Group-wise Inhibition based Feature Regularization for Robust Classification

Haozhe Liu†, Haoqian Wu†, Weicheng Xie*, Feng Liu*, Linlin Shen

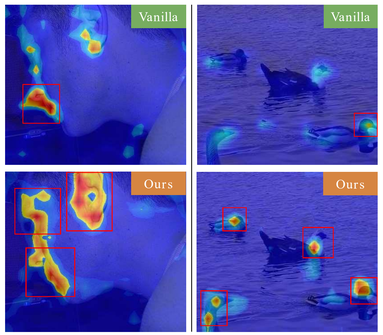
Institute of Computer Vision, College of Computer Science and Software Engineering
Shenzhen University, Shenzhen 518060, China

2021 ICCV OCTOBER 11-17 VIRTUAL

Abstract

The convolutional neural network (CNN) is vulnerable to degraded images with ev en very small variations (e.g. corrupted and adversarial samples). One of the possi ble reasons is that CNN pays more attention to the most discriminative regions, but ignores the auxiliary features when learning, leading to the lack of feature diversity for final judgment. In our method, we propose to dynamically suppress significant activation values of CNN by group-wise inhibition, but not fixedly or randomly ha ndle them when training. The feature maps with different activation distribution ar e then processed separately to take the feature independence into account. CNN is finally guided to learn richer discriminative features hierarchically for robust classi fication according to the proposed regularization. Our method is comprehensively evaluated under multiple settings, including classification against corruptions, adve rsarial attacks and low data regime. Extensive experimental results show that the p roposed method can achieve significant improvements in terms of both robustness and generalization performances, when compared with the state-of-the-art methods. ***Code is available at https://github.com/LinusWu/TENET_Training.***

Motivation



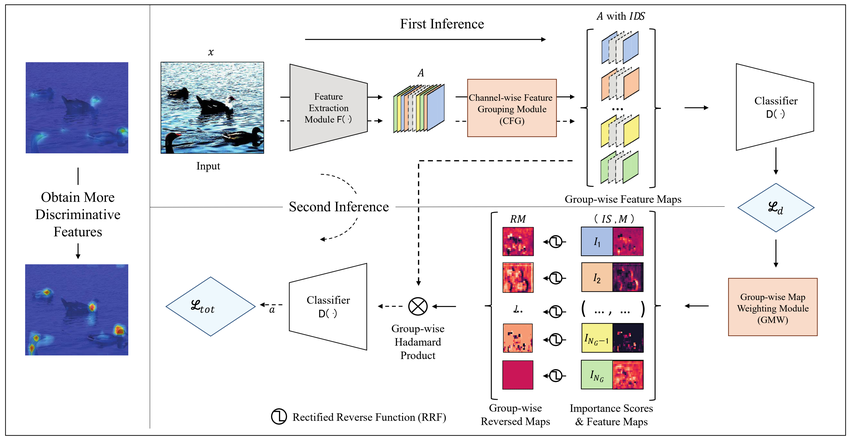
(a) Single Instance (b) Multiple Instances

GENERAL IDEA

CNNs can locate the most discriminativ e regions for both single-instance and m ulti-instance samples with the regulariza tion method, while neglecting other aux iliary features that are critical for the rec ognition. The lack of auxiliary features may lead to insufficient feature diversity, which consequently results in a feature space with low-dimension for classificat ion and limits the robustness.

The heatmap visualization of feature maps encoded with ResNet-50, based on Grad -CAM with or without the pro- posed method. Our method locates more diverse dis criminative regions (in red boxes) for both single-instance (a) and multiple-instanc e (b) samples.

Method



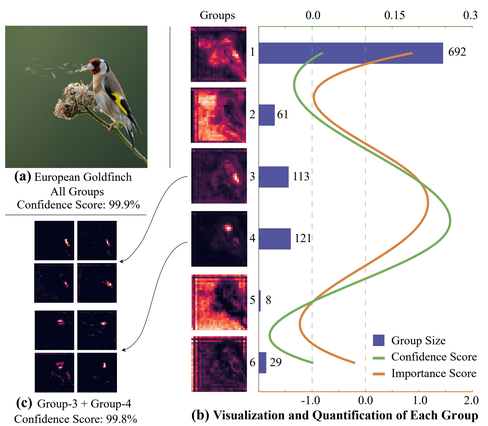
PIPELINE:
CNNs consist of the feature extr action module $F(\cdot)$ and the class ifier $D(\cdot)$. In the first inference, feature maps A encoded with $F(\cdot)$ are divided into multiple gro ups by the CFG module, and los s $\mathcal{L}_d$ is calculated based on $D(\cdot)$. Reversed maps *RM* are then der ived using GMW module and R RF. In the second inference, the Hadamard Product of A (with IDS) and *RM* is fed to $D(\cdot)$ to cal culate the loss $\mathcal{L}_{total}$.

Experimental Results

Task-[protocol]	Dataset	Previous SOTA	Gain
Standard Classification-[4]	PASCAL VOC 2012[9]	Group Orthogonal Training [4]	**2.9%**
Robustness against Adversarial Attack-[8, 24]	CIFAR-10/100 [18]	A. T. [24] Augmix[14]	**5.75%** **15.56%***
Robustness against Common Corruption-[13, 14, 21]	CIFAR-10/100-C [13] ImageNet-C [13]	Augmix[14]	**1.77%** **2.8%**†
Generalization-[2]	CUB-200 [26]	GLICO [2]	**2.75%**

To evaluate the performance of the proposed method, extensive experiments are carri ed on publicly-available data sets, including PASCAL VOC 2012, CIFAR-10/100, I mageNet-C and CUB-200.

The visualization and quantif ication of the feature maps ex tracted by the 3rd residual bl ock of ResNet-50 using TEN ET Training. (a) An input ima ge with the label of European Goldfinch. (b) The activation distribution, the correspondin g im- portance and confidenc e scores of each group cluster ed by CFG module. (c) The e xample feature maps selected from the 3rd and 4th groups.



Conclusion

- No limitation for network architecture
- Flexible to complement with other solutions
- Improving both robustness and generalization

Paper(Arxiv):

Code(Github):