

Of Non-Linearity and Commutativity in BERT

Sumu Zhao
ETH Zurich, Switzerland
Switzerland
zhaosumu@gmail.com

Damián Pascual
ETH Zurich, Switzerland
Switzerland
dpascual@ethz.ch

Gino Brunner
ETH Zurich, Switzerland
Switzerland
brunnegi@ethz.ch

Roger Wattenhofer
ETH Zurich, Switzerland
Switzerland
wattenhofer@ethz.ch

Abstract—In this work we provide new insights into the transformer architecture, and in particular, its best-known variant, BERT. First, we propose a method to measure the degree of non-linearity of different elements of transformers. Next, we focus our investigation on the feed-forward networks (FFN) inside transformers, which contain two thirds of the model parameters and have so far not received much attention. We find that FFNs are an inefficient yet important architectural element and that they cannot simply be replaced by attention blocks without a degradation in performance. Moreover, we study the interactions between layers in BERT and show that, while the layers exhibit some hierarchical structure, they extract features in a fuzzy manner. Our results suggest that BERT has an inductive bias towards layer commutativity, which we find is mainly due to the skip connections. This provides a justification for the strong performance of recurrent and weight-shared transformer models.

I. INTRODUCTION

Since the introduction of the transformer architecture [33], and in particular of BERT [6], researchers have tried to gain a deeper understanding of the inner workings of these models [28]. This research paves the way for better explanations of model decisions, thereby broadening the range of possible applications of transformers. Furthermore, based on works on understanding transformers, in [38] the authors introduce an improved self-attention mechanism, resulting in state of the art performance. This shows that deeper knowledge of transformers can not only help with interpretability, but also lead to new and better architectures. In this work, we aim to advance the understanding of the inner workings of transformers by providing novel insights from two complementary angles.

First, we assess the importance and representational power of different parts of BERT. We propose to use linear approximators to measure the degree of (non-)linearity of BERT’s building blocks. Due to the large amounts of data used for pre-training, overfitting has so far not been an issue for transformers. This is exemplified by ever larger models achieving state of the art performance [2], [27]. It thus seems clear that models with higher expressive power are beneficial for natural language tasks. We therefore hypothesise that the degree of non-linearity of a transformer can give us valuable insight into its inner workings, and in particular, it can indicate which components are modeling more complex relationships. Further, we run exhaustive experiments in which we remove or replace parts of the model and measure the impact on performance. We find that the feed-forward networks (FFN) in the transformer architecture are inefficient compared to the

self-attention operation, especially since they make up two thirds of the model parameters.

Second, we investigate why recent “recurrent transformers” such as ALBERT [17], which employs weight sharing across layers, performs on par with or even better than the original BERT. Here, we find that BERT has an inherent tendency towards weight-sharing, and that layers can be removed or swapped without causing much performance degradation. By comparing to simpler feed-forward architectures, we find that skip connections play an important role in inducing layer commutativity in neural networks.

Our main contributions are as follows:

- A diagnosis tool to measure non-linearity in Transformers that takes into account the geometry of the embedding space.
- An extensive architectural study revealing the importance of the FFNs.
- An empirical justification for why BERT is amenable to weight-sharing across layers, indicating that skip connections are largely responsible for this.

II. RELATED WORK

The field of Natural Language Processing (NLP) has undergone major developments since the introduction of the transformer architecture [33]. Since then, many different variants of transformers have been proposed [15], [19], [25]–[27], [30], [39]. One of the fundamental building blocks of transformers is self-attention, which allows the model to learn alignments between words. The possibility of visualizing attention distributions and the quest for interpretability in deep learning have sparked a lot of research into understanding transformer models. Most of this research has focused on BERT [6], a well-known transformer model, initiating a line of work now known under the term BERTology [28].

In this field, a number of studies look into attention distributions in order to understand transformers’ decisions and their internal dynamics [4], [11], [16]. However, recent work raises concerns about the interpretability of attention distributions [3], [12], [21], [24]. A complementary line of research uses probing classifiers to test for linguistic properties of different parts of the model [18], [32].

In this work, we take an architectural approach to the analysis of BERT and look into how different components of the model interact to extract features and process language. Similar to probing classifiers, we use a linear approximator

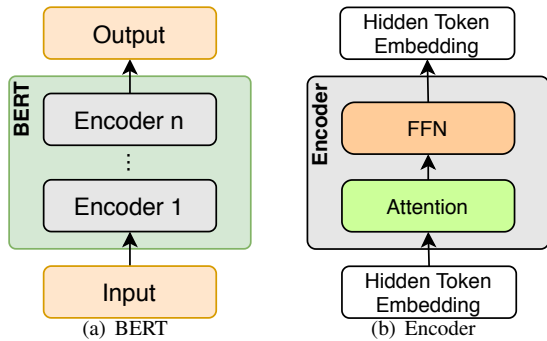


Fig. 1. Architecture of BERT and the Encoder block.

to regress the output of different architectural components in order to study their non-linearity. While existing work has focused on analyzing the self-attention heads and the impact of model depth [14], [20], [34], we direct our investigation towards the FFN blocks of BERT.

The authors of [31] and [13] show that features are extracted in BERT following the same order as in the classical NLP pipeline, with syntactic features extracted in earlier layers and semantic features in later layers. These features are extracted in a fuzzy manner, in the sense that multiple layers are responsible for solving the same tasks (e.g., Part-Of-Speech (POS) tagging). We study this phenomenon from a complementary angle by investigating the similarity between layers. Remarkably, ALBERT [17] directly shares the weights across the layers of BERT without loss of performance, and Universal Transformers [5] use recurrence over one single transformer layer. Similarly, in [23] the authors exploit layer similarity to re-order the layers and improve performance, while in [7] and [29] they analyze layer redundancy to prune BERT with minimal performance degradation. Our findings support the idea that BERT has an inductive bias towards weight sharing, which results in the layers having a strong degree of commutativity and the features being extracted in an incremental or fuzzy manner.

III. BACKGROUND

In this section, we give a brief description of BERT. For more details please refer to [6] and the official GitHub repository.¹ As Figure 1 (a) shows, BERT is composed of n encoder layers. An encoder block follows the same structure as in the original transformer [33]; it consists of an attention block and a feed-forward network (FFN), as shown in Figure 1 (b). Figure 2 shows the detailed structures of the attention and FFN blocks respectively. The input of BERT is a sequence of tokens, additionally containing special [CLS], [SEP] and [PAD] tokens. [CLS] can be used as output for classification tasks and [SEP] denotes sentence boundaries. Since BERT is trained with a fixed (maximum) sequence length, [PAD] tokens have to be appended to the actual input sequences.

In our study, we use two BERT models of different sizes (base and small); architecture details are shown in Table I.

¹<https://github.com/google-research/bert>

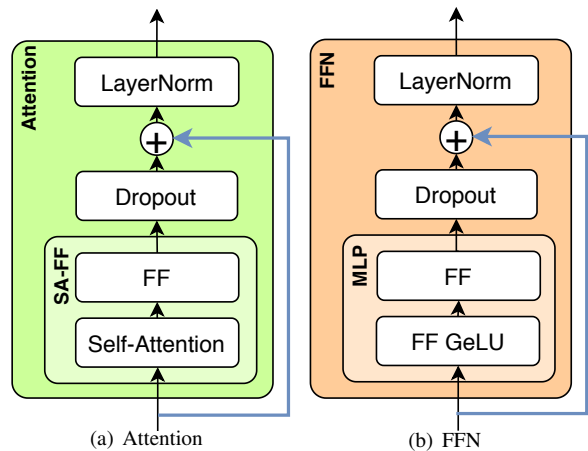


Fig. 2. Architecture of Attention and FFN blocks.

TABLE I
DETAILS OF THE BERT MODELS WE USE.

Model	Layers	Heads	Hidden Size	Params	Max. Seq Length
small	6	8	512	35M	128
base	12	12	768	110M	512

For BERT-base we use the pre-trained weights as released by the authors. All variants of BERT-small are pre-trained by us from scratch. Pre-training BERT-small takes 8 days on one Tesla V100 GPU card and 5 days on two cards. In [6] they pre-trained BERT-base on a 16GB text corpus which is a combination of ENGLISHWIKIPEDIA and BOOKCORPUS [40]. Since this dataset is not readily available, we use OPENWEBTEXT [9], which is a 38GB text corpus created using the same method as for training GPT-2 [26].

We evaluate our BERT-small models on all nine GLUE [35] tasks. For other experiments, we use the MNLI-matched (MNLI_m) dataset from GLUE. To speed up the experiments, we select a random subset, MNLI_m-sub, consisting of 39,271 examples for fine-tuning, and 983 for validation.

IV. STUDYING (NON-)LINEARITY

A reliable measure of how non-linear different components of deep neural networks are, can help guide model analysis and design. Here, we propose to use a linear approximator to quantify the non-linearity of different components of transformers.

Given an architectural component f , we train a linear approximator f^* that, given an input e , maximizes the cosine similarity between the actual output $f(e) = \mathbf{y}$ and the approximated output $f^*(e) = \mathbf{y}^*$. Cosine similarity is a common choice to measure similarity between words in an embedding space [3], [8], [22]. We define the linearity score γ_f of a component f as the average cosine similarity between actual and approximated outputs over a set of inputs: $\gamma_f = \mathbb{E}_{e \sim E}[\cos(\mathbf{y}, \mathbf{y}^*)]$. The larger γ_f , the more linear the measured component f is. Intuitively, if f is very non-linear,

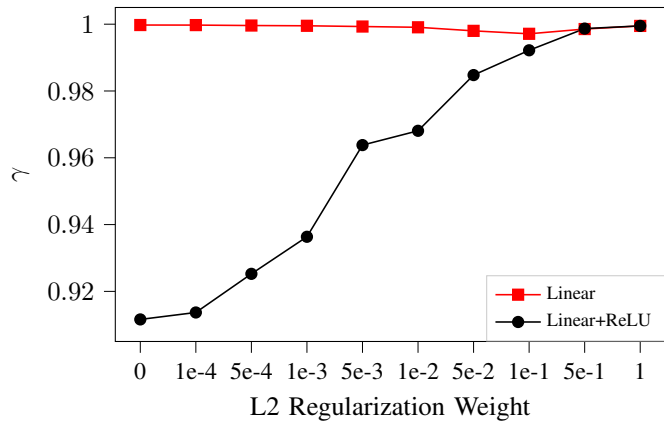


Fig. 3. Linearity score of a hidden layer of a linear MLP (Linear) and a non-linear MLP (Linear+ReLU) for different regularization strengths.

a linear approximator will not be able to approximate it accurately and the linearity score γ_f will be close to 0. If f is very linear, γ_f will be close to 1.

A. Method validation

To validate our method, we first consider a simple toy task where we can manipulate the model’s degree of non-linearity. Specifically, we use a Multi-layer Perceptron (MLP) on the MNIST handwritten digit classification task, which consists of 60,000 examples. The MLP has one hidden layer of size 768 with ReLU activations, and one softmax layer with 10 output classes. We train this model on 50,000 examples for 30 epochs with batch size 32 and a learning rate of 0.001 using the Adam optimizer, and then evaluate it on 10,000 examples. Next, we use the remaining 10,000 training examples to train a linear model with 768 neurons for 10 epochs to approximate the output activations of the MLP’s hidden layer.

To verify the linearity score γ , we train the MLP with different L2 regularization strengths and calculate γ for each model using the linear approximator. We expect that the more non-linear a layer, the smaller γ , i.e., the larger the approximation error. Since increasing the regularization strength makes the weights smaller, which causes the hidden layer to behave more linearly, the linearity score γ should increase with larger regularization weights. In Figure 3, we observe this exact behavior, with a monotonic increase in linearity with stronger regularization. Next, we run the same experiment using a linear MLP (without ReLU activation). Since this model is linear, we expect a linearity score of 1 regardless of the regularization strength; as shown in Figure 3, we observe precisely this behavior. These results show that the linear approximator can quantify the non-linearity of a simple MLP. Next, we apply this method to study the non-linearity of BERT. To do this, we first need to consider the geometry of the embedding space.

B. Measuring the Non-Linearity of BERT

In [8] it is pointed out that the hidden embeddings at each transformer layer are directional and occupy cones of different sizes in their respective spaces, i.e., the embedding

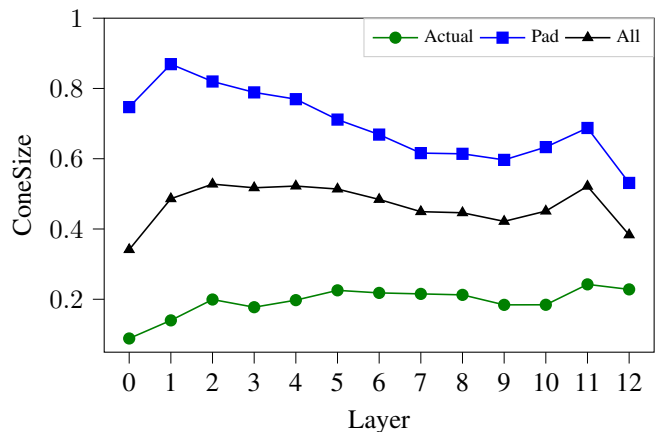


Fig. 4. Average cosine similarity of 1000 random token pairs. The pairs are generated from the validation set of MNLI-sub. Layer 0 denotes the input token embeddings, layers 1-12 denote the output token embeddings of the 12 layers of BERT.

space is anisotropic, with most hidden embeddings located within a narrow region (cone) of the space. Since the size of these cones is different at each point of the model, we need to normalize the approximation errors by the respective cone sizes in order enable comparisons across layers. This is because, for architectural components that generate smaller cones at their output, the linear approximation is easier since all the output embeddings are already close to each other.

To calculate the cone size we follow [8] and estimate the expected pair-wise cosine similarity from 1000 randomly sampled word representations. The intuition behind this is that the average similarity between pairs of randomly sampled vectors describes how narrow the cone is. This way, larger values correspond to smaller cones. At each layer l , the cone size is defined as:

$$\text{ConeSize}(l) = \mathbb{E}_{(e_i^l, e_j^l) \sim E^l} [\cos(e_i^l, e_j^l)]$$

Where E^l are the word embeddings at layer l . A value of 0 represents perfect isotropy and 1 perfect anisotropy. We then calculate the cone size at each layer of BERT for the MNLI-sub dataset. In Figure 4 we report the cone size for padding, actual and all tokens. We see that there is a significant difference between actual and padding tokens, and that the cone size for all tokens is roughly the average between both (weighted by the ratio of actual vs padding tokens). Overall, we see that the cone size for actual tokens is relatively wide².

This way, we calculate the cone size of the embeddings after different components f in BERT. These values are then used to obtain the *normalized* linearity score $\tilde{\gamma}_f^l$ of component f at layer l . To do this, we first subtract the cone size from γ_f to correct for anisotropy, following the same procedure as in [8]. Then, to allow comparison across layers, i.e., to have

²Our results differ from [8]; we tried different settings without being able to reproduce the exact results. Possible explanations are that we separate padding and actual tokens, and that we use the original pre-trained weights of BERT while [8] uses the model from the HuggingFace [36].

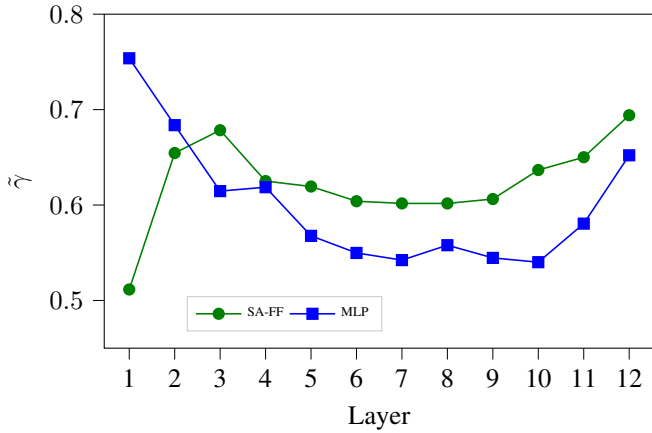


Fig. 5. Normalized linearity scores of SA-FF and MLP for each layer in BERT-base. The scores are obtained using MNLI_m-sub.

the values of $\tilde{\gamma}_f^l$ contained in a fixed range, we normalize the result by dividing by the maximum possible value of the anisotropy-corrected linearity score $1 - \text{ConeSize}(f_i)$:

$$\tilde{\gamma}_f^l = \frac{\gamma_f^l - \text{ConeSize}(f_i)}{1 - \text{ConeSize}(f_i)}$$

Using this normalized linearity score, we measure at each layer of BERT the (non-)linearity of the SA-FF and the MLPs, which are the core of the Attention and FFN blocks respectively. For both blocks, the linear approximator is a fully connected linear layer of the same size as the hidden dimension in BERT-base, i.e., 768. We train it for 3 epochs with batch size 64, using the Adam optimizer with a learning rate of 0.001 on MNLI_m-sub. In Figure 5 we show the normalized linearity score $\tilde{\gamma}_f^l$ between the approximator output and the real tokens after each of the considered blocks for each of the layers of BERT-base.

Note that, while MLPs perform a non-linear function on one input token, the self-attention operation in the SA-FF block performs a weighted sum over all tokens in the input sequence. This way, the non-linearity introduced by SA-FF comes mainly from the aggregation of context information. Given the fundamentally different non-linear mechanisms used by each block, we refrain from directly comparing their linearity scores. Instead, we analyse how the non-linearity of each of them changes along the layers.

Both MLPs and SA-FF blocks follow opposing trends in the first three layers, with the MLPs becoming less linear, and SA-FF becoming more linear. After layer 3, the linearity scores of both blocks plateaus until approximately layer 10, when both become progressively more linear. Interestingly, the U-shaped pattern followed by both blocks (with the notable exception of the SA-FF block in the first layer), indicates that layers closer to the input/output space extract more linear features.

These results show that 1) MLPs introduce strong non-linearity, especially in the middle layers; and 2) the non-linearity introduced by SA-FF follows the same trend as that of MLPs even if the non-linear mechanism is fundamentally

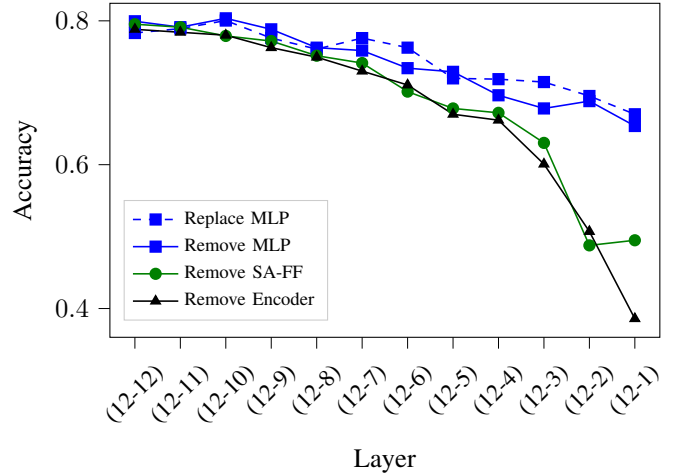


Fig. 6. Multi-layer backwards replacement/removal of different components in BERT-base, evaluated on MNLI_m-sub.

different. Next, we directly compare the impact of each block on model performance.

C. MLP vs. Self-Attention

Most research on understanding Transformer architectures has focused on the self-attention component. However, roughly two thirds of the total parameters of BERT reside within the FFN blocks, mainly in the MLPs. Here we compare the impact on performance of MLPs versus SA-FF blocks.

To this end, we first progressively replace the MLPs in all layers of BERT-base in the backwards direction, i.e., starting from the last layer, by their corresponding linear approximator. We replace backwards because this does not change the input to the earlier layers (which are not replaced). After replacement, we fine-tune and evaluate the model on MNLI_m-sub. Then we repeat the process but this time we simply remove the MLPs, i.e., without replacing them by linear approximators. We compare the performance of these experiments with the performance of progressively removing the SA-FF blocks. The results are reported in Figure 6; as reference, we also report the performance after removing the whole encoder layer.

Some surprising insights can be extracted from these experiments. First, replacing the MLPs by linear approximators yields approximately the same performance as completely removing the MLPs. This implies that it is the non-linear contribution of the MLP that is solely responsible for the performance difference. Second, the drop in performance after removing MLP blocks is smaller than when removing SA-FF blocks; especially after layer 6, removing the SA-FF blocks produces a much more severe degradation. Finally, removing SA-FF is as harmful as removing the whole encoder layer.

All of this confirms that while the majority of the parameters are in the MLPs, i.e., in the FFNs, the self-attention operation is more important for overall model performance. Furthermore, it raises the question of what happens if we remove all FFNs and instead add more attention layers. We explore this by

TABLE II
GLUE SCORES OF VARIANTS OF BERT-SMALL.

Model	GLUE score
BERT-small	74.5
no-FFN	69.6
no-FFN-SA+	70.9
no-FFN-GeLU-SA	70.0
no-FFN-GeLU-SA+	72.3

training different variants of BERT-small and evaluating them on GLUE. In particular, we train:

no-FFN: model without FFN blocks. This model has about one third of the original parameters.

no-FFN-SA+ : model without FFN blocks, but with more layers so that the total number of parameters is roughly the same as BERT-small. This model has 18 layers.

no-FFN-GeLU-SA : same as *no-FFN* but with GeLU activation functions after the SA-FF block.

no-FFN-GeLU-SA+ : same as *no-FFN-SA+* but with GeLU activation functions after the SA-FF block.

The results are shown in Table II. We find that, as expected, models with more parameters perform better; we see this from the performance of the *SA+* models. We also see that adding a GeLU non-linearity [10] to the self-attention block further improves performance. However, adding more self-attention blocks in substitution of the FFN blocks degrades the performance of the model by more than 4 points, and by more than 2 points when using GeLU non-linearities. This strong degradation is surprising and indicates that separate FFN blocks with their own skip connection and layer normalization are more effective than moving the parameters and non-linearity to the attention blocks.

Overall, we find that the FFNs are an important, albeit somewhat inefficient, building block of BERT. We believe that a redesign of these structures could lead to improved performance, or equal performance with fewer parameters. However, this redesign is not straightforward and requires a more extensive investigation.

V. BERT HAS AN INDUCTIVE BIAS TOWARDS WEIGHT SHARING

Recently there has been a trend towards transformers that feature recurrence across the layer-dimension [1], [5], [17]. For example, ALBERT achieves this by explicitly sharing the weights across all layers. Somewhat surprisingly, this does not negatively affect the performance, despite drastically reducing the number of parameters by a factor of n_{layers} . In this section we show that BERT has an inductive bias that makes it amenable for explicit weight-sharing across layers. In particular, we show that the layers in BERT are highly commutative, that is, they can be swapped with surprisingly little impact on performance.

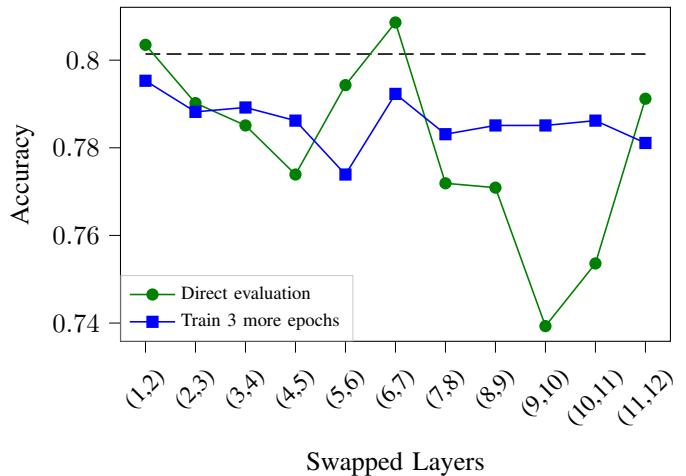


Fig. 7. Swapping adjacent layers of BERT-base fine-tuned on MNLI-sub. Baseline: 80.14% (dashed).

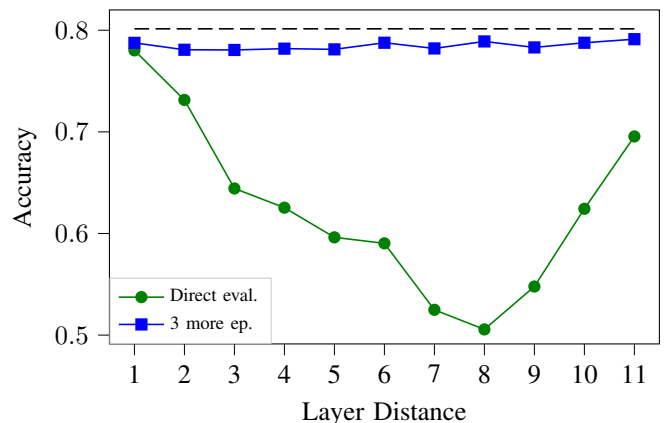


Fig. 8. Swapping all possible combinations of two layers (mean accuracy). Baseline: 80.14% (dashed).

A. Swapping Layers

Figure 7 shows the effect of swapping *adjacent* layers of BERT-base and evaluating the resulting models directly on the validation set of MNLI-sub. Surprisingly, the average performance degradation is only 2%.

Next, we investigate swapping *any two* layers by evaluating all 66 models that can be created by swapping two layers. Figure 8 shows the average accuracy for each swapping distance. Swapping distance refers to the absolute difference between the indices of the swapped layers. Note that there are fewer models for longer swapping distances and hence fewer data points. For example, swapping layers across a distance of 11 corresponds to swapping layer 1 and layer 12, and so there is only one possible model. The figure shows that, overall, BERT is surprisingly robust to this modification, and only 3 epochs of further fine-tuning can recover most of the performance loss. However, swapping over longer distances generally results in worse performance, indicating that adjacent layers are more

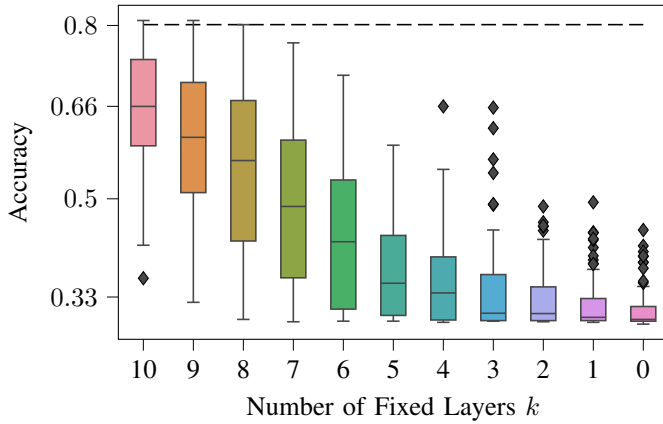


Fig. 9. Create new BERT models by keeping k layers fixed and shuffling the rest. The models are then directly evaluated on MNLI-sub without further fine-tuning. Baseline: 80.14% (dashed).

similar, as it was also observed in [37]. Interestingly, for distances above 9 this trend reverses. When examining the results in more detail, we find that this is due to the fact that swapping layer 1 with either layer 11 or 12 results in accuracies of 62% and 70% respectively. This indicates that the very first and very last layers of BERT are remarkably similar, potentially due to their “closeness” to the input/output space. This is in line with the U-shape observed in our non-linearity analysis in Section IV-B.

B. Shuffling Layers and ALBERT-like Configuration

Next, we extend our analysis from swapping two layers at a time to shuffling different numbers of layers, i.e., changing their order. Figure 9 shows the results of a series of layer-shuffling experiments on BERT; we first randomly choose k layers to be fixed, and then randomly permute all the remaining layers. For every value of k we create 100 permuted versions of BERT. We then directly evaluate each of these models on the MNLI-sub validation set without any additional fine-tuning. The results show that as we decrease the number of fixed layers, performance monotonically decreases. However, only for $k \leq 3$ does the median accuracy drop to random guessing (MNLI has three classes, thus random guessing corresponds to an accuracy of 33%). Also, for every value of k , there are many models that still perform well above random.

Figure 10 shows what happens when we fine-tune the shuffled models for another three epochs. Remarkably, for values of $k \geq 5$, there are some models that reach or even surpass the baseline score. Even for $k < 5$, much of the performance can be recovered. Note that 3 epochs of fine-tuning correspond to less than 2000 weight updates at a reduced learning rate, as is common when fine-tuning BERT. This stands in stark contrast to the over 1M weight updates during pre-training, and shows that most layers of BERT can be quickly “repurposed” after shuffling, which in turn indicates that they are very similar to each other.

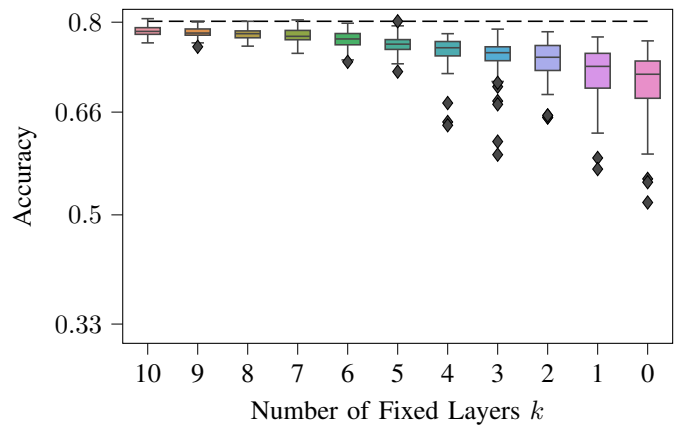


Fig. 10. Create new BERT models by keeping k layers fixed and shuffling the rest. The models are then fine-tuned again on MNLI-sub before evaluating. Baseline: 80.14% (dashed).

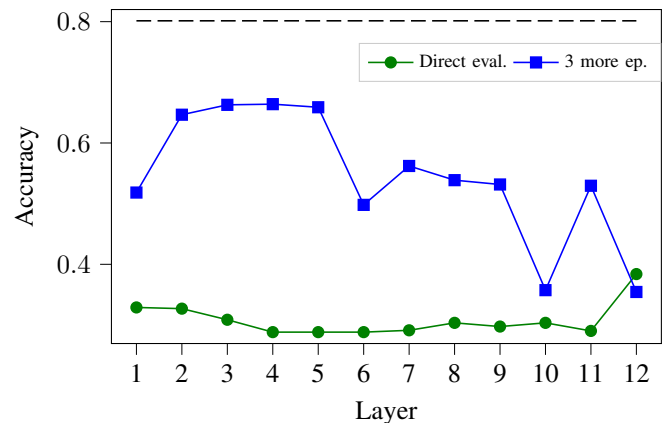


Fig. 11. Performance of ALBERT-like models on MNLI-sub. Baseline: 80.14% (dashed).

Finally, we create ALBERT-like models from the fine-tuned BERT-base by repeating each layer 12 times. Without further fine-tuning, this reduces performance to random guessing. However, if we fine-tune for another 3 epochs, performance improves significantly. The results are shown in Figure 11. Interestingly, models based on layers 2-5 seem to be best, and higher layers fail to perform well on their own. This shows that, while BERT has an inherent tendency towards weight-sharing, its layers do still exhibit a certain hierarchical structure.

So far, the results in this section suggest that BERT has an inductive bias towards learning very similar transformations at each layer, which is why layers can be swapped and shuffled without incurring a large performance loss. We therefore hypothesize that enforcing weight sharing as in ALBERT [17] makes this inherent tendency explicit, thus acting as an effective regularizer. Next, we investigate where this inherent weight-sharing tendency is coming from.

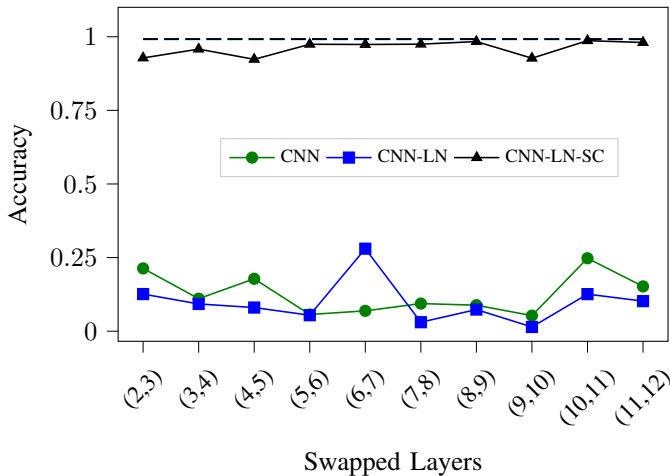


Fig. 12. Swapping adjacent layers of CNNs trained on MNIST. LN stands for layer normalization and SC for skip connections. Baseline: 99.1% (dashed).

C. What Causes the Inductive Bias Towards Weight-Sharing?

We now contrast the results on BERT with a simple CNN baseline trained on MNIST, where we find that swapping adjacent layers instantly reduces the classifier performance to random. Thus, in order to figure out which component of BERT is responsible for the strong commutativity of its layers, we train two additional CNN versions, one with layer normalization and another one with skip connections around every convolutional layer³. The results for all three CNNs are shown in Figure 12. Adding layer normalization does not make the CNN more commutative. However, the CNN with skip connections shows almost no performance drop when swapping adjacent layers. These results are very similar to the ones on BERT from Figure 7.

Analogously to the previous section, we also investigate how the CNN behaves when more than two layers are shuffled. The results are shown in Figure 13. Direct comparisons to the results on BERT are difficult due to the differences in training task and architectures. However, the figures clearly show that layers of models with skip connections are highly commutative, especially when compared to baselines without skip connections. These results suggest that much of the inherent weight-sharing bias of BERT comes from the skip connections in the FFN and Attention blocks.

Our findings in this section suggest that feature extraction in BERT is not strictly hierarchical, but happens incrementally, and that it is not the “early layers” per-se that extract low-level features, but low-level features are simply extracted first, regardless of the order in which the layers are applied. We believe that this is partly due to the commutativity of BERT’s layers, which seems to be caused mostly by the use of skip connections.

³Ablating components from BERT (e.g. skip connections) prevented convergence during training and thus, we use a simpler CNN model to analyze the behavior of the different architectural elements.

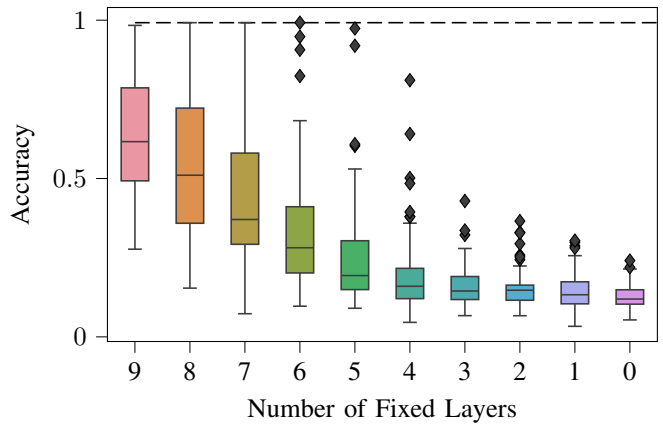


Fig. 13. Creating new skip-connection CNN models by keeping k layers fixed and shuffling the rest. The resulting models are evaluated without further fine-tuning. Baseline: 99.1% (dashed).

VI. CONCLUSION

In this work we propose a tool to measure non-linearity in transformers which takes into account the geometry of the embedding space. With this, we find that the non-linearity introduced by the MLPs and SA-FF blocks follow the same U-shaped pattern. Further, we find that although FFN blocks contain approximately two thirds of the parameters, attention has a considerably stronger impact on the performance of the model. Finally, we investigate how BERT’s layers interact with each other and observe that features are extracted in a fuzzy or incremental manner, with the layers being very similar to each other. By comparing BERT to simple CNNs we find that the commutativity of the layers is induced by the skip connections. This inductive bias towards weight sharing and recurrence provides an explanation for the strong performance of weight-shared models like ALBERT or the Universal Transformer.

We hope that this work will motivate further investigation of architectural elements of transformers other than self-attention. We believe a more efficient use of the parameters contained in the FFNs is possible, and we expect that further research will find a better way of exploiting the representational power of those parameters. Although we find that skip connections make adjacent layers commutative, the role of self-attention in enabling recurrence over layers remains unclear. Since self-attention is an input-dependent operation, it provides additional flexibility to the model that may further enforce the inductive bias towards weight-sharing. Hence, we deem the investigation of weight-shared architectures an exciting direction for future work.

REFERENCES

- [1] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 690–701. Curran Associates, Inc., 2019.
- [2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M.

- Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.
- [3] Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers. In *International Conference on Learning Representations*, 2020.
- [4] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 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, 2019.
- [5] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019.
- [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, 2019.
- [7] Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. Analyzing individual neurons in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4865–4880, 2020.
- [8] Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*. Association for Computational Linguistics, 2019.
- [9] Aaron Gokaslan and Vanya Cohen. Openwebtext corpus. <http://Skylion007.github.io/OpenWebTextCorpus>, 2019.
- [10] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.
- [11] Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R Bowman. Do attention heads in bert track syntactic dependencies? *arXiv preprint arXiv:1911.12246*, 2019.
- [12] Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 3543–3556, 2019.
- [13] Ganesh Jawahar, Benoît Sagot, and Djamel Seddah. What does bert learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.
- [14] K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*, 2020.
- [15] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.
- [16] Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 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 4356–4365, 2019.
- [17] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020.
- [18] Yongjie Lin, Yi Chern Tan, and Robert Frank. 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, 2019.
- [19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [20] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, pages 14014–14024, 2019.
- [21] Damian Pascual, Gino Brunner, and Roger Wattenhofer. Telling bert’s full story: from local attention to global aggregation. *arXiv preprint arXiv:2004.05916*, 2020.
- [22] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
- [23] Ofir Press, Noah A Smith, and Omer Levy. Improving transformer models by reordering their sublayers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2996–3005, 2020.
- [24] Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. Learning to deceive with attention-based explanations. *arXiv preprint arXiv:1909.07913*, 2019.
- [25] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. URL https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language_understanding_paper.pdf, 2018.
- [26] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 2019.
- [27] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.
- [28] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *arXiv preprint arXiv:2002.12327*, 2020.
- [29] Hassan Sajjad, Fahim Dalvi, Nadir Durrani, and Preslav Nakov. Poor man’s bert: Smaller and faster transformer models. *arXiv preprint arXiv:2004.03844*, 2020.
- [30] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.
- [31] Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, 2019.
- [32] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*, 2019.
- [33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [34] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, 2019.
- [35] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.
- [36] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.
- [37] John M Wu, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. Similarity analysis of contextual word representation models. *arXiv preprint arXiv:2005.01172*, 2020.
- [38] Zhanghao Wu, Zhijian Liu, Ji Lin, Yujun Lin, and Song Han. Lite transformer with long-short range attention. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.
- [39] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pre-training for language understanding. <https://arxiv.org/abs/1906.08237>, 2019.
- [40] Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 19–27. IEEE Computer Society, 2015.