컴퓨터 비전에서 신경망의 가중치 분포

오신모, 이효종
전북대학교 컴퓨터공학부
Chenmou0410@gmail.com, hlee@jbnu.ac.kr

Weight Distribution of Neural Networks in Computer Vision

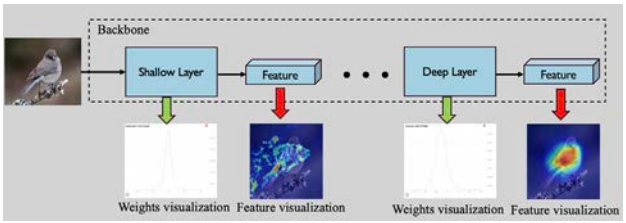
Chenmou Wu, Hyo-Jon Lee*
Dept. of Computer Science and Engineering, Jeonbuk National University
*Corresponding author

요 약

Over the last decades, deep neural networks have demonstrated significant success in various tasks. To address the special vision task, choosing a hot network as backbone to extract feature is a common way in both research and industry project. However, the choice of backbone usually requires the expert experience and affects the performance of the classification task. In this work, we propose a novel idea to support backbone decision-making by exploring the feature attribution and weights distribution of hidden layers from various backbones. We first analyze the visualization of feature maps on different size object and different depth layers to observe learning ability. Then, we compared the variance of weights and feature in last three layers. Based on analysis of the feature and wights, we summarize the traits and commonalities of existing networks.

1. Introduction

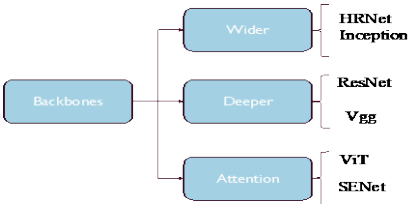
Recently, dramatic advancements in deep neural networks have opened up new capabilities for computer vision in research and applications. To address a new real-world problem, developers or researchers widely build a network by combining a suitable backbone and corresponding head. However, each network architectures focus on a particular feature extraction capability. Researchers need constant trial and error to find the applicable backbone because the choice of backbone is over relay on the expert experience. In addition, most backbones provide serval types of different parameter magnitudes.



(Figure 1) The solution of backbone analysis.

In this paper, we apply the feature attribution and weight analysis on common networks to support backbone decision-making. More specifically, the existing backbones are roughly partition into three branches based on the architecture design. From each branch, we choose two particular networks as sample for experiment. To study the fitness of backbone for different tasks, we first apply gradient camera [1] on different layer of the chosen backbone, which can visualize interested region of feature maps extracted from specified layer. The preliminary analysis can be obtained by intuitively observing the visualization results. Furthermore, we also compared learning capabilities at different depths of the network layer based on the weight distribution. For example, the features extracted by the convolutional layer with approximate weight distribution are similar. The whole solution of the proposed method as shown in Figure 1. Section 2 will introduce the chosen backbone and the visualized method. Section 3 is visualization result and analyses.

2. The Chosen Network



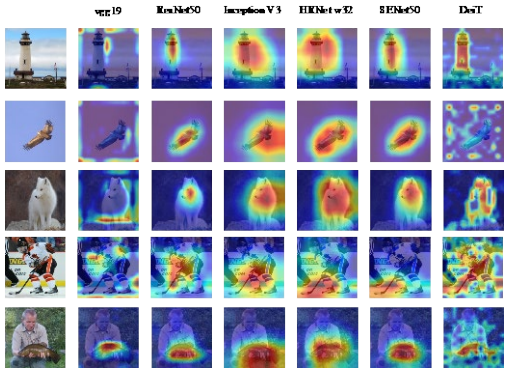
(Figure 2) Three directions of chosen backbones.

From the development of CNN, the improvement of the architecture focuses on three directions: deepening the network (add more layers), expanding the network (add more branches in one layer), and add enhancement module such as attention mechanism. As shown in Figure 2, we chose two representative networks for each branch. The Vgg19[2] and ResNet50[3] are elected as candidate on behalf for deeper architecture. The former is classical convolution neural

network while the latter is the most widely used backbone. For the wider branch, inception-V3[4] and HRNet [5] are selected for the experiment. Inception include multiple convolutional kernel size in each block and HRNet can learn multiple resolutions information in each stage. Both of them enable the network to learn information from different regions in a block or stage. For the attention part, we chose Squeeze-and-Excitation module[6] and variant of Vision-in-Transformer DeiT[7]. The following analyses only focus on these six backbones.

3. Experiments

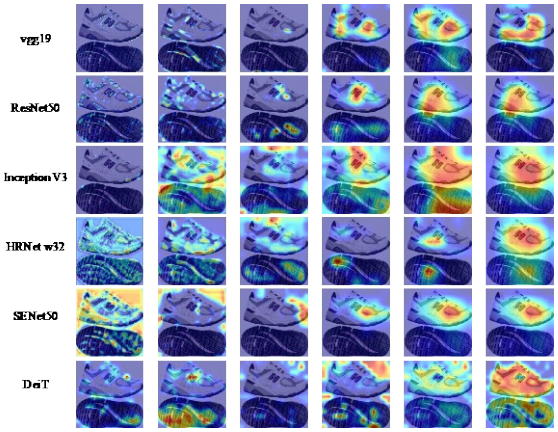
In this section, we apply gradient camera and weight histogram on the chosen backbone. The test image is from ImageNet dataset. All test models are pre-trained on the ImageNet, which is provided by PyTorch Image Models library.



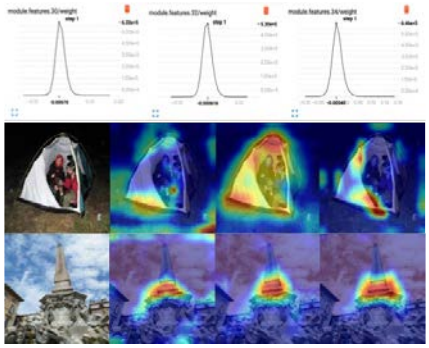
(Figure 3) The visualization on different objects of the chosen backbones.

First, we select different image on the condition that object size and background noise. As shown in Figure 3, the feature extracted from deeper backbone focus on the principal component of image. On the contrary, the wider backbones catch more information. Thus, the performance of wider backbone will be better than the deeper backbone the larger size object. However, the feature involves more information also brings more noise, especially image with complex background. The attention-based backbones are robust for complex background. But for small size object, the performance is worse than the deeper or wider backbone. Thus, the deeper architecture is more suitable for small size object while wider architecture is more suitable for larger size object. The attention module is effective against complex backgrounds but also sensitive to object size.

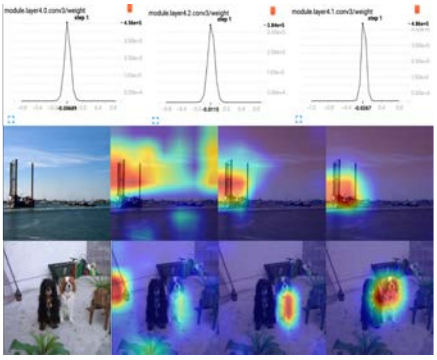
Then, we compared the different depth of the chosen backbone to view the learning process. As shown in Figure 4, we chose shallow, middle and depth layer to obvious the feature variance. The deeper and backbone learning process is similar with the process of clustering, which from edge region to region of interest. The difference mainly is receptive field which described in corresponding paper. The attention module learning is more like screening of different regions. Thus, compared with others, it is harder to finetuning for backbones including attention module.



(Figure 4) The visualization on different depth of the chosen backbones.



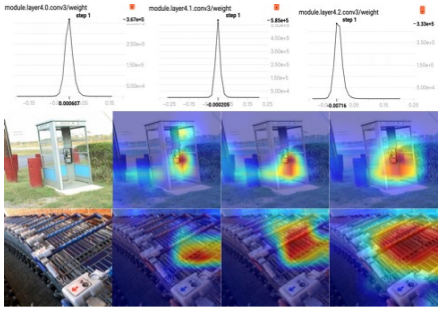
(Figure 5) The weight histogram and visualization on the last three layers of Vgg.



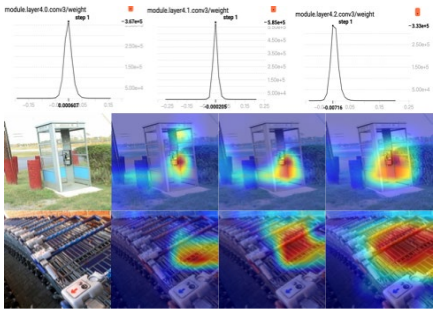
(Figure 6) The weight histogram and visualization on the last three layers of ResNet.

Furthermore, we combine feature visualization and weight histogram to analyze the last three layers of backbone. It is helpful to reduce parameters issue in the real project. The weights histogram and corresponding visualization of the deeper backbone are shown in Figure 5 and 6. The variety of weight distribution and feature are not significantly in penultimate layers. Similarly, both weight distribution and feature are change more smaller in last three layers for the attention-based backbones, which shown in Figure 7 and 8.

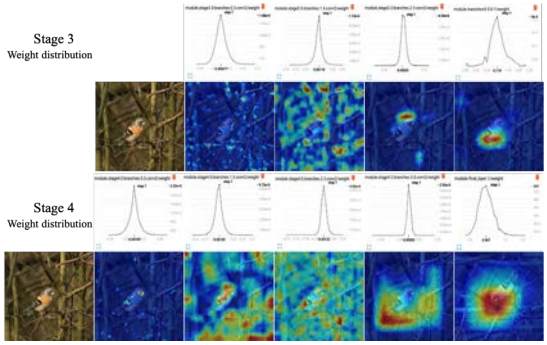
Thus, the penultimate layer is a good choice when we need slim down the deeper or attention-based backbone.



(Figure 7) The weight histogram and visualization on the last three layers of SENet.

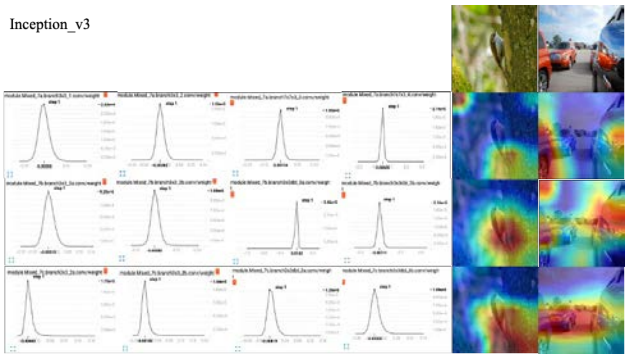


(Figure 8) The weight histogram and visualization on the last three layers of DeiT.



(Figure 9) The weight histogram and visualization on the last three layers of HRNet.

For the wider backbone, there are different situation. In the HRNet, each stage contains different branches, and the final feature of this stage is fused by these branches. As shown in Figure 9, the difference between different stages is large, and the feature of different branch in one stage is completely different. In the Inception network, each block includes different convolution kernel size. As shown in Figure 10, the variety of weight distribution is larger for last three layers. However, the variety of feature is strange. The features are similar for the small size object while are different for the large size object. Thus, the network parameters cannot be reduced by simply cutting the number of network layers for wider backbones.



(Figure 10) The weight histogram and visualization on the last three layers of Inception.

4. Conclusions

This paper summarized the traits of various backbones based on feature visualization technology and weight distribution. Based on our backbone class, the deeper architecture backbones are more suitable for small size object while wider architecture and attention-based backbones are suitable for large size object. To slim the backbone, deeper architecture and attention-based backbone can simply reduce the number of layers while wider architecture backbones need more complex process.

Acknowledgement

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant 2019R1D1A3A03103736 and in part by project for Joint Demand Technology R&D of Regional SMEs funded by Korea Ministry of SMEs and Startups in 2021 (No. S3035805).

Reference

[1] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra. Grad cam: Visual explanations from deep networks via gradient based localization. ICCV 2017, pp. 618-626.

[2] K. Simonyan, A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. ICLR, 2015.

[3] K. He, X. Zhang, S. Ren, J. Sun. Deep residual learning for image recognition. CVPR 2016, pp. 770-778.

[4] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna. Rethinking the inception architecture for computer vision. CVPR 2016, pp. 2818-2826.

[5] K. Sun, B. Xiao, D. Liu, J. Wang. Deep high-resolution representation learning for human pose estimation. CVPR 2019, pp. 5693-5703.

[6] J. Hu, L. Shen, G. Sun. Squeeze-and-excitation networks. CVPR 2018, pp. 7132-7141.

[7] H. Touvron, M. Cord, M. Douze, F. Massa, H. Jégou. Training data-efficient image transformers & distillation through attention. ICML 2021, pp. 10347-10357.