Understanding Scene in the Wild

SenseCUSceneParsing

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi
Xiaogang Wang, Tong Xiao, Jiaya Jia
Features of ADE20K Dataset

• Number of image
  • ADE20K Dataset: 20k

• Number of scene
  • Image label: 1038

• Number of semantic label
  • Wall / Building, Field / Earth, Mountain / Hill, Stair / Stairway......
Our baseline

• Pretrained Resnet 101 + FCN pixel prediction
• Result regarding to mIOU / pixel accuracy
  • Train Data: 70.20 / 90.34
  • Val Data: 35.08/76.87
• Of course, we have many pre-baselines that are not so good
Our result improves

— Evils in the details

• Various data augmentation
• Dropout to the last convolution layers
• Using dilated convolution
• Learning rate policy
• Total iteration number
• Correct way to use batch normalization
• Larger cropsize and larger receptive field
• ......

Code and model will be released later
Our result improves

— Evils in the details

• Previous baseline result by mIOU / pixel accuracy
  • Train Data: 70.20 / 90.34
  • Val Data: 35.08/76.87

• Current Resnet101 result
  • Train Data: 75.16/91.99
  • Val Data: 36.85/77.65
Our result improves

- Deeply supervise for better optimization

Resnet 101, to conv4_26

Resnet 101, to conv5_3

Convolution

Convolution

Softmax Loss

Softmax Loss

Auxiliary loss, loss weight 0.4
Our result improves

— Deep resnet improves by additional loss

• Previous result by mIOU / pixel accuracy
  • Train Data: 75.16/91.99
  • Val Data: 36.85/77.65

• Current result by deeply supervised training
  • Train Data: 77.70/93.15
  • Val Data: 38.28/78.63

• Better optimization policy improves confusion label and inconspicuous object
Image level information may help scene parsing

- A failure example
Recognize scene in image level

- State-of-the-art Image classification
  - FCN + Average Pooling
- Classical scene understanding
  - Spatial Pyramid Matching
- Better scene recognition
  - FCN + Spatial Pyramid Matching Pooling
Pixel Prediction with Image Level Information

- Utilizing image level information for scene parsing
- End to end learning
- Marginal computation cost
Our result improves

— Spatial pyramid global feature

• Previous result by mIOU / pixel accuracy
  • Train Data: 77.70/93.15
  • Val Data: 38.28/78.63

• Current result by image level information
  • Train Data: 79.51/93.65
  • Val Data: 41.29/80.04

• Global feature improves error failure to sense image label
### Deeper Pretrained Model

<table>
<thead>
<tr>
<th>Pretrained Model</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resnet 50</td>
<td>40.11/79.55</td>
</tr>
<tr>
<td>Resnet 101</td>
<td>41.29/80.04</td>
</tr>
<tr>
<td>Resnet 152</td>
<td>42.23/80.46</td>
</tr>
<tr>
<td>Resnet 269</td>
<td>43.39/80.90</td>
</tr>
</tbody>
</table>

- Better and Deeper pretrained model improves the result consistently
Testing and Ensemble

<table>
<thead>
<tr>
<th>Method</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resnet 269 Single Scale Test</td>
<td>43.39/80.90</td>
</tr>
<tr>
<td>Resnet 269 Multi Scale Test</td>
<td>44.59/81.80</td>
</tr>
<tr>
<td>Ensemble of 5 Models</td>
<td>45.95/82.48</td>
</tr>
</tbody>
</table>

- Proper testing scheme improves the result
- But it is time consuming and only useful for competitions
Summary

Baseline: 35.08/76.87
Details & Tricks: 36.85/77.65
Deep Supervise: 38.28/78.63
Global Feature: 41.29/80.04
Deeper Model: 43.39/80.90
Multi Scale Test: 44.59/81.80
Ensemble: 45.95/82.48
Final Result: 45.95/82.48

Ensemble: 1.36
Baseline: 1.77
Details & Tricks: 1.43
Deep Supervise: 3.01
Global Feature: 1.20
Deeper Model: 2.10
Multi Scale Test: 1.36
Final Result: 2.58
Visual Results
Visual Results
Visual Results
Visual Results
Visual Results
Visual Results
Future Direction

• More labeled data, and more clear definition
• Use of human semantic similarity matrix
• Small and rare object
• Scene parsing in video
• Speedup
It is not yet finished...
Learn by failure – Balance Sample

• Sample training image to uniform distribution
• Better training accuracy
• But overfitting and worse validation accuracy

Blue is baseline training accuracy, Red is training accuracy after balance sample
Learn by no significant improvement

- Hard sample mining
- CRF
- Stochastic depth
- Using predefined class correlation
- ......
Thanks & Questions