The Impact of Activation Patterns in the Explainability of Large Language Models – A Survey of recent advances

Mateus R. Figênio[1], André Santanché[2], Luiz Gomes-Jr[1]

[1]Departamento Acadêmico de Informática (DAINF)
Universidade Tecnológica Federal do Paraná (UTFPR)
80.230-901 – Curitiba – PR – Brazil

[2]Departamento de Sistemas de Informação - DSI
Universidade Estadual de Campinas (UNICAMP)
13083-970 – Campinas – SP – Brasil

mateusfigenio@alunos.utfpr.edu.br, santanch@unicamp.br, lcjunior@utfpr.edu.br

Abstract. *The performance benchmarks of Natural Language Processing (NLP) tasks have been overwhelmed by Large Language Models (LLMs), with their capabilities outshining many previous approaches to language modeling. But, despite the success in these tasks and the more ample and pervasive use of these models in many daily and specialized fields of application, little is known of how or why they reach the outputs they do. This study reviews the development of Language Models (LMs), the advances in their explainability approaches, and focuses on assessing methods to interpret and explain the neural network portion of LMs (specially of Transformer models) as means of better understanding them.*

1. Introduction

The field of Natural Language Processing (NLP), an area of research intended of enabling computers to process large natural language datasets, has been revolutionized since the introduction of deep learning (DL) models. Now, Large Language Models (LLMs) have demonstrated stellar performance on hard NLP tasks, such as text summarization, machine translation, question answering and dialog. Companies have launched many products and integrations based on new text generation models, such as OpenAI's ChatGPT or Googles's Gemini, from note-taking apps with broader self-completing capabilities, to more interactive customer service applications and integration with search engines that summarize the results of a search for the user.

However, our understanding of the inner workings of Language Models (LMs) based on neural network (NN) approaches lags behind the advancements in their size, complexity, architecture and broader use by society. To know how such models are designed and how they operate is different from understanding how its resulting properties, such as connections and weights, lead the system to a given prediction [Lillicrap and Kording 2019]. This lack of interpretability and explainability erodes trust and disavows the application of these models in contexts where the reasoning behind a decision is critical to its implementation, like those of the medical field. Meaningful explanations can also benefit the development of deep learning systems by aiding to verify if a system works as intended, can help improve the system by better understanding its

flaws, may lead to insights to specialists in its field of application and guarantee that the system complies to legislation, as argued by [Samek et al. 2017].

As such, the main goal of this study is to assess the state of explainability tools for Large Language Models (LLMs), with a specific focus on tools targeted to neuron activation explainability. In Section 2 we contextualize what language models are and the history of their development, as well as existing explainability tools and frameworks. Then in Section 3 we take a deeper look at recent papers of neuron activation explainability, which are further discussed in Section 4 where we summarize our findings, and in 5 we conclude our study.

2. Foundations

To contextualize this study and to set common ground on therms and meanings, in this section we define what are Language Models, briefly recall the development of Neural Network approaches in the field, from simple networks to Transformers, and present the development of explainability tools in the broader context of Natural Language Processing (NLP).

2.1. Language Models

The term Language Model (LM) refers to any system trained for the single task of predicting a series of tokens, whether letters, words or sentences, sequentially or not, given a previous or adjacent context [Bender and Koller 2020].

The first approaches for LMs were based on statistical patterns extracted from large corpora. These approaches used n-grams as the unit for probability estimation, with large n (i.e. large sequence of words) yielding better models [Manning and Schutze 1999]. However, these models could only capture dependencies in the n-word window, with poor coherence for larger or followings sequences.

2.1.1. Neural Networks in NLP

With the introduction of NN models in the computing field, there was a surge of initial NN approaches for language modeling, but, due to its similar implementation to those of statistical approaches, it were equally limited in the scope of sequence length. This changed with the introduction of the Long Short-term Memory (LSTM) Model [Hochreiter and Schmidhuber 1997], which was a version of the Recurrent Neural Networks (RNN) that addressed its issues and made viable its use in real world applications. It allowed for the NN to have greater persistence of information over sequence data such as time series or, in the case of NLP, text. That is accomplished through the implementation of a feedback loop, in which the output of the activation function is refeed to the cell trough summation, allowing both the current and previous values to influence its computation. This, along with LSTM new features, allowed longer sequences to be fed into the network and for the model to take a broader context into consideration, favoring its application in the context of NLP.

Despite its advances, neural networks were still hindered in the context of NLP thanks to the serial nature of its training, where one input had to be processed before the next one could be computed, that made for its training to be costly and time-consuming.

2.1.2. Transformers

Transformers discard recurrence in favor of relaying solely on the attention mechanism, which functions as a mean of assigning how an input token relates to the tokens in its surrounding context. This allows the model to take the entire context of a token in account when computing it independently, and, so, has no need for the recurrence mechanism to remember previous context and is no longer limited to a sequential training.

The model was present by [Vaswani et al. 2023] for sequence to sequence machine translation. It's architected in an encoder and decoder structure, where the encoder takes the input sequence in the original language and processes it, then the decoder takes any previous translations to the target language that occurred before, or just a begging of sentence token, computes it, joins both of these processed inputs (the current sequence to translate with the previous translated portion) and outputs the predicted final sequence to the target language. The encoder and the decoder are composed of: input word embeddings, that convert the tokens to word vectors; positional encodings, that encode the relative position of the word in the phrase; multi-head self-attention layers, that relate the token to its context; and feed-foward neural network (FFN) layers, that compute over all the previous information together. The difference between the encoder and the decoder, is that the decoder is composed by an initial encoder portion and additional encoder-decoder attention layers that combine the processed input text by the separate encoder with the processed output translated text by its own initial encoder portion. The result of this computation is then run through a final feed-foward, linear and softmax layers.

In this process, every input token is computed separately, while still considering its whole context, which enables the parallelization and distributed training of the model. With this, the Transformer model circumvents the limitations of sequential training of previous implementations, making possible greater models trained with greater datasets. Another aspect of the Transformer, that goes hand in hand with its larger training capabilities, is its capability of being able to be fine-tuned for a specific context of application after its main batch of training. This allows a base model to be extensively trained on a large corpus and then easily fine-tuned to specific applications. Finally, its structure can be modified to achieve more specific goals, either by changing the number of attention and FFN layers or by ditching the encoder or the decoder entirely. For example, BERT is composed of stacked encoders, while GPT is composed of stacked decoders.

2.1.3. Large Language Models

The innovations brought by the Transformer model, alongside its performance on various NLP tasks, led to the trend of larger and larger models such as OpenAI's GPT-3, that reaches 175 billion parameters and 570GB of training data, which is orders of magnitude above previous models developed for NLP. This earned the coinage of the term Large Language Models (LLMs) and the emergence of an area of research dedicated to them.

One of the latest developments in the LLMs field is the InstructGPT model [Ouyang et al. 2022], pioneered by OpenAI, that serves as the base for the new assistant models. Its Reinforcement Learning from Human Feedback (RLHF) utilizes a surrogate model, trained on preferred sentences ranked by human subjects, to generate sentences

that are used to refine a conventional base model to fallow instructions in more human aligned ways and to generate a more human aligned text.

With this, research of LLMs has been categorized into two training paradigms: traditional fine-tuning and prompting [Zhao et al. 2024]. Explainability works focused on the first paradigm mainly deal with question such as how the model acquires foundational understating of language from its base training and how the fine-tuning process influences its ability to solve domain specific tasks. Whereas, explainability works dealing with the second paradigm aim to understand how base models (not further trained to align with human preferences) leverage its pre-trained knowledge to respond to prompts and how assistant models (that were trained to better align with human preferences) come to be able to interact with users in open-ended conversations. The differences between the two paradigms makes so that the methods through which they can be understood are different. Beyond their size and dimension, there is the added layer of difficulty of new types of training that brings back questions about what patterns and information do exactly these models capture from a general training of token prediction and what is learned from an approach targeted to a specific task and to generate human oriented text.

2.2. Explainability and LLMs

One of the first works that specifically surveyed for tools of explainability of deep learning models in NLP and that categorized them according to a framework of thought was that of [Zini and Awad 2022], who proposed its framework based on three fundamental questions of model explanation: how they are explained, what is explained and which models are explained.

The authors justify the necessity of explainability tools specific to models that operate on language processing tasks by contrasting the particular set of challenges posed by digitally processing human language to those posed by other applications, like signal, image or data processing. Many tools have been developed for neural models of image processing, for example, but these fail to provide meaningful explanations for the predictions of language models, mostly due to the inherent differences of processing images to processing language. A letter or word has a different relationship to a sentence than a pixel has to an image, while both are the object in training and operation of models.

2.2.1. Explainability in NLP

Tackling the framework proposed by the authors, the question of "how" dictates if the explanation is *post-hoc*, after the processing of a prediction, or if the model itself is interpretable by design. This is well exemplified by decision tress, where each branching path clearly defines a rule for decision-making so that its resulting outputs are inherently explainable. A neural model, however, is much less interpretable (black-box) due to its complexity and non-linear relation of the inputs to its outputs, although even neural models can be reconfigured and retrained to be more interpretable.

The question of "what" covers what element of the model is being explained branched in three categories: input level, like the analysis of neural model embeddings and how they represent words; processing level, that focuses on the inner representations of neural models be it attention, specific weights and connections and how information is

stored and used inside the NN; and output level, which aims to explain individual model outputs in respect to input features or the models inner workings.

Lastly, "which" model is explained refers to the explanation tool being model agnostic or model specific, since some tools are tailored to a model's specific architecture and where others are independent of any models' configuration. This is better represented by output explainability, where performance tests on established datasets enable comparison between different model architectures and training forms.

2.2.2. Explainability of LLMs Inputs

The first step of a LLMs processing is constituted by word embedding, a process by which a token of a word is represented as dense vector so that it can be computed numerically. Although they can effectively and efficiently encode semantic and syntactic information, these high dimension embeddings are hard to interpret, which is not only essential for the making the whole of the model interpretable, but it is also desirable to ensure that the model is fair and efficient in its vocabulary representation.

The approaches developed for explainability of embedding revolve around: sparsification of embedding spaces that aim to remove redundancies or unnecessary dimensions, rotation of embedding spaces to understand concept dimensions, integrating external lexicon and ontological knowledge to align the embeddings to human representations and, finally, evaluating embedding interpretability in standardized tests. As a example, [Raganato and Tiedemann 2018] present a general evaluation of BERT, mainly focusing on the role of its attention mechanism in word representation, loosely fitting this category for Transformer input explainability.

2.2.3. Explainability of LLMs Inner Representations

The main approaches to explain and interpret the inner representations of LMs are divided in visualization and analysis approaches, the first pertaining to any method trough which one can visualize how the model has computed any given input or how its insides are configured, and the second refers to mathematical or statistical ways to understand the inner workings of the models.

Transformer inner representations explainability has, since its beginning, been extensively based on visualizing its attention mechanisms [Zini and Awad 2022], due to this mechanism more interpretable nature than that of the neurons of the FFN layers. The approaches vary in scale, from visualizing how individual attention heads evaluate tokens [Vaswani et al. 2023], to visualize how attention interacts across different attentions heads [Strobelt et al. 2019], how attention values flow across the model [DeRose et al. 2020] and how individual attention heads relate to concepts given by the user [Hoover et al. 2019].

Now, turning to the network portion of Transformer models, there are probing and neuron activation explainability approaches. Probing is a technique based in the idea of training a shallow classifier over a model's parameters to understand what they captured, in a sense an indirect approach to understand the model's NN. Whereas, neuron activation

is a direct approach that intends on understanding and explaining the neurons themselves, individually and collectively, and how they can be modified and deconstructed to alter their behavior. This last approach is the focus of this study.

2.2.4. Explainability of LLMs Outputs

The works in this category of explainability aim on providing evidence to support a model's decision. They are divided into *post-hoc* interpretation, where a models inputs and features are perturbed to observe changes into its output, and inherently interpretable models, which provide confidence intervals to its predication or a reasoning for its conclusion, be it by the model on prediction or a reference to an external ground truth system like Knowledge Graphs (KG). Works of inherently interpretable models are scarce, with [Zini and Awad 2022] reasoning that this is due to the large computational costs involved in retraining LLMs to make them interpretable.

3. Understanding and Manipulating Activation Patterns in LLMs

The Transformer model, currently used in the most successful LLMs, use several mechanisms to capture linguistic and task-specific patterns. The focus of this survey is the activation of the hidden layers of the deep neural network inside any transformer model. The activation patterns can be used both to understand and to manipulate the models.

3.1. Neuron Relevance Ranking

A straight forward line of work aims to identify important neurons and relate individual neurons to linguistic properties. Such is the work of [Bau et al. 2018], that, following the intuition that different Neural Machine Translation (NMT) models developed to act on the same languages will share similar properties, developed an unsupervised method to discover the neurons that relate to these shared properties. With this, the authors were able to modify the activations of individual neurons to control the model resulting outputs in predicable ways.

Another approach in this line of work is to use supervised methods to find important neurons in relation to specific language properties. Proposing a supervised method of neuron ranking, [Dalvi et al. 2019] aims to evaluate what is the impact of specific neurons in various language tests and how distributed or focused information is in NMT models. Its supervised approach of Linguistic Correlation Analysis classifies neurons in regard to their relevance to an expected linguistic property in the model input. To evaluate if their ranking was meaningful, the authors used ablation to compare how a model configured only with top neurons compared in performance to a model configure only with the bottom neurons. Although the architectures of the aforementioned methods were not that of the Transformer, the authors claim that their findings can be extended to it and to different components of NMT, such as the encoder and decoder.

The encoding of linguistic information often goes beyond a single neuron, encompassing a subset of the available dimensions. [Hennigen et al. 2020] propose a method based on Gaussian probes that identifies the subset of neurons associated with several linguistic properties. The authors focus on encodings generated by BERT and fastText, therefore limiting the analysis to the last hidden states of the networks.

3.2. Neuron Information Retrieval

Alternatively, another line of work is based on a mechanistic interpretability of neural language models, investigating neurons and their connections in similar terms as those of circuits. This interpretation was initially proposed to explain vision models, which can be intuitively understood as being a system composed of simpler building blocks. This approach was extended to explain neural networks hidden representations, such as it can be fruitfully applied to the LM context.

In [Geva et al. 2021], it is demonstrated that the FFN layers of a Transformer model acts like key value pairs that store memories related to patterns in its training data that amass the probability in favor of a specific output vocabulary. These patterns have been found not only to be human interpretable, but shown that shallow layers capture shallow patters of text, while the upper layers capture more syntactic patterns. The authors analyzed how cells are related to specific memories, and how the aggregated memories of multiple cells combine to produce a distribution different from what each cell individually could provide. Following this work, [Geva et al. 2022] explored how the FFN layers update the representation of the output vocabulary space in distilled human interpreted concepts. By decomposing the updates of the FFN layers to the vocabulary space in value vectors and analyzing how these vectors relate to concept annotated vectors, they discovered that each update can be decomposed in human interpretable concepts and that they can be altered and interfered to achieve more desirable outcomes, like less toxic models.

Understanding the patterns of activation was a foundation step for [Meng et al. 2022], which were able to identify the main layers associated with factual recalling in the LLMs tested. The authors identified the importance of the activation in middle layers during the processing of subject tokens (in a sort of priming for subsequent fact retrieval). After identifying the key activation regions, the authors introduce changes in the model to make it generate different facts. This type of post-training intervention was shown to be effective and can represent a complementary approach for LLM tuning.

4. Discussion

In revising previous surveys, the focus of the development of explainability methods appears to have shifted. In their assessment, previous to the prompting boom of late 2022, [Zini and Awad 2022] noted that its referenced works pointed to explainability development being more dedicated towards understanding the inner workings of LMs than to explaining particular outputs. Subsequently, however, the survey of [Zhao et al. 2024] shows that explainability research following the surge of new prompting models has moved to more heavily develop works concerned with output analysis, including approaches that utilized LMs to help evaluate and explain other LMs properties and outputs.

The attention mechanism is the focus of most of previous and current research towards better understanding Transformer's inner representations, either due to its more inherently more interpretable and visualizable behavior, or to its spotlight in the Transformer model development. In the latter years, new work has been developed to understand and explain the role of neural processing in the Transformer, from approaches that aim to find neurons responsible for the models' language capabilities, to works searching for where specific information and factual knowledge is encoded.

This last line of work seems to be the most promising in making the models more interpretable, their decisions more explicable, and their use more safe. The direct analysis of a model's inner representations allows the study of what information influenced its output, how it did, and enables for the alteration of this information towards more desirable outcomes, like more accurate and less toxic models. In contrast, output analysis approaches seem limited with its incapacity to remedy the model's failings, only being able to identify it.

5. Conclusion

This study presents a comprehensive background of LM explainability, firstly contextualizing the development of LMs, then presenting an explainability framework of thought for tools of explainability geared towards these models, and concluding with a focused review of works concerned with LMs hidden representation explainability and interpretability. The methods reviewed in this study present a promising line of work for making black-box LM more transparent, interpretable, explainable and safe.

References

Bau, A., Belinkov, Y., Sajjad, H., Durrani, N., Dalvi, F., and Glass, J. (2018). Identifying and controlling important neurons in neural machine translation.

Bender, E. M. and Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Dalvi, F., Durrani, N., Sajjad, H., Belinkov, Y., Bau, A., and Glass, J. (2019). What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6309–6317.

DeRose, J. F., Wang, J., and Berger, M. (2020). Attention flows: Analyzing and comparing attention mechanisms in language models.

Geva, M., Caciularu, A., Wang, K. R., and Goldberg, Y. (2022). Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space.

Geva, M., Schuster, R., Berant, J., and Levy, O. (2021). Transformer feed-forward layers are key-value memories.

Hennigen, L. T., Williams, A., and Cotterell, R. (2020). Intrinsic probing through dimension selection. *arXiv preprint arXiv:2010.02812*.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9:1735–80.

Hoover, B., Strobelt, H., and Gehrmann, S. (2019). exbert: A visual analysis tool to explore learned representations in transformers models.

Lillicrap, T. P. and Kording, K. P. (2019). What does it mean to understand a neural network?

Manning, C. and Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. (2022). Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback. Technical report, OpenAI. Disponível em: `https://arxiv.org/abs/2203.02155`.

Raganato, A. and Tiedemann, J. (2018). An analysis of encoder representations in transformer-based machine translation. In Linzen, T., Chrupała, G., and Alishahi, A., editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297, Brussels, Belgium. Association for Computational Linguistics.

Samek, W., Wiegand, T., and Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models.

Strobelt, H., Gehrmann, S., Behrisch, M., Perer, A., Pfister, H., and Rush, A. M. (2019). Seq2seq-vis: A visual debugging tool for sequence-to-sequence models. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):353–363.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., and Du, M. (2024). Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.* Just Accepted.

Zini, J. E. and Awad, M. (2022). On the explainability of natural language processing deep models. *ACM Computing Surveys*, 55(5):1–31.