Training Deeper Models by GPU Memory Optimization on TensorFlow

Alibaba Group 阿里巴巴集团

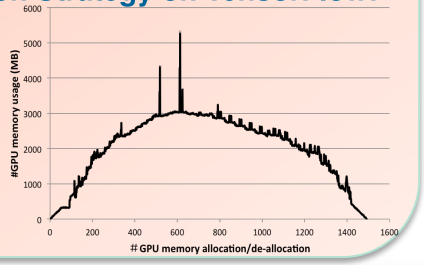
Chen Meng，Minmin Sun，Jun Yang，Minghui Qiu，Yang Gu

Introduction

1.“OOM” issue in training deeper Models.

2. The major constituents of memory usage is feature map.

3. Dynamic Allocation Strategy on TensorFlow: Tensor will be released when reference count becomes 0.



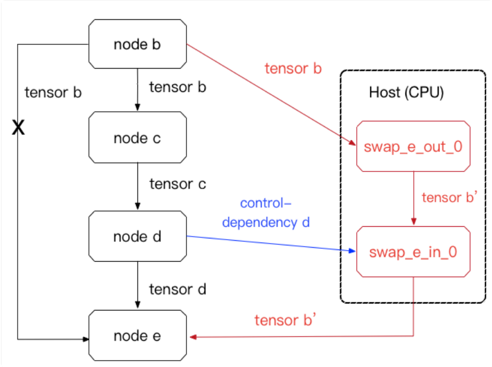


Figure 1. Atomic operation of the swap out/in optimization.

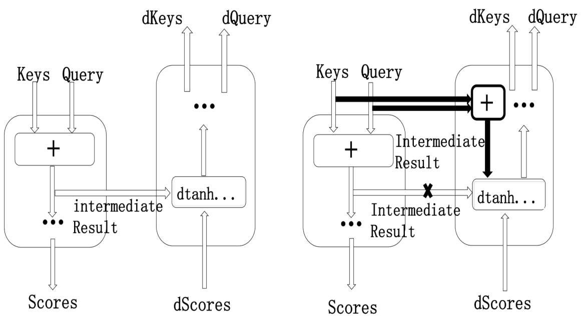


Figure 2. Optimization on Attention operation.

Our approaches

Swap out/in

➢Rewrites the dataflow graph to utilize host memory as memory pool. (Fig. 1)

- Which feature maps to be swapped out?
- When to be swapped back in?

Memory-Efficient Attention

➢Attention layer actually requires much more memory space than LSTM/GRU layers in the Seq2Seq models. (Fig. 2)

- Drop Attention intermediate results directly.
- The re-computation cost is extremely cheap, just an addition operation.

Experiments and Results

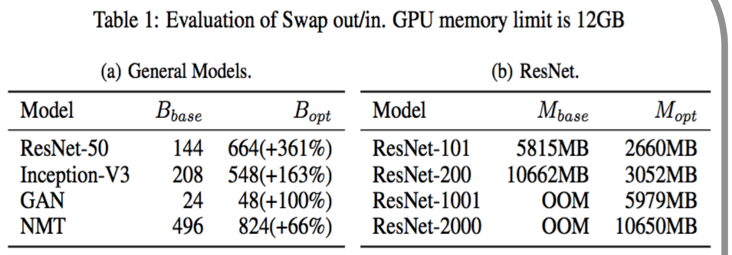


Table 1: Evaluation of Swap out/in. GPU memory limit is 12GB

(a) General Models.

Model	B_{base}	B_{opt}
ResNet-50	144	664(+361%)
Inception-V3	208	548(+163%)
GAN	24	48(+100%)
NMT	496	824(+66%)

(b) ResNet.

Model	M_{base}	M_{opt}
ResNet-101	5815MB	2660MB
ResNet-200	10662MB	3052MB
ResNet-1001	OOM	5979MB
ResNet-2000	OOM	10650MB

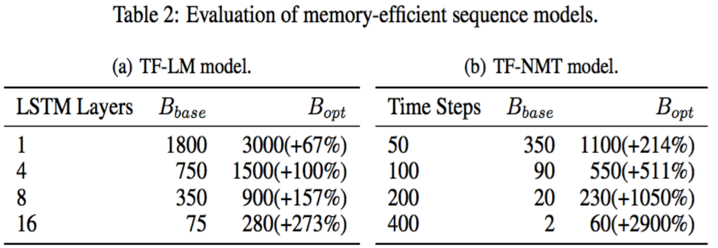


Table 2: Evaluation of memory-efficient sequence models.

(a) TF-LM model.

LSTM Layers	B_{base}	B_{opt}
1	1800	3000(+67%)
4	750	1500(+100%)
8	350	900(+157%)
16	75	280(+273%)

(b) TF-NMT model.

Time Steps	B_{base}	B_{opt}
50	350	1100(+214%)
100	90	550(+511%)
200	20	230(+1050%)
400	2	60(+2900%)

➢ B_{opt} : max batch size after applying memory optimization.

➢ M_{opt}: max memory usage after applying optimization.

Conclusions

➢ The **dataflow-graph based** Swap out/in method.

➢ Memory Efficient Attention op to save huge amount of memory for **Seq2Seq models**.

➢ All approaches are integrated into TensorFlow seamlessly **without requiring any changes** to existing model descriptions.

➢ The max training batch size can **be increased by 2 to 30 times**.