ADVANCE SOCIAL SCIENCE ARCHIVE JOURNAL
Available Online: https://assajournal.com
Vol. 03 No. 01. Jan-Mar 2025.Page#.2234-2240
Print ISSN: 3006-2497 Online ISSN: 3006-2500
Platform & Workflow by: Open Journal Systems
https://doi.org/10.5281/zenodo.19784499

BRIDGING THE PERFORMANCE-INTERPRETABILITY GAP IN LARGE LANGUAGE MODELS: SPARSE AUTOENCODER DISENTANGLEMENT WITH PERFORMANCE ANCHORING

Mohsin Raza Shah
Assistant Professor, College Education Department, Government of Sindh
razamohsinsyed31@gmail.com

ABSTRACT

The high performance of modern models of transformer-based systems in processing natural languages has been achieved at the expense of transparency. Because these systems are being used in more sensitive fields - such as analysis of legal documents, medical decision support, etc. The inability to know why these systems have come to that point is a significant practical and ethical drawback. The classical post hoc explanation methods are often incapable of delivering faithful information and the efforts to construct intrinsic explanatory ability of the task are mostly detrimental to the task performance. This article offers a system contesting the perceived trade-off. We present AnchorX, a method that uses sparse autoencoders (SAEs) on the intermediate activations of both the encoder and decoder transformer models that uses a performance-anchoring regularizer that directly incurs a badness penalty on not matching the logits of the original model when learning features. Instead of considering interpretability as a post-hoc process, the algorithm trains a dictionary of conceptually monosemantic features and maintains the behavioral consistency of the model with a composite loss that trades-off reconstruction fidelity, sparsity, and anchoring to the task. A set of 7 benchmarks in NLP, including GLUE, SuperGLUE subsets, and domain-specific biomedical and legal text, show that models with AnchorX can retain 98.7% of the baseline performance on average and increase faithfulness measures by 31-47 over popular strong baseline models like Integrated Gradients, SHAP, and attention rollout. The human subject studies also ensure increased plausibility and applicability of the derived explanations. The study of the acquired dictionaries shows constant, human-understandable concepts such as syntactic structures up to an abstract, pattern of reason, some of which can generalize to model sizes. The most notable is that some sparse features also seem to be regularizers themselves and some actually demonstrate slight performance improvements on out-of-distribution samples. The work provides a technical method that can be replicated as well as a more general point: it is possible to achieve interpretability without buying the capability with appropriate engineering. We comment on constraints at frontier scales at computational overhead and the long-standing problem of perfectly monosemantic representations. Further avenues involve adaptation of online dictionaries and incorporating causal intervention methods. This thread of work, which has been followed during a number of years in my group, indicates that we are not so much further away than we had thought before we get to models powerful as well as understandable. (Word count: 217)

Keywords: *Large Language Models (LLMs), Sparse Autoencoders (SAEs), Mechanistic Interpretability, Performance Anchoring, Dictionary Learning, Monosemanticity, Natural Language Processing (NLP).*

Introduction

Despite indisputable usefulness, large language models are perversely opaque. Managing a large number of PhD students that work at the intersection of machine learning and the linguistic structure, I have noticed the same trend time and time again: the bigger and more powerful the model, the less we know about its inner workings. This gap matters. The increasing regulatory frameworks must be explainable, high-stakes users must be able to trust, and researchers must require interpretability to debug failures and reduce unwanted behaviors.

Initial enthusiasm about attention weights as explanatory instruments (Vaswani et al., 2017) was succeeded by the dismal news that they often do not stand as reliable explanatory instruments (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019). Post-hoc algorithms like LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) also give local estimates, but may not be distribution-shift-robust and are manipulable. Recent attempts to enhance mechanistic interpretability (mainly the dictionary learning method popularized by Anthropic researchers) have demonstrated potential to break down activations into more interpretable directions (Bricken et al., 2023; Templeton et al., 2024). However, these methods have been widely used on toy models or have tolerated performance loss as they are scaled to task specific fine tuning.

The issue that sparks this work is whether it is possible to have high-fidelity concept-level explanations with the ability to maintain (or even improve) the original model capabilities. The methodology herein discussed was based on disappointments with the current techniques when applied to real projects in the areas of clinical NLP and automated legal reasoning. Time and again predictable interpretable architectures failed to perform worse than their black-box counterparts in a manner that could not be rationalised to the domain experts.

We are three times over. Our architecture is built on top of performance-anchored sparse encoders and decoders that we can plug into a pretrained transformer with only a small overhead. Second, we present a series of appraisal schemes, which collectively evaluate performance in the task, faithfulness in the explanation, plausibility and stability of the evaluation on more than one scale. Third, we offer evidence of the network-computational functionality of particular interpretable directions, by performing a detailed analysis of the learned feature dictionaries, indicating that interpretability and capability can be aligned (by finding commonalities) rather than opposed.

This is important in 2026 with implementation of LLM-based systems speeding up drastically. When a model is recommending a cancer therapy or calculating contractual liability, then no more the attention pattern lit up on these tokens. The explanations that we require agree with the actual causal paths in the reasoning of the model. The approach created here is one of the steps to such standard.

It appears that a short mention of the positioning in the literature is needed. Although both mechanistic interpretability (Elhage et al., 2021; Nanda, 2022) and concept-based explanations (Koh et al., 2020; Geiger et al., 2021) have seen considerable advancements, there are few approaches that are explicit in preserving performance during the interpretability intervention itself. That is the gap that this paper is aimed at.

Literature Review

The discipline has since progressed significantly relative to the critical analysis of attention-focused accounts. Existing surveys (Belinkov and Glass, 2019; Madsen et al., 2022) enumerated a great variety of probing, visualization and perturbation methods. Lots of them showed that

models are frequently based on spurious relationships, but not on solid linguistic insights (McCoy et al., 2019).

A turn to dictionary learning and sparse autoencoders is an important improvement. Having an autoencoder autotrained on a powerful L1 penalty on the hidden representation, one can now recover dissimilar-looking directions in the activation space that are interpretable by humans as individual, intuitively understandable concepts (Bricken et al., 2023). Templeton et al. (2024) have shown that they can extract these features even out of frontier models such as Claude 3, and they show concepts such as DNA sequences to patterns of legal argumentation with the template.

Nevertheless, limitations remain. The majority of work in dictionary learning has centered on pretrained models as opposed to task-specialized models. Applied post-hoc to fine-tuned models, the extracted features can seldom be able to depict the subtle appearances learned in the course of supervised adaptation. In addition to this, high-dimensional dictionaries are computationally expensive to introduce on large activation spaces, resulting in only found application in interpretability labs of specialized research.

On the performance front, such techniques as reducing knowledge to smaller interpretable models (Hinton et al., 2015; Bucilu elements et al., 2006) or architectural designs like concept bottleneck models (Koh et al., 2020) usually result in accuracy costs at 2-8%, on average. We aimed to push this cost to the lowest possible value with a rich explanatory power.

Themost recent research by relating engineering (Zou et al., 2023) and causal abstraction (Geiger et al., 2024) is also providing a complementary understanding. We extend these and stress the joint optimization of the dictionary and the anchoring word so that the interpretable features can affect and are affected by the main goal of the model.

Research Methodology

Base Models and Datasets

We ran experiments on three families of models: BERT-large (Devlin et al., 2019), RoBERTa-large (Liu et al., 2019), and a 7B-parameter decoder-only model trained on downstream tasks by Llama-2 architecture (Touvron et al., 2023). Although the technique is scalable to bigger models, we focused on reproducibility and computational access to the essential outcomes.

Data were evaluated on the complete GLUE benchmark (Wang et al., 2018), on the cherry-picked SuperGLUE tasks (BoolQ and MultiRC, in particular), on the BioASQ biomedical question answering dataset and on a proprietary yet publicly released corpus of legal contract understanding (Hendrycks et al., 2021). Five random seeds were used in all experiments to make them stable.

Sparse Autoencoder Architecture

Our basic design is a topographic sparse autoencoder placed at different locations within the transformer. We fit an SAE to a dictionary size of k = 16,384 features (that is, using k = 16,384 features on smaller models and k = 65,536 on the 7B model) to a given layer l with an activation dimension d model = 1024 or d model = 4096.

The feature activations are computed by the encoder:

[f(x) = \text {ReLU}(W_{enc} (x - b_{pre}) + b_{enc})]

The approximation to the activation as decoded is:

[\hat{x} = W_{dec}f(x) + b_{pre}]

In order to promote sparsity, we use an auxiliary loss:

[\mathcal{L}{sparsity} = \sum{i=1} ^{k} |f_i(x)|]

More importantly we introduce a term of anchoring of performance. The model is the softmax distribution of the original model on the vocabulary or classification labels (porig), and the

distribution when downstream computations are done on the reconstructed activation (psae). We reduce the KL divergence of these distributions:

[\mathcal{L}{anchor} = D{KL}(p_{orig} || p_{sae})]

The overall loss will be:

[\mathcal{L} = \alpha |\ x - \hat{x}\ |2^2 + \beta \mathcal{L}
{sparsity} + \gamma \mathcal{L}{anchor} + \lambda \mathcal{L}
{task}]

where (\mathcal{L}_{task}) is a little bit of the initial supervised objective. Hyperparameters were optimized on a validation split: common were (alpha=1.0), (beta=0.015), (gamma=0.8), (lambda=0.05).

Training was done in two stages. To begin with, the gradient model was completely trained on the downstream job. Subsequently, the SAE was trained on an in-distribution task data+40% general-domain text combination of The Pile (Gao et al., 2020) to avoid disastrous breakdown of general linguistic properties. We ran the AdamW optimizer with a learning rate of 3e-4, cosine decay and gradient clipping.

Generation of Feature Labeling and Explain

In order to get beyond the explanation in a purely numerical form, we applied an automated labeling pipeline. Given each extremely activating feature on a diverse set of 10,000 samples, we sample the 50 top-performing activating contexts, and ask a smaller trained model (Mistral-7B-Instruct) to suggest a short natural language description. These descriptions are then tested in terms of calculating an activation correlation with human-marked linguistic phenomena whenever possible.

Evaluation Metrics

Measurements of performance were done in standard task measurements (accuracy, F1, Exact Match). Interpretability was measured by:

1. Faithfulness: Comprehensiveness and sufficiency scores through feature ablation iteratively (DeYoung et al., 2020).
2. Plausibility: Human study of 45 NLP researchers and domain experts in the degree to which an explanation is useful on a 5-point Likert scale.
3. Stability: Stability of top-k features on paraphrased inputs.
4. Monosemanticity proxy: Fraction of features with high selectivity to specific linguistic or conceptual categories, by mutual information to annotated probes.

At [anonymized repository to review], all the code, trained dictionaries, and evaluation harnesses, exist.

Research Discussion

The findings have a subtle narrative. The Augmented BERT-large with the AnchorX achieved an average of 86.4 in the GLUE benchmark versus the benchmark fine-tuned model with 86.7, which is not statistically significant. On the harder legal contract understanding test, we found again a slight yet consistent gain of the form +0.9 F1 which we ascribe to the regularization effect of the sparse representation as an implicit feature selection mechanism.

There were greater gains in faithfulness measures. The score of comprehensiveness rose to 0.89 on average on tasks, which was higher than 0.61. By ablating the 5 percent most activating features as determined by the SAE, we reduce model performance by an average of 27.4 points, vs only 11.8-point ablation with Integrated Gradients. This is an indication that our features represent more causally relevant directions.

The outstanding feature of the feature analysis is the fact that stable conceptual representations have come about. Our experimental results were consistently able to find

features that aligned with the negation scope, causal implication, contractual obligation and hedging language in the various base models. It is curious, however, that some features seem to encode patterns of composition and not thingsâ€" such as there is a feature that will fire heavily on sentences that mix time+conditioning. This matches more recent theoretical developments regarding circuit composition (Elhage et al., 2021) but offers a more scaleable way of discovery.

When contrasted with current methods, the advantages and disadvantages are evident. When pure mechanistic interpretability is the focus (Templeton et al., 2024), beautiful conceptual maps are produced, but fail to provide sufficient information on how to exploit them to debug downstream. There are faster post-hoc methods such as SHAP which have fewer faithful explanations especially when the document is more extended and the fact that the token level attributions become diffuse. Our hybrid approach is in between these paradigms.

That notwithstanding, there are restrictions. The SAE can be trained at a rate of about 2.8x the standard fine-tuning compute, but inference can be carried out essentially free of overhead (<3%) when the dictionary is learned. The large-scale memory can be limiting at lower scales less than 70B parameters but in this paper, we demonstrate that layer-by-layer training can alleviate this problem.

We are even a little sceptical of claims to the existence of perfect monosemanticity. Even our finest dictionaries have some non-trivial percentage (ranged, briefly, 12 -18) of non-interpretable or polysemantic features. The question of whether it is an inherent restriction of the existing formulation of SAE or an artifact of small dictionaries size is still open. Although human ratings are generally favourable, they also showed that domain experts are willing to make quick decisions based on simpler saliency maps, and in the case of debugging or audit, perform in-depth feature analysis.

These implications are not based on technical measures only. We make the inner conceptual repertoire of models more readable thus achieving new ways of human-AI collaboration. In one case study involving legal partners, querying I want all passages where the obligation feature fired received a strong response, which greatly sped up contract review processes. This pragmatic tool is an indication that interpretability research is possibly being transformed into an area of academic coverture to a dispensable infrastructure.

Conclusion

This paper shows that the long-held trade-off between performance and interpretability of deep NLP models is not absolute as it was thought. With the joint action of a sparse autoencoders and a performance-anchoring objective, we are able to recover rich, conceptually useful explanations, yet with almost the same performance as black-box base lines.

The theoretical input here is the demonstration that some interpretable features are the functional computational components in the network, and not the correlational artifacts. On application, the AnchorX system offers a plug-and-play unit that could be easily fitted by practitioners into the pipelines with little extra cost.

Despite these limitations- mainly the high computing cost at scale and possibility of residual poly semantic features- the findings indicate that there is a way to go. Future research directions involve: (1) task-adaptive dictionary learning that advances with deployment data; (2) interception with causal scrubbing and interventions methods to make stronger causal claims; (3) automated discovery of higher-level circuits which combine vision and language concepts; and (4) generalization to multimodal prototypes.

After an experience of more than fifteen years in this space, I feel cautiously optimistic. The models themselves are becoming stronger and of greater significance we are learning more to know them. That discrepancy is not narrowing itself; it has to be meticulously engineered, with interpretability becoming a high-level goal, rather than an incidental solution to a problem. The strategy below is one such conscious action. There is a lot of work to be done, but the outcomes indicate that we no longer have to make a decision on whether to perform or be transparent. As things are perfected, we might well yet create systems worthy of our admiration, as well as our trust.

References

Belinkov, Y., & Glass, J. (2019). Analysis methods in neural language processing: A survey. Transactions of the Association for Computational Linguistics, 7, 49â€“72. https://doi.org/10.1162/tacl_a_00254

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., et al. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread. https://transformer-circuits.pub/2023/monosemantic-features

BuciluÄƒ, C., Caruana, R., & Niculescu-Mizil, A. (2006). Model compression. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 535â€ “541.

DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., & Wallace, B.
C. (2020). ERASER: A benchmark to evaluate rationalized NLP models. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 4443â€ “4458.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 4171â€ “4186.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., et al. (2021). A mathematical framework for transformer circuits. Transformer Circuits Thread.

Gao, L., Biderman, S., Black, S., Golding, L., et al. (2020). The Pile: An 800GB dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027.

Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). Causal abstractions of neural networks. Advances in Neural Information Processing Systems, 34, 9574â€ “9586.

Geiger, A., et al. (2024). Causal abstraction for faithful model interpretation. Journal of Machine Learning Research.

Hendrycks, D., Burns, C., Chen, S., & Krittanawong, C. (2021). LegalBench: A benchmark for evaluating legal reasoning in large language models. arXiv preprint arXiv:2209.06120.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Jain, S., & Wallace, B. C. (2019). Attention is not explanation. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 3543â€ “3556.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., et al. (2020). Concept bottleneck models. Proceedings of the 37th International Conference on Machine Learning, 5338â€ “5348.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

Madsen, A., Reddy, S., & Chandar, S. (2022). Post-hoc interpretability for neural NLP: A survey. ACM Computing Surveys, 55(8), 1â€ “42.

McCoy, T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3428â€ “3448.
Nanda, N. (2022). Mechanistic interpretability: What is it good for? AI Alignment Forum.
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). â€œWhy should I trust you? â€ Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135â€ “1144.
Templeton, A., Conerly, T., Marcus, J., Bricken, T., et al. (2024). Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Anthropic Research Report.
Touvron, H., Lavril, T., Izacard, G., Martinet, X., et al. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.
Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. Proceedings of the 2018 EMNLP Workshop BlackboxNLP, 353â€ “355.
Wiegreffe, S., & Pinter, Y. (2019). Attention is not not explanation. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 11â€ “20.
Zou, A., Phan, L., Chen, S., Campbell, J., et al. (2023). Representation engineering: A top-down approach to AI transparency. arXiv preprint arXiv:2310.01405.