Towards Explainable Image Classification

Vahidin Hasić[1]

[1]University of Sarajevo, 71210 Sarajevo, Bosnia and Herzegovina

Abstract

While Deep Neural Networks (DNNs) excel in image classification, their black-box nature necessitates the development of Explainable AI (XAI) methods. Existing XAI techniques often face limitations in balancing explainability, fidelity, and efficiency. My doctoral research addresses these limitations through an evolving series of investigations. Initially, I focused on improving the gradient-based explanations. This research led me to explorations of concept-based explanations. Currently, I am investigating sample-based explanations to attribute the importance of training samples. These seemingly disparate lines of research are connected by a common thread: the pursuit of XAI methods that are faithful to the model, understandable to humans, and computationally efficient for real-time applications.

Keywords

Explainable AI, Image Classification, Convolutional Neural Networks

1. Introduction

Deep neural networks (DNNs) revolutionize how we approach complex tasks and achieve exceptional performance in areas like image classification. Their success comes from their ability to learn complex patterns from a large amount of data. However, this performance comes at the cost of interpretability. DNNs are often described as black boxes due to their opaque internal workings, making it difficult for humans to understand the reasoning behind their predictions [1].

Consequently, the research on Explainable AI (XAI) has gained significant momentum, with the aim of making DNNs more transparent and human-understandable [2]. XAI techniques seek to provide insight into DNN decision-making processes, allowing users to understand their output and increase trust in the system. However, existing XAI methods often face trade-offs between explainability, faithfulness, and efficiency [3].

A common approach for XAI computer vision techniques is to attribute importance scores to pixels or image patches [4, 5]. However, such pixel-level explanations can be overwhelming and difficult for humans to interpret, as they lack semantic meaning and do not capture higher-level concepts relevant to the task [6, 7]. Furthermore, the faithfulness of these methods can be questionable, as they may lack sensitivity to the model and the data generating process [8]. Additionally, some methods, such as SHAP [9], can be computationally too expensive for practical application.

This research aims to develop novel explainability methods that are faithful to the model, understandable to humans, and computationally efficient for real-time applications. The main research questions investigated are:

- Can we develop explainability methods that simultaneously achieve model fidelity, human interpretability, and computational efficiency?
- What explanation type do users prefer based on their expertise?
- Are explanations at higher or lower levels of abstraction in image classification more effective for users?
- Can different abstraction levels in image classification explanation be integrated into a framework?

Late-breaking work, Demos and Doctoral Consortium, colocated with The 3rd World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey

vahidin.hasic@etf.unsa.ba (V. Hasić)
0009-0003-6306-8037 (V. Hasić)

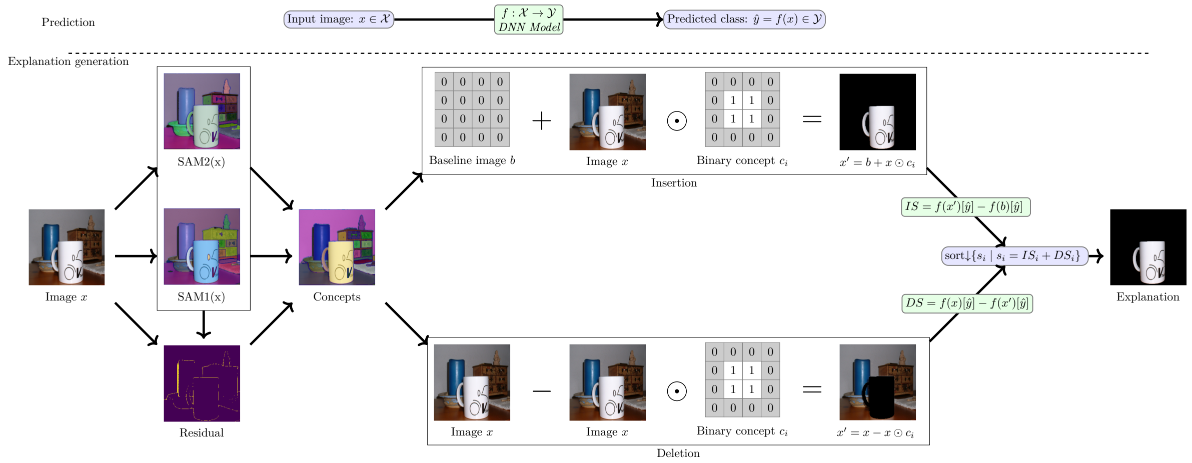


Figure 1: Any Segment Explanation (ASE) overview. An input image x is classified (f), yielding prediction $\hat{y}$. x is segmented (SAM1 [23]/SAM2 [24]; residual is unsegmented areas). Segments are treated as concepts and are transformed into binary masks. Perturbed images (x') are generated by inserting/deleting concepts. Insertion/Deletion Scores (IS/DS) measure model prediction change. The concepts are ranked by combining IS and DS scores, where the highest-scoring concept is shown as the explanation.

2. Related Work

A common approach for XAI computer vision techniques is to attribute importance scores to pixels or image patches [4]. However, such pixel-level explanations can be overwhelming and difficult for humans to interpret, as they lack semantic meaning and do not capture higher-level concepts relevant to the task [6, 7]. Concept-based and prototype-based explanations are promising alternatives [10]. Concept-based explanations are more closely aligned with human reasoning and how humans explain decisions [11]; they help identify biases and improve classification performance [12], are more stable against perturbations and more robust against adversarial attacks than traditional XAI methods [12] However, existing concept-based methods suffer from limitations such as task specificity, reliance on manual annotation of concepts, and limited automatic concept discovery. The Explain Any Concept (EAC) method [13] addresses some of these challenges using the Segment Anything Model (SAM1) for automated concept extraction. EAC assigns importance scores to SAM-generated image segments using Monte Carlo SHAP, enabling concept-level explanations without manual annotation. However, EAC's reliance on a surrogate linear model to approximate the target DNN and the computational expense of SHAP limits EAC's practical applicability.

Most of the XAI research is centered on determining the most influential input features, often referred to as feature importance [14, 9]. An alternative approach to enhancing model transparency is quantifying individual training instances' influence on the model's predictions, known as sample-based explanations (SBE) [15]. Current state-of-the-art methods for sample-based explanations are generally categorized into retraining-based and gradient-based approaches [16]. Retraining-based methods operate on the principle that a training sample's importance can be quantified by measuring its removal's impact on the model's performance after retraining [17]. Several notable works have developed methods based on this approach [18, 19]. While this approach is simple and human-understandable, its primary limitation is computational complexity. Gradient-based methods attribute training sample importance by calculating gradients over model parameters and evaluating the similarity between the gradients [20]. A fundamental limitation of gradient-based methods is the computational burden that computing the inverse Hessian poses [21, 22].

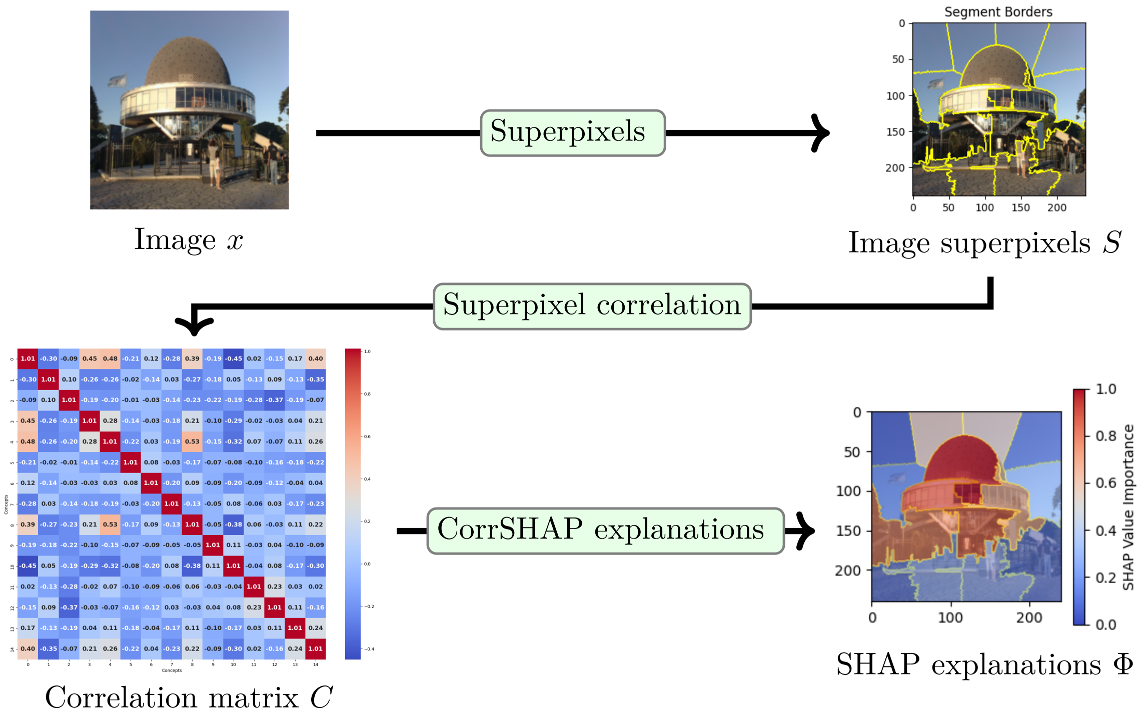


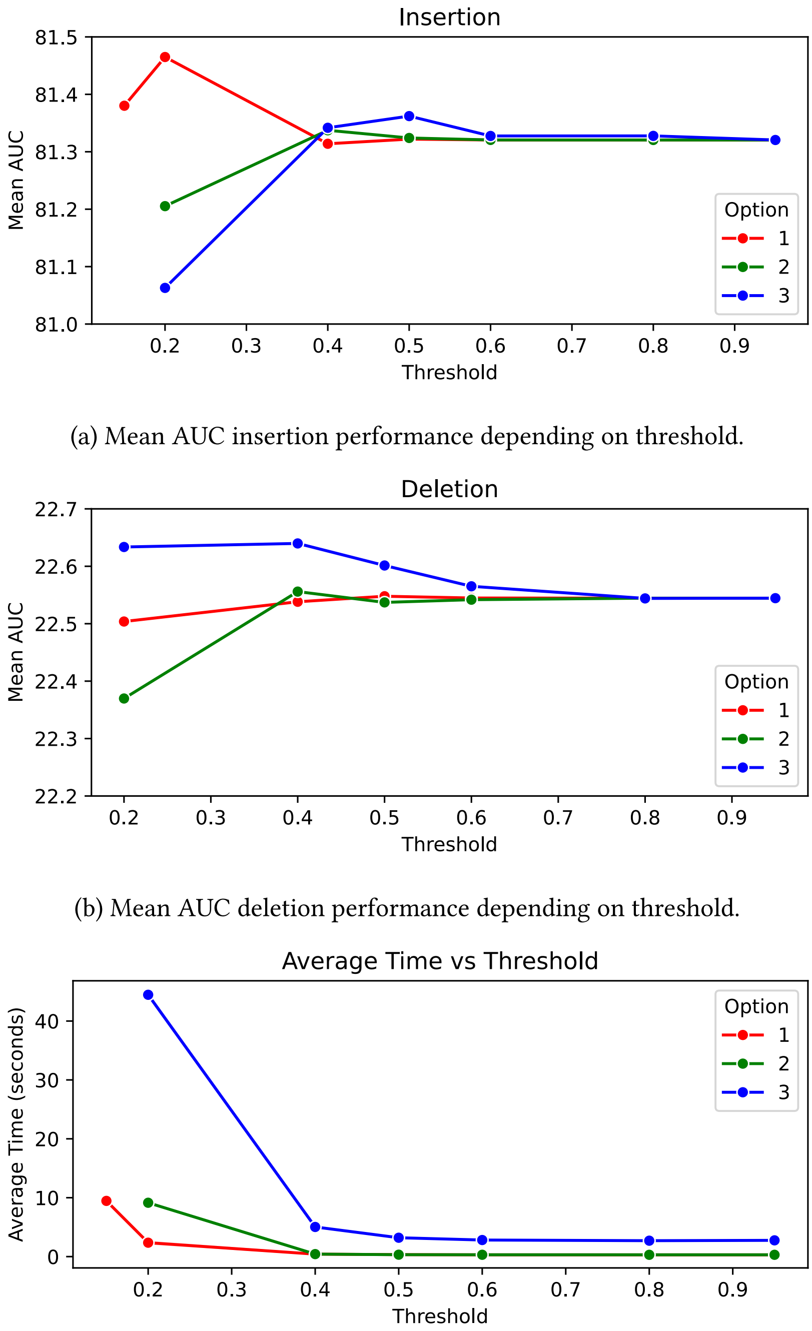
Figure 2: Framework of the proposed CorrSHAP method. Input image x is segmented into superpixels S. Superpixels are vectorized and centralized into vectors V. Correlation matrix C between superpixels is calculated using cosine similarity between individual vectors. For each superpixel, we take correlated superpixels, where correlation is higher than threshold $|C_{ij}| > \tau$, and perform perturbations on all combinations to calculate superpixel attribution.

3. Methodology

Building upon EAC Any Segment Explanations (ASE), an improved local, post-hoc, and model-agnostic explanation method was proposed. ASE overcomes EAC's limitations of model approximation and high computational cost while achieving superior model faithfulness. ASE employs state-of-the-art image segmentation algorithms Segment Anything Model 2 (SAM2) in combination with Segment Anything Model 1 (SAM1) and residual segment for concept extraction, enabling broader and more relevant concept capture. ASE employs concept insertion and deletion techniques to determine concept attributions, avoiding the need for surrogate models and the associated inaccuracies. The framework of the proposed ASE method is shown in Figure 1. This paper is currently under review.

While ASE showed good performance, it does not consider the interdependence of visual concepts. To address this limitation, a novel method that leverages the correlations between image concepts to accelerate SHAP attribution calculation Correlation SHAP (CorrSHAP) was proposed. CorrSHAP transforms image superpixels into centralized vector representations and employs the modified Pearson correlation approach to quantify superpixel relationships (Figure 2). By leveraging the concept correlation, CorrSHAP dramatically reduces the number of feature subsets that need to be evaluated for accurate SHAP value estimation, resulting in substantial computational savings. This paper has been accepted to the XAI 2025 conference.

Current state-of-the-art methods for estimating training data attribution are highly computationally expensive and have problems with scaling up. To address these limitations, a novel black-box approach leveraging kernel functions was proposed. It achieves better model faithfulness while being much faster than competing methods. This paper is under review for the ICCV conference.



(a) Mean AUC insertion performance depending on threshold.

(b) Mean AUC deletion performance depending on threshold.

(c) Execution time depending on threshold.

Figure 3: Performance analysis as a function of the threshold τ. Lower AUC deletion, higher AUC insertion scores, and lower execution time indicate superior performance.

4. Results

Table 1 shows comparative results of the proposed ASE with the state-of-the-art methods EAC [13], DeepLift [25], GradSHAP [14], IntGrad [26], KernelSHAP [9], FeatAbl [27], and LIME [28]. ASE consistently outperforms competing methods, demonstrating substantial improvements in model faithfulness. Beyond faithfulness, ASE offers significant computational advantage, averaging **1.568 seconds** per image explanation, ASE is **38.29x faster** than EAC, which requires **60.039 seconds** per explanation.

Figure 3 shows the performance of all three proposed approaches for CorrSHAP. These results demon-

		ASE(ours)	**EAC***	**DeepLIFT***	**GradSHAP***	**IntGrad***	**KernelSHAP***	**FeatAbl***	**LIME***
Insertion ↑	ResNet-50	**91.10**	83.400	75.235	64.658	68.772	64.544	70.187	76.638
	MobileNet-v2	**90.67**	74.651	34.197	47.848	48.662	60.837	59.197	61.282
	ViT-b16	**89.86**	89.594	54.455	68.125	69.480	75.152	65.656	76.161
	ResNet-18	**78.03**	73.558	47.799	38.877	36.806	50.547	43.448	50.592
Deletion ↓	ResNet-50	**8.99**	23.799	25.262	40.996	36.214	26.583	37.332	25.307
	MobileNetv2	6.61	**6.002**	26.381	14.679	13.382	7.766	8.866	7.344
	ViT-b16	**6.24**	17.298	40.784	30.948	29.903	21.825	34.191	19.254
	ResNet-18	**2.57**	6.596	8.588	11.273	11.555	6.638	8.352	6.776

Table 1

Mean AUC over 10000 random ImageNet-1k images. * indicates results reproduced from [13] using the same seed for random image sampling for direct comparison. Higher AUC (insertion) and lower AUC (Deletion) are better. ASE outperforms the other methods for all the compared models and for both evaluation circumstances.

Area Under the Curve (AUC) Insertion ↑					
Model	Superpixels	CorrSHAP 1	CorrSHAP 2	CorrSHAP 3	MCSHAP
MobileNet-v2	Quickshift	80.4	80.37	80.29	**80.89**
	SLIC	**78.12**	78.13	78.15	77.79
ResNet-18	Quickshift	60.21	60.23	60.21	**61.60**
	SLIC	54.66	54.65	54.65	**55.41**
ResNet-50	Quickshift	**82.63**	82.61	82.63	82.27
	SLIC	**81.20**	81.20	81.20	80.66
ViT-b16	Quickshift	80.83	80.83	80.74	**81.84**
	SLIC	76.36	76.33	76.41	**76.82**
Area Under the Curve (AUC) Deletion ↓					
MobileNet-v2	Quickshift	**20.14**	20.14	20.16	20.93
	SLIC	**19.48**	19.48	19.48	21.40
ResNet-18	Quickshift	8.25	8.25	8.25	**8.03**
	SLIC	9.06	9.06	9.06	**9.01**
ResNet-50	Quickshift	**22.79**	22.79	22.79	24.16
	SLIC	**22.36**	22.36	22.39	23.79
ViT-b16	Quickshift	17.20	17.17	17.13	**16.86**
	SLIC	20.53	20.53	20.53	**20.48**
Execution Time (seconds) ↓					
MobileNet-v2	Quickshift	**0.42**	0.54	1.15	16.13
	SLIC	**0.74**	0.89	1.86	14.98
ResNet-18	Quickshift	**0.43**	0.50	1.09	7.85
	SLIC	0.66	**0.51**	1.91	13.82
ResNet-50	Quickshift	**0.46**	0.58	2.76	25.16
	SLIC	**0.78**	0.93	5.93	36.49
ViT-b16	Quickshift	**0.40**	0.52	5.41	7.26
	SLIC	**0.73**	1.01	10.53	18.00

Table 2

Quantitative evaluation of explanation methods using metrics AUC Insertion, AUC Deletion, and execution time. Methods (CorrSHAP versions 1-3 and MCSHAP) are compared across architectures (MobileNet-v2, ResNet-18, ResNet-50, ViT-b16) and superpixel algorithms (Quickshift and SLIC).

strate that CorrSHAP substantially reduces execution time while maintaining explanation faithfulness (Table 2). This outcome indicates the effectiveness of our correlation method in accurately assigning correlations to superpixels. Consequently, we can restrict the computation of superpixel attributions for explanations to a smaller subset of correlated superpixels. Qualitative comparison of CorrSHAP to Monte Carlo SHAP is shown in Figure 4.

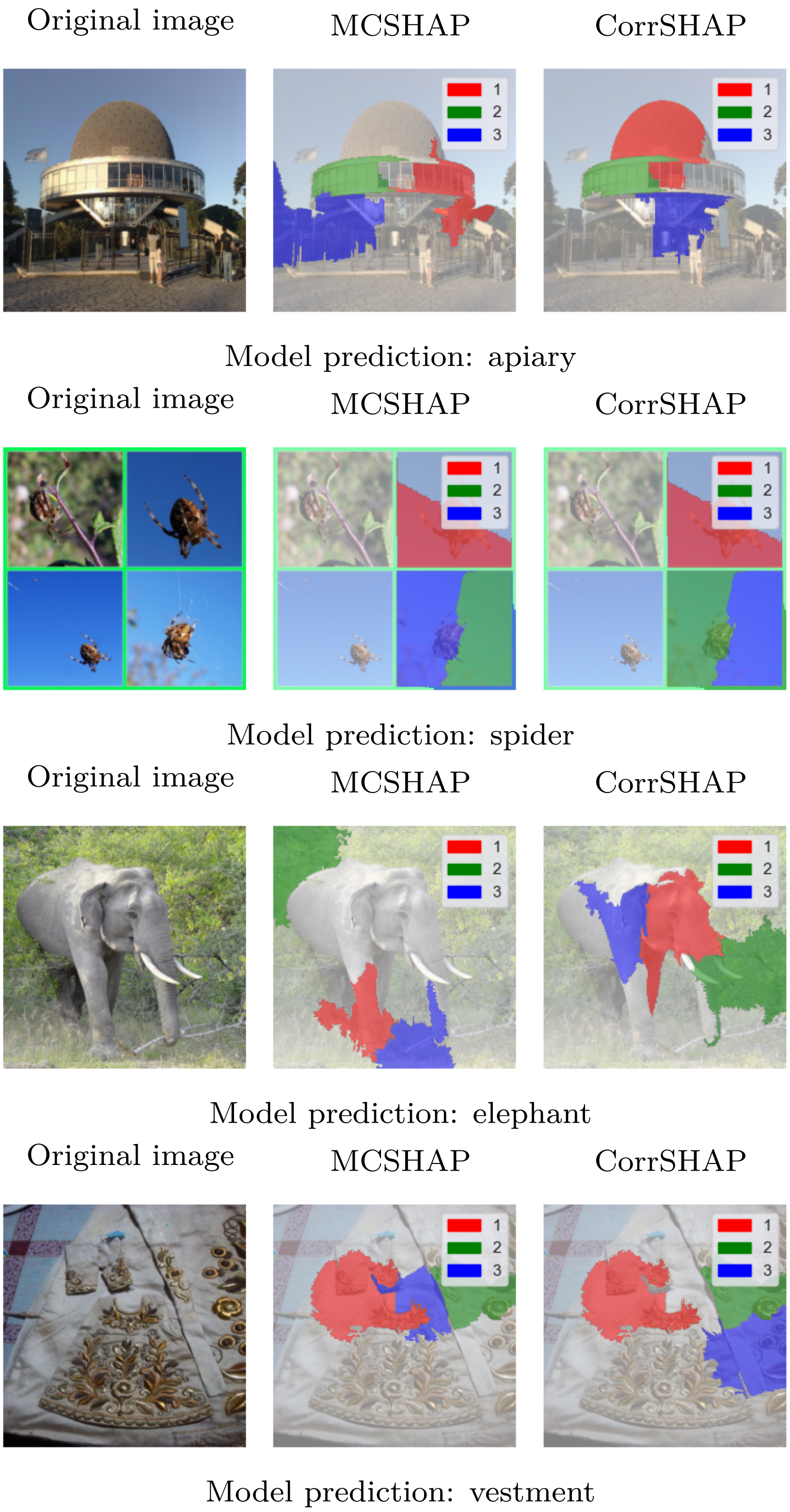


Figure 4: Qualitative comparison of explanations generated by CorrSHAP and MCSHAP. Visual inspection reveals that CorrSHAP's explanations are more consistent with human perception of important image features.

5. Research Impact and Future Work

The proposed XAI methods offer explanations that are faithful to the model, understandable to humans, and computationally efficient which enables them to be practical and applicable in real-world scenarios. The proposed sample-based XAI method has broad applicability across various domains. It can detect mislabeled data, identify data leakage, analyze memorization effects, and optimize training datasets and is applicable to other fields like control, active learning, and system identification.

Future work will explore alternative correlation measures, focus on enhancing the robustness of XAI methods, as well as conducting a deeper investigation into the interaction between kernel choice, hyperparameters, and the underlying data distribution, with the goal of developing a more stable and consistently high-performing sample based explanability method.

Declaration on Generative AI

The author has not employed any Generative AI tools.

References

[1] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, F. A. Wichmann, Shortcut learning in deep neural networks, Nature Machine Intelligence 2 (2020) 665–673.

[2] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence, Information fusion 99 (2023) 101805.

[3] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial intelligence 267 (2019) 1–38.

[4] R. C. Fong, A. Vedaldi, Interpretable explanations of black boxes by meaningful perturbation, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 3429–3437.

[5] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2921–2929.

[6] T. Fel, A. Picard, L. Bethune, T. Boissin, D. Vigouroux, J. Colin, R. Cadène, T. Serre, Craft: Concept recursive activation factorization for explainability, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2711–2721.

[7] R. Achtibat, M. Dreyer, I. Eisenbraun, S. Bosse, T. Wiegand, W. Samek, S. Lapuschkin, From attribution maps to human-understandable explanations through concept relevance propagation, Nature Machine Intelligence 5 (2023) 1006–1019.

[8] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, Advances in neural information processing systems 31 (2018).

[9] M. Scott, L. Su-In, et al., A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017) 4765–4774.

[10] M. Nauta, J. Schlötterer, M. Van Keulen, C. Seifert, Pip-net: Patch-based intuitive prototypes for interpretable image classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2744–2753.

[11] S. S. Kim, E. A. Watkins, O. Russakovsky, R. Fong, A. Monroy-Hernández, "help me help the ai": Understanding how explainability can support human-ai interaction, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–17.

[12] G. Ciravegna, P. Barbiero, F. Giannini, M. Gori, P. Lió, M. Maggini, S. Melacci, Logic explained networks, Artificial Intelligence 314 (2023) 103822.

[13] A. Sun, P. Ma, Y. Yuan, S. Wang, Explain any concept: Segment anything meets concept-based explanation, Advances in Neural Information Processing Systems 36 (2024).

[14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[15] C.-P. Tsai, C.-K. Yeh, P. Ravikumar, Sample based explanations via generalized representers, Advances in Neural Information Processing Systems 36 (2024).

[16] Z. Hammoudeh, D. Lowd, Training data influence analysis and estimation: A survey, Machine Learning 113 (2024) 2351–2403.

[17] J. Lin, A. Zhang, M. Lécuyer, J. Li, A. Panda, S. Sen, Measuring the effect of training data on deep learning predictions via randomized experiments, in: International Conference on Machine Learning, PMLR, 2022, pp. 13468–13504.

[18] J. T. Wang, R. Jia, Data banzhaf: A robust data valuation framework for machine learning, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2023, pp. 6388–6421.

[19] C. Zhang, D. Ippolito, K. Lee, M. Jagielski, F. Tramèr, N. Carlini, Counterfactual memorization in neural language models, Advances in Neural Information Processing Systems 36 (2023) 39321–39362.

[20] G. Pruthi, F. Liu, S. Kale, M. Sundararajan, Estimating training data influence by tracing gradient descent, Advances in Neural Information Processing Systems 33 (2020) 19920–19930.

[21] P. W. Koh, P. Liang, Understanding black-box predictions via influence functions, in: International conference on machine learning, PMLR, 2017, pp. 1885–1894.

[22] A. Schioppa, P. Zablotskaia, D. Vilar, A. Sokolov, Scaling up influence functions, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 8179–8186.

[23] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, et al., Sam 2: Segment anything in images and videos, arXiv preprint arXiv:2408.00714 (2024).

[24] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al., Segment anything, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4015–4026.

[25] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: International conference on machine learning, PMlR, 2017, pp. 3145–3153.

[26] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International conference on machine learning, PMLR, 2017, pp. 3319–3328.

[27] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, et al., Captum: A unified and generic model interpretability library for pytorch, arXiv preprint arXiv:2009.07896 (2020).

[28] M. T. Ribeiro, S. Singh, C. Guestrin, "why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.