Relation Networks for Visual Modeling

Han Hu
Visual Computing Group
Microsoft Research Asia (MSRA)

https://ancientmoon.github.io/

Collaborators: Zheng Zhang, Yue Cao, Jiayuan Gu, Jiarui Xu, Jifeng Dai, Yichen Wei, Stephen Lin, Liwei Wang, Zhenda Xie, Fangyun Wei
Human Brain

- Human cortex can universally perceive different senses

figure credit to J. Sharma et al.
Intelligent Machines

• A **universal** learning pipeline

![Diagram of a neural network with data flowing through layers to a prediction output, with a back-propagation arrow and a Loss function labeled as Loss(prediction, label).]
Intelligent Machines

- **Particular** basic model for different task/data

- convolution
- LSTM, GRU, convolution, self-attention, ...
- graph networks
Universal Basic Models for Intelligent Machines?
**Relation Networks:** Towards Universal Basic Models

similar things: **graph neural networks**, self-attention, ...

- graph neural networks
- (self)-attention

left figure credit to P. Battaglia et al.
Relation Networks for Graph Data

Relation Networks for NLP

A. Vaswani and et al. *Attention Is All You Need*. NIPS 2017
Relation Networks for Visual Modeling

**pixel-pixel**

Convolution Variants

Relation Networks

**object-pixel**

RoIAlign

Relation Networks

**object-object**

None

Relation Networks

---

our study timeline
Object-Object Relation Modeling

Han Hu*, Jiayuan Gu*, Zheng Zhang*, Jifeng Dai and Yichen Wei. Relation Networks for Object Detection. CVPR 2018
It is much easier to detect the *glove* if we know there is a *baseball player*.
Object Relation Module

Plug-and-Play
✓ Parallel, learnable, no additional supervision, translational invariant, stackable

Key Feature
✓ Relative Geometric Term
✓ Multiple Relation Branches
✓ Shortcut
The **First** Fully End-to-End Object Detector

## Results on COCO Object Detection

<table>
<thead>
<tr>
<th>backbone</th>
<th>setting</th>
<th>mAP</th>
<th>mAP$_{50}$</th>
<th>mAP$_{75}$</th>
<th>#. params</th>
<th>FLOPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>faster RCNN</td>
<td>2fc+SoftNMS</td>
<td>32.2/32.7</td>
<td>52.9/53.6</td>
<td>34.2/34.7</td>
<td>58.3M</td>
<td>122.2B</td>
</tr>
<tr>
<td></td>
<td>2fc+RM+SoftNMS</td>
<td>34.7/35.2</td>
<td>55.3/56.2</td>
<td>37.2/37.8</td>
<td>64.3M</td>
<td>124.6B</td>
</tr>
<tr>
<td></td>
<td>2fc+RM+e2e</td>
<td>35.2/35.4</td>
<td>55.8/56.1</td>
<td>38.2/38.5</td>
<td>64.6M</td>
<td>124.9B</td>
</tr>
<tr>
<td>FPN</td>
<td>2fc+SoftNMS</td>
<td>36.8/37.2</td>
<td>57.8/58.2</td>
<td>40.7/41.4</td>
<td>56.4M</td>
<td>145.8B</td>
</tr>
<tr>
<td></td>
<td>2fc+RM+SoftNMS</td>
<td>38.1/38.3</td>
<td>59.5/59.9</td>
<td>41.8/42.3</td>
<td>62.4M</td>
<td>157.8B</td>
</tr>
<tr>
<td></td>
<td>2fc+RM+e2e</td>
<td>38.8/38.9</td>
<td>60.3/60.5</td>
<td>42.9/43.3</td>
<td>62.8M</td>
<td>158.2B</td>
</tr>
<tr>
<td>DCN</td>
<td>2fc+SoftNMS</td>
<td>37.5/38.1</td>
<td>57.3/58.1</td>
<td>41.0/41.6</td>
<td>60.5M</td>
<td>125.0B</td>
</tr>
<tr>
<td></td>
<td>2fc+RM+SoftNMS</td>
<td>38.1/38.8</td>
<td>57.8/58.7</td>
<td>41.3/42.4</td>
<td>66.5M</td>
<td>127.4B</td>
</tr>
<tr>
<td></td>
<td>2fc+RM+e2e</td>
<td>38.5/39.0</td>
<td>57.8/58.6</td>
<td>42.0/42.9</td>
<td>66.8M</td>
<td>127.7B</td>
</tr>
</tbody>
</table>

*Faster R-CNN with ResNet-101 model are used (evaluation on minival/test-dev are reported)

- **less than 10% computation overhead on all backbones**
Object Pairs with High Relation Weights

instance recognition

duplicate removal

- reference object
- other objects contributing high weights
Class Co-Occurrence Information is Learnt

$r = 0.90$

Class Co-occurrence Probability  Learnt Attentional Weights
Extension: **Spatial-Temporal** Object Relation

Jiarui Xu, Yue Cao, Zheng Zhang and Han Hu. *Spatial-Temporal Relation Networks for Multi-Object Tracking.* Tech Report 2018
Learnable Object-Pixel Relation (vs. RoIAlign)

Image Feature to Region Feature

Geometric

Appearance

Jiayuan Gu, Han Hu, Liwei Wang, Yichen Wei and Jifeng Dai. Learning Region Features for Object Detection. ECCV 2018
Pixel-Pixel Relation Modeling

convolution

ConvNets
Question I: Can We Go Beyond Convolution?

convolution = template matching

Can we model the patterns by one channel?

Related Works: Capsule Networks

- Not aligned well with modern learning infrastructure

Figure credit by Aurélien Géron

S. Sabour et al. *Dynamic Routing Between Capsules*. NIPS2017
Related Works: Non-Local Neural Networks

• Complementary to ConvNets
Beyond Convolution: **Local Relation Layer**

\[ \text{relation network} + \text{locality} + \text{geometric prior} + \text{scalar key/query} \]

- **composability**
- **channel #1**
- **local relation**
- **appearance composability**
- **geometric composability (prior)**
Local Relation Network (LR-Net)

<table>
<thead>
<tr>
<th>stage</th>
<th>output</th>
<th>ResNet-50</th>
<th>LR-Net-50 (7×7, m=8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>res1</td>
<td>112×112</td>
<td>7×7 conv, 64, stride 2</td>
<td>1×1, 64 7×7 LR, 64, stride 2</td>
</tr>
<tr>
<td>res2</td>
<td>56×56</td>
<td>3×3 max pool, stride 2</td>
<td>3×3 max pool, stride 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1×1, 64 3×3 conv, 64 1×1, 256</td>
<td>1×1, 100 7×7 LR, 100 1×1, 256</td>
</tr>
<tr>
<td>res3</td>
<td>28×28</td>
<td>1×1, 128 3×3 conv, 128 1×1, 512</td>
<td>1×1, 200 7×7 LR, 200 1×1, 512</td>
</tr>
<tr>
<td>res4</td>
<td>14×14</td>
<td>1×1, 256 3×3 conv, 256 1×1, 1024</td>
<td>1×1, 400 7×7 LR, 400 1×1, 1024</td>
</tr>
<tr>
<td>res5</td>
<td>7×7</td>
<td>1×1, 512 3×3 conv, 512 1×1, 2048</td>
<td>1×1, 800 7×7 LR, 800 1×1, 2048</td>
</tr>
<tr>
<td></td>
<td>1×1</td>
<td>global average pool 1000-d fc, softmax</td>
<td>global average pool 1000-d fc, softmax</td>
</tr>
</tbody>
</table>

# params: $25.5 \times 10^6$  $23.3 \times 10^6$  
FLOPs: $4.3 \times 10^9$  $4.3 \times 10^9$

Totally convolution free!
Classification on ImageNet (26 Layers)
Robust to Adversarial Attacks

<table>
<thead>
<tr>
<th>network</th>
<th>adversarial train</th>
<th>regular train</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>clean</td>
<td>targeted</td>
</tr>
<tr>
<td>ResNet-26</td>
<td>44.9</td>
<td>37.9</td>
</tr>
<tr>
<td>ResNet-50</td>
<td>52.0</td>
<td>43.0</td>
</tr>
<tr>
<td>LR-Net-26</td>
<td><strong>52.1</strong></td>
<td><strong>44.2</strong></td>
</tr>
</tbody>
</table>
Question II: Do Non-local Networks Work Well Due to Relation Learning?

attention maps for different query pixels

Yue Cao*, Jiarui Xu*, Stephen Lin, Fangyun Wei and Han Hu.

\textit{GCNet: Non-local Networks meet SE-Net and Beyond.} Tech Report 2019
Explicit Query-Independent Attention Map

• Simplified Non-Local Blocks

The same accurate but significantly reducing computation!
Meet SE-Net (2017 ImageNet Champion)

(b) Simplified NL block (Eqn 3)

(c) SE block
Abstraction and New Instantiation

1. Global Attention Pooling (Simplified NL-Net)
2. Bottleneck transform (SE-Net)
3. Addition (Simplified NL-Net)

(a) Global context modeling framework

(d) Global context (GC) block
## COCO Object Detection Results

- **Baseline:** Mask R-CNN + ResNet50 + FPN

<table>
<thead>
<tr>
<th>method</th>
<th>AP (bbox)</th>
<th>AP (mask)</th>
<th>#param</th>
<th>FLOPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>37.2</td>
<td>33.8</td>
<td>44.4M</td>
<td>279.4G</td>
</tr>
<tr>
<td>NL-Net</td>
<td>38.0</td>
<td>34.7</td>
<td>46.5M</td>
<td>288.7G</td>
</tr>
<tr>
<td>SE-Net</td>
<td>38.2</td>
<td>34.7</td>
<td>46.9M</td>
<td>279.5G</td>
</tr>
<tr>
<td>GC-Net</td>
<td><strong>39.4</strong></td>
<td><strong>35.7</strong></td>
<td>46.9M</td>
<td>279.6G</td>
</tr>
</tbody>
</table>
Discussion: versus Deformable ConvNets

• Both can model content aware adaptiveness
• Verification vs. Regression
• Generality (arbitrary vs. grid)
• Partly complementary

Thanks!

pixel-pixel

object-pixel

object-object

Convolution Variants

Relation Networks

RoIAlign

Relation Networks

None

Relation Networks

Relation Network is All You Need for AI——SkyNet