A systematic approach for tracking the evolution of XAI as a field of research

Samaneh Jamshidi[0000−0001−7055−2706], Sławomir Nowaczyk[0000−0002−7796−5201], Hadi Fanaee-T[0000−0001−8413−963X], and Mahmoud Rahat[0000−0003−2590−6661]

Center for Applied Intelligent Systems Research (CAISR), Halmstad University, Halmstad, Sweden
{samaneh.jamshidi, slawomir.nowaczyk, hadi.fanaee, mahmoud.rahat}@hh.se

Abstract. The increasing use of AI methods in various applications has raised concerns about their explainability and transparency. Many solutions have been developed within the last few years to either explain the model itself or the decisions provided by the model. However, the number of contributions in the field of eXplainable AI (XAI) is increasing at such a high pace that it is almost impossible for a newcomer to identify key ideas, track the field's evolution, or find promising new research directions.

Typically, survey papers serve as a starting point, providing a feasible entry point into a research area. However, this is not trivial for some fields with exponential growth in the literature, such as XAI. For instance, we analyzed 23 surveys in the XAI domain published within the last three years and surprisingly found no common conceptualization among them. This makes XAI one of the most challenging research areas to enter. To address this problem, we propose a systematic approach that enables newcomers to identify the principal ideas and track their evolution. The proposed method includes automating the retrieval of relevant papers, extracting their semantic relationship, and creating a temporal graph of ideas by post-analysis of citation graphs.

The main outcome of our method is Field's Evolution Graph (FEG), which can be used to find the core idea of each approach in this field, see how a given concept has developed and evolved over time, observe how different notions interact with each other, and perceive how a new paradigm emerges through combining multiple ideas. As for demonstration, we show that FEG successfully identifies the field's key articles, such as LIME or Grad-CAM, and maps out their evolution and relationships.

Keywords: Field's Evolution · XAI · Explainable AI.

1 Introduction

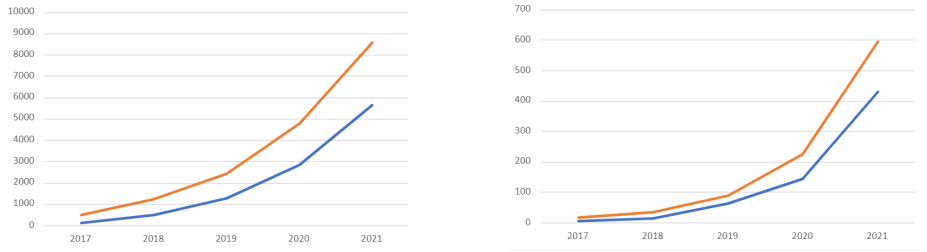
In recent years, the usage of Machine Learning (ML) and Artificial Intelligence (AI) techniques has increased greatly, especially as these methods are becoming more and more popular across all aspects of life. From the efficiency and performance standpoint, new algorithms and architectures are being continuously

proposed, providing essentially day-by-day improvements. In particular, the last decade brought the Deep Learning (DL) revolution; powered by hardware developments and enormous labeled datasets, these new models outperform, in many tasks, not only classical ML approaches but also human experts.

However, much of the new power of ML methods come at the cost of creating models of very high complexity. While traditional methods, such as (shallow) decision trees or linear regression, give the users a good understanding of how they make their decisions, the more complex methods are opaque. Often known as black boxes, they are not explainable by themselves. Although many such black box models achieve high performance, the lack of transparency that comes with it makes it so that they are not suitable in every setting. Given the desire to take advantage of new developments enabled by AI in many domains, this drawback is sometimes a deal-breaker, especially in safety-critical settings. In a domain like healthcare, it is not easy to trust a model and accept its decision without knowing the reasons for the decisions made[28]; ultimately, it is the human clinician who is responsible for the treatment, and they can only use AI-based decision support systems that provide relevant medical evidence. Prognostics and Health Management (PHM) is another interesting topic because of its high operating, maintenance, and downtime cost. So predictive remaining useful life and predictive maintenance are critical industry issues. Using AI and ML algorithms is increasing in this area, like in other areas, but lack of transparency, interpretability, understanding, and interpretation is one of the main challenges. Companies and factories cannot rely on decisions that they do not know the reasons for and can not understand why. Not only is this lack of trust related to bias and lack of representation of the training datasets, but it also includes adversarial attacks [19, 14, 34]. As an example, authors in [27] show that it is easy to produce meaningless images, unrecognizable to humans, but such that the DNNs classify them with 99.99% confidence. On the other hand, the right explanation methods can help to significantly improve the model performance or design a better architecture, as demonstrated in [52]. Generally, there are two main motivations to develop methods that make black box models explainable: 1) understanding the reasons behind a decision to make the model trustable; 2) having a better view of a model and its weakness, with the aim of debugging.

It is for those reasons that XAI is today one of the most popular and heavily researched topics in AI. It is clear that the challenges are real, but significant progress has been made in the last couple of years, in part due to cross- and inter-disciplinary collaborations. This is readily visible in the rapid growth of the number of publications within the field. For instance, more than 5500 research articles (400 for survey papers) are returned by a Google Scholar search just by using the phrase "explainable artificial intelligence" – within the year 2021 alone (Fig.1).

This explosion in popularity, however, creates unique challenges in terms of understanding the current landscape, identifying common trends, comparing solutions, and finding overlaps and gaps in state-of-the-art. This is an especially frustrating obstacle for newcomers into the field – which poses a danger of creat-



(a) articles (b) review articles

Fig. 1: The statistics appeared in Google Scholar for the number of articles and review articles published per year from 2017 to 2021, based on the search for **blue lines:** the phrase "explainable artificial intelligence" and **orange lines:** the phrase "explainable artificial intelligence" OR"explainable machine learning"OR"understanding artificial intelligence"OR"understanding machine learning", by year (Color figure online)

ing narrow, hermetic, splintered societies; spelling disaster for the field, which, by its very natures, requires broad and interdisciplinary collaborations and perspectives. Among the survey articles in the area, some focus on the XAI for rather specific topics, including medical[28, 47] or natural language processing[12]. Although domain-specific review articles have advantages, their biggest problem is missing out on the ideas that have been successful on other data or in other domains and could be applied to that specific domain. Also, among the survey articles, there are many conflicts on how to categorize the methods in the XAI field. In addition, there is no agreement on the most important articles in the field; since there are so many articles, the vast majority are only cited by a very small number of review articles.

To tackle the challenges explained above, we propose a systematic and universal approach that enables newcomers to identify the fields' main ideas and track their evolution. The remainder of the paper is organized as follows. Section 2 gives an overview of related work. Section 3 presents the proposed method. Section 4 is dedicated to experimenting with details. Section 5 demonstrates the results. Finally, Section 6 concludes the paper.

2 Related Work

In many domains, the number of published scientific papers rapidly increases every year, and some researchers have suggested automating survey generation via AI solutions. This is typically framed as a multi-document summarization, a subset of natural language processing. Abstractive [25] and extractive [48] summarizations are among the most common approaches. The idea of using citation graphs or citation links for analyzing the relations between papers has also been explored before [1, 9, 50]. One common approach is leveraging cita-

tion sentences to pinpoint important aspects of the papers. For instance, [45, 33] exploit a template-based framework and composes a template-tree. The latter crawls citation index databases such as PubMed and Semantic Scholar and analyses the citation graph. However, more advanced methods to process the citation graphs are still to be developed. For instance, it is not clear if all the citations in a paper are relevant and reliable or if they share the same level of importance in the context.

Although text summarization-based approaches have been relatively good at producing a summary of related works, they are not able to make semantic relationships between the papers or identify the evolution of the key ideas.

3 Proposed Method: Field's Evolution Graph (FEG)

Our fundamental goal in this paper is to understand the "evolution" of XAI as a field of research. We are particularly interested in identifying the key concepts and ideas that shaped further development. We aim to express these by finding a graph of relations between papers in XAI, allowing us to identify influences and concepts that have been developed and improved over time, discover groups and communities related to key ideas, etc.

The sheer volume of papers in the field makes this task infeasible if attempted manually. Therefore, we are proposing an approach that allows us to (partially) automate the task.

The key focus of our approach is analyzing the citations among papers since in the scientific world bibliographic references are the most reliable source of information about inspirations, extensions, development, and improvement of ideas. Therefore, we first extract a graph network of paper relations, second, we identify the important edges, and third, we analyze the resulting structure to uncover the thread of the evolution of key concepts in the XAI field. The key challenge, and the main focus of this section, is the discovery of different types of edges and identifying how they indicate the evolution of ideas within the field.

Algorithm 1 Field's Evolution Graph

1: Select a list of survey papers in the XAI field.
2: Extract their references using Semantic Scholar API.
3: Calculate the repetition of extracted references among those surveys **(repetition rate)**.
4: Pick those papers of step 3 that are cited by at least 25% of the survey list **(Influential papers)**.
5: Rank them based on publication year, citation number, and repetition rate.
6: Draw a graph of citations between papers of the previous step.
7: Remove unnecessary links from the graph.

3.1 Identification of Influential Papers

At first, we need a number of important and influential papers in this field. The most obvious approach would be to start with highly-cited papers. However, citations alone do not provide accurate and reliable results for several reasons. First, the number of citations depends on the year of the publication, as well as the venue, and does not necessarily accurately reflect the true importance of the contribution. More importantly, many articles belong to more than one domain, i.e., not only XAI, and their citation may be due to importance for other domains. Finally, some of the important papers just focus on a specific issue or data and, despite their importance in this area, will be referenced by a smaller number of articles. Therefore, there is a need to use other features to identify these articles.

Instead, we propose a different approach to obtaining such papers, namely, by exploring the existing surveys. This is feasible in an exploding area like XAI due to the available number of review articles published every year. We select a number of recent survey articles, based on popularity; then, we extract their references (by using Semantic Scholar API), and calculate the repetition of each paper among those survey articles. Papers with a high repetition count, i.e., those included in many surveys, are likely to be the most influential and important ones in the field. Thus, three features, including citation rate, repetition count in surveys, and publication year, have been used in identifying key articles.

3.2 Citation Importance

The next step is to find the relations between the papers we have identified as key papers, revealing a structure within the XAI field. In particular, we aim to discover how different methods have evolved in this area over time. We would like to track the evolution and incremental improvements of an idea, starting from the original paper. We also want to show how the combinations of existing methods are effective in shaping new methods and identify when it happens. Finally, we want to group the methods by revealing the different approaches in XAI in an automatic way.

The starting point is to analyze citations since they are the most direct measure of influence across papers. By considering key articles as nodes and references' status as edges, a graph of the relationships between these articles is formed. A directed graph can show these relations perfectly.

Articles refer to each other in different ways, and those references can have different meanings. For example, depending on the section (such as background, method, experiments, results, etc.) where a citation occurs, the importance and influence across papers vary greatly. Looking back at our goal, we do not consider all these types of citations. In particular, citations referring to the methodological relationships are the most important for our purpose – since it is the methodology where new ideas and solutions are formed. There are many ways of assessing citations. One of them is to do it by hand, which is time-consuming and

costly, especially in a large number of articles. The automatic alternative is Semantic Scholar, which provides high-quality citation data via API[16]. It indexes published peer-reviewed scientific literature across various disciplines, currently covering more than 187 million research papers. Semantic Scholar integrates a set of query and analytics features, several of which have been identified as useful for our study. It offers an API to pull data regarding individual records, references list, and citation data for each indexed paper. It also classifies paper references into different reference types: background, results, methods, or without a label. However, since the whole procedure is processed automatically, the accuracy of citation data does not seem perfect, and some errors are expected, thus, some manual post-processing is required.

3.3 Visualization of FEG

Visualization is a useful and efficient way in many fields, especially in analysis. It gives a higher chance of discovering insights when interacting with data. Graphs, on the other hand, are a good tool for showing the connections between the components of a set. Following the directed edges from one node to another provides useful information about the type and manner of connection between two nodes. We use FEG plots to show the relevance of articles. Although there are various methods for examining and analyzing graphs, we have used graph visualization and analysis manually at this stage of the work.

4 Experiment

We conduct a relatively small-scale experiment where we evaluate the feasibility of the proposed approach before scaling it up.

The first step toward obtaining the list of key articles in the XAI field is to analyze recent surveys. Therefore, we started from a list of 23 review articles published between 2018 and 2021. We then analyzed all the references present in those review articles, obtaining an initial list of more than 1800 potentially interesting papers. Next, we ranked the articles in this list using the three important features: publication year, citation number, and repetition rate (i.e., how many selected review articles referred to that article). There are two significant findings regarding this list:

- There is a *very* long tail of papers that were only cited by one of the selected survey papers – more than 1400 papers were only cited once among the 23 surveys. Almost 900 of them are published before 2018, which means that all of those papers were published before all the surveys, but they have been only noticed by one of them.
- Only 9 papers were cited by half (or more) of the surveys and 8 of them were published before 2018. This means that the consensus among the surveys about important papers is virtually non-existent. An extremely small ratio (half a percent) of articles has been agreed upon by the majority of review articles.

The above observations show that using these surveys for finding influential and important articles in the field is problematic, to say the least. It is very likely that, by relying on input from a handful of such papers, a new reader would get a very biased and incomplete picture of the field.

Instead, we believe that some of these issues can be diminished, even if not completely removed, by aggregating data from multiple surveys.

5 Results

Fig. 2 shows the FEG plot for all the connections of selected articles: a subset of key articles: those which are referenced by at least 25% of the reviews[2–8, 10, 11, 13, 15, 17, 18, 20–24, 26, 27, 29–32, 35–44, 46, 49, 51–55].

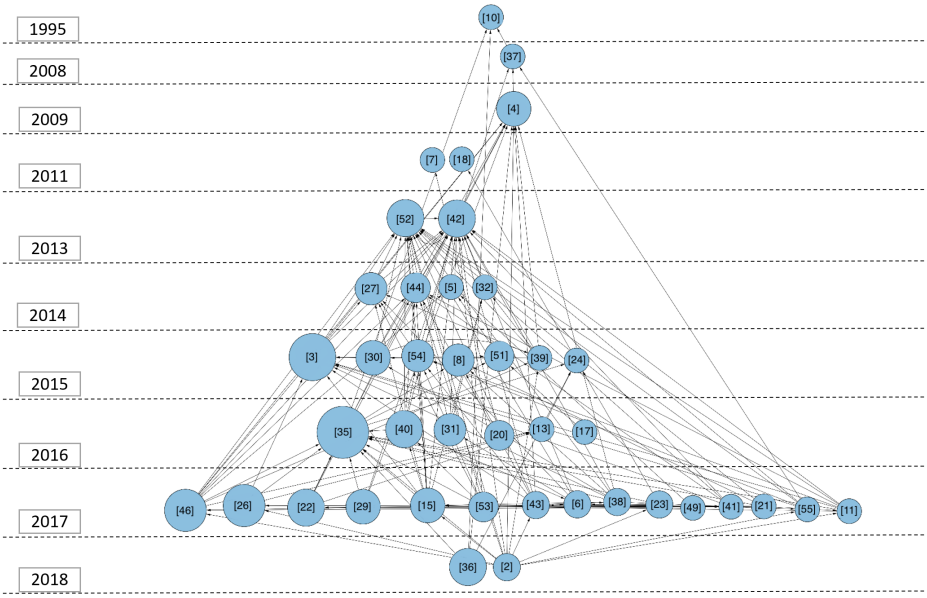


Fig. 2: FEG plot: links between extracted key articles (only those referenced by at least 25% of selected review articles). The directed edge from node A to node B means article A cites article B. The radius of each circle indicates the number of review articles referencing this paper. The vertical axis refers to the time (the upper, the older)

As discussed above, not all the links between articles are actually meaningful. For the purpose of tracking the evolution of the XAI field, we want to focus on methods that significantly influenced each other. To this end, we used semantic Scholar to label the links. A total of 158 links were found among the articles,

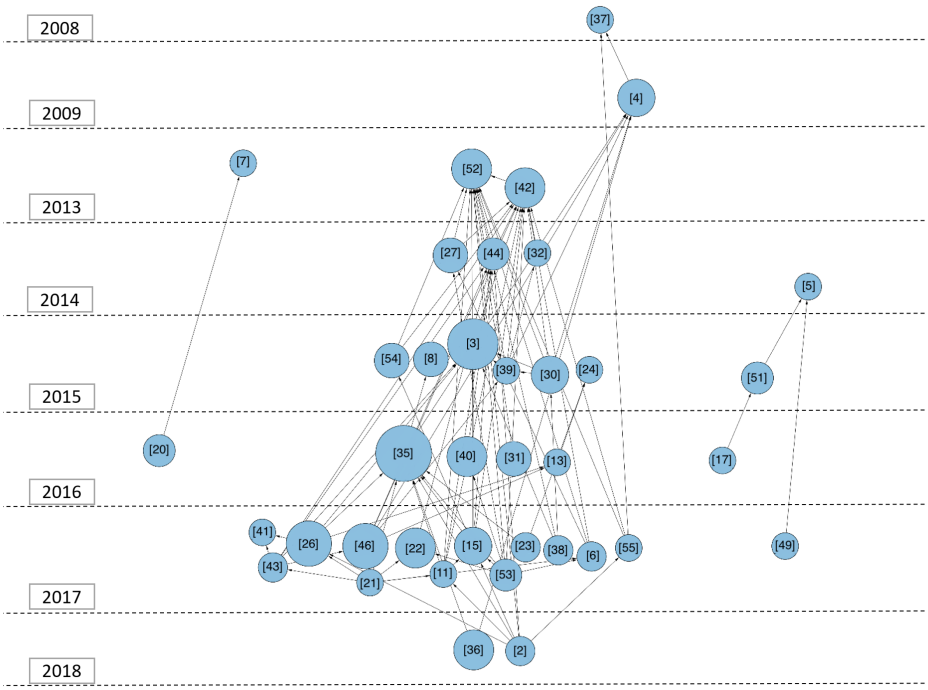


Fig. 3: FEG plot: links and labeled methods based on Semantic Scholar result, between extracted key articles referenced by at least 25% of selected review articles.

out of which 92 included methodologies, 12 included results, and 91 included backgrounds (note that some links include multiple tags). Finally, 32 of the links are unlabeled. For our work, links with the methodology label are the most important; a bit less than 60% of the links have this label. Accordingly, in Fig. 3, we keep the edges labeled "methodology" for further consideration and remove the rest.

One can immediately notice in Fig. 3 that there are two (small) disconnected sub-graphs, and a large part remains connected. Those two sub-graphs can be representative of two different types of methods in this field. By analyzing the articles of these two groups, it can be seen that one of those represents methods related to providing prototypical examples as an explanation, while the other is related to the use of image captioning as an explanation. Those findings are discussed in more detail in the following subsections.

5.1 Example-based Methods

One of the key ideas we found from FEG plots is example-based methods. The best example of this category is the work of [7] who propose to select a few

instances from the dataset; those that are a good representative of data can be a way to make a better understanding dataset. These kinds of methods, known as prototype methods, are usually used as a preprocessing part. This method suggests that the desired prototype or representative of class C should cover as many training data of class C as possible while covering as few training data as possible of classes other than C. In addition, it should be sparse. An interpretable representative of a dataset must not only contain examples of each class, but it is also necessary to provide some criticism samples. The criticism can explain what is not captured by prototypes. For instance, [20] develop the maximum mean discrepancy criticism (MMD-critic) method for prototype Selection and criticism motivated by the Bayesian model criticism framework.

5.2 NLP-based Approaches

The second class of ideas we can infer from FEG plot are natural language processing (NLP) based techniques. These methods provide a solution to explain the model decisions. The main application is creating a text to describe an image, known as image captioning. The four papers forming the rightmost sub-graph in Fig. 3 are examples of this class. Being able to describe the image from the extracted features can also be approached to make the feature production model understandable.

Inspired by attention-based models, [51] introduced a method to describe an image. Unlike other models in image captioning, which use object detectors or represent images as a single feature vector from the top layer of a pre-trained convolutional network, their model learns hidden alignments from scratch. This model extracts features used by the encoder from the lower convolutional layer instead of the fully connected layer. This way, the decoder can be more focused on the parts of the image that are important. The learned attention in the decoder can be used as a solution to visualize the model generation process and make this model interpretable. In other words, by using those attention, one can show which parts of the image are the most important contributors to producing each word; this provides an understanding of how the model works.

On the other hand, [17] discusses that a textual description of an image should not only describe that image correctly but should also be class discrimination. Explanations produced by this model are not only conditioned on the images but also conditioned on the respective classes. The authors used a discriminative loss function to encourage captioning sentences to correspond primarily to features that are class-specific. Although this model produces sentences that are discriminative as well as descriptive, it is not able to show which part of the image is related to the features mentioned in the sentences. Moreover, it is possible that some features do not appear in an image and just come to the sentences based on being class discriminative.

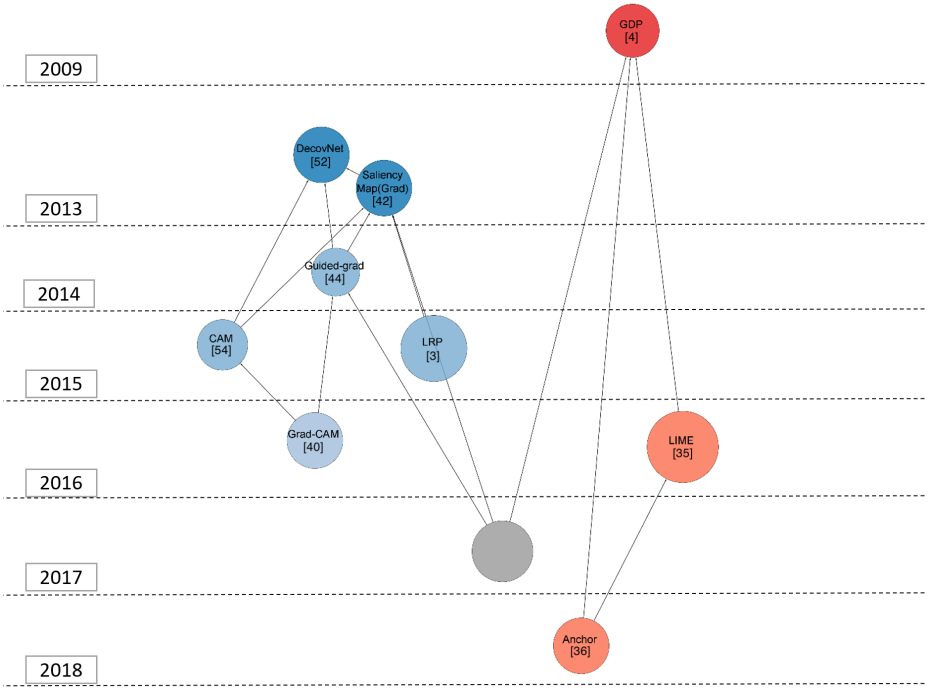


Fig. 4: FEG plot of the articles in features importance approach. The blue nodes belong to the model-specific approach, and the red nodes belong to the model-agnostic approach. The gray node represents a paper that does not fit either category; notably, it is linked to both aforementioned approaches.

5.3 Feature Importance Techniques

The largest group of ideas belongs to feature importance techniques, specified in the FEG plot as a sub-graph formed in Fig. 3 contains a number of articles that are all linked together. Disentangling those relations is going to be more challenging and requires a more in-depth analysis than sections 5.1 and 5.2. First, however, it is important to notice that almost all articles in this group use feature importance to explain either the model as a whole or individual decision. They have taken different approaches to do so; however, it is clear that they are all related. By subjective visual analysis, one can notice that papers[52] and [42], together with [4], form important "hubs". So there are two main approaches in between, which we will discuss in the following, and how the formation, expansion, and evolution of methods in these two.

By focusing on these, and the papers that cite them and ignoring the rest, one can obtain the FEG plot presented in Fig. 4. We thus now focus on analyzing this group of papers.

Model-specific One way to explain a model and its decisions are to show which features play the most critical role in output generation. Some methods are proposed on specific models to show the influential features of making a decision, which we will discuss. In 2013, [52] proposed a method to visualize the convolutional layers. They used a multi-layer DeconvNet to map the activities of each layer to the input of that layer. By doing so and displaying it in the original pixel space, one can identify the parts of the input that have the most impact on that layer. Doing this for the last layer can provide a strong visualization of the input that shows the important pixels for each decision. One of the problems with this method is that the max-pooling operator is non-invertible. The authors, therefore, approximated the inverse of this operation by producing Max Locations Switches to record the location of maximum value within each pooling area to solve this problem.

As the name implies, this method is applied to convolution layers. Following this, another method was presented by [42] to obtain the class saliency map from the gradient of the score of class $c(Y_c)$, with respect to the input image I. It can be shown that except for the RELU layer, DeconvNet effectively corresponds to the gradient backpropagation through a ConvNet. Gradient backpropagation applies to visualize the class score neurons in the final fully-connected layer. It means this method can be applied not only to a convolutional layer but also to any other type of layer. In this sense, it is seen as a generalization of [52]. In more details, this method obtains the class saliency map from the gradient of the score of class $c(Y_c$, with respect to the input image I, by taking the magnitude of it and a maximum along all its channels. If the values of the derivation of Y_c w.r.t the I is close to zero, it means that small changes in that part of the image have no effect on determining that output class. The values which are high in magnitude mean that small changes in that pixels can have a major impact on the result of score class c. Note that for obtaining that gradient, instead of back-propagating on the loss, it should be backpropagation on the score Y_c.

Later, [44] offered another improvement in [52] by eliminating the need for switches and replacing max-pooling layers with convolution' and proposed a combination of methods in [52] and [42]. The difference between' deconvolution'[52] and backpropagation [42] is handling backpropagation through the rectified linear (ReLU) non-linearity. While deconvolution computes gradient based on the top gradient signal, backpropagation computes this based on negative entries of the bottom data. In the case of the ReLU non-linearity, this amounts to setting to zero certain entries based on the top gradient in deconvolution and bottom data in backpropagation. [44] combined them and zerosout the negative gradients during backpropagation. This method, called guided Backprop or guided-grad, often produces more visually appealing and less noisy results and can be used even without' switches' (Max Location).

Class Activation Maps (CAM) is also trying to understand which pixels of an image have more contribution to the final output of the model[54]. This method replaced fully connected layers at the very end of the model with the Global Average Pooling (GAP) layer. This layer averages the activations of each

feature map and concatenates them as a vector and a weighted sum of this vector is fed to the final soft-max loss layer. According to [52, 42], each unit is expected to be activated by some visual patterns. Thus, the most relevant part of an input image (to a particular class) is identified by up-sampling CAM to the size of the image.

Although the output of the CAM is class discriminative, the network must be fine-tuned in this method. Also, fully-connected layers are replaced, so it is not applicable to all networks. Grad-CAM[40] as a combination of the saliency map[42] and CAM [54] was introduced to deal with these limitations. Grad-CAM calculates gradients of any class score with respect to the activations maps of the final convolutional layer. Then, similar to CAM, score importance is obtained by averaging the gradients across each feature map. Grad-CAM can only produce coarse-grained visualizations, therefore the authors have also combined guided-grad[44] with Grad-CAM(via element-wise multiplication) and propose Guided Grad-CAM which is able to highlight fine-gradient details.

Model-agnostic Although the methods described in the previous approach apply to a wide range of neural networks, they are all model-specific. However, several feature importance-based methods are model-agnostic and therefore can be applied to different models (the left part of Figure 4).

In particular, [4] proposed a procedure to understand the decisions for every single instance by obtaining local explanation vectors based on Gaussian Process Classification (GDP). Local gradients, as explanation vectors, determine how a data sample should be changed to change its predictive label and find the most influential features in the decision of the model for a particular instance. This technique can be applied to any classification method.

LIME[35] is also a well-known method to generate a local explanation of any black-box model. This method uses local surrogate interpretable models to approximate the prediction of the model. Its main idea is that train an accurate black-box model and then explain the model based on the simple and easy-to-understand model such as linear or logistic regression locally. LIME generates some neighborhoods of the instance that has to be explained, labels them by the black-box model, and weights them based on their vicinity to the original instance. Finally, an interpretable model applies to these weighted instances and their predicted labels to create the explanations.

6 Conclusion

We propose a systematic solution for newcomers who are interested to enter a new research area but face information overload due to the intractable number of publications. Our solution is able to efficiently identify the key group of ideas and track their evolution. This is essential in fields such as XAI that are evolving at an extremely high pace. We show how FEG can be used to uncover different key concepts in XAI, their temporal evolution, and how these ideas relate to each other. For example, the FEG created using our approach identifies three different

branches within XAI: the example-based approaches, the natural language-based approaches, and the feature importance-based approaches. FEG can also show how these ideas are formed and how mature they are. For instance, we can see how the guided-grad idea[44] evolved from the DeconvNet idea[52] or how the grad-CAM idea[40] is formed by combining CAM[54] and Guided-grad[44].

This paper is a work-in-progress. We ran the experiments on a limited number of articles in the field, removed irrelevant citations based on Semantic Scholar labeling, and analyzed the remaining graph manually. Nevertheless, we believe already these results are going to be of interest. However, at larger scales, the complete process can be automated by natural language and graph processing techniques. Another direction is the identification of the key papers in a more automatic way by using metrics and statistics in (social) network analysis. These methods can provide some important information on the relation between nodes and can also identify the important and influential nodes automatically.

Acknowledgments

This work was supported by CHIST-ERA grant CHIST-ERA-19-XAI-012 funded by Swedish Research Council.

References

1. Abu-Jbara, A., Radev, D.: Coherent citation-based summarization of scientific papers. In: Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies. pp. 500–509 (2011)
2. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. Advances in neural information processing systems **31** (2018)
3. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one **10**(7), e0130140 (2015)
4. Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., Müller, K.R.: How to explain individual classification decisions. The Journal of Machine Learning Research **11**, 1803–1831 (2010)
5. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
6. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6541–6549 (2017)
7. Bien, J., Tibshirani, R.: Prototype selection for interpretable classification. The Annals of Applied Statistics **5**(4), 2403–2424 (2011)
8. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N.: Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1721–1730 (2015)
9. Chen, J., Zhuge, H.: Automatic generation of related work through summarizing citations. Concurrency and Computation: Practice and Experience **31**(3), e4261 (2019)

10. Craven, M., Shavlik, J.: Extracting tree-structured representations of trained networks. Advances in neural information processing systems **8** (1995)
11. Dabkowski, P., Gal, Y.: Real time image saliency for black box classifiers. Advances in neural information processing systems **30** (2017)
12. Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., Sen, P.: A survey of the state of explainable ai for natural language processing. arXiv preprint arXiv:2010.00711 (2020)
13. Datta, A., Sen, S., Zick, Y.: Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In: 2016 IEEE symposium on security and privacy (SP). pp. 598–617. IEEE (2016)
14. Fink, O., Wang, Q., Svensen, M., Dersin, P., Lee, W.J., Ducoffe, M.: Potential, challenges and future directions for deep learning in prognostics and health management applications. Engineering Applications of Artificial Intelligence **92**, 103678 (2020)
15. Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: Proceedings of the IEEE international conference on computer vision. pp. 3429–3437 (2017)
16. Hannousse, A.: Searching relevant papers for software engineering secondary studies: Semantic scholar coverage and identification role. IET Software **15**(1), 126–146 (2021)
17. Hendricks, L.A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., Darrell, T.: Generating visual explanations. In: European conference on computer vision. pp. 3–19. Springer (2016)
18. Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., Baesens, B.: An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. Decision Support Systems **51**(1), 141–154 (2011)
19. Kan, M.S., Tan, A.C., Mathew, J.: A review on prognostic techniques for non-stationary and non-linear rotating systems. Mechanical Systems and Signal Processing **62**, 1–20 (2015)
20. Kim, B., Khanna, R., Koyejo, O.O.: Examples are not enough, learn to criticize! criticism for interpretability. Advances in neural information processing systems **29** (2016)
21. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: International conference on machine learning. pp. 2668–2677. PMLR (2018)
22. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: International conference on machine learning. pp. 1885–1894. PMLR (2017)
23. Lakkaraju, H., Kamar, E., Caruana, R., Leskovec, J.: Interpretable & explorable approximations of black box models. arXiv preprint arXiv:1707.01154 (2017)
24. Letham, B., Rudin, C., McCormick, T.H., Madigan, D.: Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. The Annals of Applied Statistics **9**(3), 1350–1371 (2015)
25. Li, W., Xiao, X., Liu, J., Wu, H., Wang, H., Du, J.: Leveraging graph to improve abstractive multi-document summarization. arXiv preprint arXiv:2005.10043 (2020)
26. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)
27. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5188–5196 (2015)

28. Markus, A.F., Kors, J.A., Rijnbeek, P.R.: 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**, 103655 (2021)
29. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence **267**, 1–38 (2019)
30. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.R.: Explaining nonlinear classification decisions with deep taylor decomposition. Pattern recognition **65**, 211–222 (2017)
31. Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., Clune, J.: Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. Advances in neural information processing systems **29** (2016)
32. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 427–436 (2015)
33. Nikiforovskaya, A., Kapralov, N., Vlasova, A., Shpynov, O., Shpilman, A.: Automatic generation of reviews of scientific papers. In: 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 314–319. IEEE (2020)
34. Rezaeianjouybari, B., Shang, Y.: Deep learning for prognostics and health management: State of the art, challenges, and opportunities. Measurement **163**, 107929 (2020)
35. Ribeiro, M.T., Singh, S., Guestrin, C.: " 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)
36. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
37. Robnik-Šikonja, M., Kononenko, I.: Explaining classifications for individual instances. IEEE Transactions on Knowledge and Data Engineering **20**(5), 589–600 (2008)
38. Ross, A.S., Hughes, M.C., Doshi-Velez, F.: Right for the right reasons: Training differentiable models by constraining their explanations. arXiv preprint arXiv:1703.03717 (2017)
39. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.R.: Evaluating the visualization of what a deep neural network has learned. IEEE transactions on neural networks and learning systems **28**(11), 2660–2673 (2016)
40. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
41. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: International conference on machine learning. pp. 3145–3153. PMLR (2017)
42. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: In Workshop at International Conference on Learning Representations. Citeseer (2014)
43. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)
44. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806 (2014)

45. Sun, X., Zhuge, H.: Automatic generation of survey paper based on template tree. In: 2019 15th International Conference on Semantics, Knowledge and Grids (SKG). pp. 89–96. IEEE (2019)
46. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017)
47. Tjoa, E., Guan, C.: A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE transactions on neural networks and learning systems **32**(11), 4793–4813 (2020)
48. Tohalino, J.V., Amancio, D.R.: Extractive multi-document summarization using multilayer networks. Physica A: Statistical Mechanics and its Applications **503**, 526–539 (2018)
49. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
50. Wang, J., Zhang, C., Zhang, M., Deng, S.: Citationas: A tool of automatic survey generation based on citation content. Journal of Data and Information Science **3**(2), 20–37 (2018)
51. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057. PMLR (2015)
52. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. pp. 818–833. Springer (2014)
53. Zhang, Q., Wu, Y.N., Zhu, S.C.: Interpretable convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8827–8836 (2018)
54. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)
55. Zintgraf, L.M., Cohen, T.S., Adel, T., Welling, M.: Visualizing deep neural network decisions: Prediction difference analysis. arXiv preprint arXiv:1702.04595 (2017)