

Graph-based Neural Modules to Inspect Attention-based Architectures: A Position Paper

Breno W. Carvalho^{1,2}, Artur S. d'Avila Garcez³ and Luís C. Lamb^{2,4}

¹IBM Research, Rio de Janeiro, Brazil

²UFRGS, Porto Alegre, Brazil

³Department of Computer Science, City, University of London

⁴MIT Sloan School of Management, Cambridge, MA

Abstract

Encoder-decoder architectures are prominent building blocks of state-of-the-art solutions for tasks across multiple fields where deep learning (DL) or foundation models play a key role. Although there is a growing community working on the provision of interpretation for DL models as well as considerable work in the neuro-symbolic community seeking to integrate symbolic representations and DL, many open questions remain around the need for better tools for visualization of the inner workings of DL architectures. In particular, encoder-decoder models offer an exciting opportunity for visualization and editing by humans of the knowledge implicitly represented in model weights. In this work, we explore ways to create an abstraction for segments of the network as a two-way graph-based representation. Changes to this graph structure should be reflected directly in the underlying tensor representations. Such two-way graph representation enables new neuro-symbolic systems by leveraging the pattern recognition capabilities of the encoder-decoder along with symbolic reasoning carried out on the graphs. The approach is expected to produce new ways of interacting with DL models but also to improve performance as a result of the combination of learning and reasoning capabilities.

Keywords

Neuro-symbolic models, Deep Learning explainability, Model introspection

1. Introduction

We live in hustling times for Artificial Intelligence (AI) research with many optimistic perspectives from researchers from Computer Science and other areas. We are entering an era where our models not only can generalize from examples of a given task, but given the appropriate context and conditions, they can generalize across different tasks, indicating an emergent phenomenon that is almost impossible to predict. Such models, notably deep learning (DL) ones are powerful and influential, yet the current drawbacks in design and reliability are now obvious. Those

AAAI 2022 FALL SYMPOSIUM SERIES, *Thinking Fast and Slow and Other Cognitive Theories in AI*, November 17-19, Westin Arlington Gateway in Arlington, Virginia, USA

✉ brenow@ibm.com (B. W. Carvalho); a.garcez@city.ac.uk (A. S. d. Garcez); lamb@inf.ufrgs.br (L. C. Lamb)

🌐 <https://brenowca.github.io/> (B. W. Carvalho); <https://www.staff.city.ac.uk/~aag/> (A. S. d. Garcez);

<https://www.inf.ufrgs.br/~lamb/> (L. C. Lamb)

🆔 0000-0002-2780-1735 (B. W. Carvalho); 0000-0001-7375-9518 (A. S. d. Garcez); 0000-0003-1571-165X (L. C. Lamb)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

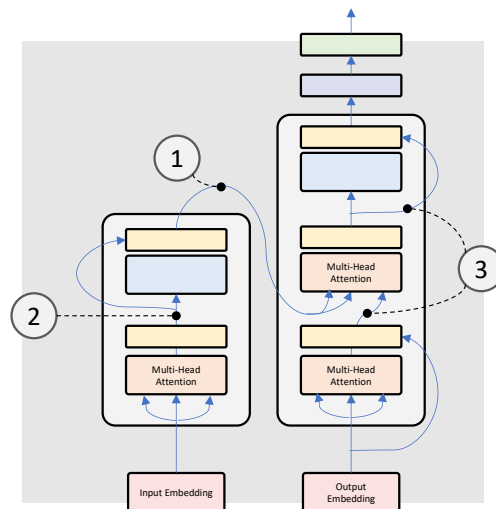


Figure 1: Standard transformer architecture example, adapted from [1] with a few possible entry points for introspection modules. Each one of those prospective entry points carries information about the model state and context. In 1) the decoder part of the transformer uses the embedded context information from the encoder to generate the output sequence in the transduction task. In 2) after the masked multi-head attention module of the *encoder* component to inspect which concepts are more strongly related than others in the construction of contextual information. In 3) information from the masked multi-head attention modules of the *decoder* component is used to inspect which concepts are more relevant to processing the context in light of the current output state.

models, like fun-house mirrors, can distort aspects of their training data and generate false affirmations even though they were fed with the correct facts necessary for the answer. They can also lean toward unethical biases and make unsupported claims. It is therefore important to be able to edit a model's answer and context to first understand the supporting facts or assumptions behind a given answer and second, correct or provide new contextual information so as to increase trust in the model's outputs.

One approach in this direction is to create introspection entry points into the models so as to *peek* into the operations of an intermediate layer (as depicted in Figure 1), interacting directly with it via graph editing. Those introspection entry points are neural layers not aimed to contribute to performance in any single task but to add an extra layer to the developer's interaction with the model by providing an abstraction to inspect specific parts of the network. There are already a few relevant methods in the literature, some of which are listed in the survey [2] and use graph representations as part of the network structure itself. At the same time, the field of neuro-symbolic (NeSy) AI research investigates ways of merging explicit symbolic knowledge with many such sub-symbolic neural approaches [3].

At the Robert S. Engelmore Memorial Lecture, during the AAAI Conference on Artificial Intelligence, New York, February 10th, 2020, in a talk entitled *The Third AI Summer*, Henry Kautz presented a taxonomy for NeSy models, with six main categories [4]. Adding those inspection nodes, which ideally should not harm model performance at the target task, falls into the *Neuro; Symbolic* category in Kautz's taxonomy. This category includes systems where a neural

component and a symbolic component mutually share information, each with a different goal. In this way, the inspection module can help edit the implicit knowledge stored in the DL model and serve even as a debugging tool. One possible approach to building such inspection modules is based on Graph Convolutional Neural Networks (GNN) [5, 6]. By combining inspection with the model one can perform a secondary task of internal state-representation of the model [2]. As DL models grow in size and complexity, we believe that those introspection tools in the network may allow for a more in-depth understanding of such models. This raises the question of the form of introspection and extensions thereof that might be expected to produce a better understanding of DL models.

2. Short-term Objectives

Our immediate objective is to design and develop an initial tool set for interpreting encoded knowledge in deep learning (large) language models and editing it. We aim to add components to infer a graph structure from specific tensors, so-called entry points, in the model’s architecture. In this way, a natural step is to evaluate the feasibility of implementing different approaches for those graph entry points (see Figure 2a). We call entry points compact and non-sparse tensors along the network pipeline that can also be connected to a GNN to produce a graph abstraction of this stage of the overall network. For now, we consider only DL architectures based on the overall encoder-decoder pattern. Those are usually models used in transduction tasks, which make up a vast class of DL models. Our short-term goals are:

To compare different alternatives for implementing graph constraints without significantly reducing performance. Creating graph constraints to specific entry points means ensuring that the underlying tensor always has a meaningful compact graph representation if the overall model was fed with valid input vectors. To this end, we contemplate adopting a few alternatives, such as having an auxiliary loss function that will signal whether the encoder was able to create a valid graph. It will also signal if the decoder can generate the expected output. Another approach is to train a GNN decoder and apply it to the output of the encoder part. Aiming at language tasks supports both kinds of strategies since there are extensive Language Resources that can be used as an initial graph structure.

To find semantic representations for the edges and nodes. A representation of one embedding without any semantics associated with it might be as daunting as looking at the model itself. We seek to adapt existing work in the literature to be able to create a representation with more meaningful names for entity nodes and connections.

To be able to edit the model to increase performance, remove biases or add new knowledge. Finally, any editing of the graph representation should be reflected in the encoder’s weights and influence the final output of the model, ideally in a way that suggests a causal relationship. This enables us to interact actively and edit context within the model to ensure that the main component of a given output is considered to be relevant and correct by an expert in the domain of interest.

To Design a Method for generating interpretations in a local scope. In the short term, we are concerned with *local interpretations* where we interpret the model components' states and context while processing a specific input. This approach is already insightful for a series of use cases and might be the stepping-stone for developing a global interpretation.

Achieving these goals will require the development of a simple but versatile tool-set that might be useful to the community for inspecting and editing models in diverse research or industrial areas (similarly to Grad-Cam [7] and Bertology [8] and many other helpful visualization tools found to be relevant by DL practitioners).

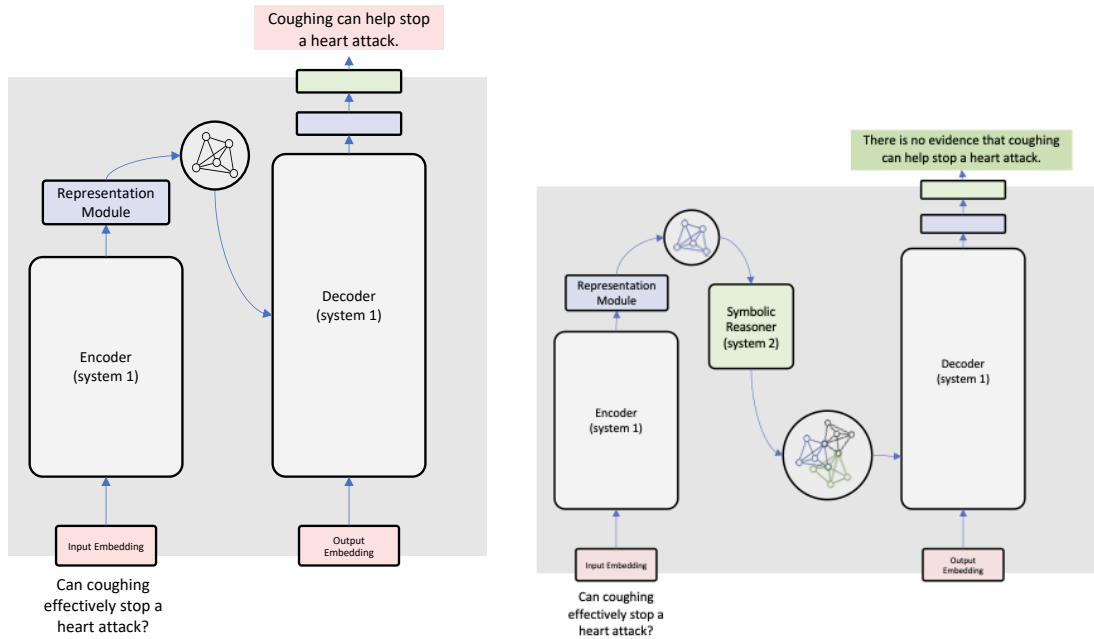
3. Long-term Objectives

Our main goal is to be able to inspect large deep learning language models through different entry points in such a way as to increase trust in their output. Transformer or transformer-inspired architectures dominate the state-of-the-art of such language models. Considering that most transformer architectures follow an encoder-decoder model, fulfilling the above short-term objectives should provide an entry point into inspecting the inner state of the model and its context. We conjecture that having a few editable graph layers attached to a transformer of considerable size will enable an expert to study the relationships of the many abstraction levels one might find by inspecting those layers. Our long-term goals are:

To have a method to enable experts to visualize, edit and interact directly with the language models. One might consider exploring multiple entry points. Manually editing context or knowledge in the network through a graph interface linked to different layers might provide much richer insight. Possibilities for graph-entry points are depicted in Figure 1. The goal is to select tensors that are most descriptive of the model's current state. One can also use this method with visualization methods such as [8].

To design a new family of neurosymbolic models based on systematic interactions with fragments of the underlying architecture. One might also leverage the power of symbolic reasoners to perform expansion and inference on the graph context representation to fix or enhance the model's output. This approach is much aligned with a fast-slow AI perspective, where the learning of the language model represents the fast, pattern-matching, system 1, and the slower, graph-based reasoning represents system 2 (see Figure 2b). In this figure, we illustrate that a symbolic expansion of the graph representation of specific parts of the network might be used to provide the model with enough information to go from a wrong answer, as depicted in Figure 2a, to the correct answer (see Figure 2b).

To explore alternative representations of the entry points that are most useful in a specific context. As we explore larger language models we might find that different entry points might benefit from different representation formats. It might even be the case that reusing smaller transformers will allow us to have natural language representation for some model entry points. An analogy would be when humans perform complex tasks and mentally talk to themselves while performing the task. Another class of representation alternatives to



- (a) Encoder-decoder example with an entry point for an interpretation module. One might explore multiple representation languages depending on the resources available. In this case, the model outputs an incorrect answer for a question from the TruthfulQA dataset, [9], due to incomplete context information, that we hope to make explicit with the graph representation.
- (b) Illustrative representation of neuro-symbolic reasoning to complement the pattern learning capabilities of attention-based models. A GNN entry point is interpreted as an AI system 2 component via the application of a symbolic reasoner.

Figure 2: Example of the graph-based inspection approach being used as an inspection element and also as a possible modification to the model architecture as well.

be explored consists of hyper-graphs that might have a more suitable role in explaining the network’s inference mechanism since they are able to convey recursive levels of abstraction in the information being represented.

To develop a method where the training data gives the representation semantics and not necessarily by external language resources. In an ideal scenario, the only information available during training should be sufficient to build meaningful representations. Even with available linguistic resources, expanding the entry point representation with tokens and concepts from the training data could allow the use of a more robust and extensible representational language such as a graph, hypergraph, or natural language.

To develop a pre-trained zoo of entry point representations for specific architectures that can be further specialized. It should be possible to use a library of visualization

modules pre-trained on commonplace text corpora as a baseline, and specialize it with specific domain corpora. This process might accelerate the inspection process and also diminish the learning curve of experts from different domains on how to use those tools. At this point, we are aiming at modularity. We intend not only to have a methodology but also specific modules that experts from diverse backgrounds can reuse.

To define a method comprising both local and global interpretation scope. In opposition to local scope interpretations, *global interpretations* aim at providing insight into the overall behavior of the model considering an entire task. For instance, understanding what parts of a large language model are responsible for summing two numbers or performing sentiment analysis.

To achieve those objectives, one should build a robust methodology to edit and interact with large language models at multiple levels of abstraction processed by the model. This methodology might enable further insight into the inner workings of such models.

4. Introspection Modules

Adding introspection modules to current deep learning architectures has the potential to enable better interpretability, but also unlock new ways to edit and condition those models. As discussed in detail by this joint initiative [10], large-scale language models (also called foundation models for their re-usability across multiple tasks) pose serious research and societal challenges. We discuss a few of them in this paper’s *Challenges* section.

4.1. Fairness and accountability goals

Although somewhat hard to quantify, fairness and accountability are important to acknowledge in what concerns trust in deep learning models of far-reaching decisions. Unfairness is characterized in [11] into two classes: prediction outcome discrimination, and prediction quality disparity. An example of the first class is when a model predicts an unfavorable treatment of a demographic group, such as bias against women. The second class refers to a model performing poorly for a determined group of individuals. Both cases raise the need to flag such models and correct them. Accountability might be understood as a capability for *post-hoc* inspection of the model in order to make it available for auditing the causes of the model’s outcome [12].

Following [10], Section 4.11, which discusses the concept of interpretability of foundation models, we contemplate three categories of interpretability: *what* (the model’s limits and capabilities), *why* (finding what in the data set is responsible for the model’s output), and *how* (gaining an understanding of how certain parts and mechanisms of the model impact the output). In this work, we focus on the *why* and *how* interpretability questions.

As well as being able to interpret those models, one should also seek to obtain better tools to assist us in accelerating the construction or adaptation of new models.

4.2. Model editing and design goals

By understanding individual model mechanisms, we seek to build a compositional understanding of the complex behavior of a foundation model. A possible way is to interconnect or build larger models from pre-trained components and reduce societal risks by building more reliable and interpretable models through “debugging” tools.

5. Challenges

Creating those representation/introspection modules raises several challenges. Some of these, such as the visual interpretation of self-attention heads, have been addressed to some extent in the literature. Others remain as yet to be explored.

The *why* and *how* interpretability questions from [10], as mentioned in the previous section, are essential for the understanding of large language models, but face many open challenges:

5.1. Challenges on describing *why* (model behavior) and *how* (model mechanisms)

Studying the influence of the model’s input might be performed by carefully studying how the behavior of the model changes with changes in the input. The ability to peek inside the model also might be used to provide insight into which parts of the input are more relevant to the model. The task of generating interpretations for the inner workings of a language model poses further challenges:

1) It is not clear how the components of those language models are interconnected. Although the overall design of a model’s architecture may be clear, the actual weights of such units have emergent behavior that might be fairly hard to predict. A central question to a fair representation of the mechanisms inside a language model is that of to what extent those mechanisms behave as one coherent model or as many models. As the model learns to generalize across multiple tasks, it might be the case that its components find weight sub-spaces that virtually correspond to many sub-models, possibly one for each task. It isn’t clear yet to which point in this spectrum current foundation models belong. This lack of understanding and clarity leads to our next challenge.

2) Local and global interpretation methods might have pitfalls related to the complexity of such language models. Given the highly complex behavior of those models and their emergent properties, it is unwise to jump to conclusions based on a single global or local interpretation method, even though there are both local[13, 14] and global[15, 16] methods for providing insights into task-specific models. For instance, considering that those large models might find weight sub-spaces for distinct tasks it might be the case that studying the model for a few of those sub-tasks doesn’t provide a fair picture of its behavior in other tasks or even for other similar inputs.

3) The inter-relationship of different components in the model might be counter-intuitive or hard to determine. It is not clear whether semantically related mechanisms, e.g. the parts of a language model responsible for summing up two numbers, are the same as the ones used to perform a related task, such as e.g. summing two numbers expressed as numeral nouns.

5.2. Challenges about the introspection modules themselves

Besides the inherent challenges of deriving interpretations from huge language models, there are also intrinsic points to be evaluated in the design of the introspection model. There are questions and decisions to be made about the desired representational language and format, and there are also considerations about the architecture of those modules and their refinement procedure.

1) Creating meaningful interpretations assumes an underlying vocabulary and semantics. This can be thought of as the symbol grounding problem, i.e. the definition of a vocabulary for the representation generated by the introspection modules. This might be achieved through the use of an underlying ontology that will constrain the interpretation space into a set of predefined concepts and vocabulary. It remains unclear how the semantics of the ontology will be transferred to the interpretation. It might be achieved as a secondary task happening during the training of the language model that we want to inspect.

2) Different abstraction levels must be considered in order to avoid overly complicated and misleading interpretations. This happens because the size of a given interpretation may be prohibitive, both in terms of meaningful visualization and interpretability. To circumvent this one needs to employ different levels of abstraction, although those model layers closer to the output layers seem to pertain to more abstract features of the data set. Walking back and forth between those different levels of abstraction reinforces the need to guarantee consistency in the interpretations' vocabulary from different entry points.

3) Interpretations should be consistent and reliable across a specific domain or data set. This is not an easy requirement to enforce. It might be the case that given the complexity of those models, consistency can be enforced only partially for most of the time. We also need to consider consistency across related input or context. For instance, it should not be the case that two very similar text prompts result in very different interpretations when looking at the same entry point in the model. A possible approach is to anchor those graph representations on concepts and constructs from public linguistic resources, such as FrameNet [17], VerbNet [18], WordNet [19] and Propbank [20], all connected by the SemLink project [21]. The process of generating a consistent interpretation must also be subject to auditing to ensure the faithfulness of the entire system.

4) Considering that the introspection module can be viewed as a black-box component, one must devise metrics of trustworthiness specific to it. The GNN or other DL model

used to generate representations from parts of the language model is also a black-box module that might pose itself some fairness questions [22]. We hope that it will prove feasible to keep those models relatively small. It is important to make sure that relying on those modules will be much less complex than the model being studied. In such case, existing visualization methods such as [8, 23] may suffice to audit them and improve trust.

In summary, the goal of adding a human-editable and interpretable representation of context and the state of specific parts of a language model has many open questions that to be solved will require drawing contributions from different areas of expertise. One way to face some of the challenges outlined here is to combine approaches that use both linguistic resources to provide the needed representation building blocks (e.g. concepts and relations) and to use tokens and concepts from the training data set as building blocks.

6. Outlook

As mentioned in previous sections, this kind of neural-symbolic work draws from a diverse mixture of research areas in AI and Computer Science more generally.

6.1. Peeking inside DL models:

While characterizing and describing the behavior of a deep learning model, one might consider the data set and the predicted outputs, or also the state variations of the models. It is also possible to build interpretations about the model behavior on specific data samples or trying to understand how it behaves in a broader sense when performing a particular task.

Local scope interpretation methods target insight into the model’s behavior on specific inputs or contrasting a given set of data points. Examples include [14], an approach that uses influence functions from robust statistics to understand how perturbations in the training set influence the model behavior; [13] uses the concept of interpretations as meta-predictors, i.e. one can use an explanation to predict the output of a model, and they instantiate this approach using image classification masks. By contrast, global methods are used to describe general aspects and behaviors of the model. Interesting examples include [15] and [16]. Both local and global methods treat the language models as black-box objects to be understood through their inputs and outputs.

Other than looking at the training set and the models’ outputs, another class of interpretation methods that is more aligned with the work proposed here, consists of models that extract information from the model weights. This includes GradCam [7], which does backward propagation of gradients to estimate which pixels of the input were more relevant for a Convolutional Network output; [8] provides a tool-set for visualizing the weights of the self-attention heads of BERT-like models. Although those methods can be very insightful when analyzing task-specific models, considering that current language models might have distinct behavior across different tasks, it is unclear how insightful those methods could be in this scenario. Most likely, the designers and analysts of such models would need to employ multiple methods of interpretation and visualization to have a fair understanding of the models’ behavior.

As an alternative to the line of work described in this paper, where we propose the use of external (and hopefully auditable) modules to peek into the language models, one can train the

model itself to generate an explanation for its output [24]. This approach faces some skepticism since language models can generate plausible, although untrue, outputs and nothing prevents this behavior from extending to the model's self-explanations.

6.2. Graph deconvolution and Neurosymbolic AI

We regard GNNs and neuro-graph algorithms in general as a promising general approach for both local and global interpretation.

Graph convolution is defined by analogy to convolutional layers over Euclidean data [25, 26] and it is an important tool to infuse deep learning models with relational knowledge [2]. Conversely, we conjecture that using specific decoder units to extract graph representations from the entry point tensors (even if having to condition the training of the underlying language model) might be a worthwhile approach to obtain meaningful interpretations from language models.

We might consider each introspection module as a simple decoder unit that maps the encoded embedding into the interpretation space. As mentioned in the Challenges section, one might initially use linguist resources conditioned on the sentence processed by the language model to assemble the vocabulary used to form the interpretations. A possible approach is to use heuristics to get concepts from the language model prompt and build a simple graph, even combining it with a syntax or dependency tree to have an initial graph to use and evaluate the decoder unit used as an introspection module. In this sense, there are a few approaches in the literature that one might draw inspiration from. For instance, there is work on graph encoders [27], and also deconvolutional graph networks [28] to derive a graph from the entry point embedding. At this stage, it might be unclear how to define the appropriate abstraction level of the interpretation.

Instead of building the modules as decoder units trained on a heuristic graph, another approach is to adapt variational graph autoencoders [29]. A variation of the encoding part could output a Graph Convolutional Network-like embedding, and the decoder would try to retrieve the original input of the language model (assuming that the entry point embedding might encode a smooth topological representation of concepts). This assumption is not guaranteed to hold, but it is an interesting issue to investigate. One might use deconvolution networks [28] to infer a graph without prior conditioning. The intuition behind this rationale is that this deconvolution approach would create representations without any supervision, which might be too complicated or obscure to be used as interpretation.

As stated in our short and long-term objectives, inferring an arbitrary graph from those entry points is not enough for our interpretation and editing purposes. Those graphs also need to have a clear semantics that correlates with the model's output and with commonsense and domain knowledge as well.

One might wonder why bother with graph representations and not try to go directly to building natural language interpretations. We conjecture that natural language will lack explicit restrictions that are needed for clarity of a given interpretation and also that the resulting natural language introspection module would need to be almost as complex as the language model itself, thus defeating its original purpose.

There are several neurosymbolic methods that aim to infuse the deep learning model archi-

tructures with symbolic or logical structures [3], including logic tensor networks (LTN) [30] and logical neural networks (LNN) [31]. They can also be used as parts of the introspection modules described here in order to make DL models more auditable and reliable.

7. Conclusion

Large language models are becoming state-of-the-art in many tasks and fields, and as they grow in popularity, they also grow in complexity. Those models usually follow a transduction encoder-decoder pattern based on self-attention mechanisms that are repeated multiple times to a point where they are so complex that surprising behavior emerges from them. In this paper, we state the need for inspection methods to both characterize and describe those models' behaviors. Despite the consolidated literature on deep learning interpretability, the community still needs to build different tools to inspect different aspects of those models. We believe that graph-based algorithms, in particular, deconvolution graph networks and variational graph auto-encoders might be key to generating formal yet flexible interpretations of parts of such large language models.

Interpretability of language models is a blooming field as those models keep growing and reaching multiple uses in society. There is a vast literature to support this field, but there is a flagrant need for new methods and approaches to deal with such models' high generalizability. Being able to audit and edit aspects of the model during development and deployment should allow the community to correct flawed inferences performed by those models and adjust to counter certain unethical biases. Thus inspection tools have the potential for tremendous impact if they can help us develop models we can trust and rely on to shape the further development of AI.

We envision a roadmap for building tools with which multidisciplinary communities can start inspecting small transduction models based on self-attention, and progressively scale up to large-scale language models. By doing so, those communities are empowered to shape the widely spread use cases of such models in society. To this end, we hope to start a conversation to unify our efforts, lower the threshold for other machine learning researchers to join us, and bring these communities closer together with a common language.

8. Acknowledgments

We thank Viviane Torres, Sandro Rama Fiorini, Emilio Ashton Brasil, and Renato Cerqueira for insightful conversations on the topic. Luis Lamb was supported in part by CAPES and CNPq, Brazil.

References

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), *Advances in Neural Information Processing*

- Systems, volume 30, Curran Associates, Inc., 2017. URL: <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>.
- [2] L. C. Lamb, A. S. d’Avila Garcez, M. Gori, M. O. R. Prates, P. H. C. Avelar, M. Y. Vardi, Graph neural networks meet neural-symbolic computing: A survey and perspective, in: *IJCAI 2020*, ijcai.org, 2020, pp. 4877–4884. URL: <https://doi.org/10.24963/ijcai.2020/679>. doi:10.24963/ijcai.2020/679.
 - [3] A. d’Avila Garcez, L. C. Lamb, Neurosymbolic AI: the 3rd wave, *CoRR abs/2012.05876* (2020). URL: <https://arxiv.org/abs/2012.05876>. arXiv:2012.05876.
 - [4] H. A. Kautz, The third AI summer: AAAI robert s. engelmore memorial lecture, *AI Mag.* 43 (2022) 93–104. URL: <https://doi.org/10.1609/aimag.v43i1.19122>. doi:10.1609/aimag.v43i1.19122.
 - [5] Z. Liu, J. Zhou, Introduction to graph neural networks, *Synthesis Lectures on Artificial Intelligence and Machine Learning* 14 (2020) 1–127.
 - [6] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, S. Y. Philip, A comprehensive survey on graph neural networks, *IEEE transactions on neural networks and learning systems* 32 (2020) 4–24.
 - [7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
 - [8] A. Rogers, O. Kovaleva, A. Rumshisky, A primer in bertology: What we know about how bert works, *Transactions of the Association for Computational Linguistics* 8 (2020) 842–866.
 - [9] S. Lin, J. Hilton, O. Evans, Truthfulqa: Measuring how models mimic human falsehoods, *CoRR abs/2109.07958* (2021). URL: <https://arxiv.org/abs/2109.07958>. arXiv:2109.07958.
 - [10] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. S. Chatterji, A. S. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. D. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. S. Krass, R. Krishna, R. Kuditipudi, et al., On the opportunities and risks of foundation models, *CoRR abs/2108.07258* (2021). URL: <https://arxiv.org/abs/2108.07258>. arXiv:2108.07258.
 - [11] M. Du, F. Yang, N. Zou, X. Hu, Fairness in deep learning: A computational perspective, *IEEE Intelligent Systems* 36 (2020) 25–34.
 - [12] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao, et al., Interpretability of deep learning models: A survey of results, in: *2017 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation (smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI)*, IEEE, 2017, pp. 1–6.
 - [13] R. C. Fong, A. Vedaldi, Interpretable explanations of black boxes by meaningful perturbation, in: *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3429–3437.

- [14] P. W. Koh, P. Liang, Understanding black-box predictions via influence functions, in: International conference on machine learning, PMLR, 2017, pp. 1885–1894.
- [15] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K.-R. Müller, Unmasking clever hans predictors and assessing what machines really learn, *Nature communications* 10 (2019) 1–8.
- [16] A. W. Thomas, H. R. Heekeren, K.-R. Müller, W. Samek, Analyzing neuroimaging data through recurrent deep learning models, *Frontiers in neuroscience* 13 (2019) 1321.
- [17] C. F. Baker, C. J. Fillmore, J. B. Lowe, The berkeley framenet project, in: COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics, 1998.
- [18] M. Palmer, K. L. Kipper, et al., Verbnnet, *The Oxford Handbook of Cognitive Science* (2004).
- [19] G. A. Miller, Wordnet: A lexical database for english, *Commun. ACM* 38 (1995) 39–41. URL: <https://doi.org/10.1145/219717.219748>. doi:10.1145/219717.219748.
- [20] M. Palmer, D. Gildea, P. Kingsbury, The proposition bank: An annotated corpus of semantic roles, *Computational linguistics* 31 (2005) 71–106.
- [21] M. Palmer, Semlink: Linking propbank, verbnnet and framenet, in: Proceedings of the generative lexicon conference, GenLex-09, Pisa, Italy, 2009, pp. 9–15.
- [22] A. Jacovi, Y. Goldberg, Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness?, *arXiv preprint arXiv:2004.03685* (2020).
- [23] X. Shi, F. Lv, D. Seng, J. Zhang, J. Chen, B. Xing, Visualizing and understanding graph convolutional network, *Multimedia Tools and Applications* 80 (2021) 8355–8375.
- [24] D. C. Elton, Self-explaining ai as an alternative to interpretable ai, in: International conference on artificial general intelligence, Springer, 2020, pp. 95–106.
- [25] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, G. Monfardini, The graph neural network model, *IEEE transactions on neural networks* 20 (2008) 61–80.
- [26] B. Sanchez-Lengeling, E. Reif, A. Pearce, A. B. Wiltschko, A gentle introduction to graph neural networks, *Distill* 6 (2021) e33.
- [27] W. L. Hamilton, R. Ying, J. Leskovec, Representation learning on graphs: Methods and applications, *arXiv preprint arXiv:1709.05584* (2017).
- [28] J. Li, J. Li, Y. Liu, J. Yu, Y. Li, H. Cheng, Deconvolutional networks on graph data, *Advances in Neural Information Processing Systems* 34 (2021) 21019–21030.
- [29] T. N. Kipf, M. Welling, Variational graph auto-encoders, *arXiv preprint arXiv:1611.07308* (2016).
- [30] L. Serafini, A. d. Garcez, Logic tensor networks: Deep learning and logical reasoning from data and knowledge, *arXiv preprint arXiv:1606.04422* (2016).
- [31] R. Riegel, A. Gray, F. Luus, N. Khan, N. Makondo, I. Y. Akhalwaya, H. Qian, R. Fagin, F. Barahona, U. Sharma, et al., Logical neural networks, *arXiv preprint arXiv:2006.13155* (2020).