CIFS: Improving Adversarial Robustness of CNNs via Channel-wise Importance-based Feature Selection

Hanshu Yan [1] Jingfeng Zhang [3] Gang Niu [3] Jiashi Feng [1]
Vincent Y. F. Tan [1,2] Masashi Sugiyama [3]

[1]ECE, NUS [2]Math, NUS

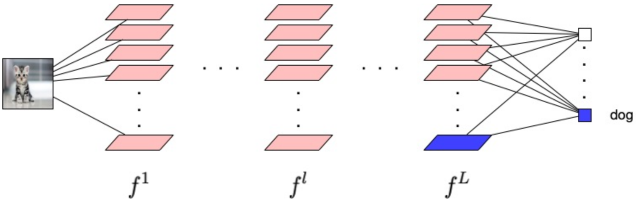
[3]RIKEN-AIP [4]GSFS, UTokyo

ICML 2021

Motivation

Different from adversarial training (AT)-based methods, this paper proposed a novel mechanism to modify CNNs, so that the robustness of CNNs can be further enhanced under AT.

- CNNs make predictions by aggregating information from various channels / feature maps
- Abnormal activated channels may result in significant prediction error

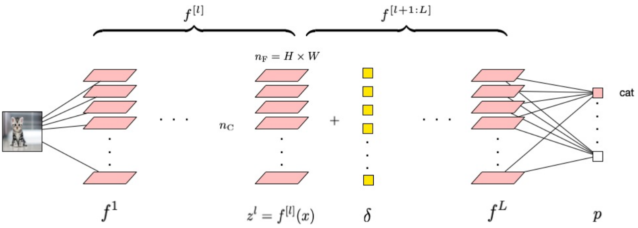


Motivation

Different from adversarial training (AT)-based methods, this paper proposed a novel mechanism to modify CNNs, so that the robustness of CNNs can be further enhanced under AT.

- It is necessary to investigate the relation between robustness and channels' activations, i.e., what types of channels are over/under activated by adversarial data.
- We can enhance the robustness of CNNs by controlling the activations of channels.

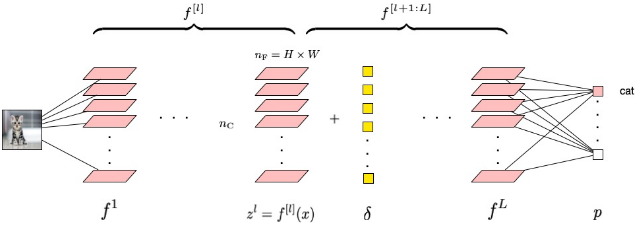
Relevances of channels to predictions



- Activation level of i^{th}: $\sum_j^{n_\mathsf{F}} (z^l_{[i][j]})/n_\mathsf{F}$
- add a channel-wise perturbation $\delta \in \mathbb{R}^{n_\mathsf{C}}$ to z^l, $z^l_\delta = z^l + \delta \cdot \mathbf{1}^\top$, where $\mathbf{1} \in \mathbb{R}^{n^l_\mathsf{F}}$
- the relevance of i^{th} channel to class y is defined as $g^l_{[i]}$

$$g^l = \nabla_\delta p^l(\delta)_{[\hat{y}]}\Big|_{\delta=\mathbf{0}} = \nabla_{z^l_\delta} f^{[l+1:L]}(z^l_\delta)_{[\hat{y}]}\Big|_{z^l_\delta = z^l} \cdot \mathbf{1}$$

Relevances of channels to predictions



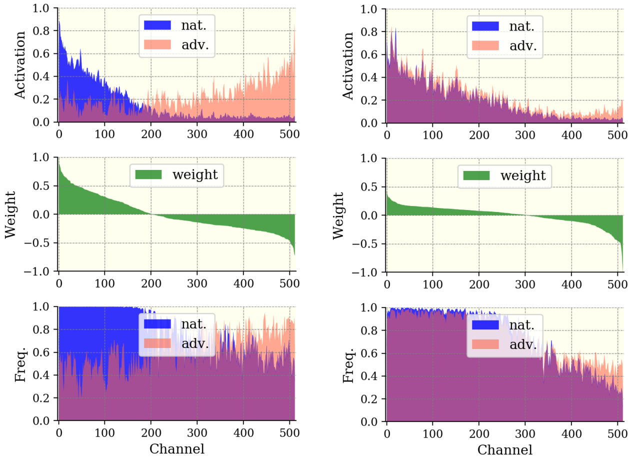
- Positively-relevant (PR) channels: $g^l_{[i]} > 0$
- Negatively-relevant (NR) channels: $g^l_{[i]} \leq 0$

Abnormal Channels in non-robust and robustified CNNs

Comparing channels' activations of non-robust and robustified CNNs

- ResNet-18, CIFAR10
- non-robust, normally trained
- robustified, adversarially trained
- feature maps of the penultimate layer (output of the last res-block before the global avg pooling and the final linear layer)
- say the true label is class k, the weights in the linear layer corresponding to class k can represent the relevances of channels

Non-robust vs. Robustified CNNs



(a) Normal / "automobile" **(b)** Adv. / "automobile"

Figure 1: Robust accuracies against PGD-20: 0% vs. 46.6%.

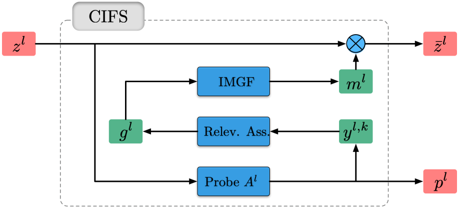
A hypothesis denoted as $\mathcal{H}$

Suppressing NR channels and promoting channels' activations based on their relevances to prediction results benefit the robustness of CNNs.

To verify hypothesis $\mathcal{H}$, we need a technique for

- Relevance assessment
- Generating importance scores to control channels' activations

CIFS: Channel-wise Importance-based Feature Selection



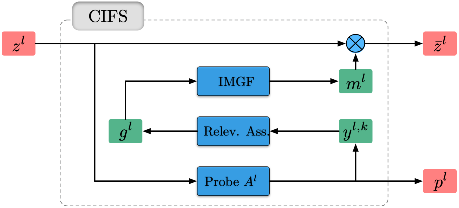
Relevance assessment

- auxiliary classifier A^l as a surrogate of $f^{[l+1:L]}$
- $p^l = A^l(z^l) \in \mathbb{R}^K$, trained under supervision of ground-truth labels.
- relevance vector g^l

$$g^l = \nabla_\delta \sum_{i \in y^{l,k}} p^l(\delta)_{[i]} \Big|_{\delta=\mathbf{0}} = \nabla_{z^l_\delta} \sum_{i \in y^{l,k}} A^l(z^l_\delta)_{[i]} \Big|_{z^l_\delta = z^l} \cdot \mathbf{1}$$

- $y^{l,k}$ denotes indices of the k largest logits of prediction p^l

CIFS: Channel-wise Importance-based Feature Selection



Importance Map Generating Function (IMGF)

- monotonic non-negative mapping (promoting PR channels)
- mapping negative values to targets close to zero (suppressing NR channels)

Options:

- softplus: $m^l_{[i]} = \frac{1}{\alpha} \cdot \log(1 + \exp(\alpha \cdot g^l_{[i]})), \quad \alpha > 0.$
- softmax: $m^l_{[i]} = \frac{\exp(g^l_{[i]}/T)}{\sum_j \exp(g^l_{[j]}/T)}, \quad T > 0.$

Adversarial Training of CIFS

In practice, we may apply the CIFS mechanism into several layers of a CNN.

- I, the set of indices of these layers
- θ_A^I, the parameters of all the probes in the CIFS-modified layers
- $|I|$ raw predictions and one final prediction $p = \bar{f}^{[L]}(x)$

$$\ell_\beta(x, y) = \frac{1}{1+\beta} \cdot \ell_{\text{ce}}(p, y) + \frac{\beta}{(1+\beta)|I|} \cdot \sum_{l \in I} \ell_{\text{ce}}(p^l, y), \tag{1}$$

- $\beta > 0$ balances the accuracy of raw predictions by CIFS and the final prediction. In practice, we set β to be $|I|$

$$\min_{\theta^{[L]}, \theta_A^I} \mathbb{E}_{P_{XY}} \left[\max_{X' \in \mathcal{B}(X, \epsilon, l_\infty)} \ell_\beta(X', Y) \right], \tag{2}$$

where $\mathcal{B}(x, \epsilon, l_\infty) = \{x' \mid \|x' - x\|_{l_\infty} \leq \epsilon\}$.

Verification of Hypothesis $\mathcal{H}$

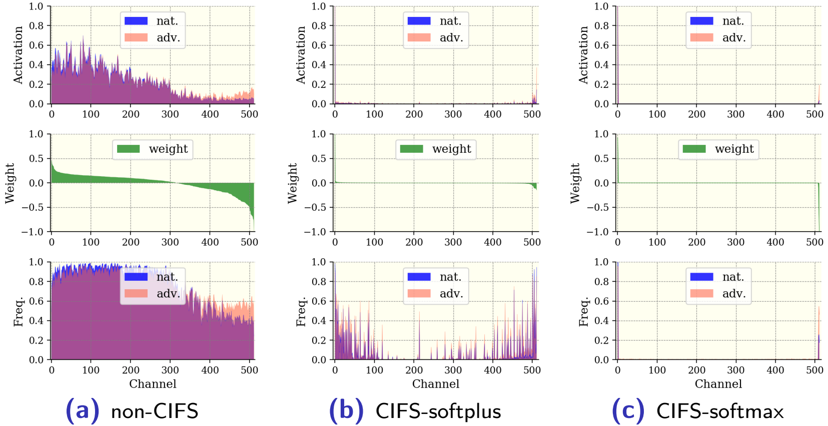


Figure 2: The robust accuracies against PGD-20 (on the whole dataset) are 46.64% for non-CIFS, 49.87% for the CIFS-sigmoid, 50.38% for the CIFS-softplus, and 51.23% for the CIFS-softmax respectively

More Experimental Results

Table 1: Robustness comparison of defense methods on CIFAR10. We report the accuracies (%) for adversarial and natural data. For each model, the results of the strongest attack are marked with an underline.

ResNet-18	Natural	FGSM	PGD-20	C&W	PGD-100
Vanilla	84.56	55.11	46.62	45.95	<u>44.72</u>
CAS	86.73	55.99	45.29	44.18	<u>43.22</u>
CIFS	83.86	**58.86**	**51.23**	**50.16**	<u>**48.70**</u>
WRN-28-10	Natural	FGSM	PGD-20	C&W	PGD-100
Vanilla	87.29	58.50	49.17	48.68	<u>47.08</u>
CAS	88.05	57.94	49.03	47.97	<u>47.25</u>
CIFS	85.56	**61.34**	**53.74**	**53.20**	<u>**51.51**</u>

Summary

- we observe that adversarial data tends to over-activate NR channels and under-activate the PR channels.
- we propose CIFS to modify the feature maps of conv layers by suppressing NR channels but promoting PR channels
- we conduct extensive experiments to verify that CIFS further enhances the robustness of CNNs under AT

Thanks !