Efficient Dense Modules of Asymmetric Convolution for Real-Time Semantic Segmentation

Shao-Yuan Lo, Hsueh-Ming Hang, Sheng-Wei Chan, Jing-Jhih Lin
National Chiao Tung University

MMAsia 2019 Oral
December 17, 2019
Introduction

Input: RGB image

Output: Semantic segmentation

CNN
Introduction

Self-driving applications

Efficiency

Accuracy
as high as possible

Inference Time
real-time

Model Size
low memory consumption
EDANet
EDANet

- Efficient Dense modules of Asymmetric convolution, the second proposed network.
EDA Module

• Point-wise convolution layer

• Dilated convolution

• Asymmetric convolution

• Dense connectivity
Asymmetric Convolution

• Factorize a standard 2D convolution kernel into two 1D convolution kernels.

• \( n \times n \rightarrow (n \times 1) \& (1 \times n) \)

\[
\sum_{i=-M}^{M} \sum_{j=-N}^{N} W(i,j)I(x - i, y - j) = \sum_{i=-M}^{M} W_x(i) \left[ \sum_{j=-N}^{N} W_y(j)I(x - i, y - j) \right]
\]
Dense Connectivity

• Each module concatenates its input and new learned features together to form final output. [Gao et al.]

• $y_m = [H_m(y_{m-1}), y_{m-1}]$

• Gather multi-scale information together.
Downsampling Block

- Two modes.
- \( \text{Win} > \text{Wout} \) \( (W_{\text{conv}} = \text{Wout}) \)
- \( \text{Win} < \text{Wout} \) \( (W_{\text{conv}} = \text{Wout} - \text{Win}) \)
Downsampling Block

- Pros: Enable network to have larger receptive fields.
- Cons: Lose spatial information.
Decoder

• No decoder.

• Use bilinear interpolation to recover resolution.
Ablation Study

• Core module

• Extra context module

• Decoder

• Downsampling Block
Cityscapes Dataset

• Class: 19

• Training data: 2975
• Validation data: 500
• Test data: 1525

• Resolution: 1024 x 2048

[Cordts et al.]
Core Module

EDA module

Non-asym variant

Non-dense variant
## Core Module

| Network             | mIoU (%) | # Param. | # Multi-Adds |
|---------------------|----------|----------|--------------|
| EDANet              | 65.10    | 0.68M    | 8.97B        |
| EDA-non-asym        | 65.11    | 0.81M    | 11.41B       |
| EDA-non-dense       | 63.92    | 0.73M    | 8.87B        |
Extra Context Module

EDA-shallow

EDA-ASPP

[Chen et al.]
## Extra Context Module

| Network        | mIoU (%) | # Param. | # Multi-Adds |
|----------------|----------|----------|--------------|
| **EDANet**     | 65.10    | 0.68M    | 8.97B        |
| EDA-shallow    | 58.09    | 0.55M    | 7.77B        |
| EDA-ASPP       | 60.64    | 3.41M    | 41.42B       |
Decoder

EDA-ERFdec

(ERFNet decoder)

[Romera et al.]
## Decoder

| Network         | mIoU (%) | # Param. | # Multi-Adds |
|-----------------|----------|----------|--------------|
| EDANet          | 65.10    | 0.68M    | 8.97B        |
| EDA-ERFdec      | 65.56    | 0.78M    | 12.95B       |
Downsampling Block

EDA-DenseDown [Gao et al.]

| Layers            | Output Size | DenseNet-121 | DenseNet-169 | DenseNet-201 |
|-------------------|-------------|--------------|--------------|--------------|
| Convolution       | 112 x 112   |              |              |              |
| Pooling           | 56 x 56     |              |              |              |
| Dense Block (1)   | 56 x 56     | 1 x 1 conv   | 1 x 1 conv   | 1 x 1 conv   |
|                   |             | x 6          | x 6          | x 6          |
| Transition Layer  | 56 x 56     | 1 x 1 conv   |              |              |
| (1)               | 28 x 28     |              |              |              |
|                   |             |              | 2 x 2 average pool, stride 2 |
| Dense Block (2)   | 28 x 28     | 1 x 1 conv   | 1 x 1 conv   | 1 x 1 conv   |
|                   |             | x 12         | x 12         | x 12         |
| Transition Layer  | 28 x 28     | 1 x 1 conv   |              |              |
| (2)               | 14 x 14     |              |              |              |
|                   |             |              | 2 x 2 average pool, stride 2 |
| Dense Block (3)   | 14 x 14     | 1 x 1 conv   | 1 x 1 conv   | 1 x 1 conv   |
|                   |             | x 24         | x 32         | x 48         |
| Transition Layer  | 14 x 14     | 1 x 1 conv   |              |              |
| (3)               | 7 x 7       |              |              |              |
|                   |             |              | 2 x 2 average pool, stride 2 |
| Dense Block (4)   | 7 x 7       | 1 x 1 conv   | 1 x 1 conv   | 1 x 1 conv   |
|                   |             | x 16         | x 32         | x 32         |
| Classification    | 1 x 1       | 7 x 7 global average pool | 1000D fully-connected, softmax |
| Layer             |             |              |              |              |
## Downsampling Block

| Network                | mIoU (%) | # Param. | # Multi-Adds |
|------------------------|----------|----------|--------------|
| EDANet                 | 65.10    | 0.68M    | 8.97B        |
| EDA-DenseDown          | 61.63    | 0.42M    | 8.51B        |
## Comparison on Cityscapes

| Method                  | Pretrained  | mIoU (%) | Speed (FPS) | # Param. |
|-------------------------|-------------|----------|-------------|----------|
|                         |             |          | Titan X     | Other GPUs |         |
|                         |             |          |             |           |         |
| **SegNet [1]**          | ImageNet    | 56.1     | 16.7        | -         | 29.5M   |
| **ENet [40]**           | No          | 58.3     | 76.9        | -         | 0.36M   |
| **SQ [56]**             | ImageNet    | 59.8     | 16.7        | -         | -       |
| **ESPNet [37]**         | No          | 60.3     | -           | 112.9+++  | 0.36M   |
| **SkipNet-MobileNet [51]** | ImageNet  | 61.5     | 45.0        | -         | -       |
| **ContextNet [42]**     | No          | 66.1     | 18.3        | -         | 0.85M   |
| **ERFNet [45]**         | No          | 68.0     | 41.7        | -         | 2.1M    |
| **BiSeNet [58]**        | ImageNet    | 68.4     | -           | 105.8++   | 5.8M    |
| **ICNet [61]**          | ImageNet    | 69.5     | 30.3        | -         | -       |
| **EDANet (ours)**       | No          | 67.3     | 81.3        | 108.7†    | 0.68M   |
## Comparison on CamVid

| Method               | mIoU (%) | Class acc. (%) | Global acc. (%) | # Param. |
|----------------------|----------|----------------|-----------------|----------|
| ENet [40]            | 51.3     | 68.3           | -               | 0.36M    |
| ESPNet [37]          | 55.6     | 68.3           | -               | 0.36M    |
| SegNet [1]           | 55.6     | 65.2           | 88.5            | 29.5M    |
| FCN-8s [36]          | 57.0     | -              | 88.0            | 134.5M   |
| FC-DenseNet56 [27]   | 58.9     | -              | 88.9            | 1.5M     |
| DeepLab-LFOV [6]     | 61.6     | -              | -               | 37.3M    |
| Dilation8 [59]       | 65.3     | -              | 79.0            | 140.8M   |
| BiSeNet [58]         | 65.6     | -              | -               | 5.8M     |
| ICNet [61]           | 67.1     | -              | -               | -        |
| EDANet (ours)        | 66.4     | 76.7           | 90.8            | 0.68M    |
Video Demo

(a) RGB  (b) Ground truth  (c) EDANet

Input

Ground truth

EDANet
Conclusion

• We develop a novel network named EDANet, which incorporates asymmetric convolution with dilated convolution and dense connectivity. It can run on high-resolution images at 108 FPS on a single GPU and achieve 67.3% mIoU on the Cityscapes dataset.

• EDANet is nearly 3 times faster than ICNet and attains comparable performance; it achieves this without any extra decoder structure, context module, post-processing scheme, and pretrained model.

• We design various types of EDANet variants to analyze the performance of different network architectures and analyze the reasons behind the results.
Thanks for your attention