深度学习的三个维度：Compactness, Speed, and Accuracy

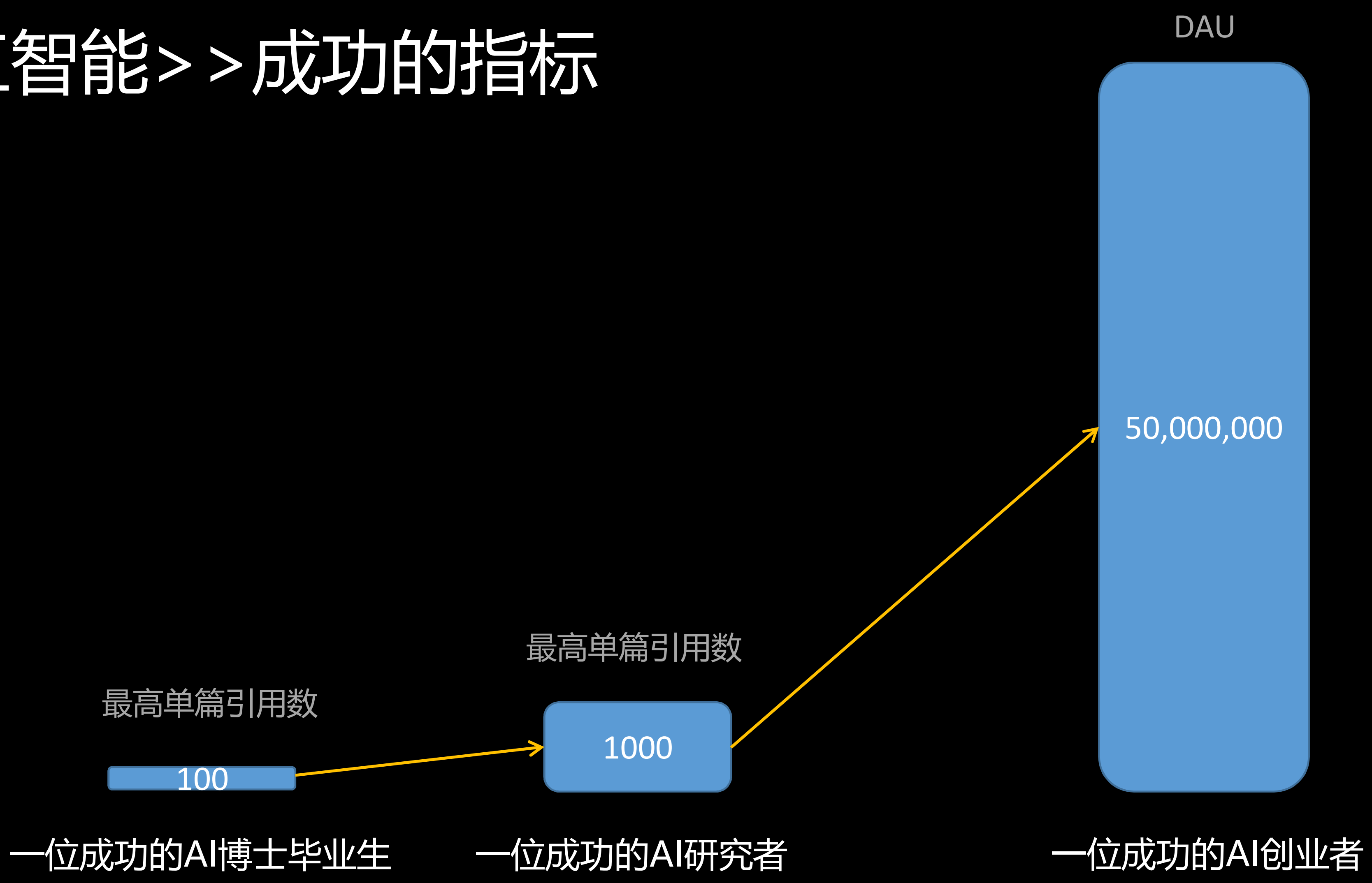
颜水成

奇虎360 副总裁、首席科学家

NUS 副教授

- 人工智能杂谈
- 深度学习研发的三个维度
  - 小、快、准
- 准: 人体与场景分割

人工智能>>成功的指标


DAU
50,000,000
最高单篇引用数
1000
最高单篇引用数
100
一位成功的AI博士毕业生
一位成功的AI研究者
一位成功的AI创业者

人工智能>>>

理想 是丰满的

现实 是骨感的

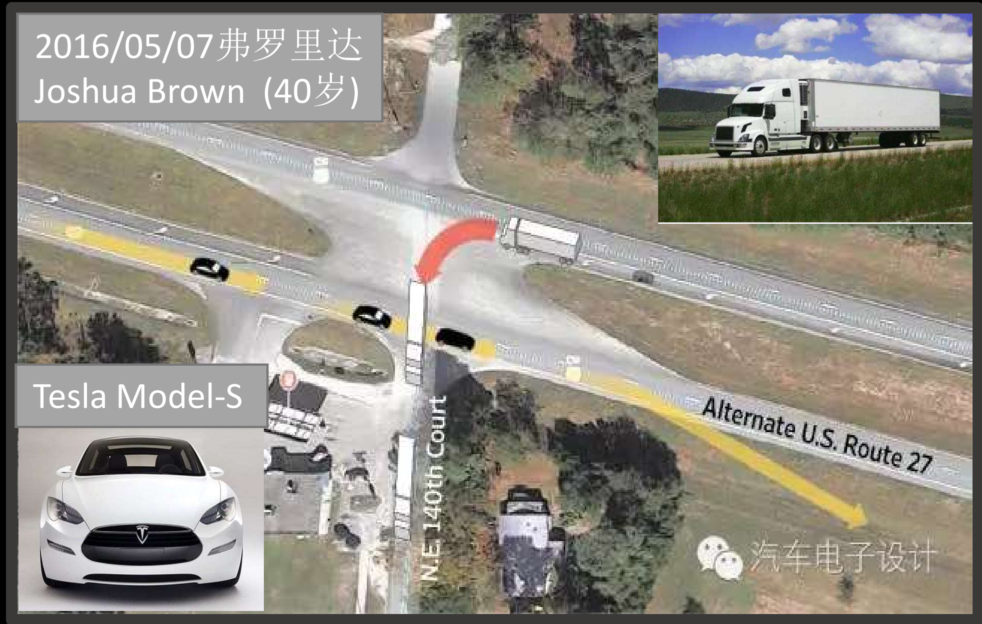
美图 629亿港币市值

Snap 336亿美金市值

自动驾驶之梦

在灾难情况下，
没有方向盘怎么办？

理想



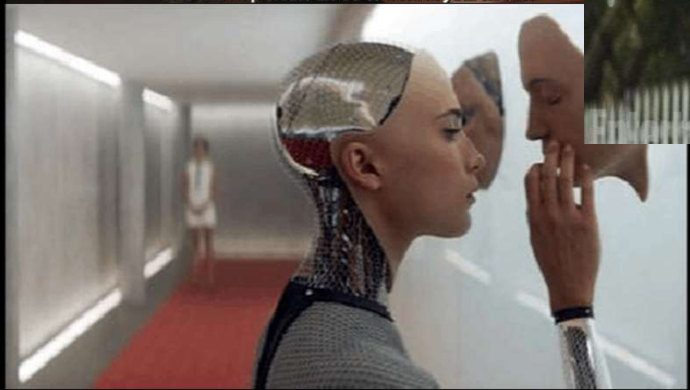
现实 [Tesla AutoPilot]

情感机器人之梦

【Her】
她

【Chappie】
超能查派

【Ex Machina】
机械姬

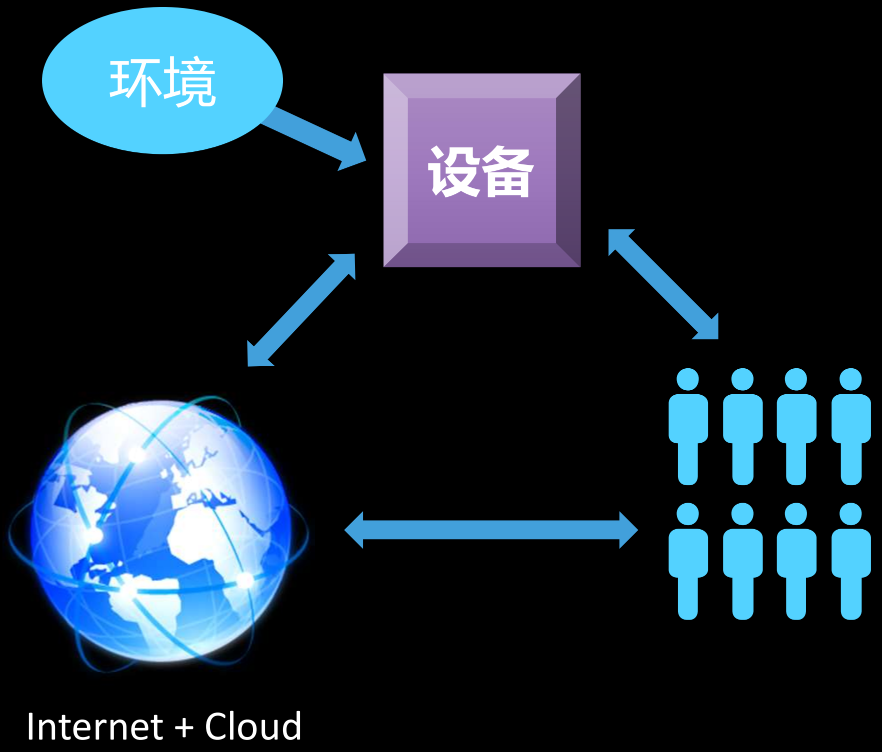


理想

现实 【ChatBot】

无辜的小胖
【我没伤人】

人工智能研发的四个主要方向



视觉分析

语音分析

语义分析(对话)

大数据分析

人工智能

人工智能研发的三种状态

初创公司

专注某一产品
或着某一领域

拼搏

百亿美金级公司

全方位支持公司多
类型的业务和创新

多维度

千亿美金级公司

每个事业群有各自的有
侧重点的人工智能团队

经常PK较量

人工智能研发的两类问题

Soft-tasks
搜索、推荐等

Hard-tasks
监控、自动驾驶等

任何新的进展都会
带来很及时的效益

必须达到一个特定
的阈值才能商业化

人工智能研发的一个现状

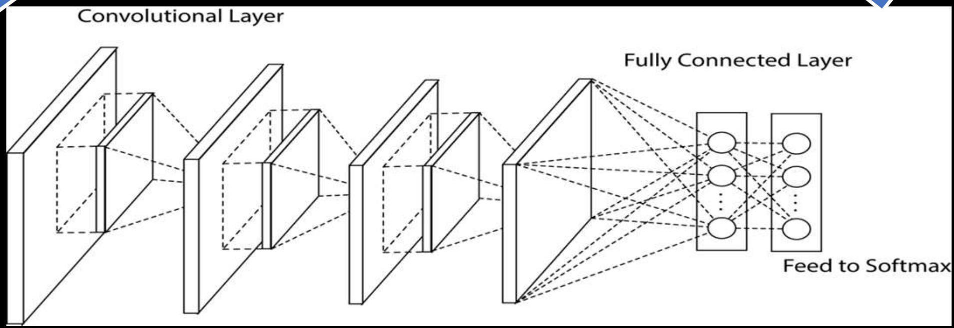
深度学习已经逐步取代各领域的传统方法

- 人工智能杂谈
- 深度学习研发的三个维度
  - 小、快、准
- 准：人体与场景分割

深度学习研发的三个维度>>小、快、准

小模型

Frequent remote upgrade



线上速度快

CPU-constrained, real-time

Background processing

预测准

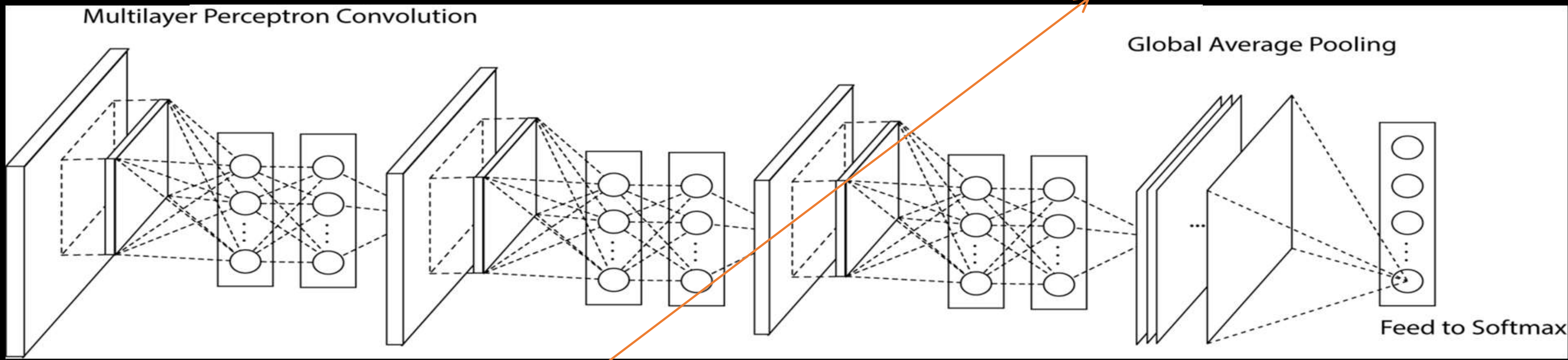
Part I: Deep Learning towards Compactness >> Model and Application

Network in Network

ICLR’ 14

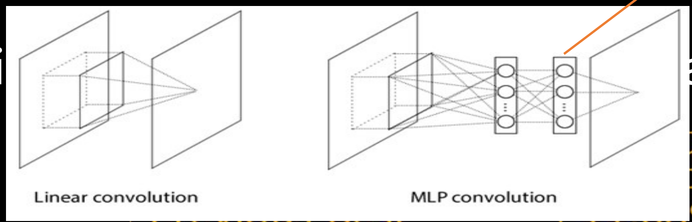
Compactness: Network in Network

NIN: complex-cell filters, pure convolutional, 1x1 convolution layers



NIN

feature maps = # classes



- Intui and more discriminative locally

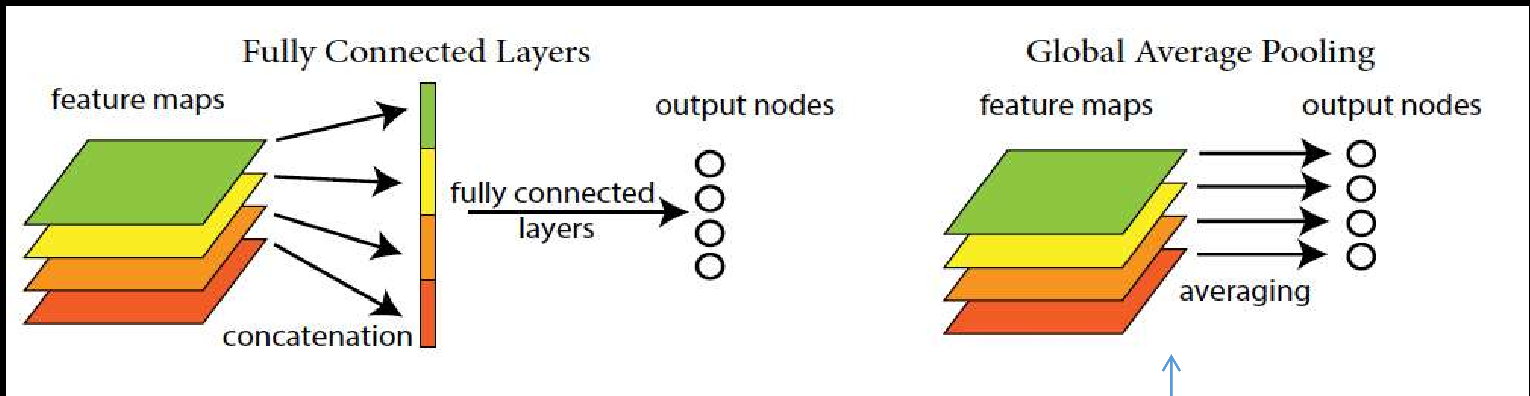
	10	Cifar-100
[illegible]	[illegible]%	38.57%
[illegible]	[illegible]%	36.30%

Can be any small networks, e.g. MLP, or others for other particular targets, but **SMALL**

Parameter # is reduced to 1/10 or less

Compactness: Network in Network

CNN NIN



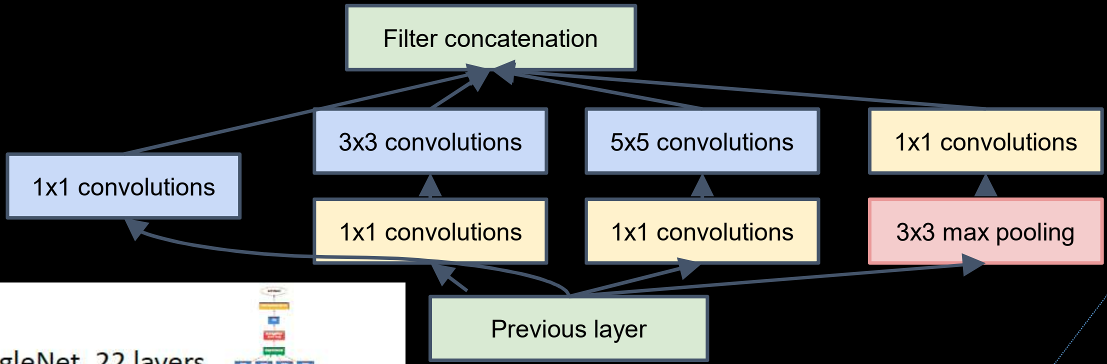
Confidence map of each category

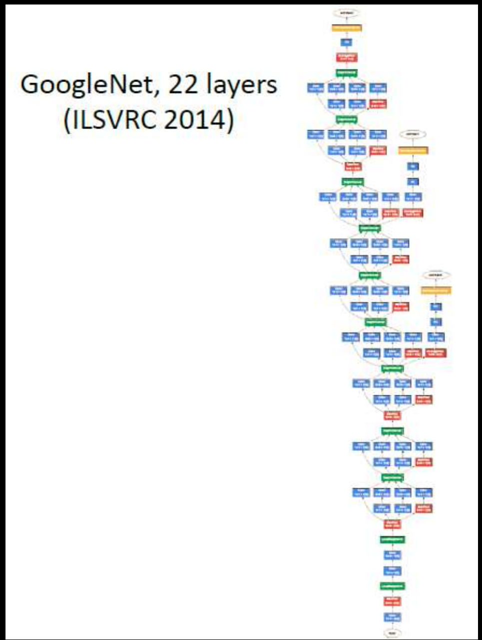
Fully convolutional [small-size model, well benefit remote model updating]

1x1 convolutional layer [complex semantic abstraction, no data matrix construction]

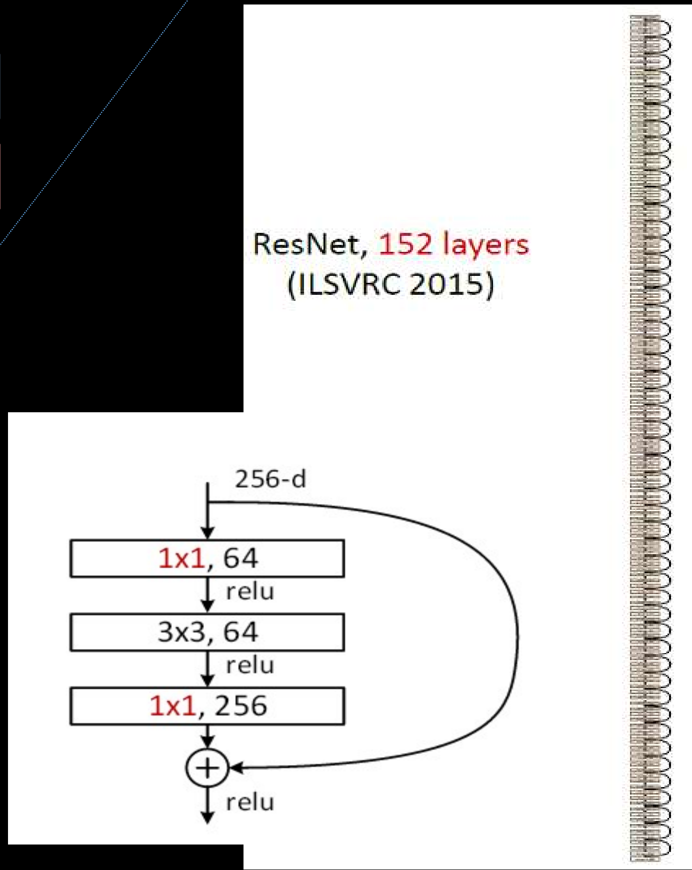
Core component for winning ImageNet Object Detection task in ILSVRC-2014

Network in Network: bring 1x1 convolution for the community





Network in a network in a network...



基于小模型的可高频更新的APP

技术原型: 准、稳、鲁棒

花椒直播: 美颜、萌颜

花椒相机: 美颜、萌颜

Part II: Deep Learning towards Efficiency>> Model and Application

More is Less

CVPR17

Efficiency: Matrix Decomposition

- Low-rank-based Acceleration

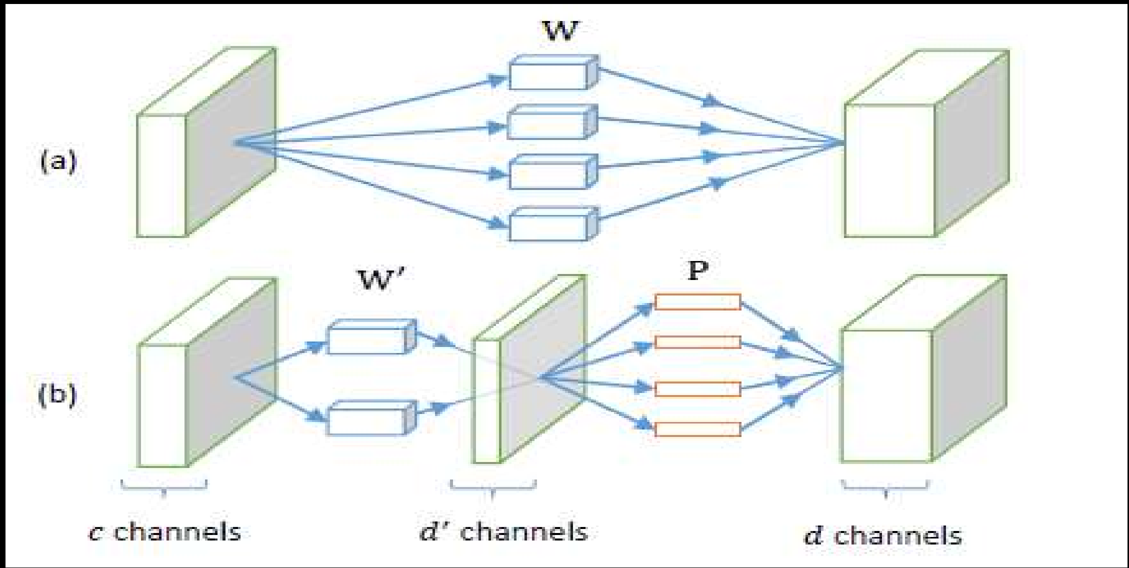
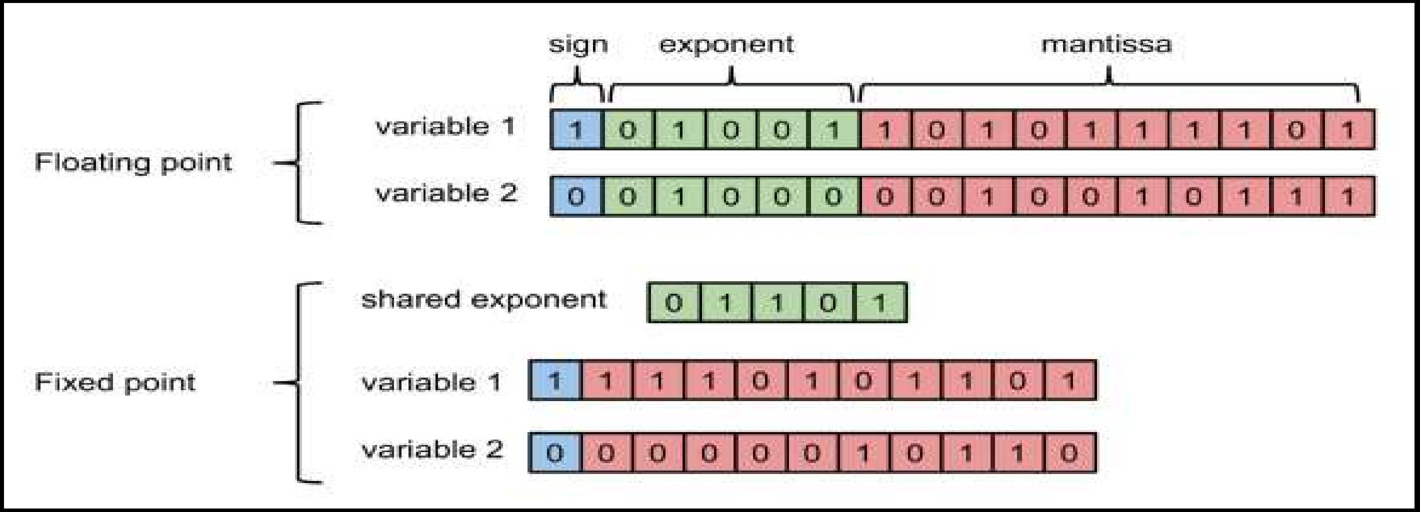


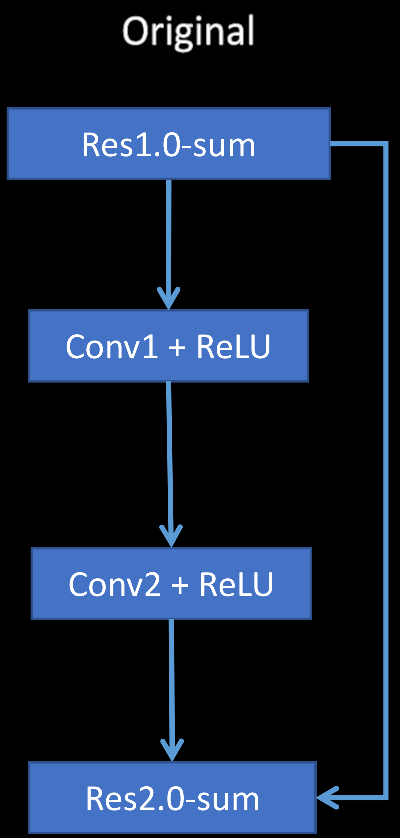
Figure: Illustration of the decomposition. (a) An original layer with complexity $O(dk^2c)$. (b) An approximated layer with complexity reduced to $O(d'k^2c) + O(dd')$.

Efficiency: Limited Numerical Precision

- Fixed-point Computation
  - 16-bit or 8-bit Integer Representation



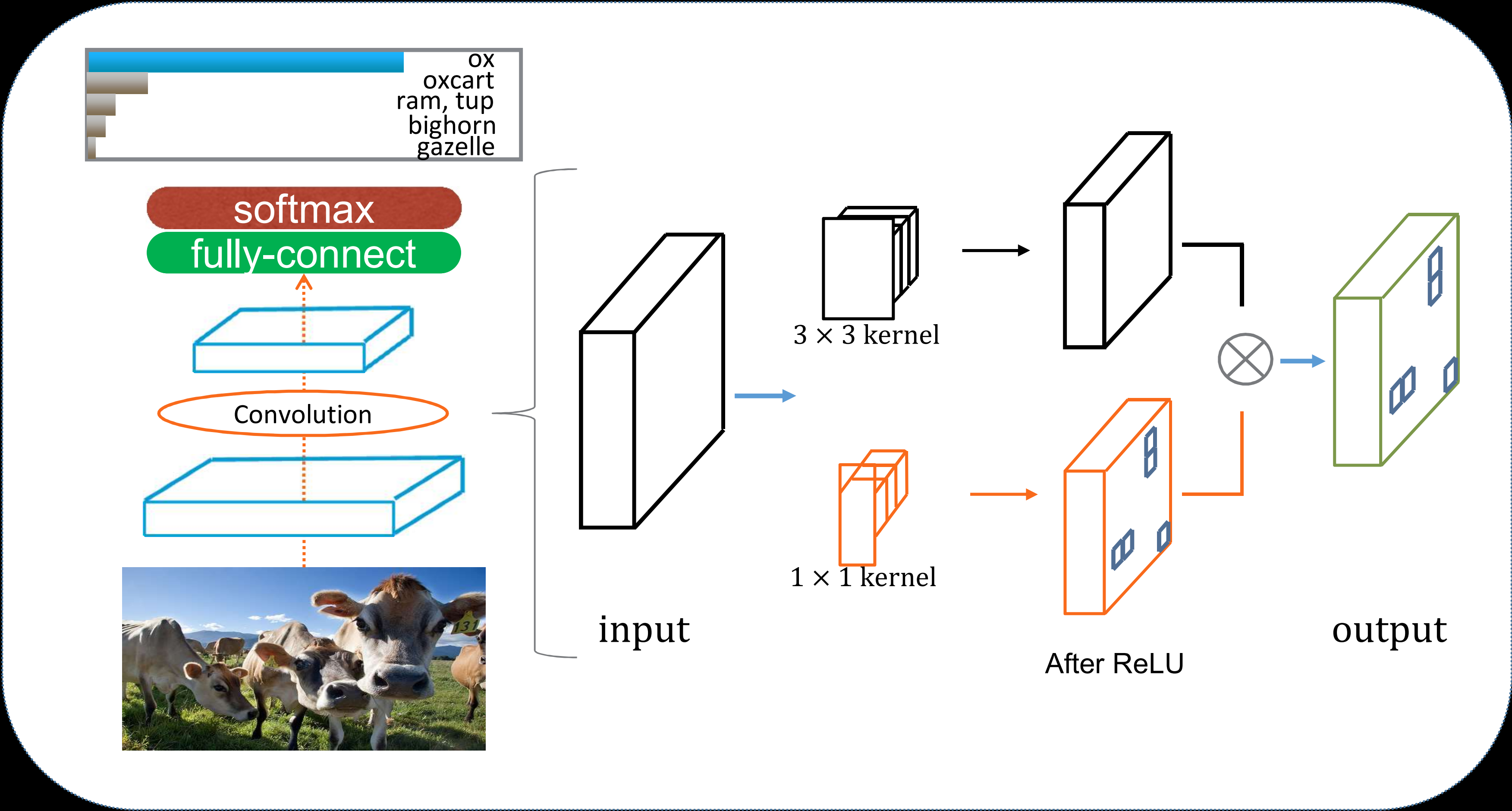
Efficiency: More (complicated structure) is Less (computation complexity)



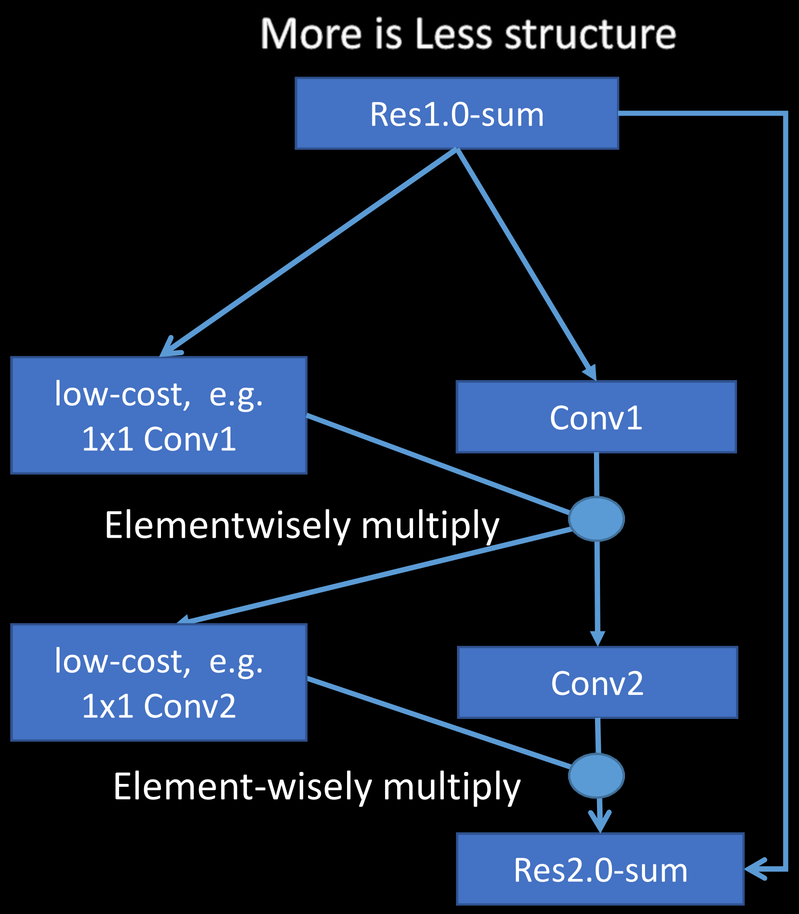
Frequently, **>40% outputs are zeros** after the ReLU 【max(0,x)】 operation, and thus their exact convolution values before ReLU are meaningless.

Can these positions be roughly estimated with very low computational cost?

Efficiency: More (complicated structure) is Less (computation complexity)



Efficiency: More (complicated structure) is Less (computation complexity)



Theoretically, model accuracy can be lossless, yet complexity is less.

If 1x1 or low-cost Conv1/2 outputs zero, then its corresponding convolution operation in conv1/2 is not required.

Efficiency: More (complicated structure) is Less (computation complexity)

CIFAR-10

	Speedup	Accuracy
ResNet-20	34.9%	91.61%
ResNet-56	41.8%	93.20%
ResNet-110	34.2%	93.69%
ResNet-164	29.1%	94.20%

CIFAR-100

	Speedup	Accuracy
WRN-40-1	36.9%	68.68%
WRN-40-2	45.6%	73.09%
WRN-52-1	25.9%	70.45%

基于快模型的应用

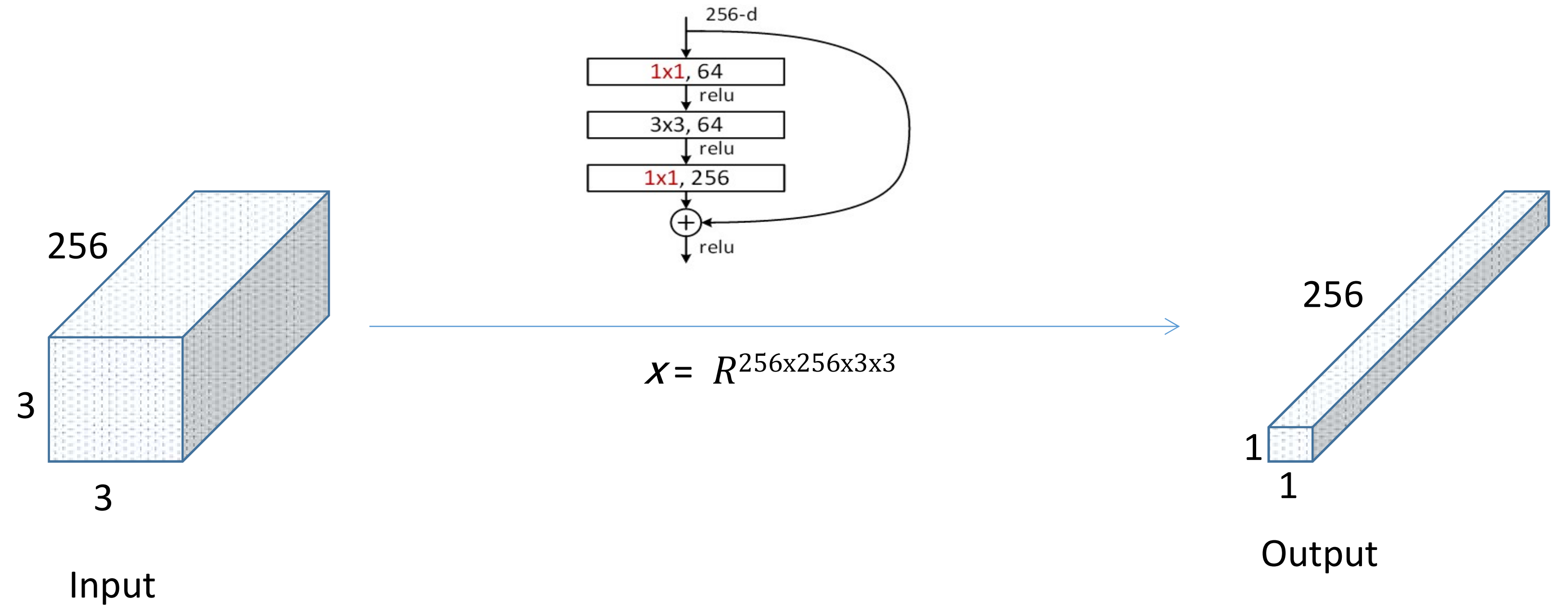
Part III: Deep Learning towards Accuracy>> Model and Application

Less is More

Cross-layer knowledge sharing towards generalization capability

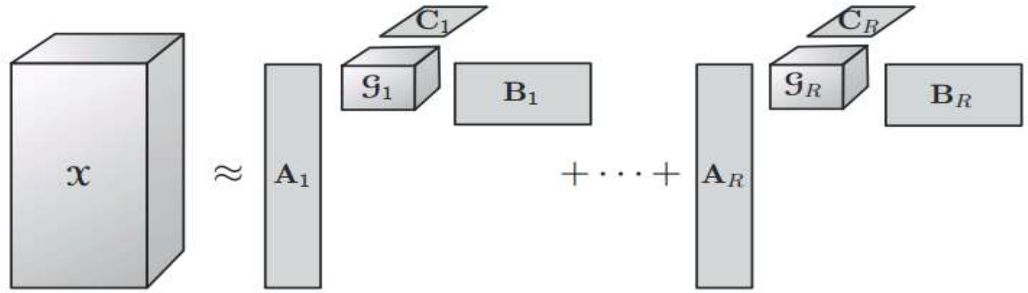
[Arxiv]

Accuracy: Secrets behind Bottleneck Structure



Accuracy: Secrets behind Bottleneck Structure

- **For a 3rd order Tensor:**



$$x \approx \sum_{r=1}^{R} \mathcal{G}_r \times_1 C_r \times_2 B_r \times_3 A_r$$

- **For a 4th order Tensor: (convolutional kernel)**

$$x = \sum_{r=1}^{R} \mathcal{G}_r \times_2 C_r \times_1 D_r$$

$$\mathcal{G}_r \in \mathbb{R}^{m \times l \times w \times h}$$

$$C_r \in \mathbb{R}^{k \times l}$$

$$D_r \in \mathbb{R}^{n \times m}$$

n=256: # output channels
k=256: # input channels
w=3: the width of the filter
h=3: the high of the filter

$x = R^{256x256x3x3}$

Improve generalization capability

Accuracy: Secrets behind Bottleneck Structure

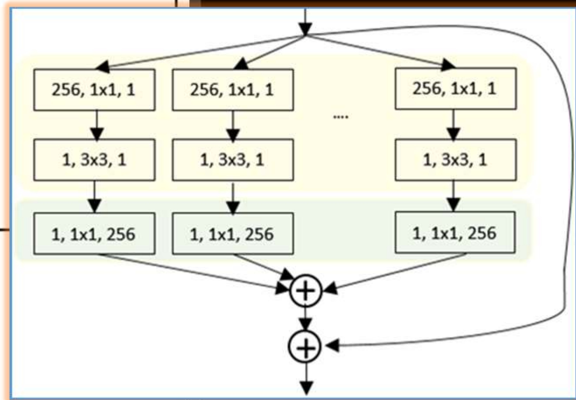
$$x = \sum_{r=1}^{R} \mathcal{G}_r \times_2 C_r \times_1 D_r$$
$$\mathcal{G}_r \in \mathbb{R}^{m\times l\times w\times h}$$
$$C_r \in \mathbb{R}^{k\times l}$$
$$D_r \in \mathbb{R}^{n\times m}$$

$R \rightarrow$ cardinality

$l = m \rightarrow$ width of bottleneck

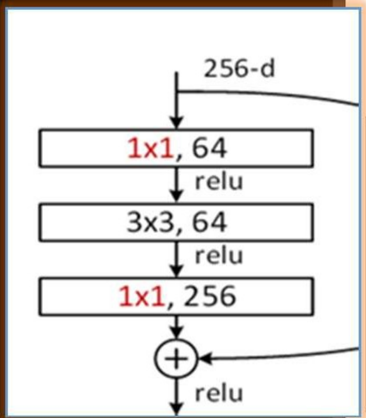
The Bottleneck in ResNeXt

$$x = \sum_{r=1}^{R} \mathcal{G}_r \times_2 C_r \times_1 D_r$$
$$\mathcal{G}_r \in \mathbb{R}^{d\times d\times w\times h}$$
$$C_r \in \mathbb{R}^{k\times d}$$
$$D_r \in \mathbb{R}^{n\times d}$$



$R \rightarrow 1$

$n \rightarrow N = \sum_{s=1}^{s=\text{number of layers}} n_s$



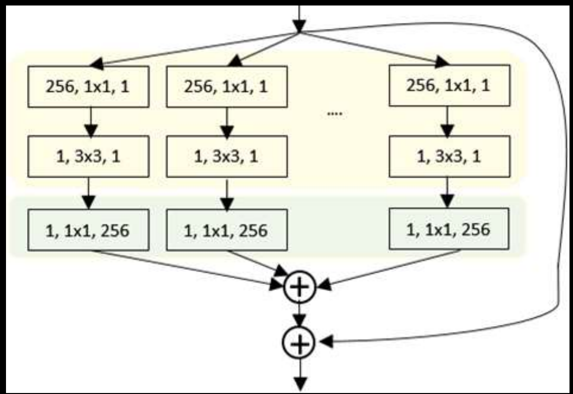
The Bottleneck in ResNet

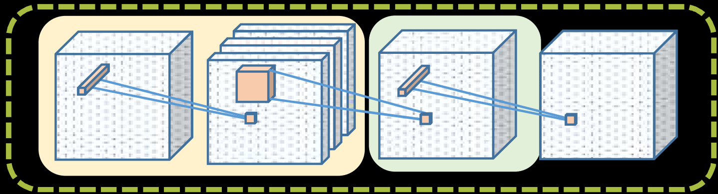
$$x = \mathcal{G} \times_2 C \times_1 D$$
$$\mathcal{G} \in \mathbb{R}^{m\times l\times w\times h}$$
$$C \in \mathbb{R}^{k\times l}$$
$$D \in \mathbb{R}^{n\times m}$$

Ours: Share across Layer

$$x = \sum_{r=1}^{R} \mathcal{G}_r \times_2 C_r \times_1 D_r$$
$$\mathcal{G}_r \in \mathbb{R}^{m\times l\times w\times h}$$
$$C_r \in \mathbb{R}^{k\times l}$$
$$D_r \in \mathbb{R}^{N\times m}$$

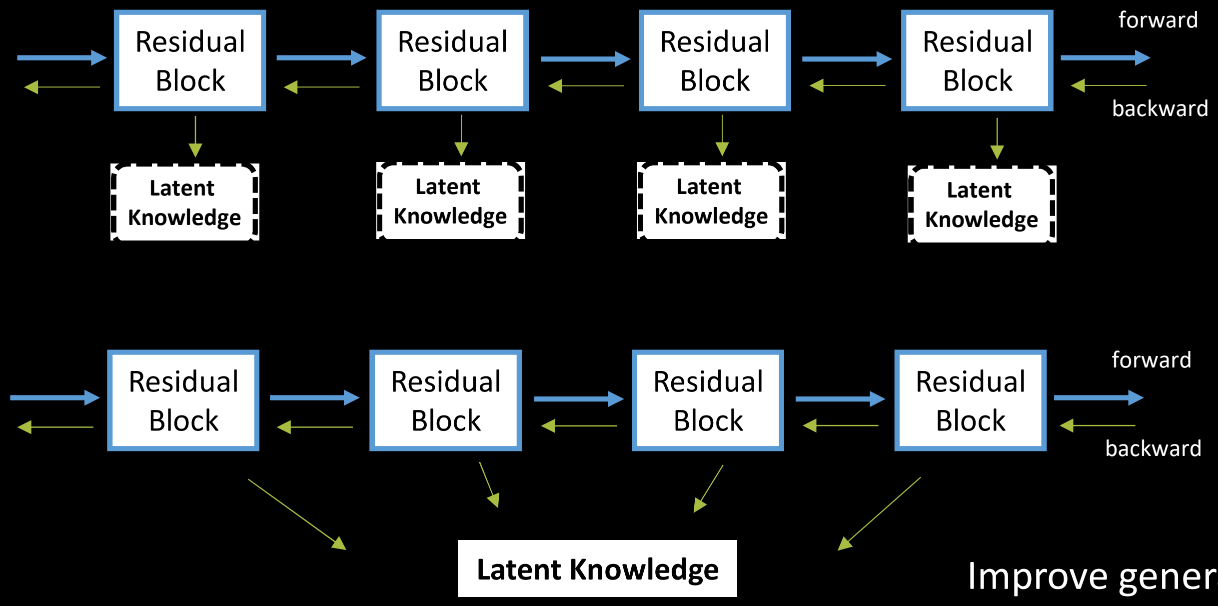
Accuracy: Secrets behind Bottleneck Structure





ResNeXt Structure

Accuracy: Share Cross-layer Knowledge



Improve generalization capability

$$x = R^{1024x256x3x3}$$

Accuracy: Share Cross-layer Knowledge

We share 6 layers @ 14x14, and fix all the other parts the same as ResNeXt to verify the effectiveness of our proposed method.

stage	output	ResNet-50	ResNeXt-50 (136x1d)	ResNeXt-50 (Nx1d)	**Proposed-50 (136x1d @x14)**
conv1	112x112	7×7, 64, stride 2	7×7, 64, stride 2	7×7, 64, stride 2	7×7, 64, stride 2
conv2	56x56	3×3 max pool, stride 2	3×3 max pool, stride 2	3×3 max pool, stride 2	3×3 max pool, stride 2
		[1x1, 64 3x3, 64, G=1 1x1, 256] $\times$ 3	[1x1, 136 3x3, 136, G=136 1x1, 256] $\times$ 3	[1x1, 136 3x3, 136, G=136 1x1, 256] $\times$ 3	[1x1, 136 3x3, 136, G=136 1x1, 256] $\times$ 3
conv3	28x28	[1x1, 128 3x3, 128, G=1 1x1, 512] $\times$ 4	[1x1, 272 3x3, 272, G=136 1x1, 512] $\times$ 4	[1x1, 272 3x3, 272, G=272 1x1, 512] $\times$ 4	[1x1, 272 3x3, 272, G=136 1x1, 512] $\times$ 4
conv4	14x14	[1x1, 256 3x3, 256, G=1 1x1, 1024] $\times$ 6	[1x1, 544 3x3, 544, G=136 1x1, 1024] $\times$ 6	[1x1, 544 3x3, 544, G=544 1x1, 1024] $\times$ 6	[**1x1, 624** **3x3, 624, G=624** 1x1, 624 1x1, 1024] $\times$ 6
conv5	7x7	[1x1, 512 3x3, 512, G=1 1x1, 2048] $\times$ 3	[1x1, 1088 3x3, 1088, G=136 1x1, 2048] $\times$ 3	[1x1, 1088 3x3, 1088, G=1088 1x1, 2048] $\times$ 3	[1x1, 1088 3x3, 1088, G=136 1x1, 2048] $\times$ 3
	1x1	global average pool 1000-d fc, softmax	global average pool 1000-d fc, softmax	global average pool 1000-d fc, softmax	global average pool 1000-d fc, softmax
# params		$\mathbf{25.5} \times 10^6$	$\mathbf{25.2} \times 10^6$	$\mathbf{24.9} \times 10^6$	$\mathbf{25.5} \times 10^6$
FLOPs		$\mathbf{4.1} \times 10^9$	$\mathbf{4.3} \times 10^9$	$\mathbf{4.3} \times 10^9$	$\mathbf{4.9} \times 10^9$

Accuracy: Share Cross-layer Knowledge

Name	Setting	Top-1
ResNet-50 [1]	1 x 64d	23.9
ResNet-200 [2]	1 x 64d	**21.7**
ResNeXt-50 [1]	2 x 40d	23.0
ResNeXt-50 [1]	32 x 4d	22.2
ResNeXt-50 (ours)	2 x 40d	22.8
ResNeXt-50 (ours)	32 x 4d	22.2
ResNeXt-50 (ours)	136 x 1d	22.1
ResNeXt-50 (ours)	N x 1d	22.5
Proposed-50	32 x 4d @x14	21.9
Proposed-50	136 x 1d @x14	**21.7**

Able to achieve comparable performance with ResNet-200 while has only the same model size as ResNet-50.

Name	Setting	Model Size	224x224		320x320 / 299x299	
			Top-1	Top-5	Top-1	Top-5
ResNet-101 [1]	1 x 64d	170 MB	22.0	6.0	-	-
ResNeXt-101 [1]	32 x 4d	170 MB	21.2	5.6	-	-
Proposed-101 @x28x14	32 x 4d	168 MB	**20.6**	**5.4**	**19.3**	**4.7**

- **1x1 convolution kernel dominates the CNNs**

Number of parameters @ conv4:

	1x1 : 3x3
ResNet-50	1 : 1
ResNext-50 (136x1d)	60 : 1
Proposed-50 (136x1d @x14)	1300 : 1

基于准模型的应用

9模型融合Top-5 错误率

2.77%

ImageNet 1000类物体识别

基于准模型的应用

1% FAR: TPR 77% → 98%

360小水滴摄像头人脸认证

- 人工智能杂谈
- 深度学习研发的三个维度
  - 小、快、准
- 准: 人体与场景分割

后备讨论：

给你一笔天使投资，你准备做款什么样的爆款APP？

1. 是不是高频刚需
2. 技术是否成熟了
3. 是否有技术壁垒

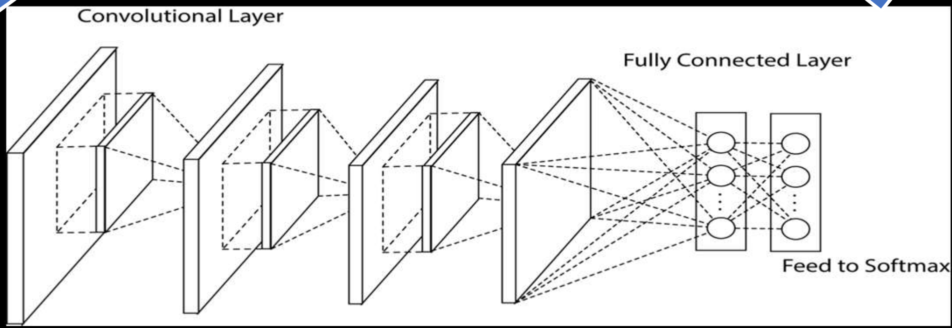
深度学习研发的三个维度>>小、快、准

小模型

Frequent remote upgrade

线上速度快

CPU-constrained, real-time



Background processing

预测准