UNSEE: Unsupervised Non-contrastive Sentence Embeddings

Anonymous EACL submission

Abstract

In this paper, we introduce UNSEE, which stands for Unsupervised Non-Contrastive Sentence Embeddings. UNSEE demonstrates better performance compared to SimCSE in the Massive Text Embedding (MTEB) benchmark. We begin by highlighting the issue of representation collapse that occurs with the replacement of contrastive objectives with non-contrastive objectives in SimCSE. Subsequently, we introduce a straightforward solution called the target network to mitigate this problem. This approach enables us to harness non-contrastive objectives while ensuring training stability and achieving performance improvements similar to those seen with contrastive objectives. We have reached peak performance in non-contrastive sentence embeddings through extensive fine-tuning and optimization. These efforts have resulted in superior sentence representation models, emphasizing the importance of careful tuning and optimization for non-contrastive objectives.

1 Introduction

Contrastive learning has been used quite extensively in the sentence embedding models (Zhang et al., 2021b; Liu et al., 2021; Reimers and Gurevych, 2019; Chuang et al., 2022; Gao et al., 2021b; Yuxin Jiang and Wang, 2022; Liu et al., 2022) which achieve remarkable results on MTEB benchmark (Muennighoff et al., 2023). The contrastive objective serves the basic purpose of regularizing the anisotropic embedding space which eventually allows the language models to be used as efficient embedding models.

On the other hand, non-contrastive methods have not gained much popularity as a main objective for training sentence embedding models despite the shown regularization power in vision (Bardes et al., 2022). The primary reason is that non-contrastive objectives perform quite poorly compared to contrastive objectives when employed in the SimCSE setting. To illustrate, SCD (Klein and Nabi, 2022) which demonstrated that Barlow Twins (Zbontar et al., 2021) only achieves 67.57 in STSBenchmark (Cer et al., 2017) test set whereas SimCSE (Gao et al., 2021b) accomplishes 76.85.

Furthermore, we show that this is not peculiar to only Barlow Twins and other well-known non-contrastive methods (Bardes et al., 2022; Ozsoy et al., 2022) also suffer from poor performance as the top evaluation scores in Figure 2 are quite worse than SimCSE which has the 82.5.

Even though the non-contrastive objectives have inferior performance as an objective in a sentence embedding framework, their inherent properties such as needlessness to negative samples and avoidance of dimensional collapse as shown in Ozsoy et al. (2022) motivate us to further explore and enhance the performance of non-contrastive objectives.

Therefore, We first provide empirical evidence for the representation collapse during the training with non-contrastive objectives, specifically those employing the siamese network, dropout as augmentation and even with additional parametrization with MLP layers and discuss the possible reasons for the poor performance in section 4.1.

Moreover, we introduce the target network as a novel augmentation method that further diversifies the embeddings which empirically avoids the collapse of the non-contrastive objectives. Moreover, we achieve the absolute best performance out of non-contrastive objectives with further finetuning and architectural refinements which we detail in section 4.2 and section 4.3.

All in all, we present a series of non-contrastive models that we gather under the name UNSEE that surpass SimCSE in the MTEB benchmark which shows the potential of non-contrastive objectives as base objectives for the training of state-of-the-art embedding models.

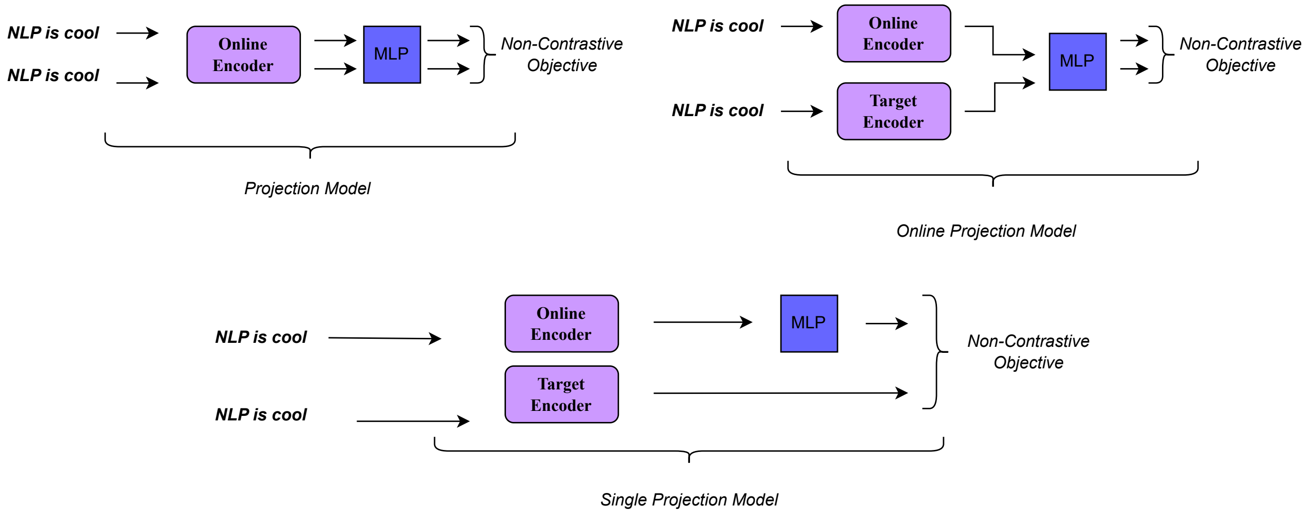

Figure 1: *Projection Model* is the same as SimCSE (Gao et al., 2021b). The Online keyword is to emphasize that the model gets gradient updates. The *Online Projection Model* is similar to the *Projection Model* except for the Target Encoder. The Target Encoder is an exponentially moving average of the Online network. Both outputs from Online and Target Encoders pass through the same MLP layer in the Online Projection Model. Target MLP is not employed due to the nature of fine-tuning which will slightly change the newly initialized MLP layer that will potentially corrupt the embeddings. In *Single Projection Model*, Target embeddings do not go through the MLP layer unlike *Online Projection Model*. *Single Projection Model* is identical to the architecture proposed in BSL (Zhang et al., 2021a). We only use BERT-base (Devlin et al., 2018) as the Encoder.

2 Related Work

Competitive models for sentence embeddings are constructed by adapting BERT (Devlin et al., 2018) with various configurations. Early sentence embedding models such as InferSent (Conneau et al., 2017) and the Universal Sentence Encoder(Cer et al., 2018) are predominantly based on LSTM (Hochreiter and Schmidhuber, 1997) or the Transformer(Vaswani et al., 2017).

The standard BERT (Devlin et al., 2018) model underperforms and operates at a slower pace. Sentence BERT abbreviated as SBERT (Reimers and Gurevych, 2019), represents a modified iteration of BERT which leverages siamese or triplet networks to generate meaningful and accurate sentence embeddings. SBERT enhances accuracy and significantly reduces the time needed to identify the most similar pair of sentences within a set of 10,000 sentences, reducing the process from 65 hours to just 5 seconds. Despite the incorporation of newer adjustments into BERT, a fundamental question arises: why are these modifications needed in the first instance?

Li et al. (2020) highlights a concern regarding BERT's sentence embeddings, specifically noting the presence of anisotropy in the embedding space. Their empirical findings demonstrate that the sentence embedding space is non-smoothing and poorly defined in certain areas, making it challenging to employ cosine similarity directly. In order to address this issue, they suggest a solution involving the transformation of sentence embeddings into a Gaussian distribution that is both smooth and isotropic, achieved through the use of normalizing flows. This flow-based generative model is trained in an unsupervised manner to maximize the likelihood of generating BERT sentence embeddings from a standard Gaussian latent variable.

Liu et al. (2021) introduce a method called MirrorBERT, which enhances sentence representations through a straightforward approach of duplicating or slightly augmenting the text input, all without relying on external supervision. These augmentations can occur either within the input space, involving actions such as random span masking, or within the feature space, employing techniques like dropout. Dropout is not only implemented within the MLP, but it also results in the deactivation of attention heads, all while maintaining the model's performance in various other tasks. Furthermore, it has been demonstrated that Mirror-BERT also enhances isotropy.

Gao et al. (2021b) introduce SimCSE, which employs conventional dropout as a means of input augmentation. By feeding a single sentence

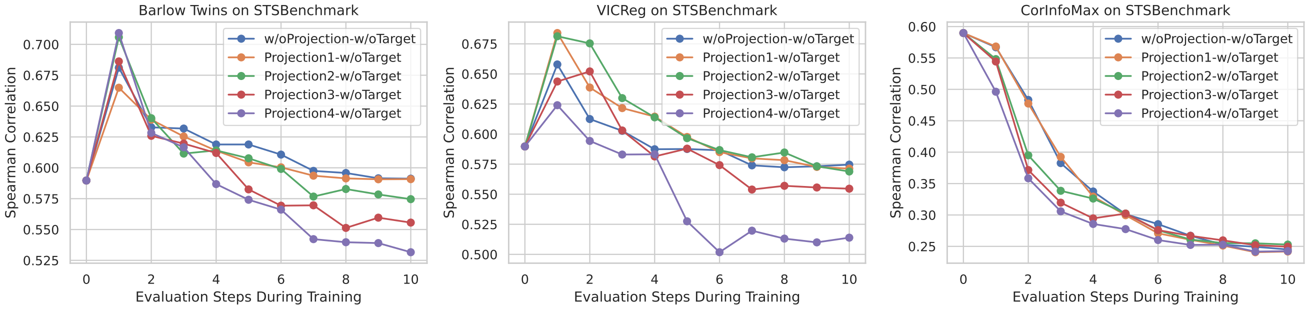

Figure 2: The performance of various non-contrastive objectives on STSBenchmark evaluation dataset (Cer et al., 2017) in the Projection Model or SimCSE setting. The difference between models is the number of MLP layers. MLP layer is adopted from BSL (Zhang et al., 2021b).

through two passes, this approach generates two distinct feature embeddings, which can be treated as similar to two separate pairs, while other sentences serve as negative samples. This dropout-based approach offers a straightforward technique for creating positive-negative pairs in contrastive learning. Impressively, it achieves superior performance compared to Mirror-BERT with only moderate modifications.

The current landscape of state-of-the-art embedding models (Xiao et al., 2023; Li et al., 2023; Su et al., 2023; Wang et al., 2022) is characterized by their training on exceedingly large and comprehensive corpora. These corpora consist of a vast volume of both unlabeled and labelled text data. This extensive and diverse training data has been instrumental in the remarkable performance achievements of these models in MTEB (Muennighoff et al., 2023) even though they are fundamentally identical with SimCSE.

In contrast, models like SimCSE operate under a significantly different paradigm, being trained on a comparatively modest dataset comprising just 1 million sentences. Given the substantial discrepancy in the scale and diversity of training data, making direct comparisons between SimCSE-like models and these state-of-the-art embedding models appears implausible and may not yield meaningful insights into their relative capabilities.

3 Background

In this section, we provide an extensive overview of non-contrastive representation learning and the methods that form the core of our research.

3.1 Non-Contrastive Representation Learning

Recent advancements in the field of self-supervised visual learning have extended beyond the traditional contrastive approach, exploring innovative avenues that reduce the reliance on negative sample pairs. These methods primarily focus on enhancing the quality of independently augmented representations, forming a subset of non-contrastive frameworks. To address challenges such as model collapse, various effective strategies have emerged within this domain. These include the adoption of asymmetric network architectures (Grill et al., 2020; Chen and He, 2020), feature decorrelation techniques (Zbontar et al., 2021; Bardes et al., 2022; Ozsoy et al., 2022; Ermolov et al., 2020), as well as clustering methods (Amrani and Bronstein, 2021; Assran et al., 2022; Caron et al., 2019, 2020), all of which contribute to the progress in self-supervised visual learning while addressing the challenges inherent to this domain.

3.2 CorInfoMax

CorInfoMax (Ozsoy et al., 2022) utilize a second-order statistics-based mutual information measure to gauge the level of correlation among its input components. The primary aims of maximizing this measure between different representations of the same input are twofold: firstly, it mitigates the risk of feature vector collapse by generating feature vectors with non-degenerate covariances. Secondly, it establishes relevance among these alternative representations by enhancing their linear interdependence.

An approximation of this information maximization objective simplifies into a Euclidean distance-based objective function, which is further regulated by the logarithm of the determinant of the feature

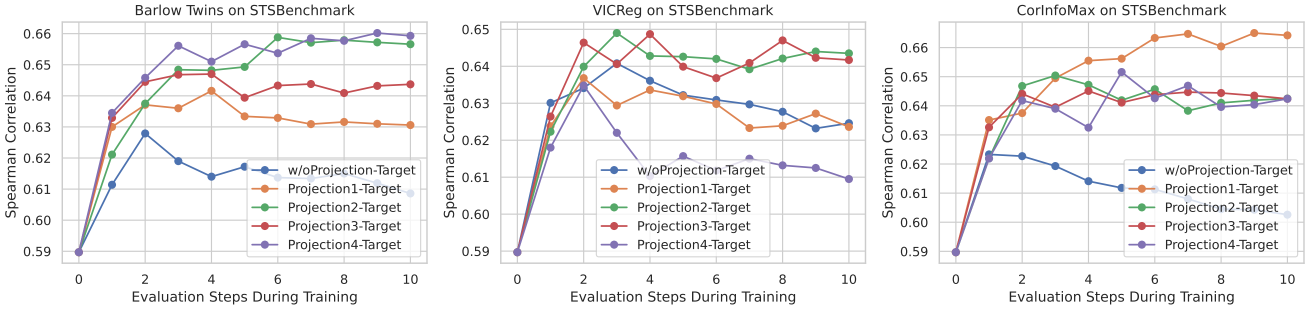

Figure 3: The performance of various non-contrastive objectives on STSBenchmark (Cer et al., 2017) in the Online Projection Model with SimCSE hyperparameters. The difference between models is the number of MLP layers. MLP layer is adopted from BSL (Zhang et al., 2021b).

covariance matrix. This regularization term serves as a natural safeguard against feature space degeneracy. Consequently, the proposed approach not only prevents complete output collapse to a single point but also effectively averts dimensional collapse by encouraging the dispersion of information across the entire feature space.

3.3 Barlow Twins

The Barlow Twins (Zbontar et al., 2021) is designed to prevent collapse in a natural way. It accomplishes this by assessing the cross-correlation matrix between the outputs of two identical networks, which are fed with altered versions of a sample. The goal is to make this cross-correlation matrix as similar to the identity matrix as possible. Consequently, this approach ensures that the embedding vectors of these distorted sample versions become more alike, all while reducing redundancy among their individual components. Importantly, Barlow Twins operates without the need for large batch sizes or introducing any disparities between the network twins, such as the inclusion of a predictor network, gradient stopping, or utilizing a moving average for weight updates.

3.4 VICReg

VICReg (Bardes et al., 2022), short for Variance-Invariance-Covariance Regularization, is an approach specifically designed to address the issue of collapse in a straightforward manner. It accomplishes this by introducing a simple regularization term that focuses on the variance of the embeddings along each dimension individually. In addition to the variance component, VICReg incorporates a mechanism that reduces redundancy and ensures decorrelation among the embeddings, achieved through covariance regularization.

3.5 BYOL

BYOL (Grill et al., 2020) hinges on the utilization of two distinct neural networks, namely the online and target networks, which collaborate and mutually enhance their learning processes. This technique operates by presenting an augmented view of an image to the online network, with the objective of training it to predict the representation of the same image as processed by the target network but under a different augmented view. Simultaneously, the target network undergoes updates through a slow-moving average mechanism based on the evolving state of the online network.

This approach essentially fosters a dynamic interplay between the online and target networks, where they iteratively adapt and refine their representations in response to the variations in augmented views. Through this collaborative learning process, BYOL aims to yield highly informative and generalized feature representations, making it particularly valuable for self-supervised learning tasks, where labelled data may be limited or unavailable.

4 From SimCSE to the UNSEE

In this section, we elaborate on the process of deriving the ultimate UNSEE models from SimCSE. We utilize the STSBenchmark evaluation dataset (Cer et al., 2017) to identify the optimal configuration. We adopt a systematic approach, incrementally discussing enhancements and providing rationales for each decision made. Lastly, SimCSE has an 82.5 score in the STSBenchmark. We exclude it intentionally from our figures since the high gap ruins the visualization in some experiments.

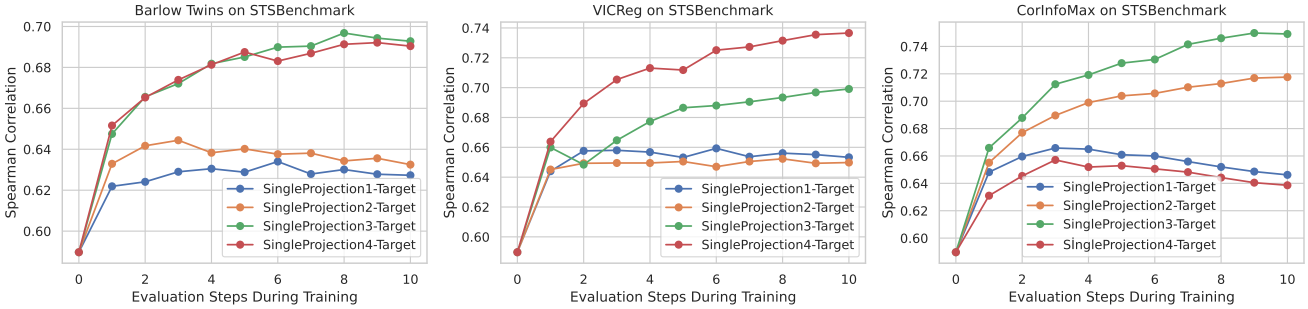

Figure 4: The performance of various non-contrastive objectives on STSBenchmark (Cer et al., 2017) in the Single Projection Model with SimCSE hyperparameters.

4.1 Projection Model

In Figure 1, the ’Projection model’ corresponds to the precise configuration outlined in SimCSE (Gao et al., 2021b), wherein dropout serves as a straightforward augmentation technique.

Figure 2 offers compelling evidence of substantial deficiencies in non-contrastive models when employed within the SimCSE framework. It’s conceivable to assert that these models undergo a representation collapse during their training phase. This leads to critical questions regarding the broader versatility and generalization capacity of such objectives, hinting at their potential effectiveness within constrained domains or contexts.

Conversely, it’s important to note that dropout augmentation emerges as a pivotal element within the SimCSE paradigm. This observation prompts us to consider the possibility of exploring alternative augmentation techniques in order to delve deeper into the potential inherent in non-contrastive objectives. This pursuit of diverse augmentation strategies could potentially unveil the true efficacy and versatility of these objectives, shedding light on their capabilities beyond their current limitations.

4.2 Online Projection Model

Given the significant underperformance of non-contrastive objectives, it’s crucial to search for new ways to enhance them. As noted by Gao (2021), most input space augmentations are not as effective as dropout. This discovery makes it doubtful that we will find an input augmentation method better than dropout.

This realization has steered our exploration toward the development of a novel augmentation technique, namely, the utilization of a target network. This approach represents a relatively straightforward feature space augmentation strategy designed to introduce greater diversity into the embeddings, surpassing the efficacy of traditional dropout. One can even draw parallels to ’lagged dropout’, wherein networks subject to dropout exhibit slight variations, and the target network operates as a slow-moving average of the online network, contributing to the diversification of embeddings.

Figure 3 illustrates that the use of a target network effectively prevents representation collapse and ensures a more stable training process. However, it is worth noting that, even in scenarios where representation collapse is avoided, the overall performance remains subpar. The introduction of additional parametrization through MLP layers has had only a marginal impact on improving performance.

One could make the argument that constructing effective sentence embeddings poses a greater challenge when non-contrastive objectives are employed, particularly when compared to vision-related tasks. In the realm of contrastive learning, the process involves explicitly pushing data samples away from each other to enhance discrimination. However, in the context of sentence embeddings with non-contrastive objectives, this process becomes implicit.

To draw a parallel, consider a problem in computer vision where every single sample is assigned a distinct label. However, these samples also share certain common labels among them. Similarly, when training a sentence embedding model using non-contrastive objectives, it mirrors this intricate situation given that we use the dataset which consists of randomly sampled Wikipedia sentences that are collected in SimCSE (Gao et al., 2021b). Each sentence in the dataset may be unique in its content, yet there exist underlying semantic or syntactic relationships among them, akin to the shared labels in the vision problem. This inherent complexity and

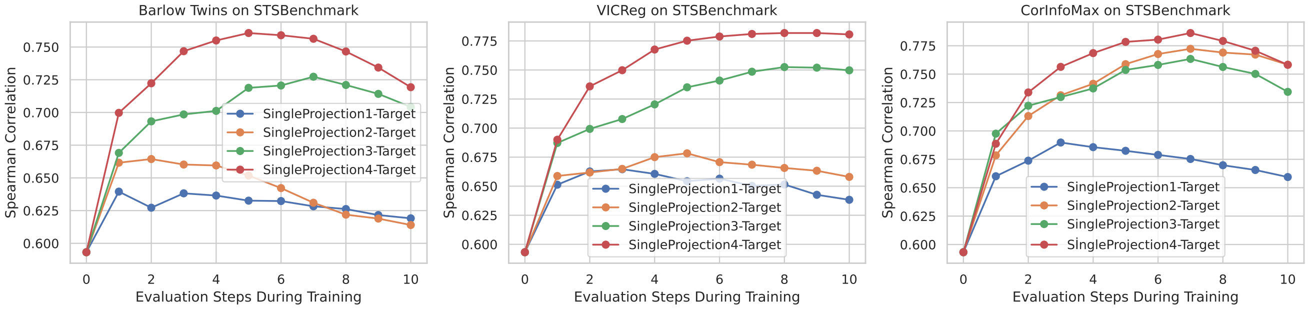

Figure 5: The performance of various non-contrastive objectives on STSBenchmark (Cer et al., 2017) in the Single Projection Model with slightly optimized hyperparameters. The difference between models is the number of MLP layers. MLP layer is adopted from BSL (Zhang et al., 2021b).

the need to implicitly capture these relationships contribute to the intricacy of the sentence embedding task when utilizing non-contrastive objectives.

4.3 Single Projection Model

Within the context of our online projection model, it is imperative to underscore the pivotal role played by MLP layers in the processing pipeline of both target and online embeddings. It is worth noting that the sentence embeddings themselves are originally derived from the BERT model

The MLP layers, however, should not be regarded as static fixtures within our model architecture; rather, they assume a dynamic and transient role during the training phase. Their purpose is instrumental in continuously shaping the embeddings to ensure effective loss minimisation. Nevertheless, it is essential to underscore that the outputs generated by these MLP layers do not constitute the definitive embeddings employed for subsequent evaluation.

This leads us to a compelling conjecture: What if we were to consider circumventing the MLP layers in the processing of the target network's embeddings? By establishing a direct, unmediated connection between the loss minimization process and the generation of embeddings, we endeavour to explore whether such architectural simplification could yield substantial advantages. This modification holds the potential to provide insights into whether a more streamlined approach might enhance both the efficiency of loss minimization and the quality of the resultant embeddings, thereby refining the overall training process.

The results in Figure 4 align closely with our hypothesis. Throughout the training process, the models consistently demonstrated incremental improvements in performance, surpassing the achievements of the previous model despite retaining identical complexities and hyperparameters.

While these findings are undeniably promising, it's important to note that they have not yet reached the level of performance exhibited by SimCSE. This indicates that further optimization efforts are warranted to bridge the gap and enable our models to match the performance of their SimCSE counterparts. Thus, there is room for refinement and enhancement in pursuit of achieving comparable or even superior performance.

We have managed to significantly enhance our model's performance by making relatively minor adjustments to certain hyperparameters, specifically focusing on the learning rate, batch size, and sequence length. The best hyperparameters are 1e-4, 32 and 64 respectively. The decay rate is 0.999 and kept the same across all experiments. Remarkably, these subtle modifications have allowed us to achieve the highest achievable scores among non-contrastive objectives, all without delving into the optimization of hyperparameters within the loss objective.

It's worth emphasizing that we have deliberately chosen to adhere to default values for the objectives, highlighting the robustness and transferability of these objectives across different domains. This observation underscores the versatility of the objectives, as they continue to perform effectively even when applied in contexts beyond their original domain.

The results depicted in Figure 5 do not represent the pinnacle of our achievement. We have achieved even better results by conducting more frequent evaluations(20 evaluations per run) during the training process and implementing a checkpointing sys-

	Class.	Clust.	PairClass.	Rerank.	Retr.	STS	Summ.	Avg.
Num. Datasets (→)	12	11	3	4	15	10	1	56
Self-supervised methods								
Glove	57.29	27.73	70.92	43.29	21.62	61.85	28.87	41.97
Komninos	57.65	26.57	**72.94**	44.75	21.22	62.47	30.49	42.06
BERT	61.66	30.12	56.33	43.44	10.59	54.36	29.82	38.33
SimCSE	62.50	29.04	70.33	46.47	20.29	**74.33**	**31.15**	45.45
UNSEE-BYOL(Ours)	62.55	27.81	65.3	46.47	23.11	73.04	30.68	45.46
UNSEE-Barlow(Ours)	62.76	**30.04**	65.7	46.9	23.06	72.15	30.25	45.82
UNSEE-CorInfoMax(Ours)	**62.85**	28.90	67.87	46.81	**24.80**	72.31	30.81	46.22
UNSEE-VICReg(Ours)	62.58	28.44	70.24	**47.23**	24.79	73.11	30.34	**46.37**

Table 1: Average of the main metric from Muennighoff et al. (2023) per task per model on MTEB English subsets. SimCSE, BERT, Komnimos, and Glove scores are taken from Muennighoff et al. (2023)

tem to capture the best-performing model. These specific runs were designed to align with our prior experiments, aimed at illustrating the efficacy of the adjustments made.

5 Evaluation Dataset

5.1 MTEB Benchmark

The primary goal of the Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2023) is to offer a comprehensive assessment of model performance across a diverse range of text embedding tasks. It serves as a valuable resource for identifying text embeddings that exhibit universal applicability across a wide spectrum of tasks. MTEB encompasses an extensive collection of 58 datasets spanning 112 languages, encompassing 8 distinct embedding tasks, including bitext mining, classification, clustering, pair classification, reranking, retrieval, STS (Semantic Textual Similarity), and summarization.

6 BYOL, BSL and Final Results

In our paper, we extensively examine and engage in discussions concerning non-contrastive objectives that incorporate a siamese network architecture. However, it's important to note that our most effective configuration closely resembles BYOL (Grill et al., 2020), and we have conducted training to incorporate this configuration into our results. The ultimate model we present is a variation of BSL (Zhang et al., 2021b) with dropout serving as an augmentation method.

Throughout our experimentation, it becomes evident that non-contrastive methods consistently outperform SimCSE as the table 1 verifies. The degree of improvement varies, with some methods showing only marginal enhancements, while others exhibit significantly more substantial gains. This overarching pattern underscores the compelling impact of non-contrastive objectives on augmenting BERT's proficiency as a sentence embedding model.

Our findings collectively reinforce the notion that non-contrastive methods contribute to a notable expansion of BERT's capabilities, effectively harnessing its potential to serve as a highly effective and versatile tool for generating sentence embeddings. This empirical evidence underscores the transformative role these methods play in enhancing the utility and adaptability of BERT across various sentence-related tasks.

7 Conclusion

UNSEE (Unsupervised Non-Contrastive Sentence Embeddings) is a simple framework for non-contrastive sentence embeddings, which outperforms SimCSE in the Massive Text Embedding Benchmark (MTEB). We address representation collapse using a simple solution called the target network, enabling stable training and achieving performance similar to contrastive objectives. Our meticulous fine-tuning leads to performant sentence embedding models, showcasing the significance of thoughtful optimization in advancing non-contrastive methods for sentence representation.

8 Limitations

UNSEE models have inherent limitations stemming from their training data, which encompasses only one million sentences. In contrast, state-of-the-art embedding models undergo training on datasets comprising over a hundred million, or even more than a billion pairs. As a result, our models are expected to exhibit inferior performance when compared to models specifically designed for sentence embedding. We recommend considering the top-performing models on the MTEB leaderboard for more effective practical use.

9 Ethics Statement

The models under examination, UNSEE-*, lack generative abilities, ensuring their incapacity to produce unfair, biased, or harmful content. The datasets utilized in this study have been meticulously selected from reputable repositories known for their safety in research applications, with strict measures in place to prevent the inclusion of personal information or offensive material.

References

Elad Amrani and Alexander M. Bronstein. 2021. Self-supervised classification network. *ArXiv*, abs/2103.10994.

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael G. Rabbat, and Nicolas Ballas. 2022. Masked siamese networks for label-efficient learning. In *European Conference on Computer Vision*.

Adrien Bardes, Jean Ponce, and Yann LeCun. 2022. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Conference on Empirical Methods in Natural Language Processing*.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2019. Deep clustering for unsupervised learning of visual features.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *ArXiv*, abs/2006.09882.

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

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Xinlei Chen and Kaiming He. 2020. Exploring simple siamese representation learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15745–15753.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Aleksandr Ermolov, Aliaksandr Siarohin, E. Sangineto, and N. Sebe. 2020. Whitening for self-supervised representation learning. *ArXiv*, abs/2007.06346.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021a. Making pre-trained language models better few-shot learners. 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 3816–3830, Online. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021b. Simcse: Simple contrastive learning of sentence embeddings. *ArXiv*, abs/2104.08821.

Jean-Bastien Grill, Florian Strub, Florent Altch'e, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. 2020. Bootstrap your own latent: A new approach to self-supervised learning. *ArXiv*, abs/2006.07733.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *ArXiv*, abs/2006.03654.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. *ArXiv*, abs/1602.03483.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Ting Jiang, Shaohan Huang, Zi qiang Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Liangjie Zhang, and Qi Zhang. 2022. Promptbert: Improving bert sentence embeddings with prompts.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *NIPS*.

Tassilo Klein and Moin Nabi. 2022. Scd: Self-contrastive decorrelation of sentence embeddings. *ArXiv*, abs/2203.07847.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning.

Fangyu Liu, Yunlong Jiao, Jordan Massiah, Emine Yilmaz, and Serhii Havrylov. 2022. Trans-encoder: Unsupervised sentence-pair modelling through self- and mutual-distillations. In *ICLR 2022*.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1442–1459, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. *ArXiv*, abs/1803.02893.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Serdar Ozsoy, Shadi Hamdan, Sercan Ö. Arik, Deniz Yuret, and Alper T. Erdogan. 2022. Self-supervised learning with an information maximization criterion.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *ArXiv*, abs/2212.03533.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *ArXiv*, abs/2005.10242.

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

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

Linhan Zhang Yuxin Jiang and Wei Wang. 2022. Improved universal sentence embeddings with prompt-based contrastive learning and energy-based learning.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*.

Yan Zhang, Ruidan He, Zuozhu Liu, Lidong Bing, and Haizhou Li. 2021a. Bootstrapped unsupervised sentence representation learning. In *Annual Meeting of the Association for Computational Linguistics*.

Yan Zhang, Ruidan He, Zuozhu Liu, Lidong Bing, and Haizhou Li. 2021b. Bootstrapped unsupervised sentence representation learning. 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 5168–5180, Online. Association for Computational Linguistics.

A Training Details

We implement UNSEE with *SentenceTransformers* from (Reimers and Gurevych, 2019). To compare our models while developing them we keep the hyperparameters as same as the SimCSE which are 64 for batch size, 3e-5 for learning rate and 32 for the sequence length. When the target network is employed, the decay rate is 0.999 throughout all experiments. Our best models have 32 for the batch size, 1e-4 for the learning rate, and 64 for the sequence length, decay rate is the same. Best BYOL and VICReg models use 3 layers of MLP. CorInfoMax and Barlow Twins use 4. We use the same MLP architecture as BSL (Zhang et al., 2021b). In Barlow Twins, we use the same λ as the original paper which is 0.0051. In VICReg, we use the same hyperparameter weights from the original paper which are 25 for invariance and variance, 1 for covariance. In CorInfoMax, we use R_ini=1, la_=0.01,la_mu=0.01, R_eps_weight=1e-6, 0.2 for covariance and 2000 for invariance loss.

B Computational Requirements

We only use Tesla T4 GPUs for our experiments.