Post-hoc Explainable Artificial Intelligence for deep learning: model specific methods

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

November 12, 2021

Abstract

Explainable Artificial Intelligence has become a necessary aspect of deep learning due to the widespread of deep neural networks in delicate matters where the decision has a big impact, the advancement pushed researchers to develop many kinds of explaining methods to fit different applications, this diversity in methods 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, more focus was given to specific post-hoc methods.

1 Introduction

The Artificial Intelligence (AI) field is continuing to grow exponentially, Deep Neural Networks (DNN) [38] new applications are being invented continuously, AI has brought a tremendous jump in both scientific and industrial fields. Applications involving computer vision [51], Natural Language Processing (NLP) [47, 11] and others [23, 48] 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 in healthcare [22], autonomous vehicles [42], or law enforcement [32, 16] 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 [44] 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 [45, 28] can cause big harm.

As a result, eXplainable Artificial Intelligence (XAI), has become an important factor to maintain 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.
- Improving the reliability by making sure the results are solely affected by the meaningful variables.

In the literature, researchers tend to have a small confusion between interpretability and explainability. Interpretability means translating to human understandable terms [13], whilst explainability focus on logical rules to make sense in the human mind [30], in this article, we use them interchangeably (as many papers did) to avoid confusion.

Related work: Many surveys have been done on this topic, different approaches were used, some focused on computer vision [46, 4], some focused on NLP [10], and many were more general [54, 33, 36, 1, 2], we do not cite all of them as they are numerous, but contrarily to these previous papers, we do not focus on quantity, we analyze deeply the most popular methods, and we focus on the post-hoc DNN methods. Additionally, we try to include some methods for a more challenging type of networks: Transformers[49].

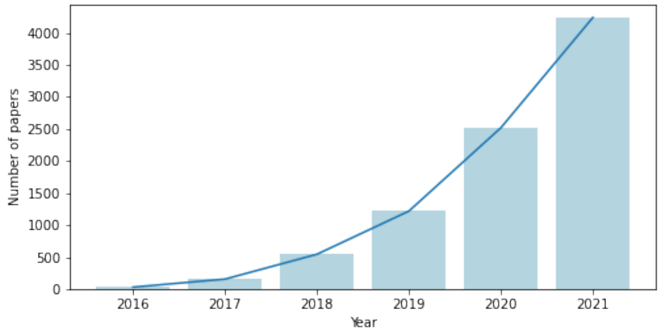

Figure 1: Evolution of the number of publications with "Explainable Artificial Intelligence" in the title or abstract or keywords during the last 6 years. Data scraped from Google Scholar (November 4th, 2021) by running research on different years' parameters. The growth in the number of publications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

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.

2.1 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[25].

2.2 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.

2.3 Local vs Global

Local methods focus on one specific result or one and explain it, it makes the majority of methods used and researched. Global methods are less common and they try to provide a justification about the model work process [18, 13].

2.4 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 [54], and distinguish 4 types of explaining methods:

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

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

Method	Type
DeconvNet [53]	Post-hoc, Model-specific, Attribution
DeepLIFT [39]	Post-hoc, Model-specific, Attribution
LRP [3]	Post-hoc, Model-specific, Attribution
LIME [34]	Post-hoc, Model-agnostic, Attribution
CAM [55]	Post-hoc, Model-specific, Attribution
Grad-CAM++ [5]	Post-hoc, Model-specific, Attribution
Anchors [35]	Post-hoc, Model-agnostic, Rules
Inversion using CNN [26]	Post-hoc, Model-specific, Hidden-semantics
Visualization [40, 15]	Post-hoc, Model-specific, Hidden-semantics
LSTMvis [43]	Post-hoc, Model-specific, Hidden-Semantics
BertViz [50]	Post-hoc, Model-specific, Hidden-Semantics
ExBert [21]	Post-hoc, Model-specific, Hidden-Semantics
Decision trees [9]	Ante-hoc, Model-specific, Attribution
Prediction Difference Analysis [56]	Ante-hoc, Model-specific, Attribution
Bayesian Rule List [24]	Ante-hoc, Model-agnostic, Rules
DeepRED [27]	Ante-hoc, Model-agnostic, Rules

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

3 Post-hoc model specific

This article focuses on post-hoc model specific in 2 use cases: Image and text data. In this section, we describe the most popular post-hoc model specific explaining methods for the two types of data. In the future, we are going to try to implement, evaluate and compare these methods with other methods.

3.1 Image data

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, however, we have seen an outbreak of transformer in this area recently. The power of CNN is inspecting the local vicinity of an input (pixels in this case). The main property of transformers is attention [49], they focus on the positional data of an input. Visual transformer (ViT) [14] divides the image into smaller image patches with positional embedding then processes them as a sequence, and transformers were proven to handle sequential data well [12]. Results show that transformers perform considerably well in this task, although they need a large quantity of data.

Now looking at some methods to interpret the results of these kinds of DNNs, we take some examples:

- DeconvNet [53]: Visualizes the input stimuli, projects feature activations, does a sensitivity Analysis of classifier output, and observes the evolution of features during training. Unpooling: we can obtain an approximate inverse by recording the locations of the maxima within each pooling region. Rectification: ConvNet uses ReLU non-linearity Filtering:DeConvNet uses the transpose of this learned filter to the rectified representation from the step above to reconstruct the deconvolved layer output.

- Smooth Grad-CAM++ [29] : this method regroups a number of methods, first the Class Activation Map (CAM) [55], this method replaces the last layer of DNN architecture with a Global Average Pooling (GAP) layer to draw a saliency map on the most influencing pixels, then variants were derived, Gradient-based CAM (Grad-CAM) [37], this method is a generalization of CAM and works on any CNN architecture, then Grad-CAM++ [5] uses the weighted combination of the positive partial derivatives of the last convolutional layer feature maps with respect to a specific class score as weights to generate a visual explanation for the class label under consideration. SMOOTHGRAD [41] sharpens gradient-based sensitivity map by taking random samples in the neighborhood. And the smooth Grad-CAM++ is the combination of the 2 latter methods, in figurewe can see the improvement of the results according to the method.

- Transformer Interpretability Beyond Attention Visualization [7] : The method employs LRP-based relevance to compute scores for each attention head in each layer. It then integrates these scores throughout the attention graph, by incorporating both relevancy and gradient

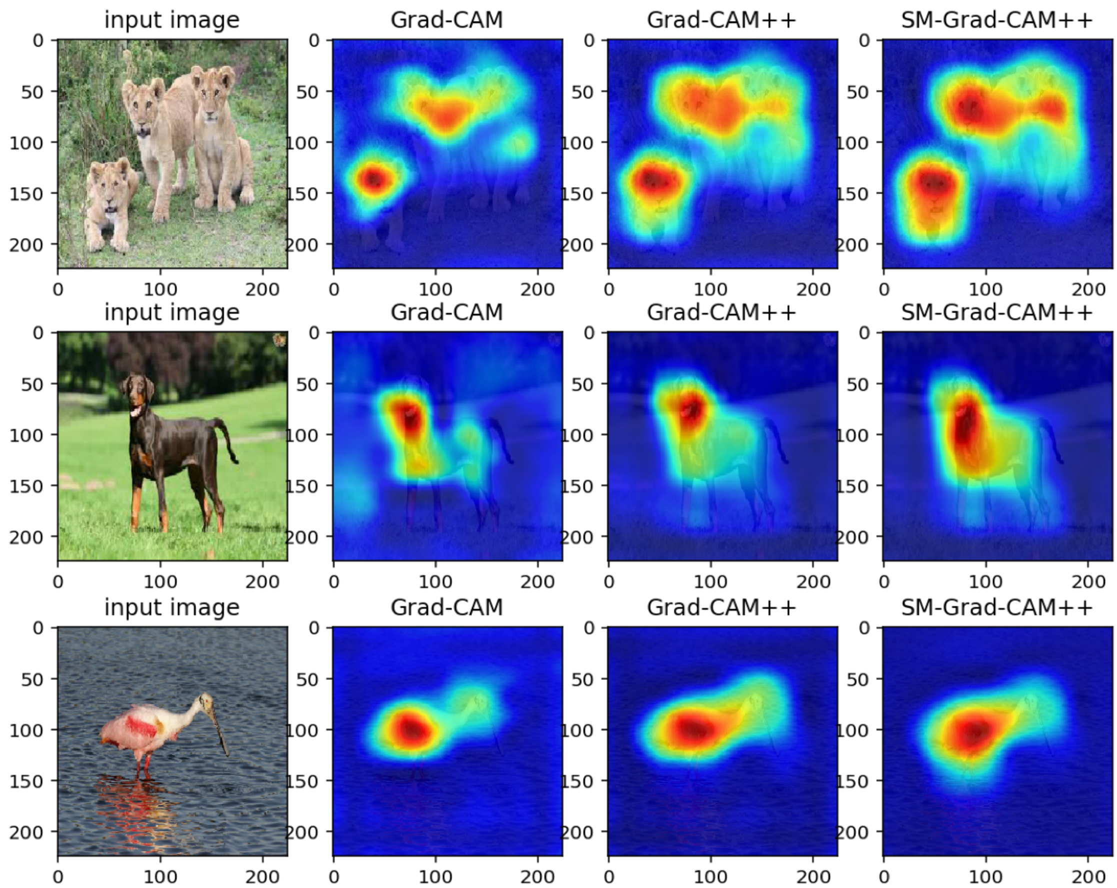

Figure 2: CAM saliency maps

information, in a way that iteratively removes the negative contributions. The result is a class specific visualization for self-attention models. As the transformers are a recent research topic, not many explainers are available but many more general and more effective methods are being explored and published currently (Example: [6, 17])

3.2 Text data

A recent XAI for NLP paper by IBM [10, 31] has presented 5 techniques for explanation that best fit the NLP tasks:

1. Feature importance: Uses the score of the importance of features.
2. Surrogate model: Using a second model to explain the predictions of the first one.
3. Example driven: Uses an example from the labels that have semantic similarities.
4. Provenance-based: illustrates the derivation of predictions.
5. Induction: uses human readable representations such as rules, trees or programs.

The study also emphasized visualization techniques, the most commonly used technique is Attention-based or saliency heat-map of an attention score. In this section we enumerate the most popular and practical methods for XAI in NLP following the post-hoc DNN specific parameters, we introduce methods for DNN in general, LSTM and transformers.

- LRP [3]: Layer-wise relevance propagation, this 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 indicating the importance of each feature 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.

Figure 3: An overview of the different components of the tool. The token "escape" is selected and masked at 0-[all]. The results from a corpus search by token embedding are shown and summarized in (d-g). Users can enter a sentence in (a) and modify the attention view through selections in (b). Self attention is displayed in (c). The blue matrices show the attention of a head (column) to a token (row). Tokens and heads that are selected in (c) can be searched over the annotated corpus (shown: Wizard of Oz) with results presented in the corpus view. Every token in Corpus view displays its linguistic metadata on hover. A colored summary of the matched token and its context is shown on its left.

- Attention [8]: the attention mechanism is an intuitive technique to humans, it creates a weighted context vector by inducing conditional distributions over inputs. It is still controversial regarding how much it explains[52, 19] as some researchers claim that it has not been formally evaluated.
- LSTMvis [43]: LSTM [20] specific method, allows a user to select a hypothesis input range to focus on local context, to match these contexts with similar patterns in a large data set. We provide data for the tool to analyze specific hidden state properties on dataset containing nesting, phrase structure, and chord progressions, and demonstrate how the tool can be used to isolate patterns for further statistical analysis.
- ExBERT [21]: an interactive tool to visualize and formulate hypothesis for the BERT model [12] reasoning process, it presents a good insight of the context and attention by matching a human labeled input to similar contexts. Figure shows an example and how the attention scores results are shown.

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] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 397–406, October 2021.

[7] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021.

[8] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. *arXiv preprint arXiv:1506.07503*, 2015.

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

[10] 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.

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

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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

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

[17] Jacob Gildenblat. Explainability for vision transformers. https://jacobgil.github.io/deeplearning/vision-transformer-explainability, 2020.

[18] 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.

[19] Christopher Grimsley, Elijah Mayfield, and Julia Bursten. Why attention is not explanation: Surgical intervention and causal reasoning about neural models. 2020.

[20] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[21] Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. exbert: A visual analysis tool to explore learned representations in transformers models. *arXiv preprint arXiv:1910.05276*, 2019.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] Kun Qian, Marina Danilevsky, Yannis Katsis, Ban Kawas, Erick Oduor, Lucian Popa, and Yunyao Li. Xnlp: A living survey for xai research in natural language processing. In *26th International Conference on Intelligent User Interfaces*, pages 78–80, 2021.

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

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

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

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

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

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

[43] 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.

[44] 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.

[45] 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.

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

[47] 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.

[48] 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.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[50] Jesse Vig. Bertviz: A tool for visualizing multihead self-attention in the bert model. In *ICLR Workshop: Debugging Machine Learning Models*, 2019.

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

[52] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*, 2019.

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

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

[55] 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.

[56] 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.