Places Challenge 2017
Scene Parsing
WinterIsComing
Riwei Chen, Qi Chen, Xinglong Wu
Yifan Lu, Yudong Jiang, Linfu Wen
Outline

• Single Model Results
• Method Overview
• Method Details
  • Model Pretraining
  • Pyramid Pooling
  • Batch Size & BN
  • Other details
  • Submissions
• Visual Results
• Future Direction
Features of ADE20K Dataset—Scene Parsing

• Number of image
  • Training: 20K
  • Validation: 2K
  • Testing: 3K

• Number of category
  • Semantic category: 150
Single Model Results on Validation Set

- Single model
  - Compared with the best single model result of 2016

<table>
<thead>
<tr>
<th>Team</th>
<th>mIoU</th>
<th>pixel accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SenseCuSceneParsing(^1)</td>
<td>43.39%</td>
<td>80.90%</td>
</tr>
<tr>
<td>Adelaide(^2)*</td>
<td>43.06%</td>
<td>80.53%</td>
</tr>
<tr>
<td>WinterIsComming(ours)</td>
<td>43.98%</td>
<td>81.13%</td>
</tr>
</tbody>
</table>

\(^1\) Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network, CVPR 2017

* The result of “Model C, 2 conv”
Method Overview

- Base Network: ResNet38
- Pyramid Pooling
- ImageNet and Places2 pretraining
- Batch Size is critical
- Ensemble models trained with different epochs
Network Structure

* Our implement is based on: https://github.com/itijyou/ademxapp
Building Blocks

(a) Res_Blk_A
(b) Res_Blk_B
(c) Res_Blk_C
Res-MobileNet

![Diagram of Res-MobileNet](image)

<table>
<thead>
<tr>
<th>Model</th>
<th>Computation (macc)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet50</td>
<td>109.4G</td>
</tr>
<tr>
<td>Res-MobileNet</td>
<td>32.5G</td>
</tr>
<tr>
<td>ResNet38</td>
<td>415.5G</td>
</tr>
<tr>
<td>VGG16</td>
<td>618.0G</td>
</tr>
</tbody>
</table>

* The computation cost of models when input size is 512x512
Model Performance

Pyramid Pooling
Pyramid Pooling

- Pyramid Pooling improves the integrity of segmentation
Pretraining

- ResNet50 without ImageNet pretraining has the lowest accuracy
- Places2 pretraining helps improve accuracy
Batch size & Batch Norm

- Training batch size is critical
- Experiment with Res-MobileNet
- ResNet38 w/o PP, batch size = 6
- After adding PP, batch size = 2
- Usually use 4 GTX 1080Ti GPUs

<table>
<thead>
<tr>
<th>Training Batch Size per GPU</th>
<th>Testing Pixel Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>68.4%</td>
</tr>
<tr>
<td>2</td>
<td>69.7%</td>
</tr>
<tr>
<td>4</td>
<td>70.7%</td>
</tr>
<tr>
<td>finetune with fixed BN</td>
<td>72.9%</td>
</tr>
<tr>
<td>finetune ImageNet pretrained model with fixed BN</td>
<td>74.1%</td>
</tr>
</tbody>
</table>
Other Details

- Training augmentation
  - Multi-scale: [0.7, 1.3]
  - Flip
  - Random crop to 512x512

- Testing augmentation
  - Flip
  - No multi-scale

- SGD solver with lr = 1e-4 for 64 epochs
Submissions

• Submit 1: train with only ADE20K training set
  • we get 81.13%/43.98% pixel accuracy/mIOU on validation set
• Submit 2-4: finetune the model with both training and validation set for 5, 22, 29 epochs respectively
• Submit 5: ensemble submit 1-4 models by voting
Summary

- Pretraining is critical and datasets of similar tasks work better
- Batch size should be large enough
- Fix BN params can further improve result (when batchsize is small)
- Pyramid Pooling can improve region integrity of segmentation
Visual Results

Image | Ground Truth | without Pyramid Pooling | with Pyramid Pooling
Future work

- Memory-efficient deep learning framework
- Well-Pretrained Res-MobileNet
- Focal loss
- Expert model
Thanks & Questions