Relatable Clothing: Detecting Visual Relationships between People and Clothing

Thomas Truong, Svetlana Yanushkevich
Department of Electrical and Computer Engineering
Presentation for ICPR 2020
Friday, November 27, 2020
Introduction

• **Motivation:**
  • Datasets for visual relationships related to clothing are lacking.
  • Consequently, detection models for clothing relationships are also lacking.

• **Research Contributions:**
  • To release a large dataset, the Relatable Clothing Dataset, which can be used for detecting visual relationships between people and worn/unworn clothing.
  • To propose and test a novel model architecture for soft attention and visual relationship detection.
Presentation Outline

• Related works
• Visual Relationship Detection
• Relatable Clothing Dataset
• Soft-attention unit
• Results
• Conclusion
Related Works

• Verbs in COCO (V-COCO) is the most popular visual relationship detection dataset.
  • Very large dataset but does not contain labels for clothing and whether they are worn/unworn.

• Open Images is another popular visual relationship detection dataset
  • Contains the label “wears” for accessories such as sunglasses, necklaces, purses, and shoes, but no clothing is labeled.

• DeepFashion2 contains labels for worn/unworn clothing.
  • There are a few problems to address first before this dataset is useable to train a worn/unworn clothing detector.
Visual Relationship Detection

• Visual Relationship Triplet, <subject, predicate, object> or <S,P,O>.
  • Derived from grammar, subject is the “who/what”, predicate is the “verb” or “relationship”, and object is often a noun which is described in conjunction with the subject and predicate.

• We are concerned with predicate prediction in this paper.
  • The subject will always be a person and the object will always be an article of clothing.
Relatable Clothing Dataset

• DeepFashion2 Dataset lacks two important features that are necessary to be used for visual relationship detection:
  • Subject segmentations. No person is segmented in this dataset.
  • Unworn articles of clothing are close-ups and do not contain any people in the image.
• We propose the Relatable Clothing Dataset for worn/unworn clothing classification problems.
  • A modified subset of the DeepFashion2 Dataset
Relatable Clothing Dataset

• Subject segmentations. No person is segmented in this dataset.
• Unworn articles of clothing are close-ups and do not contain any people in the image.
Relatable Clothing Dataset

- Subject segmentations. No person is segmented in this dataset.
- Unworn articles of clothing are close-ups and do not contain any people in the image.
Relatable Clothing Dataset

• 29852 person-clothing pairs (18726 “worn” and 11126 “unworn”) available for training
• 5705 person-clothing pairs (3604 “worn” and 2101 “unworn”) for validation and testing
Soft-Attention Unit

- A trainable unit which guides the “attention” of the network to the areas containing masks.
  - The Output is added to the output of the 3x3 convolutional layer of each bottleneck unit in ResNet.

```
256-d
1x1, 64
relu
3x3, 64
relu
1x1, 256
relu
```

```
Add
Concatenate
Resize (HxW)
```

```
Conv2D
K filters 3x3 kernel (1,1) stride relu
```

```
Output (HxWxK)
```
Baseline models

• Hard-attention model
  • Primitive masking of the input image using the masks to provide a basic attention mechanism.

• Box soft-attention model
  • Similar to previous works who use bounding box detections to do visual relationship detection.
### Results

**Performance Metrics for the Proposed Soft Attention Models.**

<table>
<thead>
<tr>
<th>Soft Attention Backbone</th>
<th>Soft Attention Units</th>
<th>Trainable Parameters</th>
<th>Accuracy (%)</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>Specificity (%)</th>
<th>F₁ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet50V2</td>
<td>1</td>
<td>26,275,713</td>
<td>96.00 ± 1.03</td>
<td>98.79 ± 0.56</td>
<td>94.83 ± 1.41</td>
<td>97.98 ± 0.96</td>
<td>96.76 ± 0.85</td>
</tr>
<tr>
<td>ResNet50V2</td>
<td>16</td>
<td>26,379,649</td>
<td>97.74 ± 0.40</td>
<td>97.76 ± 0.61</td>
<td>98.66 ± 0.46</td>
<td>96.17 ± 0.87</td>
<td>98.21 ± 0.36</td>
</tr>
<tr>
<td>ResNet101V2</td>
<td>1</td>
<td>45,285,249</td>
<td>97.97 ± 0.63</td>
<td>98.96 ± 0.33</td>
<td>97.79 ± 0.98</td>
<td>98.24 ± 0.49</td>
<td>98.37 ± 0.54</td>
</tr>
<tr>
<td>ResNet101V2</td>
<td>33</td>
<td>45,511,041</td>
<td>98.55 ± 0.35</td>
<td>99.16 ± 0.40</td>
<td>98.52 ± 0.50</td>
<td>98.58 ± 0.65</td>
<td>98.84 ± 0.29</td>
</tr>
</tbody>
</table>

**Performance Metrics for the Hard Attention Models.**

<table>
<thead>
<tr>
<th>Backbone</th>
<th>Accuracy (%)</th>
<th>Precision (%)</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>F₁ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet50V2</td>
<td>92.52 ± 1.06</td>
<td>97.17 ± 1.00</td>
<td>90.79 ± 1.48</td>
<td>95.50 ± 1.52</td>
<td>93.87 ± 0.91</td>
</tr>
<tr>
<td>ResNet101V2</td>
<td>94.11 ± 0.91</td>
<td>95.94 ± 0.72</td>
<td>94.67 ± 1.25</td>
<td>93.17 ± 0.89</td>
<td>95.30 ± 0.77</td>
</tr>
<tr>
<td>InceptionV3</td>
<td>92.59 ± 0.93</td>
<td>94.76 ± 1.14</td>
<td>93.43 ± 0.96</td>
<td>91.17 ± 1.85</td>
<td>94.08 ± 0.75</td>
</tr>
<tr>
<td>InceptionResNetV2</td>
<td>93.51 ± 0.70</td>
<td>94.27 ± 0.81</td>
<td>95.53 ± 0.82</td>
<td>90.04 ± 1.40</td>
<td>94.89 ± 0.60</td>
</tr>
</tbody>
</table>
## Results

### PERFORMANCE METRICS FOR THE PROPOSED SOFT ATTENTION MODELS.

<table>
<thead>
<tr>
<th>Soft Attention Backbone</th>
<th>Soft Attention Units</th>
<th>Trainable Parameters</th>
<th>Accuracy (%)</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>Specificity (%)</th>
<th>F₁ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet50V2</td>
<td>1</td>
<td>26,275,713</td>
<td>96.00 ± 1.03</td>
<td>98.79 ± 0.56</td>
<td>94.83 ± 1.41</td>
<td>97.98 ± 0.96</td>
<td>96.76 ± 0.85</td>
</tr>
<tr>
<td>ResNet50V2</td>
<td>16</td>
<td>26,379,649</td>
<td>97.74 ± 0.40</td>
<td>97.76 ± 0.61</td>
<td><strong>98.66 ± 0.46</strong></td>
<td>96.17 ± 0.87</td>
<td>98.21 ± 0.36</td>
</tr>
<tr>
<td>ResNet101V2</td>
<td>1</td>
<td>45,285,249</td>
<td>97.97 ± 0.63</td>
<td>98.96 ± 0.33</td>
<td>97.79 ± 0.98</td>
<td>98.24 ± 0.49</td>
<td>98.37 ± 0.54</td>
</tr>
<tr>
<td>ResNet101V2</td>
<td>33</td>
<td>45,511,041</td>
<td><strong>98.55 ± 0.35</strong></td>
<td><strong>99.16 ± 0.40</strong></td>
<td>98.52 ± 0.50</td>
<td>98.58 ± 0.65</td>
<td>98.84 ± 0.29</td>
</tr>
</tbody>
</table>

### PERFORMANCE METRICS FOR THE BOX ATTENTION MODELS.

<table>
<thead>
<tr>
<th>Backbone</th>
<th>Soft Attention Units</th>
<th>Accuracy (%)</th>
<th>Precision (%)</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>F₁ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet50V2</td>
<td>1</td>
<td>90.99 ± 1.09</td>
<td>91.90 ± 1.50</td>
<td>94.04 ± 0.88</td>
<td>85.78 ± 2.39</td>
<td>92.95 ± 0.87</td>
</tr>
<tr>
<td>ResNet50V2</td>
<td>16</td>
<td>93.99 ± 0.53</td>
<td><strong>98.81 ± 0.61</strong></td>
<td>91.58 ± 0.96</td>
<td><strong>98.09 ± 1.05</strong></td>
<td>95.05 ± 0.52</td>
</tr>
<tr>
<td>ResNet101V2</td>
<td>1</td>
<td><strong>95.37 ± 0.76</strong></td>
<td>94.79 ± 1.04</td>
<td>98.04 ± 0.61</td>
<td>90.79 ± 1.58</td>
<td>96.38 ± 0.66</td>
</tr>
<tr>
<td>ResNet101V2</td>
<td>33</td>
<td>95.14 ± 0.89</td>
<td>97.98 ± 1.03</td>
<td>94.27 ± 1.18</td>
<td>96.69 ± 1.70</td>
<td>96.08 ± 0.71</td>
</tr>
</tbody>
</table>
# Results

<table>
<thead>
<tr>
<th>Input Image</th>
<th>Person 1 Vest</th>
<th>Person 2 Vest</th>
<th>Person 3 Vest</th>
<th>Person 1 Helmet</th>
<th>Person 2 Helmet</th>
<th>Person 3 Helmet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person 1</td>
<td>1.000</td>
<td>0.003</td>
<td>0.527</td>
<td>1.000</td>
<td>0.001</td>
<td>0.889</td>
</tr>
<tr>
<td>Person 2</td>
<td>0.100</td>
<td>1.000</td>
<td>0.449</td>
<td>0.848</td>
<td>1.000</td>
<td>0.884</td>
</tr>
<tr>
<td>Person 3</td>
<td>0.069</td>
<td>0.000</td>
<td>1.000</td>
<td>0.697</td>
<td>0.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Conclusions and Future Work

• Release of the Relatable Clothing Dataset
  • 29852 person-clothing pairs for training, 5705 person-clothing pairs for validation and testing.

• Proposal of a novel soft-attention unit for visual relationship detection.
  • Demonstrated good performance for worn/unworn clothing detection on the Relatable Clothing Dataset and decent generalizability on unseen articles of clothing.

• Currently extending these works for full end-to-end object detection and visual relationship detection for applications in safety and security.
Acknowledgements

- NSERC “Biometric-enabled Identity Management and Risk Assessment for Smart Cities”
- Department of National Defence’s Innovation for Defence Excellence and Security (IDEaS)