15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Dilated Convolutions for Image Classification and Object Localization

Yasunori Kudo
Keio University, Japan
3-14-1 Hiyoshi, Kohoku-ku, Yokohama-shi,
Kanagawa 223-8522 Japan
yakudo@aoki-medialab.org

Yoshimitsu Aoki
Keio University, Japan
3-14-1 Hiyoshi, Kohoku-ku, Yokohama-shi,
Kanagawa 223-8522 Japan
aoki@elec.keio.ac.jp

Abstract

Yu et al.[1] showed that dilated convolutions are very effective in dense prediction problems such as semantic segmentation. In this work, we propose a new ResNet[2] based convolutional neural network model using dilated convolutions and show that this model can achieve lower error rate for image classification than ResNet with reduction of the number of the parameters of the network by 94% and that this model has high ability to localize objects despite being trained on image-level labels. We evaluated this model on ImageNet[5] which has 50 class labels randomly selected from 1000 class labels.

1 Introduction

Recent works showed that dilated convolutions can give good performance in semantic segmentation[1], sound generation[3] and machine translation[4]. Traditional neural networks apply pooling or convolution with 2 or more stride to decrease the feature map resolution and expand the receptive field. Dilated convolution supports exponential expansion of the receptive field without loss of feature map resolution since it applies convolution with a dilation factor instead of convolution after decreasing of the feature map resolution.

Zhou et al.[6] has shown that convolutional neural networks(CNNs) behave as object detector despite being trained on image-level labels. They use feature maps of the last convolutional layers of CNNs and localize objects. The ability to localize objects depends on the resolution of the feature maps, so they remove the last few convolution and pooling layers, then add few convolution layers to increase the feature map resolution.

In order not to decrease the feature map resolution, we replaced some 3×3 convolution layers in ResNet with 3×3 dilated convolution layers with a dilation factor and found out that ResNet with dilated convolution can reduce ImageNet classification error and localization error with reduction of the number of the parameters of the network by 94%.

2 Related Work

Recent works showed that CNNs have the ability to localize objects despite being trained on image-level labels. Oquab et al.[8] optimized the input image of CNNs to increase the predicted score of its image label and showed that the absolute value of the gradient of the input image can localize objects. Zhou et al.[6] showed that the output feature maps of the last convolutional layer can directly localize objects. Selvaraju et al.[9] combined these gradient based method and feature maps based method to localize objects.

In order to localize objects, feature maps based methods depend on the resolution of the feature maps. Long et al.[10] used backwards convolution to increase the resolution. Backward convolutions are widely used for dense prediction problems such as semantic segmentations. Yu et al.[1] proposed dilated convolutions which enable to expand the receptive field without loss of feature map resolution.

layer name	ResNet	ResNet with dilated conv
conv1	7×7,64, stride 2 output size 112×112	
conv2_x	3×3 max pooling, stride 2 $\begin{bmatrix} 1\times1,64 \\ 3\times3,64 \\ 1\times1,256 \end{bmatrix}\times3$ output size 56×56	$\begin{bmatrix} 1\times1,64 \\ 3\times3,64 \\ 1\times1,256 \end{bmatrix}\times3$ output size 56×56
conv3_x	$\begin{bmatrix} 1\times1,128 \\ 3\times3,128 \\ 1\times1,512 \end{bmatrix}\times4$ output size 28×28	$\begin{bmatrix} 1\times1,64 \\ 3\times3,64, \text{dilate } 2 \\ 1\times1,256 \end{bmatrix}\times4$ output size 56×56
conv4_x	$\begin{bmatrix} 1\times1,256 \\ 3\times3,256 \\ 1\times1,1024 \end{bmatrix}\times6$ output size 14×14	$\begin{bmatrix} 1\times1,64 \\ 3\times3,64, \text{dilate } 4 \\ 1\times1,256 \end{bmatrix}\times6$ output size 56×56
conv5_x	$\begin{bmatrix} 1\times1,512 \\ 3\times3,512 \\ 1\times1,2048 \end{bmatrix}\times3$ output size 7×7	$\begin{bmatrix} 1\times1,64 \\ 3\times3,64, \text{dilate } 8 \\ 1\times1,256 \end{bmatrix}\times3$ output size 56×56
fc	average pool, 50-d fc, softmax	
# params	23.7M	1.4M

Figure 1. Differences in the architecture of ResNet(left) and ResNet with dilated convolutions(right). Both networks have 50 layers.

3 Method

3.1 ResNet with dilated convolutions

In this section, we describe the details of ResNet with dilated convolutions. Fig.1 show the details of architecture of ResNet and ResNet with dilated convolutions. In the case of ResNet, downsampling is performed by the first 1×1 convolution layers with stride 2 in each layer blocks, conv3_x, conv4_x and conv5_x.

Instead of downsampling by convolutions with stride 2 to expand receptive field, we set stride to 1 and replace 3×3 convolutions with 3×3 dilated convolutions. We set dilation factors to 2^{N-1} in a convN_x layer block. In order to equalize the calculation amount of ResNet and ResNet with dilated convolutions, we reduce the number of output channels by dividing by dilation factors. The total amount of parameters of the networks is 23.7M for ResNet and 1.4M for ResNet with dilated convolutions.

3.2 Weakly supervised object localization

Zhou et al. used the output feature maps of the last convolution layer of the networks to localize objects. For a given image, let $f_k(x, y)$ represent the activation of unit k in the last convolutional layer at spatial location (x, y). Then, for unit k, the result of performing global average pooling, F^k is $\Sigma_{x,y} f_k(x, y)$. Thus, for a given class c, the input to the softmax, S_c is $\Sigma_k w_k^c F_k$ where w_k^c is the weight corresponding to class c for unit k. Essentially, w_k^c indicates the importance of F_k for class c. Define M_c as a class activation map(CAM) for class c, M_c for each spatial element can be describe as

$$M_c(x, y) = \sum_k w_k^c f_k(x, y). \quad (1)$$

We use the CAM to localize objects and generate bounding boxes. In order to confirm whether the performance of CAM depends on the resolution of the output feature maps of the last convolutional layers, we use 3 CNNs: (1) ResNet, (2) ResNet with dilated convolutions, (3) ResNet for CAM. (1) and (2) are described in Fig.1. About (3), inspired by Zhou et al., we simply set stride of the first convolutional layer of conv5_x block of ResNet to 1 from 2 so that downsampling is not performed in conv5_x block. The resolution of the output feature maps of the last convolutional layers of this network is 14×14 and we call this network ResNet for CAM.

3.3 Implementation

Our implementation follow the practice in [2]. The image is resized with its shorter side randomly sampled in [256, 480] for scale augmentation. A 224×224 crop is randomly sampled from an image or its horizontal flip, with the per-pixel mean subtracted. We initialize the weights as in [7] and train all nets from scratch. We use SGD with a mini-batch size of 256. The learning rate starts from 0.1 and is divided by 10 when the epochs is 100 or 150 and the models are trained for up to 200 epochs. We use a weight decay of 0.0001 and a momentum of 0.9 without accelerated gradient.

4 Experiments

4.1 ImageNet classification

We experimented with above three CNNs for image classification and localization. We evaluated these CNNs with ImageNet50, which has randomly selected 50 class labels from ImageNet. Fig.2 shows the loss curves and top-1 error curves and Table 1 shows top-1 validation error scores. While ResNet for CAM increased classification error rate by 0.32% compared with ResNet since we simply changed the stride value to increase the feature maps resolution, ResNet with dilated convolutions reduced classification error rate by 3.52% despite the reduction of the parameters of the network by 94%.

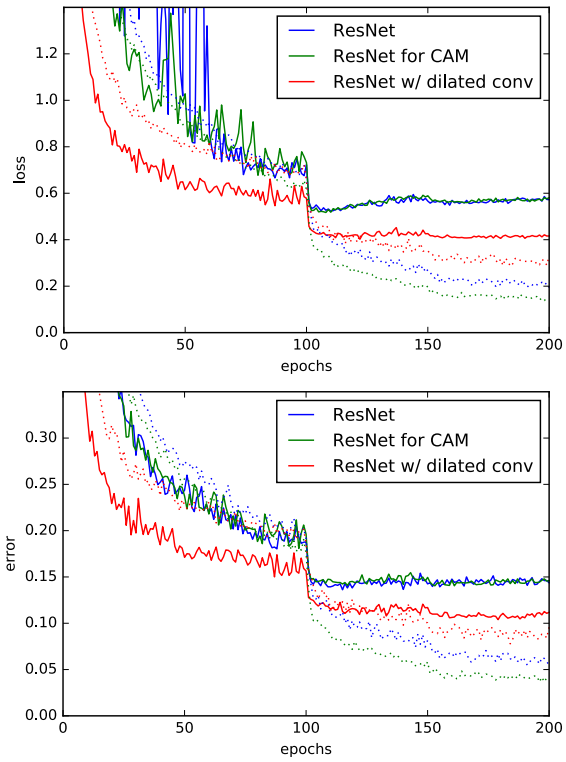

Figure 2. Loss(top) and top-1 error(bottom) curves of 3 CNNs in ImageNet50 classification. Training curves is drawn with dotted lines and validation curves solid lines.

Table 1. Top-1 classification error on ImageNet50 validation using 200 epoch parameters of CNNs.

	error(%)
ResNet	14.44
ResNet for CAM	14.76
ResNet w/ dilated conv	11.24

4.2 ImageNet localization

We also evaluated these CNNs on ImageNet50 localization. Each image in ImageNet has one label $C_k (k = 1, ..., 50)$ and bounding boxes $B_{km} (m = 1, ..., M_k)$ of objects as ground truth, where M_k is the number of instances of the k^{th} object in the current image. Given an image our CNNs predict a image label c and a object bounding box b. Let $d(c, C_k) = 0$ if $c = C_k$ and 1 otherwise. Let $f(b, B_{km}) = 0$ if b and B_{km} have more than x overlap, and 1 otherwise.

Localization top-1 error for overlap x is computed using below metrics:

$$e(x) = \frac{1}{50} \sum_k \min_m \max \left(d(c, C_k), f(b, B_{km}) \right). \quad (2)$$

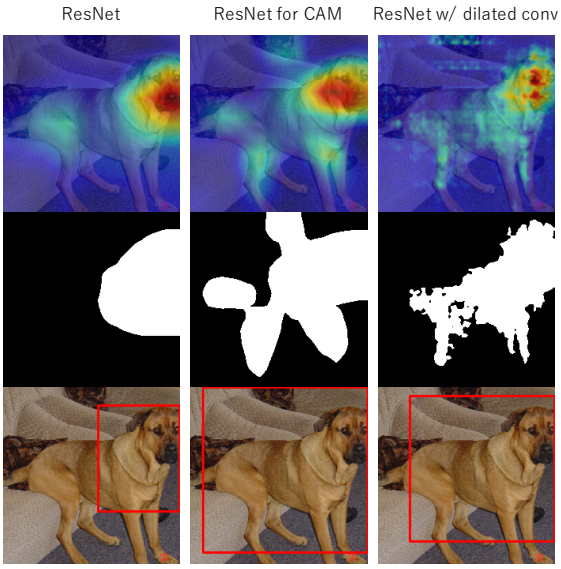

Figure 3. Visualization of the CAM(top) and segmentation masks(middle) and generated bounding boxes(bottom).

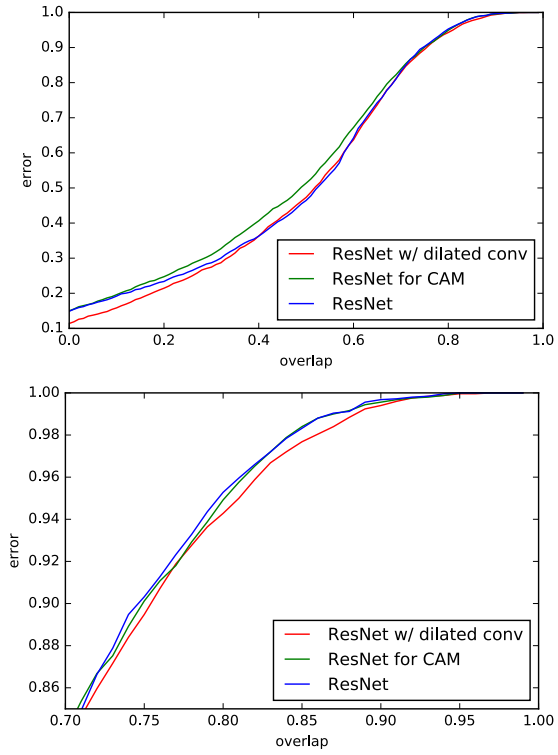

Figure 4. Localization error curves of the CNNs for overlap x. The top figure shows localization error curves in all ranges of overlap x and the bottom figure in high ranges. ResNet with dilated convolutions achieve low error rate in high ranges of overlap.

Following the practice in [6], in order to generate a bounding box, we use the CAM. Fig.3 shows the procedure of how to generate a bounding box. We first segment the regions of which the value is above 20% of

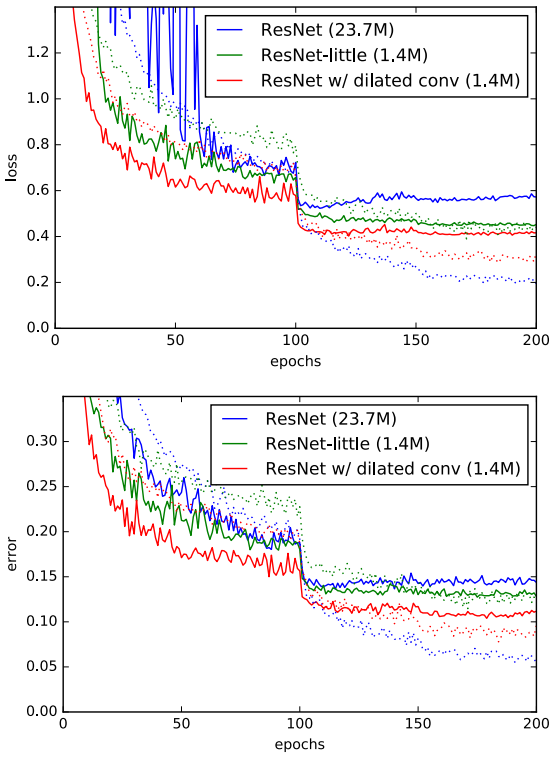

Figure 5. Loss(top) and top-1 error(bottom) curves of 3 CNNs in ImageNet50 classification. Training curves is drawn with dotted lines and validation curves solid lines.

the max value of the CAM. Then we take the bounding box that covers the largest connected component in the segmentation map.

Fig.4 shows the localization error for overlap x. We found out that in the condition of high threshold of overlap ResNet with dilated convolutions reduced localization error rate compared with ResNet and ResNet for CAM since ResNet with dilated convolutions can generate the high resolution CAM.

5 Discussion

5.1 Number of parameters of CNNs

We showed that ResNet with dilated convolutions improved the accuracy of the image classification task even if decreasing the number of parameters by 94 %. The dataset we used to evaluate our methods was ImageNet50, which has 50 class labels randomly selected from 1000 class labels, and the number of images ImageNet50 has is one twentieth of the original ImageNet has. You may assume that the original ResNet is too expressive to the small datasets like ImageNet50 and the reduction of parameters works as regularization, avoids over-fitting and improves the accuracy to test data. Therefore we trained ResNet with narrow channels (ResNet-little), which has the same amount of parameters as ResNet with dilated convolutions and compared the loss and the error with original ResNet and ResNet with dilated convolutions (Fig.5). Fig.5 reveals that one reason why ResNet with dilated convolution

succeeded is the regularization by reduction of parameters of CNN and the other is the dilated convolution itself.

6 Conclusion

In this work, We proposed new CNNs architecture, ResNet with dilated convolutions and showed that it can reduce the classification and localization error rate with reduction of the number of the parameters of the network.

References

[1] F. Yu, et al.: "Multi-Scale Context Aggregation by Dilated Convolutions", International Conference on Learning Representations(ICLR), 2016.

[2] K. He, et al.: "Deep Residual Learning for Image Recognition", The Conference on Computer Vision and Pattern Recognition(CVPR), 2016.

[3] Aaron van den Oord, et al.: "WaveNet: A Generative Model for Raw Audio", arXiv:1609.03499v2, 2016.

[4] N. Kalchbrenner, et al.: "Neural Machine Translation in Linear Time", arXiv:1610.10099v1, 2016.

[5] O. Russakovsky, et al.: "ImageNet Large Scale Visual Recognition Challenge" Journal of Computer Vision, 2015.

[6] B. Zhou, et al.: "Learning Deep Features for Discriminative Localization" The Conference on Computer Vision and Pattern Recognition(CVPR), 2016.

[7] K. He, et al.: "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification" The International Conference of Computer Vision(ICCV), 2015.

[8] M. Oquab, et al.: "Learning and transferring mid-level image representations using convolutional neural networks" The Conference on Computer Vision and Pattern Recognition(CVPR), 2014.

[9] R. Selvaraju, et al.: "Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization", arXiv:1610.02391v1, 2016.

[10] J. Long, et al.: "Fully Convolutional Networks for Semantic Segmentation", The Conference on Computer Vision and Pattern Recognition(CVPR), 2015.