南京大学
NANJING UNIVERSITY

RINC

面向卷积神经网络的模型加速研究

- 答辩人：赵加成 MG1833090
- 导　师：申富饶 教授

南京大学
NANJING UNIVERSITY

目录
CONTENTS

1 研究背景

2 提出方法
- 基于假剪枝与网络搜索的压缩方法
- 基于算子重写的工程推理加速方法

3 实际应用
- 网络加速在鸟类识别系统中的应用

4 研究生期间工作成果

5 总结与展望

南京大学
NANJING UNIVERSITY

第一部分

Research Background

研究背景

- 随着网络性能的不断提升，网络复杂度也在大幅上升，从而带来高昂的计算成本
- 网络加速：在不显著损失网络精度的前提下，对神经网络的运行进行加速

深度神经网络计算

Year	Model	Layers	Parameter	FLOPs	ImageNet Top-5 error
2012	AlexNet	5+3	60M	725M	16.4%
2013	Clarifai	5+3	60M	—	11.7%
2014	MSRA	5+3	200M	—	8.06%
2014	VGG-19	16+3	143M	19.6B	7.32%
2014	GoogLeNet	22	6.8M	1.566B	6.67%

困难与挑战

神经网络强大的学习能力对人们来说依旧是个未解之谜，对其进行加速便困难重重。

相关工作

表 2-1: 网络压缩常用方法

压缩方法	方法简介	应用结构
网络剪枝	删除对网络性能影响不大的参数及结构	卷积层和全连接层
紧凑卷积核设计	设计精巧的卷积核来搭建网络	卷积层
知识蒸馏	搭建一个紧凑网络从大型网络中蒸馏学习	卷积层和全连接层
低秩分解	使用矩阵方法对参数进行分解估计	卷积层和全连接层

现有的相关工作都需要较强的专家经验，同时基于手工操作也会耗时耗力，并且最终的优化还有一定的提升空间。

为了解决这些弊端，本文提出了基于网络搜索的自动化压缩方法，并创新性地设计了假剪枝方法进一步提升压缩性能。同时结合工程推理优化方法，在实际部署中对网络加速的性能进行较好地挖掘。

南京大学
NANJING UNIVERSITY

第二部分

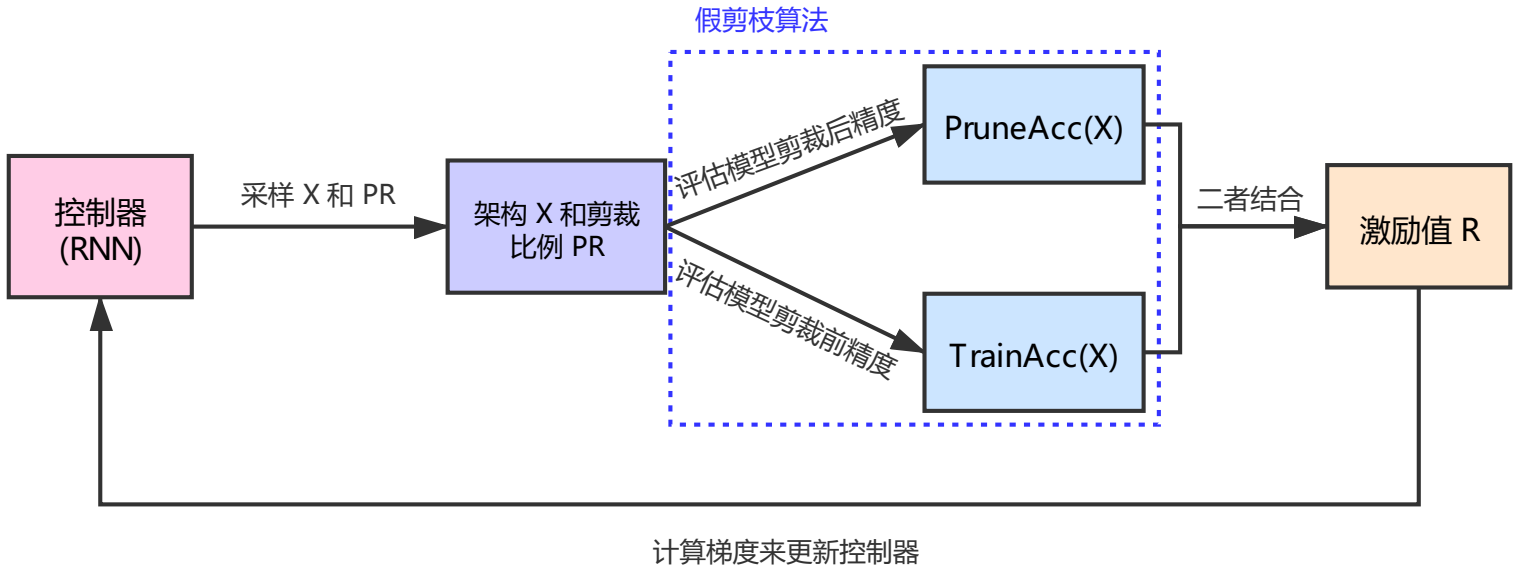
Proposed Methods

提出方法

- 基于假剪枝和网络搜索的压缩方法
- 基于算子重写的工程推理加速方法

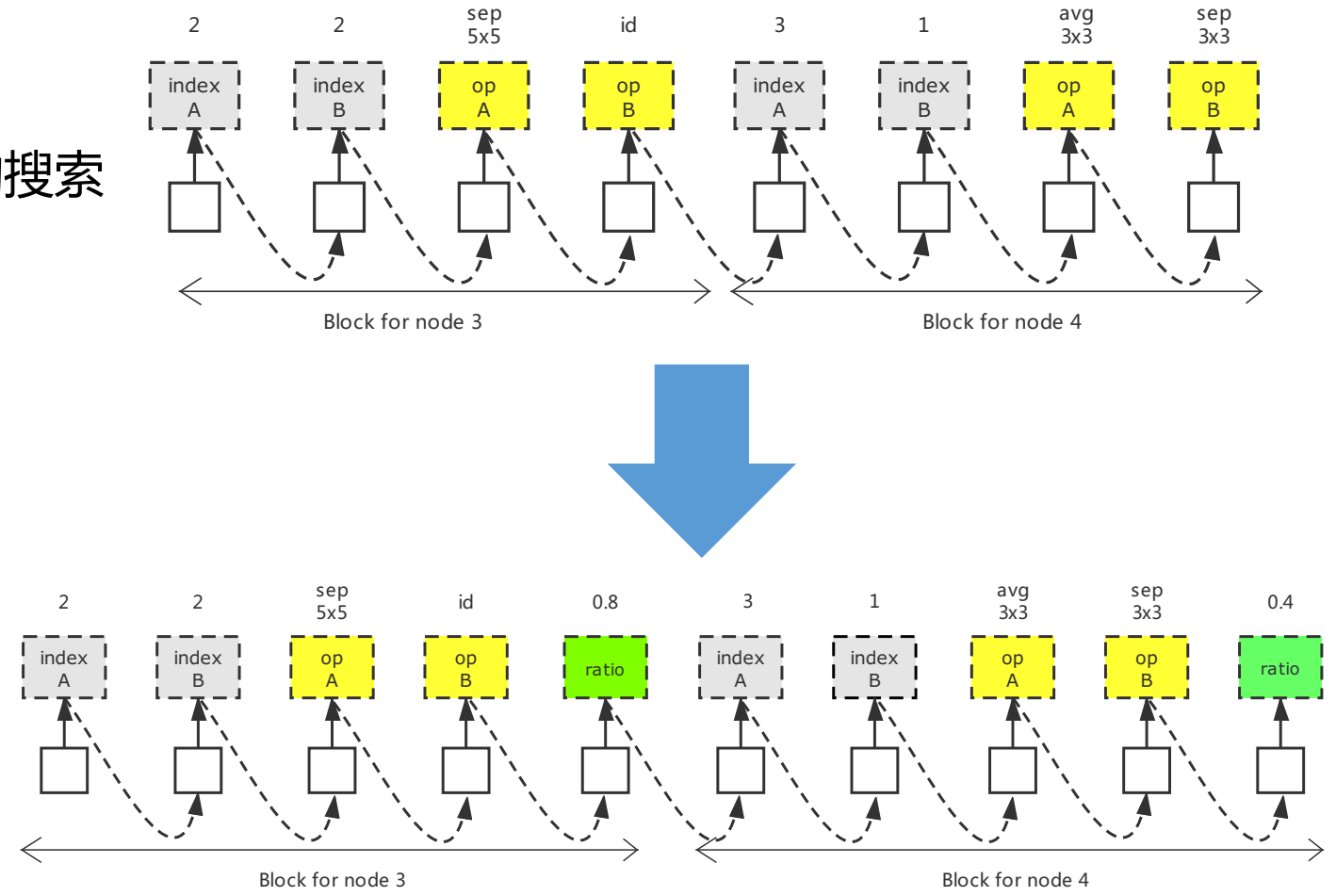
基于假剪枝和网络搜索的压缩方法

- 训练与剪裁解耦
- 剪裁比例的搜索
- 激励函数设计



剪裁比例的搜索

- 原有基础上加入每层剪裁比例ratio的搜索
  - 基于特征图通道维度裁剪
  - 按照通道绝对值总和进行排序
  - 采用补零的方式进行同层结果融合



激励函数设计

- 计算整个网络的剪裁比例总和

$$Ratio = \sum ratio(l)$$

- 计算剪裁前后的精度差

$$Gap = PruneAcc(X) - TrainAcc(X)$$

- 参数调节精度差以及剪裁比例总和对reward的影响

$$Reward = \alpha * Ratio * (1 + \beta * Gap) + TrainAcc(X)$$

实验验证

表 3-1: 算法产生的更高精度的模型与其他主流压缩方法的对比

Method	Error(%)	FLOPs(10^8)	Params(10^8)
Network Slimming(VGG)[27]	6.2	391	2.3
Network Slimming(ResNet164)	5.08	381	1.44
Filter Pruning(ResNet110)[26]	6.45	213	1.68
SSS(ResNet164)[33]	4.83	248	1.65
ENAS + micro search space[30]	3.54	751	4.6
Fake Pruning(60%)	**3.85**	**306**	**1.7**
ENAS + micro search space + CutOut	2.89	751	4.6
Fake Pruning(60%) + CutOut	**3.27**	**306**	**1.7**

表 3-2: 算法产生的更高速度的模型与其他主流压缩方法的对比

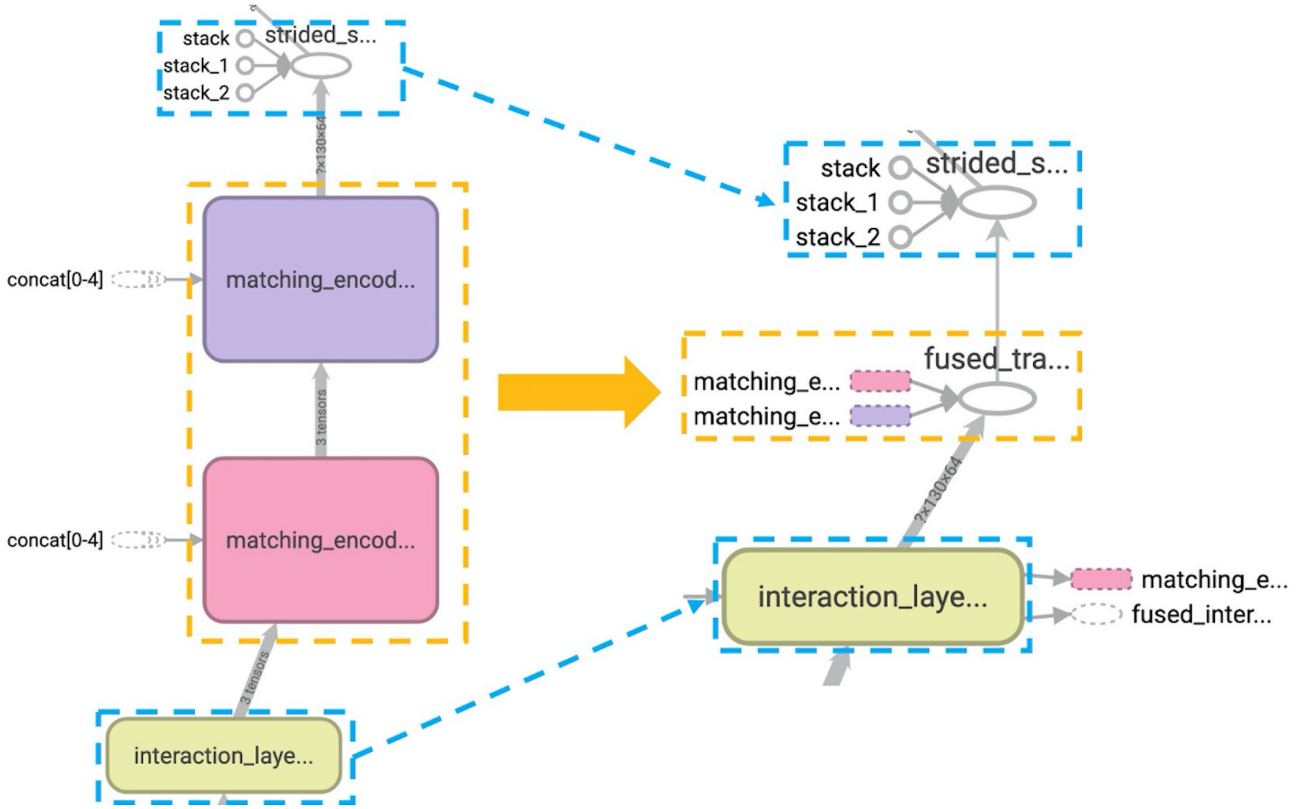
Method	Error(%)	FLOPs(M)	Params(M)
SSS(ResNeXt164)	5.43	110	1.1
AOFP-A1(VGG)[14]	6.19	215	/
SFP(ResNet110)[35]	6.17	216	/
FPGM(ResNet110)[36]	6.26	121	/
GAL-0.01(Densenet40)[37]	5.39	182.9	0.67
Fake Pruning(73%)	**4.9**	**206**	**1.1**
Fake Pruning(73%) + CutOut	**4.31**	**206**	**1.1**

表 3-3: 基于 CIFAR-10 的消融实验。

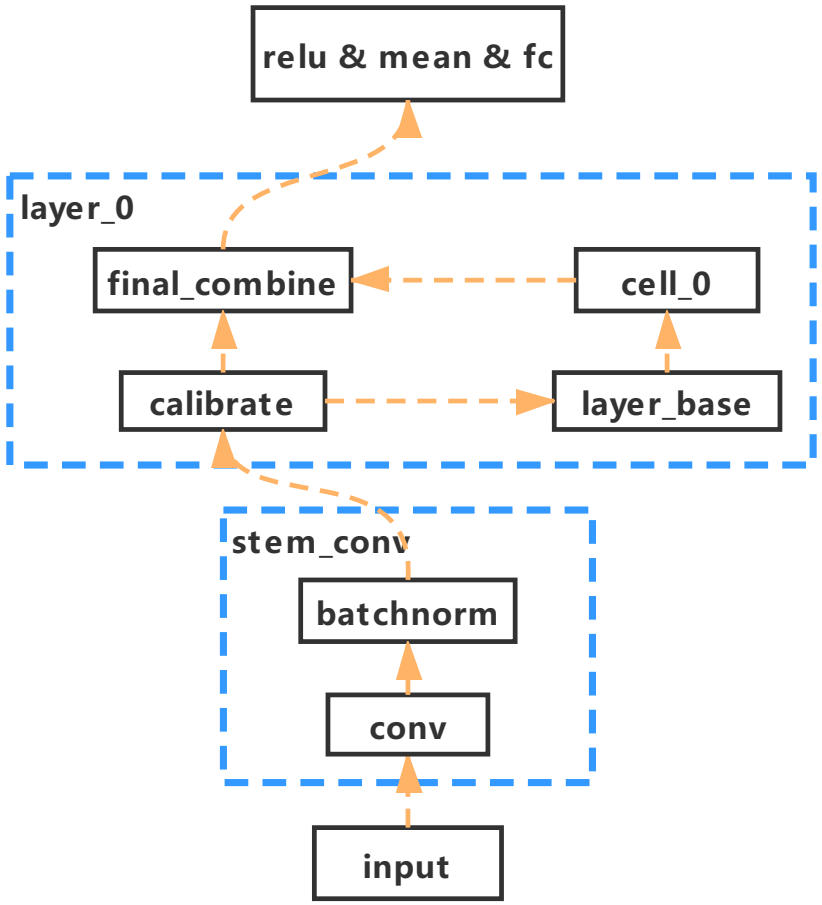
Method	Error(%)	FLOPs(M)	Params(M)
Baseline(ENAS)	3.54	751	4.6
PR + FP($\alpha = 0.02$ $\beta = 0$)	4.3	321	1.8
PR + FP($\alpha = 0.02$ $\beta = 20$)	**3.85**	**306**	**1.7**
PR + FP($\alpha = 0.5$ $\beta = 20$)	4.9	**206**	**1.1**
PR + FP($\alpha = 0.005$ $\beta = 5$)	**3.71**	422	2.4
PR + FP($\alpha = 0.005$ $\beta = 5$) + CutOut	**3.11**	422	2.4

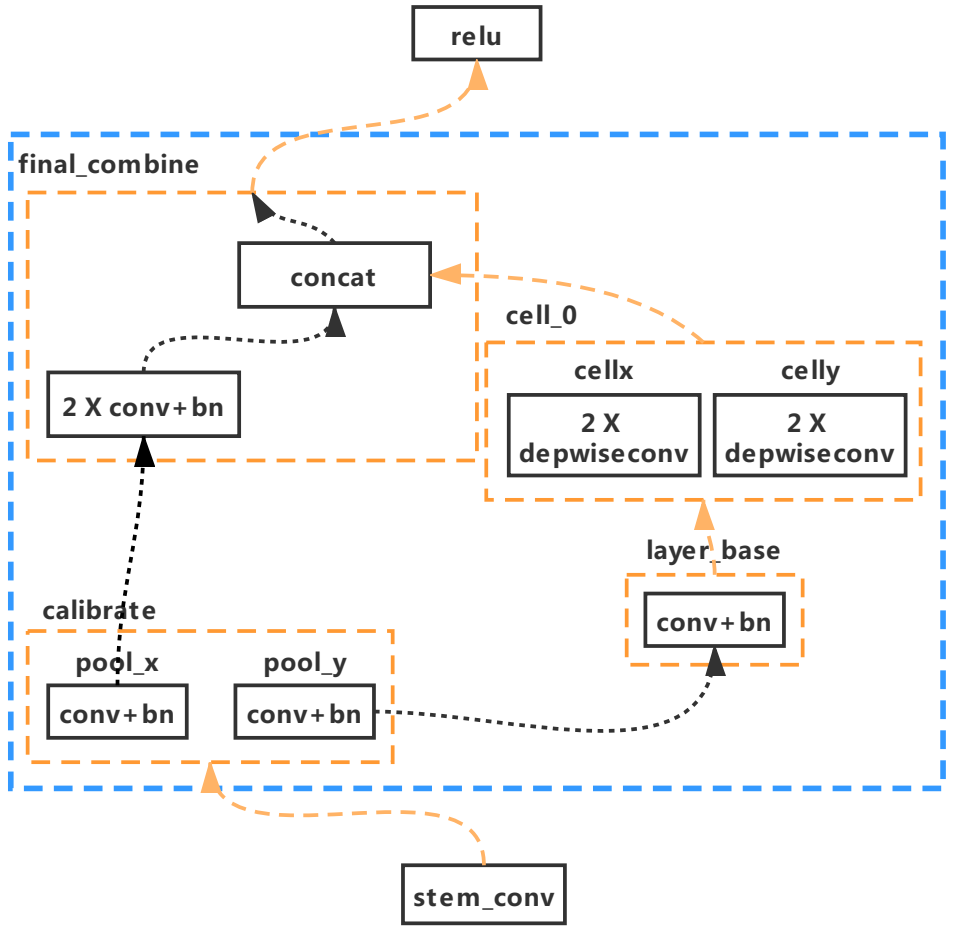
基于算子重写的工程推理加速方法

利用tensorflow算子重写的方式，将耗时模块内所有计算操作进行高性能重写并融合。

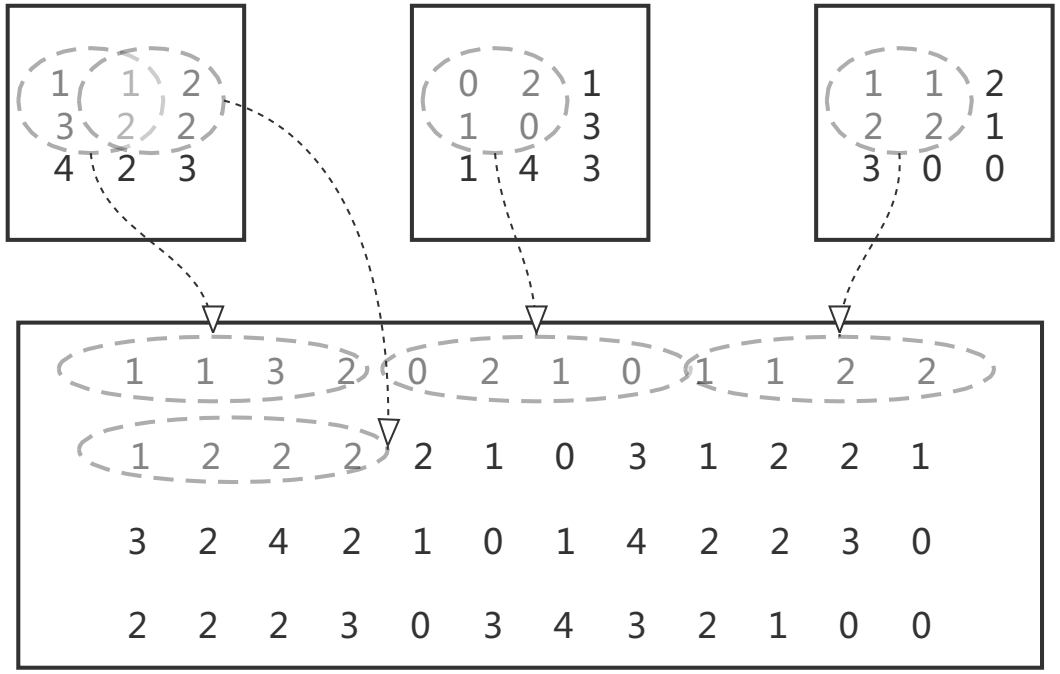


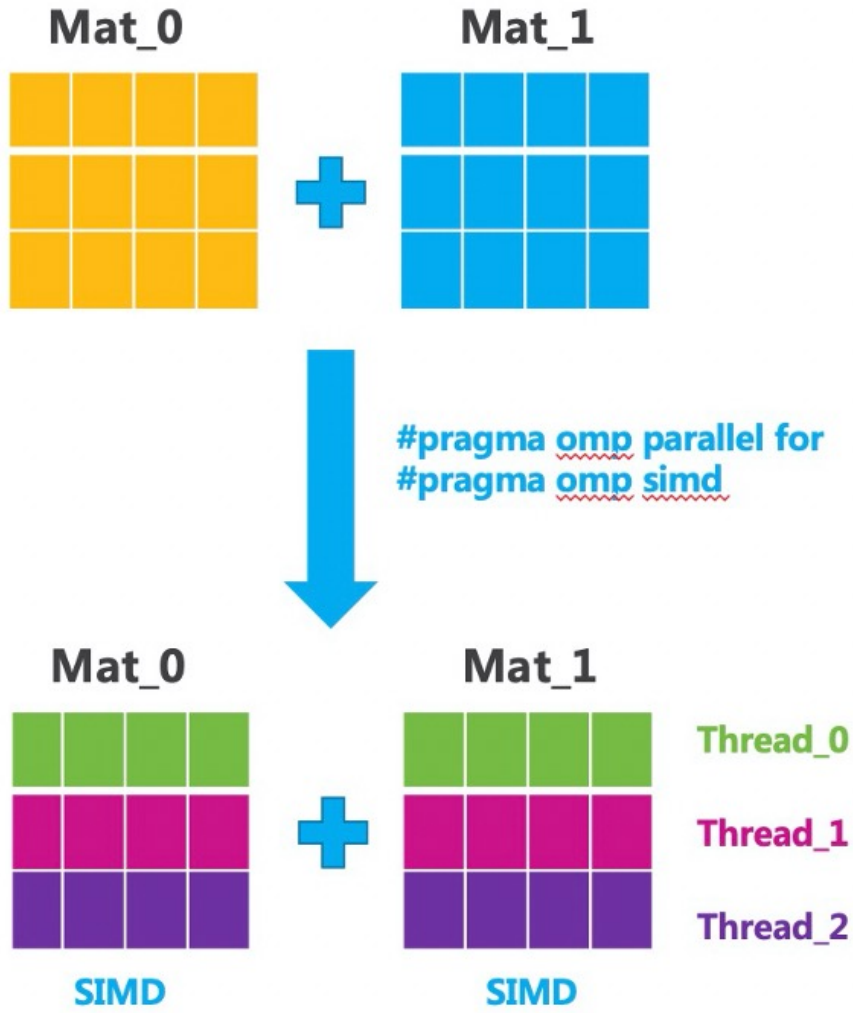
网络结构分析


relu & mean & fc
layer_0
final_combine
cell_0
calibrate
layer_base
stem_conv
batchnorm
conv
input


relu
final_combine
concat
cell_0
cellx
celly
2 X conv+bn
2 X depwiseconv
2 X depwiseconv
layer_base
conv+bn
calibrate
pool_x
pool_y
conv+bn
conv+bn
stem_conv

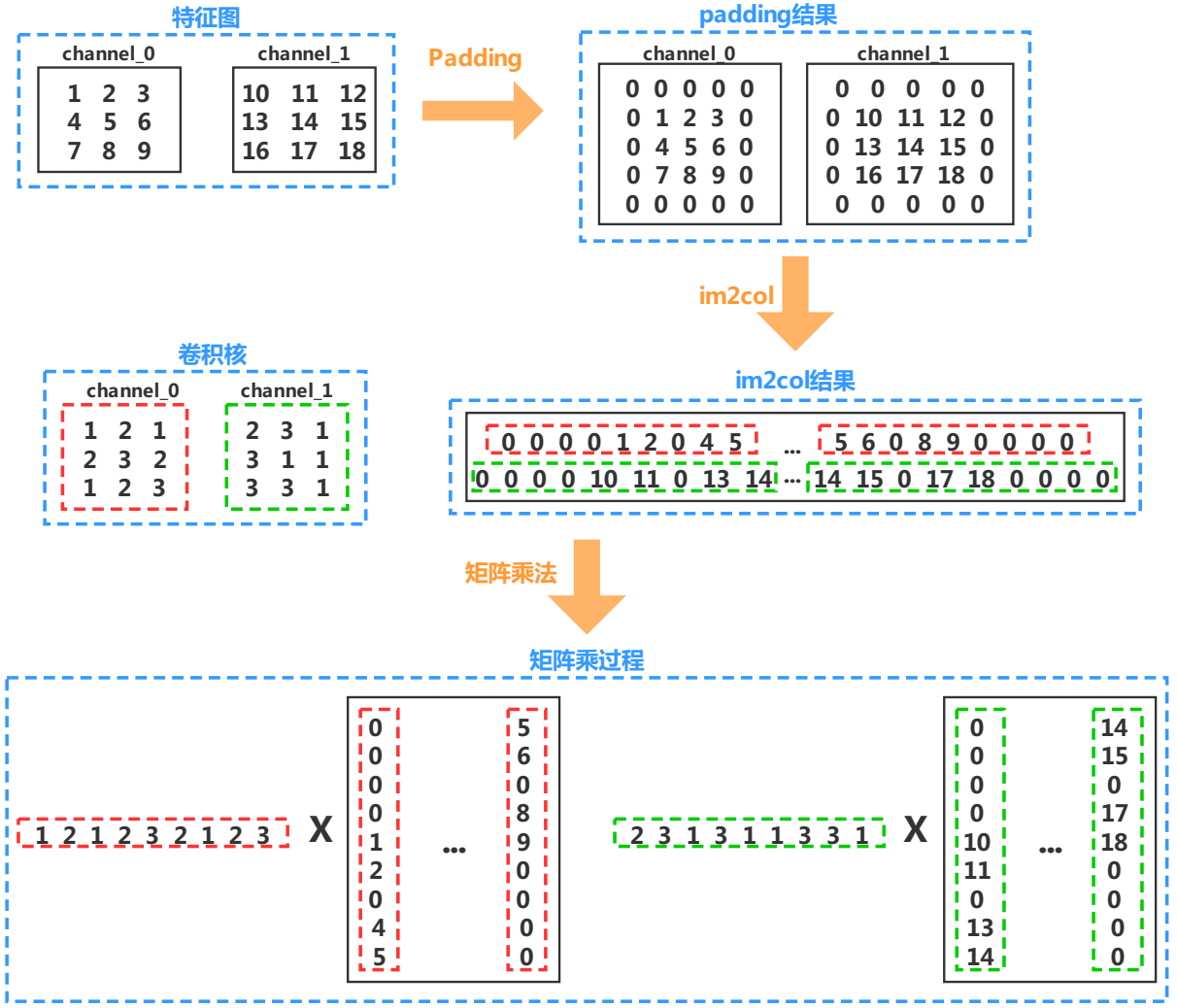
算子重写


1 1 2
3 2 2
4 2 3
0 2 1
1 0 3
1 4 3
1 1 2
2 2 1
3 0 0
1 1 3 2 0 2 1 0 1 1 2 2
1 2 2 2 2 1 0 3 1 2 2 1
3 2 4 2 1 0 1 4 2 2 3 0
2 2 2 3 0 3 4 3 2 1 0 0


Mat_0
Mat_1
#pragma omp parallel for
#pragma omp simd
Mat_0
Mat_1
Thread_0
Thread_1
Thread_2
SIMD
SIMD

逐通道卷积重写

- 首先对特征图进行padding操作
- 然后将padding结果按单通道展开
- 最终与对应的卷积核执行矩阵乘法操作



实验验证

表 4-2: stem_conv 模块优化分析

tensorflow 框架下运行耗时					
conv		bn		total	
0.257ms		0.611ms		0.868ms	
stem_conv 模块初步优化耗时					
im2col	sgemm	transpose	bn	transpose	total
0.107ms	0.082ms	0.342ms	0.276ms	0.146ms	0.953ms
stem_conv 模块最终优化耗时					
im2col	sgemm	transpose	bn	transpose	total
0.096ms	0.081ms	——	0.23ms	——	0.407ms

表 4-3: layer_0 模块优化分析

Tensorflow 框架下运行耗时						
pool_x	pool_y	layer_base	cellx	celly	final_combine	total
0.514ms	0.515ms	0.454ms	0.431ms	0.427ms	0.206	2.94ms
layer_0 模块初步优化耗时						
pool_x	pool_y	layer_base	cellx	celly	final_combine	total
0.435ms	0.294ms	0.329ms	1.328ms	1.284ms	0.131ms	3.801ms
layer_0 模块最终优化耗时						
pool_x	pool_y	layer_base	cellx	celly	final_combine	total
0.391ms	0.295ms	0.290ms	0.394ms	0.332ms	0.128ms	1.829ms

第三部分

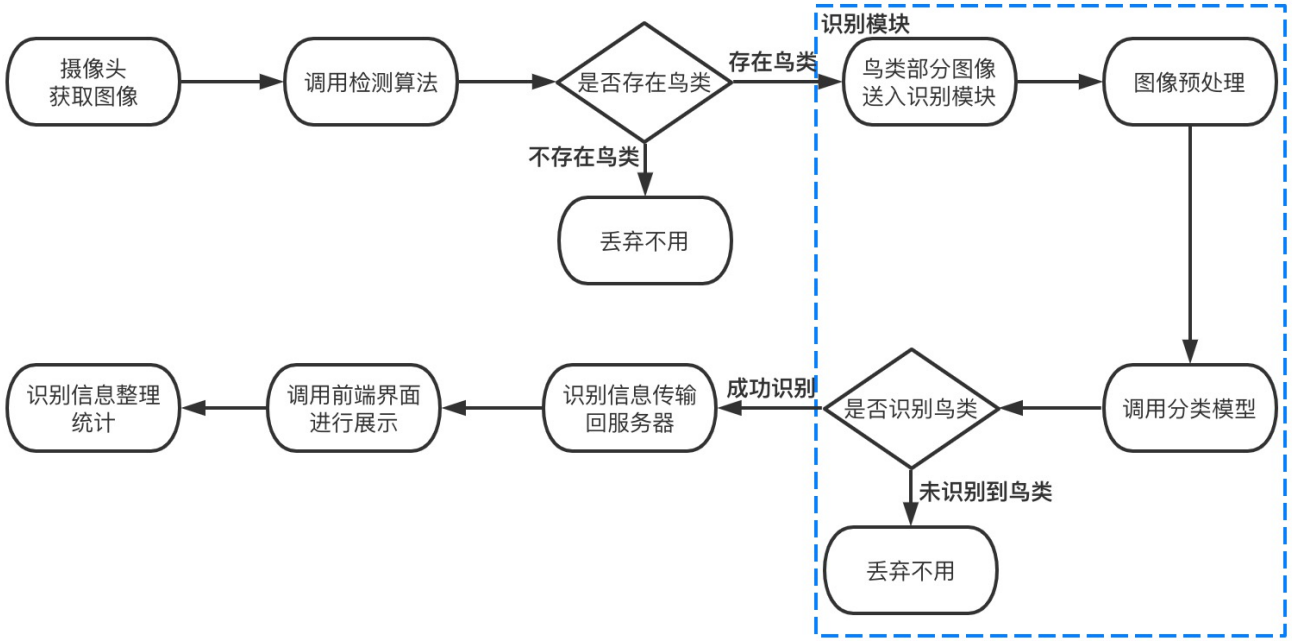
Applications

实际应用

- 鸟类识别系统

鸟类识别系统需求

- 较强的边缘计算能力
  - 减少信息流传输、降低部署成本
- 较高的鸟类识别精度
  - 需要准确追踪鸟类信息
- 较为高效的系统迭代能力
  - 应当尽可能保证系统稳定



系统效果

采用本方法加速后的模型在保证模型精度的前提下，有效提升了3.26倍的运行速度。

表 5-1: 本文提出的加速方法速度实验

模型方案	1000 次前向传播平均耗时
ENAS 搜索得到的模型 (baseline)	7.566ms
假剪枝方法 (我们提出的算法端优化)	3.808ms
假剪枝方法 + 模型推理优化 (算法与工程的结合优化)	2.320ms

南京大学
NANJING UNIVERSITY

第三部分

Summary

研究生期间工作总结

学术论文

- **Jiacheng Zhao,** Hongyan Hao, Baile Xu, Jian Zhao, Furao Shen, “Fake Pruningwith AutoML for Model Compression”

专利成果

- 申富饶，赵加成，于惠.“一种基于自编码器和tSNE降维的图片检索方法”(201910932207)

比赛奖项

- 韩峰，李雪健，赵加成，姜少魁 2019首届IKCEST“一带一路”国际大数据竞赛，国际二等奖

研究生期间工作总结 | Summary

项目

- RINC多源身份验证系统
- 基于微信小程序的签到系统

荣誉奖项

- 南京大学二等学业奖学金

南京大学
NANJING UNIVERSITY

第四部分

Summary

总结与展望

全文工作总结

基于假剪枝与网络搜索的模型压缩方法

- 提出假剪枝算法
- 搜索空间加入剪裁比例搜索
- 激励函数重新设计

基于算子重写的工程推理加速方法

- 网络结构分析
- 耗时算子高性能重写
- 针对逐通道卷积深入优化

鸟类识别系统

- 针对鸟类识别系统进行识别模型加速
- 利用假剪枝结合网络搜索得到精简模型
- 采用算子重写进行部署加速

未来工作总结 | Summary

未来工作展望

结合更多压缩方法

- 结合模型量化、网络蒸馏等方法进一步提升压缩性能

扩展压缩方法使用场景

- 将本文压缩方法拓展至目标检测、实例分割等其他视觉领域

提升算子重写效率

- 将算子重写模块化替换，提升模型部署加速效率

南京大学
NANJING UNIVERSITY

谢谢！