2024 International Conference on Intelligent and Innovative Computing Applications (ICONIC)
7th-9th November 2024, Pearle Beach Resort & Spa, Flic en Flac, Mauritius
https://doi.org/10.59200/ICONIC.2024.006

Evaluating Concept Explanations for CNNs Under Adversarial Image Transformations

Ugochukwu Ejike Akpudo*, Yongsheng Gao*, Andrew Lewis*, Edwin Kwadwo Tenagyei*, Yi Liao* and Jun Zhou*

* Integrated and Intelligent Systems, Griffith University, Australia

{*ugochukwu.akpudo, edwinkwadwo.tenagyei, yi.liao2*}*@griffithuni.edu.au,*
{*yongsheng.gao, a.lewis, jun.zhou*}*@griffith.edu.au*

Abstract
Concept-based explainers for convolutional neural networks (CNNs) provide human-understandable explanations by revealing what the CNN sees, rather than merely indicating where it looked. However, their performance is limited by the *reducer* at its core and adversarial attacks. Although CNN classification performance may be enhanced by some image transformations in small amounts whereas intense image transformations can cause noticeable variations to CNN predictions, it is uncertain how explainers perform in such cases. This paper investigates the performance of state-of-the-art concept-based explainers at different levels of adversarial attacks for the first time. We achieve this by exploring different image transformations as adversarial attacks, including Gaussian noise, elastic transform, rotation, and contrast on the ILSVRC2012 dataset. Our study shows that image transformation techniques altering only image coordinates have little impact on classifier and explainer performance, whereas methods modifying image pixels, such as elastic transform and contrast, significantly affect performance, akin to introducing Gaussian noise. Our work underscores the significance of scrutinizing explainers during their development and adoption for CNNs.

Keywords: Convolutional neural networks, adversarial attacks, concept explanations, fidelity, image transformation

1. Introduction

Convolutional Neural Networks (CNNs) have revolutionized the field of computer vision, driving advancements in image classification, object detection, and other critical applications (Rathod et al., 2022). As these models are increasingly deployed in various domains, concerns about trust, accountability, regulatory compliance, transparency, safety, etc. underscore the importance of understanding their decision-making processes through qualitative and quantitative methods (Kim et al., 2018; Poppi et al., 2021). While attribution-based methods like Gradcam offer insights into where a CNN looked by highlighting the contributing pixels to a CNN's outcome, they require domain expert intervention and make them susceptible to bias. They also produce inaccurate pixel importance estimates in large CNN models (Chakraborty et al., 2022; Preechakul et al., 2022; Salahuddin et al., 2022). Contrary to attribution-based methods that merely highlight regions of interest, concept-based explainers provide insights into what a CNN saw, revealing the specific *concept*[1] that influences the CNN's decisions (Fel et al., 2023; Ghorbani et al., 2019; Kim et al., 2018; Zhang et al., 2021). However, their performance is often constrained by the type of *reducer*[2] used and their susceptibility to adversarial attacks, which can significantly undermine their reliability (Chakraborty et al., 2022).

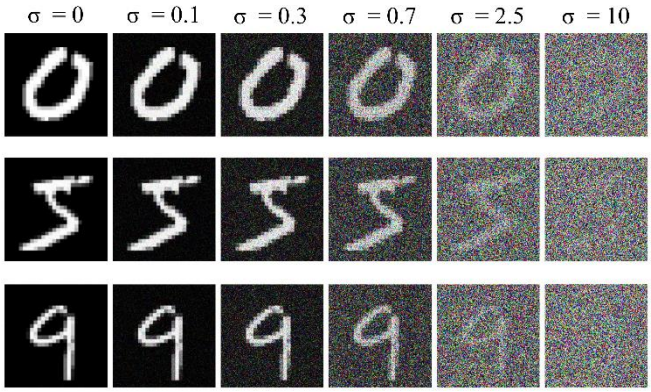

Figure 1. Impact of increasing the additive Gaussian noise levels to three dissimilar images from MNIST dataset (Xiao et al., 2017). The inter-class similarity increases as additive noise increases.

[1]A concept is defined as a high-level representation of a pattern or an abstract idea within an image class, e.g. "stripes".
[2]A reducer is a dimensionality reduction method used by a concept-based explainer for automatically discovering high-level concepts from a CNN's activation map.

Adversarial attacks pose a significant influence on the performance of CNNs (Chen et al., 2018; Xiang et al., 2021), however, to the best of the author's knowledge, the inherent impact of such adversarial attacks on explainers is understudied. While previous works suggest that introducing subtle perturbations can enhance CNN performance, extreme adversarial attacks may impede CNN performance. CNN performance can vary under different image transformations due to factors like transformation complexity, dataset size, and CNN architecture (Karimi et al., 2020). For instance, Figure 1 illustrates the impact of increasing the additive noise levels to three dissimilar images, causing the interclass variance to diminish and causing a reduced classifier performance. On a different note, there is a dearth of comprehensive studies on the robustness of concept-based explainers under such adversarial conditions. This gap in the literature is critical, as the effectiveness of explainers in adversarial scenarios directly impacts their practical utility and reliability, especially for cases where inter-class variance is low (Akpudo et al., 2023a, 2023b). Adversarial attacks undermine trust in AI systems and pose significant security risks in critical applications like autonomous vehicles or medical imaging. Explainers help in understanding why adversarial attacks succeed. If deep models can be easily fooled by imperceptible changes, users are less likely to trust their decisions, impacting adoption and deployment in real-world scenarios. By examining the model's decision-making process for adversarial attacks, users can gain insights into the vulnerabilities exploited by attackers. Conversely, adversarial attacks challenge the reliability of explanations provided by CNNs, as explanations derived from perturbed inputs may differ significantly from those derived from original inputs.

This study addresses this issue by investigating the performance of state-of-the-art concept-based explainers under varying levels of adversarial attacks and makes the following contributions:

- We demonstrate a comprehensive analysis of the impact of various adversarial image transformations on the performance of concept-based explainers. We provide empirical evaluation using the ILSVRC2012 dataset to quantify the effects of additive noise, rotation, and warping on both CNN classifiers and their explainers.
- We provide insights into the relative robustness of concept-based explainers under various adversarial conditions and guiding their development and deployment in real-world scenarios.

2. Background

2.1. Review of SOTA Concept-based Explainers

Earlier concept-based explanation methods require pre-defined concept banks, which are more challenging to learn than the target classes they aim to explain (Ramaswamy et al., 2023). This reliance limits their utility and the trust placed in them. Recent advances take a different approach: they integrate reducers into CNN architectures for automatic concept discovery without human supervision (Ghorbani et al., 2019; Nauta et al., 2023; Zhang et al., 2021). These advances have significantly improved CNN explanations, eliminating the need for pre-defined concept banks (Fel et al., 2023; Ghorbani et al., 2019; Zhang et al., 2021).

In a pioneering effort, Ghorbani et al. (Ghorbani et al., 2019) introduced the ACE framework, which involves segmenting class images into three levels, clustering similar segments based on Euclidean distance, rejecting outliers, and extracting important concepts. However, the outlier rejection phase may result in the loss of meaningful information, and the discovered concepts may assign different importance weights to instances of the same explanation case. More recent works like Zhang et al. (Zhang et al., 2021) and Fel et al. (Fel et al., 2023) both proposed the use of NMF as reducers in their ICE and CRAFT frameworks respectively. While the ICE framework offers significant performance, the CRAFT framework (Zhang et al., 2021) introduced recursivity into its concept decomposition process for producing enhanced concept explanations.

Other works (Brocki and Chung, 2019; N. Liu et al., 2023) have also been proposed with generative models at their core. Amidst their performance, generative models necessitate fine-tuning and further interpretation due to their black-box nature (Takeishi and Kawahara, 2020). Also, the perceptual similarity metrics derived from generative models often do not align with human perception (Zhang et al., 2016) and these contribute to even further interpretability issues beyond the CNN they aim to explain.

2.2. Automatic Concept Discovery

Performing classification with a CNN $E(\cdot): X \rightarrow Y$ involves a supervised learning process where images $(x_1, \ldots, x_n) \in X^{\mathrm{n}}$ are trained with their associating labels $(y_1, \ldots, y_n) \in Y^{\mathrm{n}}$ (He et al., 2016). Figure 2 illustrates the typical concept-based explanability framework for CNNs, revealing the key components of the explainer in the light blue box. $E(\cdot)$ can be split into two parts: the convolutional part $f(\cdot)$ for feature extraction and the linear classifier $C(\cdot)$ (with trainable weights t) for label predictions, such that $E(x_i) = (f \circ C)(x_i) f(x_i)$ produces the high-dimensional activation map $\mathcal{A}_l^{\mathrm{m\times c}} \equiv {h_l}^k(x_i) \subseteq \mathbb{R}^+$ at layer l of k layers ($\mathcal{A}_l^{\mathrm{m\times c}} \geq 0$) and $m = n \times h \times w$, where h, w are the feature map size, c is the number of channels, and n is the number of examples). $\mathcal{A}_l^{\mathrm{m\times c}}$ contains the most discriminative features from which concepts can be

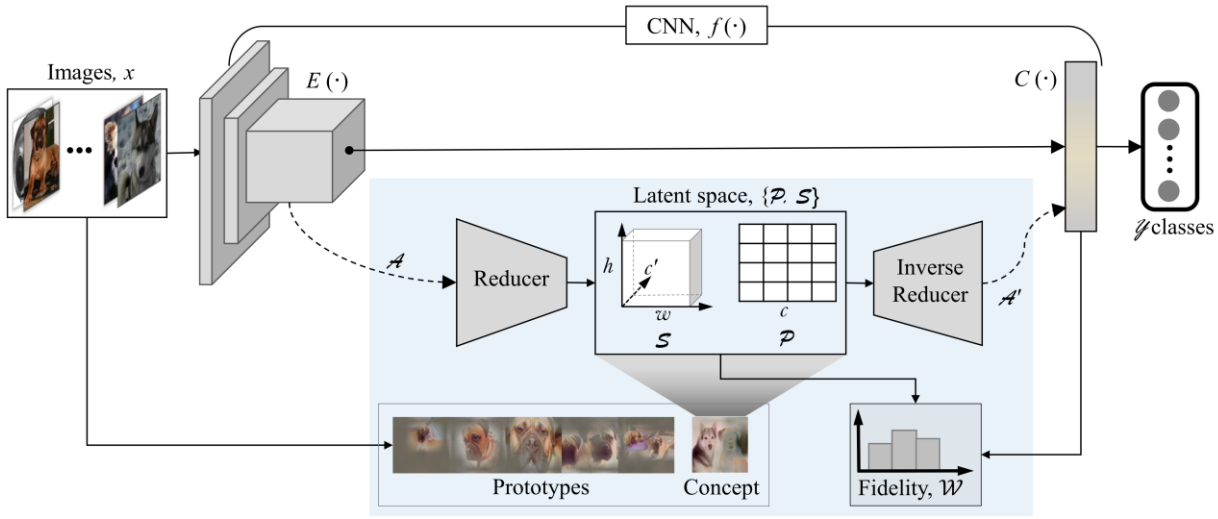

Figure 2. Illustration of a typical concept-based explanation for CNNs. The reducer generates the concepts and CAVs $\{\mathcal{S}, \mathcal{P}\}$ from the activation map ${\mathcal{A}_l}^{\mathrm{m}\times\mathrm{c}}$ produced by the concept extractor $f(\cdot)$. The inverse reducer helps compute the Fidelity $\mathcal{Z}$ of the explainer while the classifier helps compute the concept importance $\mathcal{W}y_i$.

Table 1. Adversarial attacks and their parameters.

Adversarial attack, $\mathcal{H}(x_i)$	**Parameters**
Gaussian noise	μ = 0; σ = {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}
Elastic transform	λ = 0.5, α = {20, 40, 60, 80, 100, 120, 140, 160, 180, 200}
Rotation	τ = {36, 72, 108, 144, 180, 216, 252, 288, 324, 360}
Contrast	β = {1.0, 1.2, 1.4, 1.6, 1.8, 2.0, 2.2, 2.4, 2.6, 2.8, 3.0}

discovered automatically using a concept-based explainer with a reducer N such as non-negative matrix factorization (NMF), principal component analysis (PCA), K-means, etc. at its core (S. Liu and Chen, 2021; Mendez, 2023).

A typical concept-based explainer receives $\{x_1, y_1\}$, and $E(\cdot)$ from which ${\mathcal{A}_l}^{\mathrm{m}\times\mathrm{c}}$ is produced. It then employs $\mathbb{N}$ to produce a lower dimensional representation $\mathcal{S}^{\mathrm{m}\times c'} \subseteq \mathbb{R}^+$ of ${\mathcal{A}_l}^{\mathrm{m}\times\mathrm{c}}$ and the concept activation vectors (CAVs) $P^{\mathrm{c}\times c'} \subseteq \mathbb{R}^+$, with a minimal loss u such that ${\mathcal{A}_l}^{\mathrm{m}\times\mathrm{c}} = \mathcal{S}^{\mathrm{m}\times c'}\mathcal{P}^{\mathrm{c}\times c'} + u$, such that $c' \ll c$ (c' = user-defined number of concepts). Given the complex structure of ${\mathcal{A}_l}^{\mathrm{m}\times\mathrm{c}}$, the main objective of N is to extract discriminative information stored in $= \mathcal{S}^{\mathrm{m}\times c'}$ and $\mathcal{P}^{\mathrm{c}\times c'}$ (Kim et al., 2016, 2018; Zhang et al., 2021).

2.3. Concept Importance Estimation

The rationale for concept-based explanation for CNNs is rooted in the need to make the decision-making process of CNNs more interpretable and understandable to humans. To achieve this, both qualitative and quantitative checkpoints are necessary. While qualitative explanations can be produced via prototypes, quantitative checkpoints provide further validations for the explainer's performance. The method testing with concept activation vectors (TCAV) method (Kim et al., 2018) provides a reliable approach for computing the concept importance $\mathcal{W}_{y_i} \subseteq \mathbb{R}^{c'}$ as the directional derivative of $C(\cdot)$ with respect to $\mathcal{P}$ in layer l, such that:

$$\frac{\partial C_{l,y}}{\mathcal{P}_l} = lim\,\frac{{h_l}^k({\mathcal{A}_l}^{\mathrm{m}\times\mathrm{c}} + \epsilon\mathcal{P}_l) - {h_l}^k({\mathcal{A}_l}^{\mathrm{m}\times\mathrm{c}})}{\epsilon} \quad (1)$$

where the estimated concept weight $\mathcal{W}_{y_i} \in \mathbb{R}^{c'\times 1} = \mathcal{P} \cdot t$ for CAV P for a target class y_i following a global average pooling of ${\mathcal{A}_l}^{\mathrm{m}\times\mathrm{c}}$.

2.4. Adversarial Attacks and Explainers

Adversarial attacks in intense amounts can significantly impact the reliability and trustworthiness of concept-based explainers for CNNs. Some of these impacts may include misleading explanations, reduced interpretability, erosion of trust, bias amplification, compromised explainer robustness, and uncertainty (Xiang et al., 2021). While some studies show that in small amounts of image transformations as adversarial

attacks such as additive noise, CNN classification performance may be enhanced and for certain types of image transformations, the CNN classification performance is barely affected (Chen et al., 2018). This is because image transformations that corrupt image pixels by directly altering the pixel values cause the images to deviate from looking like the classes they typically belong to.

In contrast, image transformations that only recalculate each pixel's coordinates affect the position of each pixel but do not directly alter the pixel values themselves. This boils down to saying that the performance of CNNs can vary under image transformations due to factors like transformation type, intensity of transformation, transformation complexity, dataset size, and CNN architecture (Karimi et al., 2020). It becomes imperative to study how explainers perform under these dynamics.

2.5. Performance Evaluation

We utilize the fidelity assessment method as outlined in (Zhang et al., 2021). Fidelity, in this context, is used as a measure of the explainer's faithfulness, aiming to quantitatively evaluate how closely the explainer's predictions align with those of the CNN model under different adversarial attacks $H(\cdot)$. This is achieved by calculating the average relative error between their predictions. Given E(·) and the approximate model $\hat{E}(\cdot) = C_l\left(\mathbb{N}'\left(\mathbb{N}\left(f_l(X)\right)\right)\right)$, Fidelity $\mathcal{Z}$ is measured as:

$$\mathcal{Z}(\hat{E}((x_i)) = \frac{\#\{x_i \in \mathrm{X} \mid E(\mathcal{H}(x_i)) = \hat{E}(\mathcal{H}(x_i))\}}{\#\{\mathrm{X}\}} \qquad (2)$$

3. Experiment

We utilize the ILSVRC2012 (ImageNet Large Scale Visual Recognition Challenge 2012) dataset, a benchmark dataset in computer vision consisting of over 1.2 million images across 1,000 object categories (ImageNet, 2012). The images were normalized and resized to 224 × 224 to ensure consistency and fairness in evaluating different image transformations. We also follow recommendations from (Ramaswamy et al., 2023) and choose $c' = 32$. Figure 3 shows the concept explanations for the classes Australian Kelpie and a Chihuahua without adversarial attacks. For readability, only the four most important concepts are displayed in each case. Overall, the discovered prototypes for each concept (represented by IDs) are representative of the image class being explained. For each concept explanation (row), the $\mathcal{W}$ scores and their contribution are also recorded. The $\mathcal{Z}$ reveal a high explainer faithfulness.

We explore the effects of different image transformations including Gaussian noise, rotation, and

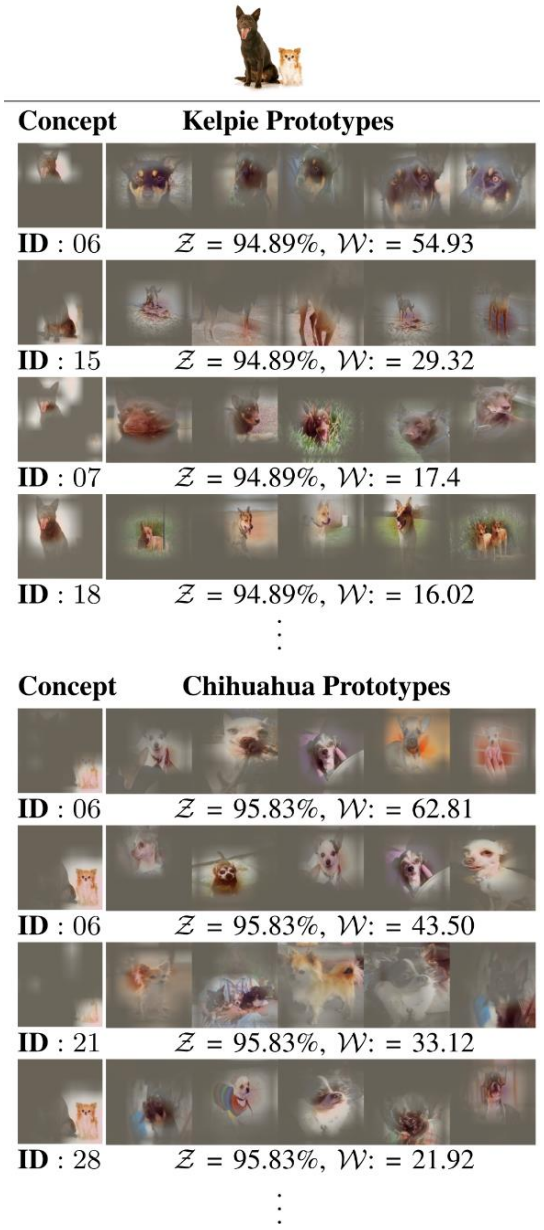

Figure 3. The four most important explanations produced by ICE (Zhang et al., 2021) explainer for Australian Kelpie (top) and Chihuahua (bottom) for a ResNet50 model.

elastic transform on the performance of ICE and CRAFT frameworks respectively. We tested them on a pre-trained ResNet50 model and then summarized the different adversarial attacks and the different ranges for each of them in Table 1. Figure 4 shows some of the concept explanations under the adversarial attacks while Figure 6 shows the overall performance of the explainer under the different adversarial attacks summarised in Table 1. We maintained $c' = 32$ and recorded the classification test accuracy of the ReNet50 model Acc, $\mathcal{W}$, and $\mathcal{Z}$ scores respectively.

As shown in Figures 4(a, b, d), similar concepts produce different prototypes with decreasing Acc, $\mathcal{W}$, and $\mathcal{Z}$, whereas the effect of rotation on the concept

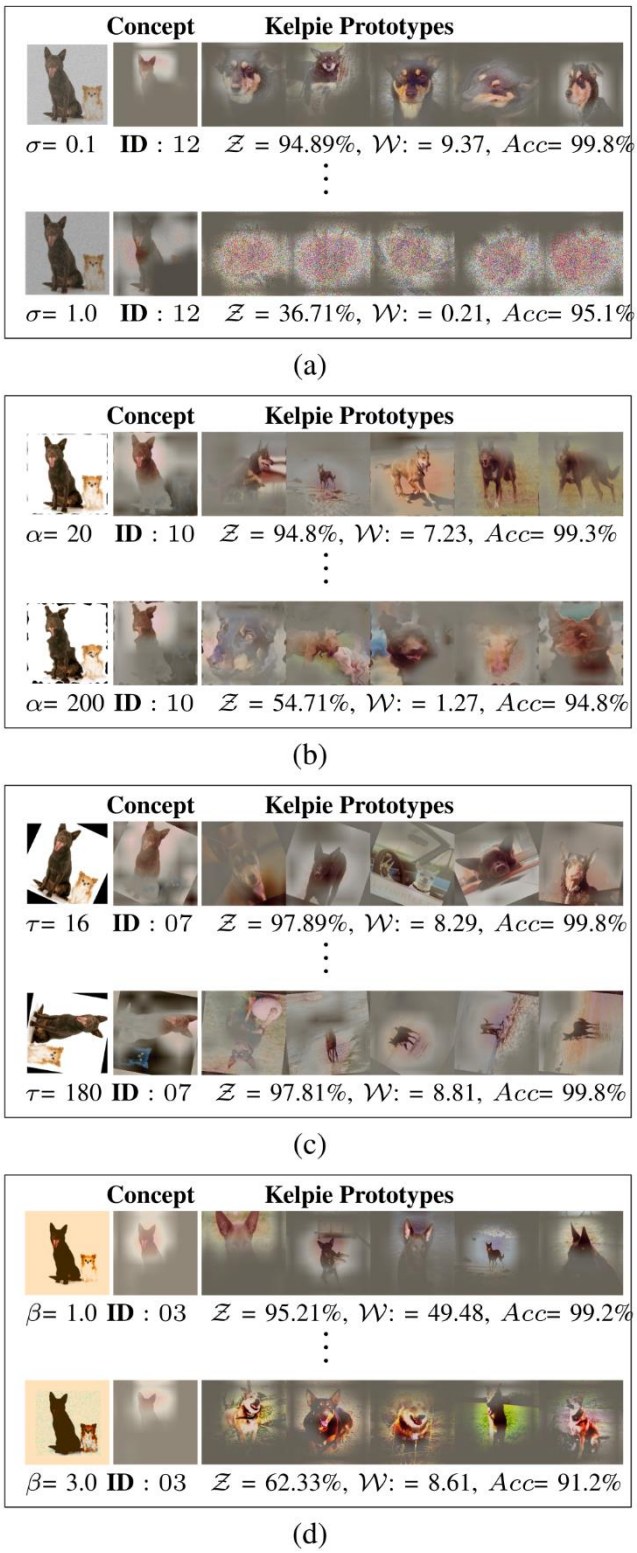

Figure 4. Australian kelpie's explanations by ICE (Zhang et al., 2021) explainer at different adversarial attacks (a) Gaussian noise, (b) elastic transform, (c) rotation, and (d) contrast.

explanations produced is unique to each case as shown in Figures 4(c), revealing a small impact of rotation on *Acc* and $\mathcal{Z}$ scores, amidst increasing

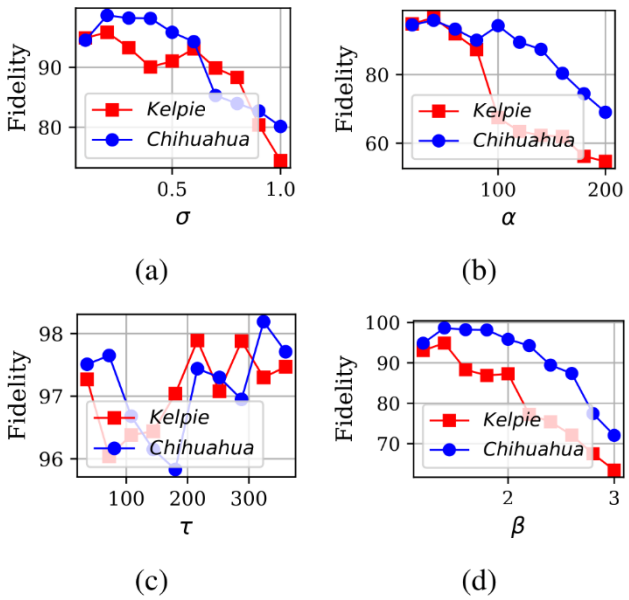

Figure 5. The impact of different adversarial attacks on ICE (Zhang et al., 2021) explainer's faithfulness for two dog breeds (a) Gaussian noise, (b) elastic transform, (c) rotation, and (d) contrast.

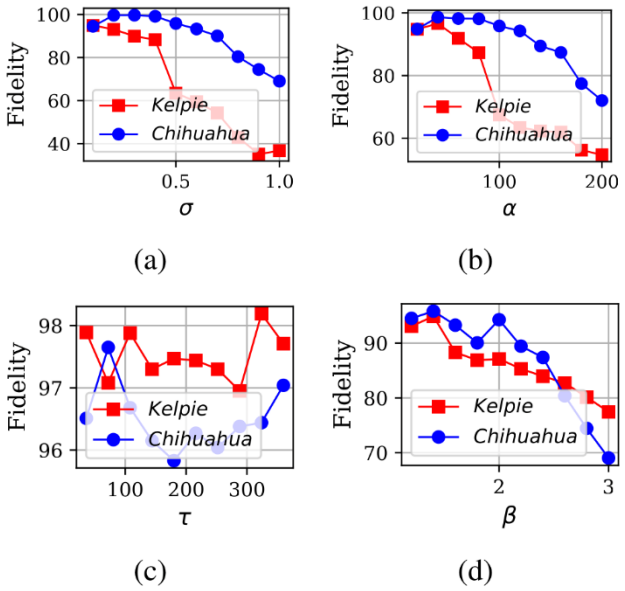

Figure 6. The impact of different adversarial attacks on CRAFT (Fel et al., 2023) explainer's faithfulness for two dog breeds (a) Gaussian noise, (b) elastic transform, (c) rotation, and (d) contrast.

degrees of rotation. The significant impact of noise, contrast and elastic transform on classifier (and explainer) performance is justified because such adversarial attacks directly alter the image pixel values while rotation recalculates each pixel's coordinates without altering them.

Figure 6 reveals that the explainer's faithfulness is affected by adversarial attacks that alter the pixel values

of images as shown in Figures 6(a, b, and d) but not affected by adversarial attacks that only affect the position of the image pixels as shown in Figure 6(c). These results highlight the necessity of thorough evaluation and scrutiny of explainers during their development to ensure their robustness against adversarial manipulations. By providing empirical evidence of the vulnerabilities in concept-based explainers, our work underscores the need for enhanced methodologies to safeguard their reliability in real-world applications.

4. Discussions, Drawn Insights and Future Works

In large amounts, image transformations that alter the pixel values introduce imperceptible changes to the images, leading to misclassification by a CNN model. Concept-based explainers might attribute the decision to features that are not semantically relevant but were perturbed. The downside is that explainability methods relying on feature importance or concept activation might incorrectly highlight non-relevant features or concepts due to adversarial perturbations (Akhtar, N., & Mian, A., 2018). Despite being in its early stages, our work offers insights into what a CNN sees and investigates such explanations by exploring the explainer's performance under different image perturbations as adversarial attacks.

Targeted attacks that aim to cause a specific misclassification by a CNN have a different effect. Explainability methods may provide explanations that reflect the targeted class rather than the true class of the adversarial example (Gilmer, J., et al., 2018). As a result, concept-based explainers might struggle to differentiate between features indicative of the true class and features highlighted by the adversarial attack. This is also the case for transferability attacks where adversarial examples crafted for one CNN model can fool other CNN models. Explainability methods might produce different explanations for the same adversarial example across different models (Dombrowski, J., et al., 2020). As a result, concept-based explainers may provide inconsistent or misleading explanations if the adversarial example is transferred to a different model architecture (Samek, W., et al., 2017).

The potential risks from adversarial attacks underscore the importance of more robust explainers for CNNs in real-world applications. Adversarial attacks undermine trust in AI systems and pose significant security risks in critical applications, impacting adoption and deployment in real-world scenarios. Therefore, developing more robust explainers for CNNs is crucial for their reliable deployment in real-world applications to ensure their resilience against adversarial attacks, robustness for interpreting CNN complexity, acceptable trade-off between interpretability and accuracy, maintain acceptable computational costs, maintain ethical and legal compliance, and provide human-centred explanations.

While these are yet to be extensively studied empirically, future works would aim at exploring in addition to and combination with image transformations, targeted and transferability attacks on concept-based explainers for CNNs. We also aim to explore paradigms towards standardizing concept-based CNN explanations.

5. Conclusion

This study investigates the performance and robustness of state-of-the-art concept-based explainers under various adversarial attacks, revealing the significant impact these perturbations have on both classifiers and their explainers. Using the ILSVRC2012 dataset, the experiments demonstrate that adversarial attacks like Gaussian noise, contrast, and elastic transformations can substantially degrade the fidelity of concept-based explainers, unlike adversarial attacks that only alter image pixel coordinates, such as rotation. Additionally, CNN classifiers and their explainers were found to be minimally affected by different transformations, with non-pixel-altering attacks having less impact on accuracy. The findings underscore the importance of rigorous adversarial testing of explainers to ensure reliability and practical utility, providing guidelines for deploying robust explainers and calling for future work on enhancing their resistance to attacks. This contributes to making deep learning models more transparent, trustworthy, and robust, addressing the demand for accountable AI systems.

6. References

Akpudo, U. E., Yu, X., Zhou, J., & Gao, Y. (2023a). Ncaf: Ntd-based concept activation factorisation framework for CNN explainability. 2023 38th International Conference on Image and Vision Computing New Zealand (IVCNZ), 1–6.

Akpudo, U. E., Yu, X., Zhou, J., & Gao, Y. (2023b). What exactly are we looking at?: Investigating for discriminance in ultra-fine-grained visual categorization tasks. 2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA), 129–136.

Akhtar, N., & Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access, 6, 14410-14430. https://doi.org/10.1109/ACCESS.2018.2803035

Brocki, L., & Chung, N. C. (2019). Concept saliency maps to visualize relevant features in deep generative models. 2019 18th IEEE International Conference On

Machine Learning And Applications (ICMLA), 1771–1778. https://doi.org/10.1109/ICMLA.2019.00287

Chakraborty, T., Trehan, U., Mallat, K., & Dugelay, J.-L. (2022). Generalizing adversarial explanations with grad-cam. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 187–193.

Chen, Y., Lyu, Z. X., Kang, X., & Wang, Z. J. (2018). A rotation-invariant convolutional neural network for image enhancement forensics. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2111–2115. https://doi.org/10.1109/ICASSP.2018.8462057

Dombrowski, J., et al. (2020). Adversarial machine learning in biomedical informatics: A survey. ACM Computing Surveys, 53(4), Article 79. https://doi.org/10.1145/3412912

Fel, T., Picard, A., Bethune, L., Boissin, T., Vigouroux, D., Colin, J., Cad`ene, R., & Serre, T. (2023). Craft: Concept recursive activation factorization for explainability. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2711–2721.

Ghorbani, A., Wexler, J., Zou, J. Y., & Kim, B. (2019). Towards automatic concept-based explanations. Advances in Neural Information Processing Systems, 32.

Gilmer, J., et al. (2018). Adversarial examples in the physical world. arXiv preprint arXiv:1802.08195. https://arxiv.org/abs/1802.08195

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778.

ImageNet Large Scale Visual Recognition Challenge (ILSVRC). (2012). ILSVRC2012. Retrieved July 7, 2024, from http://www.image-net.org/challenges/LSVRC/2012

Karimi, D., Dou, H., Warfield, S. K., & Gholipour, A. (2020). Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Medical Image Analysis, 65, 101759. https://doi.org/https://doi.org/10.1016/j.media.2020.101759

Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), Advances in neural information processing systems (Vol. 29). Curran Associates, Inc. https://proceedings.neurips.cc/paperfiles/paper/2016/file/5680522b8e2bb01943234bce7bf84534-Paper.pdf

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). International Conference on Machine Learning, 2668–2677.

Liu, N., Du, Y., Li, S., Tenenbaum, J. B., & Torralba, A. (2023). Unsupervised compositional concepts discovery with text-toimage generative models. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2085–2095.

Liu, S., & Chen, Y. (2021). Comparison of variant principal component analysis using new RNN-based framework for stock prediction. 2021 International Conference on Data Mining Workshops (ICDMW), 1047–1056. https://doi.org/10.1109/ICDMW53433.2021.00136

Mendez, M. A. (2023). Linear and nonlinear dimensionality reduction from fluid mechanics to machine learning. Measurement Science and Technology, 34(4), 042001.

Nauta, M., Schl¨otterer, J., van Keulen, M., & Seifert, C. (2023). Pip-net: Patch-based intuitive prototypes for interpretable image classification. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2744–2753.

Poppi, S., Cornia, M., Baraldi, L., & Cucchiara, R. (2021). Revisiting the evaluation of class activation mapping for explainability: A novel metric and experimental analysis. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2299–2304.

Preechakul, K., Sriswasdi, S., Kijsirikul, B., & Chuangsuwanich, E. (2022). Improved image classification explainability with high accuracy heatmaps. iScience, 25(3), 103933. https://doi.org/https://doi.org/10.1016/j.isci.2022.103933

Ramaswamy, V. V., Kim, S. S., Fong, R., & Russakovsky, O. (2023). Overlooked factors in concept-based explanations: Dataset choice, concept learnability, and human capability. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10932–10941.

Rathod, V. V., Rana, D. P., & Mehta, R. G. (2022). An extensive review of deep learning-driven remote sensing image classification models. 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT), 762–774. https://doi.org/10.1109/ICICICT54557.2022.9917583

Salahuddin, Z., Woodruff, H. C., Chatterjee, A., & Lambin, P. (2022). Transparency of deep neural networks for medical image analysis: A review of interpretability methods. Computers in Biology and Medicine, 140, 105111. https://doi.org/https://doi.org/10.1016/j.compbiomed.2021.105111

Samek, W., et al. (2017). Explainable AI: Interpreting, explaining and visualizing deep learning. Springer. https://doi.org/10.1007/978-3-319-64489-1

Takeishi, N., & Kawahara, Y. (2020). Knowledge-based regularization in generative modeling. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, 2390–2396.

Xiang, L., Zeng, X., Wu, S., Liu, Y., & Yuan, B. (2021). Computation of cnn's sensitivity to input perturbation. Neural Processing Letters, 53, 535–560.

Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashionmnist: A novel image dataset for benchmarking machine learning algorithms. CoRR, abs/1708.07747. http://arxiv.org/abs/1708.07747

Zhang, R., Isola, P., & Efros, A. A. (2016). Colorful image colorization. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14, 649–666.

Zhang, R., Madumal, P., Miller, T., Ehinger, K. A., & Rubinstein, B. I. (2021). Invertible concept-based explanations for cnn models with non-negative concept activation vectors. Proceedings of the AAAI Conference on Artificial Intelligence, 35(13), 11682–11690.