Fast and Accurate Real-Time Semantic Segmentation with Dilated Asymmetric Convolutions

Leonel Rosas-Arias¹, Gibran Benitez-Garcia², Jose Portillo-Portillo¹, Gabriel Sanchez-Perez¹ and Keiji Yanai²

¹Instituto Politecnico Nacional, Mexico City, Mexico
²The University of Electro-Communications, Tokyo, Japan
Real-time semantic segmentation

Problem:
• **High-accuracy** semantic segmentation is extremely **expensive to compute**.
• Networks for **real-time semantic segmentation** sacrifice a lot of accuracy.

Objective:
• Reduce the accuracy gap between **real-time** and **non-real-time for semantic segmentation**.

Find and classify pixels belonging to each objects in the scene.

Contributions

• **FASSD-Net:**
  – SOTA performance of **speed and accuracy** with high resolution images (1024x2048).
  – **Two additional variations** to balance the speed and accuracy trade-offs.

• **Dilated Asymmetric Pyramidal Fusion module (DAPF):**
  – Obtains feature maps rich in **contextual information**.
  – Requires considerably **fewer floating-point operations** compared with similar method, such as ASPP [1].

• **Multi-resolution Dilated Asymmetric module (MDA):**
  – Offers an improved way to **fuse two set of feature maps** of different resolution.
  – Simultaneously **refines spatial and contextual information** from input feature maps.
  – Can be used in **all decoder stages**.

FASSD-Net architecture

**FASSD-Net:**
- **Encoder:** HarDNet [2] (custom version)
- **Decoder:** DAPF + MDA

Factorized convolutions

With $\alpha=2$, our proposed module requires 50% fewer floating-point operations compared to ASPP.

The number of intermediate feature maps is controlled by the $\alpha$ parameter.
FASSD-Net architecture

Input Image (1024x2048)

- Stem Conv
- Encoder B1
- Encoder B2
- Encoder B3
- Encoder B4

Output Conv
- Decoder B3
- MDA 3
- Decoder B2
- MDA 2
- Decoder B1
- MDA 1
- DAPF

Prediction

Asymmetric branch: Context
- Refines both contextual and detail information

Symmetric branch: Details

Multi-resolution Dilated Asymmetric module (MDA)

© 2020 UEC Tokyo.
Ablation study on the Cityscapes dataset

Experimental setup:
• All networks pretrained with ImageNet.
• Training during 90k iterations.
• Batch size = 16.
• GFLOPs measured for high resolution images at 1024x2048.
• Speed in FPS measured with an Nvidia GTX 1080Ti card.
Quantitative results

1) For fair comparison, the speed of methods marked by (*) are approximated without TensorRT acceleration [3].

2) PSPNet and FC-HarDNet-L2 speeds are placed on the x-axis edges for the sake of better visualization.
Qualitative results

Input Image  Groundtruth  FC-HarDNet-70  FASSD-Net  FASSD-Net-L1  FASSD-Net-L2
• **Conclusions:**
  – We proposed two modules (DAPF & MDA) for reducing the accuracy gap between real-time and non-real-time semantic segmentation networks.
  – With FASSD-Net, we set a new SOTA accuracy for real-time semantic segmentation on the Cityscapes validation set.

• **Future work:**
  – Include more backbone networks and different datasets for evaluation.
  – Evaluate on different scenarios, such as indoor parsing and medical images.
Thank You

[QR Code for Code & Models]