Bonus Work 1

Ankita Arvind Deshmukh

SJSU ID: 016029585

Link to Colab: https://colab.research.google.com/drive/1vlc9pKbpvfhSWK5o_xnhQLOxdZMM-pD8?usp=sharing

Link to GitHub: https://github.com/ankdeshm/CMPE255_BonusWork1

Option 1: Inference via TorchScript

Introduction: With TorchScript, PyTorch aims to create a unified framework from research to production. TorchScript takes our PyTorch modules as input and convert them into a production-friendly format. It will run the models faster and independent of the Python runtime. To focus on the production use case, PyTorch uses 'Script mode' which has 2 components PyTorch JIT and TorchScript.

Example 1:

In the first example, I have utilized BERT(Bidirectional Encoder Representations from Transformers) from the transformer's library provided by HuggingFace.

Steps:

1) Initialize the BERT model/tokenizers and create a sample data for inference
2) Prepare PyTorch models for inference on CPU/GPU
3) Model/Data should be on the same device for training/inference to happen. cuda() transfers the model/data from CPU to GPU.
4) Prepares TorchScript modules (torch.jit.trace) for inference on CPU/GPU
5) Compare the speed of BERT and TorchScript

Results:

Module	Latency on CPU (ms)	Latency on GPU (ms)
BERT	88.82	18.77
TorchScript	86.93	9.32

Conclusion:

On CPU the runtimes are similar but on GPU TorchScript clearly outperforms PyTorch.

Example 2:

In the second example, I have utilized **ResNet**, short for Residual Networks.

Steps:

1) Initialize PyTorch ResNet
2) Prepare PyTorch ResNet model for inference on CPU/GPU

3) Initialize and prepare TorchScript modules (torch.jit.script) for inference on CPU/GPU
4) Compare the speed of PyTorch ResNet and TorchScript

Results:

Module	**Latency on CPU (ms)**	**Latency on GPU (ms)**
ResNet	92.92	9.04
TorchScript	89.58	2.53

Conclusion:

TorchScript significantly outperforms the PyTorch implementation on GPU.

As demonstrated in 2 different ways above, TorchScript is a great way to improve the inference improvement as compared to the original PyTorch inference.

References:

1) https://pytorch.org/tutorials/beginner/Intro_to_TorchScript_tutorial.html#basics-of-torchscript
2) https://towardsdatascience.com/pytorch-jit-and-torchscript-c2a77bac0fff