Robustness of Visual Explanations to Common Data Augmentation Methods

Lenka Tětková and Lars Kai Hansen, Technical University of Denmark

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Motivation

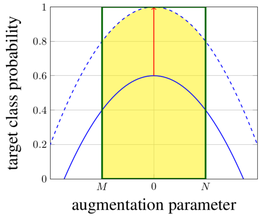
- Explanations are fragile to adversarial attacks [1], [2].
- What about naturally occuring perturbations?
- Invariant methods (changes in brightness, saturation and hue): The explanation of the augmented image should be the same as the explanation of the original image.
- Equivariant methods (rotation, translation and scaling): The explanation of the augmented image should be the same as the augmented explanations of the original image.

Assumption

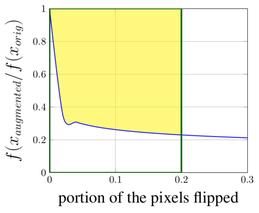
If a transformation of an image does not change the target class, the explanation should assign importance to the same part of the object as in the untransformed image.

Methods

- Intervals for augmentations are chosen so that the classification performance was reduced by 10%.
- Compare predictions (probability of the target class) and correlation of the explanations.
- We define a score S(correlation, probability) as a quotient of AUC(correlation) and AUC(probability)
- S(correlation, probability) < 1: explanations are less robust than predictions.

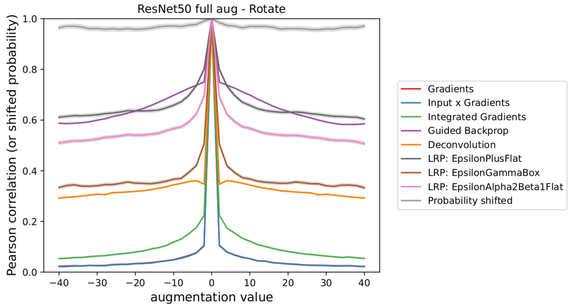


(a) Robustness score

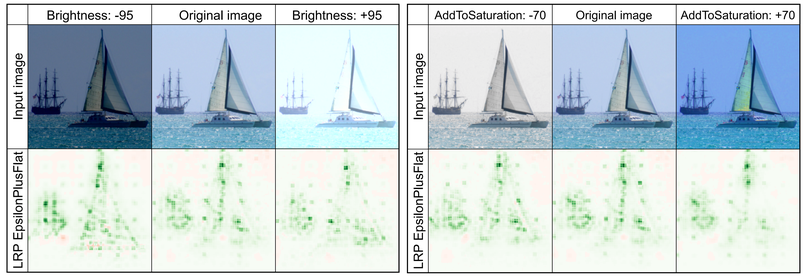


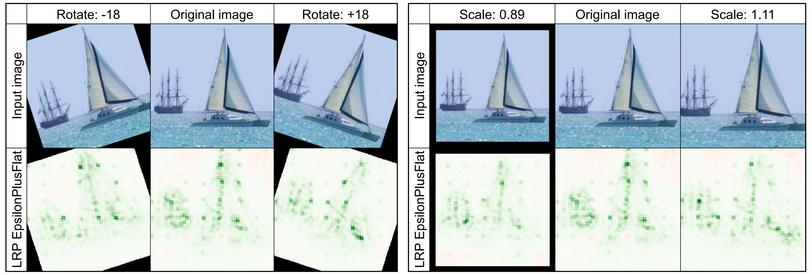
(b) Pixel flipping

- Compare models trained with full ("full aug") and limited ("lim aug") data augmentation.
- ResNet50 in the paper but similar results for EfficientNetV2 and VGG16.
- Compare stability of explanations and faithfulness (pixel flipping).

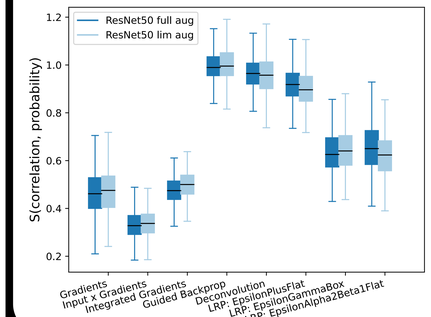


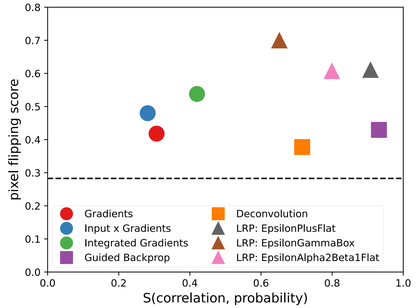
Results





	Brightness	Hue	Saturation	Rotate	Scale	Translate
Gradients	0.468	0.442	0.354	0.127	0.122	0.246
Input x Gradients	0.330	0.443	0.343	0.126	0.120	0.245
Integrated Gradients	0.478	0.636	0.546	0.209	0.229	0.327
Guided Backprop	**1.005**	1.028	0.994	**0.819**	**0.866**	**0.875**
Deconvolution	0.975	1.014	0.975	0.434	0.437	0.449
LRP: EpsilonPlusFlat	0.923	**1.053**	**1.038**	0.796	0.834	0.792
LRP: EpsilonGammaBox	0.632	0.856	0.832	0.480	0.512	0.532
LRP: EpsilonAlpha2Beta1Flat	0.662	1.006	0.972	0.691	0.722	0.706





Conclusion

- LRP composites and Guided Backprop produce are the most stable.
- Gradients and Input x Gradients are the least stable.
- Invariant (changes in brightness, hue and saturation) more stable than equivariant (rotation, translation, scaling).
- Training with data augmentation does not reduce this problem.

References

[1] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. Advances in Neural Information Processing Systems, 32, 2019.

[2] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In Proceedings of the AAAI conference on artificial intelligence, volume 33, pages 3681–3688, 2019.

Link to our paper →