تحلیل دانش و توجه
مدل‌های زبانی مبتنی بر مبدل

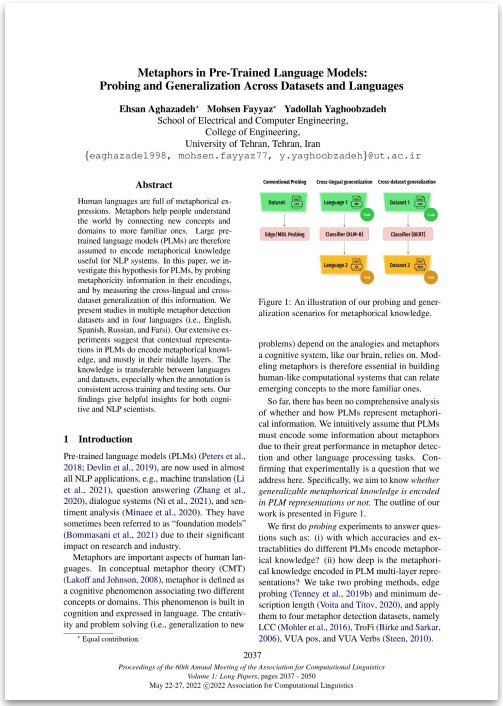
محسن فیاض

استاد راهنما

دکتر یداله یعقوب‌زاده

تابستان ۱۴۰۲

مقدمه



Metaphors in Pre-Trained Language Models: Probing and Generalization Across Datasets and Languages

Ehsan Aghazadeh* Mohsen Fayyaz* Yadollah Yaghoobzadeh
School of Electrical and Computer Engineering,
College of Engineering,
University of Tehran, Tehran, Iran
{eaghazadeh998, mohsen.fayyaz77, y.yaghoobzadeh}@ut.ac.ir

Abstract

Human languages are full of metaphorical expressions. Metaphors help people understand the world by connecting new concepts and domains to more familiar ones. Large pre-trained language models (PLMs) are therefore assumed to encode metaphorical knowledge useful for NLP systems. In this paper, we investigate this hypothesis for PLMs, by probing metaphoricity information in their encodings, and by measuring the cross-lingual and cross-dataset generalization of this information. We present studies in multiple metaphor detection datasets and in four languages (i.e., English, Spanish, Russian, and Farsi). Our extensive experiments suggest that contextual representations in PLMs do encode metaphorical knowledge, and mostly in their middle layers. The knowledge is transferable between languages and datasets, especially when the annotation is consistent across training and testing sets. Our findings give helpful insights for both cognitive and NLP scientists.

Figure 1: An illustration of our probing and generalization scenarios for metaphorical knowledge.

1 Introduction

Pre-trained language models (PLMs) (Peters et al., 2018; Devlin et al., 2019), are now used in almost all NLP applications, e.g., machine translation (Li et al., 2021), question answering (Zhang et al., 2020), dialogue systems (Ni et al., 2021), and sentiment analysis (Minaee et al., 2020). They have sometimes been referred to as "foundation models" (Bommasani et al., 2021) due to their significant impact on research and industry.

Metaphors are important aspects of human languages. In conceptual metaphor theory (CMT) (Lakoff and Johnson, 2008), metaphor is defined as a cognitive phenomenon associating two different concepts or domains. This phenomenon is built in cognition and expressed in language. The creativity and problem solving (i.e., generalization to new problems) depend on the analogies and metaphors a cognitive system, like our brain, relies on. Modeling metaphors is therefore essential in building human-like computational systems that can relate emerging concepts to the more familiar ones.

So far, there has been no comprehensive analysis of whether and how PLMs represent metaphorical information. We intuitively assume that PLMs must encode some information about metaphors due to their great performance in metaphor detection and other language processing tasks. Confirming that experimentally is a question that we address here. Specifically, we aim to know *whether generalizable metaphorical knowledge is encoded in PLM representations or not.* The outline of our work is presented in Figure 1.

We first do *probing* experiments to answer questions such as: (i) with which accuracies and extractabilities do different PLMs encode metaphorical knowledge? (ii) how deep is the metaphorical knowledge encoded in PLM multi-layer representations? We take two probing methods, edge probing (Tenney et al., 2019b) and minimum description length (Voita and Titov, 2020), and apply them to four metaphor detection datasets, namely LCC (Mohler et al., 2016), TroFi (Birke and Sarkar, 2006), VUA pos, and VUA Verbs (Steen, 2010).

* Equal contribution.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2037 - 2050
May 22-27, 2022 ©2022 Association for Computational Linguistics

GlobEnc: Quantifying Global Token Attribution by Incorporating the Whole Encoder Layer in Transformers

Ali Modarressi[1]* Mohsen Fayyaz[2]*
Yadollah Yaghoobzadeh[2] Mohammad Taher Pilehvar[3]
[1] Iran University of Science and Technology, Iran [2] University of Tehran, Iran
[3] Tehran Institute for Advanced Studies, Khatam University, Iran
m_modarressi@comp.iust.ac.ir
{mohsen.fayyaz77, y.yaghoobzadeh}@ut.ac.ir
mp792@cam.ac.uk

Abstract

There has been a growing interest in interpreting the underlying dynamics of Transformers. While self-attention patterns were initially deemed as the primary option, recent studies have shown that integrating other components can yield more accurate explanations. This paper introduces a novel token attribution analysis method that incorporates all the components in the encoder block and aggregates this throughout layers. Through extensive quantitative and qualitative experiments, we demonstrate that our method can produce faithful and meaningful global token attributions. Our experiments reveal that incorporating almost every encoder component results in increasingly more accurate analysis in both local (single layer) and global (the whole model) settings. Our global attribution analysis significantly outperforms previous methods on various tasks regarding correlation with gradient-based saliency scores. Our code is freely available at https://github.com/mohsenfayyaz/GlobEnc.

Figure 1: Aggregated attribution maps ($\tilde{A}_{\text{Enc}}$) for the [CLS] token for fine-tuned BERT on SST2 dataset (sentiment analysis). Our method (GlobEnc) is able to accurately quantify the global token attribution of the model.

1 Introduction

The stellar performance of Transformers (Vaswani et al., 2017) has garnered a lot of attention to analyzing the reasons behind their effectiveness. The self-attention mechanism has been one of the main areas of focus (Clark et al., 2019; Kovaleva et al., 2019; Reif et al., 2019; Htut et al., 2019). However, there have been debates on whether raw attention weights are reliable anchors for explaining model's behavior or not (Wiegreffe and Pinter, 2019; Serrano and Smith, 2019; Jain and Wallace, 2019). Recently, it was shown that incorporating vector norms should be an indispensable part of any attention-based analysis[1] (Kobayashi et al., 2020, 2021). However, these norm-based studies incorporate only the attention block into their analysis, whereas Transformer encoder layer is composed of more components.

Another limitation of the existing analysis techniques is that they are usually constrained to the analysis of single layer attributions. In order to expand the analysis to multi-layered encoder-based models in their entirety, an aggregation technique has to be employed. Abnar and Zuidema (2020) proposed two aggregation methods, *rollout* and *max-flow*, which combine raw attention weights across layers. Despite showing the outcome of their method to be faithful to a model's inner workings in specific cases, the final results are still unsatisfactory on a wide range of fine-tuned models.

Additionally, gradient-based alternatives (Simonyan et al., 2014; Kindermans et al., 2016; Li et al., 2016) have been argued to provide a more robust basis for token attribution analysis (Atanasova et al., 2020; Brunner et al., 2020; Pascual et al., 2021). Nonetheless, the gradient-based alternatives have not been able to fully replace attention-based counterparts, mainly due to their high computa-

* Equal contribution.

[1]We also have shown the unreliability of weights due to norm disparities in probing studies (Fayyaz et al., 2021).

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 258 - 271
July 10-15, 2022 ©2022 Association for Computational Linguistics

DecompX: Explaining Transformers Decisions by Propagating Token Decomposition

Ali Modarressi[1,2]* Mohsen Fayyaz[3]* Ehsan Aghazadeh[3]
Yadollah Yaghoobzadeh[3,4] Mohammad Taher Pilehvar[4]
[1] Center for Information and Language Processing, LMU Munich, Germany
[2] Munich Center for Machine Learning (MCML), Germany [3] University of Tehran, Iran
[4] Tehran Institute for Advanced Studies, Khatam University, Iran
amodaresi@cis.lmu.de mohsen.fayyaz77@ut.ac.ir eaghazade1998@ut.ac.ir
y.yaghoobzadeh@ut.ac.ir mp792@cam.ac.uk

Abstract

An emerging solution for explaining Transformer-based models is to use vector-based analysis on how the representations are formed. However, providing a faithful vector-based explanation for a multi-layer model could be challenging in three aspects: (1) Incorporating all components into the analysis, (2) Aggregating the layer dynamics to determine the information flow and mixture throughout the entire model, and (3) Identifying the connection between the vector-based analysis and the model's predictions. In this paper, we present *DecompX* to tackle these challenges. DecompX is based on the construction of decomposed token representations and their successive propagation throughout the model without mixing them in between layers. Additionally, our proposal provides multiple advantages over existing solutions for its inclusion of all encoder components (especially nonlinear feed-forward networks) and the classification head. The former allows acquiring precise vectors while the latter transforms the decomposition into meaningful prediction-based values, eliminating the need for norm- or summation-based vector aggregation. According to the standard faithfulness evaluations, DecompX consistently outperforms existing gradient-based and vector-based approaches on various datasets. Our code is available at github.com/mohsenfayyaz/DecompX.

Figure 1: The explanation of our method (DecompX) compared with GlobEnc and ALTI for fine-tuned BERT on SST2 dataset (sentiment analysis). Our method is able to quantify positive or negative attribution of each token as well as being more accurate.

1 Introduction

While Transformer-based models have demonstrated significant performance, their black-box nature necessitates the development of explanation methods for understanding these models' decisions (Serrano and Smith, 2019; Bastings and Filippova, 2020; Lyu et al., 2022). On the one hand, researchers have adapted *gradient-based* methods from computer vision to NLP (Li et al., 2016; Wu and Ong, 2021). On the other hand, many have attempted to explain the decisions based on the components inside the Transformers architecture (*vector-based* methods). Recently, the latter has shown to be more promising than the former in terms of faithfulness (Ferrando et al., 2022).

Therefore, we focus on the vector-based methods which require an accurate estimation of (i) the mixture of tokens in each layer (*local-level* analysis), and (ii) the flow of attention throughout multiple layers (*global-level* analysis) (Pascual et al., 2021). Some of the existing local analysis methods include raw attention weights (Clark et al., 2019), effective attentions (Brunner et al., 2020), and vector norms (Kobayashi et al., 2020, 2021), which all attempt to explain how a single layer combines its input representations. Besides, to compute the global impact of the inputs on the outputs, the local behavior of all layers must be aggregated. *Attention rollout* and *attention flow* were the initial approaches for recursively aggregating the raw attention maps in each layer (Abnar and Zuidema, 2020). By employing rollout, GlobEnc (Modarressi et al., 2022) and ALTI (Ferrando et al., 2022) significantly improved

* Equal contribution.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2649–2664
July 9-14, 2023 ©2023 Association for Computational Linguistics

Metaphors in Pre-Trained Language Models: Probing and Generalization Across Datasets and Languages

ACL 2022

Metaphors in Pre-Trained Language Models: Probing and Generalization Across Datasets and Languages

Ehsan Aghazadeh* Mohsen Fayyaz* Yadollah Yaghoobzadeh
School of Electrical and Computer Engineering,
College of Engineering,
University of Tehran, Tehran, Iran
{eaghazade1998, mohsen.fayyaz77, y.yaghoobzadeh}@ut.ac.ir

Abstract

Human languages are full of metaphorical expressions. Metaphors help people understand the world by connecting new concepts and domains to more familiar ones. Large pre-trained language models (PLMs) are therefore assumed to encode metaphorical knowledge useful for NLP systems. In this paper, we investigate this hypothesis for PLMs, by probing metaphoricity information in their encodings, and by measuring the cross-lingual and cross-dataset generalization of this information. We present studies in multiple metaphor detection datasets and in four languages (i.e., English, Spanish, Russian, and Farsi). Our extensive experiments suggest that contextual representations in PLMs do encode metaphorical knowledge, and mostly in their middle layers. The knowledge is transferable between languages and datasets, especially when the annotation is consistent across training and testing sets. Our findings give helpful insights for both cognitive and NLP scientists.

1 Introduction

Pre-trained language models (PLMs) (Peters et al., 2018; Devlin et al., 2019), are now used in almost all NLP applications, e.g., machine translation (Li et al., 2021), question answering (Zhang et al., 2020), dialogue systems (Ni et al., 2021), and sentiment analysis (Minaee et al., 2020). They have sometimes been referred to as "foundation models" (Bommasani et al., 2021) due to their significant impact on research and industry.

Metaphors are important aspects of human languages. In conceptual metaphor theory (CMT) (Lakoff and Johnson, 2008), metaphor is defined as a cognitive phenomenon associating two different concepts or domains. This phenomenon is built in cognition and expressed in language. The creativity and problem solving (i.e., generalization to new problems) depend on the analogies and metaphors a cognitive system, like our brain, relies on. Modeling metaphors is therefore essential in building human-like computational systems that can relate emerging concepts to the more familiar ones.

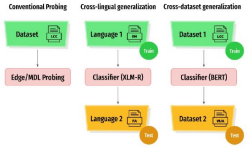


Figure 1: An illustration of our probing and generalization scenarios for metaphorical knowledge.

So far, there has been no comprehensive analysis of whether and how PLMs represent metaphorical information. We intuitively assume that PLMs must encode some information about metaphors due to their great performance in metaphor detection and other language processing tasks. Confirming that experimentally is a question that we address here. Specifically, we aim to know *whether generalizable metaphorical knowledge is encoded in PLM representations or not.* The outline of our work is presented in Figure 1.

We first do *probing* experiments to answer questions such as: (i) with which accuracies and extractablities do different PLMs encode metaphorical knowledge? (ii) how deep is the metaphorical knowledge encoded in PLM multi-layer representations? We take two probing methods, edge probing (Tenney et al., 2019b) and minimum description length (Voita and Titov, 2020), and apply them to four metaphor detection datasets, namely LCC (Mohler et al., 2016), TroFi (Birke and Sarkar, 2006), VUA pos, and VUA Verbs (Steen, 2010).

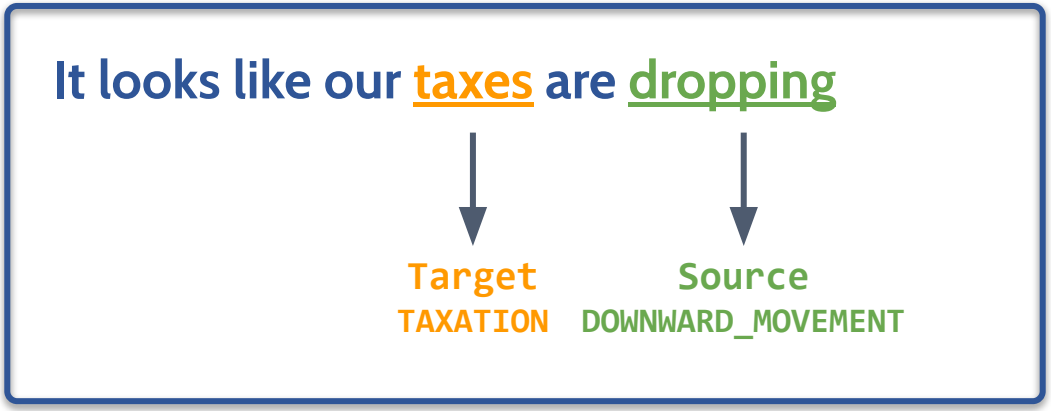
* Equal contribution.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2037 - 2050
May 22-27, 2022

۱) بررسی دانش استعاره

تعریف استعاره

- بر اساس Conceptual Metaphor Theory



[1] George Lakoff and Mark Johnson. 2008. Metaphors we live by. University of Chicago press.

۱) بررسی دانش استعاره

انگیزه بررسی استعاره

- دانش‌های زبانی پایه قبلا بررسی شده‌اند.
- استعاره‌ها در ارتباطات انسانی و ساختن سیستم‌های محاسباتی شبیه انسان ضروری هستند.

آیا مدل های زبانی پیش‌آموزش دیده استعاره‌ها را کدگذاری می‌کنند؟

۱) بررسی دانش استعاره

مجموعه‌های داده استعاره

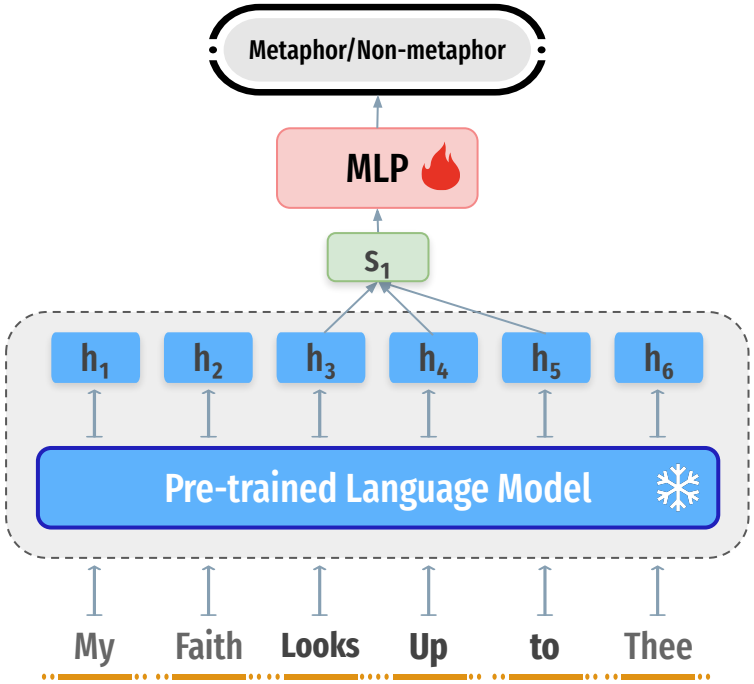
VUA Verbs	He **[finds]**$_1$ it hard to communicate with people , not least his separated parents .→ 1 He finds it hard to **[communicate]**$_1$ with people , not least his separated parents . → 0
VUA POS	They picked up power from a **[spider]**$_1$'s web of unsightly overhead wires . → 1 They picked up power from a spider 's web of unsightly overhead **[wires]**$_1$. → 0
TroFi	" Locals **[absorbed]**$_1$ a lot of losses , " said Mr. Sandor of Drexel → nonliteral Vitamins could be passed right out of the body without being **[absorbed]**$_1$ → literal
LCC	Lawful gun ownership is not a **[disease]**$_1$. → 3.0 But the Supreme Court says it's not a way to **[hurt]**$_1$ the Second Amendment → 2.0 Is he angry that gun rights **[progress]**$_1$ has been done without him? → 1.0 I mean the 2nd amendment **[suggests]**$_1$ a level playing field for all of us. → 0.0

- ★ برچسب‌های مثبت و منفی در تمام مجموعه‌ها برابرسازی شده‌اند.
- ★ مجموعه داده LCC چهار زبان انگلیسی، اسپانیایی، روسی و فارسی را شامل می‌شود.

۱) بررسی دانش استعاره

کاوند

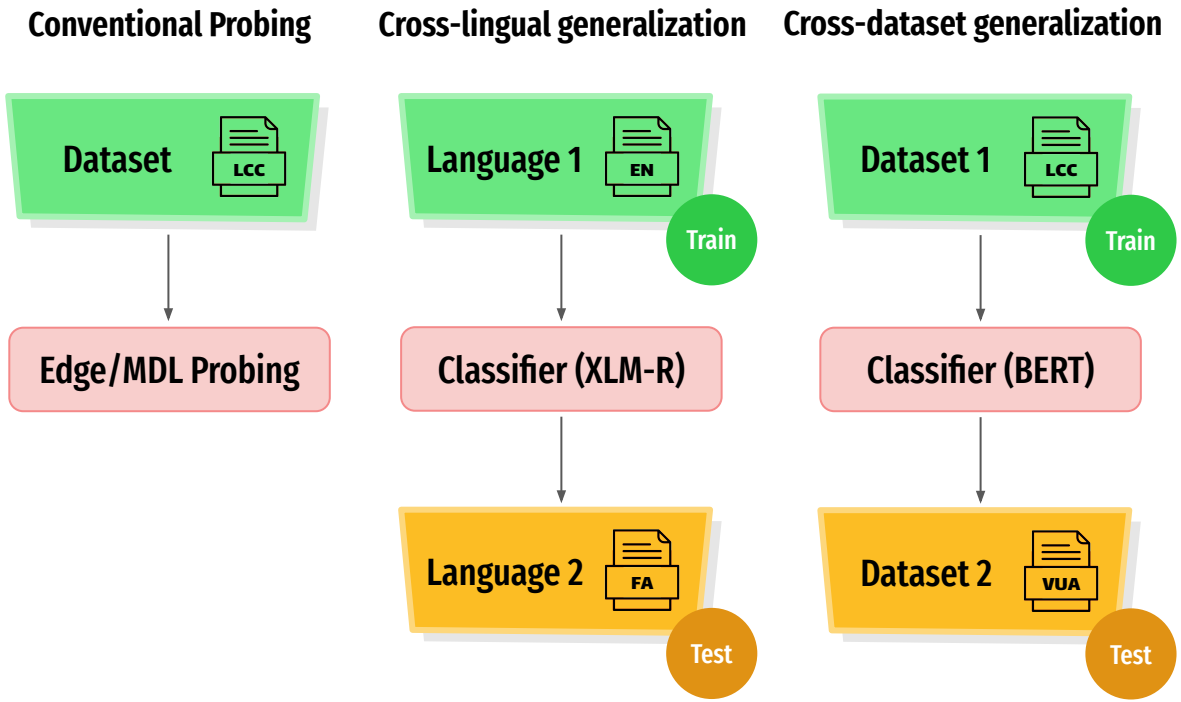
- ارزیابی دانش زبانی در بازنمایی‌های شبکه عصبی



[1] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? Probing for sentence structure in contextualized word representations. In International Conference on Learning Representations.

۱) بررسی دانش استعاره

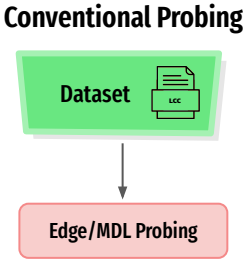
روش‌شناسی



۱) بررسی دانش استعاره

نتایج

Dataset	Baseline		BERT		RoBERTa		ELECTRA	
	Acc.	Comp.	Acc.	Comp.	Acc.	Comp.	Acc.	Comp.
LCC (en)	74.86	1.05_{2}	88.25	1.85_{6}	88.06	1.96_{5}	**89.30**	**2.05_{5}**
TroFi	67.34	1.01_{4}	**68.58**	1.07_{4}	68.46	**1.09_{6}**	68.07	1.08_{3}
VUA POS	65.92	1.03_{0}	80.32	1.43_{5}	81.72	1.48_{6}	**83.03**	**1.51_{4}**
VUA Verbs	65.97	1.04_{9}	78.29	1.28_{9}	78.88	**1.34_{5}**	**79.96**	1.31_{4}

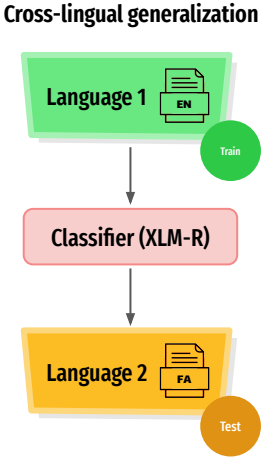


- مدل های زبانی پیش‌آموزش دیده استعاره‌ها را کدگذاری می‌کنند
- RoBERTa و ELECTRA بهتر است از BERT
  - اهداف پیش‌آموزش بهتر
  - داده‌های بیشتر پیش‌آموزش

۱) بررسی دانش استعاره

نتایج تعمیم بین زبانی

Test Lang	Train Lang: en	es	fa	ru
en	**85.14** (65.37)	79.31 (52.71)	77.59 (50.22)	80.51 (52.40)
es	79.40 (53.17)	**84.59** (66.09)	76.70 (50.32)	79.68 (53.32)
fa	75.70 (50.07)	75.29 (52.65)	**81.04** (65.91)	77.14 (50.36)
ru	83.92 (53.25)	80.54 (51.48)	76.61 (51.05)	**88.36** (67.98)

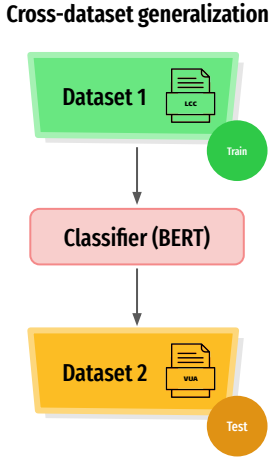


- نتایج مدل XLM-R خیلی بهتر از (مدل تصادفی) است
  - قابلیت انتقال اطلاعات استعاری بین زبان‌ها
  - ظرفیت بازنمایی‌های چندزبانه XLM-R

۱) بررسی دانش استعاره

نتایج تعمیم بین مجموعه‌ها

Test Dataset	Train Dataset: LCC(en)	TroFi	VUA POS	VUA Verbs
LCC(en)	**84.26** (54.93)	62.04 (50.05)	70.35 (50.69)	70.37 (50.14)
TroFi	59.49 (50.58)	**68.73** (64.96)	55.38 (49.45)	59.67 (53.68)
VUA POS	62.23 (51.47)	55.29 (50.47)	**76.86** (56.01)	71.6 (53.47)
VUA Verbs	60.20 (50.88)	54.55 (51.73)	72.6 (56.01)	**75.21** (60.03)



- نتایج مدل آموزش‌دیده بهتر از مدل تصادفی است
  - اطلاعات استعاری قابل تعمیم بین مجموعه‌ها است.
- شیوه‌نامه و دادگان مشابه ← نتایج بهتر

۱) بررسی دانش استعاره

نتیجه‌گیری

- بازنمایی‌های متنی در مدل‌های پیش‌آموزش‌دیده دانش استعاری را کدگذاری می‌کنند
- دانش استعاری بین زبان‌ها و مجموعه داده‌ها قابل انتقال است (متناسب با سازگاری شیوه‌نامه‌ها)

GlobEnc: Quantifying Global Token Attribution by Incorporating the Whole Encoder Layer in Transformers

NAACL 2022

GlobEnc: Quantifying Global Token Attribution by Incorporating the Whole Encoder Layer in Transformers

Ali Modarressi[1*] **Mohsen Fayyaz**[2*]
Yadollah Yaghoobzadeh[2] **Mohammad Taher Pilehvar**[3]
[1] Iran University of Science and Technology, Iran [2] University of Tehran, Iran
[3] Tehran Institute for Advanced Studies, Khatam University, Iran
m_modarressi@comp.iust.ac.ir
{mohsen.fayyaz77, y.yaghoobzadeh}@ut.ac.ir
mp792@cam.ac.uk

Abstract

There has been a growing interest in interpreting the underlying dynamics of Transformers. While self-attention patterns were initially deemed as the primary option, recent studies have shown that integrating other components can yield more accurate explanations. This paper introduces a novel token attribution analysis method that incorporates all the components in the encoder block and aggregates this throughout layers. Through extensive quantitative and qualitative experiments, we demonstrate that our method can produce faithful and meaningful global token attributions. Our experiments reveal that incorporating almost every encoder component results in increasingly more accurate analysis in both local (single layer) and global (the whole model) settings. Our global attribution analysis significantly outperforms previous methods on various tasks regarding correlation with gradient-based saliency scores. Our code is freely available at https://github.com/mohsenfayyaz/GlobEnc.

Figure 1: Aggregated attribution maps ($\mathcal{N}_{\text{ENC}}$) for the [CLS] token for fine-tuned BERT on SST2 dataset (sentiment analysis). Our method (GlobEnc) is able to accurately quantify the global token attribution of the model.

1 Introduction

The stellar performance of Transformers (Vaswani et al., 2017) has garnered a lot of attention to analyzing the reasons behind their effectiveness. The self-attention mechanism has been one of the main areas of focus (Clark et al., 2019; Kovaleva et al., 2019; Reif et al., 2019; Htut et al., 2019). However, there have been debates on whether raw attention weights are reliable anchors for explaining model's behavior or not (Wiegreffe and Pinter, 2019; Serrano and Smith, 2019; Jain and Wallace, 2019). Recently, it was shown that incorporating vector norms should be an indispensable part of any attention-based analysis[1] (Kobayashi et al., 2020, 2021). However, these norm-based studies incorporate only the attention block into their analysis, whereas Transformer encoder layer is composed of more components.

Another limitation of the existing analysis techniques is that they are usually constrained to the analysis of single layer attributions. In order to expand the analysis to multi-layered encoder-based models in their entirety, an aggregation technique has to be employed. Abnar and Zuidema (2020) proposed two aggregation methods, *rollout* and *max-flow*, which combine raw attention weights across layers. Despite showing the outcome of their method to be faithful to a model's inner workings in specific cases, the final results are still unsatisfactory on a wide range of fine-tuned models.

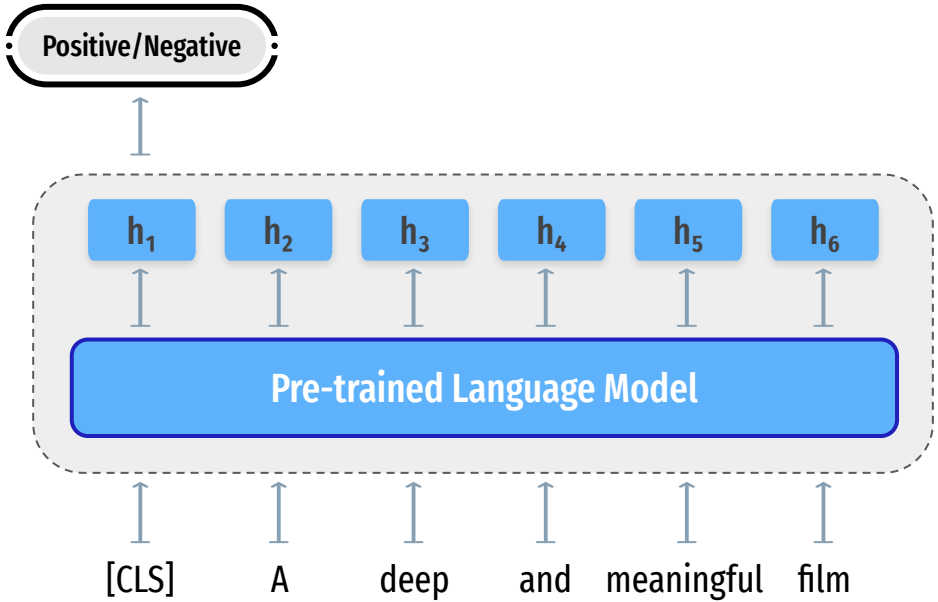
Additionally, gradient-based alternatives (Simonyan et al., 2014; Kindermans et al., 2016; Li et al., 2016) have been argued to provide a more robust basis for token attribution analysis (Atanasova et al., 2020; Brunner et al., 2020; Pascual et al., 2021). Nonetheless, the gradient-based alternatives have not been able to fully replace attention-based counterparts, mainly due to their high computa-

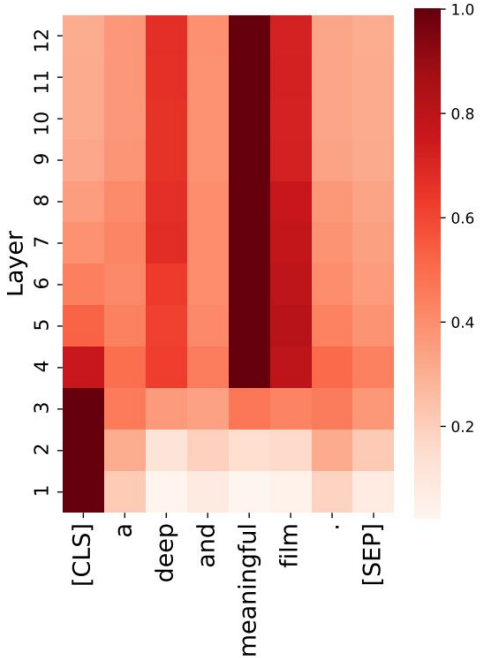
[*] Equal contribution.
[1] We also have shown the unreliability of weights due to norm disparities in probing studies (Fayyaz et al., 2021).

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 258 - 271
July 10-15, 2022 ©2022 Association for Computational Linguistics

۲) بررسی توجه با در نظر گرفتن کل لایه رمزگذار

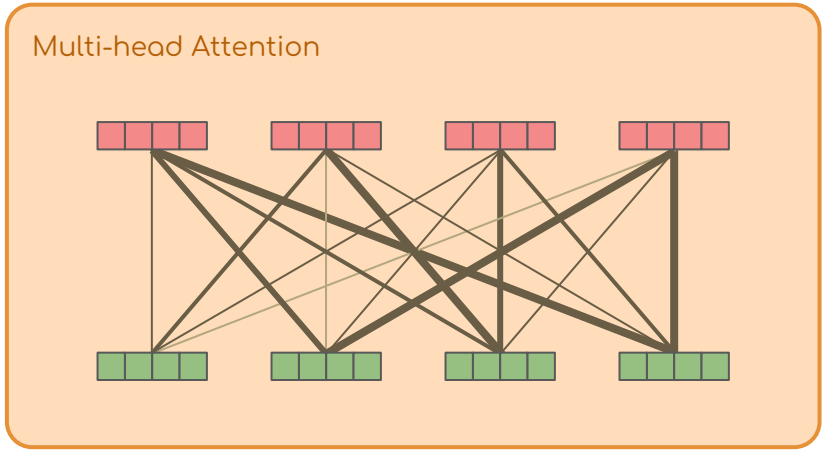
توجه مدل به ورودی‌ها





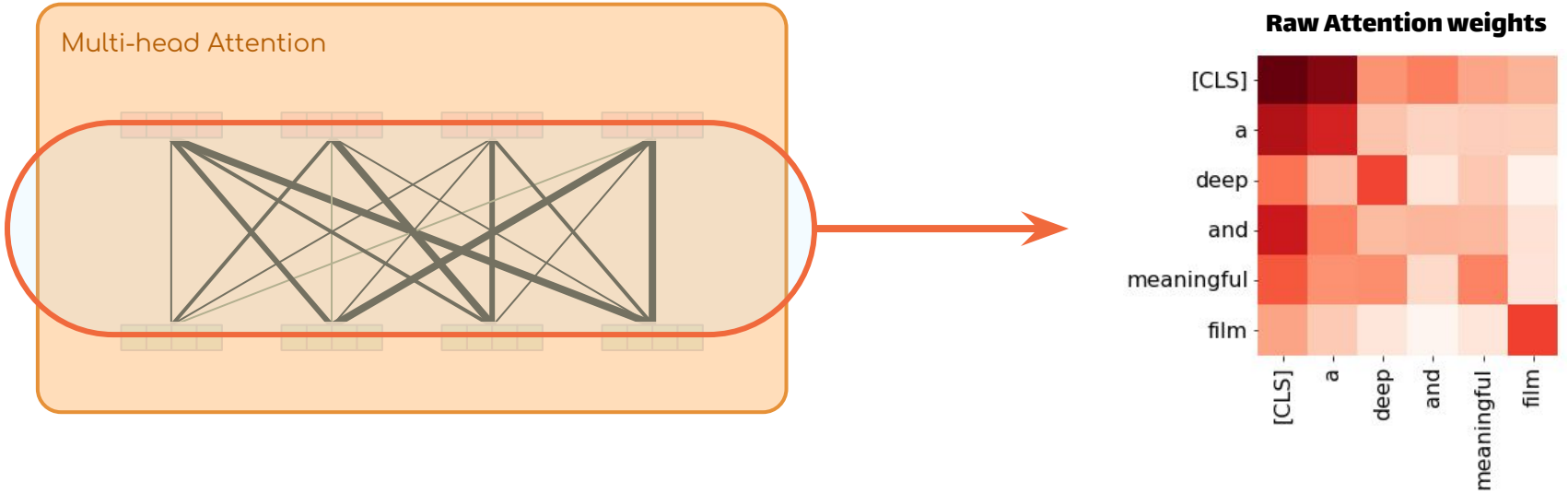
۲) بررسی توجه با در نظر گرفتن کل لایه رمزگذار

استفاده از وزن‌های خام توجه



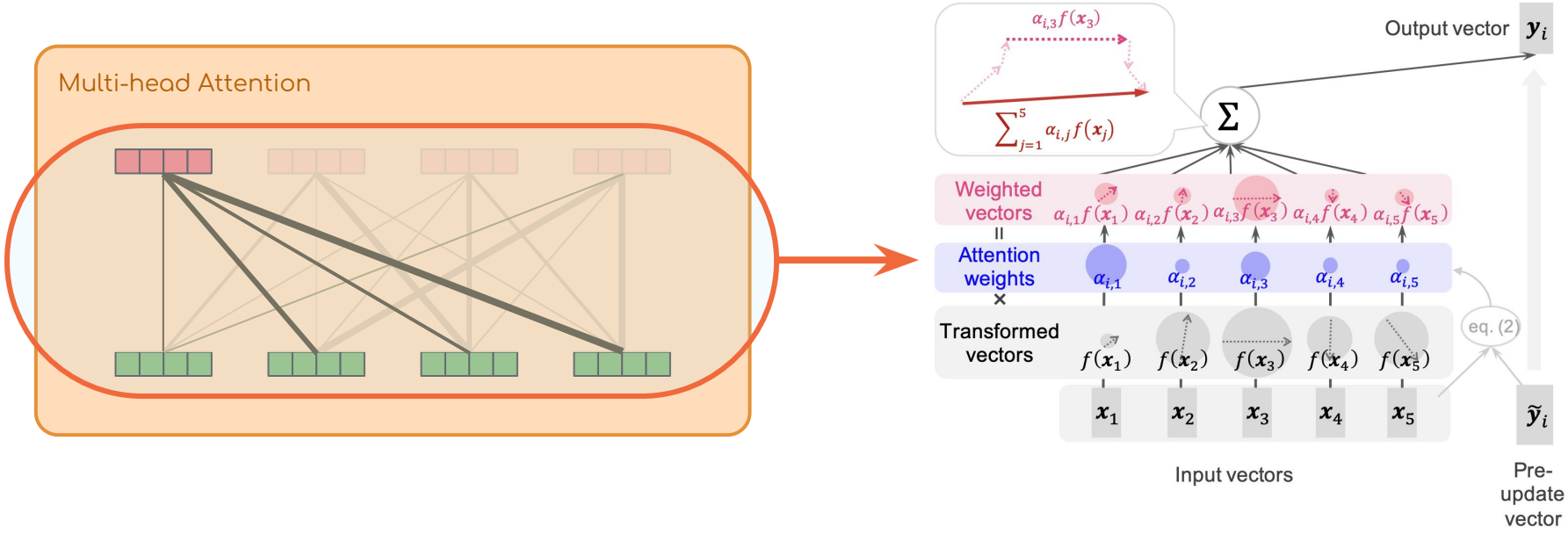
۲) بررسی توجه با در نظر گرفتن کل لایه رمزگذار

استفاده از وزن‌های خام توجه



۲) بررسی توجه با در نظر گرفتن کل لایه رمزگذار

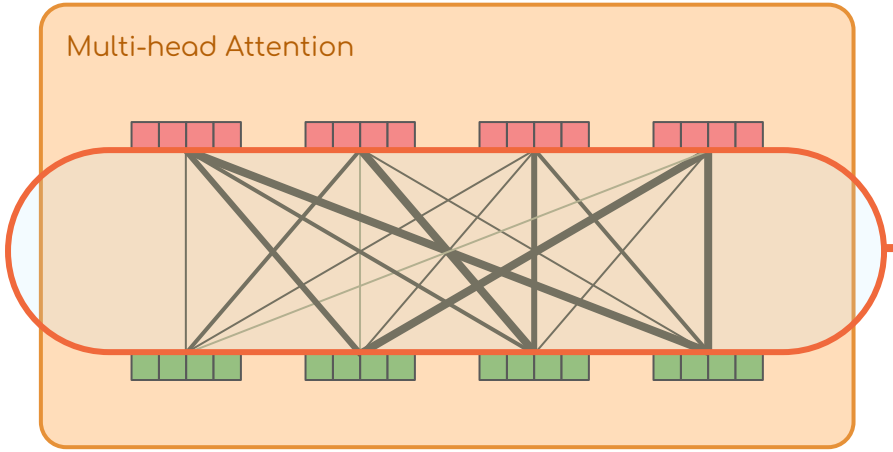
استفاده از اندازه بردارها

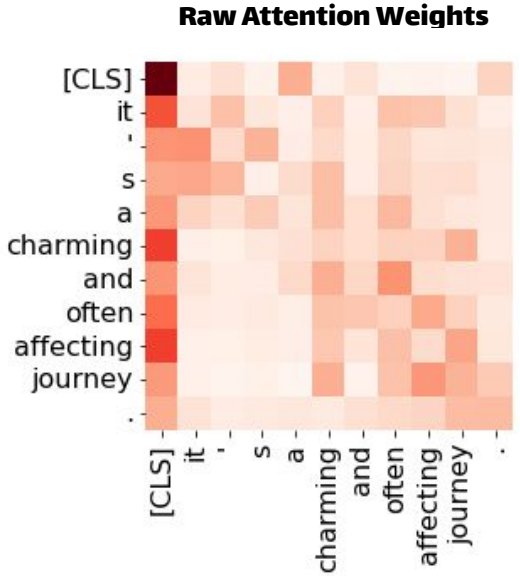


[1] Kobayashi et al,. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Pages 7057–7075, Online. Association for Computational Linguistics.

۲) بررسی توجه با در نظر گرفتن کل لایه رمزگذار

استفاده از وزن‌های خام توجه

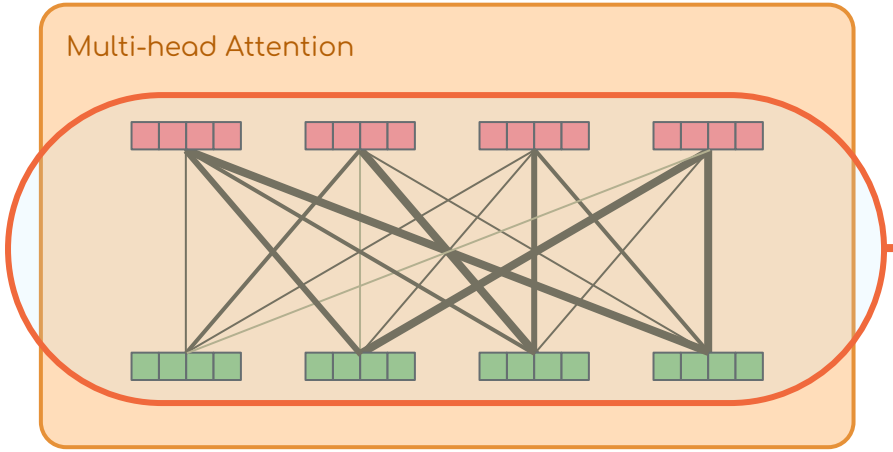


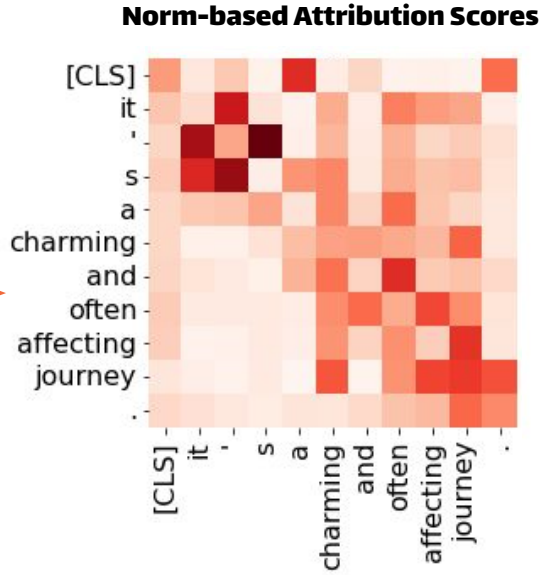


[1] Kobayashi et al,. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Pages 7057–7075, Online. Association for Computational Linguistics.

۲) بررسی توجه با در نظر گرفتن کل لایه رمزگذار

استفاده از اندازه بردارها

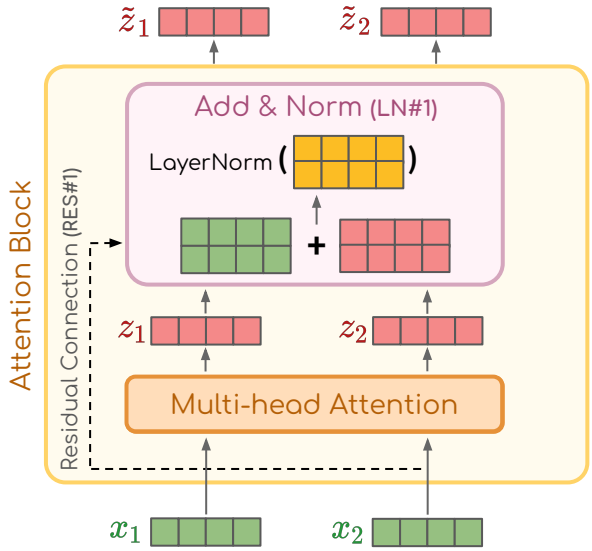


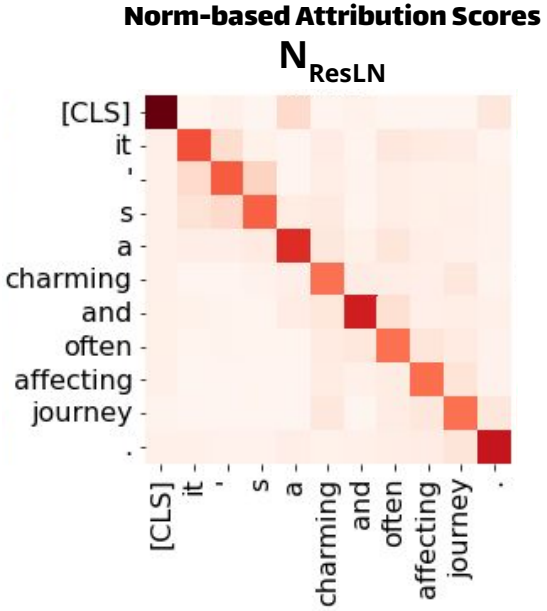


[1] Kobayashi et al,. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Pages 7057–7075, Online. Association for Computational Linguistics.

۲) بررسی توجه با در نظر گرفتن کل لایه رمزگذار

استفاده از اندازه بردارها + RES#1 + LN#1

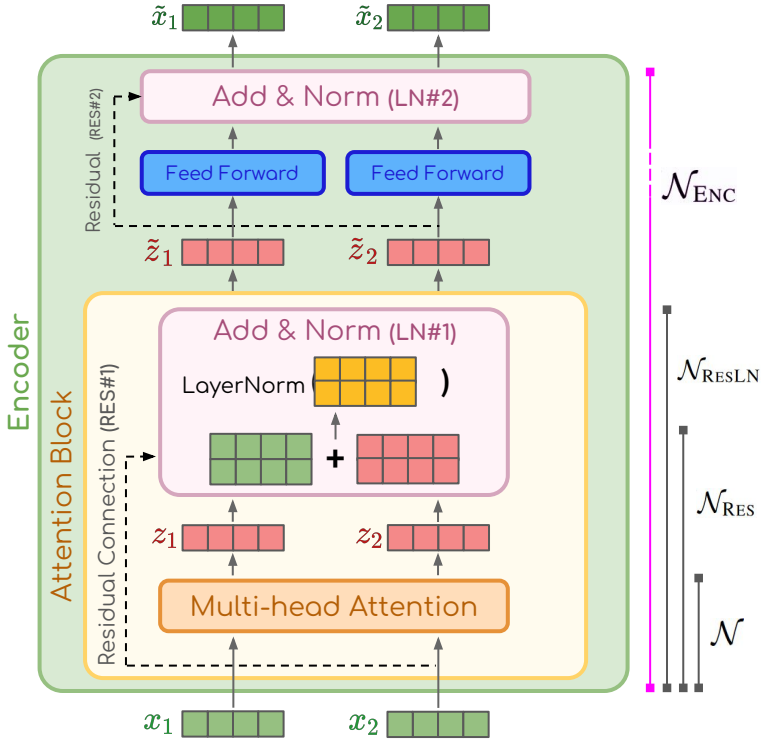




[2] Kobayashi et al,. 2021. Incorporating Residual and Normalization Layers into Analysis of Masked Language Models. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4547–4568, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

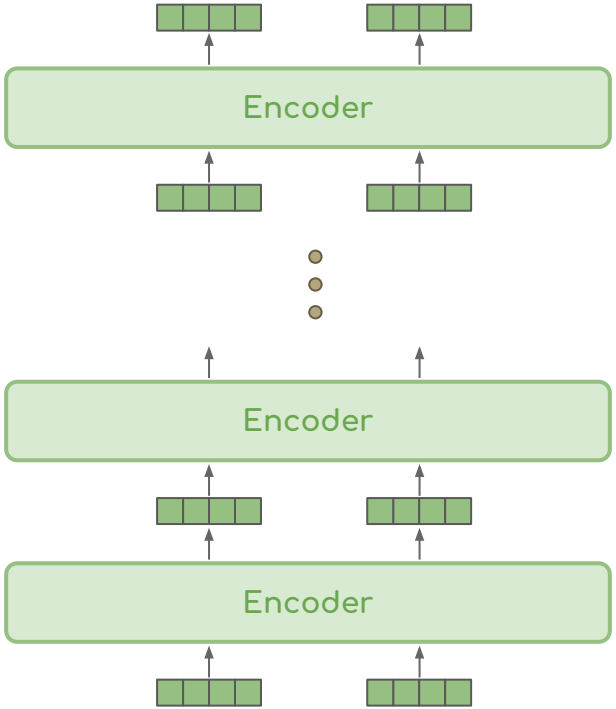
۲) بررسی توجه با در نظر گرفتن کل لایه رمزگذار

استفاده از اندازه بردارها + RES#1 + LN#1 + RES#2 + LN#2



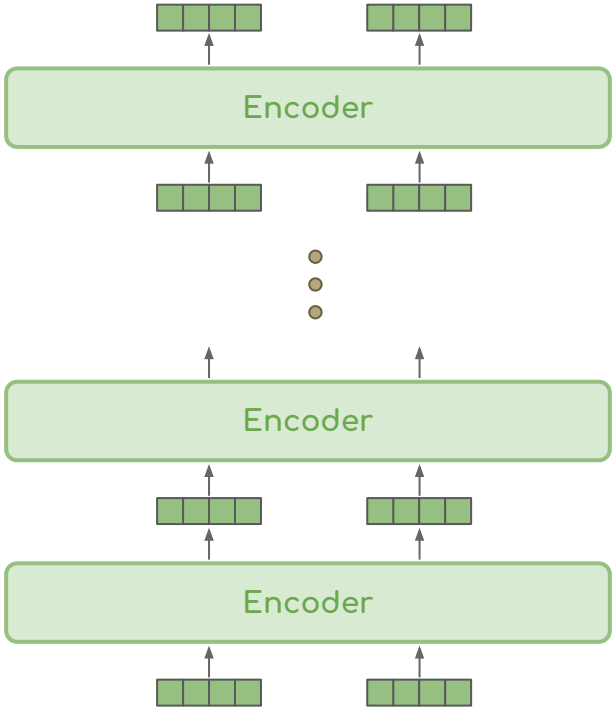
۲) بررسی توجه با در نظر گرفتن کل لایه رمزگذار

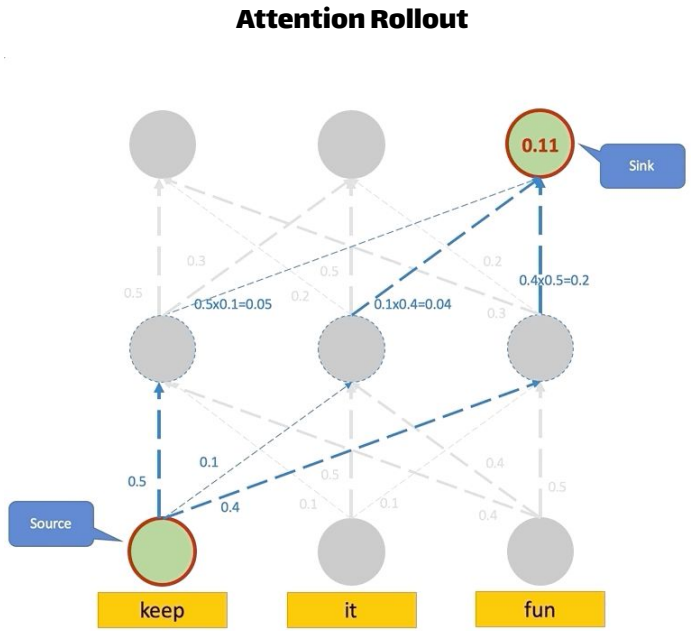
انتشار توجه در لایه‌ها



۲) بررسی توجه با در نظر گرفتن کل لایه رمزگذار

انتشار توجه در لایه‌ها



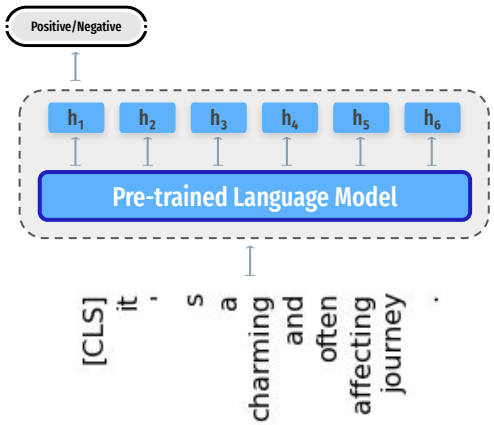


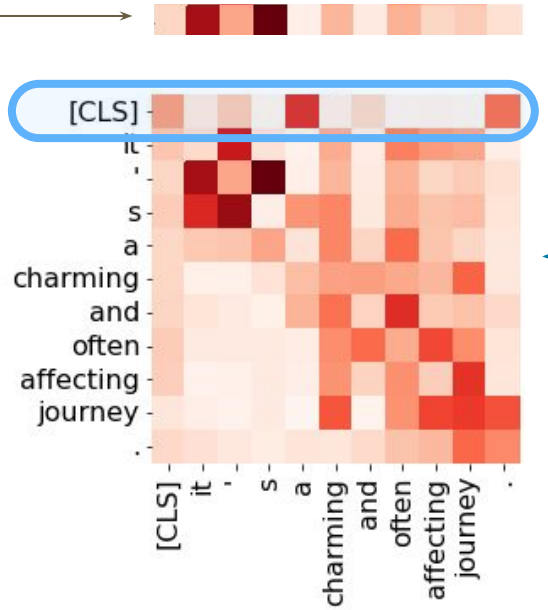
https://samiraabnar.github.io/img/flow_images/attention_rollout.gif

۲) بررسی توجه با در نظر گرفتن کل لایه رمزگذار

معیار: همبستگی با روش مبتنی بر گرادیان

$$Saliency_i = \left\| \frac{\partial y_c}{\partial e_i^0} \odot e_i^0 \right\|_2$$





Spearman Correlation

ρ

All Layers Rollout

Weight-based ($\mathcal{W}$)
w/ Fixed Residual ($\mathcal{W}_{\text{FIXEDRES}}$) [9]
w/ Residual ($\mathcal{W}_{\text{RES}}$)
Norm-based ($\mathcal{N}$)
w/ Fixed Residual ($\mathcal{N}_{\text{FIXEDRES}}$)
w/ Residual ($\mathcal{N}_{\text{RES}}$)
w/ Residual + Layer Norm 1 ($\mathcal{N}_{\text{RESLN}}$)
w/ **GlobEnc**: [Residual + Layer Norm 1, 2] ($\mathcal{N}_{\text{ENC}}$)

۲) بررسی توجه با در نظر گرفتن کل لایه رمزگذار

نتایج

	Attention Rollout		
	SST2	MNLI	HATEXPLAIN
Weight-based ($\mathcal{W}$)	−0.11 ± 0.26	−0.06 ± 0.22	0.12 ± 0.26
w/ Fixed Residual ($\mathcal{W}_{\text{FIXEDRES}}$) [9]	−0.24 ± 0.26	−0.05 ± 0.26	0.13 ± 0.28
w/ Residual ($\mathcal{W}_{\text{RES}}$)	0.19 ± 0.26	0.27 ± 0.25	0.53 ± 0.24
Norm-based ($\mathcal{N}$)	0.44 ± 0.20	0.47 ± 0.16	0.43 ± 0.22
w/ Fixed Residual ($\mathcal{N}_{\text{FIXEDRES}}$)	0.48 ± 0.20	0.55 ± 0.16	0.48 ± 0.22
w/ Residual ($\mathcal{N}_{\text{RES}}$)	0.73 ± 0.13	0.75 ± 0.10	0.66 ± 0.17
w/ Residual + Layer Norm 1 ($\mathcal{N}_{\text{RESLN}}$)	−0.21 ± 0.26	−0.06 ± 0.26	0.08 ± 0.28
w/ **GlobEnc**: [Residual + Layer Norm 1, 2] ($\mathcal{N}_{\text{ENC}}$)	**0.77 ± 0.12**	**0.78 ± 0.09**	**0.72 ± 0.17**

نتیجه‌گیری

- در نظر گرفتن RES#2 و LN#2، باعث شناسایی دقیق‌تر توجه می‌شود.
- انتشار توجه در لایه‌ها با استفاده از روش Rollout.
- دو ماژول LN#1 و LN#2 با یکدیگر مقابله می کنند.

DecompX: Explaining Transformers Decisions by Propagating Token Decomposition

ACL 2023

DecompX: Explaining Transformers Decisions by Propagating Token Decomposition

Ali Modarressi[1,2*] **Mohsen Fayyaz**[3*] **Ehsan Aghazadeh**[3]
Yadollah Yaghoobzadeh[3,4] **Mohammad Taher Pilehvar**[4]
[1] Center for Information and Language Processing, LMU Munich, Germany
[2] Munich Center for Machine Learning (MCML), Germany [3] University of Tehran, Iran
[4] Tehran Institute for Advanced Studies, Khatam University, Iran
amodaresi@cis.lmu.de mohsen.fayyaz77@ut.ac.ir eaghazade1998@ut.ac.ir
y.yaghoobzadeh@ut.ac.ir mp792@cam.ac.uk

Abstract

An emerging solution for explaining Transformer-based models is to use vector-based analysis on how the representations are formed. However, providing a faithful vector-based explanation for a multi-layer model could be challenging in three aspects: (1) Incorporating all components into the analysis, (2) Aggregating the layer dynamics to determine the information flow and mixture throughout the entire model, and (3) Identifying the connection between the vector-based analysis and the model's predictions. In this paper, we present *DecompX* to tackle these challenges. DecompX is based on the construction of decomposed token representations and their successive propagation throughout the model without mixing them in between layers. Additionally, our proposal provides multiple advantages over existing solutions for its inclusion of all encoder components (especially nonlinear feed-forward networks) and the classification head. The former allows acquiring precise vectors while the latter transforms the decomposition into meaningful prediction-based values, eliminating the need for norm- or summation-based vector aggregation. According to the standard faithfulness evaluations, DecompX consistently outperforms existing gradient-based and vector-based approaches on various datasets. Our code is available at github.com/mohsenfayyaz/DecompX.

Figure 1: The explanation of our method (DecompX) compared with GlobEnc and ALTI for fine-tuned BERT on SST2 dataset (sentiment analysis). Our method is able to quantify positive or negative attribution of each token as well as being more accurate.

1 Introduction

While Transformer-based models have demonstrated significant performance, their black-box nature necessitates the development of explanation methods for understanding these models' decisions (Serrano and Smith, 2019; Bastings and Filippova, 2020; Lyu et al., 2022). On the one hand, researchers have adapted *gradient-based* methods from computer vision to NLP (Li et al., 2016; Wu and Ong, 2021). On the other hand, many have attempted to explain the decisions based on the components inside the Transformers architecture (*vector-based* methods). Recently, the latter has shown to be more promising than the former in terms of faithfulness (Ferrando et al., 2022).

Therefore, we focus on the vector-based methods which require an accurate estimation of (i) the mixture of tokens in each layer (*local-level* analysis), and (ii) the flow of attention throughout multiple layers (*global-level* analysis) (Pascual et al., 2021). Some of the existing local analysis methods include raw attention weights (Clark et al., 2019), effective attentions (Brunner et al., 2020), and vector norms (Kobayashi et al., 2020, 2021), which all attempt to explain how a single layer combines its input representations. Besides, to compute the global impact of the inputs on the outputs, the local behavior of all layers must be aggregated. *Attention rollout* and *attention flow* were the initial approaches for recursively aggregating the raw attention maps in each layer (Abnar and Zuidema, 2020). By employing rollout, GlobEnc (Modarressi et al., 2022) and ALTI (Ferrando et al., 2022) significantly improved

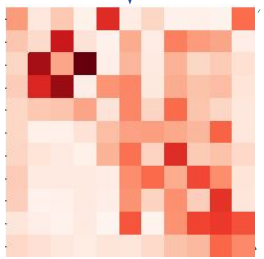
[*] Equal contribution.

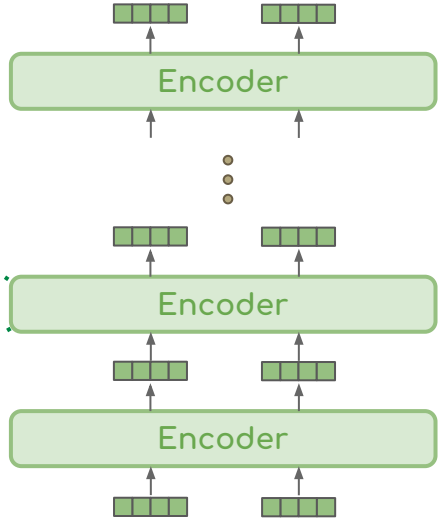
Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2649–2664
July 9-14, 2023 ©2023 Association for Computational Linguistics

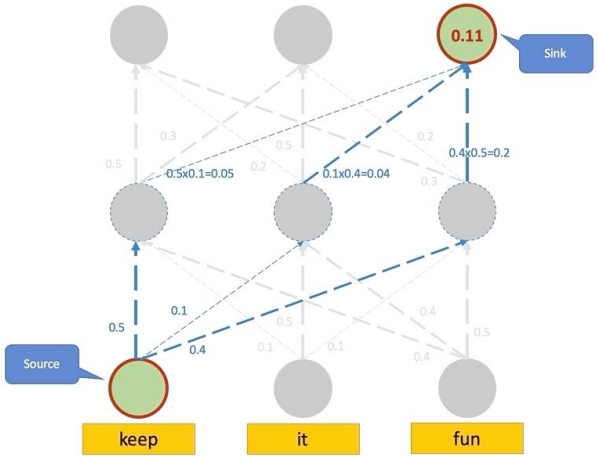
۳) بررسی توجه با انتشار تجزیه ورودی

روش‌های پیشین: نقشه توجه محلی ← انتشار توجه اسکالر

- Raw-attention[1]
- ALTI[2]
- Globenc[3]
- Value-Zeroing[4]

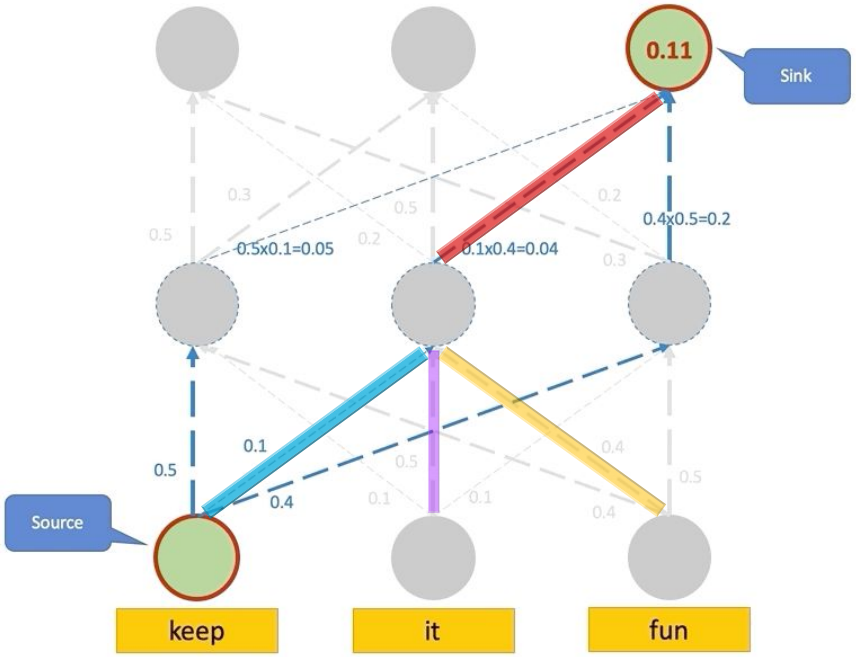






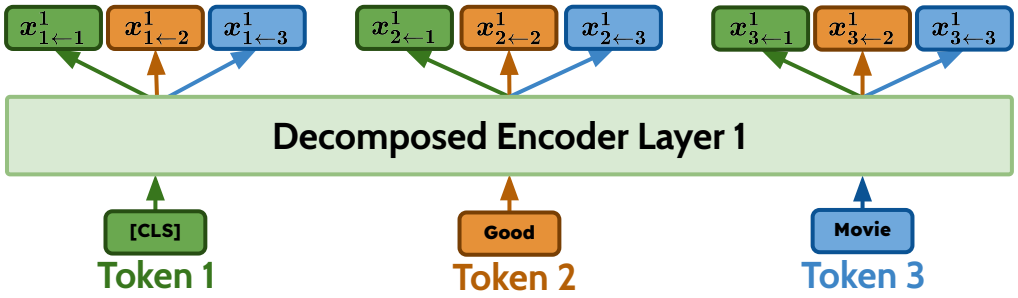
[1] Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4190–4197, Online. Association for Computational Linguistics.
[2] Javier Ferrando, Gerard I. Gállego, and Marta R. Costajussà. 2022. Measuring the mixing of contextual information in the transformer.
[3] Ali Modarressi, Mohsen Fayyaz, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. 2022. GlobEnc: Quantifying global token attribution by incorporating the whole encoder layer in transformers.
[4] Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. 2023. Quantifying context mixing in transformers. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics

روش‌های پیشین: نقشه توجه محلی ← انتشار توجه اسکالر

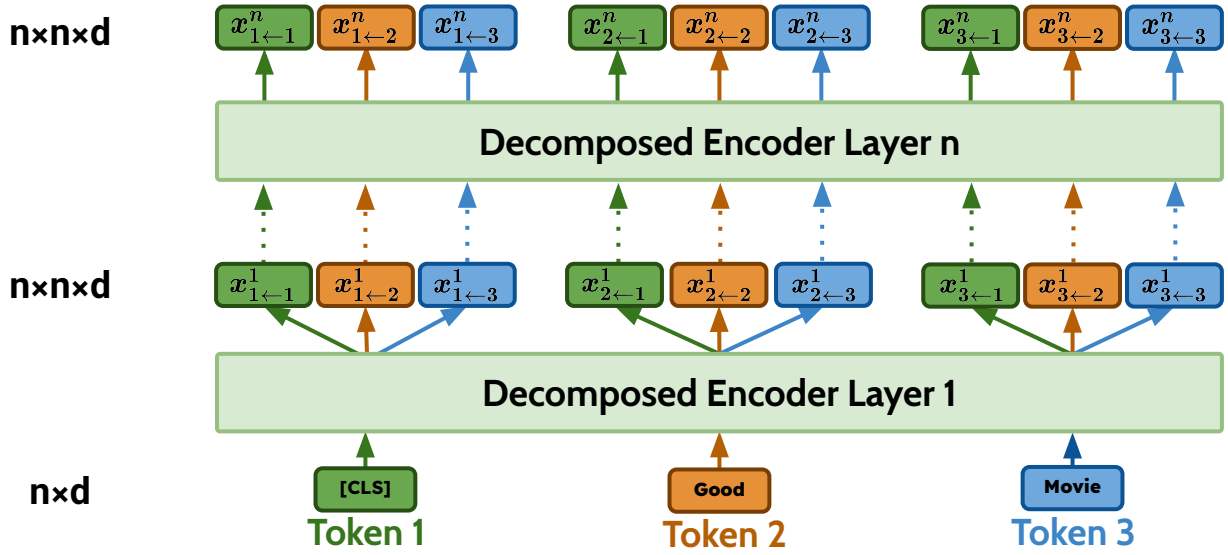


- روش‌های تجمیع توجه (مثلاً Rollout) فرض می‌کنند که تنها اطلاعات مورد نیاز برای محاسبه جریان توجه مجموعه‌ای از **اعداد اسکالر** است.
- این فرض ساده‌کننده نادیده می‌گیرد که هر بردار تجزیه شده **تأثیر چند بعدی** ورودی‌های آن را نشان می‌دهد.
- بنابراین، **از دست دادن اطلاعات** هنگام کاهش این بردارهای پیچیده به یک وزن اسکالر اجتناب ناپذیر است.

انتشار تجزیه ورودی

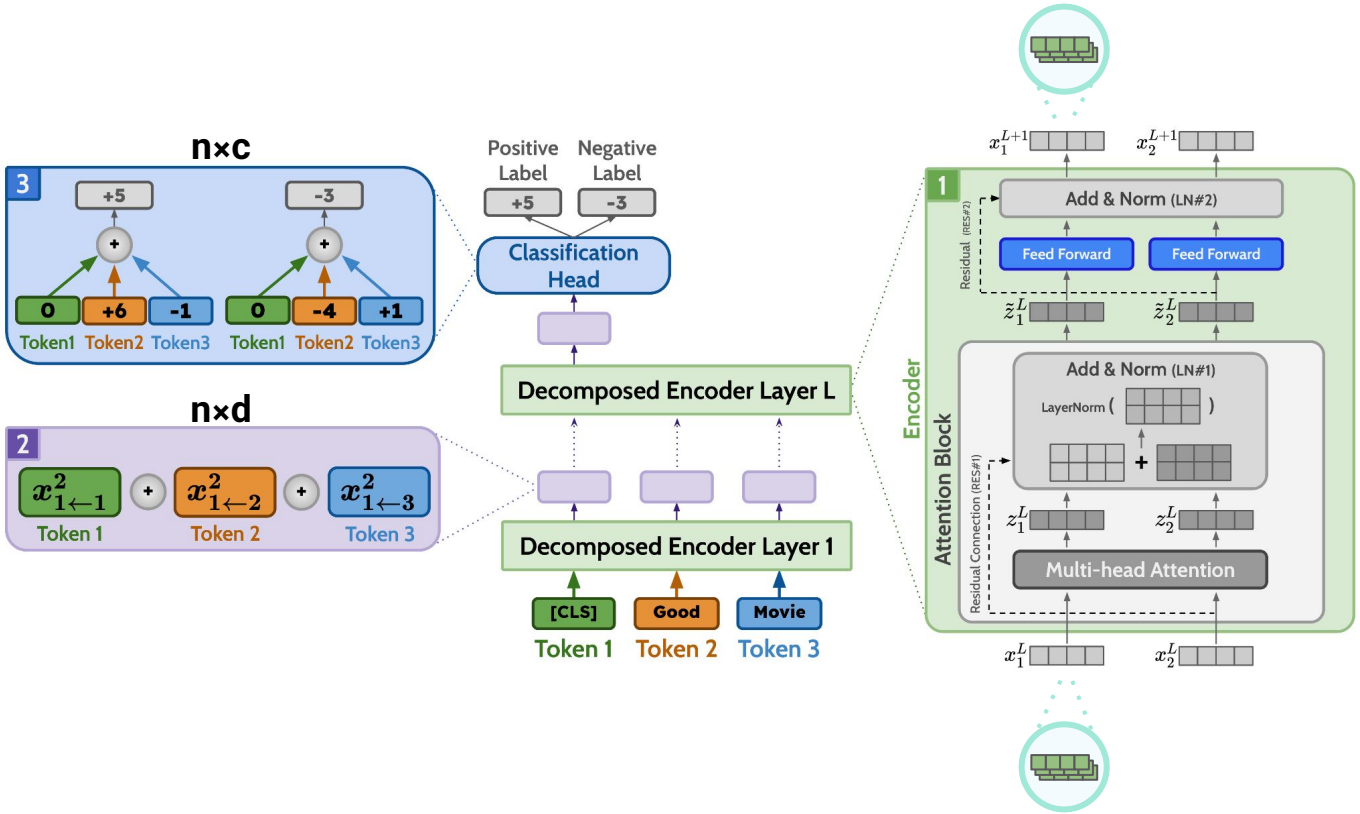


انتشار تجزیه ورودی



۳) بررسی توجه با انتشار تجزیه ورودی

DecompX



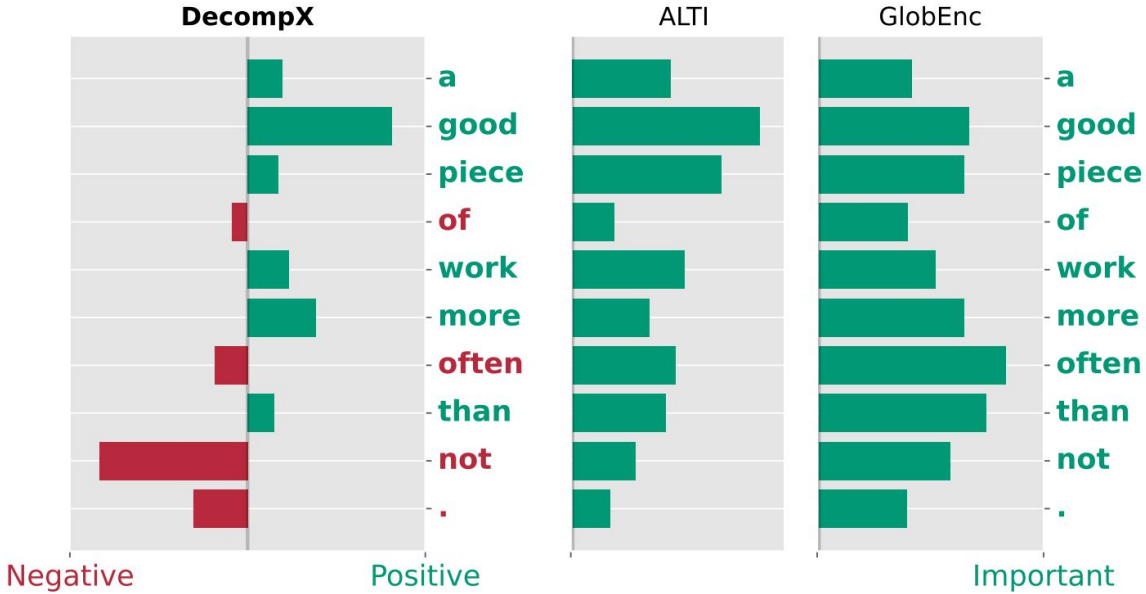
1) در نظر گرفتن همه اجزاء در لایه کدگذار مبدل، به ویژه شبکه های غیرخطی

2) انتشار بازنمایی های تجزیه شده بین لایه های مدل که از مخلوط شدن آن ها در بین لایه ها و از دست رفتن اطلاعات جلوگیری می‌کند

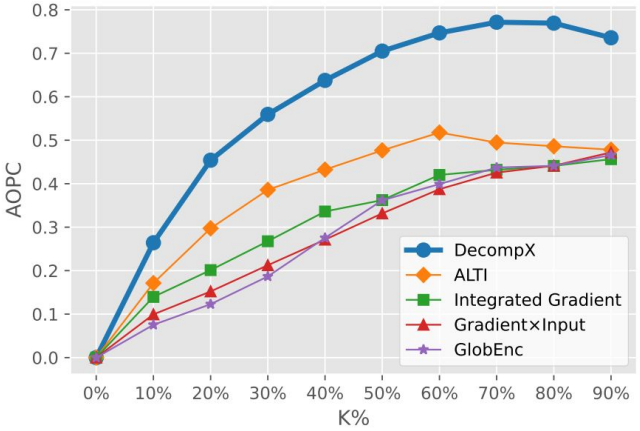
3) عبور دادن بردارهای تجزیه شده از سر طبقه‌بندی بالای مدل و در نتیجه به دست آوردن اثر مثبت یا منفی هر ورودی بر روی هر یک از کلاس های خروجی

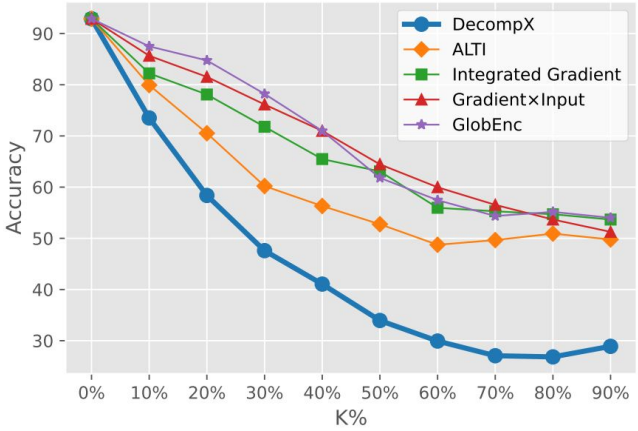
۳) بررسی توجه با انتشار تجزیه ورودی

نتایج کیفی



نتایج





$$\text{AOPC}(K) = \frac{1}{N}\sum_{i=1}^{N} p(\hat{y} \mid x_i) - p(\hat{y} \mid \tilde{x}_i^{(K)})$$

- AOPC و دقت روش‌های مختلف در SST2 با پوشاندن K% از مهم‌ترین ورودی‌ها. (AOPC بیشتر و دقت کمتر بهتر است)
  مثال: *A ~~good~~ piece of work more often than not*

- DecompX از روش‌های توضیح موجود، هم مبتنی بر بردار و هم مبتنی بر گرادیان، با یک حاشیه بزرگ در هر نسبتی بهتر عمل می‌کند.

۳) بررسی توجه با انتشار تجزیه ورودی

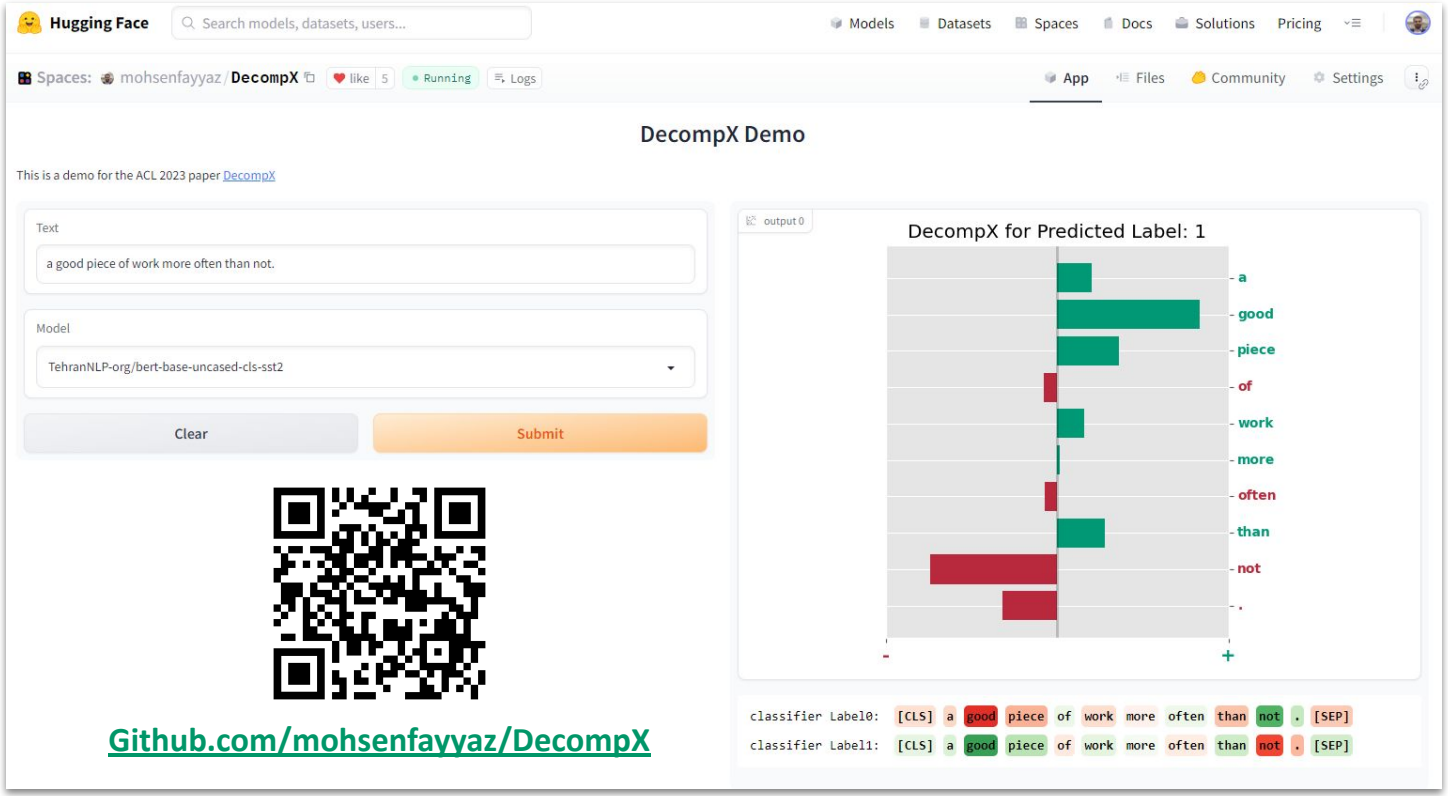
نتایج

	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**

- AOPC، Accuracy و Prediction Performance روش DecompX در مقایسه با روش‌های موجود در مجموعه داده‌های مختلف.
- DecompX همیشه از روش‌های دیگر بهتر عمل می‌کند. نتایج تأیید می‌کند که رویکردی مبتنی بر بردار و کل‌نگر می‌تواند توضیحات با کیفیت بالاتری ارائه دهد.

۳) بررسی توجه با انتشار تجزیه ورودی

دموی آنلاین



Github.com/mohsenfayyaz/DecompX

پیشنهادها

کارهای آینده

- بررسی عمیق‌تر استعاره‌ها و تأثیر شباهت‌های بین فرهنگی
- بررسی استعاره‌ها در مدل‌های مولد و کنترل تولید آن
- تفسیر توجه مدل‌ها در طول پیش‌آموزش و آموزش مدل و شناخت بایاس‌های توجه
- پیاده‌سازی روش DecompX برای معماری‌های مولد، کدگذار-کدگشا، ViT

THANK YOU!

mohsen.fayyaz77@ut.ac.ir
y.yaghoobzadeh@ut.ac.ir

با تشکر از داوران گرامی، اساتید و همکاران در این مقالات

تابستان ۱۴۰۲