Linear Decoding of Morphology Relations in Language Models

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 and productivity in linguistic and morphological aspects. 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 a middle layer representation of the base form and W is a local model derivative, is necessary and sufficient to approximate *morphological derivations*, 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 subject-object relations (Meng et al. 2022); however, linguistic competency involves a much broader range of relations. 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.

Paccanaro and Hinton (2001) introduced the concept of the linear relational embedding 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 linear relational embeddings within specific contextual relations which are linearly approximable. They choose to model the LM computation from a subject $x_s^i = \mathbf{s}$ to an to an object state $x_o^L = \mathbf{o}$ with a fixed relational context, $\mathbf{o} = F(\mathbf{s})$, by a Taylor approximation for examples s_i for $i = 1 \dots n$:

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

Then, W and b can be derived from the Jacobian of n subjects $s_1, \dots, s_n$ and their objects $F(s_1), \dots, F(s_n)$:[1]

$$W = \mathbb{E}_{s_i}\left[\left.\frac{\partial F}{\partial s}\right|_{s_i}\right]$$

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

However, the LRE $\mathbf{o} = F(\mathbf{s}) \approx \beta W\mathbf{s} + b$ diverges from its namesake by introducing bias b and scaling β terms. Assuming the relation is not only linearly approximable, 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 expected to work over truly linear relations.

We adapt the Bigger Analogy Test Set (Gladkova, Drozd, and Matsuoka 2016) [2] to a relational dataset with relation-specific prompts for each analogy. For example, the derivational morphology dataset **[verb+ment]** uses the clozed prompt *"To {} results in a {}"*. We create prompts by filling in subject to elict an object: *"To fulfill results in*

[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.

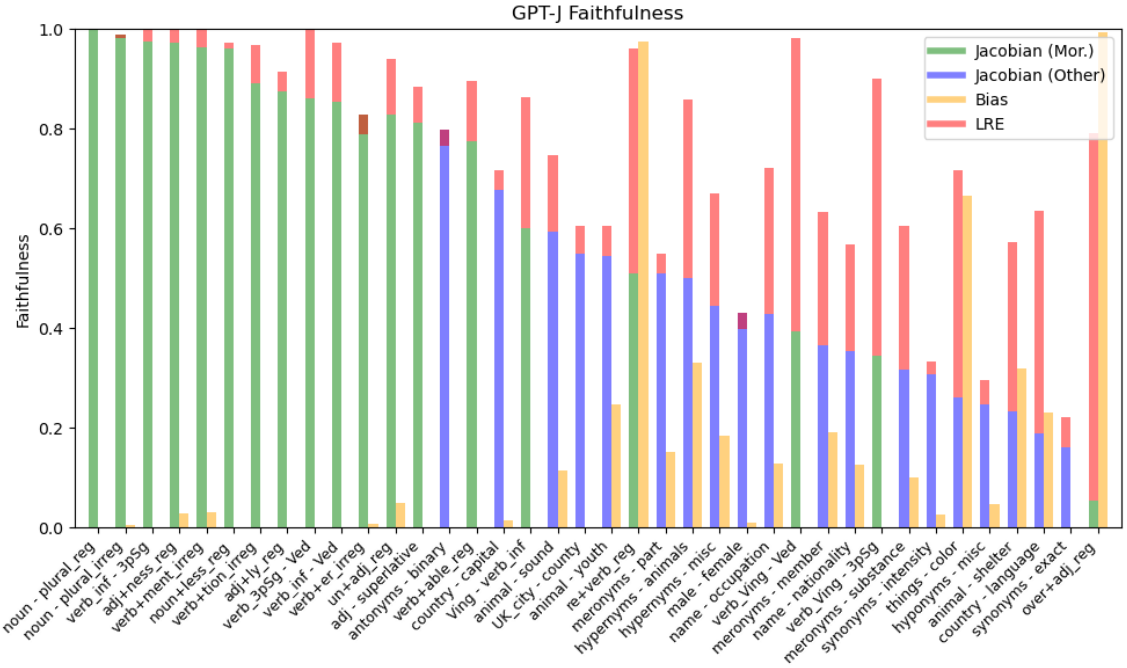

Figure 1: Breaking down the affine LRE into Jacobian ($W\mathbf{s}$) and Bias ($\mathbf{s} + b$) approximators shows that W is sufficient and necessary to approximate most morphological relations in BATS. It also suggests that W and b play complementary roles: the Jacobian is responsible for approximating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

a {fulfillment}". We use ICL to increase the probability of correct prediction, and average the Jacobian over 8 prompts to create an approximator as above. For token t and decoder head D, we calculate faithfulness as the top-1 match rate of the approximator and LM. We test the original LRE $\beta W\mathbf{s} + b$ against the Jacobian $W\mathbf{s}$, as well as Bias $\mathbf{s} + b$. We see that the Jacobian W is both *sufficient* (achieving high faithfulness) and *necessary* (removing W compromises faithfulness) to model morphological relations. [3]

Discussion

High faithfulness on inflectional tasks can be achieved by reproducing stemmed subject tokens, in which the Jacobian would not reflect morphological derivation. If this was the case, the Bias approximator $\mathbf{s} + b$ should perform as well as the LRE $W\mathbf{s} + b$. We further observe there exist many unmerged top-1 predictions such as #25303 ' sadness' and #24659 ' continuation' which replicate full derivations.

We have shown a linear transformation of a early hidden state with a model derivative approximates morphological derivations within relational contexts, suggesting that transformers encode morphology linearly. We are the first to discover a linear approximation which faithfully models an LM over a wide range of outputs. We have also shown LM relational approximation can be successfully applied to linguistic phenomena, opening avenues for further research.

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

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.

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.

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.