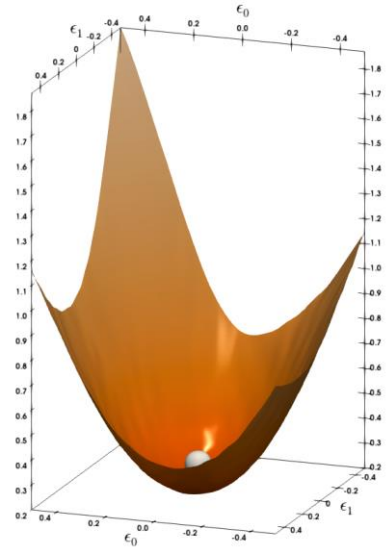
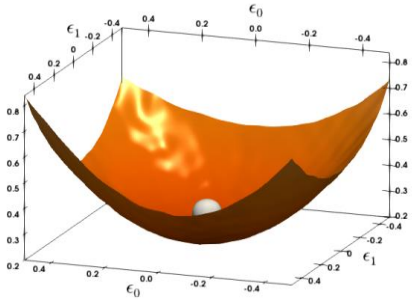
Q-BERT: Hessian-based Quantization for BERT

Shen Sheng, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael Mahoney, Kurt Keutzer

- Hessian-based ultra-low precision quantization (down to **2-bit**);
- Group-wise Quantization for multi-head attention model (BERT);
- **13x smaller** model with at most 2% accuracy loss.

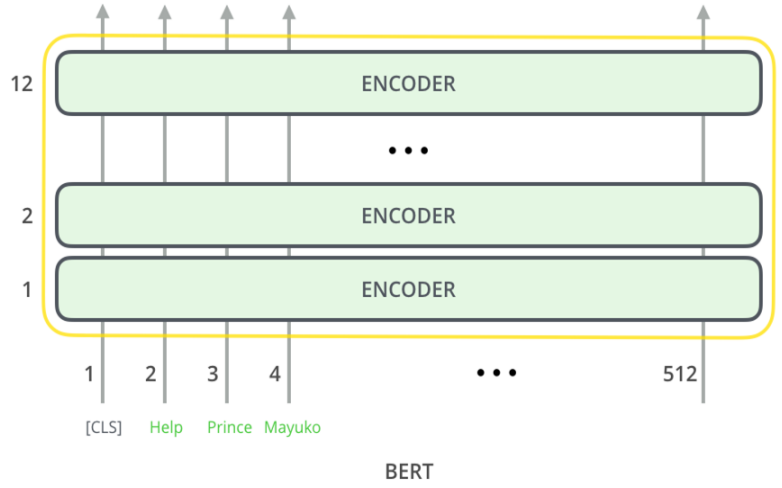


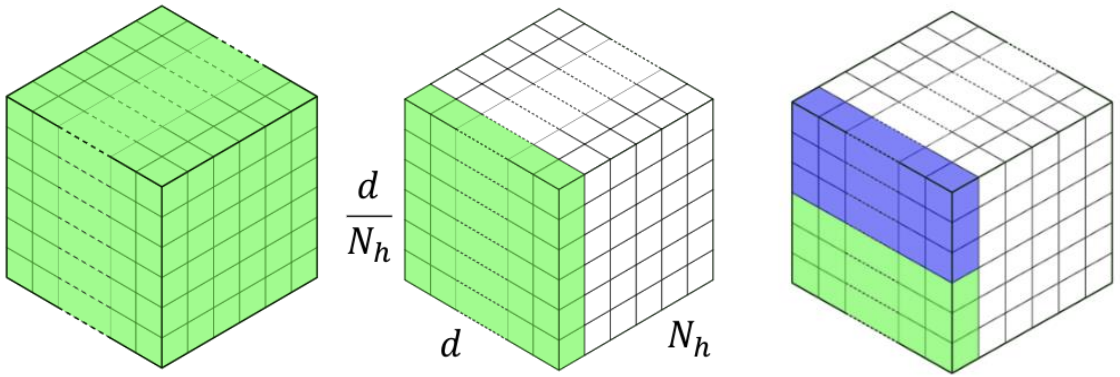
4th Layer



10th Layer

Q-BERT: Hessian-based Quantization for BERT





(a) Layer-wise

(b) Group-wise (N_h group)

(c) Group-wise ($2*N_h$ group)

(b) MNLI

Method	w-bits	e-bits	Acc m	Acc mm	Size	Size w/o-e
Baseline	32	32	84.00	84.40	415.4	324.5
Q-BERT	8	8	83.91	83.83	103.9	81.2
DirectQ	4	8	76.69	77.00	63.4	40.6
Q-BERT	4	8	**83.89**	**84.17**	63.4	40.6
DirectQ	3	8	70.27	70.89	53.2	30.5
Q-BERT	3	8	**83.41**	**83.83**	53.2	30.5
Q-BERT$_{MP}$	2/4 $_{MP}$	8	**83.51**	**83.55**	53.2	30.5
DirectQ	2	8	53.29	53.32	43.1	20.4
Q-BERT	2	8	**76.56**	**77.02**	43.1	20.4
Q-BERT$_{MP}$	2/3 $_{MP}$	8	**81.75**	**82.29**	**46.1**	**23.4**