trtexec 转换encoder.onnx模型指定动态输入：

trtexec --onnx=./encoder.onnx \

--saveEngine=encoder.plan \

--minShapes=speech:1x1x80,speech\_lengths:1 \

--optShapes=speech:64x64x80,speech\_lengths:64 \

--maxShapes=speech:256x256x80,speech\_lengths:256

parser可以解析onnx模型，但是trt不能完整支持onnxop ，报错。

[04/12/2022-01:38:05] [E] Error[2]: [myelinBuilderUtils.cpp::operator()::293] Error Code 2: Internal Error (Slice\_79requires bool I/O but node can not be handled by Myelin.)

[04/12/2022-01:38:05] [E] Error[2]: [builder.cpp::buildSerializedNetwork::609] Error Code 2: Internal Error (Assertion enginePtr != nullptr failed. )

myelin是TRT内部的cuda代码生成器

提示Slice\_79层有不能被trt支持的操作，主要是trt slice layer不支持bool类型的输入，但GreaterorEqual 和not\_30节点的输出都是bool类型，所以需要进行graph surgeon。

使用—verbose选项输出构建的详情信息，其中截取部分：

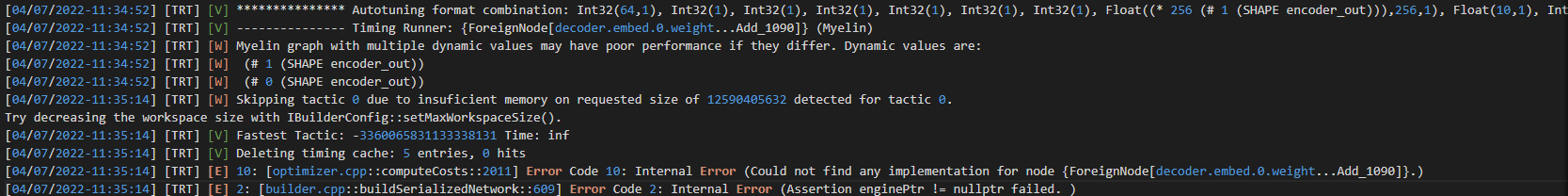
Running: ShuffleErasure

Removing (Unnamed Layer\* 2596) [Shuffle]

Running: ShuffleShuffleFusion

ShuffleShuffleFusion: Fusing Transpose\_1429 with Reshape\_1433

这段信息代表trt 做层融合，一个shuffle算子和一个reshape算子，两者合成一个shuffle算子



“你好，请问这个错误是由于显存不足引起的吗？可以提供一下解决的思路吗？”

“含有不被tensorRT支持的算子”

“我想了想，觉得不是因为这个问题，因为我固定batch size的时候是没问题的，只有动态batch size才会引发。而且这个错误应该是发生在融合优化的过程中的吧”

“固定batch，里边的shape都是静态的，对于有些算子来说，shape维度都是常量，trt可以用其他算子来组成你的算子。

动态batch，对于算子来说，维度是个变量，trt也没招了，才报错的。静态shape也有融合优化。你的日志里说是foreignnode导致的“

输入的speech有两个动态尺寸，b是speech长度，那t呢

相当于单个句子的长度

请问一个比较弱智的问题， 怎么遍历一个节点的子节点呀，或者一个节点总共有多少个子节点是哪个指令呀

我今天写一个范例放进cookbook里去哈，这个确实比较有用

“我是遍历所有节点通过输入找，就是当时感觉这里需要一个统计有多少个子节点或母节点的功能”

“我是通过.o（i）append到列表里，让i自己增加到超过索引范围就停止”

“现在的workaround就是，让它再遍历一遍计算图，去找哪些节点的输入节点就是当前这个节点”

请问一下，为什么用Polygraphy导出的模型可以通过测试，而用trtexec或者trt python API导出的模型却通不过测试，精度不足。看了一下导出的plan文件，trtexec或者trt python API导出的模型是一样的，Polygraphy则是跟他们有所不同，这有可能是哪些原因造成的呢？

polygraphy 检查了onnx文件的“每个节点”是否被 TensorRT 支持，但是没有检查节点之间的组合是否ok。

比如前面很多同学遇到的 slice\_79 的问题，看上下文 Not 节点和 Slice\_79 节点都是TensorRT支持的，但他们碰一块儿就不行。

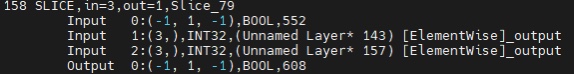
提示：各种节点能接受的数据类型可能不一样

基本思路就是这样，在 onnx 上不断修修补补（增删改一些节点）然后尝试构建，直到 TensorRT 能成功把模型吃进去构建起来

graph.cleanup().toposort() 这句代码是什么意思啊

额，就是把计算图先打扫一下（去除死节点等），然后拓扑排序一下（方便后续TensorRT解析）

请问一下，图节点的op里面是没有elementwise吗



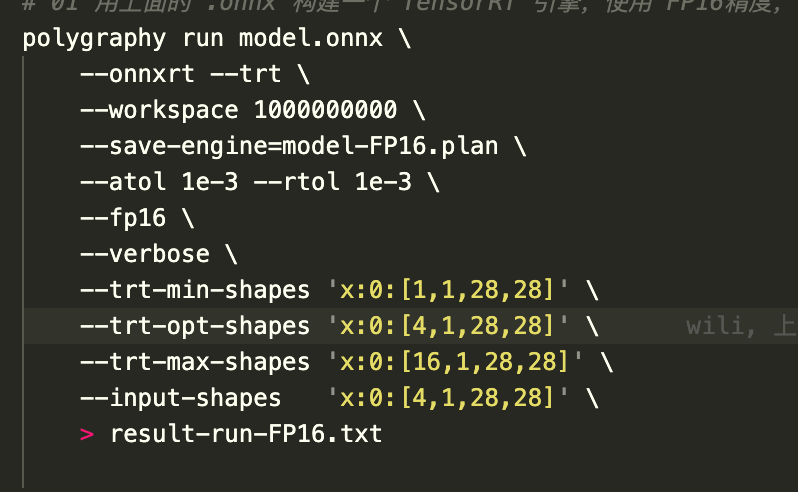
我应该如何找到intput1和input2呢

这个 Unnamed Layer\* 是你在 TensorRT 中 parse 成功后逐层打印出来的对吗？这些结构是TensorRT自己添加的，在原本的 onnx 中应该是找不到的

在TensorRT中他是一个constant layer，输出一个张量给后续某个layer用，那在onnx中他也应该是一个常量，但有可能不是单独的一个节点，二是被包含在某个节点里面了

不得不承认，TensorRT的公开文档对于无法正常解析的onnx如何处理，是缺乏描述的。李玮在前面提到，“你需要把 .onnx 的计算图进行改造，改到 TensorRT 能够支持为止”，就是这样一个大体思路，具体的细节还需要自己探索。从我们的经验来看，需要从TRT的错误输出寻找线索，比如输出中说“处理不了bool”，那么我强行转成另一种数据类型是不是就能处理了？文档并没有说这样就一定可以，或者不可以（因为TRT版本迭代很快，也许下一个版本就支持得更好了），所以这里需要做些尝试。

我鼓励大家都趁这个比赛的机会，多做些测试，多积累一些技巧。如果想问“这里我该怎么办”，应该不会得到回答；但是已经做过亲身试验，问出一些技术性很强的问题，我们总会乐意回答的。在比赛中学习，这也是我们办这个比赛的目的。

我有个问题，如果 ploygraph run 输入的shape有多个， 那么不同的shape之间分隔符是啥，用了空格，逗号，都不可以。。

https://github.com/NVIDIA/TensorRT/tree/master/tools/Polygraphy/examples/cli/convert/03\_dynamic\_shapes\_in\_tensorrt 在这

大家知道ONNX里的bool值分别是0和1这两个，还是0和-1？

如果是0和1的话，那个unsqueeze+not，没有必要添加cast（肯定会慢）。不如直接上层条件求逆（大于等于，变成less than)，然后扩充0值了。

也就是>=、unsqueeze、not整体变成一个<和填充0，不过得想办法让结果不是bool的。

我用polygraph可以对比onnx 和trt最后输出的结果，但是要对比每一层就回报bool I/O but node can not be handled by Myelin.

每一层输出会影响某些层融合吗，我又什么办法可以找到这个节点或者层吗

某层用fp32 cookbook中暂时没有这个例子，你可以搜来学习一下：BuilderFlag.STRICT\_TYPES