

# A Flexible Metric-Based Approach to Assess Neural Network Interpretability in Image Classification

Andrea Colombo<sup>1</sup>, Laura Fiorenza<sup>1,2</sup> and Sofia Mongardi<sup>1</sup>

<sup>1</sup>Dipartimento di Eletttronica, Informazione e Bioingegneria, Politecnico di Milano, Via Ponzio 34/5, 20133, Milano, Italy

<sup>2</sup>Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Via Mario Negri 2, 20156, Milano, Italy

## Abstract

Class Activation Maps (CAM) approaches have been extensively used to understand the decision-making process of neural network models when classifying images, the so-called network dissection. These approaches identify the important regions or features that mostly contribute to the model's prediction. Most studies use this tool to offer a qualitative assessment (e.g., detect biases) of models or, at most, an image-level metric of overlap (e.g., via Intersection over Union). In this work, we leverage one of the most successful tools in network dissection for image classification, the Gradient-weighted CAM, to develop a fully quantitative and simple approach based on a similarity metric, the Tversky index, that enables a flexible benchmarking analysis of the models' overall interpretability power according to a defined criterion, whenever the classification abilities are comparable. As a proof-of-concept, we apply the proposed methodology to identify which state-of-the-art neural network model is the most faithful in using object shapes when classifying images, with Grad-CAM as our saliency map tool.

## Keywords

explainable AI, semantic segmentation, Grad-CAM, network dissection

## 1. Introduction

The growing demand to understand the decision-making process of image classification models has led to the development of computational techniques to enhance the explainability of black box models. One common approach in the field of image classification is to use saliency maps [1], which highlights the parts within an image most important for the prediction. Generating Class Activation Maps (CAM) is the most popular approach to obtain such maps [2, 3, 4]. Indeed, CAMs have been developed to help visualize the regions within the input images that impact the most the prediction of a neural network (NN) model. They have been successfully used in fields such as medicine [5] or fault diagnostics [6]. The CAM approach has also been improved over the years, especially with the introduction of the Gradient-weighted CAM (Grad-CAM) [7] and its extensions, which, in general, use the gradient associated with the predicted class membership as a weight to detect the most relevant regions or pixels.

At the same time, in recent years, there has been a rush to build new neural network model architectures that can reach state-of-the-art performances in image classification tasks. With so many architectures being proposed, CAM-based approaches have been used to increase

---


XAI.it 2023 - Italian Workshop on Explainable Artificial Intelligence

✉ andrea1.colombo@polimi.it (A. Colombo); laura.fiorenza@polimi.it (L. Fiorenza); sofia.mongardi@polimi.it (S. Mongardi)

ORCID 0000-0002-7596-8863 (A. Colombo)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

trust and transparency of such models via qualitative analysis [8], to improve their overall performances [9], or even to identify biases, such as the emphasis on image texture rather than object shape [10], which, instead, has been proven to be one the most relevant and effective element for human classification strategies [11, 12]. Yet, saliency maps have typically been used only to assess a problem or domain-specific quality of a proposed NN model, where the role of evaluation metrics, such as the Intersection over Union (IoU) [13, 14], is to provide local interpretability insights, considering individual predictions. This means that they cannot be used to evaluate, within a single metric, the global model-level quality of CAMs according to a criterion, such as the object shape.

In this work, we strive to bridge this gap by proposing a quantitative approach based on a widely used and successful saliency map tool, Grad-CAM, to evaluate different modern neural network architectures' interpretability power. In particular, we focus on their ability to capture object shapes when performing image classification. We consider that such an approach can be easily extended to all tools and methods that, just like Grad-CAM, generate activation maps. Our proposed methodology aims to quantify the overall faithfulness of different models to any desired criterion, thus enabling a benchmarking analysis of NN in terms of interpretability power. Such an approach becomes relevant when it is critical to choose the most transparent classification image model (e.g., autonomous car driving [15] or medical applications [16]) among the vast amount of powerful and accurate models available.

To this aim, we adopt, on the one side, a thresholding strategy to obtain binary saliency maps, and, on the other, we employ a similarity metric, the Tversky index [17], that, unlike the IoU, can provide us the flexibility to test what happens whenever we change the pixel overlap focus, e.g., further penalizing the false negative or false positive pixels, as required by the domain of interest. This twofold strategy enables the construction of curves which can be leveraged to compute a global metric describing the interpretability power of different models. As a preliminary experiment of our approach, we provide a proof-of-concept evaluation by performing experiments on a new semantic segmented dataset that we built and that can be used as a benchmark for future CAM-based studies.

**Contribution.** Our main contributions can be summarized as follows:

- we propose an innovative quantitative approach and a metric, the *Area Under Tversky Curve*, that, based on a popular saliency map tool, Grad-CAM, enables to benchmark the interpretability power of different models in the image classification task.
- we conduct a proof-of-concept evaluation on a selection of best performing models, with object shape as our goodness-of-fit CAM criterion.
- we contribute to a new dataset of object-shape segmented images based on the Imagenette dataset [18], which can be used for future shape-based experiments.

**Overview.** The rest of this paper is organized as follows. Section 2 provides a summary of the neural network architectures considered for our proof-of-concept and briefly introduces the Grad-CAM approach. In Section 3, we present our methodology to build a metric which evaluates the models' CAMs based on object shapes. Section 4 discusses the preliminary results of our proof-of-concept while Section 5 concludes the paper and presents potential future directions and applications.

## 2. Overview of model architectures and Grad-CAM

This section briefly provides an overview of some powerful NN models commonly used in image classification. Table 1 summarizes the main differences we identified among the architectures.

	VGG16	Resnet50	InceptionV3	MobileNet	ViT Base
Architecture Style	Sequential	Residual Learning	Inception Modules	Depthwise Conv.	Transformer blocks
N. of layers	16 (including 13 Conv.)	50 (including 49 Conv.)	48 (including 42 Conv.)	28 (including 27 Conv.)	12 blocks (12 heads each)
N. of parameters	138 Mil	23 Mil	25 Mil	13 Mil	86 Mil
Conv. filters	3x3 and 1x1	7x7, 3x3, 1x1	5x5, 3x3, 1x1	3x3, 1x1,	None

**Table 1**

Overview of the neural network architectures used in the experiments.

**VGG16.** VGG16 [19] is a convolutional neural network (CNN) model with 16 layers and is one of the most popular CNN models for image classification. The VGG16 model is made up of a stack of convolutional layers and max pooling layers. Convolutional layers extract features from the input image, and max pooling layers downsample the feature maps, reducing the model's size while retaining important features.

**Resnet50.** ResNet50 [20] is a CNN model with 50 layers and is one of the most popular for image classification. It is based on the idea of residual learning, a technique that allows CNNs to learn long-range dependencies between features. This is done by adding a shortcut connection between the input and output of a convolutional block.

**Inception.** The Inception [21] model is a convolutional neural network (CNN) model originally developed by Google in 2014. Its V3 version is a deep CNN with 48 layers and is one of the most popular models for image classification. The Inception model is based on inception modules, which combine convolutional filters in a single layer. This allows the model to learn more features from the input image and can also help reduce the model's size.

**MobileNet.** MobileNet [22] is a CNN model developed by Google in 2017. It is a small, lightweight CNN that is based on the idea of depthwise separable convolutions. Depthwise separable convolutions reduce the computational complexity of a CNN by factorizing a convolutional layer into two layers: a depthwise and a pointwise convolution layer. The depthwise layer extracts features from the input image, while the pointwise layer is responsible for building new features by computing linear combinations of the input channels. This allows MobileNet to achieve high accuracy on image classification tasks while still being small and lightweight.

**ViT.** Vision Transformer (ViT) is a type of neural network based on transformers, originally developed for natural language processing tasks and recently extended to image classification [23]. ViT works by dividing the input image into a grid of patches, which are then processed by a stack of transformer blocks. Unlike the above architectures, the ViT model contains no convolutional layers. Each transformer block comprises two sub-layers: a self-attention and a feed-forward layer. The first allows the model to learn the relationships between different patches in the image while the latter allows the model to learn more complex features from the image.

## 2.1. Grad-CAM

Grad-CAM (Gradient-weighted Class Activation Mapping) is a method for visualizing an image’s regions most relevant to a particular class prediction made by a CNN [7]. It is a simple and effective method that can be used to understand how CNNs make their predictions. The Grad-CAM technique computes the gradients of the classification score for the final convolutional feature map to identify the regions of an image that most impact the classification score. The pixels with a large gradient denote the regions that most influence the final score.

$$\text{Grad-CAM}_c(i, j) = \text{ReLU}\left(\sum_k \frac{\partial Y^c}{\partial A_{ij}^k} A_{ij}^k\right) \quad (1)$$

where  $\text{Grad-CAM}_c(i, j)$  relates to the Grad-CAM activation for class  $c$  at spatial position  $(i, j)$ .  $Y^c$  denotes the output score of the target class  $c$  in the final layer of the neural network, and  $A_{ij}^k$  corresponds to the activation value of the feature map at position  $(i, j)$  in the  $k$ -th channel.

## 3. Methodology

In this section, we present our approach to quantitatively assess how models explain, through saliency maps, their prediction in image classification tasks. In particular, we focus on detecting which model most relies on object shapes for classifying images, as we identified it as one of the main drivers of human classification behavior. To this aim, we utilize a widely used benchmark dataset, Imagenette [18], and we build a set of test images via semantic segmentation, i.e., manually detecting and segmenting object shapes. Then, we define a thresholding strategy to binarize the Grad-CAM-generated images and use a similarity index to determine the degree of overlap between segmented images and the CAMs, considering distinct penalization mechanisms. Finally, we leverage the thresholds to plot the behavior of the similarity index and introduce an *Area Under Curve* criterion that captures the overall quality of CAMs.

### 3.1. Dataset, Fine-Tuning and Grad-CAM

We identify Imagenette [18], a smaller subset of 9,469 images from ImageNet [24], as our reference dataset. It includes 10 easily classifying classes (tench, English springer, cassette player, chain saw, church, French horn, garbage truck, gas pump, golf ball, and parachute). All the models in Section 2 are already pre-trained on the entire ImageNet dataset. However, we decided to perform additional fine-tuning over 20 epochs on the Imagenette subset to further mitigate classification issues. Table 2 summarizes the results of the fine-tuning <sup>1</sup>

**Grad-CAM Test Set.** Given the Imagenette dataset, we manually collected around 30 images per class from the web, resulting in a total of 307 images. These images define our *Grad-CAM Test Set*, on which we will showcase our methodology. We assume that the collected images are outside the training data as we performed a search selecting recent images only. Although some datasets with semantic segmentation already exist, such as the PASCAL dataset [25], we

---

<sup>1</sup>The codes to fine-tune the models, generating Grad-CAMs and to replicate our experiment are available at <https://github.com/SofSof98/Human-like-image-classification>

	VGG16	Resnet50	InceptionV3	MobileNet	ViT Base
Best Validation Epoch	19	18	12	15	12
Validation Accuracy	0.961	0.989	0.983	0.984	0.974

**Table 2**

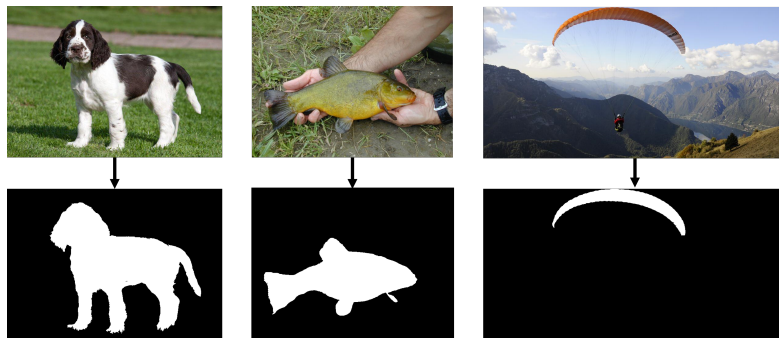
Results of the models’ fine-tuning on Imagenette with a 70-30 training-validation split.

deemed our novel dataset, based on the Imagenette classes, ideal for our analysis as it includes images belonging to rather distinct classes, thus avoiding the issue of undecidability due to the presence of multi-label or similarly-labeled images. In fact, our approach is thought to evaluate and compare the behavior of different models with similar and high classification accuracy, intending to provide an interpretability power ranking of these performing architectures.

As expected, out of our 307 test images, all the selected models in Section 2 achieve very high accuracy, with a total of 300 images that have been jointly labeled correctly by all the architectures and that can be used for generating the CAMs. While there exists a solid record of Grad-CAM applications to CNNs that proved its effectiveness [26, 27], we followed recent studies that extend Grad-CAM to be applied to a transformer-based architecture such as the popular ViT model [28, 29]. For CNN, as a general principle, we identified the last convolutional layer as the target for Grad-CAM, with adaptations depending on the model and previous CAM studies [30].

### 3.2. Proposed Approach

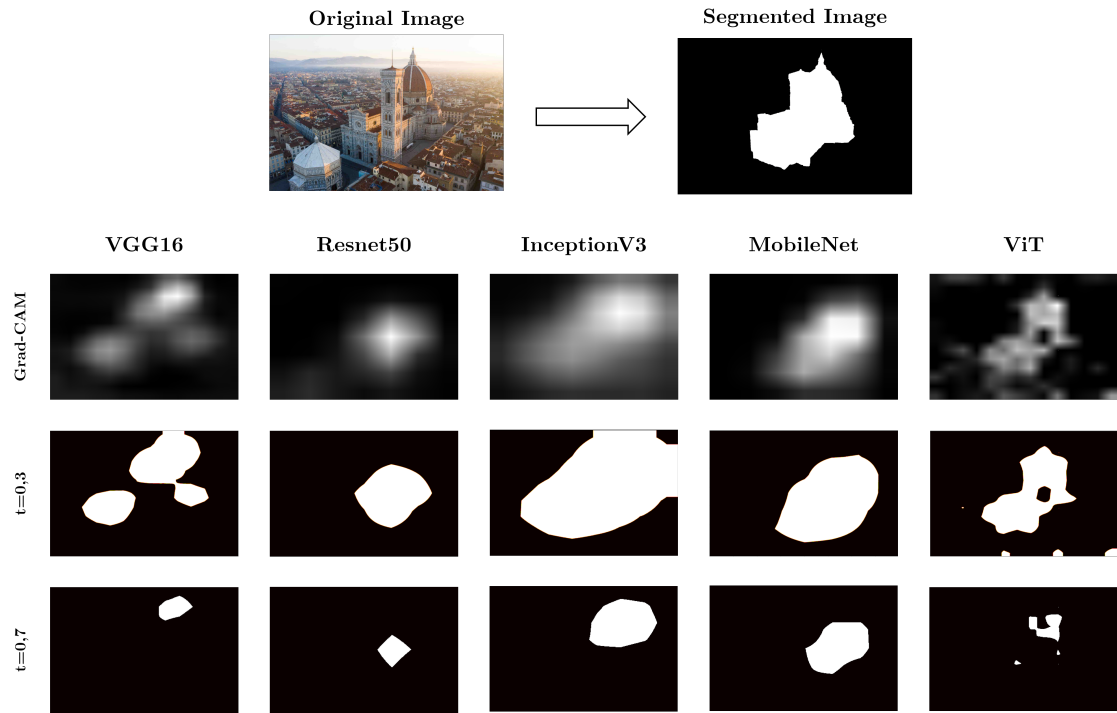
**Semantic segmentation.** The next step of our approach is to define a semantic segmentation criterion on which we want to perform our interpretability analysis. To this aim, we decided to annotate the entire object area under evaluation<sup>2</sup>, as shown in Figure 1. While this semantic segmentation approach discards elements from the background, which might result in a loss of contextual information, we follow recent studies that identify this principle as one of the main drivers that allows humans to classify images [11, 12] and, therefore, a natural choice for an interpretability criterion.



**Figure 1:** Example of semantic segmentation of test images. White area represents the object’s shape.

<sup>2</sup>The test and the segmented dataset are publicly available on HuggingFace at <https://bitly.ws/WPqN>

**Thresholding Grad-CAM images.** The segmentation criterion requires a pixel-wise comparison with the Grad-CAM output. However, the class activation map includes different intensity levels, while our maps are binary. As a manual segmentation that incorporates intensity levels is not feasible and is highly affected by inter-observer variability, we adopt a threshold strategy on Grad-CAM [31]. Based on the intensity level, each pixel will take a value of 1 if the associated value is above a certain threshold value and 0 otherwise. In our setting, we do not use a fixed threshold; instead, we test different threshold values to evaluate how models behave when focusing on more relevant regions, as shown in Figure 2.



**Figure 2:** Example of Grad-CAMs and their binarization for different thresholds  $t$ .

**The Tversky index.** To evaluate which model most resembles our criterion, we must account for different aspects. On the one hand, we have to prefer models that consider, as important regions, the ones intersecting the annotated area, i.e., high recall models. On the other hand, we need to evaluate the precision of the models and penalize those that achieve high recall by also looking at some background regions, i.e., low-precision ones. To consistently consider both aspects within a single metric, we identified the Tversky index as ideal, giving us the flexibility to compare Grad-CAMs over distinct layers of importance. The Tversky index is a similarity coefficient that measures the degree of overlap between two sets defined as:

$$T(A, B) = \frac{|A \cap B|}{|A \cap B| + \alpha|A \setminus B| + \beta|B \setminus A|} \quad (2)$$

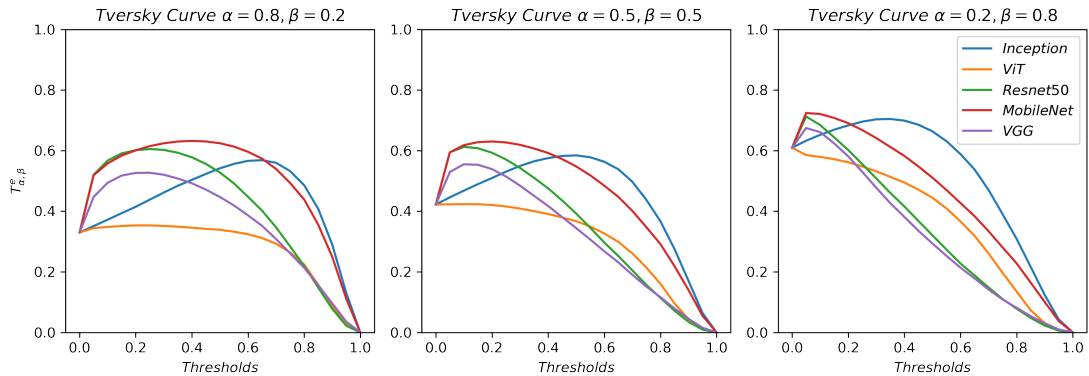
where  $A$  is the set of pixels of the segmented image and  $B$  is the one resulting from the thresholding strategy on Grad-CAM.  $|A \cap B|$  represents the cardinality of their intersection,

$|A \setminus B|$  represents the cardinality of elements in  $A$  but not in  $B$ , and  $|B \setminus A|$  represents the cardinality of elements in  $B$  but not in  $A$ . The parameters  $\alpha$  and  $\beta$  control the emphasis on the differences in the sets. Adjusting  $\alpha$  and  $\beta$  allows the Tversky index to provide a flexible similarity measure that can account for variations in the importance of shared and non-shared elements between sets. Specifically, the  $\alpha$  and  $\beta$  parameters control the relative importance of false positives and false negatives, respectively.

**Area Under Tversky Curve.** The Tversky index can be computed on an individual image level. To get a model-level metric we introduce the *Averaged Tversky Index*  $T_{\alpha,\beta}^e = \sum_i T_{\alpha,\beta}(A_i, B_i)/N$ , which represents the mean value over the entire set of the  $N$  test images. If we compute it for an increasing set of thresholds (used to derive the binary Grad-CAMs), we get a set of values for the same model that can be plotted in a graph, generating a curve that describes the behavior of the Tversky index when we focus on pixels of increased importance, given certain values of  $\alpha$  and  $\beta$ . From the curve, we can derive the *Area Under Tversky Curve*,  $AUT_{\alpha,\beta}$ , which allows us to compare the overall Grad-CAM performances between different models. In other words, with our approach, we propose replicating what happens in a binary classification problem and translating this approach to globally evaluating the interpretability performances of NN models.

## 4. Proof-of-Concept Evaluation

To evaluate the approach presented in Section 3, we conduct a proof-of-concept experiment, selecting three combinations of  $\alpha$  and  $\beta$  values for computing the Average Tversky index, such that we can focus on different levels of recall and precision. In particular, we present a case of high importance to false positives, i.e., focus on CAM precision ( $\alpha = 0.8, \beta = 0.2$ ), a balanced case ( $\alpha = \beta = 0.5$ ) and a case whose interest is on detecting which model maximize the CAM recall ( $\alpha = 0.2, \beta = 0.8$ ). In all these cases, we applied the usual relation  $\alpha + \beta = 1$ , commonly used with the Tversky index [32]. To compute the curve, we chose a set of increasing thresholds with a step size of 0.05. In Figure 3 we report the curves whose  $AUT$  values are in Table 3.



**Figure 3:** Tversky curves computed over a selection of  $\alpha$  and  $\beta$  values.

**Discussion of the results.** Analyzing the  $AUT$  values, we can clearly identify some model-specific trends. The  $AUT$  values of the Inception and ViT models increase the more we focus on maximizing the pixel overlap, i.e.,  $AUT_{0.2,0.8}$ . This means, and it can be empirically verified

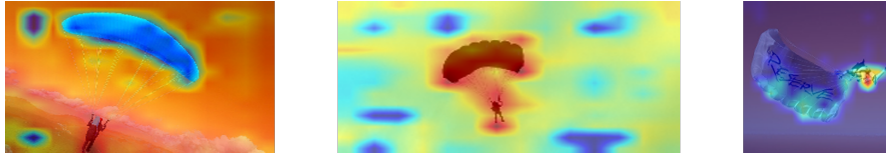
Model	$AUT_{0.8,0.2}$	$AUT_{0.5,0.5}$	$AUT_{0.2,0.8}$
InceptionV3	0.44	0.45	0.52
Resnet50	0.42	0.35	0.33
MobileNet	0.51	0.46	0.46
VGG	0.37	0.32	0.32
ViT-Base	0.28	0.30	0.37

**Table 3**

Area Under Tversky Curve for the considered architectures over different  $\alpha$  and  $\beta$  combinations.

by manually looking at their Grad-CAMs, that these models look at large sets of pixels when classifying images, generating areas bigger than the object shape. Conversely, the VGG, Resnet, and MobileNet models show a declining trend in the  $AUT$  values, meaning that they perform better in terms of precision of the overlap between the annotated area and the Grad-CAMs. Overall, in our PoC based on Grad-CAM, we identified MobileNet as the best-performing model, being the most precise in detecting object shape when classification accuracy is similar, and experiencing only a slight performance drop if we focus on recall, i.e.,  $AUT_{0.2,0.8}$ .

A dedicated discussion about the behavior of the ViT model is needed. We empirically observe that the ViT, the only non-CNN model we tested, follows a rather distinct path and seems to perform the worst according to our analysis. Further analysis revealed that the ViT model uses different criteria when classifying images, even within the same class. Cases arise where the object shape seems more relevant, while in other cases, only the background is significant, as shown in Figure 4. This might justify its higher classification accuracy in more complex cases, even outperforming all CNN-based models [33]. However, it might add a layer of complexity that undermines interpretability, with no clear patterns that can be identified in the CAMs.



**Figure 4:** Example of ViT Grad-CAMs (in the colored scale version) on images of the *Parachute* class. Note that red regions correspond to high-class relevance.

## 5. Conclusion and Future Work

In this paper, we leveraged a flexible metric, i.e., the Tversky index, to define an innovative quantitative approach to evaluate the interpretability and faithfulness of state-of-the-art neural network architectures for image classification to a defined criterion. We conducted a proof-of-concept evaluation utilizing the object shape as our criterion and a widely used saliency maps tool, Grad-CAM, and we proposed a new metric, the Area Under Tversky Curve as an overall indicator of interpretability performance. We think that, after a further, more extensive, evaluation proving its consistency, this approach can be applied to more general and complex cases. For instance, we envision its potential use in a crowdsourcing study to draw solid conclusions about the interpretability power of different NN models and detect which architectural elements might influence the ability of NN to replicate human thought.



## References

- [1] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, in: Y. Bengio, Y. LeCun (Eds.), 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings, 2014. URL: <http://arxiv.org/abs/1312.6034>.
- [2] B. N. Patro, M. Lunayach, S. Patel, V. P. Namboodiri, U-cam: Visual explanation using uncertainty based class activation maps, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7444–7453.
- [3] K. H. Sun, H. Huh, B. A. Tama, S. Y. Lee, J. H. Jung, S. Lee, Vision-based fault diagnostics using explainable deep learning with class activation maps, *IEEE Access* 8 (2020) 129169–129179. doi:10.1109/ACCESS.2020.3009852.
- [4] S. Yang, Y. Kim, Y. Kim, C. Kim, Combinational class activation maps for weakly supervised object localization, in: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE Computer Society, Los Alamitos, CA, USA, 2020, pp. 2930–2938. URL: <https://doi.ieeecomputersociety.org/10.1109/WACV45572.2020.9093566>. doi:10.1109/WACV45572.2020.9093566.
- [5] V. Gupta, M. Demirel, M. Bigelow, S. M. Yu, J. S. Yu, L. M. Prevedello, R. D. White, B. S. Erdal, Using transfer learning and class activation maps supporting detection and localization of femoral fractures on anteroposterior radiographs, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), 2020, pp. 1526–1529. doi:10.1109/ISBI45749.2020.9098436.
- [6] K. H. Sun, H. Huh, B. A. Tama, S. Y. Lee, J. H. Jung, S. Lee, Vision-based fault diagnostics using explainable deep learning with class activation maps, *IEEE Access* 8 (2020) 129169–129179. doi:10.1109/ACCESS.2020.3009852.
- [7] 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: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626. doi:10.1109/ICCV.2017.74.
- [8] M. Ribeiro, S. Singh, C. Guestrin, “why should I trust you?”: Explaining the predictions of any classifier, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, Association for Computational Linguistics, San Diego, California, 2016, pp. 97–101. URL: <https://aclanthology.org/N16-3020>. doi:10.18653/v1/N16-3020.
- [9] F. P. Caforio, G. Andresini, G. Vessio, A. Appice, D. Malerba, Leveraging grad-cam to improve the accuracy of network intrusion detection systems, in: International Conference on Discovery Science, Springer, 2021, pp. 385–400.
- [10] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, W. Brendel, Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness., in: International Conference on Learning Representations, 2019. URL: <https://openreview.net/forum?id=Bygh9j09KX>.
- [11] N. Baker, H. Lu, G. Erlichman, P. J. Kellman, Deep convolutional networks do not classify based on global object shape, *PLoS computational biology* 14 (2018) e1006613.
- [12] G. Malhotra, M. Dujmović, J. Hummel, J. S. Bowers, Human shape representations are not an emergent property of learning to classify objects (2021). URL: <https://doi.org/10.1101/>

2021.12.14.472546. doi:10.1101/2021.12.14.472546.

- [13] D. Bau, B. Zhou, A. Khosla, A. Oliva, A. Torralba, Network dissection: Quantifying interpretability of deep visual representations, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, 2017, pp. 3319–3327. URL: <https://doi.org/10.1109/CVPR.2017.354>. doi:10.1109/CVPR.2017.354.
- [14] R. Fong, A. Vedaldi, Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2018, pp. 8730–8738. URL: <https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00910>. doi:10.1109/CVPR.2018.00910.
- [15] G. Wiegand, M. Schmidmaier, T. Weber, Y. Liu, H. Hussmann, I drive - you trust: Explaining driving behavior of autonomous cars, in: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, CHI EA '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 1–6. URL: <https://doi.org/10.1145/3290607.3312817>. doi:10.1145/3290607.3312817.
- [16] Z. Salahuddin, H. C. Woodruff, A. Chatterjee, P. Lambin, Transparency of deep neural networks for medical image analysis: A review of interpretability methods, *Computers in Biology and Medicine* 140 (2022) 105111. URL: <https://www.sciencedirect.com/science/article/pii/S0010482521009057>. doi:<https://doi.org/10.1016/j.combiomed.2021.105111>.
- [17] A. Tversky, Features of similarity., *Psychological review* 84 (1977) 327.
- [18] J. Howard, Imagenette: A smaller subset of 10 easily classified classes from imagenet, 2019. URL: <https://github.com/fastai/imagenette>.
- [19] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL: <http://arxiv.org/abs/1409.1556>.
- [20] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.
- [21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9. doi:10.1109/CVPR.2015.7298594.
- [22] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, *CoRR* abs/1704.04861 (2017). URL: <http://arxiv.org/abs/1704.04861>. arXiv:1704.04861.
- [23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, 2021. URL: <https://openreview.net/forum?id=YicbFdNTTy>.
- [24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition

- Challenge, *International Journal of Computer Vision (IJCV)* 115 (2015) 211–252. doi:10.1007/s11263-015-0816-y.
- [25] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, *International Journal of Computer Vision* 88 (2010) 303–338.
- [26] R. Fu, Q. Hu, X. Dong, Y. Guo, Y. Gao, B. Li, Axiom-based grad-cam: Towards accurate visualization and explanation of cnns, *arXiv preprint arXiv:2008.02312* (2020).
- [27] M. N. Islam, M. Hasan, M. K. Hossain, M. G. R. Alam, M. Z. Uddin, A. Soylu, Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from ct-radiography, *Scientific Reports* 12 (2022) 11440.
- [28] J. Gildenblat, Advanced ai explainability for pytorch, <https://github.com/jacobgil/pytorch-grad-cam>, 2021.
- [29] H. Ma, X. Li, X. Yuan, C. Zhao, Denseformer: A dense transformer framework for person re-identification, *IET Computer Vision* 17 (2022) 527–536. URL: <https://doi.org/10.1049/cvi2.12118>. doi:10.1049/cvi2.12118.
- [30] J. Gildenblat, contributors, Pytorch library for cam methods, <https://github.com/jacobgil/pytorch-grad-cam>, 2021.
- [31] M. Xiao, L. Zhang, W. Shi, J. Liu, W. He, Z. Jiang, A visualization method based on the grad-cam for medical image segmentation model, in: *2021 International Conference on Electronic Information Engineering and Computer Science (EIECS)*, 2021, pp. 242–247. doi:10.1109/EIECS53707.2021.9587953.
- [32] S. S. M. Salehi, D. Erdogmus, A. Gholipour, Tversky loss function for image segmentation using 3d fully convolutional deep networks, in: Q. Wang, Y. Shi, H.-I. Suk, K. Suzuki (Eds.), *Machine Learning in Medical Imaging*, Springer International Publishing, Cham, 2017, pp. 379–387.
- [33] J. Maurício, I. Domingues, J. Bernardino, Comparing vision transformers and convolutional neural networks for image classification: A literature review, *Applied Sciences* 13 (2023). URL: <https://www.mdpi.com/2076-3417/13/9/5521>. doi:10.3390/app13095521.