



Shrink the Longest: Improving Latent Space Isotropy with Simplicial Geometry

Sergej Kudrjashov , Olesya Karpik , and Eduard Klyshinsky ^(✉) 

Keldysh Institute of Applied Mathematics, National Research University “Higher School of Economics”, Moscow, Russia
eklyshisky@hse.ru

Abstract. Although transformer-based models have been dominating the field of deep learning, various studies of their embedding space have shown that they suffer from “representation degeneration problem”—embeddings tend to be distributed in a narrow cone, making the latent space highly anisotropic. Increasing the isotropy has shown to improve performance in downstream tasks both in static and contextual language models. However, most of approaches either add inference overhead or require substantial amount of data for model reparametrization. We propose a novel regularization technique based on simplicial geometry to improve the isotropy of latent representations. The core idea of our method is based on maximizing the persistent entropy of barcodes obtained using Vietoris-Rips filtration from contextual embeddings in the underlying latent space. We demonstrate that the method leads to an increase in downstream performance while significantly lowering the anisotropy during fine-tuning by exploiting existing geometric structures instead of reparametrization. The code is available at github.com/Xenomirant/Shrink-the-longest.

Keywords: BERT · Latent space geometry · Topological data analysis · Isotropy

1 Introduction

Since the first introduction in Vaswani et al. [20], transformers have become a dominant SotA architecture in a variety of fields and applications. However, their impressive effectiveness and black-box nature made researchers wonder about their internal structure and seek reasons for their effectiveness and problems inherent to the class of models. Ethayarajh in [6] notes that contextual word representations in ELMO, BERT and GPT-2 are highly anisotropic. Moreover, the extent of anisotropy grows with the number of layers [17]. In other words, embeddings are not uniformly distributed in latent space, but clustered in a narrow cone, which limits the informational expressiveness of the space, making it impractical to be considered euclidean anymore. As [7] point out, it can be partly attributed to static embeddings being tied during pretraining [16], which pushes the frequent tokens in opposite direction to all the others during pretraining.

This is empirically justified by Liang et al. [13] finding high correlation between embedding $L2$ norm and frequency in the pretraining corpora. However, it is far from explaining the whole story. Kawin Ethayarajh shows that less than 5% of the variance in the word’s contextual representations could be explained by the geometry of static embeddings [6].

Despite the continuing debate on sources of anisotropy, preserving isotropy has shown to be a favorable property both theoretically and practically [9, 13, 17, 21] – not only in semantic similarity tasks, which seem to benefit the most from richer informational capacity of contextualized embeddings, but in terms of model convergence as well. It also helps to mitigate the “stolen probability effect” [3] – a consequence of low embedding norm combined with overall positive cosine distance between embeddings, which results in a failure to attribute high probability to certain token embeddings in any context.

Further investigation by Cai et al. [2] has shown that while being globally anisotropic, the latent consists of various clusters and manifolds that demonstrate significantly larger isotropy within themselves. This line of research was continued in [17] finding that most clusters are biased towards structural information, and e.g. verb tense representations are distributed in different subspaces. Thus, the clustering of the space is expressive and must be preserved, while we seek for methods to improve overall isotropy.

Our contribution is the following. We address the problem by designing an additional regularization loss function inspired by recent results in computational topology [11] that preserves the representational geometry of the latent space, while improving the isotropy by maximizing entropy of distances between clusters. Thus we can optimize for isotropy while keeping the internal clustering structure intact, which differs from most of previously proposed methods. Our method is to our knowledge the first one to leverage the differentiability of persistent barcodes to mitigate the anisotropy of the latent space. It should be noted that although we utilize it for pretrained language models, the method is in fact model-agnostic, and can be applied to any vector-based representational model.

The structure of the paper is as follows. In Sect. 2 we review the recent methods proposed for anisotropy mitigation. In Sect. 3 describe the method and give a brief overview of results in computational topology we took inspiration from. Section 4 presents an experimental setup, while Sect. 5 is dedicated to the results we have obtained. Section 6 concludes the paper.

2 Related Work

To define and compute anisotropy we leverage the approach from Razzhigaev et al. in [18] and define it by exploiting a contextual embeddings’ matrix decomposition.

Let $X \in \mathbb{R}^{N \times D}$ represent the matrix of embeddings, where $\sigma_1, \dots, \sigma_k$ are its singular values, N is the number of samples and D is the hidden size. Then the k -th anisotropy score of X is given by:

$$\text{anisotropy}_k(X) = \frac{\sigma_k^2}{\sum_{i=1}^{\min(N,D)} \sigma_i^2}$$

Latest research has shown that anisotropy is a common trait in transformer-based architectures, where the overall contribution is split between attention [8], LayerNorm [7, 23] and various other factors.

There have been multiple methods proposed recently to address these peculiarities of transformers’ latent space. Subsequent works of Gao et al. [7] and Zhang et al. [25] introduce explicit regularization of cosine distance between normalized static embeddings. However not only its computation is dependent on the vocabulary size, which appears to be costly for each gradient update, but it can hardly be directly reapplied to subsequent layers making it irrelevant to account for anisotropy induced by each of further layers.

Another line of research was taken by Wang et al. [21]. They propose to reparametrize the initial embeddings matrix by a slow-decaying prior distribution over its singular values during training, resulting in substantially smoother decrease in isotropy dimension-wise. There have been also various attempts to adopt generative approaches for the task. Li et al. [12] utilized normalizing flows to map embeddings in the transformer’s latent space to an isotropic Gaussian. Zhang et al. [25] explore the effect of using VAE with isotropic Gaussian posterior noting that it significantly improves the classification performance and robustness to input perturbations.

However, despite their utility in developing higher isotropy, all these approaches seem to have a flaw that prevents them from being fully utilized during fine-tuning of already pretrained models. They reparametrize the whole internal space of the model, completely changing its geometry, which can result in suboptimal results when the number of supervised data is small. This view is supported by results found in Ding et al. [5]. The authors evaluate aforementioned isotropy calibration methods on various GLUE tasks finding that none of them produce consistent improvements over the uncalibrated models across tasks, domains and architectures. They attribute it to the effectiveness of representations already found in local subspaces, which are utilized during fine-tuning.

3 Method

Topological data analysis has recently found lots of use cases deep learning both due to its ability to represent global features, characteristic of the dataset, and due to recent results in differentiability of persistent homology [11], which allows to construct topology aware representations [15, 19]. We apply these results to construct our own regularization loss based on barcodes obtained from Vietoris-Rips complex of the embeddings during training.

Definition. Suppose X is a metric space and $r > 0$ is a real number. The Vietoris-Rips complex at scale r is defined as having X as the set of vertices (0-

dimensional simplices) and a k -simplex for each set of k vertices having diameter less than r , thus making it equivalent to clique complex.

By varying r we obtain a filtration over the simplicial complex, which provides us with additional information about the topological structure of the metric space. This gives us a powerful tool to explore the structure of high dimensional data without performing dimensionality reduction.

Definition. Persistent barcode of dimension k is a multiset of intervals that mark “births” x_i and “deaths” y_i of k -dimensional holes detected by the k -th homology group as we vary the scale r .

As we are primarily interested in local clustering structure of the space, we only take 0-dimensional homology into account reducing the simplex to the construction of the spanning tree, where individual barcodes represent the edge lengths of the graph. The number of barcodes is obviously dependent on the number of edges and equals $N - 1$, where N is the number of vertices.

However, in general, not all the barcodes encode meaningful topological features, which motivates researchers to seek for feature selection methods. We advance from recent algorithm proposed by Atienza et al. in [1] to define a functional that will serve us a dual role: as an optimization objective and as a tool for feature selection.

Definition. Given a filtration $F = \{K(t) | t \in \mathbb{R}\}$ and the corresponding diagram $dgm(F) = \{a_i = (x_i, y_i) | 1 \leq i \leq n\}$, let $L = \{l_i = y_i - x_i | 1 \leq i \leq n\}$. The persistent entropy $E(F)$ of F is defined as an entropy of the distribution on simplex faces:

$$E(F) = - \sum_{i=1}^n p_i \log(p_i), \text{ where } p_i \text{ is the normalized bar length,}$$

$$S_L = \sum_{i=1}^n l_i, \quad l_i = y_i - x_i, \quad p_i = \frac{l_i}{S_L}$$

The maximum persistent entropy corresponds to the situation when all the bars are of equal length. On the contrary, the value decreases as the distribution becomes more diverse. Note, that this function is dependent on the number of vertices and lies in the interval $[0, \log(n)]$.

This lets us set a base to describe the algorithm for prominent topological feature selection. For the empirical distribution of bars obtained from a batch let us suppose that the longest one T and the tiniest one r represent some features essential to the observed distribution. T is always considered a feature, while r is considered a noise. This assumption is crucial for the algorithm.

Let us briefly go through the major ideas of the algorithm. As we fix the edge cases, we are interested in filtering a subset of $n - 2$ features as irrelevant noise. Thus, on each internal iteration, the entropy of the barcode obtained by replacing the first i bars of L'_i by i bars that maximize the entropy of empirical distribution. Then, the idea is to iteratively apply the procedure until the estimated bar length

Algorithm 1. Separating topological features from noise (Atienza et al., 2017)

Require: A persistent barcode $B(VR_V) = \{(x_i, y_i) | 1 \leq i \leq n\}$ with bars of finite length

$$L = \{l_i := y_i - x_i\}$$

$$r \leftarrow \min(L), T \leftarrow \max(L), \alpha = \frac{r}{T}$$

$$L \leftarrow \text{sort}(L) = \{l_1, \dots, l_n\} \text{ s.t. } (l_n = T \geq l_1) \ \& \ (l_i \geq l_j \geq l_{n-1} = r, 1 \leq i < j < n-1)$$

 \triangleright Sort in decreasing order except for the last feature

$$L'_0 \leftarrow L, n' \leftarrow n$$

procedure SELECT FEATURES(L'_0, n'):

for $i = 1$ **to** $i = n' - 2$ **do**

$$P'_i = \sum_{j=i+1}^n l_j, R'_i = \{l_{i+1}, \dots, l_{n'}\}, l'_j = \frac{P'_i}{e^{E(R'_i)}}$$

$$L'_i = \{l'_1, \dots, l'_i, l_{i+1}, \dots, l_{n'}\}$$

$$C = \frac{S_{L'_{i-1}}}{S_{L'_i}} = \frac{P'_{i-1} + (i-1) \times l'_{i-1}}{P'_i + i \times l'_i}$$

$$Q = \left\lceil \frac{\alpha n' (\alpha - 1 - \log(\alpha))}{(\alpha - 1)^2} \right\rceil$$

if $C \geq 1$ **then**
Break
end if
if $Q \leq i$ **then**

$$L'_0 = L'_0 \setminus \{l_{i+1}, \dots, l_{n'-2}\}, n' = i + 2$$

SELECT FEATURES(L'_0, n')

else
return $B(VR_V) = \{T, l_1, \dots, l_i\}$ \triangleright the set of topological features

end if
end for
end procedure
return SELECT FEATURES(L'_0, n')

l'_i leads to decrease in its length, thus increasing the probability of the longest bar T . However, if a neutralization of the bar l_i leads to increase in its neutralized estimate l'_i , then $p_n^{(i-1)} > p_n^{(i)}$, and the features from l_{i+1} to $l_{n'-2}$ are regarded as noise. At the same time, Q measures the maximum number of topological features for the observed distribution and serves as a boundary for the selection process. For much more rigorous derivation see [1].

As the selected features generally are defined as the empirical distribution of distances between selected embeddings obtained after a hidden layer, we can use it to incorporate persistent entropy of this distribution as the regularization term for the loss function.

Thus, our regularization loss is defined as follows:

$$\mathcal{L}_{ent} = - \sum_{i=1}^{N \mathbb{1}[i \in L']} \frac{l_i}{S_{L'}} \log\left(\frac{l_i}{S_{L'}}\right), L' = \{T, \dots, l_i\}$$

As the term is bounded, the overall minimization problem keeps well-defined.

4 Experiments

In order to evaluate our reasoning we test two hypothesis:

1. An addition of regularization term improves isotropy of the latent space.
2. Feature selection step is essential to the effectiveness of the approach, as it exploits the underlying structure encoded in representations.

We test them by fine-tuning bert-base-uncased¹ [4] and roberta-base² [24] on MRPC and COLA datasets from the GLUE benchmark [21] with a batch size set to 64. We collect the contextual embeddings for the regularization loss into the matrix of size $\mathbb{N} \times \mathbb{D}$ by taking the [CLS] token from the last hidden layer. Anisotropy values are tracked along the training process. We also track the anisotropy of the centered embeddings, which corresponds to the eigenvalues of the covariance matrix.

We use Adam optimizer [10] with linear warmup for 10% training steps. The exact parameters used for fine-tuning are shown in Table 1. We repeat each experiment 5 times and report the average together with the standard deviation of calculated metrics over the runs.

Table 1. Fine-tuning parameters for classification tasks

Task	Learning rate	Weight decay	Epochs
BERT-MRPC	8e-5	8e-6	10
RoBERTA-MRPC	8e-5	8e-6	10
BERT-COLA	5e-5	5e-6	10
RoBERTa-COLA	5e-5	5e-6	10

As the end task is classification, instead of applying \mathcal{L}_{ent} to the whole space, we split embeddings based on classification label, and apply the loss to them separately, thus aiming to improve isotropy group-wise. Applying it to the whole space seems more appropriate during pretraining, but not in classification task, as it will push distant clusters towards each other instead of keeping them distinct. As we want to increase the entropy of the obtained distribution, we take it with a minus sign.

The resulting objective is defined as follows:

$$\mathcal{L} = \mathcal{L}_{CE} - \sum_{i=1}^N \mathcal{L}_{ent} \mathbb{1}_{[\text{class}=i]}$$

5 Results

We observe that the anisotropy of the latent space without regularization not only persists, but raises during fine-tuning (See Fig. 1). This can be partly

¹ <https://huggingface.co/google-bert/bert-base-uncased>.

² <https://huggingface.co/FacebookAI/roberta-base>.

attributed to the emergence of certain direction, which is aligned with the normal of the hyperplane separating the classes. However, our regularization loss reduces the extent of anisotropy and serves as an obstacle that prevents embeddings from collapsing onto several few directions.

This not only leads to isotropy improvements (see. Table 3 but to slight growth in generalization as well (see Fig. 2, Table 2). The values reported in the tables are obtained by averaging over last 30% of training, when the models have already converged, over 5 different independent runs.

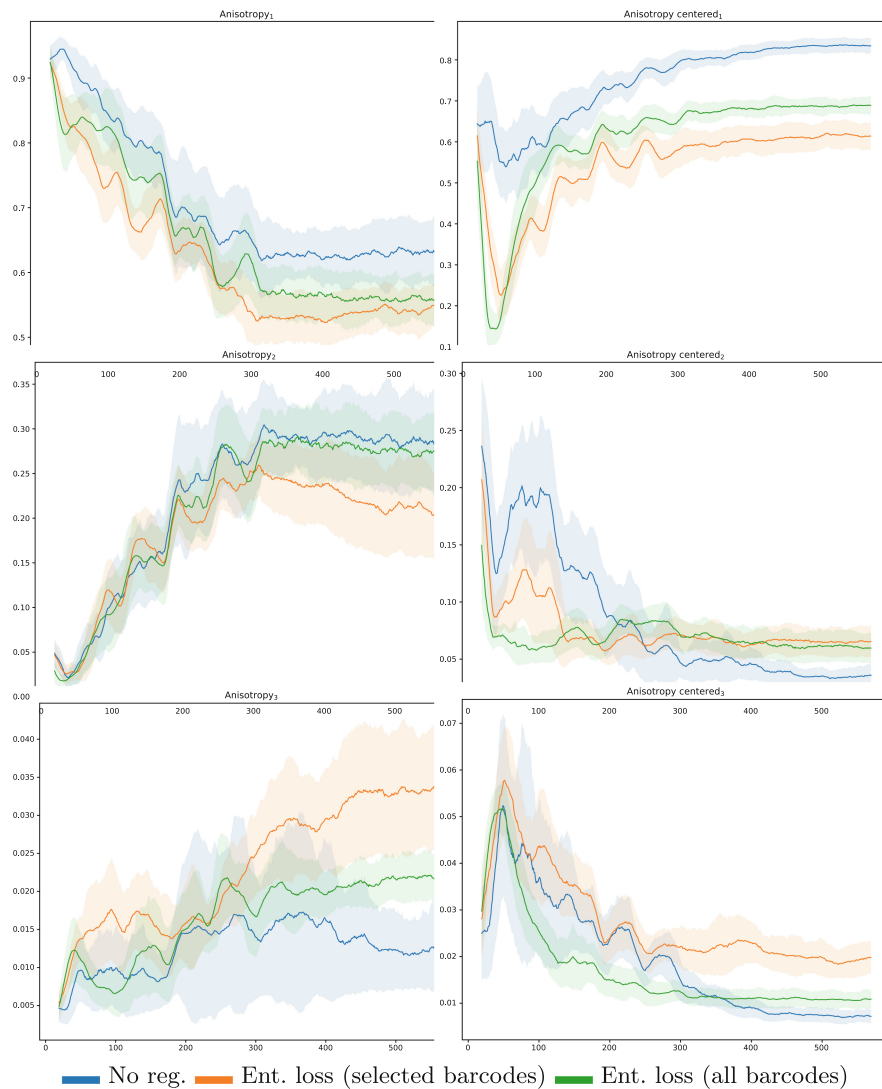


Fig. 1. Anisotropy profiles of RoBERTa models during fine-tuning on MRPC

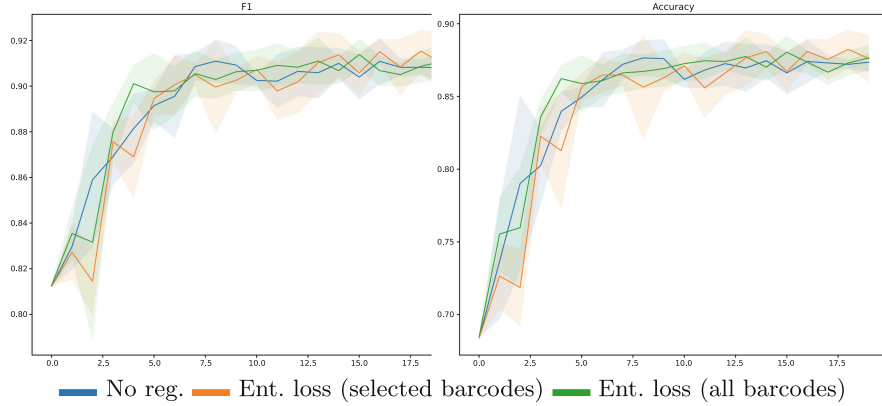


Fig. 2. MRPC validation metrics of RoBERTa models during fine-tuning

We also observe that feature selection indeed has a substantial effect on both isotropy and generalization. As we don’t discriminate between inter- and intra-cluster points if we skip the selection step, the lengths of resulting spanning tree edges obtained by entropy maximization step appear to be smaller. This leads to higher anisotropy as we fail to mitigate the contribution of most prominent directions. Destruction of the cluster-wise structure leads to poorer generalization.

Table 2. Classification metrics for the experiments

Task	No reg.	Ent. loss (sel. bars)	Ent. loss (all bars)
BERT-MRPC	0.892 ± 0.002	0.889 ± 0.001	0.892 ± 0.002
RoBERTa-MRPC	0.908 ± 0.002	0.911 ± 0.003	0.908 ± 0.002
BERT-COLA	0.575 ± 0.002	0.588 ± 0.003	0.572 ± 0.002
RoBERTa-COLA	0.608 ± 0.003	0.609 ± 0.002	0.609 ± 0.004

It is nonetheless interesting that such destruction doesn’t lead to the gradual decrease in downstream performance over the course of training, and the models perform mostly on par with the ones trained without regularization.

Overall, an addition of entropy regularization paired with topological feature selection leads to improvements both in terms of isotropy and generalization without the need of retraining the model, and is suitable for fine-tuning scenarios.

Table 3. Anisotropy values for the experiments

Task	Ani. component	No reg.	Ent. loss (sel. bars)	Ent. loss (all bars)
BERT-MRPC	Anisotropy ₁	0.679	0.485	0.541
	Anisotropy ₂	0.125	0.098	0.165
	Anisotropy ₃	0.023	0.042	0.028
	Ani. centered ₁	0.733	0.483	0.576
	Ani. centered ₂	0.034	0.057	0.050
	Ani. centered ₃	0.017	0.024	0.017
RoBERTa-MRPC	Anisotropy ₁	0.629	0.540	0.559
	Anisotropy ₂	0.287	0.214	0.276
	Anisotropy ₃	0.012	0.032	0.021
	Ani. centered ₁	0.832	0.611	0.686
	Ani. centered ₂	0.036	0.065	0.061
	Ani. centered ₃	0.007	0.019	0.010
BERT-COLA	Anisotropy ₁	0.647	0.499	0.502
	Anisotropy ₂	0.099	0.076	0.078
	Anisotropy ₃	0.028	0.028	0.023
	Ani. centered ₁	0.660	0.468	0.478
	Ani. centered ₂	0.038	0.037	0.032
	Ani. centered ₃	0.018	0.022	0.020
RoBERTa-COLA	Anisotropy ₁	0.581	0.534	0.542
	Anisotropy ₂	0.277	0.222	0.221
	Anisotropy ₃	0.016	0.021	0.018
	Ani. centered ₁	0.700	0.512	0.510
	Ani. centered ₂	0.056	0.073	0.063
	Ani. centered ₃	0.013	0.019	0.017

6 Conclusion

In our work we presented a novel regularization loss aimed to improve isotropy of the latent space, based on simplicial geometry and topological data analysis. We empirically evaluated its effect in fine-tuning scenario and found that it leads to an improvement in preserving the higher isotropy of the latent space, while improving the downstream metrics, which was reported to be unattainable for methods that rely on reparametrization of the space instead of exploiting existing structures [5]. On the contrary, our approach performs well even in absence of large training dataset and doesn't rely on retraining the model. It also doesn't add any additional inference overhead. It should be noted that although we apply our regularization loss to the language model's latent space, the general construction is actually model-agnostic and applicable to any optimization prob-

lem where the data has a point cloud structure. We leave an exploration of its usefulness on other model architectures and tasks for further work.

References

1. Atienza, N., Gonzalez-Diaz, R., Rucco, M.: Persistent entropy for separating topological features from noise in Vietoris-rips complexes. *J. Intell. Inf. Syst.* **52**, 637–655 (2019)
2. Cai, X., Huang, J., Bian, Y.-L., Church, K.W.: Isotropy in the contextual embedding space: clusters and manifolds. In: *International Conference on Learning Representations* (2021)
3. Demeter, D., Kimmel, G.J., Downey, D.: Stolen probability: a structural weakness of neural language models. In: *Annual Meeting of the Association for Computational Linguistics* (2020)
4. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: *North American Chapter of the Association for Computational Linguistics* (2019)
5. Ding, Y., Martinkus, K., Pascual, D., Clematide, S., Wattenhofer, R.: On isotropy calibration of transformer models. In: *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pp. 1–9. Association for Computational Linguistics (2022)
6. Ethayarajh, K.: 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)*, pp. 55–65. Association for Computational Linguistics (2019)
7. Gao, J., He, D., Tan, X., Qin, T., Wang, L., Liu, T.-Y.: Representation degeneration problem in training natural language generation models. In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019* (2019)
8. Godey, N., Villemonte de la Clergerie, E., Sagot, B.: Anisotropy is inherent to self-attention in transformers. In: *Conference of the European Chapter of the Association for Computational Linguistics* (2024)
9. Huang, J., et al.: WhiteningBERT: an easy unsupervised sentence embedding approach. In: *Conference on Empirical Methods in Natural Language Processing* (2021)
10. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)
11. Leygonie, J., Oudot, S., Tillmann, U.: A framework for differential calculus on persistence barcodes. *Found. Comput. Math.* **22**, 1069–1131 (2021)
12. Li, B., Zhou, H., He, J., Wang, M., Yang, Y., Li, L.: On the sentence embeddings from pre-trained language models. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9119–9130. Association for Computational Linguistics (2020)
13. Liang, Y., Cao, R., Zheng, J., Ren, J., Gao, L.: Learning to remove: towards isotropic pre-trained BERT embedding. In: Farkaš, I., Masulli, P., Otte, S., Wermter, S. (eds.) *ICANN 2021. LNCS*, vol. 12895, pp. 448–459. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86383-8_36
14. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692* (2019)

15. Moor, M., Horn, M., Rieck, B., Borgwardt, K.: Topological autoencoders. In: Proceedings of the 37th International Conference on Machine Learning, pp. 7045–7054. PMLR (2020)
16. Press, O., Wolf, L.: Using the output embedding to improve language models. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Valencia, Spain, ACL, pp. 157–163 (2017)
17. Rajaei, S., Pilehvar, M.: How does fine-tuning affect the geometry of embedding space: a case study on isotropy. In: Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 3042–3049. Association for Computational Linguistics (2021)
18. Razzhigaev, A., Mikhalechuk, M., Goncharova, E., Oseledets, I., Dimitrov, D., Kuznetsov, A.: The shape of learning: anisotropy and intrinsic dimensions in transformer-based models. In: Findings of the Association for Computational Linguistics: EACL 2024, pp. 868–874 (2024)
19. Trofimov, I., Cherniavskii, D., Tulchinskii, E., Balabin, N., Burnaev, E., Baranikov, S.: Learning topology-preserving data representations. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, 1–5 May 2023 (2023)
20. Vaswani, A., et al.: Attention is all you need. In: NEURIPS (2017)
21. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.: 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, pp. 353–355. Association for Computational Linguistics (2018)
22. Wang, L., Huang, J., Huang, K., Hu, Z., Wang, G., Gu, Q.: Improving neural language generation with spectrum control. In: Proceedings of International Conference on Learning Representations (2020)
23. Xu, J., Sun, X., Zhang, Z., Zhao, G., Lin, J.: Understanding and improving layer normalization. arXiv preprint [arXiv:1911.07013](https://arxiv.org/abs/1911.07013) (2019)
24. Yinhan, L., et al.: RoBERTa: A Robustly Optimized BERT Pretraining Approach (2019)
25. Zhang, L., Buntine, W., Shareghi, E.: On the effect of isotropy on VAE representations of text. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 694–701. Association for Computational Linguistics (2022)