Harmonizing Feature Attributions Across Deep Learning Architectures: Enhancing Interpretability and Consistency

Md Abdul Kadir[1][0000−0002−8420−2536], GowthamKrishna Addluri[1][0009−0008−0513−9016], and Daniel Sonntag[1,2][0000−0002−8857−8709]

[1] German Research Center for Artificial Intelligence (DFKI), Germany
{abdul.kadir, Gowthamkrishna.Addluri, daniel.sonntag}@dfki.de
[2] University of Oldenburg, Oldenburg, Germany

Abstract. Enhancing the interpretability and consistency of machine learning models is critical to their deployment in real-world applications. Feature attribution methods have gained significant attention, which provide local explanations of model predictions by attributing importance to individual input features. This study examines the generalization of feature attributions across various deep learning architectures, such as convolutional neural networks (CNNs) and vision transformers. We aim to assess the feasibility of utilizing a feature attribution method as a future detector and examine how these features can be harmonized across multiple models employing distinct architectures but trained on the same data distribution. By exploring this harmonization, we aim to develop a more coherent and optimistic understanding of feature attributions, enhancing the consistency of local explanations across diverse deep-learning models. Our findings highlight the potential for harmonized feature attribution methods to improve interpretability and foster trust in machine learning applications, regardless of the underlying architecture.

Keywords: Explainability · Trustworthiness · XAI · Interpretability.

1 Introduction

Deep learning models have revolutionized various domains, but their complex nature often hampers our ability to understand their decision-making processes [1, 8]. Interpretability techniques have emerged, with local and global explanations being two significant categories [7]. Local explanations focus on understanding individual predictions, highlighting the most influential features for a specific instance. This method is valuable for understanding model behavior at a granular level and providing intuitive explanations for specific predictions. On the other hand, global explanations aim to capture overall model behavior and identify patterns and trends across the entire dataset. They offer a broader perspective and help uncover essential relationships between input features and model prediction. This paper delves into interpretability in deep learning models, particularly model-agnostic feature attribution, a subset of local explanation techniques.

Feature attribution refers to assigning importance or relevance to input features in a machine learning model's decision-making process [2]. It aims to understand which features have the most significant influence on the model's predictions or outputs. Feature attribution techniques provide insights into the relationship between input features and the model's decision, shedding light on the factors that drive specific outcomes. These techniques are precious for interpreting complex models like deep learning, where the learned representations may be abstract and difficult to interpret directly [18]. By quantifying the contribution of individual features, feature attribution allows us to identify the most influential factors, validate the model's behavior, detect biases, and gain a deeper understanding of the decision-making process.

Feature attribution methods can be evaluated through various approaches and metrics [18]. Qualitative evaluation involves visually inspecting the attributions and assessing their alignment with domain knowledge. Perturbation analysis tests the sensitivity of attributions to changes in input features [17]. Sanity checks ensure the reasonableness of attributions, especially in classification problems. From a human perspective, we identify objects in images by recognizing distinct features [14]. Similarly, deep learning models are trained to detect features from input data and make predictions based on these characteristics [19]. The primary objective of deep learning models, irrespective of the specific architecture, is to learn the underlying data distribution and capture unique identifying features for each class in the dataset.

Various deep learning architectures have proven proficient in capturing essential data characteristics within the training distribution [23]. We assume that if a set of features demonstrates discriminative qualities for one architecture, it should likewise exhibit discriminative properties for a different architecture, provided both architectures are trained on the same data. This assumption forms the foundation for the consistency and transferability of feature attributions across various deep learning architectures.

Our experiments aim to explore the generalizability of features selected by a feature attribution method for one deep learning architecture compared to other architectures trained on the same data distribution. We refer to this process as harmonizing feature attributions across different architectures. Our experimental results also support our assumption and indicate that different architectures trained on the same data have a joint feature identification capability.

2   Related work

Various explanation algorithms have been developed better to understand the internal mechanisms of deep learning models. These algorithms, such as feature attribution maps, have gained significant popularity in deep learning research. They offer valuable insights into the rationale behind specific predictions made by deep learning models [4]. Notable examples of these explanation methods include layer-wise relevance propagation [12], Grad-CAM [20], integrated gradient

[22], guided back-propagation [21], pixel-wise decomposition [3], and contrastive explanations [11].

Various methods have been developed to evaluate feature attribution maps. Ground truth data, such as object-localization or masks, has been used for evaluation [5, 15]. Another approach focuses on the faithfulness of explanations, measuring how well they reflect the model's attention [17]. The IROF technique divides images into segments and evaluates explanations based on segment relevance [16]. Pixel-wise evaluations involve flipping pixels or assessing attribution quality using pixel-based metrics [3, 17].

Chen et al. [6] has demonstrated the utility of feature attribution methods for feature selection. Additionally, research conducted by Morcos et al. [13] and Kornblith et al. [10] has explored the internal representation similarity between different architectures. However, to the best of our knowledge, the generalization of feature attributions across diverse neural architectures still needs to be explored.

Motivated by the goal of evaluating feature attributions, we are investigating a novel approach that involves assessing feature attributions across multiple models belonging to different architectural designs. This method aims to provide a more comprehensive understanding of feature attribution in various contexts, thereby enhancing the overall explainability of deep neural networks.

3   Methodology

This experiment investigates the generalizability and transferability of feature attributions across different deep learning architectures trained on the same data distribution. The experimental process involves generating feature attribution maps for a pretrained model, extracting features from input images, and passing them to two models with distinct architectures. The accuracy and output probability distribution are then calculated for each architecture.

In this experiment, we employ a modified version of the Soundness Saliency (SS) method [9] for generating explanations. The primary objective with a network f, for a specific input x (Fig. 1 (a)), and label a, is to acquire a map or mask $M \in \{0,1\}^{hw}$.

This map aims to minimize the expectation $E_{\tilde{x} \sim (x,M)}[-\sum f_i(\tilde{x}) log(f_i(\tilde{x}))]$, wherein the probability assigned by the network to a modified or composite input x is maximized.

$$\tilde{x} \sim \Gamma(x,a) \equiv \overline{x} \sim \mathcal{X}, \tilde{x} = M \odot x + (1-M) \odot \overline{x} \tag{1}$$

Here, M (Fig. 1 (b)) represents the feature attribution map generated by the Soundness Saliency algorithm. The saliency map M provides information about the importance of each pixel and the extent of its contribution to the classification. If the value of M (Fig. 1 (b)) for a specific pixel is 0, it implies that the pixel has no significance in the classification process. Conversely, if the value of M for a particular pixel is high, it indicates that the pixel is highly important for the classification. We enhance the extraction of important features (Fig 1 (c))

from input by applying the Hadamard product between each input channel and the corresponding attribution map M. In addition to the Soundness Saliency (SS) algorithm, we also employ Grad-CAM [20] (GC) for feature extraction. We

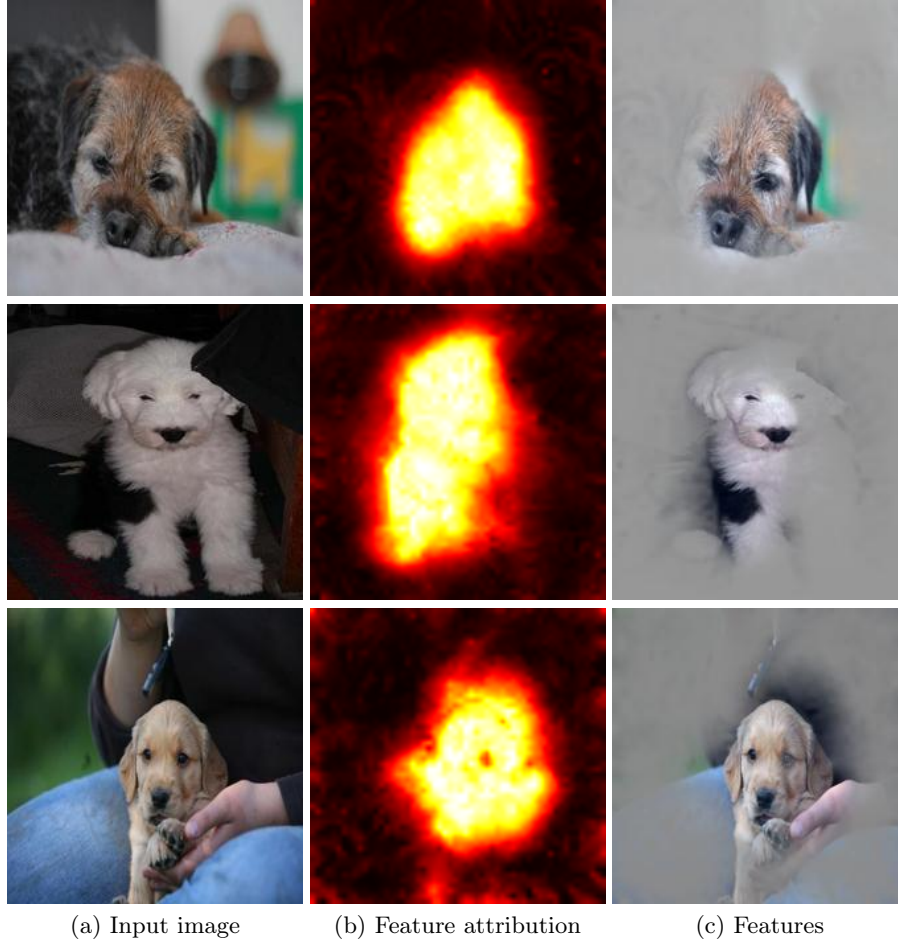(a) Input image   (b) Feature attribution   (c) Features

Fig. 1. Columns a, b, and c represent the input image, the feature attribution map generated by the soundness saliency algorithm, and the extracted features of the image based on the feature attribution map, respectively.

utilize the selected features Fig. 1 (c) extracted through feature attributions and feed them to two distinct models with different architectures, albeit trained on the same training data. Accuracy, F1 score, and output probability scores are calculated for these models. The focus is observing model prediction changes when only the selected features are inputted rather than the entire image.

4   Experiment and Results

In this study, we selected four distinct pretrained architectures: the Vision Transformer architecture (ViT) [8], EfficientNet-B7 (E-7) [23], EfficientNet-B6 (E-6) [23], and EfficientNet-B5 (E-5) [23]. To generate feature attribution maps, we first employed E-7 along with a challenging subset[3] of the ImageNet validation data, which is known to be particularly difficult for classifiers.

Subsequently, we generated feature maps for all test data and passed them to E-6 and E-5. In parallel, we also generated features for ViT and followed the same procedure. We chose to utilize both a transformer and a CNN architecture in our experiments because they are fundamentally different from one another, allowing for a comprehensive evaluation of the various architectures.

Table 1. The SS row examines model performance with image (I) and feature (F) inputs generated by the E-7 architecture and SS algorithm. It shows stable accuracy and F1 scores across the E-6 and E-5 architectures when using feature inputs. In contrast, the GC row, which uses the Grad-CAM algorithm for feature generation, demonstrates a drop in accuracy and F1 scores across the E-6, and E-5 architectures when features are used as input.

Exp.	Metric	E-7 (I)	E-7 (F)	E6 (I)	E-6 (F)	E-5 (I)	E-5 (F)
SS	Acc	78.4%	74.09%	75.90%	73.61%	77.27%	73.76%
	F1	0.87	0.84	0.85	0.84	0.86	0.84
GM	Accuracy	78.47%	58.87%	75.90%	55.74%	77.27%	57.24%
	F1	0.87	0.72	0.85	0.70	0.86	0.71

Table 2. The SS row evaluates model performance using image (I) and feature (F) inputs, generated through the ViT architecture and the SS algorithm. There's a slight decrease in accuracy and F1 score across architectures (E-6, E-5) with feature inputs. Similarly, the GC row, utilizing the Grad-CAM algorithm for feature generation, shows a comparable drop in accuracy and F1 scores across the E-6 and E-5 architectures when using feature inputs.

Exp.	Metric	ViT (I)	ViT (F)	E6 (I)	E-6 (F)	E-5 (I)	E-5 (F)
SS	Acc	89.92%	88.83%	75.90%	71.90%	77.27%	73.10%
	F1	0.94	0.94	0.85	0.83	0.86	0.83
GM	Ac	89.92%	58.56%	75.90%	44.21%	77.27%	43.47%
	F1	0.87	0.73	0.85	0.60	0.86	0.58

---

[3] https://github.com/fastai/imagenette

Our experimental results (Tables 1 and 2) indicate that features generated by a neural architecture can be detected by other architectures trained on the same data. This implies that feature attribution maps encapsulate sufficient data distribution information. Consequently, feature maps created using attribution maps on one architecture can be recognized by another architecture, provided that both are trained on the same data. As depicted in Fig. 2, when we feed only features to the model, the class probability increases (Fig. 2 (b), (d), and (f)), particularly when using similar architectures for feature generation and evaluation. When employing different types of architectures (e.g., Transformer for generating feature maps and CNN for evaluating them), there is a slight drop in accuracy (Fig. 2 (j) and (l)), but the performance remains consistent.

Accuracy decreases when features are extracted with Grad-CAM saliency maps, suggesting these maps might not capture crucial information on the data distribution. However, when examining row GC in Tables 1 and 2, it's observed that accuracy remains consistent across various architectural configurations when features are generated using Grad-CAM. This suggests that different architectures have harmony in detecting certain features from data.
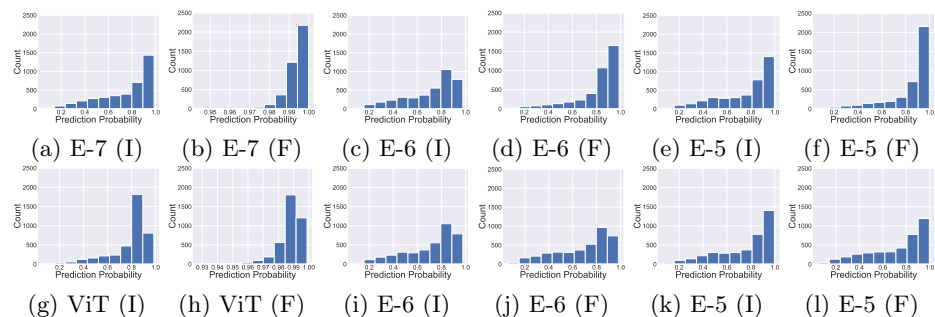

Fig. 2. Histogram of class prediction probability distribution across different architectures, comparing image-only (I) and feature-only (F) inputs for the entire test dataset.

5   Conclusion

The experiment validates our hypothesis that various architectures acquire shared features from a common data distribution. We noticed a notable rise in class prediction probabilities when utilizing selected features as inputs, particularly when employing similar neural architecture building blocks such as Convolution. Additionally, the consistency of predictions on future attribution maps across architectures demonstrates that different architectures are not randomly learning features from the data, thereby enhancing the reliability of the models. These findings underscore the potential to generalize features and emphasize the need for additional research to harmonize feature attribution maps, expanding their applicability in various domains.

References

1. Alirezazadeh, P., Schirrmann, M., Stolzenburg, F.: Improving deep learning-based plant disease classification with attention mechanism. Gesunde Pflanzen **75**(1), 49–59 (Dec 2022), https://doi.org/10.1007/s10343-022-00796-y
2. Ancona, M., Ceolini, E., Öztireli, A.C., Gross, M.H.: A unified view of gradient-based attribution methods for deep neural networks. CoRR **abs/1711.06104** (2017), https://doi.org/10.48550/arXiv.1711.06104
3. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLOS ONE **10**(7), e0130140 (Jul 2015), https://doi.org/10.1371/journal.pone.0130140, publisher: Public Library of Science
4. Burnett, M.: Explaining ai: Fairly? well? In: Proceedings of the 25th International Conference on Intelligent User Interfaces, p. 1–2, IUI '20, New York, NY, USA (2020), https://doi.org/10.1145/3377325.3380623
5. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 839–847, Lake Tahoe, Nevada (2018), https://doi.org/10.1109/WACV.2018.00097
6. Chen, J., Yuan, S., Lv, D., Xiang, Y.: A novel self-learning feature selection approach based on feature attributions. Expert Systems with Applications **183**, 115219 (2021), https://doi.org/10.1016/j.eswa.2021.115219
7. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning (2017), https://doi.org/10.48550/arXiv.1702.08608
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. CoRR **abs/2010.11929** (2020), https://doi.org/10.48550/arXiv.2010.11929
9. Gupta, A., Saunshi, N., Yu, D., Lyu, K., Arora, S.: New definitions and evaluations for saliency methods: Staying intrinsic, complete and sound. ArXiv **abs/2211.02912** (2022), https://doi.org/10.48550/arXiv.2211.02912
10. Kornblith, S., Norouzi, M., Lee, H., Hinton, G.E.: Similarity of neural network representations revisited. CoRR **abs/1905.00414** (2019), https://doi.org/10.48550/arXiv.1905.00414
11. Luss, R., Chen, P.Y., Dhurandhar, A., Sattigeri, P., Zhang, Y., Shanmugam, K., et al.: Leveraging Latent Features for Local Explanations. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 1139–1149, KDD '21, New York, NY, USA (Aug 2021), https://doi.org/10.1145/3447548.3467265
12. Montavon, G., Samek, W., Müller, K.R.: Methods for interpreting and understanding deep neural networks. Digital Signal Processing **73**, 1–15 (Feb 2018), https://doi.org/10.1016/j.dsp.2017.10.011

13. Morcos, A.S., Raghu, M., Bengio, S.: Insights on representational similarity in neural networks with canonical correlation. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems, p. 5732–5741, NIPS'18, Red Hook, NY, USA (2018), https://doi.org/10.48550/arXiv.1806.05759
14. Mozer, M.: Object recognition: Theories. In: International Encyclopedia of the Social & Behavioral Sciences, pp. 10781–10785, Pergamon, Oxford (2001), https://doi.org/10.1016/B0-08-043076-7/01459-5
15. Nunnari, F., Kadir, M.A., Sonntag, D.: On the overlap between grad-cam saliency maps and explainable visual features in skin cancer images. In: Machine Learning and Knowledge Extraction, pp. 241–253, Cham (2021), https://doi.org/10.1007/978-3-030-84060-0_16
16. Rieger, L., Hansen, L.K.: IROF: a low resource evaluation metric for explanation methods (Mar 2020), https://doi.org/10.48550/arXiv.2003.08747, arXiv:2003.08747 [cs]
17. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.R.: Evaluating the Visualization of What a Deep Neural Network Has Learned. IEEE Transactions on Neural Networks and Learning Systems **28**(11), 2660–2673 (Nov 2017), https://doi.org/10.1109/TNNLS.2016.2599820, conference Name: IEEE Transactions on Neural Networks and Learning Systems
18. Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., Müller, K.R.: Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE **109**(3), 247–278 (2021), https://doi.org/10.1109/JPROC.2021.3060483
19. Schulz, H., Behnke, S.: Deep learning. KI - Künstliche Intelligenz **26**(4), 357–363 (May 2012), https://doi.org/10.1007/s13218-012-0198-z
20. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626, Venice, Italy (2017), https://doi.org/10.1109/ICCV.2017.74
21. Springenberg, J., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. In: ICLR (workshop track), p. 10, San Diego, California (2015), https://doi.org/10.48550/arXiv.1412.6806
22. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70, p. 3319–3328, ICML'17, Sydney, Australia (2017), https://doi.org/10.48550/arXiv.1703.01365
23. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. CoRR **abs/1905.11946** (2019), https://doi.org/10.48550/arXiv.1905.11946