Exploring the Role of BERT Token Representations to Explain Sentence Probing Results

Hosein Mohebbi[⋆♡] **Ali Modarressi**[⋆♡] **Mohammad Taher Pilehvar**[♠]
[♡] Iran University of Science and Technology, Iran
[♠] Tehran Institute for Advanced Studies, Khatam University, Iran
{hosein_mohebbi, m_modarressi}@comp.iust.ac.ir
mp792@cam.ac.uk

Abstract

Several studies have been carried out on revealing linguistic features captured by BERT. This is usually achieved by training a diagnostic classifier on the representations obtained from different layers of BERT. The subsequent classification accuracy is then interpreted as the ability of the model in encoding the corresponding linguistic property. Despite providing insights, these studies have left out the potential role of token representations. In this paper, we provide a more in-depth analysis on the representation space of BERT in search for distinct and meaningful subspaces that can explain the reasons behind these probing results. Based on a set of probing tasks and with the help of attribution methods we show that BERT tends to encode meaningful knowledge in specific token representations (which are often ignored in standard classification setups), allowing the model to detect syntactic and semantic abnormalities, and to distinctively separate grammatical number and tense subspaces.[1]

1 Introduction

Recent years have seen a surge of interest in pre-trained language models, highlighted by extensive research around BERT (Devlin et al., 2019) and its derivatives. One strand of research has focused on enhancing existing models with the primary objective of improving downstream performance on various NLP tasks (Liu et al., 2019b; Lan et al., 2019; Yang et al., 2019). Another strand analyzes the behaviour of these models with the hope of getting better insights for further developments (Clark et al., 2019; Kovaleva et al., 2019; Jawahar et al., 2019; Tenney et al., 2019; Lin et al., 2019).

Probing is one of the popular analysis methods, often used for investigating the encoded knowledge in language models (Conneau et al., 2018; Tenney et al., 2018). This is typically carried out by training a set of diagnostic classifiers that predict a specific linguistic property based on the representations obtained from different layers. Recent works in probing language models demonstrate that initial layers are responsible for encoding low-level linguistic information, such as part of speech and positional information, whereas intermediate layers are better at syntactic phenomena, such as syntactic tree depth or subject-verb agreement, while in general semantic information is spread across the entire model (Lin et al., 2019; Peters et al., 2018; Liu et al., 2019a; Hewitt and Manning, 2019; Tenney et al., 2019). Despite elucidating the type of knowledge encoded in various layers, these studies do not go further to investigate the reasons behind the layer-wise behavior and the role played by token representations. Analyzing the shortcomings of pre-trained language models requires a scrutiny beyond the mere performance (e.g., accuracy or F-score) in a given probing task. This is particularly important as recent studies point out that the diagnostic classifier (applied to the model's outputs) might itself play a significant role in learning nuances of the task and hence suggest evaluating probes with alternative criteria (Hewitt and Liang, 2019; Voita and Titov, 2020; Pimentel et al., 2020; Zhu and Rudzicz, 2020).

We extend the layer-wise analysis to the token level in search for distinct and meaningful subspaces in BERT's representation space that can explain the performance trends in various probing tasks. To this end, we leverage the attribution method (Simonyan et al., 2013; Sundararajan et al., 2017; Smilkov et al., 2017) which has recently proven effective for analytical studies in NLP (Li et al., 2016; Yuan et al., 2019; Bastings and Filippova, 2020; Atanasova et al., 2020; Wu and Ong, 2021; Voita et al., 2021). Our analysis on a set of surface, syntax, and semantic probing tasks (Con-

Authors marked with a star (⋆) contributed equally.

[1]Code is available at https://github.com/hmohebbi/explain-probing-results

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 792–806
November 7–11, 2021.

neau et al., 2018) shows that BERT usually encodes the knowledge required for addressing these tasks within specific token representations, particularly at higher layers. For instance, we found that sentence-ending tokens (e.g., "[SEP]" and ".") are mostly responsible for carrying positional information through layers, or when the input sequence undergoes a re-ordering the alteration is captured by specific token representations, e.g., by the swapped tokens or the coordinator between swapped clauses. Also, we observed that the ##s token is mainly responsible for encoding noun number and verb tense information, and that BERT clearly distinguishes the two usages of the token in higher layer representations.

2 Related Work

Probing. Several analytical studies have been conducted to examine the capacities and weaknesses of BERT, often by means of probing layer-wise representations (Lin et al., 2019; Goldberg, 2019; Liu et al., 2019a; Jawahar et al., 2019; Tenney et al., 2019). Particularly, Jawahar et al. (2019) leveraged the probing framework of Conneau et al. (2018) to show that BERT carries a hierarchy of linguistic information, with surface, syntactic, and semantic features respectively occupying initial, middle and higher layers. In a similar study, Tenney et al. (2019) employed the edge probing tasks defined by Tenney et al. (2018) to show the hierarchy of encoded knowledge through layers. Moreover, they observed that while most of the syntactic information can be localized in a few layers, semantic knowledge tends to spread across the entire network. Both studies were aimed at discovering the extent of linguistic information encoded across different layers. In contrast, in this paper we explore the role of token representations in the final performance. More recently, Klafka and Ettinger (2020) investigated the extent of information that can be recovered from each word representation in a sentence about the other words. Apart from using different probing tasks and methodologies, most notably they relied solely on classifier's performance score, whereas we make conclusion based on the most contributed token representations.

Representation subspaces. In addition to layer-wise representations, subspaces that encode specific linguistic knowledge, such as syntax, have been a popular area of study. By designing a structural probe, Hewitt and Manning (2019) showed that there exists a linear subspace that approximately encodes all syntactic tree distances. In a follow-up study, Chi et al. (2020) showed that similar syntactic subspaces exist for languages other than English in the multilingual BERT and that these subspaces are shared among languages to some extent. This corroborated the finding of Pires et al. (2019) that multilingual BERT has common subspaces across different languages that capture various linguistic knowledge.

As for semantic subspaces, Wiedemann et al. (2019) showed that BERT places the contextualized representations of polysemous words into different regions of the embedding space, thereby capturing sense distinctions. Similarly, Reif et al. (2019) studied BERT's ability to distinguish different word senses in different contexts. Using the probing approach of Hewitt and Manning (2019), they also found that there exists a linear transformation under which distances between word embeddings correspond to their sense-level relationships. Our work extends these studies by revealing other types of surface, syntactic, and high-level semantic subspaces and linguistic features using a pattern-finding approach on different types of probing tasks.

Attribution methods. Recently, there has been a surge of interest in using attribution methods to open up the blackbox and explain the decision makings of pre-trained language models, from developing methods and libraries to visualize inputs' contributions (Ribeiro et al., 2016; Han et al., 2020; Wallace et al., 2019; Tenney et al., 2020) to applying them into fine-tuned models on downstream tasks (Atanasova et al., 2020; Wu and Ong, 2021; Voita et al., 2021). In particular, Voita et al. (2021) adopted a variant of Layer-wise Relevance Propagation (Bach et al., 2015) to evaluate the relative contributions of source and target tokens to the generation process in Neural Machine Translation predictions. To our knowledge, this is the first time that attribution methods are employed for layer-wise probing of pre-trained language models.

3 Methodology

Our analytical study was mainly carried out on a set of sentence-level probing tasks from SentEval (Conneau and Kiela, 2018). The benchmark consists of several single-sentence evaluation tasks. Each task provides 100k instances for training and 10k for test, all balanced across target classes. We

used the test set examples for our evaluation and in-depth analysis. Following the standard procedure for this benchmark, we trained a diagnostic classifier for each task. The classifier takes sentence representations as its input and predicts the specific property intended for the corresponding task.

In what follows in this section, we first describe how sentence representations were computed in our experiments. Then, we discuss our approach for measuring the attribution of individual token representations to classifier's decision.

3.1 Sentence Representation

For computing sentence representations for layer l, we opted for a simple unweighted averaging (h^l_{Avg}) of all input tokens (except for padding and [CLS] token). This choice was due to our observation that the mean pooling strategy retains or improves [CLS] performance in most layers in our probing tasks (cf. Appendix A.1 for more details). This corroborates the findings of Reimers and Gurevych (2019) who observed a similar trend on sentence similarity and inference tasks. Moreover, the mean pooling strategy simplifies our measuring of each token's attribution, discussed next.

Our evaluations are based on the pre-trained BERT (base-uncased, 12-layer, 768-hidden size, 12-attention head, 110M parameters) obtained from the HuggingFace's Transformers library (Wolf et al., 2020). We followed the recommended hyperparameters by Jawahar et al. (2019) to train the diagnostic classifiers for each layer. In addition to BERT, we carried out our evaluations on RoBERTa (Liu et al., 2019b, base, 125M parameters). However, we observed highly similar patterns for the two models. Hence, we only report results for the BERT model.

3.2 Gradient-based Attribution Method

We leveraged a gradient-based attribution method in order to enable an in-depth analysis of layer-wise representations with the objective of explaining probing performances. Specifically, we are interested in computing the attribution of each input token to the output labels. This is usually referred to as the *saliency* score of an input token to classifier's decision. Note that using attention weights for this purpose can be misleading given that raw attention weights do not necessarily correspond to the importance of individual token representations (Serrano and Smith, 2019; Jain and Wallace, 2019; Abnar and Zuidema, 2020; Kobayashi et al., 2020).

Using gradients for attribution methods has been a popular option in neural networks, especially for vision (Simonyan et al., 2013; Sundararajan et al., 2017; Smilkov et al., 2017). Images are constructed from pixels; hence, computing their individual attributions to a given class can be interpreted as the spatial support for that class (Simonyan et al., 2013). However, in the context of text processing, input tokens are usually represented by vectors; hence, raw feature values do not necessarily carry any specific information. Li et al. (2016)'s solution to this problem relies on the gradients over the inputs. Let w_c be the derivative of class c's output logit (y_c) with respect to the k-th dimension of the input embedding ($h[k]$):

$$w_c(h[k]) = \frac{\partial y_c}{\partial h[k]} \tag{1}$$

This gradient can be interpreted as the sensitivity of class c to small changes in $h[k]$. To have this at the level of words (or tokens), Li et al. (2016) suggests using the average of the absolute values of $w_c(h[k])$ over all of the d dimensions of the embedding:

$$Score_c(h) = \frac{1}{d}\sum_{k=1}^{d} |w_c(h[k])| \tag{2}$$

Although the absolute value of gradients could be employed for understanding and visualizing the contributions of individual words, these values can only express the sensitivity of the class score to small changes without information about the direction of contribution (Yuan et al., 2019). We adopt the method of Yuan et al. (2019) for our setting and compute the saliency score for the i^{th} representation in layer l, i.e., h^l_i, as:

$$Score_c(h^l_i) = \frac{\partial y^l_c}{\partial h^l_{Avg}} \cdot h^l_i \tag{3}$$

where y^l_c denotes the probability that the classifier assigns to class c based on the l^{th}-layer representations. Given that our aim is to explain the representations (rather than evaluating the classifier), we set c in Equation 3 as the correct label. This way, the scores reflect the contributions of individual input tokens in a sentence to the classification decision.

In what follows in the paper, we use the analysis method discussed in this section to find those tokens that play the central role in different surface (Section 4), syntactic (Sections 5 and 6.1) and semantic (Section 6.3) probing tasks. Based on these tokens we then investigate the reasons behind performance variations across layers.

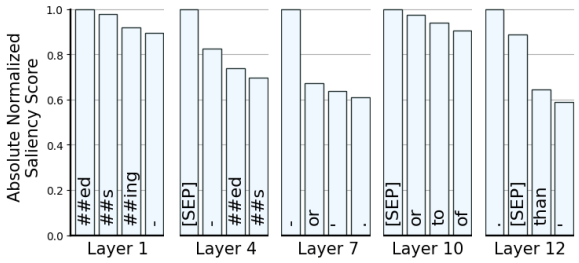

Figure 1: Absolute normalized saliency scores for the top-4 most attributed (high frequency, > 128) tokens across five different layers.[2]

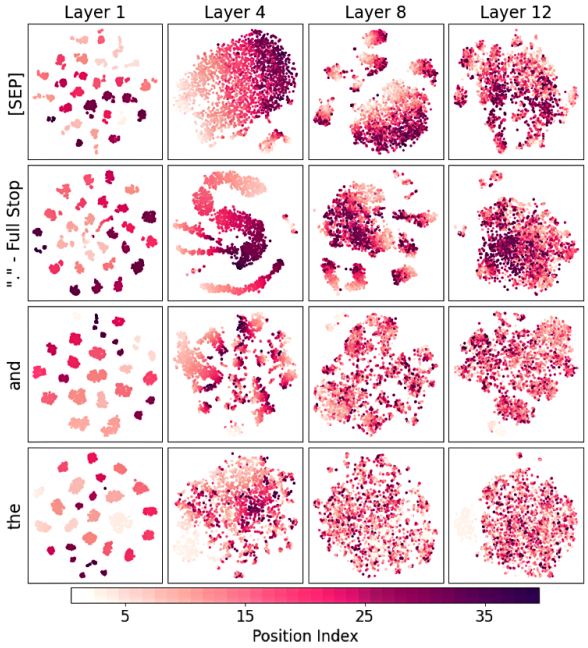

Figure 2: t-SNE plots of the representations of four selected high frequency tokens ("[SEP]", "." full stop, "the", "and") in different sentences. Colors indicate the corresponding token's position in the sentence (darker colors means higher position index). Finalizing tokens (e.g., "[SEP]", ".") preserve distinct patterns in final layers, indicating their role in encoding positional information, while other (high frequency) tokens exhibit no such behavior.

4 Sentence Length

In this surface-level task we probe the representation of a given sentence in order to estimate its size, i.e., the number of words (not tokens) in it. To this end, we used SentEval's **SentLen** dataset, but changed the formulation from the original classification objective to a regression one which allows a better generalization due to its fine-grained setting. The diagnostic classifier receives average-pooled representation of a sentence (cf. Section 3.1) as input and outputs a continuous number as an estimate for the input length.

Given that the ability to encode the exact length of input sentences is not necessarily a critical feature, we do not focus on layer-wise performance and instead discuss the reason behind the performance variations across layers. To this end, we calculated the absolute saliency scores for each input token in order to find those tokens that played pivotal role while estimating sentence length.

Rounding the regressed estimates and comparing them with the gold labels in the test set, we can observe a significant performance drop from 0.91 accuracy in the first layer to 0.44 in the last layer (cf. Appendix A.1 for details). This decay is not surprising given that the positional encodings, which are added to the input embeddings in BERT and are deemed to be the main players for such a position-based task, get faded through layers (Voita et al., 2019).

Sentence ending tokens retain positional information. Figure 1 shows tokens that most contributed to the probing results across different layers according to the attribution analysis. Finalizing tokens (e.g. "[SEP]" and ".") are the main contributors in the higher layers. We further illustrate this in Figure 2 in which we compare the representations of a finalizing token with those of another frequent non-finalizing token. Clearly, positioning information is lost throughout layers in BERT; however, finalizing tokens partially retain this information, as visible from distinct pattern in higher layers.

5 Verb Tense and Noun Number

This analysis inspects BERT representations for grammatical number and tense information. For this experiment we used the **Tense** and **ObjNum** tasks[3]: the former checks whether the main-clause verb is labeled as present or past[4], whereas the latter classifies the object according to its number, i.e., singular or plural (Conneau et al., 2018). On both tasks, BERT preserves a consistently high performance (> 0.82 accuracy) across all layers (cf. Appendix A.1 for more details).

[2]Full figures (for all layers) are available in Appendix A.2

[3]We will not discuss the SubjNum results, since we observed significant labeling issues (See Appendix A.3) that could affect our conclusions. This can also explain low human performance reported by Conneau et al. (2018) on this task.

[4]In Tense task, each sentence may include multiple verbs, subjects, and objects, while the label is based on the main clause (Conneau et al., 2018).

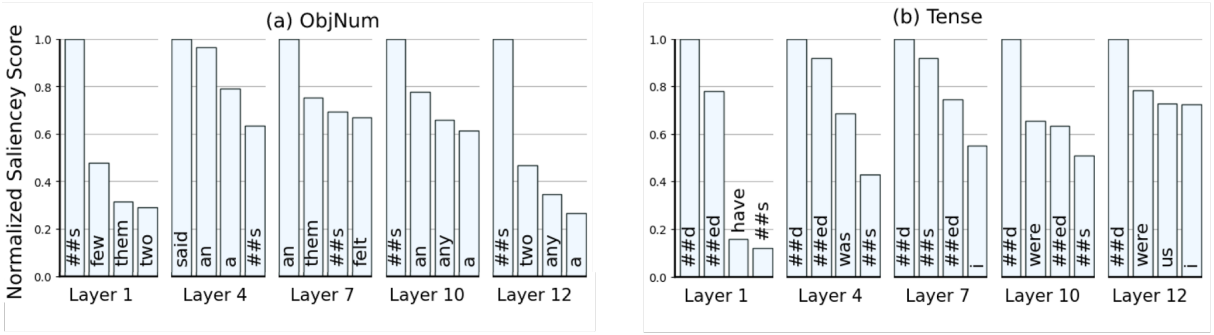

Figure 3: The top-4 most attributed (high freq.) tokens across five different layers for the ObjNum and Tense tasks.

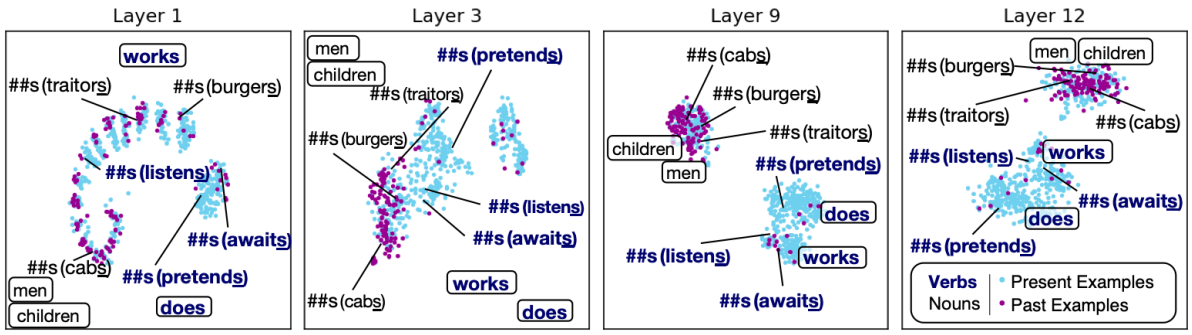

Figure 4: t-SNE plots of the layer-wise representations of the ##s token in different sentences. Colors indicate whether the token occurred in present- or past-labeled sentence in the Tense task (see Section 5). For the sake of comparison, we also include two present verbs without the ##s token[5](i.e., *does* and *works*) and two irregular plural nouns (i.e., *men* and *children*), in rounded boxes. The distinction between the two different usages of the token (noun number as well as the tense information) is clearly encoded in higher layer contextualized representations. As plural nouns can appear in both past- and present-labeled examples, the cluster belongs to the plural form of ##s token in higher layers may contain both types of examples.

Articles and ending tokens (e.g., ##s and ##ed) are key playmakers. Attribution analysis, illustrated in Figure 3(a), reveals that article words (e.g., "a" and "an") and the ending ##s token, which makes out-of-vocab plural words (or third person present verbs), are among the most attributed tokens in the ObjNum task. This shows that these tokens are mainly responsible for encoding object's number information across layers. As for the Tense task, Figure 3(b) shows a consistently high influence from verb ending tokens (e.g., ##ed and ##s) across layers which is in line with performance trends for this task and highlights the role of these tokens in preserving verb tense information.

##s — Plural or Present? The ##s token proved influential in both tense and number tasks. The token can make a verb into its simple present tense (e.g., read → reads) or transform a singular noun into its plural form (e.g., book → books). We further investigated the representation space to check if BERT can distinguish this nuance. Results are shown in Figure 4: after the initial layers, BERT recognizes and separates these two forms into two distinct clusters (while BERT's tokenizer made no distinction among different usages). Interestingly, we also observed that other present/plural tokens that did not have the ##s token aligned well with these subspaces.

6 Inversion Abnormalities

For this set of experiments, we opted for SentEval's Bi-gram Shift and Coordination Inversion tasks which respectively probe model's ability in detecting syntactic and semantic abnormalities. The goal of this analysis was to to investigate if BERT encodes inversion abnormality in a given sentence into specific token representations.

6.1 Word-level inversion

Bi-gram Shift (**BShift**) checks the ability of a model to identify whether two adjacent words within a given sentence have been inverted (Con-

[5]Tokens that were not split by the tokenizer.

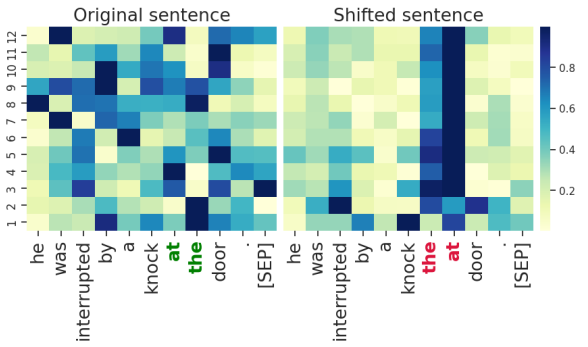

Figure 5: Normalized layer-wise attribution scores for a randomly sampled sentence from the test set (left). The right figure shows how the attribution scores changed when two words ("at" and "the") from the original sentence were inverted.

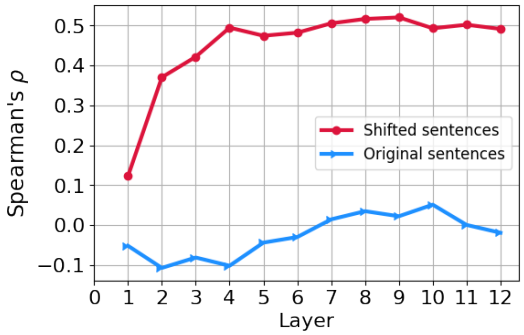

Figure 6: Spearman's ρ correlation of gradient attribution scores with the mask array M (a one-hot indicating shifted indices), averaged on all examples across all layers. High correlations indicate model's increased sensitivity to the shifted tokens, a trend which is not seen in the original sentences.

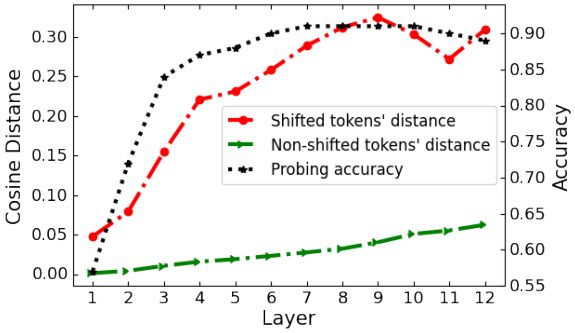

Figure 7: Average cosine distances of shifted tokens (and other tokens) to themeselves, before and after inversion in the BShift task. The trend for the shifted token distances highly correlates with that of probing performance, supporting our hypothesis of BERT encoding abnormalities in the shifted tokens.

neau et al., 2018). Probing results shows that the higher half layers of BERT can properly distinguish this peculiarity (Figure 7). Similarly to the previous experiments, we leveraged the gradient attribution method to figure out those tokens that were most effective in detecting the inverted sentences. Given that the dataset does not specify the inverted words, we reconstructed the inverted examples by randomly swapping two consecutive words in the original sentences of the test set, excluding the beginning of the sentences and punctuation marks as stated in (Conneau et al., 2018).

6.2 Results

Our attribution analysis shows that swapping two consecutive words in a sentence results in a significant boost in the attribution scores of the inverted tokens. As an example, Figure 5 depicts attribution scores of each token in a randomly sampled sentence from the test set across different layers. The classifier distinctively focuses on the token representations for the shifted words (Figure 5 right), while no such patterns exists for the original sentence (Figure 5 left).

To verify if this observation holds true for other instances in the test set, we carried out the following experiment. For each given sequence X of n tokens, we defined a boolean mask $M = [m_1, m_2, ...m_n]$ which denotes the position of the inversion according to the following condition:

$$m_i = \begin{cases} 1, & x_i \in \text{V} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where V is the set of all tokens in the shifted bigram ($|V| \geq 2$, given BERT's sub-word tokenization). Then we computed the Spearman's rank correlation coefficient of the attribution scores with M for all examples in the test set. Figure 6 reports mean layer-wise correlation scores. We observe that in altered sentences the correlation significantly grows over the first few layers which indicates model's increased sensitivity to the shifted tokens.

We hypothesize that BERT implicitly encodes abnormalities in the representation of shifted tokens. To investigate this, we computed the cosine distance of each token to itself in the original and shifted sentences. Figure 7 shows layer-wise statistics for both shifted and non-shifted tokens. Distances between the shifted token representations aligns well with the performance trend for this probing task (also shown in the figure).

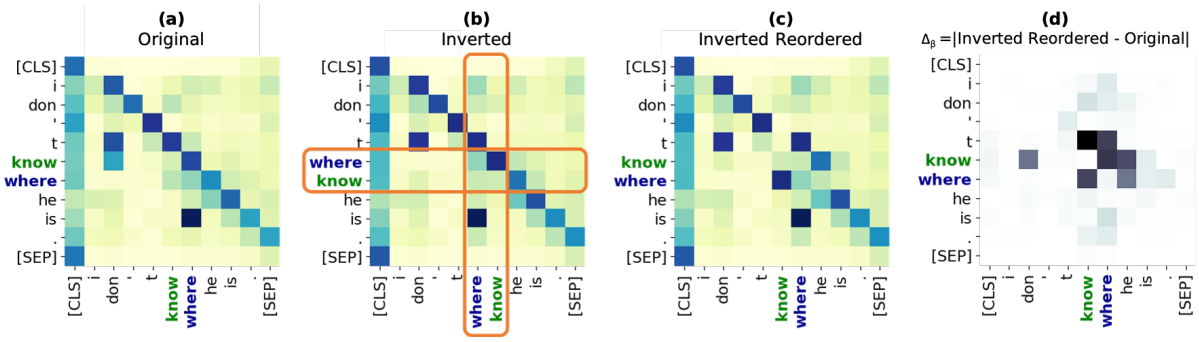

Figure 8: Evaluating the β map[6]for a single example in a specific layer (layer = 3). After computing the map for the original (a) and inverted (b) forms of the sentence, to compute the Δ_β map we need to reorder the inverted map. The corresponding columns and rows for the inverted words (orange boxes) are swapped to re-construct the original order (c). The Δ_β map (d) is the magnitude of the point-wise difference between the re-ordered and the original maps. The Δ_β map for this example clearly shows that most of the changes have occurred within the bi-gram inversion area. All values are min-max normalized.

6.2.1 Attention-norm behavior on bi-gram inversion

Our observation implies that BERT somehow encodes oddities in word order in the representations of the involved tokens. To investigate the root cause of this, we took a step further and analyzed the building blocks of these representations, i.e., the self-attention mechanism. To this end, we made use of the norm-based analysis method of Kobayashi et al. (2020) which incorporates both attention weights and transformed input vectors (the value vectors in the self-attention layer). The latter component enables a better interpretation at the token level. This norm-based metric $||\sum \alpha f(x)||$–for the sake of convenience we call it **attention-norm**–is computed as the vector-norm of the i^{th} token to the j^{th} token over all attention heads ($H = 12$) in each layer l:

$$\beta^l_{i,j} = || \sum_{head=1}^{H} \alpha^{head,l}_{i,j} f^{head,l}(\mathbf{h}^l_j)|| \qquad (5)$$

where $\alpha_{i,j}$ is the attention weight between the two tokens and $f^{head,l}(x)$ is a combination of the value transformation in layer l of the head and the matrix which combines all heads together (see Kobayashi et al. (2020)'s paper for more details).

We computed the attention-norm map in all layers, for both the original and shifted sentence. To be able to compare these two maps, we re-ordered the shifted sentence norms to match the original order. The magnitude of the difference between the original and the re-ordered map Δ_{β^l} shows the amount of change in each token's attention-norm to each token. Figure 8 illustrates this procedure for a sample instance. Given that bi-gram locations are different across each instance, to compute an overall Δ_{β^l} we centered each map based on the position of the inversion. As a result of this procedure, we obtained a Δ_{β^l} map for each layer and for all examples. Centering and averaging all these maps across layers produced Figure 9.

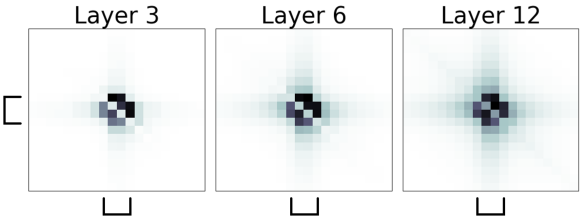

Figure 9: A cumulative view of the attention-norm changes (Δ_{β^l}) centered around the bi-gram position (the approximate bi-gram position is marked on each figure). Each plot indicates the cumulative layer-wise changes until a specific layer. Each row indicates the corresponding token's attention-norms to every token in the sentence (including itself). Although the changes slightly spread out to the other tokens as we move up to higher layers, they mostly occur in the bi-gram area. Given BERT's contextualization mechanism, variations in attention-norms in each row directly result in a change in the corresponding token's representation. Therefore, the tokens in the bi-gram undergo most changes in their representations.

Figure 9 indicates that after inverting a bi-gram, both words' attention-norms to their neighboring

[6]The value of the cell β_{ij} (i^{th} row, j^{th} column) in the map denotes the attention-norm of the i^{th} token to the j^{th} token. The contextualized embedding for the i^{th} token is constructed based on a weighted combination of their corresponding attention-norms in the i^{th} row.

tokens change and this mostly affects their own representations rather than others. This observation suggests that the distinction formed between the representations of the original and shifted tokens, as was seen in Figure 7, can be rooted back to the changes in attention heads' patterns.

6.3 Phrasal-level inversion

The Coordination Inversion (**CoordInv**) task is a binary classification that contains sentences with two coordinated clausal conjoints (and only one coordinating conjunction). In half of the sentences the clauses' order is inverted and the goal is to detect malformed sentences at phrasal level (Conneau et al., 2018). Since the phrasal-level inversion does not alter the syntax structure of the sentence, the task could be considered as a semantic one (Conneau et al., 2018). For an example:

the glass broke and i cut myself . → Original

i cut myself and the glass broke . → Inverted

While both sentences are syntactically correct, we should rely on the meaning of the sequence of the events in order to detect the abnormality in the second sentence.

BERT's performance on this task increases through layers and then slightly decreases in the last three layers. We observed that the attribution scores for "but" and "and" coordinators to be among the highest (see Appendix A.2) and that these scores notably increase through layers. We hypothesize that BERT might implicitly encodes phrasal level abnormalities in specific token representations.

Odd Coordinator Representation. To verify our hypothesis, we filtered the test set to ensure all sentences contain either a "but" or an "and" coordinator[7]. We reconstructed the original examples by inverting the order of the two clauses in the inverted instances since no sentence appears with both labels in the dataset. Feeding this to BERT, we extracted token representations and computed the cosine distance between the representations of each token in the original and inverted sentences. Figure 10 shows these distances, as well as the normalized saliency score for coordinators (averaged on all examples in each layer), and layer-wise performance for the CoordInv probing task. Surprisingly, all

[7] 9,883 of the 10K examples in the test set meet this condition.

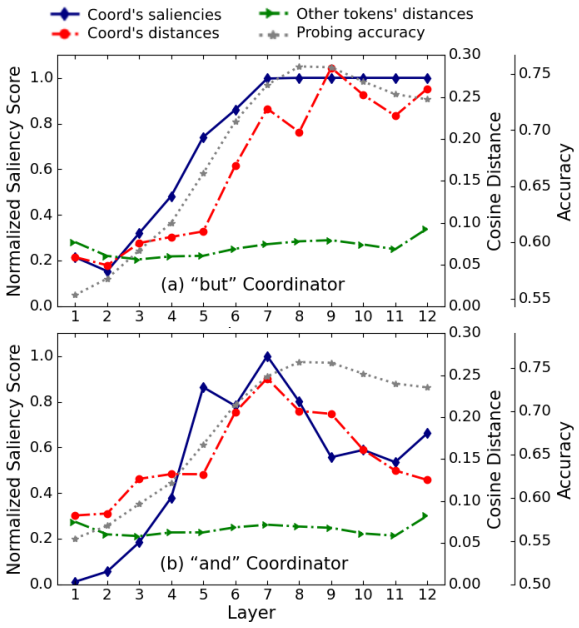

Figure 10: Averaged cosine distances between coordinators in the original and inverted sentences. We also show the normalized saliency scores for the coordinators across layers which correlate with the performance scores of the task. The distance curve for other tokens is a baseline to highlight that the representation of coordinators significantly change after inversion.

these curves exhibit a similar trend. As we can see, when the order of the clauses are inverted, the representations of the coordinators "but" or "and" play a pivotal role in making sentence representations distinct from one another while there is nearly no change in the representation of other words. This observation implies that BERT somehow encodes oddity in the coordinator representations (corroborating part of the findings of our previous analysis of BShift task in Section 6.1).

7 Control Experiments

The main motivation behind designing a control task in probing studies is to check whether it is the representations that encode linguistic knowledge or the diagnostic classifier itself which plays a significant role in learning nuances of the task (Hewitt and Liang, 2019). In this regard, most of our experiments throughout the paper (similarity curves, tSNE plots, or attention-norm analysis) all rely on fixed representations and do not need any classifier or training; hence, they all serve as control experiments or sanity checks. For example, in our attention-norm analysis (which requires no training and comes from a different perspective) we arrive at the same results as our attribution analysis.

Task	Pearson's r		Spearman's ρ	
	Mean	Max.	Mean	Max.
SentLen	0.80	0.97	0.71	0.94
ObjNum	0.60	0.91	0.63	0.98
Tense	0.84	0.93	0.67	0.91
BShift	0.84	0.92	0.79	0.87
CoordInv	0.63	0.90	0.54	0.85

Table 1: The Pearson's r and Spearman's ρ correlations averaged over all examples, reporting the mean and maximum values across all layers.[8]

Computation of attribution scores based on trained diagnostic classifiers is the only part of our experiments which involves a training procedure. Hence, we carried out a control study inspired by Talmor et al. (2020) to check the consistency of attribution patterns. The intuition behind this is in line with Voita and Titov (2020) who stated that if there is a strong regularity in the representations with respect to the labels, this can be revealed even with fewer training data points.

To this end, we used only 10% of the training data to train the diagnostic classifiers and computed the attribution scores for each task. Then, we computed the correlation between attribution scores for each sentence obtained by these classifiers and those obtained from the original classifiers (trained on full training data). After averaging the correlations over all examples, we report the mean and maximum statistics among all layers in Table 1. The strong correlations imply that a similar pattern exist in the attribution scores even when fewer training instances are used. This highlights the fact that task-specific knowledge is well encoded and regularized in the representations, nullifying the possibility of the classifier playing a major role.

8 Conclusions

In this paper we carried out an extensive gradient-based attribution analysis to investigate the nature of BERT token representations. To our knowledge, this is the first effort to explain probing performance results from the viewpoint of token representations. We found that, while most of the positional information is diminished through layers, sentence-ending tokens are partially responsible for carrying this knowledge to higher layers in the model. Furthermore, we analyzed the grammatical number and tense information throughout the model. Specifically, we observed that BERT tends to encode verb tense and noun number information in the `##s` token and that it can clearly distinguish the two usages of the token by separating them into distinct subspaces in the higher layers. Also, we found that abnormalities can be captured by specific token representations, e.g., in two consecutive swapped tokens or a coordinator between two swapped clauses.

Our approach in using a simple diagnostic classifier and incorporating attribution methods provides a novel way of extracting qualitative results in probing studies. This can be seamlessly applied to various deep pre-trained models, providing a wide range of options in sentence-level tasks and from the fine-grained viewpoint of tokens. We hope this will spur future probing studies in other evaluation scenarios. Future work might investigate how subspaces are evolved or transformed during fine-tuning and whether they are beneficial at inference time to various downstream tasks (e.g. syntactic abnormalities, grammatical number and tense subspaces in grammar-based tasks like CoLA Warstadt et al., 2019) or to check whether these behaviors are affected by different training objectives. Furthermore, our token-level analysis can provide insights for enhancing model efficiency based on token importance, something we plan to pursue in future work.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful suggestions.

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.

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.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46.

[8] Results are averaged over three runs.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online.

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.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia.

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.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. 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 2185–2194, Hong Kong, China.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Xiaochuang Han, Byron C. Wallace, and Yulia Tsvetkov. 2020. Explaining black box predictions and unveiling data artifacts through influence functions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5553–5563, Online.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota.

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.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy.

Josef Klafka and Allyson Ettinger. 2020. Spying on your neighbors: Fine-grained probing of contextual embeddings for information about surrounding words. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4801–4811, Online.

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.

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.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*.

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.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. 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 1073–1094, Minneapolis, Minnesota.

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

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. In *Advances in Neural Information Processing Systems*, pages 8594–8603.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California.

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.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

D Smilkov, N Thorat, B Kim, F Viégas, and M Wattenberg. 2017. Smoothgrad: removing noise by adding noise. arxiv. *arXiv preprint arxiv:1706.03825*.

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*, pages 3319–3328.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy.

Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. 2020. The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 107–118, Online.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2018. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. 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 4396–4406, Hong Kong, China.

Elena Voita, Rico Sennrich, and Ivan Titov. 2021. Analyzing the source and target contributions to predictions in neural machine translation. 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 1126–1140, Online.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online.

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. AllenNLP interpret: A framework for explaining predictions of NLP models. 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): System Demonstrations*, pages 7–12, Hong Kong, China.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT make any sense? interpretable word sense disambiguation with contextualized embeddings. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers*, pages 161–170, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

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.

Zhengxuan Wu and Desmond C Ong. 2021. On explaining your explanations of bert: An empirical study with sequence classification. *arXiv preprint arXiv:2101.00196*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

Hao Yuan, Yongjun Chen, Xia Hu, and Shuiwang Ji. 2019. Interpreting deep models for text analysis via optimization and regularization methods. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5717–5724.

Zining Zhu and Frank Rudzicz. 2020. An information theoretic view on selecting linguistic probes. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9251–9262, Online.

A Appendices

A.1 Probing Performance Results

Table A.1 shows the results for our average sentence representation strategy (cf. Section 3.1) for all layers and across all tasks. We trained the diagnostic classifiers three times and reported the expected test performance for each task (Dodge et al., 2019). Each task consists of 100k examples for training and 10k examples for validating the diagnostic classifiers. The test set includes 10k examples that are used for our evaluation and in-depth analysis. All dataset splits are balanced for their target classes. The performance trends in our experiments are similar to those observed by Jawahar et al. (2019).

Layer	SentLen	BShift	Tense	ObjNum	CoordInv
1	0.91	0.57	0.87	0.82	0.55
2	0.89	0.72	0.88	0.83	0.57
3	0.89	0.84	0.88	0.84	0.60
4	0.85	0.87	0.89	0.85	0.62
5	0.81	0.88	0.89	0.86	0.66
6	0.78	0.90	0.89	0.86	0.71
7	0.70	0.91	0.89	0.86	0.74
8	0.68	0.91	0.89	0.85	0.75
9	0.59	0.91	0.89	0.85	0.76
10	0.49	0.91	0.89	0.84	0.74
11	0.43	0.90	0.89	0.83	0.73
12	0.44	0.89	0.89	0.83	0.72

Table A.1: Layer-wise performance scores (accuracy) for the average sentence representation strategy on different probing tasks.

Mean Pooling vs. [CLS] Pooling. In order to show the reliability of our average-based pooling method for probing BERT, in Table A.2 we provide a comparison against the [CLS] methodology of Jawahar et al. (2019). Specifically, we show layer-wise performance differences of the two representations, with the green color indicating improvements of our strategy. The results clearly highlight that average representations are more suited to the task, providing improvements across many layers in most tasks.

A.2 Full 12-layer Figures

In this section we provide the full 12-layer version of the previous summarized layer-wise figures.

A.3 SubjNum Mislabelling

The SubjNum probing data suffers from numerous incorrect labels which are more obvious within samples which starts with a name that ends with an "s" and labelled as plural. We show five examples with this issue in Table A.3.

Layer	SentLen	BShift	Tense	ObjNum	CoordInv
1	+0.03	+0.07	+0.05	+0.07	+0.02
2	+0.05	+0.17	+0.03	+0.03	+0.02
3	+0.12	+0.19	+0.02	+0.06	+0.02
4	+0.15	+0.14	+0.02	+0.04	+0.05
5	+0.24	+0.07	0.00	+0.05	+0.03
6	+0.3	+0.08	-0.01	+0.06	+0.02
7	+0.31	+0.08	0.00	+0.05	0.00
8	+0.29	+0.07	0.00	+0.04	-0.01
9	+0.24	+0.04	0.00	+0.04	-0.02
10	+0.17	+0.04	0.00	+0.04	-0.04
11	+0.15	+0.04	0.00	+0.04	-0.04
12	+0.19	+0.02	0.00	+0.05	-0.03

Table A.2: Layer-wise performance scores comparison between average and [CLS] representations across different probing tasks. Average pooling retains or improves [CLS] performance in all layers and tasks, except for some layers in CoordInv.

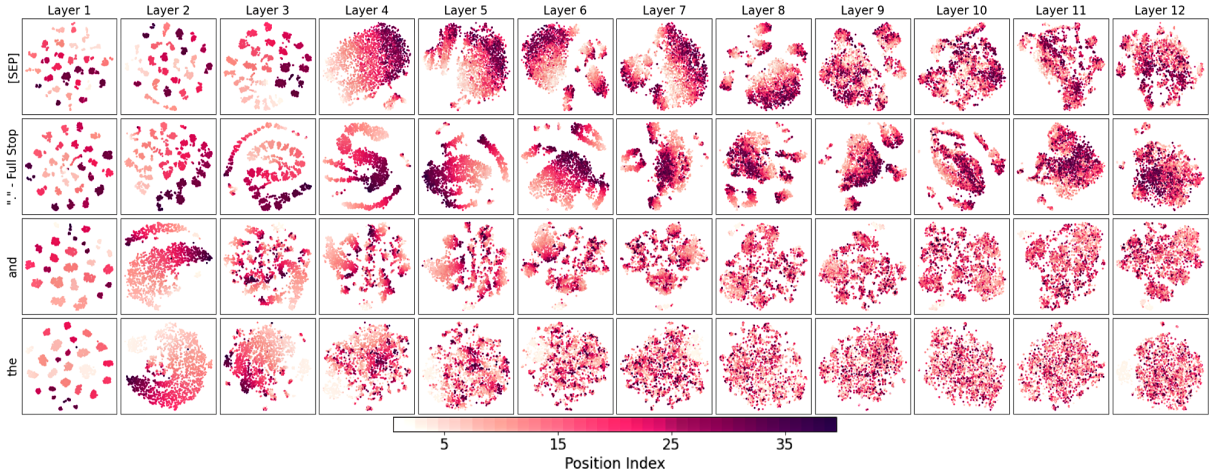

Figure A.1: Full 12 layer t-SNE plots – (Figure 2)

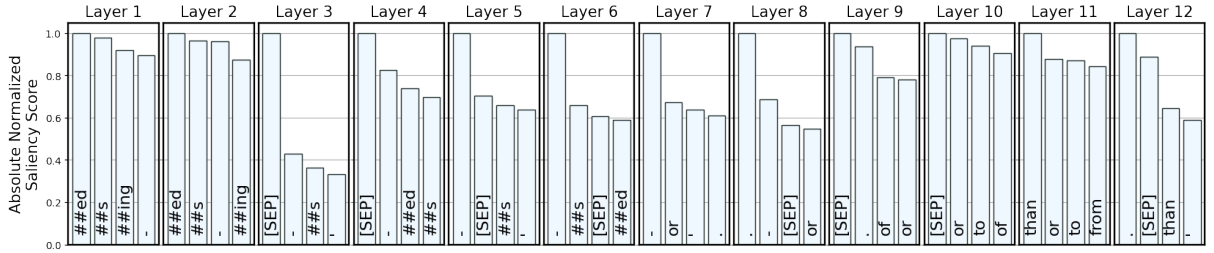

Figure A.2: Full 12 layers for the top-4 most attributed high frequency tokens in the Sentence Length task - (Figure 1)

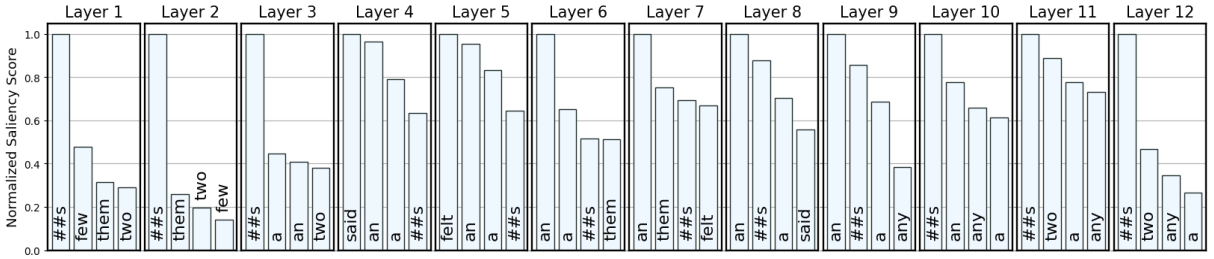

Figure A.3: Full 12 layers for the top-4 most attributed high frequency tokens in the ObjNum task - (Figure 3)

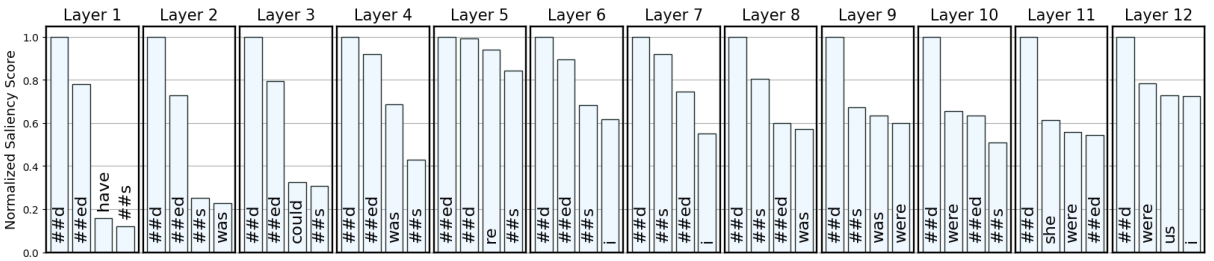

Figure A.4: Full 12 layers for the top-4 most attributed high frequency tokens in the Tense task - (Figure 3)

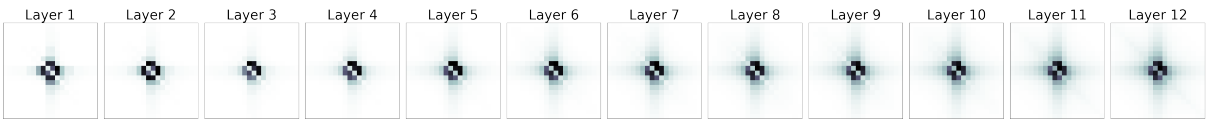

Figure A.5: Full 12 layers attention-norm changes (Δ_{β^l}) - (Figure 9)

Label	Sentence
NNS	Zeus is the child of Cronus and Rhea , and the youngest of his siblings .
NNS	Jess had never done anything this wild in her life .
NNS	Lois had stopped in briefly to visit , but didn ’t stay very long .
NNS	Tomas sank back on the seat , wonder on his face .
NNS	Justus was an unusual man .

Table A.3: Five examples from SentEval’s SubjNum data that are incorrectly labelled as plural (NNS) while the subject is clearly singular (NN). There are numerous such mislabeled instances in the test set.

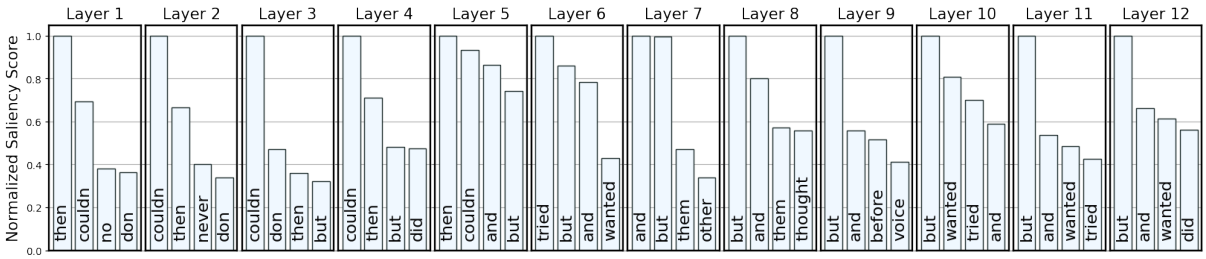

Figure A.6: Full 12 layers for the top-4 most attributed high frequency tokens in the CoordInv task