BNNs are a requisite component of artificial intelligence for providing a near-human or superhuman efficiency and accuracy with reduced weights, memory occupancy and power consumption. The objective of the project was to develop a small-sized inference model i.e., a pruned network with less complexity, by quantization, net-adaptation, and retraining in order to reduce the computation cost and maintain a better efficiency and performance. Usually a CNN consumes a huge amount of memory and power consumption. A typical ResNet-101 takes around 170MB of storage space whereas in the case of AlexNet around 250MB. This is not favorable with dense layers and increases the problem of exploding gradients or vanishing gradients which is not favorable for developing a best Neural Network model.

**Quantization - 1 bit weight, 8 bit feature map**
- Usually a CNN consumes a huge amount of memory space which is not favorable with dense layers and increased complexity.
- A typical ResNet-101 takes around 170MB of storage space whereas in the case of AlexNet around 250MB.

**Methodology**
- These weight optimizations are applied when the customized SGD optimizer is called from the train class.
- Here Bi is the binary function, with |x|, n is the scaling factor used to maintain a value range.

\[
\text{Weight hysteresis loop}
\]

Usage of CrossEntropy on loss function for quantization method with SGD optimizer with adjusting learning rate.

**NetAdapt Applied on AlexNet**
- Most of the proposed work in order to improve the efficiency of DNNs are based on ‘indirect metrics’, i.e., the weights and Multiply-Accumulate Operations duplicates with the increase in the amount of resource consumed for a pre-defined network.
- To use the NetAdapt method for ‘Direct metrics’ approach inside the optimization network. These metrics are evaluated from the empirical values obtained from the target resource.
- Most of the proposed work in order to improve the efficiency of DNNs are based on ‘indirect metrics’.

**Analysis and Results**
- CIFAR-10 and PKU-Autonomous Driving dataset were pre-processed which were further fed to each of the designed CNN models.
- Given any activation function, when a back-propagation algorithm is implemented in order to optimize the network, the larger values of input could raise the problem of exploding gradients or vanishing gradients which is not favorable for developing a best Neural Network model.

**Summary/Conclusions**
- The model were successfully implemented and verified using AlexNet and ResNet while learning their behaviors for different datasets with more accuracy along with less energy/resource consumption and computational cost.
- The Quantization of ResNet was implemented by reducing the weights from full-precision to 1-bit representation and the feature maps were represented as 8-bit values. NetAdapt was applied on a pre-trained AlexNet model analyzing the result.

**Key References**
2. Tien-Ju Yang1, Andrew Howard , Bo Chen , Xiao Zhang , Alec Go , Mark Sandler, Vivienne Sze1, and Hartwig Adam proposed “NetAdapt: Platform-Aware Neural Network Adaptation for Mobile Applications”, EVCC 2018 paper.

**Acknowledgements**
The authors are deeply indebted to Dr. Chang Choo for his invaluable patience and professional advice on this project. We also appreciate the Electrical Engineering department for granting the opportunity to presenting this project.