Scene Classification with Inception-7

Christian Szegedy
with Julian Ibarz and 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

\(^1\)Going Deeper with Convolutions, [C. Szegedy et al, CVPR 2015]
Structural changes from Inception 5 to 6

From 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

From 5 to 6

Much cheaper!
Structural changes from Inception 6 to 7

Previous layer

Filter concatenation

1x1 conv

3x3 conv

1x1 conv

3x3 conv

1x1 conv

3x3 Pooling

From 6 to 7

Previous layer
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.

**Downside:**
Needs more memory at training time
# Inception-6 vs Inception-7 Padding

Inception 6: **SAME** padding throughout:

**SAME** padding

<table>
<thead>
<tr>
<th>Input grid size</th>
<th>Patch size</th>
<th>Stride</th>
<th>Output grid size</th>
</tr>
</thead>
<tbody>
<tr>
<td>8x8</td>
<td>3x3</td>
<td>1</td>
<td>8x8</td>
</tr>
<tr>
<td>8x8</td>
<td>5x5</td>
<td>1</td>
<td>8x8</td>
</tr>
<tr>
<td>8x8</td>
<td>3x3</td>
<td>2</td>
<td>4x4</td>
</tr>
<tr>
<td>8x8</td>
<td>3x3</td>
<td>4</td>
<td>2x2</td>
</tr>
</tbody>
</table>

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

**VALID** padding

<table>
<thead>
<tr>
<th>Input grid size</th>
<th>Patch size</th>
<th>Stride</th>
<th>Output grid size</th>
</tr>
</thead>
<tbody>
<tr>
<td>7x7</td>
<td>3x3</td>
<td>1</td>
<td>5x5</td>
</tr>
<tr>
<td>7x7</td>
<td>5x5</td>
<td>1</td>
<td>3x3</td>
</tr>
<tr>
<td>7x7</td>
<td>3x3</td>
<td>2</td>
<td>3x3</td>
</tr>
<tr>
<td>7x7</td>
<td>3x3</td>
<td>4</td>
<td>2x2</td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th>Stride</th>
<th>Inception 6 padding</th>
<th>Inception 7 padding</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SAME</td>
<td>SAME (VALID on first few layers)</td>
</tr>
<tr>
<td>2</td>
<td>SAME</td>
<td>VALID</td>
</tr>
</tbody>
</table>
## Inception-6 vs Inception-7 Padding

<table>
<thead>
<tr>
<th>Stride</th>
<th>Inception 6 padding</th>
<th>Inception 7 padding</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SAME</td>
<td>SAME (VALID on first few layers)</td>
</tr>
<tr>
<td>2</td>
<td>SAME</td>
<td>VALID</td>
</tr>
</tbody>
</table>

Inception 6: 224 ➞ 112 ➞ 56 ➞ 28 ➞ 14 ➞ 7

Inception 7: 299 ➞ 147 ➞ 73 ➞ 71 ➞ 35 ➞ 17 ➞ 8

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

<table>
<thead>
<tr>
<th>Grid Size</th>
<th>Inception 5 filters</th>
<th>Inception 6 filters</th>
<th>Inception 7 filters</th>
</tr>
</thead>
<tbody>
<tr>
<td>28x28 (35x35 for Inception 7)</td>
<td>256</td>
<td>320</td>
<td>288</td>
</tr>
<tr>
<td>14x14 (17x17 for Inception 7)</td>
<td>528</td>
<td>576</td>
<td>1248</td>
</tr>
<tr>
<td>7x7 (8x8 for Inception 7)</td>
<td>1024</td>
<td>1024</td>
<td>2048</td>
</tr>
</tbody>
</table>

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

Model

Inception 5  Inception 6  Inception 7

Computational cost (Bn ops)

1.25  1.75  4.0
Computational Cost Comparison

- Inception 5
- Inception 6
- Inception 7
- VGG (max size)
LSUN specific modification

- 7x7 conv (stride 2)
- 3x3 Max Pooling (stride 2)
- 1x1 Conv (stride 2)

Accommodate low resolution images and image patches
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 contrast, 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

<table>
<thead>
<tr>
<th>Evaluation method</th>
<th>Accuracy (on validation set)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single crop</td>
<td>89.2%</td>
</tr>
<tr>
<td>Multi crop</td>
<td>89.7%</td>
</tr>
<tr>
<td>Manual score calibration</td>
<td>91.2%</td>
</tr>
</tbody>
</table>
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](mailto:yangqing.jia@stanford.edu)
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.