# Post-hoc Explainable Artificial Intelligence for computer vision: model-specific methods

Youssef OUADGHIRI
Ouadghiri.yo@gmail.com
Université Côte d'Azur

February 15, 2022

**Abstract**

Explainable Artificial Intelligence has become a necessary aspect of deep learning due to the increasing use of deep neural networks in use cases where the decision has a big impact, which has pushed researchers to develop many kinds of explaining methods to fit different applications. This diversity in approaches has in turn created a competition between methods in terms of efficiency and understandability. Due to the highly complex nature of deep neural networks, and given the importance of the results, one of the most promising categories involves post-hoc explainability which takes the specific structure of the network into account. This article gives an overview and comparison of such methods applied to the computer vision field.

## 1 Introduction

The Artificial Intelligence (AI) field is continuing to grow exponentially, Deep Neural Networks (DNN) [32] new applications are being invented continuously, AI has brought a tremendous jump in both scientific and industrial fields. Applications involving computer vision [44], Natural Language Processing (NLP) [42, 8] and others [15, 43] have proven to be a massive success and are now frequently found in daily life.

However, AI has not always been perfect, and using it in some critical fields, including healthcare [14], autonomous vehicles [37], or law enforcement [25, 12] where human lives are on the line, could be dangerous. Indeed, a bad disease diagnosis or a car accident could be lethal, a wrongful sentence could change the life of one or many individuals. Moreover, the best-performing networks have a colossal number of parameters(for example GoogleNet [39] uses 7 million parameters trained over millions or hundreds of thousands of Images), making it hard to understand and even harder to trust for people. Furthermore, sometimes, some models' unexplained dysfunctions [40, 20] can cause considerable harm in real-life decisions.

As a result, eXplainable Artificial Intelligence (XAI), has become an important factor to ensure the reliability of AI. Figure 1 emphasizes the increasing trend of the topic in research. The main role of XAI can be summed up in 3 points [2]:

- Achieve impartiality and fairness in decision making.
- Ensure the robustness of the model facing bad data.
- Improve the reliability by making sure the results are solely affected by the meaningful variables.

In the literature, a small confusion between interpretability and explainability prevails. Interpretability means translating to human-understandable terms [9], whilst explainability focus on logical rules to make sense in the human mind [24], in this article, we use them interchangeably.

**Related work:** Many surveys have been done on this topic, different approaches were used, some focused on computer vision [41, 4], some focused on NLP [7], and many were more general [47, 26, 30, 1, 2]. To differ from these previous works, our goal is to deeply analyze the most popular methods, we focus on the post-hoc DNN methods for image data.

## 2 Taxonomy

Although previous papers rely on different ways to count the types of explaining methods and sometimes intertwine them, the most general way to see it would be 4 types of differences:

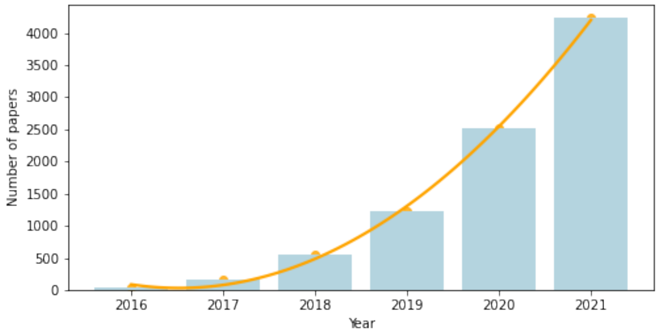

Figure 1: Evolution of the number of publications with ”Explainable Artificial Intelligence” in the title or abstract or keywords during the last 6 years. Data-scraped from Google Scholar (November 4th, 2021) by running research on different years’ parameters. The growth in the number of publications is noticeable.

### Post-hoc vs Ante-hoc

Post-hoc explanation means the method performs some operations to generate a justification after the predictions are made. This type is the most popular, although it does not seek to explain the model mechanism but rather explains the results. Ante-hoc methods, also called intrinsic methods, self-explaining methods, or transparency (as opposed to the opacity of black-boxes) are methods built on the model to generate an explanation of the model mechanism while training and generating predictions [17].

### Agnostic vs Model-specific

Agnostic methods are methods that work on any model, given predictions, whereas the model-specific methods are methods dedicated to some ML algorithms or families of algorithms.

### Local vs Global

Local methods focus on explaining one specific result and constitute the majority of methods used and researched. Global methods attempt to provide a justification for the model as a whole [13, 9].

### Types of explanation

Another way of classifying the types of methods that is less intrinsic is the type of explanation. Many papers use this factor in different ways. In this article, we take the most common and logical way of classification, presented in this paper [47], and distinguish 4 types of explaining methods:

- Explaining by example, explain by looking at a similar case.
- Explaining by attribution, evaluate the importance of certain input.
- Explaining with hidden semantics, finds the explanation for the activation of certain neurons.
- Explaining by rules, extract logical rules

Some examples of popular methods are given in table 1 along with their classification.

## 3 Methods

This article focuses on post-hoc model-specific approaches in Images. In this section, we describe the post-hoc model-specific explaining methods for images and some tools. Computer vision is one of the most important AI notions, it is largely used and made rapid progress. In the DNN area, we mostly use the Convolutional Neural Networks (CNN) for this task. The strength of CNNs is their ability to inspect the local vicinity of an input (pixels in this case). Now, looking at some methods to interpret the results of these kinds of DNNs, we take some examples:

| Method | Type |
|---|---|
| DeconvNet [46] | Post-hoc, Model-specific, Attribution |
| DeepLIFT [33] | Post-hoc, Model-specific, Attribution |
| LRP [3] | Post-hoc, Model-specific, Attribution |
| LIME [28] | Post-hoc, Model-agnostic, Attribution |
| CAM [48] | Post-hoc, Model-specific, Attribution |
| Grad-CAM++ [5] | Post-hoc, Model-specific, Attribution |
| Anchors [29] | Post-hoc, Model-agnostic, Rules |
| Inversion using CNN [18] | Post-hoc, Model-specific, Hidden-semantics |
| Visualization [34, 10] | Post-hoc, Model-specific, Hidden-semantics |
| LSTMvis [38] | Post-hoc, Model-specific, Hidden-Semantics |
| Decision trees [6] | Ante-hoc, Model-specific, Attribution |
| Prediction Difference Analysis [49] | Ante-hoc, Model-specific, Attribution |
| Bayesian Rule List [16] | Ante-hoc, Model-agnostic, Rules |
| DeepRED [19] | Ante-hoc, Model-agnostic, Rules |

Table 1: Examples of some of the most popular methods used in XAI, along with their taxonomy.

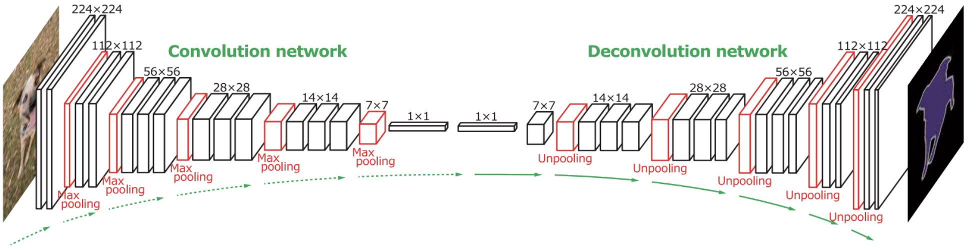

Figure 2: The structure of the proposed DeconvNet, taking the CNN's output as input, uses a sequence of unpooling, deconvolution, and rectification layers, and outputs a heatmap.

## 3.1 DeconvNet

DeconvNet [46] Visualizes the input stimuli, projects feature activation, does a sensitivity analysis of classifier output, and observes the evolution of features during training. DeconvNet was a frank success for semantic Segmentation [21], this is a subtopic of computer vision where we try to figure out the perimeter of a certain object.

The architecture of DeconvNet is illustrated in figure 2, it works through 3 steps:

- Unpooling: as seen in figure 3 allows us to obtain an approximate inverse by recording the locations of the maxima within each pooling region.
- Rectification: CNN uses ReLU non-linearity, rectification tries to invert the function.
- Filtering: DeConvNet uses the transpose of this learned filter to the rectified representation from the step above to reconstruct the deconvolved layer output.

## 3.2 LRP

Layer-wise relevance propagation LRP [3], is a more general XAI method, working for most DNNs including, fully connected layers, CNN, and RNN. It produces a heat-map in the input space

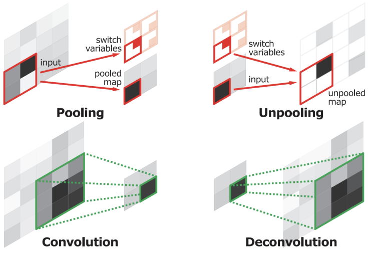

Figure 3: Illustration of how the unpooling and deconvolution steps work in the DeconvNet.

indicating the importance of pixels contributing to the final outcome. The LRP method is able to directly highlight positive contributions in the input space, this decomposition algorithm applies a redistribution rule backward to produce a relevance map. The LRP algorithm takes the output and uses the weights and neural activation and computes the relevancy by propagating the values backward until the input. The algorithm computes a relevance score $R_j$ in each layer j with this simple rule:

$$R_j = \Sigma_k \frac{a_j w_{jk}}{\Sigma_j a_k w_{jk} R_k}$$

Where j and k are neurons of consecutive layers, the first $R_k$ is taken from the output and then it is computed recursively. LRP works admirably with CNNs. The only requirement for LRP is that the DNN uses solely ReLU as an activation function. LRP can be used with transformers for computer vision as well.

## 3.3 Class Activation Map

Class Activation Map CAM [48] is one of the most popular and most efficient techniques in this field, in general, it replaces the last layer of a CNN, the dense layer, and the Softmax with a pooling layer and draws a saliency map with it. CAM methods are derived from this paper [23] where they were able to detect the top left boundaries of objects through Global Max-Pooling GMP. The CAM method uses a global average pooling GAP instead to better detect the object. As seen in figure 4, the GAP finds feature maps and averages the intensity values into scalars, we learn a simple regression model that takes us from these scalars to each of the class labels on the last layer, then it performs a weighted sum of all of the activation maps. Each activation map contains different spatial information about the input, and when the selected convolutional layer is close to the classification stage of the network, its activations are sufficiently high-level to provide a visual localization that explains the final prediction. However, the method has some limitations. It needs to rerun the CNN model, resulting in it being slow. And it does not work with all models. Consequently, this method has been developed into many versions:

- GradCAM [31] solves the retraining problem by using the calculated gradients as weights, it then uses a ReLU function to make values positive and representable on the saliency map, a guided backpropagation using the same gradients can be added for better results. In general, GradCAM is faster, and more accurate than the original CAM. Yet, it has its limitations, it performs surprisingly bad with multiple occurrences as the gradient focuses on the biggest object, moreover, it also fails to accurately capture occlusion.

- GradCAM++ [5] is an improved version of the GradCAM, it uses the weighted average of pixel-wise gradients, instead of considering all pixel gradients equally. This results in better explanations, especially for multiple instances and occlusion.

- SmoothGrad [35] Creates copies of image input and creates noise then averages the gradient, the result is sharpening in the saliency map, removing the noise on the way. The strength of this technique is the ability to be combined with other techniques. It is not very accurate nor novel as it follows the idea of the stochastic gradient descent, but it opens the door for new more powerful techniques.

- SmoothGradCAM++ [22] combines the SmoothGrad and GradCAM++ methods, it produces sharper more accurate explanations.

- FullGrad [36] decomposes the neural net response into input sensitivity, it satisfies the completeness and weak dependence properties, the FullGrad saliency map is obtained by aggregating the full-gradient components. It is a robust and precise variation of GradCAM.

- XGradCAM [11] or Axiom-based GradCAM uses another way to compute the weights which are also the gradients inside the CNN, improving the GradCAM visualizations.

- Non-gradient CAM methods are also gaining in notoriety. The gradient might suffer from a saturation resulting in diminished gradient in backpropagation and consequently, bad visualizations. Such methods include AblationCAM [27] subtracts the activation of a precise non-null region to a null region to approximate the slope to get weights, in other words, AblationCAM tests the importance of each part of the image by removing it or ablating it and seeing if it provokes a dip in the prediction score. And ScoreCAM [45] obtains the weight of each activation map through its forward passing score on target class, the final result is obtained by a linear combination of weights and activation maps.

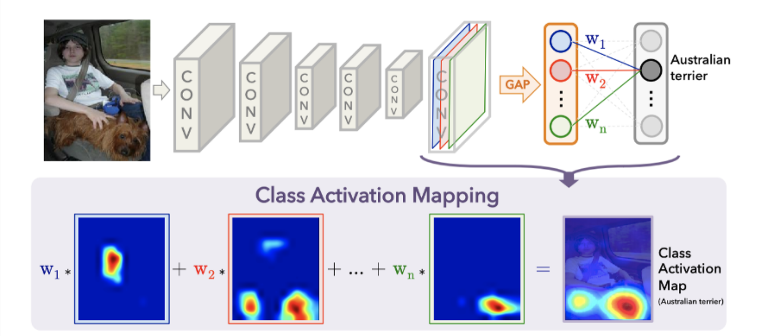

Figure 4: CAM architecture, the GAP detects feature maps, for example, the green layer, and averages all the intensity values into one single scalar for each map. Each of these scalars here represents one feature map. A weighted sum of the activation maps shows their contributions.

# 4 Experiment

## 4.1 Representation methods

To represent the results of XAI on images, we usually use a saliency map, which takes forms of pictures highlighting the most important pixels in determining the output of our DNN. A saliency map can be a heatmap, where the pixels are smoothed by area, warmer colors indicate high importance (i.e. red is extremely important, blue is indecisive). A saliency map can also take an empty frame and draw the map of the decisive pixels, the degree of influence is correlated with the opacity. The third kind of saliency map we introduce is edge detection, this method helps us see more accurately and can be great for semantic segmentation. We use these types of saliency maps interchangeably in this article.

## 4.2 Parameters

For this experiment, we use birds photographs findable on the internet and do image classification. The models we use are a pre-trained VGG16, resnet18, and resnet50 for the ImageNet dataset. The models yield great results for predicting our birds. However, the goal of the experiment is not about the quality of prediction, but the explanation of the classification or misclassification. Using more than one model is, nevertheless, important to prove the robustness of our explanator. The computer vision domain usually finds difficulties with pictures containing a high number of occurrences or occlusion. Thus, we do experiments with 1, 2, and many occluded birds' pictures. Working on Pytorch, we use some XAI available libraries like Facebook's Torchray library and Pytorch_grad_cam. The explaining methods work better with medium-sized images, we use 626x418 pixels images.

## 4.3 Results and comparison

For this application, there is no set metric to quantitatively compare the results of our methods, Annotation boxes from object detection methods can be used, however, they do not provide a reliable metric, especially with unlabeled data. Thus, the comparison will be done qualitatively by looking at images and comparing them visually.

Figure 5 shows the results, obtained for 1 occurrence, 2 occurrences, and multiple occluded occurrences. As expected, the explanators have a harder time detecting the birds in overlapping conditions. DeconvNet barely captures the bird in the one bird image, and completely fails in the other options. The DeconvNet baseline does normally not provide a clear explanation as it follows a difficult idea, indeed, the max-pooling function is non-invertible, the unpooling, rectification, and filtering try to solve the problem, but it is not always conclusive. DeconvNet, however, is a technique that was improved to work for semantic segmentation, we did not cover these variations as they are out of scope.

LRP captures the objects admirably, although we can see some noise in the first 2 images, detection of branches and flowers. The third image remains very challenging, LRP captured well the edges but a lot of noise was equally captured.

As expected again, the accuracies of the GradCAM methods are close with a little improvement in every improved method. FullGrad detects the whole bird and performs better in multi-instances

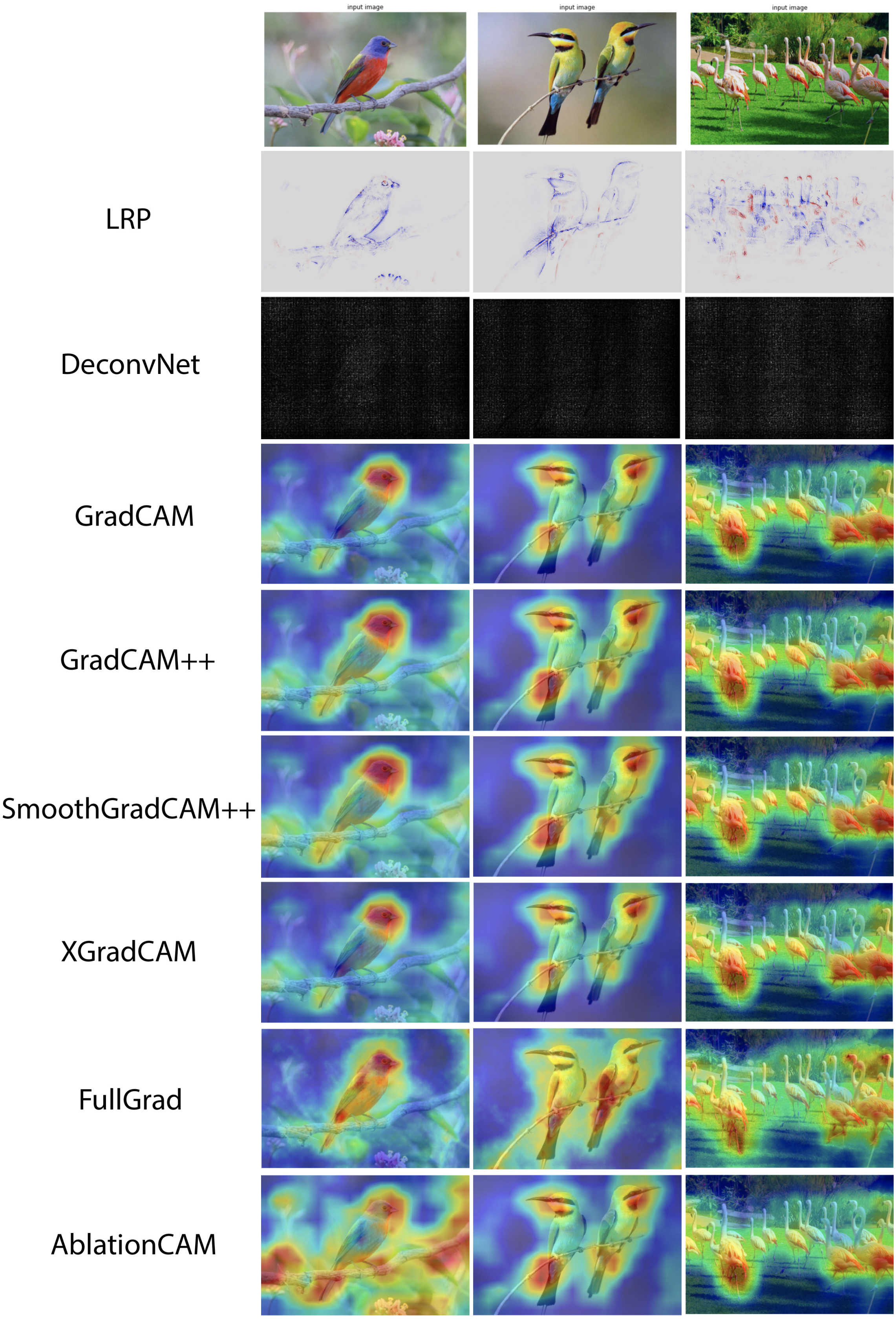

Figure 5: Saliency map for different explaining methods, first row is the input images in 626x418 size, each row is a different explaining algorithm.

images, but it also adds some noise. FullGrad uses a lot of RAM and it can be challenging for big images. Notice that some methods completely fail at capturing some situations like AblationCAM for one bird image: as mentioned before, AblationCAM is dependable on the predictor, and thus can show weaknesses sometimes. Non-gradient methods take a much longer computation time, it is proportional to the image size as there are more activation maps.

The methods operate in different ways and combining the results or using multiple explanators can be the best way to detect accurately the subject, for example, combining FullGrad which detected the whole body and noise with SmoothGrad++ which shows less noise can give a complementary explanation.

# 5 Conclusion

In this article, we defined XAI and presented its taxonomies according to multiple surveys. We then examined the most predominant type for DNNs: Post-hoc model-specific introduced some popular families of methods for computer vision explanation and tried to compare the most popular ones by implementing them. The different techniques come up with different explanations. For our experiment, SmoothGradCAM++ could extract with confidence the features that helped determine the birds in the images: head, feet, or color. On the other hand, LRP produces a good visualization of the shape of the body. In the light of the availability of different approaches, we intuitively conclude that combining the explanations or visualizing more than one method can be the best way to obtain a faithful visualization.

In the future, XAI should be used in applied research in many fields to explore the possibilities and more challenges. Accordingly, more advances will be made in this field. XAI for computer vision can also help improve prediction models: by detecting dysfunctions, we can further preprocess our images data. The focus has started to shift toward transformers for computer vision, as we have seen, some of the introduced methods can be used for them, however, more research and surveys are expected.

# References

[1] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

[2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[3] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[4] Vanessa Buhrmester, David Münch, and Michael Arens. Analysis of explainers of black box deep neural networks for computer vision: A survey. *arXiv preprint arXiv:1911.12116*, 2019.

[5] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[6] Mark Craven and Jude Shavlik. Extracting tree-structured representations of trained networks. *Advances in neural information processing systems*, 8:24–30, 1995.

[7] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. A survey of the state of explainable ai for natural language processing. *arXiv preprint arXiv:2010.00711*, 2020.

[8] Li Deng and Yang Liu. *Deep learning in natural language processing.* Springer, 2018.

[9] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[10] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

[11] Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns, 2020.

[12] Janneke Gerards and Raphaele Xenidis. Algorithmic discrimination in europe: Challenges and opportunities for gender equality and non-discrimination law. 2021.

[13] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[14] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[15] Sampo Kuutti, Richard Bowden, Yaochu Jin, Phil Barber, and Saber Fallah. A survey of deep learning applications to autonomous vehicle control. *IEEE Transactions on Intelligent Transportation Systems*, 22(2):712–733, 2020.

[16] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, and David Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350–1371, 2015.

[17] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[18] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015.

[19] Gary Mataev, Peyman Milanfar, and Michael Elad. Deepred: Deep image prior powered by red. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[20] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436, 2015.

[21] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.

[22] Daniel Omeiza, Skyler Speakman, Celia Cintas, and Komminist Weldermariam. Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models. *arXiv preprint arXiv:1908.01224*, 2019.

[23] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 685–694, 2015.

[24] Dino Pedreschi, Fosca Giannotti, Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, and Franco Turini. Meaningful explanations of black box ai decision systems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9780–9784, 2019.

[25] Stephan Raaijmakers. Artificial intelligence for law enforcement: Challenges and opportunities. *IEEE Security  Privacy*, 17:74–77, 09 2019.

[26] Arun Rai. Explainable ai: From black box to glass box. *Journal of the Academy of Marketing Science*, 48(1):137–141, 2020.

[27] Harish Guruprasad Ramaswamy et al. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 983–991, 2020.

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos 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*, pages 1135–1144, 2016.

[29] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[30] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021.

[31] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[32] Pramila P Shinde and Seema Shah. A review of machine learning and deep learning applications. In *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*, pages 1–6. IEEE, 2018.

[33] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR, 2017.

[34] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[35] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[36] Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. *Advances in neural information processing systems*, 32, 2019.

[37] Jack Stilgoe. How can we know a self-driving car is safe? *Ethics and Information Technology*, pages 1–13, 2021.

[38] Hendrik Strobelt, Sebastian Gehrmann, Hanspeter Pfister, and Alexander M Rush. Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE transactions on visualization and computer graphics*, 24(1):667–676, 2017.

[39] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[40] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, D. Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2014.

[41] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[42] Amirsina Torfi, Rouzbeh A Shirvani, Yaser Keneshloo, Nader Tavaf, and Edward A Fox. Natural language processing advancements by deep learning: A survey. *arXiv preprint arXiv:2003.01200*, 2020.

[43] Raju Vaishya, Mohd Javaid, Ibrahim Haleem Khan, and Abid Haleem. Artificial intelligence (ai) applications for covid-19 pandemic. *Diabetes& Metabolic Syndrome: Clinical Research& Reviews*, 14(4):337–339, 2020.

[44] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018.

[45] Haofan Wang, Mengnan Du, Fan Yang, and Zijian Zhang. Score-cam: Improved visual explanations via score-weighted class activation mapping. 2019.

[46] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[47] Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2021.

[48] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[49] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*, 2017.