Outlier Dimensions that Disrupt Transformers are Driven by Frequency

Giovanni Puccetti[1, 2, 4], **Anna Rogers**[3,4], **Aleksandr Drozd**[4], **Felice Dell'Orletta**[2]
[1] Scuola Normale Superiore, Pisa, Italy
[2] Istituto di Linguistica Computazionale "Antonio Zampolli", Pisa, ItaliaNLPLab - *www.italianlp.it*
[3] Center for Social Data Science, University of Copenhagen, Denmark
[4] RIKEN Center for Computational Science, Japan
giovanni.puccetti@sns.it, arogers@sodas.ku.dk,
alex@blackbird.pw, felice.dellorletta@ilc.cnr.it,

Abstract

While Transformer-based language models are generally very robust to pruning, there is the recently discovered outlier phenomenon: disabling only 48 out of 110M parameters in BERT-base drops its performance by nearly 30% on MNLI. We replicate the original evidence for the outlier phenomenon and we link it to the geometry of the embedding space. We find that in both BERT and RoBERTa the magnitude of hidden state coefficients corresponding to outlier dimensions correlates with the frequency of encoded tokens in pre-training data, and it also contributes to the "vertical" self-attention pattern enabling the model to focus on the special tokens. This explains the drop in performance from disabling the outliers, and it suggests that to decrease anisotropicity in future models we need pre-training schemas that would better take into account the skewed token distributions.

1 Introduction

The current Transformer-based language models are heavily overparametrized, which explains why it is possible to prune these models by up to 30-40% (Gordon et al., 2020; Sanh et al., 2020; Prasanna et al., 2020; Chen et al., 2020, inter alia) without a significant drop in performance. However, it has recently been shown that multiple Transformer-based language models (LMs) are highly sensitive to removal of *outlier dimensions* (Kovaleva et al., 2021): the parameters (weights and biases) in the output element of a Transformer layer, the magnitude of which is unusually large within the layer (consistently in the same dimension across the model layers). For BERT model family the output element is the LayerNorm, as shown in Figure 1.

Although these parameters constitute less than 0.0001% of the full BERT (Devlin et al., 2019) model, removing them significantly degrades BERT's performance. Puccetti et al. (2021) find that the same parameters are particularly relevant

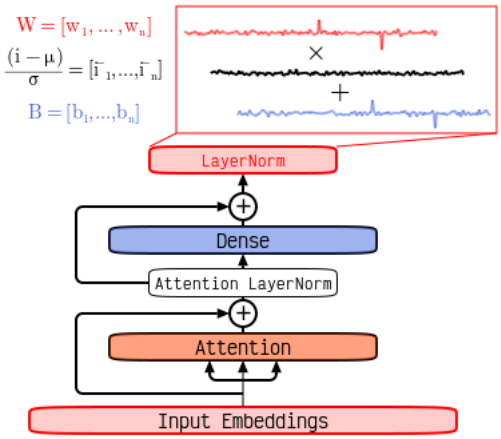


Figure 1: The Transformer Layer architecture diagram with outliers at the normalization layer (LayerNorm).

in several linguistic probing tasks. These dimensions affect the vector representation of different tokens in the same way, making the embedding space less isotropic and thus reducing its representational power (Liang et al., 2021). Outlier dimensions have also been found to make model quantization challenging (Bondarenko et al., 2021; Dettmers et al., 2022) as they need to be treated separately from others when defining quantization schemes. Thus there are both conceptual and practical reasons supporting a deeper study of this phenomenon.

What is not clear at this point is the mechanism behind the emergence of outliers. We replicate the original findings in BERT and RoBERTa, and we contribute new evidence **directly linking the outlier phenomenon with the frequency of encoded tokens in the pre-training data, as well as the self-attention pattern focusing on special tokens**. We also present evidence for two kinds of outliers: some of them affect the Masked Language Model (MLM) performance the most in the middle layers (where the correlation with token frequency is at its peak), and for others the impact grows towards

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 1286 - 1304
December 7-11, 2022

bert-base-uncased	cola	mnli	mnli-mm	mrpc	qnli	qqp	rte	sst2	stsb
baseline	56.9 +/- 1.5	84.5 +/- 0.2	84.8 +/- 0.4	84.3 +/- 1.1	91.4 +/- 0.1	91.1 +/- 0.1	66.3 +/- 1.6	92.8 +/- 0.5	89.0 +/- 0.3
1 random removed	56.5 +/- 1.5	84.5 +/- 0.2	84.8 +/- 0.4	84.5 +/- 0.9	91.3 +/- 0.1	91.1 +/- 0.1	66.6 +/- 1.7	92.8 +/- 0.4	89.0 +/- 0.3
w/o 308	47.3 +/- 1.2	81.4 +/- 1.1	82.2 +/- 1.1	54.0 +/- 12.3	88.9 +/- 0.8	89.1 +/- 1.6	62.1 +/- 3.2	90.8 +/- 1.1	56.9 +/- 17.4
w/o 381	33.8 +/- 9.4	73.2 +/- 2.1	73.8 +/- 2.0	64.6 +/- 15.2	80.3 +/- 1.5	79.8 +/- 3.2	55.8 +/- 1.6	87.9 +/- 1.0	78.1 +/- 4.8
2 random removed	56.4 +/- 1.5	84.5 +/- 0.2	84.8 +/- 0.4	84.3 +/- 0.9	91.3 +/- 0.1	91.1 +/- 0.1	66.6 +/- 1.7	92.8 +/- 0.5	88.9 +/- 0.3
w/o 308 & 381	15.9 +/- 4.2	58.4 +/- 3.3	59.0 +/- 3.5	55.1 +/- 16.3	74.5 +/- 1.3	74.6 +/- 4.6	55.3 +/- 4.5	76.0 +/- 2.4	35.7 +/- 15.3

Table 1: Average BERT scores over 5 runs on GLUE benchmarks with the effect of outlier removal. The rows *1 random removed* and *2 random removed* show the average over 5 removals of random non outliers (1 or 2 at a time respectively) for 5 different fine-tuned models

the final layers (although correlation with token frequency decreases). This work contributes to mechanistic understanding of Transfromer-based LMs, and it might be useful for future research on decreasing anisotropy in pre-trained LMs.

2 Methodology

According to Kovaleva et al., *outliers* are parameters (both weights and biases) in the final element of a Transformer layer (LayerNorm for BERT family, final MLP for GPT-2), which have unusually high magnitude[1] within the layer. *Outlier dimensions* are those dimensions at which outlier parameters are found consistently across the model layers.

The reason Kovaleva et al. study these parameters is that when they are disabled, the model performance on downstream tasks is greatly reduced. Since not all parameters that can be identified by magnitude and position criteria have that effect, we add this property to the definition. In this work the term *outlier dimension* refers to the dimensions with parameters meeting the magnitude criteria across layers and having at least 5x more damaging effect on accuracy on a representative downstream task, for which we choose MNLI (see §3.1).

To disable the outliers, unless stated otherwise, we set to zero both the LayerNorm *weight* and *bias* parameters for all layers (24 parameters in total for one outlier dimension in BERT and RoBERTa-base)[2]. See App. A for the full list of outliers identified for all models in this study.

We use the notation O to refer to specific LayerNorm outlier parameters (in BERT model family): e.g. $O381$ to indicate "an outlier with index 381". Since outlier indices are a constant for a given model, in this study we will also discuss *hidden state outlier dimensions*: the coefficient of the hidden state with the same index as the outlier.

We experiment with BERT-base (Devlin et al., 2019) (*"bert-base-uncased"*), RoBERTa-base (Liu et al., 2019b) (*"roberta-base"*) and Vision Transformer (Kolesnikov et al., 2021) (*"google/vit-base-patch16-224-in21k"*) from the *transformers* library[3]. For the experiments on pre-training dynamics we rely on the checkpoints with seed 1 provided by Sellam et al. (2022)[4].

Hardware, implementation and energy expenditure details are outlined in App. B. We release the code to replicate our experiments[5].

3 Outliers Phenomenon in Transformers

3.1 Replicating Prior Evidence

We start by replicating Kovaleva et al.'s experiments identifying the outliers for BERT- and RoBERTa-base ($O308$ and $O381$, $O77$ and $O588$ respectively), and their effect on downstream task performance.

Table 1 shows the average performance and standard deviation of BERT-base over 5 fine-tuning runs for eight GLUE[6] tasks. Thus we successfully replicate the original experiment on model degradation after[7] removal of the outliers. Since the effect is consistent across GLUE tasks, we use MNLI as a representative downstream task in the remaining

[1]The original definition of outliers is not entirely formal, and needs to be further specified for particular models: the magnitude of the outliers was within 2 standard deviations from the mean for RoBERTa, and within 3 for BERT.

[2]Note that this is equivalent to zeroing out the outlier of the hidden state generated by that layer.

[3]https://github.com/huggingface/transformers

[4]https://github.com/google-research/language/tree/master/language/multiberts

[5]https://github.com/gpucce/outliersvsfreq/tree/main

[6]We consider 8 GLUE (Wang et al., 2018) tasks: CoLA (Warstadt et al., 2018), SST (Socher et al., 2013), MRPC (Dolan and Brockett, 2005), STSB (Cer et al., 2017), MNLI (Williams et al., 2018a), QNLI (Rajpurkar et al., 2016a) and RTE (Bentivogli et al., 2009). We exclude WNLI task, which BERT is unable to "learn" (Prasanna et al., 2020).

[7]Kovaleva et al. (2021) also show that if the outliers are removed *before* fine-tuning, the model is able to recover without any negative effects.

Outliers removed	CIFAR10	CIFAR100
Full model	98.6	92.5
1 random dimension	98.6	92.5
$O759$	98.6	92.3
$O187$	98.6	90.5
2 random dimensions	98.6	92.4
$O759 + O187$	98.5	**84.9**

Table 2: Outlier removal effect for Visual Transfromer.

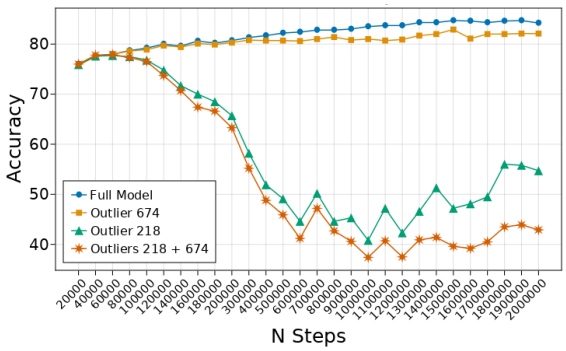


Figure 2: The accuracy on MNLI-matched of the checkpoints for BERT-base (seed 1) by Sellam et al. (2022) for full model or with each outlier removed.

experiments. We also confirm that RoBERTa-base behaves similarly (see App. C).

3.2 Outliers in Other Transformers

Kovaleva et al. (2021) focus exclusively on Transformer-based LMs. To establish whether outliers could be something specific to pre-training on language data, we investigate the presence of outliers in the Vision Transformer (ViT) (Kolesnikov et al., 2021). Table 2 shows ViT accuracy on CIFAR10 (Krizhevsky et al.) and CIFAR100 (Krizhevsky, 2009): image classification tasks with a choice between 10 and 100 possible labels respectively. Using the magnitude and position criteria we identify candidates $O759$ and $O187$, and we experiment with disabling one or both of them, as well as randomly selected dimensions as a control. For this model, the accuracy on MNLI can't be used as a measure for outliers, instead we use the accuracy on CIFAR100.

We see that, for CIFAR100, with both outliers disabled the model experiences $\approx 7\%$ loss in accuracy, but that does not happen for CIFAR10. The reason for that could be that CIFAR10 is a much simpler task, on which the model achieves above 98.5% accuracy. If the model succeeds in positioning the small number of classes sufficiently far apart in the representation space, then even the loss of outliers might be insufficient to disrupt that. If that is the reason for discrepancy between CIFAR10 and CIFAR100, then perhaps the 100-class classification is still an easier problem than the GLUE tasks, for which BERT degrades in performance significantly more (see Table 5).

We also explored two other Transformer-based models: ESM trained on protein sequences (Rao et al., 2020) and Wav2Vec trained on audio data (Baevski et al., 2020). We found no evidence for outliers there. This could be due to the fact that both of these models have a very small "vocabulary" (30-40 "tokens" vs tens of thousands for LMs).

3.3 Emergence of Outliers in Pre-Training

Kovaleva et al. (2021) pre-train a BERT-medium model for up to 250,000 steps. They find that outliers emerge relatively early in pre-training (step 50,000), and at about the same time LM perplexity starts to improve. A limitation of this experiment is a relatively small model, and the fact that both observed events coincide with the warm-up ending.

We examine the full BERT-base checkpoints released by Sellam et al. (2022), who pre-train five models from scratch with different random initializations. For each model they release the checkpoints for every 20,000 steps between 0 and 200,000 steps, and after that – for every 100,000 steps up to 2,000,000. We use the seed numbered as 1 (zero indexed). Like BERT-base and RoBERTa-base (§2), we find that this BERT also has two outliers, $O218$ and $O674$, the same for all the checkpoints for this seed.

We investigate the main outlier effect: the drop in performance of the model fine-tuned on our chosen representative downstream task, MNLI-matched (Williams et al., 2018b). Figure 2 shows the accuracy for all the checkpoints from seed 1, comparing the full model with the model with $O218$, $O674$, and both $O218$ and $O674$ removed. The expected effect is clearly observed after step 80000 for $O218$ and $O218 + O674$, but not $O674$ alone. This is consistent with the findings of Kovaleva et al. (2021) who also report various size of effects for outliers identified purely by magnitude. The results for MNLI-mismatched are similar and available in App. D.

After step 80,000 the full model steadily increases in accuracy, reaching 83.5% at step 10^6. Training for 10^6 more steps only achieves $\approx 1\%$

gain, illustrating the diminishing returns effect with further pre-training. The performance without outliers degrades over time, but at the later stages of pre-training (not observed by Kovaleva et al.) that trend is not steady: after $\approx 10^6$ steps the model accuracy with either $O218$ or $O218 + O674$ removed slowly grows over time, often with high variance between the "neighboring" checkpoints.

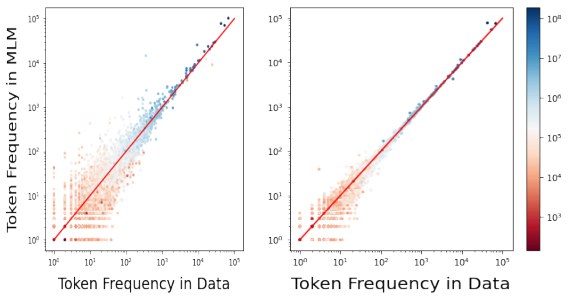
Another observation from Fig. 2 is that after the first 10^6 steps[8] , the difference between the accuracy of the model without the most disrupting outlier $O218$ and $O674$ increases. This suggests that the dynamics for the two outliers are different: while one gains importance from the early stages of pre-training, the other one rises after more optimization steps[9]. This may be related to the different behavior for the hidden state dimensions corresponding to the two outliers, which we will present in §4.

4 What Do the Outliers Impact?

4.1 Effects on Masked Language Modeling

So far we know that disabling the outliers negatively affects BERT downstream task performance (Fig. 2), but it is unclear *why* that happens. Since LMs rely on statistical patterns of token co-occurrence, token frequency in pre-training data[10] could be expected to affect the learned representations. We investigate whether outlier removal affects what kinds of tokens (in terms of their frequency in pre-training data) the MLM predicts.

Figure 3 shows the frequency of tokens predicted by the model over $200,000$ sentences from Wikipedia. We use the standard masking strategy: 15% tokens masked randomly. For BERT-base we observe that **the model with disabled outliers consistently predicts more tokens that were highly frequent in the training data**, and fewer tokens that were rare. RoBERTa shows a similar behaviour



(a) Without Outliers (b) Full Model

Figure 3: A log-log scatter plot of token generation frequency vs true token frequency in data in MLM. The x-axis represents the number of time a token has been masked and the y-axis the times it has been predicted. The color shows the token appearances in pre-training data. In (a) for the *bert-base-uncased* model with zeroed out outliers and in (b) for the full pre-trained model.

(see App. C for the details).

We also considered if the outliers impact the distribution of POS tags of the predicted tokens. We found that disabling $O381$ is the most disruptive and that, similarly to $O308$, it pushes the model towards predicting more nouns, punctuation, symbols and adpositions (see App. E for details).

4.2 Token Frequency Vs Performance

If outlier removal impacts the MLM ability to *predict* tokens it observed less often in pre-training (§4.1), could it also impact the model *encoding* of tokens more/less frequently seen in pre-training?

The LayerNorm outliers are an intrinsic property of the model itself. For this experiment we need to consider the interaction between the model and its input data. Hence we consider *the hidden state outlier dimensions*: the hidden state parameters at the dimensions corresponding to the outlier dimensions. They are the most affected by the outlier removal, since zeroing out a LayerNorm parameter removes precisely this component.

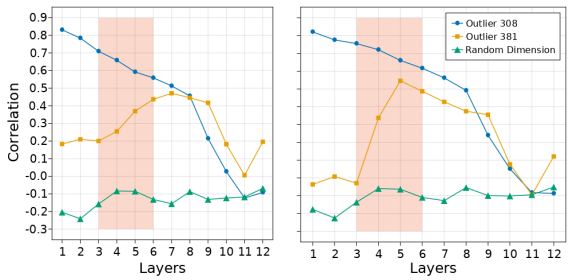
In this experiment we encode the validation set of Wikitext-v2 (Merity et al., 2016) by BERT-base, and we measure the Pearson correlation between pre-training data frequency of encoded tokens, and the magnitude of the hidden state parameters corresponding to the outlier dimensions ($O308$ and $O381$) in each layer (see App. B for more details). The results are presented in Fig. 4. We also track across all layers the main outlier effect (performance degradation when the outliers are disabled) in MLM and MNLI tasks, as shown in Fig. 5.

We find that for the hidden state parameters cor-

[8]Interestingly, the number of 10^6 steps is also the number of training steps mentioned in the original BERT paper (Devlin et al., 2019), and even the models by Sellam et al. (2022) (also from Google) do not match the originally reported performance at the original amount of pre-training. Sellam et al. (2022) state that they need to train for twice longer to reach comparable performance on all the tasks from GLUE (Wang et al., 2018) and SQuAD (Rajpurkar et al., 2016b).

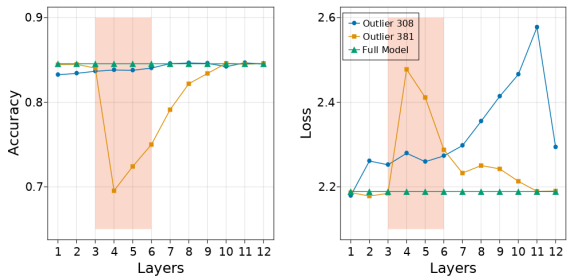
[9]Figure 2 shows the accuracy with different classification heads initialization. See App. D for the similar case with fixed initialization.

[10]To estimate the frequency in the pre-training data we use a corpus similar to BERT pre-training data: it contains the Book Corpus (Zhu et al., 2015) and Wikipedia dump from November 1st 2021.



(a) With special tokens (b) Without special tokens

Figure 4: BERT-base encoding Wikitext-v2 validation set: Pearson correlation between magnitude of hidden state parameters corresponding to outlier dimensions, and frequency of encoded tokens in pre-training data.



(a) MNLI-m performance (b) MLM loss (in wikitext-v2)

Figure 5: BERT-base: effect of disabling outliers on MNLI-matched and MLM loss.

responding to both $O308$ and $O381$ the correlation between their magnitude and encoded token frequency is much higher than for random dimensions, but they exhibit different layer-wise trends for that correlation vs impact on model performance:

Case 1: the magnitude of the hidden state parameters corresponding to the outlier dimensions is directly proportional to both its correlation with the encoded token frequency, and performance drop after removal of LayerNorm outlier parameters. For the hidden state dimension corresponding to $O381$, the correlation of the hidden state parameter magnitude with the encoded token frequency is closer to zero at the initial and final layers, and high in the middle layers (this trend continues until layer 9 when special tokens are included). Fig. 5b shows that the removal of $O381$ has largest impact (in both MNLI and MLM) in layers 4-6. Coincidentally, Fig. 4b shows that layers 4-6 are also the layers where the magnitude of hidden state dimension corresponding to $O381$ correlates with token frequency the most.

Case 2: the magnitude of the hidden state parameters corresponding to the outlier dimensions, and their correlation with the encoded token frequency are both inversely proportional to the performance drop after removal of LayerNorm outlier parameters. For $O308$ the pattern is the opposite: the magnitude of its corresponding hidden state parameter strongly correlates with encoded token frequency at the initial layers, but not in the final ones. However, Fig. 5b shows that the removal of this LayerNorm outlier has a larger impact on MLM loss on the final layers[11]. As a result, the removal of $O308$ is less harmful for most downstream tasks as shown in Table 1 because fine-tuning mostly affects the layers closer to the output (Liu et al., 2019a; Kovaleva et al., 2019), therefore it cancels a part of the effect of disabling this parameter.

To confirm that this is not a pattern specific to BERT we also perform the same experiments for RoBERTa-base, and we find that it also has the two kinds of outliers with the direct and inverse relationship to performance drop ($O588$ and $O77$ respectively). The data for these experiments is available in App. C.

Since BERT encodes sequences always starting with '[CLS]' and ending with '[SEP]', these special tokens could store positional information, and they are also highly frequent. Therefore we repeat the experiment discarding them (Fig. 4b), but the overall trend is not affected.

4.3 What Happens to Attention?

In §4.2 we showed that there is a correlation between the magnitude of the hidden state parameters corresponding to outlier dimensions, and the token frequency in the pre-training data. Prior work (Clark et al., 2019; Kovaleva et al., 2019) showed that BERT self-attention often "points" to highly frequent tokens, including the special tokens and punctuation marks. Given this, our next question is whether the outliers also affect the self-attention patterns. As argued by Dong et al. (2021), attention alone would map tokens to very low dimensional spaces, and in that case the outlier phenomenon would be consistent with such a mapping.

We find that there is indeed such an effect. To illustrate it we encode a MNLI sample with BERT-

[11]The main discrepancy in this pattern is the frequency correlation of the hidden state dimension corresponding to $O308$, and its MLM loss at the last layer. However, the lower loss can be a consequence of the parameter not affecting any following Transformer layer.

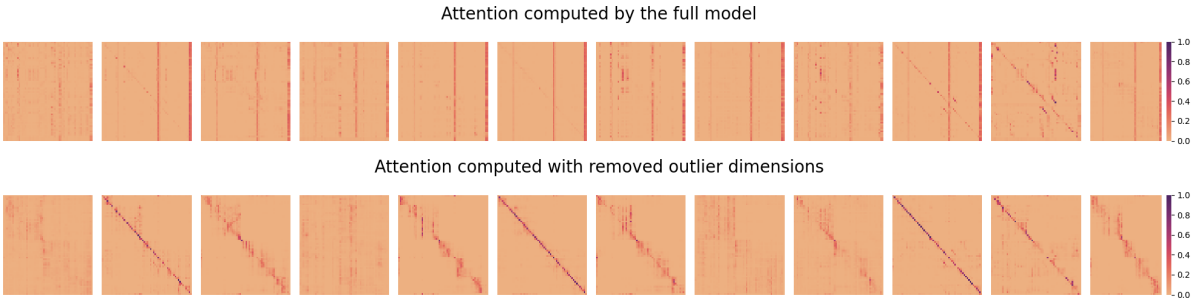


Figure 6: The self-attention patterns at the 10th layer of the full 'bert-base-uncased' pre-trained model vs the same model with removed LayerNorm outliers.
Encoded example from MNLI: *[CLS] Thebes held onto power until the 12th Dynasty, when its first king, Amenemhet I who reigned between 1980 1951 b.c. established a capital near Memphis.[SEP] The capital near Memphis lasted only half a century before its inhabitants abandoned it for the next capital. [SEP]*

base. Fig. 6 shows the self-attention maps for the 12 heads of the 10th layer[12], directly comparing the self-attention in a full model vs a model with the outlier dimensions removed.

The most conspicuous difference is the fact that the vertical bars in the self-attention maps of the full model vanish once the outliers are zeroed out. This "vertical" attention pattern has been reported before (Kovaleva et al., 2019), and in BERT it often corresponds to attention to special tokens and punctuation. It may seem that without the outliers the diagonal patterns become more salient, but in fact they are also present with the intact outliers, and their increased saliency in the plot is simply an effect of softmax normalization.

Figure 6 only shows a single example. To establish whether this effect is stable, we encode 1500 sequences from Wikitext-v2 validation set and measure the Pearson correlation between *average vertical attention value* of each token (the average over attention columns in the encoded sequence), and the magnitude of the hidden state parameters corresponding to the outlier dimensions. In cases of the "vertical" self-attention pattern, the average vertical attention value would be relatively high.

Figure 7 shows the results of this experiment, which we repeat with and without BERT special tokens ('[CLS]' and '[SEP]'). As a control, Fig. 7c and Fig. 7f show the average correlation over a sample of hidden state parameters at random dimensions. For the randomly picked weights the correlation is ≈ 0, which is expected (since these vectors have length 768, the individual dimensions of randomly sampled vectors should have a negligible contribution).

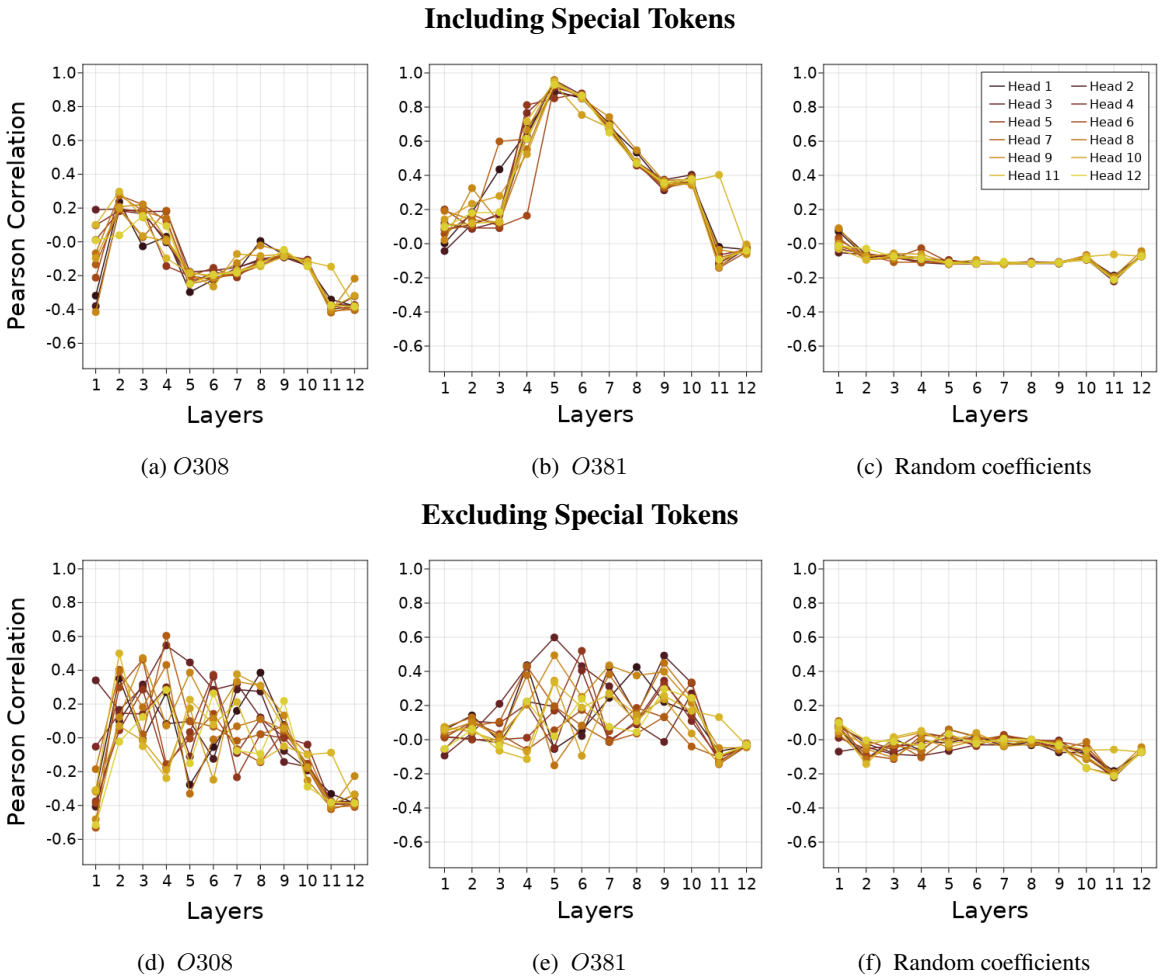
Compared to random dimensions, the hidden state parameters at dimensions corresponding to both $O308$ and $O381$ have on average a significantly higher correlation between their magnitude and average self-attention query values. This confirms that the pattern shown in Fig. 6 is prevalent, and the tokens with high hidden state outlier dimension value tend to also have high average value over attention columns, i.e. **they are attended to by most other tokens**.

An unexpected pattern is represented by the negative correlations in Fig. 7a and Fig. 7d at initial and final layers. We argue that at early layers this happens because the vertical patterns are less frequent, while at the final ones because the outliers in those layers are less relevant. The trend is similar to what we observed in Fig. 4.

We also observe several trends that mirror the observations from §4.2:

- The hidden state parameter value corresponding to $O308$ has a higher correlation with average vertical attention value since the initial layers (except the very first) which decreases at the final layers. For parameters corresponding to $O381$, the correlation grows at layer 4-5 and then vanishes at the final one. Both of these trends are consistent with Fig. 4 showing the correlation to frequency.
- Special tokens affect these trends: Fig. 7d and Fig. 7e show that excluding them does not fundamentally change the pattern, but the results become less stable across heads.
- Both Fig. 4 and Fig. 7 show large variation as the information flows through the model,

[12] We choose the 10th layer because prior work suggests that the layers closer to the output are more affected by fine-tuning (Kovaleva et al., 2019), and also encode more task-specific information (Liu et al., 2019a).



(a) $O308$ (b) $O381$ (c) Random coefficients

(d) $O308$ (e) $O381$ (f) Random coefficients

Figure 7: Each figure shows the correlation between the *average vertical attention values* in BERT-base self-attention heads, and the magnitude of hidden state parameters at the dimensions corresponding to outlier dimensions. The correlation is computed over examples from Wikitext-v2. Fig. (c) and (f) show the average over 10 random dimensions.

which suggests that the effect is not entirely formed at the model input.

Overall the results of this experiment suggest that **the relationship between the outlier phenomenon and encoded token frequencies in pre-training data also affects the self-attention mechanism of BERT**. In particular, it affects the "vertical" attention pattern in which a token is attended to by most other tokens, and which was previously reported for the high-frequency special tokens. We confirmed that RoBERTa self-attention exhibits a similar pattern (see App. C).

4.4 What Causes the Outliers?

We have now identified a correlation between the magnitude of hidden state parameters corresponding to outlier dimensions, and the frequency of the encoded tokens in the pre-training data. However, it is unclear whether the relationship is causal.

To establish that, we pre-train[13] from scratch 3 versions of BERT-medium as defined by Turc et al. (2019), with the following tokenization schemes:

- *SENTENCE*: We split sentences using a Spacy sentencizer[14], and add a '[SEP]' token at the end of each sentence and at the end of each encoded sequence. This is similar to the "full sentences" tokenization used to train RoBERTa (Liu et al., 2019b).
- *CHUNK*: We add a single '[SEP]' token at the end of each sequence of 256 tokens rather

[13] Except for the tokenization strategy, the training for each model is similar to the original BERT (Devlin et al., 2019) with two exceptions: (a) the Wikipedia corpus is a more recent version, from 01/03/2022, (b) the max sequence length is 256 (instead of 512) and batch size 128 instead of 256 due to computational constraints (these appear to have limited effect on the MNLI benchmark). All models were trained for 327,500 steps.

[14] https://spacy.io/

	SENTENCE	*CHUNK*	*SENTENCE_FREQ*
Full Model	79.6	79.2	76.8
Minus $O378$	**66.9**	**47.4**	-
Minus $O281$	78.8	-	-
Minus $O353$	-	-	75.8
Minus $O362$	-	-	**74.4**

Table 3: Accuracy on MNLI-matched for each pre-training setting of BERT-medium.

than each sentence. The main effect that we expect from this is that the amount of '[SEP]' tokens is reduced roughly by a factor of 10.

- *SENTENCE_FREQ*: Each sentence is followed by the '[SEP]' token, but we replace 50% of occurrences of regular tokens with frequency above 1.e-5 in the training corpus with a random token with a frequency below 1.e-5 in 50% of their occurrences[15].

Note that the RoBERTa-like *SENTENCE* tokenization is different from the classic BERT approach, where each encoded sequence always contains exactly 2 '[SEP]' tokens in each encoded sequence. The RoBERTa approach would make this token more frequent for the sequences containing more than one sentence, and hence also more appropriate for testing our frequency hypothesis.

Both *CHUNK* and *SENTENCE_FREQ* conditions corrupt the linguistic structure of the input, and so the trained MLM quality could be expected to drop as it acquires worse knowledge. But this setting will let us (a) identify the impact of token frequency on the outlier phenomenon, (b) disentangle the effect between frequent tokens in general and the '[SEP]' token.

All three models started from the same initialization but were fed different data according to the tokenization schemes. We find that *SENTENCE* model developed outliers $O281$ and $O378$, whereas in *CHUNK* the detrimental effect is only clear for $O378$. The *SENTENCE_FREQ* model developed two outliers: $O353$ and $O362$.

Table 3 shows all three BERT-medium models evaluated on MNLI-matched validation set as either the full model or with their respective outliers removed one by one. *SENTENCE* model is the best performing overall, but *CHUNK* is only .4 points behind as the full model. Both of them develop a very damaging outlier $O378$, whereas the effect of $O281$ is less pronounced in *SENTENCE* and insignificant in *CHUNK*. Moreover, the single outlier in *CHUNK* is more damaging for the model. One possible explanation is that when the model has only one outlier, it likely relies on it more, which would result in higher performance degradation when it is disabled.

As expected, the *SENTENCE_FREQ* model that was fed the noisiest data performs worse than the other two (by $\approx 3\%$). But interestingly, it also does not develop any outliers as damaging as $O378$ is for the other two models.

We conclude that the frequency distribution of tokens in pre-training data contributes to the outlier phenomenon, and the '[SEP]' token is a part of that effect (since high frequency is one of the factors that characterizes it).

5 Discussion

5.1 Outliers in Transformer Pre-training

Prior work (Kovaleva et al., 2021) showed that the outliers are present in a large number of Transformer-based LMs. We provide complementary evidence for the Vision Transformer (Table 2). However, we were unable to identify outliers in protein and audio Transformers, which we attribute to significantly smaller vocabulary size. This finding hints towards the training data distribution being at the core of the outlier phenomenon.

Kovaleva et al. (2021) also show that outliers emerge early in pre-training (after 50K steps for BERT-medium). We extend that experiment by investigating the fully pre-trained BERT-base by Sellam et al. (2022), we find that the impact of outliers on the model grows up steadily until step 10^6. After that step the outlier effect is inconsistent between checkpoints, and the full model performance saturates. An interesting question for future work is what happens after outlier removal stops degrading model performance (around step 10^6), and whether it could be used as an early stopping criterion.

Transformer-based language models (LMs) have been shown to exhibit anisotropic behavior in their representations of both tokens and sentences (Ethayarajh, 2019; Gao et al., 2019; Rajaee and Pilehvar, 2021; Timkey and van Schijndel, 2021). While pervasive, this is an undesirable property because it reduces the average distance between tokens embeddings, and thus makes it more difficult to distin-

[15]Due to high computational costs of BERT pre-training we only experiment with one possible value of the threshold (1.e-5). When exactly tokens become "high frequency" for BERT-type MLMs remains a question for future work.

guish between tokens in the embedding space.

One consequence of the outliers growth over pre-training is that the attempts to remove anisotropy at the downstream task level (Rajaee and Pilehvar, 2021), although effective in some cases, could be only partially addressing the problem. In that case it might be more productive to change pre-training so as to better account for the skewed token frequency distribution (Li et al., 2021b).

5.2 Outliers and Token Frequency

Li et al. (2021a) and Gao et al. (2019) show that embeddings of low frequency tokens lie further away from high frequency ones in the embedding space. In §4.2 we showed how the outlier parameters influence the hidden state geometry proportionally to token frequency, and how this is more sensitive at earlier layers. This is consistent with findings of Li et al. (2021a) who show that the different geometry of frequent and non-frequent tokens is more evident for the layers closer to the input. Indeed, we observe this effect for $O381$ in BERT-base and $O588$ in RoBERTa-base (see Fig. 4).

To the best of our knowledge, this is the first work to demonstrate the link not only between the geometry of the hidden states and frequency of encoded tokens in pre-training data, but also the model performance.

5.3 Outliers and Positional Embeddings

Concurrently with the demonstration of the outlier phenomenon by Kovaleva et al. (2021), Luo et al. (2021) attributed the high-magnitude weights to a different source: positional embeddings rather than LayerNorm weights. The positional embeddings could be expected to have more impact in the earlier layers. Our work contributes to the dispute by showing that two different behaviours are present in both BERT-base and RoBERTa-base: one outlier dimension in the hidden states is disruptive in layers 4-6 ($O381$ for BERT and $O588$ for RoBERTa) while the other one at the layers 10-11 ($O308$ for BERT and $O77$ for RoBERTa). This suggests that both mechanisms may play a role.

5.4 Outliers and Self-Attention

Kovaleva et al. (2019) identify 5 frequent self-attention patterns, 4 of which include vertical lines corresponding to special tokens. We showed (Fig. 7) that the presence of special tokens increases the correlation (in absolute value) between the average query value and the magnitude of the hidden state dimensions corresponding to outlier dimensions. This suggests that the outlier phenomenon contributes to the vertical attention patterns identified by Kovaleva et al. (2019). From the computational perspective this is consistent with the attention being a bilinear form. Moreover, the relation between the outliers and the vertical self-attention pattern (often "pointing" to the highly frequent special tokens and punctuation) also hints at the relation between outliers and the token distribution in the pre-training data.

At the same time, the correlation remains evident in the final layers even when special tokens are ignored, indicating that the outliers also contribute to the attention shape more broadly. This is in line with Kobayashi et al. (2020) who argue that vertical patterns in attention do not indicate that no other information is encoded (hence simply norming the self-attention makes other relations more salient).

6 Conclusion

To the best of our knowledge, this is the first work to directly link the outlier dimension phenomenon in Transformer-based models (in particular BERT and RoBERTa) to encoded token frequency in pre-training data. We also find that the magnitude of hidden state dimensions corresponding to outliers correlates with the vertical self-attention pattern, which enables the attention to the classification tokens. Furthermore, we find that there are two types of outliers: some of them affect the MLM performance the most in the middle layers (where the correlation with token frequency is also at its peak), and for others the impact grows towards the final layers (even though the correlation with token frequency decreases).

Our findings suggest that outliers are due not to the Transformer architecture per se, but rather to the highly skewed token frequency distribution in textual pre-training data. In that case, to mitigate anisotropy we might need to design a pre-training scheme that better accounts for such distributions.

7 Limitations

This work establishes a relation between the outlier phenomenon in Transformer-based language models and the frequency of tokens in the corpus used for pre-training. We focus on two of the most popular Transformers (BERT-base and RoBERTa-base) and show that our key observations hold for both of them, but there are hundreds of other possible

Transformer-based LM architectures and modifications to pre-training regimes and tokenization that could be explored in future work. That being said, we believe that a thorough examination of the most common models, such as presented in this paper, is pre-requisite to establishing the methods and hypotheses for a large-scale study.

Methodologically, our experiments have the following limitations:

- We identify a correlation between frequency of the encoded tokens in the pre-training corpus, and the magnitude of the coefficients within the hidden states of Transformer layers which correspond to the outlier dimensions (§4.2). A limitation of this experiment is that it only establishes that there is mutual influence between token frequency and outliers. We cannot exclude the presence of covariates or claim that the token distribution is the reason behind the outlier phenomenon.

- To establish whether there is also a causal relation, we show that pre-training a *bert-medium* model with changes in the tokenization reduces the impact of outlier removal (§4.4). This experiment does show that we can almost remove the outliers effect by changing the token distribution, however, these changes by themselves degrade the quality of the model and its downstream task performance, and therefore the reduced outlier removal effect could be partly due to the model being overall less performant. This also raises a follow-up question for future work: do the special tokens have such a connection to the outlier phenomenon partly because of their special role, or simply due to the fact that they are among the most frequent tokens in the pre-training corpus?

- We find that outliers have a strong impact on the shape of attention heads, most notably the "vertical" patterns (§4.3), and we hypothesize that the outlier removal effect on downstream task performance may thus be explained by the inability of the model to "focus" on the special tokens, which according to prior work is a key role of the "vertical" self-attention patterns. This hypothesis merits further investigation.

8 Broader Impacts

This work focuses on the analysis of two popular Transformer-based LMs of the BERT family (BERT- and RoBERTa-base). This work relies on established benchmarks, does not collect new human subjects data and presents no new models. Its broader impacts center on improving the mechanistic understanding of training Transformer-based LMs, which could lead to developing better models in the future. We also presented evidence of outlier phenomenon in Vision Transformer, which suggests that vision and multimodal Transformers may also be vulnerable to attacks involving direct modification of outlier weights.

Experiments were conducted using a private infrastructure, which has a estimated carbon efficiency of 0.37 kgCO_2eq/kWh (average carbon efficiency in Japan, where the machine is based, for the year 2020). Including experiments that were discarded and failed runs, we estimate that a cumulative of 200 hours of computation was performed on hardware of type RTX A6000 (TDP of 300W). Total emissions are estimated to be 22.2 kgCO_2eq.

Acknowledgements

We would like to thank Olga Kovaleva, Anna Rumshisky, and the anonymous reviewers for their insightful comments. This work is partially supported by JST KAKENHI grant JP22H03600 and JST CREST grant JPMJCR19F5. This work used computational resources of the supercomputer Fugaku provided by RIKEN through the HPCI Fugaku General Access (Small-Scale) Project (Project ID: hp210265).

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Luisa Bentivogli, Bernardo Magnini, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. 2009. The fifth PASCAL recognizing textual entailment challenge. In *Proceedings of the Second Text Analysis Conference, TAC 2009, Gaithersburg, Maryland, USA, November 16-17, 2009*. NIST.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. 2021. Understanding and overcoming the challenges of efficient transformer quantization.

In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7947–7969, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin. 2020. The Lottery Ticket Hypothesis for Pre-trained BERT Networks. *arXiv:2007.12223 [cs, stat]*.

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

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm.int8(): 8-bit matrix multiplication for transformers at scale.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*. Asia Federation of Natural Language Processing.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. 2021. Attention is not all you need: pure attention loses rank doubly exponentially with depth. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2793–2803. PMLR.

Kawin Ethayarajh. 2019. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Representation degeneration problem in training natural language generation models. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Mitchell Gordon, Kevin Duh, and Nicholas Andrews. 2020. Compressing BERT: Studying the effects of weight pruning on transfer learning. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 143–155, Online. Association for Computational Linguistics.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

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

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

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

Alex Krizhevsky. 2009. Learning multiple layers of features from tiny images. Technical report.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research).

Bai Li, Zining Zhu, Guillaume Thomas, Yang Xu, and Frank Rudzicz. 2021a. How is BERT surprised? layerwise detection of linguistic anomalies. 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 4215–4228, Online. Association for Computational Linguistics.

Yan Li, Dhruv Choudhary, Xiaohan Wei, Baichuan Yuan, Bhargav Bhushanam, Tuo Zhao, and Guanghui Lan. 2021b. Frequency-aware SGD for efficient embedding learning with provable benefits. *CoRR*, abs/2110.04844.

Yuxin Liang, Rui Cao, Jie Zheng, Jie Ren, and Ling Gao. 2021. Learning to Remove: Towards Isotropic Pre-trained BERT Embedding. *arXiv preprint arXiv:2104.05274*.

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. Association for Computational Linguistics.

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.

Ziyang Luo, Artur Kulmizev, and Xiaoxi Mao. 2021. Positional Artefacts Propagate Through Masked Language Model Embeddings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5312–5327, Online. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *CoRR*, abs/1609.07843.

Sai Prasanna, Anna Rogers, and Anna Rumshisky. 2020. When BERT Plays the Lottery, All Tickets Are Winning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3208–3229, Online. Association for Computational Linguistics.

Giovanni Puccetti, Alessio Miaschi, and Felice Dell'Orletta. 2021. How Do BERT Embeddings Organize Linguistic Knowledge? In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 48–57, Online. Association for Computational Linguistics.

Sara Rajaee and Mohammad Taher Pilehvar. 2021. How Does Fine-tuning Affect the Geometry of Embedding Space: A Case Study on Isotropy. *arXiv preprint arXiv:2109.04740*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016a. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016b. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Roshan M Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. 2020. Transformer protein language models are unsupervised structure learners. *bioRxiv*.

Victor Sanh, Thomas Wolf, and Alexander Rush. 2020. Movement pruning: Adaptive sparsity by fine-tuning. In *Advances in Neural Information Processing Systems*, volume 33, pages 20378–20389. Curran Associates, Inc.

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

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

William Timkey and Marten van Schijndel. 2021. All Bark and No Bite: Rogue Dimensions in Transformer Language Models Obscure Representational Quality. *arXiv:2109.04404 [cs]*.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: The impact of student initialization on knowledge distillation. *CoRR*, abs/1908.08962.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018a. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018b. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Model name	Outlier 1	Outlier 2
"bert-base-uncased"	308	381
"roberta-base"	77	588
"multiberts-seed-1"	218	674
"google/vit-095base-patch16-224-in21k"	187	759
"BERT-medium (ours)"	281	378

Table 4: The outliers identified for each model used in the paper

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

A Outliers For Each Model

As described in §2 the outliers are zeroed out in the LayerNorm layers, in §2 we mention that the definition of outliers is not entirely formal: while the weights magnitude let us identify a small subset of weight among which we can search for outliers, we need to fine tune the model on a downstream task, we use MNLI although all tasks in GLUE would work, to identify which weights are the most harmful. Table 4 lists the two most damaging outliers for each model we used in the paper.

B Replicability

In this section we describe in detail all the experiments carried out within this work, together with the code used to make them this should allow for an effective reproduction of our results. All experiments are carried out using a NVidia A6000 with 48 gbyte of memory.

- For the results in Table 1 we fine tune *bert-base-uncased* for 4 epochs on each task with a 2.e-5 learning rate and 256 maximum sequence length. We measure the respective metric for each GLUE task (as defined by Wang et al. (2018)) on the validation set. Both models and datasets are loaded through huggingface `https://huggingface.co/`. For the computation with removed outliers, what we do is we compute the same metric as for the full model after manually setting to 0 the chosen LayerNorm *weight* and *bias* parameters in all layers. A similar procedure is adopted to compute the values in Table 2. Fine tuning on the largest datasets within the glue benchmarks (mnli, qnli, qqp), with the hyperparameters described above on average requires approximately 4000 seconds. The remaining datasets among the glue benchmarks are between 10 to 100 times smaller and require a proportionally scaled amount of time.

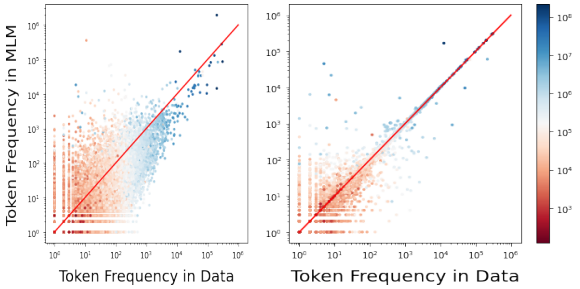
- The token counts in Fig. 8 are obtained through Wikipedia and book corpus by directly using a *bert-base-uncased* and *roberta-base* tokenizers on the whole corpus and counting each token occurrence.

- The results in Fig. 4 are obtained as follows: for each token in the data (as part of an encoded sequence) we compute the hidden states through a *bert-base-ucased* model and pick the hidden state parameter at the outlier index therefore getting a single numerical value for each token. We also associate to each token its frequency in the pre-training corpus and we measure the Pearson correlation coefficient between this two lists of values.

- The results in 5 are obtained by setting LayerNorm *weight* and *bias* parameters at the given outlier index for a given layer to 0. For Fig. 5a this is done for a model fine-tuned on MNLI train set and we measure the accuracy on MNLI matched, for Fig. 5b this is done to a pre-trained only model by measuring the MLM loss on the wikitext-v2 validation set (the masking probabilities are kept as in the original BERT paper). This process is repeated for each layer in the model.

- The results in Fig. 7 are obtained as follows: for each sample in the wikitext-v2 validation set (a single sequence containing n tokens), we encode it with *bert-base-uncased*. This provides us with attention matrices with size $n \times n$ we take the average over the columns, thus getting a single numerical value for each token. As above, for each token we also collect the hidden state value at the outlier dimension, a single numerical value (for each layer) for each token, and finally we measure the correlation between these two values. In particular since for each layer there are 12 heads we compute 12 correlations at each layer.

- The scores in Table 3 are obtained as for Table 1 on three instances of *bert-medium* architecture pre-trained with different tokenization

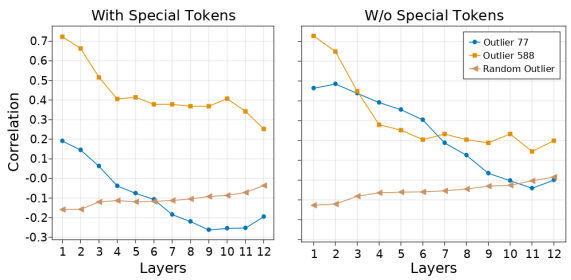


(a) Without Outliers (b) Full Model

Figure 8: A log-log scatter plot of token generation frequency vs true token frequency in data in MLM. The x-axis represents the number of time a token has been masked and the y-axis the times it has been predicted. The color shows the token frequency in pre-training data (wikipedia + book corpus). In (a) for the *roberta-base* model with zeroed out outliers and in (b) for the pre-trained model.

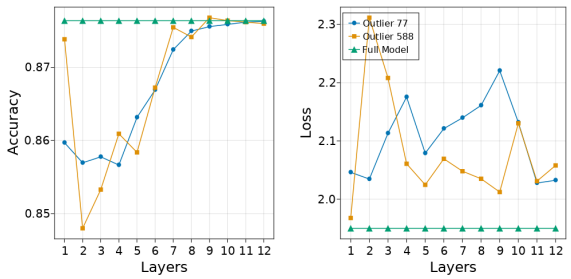
Outliers	cola	mnli-mm	mnli	mrpc	qnli	qqp	rte	sst2	stsb
baseline	58.3	87.4	87.6	87.3	92.7	91.4	69.0	95.0	89.1
77	51.5	85.4	85.5	80.1	89.8	90.4	65.0	93.9	83.7
588	7.4	61.5	59.4	70.8	56.6	64.2	54.2	70.3	19.1
588, 77	12.6	45.9	44.9	70.3	50.6	61.2	51.6	68.8	5.4

Table 5: Full RoBERTa scores on GLUE benchmarks with outlier effects.



(a) With special tokens (b) Without special tokens

Figure 9: Correlation of outlier dimension magnitude with token frequency over the Wikitext corpus for a pre-trained RoBERTa-base model. In (a) the correlations accounts for special tokens, in (b) they are excluded.



(a) MNLI-m performance (b) MLM loss (in wikitext-v2)

Figure 10: RoBERTa-base: effect of disabling outliers.

strategies. The pretraining of this model is performed with 256 max length, 128 batch size and 1.e-4 learning rate.

Experiments were conducted using a private infrastructure, which has a estimated carbon efficiency of 0.37 kgCO_2eq/kWh (average carbon efficiency in Japan, where the machine is based, for the year 2020). Including experiments that were discarded and failed runs, we estimate that a cumulative of 200 hours of computation was performed on hardware of type RTX A6000 (TDP of 300W). Total emissions are estimated to be 22.2 kgCO_2eq.

C RoBERTa Experiments

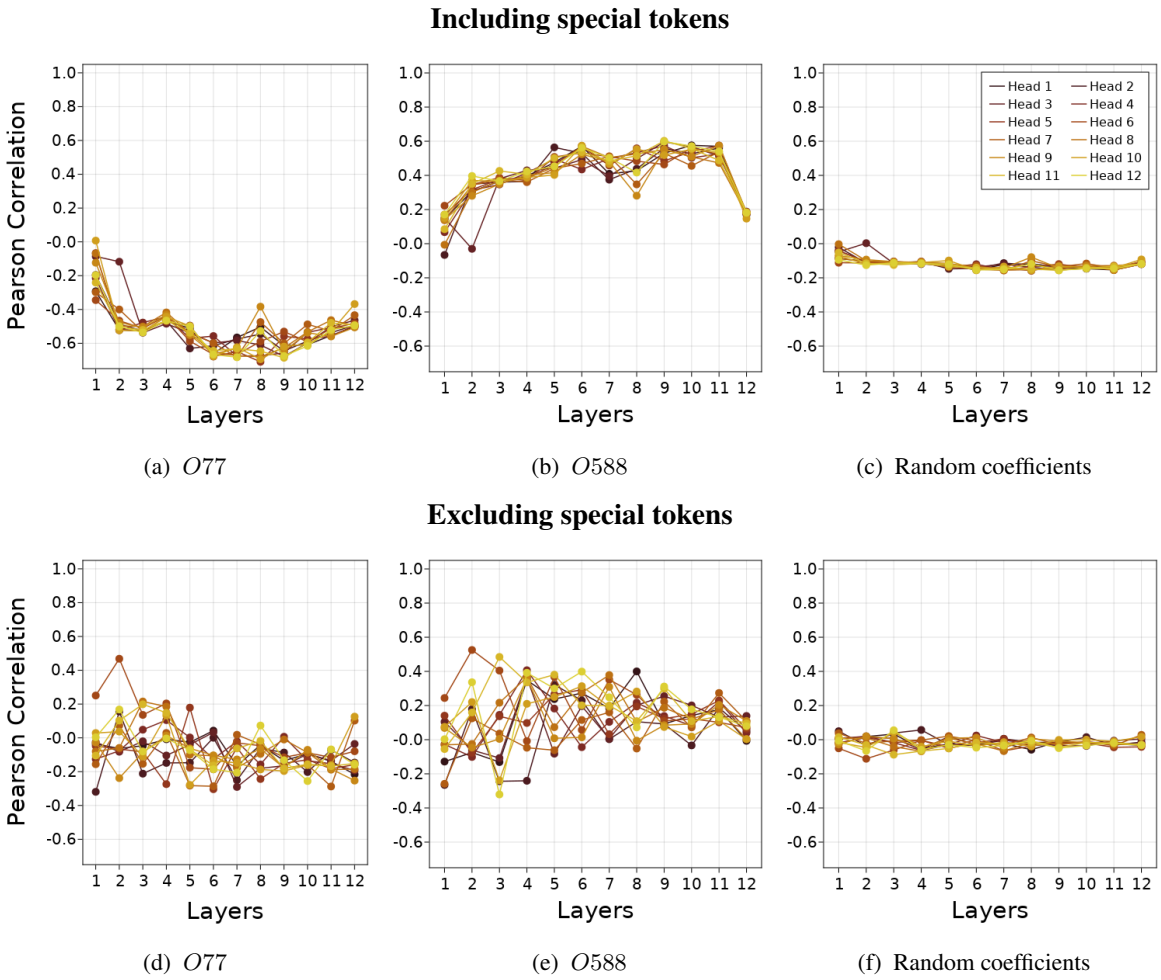
The results we showed for BERT-base similarly hold for RoBERTa-base. The generation distribution with removed outliers, Fig. 8a, shows that a single token on the top right, the "</s>" token, is generated a larger number of times (log scale), making this coherent with the results for BERT. We note that RoBERTa pre-training data is not the same as we use[16], however the core of the vocabulary is shared and therefore the qualitative results shown in Fig. 8 are reliable.

Table 5 shows the performance degradation with outliers removed on all GLUE tasks. As shown by Kovaleva et al. (2021) there is one more damaging outlier $O588$ and a less damaging one $O77$, when removed together they cause the largest performance degradation. Fig. 9a and Fig. 10 show that for RoBERTa patterns similar to those we see for BERT in Fig. 4 and Fig. 5 appear.

In particular, $O588$ is more damaging when the magnitude of the respective hidden state outlier dimension correlates the most to token frequency. In this case at layers 2-4 and at layer 10. In Fig. 9b we observe a spike in correlation with frequency, and Fig. 10b shows a similar one for MLM loss. On the other hand, $O77$ shows that the less the hidden state dimension corresponding to the outlier correlates to frequency, the more the removal of the LayerNorm outlier damages the model.

For this model we also see an anti-pattern at layer 4 (Fig. 10b): the loss with $O77$ is higher

[16]For RoBERTa the token frequency in pre-training is computed on Wikipedia + Book corpus plus an open source version of OpenWebText (`https://huggingface.co/datasets/openwebtext`), however RoBERTa pre-training data also include Stories and CC-news datasets not openly available.



(a) $O77$ (b) $O588$ (c) Random coefficients

(d) $O77$ (e) $O588$ (f) Random coefficients

Figure 11: Each figure shows the correlation between the *average vertical attention values* in RoBERTa-base self-attention heads, and the magnitude of hidden state parameters at the dimensions corresponding to outlier dimensions. The correlation is computed over examples from Wikitext-v2. Figures (c) and (f) show the average over 10 random dimensions.

and the one with $O588$ is lower. However, Fig. 9a shows that layer 4 is where the correlation including special tokens is closest to zero, possibly due to RoBERTa pre-training schedule including a larger number of special tokens (Liu et al., 2019b).

The general pattern observed for BERT is kept, however, while for BERT the worst layers in term of performance are layers 4-5, for RoBERTa this are layers closer to the input 1-2. One of the reasons behind this difference could be that RoBERTa had longer pre-training.

Finally we also replicate the analysis of attention patterns. Fig. 11 shows for RoBERTa the same patters that Fig. 7 shows for BERT: for the hidden state parameters corresponding to the outlier dimensions, the correlation values are very different when compared to random ones, both when including the special tokens or not.

D Outliers in Pre-Training

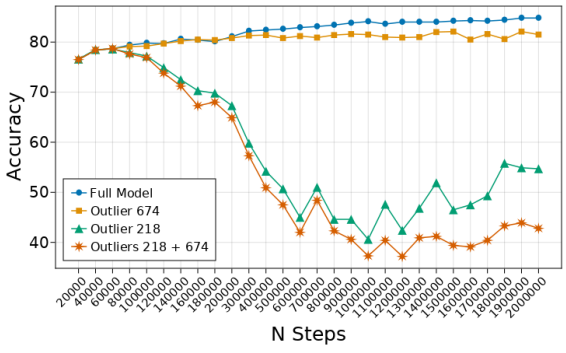


Figure 12: The accuracy on MNLI-mismatched of the checkpoints for BERT-base (seed 1), provided by Sellam et al. (2022).

Figure 12 shows the accuracy on MNLI-mismatched, at various checkpoints for BERT-base seed 1 provided Sellam et al. (2022). The results

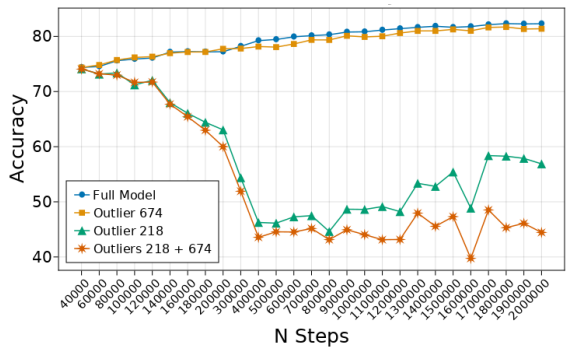


Figure 13: The accuracy on MNLI-matched of the checkpoints for BERT-base (seed 1) by Sellam et al. (2022) for full model or with each outlier removed. All classification heads are equally initialized.

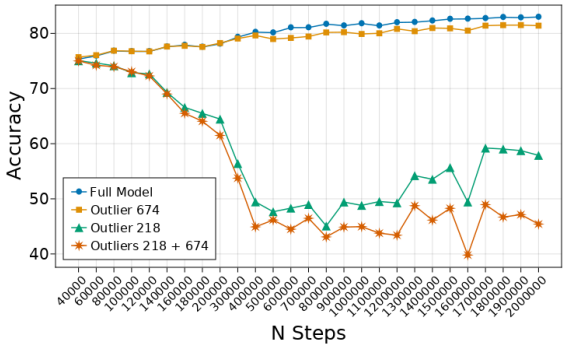


Figure 14: The accuracy on MNLI-mismatched of the checkpoints for BERT-base (seed 1), provided by Sellam et al. (2022). All classification heads are equally initialized.

are very similar to what Fig. 2 shows for MNLI-matched: early degradation around 80,000 steps, almost steadily worsening until step 1,000,000, and then fluctuating further on. The initialization of the classification layer is not fixed across checkpoints.

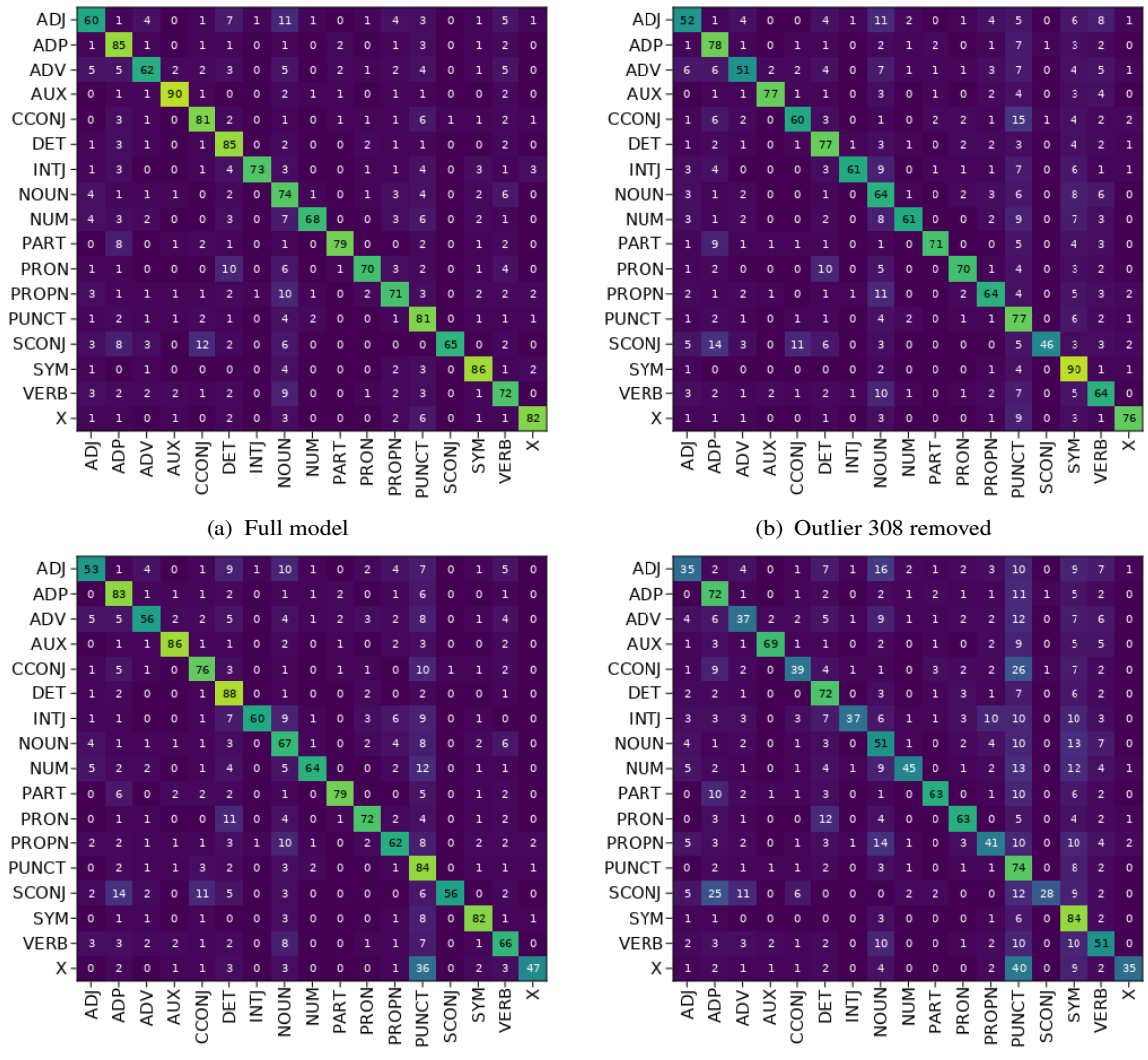
In Fig. 13 and Fig. 14 we also replicate the experiments while fixing the classification head seed at initialization. In this case as well the results are very close to those from Fig. 2 and Fig. 12. Specifically, the fluctuating behaviour appearing after 1 million steps is very distinct in this case as well. It is therefore not caused by changes in different fine-tuning initialization but confirmed to be caused by the number of pre-training steps.

An interesting question for future research on this topic is, what is the influence of longer pre-training on this phenomenon, does it get slowly cancelled? Does adding pre-training data from sources other than Wikipedia, the largest source of data for the models we investigate, make the outliers effect smaller or larger?

E POS Tag Distribution of Tokens Predicted by BERT MLM with Disabled Outliers

In this experiment we investigated the POS tags of the tokens predicted by the BERT-base MLM with disabled outlier dimensions. Fig. 15 shows the distribution of tags over the replaced tokens. Each row shows the percentage of tags of generated tokens with respect to the tag of the masked token: for example, the top row in Figure 15d shows that ADJ tokens are replaced with 16% probability by NOUN tokens, with 35% by ADJ tokens, with 10% by PUNCT tokens and so on.

We have previously shown in Table 1 that out of two outliers one damages the model performance considerably more. This pattern is also observed here. Fig. 15 shows that individually $O381$ has a much larger effect than $O308$. We can also see the qualitative difference between the outliers in the distribution of POS tags of the generated tokens: with only $O381$ disabled, the model becomes more likely to generate nouns and punctuation signs, while $O308$ does not produce so many changes. However, $O308$ has a larger effect in combination with $O381$, again pushing the model towards generating more nouns and punctuation, but also symbols and adpositions.



(a) Full model

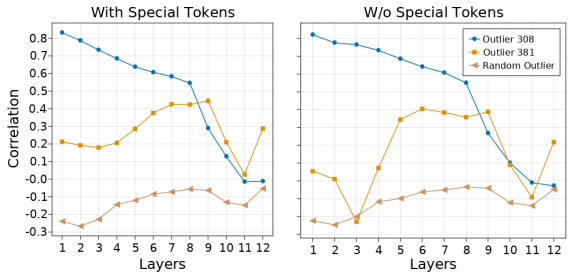
(b) Outlier 308 removed

(c) Outlier 381 removed

(d) Outliers 308 and 381 removed

Figure 15: The shift in percentage between the POS tags generated through MLM for full BERT-base model(a) and with different outliers removed, number 308 (b), number 381 (c) and together number 308 an number 381 (d).

F Outliers vs Encoded Token Frequency: the Case of Fine-Tuned Models



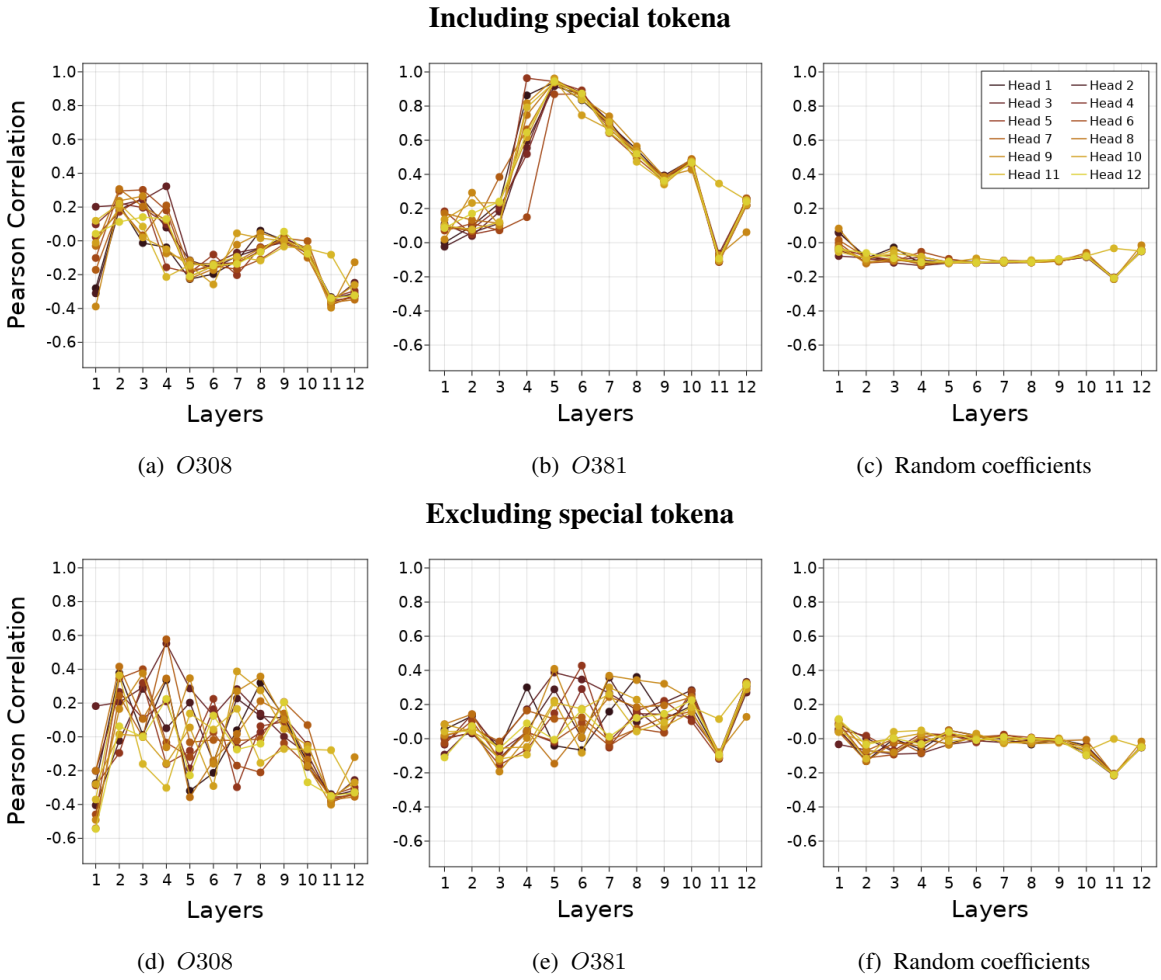
(a) With special tokens (b) Without special tokens

Figure 16: BERT-base (fine tuned on MNLI) encoding Wikitext-v2 validation set data: the correlation between magnitude of hidden state parameters corresponding to outlier dimensions, and frequency of encoded tokens in pre-training data.

To control how much fine-tuning affects the patterns we study, we repeat the experiments with models fine-tuned on MNLI, we proceed as follows: we fine-tune the model using a classification head and then extract the hidden states at each layer and use those in place of the ones of the pre-trained model.

Figure 16 shows the same information as Fig. 4, that is the correlation between the hidden states outlier dimension magnitude and the frequency of the encoded tokens in pre-training data, for a BERT-base model fine-tuned on MNLI. The overall patterns is similar to using the pre-trained model, but the correlation values generally decrease: the highest value is now 0.3 when it used to be 0.5 for the pre-trained model. This agrees with the findings from §4.4: the outliers are impacted by the model training.

Investigating attention patterns, Fig. 17 reports the same information as Fig. 7 for *bert-base-uncased* model fine-tuned on MNLI. In this case we see that the correlations stays high at layers closer to the input data, while those closer to the output have lower values, although in this case as well the values are higher than they are for random outliers Figs. 17c and 17f.



(a) $O308$ (b) $O381$ (c) Random coefficients

(d) $O308$ (e) $O381$ (f) Random coefficients

Figure 17: Each figure shows the correlation between the *average vertical attention values* in a BERT-base fine tuned on MNLI self-attention heads, and the magnitude of hidden state parameters at the dimensions corresponding to outlier dimensions. The correlation is computed over examples from Wikitext-v2. Figures (c) and (f) show the average over 10 random dimensions.