Training Quantization for 16 Bit Factorization

A) Retraining with Maximum Weight Constraint
- No Retraining
- w/Max Weight Constraint

B) Retraining with Quantization Loss
- 32 bit Floating Point
- 6 bit, No Retraining
- 6 bit, with Retraining

C) Weights distribution before Retraining

D) Weights distribution after Quantization Retraining