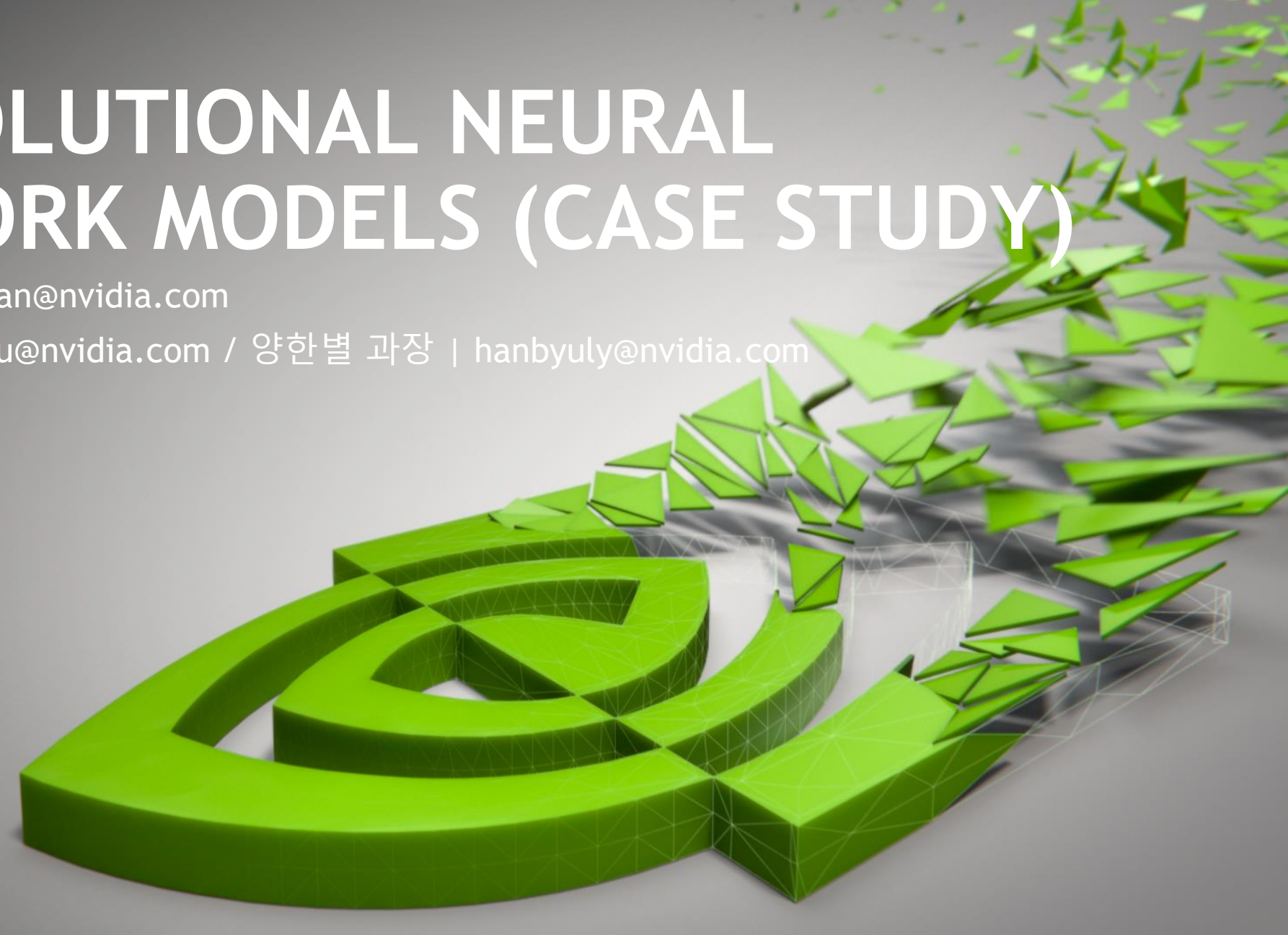
CONVOLUTIONAL NEURAL NETWORK MODELS (CASE STUDY)

한재근 과장 | jahan@nvidia.com
유현곤 부장 | hryu@nvidia.com / 양한별 과장 | hanbyuly@nvidia.com

NVIDIA



AGENDA

Imagenet Challenge

Alexnet

GoogLeNet

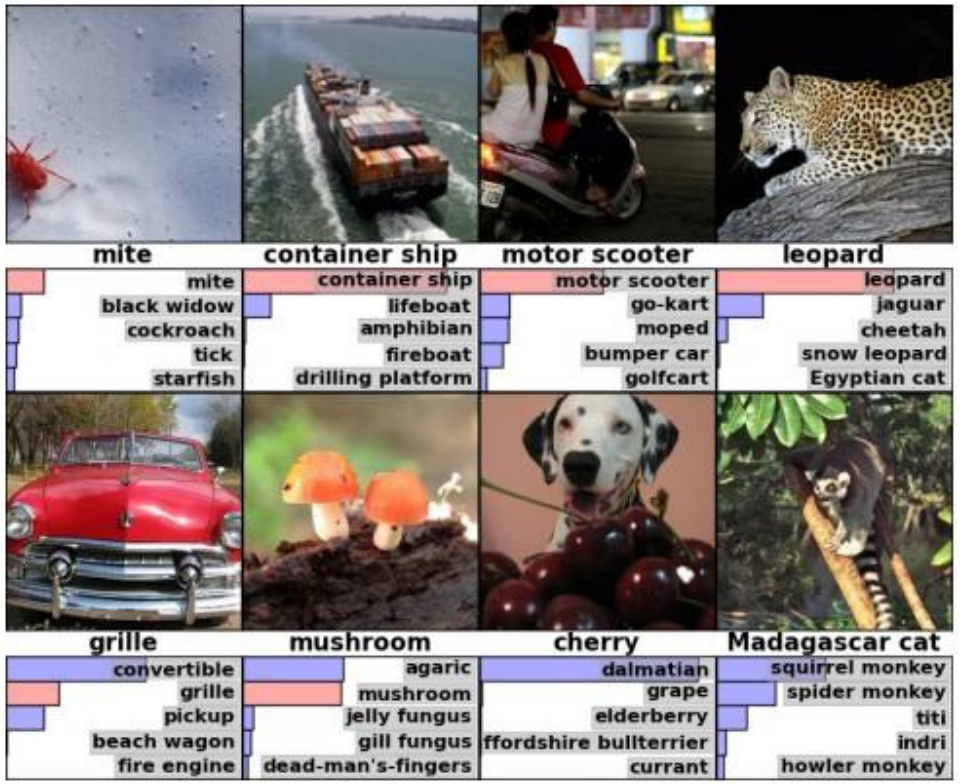
Residual Network (Resnet)

Inception

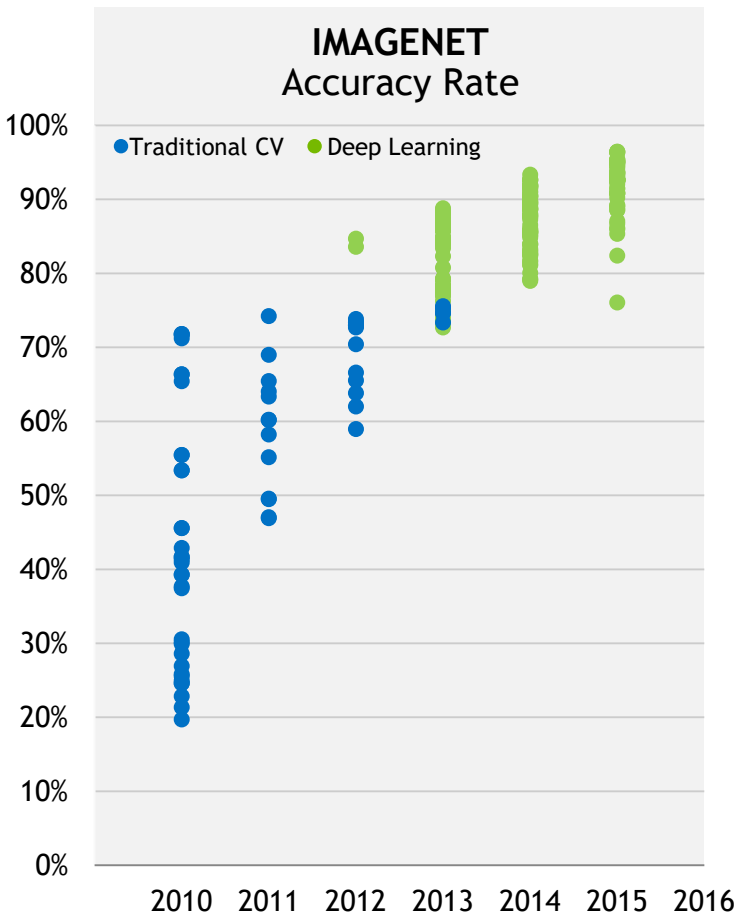
PolyNet

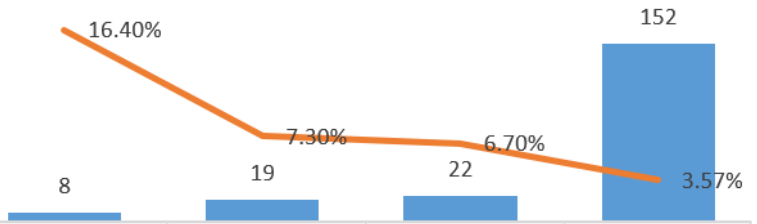
IMAGENET

- 1,000 object classes (categories)
- Images:
  - 1,281,167 train images (140G)
  - 100k test images
- Driven by Stanford Vision Lab
  - Prof. Fei-Fei Li



Deep Learning Revolution





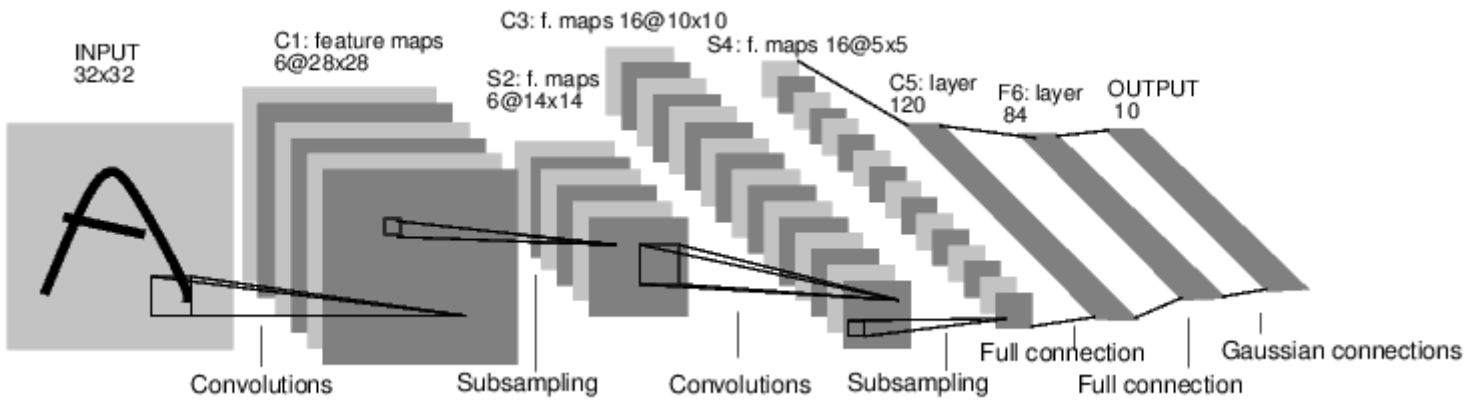
Models	AlexNet	VGG	GoogLeNet	ResNet
Challenge Year	2012	2014	2014	2015
# of layers	8	19	22	152
Top-5 error	16.4%	7.3%	6.7%	3.57%
Data Augmentation	+	+	+	+
Inception(NIN)	-	-	+	-
# of convolutions	5	16	21	151
Conv. kernel sizes	11,5,3	3	7,1,3,5	7,1,3,5
# of FCN	3	3	1	1
FCN sizes	4096,4096,1000	4096,4096,1000	1000	1000
Dropout	+	+	+	+
Local Response Normalization	+	-	+	-
Batch Normalization	-	-	-	+

NVIDIA.

LeNet-5

Once upon a time,

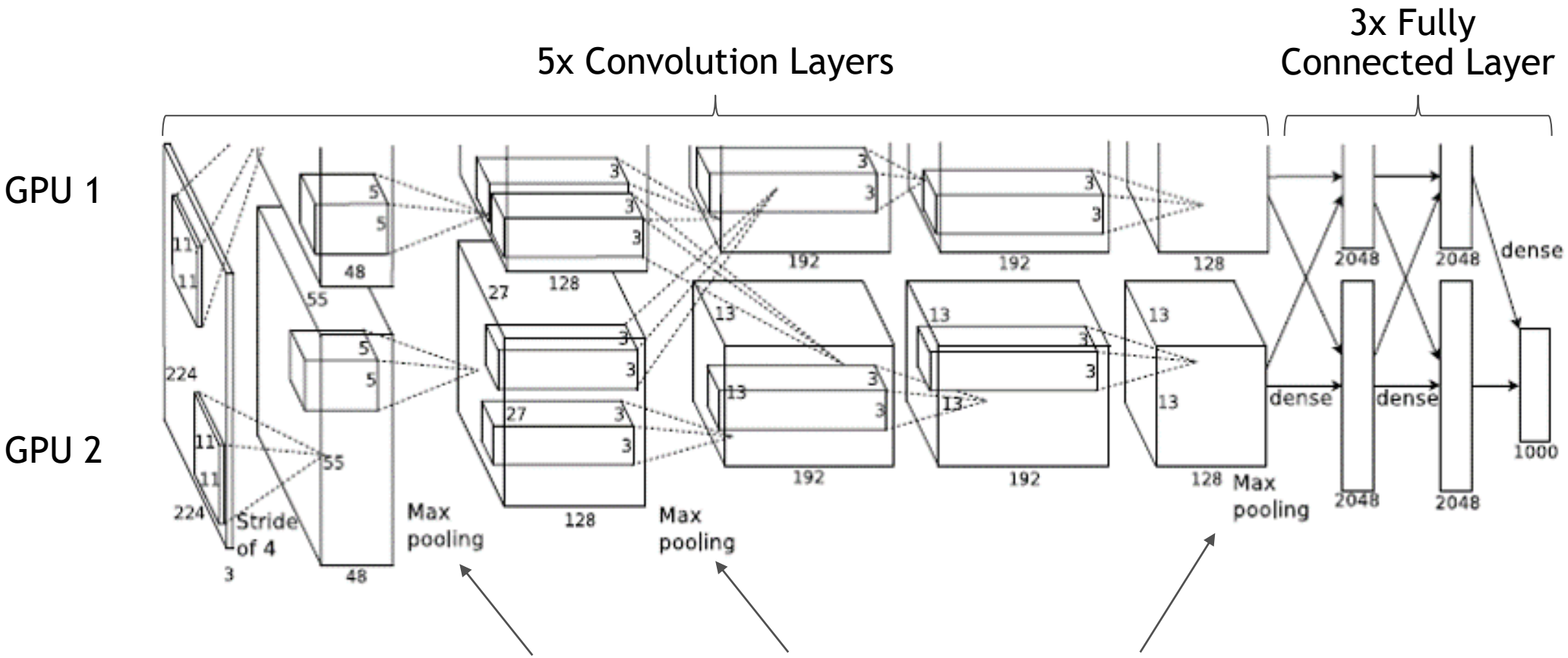
[LeCun et al., 1998]



Conv filters were 5x5, applied at stride 1
Subsampling (Pooling) layers were 2x2 applied at stride 2
i.e. architecture is [CONV-POOL-CONV-POOL-CONV-FC]

Alexnet

Convolution Neural Network with 2 GPUs

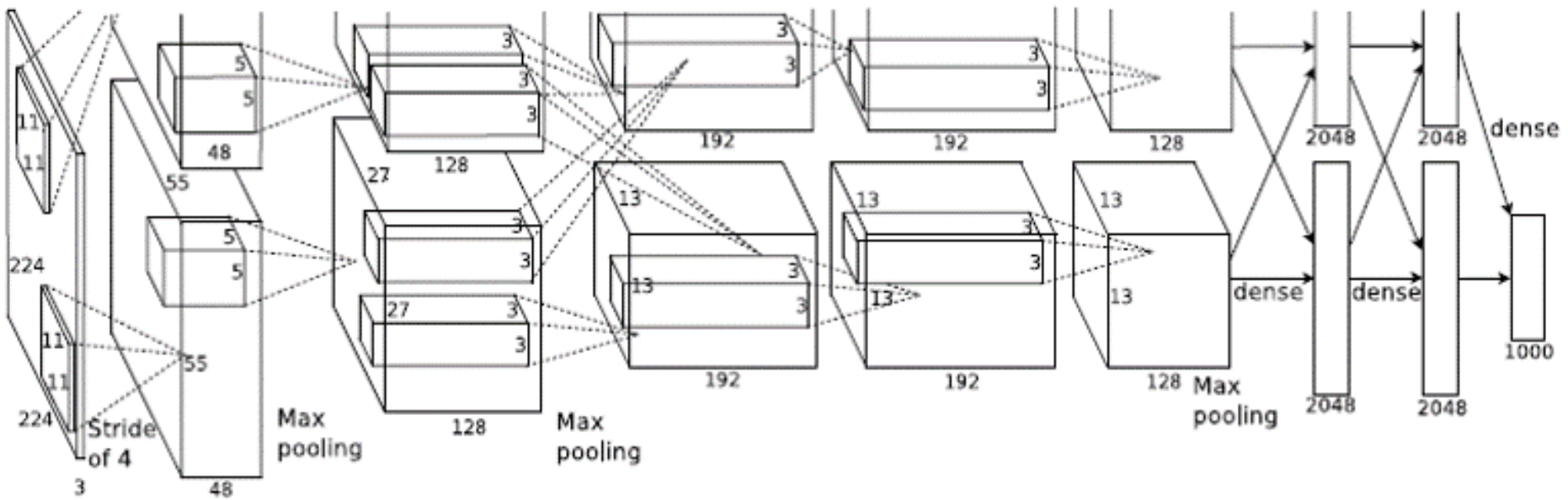


Max pooling layers / Local Response Normalization layers

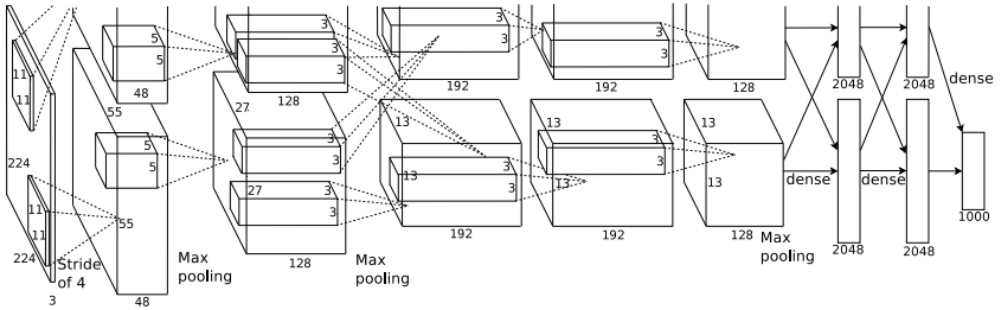
NVIDIA.

Parameter Size

Kernel size [11x11x3] [5x5x96] [3x3x256] [3x3x192] [3x3x192]



# of Kernels [96] [256] [384] [384] [256]

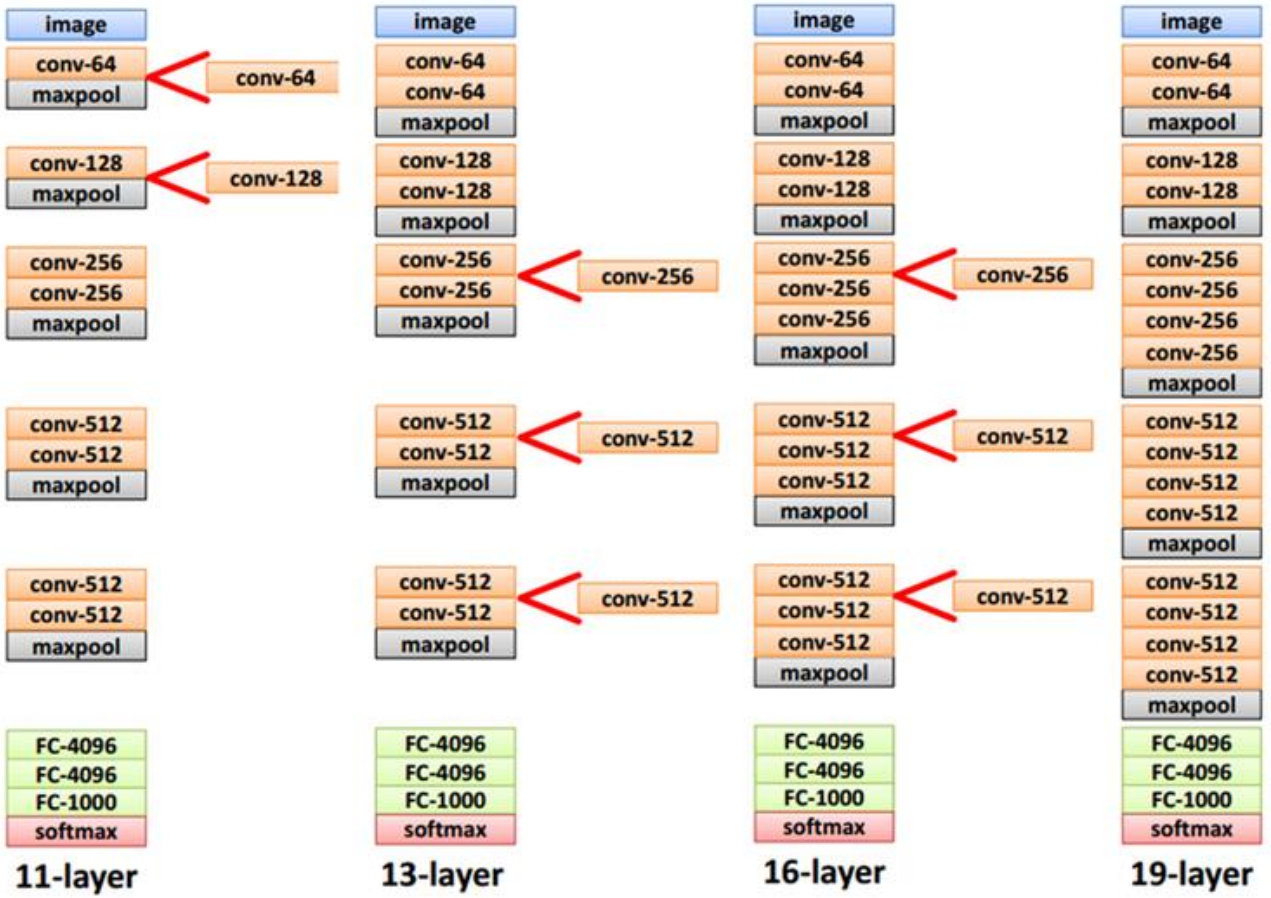


Sizing

			Parameters	Flops	Output Volume
CONV1	96 11x11x3 filter, stride 4, pad 0	[55x55x96]	35K	105M	290K
POOL1	3x3 filters, stride 2	[27x27x96]			
NORM1	Local Response Norm	[27x27x96]			
CONV2	256 5x5x96 filter, stride 1, pad 2	[27x27x256]	614K	448M	187K
POOL2	3x3 filters, stride 2	[13x13x256]			
NORM2	Local Response Norm	[13x13x256]			
CONV3	384 3x3x256 filters, stride 1, pad 1	[13x13x384]	884K	150M	65K
CONV4	384 3x3x384 filters, stride 1, pad 1	[13x13x384]	1.3M	224M	65K
CONV5	256 3x3x384 filters, stride 1, pad 1	[13x13x384]	884K	150M	65K
POOL3	3x3 filters, stride 2	[6x6x256]			
FC6	4096 neurons	[4096]			
FC7	4096 neurons	[4096]			
FC8	1000 neurons	[1000]			

VGGNet

Only 3x3 CONV stride 1, pad 1
and 2x2 MAX POOL stride 2



best model

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Table 2: **Number of parameters** (in millions).

Network	A,A-LRN	B	C	D	E
Number of parameters	133	133	134	138	144

[Simonyan and Zisserman, 2014]

NVIDIA.

VGGNet

(Not counting biases)

15M * 4 byte ~= 60M / image (inference)
138M parameters

		memory	params	flops
conv3-64	[224x224x64]	3,211,264	1,728	86,704,128
conv3-64	[224x224x64]	3,211,264	36,864	1,849,688,064
maxpool	[112x112x64]	802,816		
conv3-128	[112x112x128]	1,605,632	73,728	924,844,032
conv3-128	[112x112x128]	1,605,632	147,456	1,849,688,064
maxpool	[56x56x128]	401,408		
conv3-256	[56x56x256]	802,816	294,912	924,844,032
conv3-256	[56x56x256]	802,816	589,824	1,849,688,064
conv3-256	[56x56x256]	802,816	589,824	1,849,688,064
maxpool	[28x28x256]	200,704		
conv3-512	[28x28x512]	401,408	1,179,648	924,844,032
conv3-512	[28x28x512]	401,408	2,359,296	1,849,688,064
conv3-512	[28x28x512]	401,408	2,359,296	1,849,688,064
maxpool	[14x14x512]	100,352		
conv3-512	[14x14x512]	100,352	2,359,296	462,422,016
conv3-512	[14x14x512]	100,352	2,359,296	462,422,016
conv3-512	[14x14x512]	100,352	2,359,296	462,422,016
maxpool	[7x7x512]	25,088		
fc-4096	[1x1x4096]	4,096	102,760,448	102,760,448
fc-4096	[1x1x4096]	4,096	16,777,216	16,777,216
fc-1000	[1x1x1000]	1,000	4,096,000	4,096,000
		15,087,080	138,344,128	15,470,264,320

Heavy Compute

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

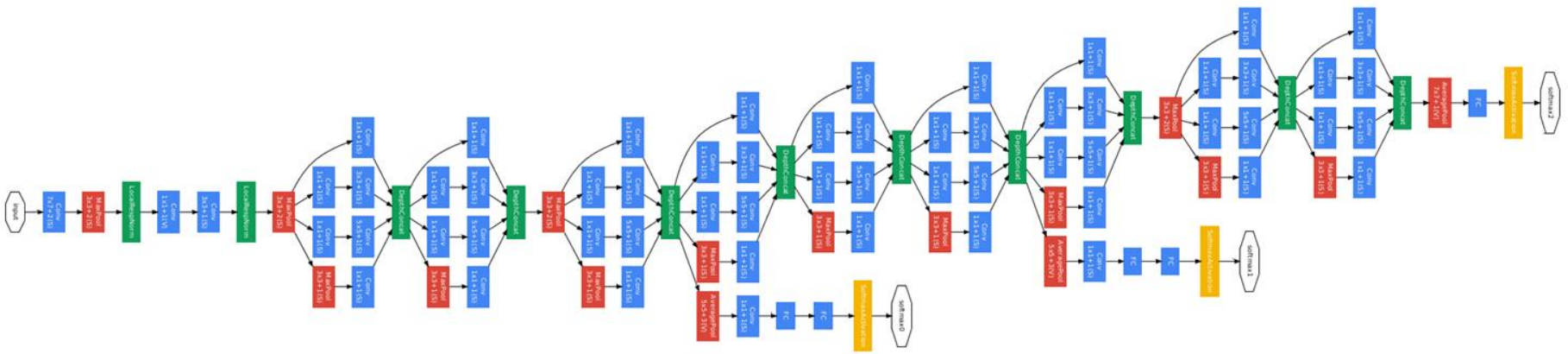
Table 2: **Number of parameters** (in millions).

Network	A,A-LRN	B	C	D	E
Number of parameters	133	133	134	138	144

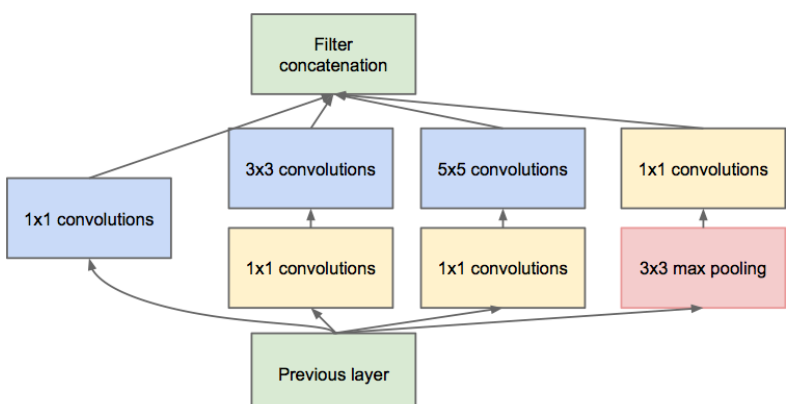
[Simonyan and Zisserman, 2014]

NVIDIA.

GoogLeNet



Inception Module



[Szegedy et al., 2014]

GoogLeNet

In depth view

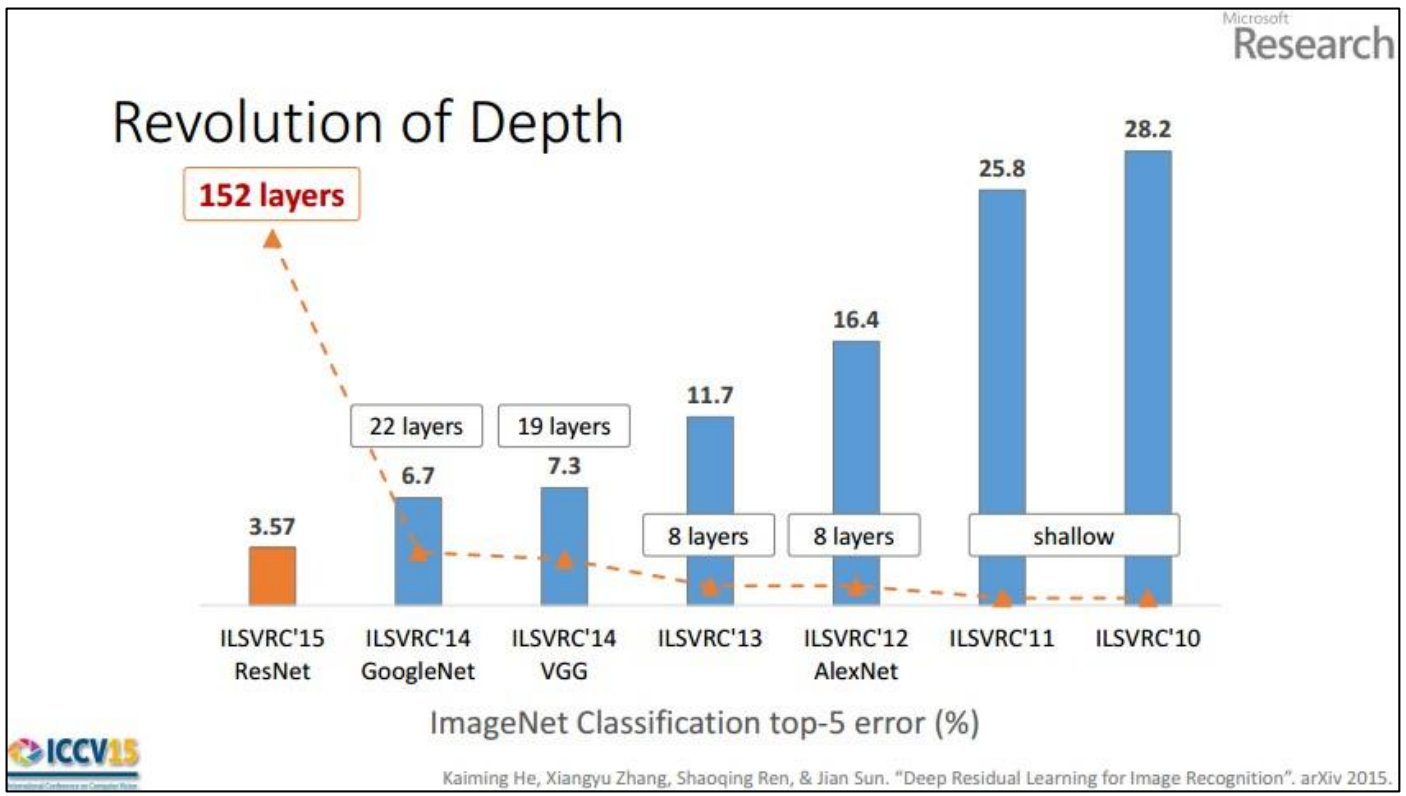
type	patch size/ stride	output size	depth	#1×1	#3×3 reduce	#3×3	#5×5 reduce	#5×5	pool proj	params	ops
convolution	7×7/2	112×112×64	1							2.7K	34M
max pool	3×3/2	56×56×64	0								
convolution	3×3/1	56×56×192	2		64	192				112K	360M
max pool	3×3/2	28×28×192	0								
inception (3a)		28×28×256	2	64	96	128	16	32	32	159K	128M
inception (3b)		28×28×480	2	128	128	192	32	96	64	380K	304M
max pool	3×3/2	14×14×480	0								
inception (4a)		14×14×512	2	192	96	208	16	48	64	364K	73M
inception (4b)		14×14×512	2	160	112	224	24	64	64	437K	88M
inception (4c)		14×14×512	2	128	128	256	24	64	64	463K	100M
inception (4d)		14×14×528	2	112	144	288	32	64	64	580K	119M
inception (4e)		14×14×832	2	256	160	320	32	128	128	840K	170M
max pool	3×3/2	7×7×832	0								
inception (5a)		7×7×832	2	256	160	320	32	128	128	1072K	54M
inception (5b)		7×7×1024	2	384	192	384	48	128	128	1388K	71M
avg pool	7×7/1	1×1×1024	0								
dropout (40%)		1×1×1024	0								
linear		1×1×1000	1							1000K	1M
softmax		1×1×1000	0								

- Only 5 million params! (Removes FC layers completely)

Compared to AlexNet:
- 12X less params
- 2x more compute
- 6.67% (vs. 16.4%)

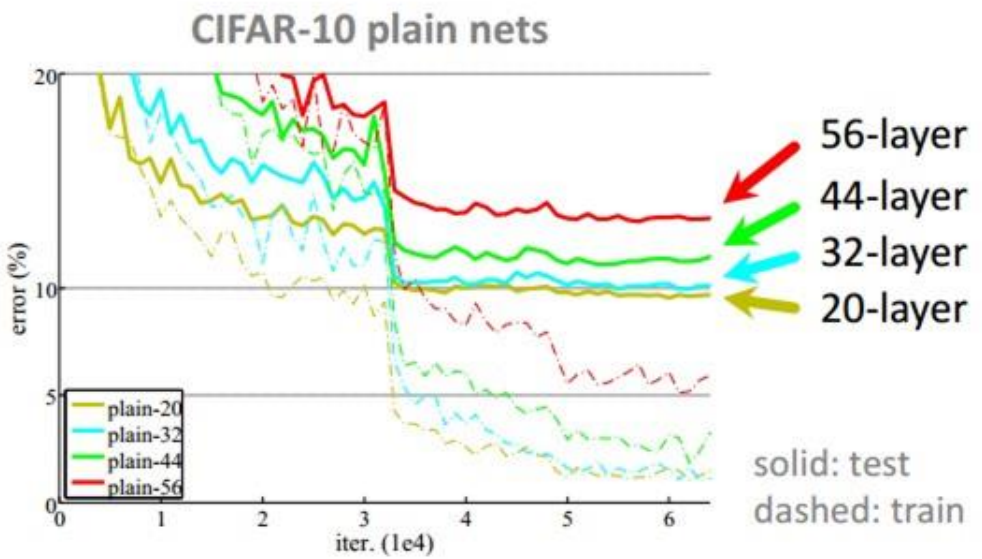
ResNet

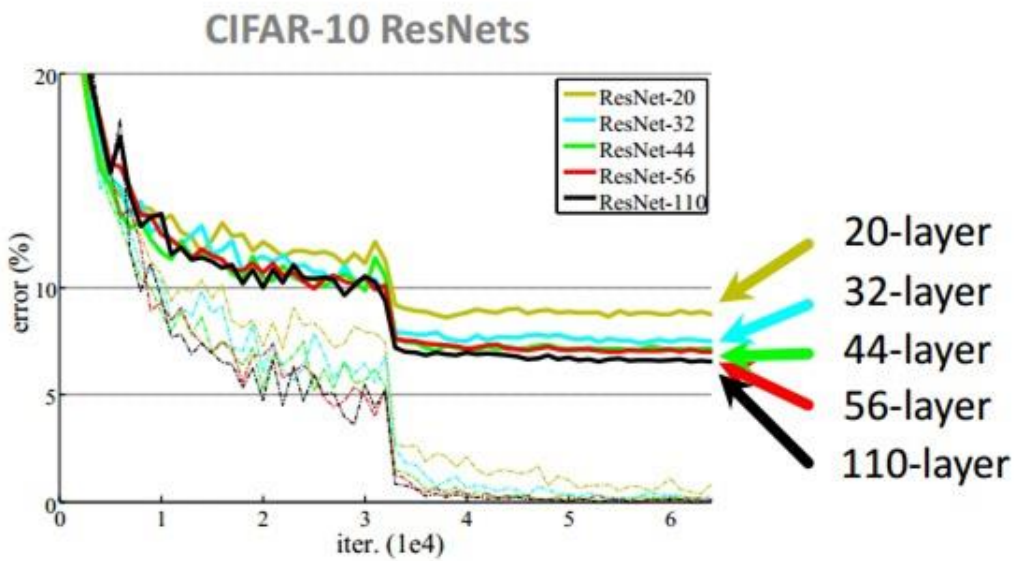
152 layers !!



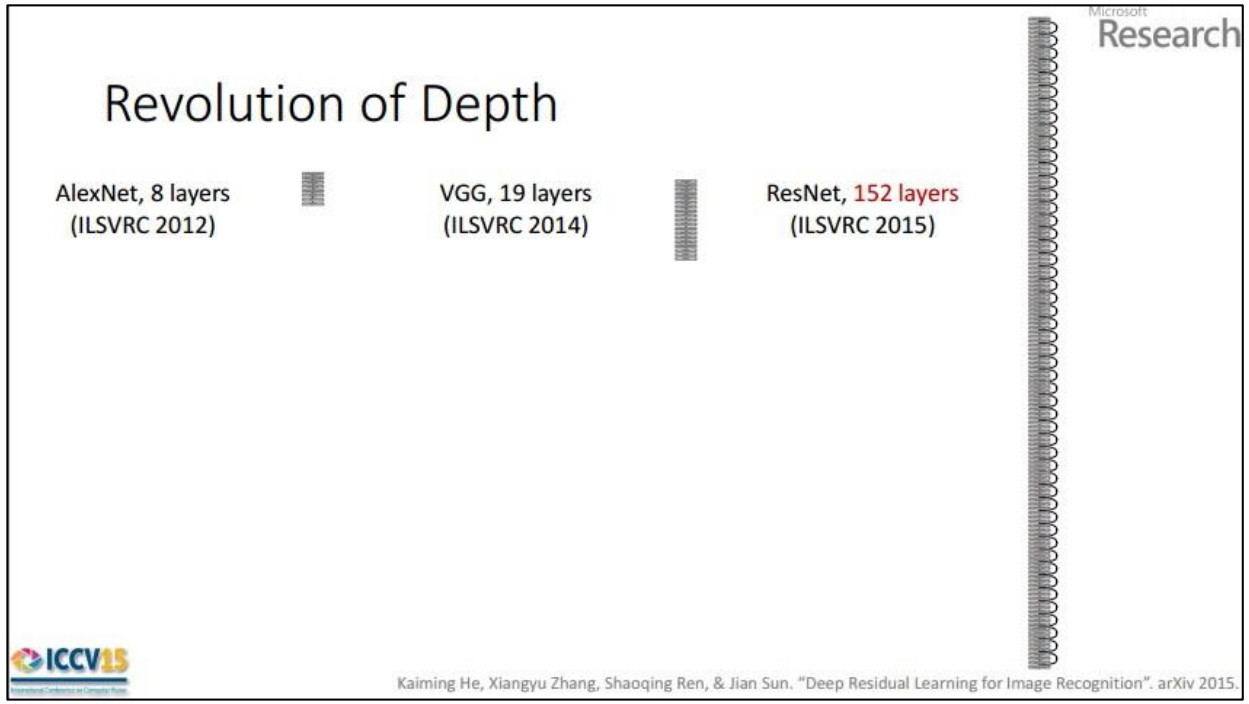
[He et al., 2015]

CIFAR10 Experiments





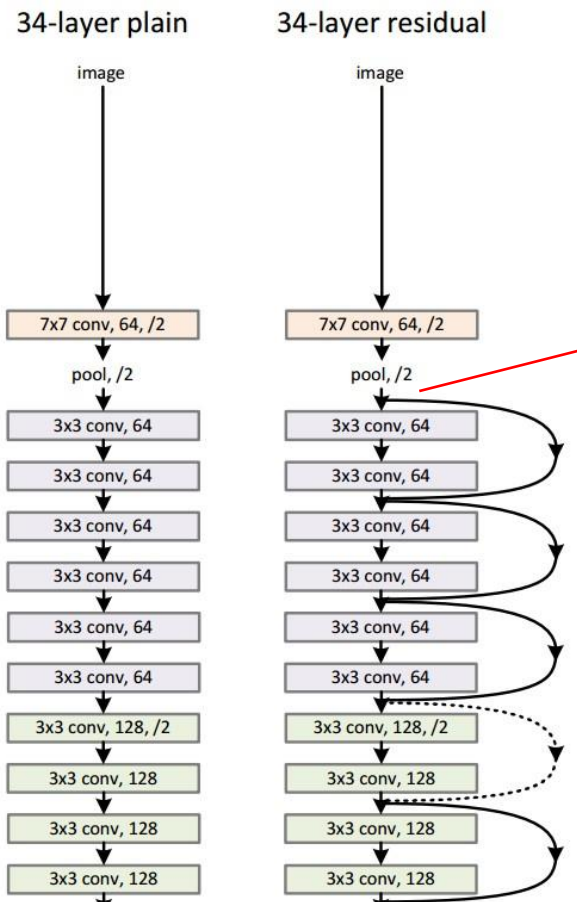
ResNet Running time



2-3 weeks of training on 8 GPU machine

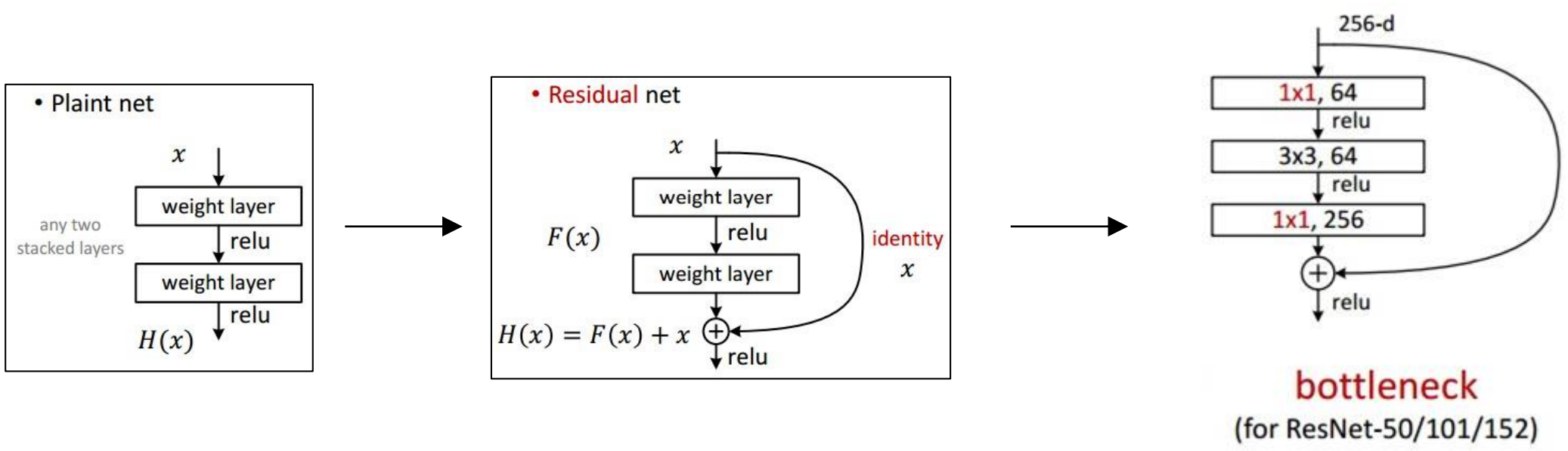
at runtime: faster than a VGGNet! (even though it has 8x more layers)

ResNet Technique



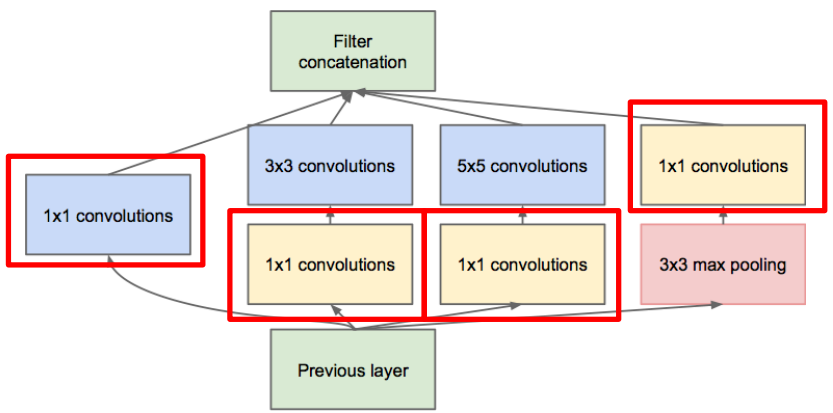
spatial dimension
only 56x56!

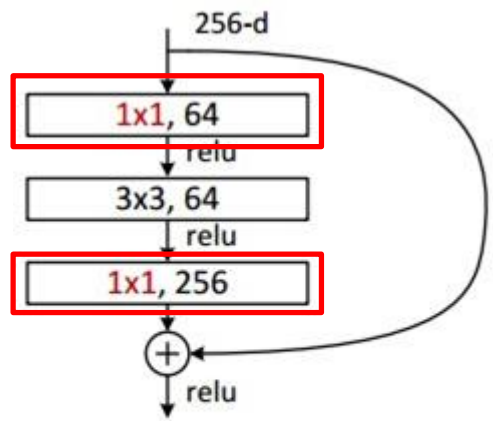
ResNet: Residual net / bottleneck



1x1 Convolution?

Convolution for channel direction only



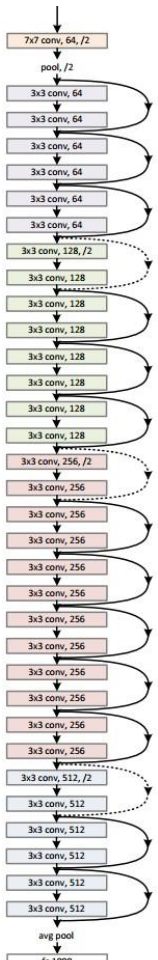


bottleneck

(for ResNet-50/101/152)

NVIDIA.

ResNet: Depth View



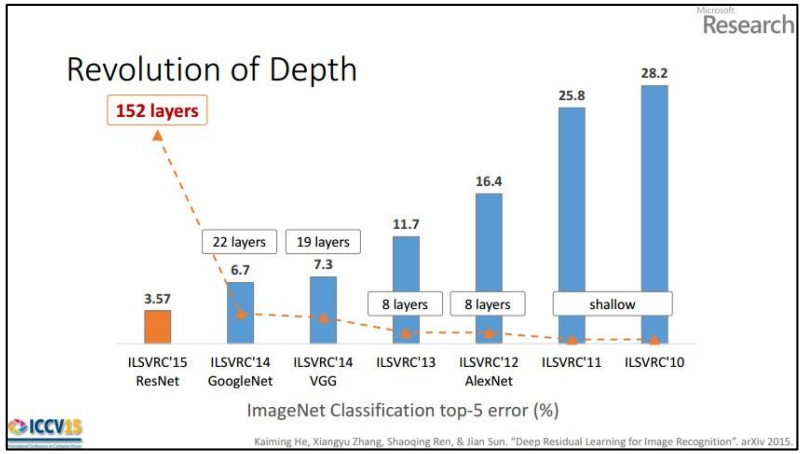
layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

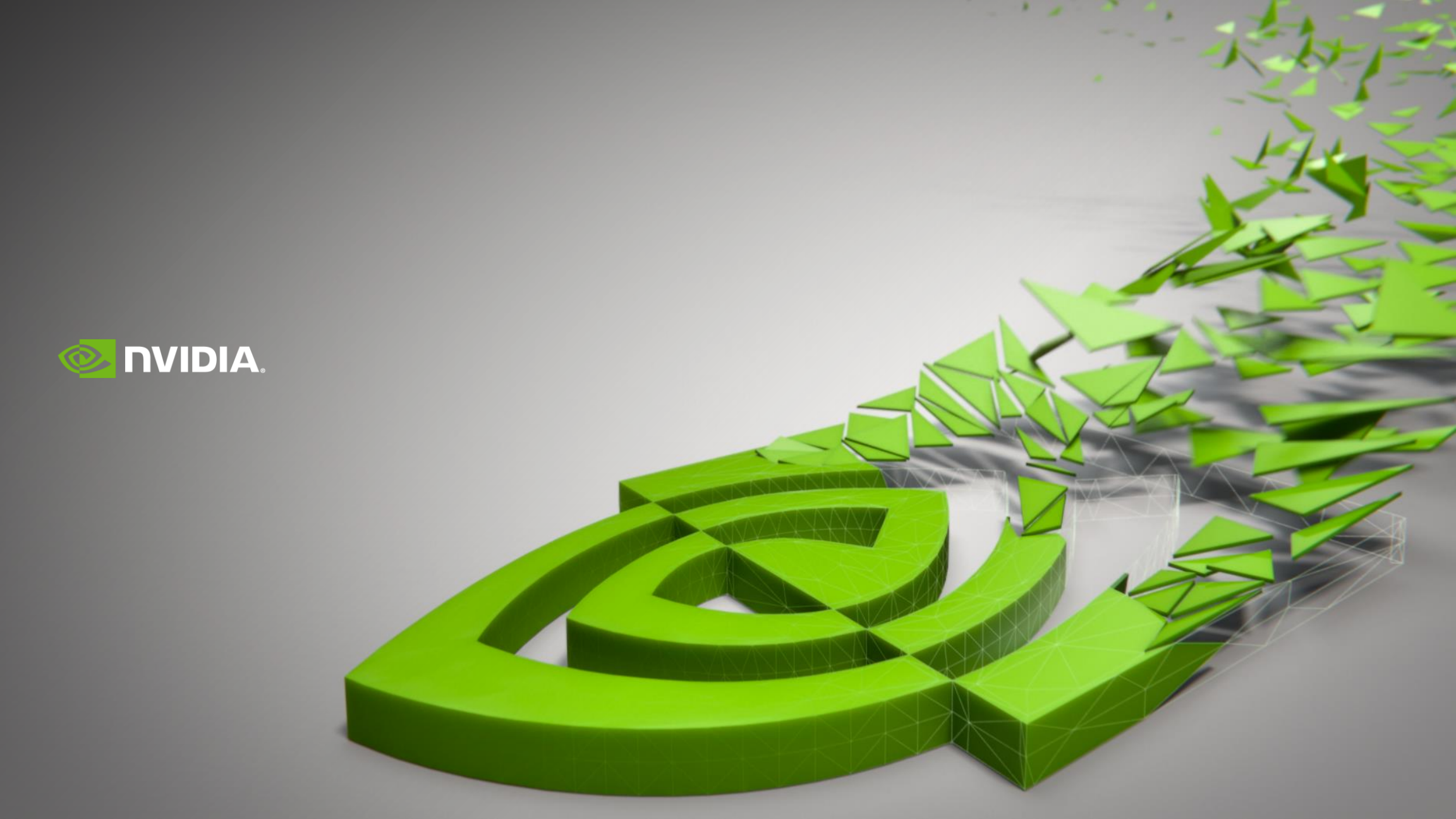
NVIDIA.

ILSVRC2016

16% better performance

Team name	Entry description	Classification error	Localization error
Trimps-Soushen	Ensemble 2	0.02991	0.077668
Trimps-Soushen	Ensemble 3	0.02991	0.077087
Trimps-Soushen	Ensemble 4	0.02991	0.077429
ResNeXt	Ensemble C, weighted average, tuned on val. [No bounding box results]	0.03031	0.737308
CU-DeepLink	GrandUnion + Fused-scale EnsembleNet	0.03042	0.098892
CU-DeepLink	GrandUnion + Multi-scale EnsembleNet	0.03046	0.099006
CU-DeepLink	GrandUnion + Basic Ensemble	0.03049	0.098954
ResNeXt	Ensemble B, weighted average, tuned on val. [No bounding box results]	0.03092	0.737484
CU-DeepLink	GrandUnion + Class-reweighted Ensemble	0.03096	0.099369
CU-DeepLink	GrandUnion + Class-reweighted Ensemble with Per-instance Normalization	0.03103	0.099349
ResNeXt	Ensemble C, weighted average. [No bounding box results]	0.03124	0.737526
Trimps-Soushen	Ensemble 1	0.03144	0.079068
ResNeXt	Ensemble A, simple average. [No bounding box results]	0.0315	0.737505
SamExynos	3 model only for classification	0.03171	0.236561
ResNeXt	Ensemble B, weighted average. [No bounding box results]	0.03203	0.737681
KAISTNIA_ETRI	Ensembles A	0.03256	0.102015
KAISTNIA_ETRI	Ensembles C	0.03256	0.102056
KAISTNIA_ETRI	Ensembles B	0.03256	0.100676
DeepIST	EnsembleC	0.03291	1.0
DeepIST	EnsembleD	0.03294	1.0
DGIST-KAIST	Weighted sum #1 (five models)	0.03297	0.489969
DGIST-KAIST	Weighted sum #2 (five models)	0.03324	1.0




NVIDIA