Different Descriptors for Squeeze and Excitation Attention Block

Umar Masud

Abstract

External attention modules such as the Squeeze and Excitation (SENet) block, Efficient Channel Attention (ECA) block, Convolutional Block Attention Module (CBAM), etc. have been recently proposed for various image related tasks. They have particularly exploited the channel wise information to improve overall accuracy of the network. In this work, different descriptors such as standard deviation, trace, largest singular value and DC term of discrete cosine transform (DCT) are used for extracting channel information instead of the usual global average. The experiments were tested on ResNet models of depth 20 and 32 with SENet using the CIFAR-10 dataset. The results showed that there is a slight increase in the accuracy of classification when standard deviation and largest singular value are used as channel descriptors with minimal increase in the training time.

1 Introduction

Continuous efforts have been made to extract more useful representations in convolutional neural networks (CNN) to better capture the hierarchical patterns and information for visual tasks. Specific architectural units have been proposed by integrating learning mechanisms into the network that help capture spatial correlations between features. In our context of computer vision, incorporating these spatial dependencies is what has been been termed as bringing *attention* in the network.

Developing on this idea, various attention modules have been proposed that are used as sub-networks to enhance the feature representation in the convolutional network pipeline. Squeeze and Excitation Network (SENet) by Hu et al. [1], Efficient Channel Attention (ECA) block by Wang et al. [2] Convolutional Block Attention Module (CBAM) by Woo et al. [3], etc. are some the works that focus on improving the representations learnt by the CNN architecures. SE block and ECA block have utilised channel wise information while CBAM has used both spatial as well as channel wise information.

In this work, we investigate the design of these attention modules itself and experiment with the methods of extracting those spatial or channel-wise information. Usually, for channel wise description, global average pooling or global max pooling is used as a representative value for a particular channel and then these values are further used in dense networks or 1D convolutions for further processing. However, we had the intuition that if we use a better representative value for channel description, a better summary statistic than just the average or maximum, we might be able to further improve the performance of the network. Instead of the global average or maximum, we tried using the standard deviation, trace, largest singular value and DC term of discrete cosine transform (DCT) to get a summary statistic from the channel. Our results showed largest singular value and standard deviation can slightly improve the performance, with minimal cost of training time or extra space.

2 Related Work

Attention mechanisms have been one of the important research interests in recent years. In computer vision, attention modules are being designed to assist deep neural networks to suppress less salient pixels or channels. These methods successfully exploit the mutual information from different dimensions of features. Squeeze-and-Excitation Networks (SENet) (Hu et al. [1]) integrate the spatial information into channel-wise feature responses and compute the corresponding attention with two dense feed-forward layers. Efficient Channel Attention (ECA) block (Wang et al. [2]) uses 1-D convolutional on the channel-wise features instead of using dense feed-forward network. Bottleneck Attention Module (BAM) (Park et al. [4]) builds separated spatial and channel sub-modules in parallel. Convolutional Block Attention Module (CBAM) (Woo et al. [3]) provides a solution that embeds the channel and spatial attention sub-modules sequentially. Global Attention Mechanism (GAM) (Liu et al. [5]) also works on a similar approach. Considering the significance of the cross-dimension interactions, Triplet Attention Module (TAM) (Misra et al. [6]) takes account of dimension correlations by rotating the feature maps. Normalization-based Attention Module (NAM) (Liu et al. [7]) use a scaling factor of batch normalization which uses the standard deviation to represent the importance of weights.

These methods helps in refining perceived information while retaining the context of it. They have efficiently incorporated this attention mechanism in deep convolution neural network (CNN) architectures to improve

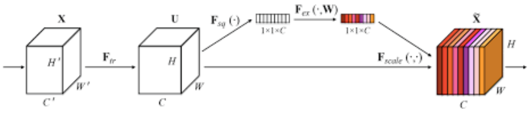

Figure 1: Squeeze and Excite Block (taken from [1])

performance on large-scale vision tasks.

3 Experiment and Results

3.1 Hpyothesis

In our experiments we have used different methods to extract a summary statistic for the channel-wise feature. The choice of these methods were based on properties of matrix decomposition and image processing (treating each channel as an image). Standard deviation relates to mutual information of the values in the channel matrix. The trace of a matrix is the sum of all its eigenvalues. Eigenvalues can play an important role for extracting features of an image, from a viewpoint of matrix decomposition. Largest singular value is in fact the largest eigen value which we obtain after Singular Value Decomposition (SVD) of a matrix. It is known that only the few large eigenvalues actually contain all the necessary information of an image matrix. Discrete Cosine Transform (DCT) converts an image's spatial information into its frequency or spectral components. The DC coefficient of a DCT is the largest value of the transformed matrix and contains the most information about the image. Thus, our hypothesis was that these values seem to be a better representative of the channel matrix than the average or maximum value and should give better results.

3.2 Experiment Setup

Due to computational constraints, we only ran the experiments on ResNet [8] models of depth 20 and 32 with the CIFAR-10 dataset [9]. We only used Squeeze and Excitation Network (SENet) [1] in our study because it focuses solely on channel-wise information and does not require much computing. The architecture of SENet can be seen in figure 1. We have just changed the way 1x1xC feautures are computed by replacing GlobalAvgPool2d with our own statistic measures. The rest of the architecture and processing remains the same as in the original paper.

The CIFAR-10 dataset was split into three parts - train set containing 45000 samples, validation set containing 5000 samples and test set having 10000 samples. The data was minimally augmented using keras ImageDataGenerator on the fly, with zca_epsilon $= 1e-06$, width_shift_range $= 0.1$ for random horizontal shift, and height_shift_range $= 0.1$ for vertical shift. The batch size was 16, epochs were set to 50 with a patience of 3 on val_loss. The optimiser was Adam having initial lr $= 0.001$ with ReduceLROnPlateau $= 0.3$.

Model	Top 1 Acc.	Top 5 Acc.	Approx. Time Per Epoch
Resnet20_v1 + SE	76.64	98.68	15 min
Resnet20_v1 + SE_std	**76.83**	98.77	15 min
Resnet20_v1 + SE_trace	73.29	98.45	15 min
Resnet20_v1 + SE_svd	**77.03**	98.94	25 min
Resnet20_v1 + SE_dct	74.92	98.02	25 min

Table 1: ResNet20 models classification on CIFAR-10

Model	Top 1 Acc.	Top 5 Acc.	Approx. Time Per Epoch
Resnet32_v1 + SE	79.23	98.72	25 min
Resnet32_v1 + SE_std	**79.48**	98.79	25 min
Resnet32_v1 + SE_trace	77.96	98.04	25 min
Resnet32_v1 + SE_svd	**79.85**	98.95	45 min
Resnet32_v1 + SE_dct	78.14	98.83	60 min

Table 2: ResNet32 models classification on CIFAR-10

All the experiments were run using Intel i5-5300U CPU @ 2.30GHz × 4, 12Gb RAM, Ubuntu 20.04.3 LTS, and Python 3.8.10 environment.

3.3 Results

The results for ResNet-20 model are shown in table 1. The best Top-1 accuracy was with largest singular value of 77.03%. Standard deviation also slightly imporved the accuracy. There was however 10 minutes of additional training time per epoch with SE_SVD. The worst performance was by trace followed by DC coefficient of DCT, which gave below baseline accuracy. SE_DCT also took additional 10 minutes per epoch while training.

Table 2 shows the results for ResNet-32 model. Again, largest singular value gave the best Top-1 accuracy of 79.85% but with a cost of 20 extra minutes per epoch during training. Trace gave sub-optimal results with ResNet32 as well, dropping by around 1.4% below baseline. SE_DCT took more than double the time to train per epoch than the baseline and other methods.

The code can be found here: https://github.com/umar07/SENet-Descriptors.

Conclusions

In this work we experimented with different descriptors for channel information. Our intuition that some better summary statistic for a channel than global average or maximum might give better results was proved right to some extent. There was however an additional cost of training time. The future work of this study can be to extend the experiments to other CNN models, different attention mechanisms, and more datasets.

References

[1] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Con-*

ference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[2] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks, 2020.

[3] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[4] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module, 2018.

[5] Yichao Liu, Zongru Shao, and Nico Hoffmann. Global attention mechanism: Retain information to enhance channel-spatial interactions, 2021.

[6] Diganta Misra, Trikay Nalamada, Ajay Uppili Arasanipalai, and Qibin Hou. Rotate to attend: Convolutional triplet attention module, 2020.

[7] Yichao Liu, Zongru Shao, Yueyang Teng, and Nico Hoffmann. Nam: Normalization-based attention module, 2021.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[9] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.