Can Text Encoders be Deceived by Length Attack?

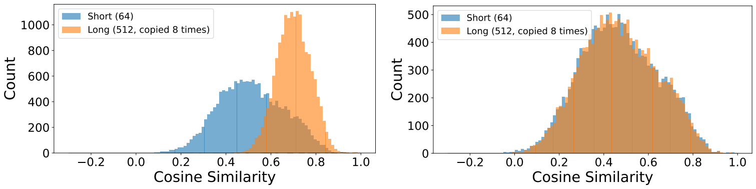
Chenghao Xiao[1,*], **Zihuiwen Ye**[2,*], **G Thomas Hudson**[1], **Zhongtian Sun**[1], **Phil Blunsom**[2], **Noura Al Moubayed**[1]
[1]University of Durham [2]University of Oxford

Abstract

Albeit *de facto* to use in training dense retrieval models, we observe that contrastive learning is prone to length overfitting, making it vulnerable to adversarial length attacks. We examine the behaviour of this phenomenon and propose an editing method to mitigate this problem. We find that our method can effectively improve the robustness of models against length attacks. Its effectiveness can be attributed to reduced length information in the embeddings, more robust intra-document token interaction, and enhanced isotropy at trained length range.

1 Introduction

Contrastive learning aims to learn meaningful representation by minimizing the differences between similar pairs and maximizing them between dissimilar pairs in the embedding space. We investigate the tendency of contrastive learning to length-based overfitting to the training set, and find that such models are vulnerable to *length attack*, i.e. we can deceive the models to perceive a higher similarity between two documents by simply using longer documents. We conduct extensive experiments to examine this phenomenon, and propose an editing method to mitigate this problem, significantly smoothing the distribution discrepancy among inputs of varied sequence lengths.



(a) Default Training

(b) Ours (Length Adversarial Training)

Figure 1: (a) Contrastive text encoders are misled to attribute greater similarity to document pairs when taking in long version of them. (b) Our method effectively mitigates this misalignment.

2 Method

To study how sentence encoders behave under *length attacks*, we construct *short* and *long* inference datasets, that have identical semantics, by concatenating each example in the *short* inference dataset multiple times to reach longer lengths. We then train a sentence encoder on *short* training dataset and test its performance on both *short* and *long* inference datasets. We use QQP for training and ArguAna for inference. We hypothesise that models trained only on *short* datasets would assign different similarity scores to *short* and *long* inference datasets, even if they share identical semantic meaning (see Appendix C for *short* and *long* dataset definitions, criteria, and preprocessing details).

We propose a copy-and-concat method to augment each *short* training example to align with the length range of the *long* inference dataset, which could be formalised as length-based adversarial training. We hypothesise that models trained with this method are more robust to deception, resulting in a smaller distributional shift in the similarity scores produced. We quantitively measure the inter-distributional shift on the similarity metric (cosine similarity) with Jenson Shannon Distance (JSD).

[1] Equal contribution. `chenghao.xiao@durham.ac.uk, zihuiwen.ye@cs.ox.ac.uk`

3 Results

We report the main results in Table 1 (left) and visualise the distributional difference in Figure 1. Figure 1 shows that the proposed method can effectively mitigate the spurious misalignment brought by document length. In Table 1 (left), we observe that in all settings, JSD decreases after length-based adversarial training, although there exists a behavioral gap when training with different pooling methods. We additionally perform experiments to measure inter-distributional shift on semantically-altered texts and observe similar behaviours. Details are in Appendix F.

Table 1: *Length attacked* distributional shift measured by JSD (left); Performance of length predictions on inference dataset measured by R^2x100 (right).

	Default JSD	**+ Ours**	**JSD difference**	**Default R^2**	**+ Ours**	**Length Resilience ↑**
Mean pooling 1 epoch	0.67	0.068	0.60	99.47	98.87	0.60
Mean pooling 3 epochs	0.64	0.048	0.59	99.43	98.05	1.38
[cls] pooling 1 epoch	0.52	0.127	0.39	92.57	86.47	6.10
[cls] pooling 3 epochs	0.43	0.075	0.36	92.92	84.70	8.22

Inner-workings We decouple the inner-workings of our method through three lenses:

1. **Probing embeddings for length prediction.** We construct a proxy indicator of how much length information is remaining, through probing embeddings from models trained with varied settings to predict sequence length of the inference set. Table 1 (right) shows that length-agnostic encoding ability brought by our method could partly be attributed to removed length information.
2. **More robust intra-document similarity.** Since models are typically deceived to attribute greater similarity to longer text inputs, it is of interest how tokens interact within the same document, and whether our method can modify this behaviour. We thus inspect intra-document similarity of embeddings (Ethayarajh, 2019) to understand the mechanism of length attack (Figure 2).

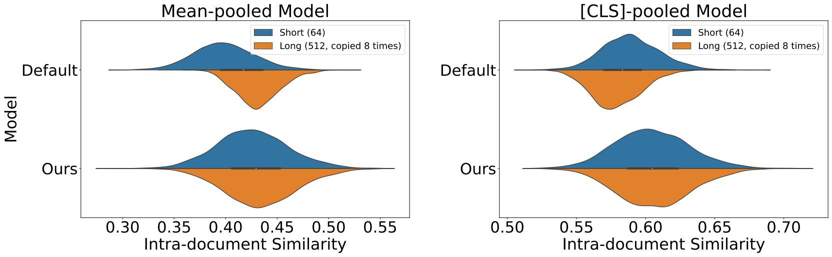


Figure 2: Intra-document similarity of tokens. After adversarial length training, tokens interact in a more robust way (intra similarity does not shift when taking in long copied text).

3. **Length-based enhanced isotropy.** Drawing on the known representation degeneration problem in PLMs (Gao et al., 2019; Ethayarajh, 2019) and the validated uniformity promise brought by contrastive learning (Wang & Isola, 2020; Gao et al., 2021), we hypothesise that if a model is trained on inputs of certain length, it would produce more isotropic embeddings at the corresponding length range and less at unseen range. We train an extra group of models on only *long-long* pairs, and conduct the same scoring on *short-long* pairs. In contrast with models that perceive longer documents to be more similar, we find that models that have only been exposed to long pairs tend to assign greater similarity to shorter pairs (see Figure 4 in Appendix).

4 Conclusion

This paper highlights a novel perspective in contrastive learning, which has yet rarely been observed. We examine the vulnerability of contrastive text encoders to adversarial length attack, and propose an effective editing method to mitigate it. We envision this method to be crucial in tasks and domains where only training sets of narrow length distribution are available, while the models yet have to be exposed to unexpected text length in production.

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

References

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 55–65, 2019.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*, 2019.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910, 2021.

Seiichiro Kondo, Kengo Hotate, Tosho Hirasawa, Masahiro Kaneko, and Mamoru Komachi. Sentence concatenation approach to data augmentation for neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pp. 143–149, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-srw.18. URL https://aclanthology.org/2021.naacl-srw.18.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models, 2021. URL https://arxiv.org/abs/2104.08663.

Dusan Varis and Ondřej Bojar. Sequence length is a domain: Length-based overfitting in transformer models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8246–8257, 2021.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 241–251, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1023. URL https://aclanthology.org/P18-1023.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020. URL http://proceedings.mlr.press/v119/wang20k.html.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020.

Chenghao Xiao, Yang Long, and Noura Al Moubayed. On isotropy and learning dynamics of contrastive-based sentence representation learning. *arXiv preprint arXiv:2212.09170*, 2022.

A Related Work

Length preference for dense retrieval models is observed in Thakur et al. (2021), who then show that models trained with dot-product and cosine similarity exhibit different length preferences. However, this phenomenon has not been attributed to the distributional misalignment of length between training and inference set, and it remains unknown what abilities of the model are enhanced and deteriorated when trained with certain length range.

We bridge this gap by inspecting how the length-agnostic pattern is reflected in a model's intra-document similarity (Ethayarajh, 2019; Xiao et al., 2022), and the model's diminished ability to encode length information. We further formalize it to be due to the eased anisotropy (Wang & Isola, 2020; Gao et al., 2021; Xiao et al., 2022) at certain length range.

Similar editing methods have been proved effective in machine translation (Kondo et al., 2021; Varis & Bojar, 2021), enhancing translation performance on the targeted length range. We show that a similar method applies well to contrastive text representation learning, and most importantly disentangle why it works with three theoretical perspectives.

B Experiment Details

Backbone We use MiniLM (Wang et al., 2020) as the backbone in all experiments. This model has achieved state-of-the-art performance on many semantic tasks when fine-tuned with contrastive loss [1], while being computationally cheap [2].

Training settings For each setting, we experiment with both [cls] and mean pooling. The models are all trained with both 1 epoch and 3 epochs. We train the models with the standard InfoNCE loss (Oord et al., 2018) as the contrastive loss:

$$\mathcal{L} = -\log \frac{\exp(\text{sim}(q, d_+)/\tau)}{\sum_{i=0}^{n} \exp(\text{sim}(q, d_i)/\tau)},$$

where the goal is to maximize the similarity between the query q and the corresponding document d_+, while discriminate against the rest of the documents in the batch, where $d_i \neq d_+$. In all settings, we set the temperature τ to 0.05.

C Datasets

Short and Long Datasets: Selection Criteria To construct *short* training sets, we use datasets that only consist of short document pairs, which could then be easily augmented using our method without complicated rules. For inference set, we look for datasets that have long documents, granting us the flexibility to truncate and replicate at diverse lengths, with varied amount of semantics remained. In auxiliary experiments, we also include training datasets that have only *long* document pairs and test the models' performance on *short* inference dataset, to study whether the observed pattern could be flipped on the selected inference set.

Training Set We use positive pairs in Quora Question Pairs[3] dataset (around $104k$) as the training set, as the dataset consists of mainly *short* document pairs. In the default training setting, we train models with the original dataset.

In our augmented setting (length adversarial training), given each input, we artificially augment it to be *long* by sampling a multiple to copy and concatenate it n times, making its augmented length fall into the range between the original and the maximum length (512). For instance, if an input has 20 tokens, we sample a multiple accordingly from $(1, 25)$. Following the augmentation, the dataset is aligned with the length range of inference set, allowing us to validate whether the length attack problem can be addressed.

We also include an extra group of experiments only on long-long document pairs to investigate whether the model's behavior would be reversed. In this group of experiments, we use the concate-

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[2] Notably, we use a 6-layer version of the model by taking every second layer. https://huggingface.co/nreimers/MiniLM-L6-H384-uncased

[3] https://huggingface.co/datasets/sentence-transformers/embedding-training-data

nation of title and question body from StackExchange dataset as document pairs. Each pair consists of a (title + question body, title + question body) pair.

Inference Set We use ArguAna (Wachsmuth et al., 2018) dataset in BEIR benchmark (Thakur et al., 2021) for inference. ArguAna is a corpus built for argument retrieval task collected from *idebate.org*. We extract arguments sharing the same title as inference pairs. To ensure that the inference document contains only *long* texts, we retain only pairs with length over 256 tokens. For *length attack* experiments, we construct *short* dataset by taking the first 64 tokens of each pair, and *long* dataset by concatenating them 8 times to reach token length 512.

The original corpus size for ArguAna is 8,674 with an average token length of 205. There are 541 topics for all the arguments in total. We take all possible pairs within each topic as inference pairs, resulting in 135,632 pairs of arguments. Then, we filter out the shorter pairs (token length $\leq$ 256), and end up with 163,46 pairs with an average token length of 365.

We plot the histograms of token lengths for ArguAna, Quora Question Pairs, and StackExchange[4] in Figure 3. Note that the training Quora Question Pair dataset has a significantly different length distribution compared to the inference ArguAna dataset.

Table 2: Statistics of datasets

Dataset	**#Pair**	**Length Range**	**Avg. Length**
Training			
Quora Question Pairs (only positive pairs)	103,663	[4,104]	14
Quora Question Pairs (only positive pairs, augmented)	103,663	[4,512]	227
StackExchange (question title + question body)	250,519	[7, 2702]	186
Inference			
ArguAna (original corpus)	8,674	[4, 1401]	205
ArguAna (processed corpus)	163,46	[256, 1401]	365

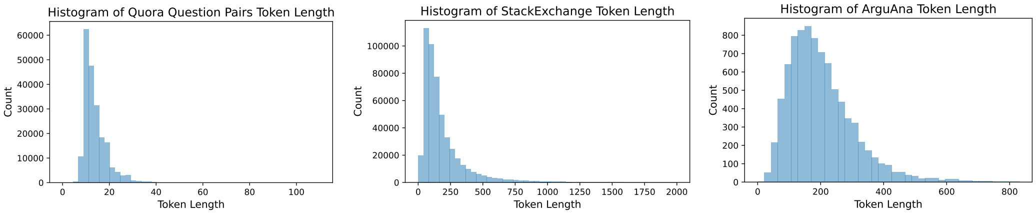


Figure 3: Token length distributions for train and inference datasets.

D Probing Experiment

As mentioned in Section 3, we investigate how much length information is remained in the embeddings, training with default setting and our method.

Specifically, we first encode all documents in the inference set (ArguAna) with all models. If a model is trained with [cls] pooling, we take the [cls] token as the document embedding; while if it is trained with mean pooling, we take the mean-pooled embedding as the document embedding.

For each setting, we use 80% of the (embedding, document length) pairs to train a linear regression model, and measure the R^2 of its prediction on the held-out 20% test set. We report the results of 5-fold cross validation averaged for each model.

[4] https://huggingface.co/datasets/sentence-transformers/embedding-training-data

E Reversed Pattern by Training on Long-Long Pairs

In Figure 4, we observe that when trained only on long-long document pairs, model behavior could be flipped. When scoring similarity on 64-token inference inputs and the long copied version of them (512, copied 8 times), models now favour shorter documents over longer documents. This again validates our hypothesis that isotropy of embeddings that models produce is enhanced more at the length range it has been trained on (the embeddings are pushed closer to 0 on that length range). Also, models could learn better nuances of text at the seen length range.

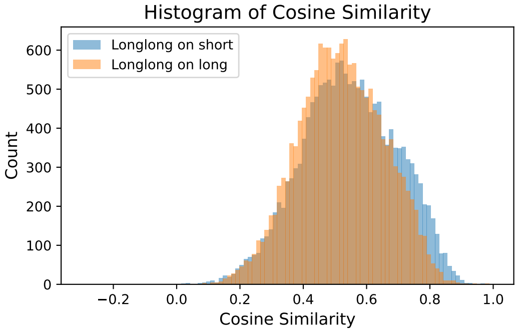
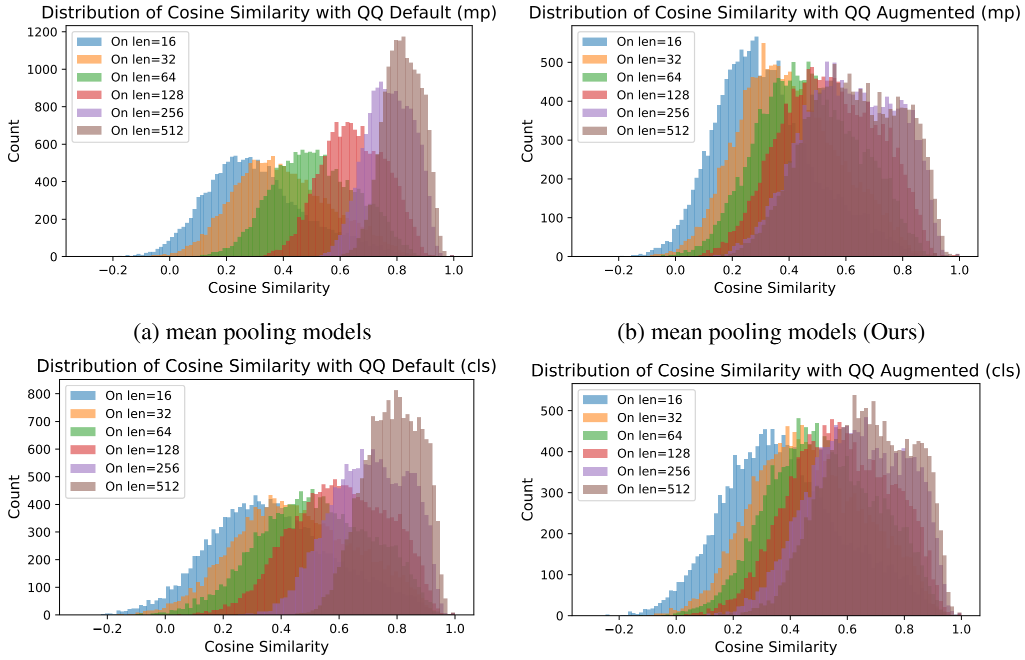


Figure 4: Similarity scoring on ArguAna pairs of 64 tokens and 512 tokens (copied 8 times) after training on long-long pairs. The model perceives short documents to be more similar.

F Results on semantic-altered texts

In addition to the semantic-preserved setting (Section C) in which the first 64 tokens of each inference dataset example (*short*) are concatenated 8 times to reach length 512 (*long*), we construct inference datasets of varying lengths by taking the first k tokens of each inference example, where $k \in [16, 32, 64, 128, 256, 512]$. We call this setting semantic-altered, as the first k tokens of each document preserve different levels of semantic meaning in the original document. We use the same Quora Question Pair for training and ArguAna for inference. We plot the cosine similarity scores for each k and show the result in Figure 5. We observe that similar to the semantic-preserved setting, the similarity score produced has smaller inter-distributional shift after length adversarial training.



(a) mean pooling models

(b) mean pooling models (Ours)

(c) cls pooling models

(d) cls pooling models (Ours)

Figure 5: Similarity scores on semantic-altered inference datasets with different pooling methods.