An Approach to Selecting the Best Explainability Method for CNNs Based on Distance Metrics and Human Perception

Daniel da Silva Costa[1,*]

[1]Graduate Program in Computer Science (PPGI), Federal University of the State of Rio de Janeiro (UNIRIO), Rio de Janeiro, Brazil

Abstract

Deep Learning models have been increasingly used in numerous applications, including those sensitive to human life, which require clear explanations and justifications. This has encouraged several investigations into the explainability of neural networks. Various explainability methods have been proposed, but not many metrics have been developed to evaluate them. The Intersection over Union is the most common metric. It is applied between two bounding boxes to measure their similarity. The saliency maps generated by the explainability methods have no known shape. We assume that there must be a better metric that allows us to find an explainability method that best resembles human perception. To this end, we propose applying distance metrics to assess the similarity between human perception and explanation saliency maps. We conducted an investigation using images of chihuahuas from the ImageNet dataset. Several CAM-based explainability methods were used to generate saliency maps, which were compared with human perception of the most important parts of each image. With the results of the distance metrics, rankings of the best saliency maps were created and compared to the ranking obtained using people's choices for each selected image. This comparison was performed using the Rank-Biased Overlap (RBO) metric. The results obtained thus far indicate the feasibility of our method in finding the explainability method that best resembles human perception. In our experiments, the two best metrics were Manhattan and Correlation. Besides, the best explainability methods regarding human perception were LayerCAM, Score-CAM, and IS-CAM.

Keywords

Explainability, Deep Learning, Computer Vision, Human Perception

1. Context and Motivation

Recently, Deep Learning (DL) models, which are based on artificial neural networks (ANNs), have been used in a wide variety of contexts. ANNs can internally and automatically build a complex model of the input data without the need for human operators to describe the characteristics of the data [1].

Despite significant advances in DL, some application areas require greater care and understanding of the information on which ANNs are based when learning internal representations of input data, since they deal with situations potentially harmful to human life, such as medicine and law, which usually require clear and unequivocal explanations of the systems' decisions. The confidence in the results and the usefulness of ANNs will be compromised if the user does not understand the rules behind the decision process of these models [2].

Generally speaking, Explainability aims to create an explanation that can help the user understand why a given model has a specific decision pattern, whether regarding the entire dataset or an instance of this set [2], and can help increase the reliability of the use of ANNs models [2].

For the task of image classification in Computer Vision (CV), explainability methods typically try to show the user which information from the input the model considers most relevant for a specific classification, generating a saliency map (or heatmap) that can be superimposed on the original image.

Late-breaking work, Demos and Doctoral Consortium, colocated with the 4th World Conference on eXplainable Artificial Intelligence: July 01–03, 2026, Fortaleza, Brazil

*Corresponding author.

daniel.scosta@edu.unirio.br (D. d. S. Costa)

https://ppgi.uniriotec.br/en/home-page/ (D. d. S. Costa)

0009-0002-2952-5129 (D. d. S. Costa)

Researchers often assess the quality of their explainability methods by visually comparing their heatmaps with the heatmaps of other methods, which can bias the evaluation. Although objective metrics have been proposed, they are more common than human perception-based metrics and, to the best of our knowledge, evaluating the quality and reliability of these methods remains an open problem [3].

2. Related Works

Explainability methods based on the Class Activation Maps (CAM) [4] method are among the most traditional methods [2, 5] used with convolutional neural networks (CNN) [6]. In general, these methods aim to highlight which pixels of the images the model gave more importance to when classifying those images [2, 5]. The original CAM method allows the image to be visualised with a superimposed heatmap of the discriminating areas used by CNN in the inference phase.

The CAM-based methods can be divided into two groups: activation-based and gradient-based methods. Gradient-based (or backpropagation-based [2]) methods calculate the importance of each input feature based on some evaluation of the gradients of the trained neural network. The activation-based (or perturbation-based [2]) methods calculate the importance of each input feature based on the differences in the outputs for an image and modified copies of it. Despite many proposals for CAM-based methods in recent years [4, 7, 8, 9, 10, 11, 12, 13, 14], to date, there have not been many papers dedicated to exploring and conducting an exhaustive comparison of these methods.

A recent work similar to the one proposed in this paper is that of [15]. Although the authors carried out alignment experiments between human perception and the results of different explainability methods and network architectures, they did not take into account that alignment can have specific results depending on whether the explainability method is gradient-based or not. They only evaluated the results of gradient-based methods: Grad-CAM and XRAI. The Score-CAM method, for example, is not based on gradients and has been one of the most cited in recent works. In this sense, a novelty presented by this work corresponds to the evaluation of the results of CAM-based explainability methods using human perception as ground truth, while considering both gradient-based and activation-based methods.

3. Research Question, Hypothesis and Objectives

During the literature review, it was observed that, in general, the authors of several research studies aim to produce new explainability methods and usually justify that the results of their methods are better than others by visually comparing the heatmaps produced by their methods with those found in the literature, but without any different or more robust evidence; for example, by objectively comparing the results or collecting data generated by human beings to consider, at least, human perception.

Generally speaking, the authors merely present the heatmaps generated by their newly proposed methods, alongside the results of other methods, and conclude that they obtained results compatible with those in the literature. In this way, they ended up performing a subjective and, therefore, biased evaluation. Thus, the research of the present work started with the following research question: **"How Can One Choose the Best CAM-Based Explainability Method?"** and aims to promote the discussion around the importance of developing methodologies and metrics that can help address the evaluation problem of explainability methods.

The hypothesis of this study is based on the idea that the more similar human explanations are to those produced by explainability methods, the greater the trust humans will place in the explanations provided by these methods [16, 17]. Therefore, this study aims to evaluate and/or develop metrics that enable, with greater certainty, the identification of explainability methods that produce results most closely resembling human perception.

We seek to understand and evaluate the alignment [16] between the CAM-based explainability methods and the perception of people regarding the most important area of each image in the task of image classification.

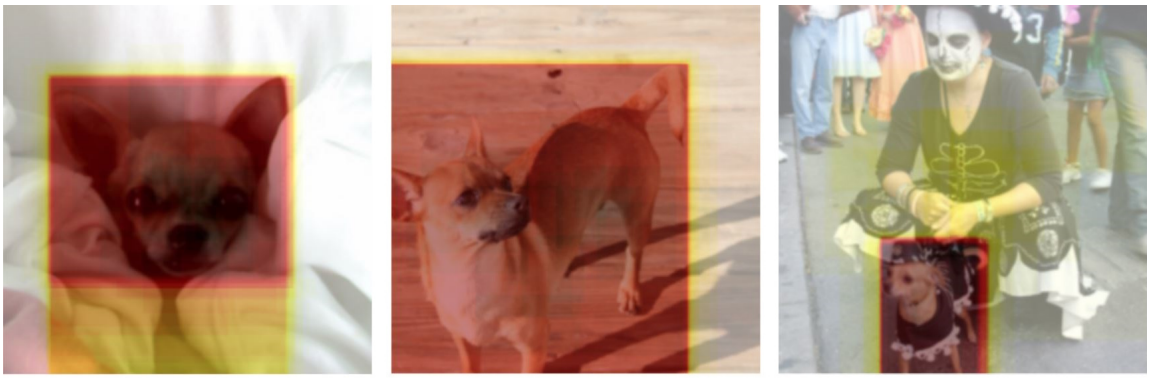

Figure 1: Examples of the final annotation heatmap for three images: n02085620_1312, n02085620_5542 and n02085620_8174. The darker the shade of red, the more important the area. The lighter the yellow, the less important the area.

4. Research Approach and the Method

It is common to apply the Intersection over Union (IoU) metric (also called Jaccard similarity) [4] to calculate the similarity between two boxes. The problem is determining the limit of the heatmap, which is not necessarily a box. Therefore, the present research ought to use different metrics that allow for an investigation of the similarity between human perception and the heatmaps generated by the explainability methods, using all the information from the heatmaps. We randomly selected 20 Chihuahua images from the ImageNet [18] dataset and classified each using the pre-trained ResNet50 model [19]. Three images were incorrectly classified. We removed them from the experiments.

The remaining 17 images were used in an "annotation experiment" on the Appen crowdsourcing platform[1] where 50 people drew a bounding box around the chihuahua in each image. All the 50 annotations were combined into one "annotation heatmap" to allow us to better capture the importance of each pixel with respect to people's perception. Consequently, if a pixel is present in, for example, five bounding boxes, it will be assigned a weight of five in the final annotation heatmap. The values of all heatmaps were normalised between 0 and 1. Figure 1 shows three examples of the annotation heatmap.

The CAM-based methods available in the TorchCAM[2] library were applied to the model to generate "explanation heatmaps". The methods are shown in Table 1. Figure 2 shows the preprocessed image n02085620_7243 and the explanation heatmaps generated by the CAM-based methods. The heatmaps are similar but not equal. For example, in the case of Score-CAM, the heatmap covers a broader area of the dog, highlighting parts of the ears and neck; and, in the case of Smooth Grad-CAM++, the generated heatmap is more concentrated around the dog's snout.

We calculate the distance between the annotation heatmaps and the explanation heatmaps using the following metrics: Weighted Jaccard (WJ); Wasserstein (WA); Bray-Curtis (BC); Canberra (CA); Chebyshev (CY); Manhattan (MA); Correlation (CR); Cosine (CS); Euclidean (EU); Jensen-Shannon (JS); Minkowski (MI); and squared Euclidean (SE). We will not describe in detail all the distance metrics used in this research for the sake of space.

We used the crowdsourcing platform called Prolific[3] to create an "validation experiment" where we asked people to select the heatmap that best explained the most important parts of the chihuahua. The images were shown with each overlapped heatmap, similar to Figure 2 but without the methods names.

[1] https://www.appen.com/
[2] https://github.com/frgfm/torch-cam
[3] https://www.prolific.com/

Table 1
The Selected Explainability Methods.

Method	Acronym	Type	Paper	Year
CAM	CAM	Activation	[4]	2016
SS-CAM	SSCAM	Activation	[8]	2020
IS-CAM	ISCAM	Activation	[9]	2020
Score-CAM	ScCAM	Activation	[7]	2020
Grad-CAM	GCAM	Gradient	[10]	2017
Grad-CAM++	GCAM++	Gradient	[11]	2018
Smooth Grad-CAM++	SGCAM++	Gradient	[12]	2019
X-Grad-CAM	XGCAM	Gradient	[13]	2020
LayerCAM	LCAM	Gradient	[14]	2021

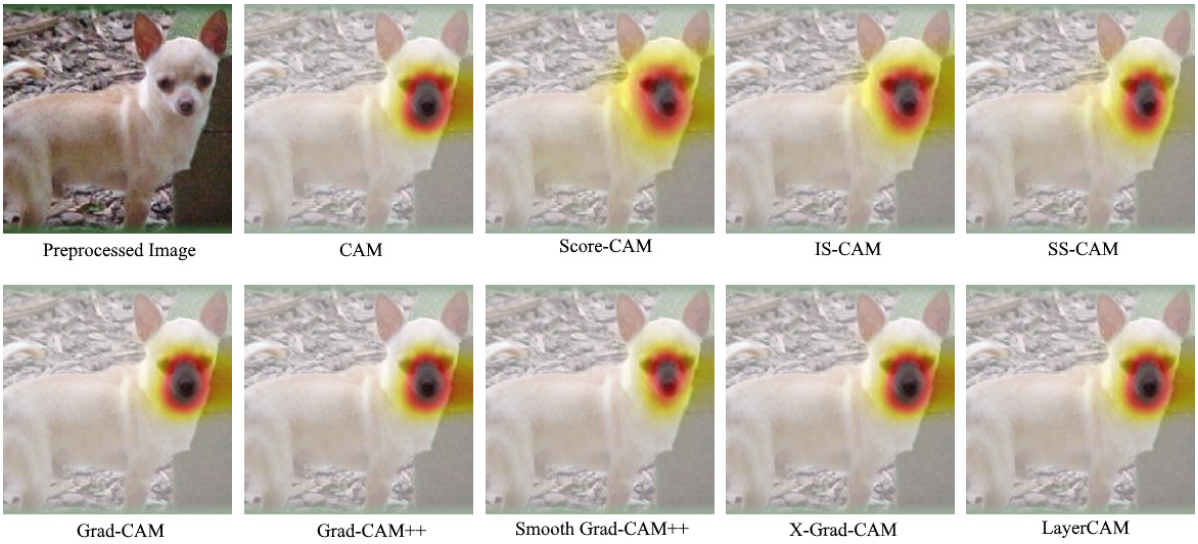

Figure 2: The preprocessed image n02085620_7243 on the top-left and the overlapped explanation heatmap of each explainability method.

Each person should choose one and only one option per chihuahua image. The results were used to create a human ranking of the best explainability methods and in the same way the rankings with the results of each distance metric ("metric ranking"). We applied the Rank-Biased Overlap (RBO) [20] metric to compare the rankings. RBO is a similarity measure for indefinite rankings. The main goals of this metric are: (i) handle non-conjointness; (ii) weight high rankings more heavily than low; (iii) and be monotonic with increasing depth of evaluation.

5. Preliminary Results

The Table 2 shows the scores for each distance metric calculated between the annotation heatmap and each explanation heatmap for the image n02085620_8174. The lower the value, the better, which means that the explanation heatmap is closest to the ground truth for a specific distance metric. The best results were highlighted in bold. The values were normalised between 0 and 1. We conducted the same analysis for all the 17 images. For each method, the best result was highlighted in bold. For this image, the Correlation (CR) metric presented values closest to zero in most cases: five out of nine explainability methods. SSCAM obtained the best result in 11 metrics. It is also possible to see that the XGCAM scores were the same as the GCAM scores for all metrics. This was observed for all 17 chihuahua images.

Table 2
Calculated distance normalized values between the annotation heatmaps and the explanation heatmaps, for the image n02085620_8174.

	Explainability Methods								
Metrics	GCAM	GCAM++	SGCAM++	XGCAM	LCAM	CAM	ScCAM	SSCAM	ISCAM
WJ	0.3380	0.6401	0.5812	0.3380	0.1744	0.3239	**0.0000**	1.0000	0.3174
WA	0.2458	0.4743	**0.2121**	0.2458	**0.0068**	0.2397	**0.0000**	1.0000	0.5324
BC	0.3052	0.6047	0.5441	0.3052	0.1538	0.2918	**0.0000**	1.0000	0.2857
CA	0.6418	1.0000	0.2491	0.6418	0.0073	0.5893	**0.0000**	**0.2891**	0.0328
CY	0.7282	0.8369	0.6195	0.7282	0.6630	0.7391	0.2282	1.0000	**0.0000**
MA	0.0735	0.2081	0.2715	0.0735	0.0341	0.0675	**0.0000**	1.0000	0.3667
CR	**0.0543**	**0.0979**	0.2483	**0.0543**	0.0526	**0.0541**	**0.0000**	1.0000	0.1498
CS	0.0830	0.1627	0.2771	0.0830	0.0609	0.0813	**0.0000**	1.0000	0.1730
EU	0.1557	0.2652	0.4170	0.1557	0.1182	0.1533	**0.0000**	1.0000	0.2515
JS	0.3221	0.9021	0.4313	0.3221	0.0812	0.2951	**0.0000**	1.0000	0.3849
MI	0.2176	0.3219	0.4984	0.2176	0.1830	0.2163	**0.0000**	1.0000	0.1862
SE	0.1122	0.2007	0.3365	0.1122	0.0837	0.1103	**0.0000**	1.0000	0.1892

The Table 3 shows the human ranking and the metrics rankings for the case of the image n02085620_8174. In the column "Rankings", "H" is for the human ranking, and the others are for the distance metrics. The column "1st" is the most important position in that ranking, and "9th" is the least important. The score achieved by each metric ranking, compared to the human ranking, is shown in the column "RBO". An RBO score close to 0 indicates proximity to the human ranking, while a score close to 1 indicates the contrary. In this case, the Minkowski (MI) ranking, in bold, is the most similar to the human ranking, with an RBO score equal to 0.4871.

The parameter p was set to 0.5 in this case, which means that in the comparison of the rankings using the RBO metric, 50% importance was given to the first position of each ranking, followed by 25% for the second and 12.5% for the third, as shown in Table 4. The importance is distributed in geometric progression; therefore, the sum of the weights given to all positions in a ranking is equal to 1. The lower the value of p, the greater the importance attributed to the first items in the ranking; and the higher the value of p, the more evenly the importance is distributed among the other items, as can be seen in Table 4.

Table 3
Rankings of explainability methods for the image n02085620_8174.

Rankings	**1st**	**2nd**	**3rd**	**4th**	**5th**	**6th**	**7th**	**8th**	**9th**	**RBO**
H	**SSCAM**	**LCAM**	**ISCAM**	**GCAM++**	**XGCAM**	**GCAM**	**ScCAM**	-	-	0.0000
WJ	ScCAM	LCAM	ISCAM	CAM	GCAM	XGCAM	SGCAM++	GCAM++	SSCAM	0.5075
WA	ScCAM	LCAM	SGCAM++	CAM	GCAM	XGCAM	GCAM++	ISCAM	SSCAM	0.6432
BC	ScCAM	LCAM	ISCAM	CAM	GCAM	XGCAM	SGCAM++	GCAM++	SSCAM	0.5075
CA	ScCAM	LCAM	ISCAM	SGCAM++	SSCAM	CAM	GCAM	XGCAM	GCAM++	0.5027
CY	ISCAM	ScCAM	SGCAM++	LCAM	GCAM	XGCAM	CAM	GCAM++	SSCAM	0.6265
MA	ScCAM	LCAM	CAM	GCAM	XGCAM	GCAM++	SGCAM++	ISCAM	SSCAM	0.5908
CR	ScCAM	LCAM	CAM	GCAM	XGCAM	GCAM++	ISCAM	SGCAM++	SSCAM	0.5704
CS	ScCAM	LCAM	CAM	GCAM	XGCAM	GCAM++	ISCAM	SGCAM++	SSCAM	0.5704
EU	ScCAM	LCAM	CAM	GCAM	XGCAM	ISCAM	GCAM++	SGCAM++	SSCAM	0.5704
JS	ScCAM	LCAM	CAM	GCAM	XGCAM	ISCAM	SGCAM++	GCAM++	SSCAM	0.5908
MI	**ScCAM**	**LCAM**	**ISCAM**	**CAM**	**GCAM**	**XGCAM**	**GCAM++**	**SGCAM++**	**SSCAM**	**0.4871**
SE	ScCAM	LCAM	CAM	GCAM	XGCAM	ISCAM	GCAM++	SGCAM++	SSCAM	0.5704

The best metric according to RBO distance (p = 0.5): **Minkowski (MI)**.

Table 4
The weights for each of the first three positions in a ranking according to some examples of values to the parameter p on the RBO metric.

p	First Position	Second Position	Third Position
0.0	100.0%	0.0%	0.0%
0.5	50%	25%	12.5%
0.8	20%	16%	12.8%
0.9	10%	9%	8.1%
1.0	0.0%	0.0%	0.0%

The Table 5 shows the number of times each metric ranking was most similar to the human ranking according to RBO. The header in each column p indicates the value chosen for the parameter p in the RBO. The values in each cell show the total number of images where that metric obtained the best RBO. The best values are highlighted in bold. For some images, there were two or more rankings with the same score.

Table 5
The total of images where each metric obtained the best RBO score.

Metric Ranking	p=0.0	p=0.5	p=0.8	p=0.9	p=1.0
WJ	7	4	4	2	3
WA	5	5	5	3	4
BC	7	4	4	2	3
CA	4	3	2	1	2
CY	1	0	0	1	1
MA	6	**6**	**6**	4	5
CR	**8**	**6**	**6**	**6**	5
CS	6	4	4	4	**6**
EU	6	4	4	2	3
JS	4	2	2	2	4
MI	6	4	5	4	4
SE	6	4	4	2	3

6. Next Research Steps

The objective of this research is to find a better way to choose the best explainability method; therefore, this work proposed a method in which different distance metrics were applied between the annotation heatmap created on several bounding boxes made by humans and the explanation heatmaps created using the CAM-based methods. People were asked to choose the best explanations in a conducted experiment, and a ranking with the results was created to compare with the results of the distance metrics using the RBO metric. Our results suggest that: (i) the best metrics are two: Manhattan and Correlation; and (ii) the best explainability methods were LayerCAM, Score-CAM, and IS-CAM for 4 images each, and SS-CAM for 3, out of the total of 17 chihuahuas images randomly selected from the ImageNet dataset. For the next steps, we want to conduct diverse experiments with new settings to test our method in different aspects; for example: (i) using different CNN architectures; (ii) collecting annotations in polygon shapes; (iii) collecting annotations created using Vision Language Models; and (iv) using images from different classes or other datasets.

Acknowledgments

The authors would like to thank the Brazilian funding agency Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Finance Code 001, for the partial financial support of this research.

Declaration on Generative AI

During the preparation of this work, the author(s) used the spellcheck language tool of Overleaf to check grammar and spelling. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, Cambridge, MA, USA, 2016. http://www.deeplearningbook.org.

[2] G. Ras, N. Xie, M. van Gerven, D. Doran, Explainable deep learning: A field guide for the uninitiated, Journal of Artificial Intelligence Research (JAIR) 73 (2022) 329–397.

[3] J. Colin, T. Fel, R. Cadene, T. Serre, What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods, in: Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 2832–2845.

[4] 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 (CVPR), IEEE, 2016, pp. 2921–2929.

[5] R. Ibrahim, M. O. Shafiq, Explainable convolutional neural networks: A taxonomy, review, and future directions, ACM Computing Surveys (CSUR) 55 (2023) 1–40.

[6] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Computation 1 (1989) 541–551.

[7] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, X. Hu, Score-cam: Score-weighted visual explanations for convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2020, pp. 111–119.

[8] H. Wang, R. Naidu, J. Michael, S. S. Kundu, Ss-cam: Smoothed score-cam for sharper visual feature localization, arXiv preprint arXiv:2006.14255 (2020). URL: https://arxiv.org/abs/2006.14255.

[9] R. Naidu, A. Ghosh, Y. Maurya, S. R. Nayak K, S. S. Kundu, Is-cam: Integrated score-cam for axiomatic-based explanations, arXiv preprint arXiv:2010.03023 (2020). URL: https://arxiv.org/abs/2010.03023.

[10] 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 (ICCV), 2017, pp. 618–626.

[11] A. Chattopadhay, A. Sarkar, P. Howlader, V. N. Balasubramanian, Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2018, pp. 839–847.

[12] D. Omeiza, S. Speakman, C. Cintas, K. Weldermariam, Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models, arXiv preprint arXiv:1908.01224 (2019).

[13] 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).

[14] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, Y. Wei, Layercam: Exploring hierarchical class activation maps for localization, IEEE Transactions on Image Processing 30 (2021) 5875–5888.

[15] K. Morrison, A. Mehra, A. Perer, Shared interest... sometimes: Understanding the alignment between human perception, vision architectures, and saliency map techniques, in: Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3775–3780.

[16] G. Prasad, Y. Nie, M. Bansal, R. Jia, D. Kiela, A. Williams, To what extent do human explanations of model behavior align with actual model behavior?, arXiv preprint arXiv:2012.13354 (2020).

[17] Z. Zhang, J. Singh, U. Gadiraju, A. Anand, Dissonance between human and machine understanding, Proceedings of the ACM on Human-Computer Interaction 3 (2019) 1–23.

[18] 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.

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

[20] W. Webber, A. Moffat, J. Zobel, A similarity measure for indefinite rankings, ACM Transactions on Information Systems (TOIS) 28 (2010) 1–38.