图算融合论文分享：Apollo

嘉宾：赵薇

论文信息

APOLLO: AUTOMATIC PARTITION-BASED OPERATOR FUSION THROUGH LAYER BY LAYER OPTIMIZATION

Jie Zhao [1] Xiong Gao [2] Ruijie Xia [2] Zhaochuang Zhang [2] Deshi Chen [2] Lei Chen [3] Renwei Zhang [2]
Zhen Geng [2] Bin Cheng [2] Xuefeng Jin [2]

ABSTRACT

We study fusion for deep neural networks (DNNs) in a just-in-time (JIT) compilation framework APOLLO. It considers both memory- and compute-bound tensor operators for fusion, and integrates graph-level node grouping and operator-level loop fusion closely, widening the fusion search space. APOLLO enables the upward feedback from the downstream loop optimizer, enforcing the graph engine to regenerate partition patterns amenable to the downstream pass and thus resolving the scalability issue. Besides data locality, APOLLO also exploits the parallelism between independent tensor operators, further improving the performance of DNN workloads. Experimental results on training workloads show that APOLLO outperforms TensorFlow and XLA by 1.86× and 1.37× on a single GPU, and 1.96× and 1.18× on multiple GPUs. APOLLO also improves the performance of a vendor-provided DNN framework by 19.7% on a domain-specific accelerator. In addition, the results of inference workloads demonstrate the general applicability of our fusion framework.

先进编译实验室
Advanced Compiler

- **作者信息**

华为MindSpore图算融合团队和赵捷、陈雷合作出品

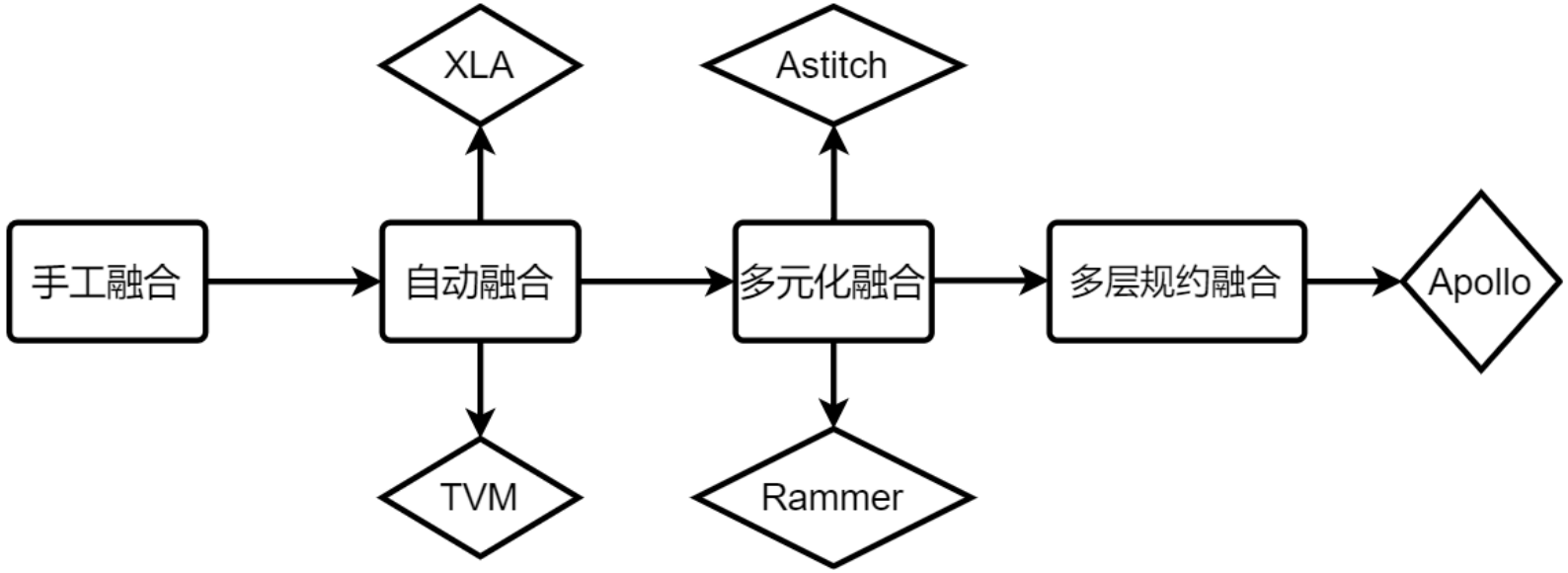
- **论文来源**

机器学习系统研究方向顶级会议 MLSys 2022

目录

算子融合技术发展

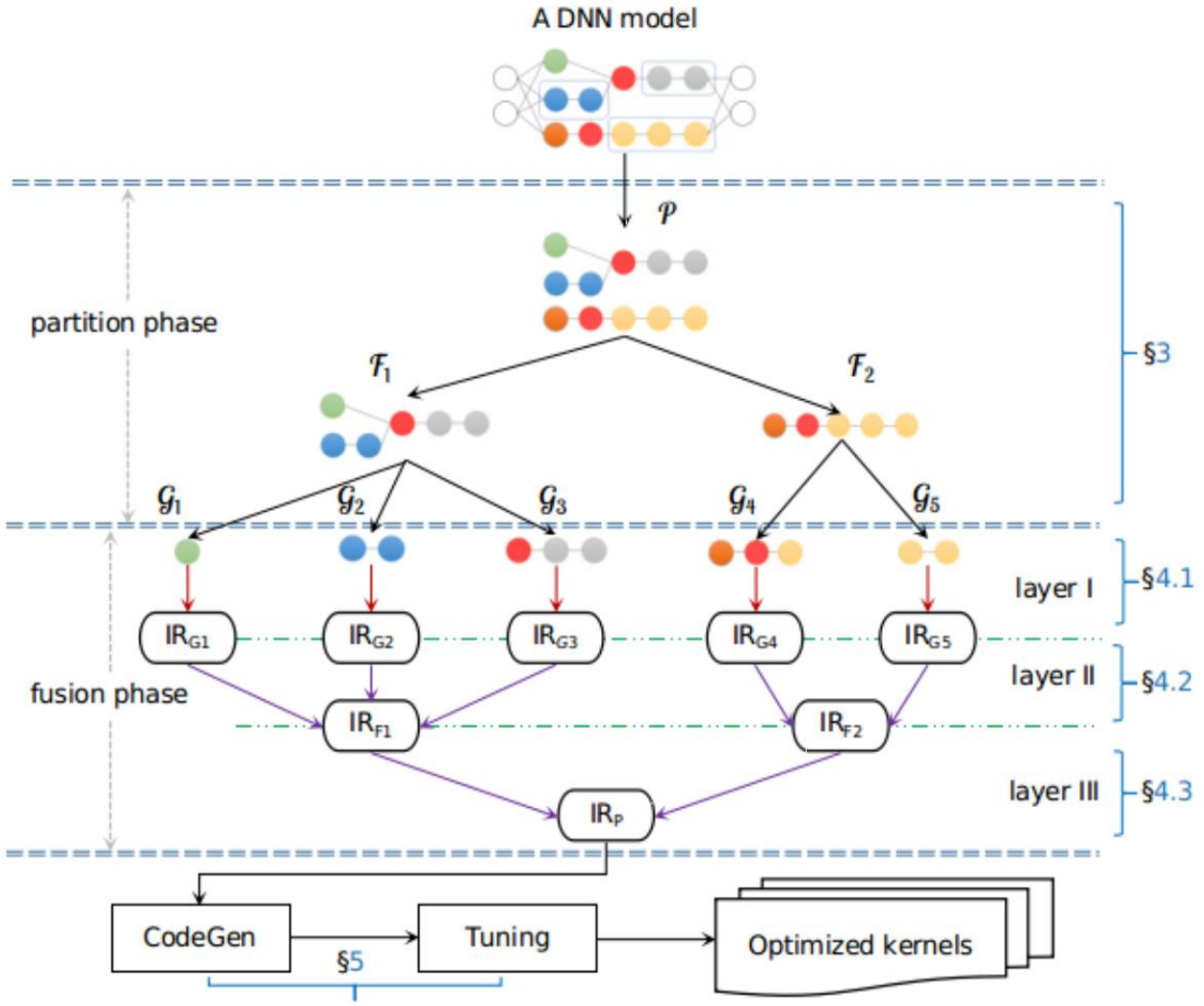
- 早期的AI框架，主要通过手工方式实现固定Pattern的算子融合
- 以XLA、TVM为代表的AI编译框架开始转向自动融合优化技术
- Rammer，Astitch的发布更好地解决了内存墙和并行墙的问题
- Apollo提出了“多层规约融合”，将不同优化角度的融合技术纳入统一的框架下



Apollo技术实现

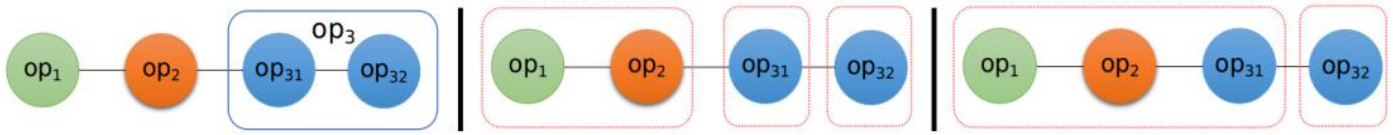
Apollo的两个主要阶段：

- 图划分
- 图融合



先进编译实验室
Advanced Compiler

Apollo—图划分



图划分的具体步骤：

- 将所有合适做融合的算子提取出来
- 分解复合算子
- 聚合基本算子

Rules	$\mathcal{G}_p$	$\mathcal{G}_c$	$\mathcal{G}_a$
①	element-wise	element-wise	element-wise
②	broadcast	element-wise	broadcast
③	broadcast	broadcast	broadcast
④	element-wise	reduction	reduction
⑤	broadcast	reduction	reduction
⑥-transpose	element-wise/broadcast	transpose	transpose
⑥-matmul	matmul	element-wise	matmul
⑥-matmul	element-wise	matmul	matmul
⑥-conv	conv	element-wise	conv
⑥-conv	element-wise	conv	conv

Apollo—图融合

图融合的三层设计：

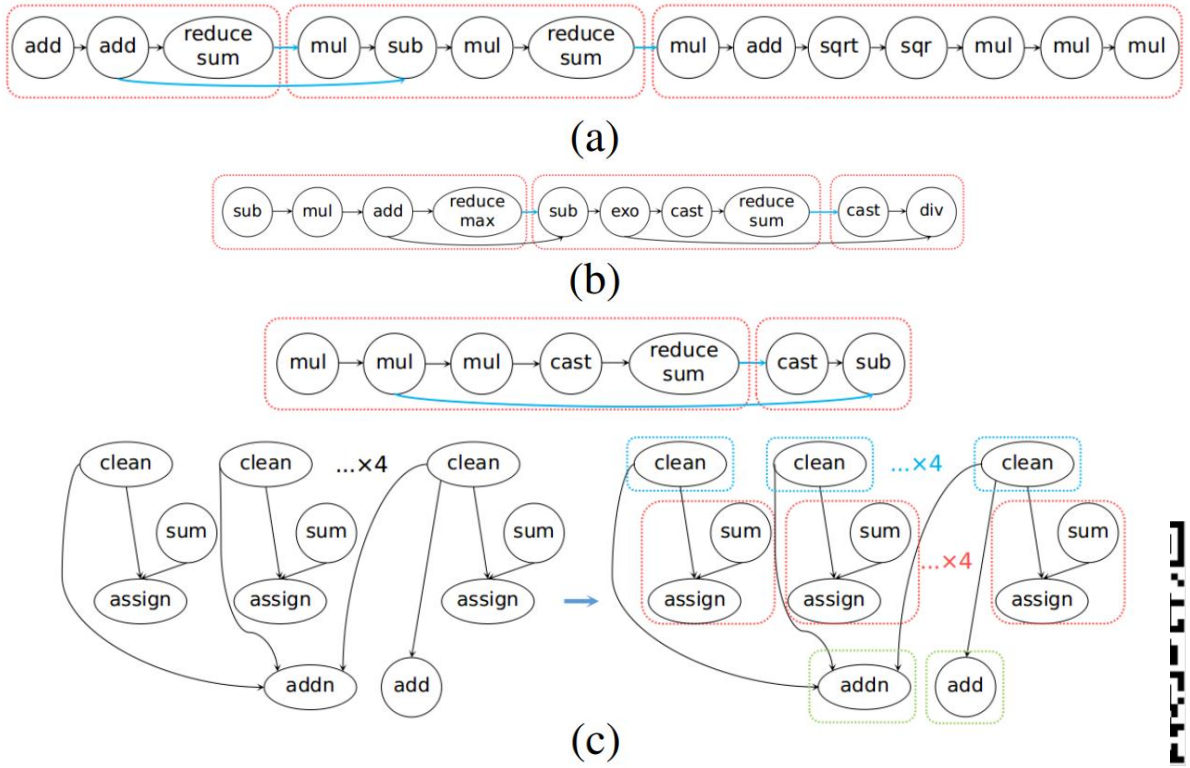
Lay I：对micro-graph进行循环融合

Lay II：在layer I基础上，进一步做Stitch Fusion

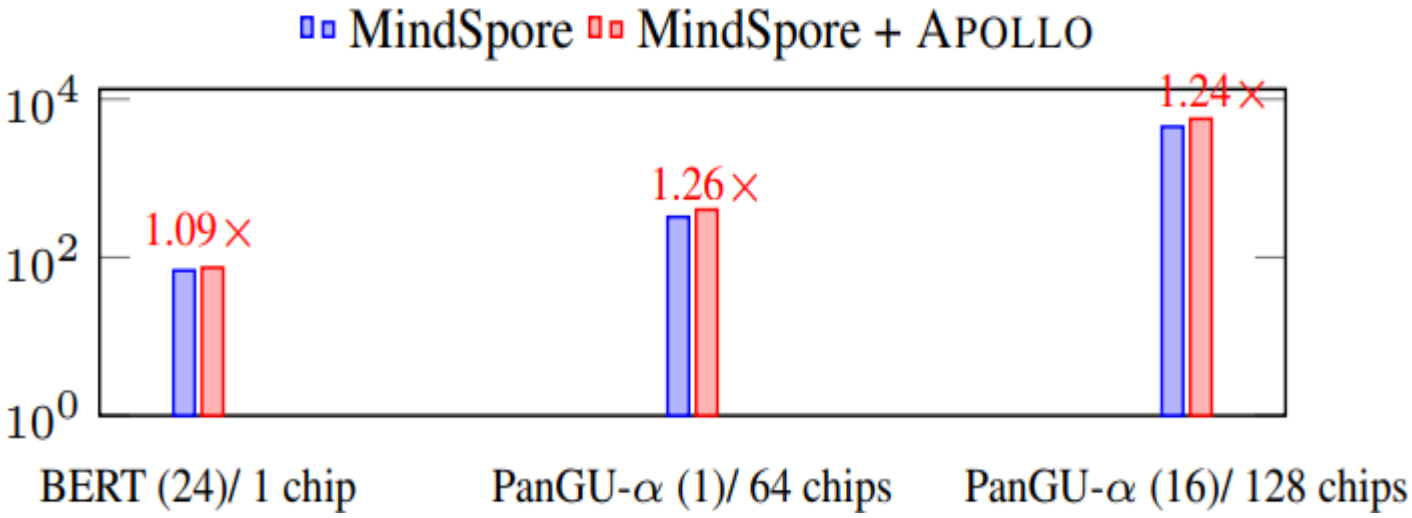
Lay III：识别其中的无依赖融合算子进行融合

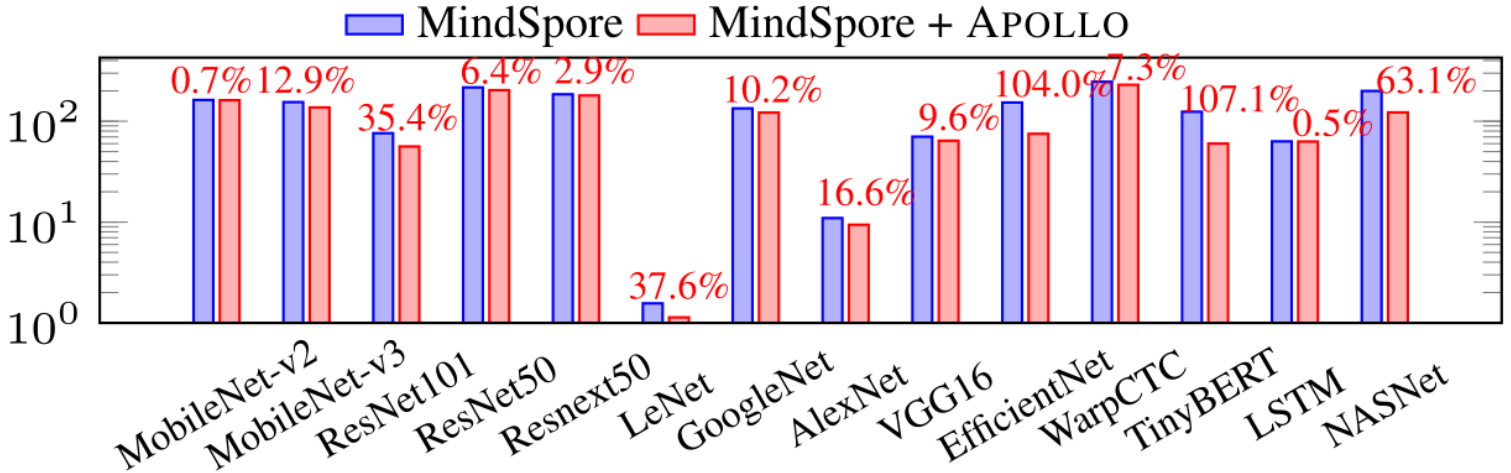

```
for i in [0,M)
  for j in [0,N)
    a(i,j)=a(i,j)+bias; //S1
for i in [0,M/2)
  for j in [0,N/2)
    pool(i,j)=max(a(2i,2j),
      a(2i,2j+1),
      a(2i+1,2j),
      a(2i+1,2j+1)); //S2
```

```
for i in [0,M)
  for j in [0,N){
    a(i,j)=a(i,j)+bias;
    if(i+1) mod 2 = 0 and
      (j+1) mod 2 = 0
      pool((i-1)/2,(j-1)/2)=
        max(a(i-1,j-1),a(i,j-1),
        a(i-1,j),a(i,j)); //S2
  }
```



Apollo技术效果


MindSpore
MindSpore + APOLLO
10^4
10^2
10^0
1.09×
1.26×
1.24×
BERT (24)/ 1 chip
PanGU-α (1)/ 64 chips
PanGU-α (16)/ 128 chips


MindSpore
MindSpore + APOLLO
10^2
10^1
10^0
0.7%
12.9%
35.4%
6.4%
2.9%
37.6%
10.2%
16.6%
9.6%
104.0%
7.3%
107.1%
0.5%
63.1%
MobileNet-v2
MobileNet-v3
ResNet101
ResNet50
Resnext50
LeNet
GoogleNet
AlexNet
VGG16
EfficientNet
WarpCTC
TinyBERT
LSTM
NASNet

参考资料

论文链接：

Apollo: Automatic Partition-based Operator Fusion through Layer by Layer Optimization (mlsys.org)

AdvancedCompiler

Tel: 13839830713