Context-Gated Convolution – Supplementary Material

Xudong Lin[1⋆], Lin Ma[2], Wei Liu[2], and Shih-Fu Chang[1]

[1] Columbia University
{xudong.lin, shih.fu.chang}@columbia.edu
[2] Tencent AI Lab
forest.linma@gmail.com wl2223@columbia.edu

A Appendix

A.1 Analysis of the Gate

To further validate that our CGC uses context information of the target objects to guide the convolution process, we calculate the average modulated kernel (in the last CGC of the model) for images of each class in the validation set. Then we calculate inter-class $L-2$ distance between every two average modulated kernels, i.e., class centers, and the intra-class $L-2$ distance (mean distance to the class center) for each class. As is shown in Fig. 1, we visualize the difference matrix between inter-class distances and intra-class distances. In more than 93.99% of the cases, the inter-class distance is larger than the corresponding intra-class distance, which indicates that there are clear clusters of these modulated kernels and the clusters are aligned very well with the classes.

This observation strongly supports that our CGC successfully extracts class-specific context information and effectively modulates the convolution kernel to extract representative features. Meanwhile, the intra-class variance of the modulated kernels indicates that our CGC dynamically modulates convolution kernels according to different input contexts.

A.2 Details of training settings on ImageNet and CIFAR-10

For the default setting on ImageNet, we use 224×224 random resized cropping and random horizontal flipping for data augmentation. Then we standardize the data with mean and variance per channel. We use a traditional cross-entropy loss to train all the networks with a batch size of 256 on 8 GPUs by SGD, a weight decay of 0.0001, and a momentum of 0.9 for 100 epochs. We start from a learning rate of 0.1 and decrease it by a factor of 10 every 30 epochs. The last normalization layers in the module are zero-initialized to make the gates start from constants. All the extra layers in Context-Gated Convolution have a learning rate ten times smaller than convolutional kernels.

⋆ This work was done when Xudong Lin interned at Tencent AI Lab.

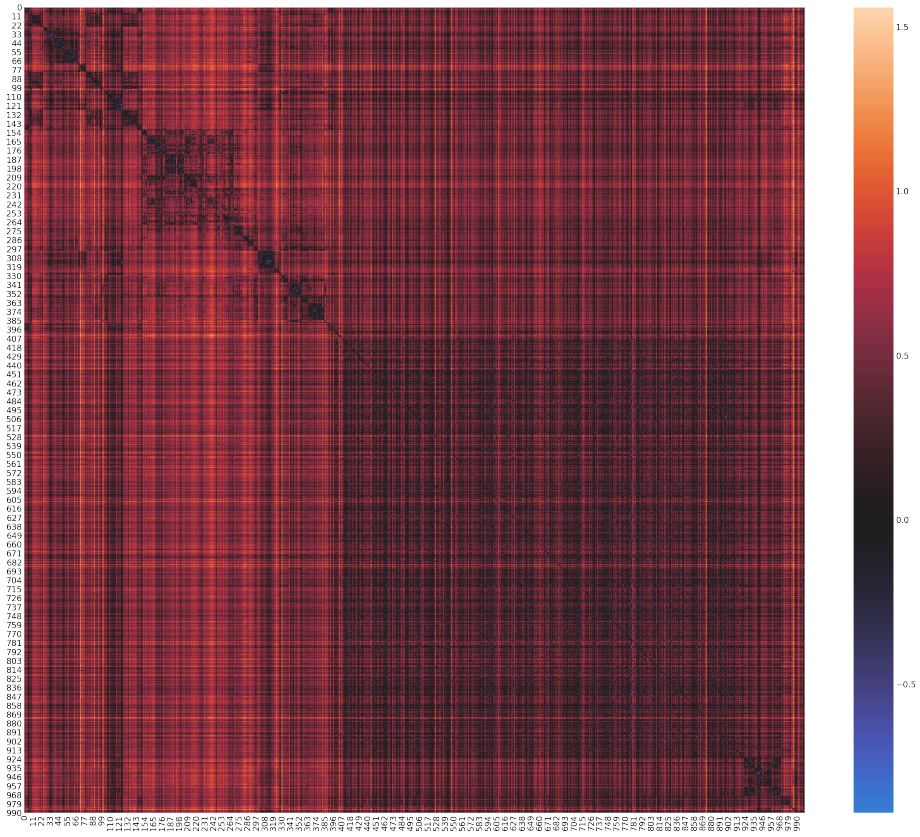

Fig. 1: Visualization of the difference matrix between inter-class distances and intra-class distances of the last gate in the network on the ImageNet validation set. (Best viewed on a monitor when zoomed in)

For the advanced setting, we also use mixup [4] for data augmentation, and we follow [1] to use learning rate warm-up in the first 5 epochs of training. We train the networks with the cosine learning rate schedule [1] for 120 epochs. The other hyper-parameters are set to be same with the default setting.

For CIFAR-10, we use 32×32 random cropping with a padding of 4 and random horizontal flipping. We use a batch size of 128 and train on 1 GPU. We decrease the learning rate at the 81st and 122nd epochs, and halt training after 164 epochs. For the ablation study, the result is an average of 3 runs.

A.3 Details about P3D-A

Based on ResNet-50, we add a temporal convolution with $k = 5, stride = 2$ after the first convolutional layer. For convolutional layers in residual blocks, we

follow [3] to add $3 \times 1 \times 1$ convolution (stride is 1) after every two $1 \times 3 \times 3$ convolutions. The added temporal convolutional layers are initialized to imitate the behavior of TSM [2] to ease the training process. We only inflate the max pooling layer after the first convolutional layer with a temporal kernel size of 3 and a stride of 2 without adding any other temporal pooling layers. Note that all the aforementioned convolutional layers come with a Batch Normalization layer and a ReLU activation function.

References

1. He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., Li, M.: Bag of tricks for image classification with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 558–567 (2019)
2. Lin, J., Gan, C., Han, S.: Temporal shift module for efficient video understanding. arXiv preprint arXiv:1811.08383 (2018)
3. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. arXiv preprint arXiv:1711.07971 **10** (2017)
4. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)