Anisotropy Is Inherent to Self-Attention in Transformers

Nathan Godey [1,2] **Éric de la Clergerie**[1] **Benoît Sagot**[1]
[1]Inria, Paris, France
[2]Sorbonne Université, Paris, France
{nathan.godey,eric.de_la_clergerie,benoit.sagot}@inria.fr

Abstract

The representation degeneration problem is a phenomenon that is widely observed among self-supervised learning methods based on Transformers. In NLP, it takes the form of *anisotropy*, a singular property of hidden representations which makes them unexpectedly close to each other in terms of angular distance (cosine-similarity). Some recent works tend to show that anisotropy is a consequence of optimizing the cross-entropy loss on long-tailed distributions of tokens. We show in this paper that anisotropy can also be observed empirically in language models with specific objectives that should not suffer directly from the same consequences. We also show that the anisotropy problem extends to Transformers trained on other modalities. Our observations suggest that anisotropy is actually inherent to Transformers-based models.

1 Introduction

In recent years, deep learning models based on Transformers have led to significant breakthroughs in the field of natural language processing (NLP). These models have demonstrated state-of-the-art performance across a range of tasks, such as language modeling, machine translation, and sentiment analysis. However, despite their successes, they suffer from a phenomenon known as the representation degeneration problem. Specifically, this degeneration is characterized by anisotropy, a property of hidden representations that makes them all close to each other in terms of angular distance (cosine-similarity).

Anisotropy has been widely observed among self-supervised models based on Transformers, and literature currently suggests that it may be a consequence of optimizing the cross-entropy loss on long-tailed distributions of tokens (Gao et al., 2019; Biś et al., 2021). However, it remains uncertain whether anisotropy is a fundamental property of Transformers-based models or a consequence of the pre-training process.

In this paper, we investigate the anisotropy problem in depth, and we make several contributions:

- We demonstrate empirically that anisotropy can be observed in language models with character-aware architectures that should not suffer directly from the same consequences as token-based models. We extend our observations to Transformers trained on other modalities, such as image and audio data, and show that anisotropy cannot be explained solely based on linguistic properties;
- We provide empirical observations on the anisotropic properties of the Transformer block by studying untrained layers, and establish a relation between anisotropy and the general sharpness of the self-attention mechanism;
- We conduct an analysis of the representations used in self-attention (queries and keys) along training and show that anisotropy appears intrinsically in the self-attention mechanism, when training pushes for sharp patterns.

2 Related Work

The general phenomenon of anisotropy in token-based Transformers for language models has been shown in Ethayarajh (2019). Figure 1 extends one of their experiment to more architectures. Gao et al. (2019) shows that the degeneration of representations comes from the distributions of subwords in natural language, namely the existence of unused and rare tokens that tend to push all representations away from the origin towards a specific direction.

Other works have established a connection between word frequency and distortions of the latent spaces (Yu et al., 2022; Puccetti et al., 2022; Rajaee and Pilehvar, 2022). Biś et al. (2021) have shown

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 35–48

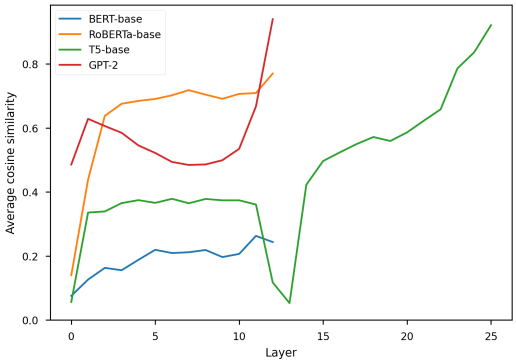


Figure 1: Average cosine-similarity between hidden representations across layers for token-level NLP models. For T5-base, we concatenate encoder and decoder results.

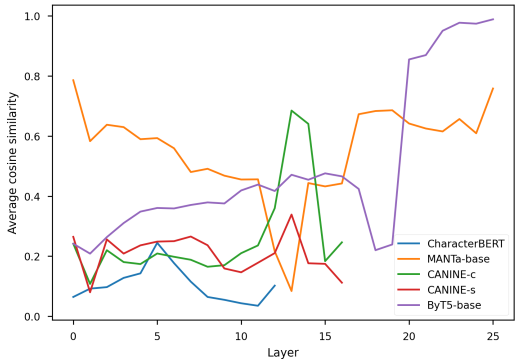


Figure 2: Average cosine-similarity between hidden representations across layers for character-level models.

that anisotropy in LMs could be explained by a global *drift* of the representations in the same direction, thus unifying conclusions from Ethayarajh (2019) and Gao et al. (2019). The authors propose that this drift is caused by the persistent updating of the representation of rare and unused tokens in a consistent direction, due to the nature of the softmax operation in the cross-entropy loss. They show that removing the average component to all representations leads to a nearly perfect isotropy.

Several methods have been proposed to reduce anisotropy in Transformers-based LMs at token-level (Rajaee and Pilehvar, 2021; Wang et al., 2020), or at sentence-level (Gao et al., 2021; Yan et al., 2021; Su et al., 2021). They usually consist in post-processing the representations, and lead to downstream performance boosts. We argue that these positive results are paving the way for the search of pre-training objectives that do not introduce anisotropy in the first place, in the hope that the resulting models will also perform better without any post-processing, and potentially be trained more efficiently. This motivates us to gain a deeper understanding of the underlying factors that induce anisotropy, whether they belong in data, architectures, or training procedures.

3 Anisotropy in pre-trained Transformers

3.1 Character-based NLP

To assert whether the cross-entropy objective applied on vocabularies containing rare tokens is the sole cause for the common drift issue, we explore anisotropy in character-based models. We study different architectures:

- CharacterBERT (El Boukkouri et al., 2020) is constructing whole word representations from character embeddings put through convolutions and highway layers, before feeding them to a Transformers architecture.
- CANINE (Clark et al., 2022) is downsampling contextualized character representations via a strided convolution before feeding them to a Transformers. It can be trained either with a subword-based objective (CANINE-s) or with a character-level one (CANINE-c).
- MANTa-LM (Godey et al., 2022) is based on a differentiable segmentation and embedding module added before an encoder-decoder model in the style of T5 (Raffel et al., 2020). It takes bytes as inputs and outputs, but builds internal representations that are usually based on several bytes.
- ByT5 (Xue et al., 2022) is a version of T5 that is trained at byte-level. To afford for more complex encoding, the authors resize the encoder-decoder architecture.

Neither of these architectures should suffer from out-of-vocabulary tokens in the process of creating representations. The models that predict at word or sub-word level (CharacterBERT and CANINE-s) could have the cross-entropy loss systematically pushing away rare item representations. However, it is rather unclear why it would imply an embedding drift at deeper layers. Hence, if anisotropy was only caused by the presence of unused or rare subwords, those character-level models should be much less prone to this issue.

To verify this hypothesis, we compute hidden representations for the validation set of the WikiText-103 corpus (Merity et al., 2016). We then

compute the average cosine-similarity between two representations, uniformly taken in the whole validation corpus.

In fact, as shown in Figure 2, those models all display significant levels of anisotropy in at least one of their layers. Interestingly, the models that are based solely on characters or bytes for input and prediction (ByT5, CANINE-c, and MANTA-LM) seem to display even higher levels of anisotropy. We note, as it is the case for the T5 model, that the ByT5 decoder displays extremely high levels of anisotropy.

3.2 Other modalities

We've shown in the previous section that character-level language models suffer from anisotropy similarly to token-level ones, hinting that subword token distributions are not solely responsible for anisotropy. However, it may be argued that anisotropy is related to linguistic properties. Thus, we proceed to explore the anisotropy problem for Transformers-based models in other modalities, specifically speech and vision.

For speech models, we consider wav2Vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and Whisper (Radford et al., 2022) with the Common Voice 11.0 dataset (Ardila et al., 2020). For vision models, we use ViT (Wu et al., 2020), BEiT (Bao et al., 2021), MiT (Xie et al., 2021), and DEiT (Touvron et al., 2021) on the ImageNet dataset (Russakovsky et al., 2015).

As in subsection 3.1, we infer hidden representations on the validation sets for each modality. We then uniformly sample pairs of vectors to get cosine-similarity values for every layer of every model. The averaged results are displayed in Figure 3.

Once again, almost every model shows a significant level of anisotropy on some of its layers. Notably, speech models seem to have very anisotropic representations, as every layer of every model outputs an average cosine-similarity of at least 0.2. We find some exceptions among vision models, since the MiT model seems to use isotropic representation spaces and the ViT model has a low average cosine-similarity for all its layers.

We also conduct the same experiment for convolution-based networks in the vision modality. The models at glance are ResNet (He et al., 2016), EfficientNet (Tan and Le, 2019), CvT (Wu et al., 2021), ConvNeXt (Liu et al., 2022), and VAN (Guo et al., 2022). For these networks, we flatten convolution maps to vectors before computing the cosine-similarity.

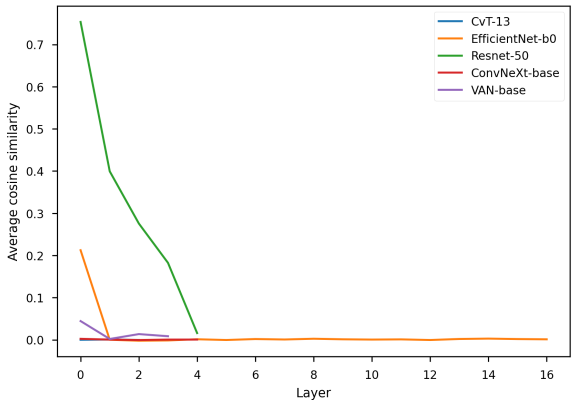


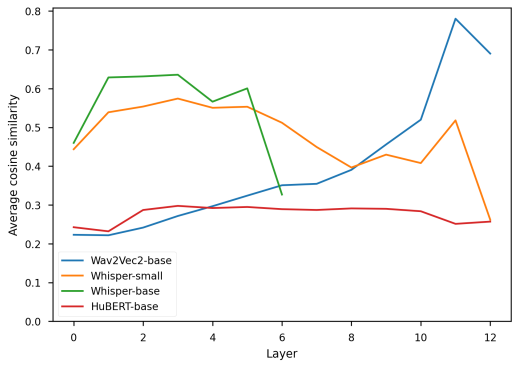
Figure 4: Average cosine-similarity between hidden representations across layers for convolution-based vision models.

We observe in Figure 4 that most of the convolution-based models are isotropic. Interestingly, the only exception is ResNet-50, whose representations become more and more isotropic as one explores deeper layers. This could partially be explained by the fact that the batch normalization (Ioffe and Szegedy, 2015) used in some of these models mitigates *a posteriori* the drift effect by removing the mean component of the representations. However, the ConvNeXt model also seems to use isotropic representations while not using batch normalization, which shows that this is not the only factor in the isotropic behavior of these models.

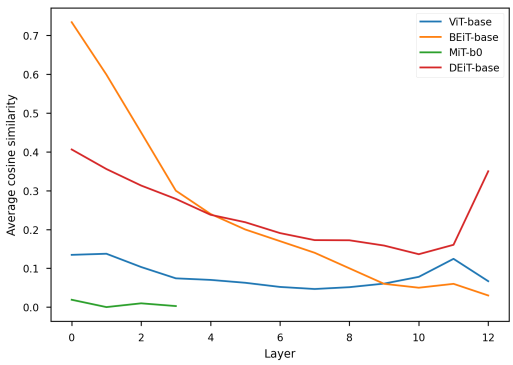
3.3 To drift or not to drift?

Related works (Biś et al., 2021; Gao et al., 2019) show that anisotropy in subword-level language models is caused by a drift of the hidden representations in a shared direction. In this section, we try to extend this observation to other modalities.

We study the correlation between the uniformly measured cosine-similarity, and the norm of the average hidden representation $||\bar{x}||_2$ for each layer. If anisotropy could be directly explained by the drift effect, we would expect a monotonic relation between $||\bar{x}||_2$ and the average cosine-similarity. To verify this, we apply a Spearman correlation test on these two metrics for every model from subsection 3.1 and subsection 3.2, along with some token-level language models, namely T5 (Raffel et al., 2020), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and GPT-2 (Radford et al., 2019).



(a) Speech



(b) Vision

Figure 3: Average cosine-similarity between hidden representations across layers for Speech and Vision modalities. We observe that across both modalities, several models display significant levels of anisotropy.

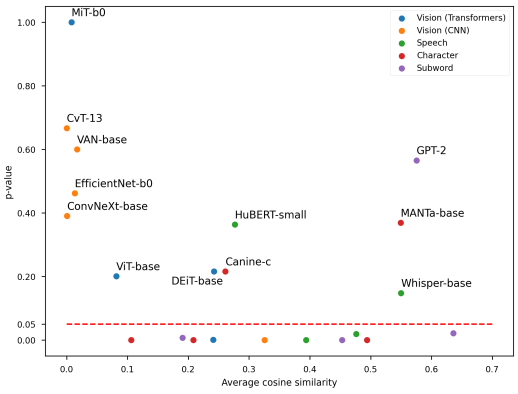


Figure 5: p-value of the Spearman correlation test between the norm of the average representation and the cosine-similarity averaged over all layers, across modalities. For models above the red dotted line, there is no significant ($p > 0.05$) correlation between the drift effect and the anisotropy level.

In Figure 5, we observe that we can correlate the anisotropy level and the magnitude of the drift component across layers for several models. The anisotropy of subword-based models can generally be correlated with the drift effect, except for GPT-2 for which the Spearman correlation metric may not be appropriate. We provide a similar analysis based on the Pearson correlation test and discuss the relevance of each statistic in Appendix A.

Interestingly, we notice that the anisotropy affecting most CNN-based vision models is generally not correlated with the drift effect, contrary to Tranformers-based models in the same modality. Some speech models (HuBERT and Whisper-base) also display signs of anisotropy that cannot be correlated with the drift effect. Figure 5 also shows a correlation for all character-based models but Canine-C and MANTa-base.

4 Exploring the representation drift

In this section, we focus on some intrinsic properties of the Transformer block in a modality-agnostic fashion, i.e. with minimal assumptions on the data distribution, and without training. We analyze experimentally the behavior of the untrained Transformer block T when a common bias term b is added to untrained input representations $\mathbf{x}$. This allows us to mimic the common drift as mentioned in Biś et al. (2021) and to identify some properties induced by this artificial drift on the output representations.

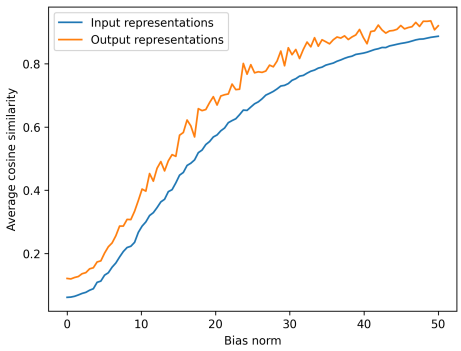
4.1 Experimental setup

We consider an embedding lookup table E and a Transformer block T with weights initialized as in BERT (Devlin et al., 2019). We then draw 16 input embedding sequences $\mathbf{x}$ of length 512 uniformly from E. To account for a drift component of norm $N \in \mathbb{R}$, we generate a vector $b_u \sim \mathcal{N}(0, I_d)$, which we normalize into $b = \frac{b_u}{||b_u||_2} \times N$. We finally compute $T(\mathbf{x}_i + b)$ for every sequence x_i, and study the resulting distributions.

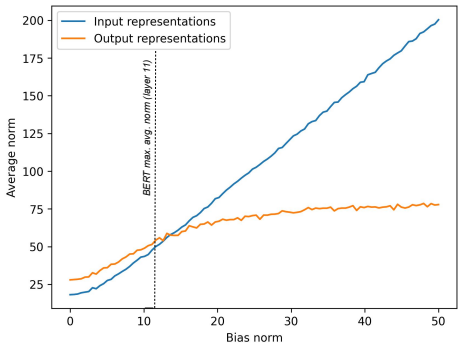
Specifically, we study the average norm of the input representations $\mathbb{E}(||\mathbf{x}_i + b||_2)$ against the average norm of the output representations $\mathbb{E}(||T(\mathbf{x}_i + b)||_2)$ in Figure 6b. We also retrieve the self-attention scores before the softmax operation, namely $\frac{QK^T}{\sqrt{d_k}}$, along with the corresponding Q and K matrices. We study some of their properties in Figure 7 and Figure 8.

4.2 Input vs. output analysis

In Figure 6a, we observe that the output representations have an average cosine-similarity value that



(a) Cosine similarity



(b) Norm

Figure 6: Input/Output comparison of a Transformer block from BERT-base as the bias norms increases.

is slightly higher than the one of the input representations, no matter the level of input bias. We also notice that while the norm of the average output representation increases with the bias norm, it seems to meet the corresponding input measure for a given bias norm.

Interestingly, this shows that there is a *fixed point* in terms of norm in the Transformers function with biased input. More formally, there seems to exist a bias norm $N^* \in \mathbb{R}_+$ such that:

$$\mathbb{E}_{x,b_{N^*}}(||x_i + b_{N^*}||) = \mathbb{E}_{x,b_{N^*}}(||T(x_i + b_{N^*})||)$$

Moreover, this fixed point level N^* is in the order of magnitude of the average hidden state norms of the layers of the trained BERT model. This hints that the model's representations stabilize when their norm is close to this fixed point. We leave a more thorough analysis of this hypothesis for future work.

4.3 Exploring the Transformer block

To understand the effect of the drift effect on the inner workings of the Transformer layer, we take a closer look at the self-attention operation as the average input representation drifts away.

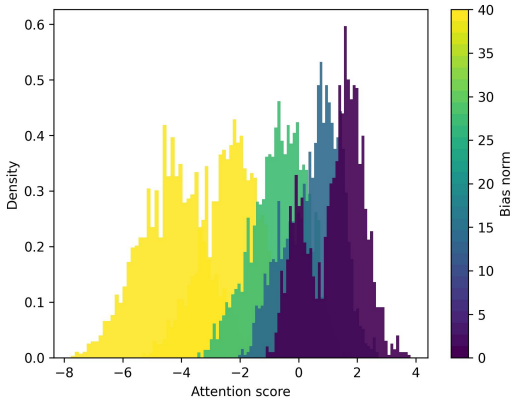
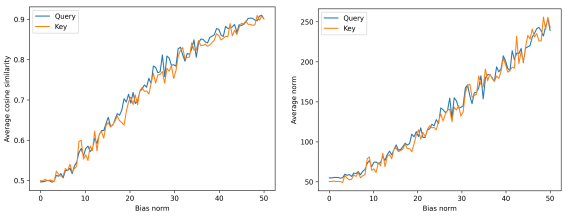


Figure 7: Histograms of the pre-softmax attention scores as the input bias norm increases. Other initializations of the layer and of the bias direction b_u led to a general *increase* of the attention scores instead.

Figure 7 shows that the attention scores tend to move away from zero as the input bias norm increases. Indeed, as the norm of the average $\bar{x}$ of the input embeddings increases, we can expect the query and key vectors Q and K to also display signs of anisotropy. Actually, for each self-attention head, and for all position $i \in [1, L]$, we have:

$$\begin{cases} \mathbb{E}_x(Q_i) = W_Q\bar{x} + b_Q \\ \mathbb{E}_x(K_i) = W_K\bar{x} + b_K \end{cases} \quad (1)$$

We can observe in Figure 8 that query and key representations indeed increase in norm with the input bias norm. We also notice that the corresponding distributions are anisotropic even when no bias is added, which may be a consequence of BERT's initialization parameters.



(a) Cosine sim. (b) Norm

Figure 8: Analysis of the self-attention query and key distributions

4.4 Impact of the drift

After exploring the consequences of the drift of input representations on the query-key product in self-attention, we identify in this section the implications of this drift at a more explainable level, by observing the resulting post-softmax distributions.

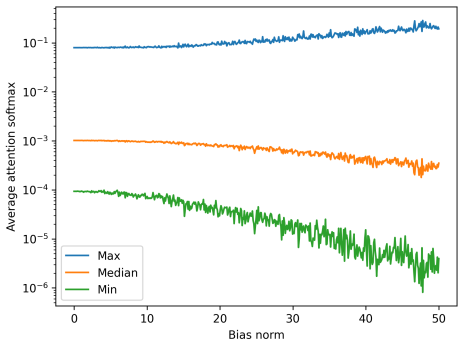
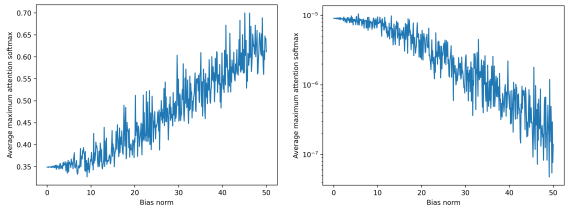


Figure 9: Evolution of the self-attention softmax values as the input bias norm increases.

In Figure 9, we retrieve softmax values in the self-attention block and for each position, we extract the maximum, the median and the minimum. We then average these values over the whole batch, and repeat for various input bias norm levels. We notice that as the input bias norm increases, the self-attention softmax distributions tend to become less entropic, evolving towards higher maximal probabilities and lower minimal probabilities. In the following analysis, we'll use the term *sharpness* to discuss entropy levels of the self-attention distributions.



(a) Maximum (b) Minimum

Figure 10: Comparison of the extreme values of each sequence averaged over the batch as the bias norm increases.

This sharpening effect of the attention distributions becomes even clearer if we consider the maximum and minimum values over the whole sequences, as in Figure 10.

However, at low anisotropy levels, i.e. when the bias norm is low, we see that the effect is not very important. Figure 9 and Figure 10 only hint at the fact that the drift of embeddings may help the self-attention to be sharper. Another explanation could be that training favors sharp self-attention patterns, as has been pointed out in previous works (Clark et al., 2019), which in turn induces a drift in the models' representations. In order to account for that, we need to study the evolution of latent spaces at the self-attention level along training.

5 Queries and keys: training dynamics

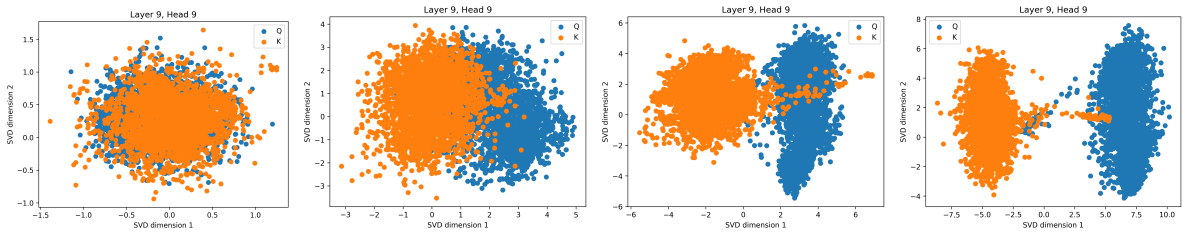
We have established that manually pushing for drift-based anisotropy on *untrained* Transformers models leads to sharper (i.e. low-entropy) self-attention patterns. In this section, we show that this evolution of self-attention values actually takes place during training, and we explore the mechanism behind their appearance. As pointed out in section 4, the self-attention scores result from the QK^T operation, which computes scalar products between query and key representations corresponding to each pair of positions. Thus, in this section, we study the evolution of these query and key representations *along training*, and explore the mechanism behind the increase of the scalar products leading to self-attention scores.

We use the MultiBERT checkpoints (Sellam et al., 2021) with seed 0 to retrieve Q and K distributions at different pretraining steps, and we use 128 samples from Wikitext-103 as input data. Along this section, Q_s and K_s refer to query and key representations extracted at a specific layer and head at a given step s, and $\hat{Q}_s$ and $\hat{K}_s$ are the average representations, taken over all tokens in the sampled batch. By studying $\bar{Q}_s$ and $\bar{K}_s$, we aim at exploring the common (or context-agnostic) drifts of keys and queries distributions.

In Figure 11 and Figure 12, we compute a SVD of the union of Q_s and K_s for all steps s, so that the projection makes sense for both distributions across steps for visualization purposes [1]. As shown in our selected examples, we observe that the dynamics of $\bar{Q}_s$ and $\bar{K}_s$ tend to align along training, making the average of the distributions drift in either similar or opposite directions. The first dimension of the SVD seems to describe this common drift. Note that in $\mathbb{R}^{d_h}$ ($d_h = 64$ being the head dimension), such an alignment is very unlikely to happen randomly. Interestingly, Figure 12a shows that the common direction dynamics appear in the first few steps, while the opposite direction dynamics of Figure 12b only starts after 8% of the total training steps.

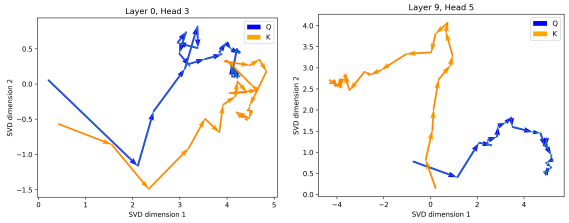
To consolidate our observations, we compute the evolution of the cosine-similarity between $\bar{Q}_s$ and $\bar{K}_s$ along training in Figure 13. We also display

[1] We actually uniformly sample 20% of the whole set of representations to compute the SVD under reasonable memory constraints.



(a) Step 0 (b) Step 40k (c) Step 200k (d) Step 2M (final)

Figure 11: Evolution of Q_s and K_s distributions along training. Vectors are projected using a common SVD.



(a) Similar (b) Opposite

Figure 12: Evolution of $\bar{Q}_s$ and $\bar{K}_s$ along training for two different heads in the network, projected via common SVD. Each arrow represents a checkpoint in the MultiBERT suite. We display typical examples of dynamics in same/opposite direction.

some projected Q_s and K_s distributions for several s steps in Figure 11.

Figure 13 shows that the first layers display a common direction dynamic, as the cosine-similarity tends to increase, thus showing that **the key and query distributions drift along a similar direction** in average. The last layers seem to adopt an opposite direction dynamic, as the cosine-similarity between their mean key and query representations gets negative along training.

As shown in Figure 14, this drift induces an increase in the magnitude of scalar products obtained in the self-attention QK^T operation, thus facilitating the emergence of sharp patterns where attention focuses on specific tokens.

Finally, Figure 15 describes the evolution of the average entropy in self-attention distributions. We observe that training induces an overall decay of the entropy for all layers, with different dynamics. This corresponds to sharper self-attention distributions. It is interesting to notice that the distributions in the first layers remain sharper than the ones in the last layers.

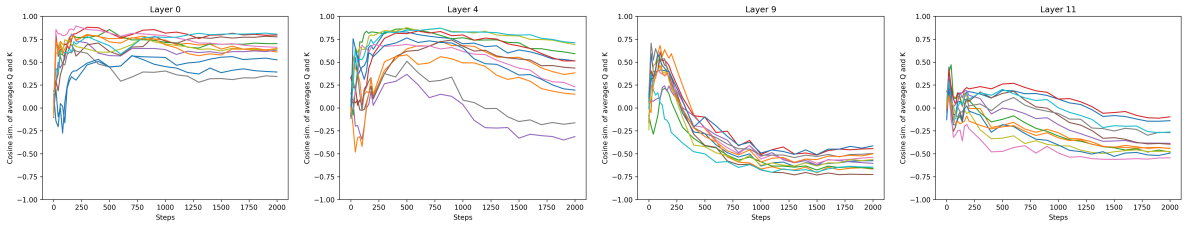
Overall, this section shows that drift anisotropy emerges in the query and key representations during the training of MultiBERT, as self-attention distributions become sharper. The drifts of queries and keys tend to align, thus increasing the magnitude of scalar products, and the general sharpness of self-attention.

Although this section focuses on the case of token-based NLP, we believe that strong attention patterns may be required when training Transformers across all modalities, potentially generating distortions in query and key distributions that account for the final observed anisotropy of the models. However, we could not extend experiments to other modalities due to the lack of released intermediate checkpoints, to the best of our knowledge.

6 Discussion

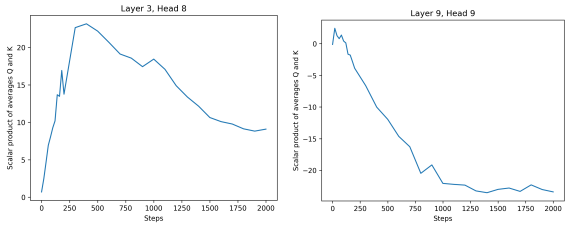
In this work, we argue that the nature of data distributions is not solely responsible for the anisotropy observed in most hidden representations of Transformers-based models across modalities. As section 4 shows, untrained Transformers layers display a tendency towards anisotropy. Biased inputs tend to increase the variance of the attention scores and thus facilitate the emergence of sharp patterns in the self-attention mechanisms. We also show in section 5 that along training, query and key distributions drift in parallel directions, which increases anisotropy in the inner representations of the Transformer layers, while allowing sharper attention patterns. As discussed in Puccetti et al. (2022), outlier dimensions in Transformers are also involved in the emergence of strong attention patterns.

Consistency of the SVD In section 5, we use an SVD on the *union* of Q_s and K_s for visualization purposes (see Figure 11 and Figure 12). It may be argued that this approach favors the emergence of a discriminative singular direction, that helps distinguish between keys and queries, thus supporting the findings in a less convincing way. To address this concern, we display alternative projections in Appendix C, where we compute the SVD on Q_s or



(a) Layer 0 (b) Layer 4 (c) Layer 9 (d) Layer 11

Figure 13: Evolution of cosine-similarity between $\bar{Q}_s$ and $\bar{K}_s$ along training. Each color represents one self-attention head. Steps are counted in thousands. We generally observe that almost all heads see $\bar{Q}_s$ and $\bar{K}_s$ align in common or opposite directions along training. In other words, the average components of keys and queries representations tend to align in self-attention heads, which maximizes the magnitude of the scalar product between two average representations. We run a similar experiment on all MultiBERT seeds in Figure 23, and obtain comparable results.



(a) Similar (b) Opposite

Figure 14: Evolution of the scalar product between $\bar{Q}_s$ and $\bar{K}_s$ along training. Steps are in thousands.

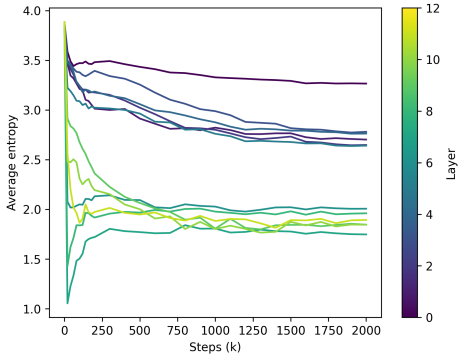


Figure 15: Average entropy of the probability distributions corresponding to self-attention rows along training. Each curve corresponds to one layer.

K_s only, and then project all representations using this SVD. Our observations show that our findings are consistent for these alternative projections.

Harmfulness of anisotropy Even though anisotropy has not been shown to be an issue in language modeling, previous works have advocated that removing anisotropy in output representations leads to better sense disambiguation abilities (Bihani and Rayz, 2021; Biś et al., 2021). Isotropic models could also improve cross-lingual alignment in multilingual language models (Hämmerl et al., 2023). Nevertheless, concurrent works have suggested that anisotropy may not hurt the quality of the representations (Ait-Saada and Nadif, 2023; Rudman and Eickhoff, 2023). We argue that anisotropy in the Transformer architecture may actually help models by allowing sharp attention patterns, but we also believe that our work can pave the way for new isotropic architectures that can easily use sharp attention patterns.

Conclusion

In this paper, we investigated the anisotropy problem through the lens of the drift effect, and made several contributions to the understanding of this phenomenon. We demonstrated that anisotropy can be observed in language models with character-aware architectures, extended our observations to Transformers trained on other modalities, and studied anisotropy in untrained Transformers layers. We finally explored the training dynamics of the query and key distributions, and found that they drift along a shared direction hence maximizing QK^T scalar products in absolute value, allowing stronger attention patterns as a result.

We conclude that anisotropy almost systematically affects Transformers on all modalities, in a way that is not always correlated with the drift of the representations. We also provide empirical evidence that anisotropy appears as an inherent property of latent distributions used in the self-attention mechanism when modeling sharp attention patterns. We hypothesize that a revision of the self-attention operation could help reduce anisotropy by facilitating the emergence of sharp attention softmax distributions without distorting the geometry of the hidden representations.

Limitations

As mentioned in the Discussion section, we acknowledge that section 4 does not take into account the training dynamics, and only exposes some properties of the Transformer layer at initialization. We also notice that the Spearman correlation test used in Figure 5 may not be well-suited for such noisy observations, as the high p-value of the GPT-2 model shows. We provide a similar graph based on the Pearson correlation in Appendix A.

Moreover, we are aware that our approach is not theoretically rigorous in some aspects. For instance, we don't prove that sharp self-attention patterns *cannot* emerge without anisotropy in keys and queries representations. In other words, this article is focusing on exposing and *correlating* factors that explain anisotropy, but we do not demonstrate theoretical properties that would help identify the *causes* of anisotropy. Nevertheless, we believe that our work can pave the way for such theoretical exploration in the future.

Ethics Statement

To the best of our knowledge, our work does not raise any ethical concern. However, as noted in Zhou et al. (2021), we believe that distortions in the embedding space may be related to bias in the training data, whether it is inherent to the structure of the modality (e.g. the Zipfian distribution of words), or due to human factors (e.g. geographical considerations).

Acknowledgements

This work was funded by the last authors' chair in the PRAIRIE institute funded by the French national agency ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001. This work was granted access to the HPC resources of IDRIS under the allocation 2023-AD011013680R1 made by GENCI.

We would like to thank Roman Castagné for useful discussions that led to focusing on observing the effect of anisotropy in the self-attention process.

References

Mira Ait-Saada and Mohamed Nadif. 2023. Is anisotropy truly harmful? a case study on text clustering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1194–1203, Toronto, Canada. Association for Computational Linguistics.

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

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.

Hangbo Bao, Li Dong, and Furu Wei. 2021. Beit: BERT pre-training of image transformers. *CoRR*, abs/2106.08254.

Geetanjali Bihani and Julia Rayz. 2021. Low anisotropy sense retrofitting (LASeR) : Towards isotropic and sense enriched representations. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 81–95, Online. Association for Computational Linguistics.

Daniel Biś, Maksim Podkorytov, and Xiuwen Liu. 2021. Too much in common: Shifting of embeddings in transformer language models and its implications. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5117–5130, Online. Association for Computational Linguistics.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

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

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

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Jun'ichi Tsujii. 2020. CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6903–6915, Barcelona, Spain (Online).

International Committee on Computational Linguistics.

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.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nathan Godey, Roman Castagné, Éric de la Clergerie, and Benoît Sagot. 2022. MANTa: Efficient gradient-based tokenization for end-to-end robust language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2859–2870, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. 2022. Visual attention network. *arXiv preprint arXiv:2202.09741*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Katharina Hämmerl, Alina Fastowski, Jindřich Libovický, and Alexander Fraser. 2023. Exploring anisotropy and outliers in multilingual language models for cross-lingual semantic sentence similarity.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France. PMLR.

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

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Giovanni Puccetti, Anna Rogers, Aleksandr Drozd, and Felice Dell'Orletta. 2022. Outlier dimensions that disrupt transformers are driven by frequency. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1286–1304, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sara Rajaee and Mohammad Taher Pilehvar. 2021. A cluster-based approach for improving isotropy in contextual embedding space. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 575–584, Online. Association for Computational Linguistics.

Sara Rajaee and Mohammad Taher Pilehvar. 2022. An isotropy analysis in the multilingual BERT embedding space. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1309–1316, Dublin, Ireland. Association for Computational Linguistics.

William Rudman and Carsten Eickhoff. 2023. Stable anisotropic regularization.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Thibault Sellam, Steve Yadlowsky, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, Dipanjan Das, Ian Tenney, and Ellie Pavlick. 2021. The multiberts: Bert reproductions for robustness analysis. *arXiv preprint arXiv:2106.16163*.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Mingxing Tan and Quoc V. Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. *ArXiv*, abs/1905.11946.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. 2021. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Lingxiao Wang, Jing Huang, Kevin Huang, Ziniu Hu, Guangtao Wang, and Quanquan Gu. 2020. Improving neural language generation with spectrum control. In *International Conference on Learning Representations*.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Masayoshi Tomizuka, Kurt Keutzer, and Péter Vajda. 2020. Visual transformers: Token-based image representation and processing for computer vision. *ArXiv*, abs/2006.03677.

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. 2021. Cvt: Introducing convolutions to vision transformers. *arXiv preprint arXiv:2103.15808*.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. 2021. Segformer: Simple and efficient design for semantic segmentation with transformers. *CoRR*, abs/2105.15203.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. 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 5065–5075, Online. Association for Computational Linguistics.

Sangwon Yu, Jongyoon Song, Heeseung Kim, Seongmin Lee, Woo-Jong Ryu, and Sungroh Yoon. 2022. Rare tokens degenerate all tokens: Improving neural text generation via adaptive gradient gating for rare token embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29–45, Dublin, Ireland. Association for Computational Linguistics.

Kaitlyn Zhou, Kawin Ethayarajh, and Dan Jurafsky. 2021. Frequency-based distortions in contextualized word embeddings. *CoRR*, abs/2104.08465.

A Pearson correlation of the drift norm and anisotropy

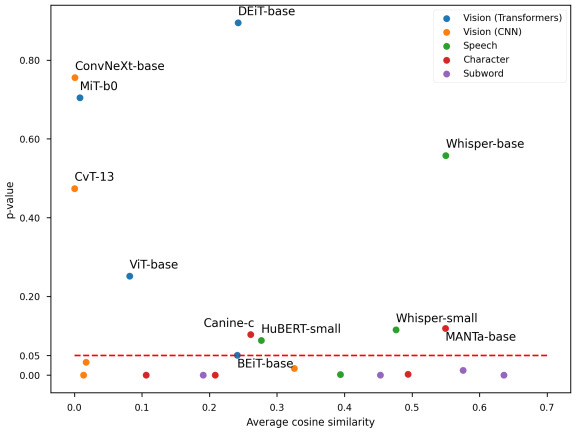


Figure 16: p-value of the Pearson correlation test between the norm of the average representation and the cosine-similarity averaged over all layers, across modalities. Models above the red dotted line are not significantly affected by the drift effect.

The Pearson test measures a linear correlation between random variables, while the Spearman test measures a monotonic correlation. As there is no specific argument in favor of a linear relationship between the measured distributions (average cosine-similarity and norm of the average representation), we decided to use the Spearman correlation test in order to take into account more complex relation patterns.

Nevertheless, this metric is based on the rank of each observation, and is thus not robust to fluctuations due to sample variance, specifically when working with such small samples. This is reflected by the discrepancy between Pearson and Spearman p-values for some models (e.g. GPT-2).

B Cosine-similarity and anisotropy

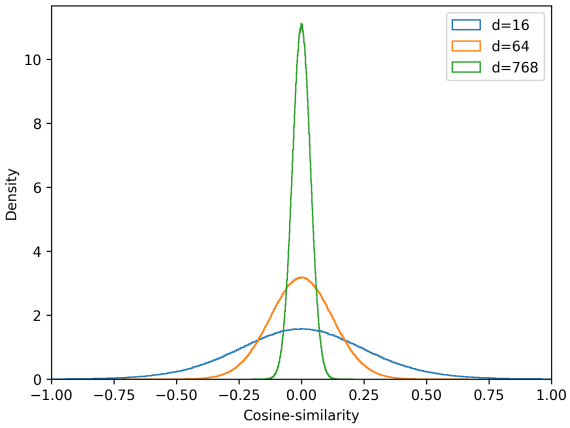


Figure 17: Density function of cosine-similarity for a normal distribution as the dimension increases.

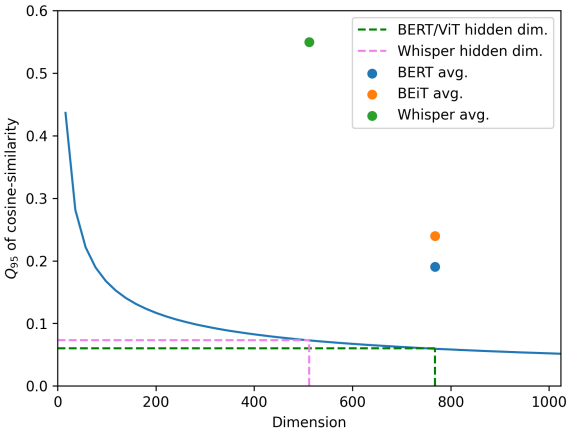


Figure 18: 95th quartile of the cosine-similarity distribution on a normal distribution as the dimension increases. We add points for the average cosine-similarity level of Transformers models for several modalities.

It can be argued that describing anisotropy as the observation of "high" cosine-similarity values is not a convincing definition. This section aims at showing which ranges of cosine-similarity values are characteristic of anisotropic distributions. In Figure 17, we show the density function of the cosine-similarity values obtained when drawing pairs of samples from isotropic normal distributions in $\mathbb{R}^d$ as d increases.

For smaller dimensions ($d = 16$), we see that the range of cosine-similarity values that are attained between isotropic distributions is relatively broad compared to the possible spectrum ($[-1, 1]$). As d increases, the support of the observed distributions seems to become smaller, due to the curse of dimensionality.

We analyze this effect more in-depth in Figure 18, where we plot the 95th quantile of the cosine-similarity distribution in the isotropic scenario. We also add values for the layer-wise average cosine-similarity levels of typical models in several modalities for comparison. We can clearly observe that the levels of cosine-similarity observed in the representations of Transformers-based models are significantly unlikely to be observed in between samples drawn in isotropic normal distributions.

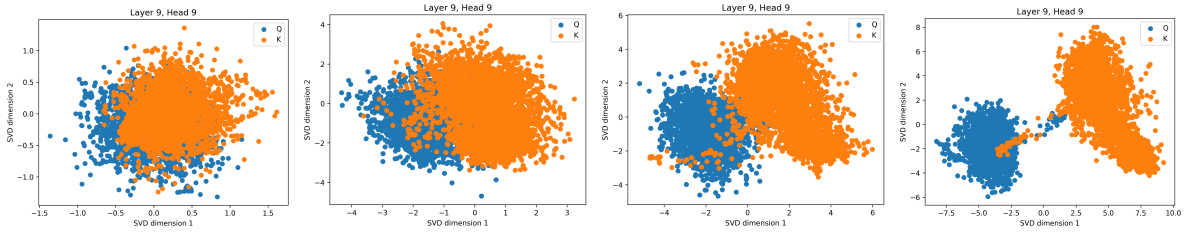
Nevertheless, as we go towards higher dimensional spaces for bigger models (e.g. Llama-65B from Touvron et al. (2023) has 8192 hidden dimensions), we believe that it may be relevant to introduce isotropy metrics that are grounded to isotropic cosine-similarity distributions. We leave this question for future works.

C Other projections for Q_s and K_s

As mentioned in the Discussion (section 6), we reproduce visualizations from section 5 using different projection choices. Namely, we compute the SVD on K_s only in Figure 19 and Figure 21, and on Q_s only in Figure 20 and Figure 22.

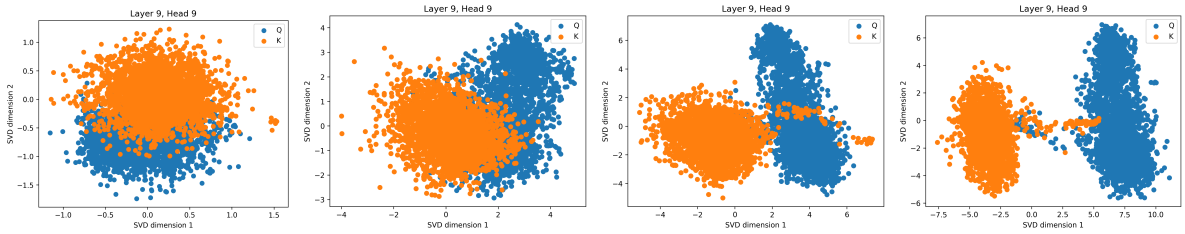
The plots show that not only does the distribution used for the SVD drifts away from the origin along training, but also that the other distribution drifts away from the origin in an opposite direction. In other words, the singular components of each distribution are also relevant to describe the drift of the other distribution. Hence, Figure 19 and Figure 20 support our conclusion that the drift directions of keys and queries are aligned during training.

D Stability across MultiBERT seeds



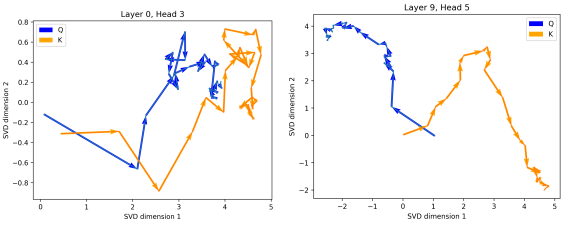
(a) Step 0 (b) Step 40k (c) Step 200k (d) Step 2M (final)

Figure 19: Evolution of Q_s and K_s distributions along training. Vectors are projected using the SVD computed on K_s.



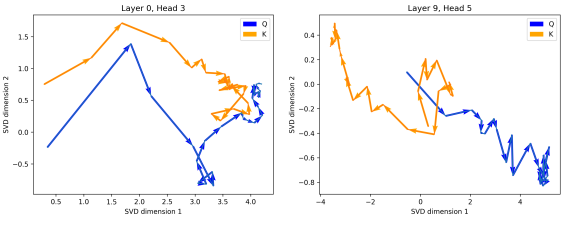
(a) Step 0 (b) Step 40k (c) Step 200k (d) Step 2M (final)

Figure 20: Evolution of Q_s and K_s distributions along training. Vectors are projected using the SVD computed on Q_s.



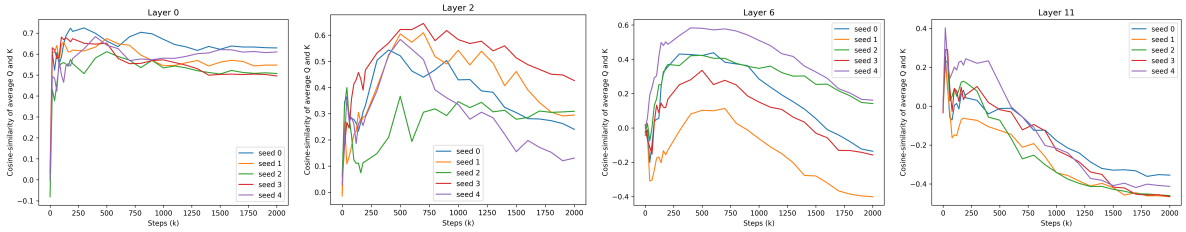
(a) Similar (b) Opposite

Figure 21: Evolution of $\bar{Q}_s$ and $\bar{K}_s$ along training for two different heads in the network, projected via the SVD of K_s.



(a) Similar (b) Opposite

Figure 22: Evolution of $\bar{Q}_s$ and $\bar{K}_s$ along training for two different heads in the network, projected via the SVD of Q_s.



(a) Layer 0 (b) Layer 2 (c) Layer 6 (d) Layer 11

Figure 23: Evolution of cosine-similarity between $\bar{Q}_s$ and $\bar{K}_s$ along training for various initialization seeds. Representations are concatenated across heads, and each color represents one seed of the MultiBERT models. We observe similar trends across seeds.