PANDA: Pose Aligned Networks for Deep Attribute Modeling

Ning Zhang\textsuperscript{1,2} \quad Manohar Paluri\textsuperscript{1} \quad Marc’Aurelio Ranzato\textsuperscript{1} \quad Trevor Darrell\textsuperscript{2} \quad Lubomir Bourdev\textsuperscript{1}

\textsuperscript{1} Facebook AI Research \quad \textsuperscript{2} EECS, UC Berkeley
Why is attribute classification challenging?
Low resolution

Pose variations

Occlusion
Toward attribute classification

Transfer knowledge

polar bear
black: no
white: yes
brown: no
stripes: no
water: yes
eats fish: yes

[Lampert et al. (CVPR 09), Farhadi et al. (CVPR 09)]

Facial Attribute

[Kumar et al. (ICCV 09)]

Part-based approach

[Bourdev et al. (ICCV11), Zhang et al. (ICCV 13) Joo et al. (ICCV 13)]
Progress in deep learning

**Image classification**

- Krizhevsky et al. NIPS 12, Zeiler et al. ICLR 14

**Object detection**

- R-CNN: Regions with CNN features
  - Girshick et al. CVPR 14

**Human pose estimation**

- Toshev et al. CVPR 14

**Face verification**

- Taigman et al. CVPR 14
Can we train CNN from scratch?

<table>
<thead>
<tr>
<th>method</th>
<th>Joo et al. ICCV 2013</th>
<th>CNN from scratch</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean AP</td>
<td>70.7</td>
<td>58.11</td>
</tr>
</tbody>
</table>

Lack of training data!
What if we finetune from ImageNet?

<table>
<thead>
<tr>
<th>method</th>
<th>Joo et al</th>
<th>from scratch</th>
<th>from ImageNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean AP</td>
<td>70.7</td>
<td>58.11</td>
<td><strong>67.49</strong></td>
</tr>
</tbody>
</table>

How can we simplify the task?

[Donahue et al. ICML 2014]
Decompose the image into parts

Part-based approach

[Bourdev et al. (ICCV11), Zhang et al. (ICCV 13) Joo et al. (ICCV 13)]
Decompose the image into parts

gender
long hair
wear long pants
wear jeans
wear hat
Wear glasses
is sitting
is playing tennis
is dancing
...
Our approach

Pose Aligned Networks for Deep Attribute modeling (PANDA)
Poselets capture part of the pose from a given viewpoint
PANDA

whole person region representation

poselet 1

poselet 2

CNN

CNN

CNN

SVM Classifier

gender
short sleeves
wear hat
wear shorts
long hair

Final representation
Part-Level CNN

Input: Poselet RGB patches

Each poselet → CNN → Output: [attribute 1, attribute 2, ..., attribute N]

Convolutional Neural Network (CNN) architecture:
- Each poselet goes through a series of convolutional layers with increasing sizes.
- The final output is a set of attributes.

Generic attribute layer:
- The attributes are then used in a downstream task, such as image classification or object detection.
Input: Poselet RGB patches

HolisLc

Final Representation

Linear SVM

gender
short sleeves
wear hat
wear shorts
long hair
Dataset: Attribute 25k

Distribution of ground truth labels

- Baby
- Short sleeves
- Sunglasses
- Dress
- Glasses
- Hat
- Long hair
- Male

Positive: Red
Negative: Pink
Unspecified: Green

2061 training examples per poselet on average
RESULTS
Average Precision (AP) on Attribute 25k
Average Precision (AP) on Attribute 25k

- Male
- Long hair
- Hat
- Glasses
- Dress
- Sunglasses
- Short sleeves
- Baby

100% improvement

- Bourdev et al. ICCV13
- DPD Zhang et al. ICCV 13
- PANDA
## Component Evaluation

<table>
<thead>
<tr>
<th>method</th>
<th>mean AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>PANDA (Holistic + Poselets)</td>
<td>70.74</td>
</tr>
</tbody>
</table>
## Component Evaluation

<table>
<thead>
<tr>
<th>method</th>
<th>mean AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>PANDA (Holistic + Poselets)</td>
<td>70.74</td>
</tr>
<tr>
<td>Holistic only</td>
<td>44.97</td>
</tr>
<tr>
<td>Poselets only</td>
<td>64.72</td>
</tr>
</tbody>
</table>
Component Evaluation

<table>
<thead>
<tr>
<th>method</th>
<th>mean AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>PANDA (Holistic + Poselets)</td>
<td>70.74</td>
</tr>
<tr>
<td>Holistic only</td>
<td>44.97</td>
</tr>
<tr>
<td>Poselets only</td>
<td>64.72</td>
</tr>
<tr>
<td>Holistic + DPM</td>
<td>61.20</td>
</tr>
</tbody>
</table>
Poselets vs DPM

Frontal face poselet

Head DPM

Forced to fire no matter what

Mixes different poses

Alignment noise
Transfer learning

Adding new attributes and retrain CNNs

Use the same CNNs only retrain SVM classifier

**smiling**: AP 84.7% (frequency baseline 40.67%)

**walking**: AP 26.0% (frequency baseline 4.34%)

**sitting**: AP 25.70% (frequency baseline 7.65%)
AP on Berkeley Attributes of People Dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bourdev et al.</td>
<td>65.18</td>
</tr>
<tr>
<td>DPD</td>
<td>69.88</td>
</tr>
<tr>
<td>Joo et al.</td>
<td>70.7</td>
</tr>
</tbody>
</table>
AP on Berkeley Attributes of People Dataset

- Bourdev et al.: 65.18
- DPD: 69.88
- Joo et al.: 70.7
- DL-Pure: 58.11
- Holistic (DeCAF): 67.49
The part-level CNNs are trained using Attribute 25k data.
Top scoring examples

wear hat

wear shorts

wear jeans
Failure Cases

Predicted: Long sleeves, Ground truth: short sleeves

Hard to see skin

Predicted: short pants, ground truth: Long pants

Unusual pose

Annotation errors

ambiguous
Gender Recognition on Labeled Faces in the Wild

Much easier dataset – no occlusion, high resolution, centered frontal faces

<table>
<thead>
<tr>
<th>Method</th>
<th>Gender AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kumar et al</td>
<td>95.52</td>
</tr>
<tr>
<td>Frontal face poselet</td>
<td>96.43</td>
</tr>
</tbody>
</table>

[Kumar et al, ICCV 2009]
Gender Recognition on Labeled Faces in the Wild

Much easier dataset – no occlusion, high resolution, centered frontal faces

<table>
<thead>
<tr>
<th>Method</th>
<th>Gender AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kumar et al</td>
<td>95.52</td>
</tr>
<tr>
<td>Frontal face poselet</td>
<td>96.43</td>
</tr>
<tr>
<td><strong>PANDA</strong></td>
<td><strong>99.54</strong></td>
</tr>
</tbody>
</table>

Male of female?

[Kumar et al, ICCV 2009]
Does more data help?

![Graph showing the relationship between the number of training examples in k and Average Precision (AP) for DL-Pure and DL-Poselets. The graph indicates that as the number of training examples increases, the average precision also increases for both models.](image-url)
Comparison

Bourdev et al. ICCV 11

- Use poselet as part-based model
- Has context-level attribute classifier
- Use HOG+color+skin+part masks

PANDA

- Use poselets as part-based model
- Attributes are jointly trained
- Training part-level CNN for powerful discriminative feature
- Generalized much better to new attributes
Conclusion

• Pose-normalization significantly helps deep convolutional networks in the task of attribute classification.

• Mid-level parts remain important in the context of CNNs.
Thanks!

- Code and pre-trained models will be released soon.

*None of the images in this slides are taken from Facebook.*
Running time

- Single CPU
- 13s (poselet detection) +2s (feature extraction)