Q-BERT: Hessian Based Ultra Low Precision Quantization of BERT

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

BAIR Lab, University of California at Berkeley

The Question

- How to build an efficient in-car human interaction system with SOTA NLP models?
- How to adjust well-known efficient methods to NLP models without performance degradation?
- What are the difficulties/bottlenecks when compressing the SOTA NLP model like BERT [3]?

Contributions

We introduce a novel quantization scheme for BERT:

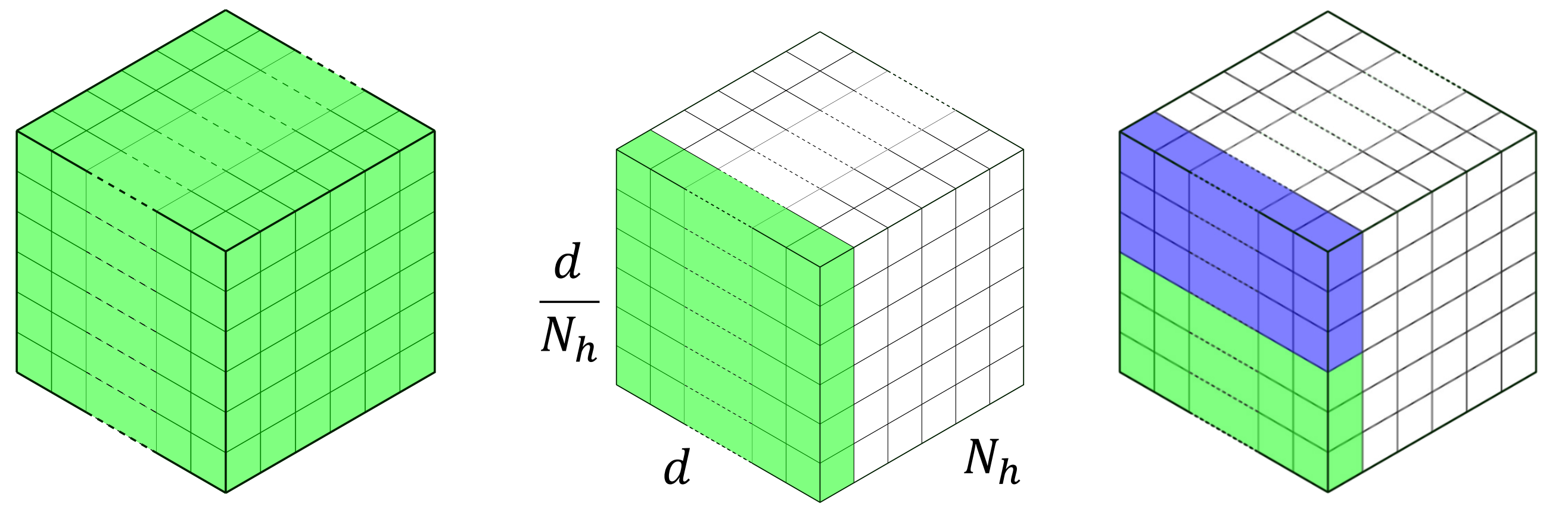
- We apply mixed-precision quantization on BERT, guided by extensive layer-wise analysis of second-order information (i.e., Hessian information). BERT exhibits drastically different Hessian behavior, as compared with NN models for computer vision. Therefore, We propose a sensitivity measurement based on both mean and variance of the top eigenvalues in order to achieve better mixed-precision quantization.
- We propose a new quantization scheme, named the group-wise quantization, which can alleviate accuracy degradation, without a significant increase in hardware complexity.
- We investigate the bottlenecks in BERT quantization, namely how different factors such as quantization scheme and modules such as embedding, self-attention, and fully-connected layers affect the trade-off between NLP performance and the model compression ratio.

Problem Description

Natural language processing (NLP) has recently become a core capability for many consumer devices. This is especially true in safety-critical applications, such as in cars, when a distraction due to a complex user interface can potentially result in a collision or even death. Recently, BERT [3] has become widely used in NLP, having achieved state-of-the-art accuracy in essentially every domain, including user intent classification, question answering, and sentiment classification. The extreme memory footprint required for BERT prevents its adaption to in-car embedded systems. We pose the first attempt to quantize BERT by 13× smaller, which allows for negligible performance degradation and makes efficient edge deployment on systems with limited cache and FLOPs capability feasible.

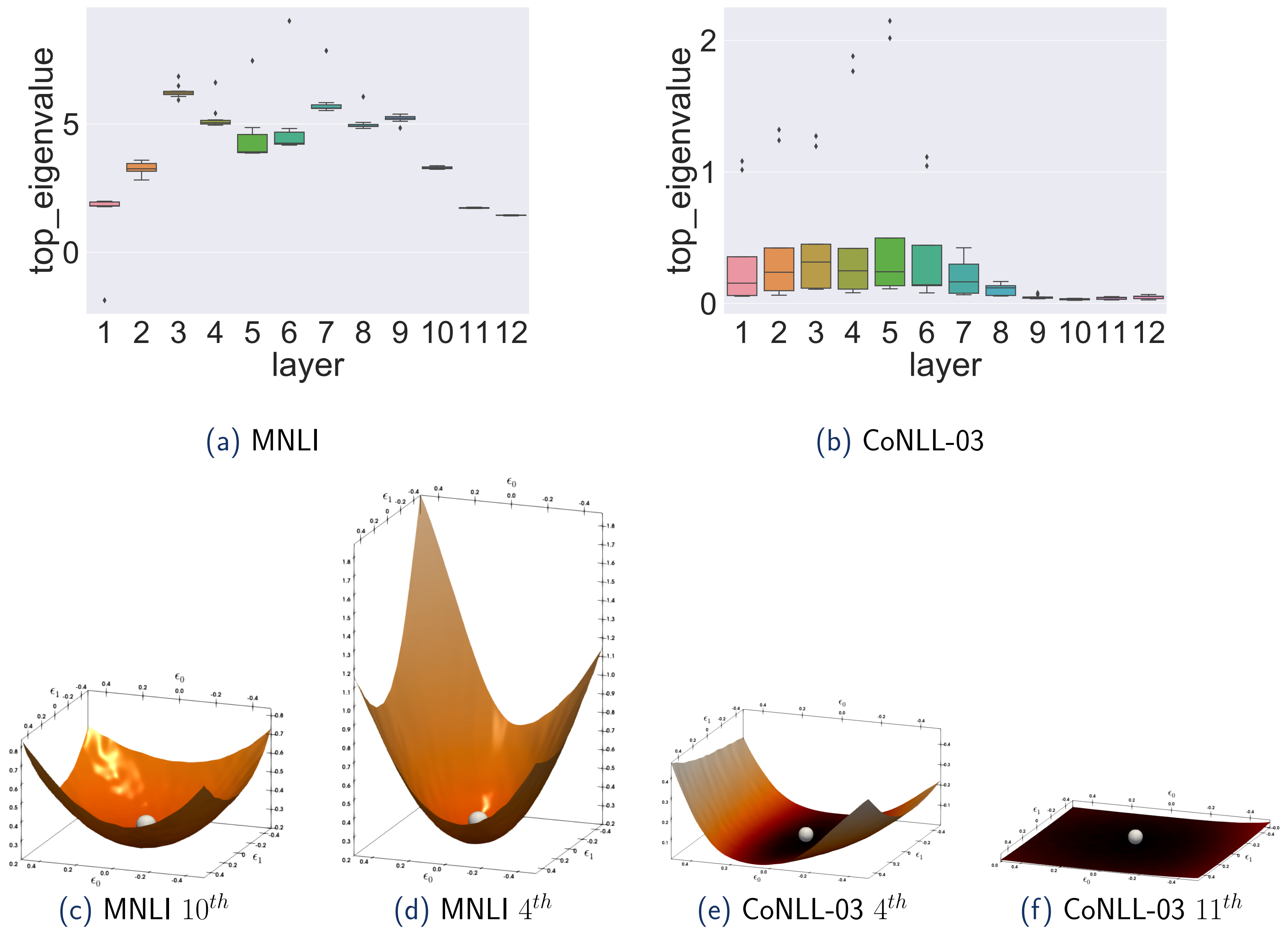
Method

We first present group-wise quantization in Fig. 1. Specifically, we partition each matrix to different groups, each with its unique quantization range and lookup table.



(a) Layer-wise (b) Group-wise (N_h group) (c) Group-wise ($2N_h$ group)

Figure 1: The overview of the Group-wise Quantization Method. We illustrate this with value matrices of a multi-head self-attention layer. Here N_h(number of heads) value matrices W_v are concatenated together, which results in a 3-d tensor. The same color denotes the same group with a shared quantization range. As shown in (a), for layer-wise quantization, the entire 3-d tensor will be quantized from a universal quantization range into discrete unsigned integers. A special case of group-wise quantization in (b) is that we treat each dense matrix as a group, and every matrix can have its own quantization range. We show a more general case in (c), where we partition each dense matrix w.r.t. output neuron and bucket every continuous $\frac{d}{2N_h}$ output neurons as a group.



(a) MNLI (b) CoNLL-03

(c) MNLI 10^{th} (d) MNLI 4^{th} (e) CoNLL-03 4^{th} (f) CoNLL-03 11^{th}

Figure 2: From (a) to (b): Top eigenvalue distributions for different encoder layers for MNLI, CoNLL-03, respectively. For each task, 10% of the data is used to compute the top eigenvalue, and we perform 10 individual runs to plot the top eigenvalue distribution. It can be seen that layers in the middle have higher mean values, and they also tend to have larger variance than the others. The last three layers have the smallest variance as well as mean values among all layers.

Observed the large variance in top eigenvalue distribution in Fig. 2, we use the following metric instead of just using mean value as in [1],

$$\Omega_i \triangleq |\text{mean}(\lambda_i)| + \text{std}(\lambda_i), \tag{1}$$

where λ_i is the distribution of the top eigenvalues of H_i, calculated with 10% of the training dataset.

Results

Table 1: Quantization results for Q-BERT on MNLI.

Method	w-bits	e-bits	Acc-m	Size w/o-e
Baseline	32	32	84.00	324.5
Q-BERT	8	8	83.91	81.2
DirectQ	4	8	76.69	40.6
Q-BERT	4	8	**83.89**	40.6
DirectQ	3	8	70.27	30.5
Q-BERT	3	8	**83.41**	30.5
Q-BERT MP	2/4 MP	8	**83.51**	30.5
DirectQ	2	8	53.29	20.4
Q-BERT	2	8	**76.56**	20.4
Q-BERT MP	2/3 MP	8	**81.75**	**23.4**

Table 2: Quantization results for Q-BERT on CoNLL-03.

Method	w-bits	e-bits	F_1	Size-w/o-e
Baseline	32	32	95.00	324.5
Q-BERT	8	8	94.79	81.2
DirectQ	4	8	89.86	40.6
Q-BERT	4	8	**94.90**	40.6
DirectQ	3	8	84.92	30.5
Q-BERT	3	8	**94.78**	30.5
Q-BERT MP	2/4 MP	8	**94.55**	30.5
DirectQ	2	8	54.50	20.4
Q-BERT	2	8	**91.06**	20.4
Q-BERT MP	2/3 MP	8	**94.37**	**23.4**

Table 3: Quantization results for Q-BERT on SST-2.

Method	w-bits	e-bits	Acc	Size-w/o-e
Baseline	32	32	93.00	324.5
Q-BERT	8	8	92.88	81.2
DirectQ	4	8	85.67	40.6
Q-BERT	4	8	**92.66**	40.6
DirectQ	3	8	82.86	30.5
Q-BERT	3	8	**92.54**	30.5
Q-BERT MP	2/4 MP	8	**92.55**	30.5
DirectQ	2	8	80.62	20.4
Q-BERT	2	8	**84.63**	20.4
Q-BERT MP	2/3 MP	8	**92.08**	**25.4**

Table 4: Quantization results for Q-BERT on SQuAD.

Method	w-bits	e-bits	F_1	Size-w/o-e
Baseline	32	32	88.69	324.5
Q-BERT	8	8	88.47	81.2
DirectQ	4	8	77.10	40.6
Q-BERT	4	8	**88.36**	40.6
DirectQ	3	8	59.83	30.5
Q-BERT	3	8	**87.66**	30.5
Q-BERT MP	2/4 MP	8	**87.49**	30.5
DirectQ	2	8	10.32	20.4
Q-BERT	2	8	**79.60**	20.4
Q-BERT MP	2/3 MP	8	**86.95**	**25.4**

As shown above, Q-BERT consistently outperforms the Direct quantization (DirectQ) by a large margin. In particular, under low-bit settings like 3-bits, the gap between Q-BERT and DirectQ increases even further to 9.68-27.83% for various tasks.

Conclusions

In this work, we summarize our conclusions as follows:

- Q-BERT, a new layer-wise Hessian based method which captures both the average and the variance of the eigenvalues [4]. It can aggressively reduce the model size by mixed-precision quantization with negligible performance degradation.
- A new group-wise quantization is proposed to perform fine-grained quantization inside each encoder layer for BERT [3], which add extra efficiency again.
- In four downstream tasks, equipped with the aforementioned methods, Q-BERT achieves 13× compression ratio in weights, 4× smaller activation size, and 4× smaller embedding size, with at most 2.3% accuracy loss.
- Throughout analysis is in the paper [2], which includes how different factors will affect the trade-off between performance and the model compression ratio in Q-BERTwith quantizing different modules in BERT, respectively.

References

[1] Z. Dong, Z Yao, A Gholami, MW Mahoney, K Keutzer. *H*AWQ: Hessian AWare Quantization of Neural Networks with Mixed-Precision. ICCV'19.

[2] S. Shen, Z. Dong, J. Ye, L. Ma, Z. Yao, A. Gholami, M. Mahoney, K. Keutzer, *Q*-BERT: Hessian Based Ultra Low Precision Quantization of BERT. AAAI'20.

[3] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., *B*ERT: Pre-training of deep bidirectional transformers for language understanding. NAACL'19.

[4] https://github.com/amirgholami/PyHessian.