DecompX: Explaining Transformers Decisions by Propagating Token Decomposition

Ali Modarressi*, Mohsen Fayyaz*, Ehsan Aghazadeh, Yadollah Yaghoobzadeh, Mohammad Taher Pilehvar

ACL 2023

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

mcml Munich Center for Machine Learning

TeIAS | Tehran Institute for Advanced Studies

Introduction

Summary

➔ We introduce *DecompX*, an explanation method based on propagating decomposed token vectors up to the classification head, which addresses the major issues of the previous vector-based methods.

➔ Our contributions:

1) We incorporated all the encoder layer components including nonlinear functions

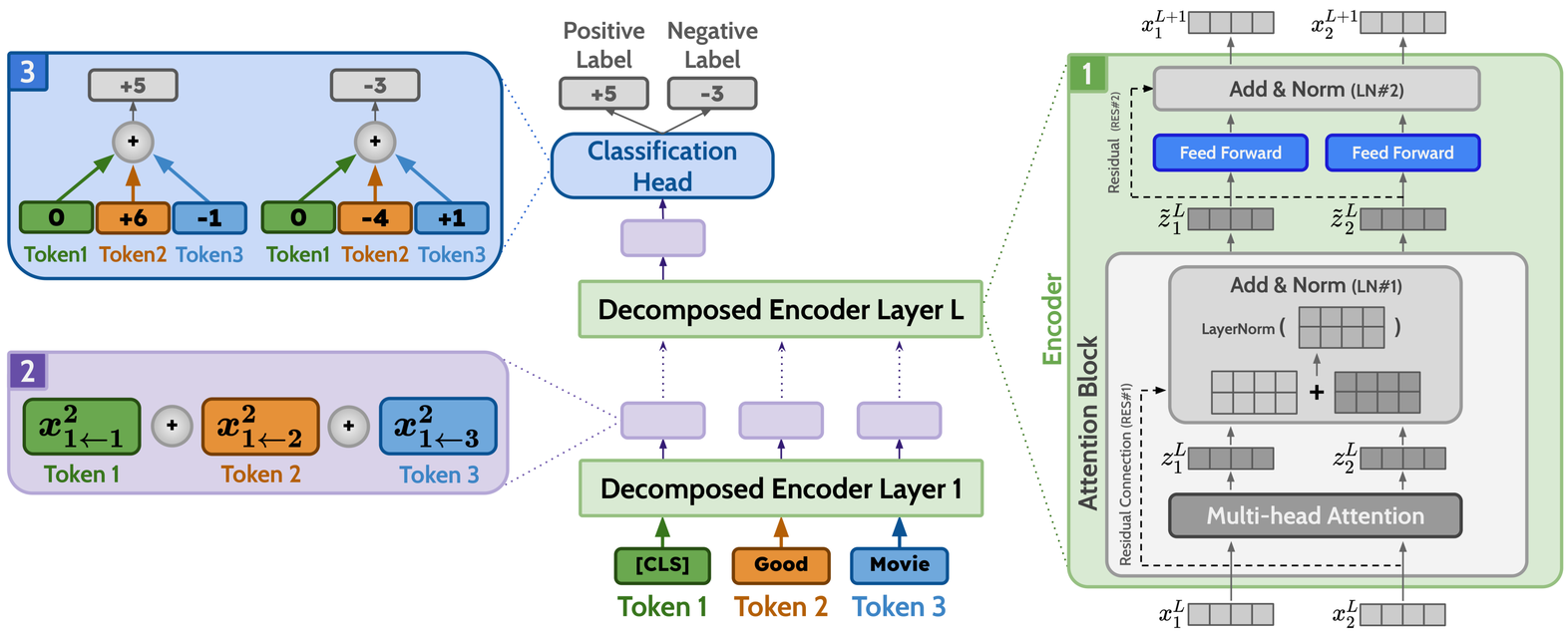
2) Propagated the decomposed vectors throughout the whole model

3) Incorporated the classification head

➔ Our results:

- DecompX is consistently better than existing vector- and gradient-based methods by a wide margin.

The Overall Workflow of DecompX



Methodology

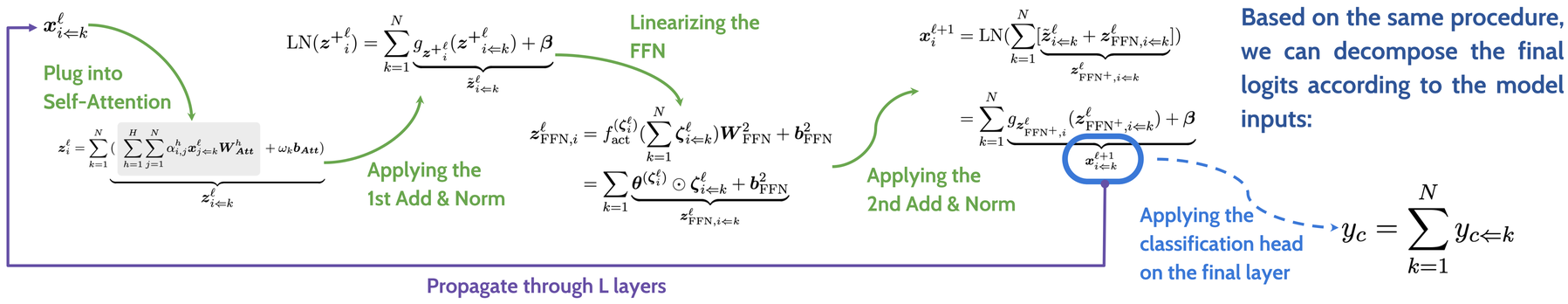
Decomposing token representations into their constituent vectors based on model inputs (k^{th} token):

$$\boldsymbol{x}_i^\ell = \sum_{k=1}^{N} \boldsymbol{x}_{i \Leftarrow k}^\ell$$

The goal is to propagate these vectors throughout the multilayer model. So we should determine $\boldsymbol{x}_{i \Leftarrow k}^{\ell+1}$ by having $\boldsymbol{x}_{i \Leftarrow k}^{\ell}$:

$$\boldsymbol{x}_{i \Leftarrow k}^{\ell} \rightarrow \text{Encoder}^\ell \rightarrow \boldsymbol{x}_{i \Leftarrow k}^{\ell+1}$$

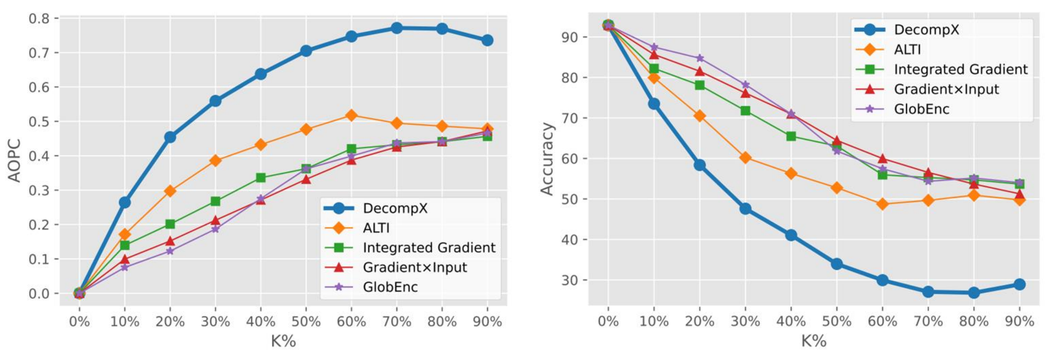
Propagating Token Decomposition



Based on the same procedure, we can decompose the final logits according to the model inputs:

Results

Faithfulness Results



➔ AOPC and Accuracy of different explanation methods on SST2 upon masking K% of the most important tokens

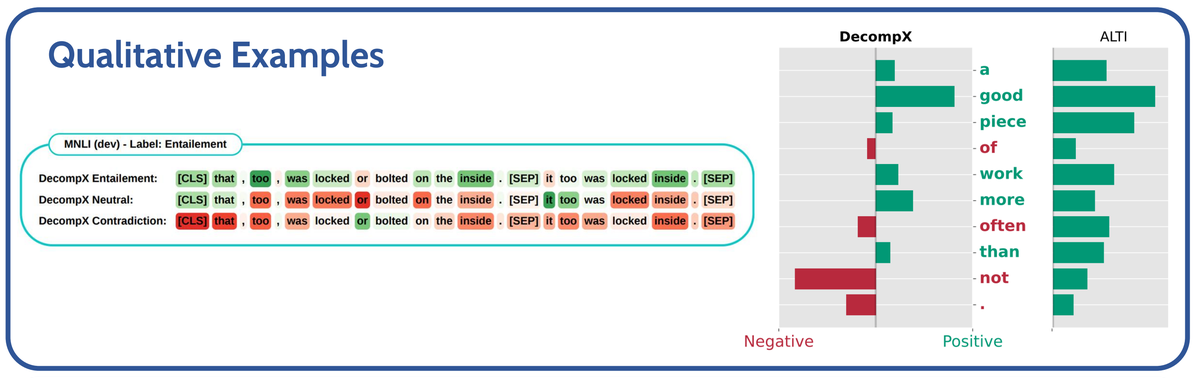
➔ DecompX outperforms existing explanation methods, both vector- and gradient-based, by a large margin at every corruption ratio.

	SST2			MNLI			QNLI			HATEXPLAIN		
	ACC↓	AOPC↑	PRED↑	ACC↓	AOPC↑	PRED↑	ACC↓	AOPC↑	PRED↑	ACC↓	AOPC↑	PRED↑
GlobEnc (Modarressi et al., 2022)	67.14	0.307	72.36	48.07	0.498	70.43	64.93	0.342	84.00	47.65	0.401	56.50
+ FFN	64.90	0.326	79.01	45.05	0.533	75.15	63.74	0.354	84.97	46.89	0.406	59.52
ALTI (Ferrando et al., 2022)	57.65	0.416	88.30	45.89	0.515	74.24	63.85	0.355	85.69	43.30	0.469	64.67
Gradient×Input	66.69	0.310	67.20	44.21	0.544	76.05	62.93	0.366	86.27	46.28	0.433	60.67
Integrated Gradients	64.48	0.340	64.56	40.80	0.579	73.94	61.12	0.381	86.27	45.19	0.445	64.46
DecompX	**40.80**	**0.627**	**92.20**	**32.64**	**0.703**	**80.95**	**57.50**	**0.453**	**89.84**	**38.71**	**0.612**	**66.34**

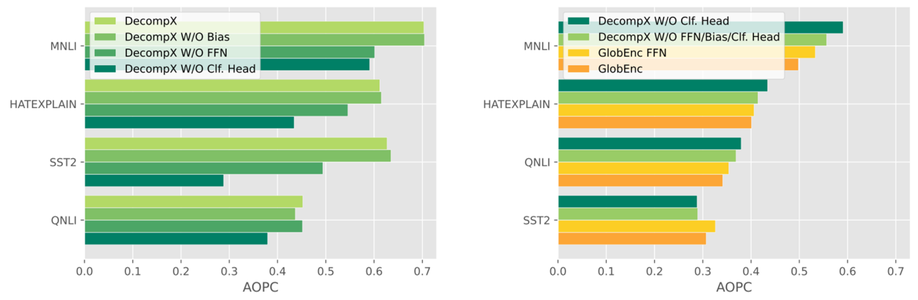
➔ Accuracy, AOPC, and Prediction Performance of DecompX compared with the existing methods on different datasets.

➔ DecompX consistently outperforms other methods, which confirms that a holistic vector-based approach can present higher-quality explanations.

Qualitative Examples



Ablation Studies



➔ Our experiments reveal that removing FFN significantly decreases the AOPC

➔ Even though considering bias in the analysis only has a slight effect, it is important to add biases for the human interpretability of DecompX.

➔ AOPC drastically drops when we do not consider the classification head, even more than neglecting bias and FFN, highlighting the important role played by the classification head.

➔ Even without the FFN and bias, decomposition can outperform the rollout-based GlobEnc. These results demonstrate that aggregation in-between layers causes information loss and the final attributions are susceptible to this simplifying assumption.

Conclusions

- We incorporated all the encoder layer components, propagated the decomposed vectors throughout the whole model, and for the first time, incorporated the classification head.
- We demonstrated that our method is consistently better than existing vector- and gradient-based methods by a wide margin.
- Check out our demo at: https://github.com/mohsenfayyaz/DecompX

References

- [1] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 7057–7075, Online. Association for Computational Linguistics
- [2] Ali Modarressi, Mohsen Fayyaz, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. 2022. GlobEnc: Quantifying global token attribution by incorporating the whole encoder layer in transformers. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 258–271, Seattle, United States. Association for Computational Linguistics
- [3] Javier Ferrando, Gerard I. Gállego, and Marta R. Costajussà. 2022. Measuring the mixing of contextual information in the transformer. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 8698–8714, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.