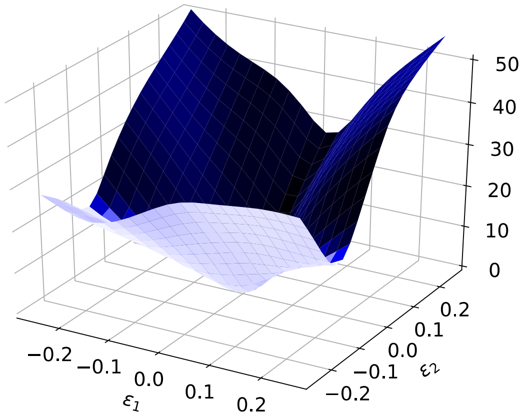
Beyond 250KB: Hessian AWare Quantization

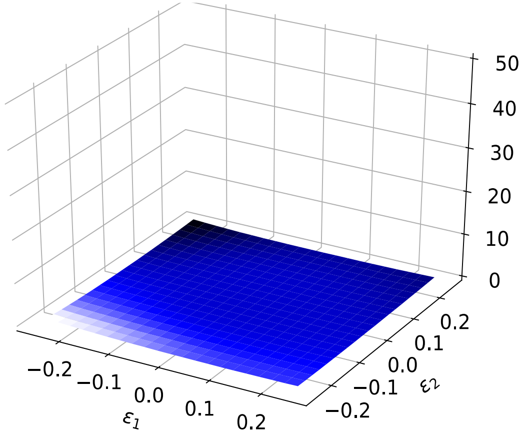
Team members: Zhen Dong, Yaohui Cai, Amir Gholami, Tianjun Zhang, Kurt Keutzer

- Modified shufflenetv2 with size 680KB, peak memory 1296KB.
- **Automated Hessian Aware Mixed-precision and Fine-tuning:**
- Based on Second-order information **with no searching**.



Results (using only 4 GPUs):

- Accuracy: **91.55%**, Model Size: **129KB**, MACs: **23M**.



Dong, Z., et al. HAWQ: Hessian AWare Quantization of Neural Networks with Mixed-Precision. *arXiv preprint arXiv:1905.03696*.