Measuring Augmentation Effects with Central Kernel Alignment

Bumjin Park

Abstract

Data augmentation is known to be effective for generalization. However, it is unclear how much the model is affected by the augmentation. To quantitatively measure the influence, we compute the Central Kernel Alignment of CNN model features for increased data augmentation. We empirically show that the deeper layers have lower similarity with the model without augmentation.

1. Introduction

Regularization on input space is particularly important for robustness and generalization. Because a neural network is a block-box, users could not compare the regularized models in quantitative manner. In addition, it is not clear whether the internal representations are clearly different based on the regularization. In this work, we demonstrate the properties of Central Kernel Alignment with data augmentation.

2. Method

2.1. Central Kernel Alignment

Consider a pair of hidden representations generated by two representations $\Theta_1 : \mathcal{X} \to \mathbb{R}^{d_1}$ and $\Theta_2 : \mathcal{X} \to \mathbb{R}^{d_1}$, **Central Kernal Alignment (CKA)** is defined by

$$\text{CKA}(\Theta_1, \Theta_2) = \frac{||\text{Cov}(\Theta_1(x), \Theta_2(x))||_F^2}{||\text{Cov}(\Theta_1(x))||_F \cdot ||\text{Cov}(\Theta_2(x))||_F} \tag{1}$$

CKA provides similarity score from 0 to 1 and the $\text{Cov}(\Theta_1(x))$ is defined by the variance obtained by features with kernels such as the linear kernel and the RBF kernel.

2.2. Perturbation on Input Space

Given image x, attacking is done by slightly moving the image to the direction that changes the classification result. In the same manner, we can easily achieve the robustness by augmenting the data with random noise on it.

$$\hat{x} = x + \lambda\epsilon \tag{2}$$

where $\epsilon \sim \mathcal{U}(-1, 1)$ is random variable in the normal distribution. As ϵ increase, the regularization covers more regions of the perturbed input space.

3. Result

To evaluate the similarity on feature space, we use two blocks CNN model with CIFAR10 dataset. 1 shows the accuracy score over perturbation λ. The result show that $\lambda = 0.1$ and $\lambda = 0.2$ have better effect on the performance than the model without augmentation $\lambda = 0.0$.

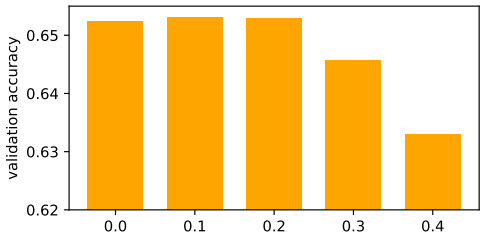

Figure 1. CIFAR10 validation accuracy over perturbation ratio λ.

Next, we measure the similarity between the features space for each layer (Conv1, Conv2, and the linear layer). Figure 2 shows the CKA with non-augmented model and the augmented models. The result show that there is decreasing trend over layers. That is, the similarity decrease as the layers are deeper. However, we could not observe clear distinction by perturbation. Therefore, we could conclude that the feature space is not clearly different.

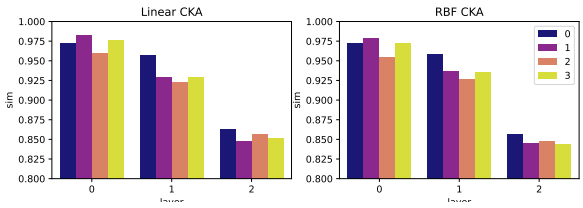

Figure 2. CKA with the non-augmented and augmented models. X-axis indicates the layer index. There is decreasing trend over layers, yet no clear trend by the perturbation ratio.

4. Conclusion

In this work, we show that (1) models have more different semantics as the layers are deeper (2) models with data augmented model has no clear CKA score difference.

Layer	Description
Convolution	(3,16,3,1,1)
ReLU	-
MaxPool	2
Convolution	(16,16,3,1,1)
ReLU	-
MaxPool	2
Flatten	-
Linear	(1024,512)
ReLU	-
Linear	(512,10)
MaxPool	2

Table 1. Model description. The solid lines indicate the end of blocks. The numbers in the convolution layers indicate (in-channels, out-channels, kernel size, stride, and padding).

References