Scene Classification with Inception-7

Christian Szegedy
with Julian Ibarz and Vincent Vanhoucke

Julian Ibarz

Vincent Vanhoucke

Task

Classification of images into 10 different classes:

- Bedroom
- Bridge
- Church Outdoor
- Classroom
- Conference Room
- Dining Room
- Kitchen
- Living Room
- Restaurant
- Tower

Training/validation/test set

Classification of images into 10 different classes:

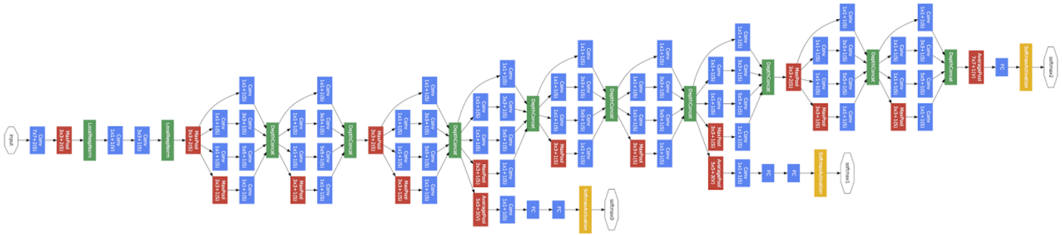
- ~9.87 million **training** images
- 10 thousand **test** images
- 3 thousand **validation** images

Task

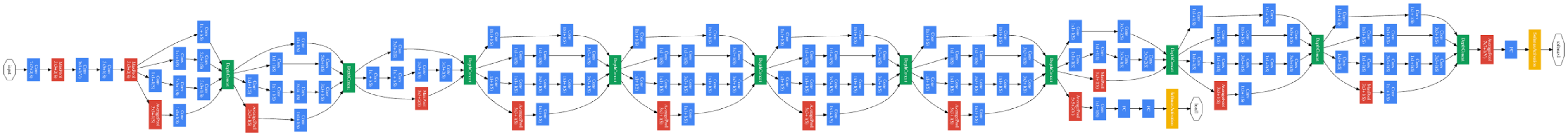
Classification of images into 10 different classes:

- Bedroom
- Bridge
- Church Outdoor
- Classroom
- Conference Room
- Dining Room
- Kitchen
- Living Room
- Restaurant
- Tower

Evolution of Inception



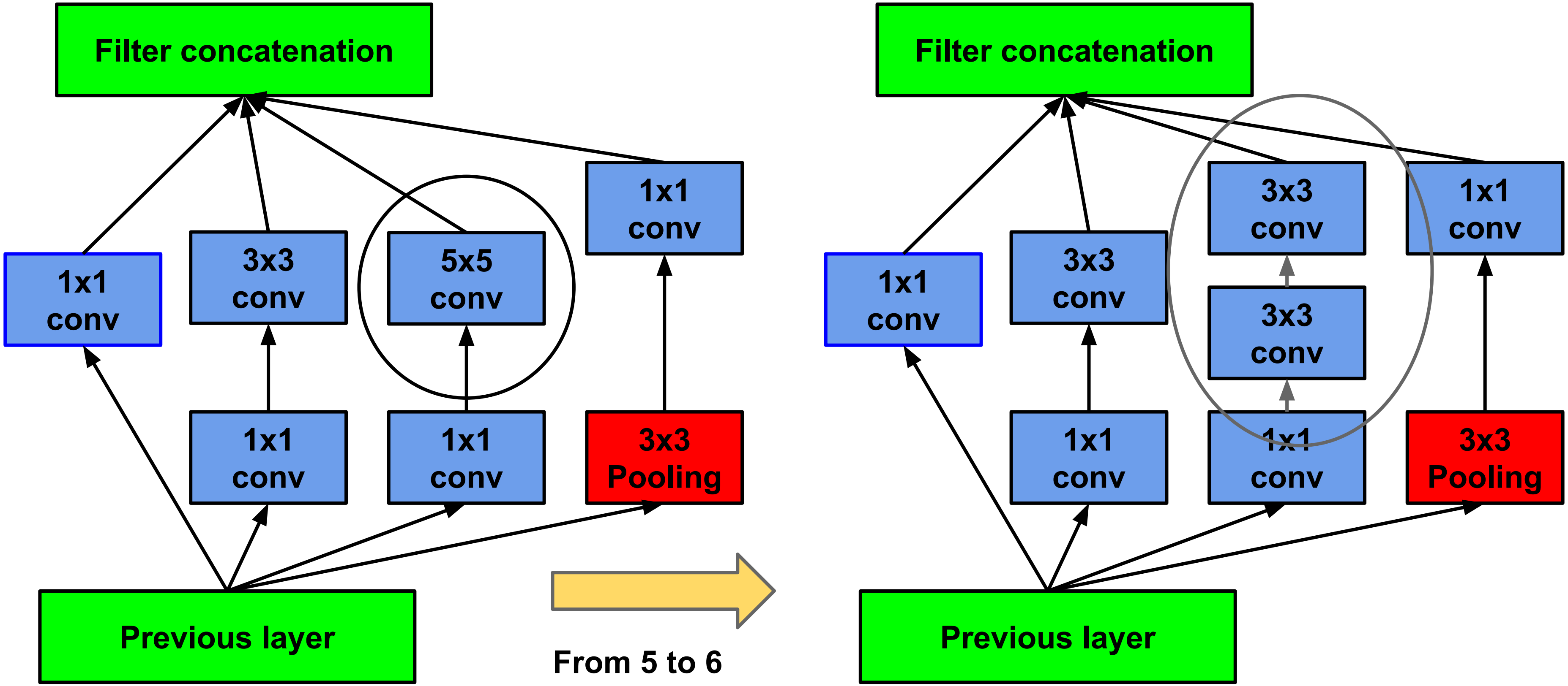
[1]Inception 5 (GoogLeNet)

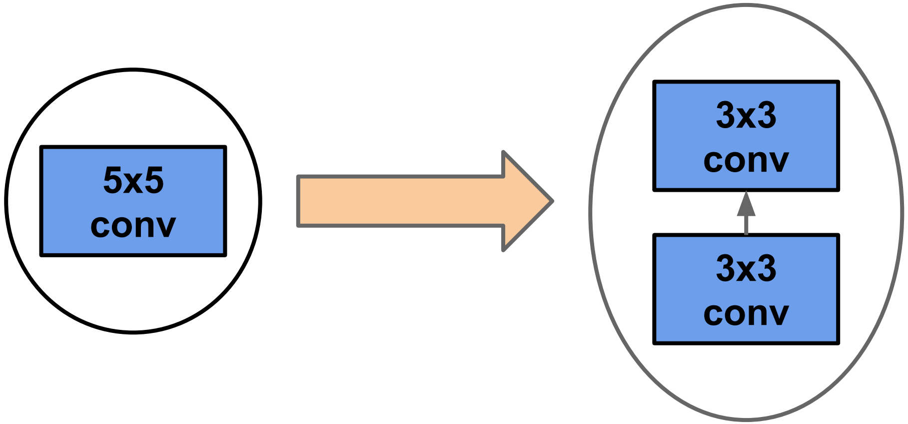


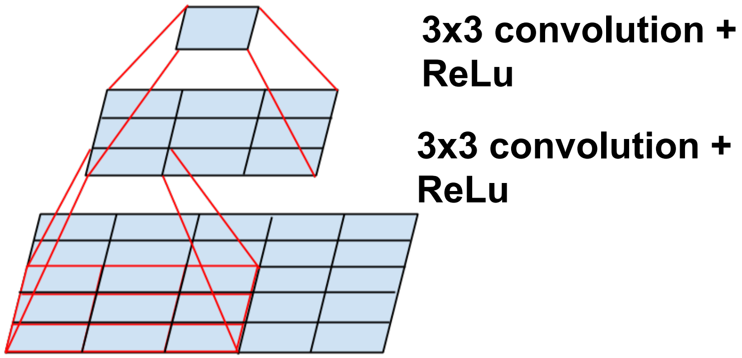
Inception 7a

[1]Going Deeper with Convolutions, [C. Szegedy et al, CVPR 2015]

Structural changes from Inception 5 to 6





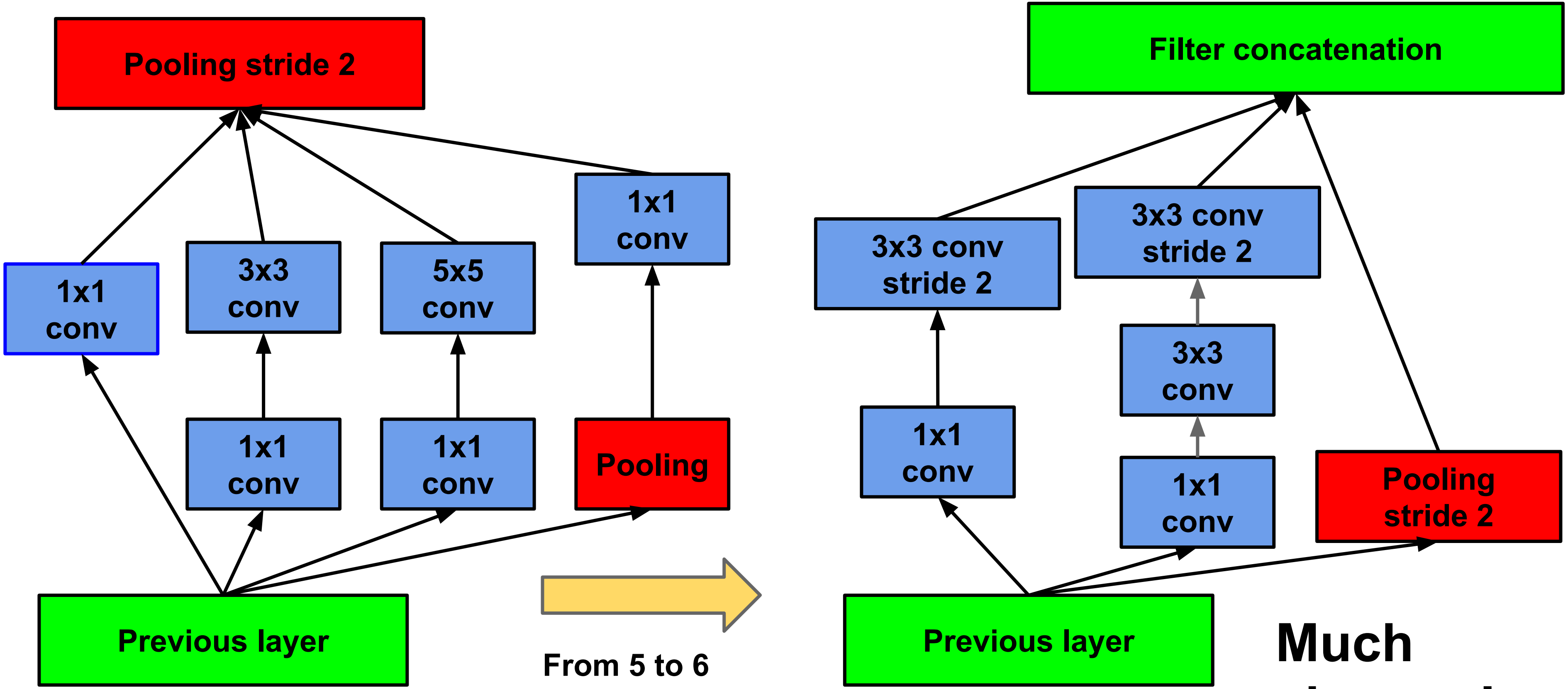


- Each mini network has the same receptive field.
- Deeper: more expressive (ReLu on both layers).
- 25 / 18 times (~28%) cheaper (due to feature sharing).
- Computation savings can be used to increase the number of filters.

Downside:

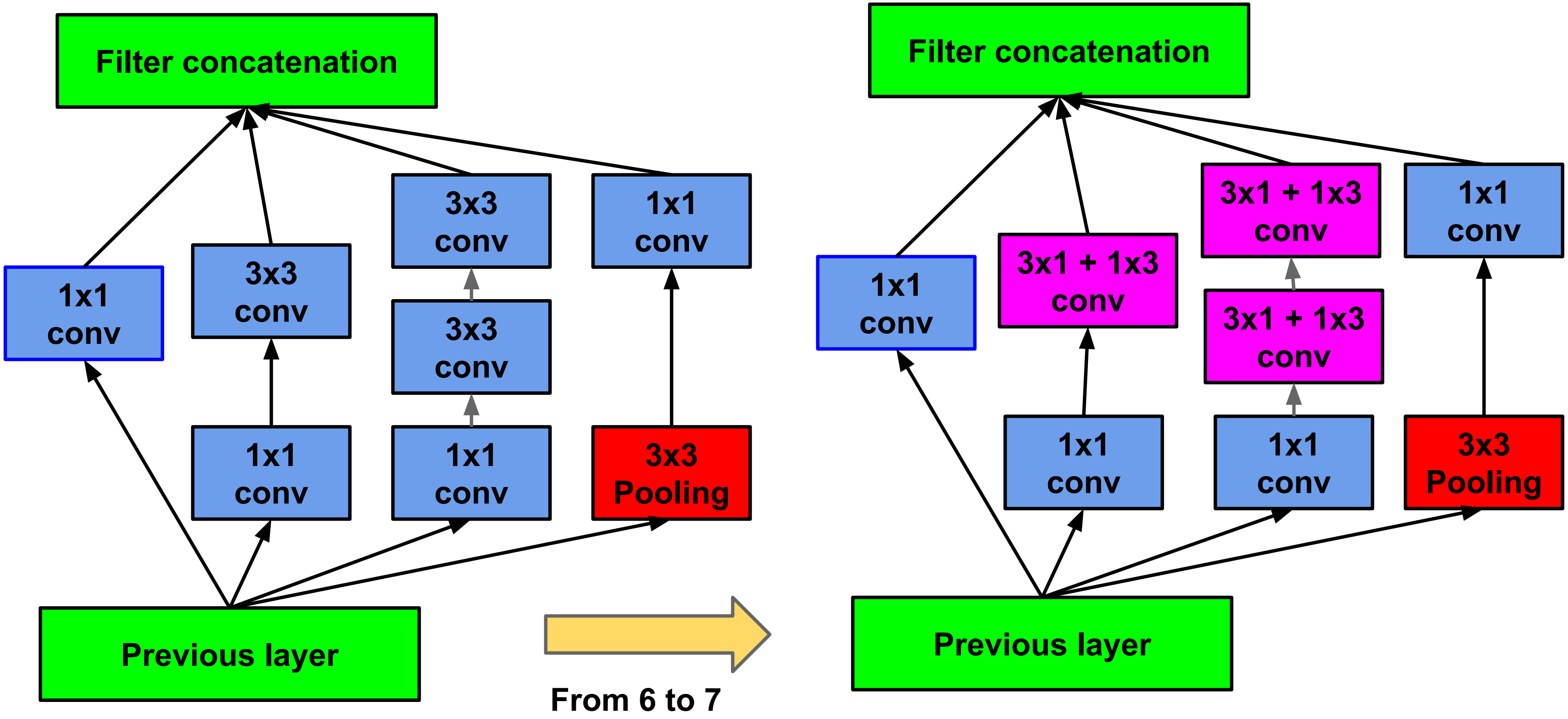
Needs more memory at training time

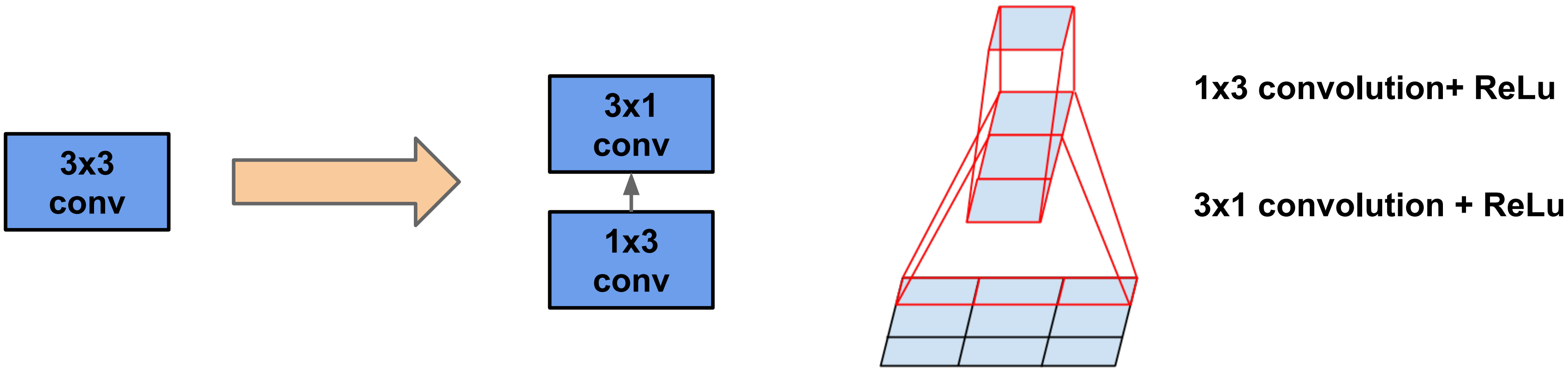
Grid size reduction Inception 5 vs 6



Much cheaper!

Structural changes from Inception 6 to 7





- **Each mini network has the same receptive field.**
- **Deeper: more expressive (ReLu on both layers).**
- **9 / 6 times (~33%) cheaper (due to feature sharing).**
- **Computation savings can be used to increase the number of filters.**

<u>Downside:</u>

Needs more memory at training time

Inception-6 vs Inception-7 Padding

Inception 6: **SAME** padding throughout:

SAME padding

Input grid size	Patch size	Stride	Output grid size
8x8	3x3	1	8x8
8x8	5x5	1	8x8
8x8	3x3	2	4x4
8x8	3x3	4	2x2

- Output size is independent of patch size
- Padding with zero values

VALID padding

Input grid size	Patch size	Stride	Output grid size
7x7	3x3	1	5x5
7x7	5x5	1	3x3
7x7	3x3	2	3x3
7x7	3x3	4	2x2

- Output size depends on the patch size
- No padding: each patch is fully contained

Inception-6 vs Inception-7 Padding

Advantages of padding methods

SAME padding

- More equal distribution of gradients
- Less boundary effects
- No tunnel vision (sensitivity drop at the border)

VALID padding

- More refined: higher grid sizes at the same computational cost

Stride	Inception 6 padding	Inception 7 padding
1	**SAME**	**SAME (VALID on first few layers)**
2	**SAME**	**VALID**

Inception-6 vs Inception-7 Padding

Stride	Inception 6 padding	Inception 7 padding
1	**SAME**	**SAME (VALID on first few layers)**
2	**SAME**	**VALID**

Inception 6: **224 $\Rightarrow$ 112 $\Rightarrow$ 56 $\Rightarrow$ 28 $\Rightarrow$ 14 $\Rightarrow$ 7**

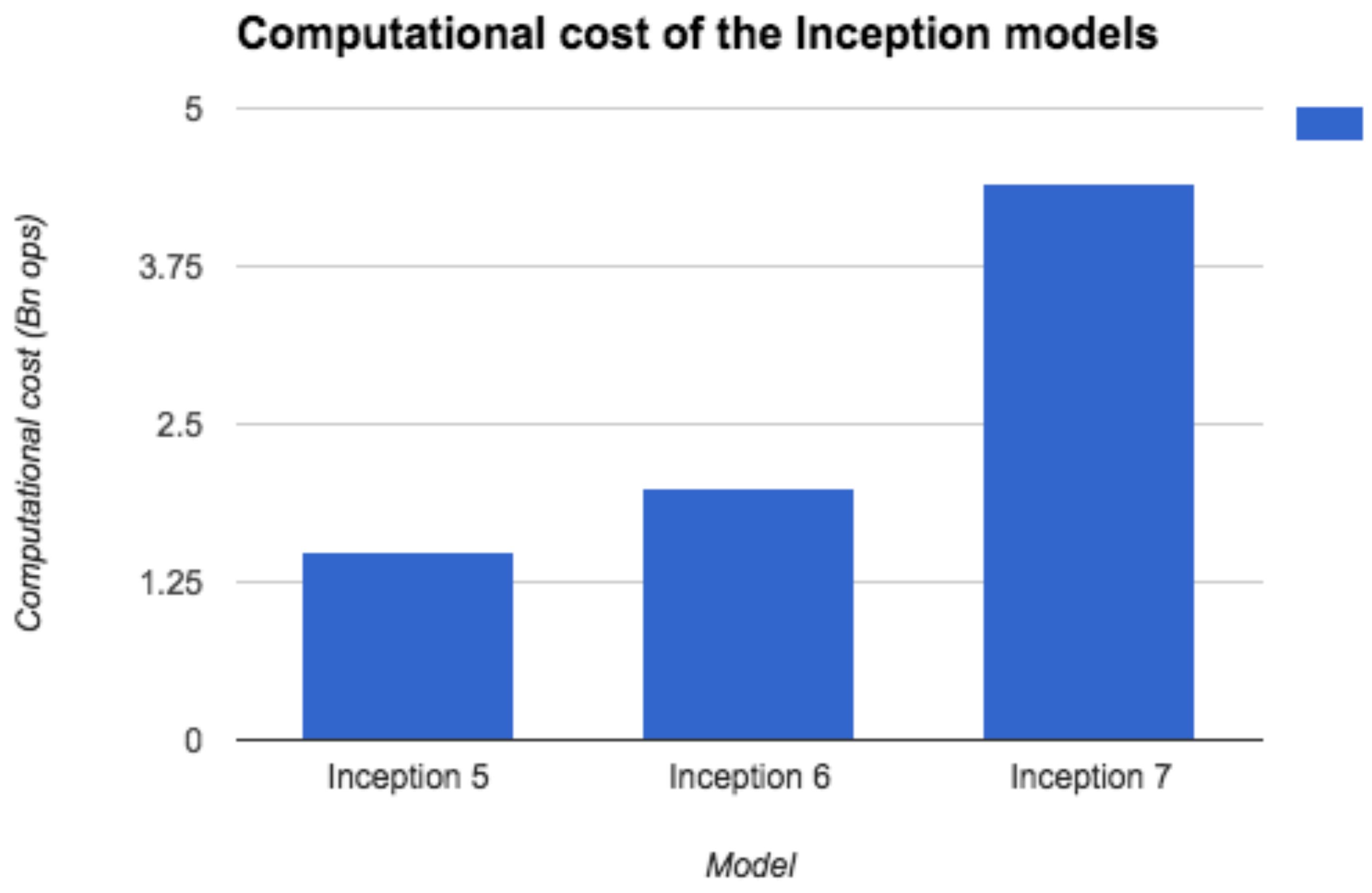
Inception 7: **299 $\Rightarrow$ 147 $\Rightarrow$ 73 $\Rightarrow$ 71 $\Rightarrow$ 35 $\Rightarrow$ 17 $\Rightarrow$ 8**

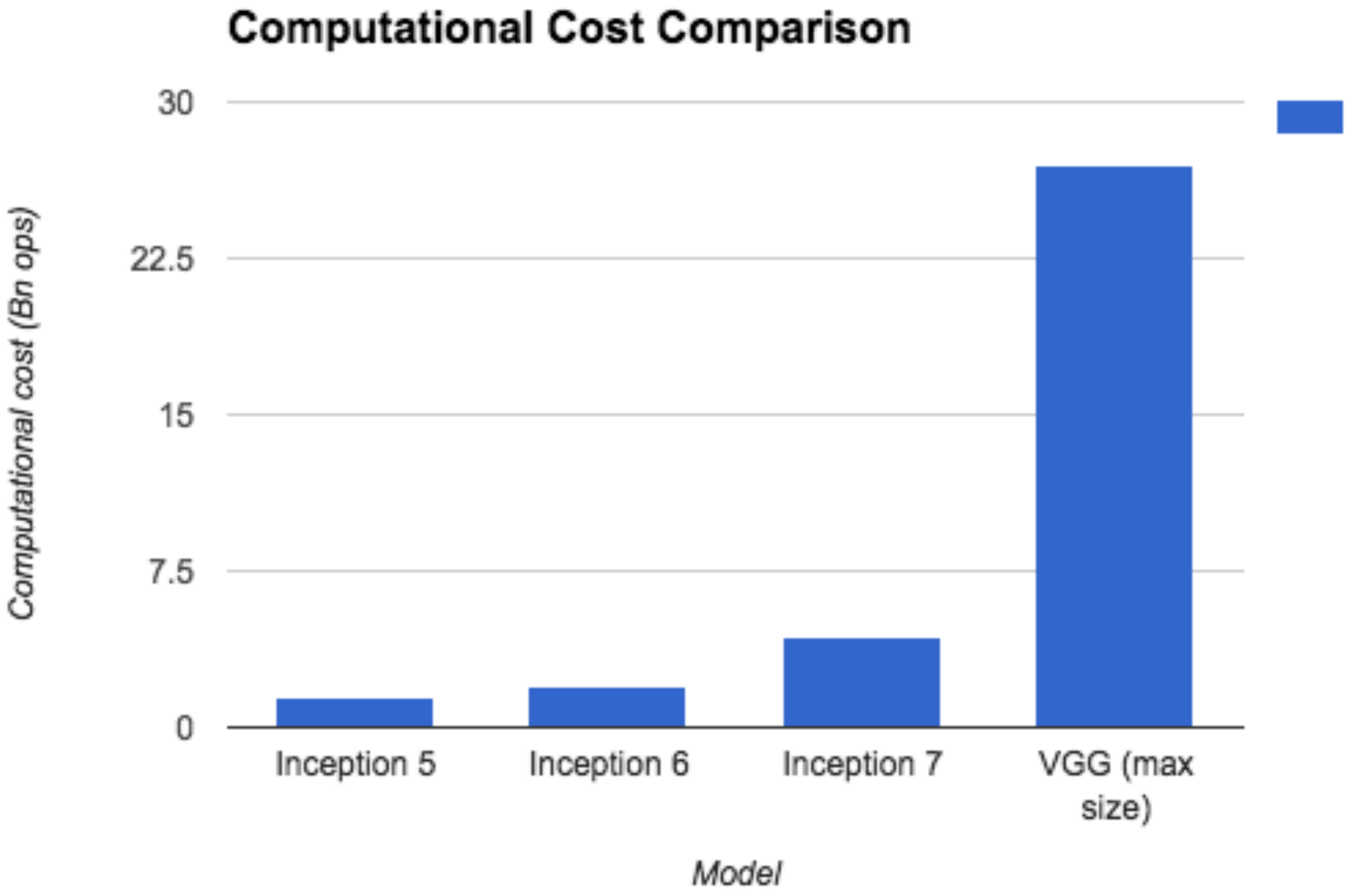
30% reduction of computation compared to a 299x299 network with SAME padding throughout.

Spending the computational savings

Grid Size	Inception 5 filters	Inception 6 filters	Inception 7 filters
28x28 (35x35 for Inception 7)	256	320	288
14x14 (17x17 for Inception 7)	528	576	1248
7x7 (8x8 for Inception 7)	1024	1024	2048

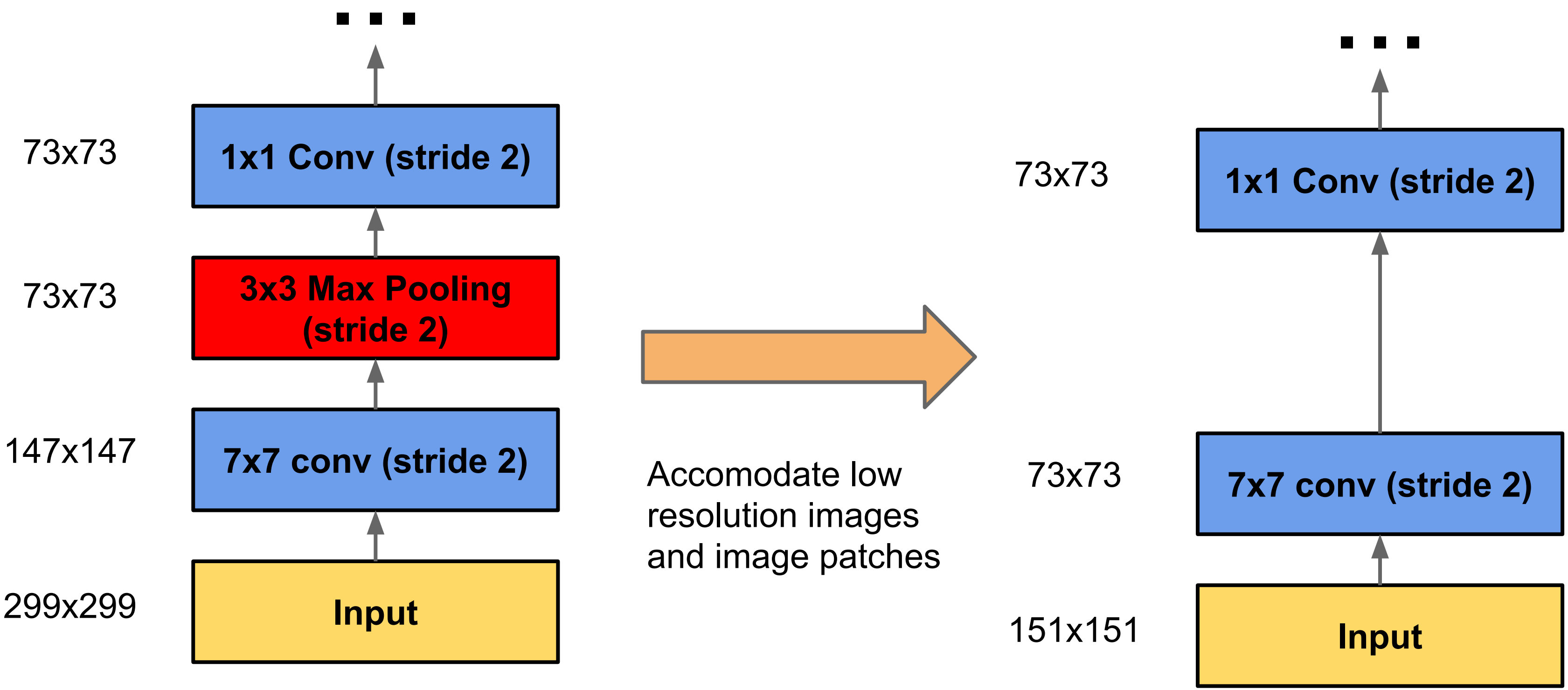
Note: filter size denotes the maximum number of filters/grid cell for each grid size. Typical number of filters is lower, especially for Inception 7.


Computational cost of the Inception models
Computational cost (Bn ops)
5
3.75
2.5
1.25
0
Inception 5
Inception 6
Inception 7
Model



Computational Cost Comparison
30
22.5
15
7.5
0
Computational cost (Bn ops)
Inception 5
Inception 6
Inception 7
VGG (max size)
Model

LSUN specific modification



Training

- Stochastic gradient descent
- Momentum (0.9)
- Fixed learning rate decay of 0.94
- Batch size: 32
- Random patches:
  - Minimum sample area: 15% of the full image
  - Minimum aspect ratio: 3:4 (affine distortion)
  - random constrast, brightness, hue and saturation
- Batch normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, S. Ioffe, C.Szegedy, ICML 2015)

Task

Classification of images into 10 different classes:

- Bedroom
- Bridge
- Church Outdoor
- Classroom
- Conference Room
- Dining Room
- Kitchen
- Living Room
- Restaurant
- Tower

Manual Score Calibration

- Compute weights for each label that maximizes the score on half of the validation set
- Cross-validation on the other half of the validation set
- Simplify weights after error-minimization to avoid overfitting to the validation set.

Final score multipliers:

- 4.0 for church outdoor
- 2.0 for conference room

Probable reason: classes are under-represented in the training set.

Evaluation

- Crop averaging at 3 different scales (Going Deeper with Convolutions, Szegedy et al, CVPR 2015): score averaging of 144 crops/image

Evaluation method	Accuracy (on validation set)
Single crop	89.2%
Multi crop	89.7%
Manual score calibration	91.2%

Releasing Pretrained Inception and MultiBox

Academic criticism: Results are hard to reproduce

We will be releasing pretrained **Caffe** models for:

- **GoogLeNet** (Inception 5)
- **BN-Inception** (Inception 6)
- **MultiBox-Inception** proposal generator (based on Inception 6)

Contact: Yangqing Jia

Acknowledgments

We would like to thank:

Organizers of LSUN

DistBelief and Image Annotation teams at Google for their support for the machine learning and evaluation infrastructure.