Layer by Layer - Examining BERT's Syntactic and Semantic Representation

Harsh Dubey [1] **Yulu Qin** [2] **Rahul Meghwal** [3]

Abstract

BERT has achieved remarkable results across various language understanding benchmarks, suggesting that it captures structural information about language. In this work, we conduct a series of experiments to examine the elements of English language structure learned by BERT. Our findings reveal that BERT's lower layers capture phrase-level information in its phrasal representations. Additionally, BERT's intermediate layers encode a rich hierarchy of linguistic information, with surface features at the bottom, syntactic features in the middle, and semantic features at the top. Furthermore, BERT requires deeper layers to handle long-distance dependency information, such as tracking subject-verb agreement. This study contributes to the growing body of research aiming to enhance our understanding of BERT's inner workings and the extent of its language understanding capabilities.

1. Introduction

As we find ourselves at the cutting edge of AI advancement, ChatGPT has attracted considerable attention due to its remarkable performances. Before debating whether a Large Language Model (LLM) like ChatGPT exhibits human-level intelligence or language learning capabilities, we must first investigate what representations it has acquired and its learning process.

Noam Chomsky (1957) introduced the groundbreaking concept of "transformational-generative grammar," suggesting that humans inherently possess the capacity to process and generate hierarchical structures in language. With the evolution of neural networks, researchers have explored whether these networks can learn implicit hierarchical structures for syntactic and semantic processing from sequential inputs.

Looking back at the history of neural networks for language processing, Elman initially introduced the idea of incorporating time implicitly into representational structures, developing the Simple Recurrent Network (SRN) for efficient sequence processing. Subsequently, enhanced versions of Recurrent Neural Networks (RNNs) allowed models to process text sequences while retaining memory of prior inputs. These architectures achieved considerable success in sequence-to-sequence tasks. Bahdanau et al. (2014) introduced attention mechanisms, revolutionizing NLP by enabling models to assign varying importance to different parts of the input sequence, thereby improving their ability to capture long-range dependencies. Vaswani et al. (2017) presented the transformer architecture, which relied exclusively on self-attention mechanisms and showcased exceptional performance in NLP tasks. This architecture laid the groundwork for many contemporary LLMs, including ChatGPT.

Understanding transformer-based models has become increasingly complex compared to its predecessors due to the exponential growth in size. Thus, interpretability of transformer-based models has surfaced as a crucial area of research in the field.

In this project[1], we investigate BERT(Bidirectional Encoder Representations from Transformers), a prominent transformer-based model that received substantial attention before ChatGPT's debut. (Devlin et al., 2018) Although it utilizes distinct training approaches compared to ChatGPT, the underlying architecture is similar, providing valuable insights into how transformer-based language models learn hierarchical representations during the training phase.

2. Related Work

In Belinkov's survey (2019), various methods have been explored to study the syntactic representation within neural networks. One widely used approach is visualization. Elman was the first to demonstrate that Simple Recurrent Networks (SRNs) can acquire word representations reflecting both lexical and syntactic categories via hierarchical clustering (Elman, 1990). He also pioneered the visualization of hidden unit activations in SRNs and their correspondence to specific grammatical relations, such as number agreement

[1]hd2225 [2]yq810 [3]rm5707. Correspondence to: Yulu Qin <yq810@nyu.edu>.

[1]Code is publicly accessible at https://github.com/LLM-CCM/Lexical-Syntactic-Structure-LLM

(Elman, 1991). Subsequent research has built upon this tradition. Brunner et al. (2018) trained an RNN encoder in a multitask learning setup and visualized the clustering of sentence embeddings. Karpathy et al. (2015) presented analysis and visualization techniques for character-level RNNs. Linzen et al. (2016) studied the number agreement Long Short-Term Memory (LSTM) network's inner mechanisms by examining its activation in response to particular syntactic structures. However, it is worth noting that assessing the quality of these visualizations remains a challenge. (Belinkov & Glass, 2019)

Probing is another widely-used method for examining a language model's ability to understand syntax by diagnostic classifiers known as probes. These classifiers are designed to predict specific linguistic properties such as parts-of-speech, from word, phrase, or sentence representations of a pretrained model. Probes trained on various representations have demonstrated remarkable accuracy in tasks involving morphological and part-of-speech information, as well as syntactic and semantic information, highlighting the effectiveness of these approaches in language understanding tasks.(Belinkov & Glass, 2017; Peters et al., 2018; Tenney et al., 2019). To assess the quality of a probe, Zhang and Bowman (2018) introduced random representation baselines. Hewitt and Liang (2019) further suggested control tasks and argued a proficient probe should exhibit selectivity, characterized by high linguistic task accuracy and low control task accuracy. Later, Conneau and Kiela (2018) from Facebook AI Research presented SentEval, a widely accepted framework for evaluating the quality of universal sentence representations. In our study, we employ the SentEval toolkit to construct a binary probe classifier, adhering to the suggested hyperparameter space.

Everaert et al. (2015) suggest that subject-verb agreement exemplifies the concept that words in sentences follow "structures, not strings." Early analysis on the syntactic capabilities of neural networks primarily focused on the long-distance agreement between subjects and verbs, as described in (Linzen & Baroni, 2021). Linzen (2016) first devised the number prediction task to examine syntactic representation in a straightforward manner. This involved introducing experiments with an increasing number of intervening nouns between the subject and the verb. These intervening nouns, called attractors, were identified as the primary source of occasional errors in language production (Bock & Miller, 1991) and comprehension (Nicol et al., 1997) for both models and humans. To accurately predict the verb's number, the neural network must implicitly analyze the sentence structure and avoid being misled by the nearby, yet irrelevant attractor.

Linzen's experiments revealed that as the number of attractors increased, accuracy decreased, but still remained robust at 82%. (Linzen et al., 2016) Bernardy and Lappin (2017) demonstrated that GRU and CNN models could also successfully handle the number prediction task, suggesting that this outcome is generalizable across various deep neural network models.

Subsequently, Goldberg (2019) took one important step forward to examine BERT, an attention-based model, and found that the model effectively captures syntactic information for subject-verb agreement. Building on Goldberg and Linzen's work, Jawahar et al.(2019) performed layer-by-layer tests on BERT, accounting for the number of attractors. We continued this line of research and carried out a more comprehensive analysis.

3. Phrasal Representation

(Peters et al., 2018) attempted to dissect contextual word embeddings to understand the nature of information captured at both syntax and semantics levels. These embeddings have gained popularity due to their ability to comprehend and capture meanings based on contextual cues from their surroundings. However, the results of their study do not directly apply to the newer modeling architectures supported by Transformers, such as BERT. Even if we assume that these results may be applicable, given the complexity of these models, it becomes challenging to analyze where and how this knowledge is being captured.

The authors proposed various methods to investigate, analyze, and comprehend the structure, functioning, and superior performance of these models in linguistic tasks. They utilized span representations to capture phrase-level or span-level information. Building upon these investigations, our approach follows the same idea of extracting span representations from each layer of BERT. This approach can help us answer the crucial question we aim to investigate: whether Transformer-based models like BERT capture span-level information and how this information is distributed across layers.

3.1. Methodology

We utilize the CoNLL 2000 chunking dataset, randomly selecting 3000 labeled chunks and 500 non-chunk spans. The investigation method begins with the process of extracting span representations. This process involves capturing the representation at each layer l for a token sequence $s_i, ...s_j$, denoted as $s(s_i, s_j)$. To achieve this, we concatenate the first and last hidden vector representations at each layer, along with their element-wise multiplication and difference.

To visualize the span representations layer by layer, we employ t-SNE, a dimensionality reduction algorithm that represents the embeddings in a 2D space. The resulting t-SNE visualizations per layer are depicted in the attached

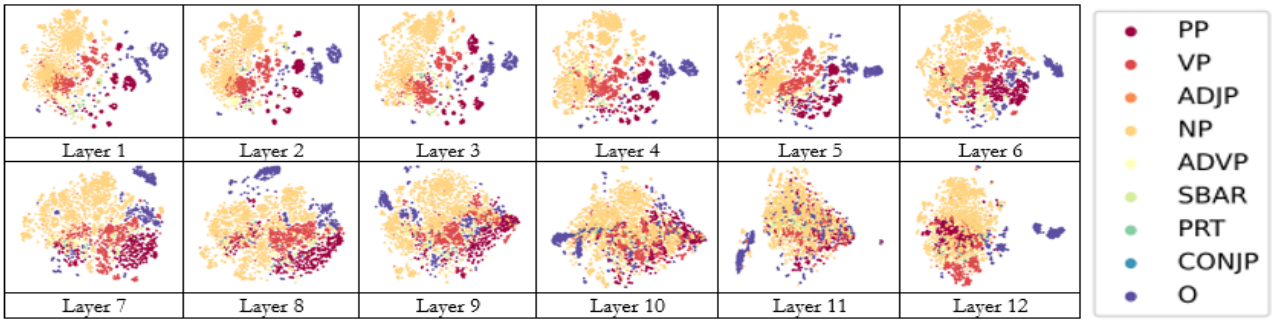

Figure 1. 2D t-SNE plot of span embeddings computed from the layers of BERT.

Layer	1	2	3	4	5	6	7	8	9	10	11	12
NMI	0.42	0.35	0.35	0.26	0.23	0.21	0.19	0.15	0.16	0.18	0.17	0.19

Figure 2. Clustering performance of span representations obtained from different layers of BERT.

Figure 1. Additionally, to support our observations, we perform k-means clustering on the span representations, with k representing the different chunk types.

The final step involves using the NMI (Normalized Mutual Information) metric to compare the information captured among layers.

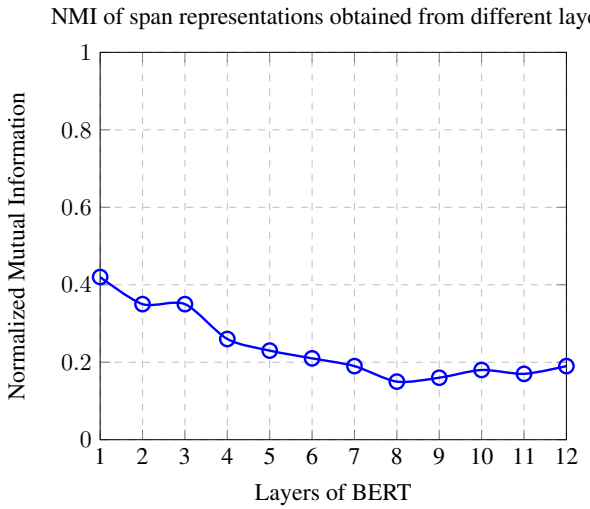

Figure 3. Clustering performance visualization from different layers of BERT.

3.2. Results

Our observations from Figure 1 and 3 indicate that BERT predominantly captures phrase-level information in the lower layers, and this information gradually diminishes as we move to higher layers. In the lower layers, the span representations exhibit a tendency to map chunks together with their corresponding underlying category. To quantify this observation, we conducted k-means clustering on the span representations using k = 10, representing the number of distinct chunk types. The evaluation of the resulting clusters using the NMI metric in Figure 2 reaffirms that the lower layers of BERT better encode phrasal information compared to the higher layers.

3.3. Limitation:

While the result provides valuable insights into visualizing and qualitatively analyzing span representations, it would benefit from incorporating quantitative analysis methods to strengthen the analysis of the captured information. Additionally, the limitations of the dataset used in the study may restrict the generalizability of the results. To enhance the arguments regarding linguistic structure representation, further analysis using diverse and more comprehensive data sources is recommended. It is important to note that the current method primarily provides a superficial examination of the representations and may not fully capture syntactic and semantic dependencies across longer spans. Therefore, it is necessary to combine these findings with other methods to draw more conclusive insights.

4. Syntactic and Semantic Understanding via Probing

In the past, traditional machine learning methods for language understanding and natural language processing relied on incorporating linguistic features, such as part-of-speech

tags, syntactic dependencies, and frequency metrics. These features were well-established and provided interpretable insights into the underlying linguistic properties of text. However, with the advent of large language models, the use of traditional linguistic features has diminished.

Contemporary NLP approaches leverage word representations derived from large language models. These word representations, despite their superior performance, have one drawback: these models are black boxes, lacking interpretability of the complex patterns they capture. To address these issues, researchers have advocated for the use of probing techniques.

Probing involves employing specific diagnostic tasks to extract hidden linguistic knowledge from these black box representations.

4.1. Overview of probing task

Probing tasks, also known as diagnostic tasks, have been widely used in previous studies by (Yossi Adi, 2018), and (Conneau et al., 2018) to uncover the linguistic features encoded by neural networks. These tasks involve setting up auxiliary classification tasks where the final output of a model is used as features to predict a specific linguistic phenomenon. If the auxiliary classifier performs well in predicting the linguistic property, it indicates that the original model likely encodes that property. In this study, we will employ probing tasks to evaluate the ability of individual layers in language models to encode different types of linguistic features.

4.2. Exploring Syntactic Structure with Basic Probes: Initial Intuition

(Guillaume Alain, 2016) suggested that it is possible to explore the interpretability of deep neural networks using Linear Classifier Probes, which can be used to investigate the hidden representations being captured by the layers. They suggested that early layers capture low-level features, while deeper layers capture high-level features and dependencies.

(Tenney et al., 2019) investigated the extent to which contextualized word representations encode sentence-level structural information. They proposed a probing framework that required the model to predict linguistic properties related to sentence structure, including part-of-speech tags, constituent labels, and dependencies. They compared the representations at different layers of the models to investigate whether sentence structure information is present across layers. The models achieved high accuracy on the probing tasks, indicating their ability to capture syntactic and semantic properties.

(Hewitt & Liang, 2019) endorsed probing as a method to interpret representations learned by machine learning models but raised concerns about the accuracy of its interpretations. They posed a serious question about whether the probes are simply getting good at learning the task itself. It is important to avoid overestimating the extent to which high probe accuracy reflects the properties of the representation. They proposed an approach called control tasks to enhance the interpretability of the probing results.

Control Tasks aim to measure the extent to which a probe's performance reflects the linguistic property being probed. The Probe Confounder Problem suggests that complex neural networks are capable of memorizing a large number of labeling decisions independent of probing tasks. As a result, probes may receive high accuracy without truly reflecting the represented property. To address this issue, Hewitt and Liang (2019) introduced the concept of selectivity, which helps evaluate probes and raises concerns about representational qualities.

4.2.1. Methodology

In this paper, we design a low-dimensional MLP probe that can be used to capture part-of-speech tagging linguistic tasks while being both selective and non-selective.

We have selected part-of-speech (POS) tagging as the specific NLP task for our study. To accomplish this, we will utilize a portion of the English Web dependency treebank from the Universal Dependencies project. This dataset provides valuable information such as POS labels, morphological features (tense, gender, number, etc.), and dependency labels (subject, object, etc.), making it suitable for exploring different aspects of language.

In order to measure the selectivity of layers in capturing linguistic information, we introduce a control task that is unrelated to the part-of-speech (POS) tagging task. Following the approach suggested by Hewitt and Liang (2019), we assigned a random POS tag to each word identity based on the distribution of these tags in the dataset. Importantly, each word identity consistently receives the same tag every time it appears.

During training and testing, the layers of the model will be tasked with predicting the assigned tag based on the word embeddings. It is worth noting that the assigned tag is solely determined by the word identity and is therefore a deterministic function. Consequently, if a layer exhibits high selectivity, it indicates that the embedding has forgotten or disregarded certain information related to the word identity, as the predicted tag accuracy on the control task will differ from that on the actual POS tagging task. By comparing the probed accuracy of the layers on the POS task and the control task, we can assess the selectivity of each layer in capturing POS-specific information.

4.2.2. RESULTS

The findings in Figure 4 and 5 from the probing tasks indicate that BERT's representations do capture structural information to some extent. However, the results vary depending on the experimental design employed. In one design, using a probing classifier without control tasks, it is observed that all layers of BERT exhibit high levels of syntactic information, with the middle layers demonstrating the most accurate representation of the structure. On the other hand, another design incorporates a control task to guide the probing classifier and reveals that although the classification accuracy of the probes decreases with deeper layers, the selectivity score increases. It is worth noting that popular probe design choices tend to yield high accuracy in the control task but low selectivity in the probes, indicating that they are capable of memorizing a large number of decisions without being motivated by the underlying representation. This suggests that the deeper layers, which exhibit high selectivity, contain the most informative representations.

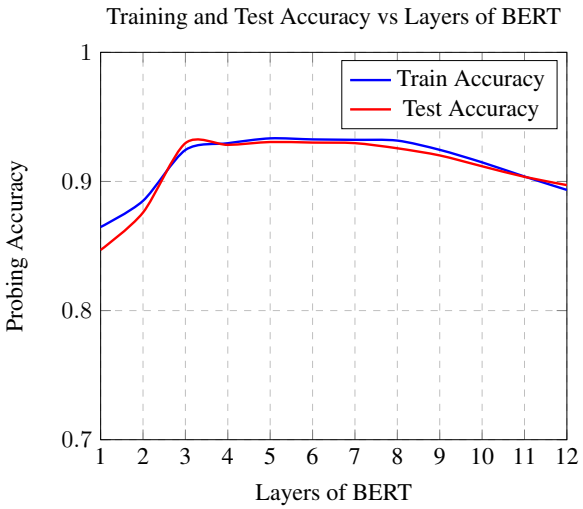

Figure 4. Probing Task performance visualization without Selectivity.

4.2.3. LIMITATIONS

The probing method used in this study, offers valuable insights into enhancing the interpretability of probing results. However, it is important to acknowledge that this approach has certain limitations. While POS tags contribute to the syntactic structure of a sentence, the method may not be applicable or generalizable to all types of probing tasks and linguistic properties.

The results obtained from the probing tasks demonstrate that some structural information is indeed captured by BERT's representations. Additionally, the findings indicate that the initial layers of BERT contain more syntactic information

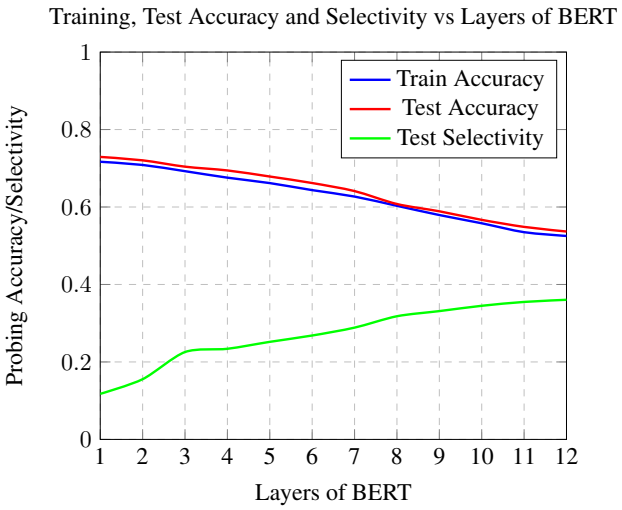

Figure 5. Probing Task performance visualization with Selectivity.

compared to deeper layers for parts-of-speech tasks. However, these results alone may not be satisfactory for fully understanding the syntax and semantics of BERT for language understanding.

Therefore, there is a need for better probes or probing frameworks that can delve into the hidden representations of BERT from multiple dimensions. To address this, the study employs complicated probes from the Facebook research SentEval tool to extract hidden representations from each layer of BERT, aiming to gain a more comprehensive understanding of the model's internal structure. (Conneau & Kiela, 2018)

4.3. Exploring Representation Interpretability with Complex Probes: Further Analysis

(Conneau et al., 2018) proposed ten probing tasks to assess the linguistic properties captured by sentence embeddings. These tasks encompass surface, syntactic, and semantic dimensions, providing a comprehensive evaluation of language models. The surface tasks include SentLen (sentence length) and WC (word content). Syntactic tasks involve BShift (word order sensitivity), TreeDepth (depth of syntactic trees), and TopConst (sequence of top-level constituents). Semantic tasks encompass Tense (tense information), SubjNum and ObjNum (subject and direct object number), SOMO (sensitivity to noun/verb replacements), and CoordInv (swapping coordinated clauses).

To evaluate the layers of BERT, the authors utilized the SentEval toolkit Conneau et al. 2018 and optimized hyperparameters to find the best probing classifier. The findings reveal that BERT exhibits a hierarchical organization of linguistic signals across its layers. Lower layers primar-

ily encode surface-level information, middle layers capture syntactic properties, and higher layers focus on semantic information. This suggests that BERT's representations progressively capture more abstract and complex linguistic features as we move up the layers.

Using the probing tasks highlights the capability of BERT to encode a diverse range of linguistic properties. It demonstrates that the model's layers have distinct roles in representing different linguistic dimensions, contributing to our understanding of how information is captured and organized in sentence embeddings.

4.3.1. Methodology

In our study, we have selected two probing tasks from the set of ten tasks proposed by (Conneau et al., 2018) to investigate the linguistic signals captured by BERT. Our focus is on examining both syntactic and semantic understanding, so we have chosen the TreeDepth probing task for syntax and the SubjNum probing task for semantics.

For training the probes, we followed the same dataset guidelines as suggested by (Conneau et al., 2018). The datasets consist of 100,000 training instances, 10,000 validation instances, and 10,000 test instances, ensuring a balanced distribution across the target classes.

The TreeDepth probing task is a classification task where the objective is to predict the maximum depth of the syntactic tree for a given sentence. The values range from 5 to 12, and since sentence length is naturally correlated with sentence depth, we designed a target bivariate Gaussian distribution that establishes the relationship between the two. To create a decorrelated sample, we selected a subset of sentences that match this distribution.

In the SubjNum task, we focus on the number of subjects in the main clause, which involves binary classification between singular (NN) and plural or mass forms (NNS). We limited our selection to target noun forms with corpus frequencies between 100 and 5,000, ensuring a balanced distribution of noun forms across the dataset partitions.

By implementing these two probing tasks, we aim to evaluate how well BERT captures sentence-depth information and its ability to discern the number of subjects in a sentence.

4.3.2. Results

The results obtained from the probing tasks confirm that BERT's representations encompass both structural and semantic information. Consistent with the original paper's proposal, the findings suggest that the middle layers of BERT contain a higher degree of syntactic information, while the deeper layers exhibit stronger semantic representations. This implies that certain layers are more specialized in capturing specific linguistic aspects for performing particular tasks. However, it also highlights that the other layers maintain a contextual understanding of linguistic structures and rules. The distribution of information for linguistic comprehension, spanning surface-level features, syntax, and semantics, is observed across multiple layers. Table 1 provides evidence that the middle layer, particularly layer 7, achieves the highest probing accuracy in capturing syntactic representations, while the deeper layers, such as layer 9, excel in representing semantic features.

Layer	TreeDepth (Syntactic)	SubjNum (Semantic)
1	32.08	73.97
2	34.67	79.02
3	34.43	79.27
4	33.73	79.49
5	34.19	82.6
6	35.03	86.25
7	**35.35**	86.09
8	34.46	86.24
9	34.68	**86.83**
10	34.24	86.21
11	33.2	84.09
12	31.47	82.37

Table 1. In this table, *Layer* refers to the current layer for which the representations are being captured, while "TreeDepth", and "SubjNum" denote the test accuracy of these probing tasks for each BERT *Layer*.

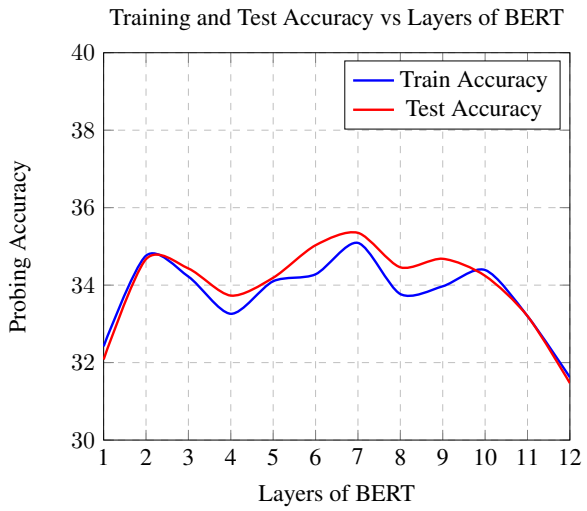

Figure 6. Probing Task performance visualization for TreeDepth Probing.

4.3.3. Limitations

The study demonstrates that probing tasks effectively capture linguistic properties in sentence embeddings. However,

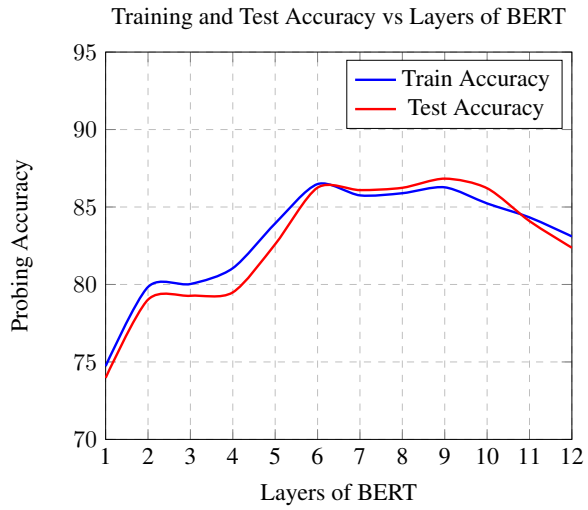

Figure 7. Probing Task performance visualization for SubjNum Probing.

the analysis can be improved by expanding the range of probing tasks to cover a broader set of linguistic properties. While probing tasks provide valuable insights, they are not sufficient on their own to fully understand how language models learn and represent language. Alternative methods should be explored to gain deeper insights into the interpretability of these models. Additionally, investigating linguistic properties at the word-level and document-level representations would provide a more comprehensive understanding. Developing enhanced representational methods is crucial to extract and interpreting the knowledge obtained from probing tasks, leading to deeper insights into the encoded linguistic properties. In summary, while probing tasks are valuable, further advancements can be achieved by expanding the range of tasks, exploring different levels of linguistic analysis, and developing sophisticated representational methods.

5. Subject-verb agreement

5.1. Background

While subject-verb agreement is not the only structure-sensitive dependency, a model's success in mastering it can strongly suggest its ability to learn hierarchical structures.(Linzen et al., 2016) A key issue that arises for models lacking sensitivity to the structure is the potential for agreement attraction errors. (Bock & Miller, 1991) Agreement attractors refer to intervening nouns that have a different number than the subject. The presence of such attractors necessitates that the model correctly identifies the syntactic subject's head corresponding to a specific verb, in order to select its correct inflected form. It is plausible that selecting correct forms through simple strategies, such as "agreeing with the most recent noun," could be easily applied by sequence models. In numerous datasets, the model might still exhibit good performance, as the majority of sentences are simple and follow this heuristic. However, this approach can be unreliable because agreement attractors may appear between the subject and verb within the linear arrangement of a sentence.

Linzen et al. conducted an experiment to evaluate LSTM's syntactic capabilities by increasing the number of intervening nouns, resulting in an 82% overall test accuracy. Goldberg (2019) later demonstrated that BERT effectively learns syntactic structures for subject-verb agreement, using various stimuli. Jawahar et al.(2019) expanded on this work by testing each layer of BERT and controlling the number of attractors. Our research built upon Jawahar's method and carried out a more extensive analysis.

Layer	Overall	0	1	2	3	4
1	87.51	90.58	35.4	21.74	21.95	23.36
2	88.98	91.89	40.37	24.11	23.41	22.65
3	89.92	92.64	44.78	28.91	25.02	27.96
4	91.9	94.29	52.93	34.6	30.69	31.15
5	93.1	95.05	62.48	42.4	37.1	35.4
6	93.56	95.38	65.59	44.63	37.1	32.57
7	93.92	95.46	70.55	50.91	40.91	38.23
8	**94.23**	**95.66**	**72.71**	**53.51**	44.77	41.24
9	93.96	95.46	70.86	53.44	**45.7**	**44.96**
10	93.39	94.98	68.69	51.34	43.01	41.59
11	92.49	94.28	64.21	47.41	39.64	35.58
12	91.78	93.71	60.38	46.96	39.93	38.05

Table 2. In this table, $Layer$ refers to the current layer for which the representations are being captured, while "Overall" represents the overall test accuracy for the sva task, and "0-4" denotes the test accuracy given that number of intervening nouns in the inputs for each BERT $Layer$.

5.2. Methodology

We employed the stimuli developed by Linzen et al. (2016) and the SentEval toolkit (2018) to optimally configure binary classifiers. As we increased the number of intervening nouns, we documented the classifier's performance based on representations extracted from each layer of BERT. The training process consisted of two phases. In the first phase, we extracted features from the pre-trained BERT model using the stimuli. In the second phase, we trained the classifier and assessed its performance.

5.3. Results And Analysis

After conducting the experiments layer by layer, we present the results in Figure 8 and highlights the best results in Table 2. The graph demonstrates that as the number of intervening nouns increases, the probe's performance declines signif-

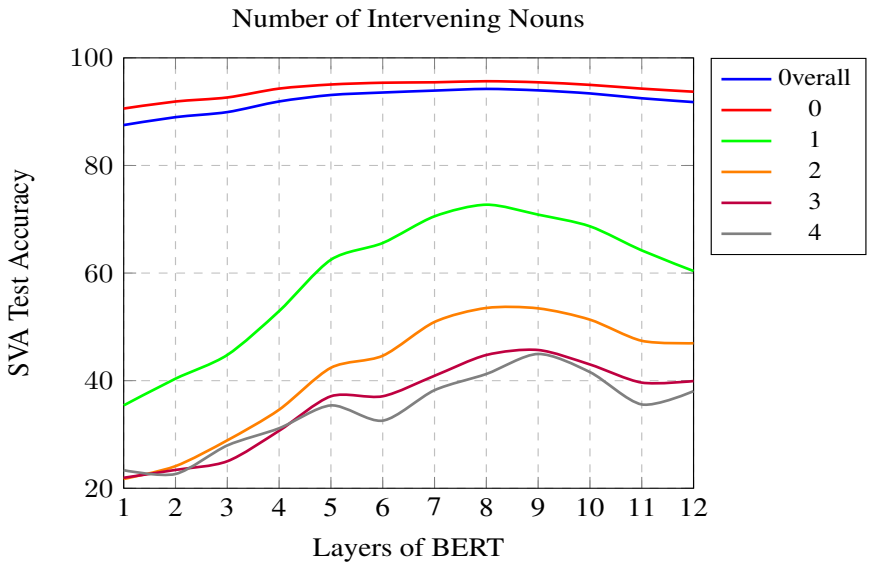

Figure 8. Layber by Layer Performance on Subject-Verb Agreement

icantly. In addition, the middle layers exhibit better test accuracy than both the lower and higher layers, aligning with the findings discussed in Section 4. Within the middle layers, deeper layers (layer 8 and 9) display higher accuracy than relatively lower ones, which indicates deeper layers may be necessary to learn long-distance dependency. An intriguing observation is that while the overall performance remains strong, the test accuracy plummets when an intervening noun is present. For instance, when one intervening noun is introduced, the accuracy for layer 1 drops to 40% compared to 90% with no intervening nouns. But the overall performance remains largely unaffected by the poorer performance resulting from increased intervening nouns, since the number of sentences without intervening nouns far exceeds the rest, as highlighted in Linzen et al.'s (2016) analysis.

6. Conclusion

In this paper, we showed that BERT's phrasal representations better capture phrase-level information in the lower layers, and through control tasks in probing and subject-verb agreement tasks, we found that deeper layers are essential for effectively modeling syntactic information. Furthermore, we discovered that BERT creates a hierarchy of linguistic cues ranging from surface to semantic features. Our results indicate that BERT's middle layers possess a greater amount of syntactic information, while the deeper layers display more robust semantic representations, in line with the findings of (Jawahar et al., 2019).

Connection to Computational Cognitive Modeling Our study aligns with computational cognitive modeling and human learning by exploring how BERT's internal representations resemble the compositional modeling approach found in human language learning. Compositional modeling involves combining smaller linguistic units to create larger meaningful structures, and we demonstrates that BERT exhibits this capacity.

We further emphasizes the significance of deeper layers in BERT for capturing long-range dependency information. This observation is in line with the understanding that human language learning entails integrating information from various levels and scales, encompassing local dependencies and global structures. BERT's reliance on deeper layers to capture these dependencies aligns with the idea that humans also rely on integrating information from multiple levels to comprehend complex linguistic phenomena. Overall, these insights advance our understanding of the relationship between computational models and human cognition.

7. Acknowledgement

We express our gratitude to Professor Lake and Gureckis for their assistance with HPC accounts, which enabled our feature extraction and training. We appreciate Linzen et al.'s work on subject-verb agreement, which laid the groundwork for subsequent syntax-related research. Additionally, we are thankful for Jawahar et al.'s systematic analysis of BERTology, which provided the basis for most of our experiments.

References

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Belinkov, Y. and Glass, J. Analyzing hidden representations in end-to-end automatic speech recognition systems. *Advances in Neural Information Processing Systems*, 30, 2017.

Belinkov, Y. and Glass, J. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 2019.

Bernardy, J.-P. and Lappin, S. Using deep neural networks to learn syntactic agreement. *Linguistic Issues in Language Technology*, 2017.

Bock, K. and Miller, C. A. Broken agreement. *Cognitive psychology*, 23(1):45–93, 1991.

Brunner, G., Wang, Y., Wattenhofer, R., and Weigelt, M. Natural language multitasking: analyzing and improving syntactic saliency of hidden representations. *arXiv preprint arXiv:1801.06024*, 2018.

Chomsky, N. Syntactic structures. 1957.

Conneau, A. and Kiela, D. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*, 2018.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., and Baroni, M. 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)*, pp. 2126–2136, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1198. URL https://aclanthology.org/P18-1198.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Elman, J. L. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Elman, J. L. Distributed representations, simple recurrent networks, and grammatical structure. *Machine learning*, 7:195–225, 1991.

Everaert, M. B., Huybregts, M. A., Chomsky, N., Berwick, R. C., and Bolhuis, J. J. Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in cognitive sciences*, 19(12):729–743, 2015.

Goldberg, Y. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*, 2019.

Guillaume Alain, Y. B. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Hewitt, J. and Liang, P. Designing and interpreting probes with control tasks. *arXiv preprint arXiv:1909.03368*, 2019.

Jawahar, G., Sagot, B., and Seddah, D. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Karpathy, A., Johnson, J., and Fei-Fei, L. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*, 2015.

Linzen, T. and Baroni, M. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7:195–212, 2021.

Linzen, T., Dupoux, E., and Goldberg, Y. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016.

Nicol, J. L., Forster, K. I., and Veres, C. Subject–verb agreement processes in comprehension. *Journal of Memory and Language*, 36(4):569–587, 1997.

Peters, M. E., Neumann, M., Zettlemoyer, L., and Yih, W.-t. Dissecting contextual word embeddings: Architecture and representation. *arXiv preprint arXiv:1808.08949*, 2018.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S. R., Das, D., et al. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yossi Adi, Einat Kermany, Y. B. O. L. Y. G. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*, 2018.

Zhang, K. W. and Bowman, S. R. Language modeling teaches you more syntax than translation does: Lessons learned through auxiliary task analysis. *arXiv preprint arXiv:1809.10040*, 2018.