Interpretability in Activation Space Analysis of Transformers: A Focused Review

Soniya Vijayakumar
German Research Institute of Artificial Intelligence (DFKI), Saarland Informatics Campus,
Saarland, Germany

AIMLAI: Advances in Interpretable Machine Learning and Artificial Intelligence Workshop, CIKM 2022

Agenda

1. Introduction
2. Concepts
3. Observation and Summary

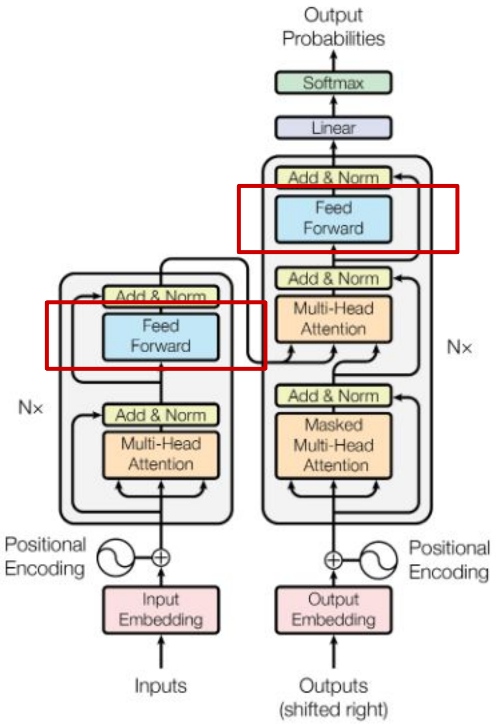
Activation Space

Transformer Models

- **Focus**: interpretability methods to understand the learnings in feed-forward layers.
- The latent space, that comprises of the activations extracted from these layers, as the ***Activation Space.***

Survey (2018 - 2022)

- Explainability methods in the NLP domain in the transformer architecture.
- Feed-forward neuron-level, individual vs global, within the transformer model.



The Transformer - model architecture. [1]

Concepts

Linguistic Phenomena

Neural Memory Cells

Knowledge Illusion

Concepts

Presence of various linguistic features such as word morphology, lexical semantics, syntax or linguistic knowledge such as parts-of-speech, grammar, coreference, lemmas.

Linguistic Phenomena

using diagnostic classifiers to understand the knowledge captured in neural representations is another common method for associating model components with linguistic properties

understand what model learned about linguistic features and determining those neurons that explicitly focus on such phenomena

Linguistic Correlation Analysis

Probing

- Fine-grained neuron level analysis using Logistic Regression probing classifier [15].
- Ablation study on ELMO-T-ELMo, BERT, XLNET.
- **Higher distribution of linguistic information** across the network when underlying **task** is more **complex** -> information redundancy.
- **Number of neurons** required to achieve the Oracle accuracy **varies** and is **dependent** on **task complexity.**

- Probe feed-forward neuron activations for POS information [16].
- GPT2, BERT, RoBERTa.
- **POS** information at **levels** comparable to **BERT's hidden states**.
- Non-negative matrix factorization method -> identify **patterns in neuron activations** that **correspond** to **syntactic** and **semantic properties** of the input text.

- Layer-wise and neuron-level diagnostic classifiers [14].
- BERT, RoBERTa, XLNET.
- Predict a certain linguistic property (GLUE tasks).
- The **morpho-syntactic linguistic phenomenon** that is **preserved**, post fine-tuning, in the **higher layers** is **task dependent.**
- **Different architectures** preserve linguistic information **differently post fine-tuning**.

Concepts

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

Neural Memory Cells

Key-Value Pairs

- **Mathematical similarity** between feed-forward and key-value memories [12].
- Key correlates with textual patterns in training examples.
- Each value induces a distribution over the output vocabulary.
- **Memories** are associated with **human-recognizable** patterns.
- **Shallow layers** detect **shallow patterns** & **upper layers** learn more **semantic patterns.**
- Intra-Layer Memory Composition & Inter-Layer Prediction Refinement.

Knowledge Neurons

neurons that express a fact

- **Neurons** that **express facts** and how their **activations correlate in expressing these facts** [8].
- Knowledge attribution method: based on integrated gradients, evaluates the contribution of each neuron.
- Fill-in-the-blank cloze tasks-> recall factual knowledge.
- **Two use cases: fact updation and erasing.**
- Indicate that **changes in very few neurons** in the transformers **can affect certain facts.**
- **Post fact erasing operation**, i.e. setting knowledge neuron to zero vectors, the **perplexity** of the moved knowledge **increased**.

Concepts

Bolukbasi et al. [17] describe a surprising phenomenon "interpretability illusion"

Knowledge Illusion

- BERT-base-uncased model.
- Determine if **individual neurons** contained **human-interpretable meaning**.
- **Identify Patterns**: a single property such as sentence length or lexical similarity shared by a set of sentences.
- **Illusion sources**: dataset idiosyncrasy, local semantic coherence in BERT's embedding space, and annotator error.

- Top-10 activating sentences for the neuron.
- Top-10 activating sentences in random direction.
- 10 random sentences.
- **illusion further explored** by studying
  - Regions of activation space the input data occupies.
  - Influence of top activating sentences on patterns from both local semantic coherence and global directions.
  - Annotation Error.
- Qualitative analysis is conducted through visualization -> **sentences cluster in accordance with datasets**.

○

Observations

- **Local interpretability methods** and **limit** themselves to the **top N salient neurons.**

- **Alternates** between identifying neurons that capture the relevant linguistic information and neuron subsets that affect the prediction accuracy.

- Some interpretability methods evaluated through **user studies** whereas others in terms of how they **satisfy some properties**, either quantitatively or qualitatively, without real users' evaluations.

- **Lack** of both **theoretical foundations and empirical considerations** in evaluations. -> **Confined scope**: specific model architectures or task-related domains.
  - **Fixed model architecture:** fixed set of neurons are examined, each set of neurons encode different information, dependent on the input dataset.
  - **Wider model architectures:** same neurons set encode similar information at lower and higher layers across architectures, dependent on the underlying task.

Summary

Emphasize the **dependency** on the **input data** and the **underlying task** of interpreting the linguistic information encoded in the activation space.

Gap in human-understandable linguistic concepts and linguistic features captured in the network. **Inclusion of Domain Expertise**

Extend the interpretability techniques from image processing to the natural language processing domain through transfer learning.

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:1706.0376

[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

[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

[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, As sociation 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

Thank You!

In Short

Methods	**Properties**	**NLP Tasks**	**Evaluation Metric**	**Human Evaluation: Are there human understandable concepts -> Q vs Q**
Linguistic Phenomena	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	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 in single memory cells and aggregated multiple cells in multiple layers
Knowledge Illusion	Lexical, Geometric Properties (Local Semantic Coherence)	Next Sequence Prediction	Projection Score, Activation Quantile and Word Frequency Correlation	Human annotations for patterns using visualization