Joint COCO and Mapillary Recognition Challenge Workshop

Sunday, September 9th, ECCV 2018

Alexander Kirillov, Facebook AI Research
COCO Panoptic Dataset
2018 Panoptic Segmentation Dataset

- For each pixel $i$ predict semantic label $l$ and instance id $z$
- no overlaps between segments by design
COCO annotations have overlaps
Most overlaps can be resolved automatically
~25k overlaps require manual resolution
2018 Panoptic Segmentation Dataset

Instructions:

In each row click on the image with better objects layout.

Draw cup in front

Draw table in front

Click to see more examples.

Task:

Draw person in front

Draw tennis racket in front
2018 Panoptic Segmentation Dataset

- train: 118k, val: 5k, test-dev: 20k, test-challenge: 20k
- 80 things categories, 53 stuff categories
**Panoptic Quality Measure**

PQ Computation:
- Step 1: Matching
- Step 2: Calculation
Theorem: For panoptic segmentation problem each ground truth segment can have at most one corresponding predicted segment with IoU greater than 0.5.

Proof sketch:

If \( \text{IoU} > 0.5 \), then there is no other non-overlapping object that has IoU \( > 0.5 \).
Panoptic Quality (PQ): Matching

Ground Truth

Prediction

\[ TP = \{(\text{ }, \text{ }),(\text{ }, \text{ })\} \]

\[ FP = \{\text{ }\} \]

\[ FN = \{\text{ }\} \]
Panoptic Quality (PQ): Calculation

\[ PQ = \sum_{(p, g) \in TP} \text{IoU}(p, g) \]
\[ \frac{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \]
Panoptic Quality (PQ): Calculation

Ground Truth

\[ \text{PQ} = \frac{\sum_{(p,g) \in TP} \text{IoU}(p,g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} = \frac{\sum_{(p,g) \in TP} \text{IoU}(p,g)}{|TP|} \times \frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \]

Segmentation Quality (SQ)

Recognition Quality (RQ)

Prediction
COCO Panoptic Metrics

Average Panoptic Metrics:
- PQ: % Panoptic Quality (primary challenge metric)
- SQ: % Segmentation Quality component of PQ
- RQ: % Recognition Quality component of PQ

Panoptic Metrics for Things Categories:
- $PQ_{th}^t$: % PQ for things categories only
- $SQ_{th}^t$: % SQ for things categories only
- $RQ_{th}^t$: % RQ for over things categories only

Panoptic Metrics for Stuff Categories:
- $PQ_{st}^t$: % PQ for stuff categories only
- $SQ_{st}^t$: % SQ for stuff categories only
- $RQ_{st}^t$: % RQ for stuff categories only
5000 COCO images were annotated independently twice
COCO Annotations Consistency

5000 COCO images were annotated independently twice

<table>
<thead>
<tr>
<th></th>
<th>PQ</th>
<th>SQ</th>
<th>RQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>53.5</td>
<td>82.6</td>
<td>63.9</td>
</tr>
<tr>
<td>Things</td>
<td>57.8</td>
<td>81.4</td>
<td>69.7</td>
</tr>
<tr>
<td>Stuff</td>
<td>47.1</td>
<td>84.3</td>
<td>55.2</td>
</tr>
</tbody>
</table>

- Crowd sourced annotations are very noisy
5000 COCO images were annotated independently twice.

<table>
<thead>
<tr>
<th></th>
<th>PQ</th>
<th>SQ</th>
<th>RQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>53.5</td>
<td>82.6</td>
<td>63.9</td>
</tr>
<tr>
<td>Things</td>
<td>57.8</td>
<td>81.4</td>
<td>69.7</td>
</tr>
<tr>
<td>Stuff</td>
<td>47.1</td>
<td>84.3</td>
<td>55.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>PQ</th>
<th>SQ</th>
<th>RQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>25.2</td>
<td>62.1</td>
<td>32.8</td>
</tr>
<tr>
<td>Medium</td>
<td>53.5</td>
<td>81.7</td>
<td>64.6</td>
</tr>
<tr>
<td>Large</td>
<td>69.6</td>
<td>87.5</td>
<td>78.3</td>
</tr>
</tbody>
</table>

- Crowd sourced annotations are very noisy
- Annotations are highly inconsistent for small objects
Annotations Consistency $<$ Human Performance
Annotations Consistency < Human Performance

real GT = [0, 0, 0, 0, 0, 1, 1, 1, 1, 1]
noisy annotator 1 = [0, 0, 0, 0, 1, 1, 1, 1, 1, 1]
noisy annotator 2 = [0, 0, 0, 0, 0, 0, 1, 1, 1, 1]
Annotations Consistency $<$ Human Performance

real GT $= [0, 0, 0, 0, 0, 1, 1, 1, 1, 1]
noisy annotator 1 $= [0, 0, 0, 0, 1, 1, 1, 1, 1, 1]
noisy annotator 2 $= [0, 0, 0, 0, 0, 0, 1, 1, 1, 1]

Accuracy(real GT, annotator_1) $= 0.9$
Accuracy(real GT, annotator_2) $= 0.9$
Accuracy(annotator_1, annotator_2) $= 0.8$
Annotations Consistency $<$ Human Performance

<table>
<thead>
<tr>
<th>Annotation Type</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real GT</td>
<td>0.9</td>
</tr>
<tr>
<td>Noisy Annotator 1</td>
<td>0.9</td>
</tr>
<tr>
<td>Noisy Annotator 2</td>
<td>0.8</td>
</tr>
<tr>
<td>Ideal Annotator</td>
<td>?</td>
</tr>
</tbody>
</table>

real GT $= [0, 0, 0, 0, 0, 1, 1, 1, 1, 1]$  
noisy annotator 1 $= [0, 0, 0, 0, 1, 1, 1, 1, 1, 1]$  
noisy annotator 2 $= [0, 0, 0, 0, 0, 0, 1, 1, 1, 1]$
11 teams joined the competition
11 teams joined the competition
4 teams achieved better performance than the baseline (RN50 Mask R-CNN + RN50 FPN-FCN)
11 teams joined the competition
4 teams achieved better performance than the baseline (RN50 Mask R-CNN + RN50 FPN-FCN)
Summary of Findings

2018 COCO Panoptic Challenge Take-aways

- All submission above the baseline combined the outputs of two separate networks for stuff and things
Summary of Findings

2018 COCO Panoptic Challenge Take-aways

- All submission above the baseline combined the outputs of two separate networks for stuff and things
- Best submission showed better PQ for things categories than the human consistency experiment

<table>
<thead>
<tr>
<th></th>
<th>PQ</th>
<th>SQ</th>
<th>RQ</th>
<th>PQ&lt;sup,Th&lt;/sup&gt;</th>
<th>SQ&lt;sup,Th&lt;/sup&gt;</th>
<th>RQ&lt;sup,Th&lt;/sup&gt;</th>
<th>PQ&lt;sup,St&lt;/sup&gt;</th>
<th>SQ&lt;sup,St&lt;/sup&gt;</th>
<th>RQ&lt;sup,St&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Consistency</td>
<td>53.5</td>
<td>82.6</td>
<td>63.9</td>
<td>57.8</td>
<td>81.4</td>
<td>69.7</td>
<td>47.1</td>
<td>84.3</td>
<td>55.2</td>
</tr>
<tr>
<td>Megvii (Face++)</td>
<td>53.8</td>
<td>83.4</td>
<td>63.6</td>
<td>62.8</td>
<td>85.7</td>
<td>73.1</td>
<td>40.2</td>
<td>80.0</td>
<td>49.2</td>
</tr>
</tbody>
</table>
2018 COCO Panoptic Challenge Take-aways

- All submission above the baseline combined the outputs of two separate networks for stuff and things.
- Best submission showed better PQ for things categories than the human consistency experiment.

<table>
<thead>
<tr>
<th>Things</th>
<th>PQ\text{Th}</th>
<th>SQ\text{Th}</th>
<th>RQ\text{Th}</th>
<th>all</th>
<th>TP</th>
<th>all</th>
<th>FP</th>
<th>all</th>
<th>FN</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Consistency</td>
<td>57.8</td>
<td>81.4</td>
<td>69.7</td>
<td>24890</td>
<td>8860</td>
<td>9628</td>
<td>72.6%</td>
<td>69.5%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Megvii (Face++)</td>
<td>62.8</td>
<td>85.7</td>
<td>73.1</td>
<td>24205</td>
<td>4929</td>
<td>10313</td>
<td>81.1%</td>
<td>67.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- The result suggests ability to learn models with low noise level from large-scale noisy data.
- Accuracy of the test set ground truth needs to be improved in the future.
Prediction Examples

Image

Prediction
Megvii (Face++)

GT
Prediction Examples

Image

Prediction
Megvii (Face++)

GT
Prediction Examples

Image

Prediction
Megvii (Face++)

GT
Prediction Examples

Image

Prediction
Megvii (Face++)

GT

dog
Prediction Examples

Image

Prediction
Megvii (Face++)

GT
11 teams joined the competition
4 teams achieved better performance than the baseline (RN50 Mask R-CNN + RN50 FPN-FCN)
11 teams joined the competition
4 teams achieved better performance than the baseline (RN50 Mask R-CNN + RN50 FPN-FCN)

*External segmentation datasets were used
2018 COCO Panoptic Challenge

**Invited Speakers:**

Team Megvii / (9:15am – 9:45am)
Team Caribbean / (11:30am – 11:45am)
Team PKU_360 / (11:45am – 12:00pm)

<table>
<thead>
<tr>
<th>Team</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Megvii (Face++)</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
</tr>
<tr>
<td>Caribbean</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
</tr>
<tr>
<td>PKU_360</td>
<td>3&lt;sup&gt;rd&lt;/sup&gt;</td>
</tr>
</tbody>
</table>