Preprint and Pre-review Version

Collected work provisionally entitled: Applications of Fuzzy Techniques: Proceedings of the 2023 Annual Conference of the North American Fuzzy Information Processing Society NAFIPS Book series: Advances in Intelligent Systems and Computing

Forcing the Network to use Human Explanations in its Inference Process

Javier Viaña and Andrew Vanderburg

Abstract We introduce the concept of ForcedNet, a neural network that has been trained to generate a simplified version of human-like explanations in its hidden layers. The main difference with a regular network is that the ForcedNet has been educated such that its inner reasoning reproduces certain patterns that could be somewhat considered as human-understandable explanations. If designed appropriately, a ForcedNet can increase the model's transparency and explainability. We also propose the use of support features, hidden variables that complement the explanations and contain additional information to achieve high performance while the explanation contains the most important features of the layer. We define the optimal value of support features and what analysis can be performed to select this parameter. We demonstrate a simple ForcedNet case for image reconstruction using as explanation the composite image of the saliency map that is intended to mimic the focus of the human eye. The primary objective of this work is to promote the use of intermediate explanations in neural networks and encourage deep learning development modules to integrate the possibility of creating networks like the proposed ForcedNets.

Keywords: Artificial intelligence, deep learning, neural networks, explainable AI, explainability, AI ethics, image processing, machine learning

Javier Viaña and Andrew Vanderburg
MIT Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge 02139, USA
e-mail: `vianajr@mit.edu`

1 Introduction

We are in a time when the exponential use of artificial intelligence is prioritizing performance over the ability to explain and justify the outcomes from a human perspective. The field of eXplainable AI (XAI), which started becoming popular in 2016 [1], has been slowly developing over the last decade, but there is not yet a generalized method to understand the internal inference process of the widely used deep neural networks [2, 3, 4, 5]. Nevertheless, some essential ideas such as extracting meaning from neural networks or relying on existing prior expert knowledge were already studied in the 90's [6, 7], which years later evolved into the field of XAI.

This need for understanding becomes more acute with the entry of legislation such as the General Data Protection Regulation in the European Union on algorithmic decision-making and the "right to explanation" when it comes to human data [8, 9, 10]. However, many of the XAI technologies used for tasks that leverage human data, such as user experience enhancement or travel demand analysis, do not meet yet these requirements [11, 12].

In healthcare, the integration of AI largely depends on the trustworthiness of the algorithm chosen. Explainability is playing a critical role in order to achieve the validation and verification capabilities desired [13]. In fact, this need for trustworthiness has fostered the development of novel XAI for specific medical applications that can later be extended to other areas [14, 15, 16].

In engineering, the opposite happened, where human supervision became obsolete in those processes that were automated with black box AI and it was not until the lack of transparency was evident that the awareness of XAI began to be raised. Over time, several methods have appeared that claim to be explainable in the field, e.g., from the generative design of motors [17], to their prognosis, health monitoring and fault diagnosis [18, 19].

Many of the techniques used to generate explanations in neural networks focus on what are known as post-hoc explanations. This means that once the system is trained, we add different algorithms that can extract posterior reasoning [20, 21, 22]. Some examples include, semantic web technologies [23], contrastive sample generation (GRACE) [24], or extracting global symbolic rules [25]. Nonetheless, opting for methods that extract explanations after the training implies that the network is not "aware" of our intention to explain its reasoning. In other words, the opportunity to add this information during training is lost, which is not only attractive for performance reasons, but also to educate the pipeline so that its inner process is more human understandable.

Some techniques such as DeConvNet [26], Guided BackProp [27], Layer-wise Relevance Propagation [28] and the Deep Taylor Decomposition [29] seek to explain the classifier decisions by propagating the output back through the network to map the most relevant features of the encoding process. However, more recent work [30] has shown that these methods do not produce valid explanations for simple linear models. An alternative approach that has demonstrated to be very successful is generating more interpretable latent spaces in autoencoders [31, 32, 33, 34]. These often encode the information taking into account expert knowledge in order to improve the understanding of the latent variables. Nevertheless, their training does not rely on any ground truth latent explanation. Symmetric autoencoders have also been utilized to represent coherent information via reordering of the data [35], which can certainly help to understand the underlying decisions of the machine.

Another option is to develop new algorithms that are more transparent by nature and also high performing [36]. The main drawback is that such development implies a bottom-up reformulation of the neural networks and the learning formulas of the backpropagation. CEFYDRA is an example of these novel net-based algorithms that, in its case, replaces the neural unit with a fuzzy inference system in order to reason its outputs [37, 38, 39]. Combining case-based reasoning with deep learning has made significant advances in XAI as well, where the networks leverage already seen scenarios and adapt their solutions to solve new problems [40, 41, 42]. Other researchers have considered adding constrains to the learning process [43, 44, 45] but it has not been done with the intention of forcing the machine to think or replicate human reasoning. On the other hand, there have also been attempts to improve explainability by simulating human reasoning, but not within the neural network itself [46].

2 Architecture

2.1 The ForcedNet

The growth of XAI has raised concerns about the quality of the explanations used to explain the algorithms [47, 48, 49, 50]. Logically, an explanation is valid as long as it's understandable for the person who digests it. Therefore, its validity depends on the recipient, which makes that assessment a highly subjective task. Regardless, if we consider the explanations as one more feature of the dataset, we could even integrate them into the training process, to educate the machine in a human manner. Such human-like reasoning could possibly help opening the black box [51], [52], [53].

In this work we introduce the concept of ForcedNet, a neural network that has been "forced" to produce a simplified version of a human-like explanation in one of its intermediate layers and then leverages fully or partially this explanation to generate the desired output. We define these simplified versions human-like explanations as any type of information that can help understand the reasoning process of the algorithm from the human perspective. The choice of the best explanation format is problem specific, and even for one same task there might be several types of useful explanations, such as visual or textual. As it was mentioned in the introduction, the validity of an explanation is subjective since it involves the perception and assessment of a human, which might vary.

Fig. 1 is the depiction of an example ForcedNet architecture. In addition to the usual inputs and outputs, we also have the desired explanations, and the support

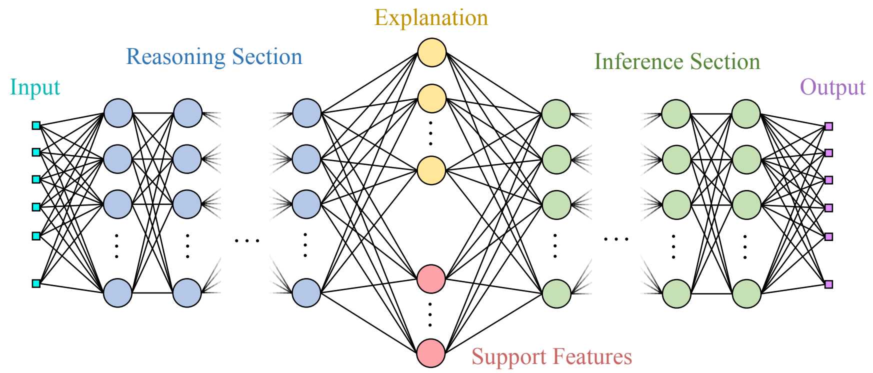

Fig. 1 Schematic representation of the proposed network, ForcedNet, composed of a central explanation layer that contains the human-understandable explanation and the support features.

features (whose purpose will be described later) in what we identify as the explanation layer of the network. This layer divides the left and right sections which are the reasoning and inference sections, respectively.

The reasoning section is responsible for producing the explanation of the input. This explanation should be a description that not only encodes the information of the input but at the same time is understandable from the human perspective. The inference section is responsible for generating the desired output from the human-understandable explanation.

Generating the final output using only the information encoded in the explanation might be a difficult task. Indeed, the performance of the inference section depends on the quality of the explanation chosen for the problem. For that matter, we introduce the concept of support features, which precisely support the explanation features encapsulating additional information of the input that is not present in the explanation. These features are not understandable from the human perspective as the explanation features, but in some problems might be necessary. To grasp their importance, let us consider the following image reconstruction task depicted in Fig. 2, where $\boldsymbol{x}$, and $\boldsymbol{y}$ denote the input and output respectively. The explanation chosen, $\boldsymbol{t}$, is a short linguistic description of the image. The support vector (support features arranged in a vector format), $\boldsymbol{s}$, has S features that have no apparent meaning, but together with the embedded explanation, the inference section is able to retrieve most of the image. In Fig. 2, $\hat{\boldsymbol{v}}$ denotes a prediction of the variable $\boldsymbol{v}$.

The training of such a system can leverage a triple backpropation method, where in each step we train the reasoning section, the inference section, and the full system separately as shown in Fig. 3. Note that the backpropagation in the reasoning and the inference sections excludes the weights associated to the neurons of the support vector because we have no prior data for these features.

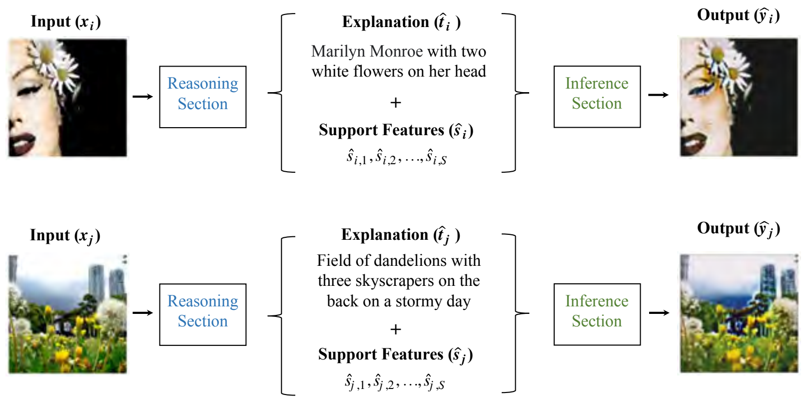

Fig. 2 Two examples of a ForcedNet that uses the image embedding as the explanation and performs the image reconstruction task based on this textual definition of the image. Images publicly available at Flickr [54]. This example is schematic and only serves to understand better the idea behind a ForcedNet.

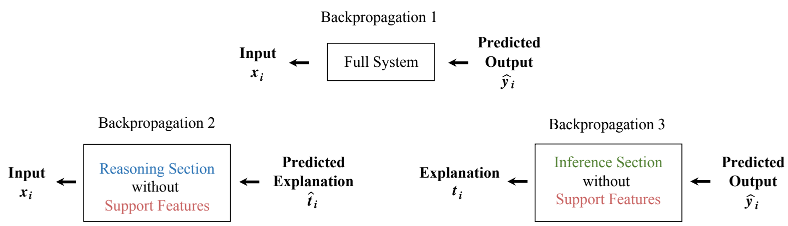

Fig. 3 Schematic representation of the triple backpropagation method for the training of a ForcedNet, given a single explanation layer.

2.2 Design Considerations

Choosing the best number of support features should be evaluated carefully and is problem specific. There is an optimal trade-off between explainability (low S) and performance (high S). Alternatively, one can study the evolution in performance of the following two pipelines as we vary the value of S to make a better decision:

- Network A, which uses the explanation of T fixed number of features and S support features, i.e., a total of $T+S$ hidden features.
- Network B, which only uses S hidden features, and it does not produce any explanation.

Fixing the same training hyperparameters, we increase the value of S in both networks and keep track of their performance, which we denote by ρ_A and ρ_B respectively (same figure of merit must be the chosen, e.g., for regression tasks: Root Mean Squared Error, or Mean Absolute Error, etc.). When the $\rho_A \approx \rho_B$ for a given value S, it means that network A is not using the explanation features to produce the

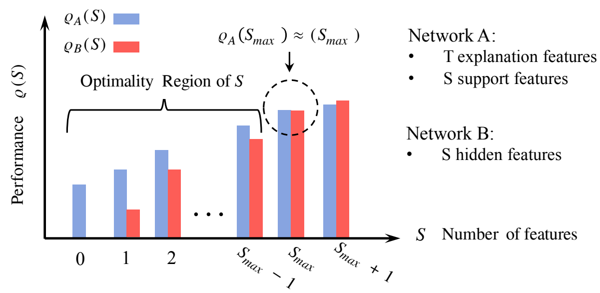

Fig. 4 Performance comparison for two different networks as we increase the value of S while we fix the value of T. Network A is a ForcedNet that utilizes T explanation features together with S support features. Network B is a standard autoencoder that uses S hidden features.

outputs. We denote this limiting value of S with S_{max} (Fig. 4). In other words, the network A has sufficient information with the S_{max} support features to fulfill its task, since network B has been able to obtain similar performance without the use of the F features. Thus, the explanations generated by A are meaningless because pipeline A may not be using them at all. The optimal value S, is $S_{opt} \in \mathbb{Z}^+ : 0 \leqslant S_{opt} < S_{max}$, which marks the trade-off between performance and explainability, because it ensures that the explanation features are indeed used in the reasoning to generate the output.

3 Case Study

We chose the image reconstruction task to exemplify a simple proof of concept of a ForcedNet. To generate the intermediate explanations, we decided to leverage the saliency map ($\boldsymbol{m}$) of the image, which defines the regions of the image on which the human eye focuses first. Our explanation is composed of both the original image and the saliency map. We injected Gaussian noise and increased the transparency proportionally to those pixels that were less important according to the saliency map, while we kept resolved those areas that were more relevant. In other words, in the explanation we tried to capture the way in which the human eye focuses on an image, distorting the surrounding information while locating the machine's center of attention on the vital pixels.

For the reasoning section of the network, we utilized DeepGaze II [55], a pre-trained model that has demonstrated the best performance to date predicting saliency maps on datasets like MIT300 saliency benchmark [56], where it reported the following metrics: AUC = 88%, sAUC = 77%, NSS = 2.34. DeepGaze II was trained in two phases. In its pre-training phase it used the SALICON dataset [57], which consists of 10000 images with pseudofixations from a mouse-contingent task, and was fine-tunned using the MIT1003 dataset [58].

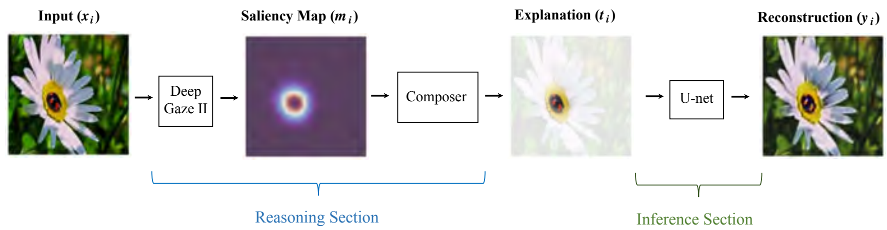

Fig. 5 Processing pipeline of the image reconstruction task fully based on the human-like attention map (the explanation).

Because we are using a pre-trained model for the reasoning section we cannot consider the use of support features. However, the scope of this research was simply to introduce the idea of ForcedNets and to illustrate a simple example with the reconstruction task chosen. In future work, we will study the effect of different support features in a ForcedNet trained back-to-back.

For the inference section, we chose a shallow convolutional autoencoder. The specifications of the architecture are shown in Table 1.

Table 1 Architecture of the convolutional autoencoder chosen.

Layer	Filter	Kernel Size	Activation Function	Padding	Strides
Convolution 2D	16	3x3	ReLu	Same	2
Convolution 2D	8	3x3	ReLu	Same	2
Deconvolution 2D	8	3x3	ReLu	Same	2
Deconvolution 2D	16	3x3	ReLu	Same	2
Convolution 2D	3	3x3	Sigmoid	Same	2

We chose a dataset of 1,000 RGB flower images of 128x128 pixels obtained from Flickr [54]. This selection of images is publicly available in [59].

First, we predicted the feature map of all the images using DeepGaze II, and then we obtained the composed image that served as the explanation matrix. We then trained the convolutional autoencoder with the generated explanations and the desired outputs. Fig. 5 shows the different steps of the architecture.

For the training of the inference section, we chose a learning rate of 10^{-4} for $3,000$ epochs with 20 steps each. We used 0.7, 0.2, 0.1, training, validation and testing splits.

Since the reasoning section of the ForcedNet was already trained we did not require the triple backpropagation method. However, for future reference, the reasoning and the inference sections should be trained simultaneously so that both can learn from all the three types of data (inputs, outputs, and explanations).

Fig. 6 Eight training instances of the dataset. First row represents the original image, second row the saliency map predicted by DeepGaze II, third row the composed attention map (the human-like explanation), row four the output of the convolutional autoencoder (reconstructed image), row five the difference image between the input and the output of the system in normalized space to appreciate better the errors.

Additionally, DeepGaze II required an additional reference image map that measured the human bias to look at pixels on the central region of the image. This map was generated accordingly as described in [55] for the chosen dataset.

4 Results

After training, the Mean Absolute Error (MAE) for 10-fold cross validation was 0.0107, 0.0109, 0.0108 for each set, respectively. To calculate the MAE we used the average error for all the pixels of the image.

Fig. 6 and Fig. 7 show some prediction examples for both the training and the testing datasets of the ForcedNet considered. It can be appreciated that the convolutional autoencoder is able to reconstruct most of the image using only the information available on the explanation data. At the same time, the explanation layer gives us that additional knowledge in the prediction of the areas in the image where the machine focuses more.

Fig. 7 Eight testing instances of the dataset. First row represents the original image, second row the saliency map predicted by DeepGaze II, third row the composed attention map (the human-like explanation), row four the output of the convolutional autoencoder (reconstructed image), row five the difference image between the input and the output of the system in normalized space to appreciate better the errors.

5 Discussion

With this study, we are trying to demonstrate that is viable to force the network to have intermediate layers where we have explanations, and that there are different but simple ways in which we can conceive simplified behaviors of human reasoning that might even help the machine learning process.

Having the explainability requirement in the middle of the pipeline might increase the difficulty of the learning process, as it imposes a constrain in the machine's internal inference process. However, there might be scenarios where we see the opposite, e.g., if the human description contains useful information for the task chosen, it could even help the training guiding the weights to more optimal combinations. On the other hand, it does require an extra effort from the human to generate the dataset of explanation training instances.

The architecture explained in this study is scalable to designs with several explainable layers. In other words, one could stack different reasoning sections together to form a chain of explanations as shown in Fig. 8. That kind of pipeline would guide the thought process of the network even more than the case studied. The motivation behind such a system lies in validation and verification purposes or simply because we decide to constrain the internal inference process of the network in a smarter way than letting the optimization to brute force input-to-output

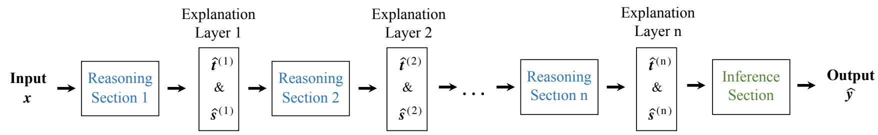

Fig. 8 Schematic representation of a ForcedNet composed of several explanation layers. Each explanation layer is generated by the corresponding reasoning section, but there is only one inference section.

backpropagation. The number of backpropagation methods that one can perform simultaneously in each step given n explanation layers is $\frac{n^2+3n+2}{2}$, which comes from the combinations of $n+2$ elements (n explanation layers and both the extremes of the pipeline) grouped in pairs. Note that n^2+3n will always generate a positive even number if $n \in \mathbb{Z}^+$.

Future work could focus on demonstrating the use of ForcedNets in image reconstruction leveraging the image embedding as the explanation. Further experiments should also be conducted to test the benefit of support features and the optimal choice of S.

The authors believe that the community would be benefited from an open-source module for neural network development that could automate plugging explanations in the middle of the network and choosing for the right number of support features.

6 Conclusions

We have presented the concept of ForcedNets as a tool to guide the learning of a neural network to generate intermediate human-understandable explanations. This technique has been demonstrated with a simple image reconstruction example where the explanation is a composition image of the saliency map. The use of support features has been discussed as a method to ensure high performance even when the explanations are too simple or do not capture all the necessary information of the previous layers. The authors believe that this technique could significantly help to obtain the desired explainability in neural networks, as long as there are explanations for the chosen problem and that these can be easily incorporated into the learning process.

7 Acknowledgments

For their insightful discussions on the topic and comments on the paper, we thank Faraz Faruqi and Mariona Badenas from Massachusetts Institute of Technology, Jagoba Zuluaga and Eric del Sastre from Harvard Medical School. The authors would also like to extend special thanks to the anonymous reviewers that peer-reviewed this work.

8 Authors' contributions

Made the conception and design of the study and performed data analysis and interpretation: Javier Viaña. Supervised the work: Andrew Vanderburg.

9 Availability of data and materials

All images in this archive are licensed under the Creative Commons By-Attribution License. Data publicly available in Flickr [54], selection of training data stored in GitHub [59].

Code developed for the project publicly available in GitHub[59]. Pre-trained DeepGaze II model available in GitHub[60].

10 Financial support and sponsorship

This work was supported by two NASA Grants, the NASA Extremely Precise Radial Velocity Foundation Science Program (No. 80NSSC22K0848) and the NASA Astrophysical Data Analysis Program (No. 80NSSC22K1408).

References

1. David Gunning. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2(2), 2017.
2. Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. Explainable AI: A brief survey on history, research areas, approaches and challenges. In Jie Tang, Min-Yen Kan, Dongyan Zhao, Sujian Li, and Hongying Zan, editors, *Natural Language Processing and Chinese*, pages 563–574, Cham, Computing 2019. Springer International Publishing. ID: 10.1007/978-3-030-32236-6_51.
3. Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.
4. Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.
5. Plamen Angelov and Eduardo Soares. Towards explainable deep neural networks (xDNN). *Neural Networks*, 130:185–194, 2020. ID: 271125.
6. Harry Turner and Tamás D. Gedeon. Extracting meaning from neural networks. In *Proceedings 13th International Conference on AI*, volume 1, pages 243–252, 1993.

7. Sebastian Thrun. *Explanation-Based Neural Network Learning*, pages 19–48. Explanation-Based Neural Network Learning: A Lifelong Learning Approach. Springer US, Boston, MA, 1996. ID: Thrun1996.
8. Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57, 2017.
9. Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2):76–99, 2017.
10. Ronan Hamon, Henrik Junklewitz, Ignacio Sanchez, Gianclaudio Malgieri, and Paul De Hert. Bridging the gap between AI and explainability in the GDPR: Towards trustworthiness-by-design in automated decision-making. *IEEE Computational Intelligence Magazine*, 17(1):72–85, 2022.
11. Juliana J. Ferreira and Mateus S. Monteiro. What are people doing about XAI user experience? A survey on AI explainability research and practice. In *International Conference on Human-Computer Interaction*, pages 56–73. Springer, 2020.
12. Ahmad Alwosheel, Sander van Cranenburgh, and Caspar G. Chorus. Why did you predict that? Towards explainable artificial neural networks for travel demand analysis. *Transportation Research Part C: Emerging Technologies*, 128:103143, 2021. ID: 271729.
13. Aniek F. Markus, Jan A. Kors, and Peter R. Rijnbeek. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113, 2021.
14. Ran Gu, Guotai Wang, Tao Song, Rui Huang, Michael Aertsen, Jan Deprest, Sébastien Ourselin, Tom Vercauteren, and Shaoting Zhang. CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE Transactions on Medical Imaging*, 40(2):699–711, 2021.
15. Chandra Mohan Dasari and Raju Bhukya. Explainable deep neural networks for novel viral genome prediction. *Applied Intelligence*, 52(3):3002–3017, 2022.
16. Carlo Biffi, Ozan Oktay, Giacomo Tarroni, Wenjia Bai, Antonio De Marvao, Georgia Doumou, Martin Rajchl, Reem Bedair, Sanjay Prasad, Stuart Cook, Declan O'Regan, and Daniel Rueckert. Learning interpretable anatomical features through deep generative models: Application to cardiac remodeling. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 20*, pages 464–471, Cham, 18 2018. Springer International Publishing.
17. Hidenori Sasaki, Yuki Hidaka, and Hajime Igarashi. Explainable deep neural network for design of electric motors. *IEEE Transactions on Magnetics*, 57(6):1–4, 2021.
18. John Grezmak, Peng Wang, Chuang Sun, and Robert X. Gao. Explainable convolutional neural network for gearbox fault diagnosis. *Procedia CIRP*, 80:476–481, 2019.
19. Min Su Kim, Jong Pil Yun, and PooGyeon Park. An explainable convolutional neural network for fault diagnosis in linear motion guide. *IEEE Transactions on Industrial Informatics*, 17(6):4036–4045, 2020.
20. Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, and Mani Srivastava. How can I explain this to you? an empirical study of deep neural network explanation methods. *Advances in Neural Information Processing Systems*, 33:4211–4222, 2020.
21. Mark T. Keane and Eoin M. Kenny. How case-based reasoning explains neural networks: A theoretical analysis of XAI using post-hoc explanation-by-example from a survey of ANN-CBR twin-systems. In Kerstin Bach and Cindy Marling, editors, *Case-Based Reasoning Research and*, pages 155–171, Cham, Development 2019. Springer International Publishing. ID: 10.1007/978-3-030-29249-2_11.
22. Quanshi Zhang, Yu Yang, Yuchen Liu, Ying Nian Wu, and Song-Chun Zhu. Unsupervised learning of neural networks to explain neural networks. *arXiv preprint arXiv:1805.07468*, 2018.
23. Md Kamruzzaman Sarker, Ning Xie, Derek Doran, Michael Raymer, and Pascal Hitzler. Explaining trained neural networks with semantic web technologies: First steps. *arXiv preprint arXiv:1710.04324*, 2017.

24. Thai Le, Suhang Wang, and Dongwon Lee. GRACE: Generating concise and informative contrastive sample to explain neural network model's prediction. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 238–248, 2020.
25. Zhi-Hua Zhou, Yuan Jiang, and Shi-Fu Chen. Extracting symbolic rules from trained neural network ensembles. *Ai Communications*, 16(1):3–15, 2003.
26. Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 20*, pages 818–833, Cham, 14 2014. Springer International Publishing. ID: 10.1007/978-3-319-10590-1_53.
27. Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015.
28. Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In *International Conference on Artificial Neural Networks*, pages 63–71. Springer, 2016.
29. Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.
30. Pieter-Jan Kindermans, Kristof T. Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne. Learning how to explain neural networks: PatternNet and PatternAttributionn. *arXiv preprint arXiv:1705.05598*, 2017.
31. M. Neumeier, M. Botsch, A. Tollkühn, and T. Berberich. Variational autoencoder-based vehicle trajectory prediction with an interpretable latent space. pages 820–827, 2021.
32. Jin-Young Kim and Sung-Bae Cho. Explainable prediction of electric energy demand using a deep autoencoder with interpretable latent space. *Expert Systems with Applications*, 186:115842, 2021. ID: 271506.
33. Francesco Bodria, Riccardo Guidotti, Fosca Giannotti, and Dino Pedreschi. Interpretable latent space to enable counterfactual explanations. In Poncelet Pascal and Dino Ienco, editors, *Discovery*, pages 525–540, Cham, Science 2022. Springer Nature Switzerland. ID: 10.1007/978-3-031-18840-4_37.
34. Kutay Bölat and Tufan Kumbasar. Interpreting variational autoencoders with fuzzy logic: A step towards interpretable deep learning based fuzzy classifiers. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–7, 2020.
35. Pawan Bharadwaj, Matthew Li, and Laurent Demanet. Redatuming physical systems using symmetric autoencoders. *Physical Review Research*, 4(2), 2022.
36. Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, May 1, 2019.
37. Javier Viaña, Stephan Ralescu, Vladik Kreinovich, Anca Ralescu, and Kelly Cohen. Single hidden layer CEFYDRA: Cluster-first Explainable FuzzY-based Deep self-Reorganizing Algorithm. In Scott Dick, Vladik Kreinovich, and Pawan Lingras, editors, *Applications of Fuzzy Techniques*, pages 298–307, Cham, 2023. Springer International Publishing. ID: 10.1007/978-3-031-16038-7_29.
38. Javier Viaña, Stephan Ralescu, Vladik Kreinovich, Anca Ralescu, and Kelly Cohen. Multiple hidden layered CEFYDRA: Cluster-first Explainable FuzzY-based Deep self-Reorganizing Algorithm. In Scott Dick, Vladik Kreinovich, and Pawan Lingras, editors, *Applications of Fuzzy Techniques*, pages 308–322, Cham, 2023. Springer International Publishing. ID: 10.1007/978-3-031-16038-7_30.
39. Javier Viaña, Stephan Ralescu, Vladik Kreinovich, Anca Ralescu, and Kelly Cohen. Initialization and plasticity of CEFYDRA: Cluster-first Explainable FuzzY-based Deep self-Reorganizing Algorithm. In Scott Dick, Vladik Kreinovich, and Pawan Lingras, editors, *Appli-*

cations of Fuzzy, pages 323–335, Cham, Techniques 2023. Springer International Publishing. ID: 10.1007/978-3-031-16038-7_31.
40. Jae Heon Park, Chung-Kwan Shin, Kwang Hyuk Im, and Sang Chan Park. A local weighting method to the integration of neural network and case based reasoning. In *Neural Networks for Signal Processing XI: Proceedings of the 2001 IEEE Signal Processing Society Workshop (IEEE Cat. No. 01TH8584)*, pages 33–42. IEEE, 2001.
41. Kareem Amin, Stelios Kapetanakis, Klaus-Dieter Althoff, Andreas Dengel, and Miltos Petridis. Answering with cases: A CBR approach to deep learning. In *International Conference on Case-Based Reasoning*, pages 15–27. Springer, 2018.
42. Lisa Corbat, Mohammad Nauval, Julien Henriet, and Jean-Christophe Lapayre. A fusion method based on deep learning and case-based reasoning which improves the resulting medical image segmentations. *Expert Systems with Applications*, 147:113200, 2020.
43. Zebin Yang, Aijun Zhang, and Agus Sudjianto. Enhancing explainability of neural networks through architecture constraints. *IEEE Transactions on Neural Networks and Learning Systems*, 32(6):2610–2621, 2021.
44. Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge. In *International conference on machine learning*, pages 8116–8126. PMLR, 2020.
45. Jude W. Shavlik and Geoffrey G. Towell. Combining explanation-based learning and artificial neural networks. In *Proceedings of the Sixth International Workshop on Machine Learning*, pages 90–92. Elsevier, 1989.
46. Paul J. Blazek and Milo M. Lin. Explainable neural networks that simulate reasoning. *Nature Computational Science*, 1(9):607–618, 2021. ID: Blazek2021.
47. Thomas Fel, David Vigouroux, Rémi Cadène, and Thomas Serre. How good is your explanation? Algorithmic stability measures to assess the quality of explanations for deep neural networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 720–730, 2022.
48. Adam J. Johs, Meaghan Lutts, and Rosina O. Weber. Measuring explanation quality in XCBR. In *Proceedings of the 26th International Conference on Case-Based Reasoning*, page 75. Springer International Publishing, 2018.
49. 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.
50. Ann-Kathrin Dombrowski, Christopher J. Anders, Klaus-Robert Müller, and Pan Kessel. Towards robust explanations for deep neural networks. *Pattern Recognition*, 121:108194, 2022.
51. Davide Castelvecchi. Can we open the black box of AI? *Nature News*, 538(7623):20, 2016.
52. Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, Feb 2019.
53. Eoin M. Kenny, Courtney Ford, Molly Quinn, and Mark T. Keane. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in xai user studies. *Artificial intelligence*, 294:103459, May 2021.
54. Flickr. `www.flicker.com`. Accessed: 2022-11-16.
55. Matthias Kümmerer, Thomas S.A. Wallis, Leon A. Gatys, and Matthias Bethge. Understanding low- and high-level contributions to fixation prediction. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4799–4808, 2017.
56. Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. MIT saliency benchmark.
57. Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. SALICON: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1072–1080, 2015.
58. Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.
59. Javier Viaña. ForcedNet for image reconstruction. `https://github.com/JavierVianaAi/forcednets-image-reconstruction`, 2022.
60. Matthias Kummerer. DeepGaze. `https://github.com/matthias-k/DeepGaze`, 2022.