**Introduction**

- As the design complexities of CNN have increased, the implementation of such models on embedded devices has become harder.
- The right choice of numeric representation format can either increase the performance of the overall system or degrade it.
- Here the use of fixed point which is hardware friendly is compared with IEEE754 floating point format on several CNNs.

**Methodology**

**Convolutional Neural Network Architectures**

1. **LeNet**
   - Figure 3: LeNet Model (the first CNN for digit recognition)

2. **AlexNet**
   - Figure 4: AlexNet Model

3. **VGGNet**
   - Figure 5: VGGNet

**MNIST dataset**

For performance analysis of each CNN architecture the MNIST digit recognition is implemented consisting 60,000 training images and 10,000 testing images.

**Processor Specifications**

- CPU: ARM64/AArch64
- Instruction set: 64 bit
- Architecture: AArch64
- Core: 8
- OS: Android
- Platform: Python3
- Code editor: Jupyter
- Notebook

**Analysis and Results**

1. Embedded devices performance also depends on the use of number representation used to run the computation. Fixed point is hardware and power efficient but only at register transfer level.
2. Floating point provides better accuracy but requires more power.
3. The comparison of these two formats are tested against multiple CNNs for comparison.

Various CNN architectures are analyzed on CPU/GPU for 20 epochs each and the performance metrics considered are: Loss, Accuracy, Power usage, CPU/GPU usage.

**Figure 6:** Identity block and skip connection of ResNet

**Figure 7:** MNIST dataset

**Figure 8:** Processor specification obtained from Android

**Summary/Conclusions**

1. Fixed point number representation format is compared with floating point representation to attain higher accuracy over the CNN architectures.
2. IEEE754 floating point various precisions are also compared for same CNN architectures for change in accuracy, power usage and memory utilization.
3. The performance of the CNN increased by 0.7X when shifted to 32 bit fixed point from fixed and 0.25X when switched with 16bit over 32 bit floating point.

**Key References**


**Acknowledgements**

I would like to express my sincere gratitude to Electrical Engineering Dept., SJSU for the continuous support of our Master’s project and related research.