北京大學
PEKING UNIVERSITY

Paper tag:THU-AM-387

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

AGAIN: Adversarial Training with Attribution Span Enlargement and Hybrid Feature Fusion

Shenglin Yin[1], Kelu Yao[2, 3, *], Sheng Shi[4, 5], Yangzhou Du[5], Zhen Xiao[1, *]

[1]School of Computer Science, Peking University, China
[2]Zhejiang Laboratory, Hangzhou 311100, China
[3]Institute of Computing Technology, Chinese Academy of Sciences, China
[4]Northwest University, Xi’ an 710127, P. R. China
[5]AI Lab, Lenovo Research, Beijing 100094, P. R. China
*Correspondence authors

Overview

Problem: The deep neural networks (DNNs) trained by adversarial training (AT) usually suffered from significant robust generalization gap, i.e., DNNs achieve high training robustness but low test robustness.

Motivation: We aim to improve the generalizability of the adversarially trained model and to improve the accuracy of the model on both clean data and adversarial examples.

Our approach:

We find that the generalization ability of models trained by adversarial training is related to their attribution span.

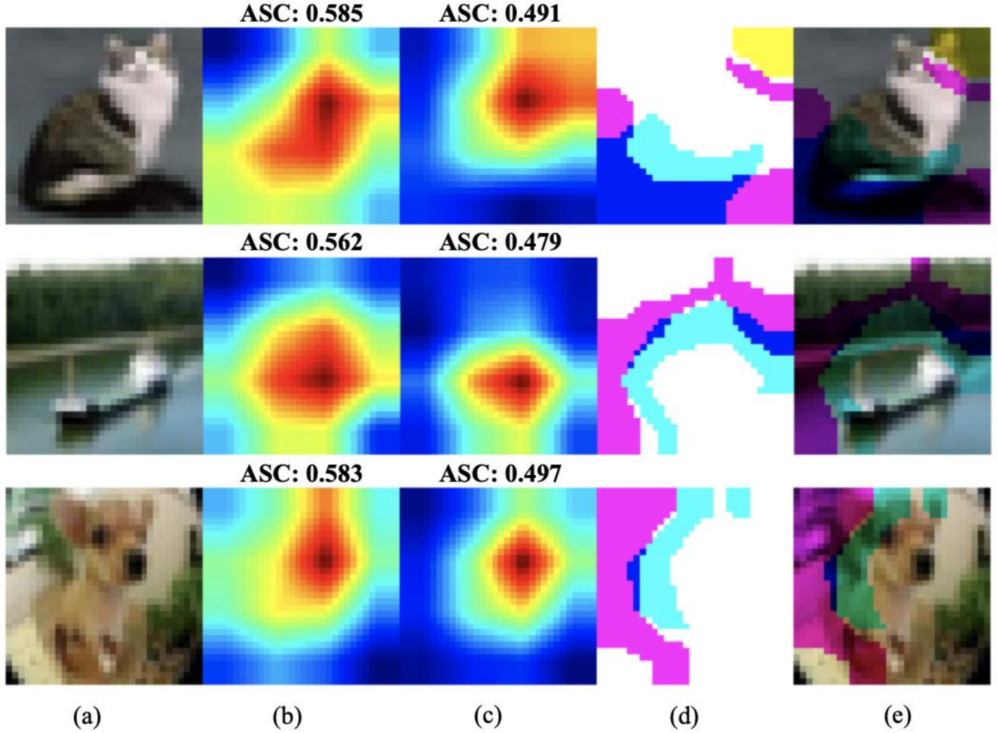
We propose a method to boost AT, called AGAIN, which is short for Attribution Span EnlarGement and Hybrid FeAture FusIoN. It improves the model's generalization by enlarging the learned attribution span.

Introduction

Methods 1[1, 2]: narrow the generalization gap from perspective of weight loss landscapes.

Methods 2[3, 4]: narrow the generalization gap from perspective of training strategies.

Our approach: narrow the generalization gap from perspective of attribution span.



[1] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. neural in- formation processing systems, 2020. 2, 6, 7

[2] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In International Conference on Learning Representations, 2019. 2, 3, 6, 7

[3] Xiaojun Jia, Yong Zhang, Baoyuan Wu, Ke Ma, Jue Wang, and Xiaochun Cao. Las-at: Adversarial training with learn- able attack strategy. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13398–13408, 2022. 2, 3, 6, 7

[4] Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In International conference on machine learning, pages 11278– 11287. PMLR, 2020. 2, 3

Method

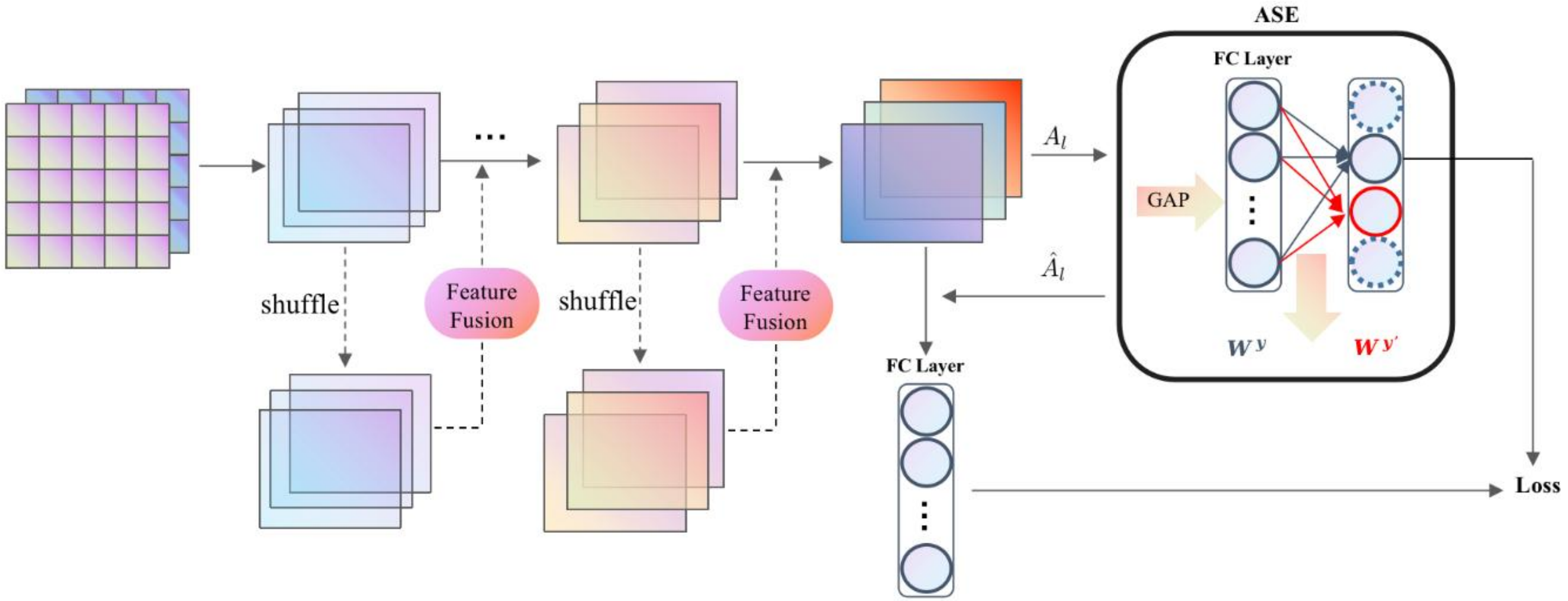


Figure 2. A visual illustration of our proposed method. It consists of two parts: attribution span enlargement (ASE) and hybrid feature fusion (HFF). ASE is used in the second last layer of the model; HFF is used before ASE.

Method

Attribution Span Enlargement (ASE)

$$\hat{A}_l = \alpha \cdot A_l \cdot W_{ASE}^{y} + (1-\alpha) \cdot A_l \cdot W_{ASE}^{y'}$$

Hybrid Feature Fusion (HFF)

$$\hat{\mu} = \gamma_1 \cdot \mu(\mathbf{A}_l) + \gamma_2 \cdot \mu(\mathbf{A}_l') + \gamma_3 \cdot \mu(\mathbf{A}_l^1) + \gamma_4 \cdot \mu(\mathbf{A}_l^2)$$

$$\hat{\sigma} = \gamma_1 \cdot \sigma(\mathbf{A}_l) + \gamma_2 \cdot \sigma(\mathbf{A}_l') + \gamma_3 \cdot \sigma(\mathbf{A}_l^1) + \gamma_4 \cdot \sigma(\mathbf{A}_l^2)$$

$$FeatureFusion(\mathbf{A}_l) = \hat{\sigma}\frac{\mathbf{A}_l - \mu(\mathbf{A}_l)}{\sigma(\mathbf{A}_l)} + \hat{\mu}$$

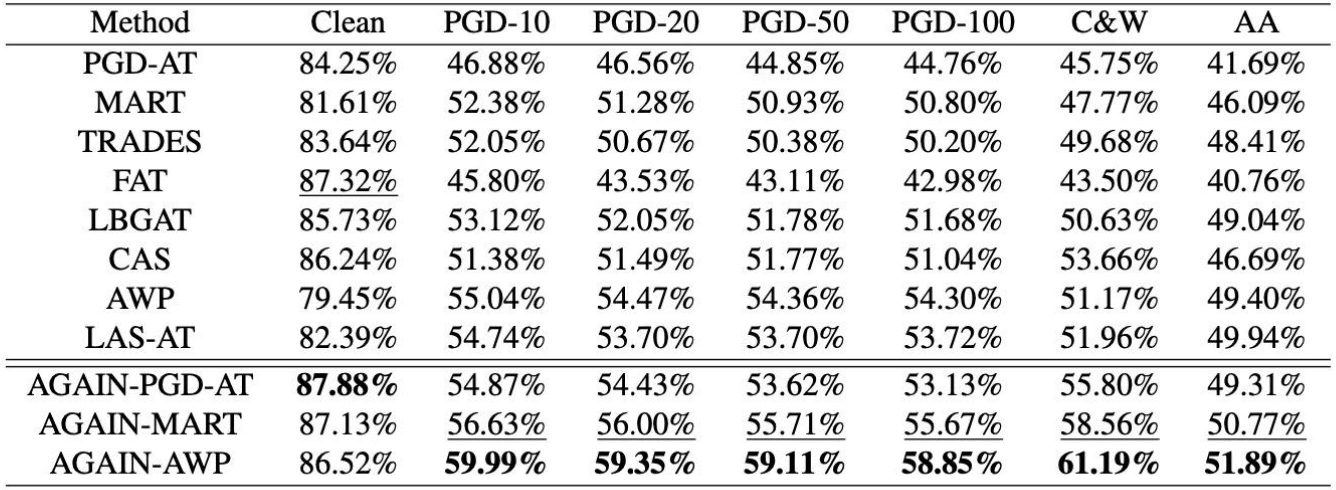
Loss Function

$$L = L_{CE}(F_{ori}(\mathbf{x}_{adv}), y) + L_{CE}(F_{ASE}(F_{ori}^{l}(\mathbf{x}_{adv})), y)$$

Experiments

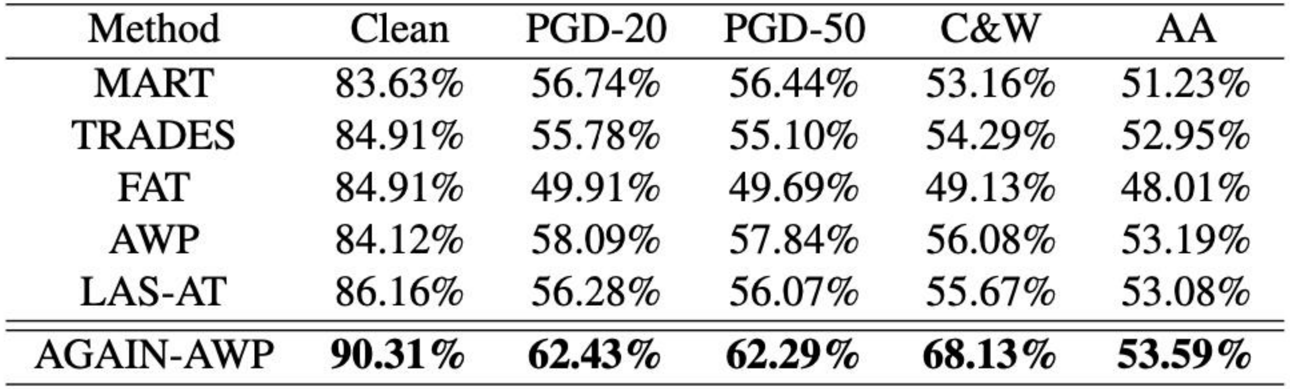
White-box Attack

- Results on ResNet-18



Method	Clean	PGD-10	PGD-20	PGD-50	PGD-100	C&W	AA
PGD-AT	84.25%	46.88%	46.56%	44.85%	44.76%	45.75%	41.69%
MART	81.61%	52.38%	51.28%	50.93%	50.80%	47.77%	46.09%
TRADES	83.64%	52.05%	50.67%	50.38%	50.20%	49.68%	48.41%
FAT	87.32%	45.80%	43.53%	43.11%	42.98%	43.50%	40.76%
LBGAT	85.73%	53.12%	52.05%	51.78%	51.68%	50.63%	49.04%
CAS	86.24%	51.38%	51.49%	51.77%	51.04%	53.66%	46.69%
AWP	79.45%	55.04%	54.47%	54.36%	54.30%	51.17%	49.40%
LAS-AT	82.39%	54.74%	53.70%	53.70%	53.72%	51.96%	49.94%
AGAIN-PGD-AT	**87.88%**	54.87%	54.43%	53.62%	53.13%	55.80%	49.31%
AGAIN-MART	87.13%	56.63%	56.00%	55.71%	55.67%	58.56%	50.77%
AGAIN-AWP	86.52%	**59.99%**	**59.35%**	**59.11%**	**58.85%**	**61.19%**	**51.89%**

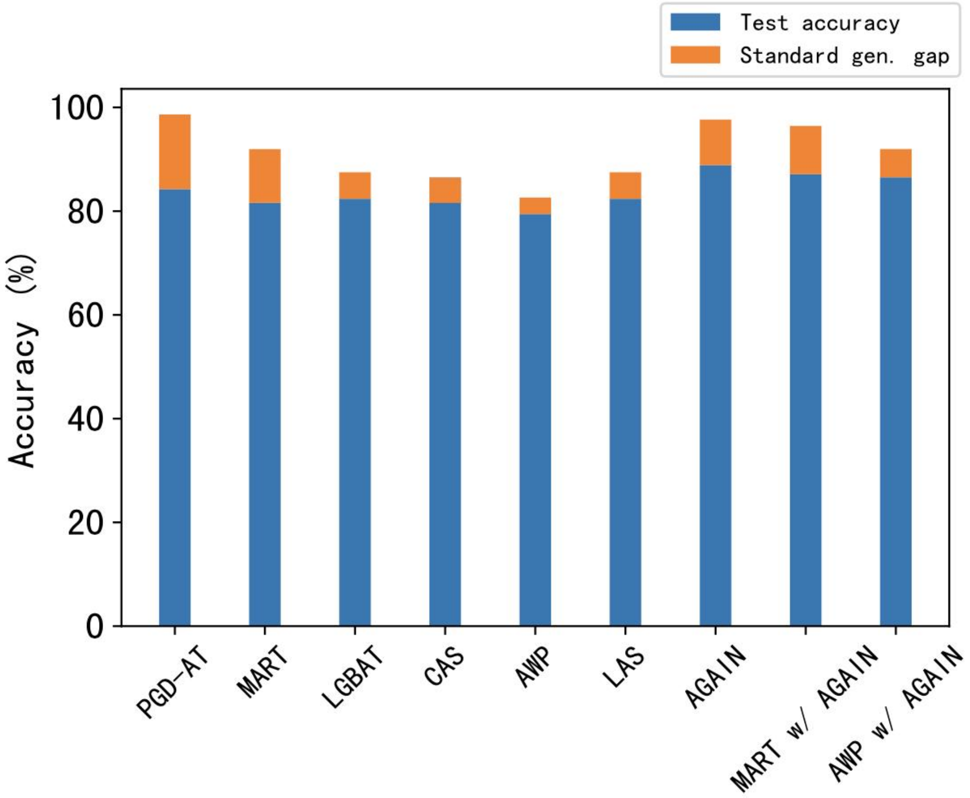
- Results on WRN-34-10



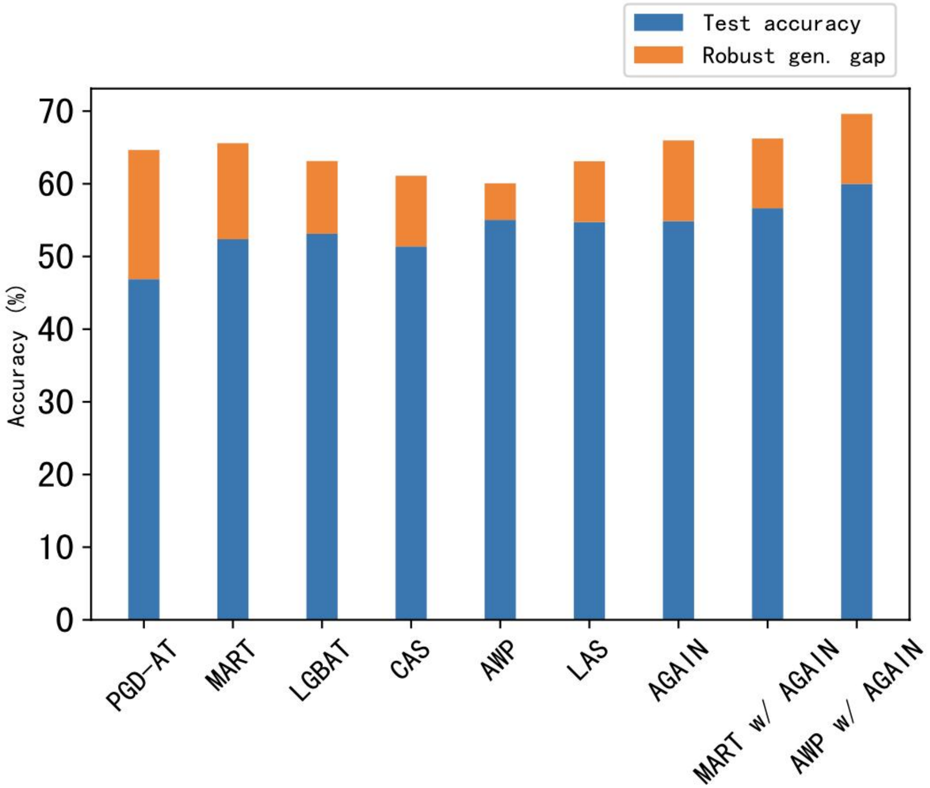
Method	Clean	PGD-20	PGD-50	C&W	AA
MART	83.63%	56.74%	56.44%	53.16%	51.23%
TRADES	84.91%	55.78%	55.10%	54.29%	52.95%
FAT	84.91%	49.91%	49.69%	49.13%	48.01%
AWP	84.12%	58.09%	57.84%	56.08%	53.19%
LAS-AT	86.16%	56.28%	56.07%	55.67%	53.08%
AGAIN-AWP	**90.31%**	**62.43%**	**62.29%**	**68.13%**	**53.59%**

Experiments

Analysis of the Generalization



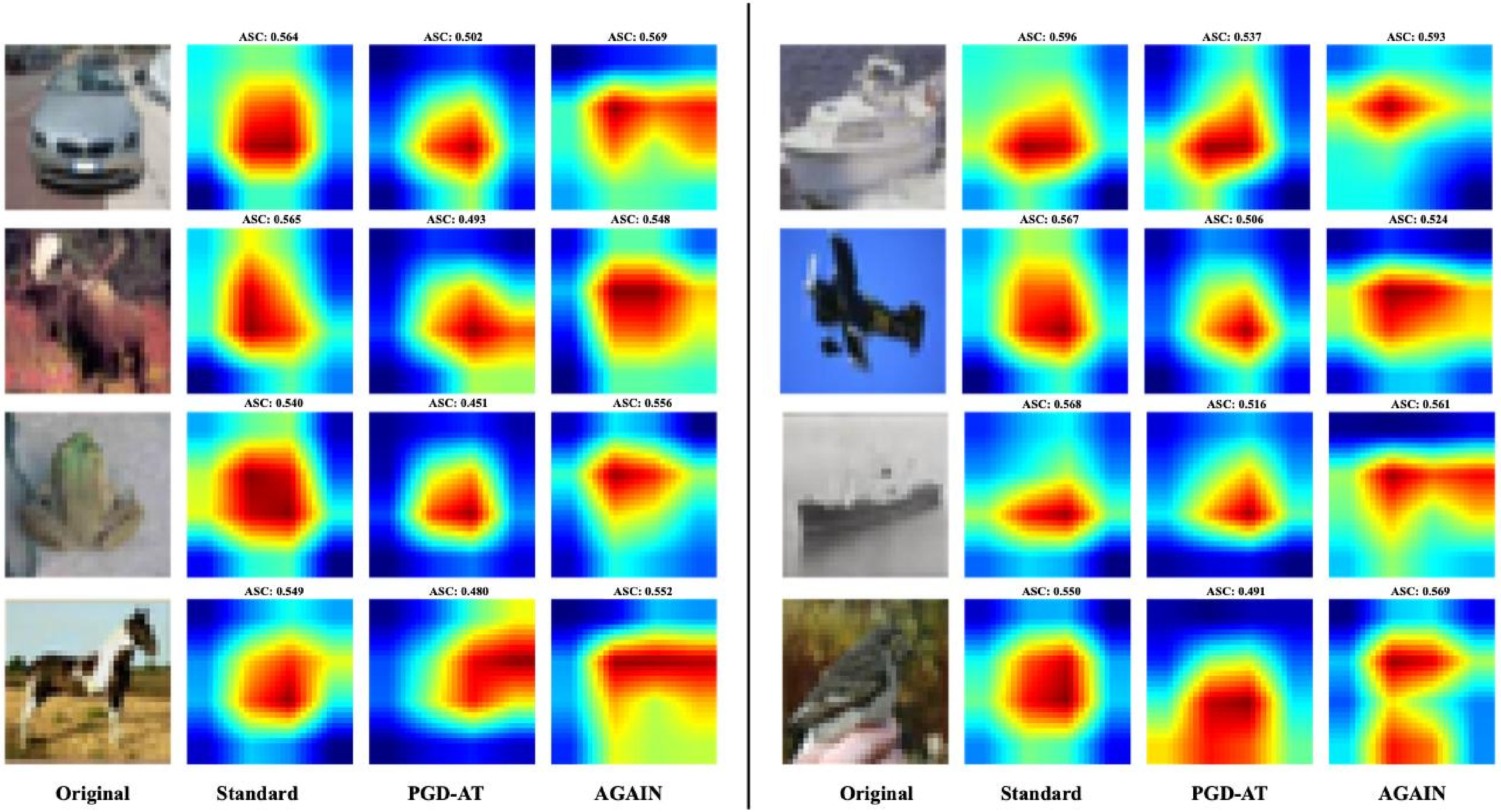
(a) Standard Generalization



(b) Robust Generalization

Experiments

Visualization



Conclusion

In this paper, we propose an AT approach based on attribution span enlargement and hybrid feature fusion. The method ensures that the model learns robust features while paying extra attention to features in other spans, and combines feature fusion to improve the accuracy of the model on clean data and adversarial examples. Comprehensive experiments show that our method is effective and general enough to improve the robustness of the model across different AT methods, network architectures and datasets.

Thank you!