Learning Adaptive Receptive Fields for Deep Image Parsing Network

Zhen Wei[1,4], Yao Sun[1], Jinqiao Wang[2], Hanjiang Lai[3], Si Liu[1]

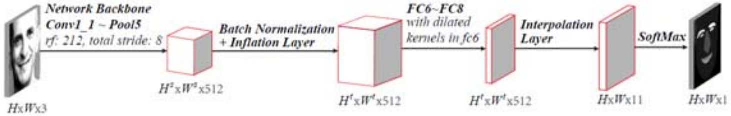
1. IIE, CAS, Beijing, China 2. IA, CAS, Beijing, China 3. SYU, Guangzhou, China 4. UCAS, Beijing, China

IEEE 2017 Conference on Computer Vision and Pattern Recognition

Introduction:

- In this paper, we propose a learning based, data-driven method for regulating receptive field in deep image parsing network automatically..

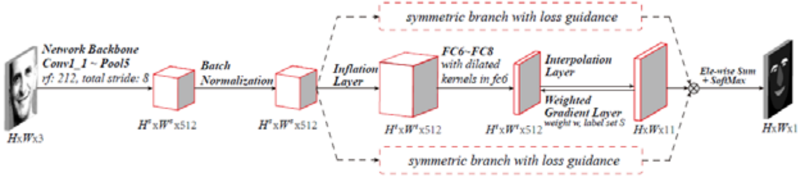
- Framework:
  - Two derivable affine transformation layers are introduced into the network;
  - The new layers are inserted *before* (inflation layer) and *after* (interpolation layer) the convolutional layer whose receptive fields need to be regulated (*e.g. fc6 – fc8 layers*);



  - The two new layers share one parameter f, indicating the *resize factor*;
  - f is derivable and is trained end-to-end with the network.

For multi-path networks:

  - Initialize each path with symmetric structures and dilation rate;
  - Use a *weighted gradient layer* to guide each path to learn discriminative f and thus focus on different semantic labels with different scales.



Method:

- Detailed explanation of the two affine transformation layers:

- Forward (*affine transformation on feature maps*):
  - inflation layer: resize each feature maps with the factor of f;
  - Interpolation layer: resize feature maps back to a fixed size, the factor $f' = {}^{F}/_{f}$, where F is a pre-defined constant.

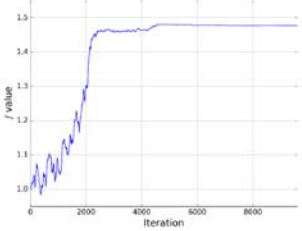
- Backward (*the gradient computation of f*):
  - inflation layer: the gradient w.r.t. f: $G_{inf} = \frac{\partial Loss}{\partial f}$
  - interpolation layer: the gradient w.r.t. f: $G_{inter} = \frac{\partial Loss}{\partial f'}\frac{\partial f'}{\partial f} = \frac{\partial Loss}{\partial f'}(\frac{-F}{f^2})$

- In implementation, the two gradients of f are added together:

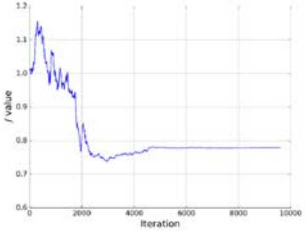
$$\frac{\partial Loss}{\partial f}=G_{inf}+G_{inter}$$

- The fluctuation of f during training on VOC dataset:
  - Best receptive field $rf = 404$ (*by manually grid search*).
  - If networks are initialized with bad receptive fields, the learned f will regulate receptive fields automatically:

$rf = 436$ (*initial*) → 396 (*after training*)



rf =308 (*initial*) → 364 (*after training*)



Experiments:

- Here we present results from the single-path networks trained on VOC dataset (*general image parsing task*) and Helen dataset (*face parsing task*).

- Image parsing results: