Measuring the Mixing of Contextual Information in the Transformer

Javier Ferrando, Gerard I. Gállego and **Marta R. Costa-jussà**
TALP Research Center, Universitat Politècnica de Catalunya
{javier.ferrando.monsonis,gerard.ion.gallego,marta.ruiz}@upc.edu

Abstract

The Transformer architecture aggregates input information through the self-attention mechanism, but there is no clear understanding of how this information is mixed across the entire model. Additionally, recent works have demonstrated that attention weights alone are not enough to describe the flow of information. In this paper, we consider the whole attention block –multi-head attention, residual connection, and layer normalization– and define a metric to measure token-to-token interactions within each layer. Then, we aggregate layer-wise interpretations to provide input attribution scores for model predictions. Experimentally, we show that our method, ALTI (Aggregation of Layer-wise Token-to-token Interactions), provides more faithful explanations and increased robustness than gradient-based methods.

1 Introduction

The Transformer (Vaswani et al., 2017) has become ubiquitous in different tasks across multiple domains, becoming the architecture of choice for many NLP (Devlin et al., 2019; Brown et al., 2020) and computer vision (Dosovitskiy et al., 2021) tasks. The self-attention mechanism inside the Transformer is in charge of combining contextual information in its intermediate token representations. Attention weights offer a straightforward layer-wise interpretation, as they provide a distribution over input units, which is often presented as giving the relative importance of each input.

A prominent line of research has investigated the faithfulness of attention weights (Jain and Wallace, 2019; Serrano and Smith, 2019; Pruthi et al., 2020; Wiegreffe and Pinter, 2019; Madsen et al., 2021b) with contradictory conclusions. Some works have studied layer-wise attention patterns by analyzing standard attention (Kovaleva et al., 2019; Clark et al., 2019; Vig and Belinkov, 2019) and effective

$Grad_{\ell_2}$
went here just before a movie . the service was fast but that ' s it . i ordered the mango and shrimp quesadilla . my friend ordered nachos . the food was not good . i and my friend could not finish our food and we had stomach aches immediately .

IG_{ℓ_2}
went here just before a movie . the service was fast but that ' s it . i ordered the mango and shrimp quesadilla . my friend ordered nachos . the food was not good . i and my friend could not finish our food and we had stomach aches immediately .

ALTI
went here just before a movie . the service was fast but that ' s it . i ordered the mango and shrimp quesadilla . my friend ordered nachos . the food was not good . i and my friend could not finish our food and we had stomach aches immediately .

Table 1: Saliency maps of BERT generated by two common gradient methods and by our proposed method, ALTI, for a **negative** sentiment prediction example of Yelp dataset.

attention (Brunner et al., 2020; Sun and Marasović, 2021), but explaining the Transformer beyond attention weights needs further investigation (Lu et al., 2021).

Kobayashi et al. (2020) extended the explainability of the model by also considering the magnitude of the vectors involved in the attention mechanism, and Kobayashi et al. (2021) went as far as incorporating the layer normalization and the skip connection in their analysis. While these works have helped better understand the layer-wise behavior of the Transformer, there is a mismatch between layer-wise attention distributions and global input attributions (Pascual et al., 2021) since intermediate layers only attend to a mix of input tokens. Brunner et al. (2020) quantified the aggregation of contextual information throughout the model with a gradient attribution method. Although they found the self-attention mechanism greatly mixes the information of the model input, they were able to recover the token identity from hidden layers with high accuracy with a learned linear mapping. This phenomenon is partially explained by Kobayashi

et al. (2021) and Lu et al. (2021), who have shown the relatively small impact of the multi-head attention, which loses influence with respect to the residual connection, consequently revealing a reduced entanglement of contextual information in BERT. Finally, Abnar and Zuidema (2020) proposed the attention rollout method, which measures the mixing of information by linearly combining attention matrices, a method that has been extended to Transformers in the visual domain (Chefer et al., 2021a,b). A drawback of this method is that it assumes an equal influence of the skip connection and the attention weights.

In this work, we propose ALTI, an interpretability method that provides input tokens relevancies[1] to the model predictions by measuring the aggregation of contextual information across layers. We use the attention block decomposition proposed by Kobayashi et al. (2021) and refine the measure of the contribution of each input token representation to the attention block output (layer-wise token-to-token interactions), based on the properties of the representation space and the limitations of previously proposed metrics. We then aggregate the layer-wise explanations and track the mixing of information in each token representation, yielding input attributions for the model predictions. Finally, in the Text Classification and Subject-Verb Agreement tasks, we show ALTI scores higher than gradient-based methods and previous similar approaches in two common faithfulness metrics, while showing greater robustness. The code to reproduce the experiments is publicly available.[2]

2 Background

2.1 Attention Block Decomposition

The attention block[3] computations in each layer (highlighted parts in Figure 1) can be reformulated (Kobayashi et al., 2021) as a simple expression of the layer input representations. Given a sequence of token representations $\mathbf{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_J) \in \mathbb{R}^{d \times J}$, and a model with H heads and head dimension $d_h = d/H$, the attention block output of the i-th token $\boldsymbol{y}_i$ is computed by applying the layer normalization (LN) over the sum of the residual vector $\boldsymbol{x}_i$, and the output of the multi-head attention module
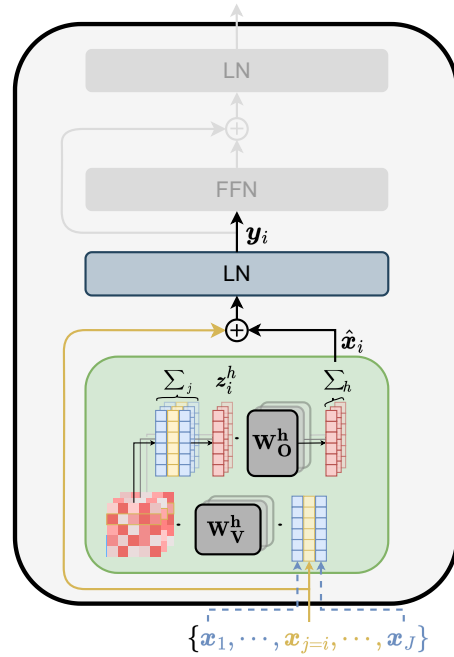

Figure 1: Transformer layer with the modules considered in the analysis. We compute layer-wise token-to-token interactions by measuring the contributions of each input token representation $\boldsymbol{x}_j$ to the attention block output $\boldsymbol{y}_i$.

(MHA) $\hat{\boldsymbol{x}}_i$:

$$\boldsymbol{y}_i = \text{LN}(\hat{\boldsymbol{x}}_i + \boldsymbol{x}_i) \tag{1}$$

Each head inside MHA computes[4] $\boldsymbol{z}_i^h \in \mathbb{R}^{d_h}$:

$$\boldsymbol{z}_i^h = \sum_{j}^{J} \mathbf{A}_{i,j}^h \mathbf{W}_V^h \boldsymbol{x}_j \tag{2}$$

with $\mathbf{A}_{i,j}^h$ referring to the attention weight where token i attends token j, and $\mathbf{W}_V^h \in \mathbb{R}^{d_h \times d}$ to a learned weight matrix. $\hat{\boldsymbol{x}}_i$ is calculated by concatenating each $\boldsymbol{z}_i^h$ and projecting the joint vector through $\mathbf{W}_O \in \mathbb{R}^{d \times d}$:

$$\hat{\boldsymbol{x}}_i = \mathbf{W}_O \, \text{Concat}(\boldsymbol{z}_i^1, \cdots, \boldsymbol{z}_i^H) \tag{3}$$

This is equivalent to a sum over heads where each $\boldsymbol{z}_i^h$ is projected through the partitioned weight matrix $\mathbf{W}_O^h \in \mathbb{R}^{d \times d_h}$ and adding the bias $\boldsymbol{b}_O \in \mathbb{R}^d$:

$$\hat{\boldsymbol{x}}_i = \sum_{h}^{H} \mathbf{W}_O^h \boldsymbol{z}_i^h + \boldsymbol{b}_O \tag{4}$$

---

[1] We use 'relevancies', 'attributions', and 'importances' interchangeably.

[2] https://github.com/mt-upc/transformer-contributions.

[3] We refer to 'attention block' as the multi-head attention, residual connection and layer normalization components.

[4] The bias vector associated with $\mathbf{W}_V^h$ is omitted for the sake of simplicity.
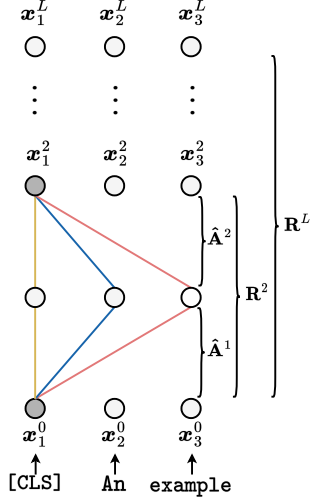
Figure 2: Example of attention graph. The relevance $\left(\mathbf{R}^2_{\texttt{[CLS]}}\right)$ of the input token $\texttt{[CLS]}$ token to its second layer representation $\boldsymbol{x}^2_1$ is obtained by summing all possible paths (coloured).

By swapping summations we can now rewrite Eq. 1 as:

$$\boldsymbol{y}_i = \text{LN}\left(\sum_j^J \sum_h^H \mathbf{W}^h_O \mathbf{A}^h_{i,j} \mathbf{W}^h_V \boldsymbol{x}_j + \boldsymbol{b}_O + \boldsymbol{x}_i\right) \tag{5}$$

Given a vector $\boldsymbol{u}$, $\text{LN}(\boldsymbol{u})$ can be reformulated as $\frac{1}{\sigma(\boldsymbol{u})}\mathbf{L}\boldsymbol{u} + \beta$ (see Appendix A), where $\mathbf{L}$ is a linear transformation. Thanks to the linearity of $\mathbf{L}$, we can express $\boldsymbol{y}_i$ as:

$$\boldsymbol{y}_i = \sum_j^J T_i(\boldsymbol{x}_j) + \frac{1}{\sigma\left(\hat{\boldsymbol{x}}_i + \boldsymbol{x}_i\right)}\mathbf{L}\boldsymbol{b}_O + \beta \tag{6}$$

where the transformed vectors $T_i(\boldsymbol{x}_j)$ are:

$$T_i(\boldsymbol{x}_j) = \begin{cases} \frac{1}{\sigma(\hat{\boldsymbol{x}}_i + \boldsymbol{x}_i)}\mathbf{L}\sum_h^H \mathbf{W}^h_O \mathbf{A}^h_{i,j} \mathbf{W}^h_V \boldsymbol{x}_j & \text{if } j \neq i \\ \frac{1}{\sigma(\hat{\boldsymbol{x}}_i + \boldsymbol{x}_i)}\mathbf{L}\left(\sum_h^H \mathbf{W}^h_O \mathbf{A}^h_{i,j} \mathbf{W}^h_V \boldsymbol{x}_j + \boldsymbol{x}_i\right) & \text{if } j = i \end{cases}$$

Kobayashi et al. (2021) stated that the contribution $c_{i,j}$ of each input vector $\boldsymbol{x}_j$ to the layer output $\boldsymbol{y}_i$ can be estimated by how much its transformed vector $T_i(\boldsymbol{x}_j)$ affects the result in Eq. 6. They propose using the Euclidean norm of the transformed vector as the metric of contribution:

$$c_{i,j} = \|T_i(\boldsymbol{x}_j)\|_2 \tag{7}$$

2.2 Attention Rollout

Abnar and Zuidema (2020) proposed to measure the mixing of contextual information across the model by relying on attention weights, creating an
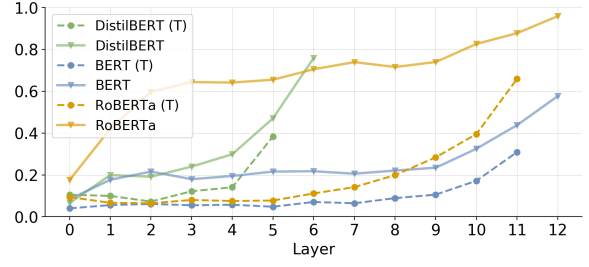

Figure 3: Average cosine similarity of attention block output representations (solid line) and transformed representations (dashed line) in 500 random samples of SST-2 dataset.

"attention graph" where nodes represent tokens and hidden representations, and edges attention weights. Two nodes in different layers are connected through multiple paths. To add the residual connection, the attention weights matrix gets augmented with the identity matrix $\hat{\mathbf{A}}^l = 0.5\mathbf{A}^l + 0.5\mathbf{I}$.

We can compute the amount of information flowing from one node to another in different layers by multiplying the edges in each path, and summing over the different paths. In the example of Figure 2, the amount of input information of $\texttt{[CLS]}$ in its second layer representation $\boldsymbol{x}^2_1$ can be obtained as $\hat{\mathbf{A}}^2_{1,1} \cdot \hat{\mathbf{A}}^1_{1,1} + \hat{\mathbf{A}}^2_{1,2} \cdot \hat{\mathbf{A}}^1_{2,1} + \hat{\mathbf{A}}^2_{1,3} \cdot \hat{\mathbf{A}}^1_{3,1}$. This is equivalent to the dot product between $\hat{\mathbf{A}}^2_{1,:}$ and $\hat{\mathbf{A}}^1_{:,1}$, which generalizes to the matrix multiplication when considering all tokens, giving the input relevance matrix at layer l, $\mathbf{R}^l \in \mathbb{R}^{J \times J}$:

$$\mathbf{R}^l = \hat{\mathbf{A}}^l \cdot \hat{\mathbf{A}}^{l-1} \cdot \cdots \cdot \hat{\mathbf{A}}^1 \tag{8}$$

3 Proposed Approach

The decomposition of the attention block, represented as a sum of vectors in Eq. 6, allows us to interpret token-to-token interactions within each layer. Kobayashi et al. (2021) proposed to measure the influence of each input token with the ℓ_2 norm of the transformed vectors (Eq. 7). We present two reasons why this estimation may not be accurate:

1. A property of the contextual representations in Transformer-based models is that they are highly anisotropic (Ethayarajh, 2019), i.e. the expected cosine similarity of randomly sampled token representations tend to be close to 1 (solid lines in Figure 3). However, transformed representations exhibit reduced anisotropy, especially for the first layers, where there is almost isotropy (dashed lines in Figure 3), i.e. they are more randomly
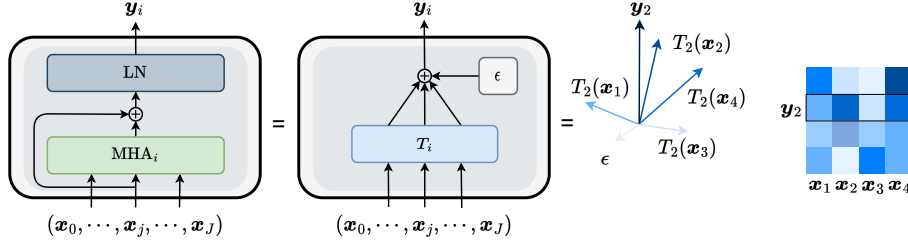
Figure 4: The self-attention block (left) at each position i can be decomposed as a summation of transformed input vectors (middle). The closest vector ($T_2(\boldsymbol{x}_2)$) *contributes* the most to $\boldsymbol{y}_2$. We obtain a matrix of contributions **C** (right) reflecting layer-wise token-token interactions.

spread across the space. This reinforces the need of accounting for the vector's orientation in space, as opposed to solely relying on their norm.

2. Recent studies (Kovaleva et al., 2021; Cai et al., 2021; Luo et al., 2021) have found that some embedding dimensions acquire disproportionately large values, dominating the similarity measures (Timkey and van Schijndel, 2021). ℓ_2 normalized metrics, since they square each vector component, unavoidably weigh heavily the outlier dimensions.

We can analyze the expression in Eq. 6 as $T_i(\boldsymbol{x}_j)$ vectors *contributing* to the sum resultant $\boldsymbol{y}_i$. We propose to measure how much each transformed vector contributes to the sum by means of its distance to the output vector $\boldsymbol{y}_i$. We expect that the closer the vector is to $\boldsymbol{y}_i$, the higher its contribution (Figure 4). In this way, we take into account where each transformed vector lies in the representation space (Reason 1). Due to its robustness to the aforementioned idiosyncratic dimensions (Reason 2), we use ℓ_1 norm, i.e. the Manhattan distance between the attention block output and the transformed vector:

$$d_{i,j} = \|\boldsymbol{y}_i - T_i(\boldsymbol{x}_j)\|_1 \qquad (9)$$

The level of contribution of $\boldsymbol{x}_j$ to $\boldsymbol{y}_i$, $c_{i,j}$, is proportional to the proximity of $T_i(\boldsymbol{x}_j)$ to $\boldsymbol{y}_i$. The closer the transformed vector is to $\boldsymbol{y}_i$, the larger its contribution. We measure proximity as the negative of Manhattan distance $-d_{i,j}$. Finally, we neglect the contributions of those vectors lying beyond the ℓ_1 length of $\boldsymbol{y}_i$:

$$c_{i,j} = \frac{\max(0, -d_{i,j} + \|\boldsymbol{y}_i\|_1)}{\sum_k \max(0, -d_{i,k} + \|\boldsymbol{y}_i\|_1)} \qquad (10)$$

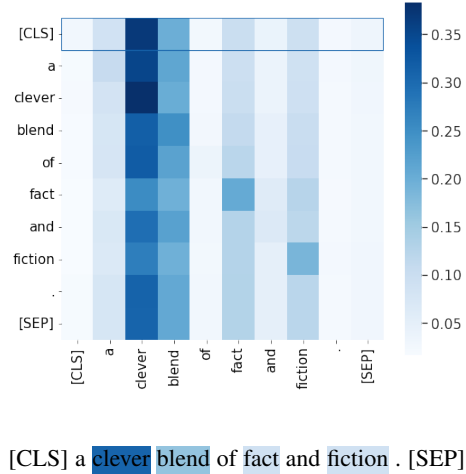[CLS] a clever blend of fact and fiction . [SEP]

Figure 5: Top: input (columns) attribution scores $\mathbf{R}^L$ provided by ALTI in BERT's last layer token representations (rows). Bottom: attribution scores corresponding to [CLS] token ($\mathbf{R}^L_{[CLS]}$).

Computing Eq. 9 and Eq. 10 for all $\boldsymbol{y}_i$ gives us the contributions matrix $\mathbf{C} \in \mathbb{R}^{J \times J}$ containing every token-to-token interaction within the layer.

We also propose to consider contributions across the model as a "contribution graph", similar to the attention graph in Section §2.2 but using the obtained contributions instead of attention weights. We can then track the amount of contextual information from the input tokens in intermediate token representations, which we use as input attribution scores. By combining linearly the contribution matrices up to layer l (Figure 5 bottom) we get:

$$\mathbf{R}^l = \mathbf{C}^l \cdot \mathbf{C}^{l-1} \cdot \,\cdots\, \cdot \mathbf{C}^1 \qquad (11)$$

4 Experimental Setup

We perform our experiments in the Text Classification (TC) and the Subject-Verb Agreement (SVA) tasks. The former evaluates how models classify

an entire input sequence, the latter assesses the ability of a model to capture syntactic phenomena (Linzen et al., 2016; Goldberg, 2019). For the TC task, we use the Stanford Sentiment Treebank v2 (SST-2) (Socher et al., 2013) composed of short sentences, IMDB (Maas et al., 2011) with movies reviews and longer inputs than SST2, and Yelp Dataset Challenge (Zhang et al., 2015) containing user's reviews from businesses of similar length than in IMDB. All of them have positive and negative sentiment sentences. For the SVA task, we use Linzen et al. (2016) dataset, which includes sentences from Wikipedia containing a present-tense verb that agrees in grammatical number (singular/plural) with the head of the subject. The sentence is fed into the model with its verb masked, and the model is asked to predict if the masked verb is singular or plural[5] (binary classification):

*At least four **players** from the 1983 draft now* [MASK] *as coaches.*
plural

4.1 Models

For our experiments we consider three common Transformer pre-trained models[6] with different sizes and pre-training procedures: BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2019) and RoBERTa (Liu et al., 2019).

In the TC task, we use fine-tuned models provided by TextAttack (Morris et al., 2020). For the robustness analysis in Section §5.2, we fine-tune 10 pre-trained BERT models on SST-2 with the recommended hyperparameters in Devlin et al. (2019). We compute attribution scores from the row $\mathbf{R}^L_{\texttt{[CLS]}} \in \mathbb{R}^J$ (Figure 5 bottom) that corresponds to the final layer [CLS] embedding, considered a sentence representation for classification tasks. Regarding the SVA task, we split Linzen et al. (2016) dataset into 60%/20%/20% for training, validation, and testing respectively, and fine-tune a pre-trained BERT model. We use the input relevances of $\mathbf{R}^L_{\texttt{[MASK]}}$.

4.2 Faithfulness Metrics

An interpretation is considered to be faithful if it accurately reflects a model's decision-making pro-

cess. A well-established method for measuring faithfulness is by deleting parts of the input sentence $\mathbf{x}$ and observing the change in the predicted probability. Two common erasure-based metrics are comprehensiveness (comp.) and sufficiency (suff.) (DeYoung et al., 2020). Chan et al. (2022) have demonstrated that they have higher diagnosticity, i.e. they favor faithful interpretations over randomly generated ones, and lower time complexity than other well-known faithfulness metrics. Comprehensiveness and sufficiency are defined as:

Comprehensiveness. Measures the change in probability of the predicted class after removing important tokens:

$$\text{Comp.} = \frac{1}{|B| + 1} \sum_{k \in B} (f(\mathbf{x}) - f(\mathbf{x} \backslash \mathbf{r}_{:k\%})) \quad (12)$$

where $\mathbf{r}_{:k\%}$ refers to the top-$k\%$ most important tokens obtained by an interpretability method. The higher the drop in the probability, the more faithful the interpretation.

Sufficiency. Captures if important tokens are enough to retain the original prediction:

$$\text{Suff.} = \frac{1}{|B| + 1} \sum_{k \in B} (f(\mathbf{x}) - f(\mathbf{r}_{:k\%})) \quad (13)$$

Lower values of sufficiency indicate a more faithful interpretation, since, in that case, the prediction doesn't change when considering only the important tokens. As in the original paper, for both metrics we use $B = \{0, 5, 10, 20, 50\}$.

4.3 Input Attribution Methods

Input attribution methods rank input tokens in accordance with how they impact model predictions. They can be divided into: gradient-based methods, perturbation-based, and those relying on the attention mechanism. The gradient of the model's output with respect to the input embeddings is often used as a baseline of faithfulness interpretation (Jain and Wallace, 2019). Atanasova et al. (2020); Zaman and Belinkov (2022) show that gradient-based methods perform better than other interpretability methods, regarding different faithfulness metrics. Finally, perturbation-based methods (Zeiler and Fergus, 2014; Ribeiro et al., 2016) compute attributions by replacing the original sentence with a modification. Zaman and Belinkov (2022)

---

[5]As a general rule, a singular verb has an 's' added to it in the present tense, such as eats, plays, is, has. A plural verb does not have an 's' added to it.

[6]We use the models available at `https://github.com/huggingface/transformers` (Wolf et al., 2020).

show that erasure-based methods, such as comprehensiveness and sufficiency favor perturbation-based methods attributed to noise due to the OOD perturbations.

Gradient. Considering the model f taking as input a sequence of embeddings $\mathbf{X}^0 \in \mathbb{R}^{d \times J}$, f can be approximated by the linear part of the Taylor-expansion at a baseline point (Simonyan et al., 2014), $f(\mathbf{X}^0) \approx \nabla f(\mathbf{X}^0) \cdot \mathbf{X}^0$. Then, $\nabla f(\mathbf{X}^0)$ gives a score per embedding dimension, which is often considered as how sensitive the model is to each input dimension when predicting a certain class. To get per token saliency scores (Li et al., 2016), we obtain the gradient vector corresponding to the j-th token $\nabla_{\boldsymbol{x}_j^0} f(\mathbf{X}^0) = \frac{\partial f(\mathbf{X}^0)}{\partial \boldsymbol{x}_j^0}$. Then, we aggregate the gradient vector into a scalar using the ℓ_2 norm (Grad_{ℓ_2}):

$$\mathrm{attr}(x_j) = \left\| \nabla_{\boldsymbol{x}_j^0} f(\mathbf{X}^0) \right\|_2 \qquad (14)$$

Recently, Bastings et al. (2021) showcased (in BERT and SST-2) the high degree of faithfulness of Grad_{ℓ_2} method.

Gradient × input. This method (Shrikumar et al., 2016) performs the multiplication of the gradient and the corresponding input embedding. Each component of the gradient vector gets multiplied by the corresponding component of the embedding. Following (Atanasova et al., 2020; Zaman and Belinkov, 2022), we aggregate the component scores into a single scalar by taking the ℓ_2 norm (G × I$_{\ell_2}$) as in Eq. 14 or by taking the mean (G × I$_\mu$) as follows:

$$\mathrm{attr}(x_j) = \frac{1}{N} \sum_{k=1}^{d} |\nabla_{x_{jk}^0} f(\mathbf{X}^0) \cdot x_{jk}^0| \qquad (15)$$

Integrated Gradients. Integrated gradients (Sundararajan et al., 2017) approximates the integral of gradients of the model's output with respect to the inputs along the straight line path from a baseline input $\boldsymbol{B}$, to the actual input. The attribution score for each embedding dimension is defined as:

$$(x_{jk}^0 - \boldsymbol{b}_{jk}) \cdot \frac{1}{m} \sum_{c=1}^{m} \nabla_{\hat{x}_{jk}^0} f(\hat{\boldsymbol{X}}_c^0) \qquad (16)$$

where $\hat{\boldsymbol{X}}_c^0 = \boldsymbol{B} + \frac{c}{m}(\mathbf{X}^0 - \boldsymbol{B})$, and m number of steps. As baseline, we use repeated [MASK] vectors for each word except for [CLS] and [SEP] (Sajjad et al., 2021), and 100 steps. We aggregate (IG$_{\ell_2}$

and IG$_\mu$) the attribution scores of the embedding dimensions of Eq. 16 to obtain $\mathrm{attr}(x_j)$.

Finally, we normalize the obtained attribution scores in the range so that they sum 1. We use the Captum library implementations (Kokhlikyan et al., 2020).

Attention. Attention-based methods that provide input attributions include Attention *Rollout* (Abnar and Zuidema, 2020), as described in Section §2.2. Concurrent to this work, Modarressi et al. (2022) propose *Globenc*, which combines Attention Rollout aggregation technique with Kobayashi et al. (2021) layer-wise contributions (Eq. 7), with the addition of the layer normalization of the FFN module. In Section §5.4 we compare ALTI to Globenc.

5 Results

In this section, we present quantitative and qualitative results comparing ALTI with other input attribution methods.

5.1 Faithfulness Results

In Table 2 we show comprehensiveness and sufficiency results for the three models and four datasets. It can be seen that across every different configuration, our proposed ALTI method outperforms other input attribution methods. Regarding comprehensiveness, datasets with short sentences like SST-2 and SVA (Figure 6 (c)) provide small differences between methods. This is expected since these datasets are simpler, and therefore, interpretations can more easily find the important tokens. However, for datasets containing longer inputs with multiple sentences, like IMDB and Yelp, ALTI clearly stands out. This can be observed in Figure 6 (a) and (b), where the probability drop in the model prediction is shown when removing one token at a time. We observe small differences in performance within gradient-based methods across datasets and models, with IG$_{\ell_2}$ performing the best on average among them, agreeing with the observations of Atanasova et al. (2020). However, ALTI outperforms IG$_{\ell_2}$ by 58% on average in comprehensiveness, and by 38% in sufficiency. Results of RoBERTa and DistilBERT on every dataset can be found in Appendix B.

Previous research concluded that faithfulness results for evaluating different interpretability methods are task and model dependent (Bastings et al., 2021; Madsen et al., 2021a). Interestingly, although for the rest of the methods results vary

Methods	BERT								RoBERTa		DistilBERT	
	SST-2		Yelp		IMDB		SVA		SST-2		SST-2	
	Comp.↑	Suff.↓	Comp.	Suff.	Comp.	Suff.	Comp.	Suff.	Comp.	Suff.	Comp.	Suff.
Grad_{ℓ_2} --	0.204	0.076	0.083	0.101	0.192	0.052	0.284	0.145	0.190	0.075	0.230	0.066
IG_{ℓ_2} --	0.223	0.084	0.111	0.024	0.214	0.06	0.315	0.171	0.223	0.074	0.295	0.048
IG_{μ} --	0.211	0.112	0.106	**0.022**	0.179	0.063	0.317	0.174	0.231	0.067	0.279	0.064
$\text{G} \times \text{I}_{\ell_2}$ --	0.199	0.080	0.081	0.104	0.197	0.056	0.279	0.149	0.187	0.081	0.235	0.065
$\text{G} \times \text{I}_{\mu}$ --	0.207	0.073	0.087	0.098	0.213	0.054	0.285	0.145	0.191	0.078	0.237	0.065
Rollout --	0.074	0.270	0.076	0.102	0.09	0.185	0.108	0.292	0.076	0.179	0.152	0.147
Globenc --	0.174	0.125	0.118	0.053	0.207	0.117	0.251	0.178	0.161	0.119	0.24	0.072
ALTI --	**0.317**	**0.044**	**0.255**	**0.022**	**0.308**	**0.031**	**0.372**	**0.088**	**0.269**	**0.057**	**0.332**	**0.034**

Table 2: Faithfulness results of the different interpretability methods for BERT, RoBERTa and DistilBERT on four different datasets. ↑ means a higher number indicates better performance, while ↓ means the opposite.
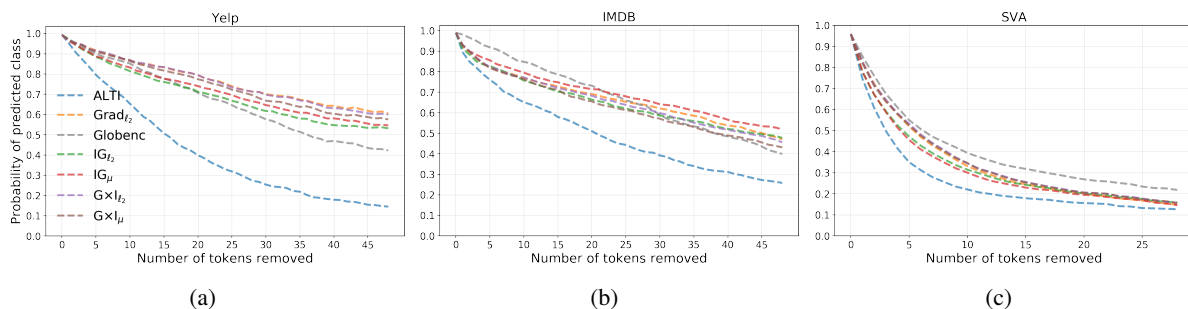

Figure 6: Probability drop in BERT predictions when removing important tokens, obtained by different interpretability methods. We show results on three datasets.

across models and tasks, we observe ALTI repeatedly performs the best across different tasks and models.

In the qualitative examples in Tables 1 and 3 we can observe that gradient-based methods often miss the relevant tokens that drive the model's negative prediction. ALTI consistently assigns high relevance to spans of text that have a negative connotation, such as 'depressing', 'don't plan on returning' in Table 3, or 'food not good', 'stomach aches' in Table 1, as expected from a negative sentiment prediction. We observe that, as opposed to ALTI, gradient-based methods become less accurate with longer sequences. A very large example with a positive sentiment prediction can be found in Appendix C Table 9, where ALTI accurately picks as important tokens those with positive meanings.

5.2 Robustness Analysis

We perform a study to investigate the robustness of different interpretability methods based on the *implementation invariance* property defined by Sundararajan et al. (2017). Given a set of models with the same architecture, and trained with the same data, but only differing in their random weight initialization, it compares how different are the input
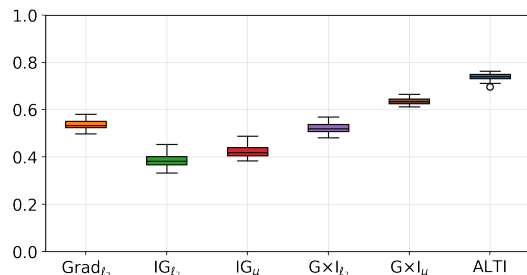

Figure 7: Jaccard-25% similarity score between the interpretations of each method in 10 BERT's random seeds.

attribution scores between the models, for the same interpretability method. If the predictions of the models are also identical, i.e. models are *functionally equivalent*, we would expect input attributions also to be identical. Zafar et al. (2021) perform this test on two identical neural text classifiers (i, j), differing in their random weight initialization. Since the vast majority of the predictions are the same for both models, they consider them to be *almost functionally identical* models. Then, they measure the Jaccard similarity score between the top-25% tokens ranked based on their importance as specified by an input attribution method, for model i and

Grad$_{\ell_2}$

um . it ' s okay , i guess . they have food at decent prices , but the isles are narrow , everything needs a good cleaning and repainting , and it just felt dark and depressing . otherwise it ' s all right , but i don ' t plan on returning here .

IG$_{\ell_2}$

um . it ' s okay , i guess . they have food at decent prices , but the isles are narrow , everything needs a good cleaning and repainting , and it just felt dark and depressing . otherwise it ' s all right , but i don ' t plan on returning here .

ALTI

um . it ' s okay , i guess . they have food at decent prices , but the isles are narrow , everything needs a good cleaning and repainting , and it just felt dark and depressing . otherwise it ' s all right , but i don ' t plan on returning here .

Table 3: Saliency maps of BERT generated by two common gradient methods and by our proposed method, ALTI, for a **negative** sentiment predictions of Yelp dataset.
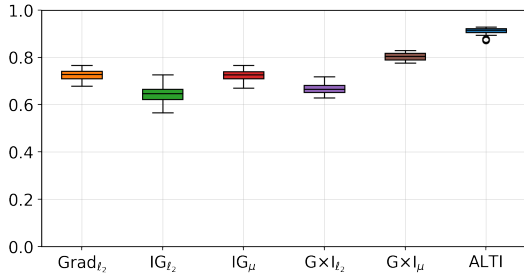
Figure 8: Spearman's rank correlation between the interpretations of each method in 10 BERT's random seeds.

Dataset	Metric	BERT		RoBERTa		DistilBERT	
		ALTI ℓ_2	Norms	ALTI ℓ_2	Norms	ALTI ℓ_2	Norms
SST-2	Comp.↑	**0.3**	0.134	**0.23**	0.168	**0.292**	0.157
	Suff.↓	**0.045**	0.18	**0.075**	0.135	**0.051**	0.152
IMDB	Comp.	**0.286**	0.148	**0.203**	0.133	**0.254**	0.175
	Suff.	**0.031**	0.115	**0.086**	0.162	**0.053**	0.104
Yelp	Comp.	**0.212**	0.082	**0.108**	0.064	**0.2**	0.093
	Suff.	**0.02**	0.057	**0.104**	0.18	**0.031**	0.114
SVA	Comp.	**0.374**	0.273	**0.379**	0.275	**0.357**	0.315
	Suff.	**0.088**	0.152	**0.114**	0.178	**0.096**	0.117

Table 4: Faithfulness results of BERT, RoBERTa and DistilBERT comparing ALTI ℓ_2 with the Norms approach.

j. If the top-25% tokens by the two attributions coincide, Jaccard-25%$(i, j) = 1$. In case tokens don't overlap, Jaccard-25%$(i, j) = 0$.

We perform the robustness test with 10 pre-trained BERT models from the MultiBERTs (Sellam et al., 2022), which only differ in their random weight initialization. For each interpretability method we compute Jaccard-25% score between all the different pairs of models. In Figure 7 we show the distribution of the obtained scores. We also compute the Spearman's rank correlation coefficient (Figure 8), which evaluates how well the relationship between the two ranked interpretations can be described using a monotonic function. We can observe that ALTI provides more homogeneous interpretations across identical models in terms of similarity and correlation, suggesting that it is a more robust interpretability method than gradient-based methods.

5.3 Ablation Study

We inspect the importance of the different components conforming ALTI.

Layer-wise token contributions. We compare the effect of our token contributions' measurement in Eq. 9 and Eq. 10 against previous approach by Kobayashi et al. (2021) (Norms) described in Eq. 7

by aggregating each type of contributions with the Rollout method. To isolate the influence of the norm choice, we use the ℓ_2 in Eq. 9 and Eq. 10 (ALTI ℓ_2). Results in Table 4 show our proposed layer-wise contribution measurement largely improves previous approach.

Norm choice in ALTI. We also evaluate the election of the norm in our proposed approach. In Table 5 we show faithfulness results considering ℓ_1 and ℓ_2. In almost every setting ℓ_1 outperforms ℓ_2. Remarkably, the advantage of the ℓ_1 is less noticeable on BERT, which we hypothesize is explained by the reduced anisotropy of its representations (Figure 3).

Dataset	Metric	BERT		RoBERTa		DistilBERT	
		ℓ_1	ℓ_2	ℓ_1	ℓ_2	ℓ_1	ℓ_2
SST-2	Comp.↑	**0.317**	0.3	**0.269**	0.23	**0.332**	0.292
	Suff.↓	**0.044**	0.045	**0.057**	0.075	**0.034**	0.051
IMDB	Comp.	**0.308**	0.286	**0.266**	0.203	**0.304**	0.254
	Suff.	**0.031**	0.031	**0.05**	0.086	**0.039**	0.053
Yelp	Comp.	**0.221**	0.212	**0.138**	0.108	**0.237**	0.2
	Suff.	**0.02**	0.02	**0.07**	0.104	**0.017**	0.031
SVA	Comp.	0.372	**0.374**	**0.39**	0.379	**0.382**	0.357
	Suff.	**0.088**	0.088	**0.109**	0.114	**0.084**	0.096

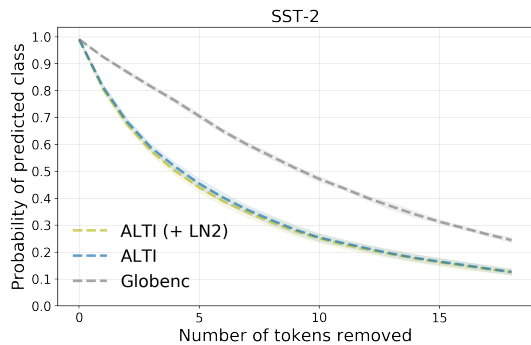Table 5: Faithfulness results of BERT, RoBERTa and DistilBERT using ℓ_1 and ℓ_2 norms.

Figure 9: Probability drop in BERT predictions when removing important tokens, results show mean and SD in BERT across 10 seeds in SST-2 dataset.

5.4 Addition of Layer Norm 2

Concurrent work (Modarressi et al., 2022) present Globenc method, which aggregates the contributions obtained by (Kobayashi et al., 2021) in Eq. 7 with Attention Rollout method. Moreover, they add the layer normalization (LN2) of the Feed-forward module of the Transformer layer into their method. We evaluate the faithfulness of the interpretations provided by Globenc in Table 2, and although it improves the Rollout baseline, is far from the results obtained with ALTI. We consider analyzing the influence of the second layer normalization by including it in ALTI method. The probability drop in SST-2 across 10 BERT seeds (Figure 9) shows the influence of LN2 is negligible. We observe similar patterns across models and datasets.

6 Conclusions

In this paper, we have presented ALTI, an input attribution method that quantifies the mixing of information in the Transformer. We have demonstrated that with accurate layer-wise token-to-token contribution measurements relying on ℓ_1-based metrics, the interpretable attention decomposition of the attention block is a powerful tool when combined with the rollout method. Empirically, we show that ALTI outperforms every input attribution method we have experimented with in two common faithfulness metrics, while showing greater robustness. Overall, we believe this opens new possibilities for studying contextual information aggregation across the Transformer.

Limitations

ALTI measures the amount of contextual information in each layer representation of the Transformer.

From the influence of each input token to the last layer representation we extract input attributions for the model prediction. However, our method does not consider the classifier on top of the Transformer. Therefore, our proposed method doesn't provide explanations for each of the output classes, as opposed to gradient-based methods. We also underline that faithfulness in this work is evaluated via sufficiency and comprehensiveness metrics.

Ethical Considerations

ALTI provides explanations about input attributions in the Transformer. By itself, we are not aware of any ethical implications of the methodology, which does not take into account any subjective priors. To prove its usefulness, we have used two different benchmarks, text classification, and subject-verb agreement. As far as we are concerned, these benchmarks have been used in the past without raising major ethical considerations. Therefore, we do not have any major issue to report in this section.

7 Acknowledgements

References

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.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Jasmijn Bastings, Sebastian Ebert, Polina Zablotskaia, Anders Sandholm, and Katja Filippova. 2021. "will you find these shortcuts?" a protocol for evaluating the faithfulness of input salience methods for text classification.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss,

Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2020. On identifiability in transformers. In *International Conference on Learning Representations*.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2021. Isotropy in the contextual embedding space: Clusters and manifolds. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Chun Sik Chan, Huanqi Kong, and Liang Guanqing. 2022. A comparative study of faithfulness metrics for model interpretability methods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5029–5038, Dublin, Ireland. Association for Computational Linguistics.

Hila Chefer, Shir Gur, and Lior Wolf. 2021a. Generic attention-model explainability for interpreting bimodal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 397–406.

Hila Chefer, Shir Gur, and Lior Wolf. 2021b. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *CoRR*, abs/1901.05287.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 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.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch.

Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. 2021. BERT busters: Outlier dimensions that disrupt transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3392–3405, Online. Association for Computational Linguistics.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Trans. Assoc. Comput. Linguistics*, 4:521–535.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Kaiji Lu, Zifan Wang, Piotr Mardziel, and Anupam Datta. 2021. Influence patterns for explaining information flow in BERT. In *Advances in Neural Information Processing Systems*.

Ziyang Luo, Artur Kulmizev, and Xiaoxi Mao. 2021. Positional artefacts propagate through masked language model embeddings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5312–5327, Online. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Andreas Madsen, Nicholas Meade, Vaibhav Adlakha, and Siva Reddy. 2021a. Evaluating the faithfulness of importance measures in nlp by recursively masking allegedly important tokens and retraining.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2021b. Post-hoc interpretability for neural nlp: A survey.

Ali Modarressi, Mohsen Fayyaz, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. 2022. Globenc: Quantifying global token attribution by incorporating the whole encoder layer in transformers.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.

Damian Pascual, Gino Brunner, and Roger Wattenhofer. 2021. Telling BERT's full story: from local attention to global aggregation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 105–124, Online. Association for Computational Linguistics.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793, Online. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Hassan Sajjad, Narine Kokhlikyan, Fahim Dalvi, and Nadir Durrani. 2021. Fine-grained interpretation and causation analysis in deep NLP models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*, pages 5–10, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Thibault Sellam, Steve Yadlowsky, Ian Tenney, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Raluca Turc, Jacob Eisenstein, Dipanjan Das, and Ellie Pavlick. 2022. The multiBERTs: BERT reproductions for robustness analysis. In *International Conference on Learning Representations*.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2016. Not just a black box: Learning important features through propagating activation differences. *CoRR*, abs/1605.01713.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Kaiser Sun and Ana Marasović. 2021. Effective attention sheds light on interpretability. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4126–4135, Online. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328, International Convention Centre, Sydney, Australia. PMLR.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing.

In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Muhammad Bilal Zafar, Michele Donini, Dylan Slack, Cedric Archambeau, Sanjiv Das, and Krishnaram Kenthapadi. 2021. On the lack of robust interpretability of neural text classifiers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3730–3740, Online. Association for Computational Linguistics.

Kerem Zaman and Yonatan Belinkov. 2022. A multilingual perspective towards the evaluation of attribution methods in natural language inference.

Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pages 818–833, Cham. Springer International Publishing.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

A   Layer Normalization decomposition

Layer normalization acting over an input $\boldsymbol{u}$ can be defined as: $\text{LN}(\boldsymbol{u}) = \frac{\boldsymbol{u} - \mu(\boldsymbol{u})}{\sigma(\boldsymbol{u})} \odot \gamma + \beta$, where μ and σ compute the mean and standard deviation of u, and γ and β refer to the element-wise transformation and bias respectively. $\text{LN}(\boldsymbol{u})$ can be decomposed into $\frac{1}{\sigma(\boldsymbol{u})}\mathbf{L}\boldsymbol{u} + \beta$, where $\mathbf{L}$ is a linear transformation:

$$
\mathbf{L} :=
\begin{bmatrix}
\gamma_1 & 0 & \cdots & 0 \\
0 & \gamma_2 & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots \\
0 & 0 & \cdots & \gamma_n
\end{bmatrix}
\begin{bmatrix}
\frac{n-1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\
-\frac{1}{n} & \frac{n-1}{n} & \cdots & -\frac{1}{n} \\
\cdots & \cdots & \cdots & \cdots \\
-\frac{1}{n} & -\frac{1}{n} & \cdots & \frac{n-1}{n}
\end{bmatrix}
$$

The linear map on the right subtracts the mean to the input vector, $\boldsymbol{u}' = \boldsymbol{u} - \mu(\boldsymbol{u})$. The left matrix performs the hadamard product with the layer normalization weights ($\boldsymbol{u}' \odot \gamma$).

B   RoBERTa and DistilBERT Results

Methods	IMDB		Yelp		SVA	
	Comp.↑	Suff.↓	Comp.	Suff.	Comp.	Suff.
Grad_{ℓ_2}	0.266	0.054	0.111	0.073	0.338	0.107
IG_{ℓ_2}	0.275	0.055	0.128	0.054	0.372	0.108
IG_μ	0.244	0.059	0.108	0.056	0.365	0.118
$\text{G} \times \text{I}_{\ell_2}$	0.281	0.053	0.115	0.075	0.333	0.111
$\text{G} \times \text{I}_\mu$	0.282	0.053	0.117	0.074	0.335	0.11
Rollout	0.183	0.089	0.092	0.114	0.241	0.179
Globenc	0.252	0.066	0.124	0.066	0.341	0.108
ALTI	**0.304**	**0.039**	**0.237**	**0.017**	**0.382**	**0.084**

Table 6: Faithfulness results of the different interpretability methods for DistilBERT on IMDB, Yelp and SVA datasets. ↑ means a higher number indicates better performance, while ↓ means the opposite.

Methods	IMDB		Yelp		SVA	
	Comp.↑	Suff.↓	Comp.	Suff.	Comp.	Suff.
Grad_{ℓ_2}	0.216	0.083	0.075	0.122	0.273	0.165
IG_{ℓ_2}	0.2	0.084	0.087	0.11	0.363	0.163
IG_μ	0.215	0.083	0.102	0.094	0.351	0.161
$\text{G} \times \text{I}_{\ell_2}$	0.183	0.109	0.066	0.149	0.28	0.169
$\text{G} \times \text{I}_\mu$	0.225	0.08	0.083	0.12	0.27	0.167
Rollout	0.077	0.197	0.031	0.208	0.223	0.183
Globenc	0.154	0.086	0.065	0.12	0.305	0.17
ALTI	**0.266**	**0.05**	**0.138**	**0.07**	**0.39**	**0.109**

Table 7: Faithfulness results of the different interpretability methods for RoBERTa on IMDB, Yelp and SVA datasets. ↑ means a higher number indicates better performance, while ↓ means the opposite.

C   Qualitative Examples

Grad_{ℓ_2}
friendly staff and nice selection of vegetarian options . food is just okay , not great . makes me wonder why everyone likes food fight so much .

$\text{G}\times\text{I}_{\ell_2}$
friendly staff and nice selection of vegetarian options . food is just okay , not great . makes me wonder why everyone likes food fight so much .

IG_{ℓ_2}
friendly staff and nice selection of vegetarian options . food is just okay , not great . makes me wonder why everyone likes food fight so much .

ALTI
friendly staff and nice selection of vegetarian options . food is just okay , not great . makes me wonder why everyone likes food fight so much .

Table 8: Saliency maps of BERT generated by three common gradient methods and by our proposed method, ALTI, for a **negative** sentiment prediction example of Yelp dataset.

Grad$_{\ell_2}$

low budget horror movie . if you don ' t raise your expectations too high , you ' ll probably enjoy this little flick . beginning and end are pretty good , middle drags at times and seems to go nowhere for long periods as we watch the goings on of the insane that add atmosphere but do not advance the plot . quite a bit of gore . i enjoyed bill mcghee ' s performance which he made quite believable for such a low budget picture , he managed to carry the movie at times when nothing much seemed to be happening . nurse charlotte beale , played by jesse lee , played her character well so be prepared to want to slap her toward the end ! she makes some really stupid mistakes but then , that ' s what makes these low budget movies so good ! i would have been out of that place and five states away long before she even considered that it might be a good idea to leave ! if you enjoy this movie , try committed from 1988 which is basically a rip off of this movie .

G×I$_{\ell_2}$

low budget horror movie . if you don ' t raise your expectations too high , you ' ll probably enjoy this little flick . beginning and end are pretty good , middle drags at times and seems to go nowhere for long periods as we watch the goings on of the insane that add atmosphere but do not advance the plot . quite a bit of gore . i enjoyed bill mcghee ' s performance which he made quite believable for such a low budget picture , he managed to carry the movie at times when nothing much seemed to be happening . nurse charlotte beale , played by jesse lee , played her character well so be prepared to want to slap her toward the end ! she makes some really stupid mistakes but then , that ' s what makes these low budget movies so good ! i would have been out of that place and five states away long before she even considered that it might be a good idea to leave ! if you enjoy this movie , try committed from 1988 which is basically a rip off of this movie .

IG$_{\ell_2}$

low budget horror movie . if you don ' t raise your expectations too high , you ' ll probably enjoy this little flick . beginning and end are pretty good , middle drags at times and seems to go nowhere for long periods as we watch the goings on of the insane that add atmosphere but do not advance the plot . quite a bit of gore . i enjoyed bill mcghee ' s performance which he made quite believable for such a low budget picture , he managed to carry the movie at times when nothing much seemed to be happening . nurse charlotte beale , played by jesse lee , played her character well so be prepared to want to slap her toward the end ! she makes some really stupid mistakes but then , that ' s what makes these low budget movies so good ! i would have been out of that place and five states away long before she even considered that it might be a good idea to leave ! if you enjoy this movie , try committed from 1988 which is basically a rip off of this movie .

ALTI

low budget horror movie . if you don ' t raise your expectations too high , you ' ll probably enjoy this little flick . beginning and end are pretty good , middle drags at times and seems to go nowhere for long periods as we watch the goings on of the insane that add atmosphere but do not advance the plot . quite a bit of gore . i enjoyed bill mcghee ' s performance which he made quite believable for such a low budget picture , he managed to carry the movie at times when nothing much seemed to be happening . nurse charlotte beale , played by jesse lee , played her character well so be prepared to want to slap her toward the end ! she makes some really stupid mistakes but then , that ' s what makes these low budget movies so good ! i would have been out of that place and five states away long before she even considered that it might be a good idea to leave ! if you enjoy this movie , try committed from 1988 which is basically a rip off of this movie .

Table 9: Saliency maps of BERT generated by three common gradient methods and by our proposed method, ALTI, for a **positive** sentiment prediction example of IMDB dataset.
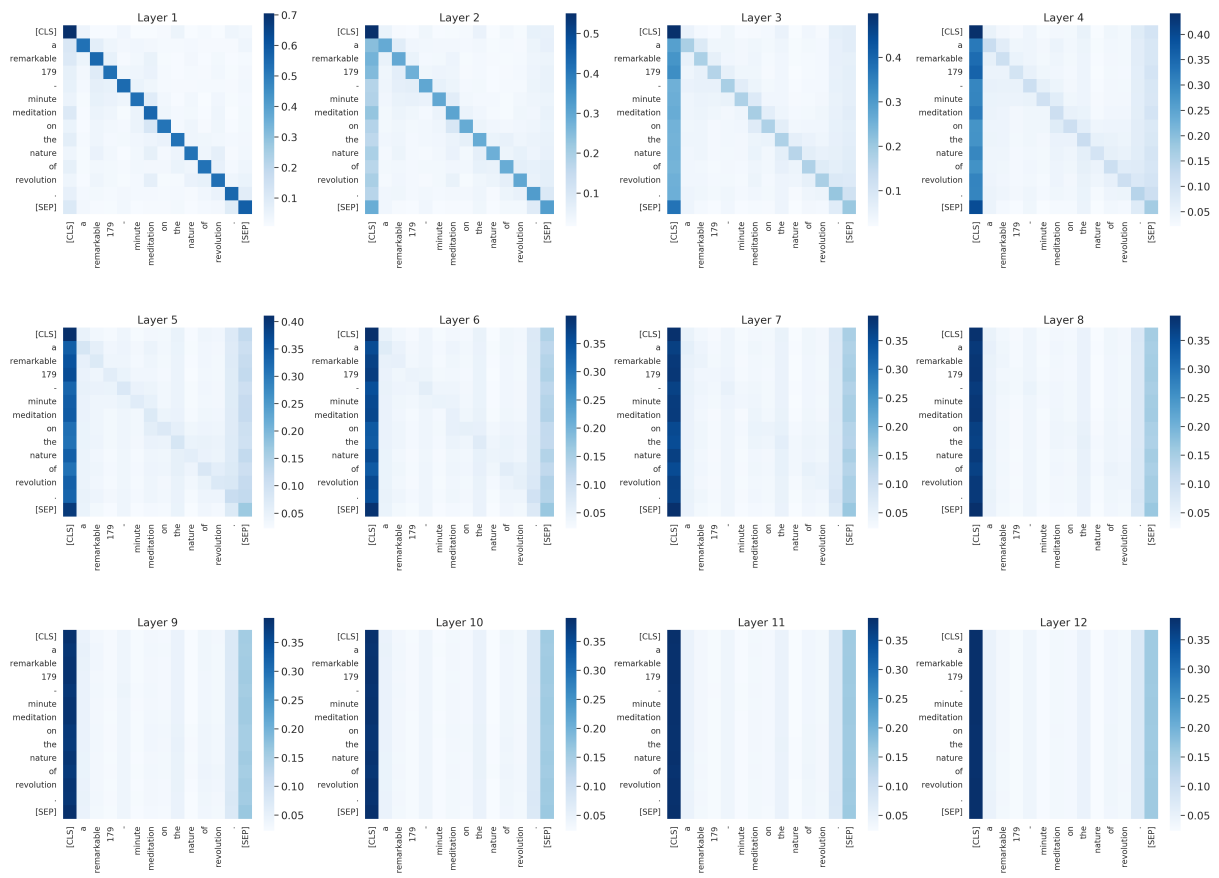
Figure 10: Attn Rollout relevancies $\mathbf{R}^l$ in BERT across layers.

Figure 11: Globenc relevancies $\mathbf{R}^l$ in BERT across layers.
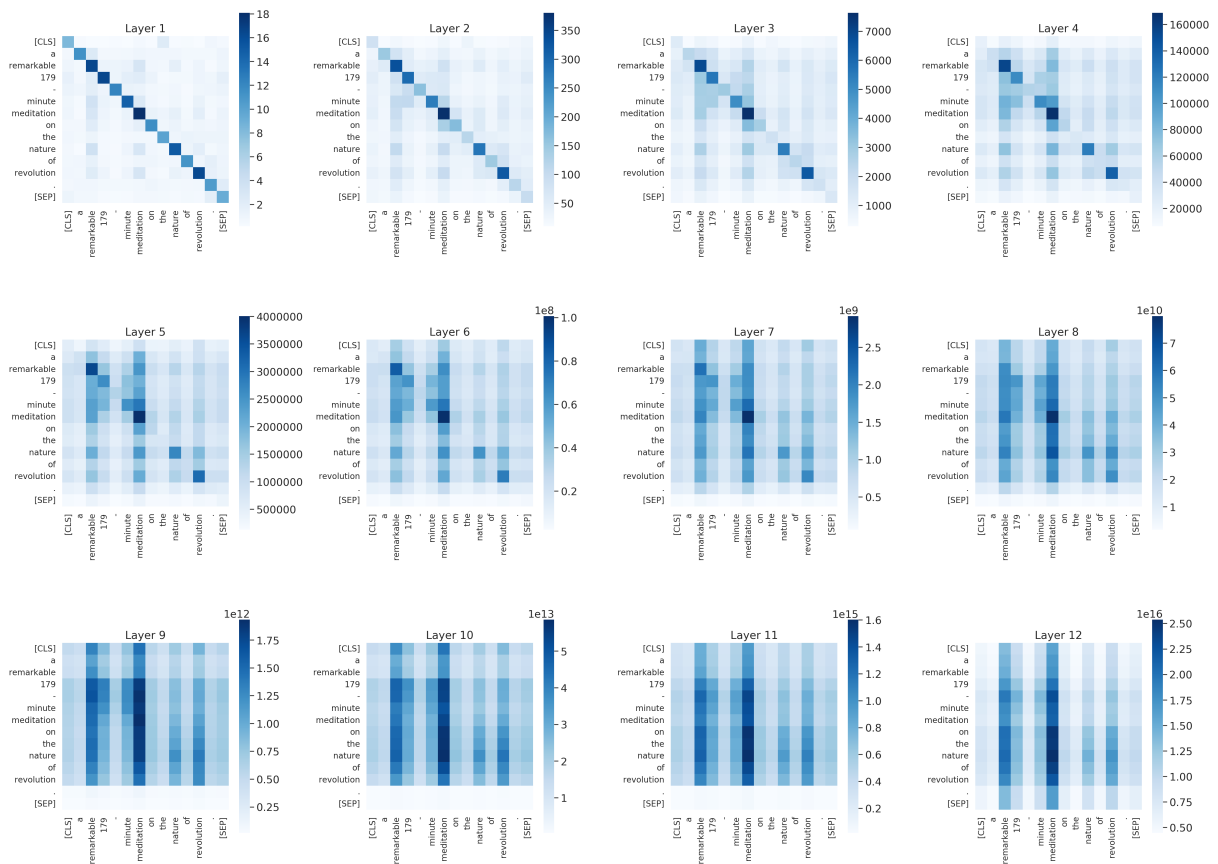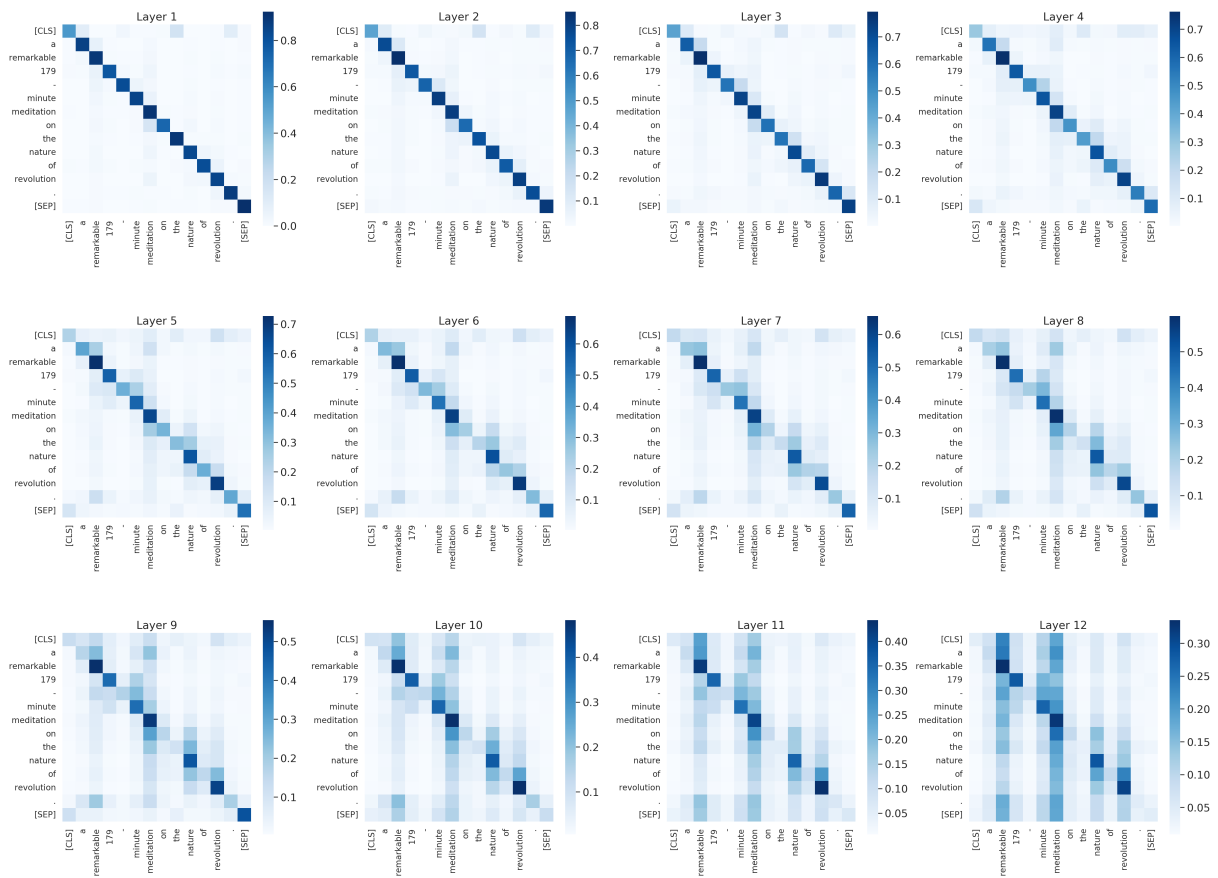
Figure 12: ALTI method relevancies $\mathbf{R}^l$ in BERT across layers.