# Analysis of Numeric Formats in Artificial Intelligence: Balancing Accuracy and Resource Usage in Depth-Wise Convolutions

Dayoung Lee, Jaeseong Kim, Chaebin Lee, Joungmin Park, Raehyeong Kim and Seung Eun Lee*
*Dept. of Electronic Engineering, Seoul National University of Science and Technology, Seoul, Republic of Korea*

The choice of numeric formats in artificial intelligence (AI) computations affects accuracy and resource usage, which are crucial for implementing AI [1]. In this paper, we analyze four numeric formats of the convolution operations by comparing their effects on accuracy and resource usage. For comparison, a depth-wise convolution layer in MobileNetV1 was designed based on a systolic array architecture in each format by Verilog HDL and implemented on a Cyclone-IV: EP4CE115F29C7 for verification. Fig 1 shows the experimental environment via UART. Parameters of a pre-trained model with CIFAR-10 dataset were converted to floating point (FP32/FP16), bfloat16(BF16), and fixed32 formats for inference.

The results are indicated in Fig 2. FP16 and BF16 had lower accuracy than FP32 but halved register and memory usage due to their bit width. BF16 reduces hardware size by adopting an 8-bit exponent and a 7-bit mantissa. This broadens the representable range while reducing circuit complexity. FP16, by contrast, provides a more precise decimal representation with a 5-bit exponent and a 10-bit mantissa. The additional mantissa bits in FP16 require an extra multiplication and normalization process, which increases the use of LUTs. Thus, BF16 is the optimal choice for resource-constrained environments, and this paper provides criteria for selecting numeric formats based on accuracy and resource requirements.

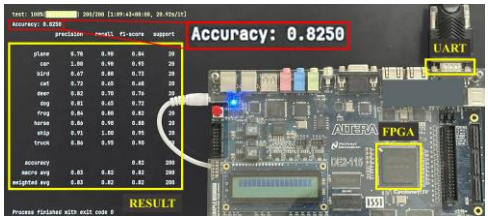


Fig 1. The experimental environment

| Data Type | LUT | Register | Memory Usage |
|---|---|---|---|
| FP32 | 29,109 | 1,998 | 2,656,032 / 3,981,312(67%) |
| FP16 | 13,490 | 1,094 | 1,328,096 / 3,981,312(33%) |
| BF16 | 11,255 | 1,094 | 1,328,096 / 3,981,312(33%) |
| fixed32 | 20,126 | 1,998 | 2,656,032 / 3,981,312(67%) |

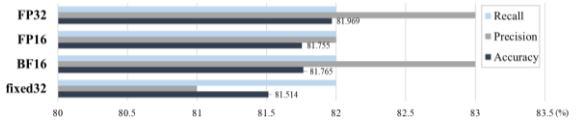


Fig 2. Comparison of inference accuracy and resource

**Acknowledgments** This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2024-RS-2022-00156295) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

**References** [1] A. Y. Romanov et al., "Analysis of Posit and Bfloat Arithmetic of Real Numbers for Machine Learning," in IEEE Access, vol. 9, pp. 82318-82324, 2021.