Interpretability in Activation Space Analysis of Transformers: A Focused Survey

Soniya Vijayakumar[1,*]

[1]German Research Center for Artificial Intelligence (DFKI), Saarland Informatics Campus, Saarland, Germany

Abstract

The field of natural language processing has reached breakthroughs with the advent of transformers. They have remained state-of-the-art since then, and there also has been much research in analyzing, interpreting, and evaluating the attention layers and the underlying embedding space. In addition to the self-attention layers, the feed-forward layers in the transformer are a prominent architectural component. From extensive research, we observe that its role is under-explored. We focus on the latent space, known as the *Activation Space*, that consists of the neuron activations from these feed-forward layers. In this survey paper, we review interpretability methods that examine the learnings that occurred in this activation space. Since there exists only limited research in this direction, we conduct a detailed examination of each work and point out potential future directions of research. We hope our work provides a step towards strengthening activation space analysis.

Keywords

explainability, interpretability, machine learning, activation space analysis, linguistic information, transformers, feed-forward layers

1. Introduction

Through thick and thin, there is evidence that transformers have established itself as the state-of-the-art in various Natural Language Processing (NLP) tasks since their conception and realization in 2017. BERT, the most well-known transformer language model [1], consists of two major architectural components: self-attention layers and feed-forward layers. Much work has been done in analyzing the functions of self-attention layers [2, 3, 4]. In our survey, we focus on interpretability of the feed-forward layers. Each layer in the encoder and decoder contains a fully connected position-wise feed-forward network. The feed-forward network contains two linear transformations with a rectified linear activation function. Even though existing works highlight the importance of such feed-forward layers in transformers [5, 6, 7], still, to date, the role of feed-forward layers remains under-explored [8]. Our review focuses on the research that uses interpretability methods to understand the learnings in these feed-forward layers. We define the latent space, that comprises of the activations extracted from these layers, as the *Activation Space*. Many methods already exist for aggregating these representations including the default Huggingface[1] pipeline used in the original BERT paper [9].

Several methods for explaining and interpreting deep neural networks have been devised and we observe that much of the focus is in the domain of image processing [10]. A challenge that exists is the gap between the low-level features that the neural networks compute and the high-level concepts that are human-understandable. Furthermore, we observe that there have been relatively fewer research methods applied in understanding the internal learnings of networks in comparison to analyzing the functions of self-attention layers.

The core focus of our review is directed towards those methods that unfold the learnings in the internal representations of the neural network, i.e, we look at those methods that answer the question: "What does the model learn?" We further refine our focus on understanding specifically the feed-forward layers in transformer models. The motivation for this study is two-fold:

- The inputs undergo a non-linear transformation when passing through the activation functions in the feed-forward layers of deep neural networks [11].
- The parameters in the position-wise feed-forward layers of the transformer account for two-thirds of the total model's parameters ($8d^2$ per layer, d is the model's hidden dimension). This also implies that there is a considerable amount of computational budget involved in training these parameters to achieve the state-of-the-art performance they deliver today [12].

From recent research, the methods that focus on understanding the feed-forward layers show substantial evidence that the feed-forward layer activation space embeds useful information (see Section 5). We find that the learnings in the feed-forward layer remain under-explored. With our methodological survey, our objective

Proceedings of the CIKM 2022 Workshops, October 17 - 22, 2022

*Corresponding author.

soniya.vijayakumar@dfki.de (S. Vijayakumar)

[1]https://huggingface.co/

Table 1
Major attributes of the methods explored in the activation space analysis methods

Method	Properties	NLP Tasks	Quantitative Evaluation	Qualitative Evaluation
Linguistic Phenomena [13, 14, 15, 16]	Word Morphology, Lexical Semantics, Sentence Length, Parts-of-Speech	Parts-of-Speech, Semantic and Syntax Tagging and Prediction, Syntactic Chunking	Sensitivity, Prediction Accuracy, Selectivity Score	Human-expert visual inspection of selected neurons
Neural Memory Cells [12, 8]	Vocabulary Distribution, Human-Interpretable Patterns, Factual Knowledge	Next Sequence Prediction, Fill-in-the-blank Cloze Task	Agreement Rate, Prediction Probability, Attribution Score, Perplexity, Change and Success Rate	Pattern search by human experts
Knowledge Illusion [17]	Lexical, Geometric Properties (Local Semantic Coherence)	Next Sequence Prediction	Projection Score, Activation Quantile, Word Frequency Correlation	Human annotations for patterns using visualization

is to understand the internal mechanisms of transformers by exploring the activation space of the feed-forward network. Further, we consider this paper as a focused starting point for facilitating future research in activation space analysis. Finally, we also conduct a comparative study of these methods, their evaluation techniques and report our observations, understandings, and potential future directions (see Section 7). Table 1 summarizes the methods and its attributes that we have explored.

2. Related Surveys

As the interest in the Explainable Artificial intelligence (XAI) field grows, various survey articles were published, trying to consolidate and categorize the approaches. We segregate the reviews into two categories: Surveys that give a general overview of existing explainability methods [18, 19, 20, 21, 22] and surveys that focus on explainability methods in the NLP domain. We narrow our surveys to the NLP domain as this is the core focus of this review paper.

A survey that acts as a prior to ours is from Belinkov and Glass [23], where the authors review the various analysis methods used to conduct novel and fine-grained neural network interpretation and evaluation. The primary question that has been relevant while formulating these interpretation methods is: What linguistic information is captured in neural networks? The authors emphasize three aspects of the language-specific analysis, namely, methods used for conducting the analysis, linguistic information sought, and neural network parts investigated. They also identify several gaps and limitations in the surveys.

Danilevsky et al. [24] presents a broader overview of the state of XAI over a span of 7 years (until 2020), with a focus on the NLP domain. This work focuses on outcome explanation problems which help end users understand the model's operation and thereby build trust in these NLP-based AI systems. Along with the high-level classification of explanations, the work introduces two additional aspects: techniques that derive the explanation and techniques to present to the end user. The explainability techniques are categorized into feature importance, surrogate models, example-driven, provenance-based and declarative induction. A set of operations such as first-derivative salience, layer-wise relevance propagation, input perturbations, attention mechanism, and Long-Short-Term-Memory (LSTM) gating signal and explainability-aware architectures enable explainability. An interesting observation is the consideration of adding attention layers to neural network architectures as a strategy to enable explanations.

The closest survey related to our work is from Sajjad et al. [25], where the survey is on fine-grained neuron analysis. While there have been two previous surveys that cover Concept Analysis [26] and Attribution Analysis [24], their focus is on analyzing individual neurons to better understand the inner workings of neural networks. They refer to this as Neuron Analysis and categorized these reviewed methods into visualization, corpus-based, neuron probing, and unsupervised methods. The work further discusses findings and applications of neuron interpretation and summarizes open issues.

We observe that, from the various existing surveys, there are different dimensions to be considered. We narrow down our survey into the following dimensions:

- *Analysis methods* that focus on the internal interpretation of the activation space.
- *Linguistic Information* such as parts-of-speech, syntactic, semantic and *Non-linguistics Informa-*

tion such as sentence length, factual knowledge, geometric properties.
- *Neural network object* neurons and its activations as the Activation Space in the transformer language model.

We believe that interpretability alone is not sufficient in understanding the inner workings of the transformers, we also need explainability to summarize the reason for the model's behaviour in a human-comprehensible manner. One has to keep in mind that, explainability and interpretability have distinguishable meanings [27] and our review focuses only on interpretability methods because the research works reviewed focus on the same.

3. Survey Methodology

Our survey aim to cover the advances in NLP XAI research focusing on neuron interpretation. As defined earlier, we define this latent dimension as *Activation Space* and consider the reviewed techniques as *Activation Space Analysis* methods. We filtered to those methods that work at the feed-forward neuron-level, individual vs global, within the transformer model. We identified relevant papers published in NLP and AI conferences (AAAI, ACL, IJCNLP, EMNLP) between 2018 and 2022. With the limited scope of neuron-level analysis, we arrived at seven contemporary papers. With a limited number of work in this direction, we decided to take a deeper look into each of these methods, analyze its benefits, limitations, and gaps and present this study as our review paper. We are aware that this is an ongoing and relatively new research field and our focus is extremely limited; we acknowledge that we might have omitted certain papers. We also assume that if the authors have focused on explainability, they are more likely to cover the relevant related taxonomies, categories, and methods. Another common observation is that explanations are generated in an NLP task-oriented setting and remain relevant to the task context. Even though we summarize the tasks on which these researches are based, the task definitions are not relevant in our review process of understanding the activation space.

4. Taxonomies and Categorization

There still exists a reasonably vague understanding and lack of concrete mathematical definition between the two commonly used terms: explainability and interpretability. Interpretability has been defined as "the degree to which a human can understand the cause of a decision" [28] or the degree to which a human can consistently predict the model's result [29]. A broader definition exists for the term *interpretable machine learning* as the extraction of relevant knowledge from a machine-learning model concerning relationships either contained in the data or learned by the model. This definition rather focuses on understanding what the model learns either from an input-output mapping perspective or what the model itself learns. On the other hand, explainability directs the focus back to human understanding by examining the relationship between input features and model predictions in a human-understandable format [21].

After reviewing numerous relevant existing literature, we observed that explainability techniques broadly fall into three major classes. The first differentiates between understanding a model's individual prediction process versus prediction process as a whole [24]. A second differentiation is made in self-explaining or post-hoc methods, where the former generates explanations along with the model's prediction process whereas the latter requires post-processing of elements extracted during the model prediction process. The third major distinction corresponds to methods that are model specific or agnostic in nature. We also observed the existence of various other categorizations like outcome-based explanations, visual explanation methods, operations, and conceptual vs attribution. Visualization methods play a salient role in further understanding any interpretation method [30, 31, 32, 33]. These methods are inherent to interpretability and is been widely reviewed, we leave this to the reader to explore the relevant literature.

5. Activation Space Analysis Methods

There are two types of interpretability analysis that are carried out in the related research work: 1) Analyze individual neurons and 2) Analyze the entire set of neurons of the feed-forward layer. We look into both approaches from four perspectives: categorization, linguistic knowledge sought for, methodology, and evaluations, and conduct a comparative analysis of these methods.

Linguistic Phenomena: Investigating the linguistic phenomena that occurs within the activations of pretrained models, when trained for a specific task set, using various interpretability analysis methods, is a common way to interpret the features learned by these models. The linguistic phenomenon refers to the presence of various linguistic features such as word morphology, lexical semantics, syntax or linguistic knowledge such as parts-of-speech, grammar, coreference, lemmas. Linguistic Correlation Analysis (LCA) is one such method that focuses on understanding what the model learned about linguistic features and determining those neurons that explicitly focus on such phenomena. A toolkit with three major methods, Individual Model Analysis, Cross-model Analysis and LCA, to identify salient neurons within

the model or related to a task under consideration, is presented by Dalvi et al. [13].

Probing using diagnostic classifiers to understand the knowledge captured in neural representations is another common method for associating model components with linguistic properties [34, 35, 36]. This involves extracting feature representations from the network and training an auxiliary classifier to predict the linguistic property. Layer-wise and neuron-level diagnostic classifiers that probe representation from individual layers w.r.t linguistic properties and find neurons that capture salient features, respectively, are used to conduct analysis on pre-trained models BERT, RoBERTa and XLNet [14]. The task of predicting a certain linguistic property is defined. A diagnostic classifier (logistic regression) is trained on generated activations, for both layer-wise and neuron-wise probes, to predict the existence of this linguistic property. An LCA is conducted to generate neuron ranking based on weight distribution. Additionally, an elastic-net regularization is fine-tuned using grid-search to balance between focused and distributed neurons. The top N salient neurons extracted from this ranked list are used to retrain the classifier until an Oracle accuracy is achieved.

Durrani et al. [15] and Alammar [16] conducts similar experiments, where the entire neuron activations from the feed-forward layers are used to train an external classifier. Durrani et al. [15] uses a probing classifier (logistic regression) with the additional elastic-net regularization to conduct a fine-grained neuron level analysis on pre-trained models ELMo, T-ELMo, BERT, and XLNET. This variance of models, in this study, covers different modeling choices of the blocks, optimization objectives, and model architectures. The case study conducted by Alammar [16] uses probing the feed-forward neuron activations for Parts-of-Speech (POS) Information. A control task is created where each token is assigned to a random POS tag and a separate probe is trained on this control set. This allows us to measure the difference in prediction accuracy between the actual and control dataset, selectivity score, thereby concluding if the probe really extracts the POS information. The author collects existing methods that examines input saliency, hidden state evolution, neuron activations, and non-negative matrix factorization of neuron activations, along with dimensionality reduction methods to extract patterns into an open-source library known as Ecco [16]. These methods can be directly employed on pre-trained models such as GPT2, BERT, RoBERTa.

Neural Memory Cells: In the context of a neural network with a recurrent attention model, Sukhbaatar et al. [37] introduced input and output memory representations. A recent work extends this neural memory concept and shows that the feed-forward layers in the transformer models operate as key-value memories, where keys correlate to specific human-interpretable input pattern sets and simultaneously, values induce a distribution over the output vocabulary [12]. The work analyzes these memories present in the feed-forward layers and further explores the function of these layers in transformer-based language models.

A neural memory is defined as a key-value pair, where each key value is a d-dimensional vector. The emulation, mathematical similarity between feed-forward and key-value neural memories, allows the hidden dimension to be considered as number of memories in each layer and the activations as vectors containing un-normalized non-negative memory coefficients. Using this similarity, the study posits that the key vectors act as pattern detectors. This hypothesis is tested by looking for the highest memory coefficient that is associated with the input text, retrieving those input examples, and conducting human evaluations to identify patterns. The study further explores intra-layer memory composition and inter-layer prediction refinement.

The concept of knowledge neurons, neurons that express a fact, is introduced by Dai et al. [8]. The authors propose a method to find the neurons that express facts and how their activations correlate in expressing these facts. The evaluations on pre-trained models for fill-in-the-blank cloze tasks show that these models have the ability to recall factual knowledge even without fine-tuning. The work considers feed-forward layers as key-value memories, hypothesize that these key-value memories store factual knowledge and proposes a knowledge attribution method. The knowledge attribution method, based on integrated gradients, evaluates the contribution of each neuron, in BERT-base-cased transformer, to knowledge predictions by assigning them an attribution score. Those neurons with a higher gradient i.e attribution score are identified as those contributing to factual expressions. Further refinement of these neurons is done under the hypothesis that there are chances that the same fact can share the same set of true positive knowledge neurons. This refinement allows in retaining only those knowledge neurons that are shared by a certain percentage of input prompts.

Knowledge Illusion: Based on the generalization of the hypothesis that concepts are encoded in the linear combinations of neural activations, Bolukbasi et al. [17] describe a surprising phenomenon "interpretability illusion". Probing experiments conducted on BERT-base-uncased model determines if individual neurons contained human-interpretable meaning. The final layer creates embeddings for four datasets (QQP, QNLI, Wiki, and Books) and top 10 activating sentences for a neuron are annotated to determine a pattern. Here a pattern is defined as a single property such as sentence length or lexical similarity shared by a set of sentences. By proposing three sources: dataset idiosyncrasy, local semantic coherence in BERT's embedding space, and annotator

error, the authors explain this illusion. The same experiment is repeated, by keeping a set of target neurons constant, on various datasets to reveal the illusion as described by the authors. The work further explores the causes of this illusion by investigating local, global and dataset-level concepts.

6. Evaluations

Linguistic Phenomena: A layer-wise probing is conducted to understand the redistribution of linguistic knowledge (syntactic chunking, POS, and semantic tagging) when fine-tuned for downstream tasks [14]. Using this probing across three fine-tuned models BERT, RoBERTa, and XLnet, on GLUE tasks and architectures reveal the following observations: The morpho-syntactic linguistic phenomenon that is preserved, post fine-tuning, in the higher layers is dependent on the task; Different architectures preserve linguistic information differently post fine-tuning. The neuron-wise probing further refines to the fine-grained neuron level, where the most salient neurons are extracted and their distribution across architecture and variations in downstream tasks are studied. An alignment of findings is found with Merchant et al. [38], where the fine-tuning affects only the top layer. In comparison with Mosbach et al. [39], which is focused on sentence level probing, Durrani et al. [14] studies core-linguistic phenomena. Additionally, their findings from fine-grained neuron analysis extend the core-linguistic task layer-wise analysis, along with fine-tuning effects on these neurons. Another interesting observation made is the different patterns that are entailed when these networks are pruned from top or bottom.

An ablation study conducted by Durrani et al. [15] on the top salient neurons, from four pre-trained models ELMo, T-ELMo, BERT, and XLNet, indicates higher distribution of linguistic information across the network when the underlying task is more complex (CCG supertagging), revealing information redundancy. Further refined study, considering only a minimal set of neurons, to identify the network parts that predominantly capture the linguistic information and understand the localization or distribution of this information, indicate that the number of neurons required to achieve the Oracle accuracy varies and is dependent on the complexity of the task. By employing a selectivity score next to the prediction accuracy score, and training separate POS probes for the actual dataset and a control task, Alammar [16] observes that the activation space encodes POS information at levels comparable to BERT's hidden states. The non-negative matrix factorization method helps in identifying those patterns in neuron activations that correspond to syntactic and semantic properties of the input text. The NeuroX toolkit is compared with the What-if tool from Google, that inspects trained models based on prediction and Seq2Seq-Vis [40], that can trace back prediction decisions in Neural Machine Translation input models [13].

Neural Memory Cells: Relating the patterns identified by human experts (NLP graduate students) to human understanding, the patterns are classified as shallow or semantic and are associated with lower layers and upper layers of a 16-layer transformer model, respectively [8]. Further analysis of the corresponding values from the key-value memories complements the patterns observed in the respective keys. The agreement rate, the fraction of memory cells that match the corresponding keys and values, is seen to increase in higher layers. The authors suggest that the memory cells in the higher layers contribute to the output whereas the lower layers do not show such a clear key-value correlation to contribute toward the output distribution of the next word. A qualitative analysis, by manually analyzing a few random cases, is conducted on the layer-wise distribution of memory cells and how the model refines its prediction from layer to layer using residual connections. The work is an extension of Sukhbaatar et al. [37], which suggests a theoretical similarity between feed-forward layers and key-value memories. Additionally their observations, of shallow feature encoding, confirms with recent findings from Peters et al. [41], Jawahar et al. [42], Liu et al. [43].

The BERT-base-cased model is experimented with the knowledge attribution, where activation value is considered as the attribution score for a neuron, to measure neuron sensitivity towards input. Similar observations to Geva et al. [12] and Tenney et al. [44] are identified: fact-related neurons are distributed in the higher layers of the transformer. Further, the authors investigate how these neurons contribute to expressing the knowledge either by suppressing or amplifying their activations. Two additional use cases, updating facts and erasing relations, are presented, where the authors demonstrate the potential application of these identified knowledge neurons. Two evaluation metrics are used: change and success rate for measuring fact updating and inter/intra-relation perplexity for measuring the influence on other knowledge. These evaluations indicate that changes in very few neurons in the transformers can affect certain facts. Erasing of facts is also measured using perplexity and is observed that post fact erasing operation, i.e. setting knowledge neuron to zero vectors, the perplexity of the moved knowledge increased. The knowledge attribution method, built on integrated gradients, is inspired by Hao et al. [45] and Sundararajan et al. [46].

Knowledge Illusion: A qualitative evaluation is conducted by annotating three sets of sentences for a neuron in consideration: 1) top ten activating sentences for the neuron, 2) top ten activating sentences in random direction and 3) ten random sentences [17]. The objective

of this annotation is to find patterns, where a pattern is defined as a property shared by a set of sentences. A pattern is considered as a proxy for a learned concept by the model. For each neuron under consideration, an average of 2.5 distinct patterns across four datasets are observed. This illusion is further explored by studying the regions of activation space the input data occupies, the influence of top activating sentences on patterns from both local semantic coherence and global directions, and annotation error. Qualitative analysis is conducted through (UMAP dimensionality reduction) visualization and it is observed that sentences cluster in accordance with datasets. Additionally, the high accuracy of a Support Vector Machine classifier distinguishes between these datasets and provides quantitative evidence for this observation. This indicates the dependence of information encoded within neurons on the idiosyncrasies of the natural language datasets, even though they have similar activation values. The analysis of global directions in BERT's activation space using activation quantiles helps in understanding the correlation between word frequency change and its monotonicity in each combination of datasets. This correlation indicated that despite BERT's illusionary effect, there still exists meaningful global direction in its activation space. While comparing the observed illusions with previous works, it is in alignment with Aharoni and Goldberg [47], where they demonstrate the usage of BERT representations to disambiguate datasets. This explains the existence of patterns in datasets, further experiments are conducted to understand the cause of such pattern existence.

We observe that all the methods that we reviewed so far fall under the local interpretability methods and limit themselves to the top N salient neurons (see Table 1). From reviewing these studies, we observe dimensionality reduction is required to understand the properties under consideration. Dimensionality reduction is associated with information loss and this loss is not accounted for in these studies. Another observation is that the focus of these studies alternates between identifying the neurons that capture the relevant linguistic information and those subsets of these neurons that affect the prediction accuracy. Moreover, some interpretability methods are evaluated through user studies (where users subjectively evaluate the explanations), whereas others are evaluated in terms of how they satisfy some properties, either quantitatively or qualitatively, without real users' evaluations. In the next section, we further discuss our observations and present our insights and future detections.

7. Insights and Future Directions

A common observation that we see in the contemporary general surveys and from our focused reviews is the lack of both theoretical foundations and empirical considerations in evaluations [25, 23, 24]. Even though each method has quantitative measures for evaluation, there is no standard set of metrics for comparing various observations, hence, confining the scope of respective interpretability technique results to specific model architectures or task-related domains. Studies have proposed various desiderata for interpretable concepts such as Fidelity, Diversity and Grounding [48] for qualitative consistency Additionally, a few studies employ human experts for qualitative analysis such as pattern annotation and identifications, but again lack a standard framework for a comparative study and consistent explanations. Moreover, the subjective nature of interpretability and the lack of existence of ground truth in qualitative analysis makes it even more challenging to evaluate these methods.

By reviewing the above works, that focus on activation space, we observe the following from the model perspective: For a fixed model architecture and when a fixed set of neurons are examined, each set of neurons encode different information, dependent on the input dataset; On the contrary, when a wider set of model architectures are considered, the same set of neurons encode similar information at lower and higher layers across these architectures but the information encoded is dependent on the underlying task. These observations emphasize the dependency on the input data and the underlying task of interpreting the linguistic information encoded in the activation space.

Experiments conducted align with the definition of interpretability and explainability in understanding the rationale behind the model's decision but lack human understandable explanations. In the context of explainability, we observe that there is a gap in human-understandable linguistic concepts and linguistic features captured in the network. We make a clear distinction between linguistic features and concepts: features consist of linguistic properties such as parts-of-speech, syntactic and semantic properties, and word morphology whereas the linguistic concepts, from a human understandable perspective, encode general human knowledge and how it is expressed in natural language. Various contemporary methods such as Concept Relevant Propagation [49], Testing Concept Activation Vector [50], Integrated Conceptual Sensitivity [51] that are based on human understandable local and global concept-based explanations exist. These methods are applied and evaluated in the image processing domain and are yet to be explored in understanding linguistic concepts. It is evident that exploring activation space is a promising research direction and we propose a potential future direction: extend the interpretability techniques from image processing to the natural language processing domain through transfer learning.

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful feedback. The work was partially funded by the German Federal Ministry of Education and Research (BMBF) through the project XAINES (01IW20005).

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, CoRR abs/1706.03762 (2017). URL: http://arxiv.org/abs/1706.03762. arXiv:1706.03762.

[2] M. Nasr, R. Shokri, A. Houmansadr, Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning, in: 2019 IEEE Symposium on Security and Privacy (SP), 2019, pp. 739–753. doi:10.1109/SP.2019.00065.

[3] K. Clark, U. Khandelwal, O. Levy, C. D. Manning, What does BERT look at? an analysis of bert's attention, CoRR abs/1906.04341 (2019). URL: http://arxiv.org/abs/1906.04341. arXiv:1906.04341.

[4] J. Vig, Y. Belinkov, Analyzing the structure of attention in a transformer language model, in: Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Florence, Italy, 2019, pp. 63–76. URL: https://aclanthology.org/W19-4808. doi:10.18653/v1/W19-4808.

[5] O. Press, N. A. Smith, O. Levy, Improving transformer models by reordering their sublayers, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 2996–3005. URL: https://aclanthology.org/2020.acl-main.270. doi:10.18653/v1/2020.acl-main.270.

[6] B. Pulugundla, Y. Gao, B. King, G. Keskin, H. Mallidi, M. Wu, J. Droppo, R. Maas, Attention-based neural beamforming layers for multi-channel speech recognition, 2021. URL: https://arxiv.org/abs/2105.05920. doi:10.48550/ARXIV.2105.05920.

[7] H. Xu, Q. Liu, D. Xiong, J. van Genabith, Transformer with depth-wise LSTM, CoRR abs/2007.06257 (2020). URL: https://arxiv.org/abs/2007.06257. arXiv:2007.06257.

[8] D. Dai, L. Dong, Y. Hao, Z. Sui, F. Wei, Knowledge neurons in pretrained transformers, CoRR abs/2104.08696 (2021). URL: https://arxiv.org/abs/2104.08696. arXiv:2104.08696.

[9] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[10] A. Das, P. Rad, Opportunities and challenges in explainable artificial intelligence (XAI): A survey, CoRR abs/2006.11371 (2020). URL: https://arxiv.org/abs/2006.11371. arXiv:2006.11371.

[11] S. Zhao, D. Pascual, G. Brunner, R. Wattenhofer, Of non-linearity and commutativity in BERT, CoRR abs/2101.04547 (2021). URL: https://arxiv.org/abs/2101.04547. arXiv:2101.04547.

[12] M. Geva, R. Schuster, J. Berant, O. Levy, Transformer feed-forward layers are key-value memories, CoRR abs/2012.14913 (2020). URL: https://arxiv.org/abs/2012.14913. arXiv:2012.14913.

[13] F. Dalvi, A. Nortonsmith, D. A. Bau, Y. Belinkov, H. Sajjad, N. Durrani, J. Glass, Neurox: A toolkit for analyzing individual neurons in neural networks, Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2019).

[14] N. Durrani, H. Sajjad, F. Dalvi, How transfer learning impacts linguistic knowledge in deep NLP models?, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 4947–4957. URL: https://aclanthology.org/2021.findings-acl.438. doi:10.18653/v1/2021.findings-acl.438.

[15] N. Durrani, H. Sajjad, F. Dalvi, Y. Belinkov, Analyzing individual neurons in pre-trained language models, CoRR abs/2010.02695 (2020). URL: https://arxiv.org/abs/2010.02695. arXiv:2010.02695.

[16] J. Alammar, Ecco: An open source library for the explainability of transformer language models, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2021, pp. 249–257. URL: https://aclanthology.org/2021.acl-demo.30. doi:10.18653/v1/2021.acl-demo.30.

[17] T. Bolukbasi, A. Pearce, A. Yuan, A. Coenen, E. Reif, F. B. Viégas, M. Wattenberg, An interpretability illusion for BERT, CoRR abs/2104.07143 (2021). URL: https://arxiv.org/abs/2104.07143. arXiv:2104.07143.

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

[19] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51

(2018). URL: https://doi.org/10.1145/3236009. doi:10.1145/3236009.

[20] V. Arya, R. K. E. Bellamy, P. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilovic, S. Mourad, P. Pedemonte, R. Raghavendra, J. T. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, Y. Zhang, One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques, CoRR abs/1909.03012 (2019). URL: http://arxiv.org/abs/1909.03012. arXiv:1909.03012.

[21] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis, Explainable ai: A review of machine learning interpretability methods, Entropy 23 (2021). URL: https://www.mdpi.com/1099-4300/23/1/18. doi:10.3390/e23010018.

[22] A. Krajna, M. Kovac, M. Brcic, A. Šarčević, Explainable artificial intelligence: An updated perspective, in: 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO), 2022, pp. 859–864. doi:10.23919/MIPRO55190.2022.9803681.

[23] Y. Belinkov, J. Glass, Analysis Methods in Neural Language Processing: A Survey, Transactions of the Association for Computational Linguistics 7 (2019) 49–72. URL: https://doi.org/10.1162/tacl_a_00254. doi:10.1162/tacl_a_00254.

[24] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, P. Sen, A survey of the state of explainable AI for natural language processing, in: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2020, pp. 447–459. URL: https://aclanthology.org/2020.aacl-main.46.

[25] H. Sajjad, N. Durrani, F. Dalvi, Neuron-level interpretation of deep NLP models: A survey, CoRR abs/2108.13138 (2021). URL: https://arxiv.org/abs/2108.13138. arXiv:2108.13138.

[26] Y. Belinkov, N. Durrani, F. Dalvi, H. Sajjad, J. R. Glass, On the linguistic representational power of neural machine translation models, CoRR abs/1911.00317 (2019). URL: http://arxiv.org/abs/1911.00317. arXiv:1911.00317.

[27] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. A. Specter, L. Kagal, Explaining explanations: An approach to evaluating interpretability of machine learning, CoRR abs/1806.00069 (2018). URL: http://arxiv.org/abs/1806.00069. arXiv:1806.00069.

[28] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, CoRR abs/1706.07269 (2017). URL: http://arxiv.org/abs/1706.07269. arXiv:1706.07269.

[29] B. Kim, R. Khanna, O. O. Koyejo, Examples are not enough, learn to criticize! criticism for interpretability, in: D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 29, Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper/2016/file/5680522b8e2bb01943234bce7bf84534-Paper.pdf.

[30] P. Pezeshkpour, Y. Tian, S. Singh, Investigating robustness and interpretability of link prediction via adversarial modifications, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 3336–3347. URL: https://aclanthology.org/N19-1337. doi:10.18653/v1/N19-1337.

[31] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, J. Eisenstein, Explainable prediction of medical codes from clinical text, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1101–1111. URL: https://aclanthology.org/N18-1100. doi:10.18653/v1/N18-1100.

[32] D. Croce, D. Rossini, R. Basili, Auditing deep learning processes through kernel-based explanatory models, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 4037–4046. URL: https://aclanthology.org/D19-1415. doi:10.18653/v1/D19-1415.

[33] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, 2014. URL: https://arxiv.org/abs/1409.0473. doi:10.48550/ARXIV.1409.0473.

[34] D. Hupkes, S. Veldhoen, W. Zuidema, Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure (2017). URL: https://arxiv.org/abs/1711.10203. doi:10.48550/ARXIV.1711.10203.

[35] A. Conneau, G. Kruszewski, G. Lample, L. Barrault, M. Baroni, What you can cram into a single vector: Probing sentence embeddings for linguistic properties, CoRR abs/1805.01070 (2018). URL: http://arxiv.org/abs/1805.01070. arXiv:1805.01070.

[36] Y. Belinkov, J. R. Glass, Analysis methods in neural language processing: A survey, CoRR abs/1812.08951 (2018). URL: http://arxiv.org/abs/

1812.08951. arXiv:1812.08951.

[37] S. Sukhbaatar, E. Grave, G. Lample, H. Jégou, A. Joulin, Augmenting self-attention with persistent memory, CoRR abs/1907.01470 (2019). URL: http://arxiv.org/abs/1907.01470. arXiv:1907.01470.

[38] A. Merchant, E. Rahimtoroghi, E. Pavlick, I. Tenney, What happens to BERT embeddings during fine-tuning?, in: Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Online, 2020, pp. 33–44. URL: https://aclanthology.org/2020.blackboxnlp-1.4. doi:10.18653/v1/2020.blackboxnlp-1.4.

[39] M. Mosbach, A. Khokhlova, M. A. Hedderich, D. Klakow, On the Interplay Between Fine-tuning and Sentence-level Probing for Linguistic Knowledge in Pre-trained Transformers, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 2502–2516. URL: https://aclanthology.org/2020.findings-emnlp.227. doi:10.18653/v1/2020.findings-emnlp.227.

[40] H. Strobelt, S. Gehrmann, M. Behrisch, A. Perer, H. Pfister, A. M. Rush, Seq2seq-vis: A visual debugging tool for sequence-to-sequence models, CoRR abs/1804.09299 (2018). URL: http://arxiv.org/abs/1804.09299. arXiv:1804.09299.

[41] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 2227–2237. URL: https://aclanthology.org/N18-1202. doi:10.18653/v1/N18-1202.

[42] G. Jawahar, B. Sagot, D. Seddah, What does BERT learn about the structure of language?, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3651–3657. URL: https://aclanthology.org/P19-1356. doi:10.18653/v1/P19-1356.

[43] N. F. Liu, M. Gardner, Y. Belinkov, M. E. Peters, N. A. Smith, Linguistic knowledge and transferability of contextual representations, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 1073–1094. URL: https://aclanthology.org/N19-1112. doi:10.18653/v1/N19-1112.

[44] I. Tenney, D. Das, E. Pavlick, BERT rediscovers the classical NLP pipeline, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4593–4601. URL: https://aclanthology.org/P19-1452. doi:10.18653/v1/P19-1452.

[45] Y. Hao, L. Dong, F. Wei, K. Xu, Self-attention attribution: Interpreting information interactions inside transformer, 2020. URL: https://arxiv.org/abs/2004.11207. doi:10.48550/ARXIV.2004.11207.

[46] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: D. Precup, Y. W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 3319–3328. URL: https://proceedings.mlr.press/v70/sundararajan17a.html.

[47] R. Aharoni, Y. Goldberg, Unsupervised domain clusters in pretrained language models, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7747–7763. URL: https://aclanthology.org/2020.acl-main.692. doi:10.18653/v1/2020.acl-main.692.

[48] D. Alvarez Melis, T. Jaakkola, Towards robust interpretability with self-explaining neural networks, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 31, Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper/2018/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf.

[49] R. Achtibat, M. Dreyer, I. Eisenbraun, S. Bosse, T. Wiegand, W. Samek, S. Lapuschkin, From "where" to "what": Towards human-understandable explanations through concept relevance propagation, 2022. URL: https://arxiv.org/abs/2206.03208. doi:10.48550/ARXIV.2206.03208.

[50] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, R. Sayres, Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav) (2017). URL: https://arxiv.org/abs/1711.11279. doi:10.48550/ARXIV.1711.11279.

[51] J. Schrouff, S. Baur, S. Hou, D. Mincu, E. Loreaux, R. Blanes, J. Wexler, A. Karthikesalingam, B. Kim, Best of both worlds: local and global explanations with human-understandable concepts, CoRR abs/2106.08641 (2021). URL: https://arxiv.org/abs/2106.08641. arXiv:2106.08641.

A. Evaluation Metrics Definitions

- *Selectivity:* The difference between linguistic task accuracy and control task accuracy
- *Prediction Accuracy:* Performance measure of the model on a given task
- *Agreement Rate:* The fraction of memory cells (dimensions) where the value's top prediction matches the key's top trigger example
- *Value Probability:* Probability of the values' top prediction
- *Projection Score:* The dot product between a sentence embedding and a direction
- *Activation Quantile:* Equally sized smaller subsection of the activation space
- *Word Frequency Correlation:* The correlation between directions and words in the embedding space
- *Attribution Score:* Measures the contribution of the neuron to the factual expressions
- *Perplexity:* Measurement of how well a probability model predicts a sample, degree of 'uncertainty' a model has in predicting
- *Change Rate:* The ratio that the original prediction is modified to another
- *Success Rate:* The ratio that becomes learned prediction the top predictions