Laboratory of Machine Learning and Popular Computing
School of Computer Science, Nankai University, Tianjin, China

How Does DCNN Make Decisions?

Yi Lin, Namin Wang, Xiaoqing Ma, Ziwei Li, Gang Bai

Abstract

In this paper, the major contributions we made are: firstly, provide the hypothesis, "point-wise activation" of convolution function, according to the analysis of DCNN' s architectures and training process; secondly, point out the effect of "point-wise activation" on DCNN' s uninterpretable classification and pool robustness, and then suggest, in particular, the contradiction between the traditional and DCNN' s convolution kernel functions; finally, distinguish decision-making interpretability from semantic interpretability, and indicate that DCNN' s decision-making mechanism need to evolve towards the direction of semantics in the future. Besides, the "point-wise activation" hypothesis and conclusions proposed in our paper are supported by extensive experimental results.

Pixel-Wise Activaion

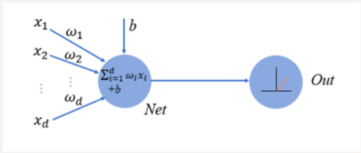

Fig.1. The architectures of the convolutional layers and function ReLU

According to the architectures of the convolutional layers and function ReLU, we provide the "Point-Wise Activation" hypothesis.

The "Picel-Wise Activation" hypothesis: pixels(points) used to be combined as features to distinguish objects can be rather few in DCNN. (Unless otherwise stated, pixels(points) in our paper are of single-channel.)

Experiments

Network Training Experiments

Model: VGG16/Resnet18

We find that the distributions of parameters in DCNN' s convolutional layers are roughly presenting a normal distribution centered on zero.

Conclusion: Due to the 3σ principle of the normal distribution, we consider trimming out parameters of small contributions in order to acquire a better understanding of the limits of current DCNN architectures.

Network Compression Experiments

Model: VGG16/Resnet18

We find that the the model accuracy remains above 85% when the pruning rate has reached 90%.

Conclusion: DCNN' s decision-making mainly determined by a few critical pixels. These pixels are diifficult to construct enough features that people can understood. In orther words, DCNN just distinguish categories instead of recognizing them.

Adversarial Attack Experiments

One- pixel attack[1]: Generating one-pixel adversarial perturbations based on differential evolution (DE) .

Conclusion: DCNN enables very few pixels play a key role in classifications. The results verify the conclusion of network compression experiments again.

Single Value Experiments

Model: VGG16/Resnet18

Conclusion: As long as the pixel exists, DCNN can classify regardless of whether it has intelligible features such as shapes or not.

Hard Sample Experiments

Model: VGG16-like/Resnet18-like

Conclusion: It is rather difficult for human observers to recognize the distribution of pixel sets used by DCNN.

Discussions and Conclusion

Discussions

No1. The convolution kernel function in DCNN is different from the others used in the field of traditional computer vision for the methods to determine the parameters differ.

No2. We consider "point-wise activation" as the main reason for the poor robustness of DCNN.

No3. For interpretability, we consider dividing it into two types. One is decision-making interpretability, the other is semantic interpretable.

Conclusion

In order to make credible decisions, DCNN' s decision-making mechanism need to evolve towards the direction of semantics in the future.

References

[1]J. Su, D. V.Vargas, and S. Kouichi. "One pixel attack for fooling deep neural networks," IEEE Transactions on Evolutionary Computation, 2017.

ICPR2020 Milan 25th International Conference on Pattern Recognition