Task Generalizable Spatial and Texture Aware Image Downsizing Network

Lin Ma[1], Weiming Li[1], Hongsheng Li[2], Qiang Wang[1] and Ji-Yeon Kim[3]
1. Samsung Research Institute China - Beijing (SRC-B), Beijing, China
2. Chinese University of Hong Kong, Hong Kong, China
3. Samsung Advanced Institute of Technology, Suwon, South Korea
Email: malin_u@126.com

Motivation

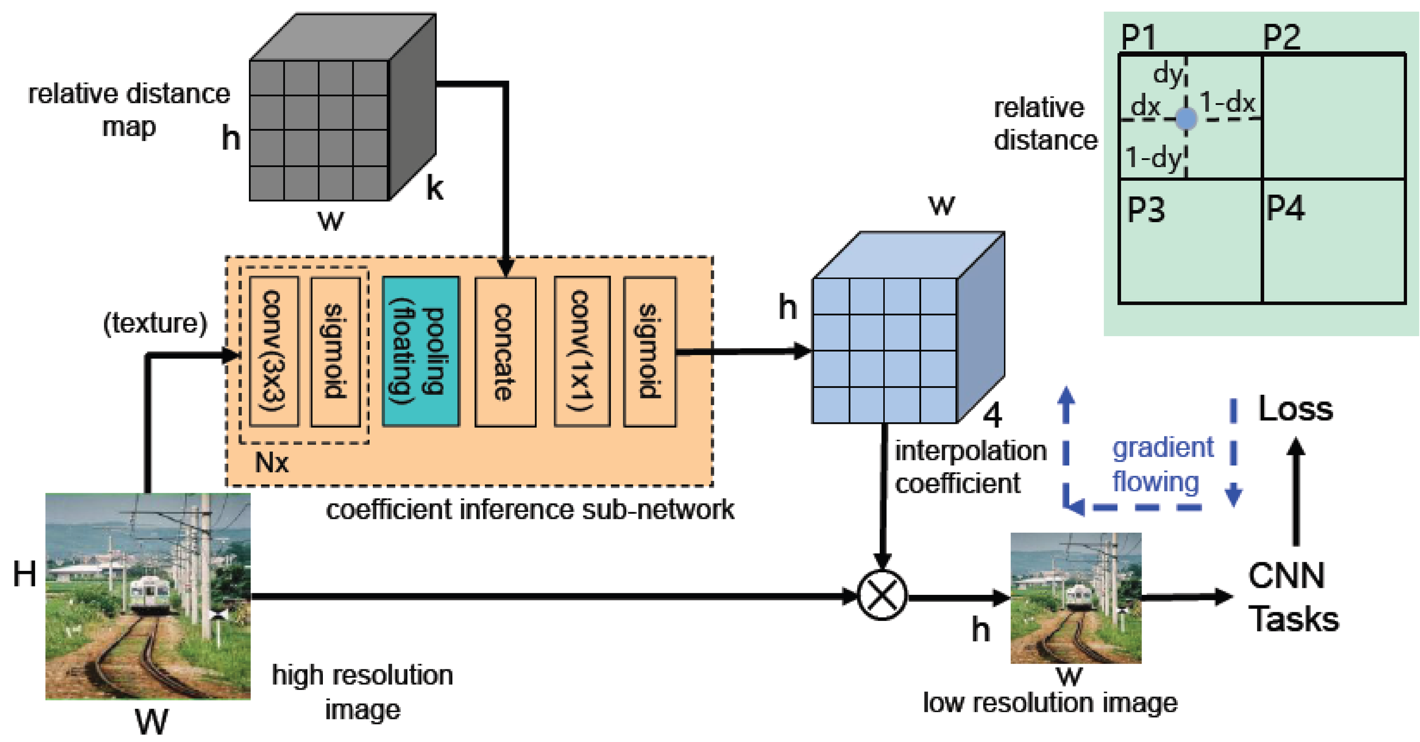
- Widely used bilinear interpolation, cubic interpolation, etc. are handcrafted, and only consider relative distance in defining the interpolation coefficients.
- Handcrafted interpolation losses much information in downsizing image, which reduces the network performance.

Contributions

- We propose a new interpolation method DownsizeNet aiming to preserve image information in image downsizing with a sub-network interpolation module at the image pre-processing stage. The interpolation sub-network can be easily used in various vision tasks, and has good generalization ability to different pipelines by making the downsizing result retain a real image.
- We introduce texture feature into CNN based interpolation coefficient inference. A special floating type pooling layer is designed to spatially align the CNN texture feature and the encoded relative distance map, and this facilitates the pixel-wise interpolation coefficient inference.
- Experiments on seven pipelines in detection and segmentation tasks demonstrate that our method consistently reduces accuracy drop than widely used bilinear interpolation. Besides, we have demonstrated that our method also outperforms texture only based interpolation, cubic interpolation and area interpolation.

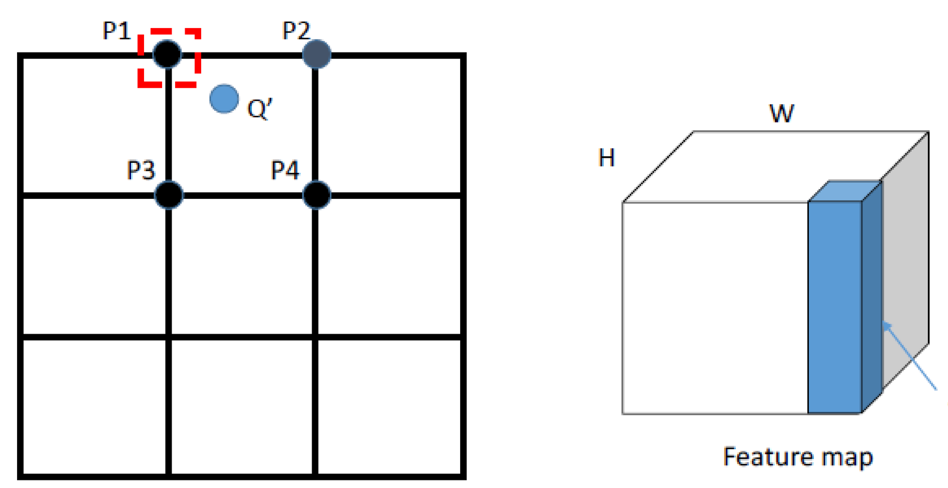
Network architecture

- Given the high resolution image, and the high resolution HxW and low resolution value hxw, output a low resolution image.
- The interpolation network is simple, and contains only 2~3 3*3 convolution layer and 1 1*1 convolution layer. Besides, two efficient special layers are proposed. This makes the network efficient.



Two special layers

- Floating pooling layer. Given a projected point Q' on high resolution image, the feature vector of P1 is selected as the feature of Q'.



- Interpolation layer. The last sigmoid can be replaced by softmax, then the interpolation operation is more convenient.

$$\tilde{q} = [q_1, \ldots, q_4][a_1, \ldots, a_4]^T.$$
$$s.t. \quad \sum_i a_i = 1, \quad 0 \leq a_i \leq 1.$$

Experimental results

Table 1: **The performance (mAP) comparison between Bilinear and ours on four detection pipelines.** RefineDet is implemented with Caffe, and other networks are implemented with PyTorch. ~ 300 denotes that one side of the image is resized to 300 while the height and width ratio is fixed.

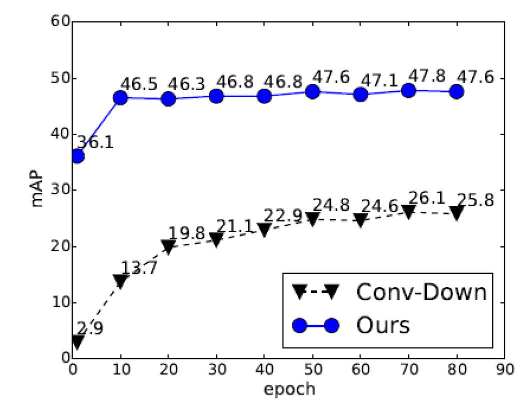
Pipeline	Attribute	Dataset	Backbone	Resolution	Bilinear	Ours
RefineDet [48]	One-stage	VOC2007	VGG16	$300\times300->160\times160$	68.0	**68.9**
CenterNet [51]	Anchor-Free	VOC2007	ResDCN18	$300\times300->160\times160$	41.0	**41.9**
DETR [3]	Transformer	COCO	ResNet101	$300\times300->165\times165$	16.1	**16.3**
LightHead [22]	Two-stage	VOC2007	VGG16	$\sim300->\sim165$	50.0	**50.5**

Table 2: **The performance (mean IoU) comparison between Bilinear and ours on three segmentation pipelines.** FCN is implemented with Caffe, and other networks are implemented with PyTorch. SPP: Spatial Pyramid Pooling. For DeepLabv3+, the image is resized to 0.35 times of the original resolution. For SPNet, the image is resized to 0.3 times of the high resolution. The resolutions 513 and 768 are used in the original codes of DeepLabv3+ and SPNet respectively.

Pipeline	Attribute	Dataset	Backbone	Resolution	Bilinear	Ours
FCN [29]	Fully convolution	VOC2012Seg	VGG16	$Random->100\times100$	25.8	**26.3**
DeepLabv3+ [4]	SPP, encode-decode	Cityscapes	Resnet101	$513\times513->179\times179$	46.4	**47.3**
SPNet [15]	Strip pooling	Cityscapes	Resnet50	$768\times768->230\times230$	58.6	**59.2**

Table 3: **Test on LightHead about the generalization ability of the pretrained DownsizeNet model** on Pascal VOC2007 (mAP metric). The interpolation sub-network is pre-trained on DeepLabv3+ for 100, 200 and 300 epochs separately. The LightHead is re-trained with 10 epochs. The image is resized to 0.3 times of the high image resolution for both LightHead and DeepLabv3+.

Training epochs on DeepLabv3+	Bilinear	Ours
100	34.7	**34.8**
200	34.7	**35.1**
300	34.7	**35.2**



Conv-Down uses traditional stride 2 conv to downsize the image. The image is resized from 300x300 to 150x150.

Figure 5. **Test on LightHead about the influences of pretrained downsizing model on Pascal VOC2007.** The downsizing modules are pretrained on DeepLabv3+ with 200 epochs. Results at epoch 1 and $10 \times k, k = 1, 2, ..., 8$ are given.

Table 9. **Time cost (mili-second) for each frame for *Bilinear* and ours** using RefineDet as baseline architecture on Pascal VOC2007. The time cost is tested on 1 K80 GPU.

convolution layer number	Bilinear	Ours
2	31.4	34.2
3	31.4	35.2
4	31.4	37.6

Discussion

- DownsizeNet can also be used in feature map resizing (downsizing and upscaling).

BMVC 2022