Linear Decoding of Morphology Relations in Language Models (Student Abstract)

Eric Xia[1], **Jugal Kalita** [2]

[1]Brown University
[2]University of Colorado Colorado Springs
eric_xia@brown.edu, jkalita@uccs.edu

Abstract

The recent success of transformer language models owes much to their conversational fluency, which includes linguistic and morphological proficiency. An affine Taylor approximation has been found to be a good approximation for transformer computations over certain factual and encyclopedic relations. We show that the truly linear approximation $W\mathbf{s}$, where $\mathbf{s}$ is an early layer representation of the base form and W is a local model derivative, is necessary and sufficient to approximate *morphological derivation*, achieving above 80% top-1 accuracy across most morphological tasks in the Bigger Analogy Test Set. We argue that many morphological forms in transformer models are likely linearly encoded.

Code — https://github.com/rkique/linear-morphology

Introduction

Transformer language models (LMs) display impressive capabilities for factual recall, which commonly involve relations between entities (Brown et al. 2020). Work to date around relational representation in LMs have focused on factual relations (Meng et al. 2022); however, linguistic competency involves a much broader range of relations between subjects and objects. The impressive conversational ability of LMs depends on their lexical and morphological productivity, and uncovering how models are able to achieve this is an important aspect of model interpretability.

Approach

In an LM, input text is converted to a sequence of tokens $t_1 \dots t_n$ embedded as $x_1 \dots x_n \in \mathbb{R}^d$. The hidden states $x_1 \dots x_n$ are then passed through L transformer layers, each composed of a self-attention layer a^l and an multi-layer perceptron (MLP) layer m^l, and then decoded by the decoder head D to a probability distribution over tokens. The representation state x_i^l of the i^{th} token at layer l is then obtained as

$$x_i^l = x_i^{l-1} + a_i^l + m_i^l$$

, where a_i^l is multi-headed Key-Value Query attention over x^{l-1}(Vaswani et al. 2017) and m_i^l the i^{th} output of the l^{th} MLP sublayer. Following the insights of Meng (2022) and Geva (2023) that the last subject token state in middle layers are strongly casual, we are interested in utilizing the gradient between the last token position of the subject s, and the last token position overall, the object prediction o. Paccanaro and Hinton (2001) introduced the concept of the linear relational embedding (LRE) for learning relational knowledge from subject-relation-object triples. We directly build off the affine LRE (Hernandez et al. 2023), which assumes that the LM is implicitly learning affine embeddings within specific contextual relations. They choose to approximate the LM computation from an early subject state $x_s^i = \mathbf{s}$ to a final-layer object state $x_o^L = \mathbf{o}$ with a fixed relational context, $\mathbf{o} = F(\mathbf{s})$, by a first-order Taylor approximation across examples s_i, for $i = 1 \dots n$:

$$\begin{aligned} F(\mathbf{s}) &\approx F(s_i) + W(\mathbf{s} - s_i) = F(s_i) + W\mathbf{s} - Ws_i \\ &= W\mathbf{s} + b, \\ \text{where } b &= F(s_i) - Ws_i \end{aligned}$$

They find W and b can be derived from the *model* Jacobian of n subjects $s_1, \dots, s_n$ and their objects $F(s_1), \dots, F(s_n)$:

$$W = \mathbb{E}_{s_i}\left[\left.\frac{\partial F}{\partial s}\right|_{s_i}\right] \qquad b = \mathbb{E}_{s_i}\left[\left.F(s) - \frac{\partial F}{\partial s}\, s\right|_{s_i}\right]$$

However, the affine LRE [1]:

$$\mathbf{o} = F(\mathbf{s}) \approx \beta W\mathbf{s} + b$$

diverges from its namesake by introducing bias b and scaling β terms. Assuming the relation is not only linearly approximable e.g. continuous around s_i, but truly linear, we would expect the following to be valid:

$$F(\mathbf{s}) \approx F'(s_i)\mathbf{s}$$

Then $\mathbf{o} = F(\mathbf{s}) \approx W\mathbf{s}$ is faithful to the original form (2001), and is expected to work over truly linear relations. We adapt the Bigger Analogy Test Set – BATS – (Gladkova, Drozd, and Matsuoka 2016) [2] to a relational format by introducing prompts for each analogy which are compatible with

[1]The scalar β is introduced to account for differences in magnitude due to layer normalization in the decoder head.

[2]BATS is 50 pairs for 40 analogies across derivational/inflectional morphology, encyclopedic, and semantic categories

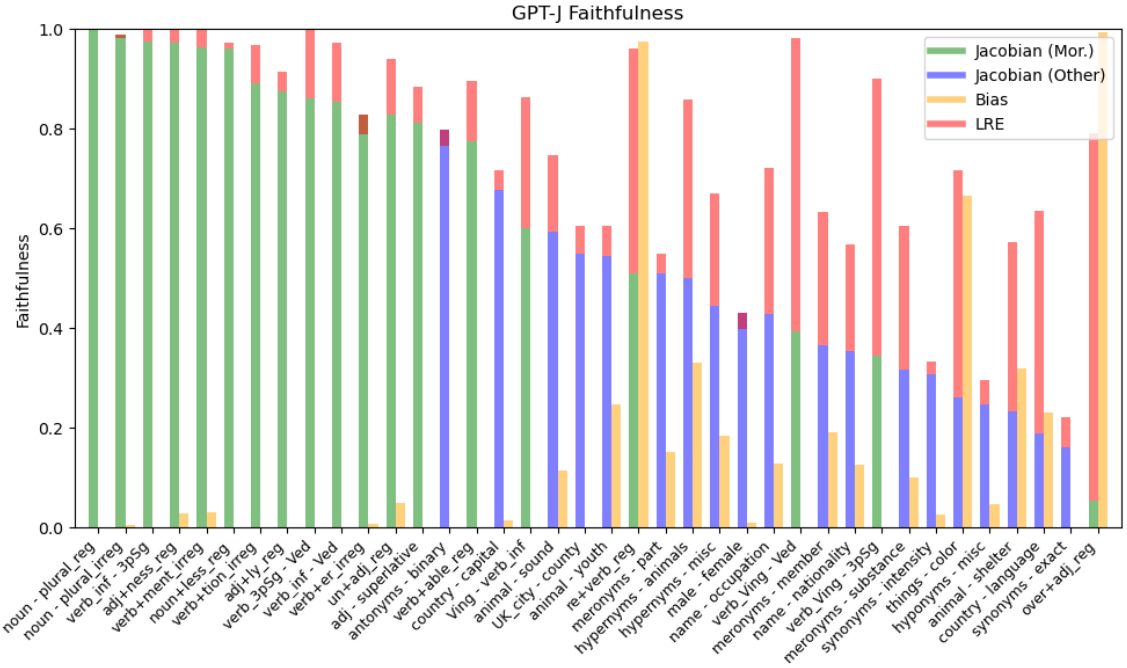

Figure 1: Breaking down the affine LRE ($\beta W\mathbf{s}+b$) into Jacobian ($W\mathbf{s}$) and Bias ($\mathbf{s}+b$) approximators shows that W is sufficient and necessary to model most morphological relations. It also suggests that W and b play complementary roles: the Jacobian is responsible for approximating alternate forms, while the bias is responsible for conceptual shifts.

all instances of the analogy. For example, the [**verb+ment**] dataset comprises pairs of words with base ("fulfill") and derived ("fulfillment") forms. The prompt given to the LM then elicits the derived form from the base by a clozed semantic relation: "To fulfill results in a _____". To increase the probability of a correct prediction through in-context learning, we utilize ICL, prepending 7 complete examples of the phrase. The weight W and bias b are then determined by the average model Jacobian and true object states, as formulated above. For token t and decoder head D, we calculate faithfulness as the top-1 match rate of the approximator and LM:

$$\underset{t}{\text{argmax}}\, D(F(\mathbf{s}))_t \stackrel{?}{=} \underset{t}{\text{argmax}}\, D(\tilde{F}(\mathbf{s}))_t$$

We test the original LRE $\beta W\mathbf{s}+b$ against the Jacobian $W\mathbf{s}$, as well as Bias $\mathbf{s}+b$. Comparing these approximations, we find the Jacobian approximation is both *sufficient* (achieving high faithfulness) and *necessary* (removing W compromises faithfulness) to model morphological relations. In other categories, the LRE improves significantly past the Jacobian, suggesting that other relations which are affinely approximable are not necessarily linear as well. [3]

Discussion

Following the affine LRE (Hernandez et al. 2023), we have uncovered an implicit linear relation between transformer states, in which multiplying a hidden state with a model derivative faithfully approximates the linguistic process of morphological derivation. This decoding suggests that transformers encode morphology in a linear manner depending on a single token state.

Morphological relations involve well characterized concepts between words. However, contrary to theories such as the Linear Representational Hypothesis (Park et al. 2024), which broadly posit that directions in the representation space of a language model encode high-level concepts, we have found that morphological derivations are well-approximated with a linear transformation, and not by concept vectors (See the Bias approximator in Figure 1, or the TRANSLATION approximator in the supplementary material). This highlights regular structure in language models which has gone unexplored to date.

In summary, this work illuminates a linear decoding for transformer LM computations in morphological contexts, in which transforming a subject hidden state by a single derivative is able to approximate object hidden states. The existence of this approximation reflects regularity in the inputs and outputs in the model representation space.

This finding helps explain a crucial aspect of how LMs achieve linguistic fluency, and has implications in downstream applications of LMs which require precise control of outputs. More broadly, it indicates that linear embeddings can be implicitly present in language models in various settings, and are a promising avenue for future research.

[3] Results for GPT-J are shown in Figure 1; Llama-7b yields similar results. See supplementary material for a detailed analysis.

Acknowledgements

The research done here was supported by the National Science Foundation under award number #2349452. Any opinion, finding, or conclusion in this study is that of the authors and does not necessarily reflect the views of the National Science Foundation.

The authors extend their gratitude to the VAST Lab at University of Colorado Colorado Springs for supplying the hardware necessary to run the experiments.

They would also like to thank Jack Merullo and Ellie Pavlick at the Brown Language Understanding and Representation Lab for their guidance throughout the research process.

References

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Geva, M.; Bastings, J.; Filippova, K.; and Globerson, A. 2023. Dissecting Recall of Factual Associations in Auto-Regressive Language Models. ArXiv:2304.14767 [cs].

Gladkova, A.; Drozd, A.; and Matsuoka, S. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In Andreas, J.; Choi, E.; and Lazaridou, A., eds., *Proceedings of the NAACL Student Research Workshop*, 8–15. San Diego, California: Association for Computational Linguistics.

Hernandez, E.; Sharma, A. S.; Haklay, T.; Meng, K.; Wattenberg, M.; Andreas, J.; Belinkov, Y.; and Bau, D. 2023. Linearity of relation decoding in transformer language models. *arXiv preprint arXiv:2308.09124*.

Meng, K.; Sharma, A. S.; Andonian, A. J.; Belinkov, Y.; and Bau, D. 2022. Mass-Editing Memory in a Transformer. In *The Eleventh International Conference on Learning Representations*.

Paccanaro, A.; and Hinton, G. E. 2001. Learning distributed representations of concepts using linear relational embedding. *IEEE Transactions on Knowledge and Data Engineering*, 13(2): 232–244.

Park, K.; Choe, Y. J.; Jiang, Y.; and Veitch, V. 2024. The Geometry of Categorical and Hierarchical Concepts in Large Language Models. arXiv:2406.01506.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, u.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.