Analysis of Explainability of Deep Learning Models for Medical Applicability

Minds Brains and Machines (MBM) — Master in Artificial Intelligence
Universitat Politècnica de Catalunya

Gonzalo Recio Domènech (gonzalo.recio@fib.upc.edu) — June 2020

Abstract: *Deep learning is achieving outstanding results recently in neuroscience and health diagnosing. However, due to the lack of explainability of these kind of models, they cannot be used in clinical environments because they arise many legal and ethical concerns. This work aims to explore the explainability and interpretability techniques that can be used to extract explanations from deep network models and analyze up to which point they are explainable. Also, future research directions are discussed.*

1 Introduction

Model explainability is one of the most important problems in machine learning nowadays. It is often the case that certain "black box" models such as deep neural networks (DNN) are deployed to production and are running critical systems. It can be scary to think that not even the developers of these algorithms understand why exactly the algorithms make the decisions they do - or even worse, how to prevent an adversary from exploiting them. Many classical machine learning models (ML) can deal with explainability due to their simplicity (e.g. rule-based methods) but with the emergence of deep learning (DL) models interpretability issue becomes worse.

Deep learning methods have proven to achieve state-of-the-art results in Artificial Intelligence tasks such as pattern recognition and problem modeling. Deep learning models are usually based in connectionism theory, which hopes to explain mental phenomena using artificial neural networks (ANN). Learning is achieved through an iterative process of adjustment of the interconnections between the artificial neurons within the network, much like in the human brain [1]. Some advantages of the connectionist approach include its applicability to a broad amount of functions, structural approximation to biological neurons and low requirements for innate structure. Besides, some disadvantages include the difficulty in interpreting how ANNs process information and explaining phenomena at a higher level [2].

Despite AI and deep learning's ability to create accurate predictions and classifications, in most cases it lacks the ability to provide a mechanistic understanding of how inputs and outputs relate to each other: the complex patterns learned by these models are very difficult to interpret or explain. In domains where a decision can have serious consequences (e.g., medical diagnosis, autonomous driving, criminal justice, finance decisions, etc.), it is especially important that the decision-making models are transparent. Otherwise, the use of opaque models in these settings may arise serious legal and ethical concerns. There is extensive evidence for the importance of explanation towards understanding and building trust, not only in machine learning research [3, 4], but also in cognitive psychology [5] and philosophy [6]. This has led to a new research area called Explainable Artificial Intelligence (XAI)

[7, 8], which aims to produce more explainable models, while maintaining a high level of learning performance (prediction accuracy), and enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent technologies.

1.1 Can AI be explainable?

The field is split about the potential and need for AI to be explainable and/or interpretable [9, 10]. Some view AI as a tool for solving a technical problem but not necessarily useful for answering a scientific question. Others think it may indeed be possible for AI actions to be interpreted and/or understood by humans, but it depends on the level of understanding being sought.

A related question is whether AI solutions can be explainable to the point of providing mechanistic insights into how the brain is accomplishing a particular function or a set of complex behaviors. Currently, there is a significant gap between the performance of explainable biophysical models for prediction and that of more opaque ANNs. Maybe the point is to wonder if it is reasonable to expect synthetic AI algorithms and architectures to be informative of the underlying biological process. Perhaps the AI process are not the same as the brain process. It may be that AI solutions are explainable (in abstraction) but inherently uninterpretable in the context of the underlying biology.

In any case, explanations can at least give insights and help improve the AI performance [11], and at least least achieving a certain degree of explainability is more than enough for introducing the use of complex "black-box" models into clinical settings.

2 Machine Learning and Deep Learning in Neuroscience

Many classic machine learning (ML) algorithms have a lot of techniques for interpreting the model behavior. For instance, decision trees can be seen as a set of rules over the input features. Also, in linear models, several feature importance algorithms can be applied to get relative interpretation of what the model expects for doing predictions. Even though, this is not that case in deep learning, and in this work we are going to focus on explainability and interpretability in DL models.

Over the last decade, deep learning has been very successful, especially in tasks that involve images and texts such as image classification and language translation. The success story of deep neural networks (DNN) began in 2012, when the ImageNet image classification challenge was won by a deep learning approach (AlexNet [12]). Since then, we have witnessed a significant explosion of deep neural network architectures, with a trend towards deeper networks with more and more weight parameters.

Besides, in the recent years, deep learning models have achieved the state-of-the-art results and accuracies in diagnosing for many diseases [13, 14], for instance, a survey made in 2020 about diagnosing Alzheimer's disease shows that deep learning models are leading in prediction effectiveness [15]. At the beginning, deep learning focused on tasks such as radiological image classification and segmentation. These tasks are uniquely suited to deep learning due to the high-dimensional nature of neuroimaging data which is unfavorable to manual analysis, combined with the digital nature of modern imaging. Later, deep learning has been applied to functional brain mapping and correlational studies using functional magnetic resonance imaging (fMRI) data for tasks such as prediction of postoperative seizure [14]. Lastly, diagnostic prognostication with deep learning using multiple data types, including laboratory values, images, speech recordings, time-series, among others, has been used to diagnose disease risk.

Most of this models are trained with neuroimaging using MRI, fMRI, volumetric images, electroencephalography, X-rays, etc. [14]. Figure 1 illustrates the most common machine learning algorithms present in medical AI research, and we can see that DL models are gaining considerable presence.

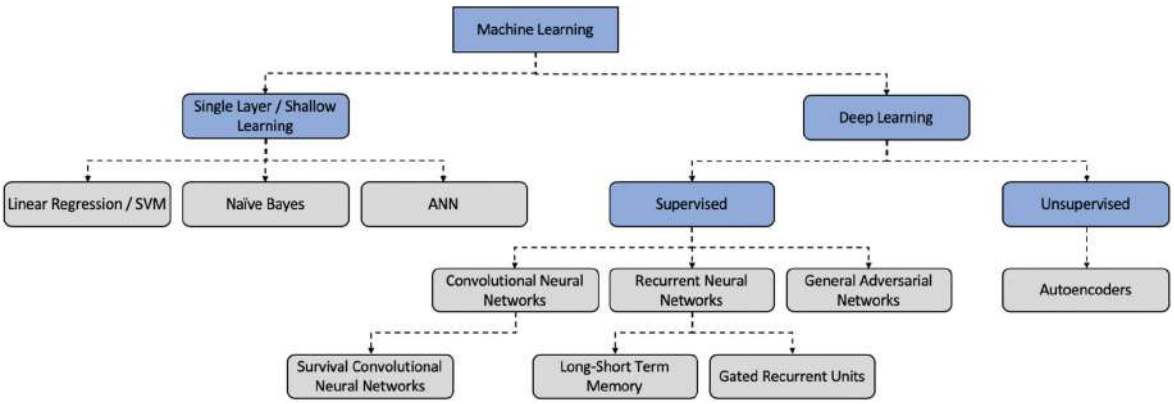


Figure 1: Breakdown of algorithm types in the machine learning family that are commonly used in medical subdomain research and analyses [13].

As mentioned before, in the literature we can find plenty of multimodal models; that is, they are fed with data of different types, for instance, neuroimaging with language speech [16]. In these cases researchers usually make use of combined DL architectures to build the models (e.g. CNNs with RNNs). Although the performance of these models is reaching new accuracy horizons, the main concern is the lack of explainablility due to the large amount of tuneable weights and parameters, which gets even more difficult when different DL architectures are combined. In medical settings, the responsibility of a diagnose or clinical conclusion cannot be accepted if it comes from a "black-box": it generates an accountability gap and it is not legal and ethical to not having a grounded explanation for medical decisions [17].

2.1 Deep Neural Network models

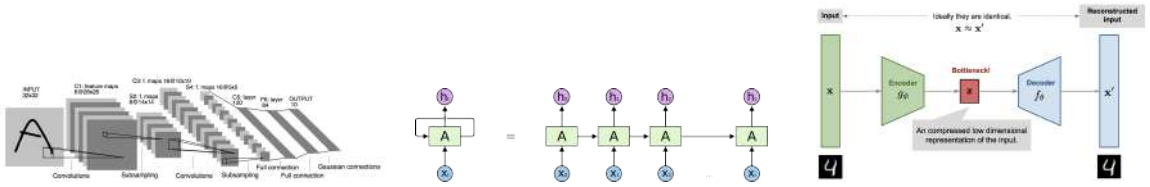
Deep learning differs from traditional machine learning in how representations are automatically discovered from raw data. In contrast to ANNs, which are shallow feature learning techniques, deep learning algorithms employ multiple, deep layers of perceptrons that capture both low- and high-level representations of data, enabling them to learn richer abstractions of inputs [18]. This obviates the need for manual engineering of features and allows deep learning models to naturally discover previously unknown patterns and generalize better to novel data. Variants of these algorithms have been employed across numerous domains in engineering and medicine.

Convolutional neural networks (CNNs) have gained particular attention within computer vision and imaging-based medical research [19]. CNNs gather representations across multiple layers, each of which learns specific features of the image, much like the human visual cortex is arranged into hierarchical layers, including the primary visual cortex (edge detection), secondary visual cortex (shape detection), and so on [20]. CNNs consist of convolutional layers in which data features are learned: pooling layers, which reduce the number of features, and therefore computational demand, by aggregating similar or redundant features; dropout layers, which selectively turn off perceptrons to avoid over-reliance on a single component of the network; and a final output layer, which collates the learned features into a score or class decision, i.e., whether or not a given radiograph shows signs of a certain disease. These algorithms have achieved rapid profound success in image classification tasks and, in some cases, have matched expert human performance [21, 22].

Recurrent neural networks and variants, such as long short-term memory (LSTM) and gated recurrent units, have revolutionized the analysis of time-series data that can be found in videos, speech, and texts [23]. These algorithms sequentially analyze each element of input data and employ

a gating mechanism to determine whether to maintain or discard information from prior elements when generating outputs. In this manner, they efficiently capture long-term dependencies and have revolutionized machine translation, speech processing, and text analysis.

Autoencoders (AEs) are a class of unsupervised learning algorithms that discover meaningful representations of data by learning a lower-dimensional mapping from inputs to outputs [24]. They are composed of an encoder, which learns a latent representation of the input, and a decoder, which reconstructs the input from the latent representation. By constraining the latent representation to a lower dimensionality than the input, AEs are able to learn a compressed representation of data that contains only the features necessary to reconstruct the input. Such algorithms are often employed to learn features that can be subsequently utilized in conjunction with the deep learning techniques previously discussed.



CNN RNN AE

Figure 2: Schemes of most common DNN architectures applied to medicine and neuroscience AI.

3 Explainability and Interpretability

Deep neural networks (DNN) have provided a new burst of research in the machine learning community. However, their complexity obfuscates the underlying processes that drive their inferences. This has lead to a new field of *explainable AI* (XAI) [13]. A variety of tools are being developed to enable researchers to peer into the inner workings of DNNs. In this section we are going to overview the main explainability and interpretability techniques currently used in the field. Firstly, we are going to define the related terms to this topic according to the scientific community [25]:

- **Interpretability**: It is defined as the ability to explain or to provide the meaning in understandable terms to a human.
- **Explainability**: Explainability is associated with the notion of explanation as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans
- **Transparency**: A model is considered to be transparent if by itself it is understandable.

There is a growing sense that neural networks need to be interpretable to humans. The field of neural network interpretability has formed in response to these concerns. As it matures, two major threads of research have begun to join together: *feature visualization* and *attribution/saliency*. However, in the literature we can find a more general categorization for the explainability techniques as proposed in Tjoa and Guan (2019) [26]:

- *Perceptive interpretability*: include in this category interpretabilities that can be humanly perceived, often one that will considered as obvious.
  - *Saliency*: Saliency (or attribution) method explains the decision of an algorithm by assigning values that reflect the importance of input components in their contribution to that decision.
  - *Signal Method*: Methods of interpretability that observe the stimulation of neurons or a collection of neurons are called signal methods.

  - *Verbal interpretability*: This form of interpretability takes the form of verbal chunks that human can grasp naturally. Examples include sentences that indicate causality.

- *Interpretability by mathematical structures*: Mathematical structures have been used to reveal the mechanisms of ML and NN algorithms. Deeper layers of NN are shown to store complex information while shallower layers store simpler information. Some examples would be *feature extraction* and *activation maximization*.

Usually, most practical methods combine some of these explanation techniques categories in the same approach (e.g. [27] combining saliency maps and attention layers on text). In the following sections we aim to explore some of the most successful methods that have shown interesting results in the recent literature of deep learning.

3.1 Feature visualization

Feature visualization answers questions about *what* a network - or parts of a network - are looking for by generating examples. It consists in generating the input that maximizes certain activations of the neural networks. First approach related to feature visualization can be found in the work of Erhan et al. (2009) [28] where they aim to explore which features are learned by a CNN on MNIST dataset. In a ANN, given a unit i and a layer j; h_{ij} is a function of both θ and the input sample x. Assuming a fixed θ (for instance, the parameters after training the network), we can view the idea as looking for

$$x^* = \underset{x}{\text{argmax}}\ h_{ij}(\theta, x)$$

where input x^* can be obtained using gradient ascend in the input space, i.e. computing the gradient of $h_{ij}(\theta, x)$ and moving x in the direction of this gradient (maximize h_{ij}). However, this does not work for complex deep networks and, in order to obtain realistic and meaningful generated inputs, some operations, for instance, of iterative input blurring, need to be performed.

This technique is called *activation maximization*, and it can be performed on different parts of a network and depending on the optimization objectives, we might get different feature visualization levels. For example, in figure 3 we can observe that feature can be visualized from individual neuron activation level to feature/channel level, layer level, class logits level or class probability level [29].

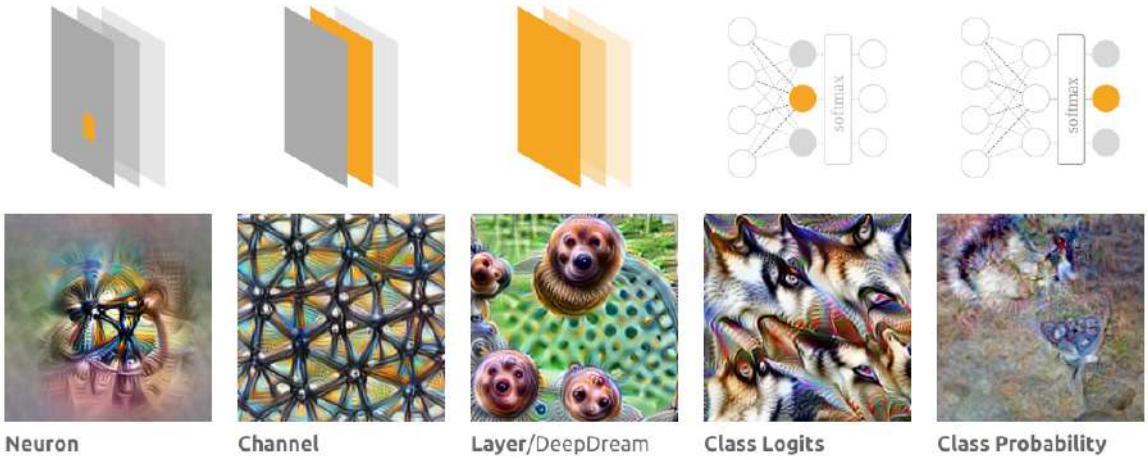


Figure 3: Activation maximization from different network levels [30].

We can see that this technique can offer us interpretability of what the model is looking for at different levels of the input. In order to get an idea of the input that the network expects, if we want to understand individual features, we can search for examples where they have high values - either for a neuron at an individual position, or for an entire channel. Since we are interested in understanding and find explanations about the 2-dimensional input, we have focused in the channel objective for activation maximization. The following images present what GoogLeNet network [31] has learned from training with ImageNET dataset by using activation maximization method [30]. Figure 4 illustrates some of the features that captures one of the first convolutional layers, which shows some interesting textures and basic patterns. Each neuron only looks at a small receptive field, so these channel visualizations show a tiling of them. In deeper layers, figure 5, patterns get complex enough that it can often help to look at the neuron objective rather than the channel objective. We can find neurons responding house shapes, to dogs with leashes, many wheel detectors, and many more.

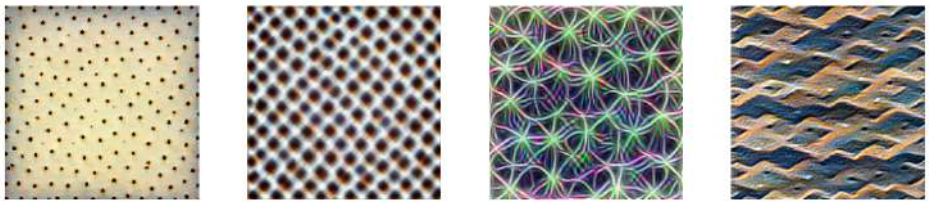


Figure 4: Features captured by the 3rd block of convolutional layers of GoogLeNet.

Figure 5: Features captured by the 4th block of convolutional layers of GoogLeNet.

This shows that with activation maximization we can have access to the explainability at different levels of the networks. This can somehow explain what the different parts of the network expects to fire a category and at different levels of abstraction (from simple features to complex ones). This method introduces some principle of transparency to DNNs and helps to have a clearer insight of what's happening inside the "black box".

3.2 Saliency Maps

Saliency maps (also called *attribution*) study *which part* of an input example is responsible for the network activating a particular way for deep convolutional neural networks [32]. Saliency maps are generally used with image or video processing applications and they are supposed to show what parts of an image or video frame are most important to a network's decisions. For example, figure 6 shows the saliency maps of a CNN showing what the module focused for predicting a duck. New approaches are appearing in the very recent years [33, 34] showing great results and performance for explaining the *why* of the network output prediction.

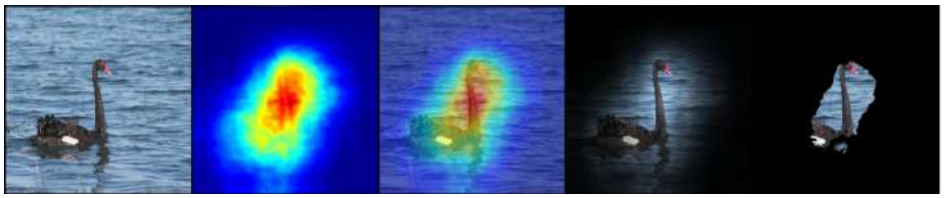


Figure 6: Saliency Maps to detect where the CNN focuses for prediction [33]

In figure 7, we can observe an example architecture to compute saliency maps from a recent proposed work [34]. Given a "black box" model (assuming we cannot access to its parameters, features or gradients), the key idea is to probe the base model by sub-sampling the input image via random masks and recording its response to each of the masked images. The final importance or saliency map is generated as a linear combination of the random binary masks where the combination weights come from the output probabilities predicted by the base model on the masked images. This apparently simple yet surprisingly powerful approach allows to peek inside an arbitrary network without accessing any of its internal structure.

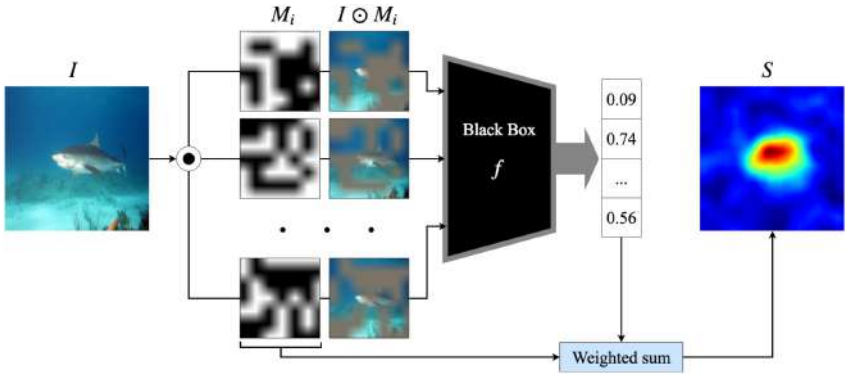


Figure 7: Input image I is element-wise multiplied with random masks M_i and the masked images are fed to the base model. The saliency map is a linear combination of the masks where the weights come from the score of the target class corresponding to the respective masked inputs [34]

3.3 Explainable Deep Models

An other interesting approach for obtaining model explanation is to directly train it for solving the task of explanation along the task of classification. Several works have been proposed using recurrent neural networks combined with existing classification architectures to generate natural language explanations of the classifications [35, 36, 37]. The success of these architectures rely on the assumption that network's computation and reasoning is represented in its internal layer activations. Figure 8 shows the architecture of InterpNET [35], which consists of a combination of a CNN with a RNN to be able to perform both image classification and natural language explanation task. As it shown, the output of these models is the image category along and the explanation of *why* that sample has been classified to that category.

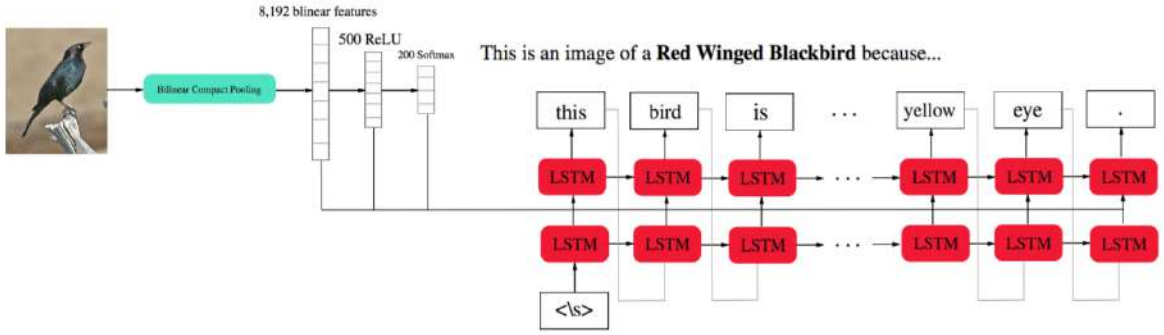


Figure 8: Example if a XAI model architecture (InterpNET [35]).

3.4 Attention in RNNs

We have seen that RNNs are another common DNN architectures that are very present in for solving medical and neurological challenging problems, for instance, for analyzing speech recordings or time-series data [38, 39]. Over the last years, attention layers in RNNs have been proved to be very effective in natural language processing and improve significantly prediction results [40]. Interestingly, attention layers can provide a key to partially interpret and explain neural network behavior, even if it cannot be considered a reliable means of explanation [41]. For instance, the weights computed by attention can point us to relevant information discarded by the neural network or to irrelevant elements of the input source that have been factored in and could explain a surprising output of the neural network. Therefore, visual highlights of attention weights could be instrumental to analyzing the outcome of neural networks, and a number of specific tools have been devised for such a visualization [42]. Figure 9 shows an example of attention visualization in the context of aspect-based sentiment analysis achieving successful results.

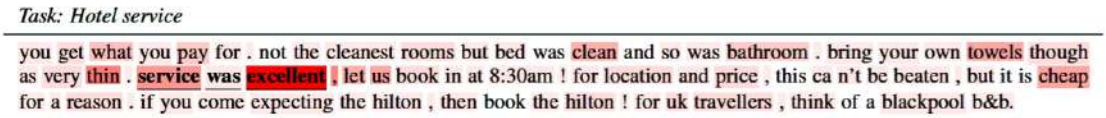


Figure 9: Example of attention visualization for an aspect-based sentiment analysis task. Words are highlighted according to attention scores. Phrases in bold are the words considered relevant for the task, or human rationales [43].

3.5 Autoencoders for feature importance

Autoencoders (AE) are able to map input data non-linearly into a low-dimensional explanation space while retaining most of the relevant information, so that the original deep learning predictions can be constructed from the few concepts extracted by the explanation network. Combined with CNNs, we can map 2-dimensional images into a 1-dimensional array preserving important information (most relevant features). There are works [44] that propose frameworks for visualizing the encoded relevant information that deep learning is using to make decisions as concepts for human to learn about the high-level keypoints. For example, in figure 10, we can see this framework applied to a bird dataset. First row shows a bird with red mark in the head and the proposed AE has detected that it is the most relevant feature for detecting that class. Same happens with the second row in figure 10, where the the AE detects that the main feature for the bird is the orange color of its paws. This is somehow

how a human would act when classifying bird spices, by focusing in distinguishable keypoints, and the most interesting thing is that these keypoints are discovered by the AE in an unsupervised way.

Figure 10: The most important x-feature for several categories [44].

To sum up all these techniques applied to CNNs, we can see that there is a research direction for extracting explanations from these complex models. Feature visualization for explaining the *what* and saliency maps for explaining the *why*. Also, we have observed that networks (such as InterpNET) open a new paradigm that forces ML methods to fit models with specific and explicit interpretations. The applications of this techniques are limitless and could be really helpful, for instance, to determine a justification of diagnoses outputted by DNNs (or at least to have a clue of what the model is focusing from the input).

4 Examples of applications in neuroscience

In the neuroscience community, we can already find some very recent published works that are staring to put into practice some of the mentioned techniques for improving model transparency. This shows that model explainability is currently a matter of study and a trend in deep learning for solving medical challenges.

In the work of Lee et al. (2019) [45] they present the development of an understandable deep-learning system that detects acute intracranial haemorrhage (ICH) and classifies five ICH subtypes from unenhanced head computed-tomography scans. The system includes an attention map and a prediction basis retrieved from training data to enhance explainability, and an iterative process that mimics the workflow of radiologists. Using small datasets they manage to achieve +90% of accuracy.

Abbasi-Asl et al. (2018) [46] introduce DeepTune, a visualization framework for CNNs, for applications in neuroscience. DeepTune consists of an ensemble of CNNs that learn multiple complementary representations of natural images. The features from these CNNs are fed into regression models to predict the firing rates of neurons in the visual cortex. The interpretable deepTune images, that means representative images of the visual stimuli for each neuron, are generated from an optimization process and pooling over all ensemble members.

In the work of Tomita et al. (2019) [47] and Schlemper et al. (2019) [48], attention-based NN models are employed to classify and segment histological images, e.g., microscopic tissue images, magnetic resonance imaging (MRI), or computed tomography (CT) scans. Tomita et al. (2019) found that the employed modules turned out to be very attentive to regions of pathological, cancerous tissue and

non-attentive in other regions. Furthermore, Schlemper et al. (2019) built an attentive gated network which gradually fitted its attention weights with respect to targeted organ boundaries in segmenting tasks.

Iten et al. (2020) [49] introduce SciNet, a modified variational autoencoder which learns a representation from experimental data and uses the learned representation to derive physical concepts from it rather than from the experimental input data. The learned representation is forced to be much simpler than the experimental data, for example by being captured in a few neurons, and contains the explanatory factors of the system such as the physical parameters.

Yan et al. (2019) [50], for example, introduce a grouping layer in a graph-based NN called GroupINN to identify subgroups of neurons in an end-to-end model. In their work, they build a network for the analysis of time-series of functional magnetic resonance images of the brain, which are represented as functional graphs in order to reveal relationships between highly predictive brain regions and cognitive functions.

This latest application example [50] shows that with these deep learning tools and techniques, not only we can find a certain degree of explainability, but they also can be used to discover new knowledge due to the unsupervised nature of DL models to extract good features and to find relevant low-dimensional representations of the data.

5 Discussion

We have seen that deep learning models are very effective for challenging tasks and in some problems there is not any efficient method currently other than deep learning models (i.e., medical image segmentation). DL models have proven a to have a positive impact on results improvement for solving numerous medical challenges and we cannot deny their usefulness in clinical diagnoses and research. DL models are definitely here to stay, and the research directions should be focused on their explainability and transparency, rather than proving them as uninterpretable opaque boxes.

Furthermore, deep learning has been broadly studied in the scientific community over the last years and backed up with solid mathematical theory, which has motivated the putting in practice of DL models to solve specific use cases. It is interesting to note the latent need for interpretable AI models over time (which is understandable, as interpretability is a requirement in many scenarios), yet it has not been until 2017 when the interest in techniques to explain AI models has permeated throughout the research community (figure 11).

Training DL models (such as CNNs, RNNs, or AEs) is not only more likely to produce competitive performances and better results compared to other methods, but also to provide insight into the task-relevant features through visualization techniques like saliency maps. These explainability techniques that we have reviewed here cannot be applied to simple ML models: the initial drawback of lack of explainability in DNN due to the complex nature of these models, now it turns out to be an advantage for being able to apply interpretability techniques such as feature visualization and saliency maps.

An other significant advantage of deep learning models is that unsupervised approaches (e.g., autoencoders, feature extraction) can automatically discover new information for solving medical challenges and point out where experts should put the attention to find solution for problems yet unknown.

Nonetheless, the techniques mentioned in this work are not enough to consider the DNN model predictions to be rigorously justified and explained. That is the reason why it is not likely to get rid of the human in-the-loop when using DL models. However, we have seen that there are currently a lot of methods to extract significant explainability from them. This opens a new research direction and a potential first step for introducing DL models into clinical settings (at least as a tool with expert supervision).

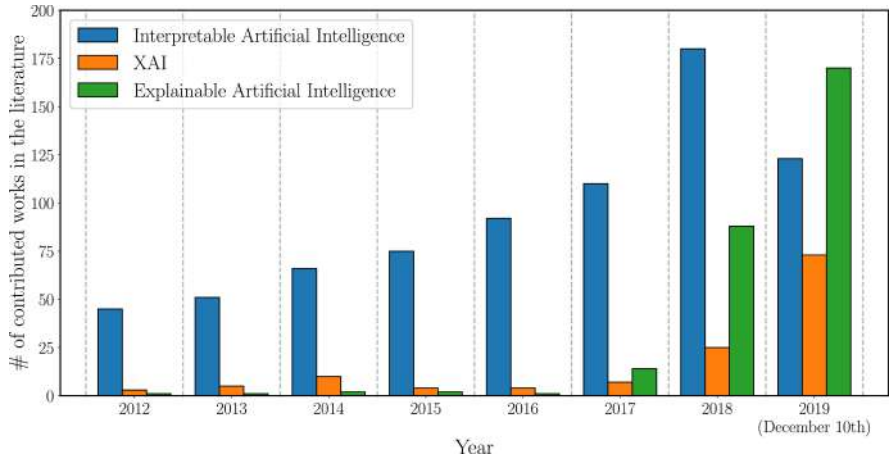


Figure 11: Evolution of the number of total publications whose title, abstract and/or keywords refer to the field of XAI during the last years. Data retrieved from Scopus [25].

More works can be found exploring explainability in this direction for deep learning models in medical problems [51], which is an evidence of the scientific community intentions to break interpretability barrier of deep learning models in order to start using them as valid and consistent tools in neuroscience and medicine fields.

6 Conclusions

Due to the deep learning models design, scientific consistency and plausibility is enforced, even if not as a primary goal. The explanation of specific model components are meant to lead to novel scientific discoveries or insights.

We have reviewed successful explainability techniques for deep learning models that are already being applied to neuroscience and medical problems over the last years. This is clearly a new trend in AI that is gaining the attention it requires and will help to introduce deep learning methods into medical environments. However, important challenges remain a barrier to integration of deep learning tools in the clinical setting.

As a data science student and practitioner in AI, explainability is a common problem that is not usually treated or given the attention it really needs. Main efforts have been focused in achieving models with great accuracy to prove the potential of deep learning, but all these models might be useless in many fields if they have a "black-box" interpretability. It was a research topic I really wanted to explore as an AI student and opens to me as a new field of possible future research.

References

[1] Y. Bengio, *Learning deep architectures for AI*. Now Publishers Inc, 2009.

[2] P. Smolensky, "Grammar-based connectionist approaches to language," *Cognitive Science*, vol. 23, no. 4, pp. 589–613, 1999.

[3] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *International journal of human-computer studies*, vol. 58, no. 6, pp. 697–718, 2003.

[4] M. T. Ribeiro, S. Singh, and C. 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*, pp. 1135–1144, 2016.

[5] T. Lombrozo, "The structure and function of explanations," *Trends in cognitive sciences*, vol. 10, no. 10, pp. 464–470, 2006.

[6] T. Lombrozo, "The instrumental value of explanations," *Philosophy Compass*, vol. 6, no. 8, pp. 539–551, 2011.

[7] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[8] J. Choo and S. Liu, "Visual analytics for explainable deep learning," *IEEE Computer Graphics and Applications*, vol. 38, p. 84–92, Jul 2018.

[9] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable ai systems for the medical domain?," *arXiv preprint arXiv:1712.09923*, 2017.

[10] H. Jones, "Geoff hinton dismissed the need for explainable ai: 8 experts explain why he's wrong," 2018.

[11] J.-M. Fellous, G. Sapiro, A. Rossi, H. S. Mayberg, and M. Ferrante, "Explainable artificial intelligence for neuroscience: Behavioral neurostimulation," *Frontiers in Neuroscience*, vol. 13, p. 1346, 2019.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[13] A. Valliani, D. Ranti, and E. K. Oermann, "Deep learning and neurology: A systematic review," *Neurology and therapy*, pp. 1–15, 2019.

[14] M. Biswas, V. Kuppili, L. Saba, D. R. Edla, H. S. Suri, E. Cuadrado-Godia, J. R. Laird, R. T. Marinhoe, J. M. Sanches, A. Nicolaides, *et al.*, "State-of-the-art review on deep learning in medical imaging.," *Frontiers in bioscience (Landmark edition)*, vol. 24, pp. 392–426, 2019.

[15] M. A. Ebrahimighahnavieh, S. Luo, and R. Chiong, "Deep learning to detect alzheimer's disease from neuroimaging: A systematic literature review," *Computer Methods and Programs in Biomedicine*, vol. 187, p. 105242, 2020.

[16] H. Choi, S. Ha, H. J. Im, S. H. Paek, and D. S. Lee, "Refining diagnosis of parkinson's disease with deep learning-based interpretation of dopamine transporter imaging," *NeuroImage: Clinical*, vol. 16, pp. 586–594, 2017.

[17] Z. C. Lipton, "The doctor just won't accept that!," *arXiv preprint arXiv:1711.08037*, 2017.

[18] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Advances in neural information processing systems," *Neural Information Processing Systems Foundation*, vol. 1269, 2012.

[20] L. Saba, M. Biswas, V. Kuppili, E. C. Godia, H. S. Suri, D. R. Edla, T. Omerzu, J. R. Laird, N. N. Khanna, S. Mavrogeni, *et al.*, "The present and future of deep learning in radiology," *European journal of radiology*, vol. 114, pp. 14–24, 2019.

[21] H. A. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. B. H. Hassen, L. Thomas, A. Enk, *et al.*, "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Annals of Oncology*, vol. 29, no. 8, pp. 1836–1842, 2018.

[22] R. Poplin, A. V. Varadarajan, K. Blumer, Y. Liu, M. V. McConnell, G. S. Corrado, L. Peng, and D. R. Webster, "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning," *Nature Biomedical Engineering*, vol. 2, no. 3, p. 158, 2018.

[23] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," 2015.

[24] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[25] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[26] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Towards medical xai," 2019.

[27] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," 2016.

[28] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *University of Montreal*, vol. 1341, no. 3, p. 1, 2009.

[29] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization," *Distill*, 2017. https://distill.pub/2017/feature-visualization.

[30] D. Wei, B. Zhou, A. Torrabla, and W. Freeman, "Understanding intra-class knowledge inside cnn," 2015.

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2015.

[32] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013.

[33] T. N. Mundhenk, B. Y. Chen, and G. Friedland, "Efficient saliency maps for explainable ai," 2019.

[34] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," 2018.

[35] S. Barratt, "Interpnet: Neural introspection for interpretable deep learning," 2017.

[36] L. A. Hendricks, R. Hu, T. Darrell, and Z. Akata, "Grounding visual explanations (extended abstract)," 2017.

[37] D. H. Park, L. A. Hendricks, Z. Akata, B. Schiele, T. Darrell, and M. Rohrbach, "Attentive explanations: Justifying decisions and pointing to the evidence," 2016.

[38] F. Çevik and Z. Kilimci, "Analysis of parkinson's disease using deep learning and word embedding models," *Academic Perspective Procedia*, vol. 2, pp. 786–797, 11 2019.

[39] K. M. Tsiouris, V. C. Pezoulas, M. Zervakis, S. Konitsiotis, D. D. Koutsouris, and D. I. Fotiadis, "A long short-term memory deep learning network for the prediction of epileptic seizures using eeg signals," *Computers in biology and medicine*, vol. 99, pp. 24–37, 2018.

[40] P. Chen, Z. Sun, L. Bing, and W. Yang, "Recurrent attention network on memory for aspect sentiment analysis," in *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 452–461, 2017.

[41] S. Jain and B. C. Wallace, "Attention is not explanation," *arXiv preprint arXiv:1902.10186*, 2019.

[42] S. Liu, T. Li, Z. Li, V. Srikumar, V. Pascucci, and P.-T. Bremer, "Visual interrogation of attention-based models for natural language inference and machine comprehension," tech. rep., Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States), 2018.

[43] A. Galassi, M. Lippi, and P. Torroni, "Attention in natural language processing," 2019.

[44] Z. Qi and F. Li, "Learning explainable embeddings for deep networks," in *NIPS Workshop on Interpreting, Explaining and Visualizing Deep Learning. Long Beach, CA, December*, vol. 9, 2017.

[45] H. Lee, S. Yune, M. Mansouri, M. Kim, S. H. Tajmir, C. E. Guerrier, S. A. Ebert, S. R. Pomerantz, J. M. Romero, S. Kamalian, *et al.*, "An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets," *Nature Biomedical Engineering*, vol. 3, no. 3, p. 173, 2019.

[46] R. Abbasi-Asl, Y. Chen, A. Bloniarz, M. Oliver, B. D. Willmore, J. L. Gallant, and B. Yu, "The deeptune framework for modeling and characterizing neurons in visual cortex area v4," *bioRxiv*, p. 465534, 2018.

[47] N. Tomita, B. Abdollahi, J. Wei, B. Ren, A. Suriawinata, and S. Hassanpour, "Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides," *JAMA network open*, vol. 2, no. 11, pp. e1914645–e1914645, 2019.

[48] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical image analysis*, vol. 53, pp. 197–207, 2019.

[49] R. Iten, T. Metger, H. Wilming, L. del Rio, and R. Renner, "Discovering physical concepts with neural networks," *Physical Review Letters*, vol. 124, Jan 2020.

[50] Y. Yan, J. Zhu, M. Duda, E. Solarz, C. Sripada, and D. Koutra, "Groupinn: Grouping-based interpretable neural network for classification of limited, noisy brain data," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 772–782, 2019.

[51] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *IEEE Access*, vol. 8, p. 42200–42216, 2020.