Capstone Bi-Monthly Update

Date: 10/11/2023

Team: Aditya, Elisa, Jenny, Zhanyi

Agenda

1) Network Acceleration Experiment Discussion
    a) Updates & Results
    b) Challenges
2) Next Steps

BERT Distillation Results

Model: BERT
Task: Multilabel-Classification
Number of Epochs: 5
Dataset: multilabel emotion classification for tweets (default Huggingface dataset)
Machine: GPU

Metric	Baseline (Fine-Tuned)	Distilled
Accuracy	0.276	0.268
Inference Time	7.3s	4.8s
Training Time	953s	1699s

BERT Distillation Experiment - Challenges

1) OutOfMemory Issue with GPUs

2) Hard to generalize code across different datasets that require varying degree of preprocessing

3) Mismatch between teacher model and student model word embedding happens sometime（With TextBrewer Package）

4) Appropriate choice of student architecture requires more research

Bloomz 3B Quantization- Results

Model: Bloomz
Task: Binary Classification
Number of Epochs: pre-trained
Dataset: Human vs LLM generated wikipedia articles (“NicolaiSivesind/human-vs-machine”)
Machine: CPU

Metric	Baseline (Fine-Tuned)	Quantized
Accuracy	78.4	72.4
Inference Time (s)	22.19 (mean)	20.3 (mean)
Training Time	953s	TBD
Model Size (MB)	12010	4932

Bloomz Quantization challenges

1) Libraries are not mature enough. Not enough debugging support.
2) Libraries are not stable, works for some but does not work for either (Quantization code did not work for distilled BERT)

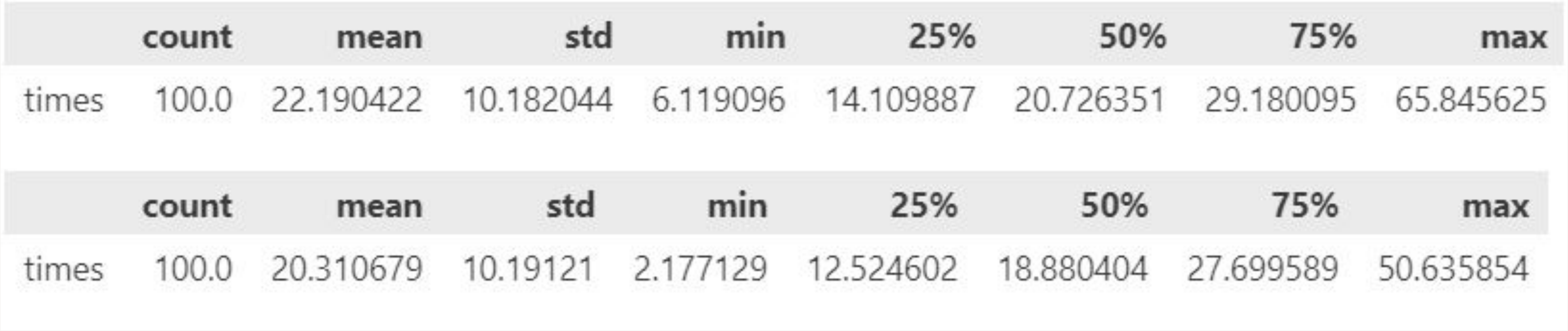
Next Steps

1) Network Acceleration Experiment
    - Expand framework to other LLMs
    - Refactor code to generalize across other LLMs and datasets
    - Hyperparameter Tuning (deprioritize)
    - Distilled + Quantization

2) Literature Review

Appendix:

Quantized Bloomz inference metrics



Base Model

	count	mean	std	min	25%	50%	75%	max
times	100.0	22.190422	10.182044	6.119096	14.109887	20.726351	29.180095	65.845625

Quantized Model

	count	mean	std	min	25%	50%	75%	max
times	100.0	20.310679	10.19121	2.177129	12.524602	18.880404	27.699589	50.635854

Duke