ExFuse: Enhancing Feature Fusion for Semantic Segmentation

ECCV 2018
Zhenli Zhang\textsuperscript{1} Xiangyu Zhang\textsuperscript{2} Chao Peng\textsuperscript{2} Dazhi Cheng\textsuperscript{3} Jian Sun\textsuperscript{2}
\textsuperscript{1}Fudan University \textsuperscript{2}Megvii Inc. (Face++) \textsuperscript{3}Beijing Institute of Technology
Motivation

An ordinary U-Net Structure
Motivation

An ordinary U-Net Structure

• Low-level features:
  • Many structure details
  • Hardly any semantic information

• High-level features:
  • Rich semantic information
  • Hardly any structure details
Motivation

The fusing problem of U-Net Structure

• Large semantic and resolution gap
  • Low level feature: the inner details may contain noises
  • High level feature: after upsample, the boundary may be not correct
Solution

Reduce the fusing gap

• Use high level feature to refine the low level feature map, discarding some noises.

• Embed more spatial information into high-level features.
Solution

Reduce the fusing gap
Method

The network architecture
Method
The network architecture

• Basically a U-Net structure:
  • Encoder: down-sample
  • Decoder: up-sample

• Backbone:
  • ResNet
Method

The network architecture
Network Details

**SS: semantic supervision**

- Also called auxiliary supervisions, only used in training process.
- take account of the loss of all the auxiliary branches
- force the low level features to contain more semantic information.
Network Details

Layer Rearrangement

• To make low level features (res-2 or res-3) 'closer' to the supervisions, arrange more layers in the early stages rather than the latter.

• ResNeXt 101 model has \{3; 4; 23; 3\} building blocks for Stage 2-5 respectively; we rearrange the assignment into \{8; 8; 9; 8\}
Method

The network architecture
Network Details

**SEB: Semantic Embedding Branch**

- Use rich semantic information to refine the low-level feature
- Just like attention
Method
The network architecture

Global Convolutional Network
Network Details

**GCN: Global Convolutional Network**

- Take apart the convolution kernel
- \(k \times k \to k \times 1 + 1 \times k\)
- Large kernel size with low computation cost
  - Larger receptive field
Method
The network architecture

Explicit Channel Resolution Embedding
Network Details

ECRE: Explicit Channel Resolution Embedding

• Method from ESPCN
• Group sub-pixels from 4 channels to up-sample

Method

The network architecture
Network Details

Densely Adjacent Prediction

- To get the final segmentation map, result at each position can be generated by averaging the associated scores.
- Neighboring channels are associated
- Output channel: 189 -> 21*3*3
  - 21 is the PASCAL VOC dataset classes number
  - 3*3 channels are contained in each group for average


2019/4/15    Zhixuan Li
Experiments Settings

• Pretrain: Microsoft COCO dataset
• Finetune: PASCAL VOC 2012 (Enhanced with SBD)
• evaluation metric:

\[ mIoU = \frac{TP}{FN + TP + FP} \]
### Ablation Study

<table>
<thead>
<tr>
<th>Index</th>
<th>Baseline</th>
<th>SS</th>
<th>LR</th>
<th>ECRE</th>
<th>SEB</th>
<th>DAP</th>
<th>mIoU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>76.0</td>
</tr>
<tr>
<td>2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>77.5</td>
</tr>
<tr>
<td>3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>78.3</td>
</tr>
<tr>
<td>4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>78.8</td>
</tr>
<tr>
<td>5</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>79.0</td>
</tr>
<tr>
<td>6</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>79.6</td>
</tr>
<tr>
<td>7</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>80.0</td>
</tr>
</tbody>
</table>

Table 3. Ablation experiments of the methods in Sec 3. Performances are evaluated by standard mean IoU(%) on PASCAL VOC 2012 validation set. The baseline model is [8] (our impl.) SS – semantic supervision. LR – layer rearrangement. ECRE – explicit channel resolution embedding. SEB – semantic embedding branch. DAP – densely adjacent prediction.
## Ablation Study

<table>
<thead>
<tr>
<th>Module</th>
<th>Index compared</th>
<th>Improvement(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS - semantic supervision</td>
<td>1, 2</td>
<td>1.5</td>
</tr>
<tr>
<td>LR - layer rearrangement</td>
<td>2, 3</td>
<td>0.8</td>
</tr>
<tr>
<td>ECRE - explicit channel resolution embedding</td>
<td>3, 4</td>
<td>0.5</td>
</tr>
<tr>
<td>ECRE</td>
<td>6, 7</td>
<td>0.4</td>
</tr>
<tr>
<td>SEB - semantic embedding branch</td>
<td>3, 5</td>
<td>0.7</td>
</tr>
<tr>
<td>DAP - densely adjacent prediction</td>
<td>5, 6</td>
<td>0.6</td>
</tr>
</tbody>
</table>
Rank

Achieved state-of-the-art with the same experiment setting.

<table>
<thead>
<tr>
<th>Method</th>
<th>mIOU%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep Layer Cascade (LC) (Li et al. 2017)</td>
<td>82.7</td>
</tr>
<tr>
<td>TuSimple (Wang et al. 2017)</td>
<td>83.1</td>
</tr>
<tr>
<td>Large Kernel Matters (Peng et al. 2017)</td>
<td>83.6</td>
</tr>
<tr>
<td>Multipath-ReneNet (Lin et al. 2017)</td>
<td>84.2</td>
</tr>
<tr>
<td>ResNet-38 MS COCO (Wu et al. 2016)</td>
<td>84.9</td>
</tr>
<tr>
<td>PSPNet (Zhao et al. 2017)</td>
<td>85.4</td>
</tr>
<tr>
<td>ExFuse ResNet101 (Zhang et al. 2018)</td>
<td>86.2</td>
</tr>
<tr>
<td>IDW-CNN (Wang et al. 2017)</td>
<td>86.3</td>
</tr>
<tr>
<td>CASIA IVA SDN (Fu et al. 2017)</td>
<td>86.6</td>
</tr>
<tr>
<td>DIS (Luo et al. 2017)</td>
<td>86.8</td>
</tr>
<tr>
<td>DeepLabv3 (Chen et al. 2017)</td>
<td>85.7</td>
</tr>
<tr>
<td>DeepLabv3-JFT (Chen et al. 2017)</td>
<td>86.9</td>
</tr>
<tr>
<td>DeepLabv3+ (Xception) (Chen et al. 2018)</td>
<td>87.8</td>
</tr>
<tr>
<td>ExFuse ResNeXt131 (Zhang et al. 2018)</td>
<td>87.9</td>
</tr>
<tr>
<td>DeepLabv3+ (Xception-JFT) (Chen et al. 2018)</td>
<td>89</td>
</tr>
</tbody>
</table>
Conclusions

Advantages

• It’s the first paper that reveals the importance of fusing process
Conclusions

Disadvantages

• Actually the SEB module only gains 0.7% performance increase.
• Most of the improvement are from the methods of other papers.

<table>
<thead>
<tr>
<th>Module</th>
<th>Index compared</th>
<th>Improvement(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS - semantic supervision</td>
<td>1, 2</td>
<td>1.5</td>
</tr>
<tr>
<td>LR - layer rearrangement</td>
<td>2, 3</td>
<td>0.8</td>
</tr>
<tr>
<td>ECRE - explicit channel resolution embedding</td>
<td>3, 4</td>
<td>0.5</td>
</tr>
<tr>
<td>ECRE</td>
<td>6, 7</td>
<td>0.4</td>
</tr>
<tr>
<td>SEB - semantic embedding branch</td>
<td>3, 5</td>
<td>0.7</td>
</tr>
<tr>
<td>DAP - densely adjacent prediction</td>
<td>5, 6</td>
<td>0.6</td>
</tr>
</tbody>
</table>
Conclusions

Disadvantages

• Only up-sampled the high level features, but the low level features are not down-sampled to correct the semantic mask boundary.
THANK YOU!