Token-wise Decomposition of Autoregressive Language Model Hidden States for Analyzing Model Predictions

Byung-Doh Oh William Schuler

The Ohio State University

oh.531@osu.edu
github.com/byungdoh/llm_decomposition

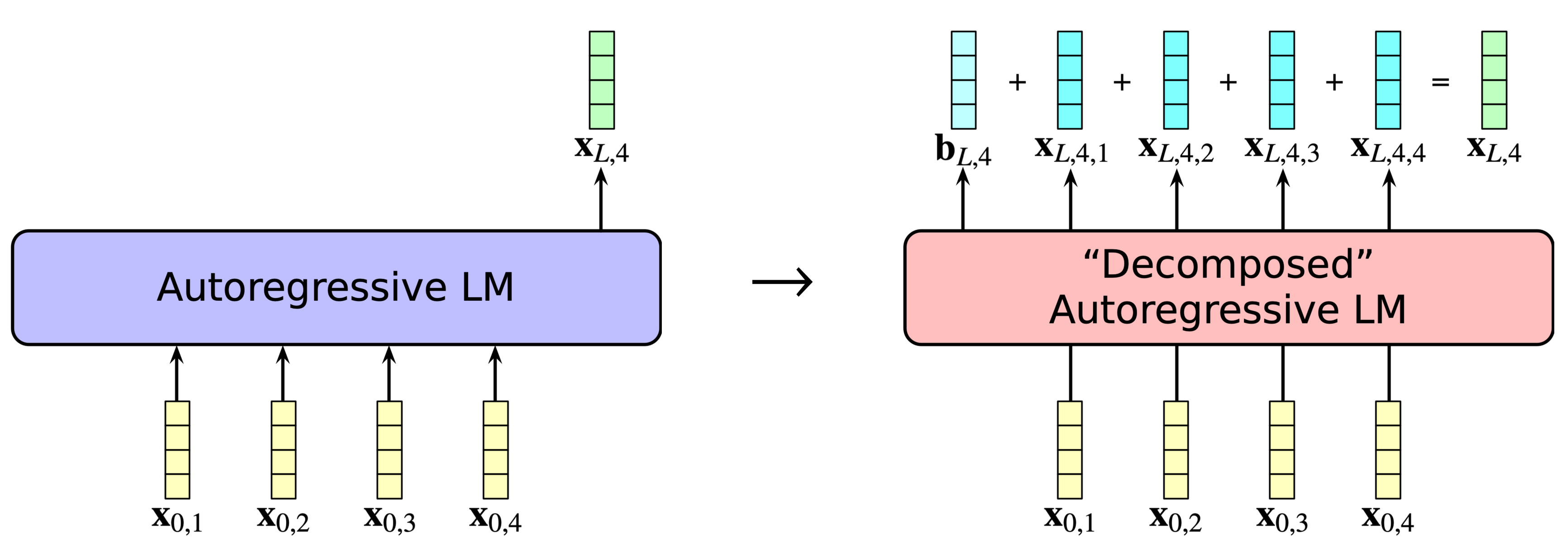
PAPER

Introduction

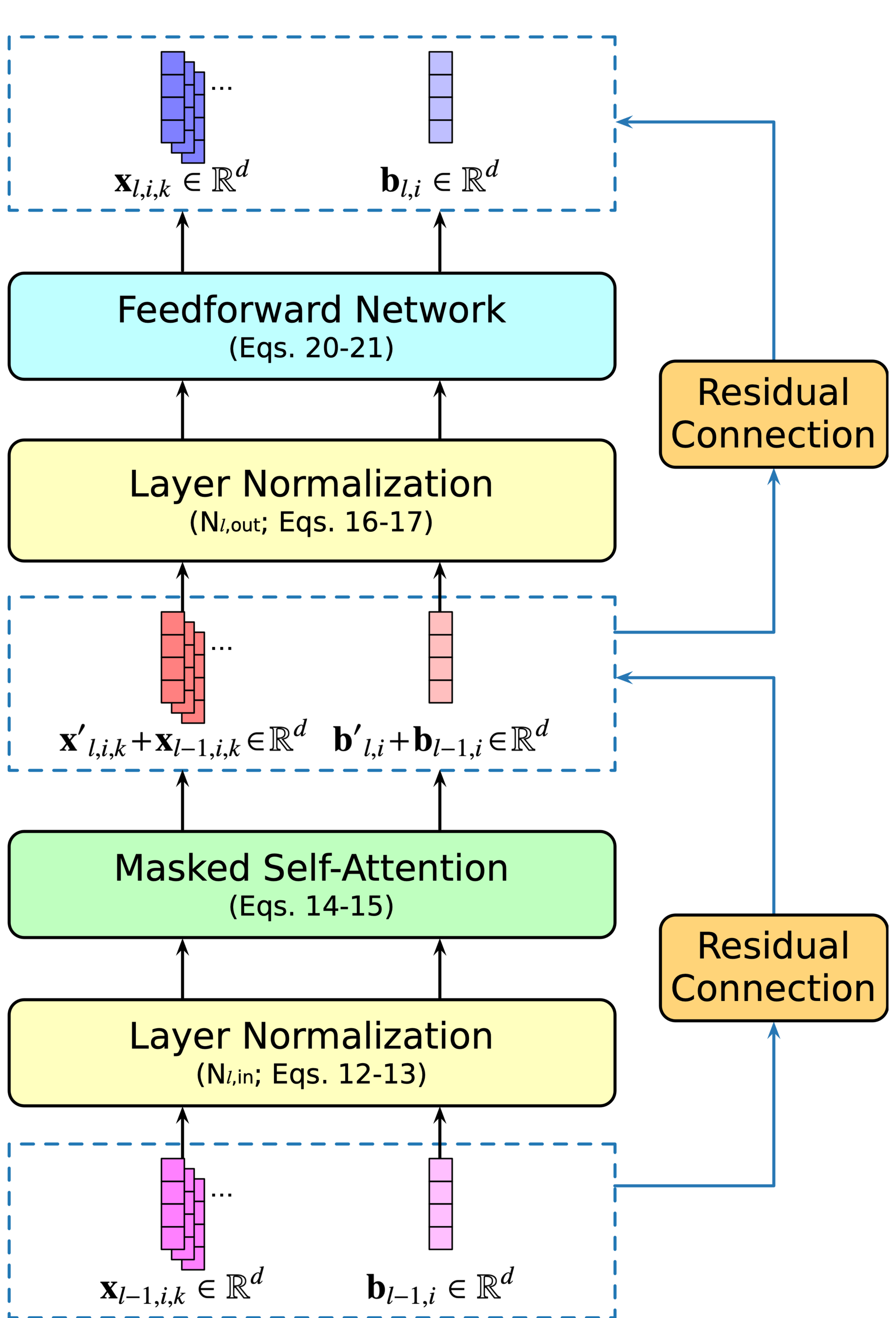
- There is much recent interest in interpreting the predictions of Transformer-based large language models [1, 6]
- However, analysis has been limited to studying the self-attention mechanism and feedforward network independently [2, 3]
- Additionally, widely used attribution methods (e.g. gradient norms) do not yield measures that are interpretable in terms of model probabilities
- This work presents a linear decomposition of hidden states that preserves the contribution of each input token, and an associated importance measure defined in terms of change in next-word probability

Token-wise Decomposition of LM Hidden States

- Vector of hidden states $\mathbf{x}_{L,i}$ is decomposed exactly into the sum of output representations of each input token $\mathbf{x}_{L,i,k}$ and a cumulative bias $\mathbf{b}_{L,i}$:



- This is achieved by maintaining input-specific vectors $\mathbf{x}_{l,i,k}$ and a bias-like vector $\mathbf{b}_{l,i}$ throughout the network:



- Layer normalization is applied using standard deviation of undecomposed representation
- Attention weights update total representation from source position k to target position i
- Activation function within the feedforward network is approximated using tangent slopes $\mathbf{s}$ and intercepts $\mathbf{i}$

$$\begin{aligned}\mathrm{FF}(\mathbf{y}) &= \mathbf{F}_2\,\sigma(\mathbf{F}_1\,\mathbf{y} + \mathbf{f}_1) + \mathbf{f}_2 \\ &= \mathbf{F}_2(\mathbf{s} \odot (\mathbf{F}_1\,\mathbf{y} + \mathbf{f}_1) + \mathbf{i}) + \mathbf{f}_2\end{aligned} \tag{1}$$

- All bias vectors are accumulated by $\mathbf{b}_{l,i}$

Importance Measure ΔLP: Change in Probabilities

The importance of w_k to the prediction of w_{i+1} is calculated as the difference between log probabilities of w_{i+1} given the context with and without w_k:

$$\Delta\mathrm{LP}(w_{i+1} \mid w_{1..i}, w_{k\in\{1,...,i\}}) = \log_2 \mathrm{P}(w_{i+1} \mid w_{1..i}) - \log_2 \mathrm{P}(w_{i+1} \mid w_{1..i\setminus\{k\}}), \tag{2}$$

$$\mathrm{P}(w_{i+1} \mid w_{1..i\setminus\{k\}}) = \underset{w_{i+1}}{\textsc{SoftMax}}(\mathbf{z}_i - \mathbf{z}'_{i,k}), \tag{3}$$

where $\mathbf{z}_i$ and $\mathbf{z}'_{i,k}$ are the vector of logit scores calculated using $\mathbf{x}_{L,i}$ and $\mathbf{x}_{L,i,k}$ respectively.

Correlation with Other Importance Measures

- Evaluation on the CoNLL-2012 corpus [5] and the WSJ corpus [4]
- ΔLP calculated using OPT-125M model [7] for each context token

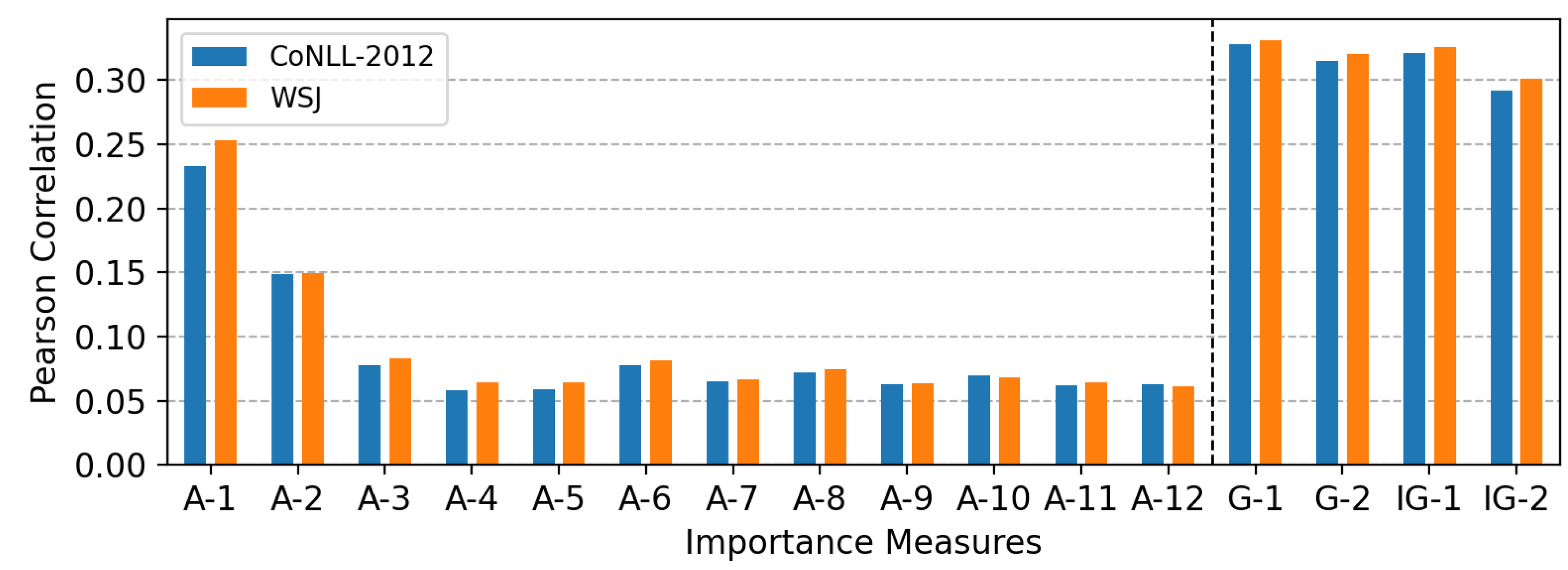


Figure 1: Pearson correlation between ΔLP and other importance measures. A-l is average attention at layer l; G-n is n-norm of gradient; IG-n is n-norm of input × gradient.

Characterizing High-Importance Context Words

- Stepwise regression models fit to the highest ΔLP value at each timestep on CoNLL-2012 [5]
- Baseline predictors: index of predicted word, linear distance from context word, log probability
- Predictors of interest: document PMI, bigram PMI, syntactic dependency, coreference relationship

Predictor	β	t-value	ΔLL
Word index	0.034	1.919	-
Distance	1.126	62.755	-
Log prob.	-0.083	-5.350	-
$\mathrm{PMI}_{\mathrm{bigram}}$	1.220	70.857	6151.262*
$\mathrm{PMI}_{\mathrm{doc}}$	1.286	73.952	3194.815*
Dependency	1.055	63.720	1981.778*
Coreference	0.123	7.195	25.883*

Table 1: Regression coefficients from the final regression model and increase in regression model likelihood (ΔLL) from including each predictor of interest.
*: $p < 0.001$.

Dependency and Coreference Prediction Using ΔLP

- Precision scores of syntactic dependency and coreference prediction calculated using high-importance words identified through ΔLP

Relation	ΔLP	Base.	$\mathrm{PMI}_{\mathrm{b}}$	$\mathrm{PMI}_{\mathrm{d}}$
Nom. subj.	61.15	39.79	1.38	1.44
Direct obj.	70.43	22.01	0.91	1.57
Oblique	52.54	24.31	-0.68	1.54
Compound	80.44	39.56	4.97	2.93
Nom. mod.	53.84	26.09	-0.41	1.84
Adj. mod.	82.55	36.02	4.36	2.17
Determiner	52.03	36.52	1.51	1.08
Case marker	52.38	27.96	-0.29	1.08
Microavg.	56.20	29.22	1.11	1.58

Table 2: Precision scores calculated using ΔLP, random word baseline, and average PMI of frequent syntactic dependency relations in the WSJ corpus.

Mention head POS	ΔLP	Base.	Rep.%
Personal pronoun	26.55	36.80	30.92
Possessive pronoun	23.29	36.45	30.59
Proper noun (sg.)	61.21	23.19	68.80
Proper noun (pl.)	70.67	57.33	68.00
Common noun (sg.)	43.39	12.55	48.75
Common noun (pl.)	47.01	24.73	55.03
Possessive ending	46.28	30.58	40.91
Microavg.	38.21	28.65	43.26

Table 3: Precision scores calculated using ΔLP, most recent head POS baseline, and proportion of repeated head words of frequent coreferent spans in the CoNLL-2012 corpus.

Conclusion

Results suggest that collocational association (PMI) strongly drives the predictions of Transformer-based autoregressive LMs

References

[1] Belinkov, Y. 2022. Probing classifiers: Promises, shortcomings, and advances.
[2] Geva, M., Caciularu, A., Wang, K., et al. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space.
[3] Kobayashi, G., Kuribayashi, T., Yokoi, S., et al. 2021. Incorporating residual and normalization layers into analysis of masked language models.
[4] Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. 1993. Building a large annotated corpus of English: The Penn Treebank.
[5] Pradhan, S., Moschitti, A., Xue, N., et al. 2012. CoNLL-2012 Shared Task: Modeling multilingual unrestricted coreference in OntoNotes.
[6] Rogers, A., Kovaleva, O., & Rumshisky, A. 2021. A primer in BERTology: What we know about how BERT works.
[7] Zhang, S., Roller, S., Goyal, N., et al. 2022. OPT: Open pre-trained Transformer language models.