SeeFood

CS231N (SPRING 2022)

Wai Tong Chung (wtchung@stanford.edu)
Mahesh Saboo (mssaboo@stanford.edu)
Chiin-Zhe Tan (chiin@stanford.edu)
Introduction

Objective
- Food detection from a photo
- Identify food objects present
- Feed ingredients directly into recipe database
- Automate and augment decision process

Background
- Universality of shopping and cooking experiences
- Significant potential for technological facilitation
- Existing models have a more general scope
Problem Statement

- Using computer vision techniques, build a food detector
- Train a model on a dataset of food objects
- Compare with baseline model using qualitative and quantitative metrics

Quantitative metric: Mask Average Precision

- Average Precision (using IoU and a threshold)
- Compute for thresholds from 0.5 to 0.95 (at 0.05 increments)
- Average over classes and thresholds
Mask R-CNN

- Mask R-CNN for Instance Segmentation
- Extension to Faster R-CNN
- Additional loss term for predicting segmentation mask on each RoI
- Used the model `mask_rcnn_R_50_FPN_3x` from detectron2
- Backbone of ResNet-50-FPN
- ResNet-FPN backbones for feature extraction with Mask R-CNN have been known to give excellent gains in both accuracy and speed

Overview of Mask R-CNN
Results: Baseline (Pretrained on MSCOCO)

<table>
<thead>
<tr>
<th>Class</th>
<th>banana</th>
<th>apple</th>
<th>sandwich</th>
<th>orange</th>
<th>broccoli</th>
<th>carrot</th>
<th>pizza</th>
<th>donut</th>
<th>cake</th>
<th>hotdog</th>
</tr>
</thead>
<tbody>
<tr>
<td>seg AP</td>
<td>25.6</td>
<td>24.3</td>
<td>42.3</td>
<td>32.5</td>
<td>28.9</td>
<td>24.7</td>
<td>56.9</td>
<td>55.1</td>
<td>37.5</td>
<td>28.6</td>
</tr>
<tr>
<td>box AP</td>
<td>29.7</td>
<td>25.1</td>
<td>42.2</td>
<td>32.9</td>
<td>30.7</td>
<td>27.9</td>
<td>58.6</td>
<td>55.1</td>
<td>37.6</td>
<td>31.7</td>
</tr>
</tbody>
</table>
Results: Our Model (Trained with a processed LVIS)

<table>
<thead>
<tr>
<th>Class</th>
<th>banana</th>
<th>apple</th>
<th>sandwich</th>
<th>orange</th>
<th>broccoli</th>
<th>carrot</th>
<th>pizza</th>
<th>donut</th>
<th>cake</th>
<th>hotdog</th>
</tr>
</thead>
<tbody>
<tr>
<td>seg AP</td>
<td>11.1</td>
<td>13.3</td>
<td>11.0</td>
<td>14.6</td>
<td>15.2</td>
<td>11.5</td>
<td>4.1</td>
<td>28.6</td>
<td>22.5</td>
<td>7.8</td>
</tr>
<tr>
<td>box AP</td>
<td>12.7</td>
<td>12.8</td>
<td>10.8</td>
<td>14.0</td>
<td>11.2</td>
<td>15.3</td>
<td>8.6</td>
<td>28.4</td>
<td>21.7</td>
<td>7.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>lettuce</th>
<th>mushroom</th>
<th>nut</th>
<th>onion</th>
<th>pastry</th>
<th>potato</th>
<th>strawberry</th>
<th>tomato</th>
<th>wine</th>
</tr>
</thead>
<tbody>
<tr>
<td>seg AP</td>
<td>1.8</td>
<td>1.0</td>
<td>0.6</td>
<td>0.5</td>
<td>0.6</td>
<td>0.4</td>
<td>1.3</td>
<td>4.7</td>
<td>13.4</td>
</tr>
<tr>
<td>box AP</td>
<td>1.8</td>
<td>1.0</td>
<td>0.9</td>
<td>0.5</td>
<td>0.6</td>
<td>0.3</td>
<td>1.1</td>
<td>4.7</td>
<td>13.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>beer</th>
<th>bell pepper</th>
<th>blueberry</th>
<th>bread</th>
<th>crumb</th>
<th>cupcake</th>
<th>grape</th>
<th>lemon</th>
</tr>
</thead>
<tbody>
<tr>
<td>seg AP</td>
<td>2.7</td>
<td>0.1</td>
<td>0.6</td>
<td>4.7</td>
<td>0.3</td>
<td>9.6</td>
<td>1.4</td>
<td>0.2</td>
</tr>
<tr>
<td>box AP</td>
<td>2.1</td>
<td>0.1</td>
<td>0.5</td>
<td>4.9</td>
<td>0.3</td>
<td>8.5</td>
<td>1.4</td>
<td>0.2</td>
</tr>
</tbody>
</table>
Conclusions

- Assembled comprehensive food database
- Used transfer learning to train model to detect 17 new classes
- Qualitative results: promising, with failures on similar food objects
- Quantitative results: less promising
- Future work: further annotation, focusing on mispredicted classes