which helped me to learn and explore new areas of research.

I would also like to thank my family and friends for their continuous support and encouragement.