HW3 使用 ONNX parser 将 Bert 模型转换成 TRT 格式

1. 下载模型

点击链接：https://huggingface.co/models，选择 bert-base-uncased 模型，并学习 bert 模型结构（参考资料见“BERT 学习资料.pdf”）。

如何下载和在本地使用 Bert 预训练模型：

https://blog.csdn.net/weixin_38481963/article/details/110535583

2. 将模型转成 onnx 格式

（1）不考虑 BertTokenizer 结构；

（2）模型输入，格式为[batch_size, max_seq_len]，

input_ids： [1, max_seq_len]

token_type_ids： [1, max_seq_len] # 全 0

input_mask: [1, max_seq_len]

PS：固定 batch_size = 1，以降低作业难度。

模型转成 onnx 格式的实现代码见 Bertmodel 2 ONNX.py 文件 。

进阶任务：使用 onnxruntime gpu 库，做 infer，得到运行时间 Tort，与后面的 trt 时间进行对比。

3. 使用 onnxparser 将 onnx 模型转成 trt plan 模型

备注：建议使用 python api，不建议使用 trtexec，太黑盒，不利于学习。

（1）下载 TensorRT：C++ api 直接使用库就行，python api 需要安装对应的 whil；

（2）使用 onn-simplifer 模型对 onnx 模型进行优化，得到 model-sim.onnx。需要进行此步，否则后面的转换会失败。

onnxsim bert-base-uncased/model.onnx bert-base-uncased/model-sim.onnx --

input-shape input_ids:1,12 token_type_ids:1,12 input_mask:1,12 --dynamic-input-shape

（3）调用 onnx parser python or c++ api，将 model-sim.onnx 转换成 model.plan；

（4）测速

使用 c++ 或者 python api 编写测速代码，得到时间 Ttrt。建议使用 c++ api，毕竟一般上线都是用 c++。

可参考：

https://hemanths933.medium.com/convert-onnx-bert-model-to-tensorrt-e809276b01b6