Structured Deep Learning of Human Motion

Christian Wolf

Fabien Baradel  Natalia Neverova  Julien Mille  Graham W. Taylor  Greg Mori
Deep Learning of Human Motion

Recognition of group activities

Pose estimation
Combining real and simulated data

Joint positions (NYU Dataset)  Synthetic data (part segmentation)

Natalia Neverova  Phd @ LIRIS, Now at Facebook
Christian Wolf  LIRIS INSA-Lyon
Graham W. Taylor  University of Guelph Canada
Florian Nebout Awabot
Semantic Segmentation with GridNetworks

Residual Conv-Deconv Grid Network for Semantic Segmentation

Damien Fourure, Rémi Emonet, Elisa Fromont, Damien Muselet, Alain Tremeau & Christian Wolf

[Fourure, Emonet, Fromont, Muselet, Tremeau, Wolf, BMVC 2017]
Activity recognition

Unconstrained internet/youtube videos
No acquisition
E.g. Youtube-8M dataset: 7M videos, 4716 classes, ~3.4 labels per video. > 1PB of data.

Videos with human activities, from youtube
No acquisition
E.g. ActivityNet/Kinetics dataset: ~300k videos, 400 classes.

Human activities shot with depth sensors
Acquisition is time consuming!
E.g. NTU RGB-D dataset, MSR dataset, ChaLearn/Montalbano dataset, etc.
Deep Learning is mostly based on global models.

- [Baccouche, Mamalet, Wolf, Garcia, HBU 2011]
- [Baccouche, Mamalet, Wolf, Garcia, Baskur, BMVC 2012]
- [Carreira and Zisserman, CVPR 2017]
- [Ji et al., ICML 2010]
The role of articulated pose

Reading

Writing
The role of articulated pose

Reading

Writing

Appearance is helpful

[Neverova, Wolf, Taylor, Nebout, PAMI 2016]

[Baradel, Wolf, Mille, Taylor, BMVC 2018]
We need put attention to places which are not always determined by pose
Context

We need put attention to places which are not always determined by pose.
Context

Frame from the NTU RGB-D Dataset
Images, objects and activities have often been represented as collections of local features, e.g. through DPMs.

\[ \sum_{i=0}^{n} F'_i \cdot \phi(H, p_i) - \sum_{i=1}^{n} d_i \cdot \phi_d(dx_i, dy_i) + b, \]

Local appearance \hspace{2cm} \text{Deformation}

[Felzenszwalb et al., PAMI 2010]
Structured Deep Learning

**Visual recognition**
(activities, gestures, objects)

**Representation learning**
Local context

**Structured and semi-structured models**
Complex relationships, Global context

**Structured deep learning**

**Deep Learning**

- **maps**
- **feature maps**

- **Structured and semi-structured models**
  - $F_1$
  - $F_2$
  - $F_3$
  - $F_4$

**Fig. 4.**
The proposed deep convolutional architecture of a single learner.

convolutional layers $F_1$, $F_2$, and $F_3$ with rectified linear activation units (ReLU).

Layers $F_1$ and $F_2$ are followed by $2 \times 2$ max pooling and reduction.

As opposed to most existing methods for scene labeling, instead of randomly sampling pixels (or patches), training is performed image-wise, i.e. all pixels from the given image are provided to the classifier at once and each pixel gets assigned with an output class label based on information extracted from its neighborhood.

Applying of the convolutional classifier with pooling/reduction layers to an image in the traditional way would lead to loss in resolution by a factor of 4 (in the given configuration). On the other hand, simply not reducing the image resolution will prevent higher layers from learning higher level features, as the size of the filter support does not grow with respect to the image content.

To avoid this dilemma, we employ specifically designed splitting functions originally proposed for image scanning in [27] and further exploited in OverFeat networks [28]. Intuitively speaking, each map at a given resolution is reduced to four different maps of lower resolution using max pooling. The amount of elements is preserved, but the resolution of each map is lower compared to the maps of previous layers.

In more detail, let us consider the output of the first convolutional layer $F_1$ of the network. Once the output feature maps are obtained, 4 virtual extended copies of them are created by zero padding with 1) one column on the left, 2) one column on the right, 3) one row on top, 4) one row in the bottom. Therefore, each copy will contain the original feature map but shifted in 4 different directions.

On the next step, we apply max pooling ($2 \times 2$ with stride $2 \times 2$) to each of the extended maps producing 4 low-resolution maps. By introducing the shifts, pixels from all extended maps combined together can reconstruct the original feature map as if max pooling with stride $1 \times 1$ had been applied.

This operation allows the network to preserve results of all computations for each pixel on each step and, at the same time, perform the reduction necessary...
Human attention: gaze patterns

[Johansson, Holsanova, Dewhurst, Holmqvist, 2012]
Local representations
(Before 2012)

Deep Learning (Global)
(Mostly after 2012)

Deep Learning (attention maps)
(~2016)

Hard attention
Attention on joints
Soft attention in feature maps

[Mnih et al., NIPS 2015]  [Song et al., AAAI 2016]  [Sharma et al., ICLR 2016]
Objective: fully trainable high-capacity local representations

1. Learn where to attend
2. Learn how to track attended points
3. Learn how to recognize from a local distributed representation

[Baradel, Wolf, Mille, Taylor, CVPR 2018]
Attention in feature space

RGB input video

3D Global model: Inflated Resnet 50

Feature space

Time

Baradel, Wolf, Mille, Taylor, CVPR 2018
Unconstrained differentiable attention

\[ h_g = \Omega(h_{g-1}, [z_{g-1}, r_t] \mid \theta) \]

\[ l_g = W_l^T [h_g, c_t] \]

“Differentiable crop”
(Spatial Transformer Network)

Hidden state from recurrent recognizers (workers)

Frame context

[Baradel, Wolf, Mille, Taylor, CVPR 2018]
Distributed recognition

RGB input video

3D Global model: Inflated Resnet 50

Unconstrained Attention in feature space

Spatial Attention process

Distributed tracking/recognition
Results
Dynamic visual attention

Baradel, Wolf, Mille, Taylor (under review)

SOTA results on two datasets NTU and N-UCLA
Larger difference between Glimpse clouds and global model on N-UCLA

[Baradel, Wolf, Mille, Taylor, CVPR 2018]
# Ablation study

<table>
<thead>
<tr>
<th>Glimpse</th>
<th>Type of attention</th>
<th>CS</th>
<th>CV</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D tubes</td>
<td>Attention</td>
<td>85.8</td>
<td>92.7</td>
<td>89.2</td>
</tr>
<tr>
<td>Seq. 2D</td>
<td>Random sampling</td>
<td>80.3</td>
<td>87.8</td>
<td>84.0</td>
</tr>
<tr>
<td>Seq. 2D</td>
<td>Saliency</td>
<td>86.2</td>
<td>92.9</td>
<td>89.5</td>
</tr>
<tr>
<td>Seq. 2D</td>
<td><strong>Attention</strong></td>
<td><strong>86.6</strong></td>
<td><strong>93.2</strong></td>
<td><strong>89.9</strong></td>
</tr>
</tbody>
</table>

Table 3. Results on the NTU: different attention and alternative strategies.

<table>
<thead>
<tr>
<th>Methods</th>
<th>$L_D$</th>
<th>$L_P$</th>
<th>$L_G$</th>
<th>CS</th>
<th>CV</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global model</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>84.5</td>
<td>91.5</td>
<td>88.0</td>
</tr>
<tr>
<td>Global model</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>85.5</td>
<td>92.1</td>
<td>88.8</td>
</tr>
<tr>
<td>Glimpse Clouds</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
<td>85.7</td>
<td>92.5</td>
<td>89.1</td>
</tr>
<tr>
<td>Glimpse Clouds</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>86.4</td>
<td>93.0</td>
<td>89.7</td>
</tr>
<tr>
<td>Glimpse Clouds</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>86.1</td>
<td>92.9</td>
<td>89.5</td>
</tr>
<tr>
<td>Glimpse Clouds + Global model</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td><strong>86.6</strong></td>
<td><strong>93.2</strong></td>
<td><strong>89.9</strong></td>
</tr>
</tbody>
</table>

Table 1. Results on NTU: ablation study.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Global model</th>
<th>Spatial Attention</th>
<th>Soft Workers</th>
<th>Loss on Pose</th>
<th>CS</th>
<th>CV</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global model only</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>84.5</td>
<td>91.5</td>
<td>88.0</td>
</tr>
<tr>
<td>Global model only</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>85.5</td>
<td>92.2</td>
<td>88.8</td>
</tr>
<tr>
<td>Glimpse Clouds + GRU</td>
<td>-</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
<td>85.8</td>
<td>92.4</td>
<td>89.1</td>
</tr>
<tr>
<td>Glimpse Clouds</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td><strong>86.6</strong></td>
<td><strong>93.2</strong></td>
<td><strong>89.9</strong></td>
</tr>
<tr>
<td>Glimpse clouds + Global model</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>86.6</td>
<td>93.2</td>
<td>89.9</td>
</tr>
</tbody>
</table>

Table 2. Results on NTU: ablation study.
Pose conditioned attention

Joint important for activity → high attention

Joint wrongly located → low attention

[Baradel, Wolf, Mille, Taylor, BMVC 2018]
AI vs. NI

2014 Nobel Prize in Medecine

Head direction

Border cells
AI vs. NI

2014 Nobel Prize in Medicine
AI vs. NI

2018: discovery of the same cells in neural networks trained on similar tasks.

[Cueva, Wei, ICLR 2018]
AI vs. NI

Emergence of the different types of cells in the same order.

[Cueva, Wei, ICLR 2018]
Reasoning : what happened?
Human psychology

- Daniel Kahnemamm (Nobel prize in 2002)
- Book: "Thinking Fast and Slow"
Cognitive tasks

24*17 = ?

Two systems

System 1
- Continuously monitors environment (and mind)
- No specific attention
- Continuously generates assessments / judgments w/o efforts, even in the presence of low data. Jumps to conclusions
- Prone to errors. No capabilities for statistics

System 2
- Receives questions or generates them
- Directs attention and searches memory to find answers
- Requires (eventually a lot of) effort
- More reliable
Where is ML today?

Claim: AI requires a combination of
- Extraction of high-level information from high-dimensional input (visual, audio, language): machine learning
- High-level reasoning: compare, assess, focus attention, perform logical deductions

Roadmap:

- Estimating semantics from low level information (Vision & Learning)
- Estimating causal relationships from data
- Reasoning: Logic + Statistics
Object level Visual Reasoning

[Baradel, Neverova, Wolf, Mille, Mori, ECCV 2018]
Object level Visual Reasoning

[Baradel, Neverova, Wolf, Mille, Mori, ECCV 2018]
Object level Visual Reasoning

[Baradel, Neverova, Wolf, Mille, Mori, ECCV 2018]
Learned interactions

Class: person-book interaction
Failure cases

Confusion between semantically similar objects
- prediction of `hand-cup-contact` instead of `hand-glass-contact`

Small size object
- `hand-cell-phone` contact not detected
## Results

<table>
<thead>
<tr>
<th>Methods</th>
<th>Top1</th>
</tr>
</thead>
<tbody>
<tr>
<td>I3D [5]</td>
<td>27.63</td>
</tr>
<tr>
<td>MultiScale TRN [39]</td>
<td>33.60</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>34.32</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods</th>
<th>Top1</th>
</tr>
</thead>
<tbody>
<tr>
<td>R50 [45]</td>
<td>40.5</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>41.7</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods</th>
<th>Top1</th>
</tr>
</thead>
<tbody>
<tr>
<td>R18 [44]*</td>
<td>32.05</td>
</tr>
<tr>
<td>I3D-18 [3]*</td>
<td>34.20</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>40.89</strong></td>
</tr>
</tbody>
</table>

### Something-something dataset

<table>
<thead>
<tr>
<th>Nb. head</th>
<th>Object type</th>
<th>$f_\phi$</th>
<th>Pairwise relations</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pixel</td>
<td>COCO</td>
<td>RNN</td>
<td>MLP</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Baseline | -   | -   | -   | -   | -   | 29.92 | 33.43 |
| Variant 1| -   | -   | -   | -   | -   | 32.01 | 35.09 |
| Variant 2| -   | -   | -   | -   | -   | 31.36 | 35.15 |
| Variant 3| -   | -   | -   | -   | -   | 32.38 | 34.15 |
| Variant 4| -   | -   | -   | -   | -   | 31.82 | 34.65 |
| **Ours** | -   | -   | -   | -   | -   | **33.75** | **36.12** |
Conclusion

- We propose a models which recognize activities from
  - a cloud of unconstrained feature points
  - Interactions between spatially well defined objects

- Visual spatial attention is useful and competitive compared to pose

- State of the art performance on 5 datasets (NTU RGB-D, Northwestern UCLA, VLOG, Something-Something, Epic Kitchen)

- Reasoning is key component of human cognition, also important for IA systems