On the role of knowledge graphs in AI-based scientific discovery

Mathieu D'aquin

Web Semantics: Science, Services and Agents on the World Wide Web 84 (2025) 100854

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

On the role of knowledge graphs in AI-based scientific discovery

Mathieu d'Aquin

LORIA, Université de Lorraine, CNRS, Nancy, France

ARTICLE INFO

Keywords:
Scientific discovery
Knowledge graphs
Machine learning
Interpretability

ABSTRACT

Research and the scientific activity are widely seen as an area where the current trends in AI, namely the development of deep learning models (including large language models), are having an increasing impact. Indeed, the ability of such models to extrapolate from data, seemingly finding unknown patterns relating implicit features of the objects under study to their properties can, at the very least, help accelerate and scale up those studies as demonstrated in fields such as molecular biology and chemistry. Knowledge graphs, on the other hand, have more traditionally been used to organize information around the scientific activity, keeping track of existing knowledge, of conducted experiments, of interactions within the research community, etc. However, for machine learning models to be truly used as a tool for scientific advancement, we have to find ways for the knowledge implicitly gained by these models from their training to be integrated with the explicitly represented knowledge captured through knowledge graphs. Based on our experience in ongoing projects in the domain of material science, in this position paper, we discuss the role that knowledge graphs can play in new methodologies for scientific discovery. These methodologies are based on the creation of large and opaque neural models. We therefore focus on the research challenges we need to address to support aligning such neural models to knowledge graphs for them to become a knowledge-level interface to those neural models.

1. Introduction

Let us start with a quote from the very beginning of Isaac Asimov's "True Love" short story [1]:

> My name is Joe. That is what my colleague, Milton Davidson, calls me. He is a programmer and I am a computer program. I am part of the Multivac-complex and am connected with other parts all over the world. I know everything. Almost everything. I am Milton's private program. His Joe. He understands more about programming than anyone in the world, and I am his experimental model. He has made me speak better than any other computer can. "It is just a matter of matching sounds to symbols, Joe", he told me. "That's the way it works in the human brain even though we still don't know what symbols there are in the brain. I know the symbols in yours, and I can match them to words, one-to-one". So I talk. I don't think I talk as well as I think, but Milton says I talk very well.

There are many things that are interesting in this short description of a system capable of speech according to this work of fiction written in 1977. The most obvious one, which is not directly relevant to the topic of this article but is still worth noticing, is that the most prominent Artificial Intelligence (AI) systems today (2024) are developed in the opposite direction. When the protagonist declares "I don't think I talk as well as I think", what could be more accurately said of current systems based on large language models (chatGPT, Claude, Gemeni, Meta-AI, etc.) is that they do not think as well as they talk.

To get closer to the topic of this article, another point on which the story differs from our current reality is that symbols in large neural networks are not more identifiable than those in the human brain. In other novels and short stories by Isaac Asimov relating to the Mulivac, a key benefit of such systems is that they have accelerated scientific discoveries, being able to process large amounts of data, rich information, and to integrate existing knowledge to come up with answers to complex questions. Current AI systems based on deep learning and large language models are expected to realize something similar. In many areas of all the fundamental sciences, neural networks of various sizes and complexities are being built to analyze phenomena ranging in scale from full eco-systems to the level of particles. Hence, we might want to ask the question: Would we need to understand symbols in large neural networks in order for them to be truly useful in advancing scientific discovery?

The reflections in this article are based on our ongoing experience in one such domain: material science. The task of discovering new materials can be summarized as trying to find, among the very large number of materials that could potentially exist, ones that are likely to

E-mail address: mathieu.daquin@loria.fr.

https://doi.org/10.1016/j.websem.2024.100854
Received 14 November 2024; Received in revised form 7 December 2024; Accepted 17 December 2024
Available online 26 December 2024

have the desired property and are therefore worth trying to synthesize. In this case, as in many others, an increasingly popular approach is to build machine learning models to predict the values of these macroscopic properties (for example, thermal conductivity, which is the one on which we focus) based on other properties that are easier to obtain or the atomic structure of the material (see, for example, [2,3]).

In such a context, the first step of our method is to build a predictive model. For example, in [4], we describe how we used a transfer learning approach based on a graph neural network originally trained on a large dataset to predict the formation energy of materials. Fine-tuning this model using smaller datasets of materials labeled with their thermal conductivity, we obtained a model that achieved promising results in the notoriously difficult task of predicting thermal conductivity. The direct goal of creating this model is to be able to use it to scan a large number of materials and to quickly identify those likely to have a low thermal conductivity, as candidates to be synthesized, since alternative approaches to estimating (with an acceptable level of accuracy) this particular property of a material takes thousands of CPU-hours, carrying out *ab initio* computations.

However, although having a predictive model is useful, it remains unsatisfactory: If valid, the model must have found something in the data that allows to at least approximate the relation between, in our case, the atomic structure of a material and its thermal conductivity. In addition, once the model is trained, this approximation is done in a few milliseconds, where the alternatives could potentially take weeks of heavily theory-backed computations. Is what it found a shortcut or a valid rule? How does it relate to existing knowledge on the physics of materials? If what is found is valid, previously unknown, and consistent with existing knowledge, could it be that the model found in the data a pattern indicative of a mechanism or phenomena that physics had not yet discovered?

To answer these questions, two components are required. First, it needs to be possible to inspect, explore, and extract higher-level structures that are representative of processes and concepts from within the neural network, i.e., to interpret it beyond the millions of tiny computations carried out within its neurons. Second, it should be possible to connect those structures to existing knowledge in the field that is formalized and represented in a way that is both computationally and humanly processable.

As we will describe later, research is very active in the field of machine learning on the task of *mechanistic interpretation*, the goal of which is to provide the first component. Concerning the second component, it could potentially become the most impactful realization of the vision of the Semantic Web as materialized through knowledge graphs: vast, rich, shared, and computationally exploitable knowledge networks that are conceptually structured, meaningful, and available globally through the Web. Knowledge graphs encoding current knowledge in scientific disciplines could become the foundation on which AI-led discovery is carried out by providing the necessary conceptual framework to ground machine learning models in shared and evolving knowledge.

2. On knowledge graphs in science

Providing a precise definition of a knowledge graph here would not be particularly relevant, and we direct the reader to works such as [5] for details on the specific challenges and associated technologies. What is important regarding knowledge graphs in our context is some of the properties they hold: they are flexible, rich representations of knowledge in a domain, that are widely accessible, dynamic, and conceptually defined (through ontologies).

Unsurprisingly, scientific domains, especially biomedicine, have been among the earliest and strongest adopters of knowledge graphs (and, by extension, ontologies). For this reason, these communities have been at the forefront of constructing tools (such as Protégé [6] or bioportal [7]), specific knowledge graphs (such as Bio2RDF [8]) or ontologies (such as the Gene Ontology [9]). Those are commonly used in various systems, in particular, to integrate and exchange information in a way that is interoperable and unambiguous. They also form a basis for describing available data that can be analyzed and processed.

An idea that has been explored extensively is that knowledge graphs can help in conducting scientific activities. In particular, there are several works that focus on knowledge graphs to support organizing, exploring, understanding, or extracting information from the scientific literature [10,11]. Other approaches have looked at ways in which knowledge graphs can help share meaningful information about experiments and scientific workflows (for example, [12]) or keep track of claims and results (for example, [13]). Through this, tools (such as [14]) can be built that can help scientists in many of their daily tasks around the scientific activity (such as finding collaborators or summarizing the literature) without necessarily directly contributing to scientific discovery.

More recently, much research has been published on the use of knowledge graphs as input to deep learning models in a more direct support for scientific discovery. Most of these works involve graph neural networks (GNN, [15]) and knowledge graph embedding [16] approaches, that can learn from rich graph structures for tasks such as link prediction [17] or node classification [18]. Methods using link prediction are particularly relevant in this space, since the idea, presented naively, is to try to learn from the known relations between entities in existing knowledge graphs (often extracted from the relevant literature) what relations might exist between other entities. This has been used, for example, in the context of drug discovery to find unknown interactions between proteins, genes, and diseases [19].

To summarize, we could argue that in most of the works described above and the many others that exist, knowledge graphs play a "service" role: they serve the scientific activity as a way to support its structuring, organization, and description, or as a format/a structure for the data being exploited by machine learning models. As discussed in [20], if we want to support scientific *discovery*, an important aspect to take into account is that discoveries are fundamentally dependent on and connected to existing knowledge of the scientific domain. When it comes to AI-led scientific discoveries (see, for example, [21]), this is even more critical. Any model, and any result obtained from applying it, might be based on a hidden discovery that can only become apparent once the model itself is aligned to existing scientific knowledge. The opportunity becomes clear for knowledge graphs to play a role in this, as wide, rich, and shared representations of such scientific knowledge.

3. On mechanistic interpretation

The need to make machine learning models interpretable no longer needs to be explained in detail. Recent AI works relying increasingly on large neural networks might be very efficient (in terms of pure performance), but remain unscrutinizable due to their complexity and the fact that they construct answers in a way which is distributed over millions (and sometimes billions) of small computations. Closely related to explainable AI, the field of interpretable AI [22] aims to find ways in which the general behavior of a model, the way it obtains results, and the aspects of the data on which it relies can be understood. There are many approaches to interpretability, including the "distillation" of simpler, intrinsically interpretable surrogate models, such as decision trees, from their output (see, for example, [23]) or employing feature-based explainability methods globally [24]. It is worth mentioning that such methods have already been made to rely on knowledge graphs, providing external knowledge on which to build abstractions of features or interpretable models (see, for example, [25,26]).

Mechanistic interpretation is a particular method for machine learning models (especially neural networks) interpretability that relies on inspecting and analyzing the inner workings of the model itself [27]. Such approaches have become particularly interesting due to the increasing deployment of large language models (LLMs) having apparent emerging capabilities that cannot be easily explained from their

training and application [28]. Both [27,28] provide a taxonomy of approaches and techniques for achieving mechanistic interpretability, but we can mention two that appear particularly relevant here: the abstraction of sub-networks within the neural network and the abstraction of features as represented by activations and weights within the network. Circuit finding is an example of the former, where modules of neurons and connections are identified that appear to carry a particular function within the model (see, for example, [29]). Recent works carried out on LLMs on the topic of "monosemanticity" are an example of the latter, where methods such as dictionary learning through sparse autoencoders are used to disentangle the way neurons play a role in the representation of multiple features (polysemanticity), to identify those features and analyze how they contribute to the results (see, for example, [30,31]).

4. On aligning knowledge graphs to mechanistic interpretations

Mechanistic interpretations help us to inspect the inner workings of neural networks to abstract features used by those networks or the functions they implement. Knowledge graphs provide conceptual knowledge about the domain in which these neural networks operate. When presented in this way, the notion that the abstracted features and functions of mechanistic interpretations would benefit from being aligned to knowledge in the form of concepts, relations, and entities in knowledge graphs appears evident. In other words, relating in this way "interpretations" that remain low level and focused on the model itself to broader conceptual entities would achieve to place the learned understanding of the data acquired by the neural network within the knowledge system of the domain of interest.

This would require understanding how features and representations within the neural network somehow correspond to identified conceptual entities. At the forefront of this are works in machine vision, where correlations are found between neuron activations and the appearance of specific elements in an image [32,33]. Much more explicit on finding "concepts", [34] proposes a more general approach (Testing Concept Activation Vectors, TCAV) to identify whether visual concepts exist (e.g., "stripes on a zebra") within a neural network for image classification. Our own work took this idea a step further and tested how self-organizing maps can be used to visualize and assess the presence and influence of concepts, possibly coming from knowledge graphs, within the activations of different layers of a neural network [35]. In another approach, [36] relies on concept induction and a 2 million concept hierarchy from Wikipedia to derive explanations from the neuron activations of a CNN-based model.

In our domain of interest, this could lead to the ability to probe a neural network for the contribution of specific aspects to the results it obtained. We could answer questions of the form "is it important to know that the material contains metals to calculate its thermal conductivity?" or "is knowing that this is a high-energy material more or less useful than knowing that it has the structure of a rocksalt?". We could further verify whether or not the model effectively relies on notions that exist in more or less formalized theories in the domain, for example that materials with low thermal conductivity would generally be found in those with a bandgap lower than 1.5 eV. Of course, this could also be done by statistically analyzing the input and output of the model to find correlation with the relevant concepts in knowledge graphs (as done, for example, in [25]). However, only by identifying these concepts within the structures and activations of the model itself can we actually verify that they play a role in the prediction and how they relate in this task to other identifiable concepts.

However, those techniques so far have only allowed us to find out how known concepts, i.e. those we are specifically looking for, appear within neural networks. In most mechanistic interpretation approaches, the method used generally goes in the opposite direction: structures, groups or patterns are found that are expected to correspond to features or functions, which are then inspected and analyzed to understand to what human-interpretable notion they might correspond. In other words, they are manually named in a post hoc manner.

A key point here is that we could rely on the availability of broad knowledge graphs, explore and search them, as a way to provide a conceptual understanding of what these structures, groups, and patterns represent. In other words, mechanistic interpretation can show how an emerging, undefined concept is expressed and used in a neural network, and knowledge graphs can be used to find which concept it is and how it connects to broader knowledge in the domain.

To achieve that, techniques would have to be developed to align mechanistic interpretations to the entities and properties in knowledge graphs. In [37], an approach was presented where online knowledge graphs are crawled to find combinations of properties of entities that could best explain a subset of the entities in a larger dataset. Similarly, in [25], a knowledge graph was used to abstract the entities representing the input of a neural network and provide interpretations on the basis of rules connecting these properties to explanations of the results. In both cases, the connection is based on first having aligned the inputs of the network to entities of relevant knowledge graphs, and then using those to create extensional descriptions of the features or patterns identified from the behavior of the model to explore. A similar approach could be used, provided that the initial alignment between the input samples of the network and entities of knowledge graphs is carried out, if achievable.

Using such an approach could represent a starting point to aligning patterns from mechanistic interpretations, provided that they can be extensionally described, to concepts and properties in knowledge graphs, provided that those exist and are already connected to the data relevant to the model. However, these are strong constraints that might not be easily fulfilled. As a key challenge towards a more effective ability to align patterns from mechanistic interpretation to concepts and properties of knowledge graphs, the availability of representations of those concepts and properties that are directly comparable to those representing patterns of mechanistic interpretations (e.g. activation vectors) might be required. In this sense, emerging work around knowledge graph embeddings already mentioned above, as well as neural models of ontologies (see, for example, [38,39]) could play a key role. Whether available knowledge graphs can be indexed so that a mapping between a specific model's representation of patterns and the generic neural representation of knowledge graph entities can easily be found remains an open question.

Now, what if it fails? What if there seems to be, according to mechanistic interpretation, a concept or a group of interacting concepts that is strongly involved in the prediction and that we cannot find in knowledge graphs? In our example, are there abstract features involved in whether a material has high or low thermal conductivity that do not relate to any entity even in the best knowledge graphs we could build to capture current knowledge in physics? Could that mean that our model, through its ability to analyze large quantities of data, has found a mechanism, a property, a phenomena of which we might not be aware? It is a possibility, and the reason why we believe that aligning the analysis of machine learning models from the point of view of mechanistic interpretation to knowledge graphs could become one of the most significant approaches to (scientific) (knowledge) discovery: finding what the model might have relearned from what we know and where there might be gaps in our knowledge.

5. Conclusion

This paints a picture of science where machine learning is not only used to predict but where learning to predict is a step in a process of knowledge extraction and discovery. Machine learning models find patterns, and there is a chance that among these are phenomena not yet understood or formalized.

There are naturally many challenges in achieving this in a way that can actually integrate in the process of scientific discovery, including:

Scale: As described in [31], some approaches to mechanistic interpretation are already very time-consuming and resource-consuming activities when performed on large models, and it is only half of the process. What is described here also assumes that there are a large number of large knowledge graphs that could include concepts and properties aligning with the patterns found through mechanistic interpretation. Exploring those concepts and properties, identifying the most relevant, mapping them to patterns identified through mechanistic interpretations, ranking them, and presenting them in their context, with their connections to others are all tasks for which new solutions are required, if only because of the scale at which they need to operate.

Accessibility of knowledge graphs: Related to the point above, in some way, what we envision in this article is also assuming that those large numbers of large knowledge graphs are easily accessible, which is far from being the case today considering that existing knowledge graphs are naturally distributed over the web, often not easily findable, and provided through systems suffering from availability issues. Several indexes and search engines for semantic web resources [40] (ontologies and entities) have been developed in the past, but most of them have disappeared and would need to be updated to operate in the manner required here. In addition, as mentioned above, to support the suggested process, knowledge graphs would have to be indexed through representations that are suitable to be mapped to neural patterns found by mechanistic interpretation.

Completeness and suitability of knowledge graphs: While many knowledge graphs exist in many scientific disciplines, not all the needs of the process described in this article are currently covered for all domains and areas to which they could apply. For instance, to the best of our knowledge, there is no material knowledge graph that would currently cover the examples mentioned in relation to our use case. In addition, even when they exist, science-related knowledge graphs might be built for the purpose of information exchange, accessibility, and interoperability and, as a result, may not be suitable to be used as a deposit of current, up-to-date theoretical and experimental knowledge in the corresponding discipline.

As a result of those challenges and of the inherent limitations of machine learning models, we expect most of what is discovered through the suggested process to be either shortcuts taken by the model (invalid results from biased datasets) or representative of our inability to obtain the appropriate knowledge from insufficiently available knowledge graphs. Those failures are also important since they might help us building better datasets, better knowledge graphs, and better models.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Mathieu d'Aquin reports was provided by Lorraine Research Laboratory in Computer Science and its Applications. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] I. Asimov, True Love, 1977, short story.

[2] A. Dunn, Q. Wang, A. Ganose, D. Dopp, A. Jain, Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm, npj Comput. Mater. 6 (1) (2020) 138.

[3] A. Toner-Rodgers, Artificial intelligence, scientific discovery, and product innovation, 2024, URL https://aidantr.github.io/files/AI_innovation.pdf.

[4] L. Klochko, M. d'Aquin, A. Togo, L. Chaput, Transfer learning for deep learning-based prediction of lattice thermal conductivity, 2024, URL https://arxiv.org/abs/2411.18259. arXiv:2411.18259.

[5] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G.D. Melo, C. Gutierrez, S. Kirrane, J.E.L. Gayo, R. Navigli, S. Neumaier, et al., Knowledge graphs, ACM Comput. Surv. (Csur) 54 (4) (2021) 1–37.

[6] M.A. Musen, The protégé project: a look back and a look forward, AI matters 1 (4) (2015) 4–12.

[7] N.F. Noy, N.H. Shah, P.L. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D.L. Rubin, M.-A. Storey, C.G. Chute, et al., BioPortal: ontologies and integrated data resources at the click of a mouse, Nucleic Acids Res. 37 (suppl_2) (2009) W170–W173.

[8] M. Dumontier, A. Callahan, J. Cruz-Toledo, P. Ansell, V. Emonet, F. Belleau, A. Droit, Bio2RDF release 3: a larger connected network of linked data for the life sciences, in: Proceedings of the 2014 International Conference on Posters & Demonstrations Track, 2014, pp. 401–404.

[9] Gene Ontology Consortium, The gene ontology resource: 20 years and still going strong, Nucleic Acids Res. 47 (D1) (2019) D330–D338.

[10] S. Auer, V. Kovtun, M. Prinz, A. Kasprzik, M. Stocker, M.E. Vidal, Towards a knowledge graph for science, in: Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, 2018, http://dx.doi.org/10.1145/3227609.3227689.

[11] D. Dessí, F. Osborne, D. Reforgiato Recupero, D. Buscaldi, E. Motta, CS-kg: A large-scale knowledge graph of research entities and claims in computer science, in: International Semantic Web Conference, Springer, 2022, pp. 678–696.

[12] J. Wen, Z. Zhou, Y. Wang, W. Gaaloul, Y. Duan, Discovering crossing-workflow fragments based on activity knowledge graph, in: On the Move To Meaningful Internet Systems: OTM 2019 Conferences: Confederated International Conferences: CoopIS, ODBASE, C&TC 2019, Rhodes, Greece, October 21–25, 2019, Proceedings, Springer, 2019, pp. 515–532.

[13] R. Han, S. Byna, H. Tang, B. Dong, M. Zheng, PROV-IO: An I/O-centric provenance framework for scientific data on hpc systems, in: Proceedings of the 31st International Symposium on High-Performance Parallel and Distributed Computing, 2022, pp. 213–226.

[14] L. Kovriguina, L. Aung, P. Haase, N. Heist, D. Lamprecht, S. Heiß, Enhancing scientific discovery and decision-making: A knowledge graph-based research support system, in: A. Salatino, A. Mannocci, F. Osborne, S. Schimmler, G. Rehm (Eds.), Proceedings of the 4th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment, 2024.

[15] X. Liu, J. Chen, Q. Wen, A survey on graph classification and link prediction based on gnn, 2023, arXiv preprint arXiv:2307.00865.

[16] Q. Wang, Z. Mao, B. Wang, L. Guo, Knowledge graph embedding: A survey of approaches and applications, IEEE Trans. Knowl. Data Eng. 29 (12) (2017) 2724–2743.

[17] A. Kumar, S.S. Singh, K. Singh, B. Biswas, Link prediction techniques, applications, and performance: A survey, Phys. A 553 (2020) 124289.

[18] S. Xiao, S. Wang, Y. Dai, W. Guo, Graph neural networks in node classification: survey and evaluation, Mach. Vis. Appl. 33 (1) (2022) 4.

[19] J. Hu, R. Lepore, R.J. Dobson, A. Al-Chalabi, D. M. Bean, A. Iacoangeli, DGLinker: flexible knowledge-graph prediction of disease–gene associations, Nucleic Acids Res. 49 (W1) (2021) W153–W161.

[20] P. Langley, Integrated systems for computational scientific discovery, in: Proceedings of the AAAI-24 Special Track AI for Social Impact, Senior Member Presentations, New Faculty Highlights, Journal Track, 2024.

[21] H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Van Katwyk, A. Deac, A. Anandkumar, K. Bergen, C.P. Gomes, S. Ho, P. Kohli, J. Lasenby, J. Leskovec, T.-Y. Liu, A. Manrai, D. Marks, B. Ramsundar, L. Song, J. Sun, J. Tang, P. Veličković, M. Welling, L. Zhang, C.W. Coley, Y. Bengio, M. Zitnik, Scientific discovery in the age of artificial intelligence, Nature 620 (7972) (2023) 47–60, http://dx.doi.org/10.1038/s41586-023-06221-2.

[22] Y. Zhang, P. Tiňo, A. Leonardis, K. Tang, A survey on neural network interpretability, IEEE Trans. Emerging Top. Comput. Intell. 5 (5) (2021) 726–742.

[23] J. Lu, Q. Wang, Z. Zhang, J. Tang, M. Cui, X. Chen, Q. Liu, Z. Fei, X. Qiao, Surrogate modeling-based multi-objective optimization for the integrated distillation processes, Chem. Eng. Process.-Process Intensif. 159 (2021) 108224.

[24] Y. Yang, Y. Yuan, Z. Han, G. Liu, Interpretability analysis for thermal sensation machine learning models: An exploration based on the SHAP approach, Indoor Air 32 (2) (2022) e12984.

[25] A. Nikolov, M. d'Aquin, Uncovering semantic bias in neural network models using a knowledge graph, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 1175–1184.

[26] I. Tiddi, F. Lécué, P. Hitzler, Knowledge Graphs for Explainable Artificial Intelligence: Foundations, Applications and Challenges, IOS Press, 2020.
[27] T. Räuker, A. Ho, S. Casper, D. Hadfield-Menell, Toward transparent AI: A survey on interpreting the inner structures of deep neural networks, in: Proceedings of the 2023 IEEE Conference on Secure and Trustworthy Machine Learning, SaTML, 2023, http://dx.doi.org/10.48550/arXiv.2207.13243.
[28] D. Rai, Y. Zhou, S. Feng, A. Saparov, Z. Yao, A practical review of mechanistic interpretability for transformer-based language models, 2024, http://dx.doi.org/10.48550/arXiv.2407.02646, URL http://arxiv.org/abs/2407.02646, arXiv:2407.02646 [cs].
[29] A. Conmy, A. Mavor-Parker, A. Lynch, S. Heimersheim, A. Garriga-Alonso, Towards automated circuit discovery for mechanistic interpretability, Adv. Neural Inf. Process. Syst. 36 (2023) 16318–16352.
[30] A.S. Jermyn, N. Schiefer, E. Hubinger, Engineering monosemanticity in toy models, 2022, arXiv preprint arXiv:2211.09169.
[31] A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, et al., Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. Transformer circuits thread, 2024.
[32] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, Springer, 2014, pp. 818–833.
[33] R. Fong, A. Vedaldi, Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8730–8738.
[34] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al., Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), in: International Conference on Machine Learning, PMLR, 2018, pp. 2668–2677.
[35] M. d'Aquin, Finding concept representations in neural networks with self-organizing maps, in: Proceedings of the 12th Knowledge Capture Conference 2023, 2023, pp. 53–60.
[36] A. Dalal, R. Rayan, A. Barua, E.Y. Vasserman, M.K. Sarker, P. Hitzler, On the value of labeled data and symbolic methods for hidden neuron activation analysis, in: International Conference on Neural-Symbolic Learning and Reasoning, Springer, 2024, pp. 109–131.
[37] I. Tiddi, M. d'Aquin, E. Motta, Dedalo: Looking for clusters explanations in a labyrinth of linked data, in: The Semantic Web: Trends and Challenges: 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings 11, Springer, 2014, pp. 333–348.
[38] J. Chen, P. Hu, E. Jimenez-Ruiz, O.M. Holter, D. Antonyrajah, I. Horrocks, OWL2Vec*: embedding of OWL ontologies, Mach. Learn. 110 (7) (2021) 1813–1845, http://dx.doi.org/10.1007/s10994-021-05997-6.
[39] M. d'Aquin, E. Nauer, Ontological relations from word embeddings, 2024, http://dx.doi.org/10.48550/arXiv.2408.00444, arXiv:2408.00444. URL http://arxiv.org/abs/2408.00444.
[40] M. d'Aquin, L. Ding, E. Motta, Semantic web search engines, in: Handbook of Semantic Web Technologies, Springer, 2011.