PB-LLM: Partially Binarized Large Language Models

Anonymous authors
Paper under double-blind review

A Supplemental Materials

A.1 Exisiting Binarization Methods on LLM Quantization

Method	BoolQ	PIQA	HellaSwag	WinoGrande	ARC-Easy	ARC-Challenge	OBQA	Mean
Random Performance	0.5	0.5	0.25	0.5	0.25	0.25	0.25	0.36
FP	0.595	0.63	0.415	0.595	0.54	0.22	0.25	0.46
BNN	0.38	0.545	0.235	0.46	0.195	0.165	0.15	0.30
XNOR	0.37	0.525	0.265	0.49	0.195	0.165	0.16	0.31
Bi-Real	0.395	0.5	0.25	0.505	0.235	0.185	0.165	0.32
ReCU	0.39	0.515	0.24	0.51	0.255	0.185	0.175	0.32
FDA	0.39	0.485	0.265	0.49	0.265	0.19	0.17	0.32

Table 1: Table corresponds to Figure 2 in the main paper: We implement five renowned binarization methods on LLMs and assess the resultant binarized LLMs across seven zero-shot common sense reasoning tasks.

We first investigate the possibility of implementing binarization to LLM quantization. Specifically, following the binarization benchmark in BiBench [Qin et al., 2023], we generalize some representative binarization methods into LLM quantization scenarios. BNN [Hubara et al., 2016], XNOR [Rastegari et al., 2016], Bi-Real [Liu et al., 2020], ReCU [Xu et al., 2021a] and FDA [Xu et al., 2021b] are re-implemented to quantize LLMs, particularly to OPT [Zhang et al., 2022]. Training details are illustrated in the Sec. 4. The results evaluated on seven zero-shot common sense reasoning tasks are shown in the above table. We can see that the LLMs binarized via the existing popular binarization algorithms perform worse than random guesses, showing that the existing binarization methods are not suitable for LLM binarization.

A.2 Codes

Codes can be found anomalously in PB-LLM.

References

Haotong Qin, Mingyuan Zhang, Yifu Ding, Aoyu Li, Zhongang Cai, Ziwei Liu, Fisher Yu, and Xianglong Liu. Bibench: Benchmarking and analyzing network binarization. *ICML*, 2023.

Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In *NeurIPS*, 2016.

Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *ECCV*, 2016.

Zechun Liu, Wenhan Luo, Baoyuan Wu, Xin Yang, Wei Liu, and Kwang-Ting Cheng. Bi-real net: Binarizing deep network towards real-network performance. *IJCV*, 2020.

Zihan Xu, Mingbao Lin, Jianzhuang Liu, Jie Chen, Ling Shao, Yue Gao, Yonghong Tian, and Rongrong Ji. Recu: Reviving the dead weights in binary neural networks. In *ICCV*, 2021a.

Yixing Xu, Kai Han, Chang Xu, Yehui Tang, Chunjing Xu, and Yunhe Wang. Learning frequency domain approximation for binary neural networks. In *NeurIPS*, 2021b.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.