Deformation Modeling in ConvNets

Jifeng Dai
Visual Computing Group
Microsoft Research Asia
Content

• Background

• Spatial Transformer Networks

• Deformable ConvNets v1

• Deformable ConvNets v2

• Related Work

• Conclusion
Modeling Spatial Transformations

• A long standing problem in computer vision

Part deformation: 

Scale: 

Viewpoint variation: 

Intra-class variation: 

(Some examples are taken from Li Fei-fei’s course CS223B, 2009-2010.)
Traditional Approaches

• 1) To build training datasets with sufficient desired variations

- Drawbacks: geometric transformations are assumed fixed and known, hand-crafted design of invariant features and algorithms

• 2) To use transformation-invariant features and algorithms

Scale Invariant Feature Transform (SIFT)  Deformable Part-based Model (DPM)
Spatial transformations in CNNs

• Regular CNNs are inherently limited to model large unknown transformations
  • The limitation originates from the fixed geometric structures of CNN modules
Content

- Background

- Spatial Transformer Networks

- Deformable ConvNets v1

- Deformable ConvNets v2

- Related Work

- Conclusion

Spatial Transformer Networks
Spatial Transformer Networks

- Parameterized Sampling Grid

\[
\begin{pmatrix}
x_i^s \\
y_i^s
\end{pmatrix} = \begin{bmatrix}
\theta_{11} & \theta_{12} & \theta_{13} \\
\theta_{21} & \theta_{22} & \theta_{23}
\end{bmatrix} \begin{pmatrix}
x_i^t \\
y_i^t \\
1
\end{pmatrix}
\]
Spatial Transformer Networks

- Differentiable Image Sampling

\[ V_i^c = \sum_{n} \sum_{m} U_{nm}^c \max(0, 1 - |x_i^s - m|) \max(0, 1 - |y_i^s - n|) \]

\[ U = \begin{bmatrix}
1 & 2 & 3 & 4 \\
5 & 6 & 7 & 8 \\
9 & 10 & 11 & 12 \\
13 & 14 & 15 & 16 \\
\end{bmatrix} \]

\[ V = \begin{bmatrix}
\end{bmatrix} \]

\[ x_i^s = 3.7 \quad y_i^s = 3.1 \]

\[ i = 6 \]

\[ 0.27 \times 11 + 0.63 \times 12 + 0.03 \times 15 + 0.07 \times 16 = 12.1 \]
Spatial Transformer Networks

• Learning a global, parametric transformation on feature maps
  • Prefixed transformation family, infeasible for complex vision tasks
Content

• Background

• Spatial Transformer Networks

• Deformable ConvNets v1

• Deformable ConvNets v2

• Related Work

• Conclusion

Highlights

• Enabling effective modeling of spatial transformation in ConvNets

• No additional supervision for learning spatial transformation

• Significant accuracy improvements on sophisticated vision tasks

Code is available at https://github.com/msracver/Deformable-ConvNets
Deformable Convolution

- Local, dense, non-parametric transformation
  - Learning to deform the sampling locations in the convolution/RoI Pooling modules

regular  |  deformed  |  scale & aspect ratio  |  rotation
Deformable Convolution

Regular convolution

\[ y(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n) \]

Deformable convolution

\[ y(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n + \Delta p_n) \]

where \( \Delta p_n \) is generated by a sibling branch of regular convolution
Deformable RoI Pooling

Regular RoI pooling
\[ y(i, j) = \sum_{p \in \text{bin}(i,j)} x(p_0 + p) / n_{ij} \]

Deformable RoI pooling
\[ y(i, j) = \sum_{p \in \text{bin}(i,j)} x(p_0 + p + \Delta p_{ij}) / n_{ij} \]

where \( \Delta p_{ij} \) is generated by a sibling fc branch.
Deformable ConvNets

• Same input & output as the plain versions
  • Regular convolution -> deformable convolution
  • Regular RoI pooling -> deformable RoI pooling

• End-to-end trainable without additional supervision
Sampling Locations of Deformable Convolution

(a) standard convolution

(b) deformable convolution
Part Offsets in Deformable RoI Pooling
Object Detection on COCO (Test-dev)

- Deformable ConvNets v.s. regular ConvNets
  - Noticeable improvements for varies baselines
  - Marginal parameter & computation overhead

<table>
<thead>
<tr>
<th>CLASS AWARE RPN (RESNET-101)</th>
<th>FASTER R-CNN, 2FC (RESNET-101)</th>
<th>R-FCN (ALIGNED-INCEPTION-RESNET)</th>
<th>R-FCN (ALIGNED-XCEPTION)</th>
<th>FPN+OHEM (ALIGNED-XCEPTION)</th>
<th>FPN+OHEM (RESNET-101)</th>
<th>FPN++ (ALIGNED-XCEPTION)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deformable</td>
<td>23.2</td>
<td>30.3</td>
<td>32.1</td>
<td>37.4</td>
<td>40.5</td>
<td>45.2</td>
</tr>
<tr>
<td>Regular</td>
<td>25.8</td>
<td>35</td>
<td>35.7</td>
<td>37.5</td>
<td>40.2</td>
<td>43.3</td>
</tr>
</tbody>
</table>

mAP (%)
Content

• Background

• Spatial Transformer Networks

• Deformable ConvNets v1

• Deformable ConvNets v2

• Related Work

• Conclusion

Highlights

• Better understanding of deformation modeling in CNNs

• Reformulation of Deformable ConvNets to strengthen its deformation modeling capability

• To harness the enhanced modeling capability, guide network training via R-CNN feature mimicking

Core operators are available at https://github.com/msracver/Deformable-ConvNets
Analysis of Deformable ConvNet Behavior

- DCN v1 visualization: theoretical spatial support (sampling / bin location only)

- DCN v2 visualization: effective spatial support (sampling / bin location & learnable network weights)
  - Effective sampling / bin locations
  - Effective receptive fields [Luo et al., NIPS 2016]
  - Error-bounded saliency regions

\[
\min ||M||_1 \\
\text{s.t. } L_{rec}(\mathcal{N}(I), \mathcal{N}(I \odot M)) < \epsilon,
\]
Analysis of Deformable ConvNet Behavior

• Spatial support of nodes in the last layer of the conv5 stage of ResNet-50
  • Regular ConvNets can model geometric variations to some extent.
  • By introducing deformable convolution, the network’s ability to model geometric transformation is considerably enhanced, but still lacks.
Analysis of Deformable ConvNet Behavior

• Spatial support of the 2fc node in the per-RoI detection head
  • By introducing deformable RoI pooling, the network’s ability to model geometric transformation is enhanced, **but still lacks**.
Analysis of Deformable ConvNet Behavior

• Observations
  • Regular ConvNets can model geometric variations to some extent.
  
  • By introducing deformable convolution & deformable RoI pooling, the network’s ability to model geometric transformation is considerably enhanced, but still lacks.
  
  • The three presented types of spatial support visualizations are more informative than the sampling locations used in Deformable ConvNets v1 paper.

• What’s next?
  • To upgrade Deformable ConvNets so that they can better focus on pertinent image content and deliver greater accuracy
Stacking More Deformable Conv Layers

• To strengthen the geometric transformation modeling capability of the entire network
Modulated Deformable Modules

• Not only adjust offsets in perceiving input features, but also modulate the input feature amplitudes from different spatial locations / bins

• Modulated deformable Convolution

\[ y(p) = \sum_{k=1}^{K} w_k \cdot x(p + p_k + \Delta p_k) \cdot \Delta m_k, \]

• Modulated deformable RoIpooling

\[ y(k) = \sum_{j=1}^{n_k} x(p_{kj} + \Delta p_k) \cdot \Delta m_k / n_k, \]
R-CNN Feature Mimicking

• Motivation
  • Even with the strong geometry modeling capability, the spatial support of the per-RoI node can still not focus on the RoI
  • Additional guidance is needed to steer the training
R-CNN Feature Mimicking

- Applied at training time only, no additional overhead for inference

- Feature mimicking loss enforced on sampled positive RoIs

\[ L_{mimic} = \sum_{b \in \Omega} [1 - \cos(f_{RCNN}(b), f_{FRCNN}(b))], \]
R-CNN Feature Mimicking

(c) modulated deformable RoIpooling, with modulated deformable conv@conv3−5 stages

(d) with R-CNN feature mimicking on setting (c) (DCNv2)
Ablation Experiments on Enriched Deformation

• Stacking more deformable conv layers and exploitation of modulation mechanism effectively improve the accuracy

<table>
<thead>
<tr>
<th>method</th>
<th>setting (shorter side 1000)</th>
<th>Faster R-CNN</th>
<th>Mask R-CNN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AP_{bbox}</td>
<td>AP_{bbox}</td>
</tr>
<tr>
<td>baseline</td>
<td>regular (Rolpooling)</td>
<td>32.1</td>
<td>14.9</td>
</tr>
<tr>
<td></td>
<td>regular (aligned Rolpooling)</td>
<td>34.7</td>
<td>19.3</td>
</tr>
<tr>
<td></td>
<td>dconv@c5 + dpool (DCNv1)</td>
<td>38.0</td>
<td>20.7</td>
</tr>
<tr>
<td>enriched deformation</td>
<td>dconv@c5</td>
<td>37.4</td>
<td>20.0</td>
</tr>
<tr>
<td></td>
<td>dconv@c4~c5</td>
<td>40.0</td>
<td>21.4</td>
</tr>
<tr>
<td></td>
<td>dconv@c3~c5</td>
<td>40.4</td>
<td>21.6</td>
</tr>
<tr>
<td></td>
<td>dconv@c3~c5 + dpool</td>
<td>41.0</td>
<td>22.0</td>
</tr>
<tr>
<td></td>
<td>mdconv@c3~c5 + mdpool</td>
<td>41.7</td>
<td>22.2</td>
</tr>
</tbody>
</table>

Table 1. Ablation study on enriched deformation modeling. The input images are of shorter side 1,000 pixels (default in paper). In the setting column, “(m)dconv” and “(m)dpool” stand for (modulated) deformable convolution and (modulated) deformable Rolpooling, respectively. Also, “dconv@c3~c5” stands for applying deformable conv layers at stages conv3~conv5, for example. Results are reported on the COCO 2017 validation set.
### Ablation Experiments of R-CNN Feature Mimicking

<table>
<thead>
<tr>
<th>setting</th>
<th>regions to mimic</th>
<th>Faster R-CNN</th>
<th>Mask R-CNN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AP\textit{bbox}</td>
<td>AP\textit{bbox}</td>
</tr>
<tr>
<td>mdconv3~5 + mdpool</td>
<td>None</td>
<td>41.7</td>
<td>43.1</td>
</tr>
<tr>
<td></td>
<td>FG &amp; BG</td>
<td>42.1</td>
<td>43.4</td>
</tr>
<tr>
<td></td>
<td>BG Only</td>
<td>41.7</td>
<td>43.3</td>
</tr>
<tr>
<td></td>
<td>FG Only</td>
<td>43.1</td>
<td>44.3</td>
</tr>
<tr>
<td>regular</td>
<td>None</td>
<td>34.7</td>
<td>36.6</td>
</tr>
<tr>
<td></td>
<td>FG Only</td>
<td>35.0</td>
<td>36.8</td>
</tr>
</tbody>
</table>

Table 3. Ablation study on R-CNN feature mimicking. Results are reported on the COCO 2017 validation set.
Content

• Background

• Spatial Transformer Networks

• Deformable ConvNets v1

• Deformable ConvNets v2

• Related Work

• Conclusion
Related Work

• Deformation Modeling
  • SIFT [Lowe, ICCV 1999], ORB [Rublee et al., ICCV 2011], DPM [Felzenszwalb et al., TPAMI 2010]
  • Spatial Transformer Networks [Jaderberg et al., NIPS 2015], DeepID-Net [Ouyang et al., CVPR 2015], etc.

• Relation Networks and Attention Modules
  • Relation Modules in NLP [Gehring et al., ACL 2017], physical system modeling [Battaglia et al., NIPS 2016]
  • Relation networks for object detection [Hu et al., CVPR 2018], non-local networks [Wang et al., CVPR 2018], Learning region features for object detection [Gu et al., ECCV 2018]
Related Work

• Spatial Support Manipulation
  • Atrous convolution [Chen et al., ICLR 2015], active convolution [Jeon and Kim, CVPR 2017], multi-path network [Zagoruyko et al., BMVC 2016]

• Network Mimicking and Distillation
  • [Ba and Caruana, NIPS 2014], [Hinton et al., STAT 2015], [Li et al., CVPR 2017]
Content

• Background

• Spatial Transformer Networks

• Deformable ConvNets v1

• Deformable ConvNets v2

• Related Work

• Conclusion
Conclusion

- Standard CNNs are not very well equipped to model deformations, and transformations of the objects.

- Spatial Transformer Networks and Deformable ConvNets enabled effective modeling of geometric deformation in CNNs

Open questions:
- More effective manner to capture geometric deformation
- Disentangle different factors in geometric deformation
- Many more...
Q & A