UNSEE: Unsupervised Non-contrastive Sentence Embeddings

Ömer Veysel Çağatan
Koç University
Rumelifeneri, Sarıyer Rumeli Feneri Yolu
34450 Sarıyer/İstanbul,Turkey
ocagatan19@ku.edu.tr

Abstract

We present UNSEE: Unsupervised Non-Contrastive Sentence Embeddings, a novel approach that outperforms SimCSE in the Massive Text Embedding benchmark. Our exploration begins by addressing the challenge of representation collapse, a phenomenon observed when contrastive objectives in SimCSE are replaced with non-contrastive objectives. To counter this issue, we propose a straightforward solution known as the target network, effectively mitigating representation collapse. The introduction of the target network allows us to leverage non-contrastive objectives, maintaining training stability while achieving performance improvements comparable to contrastive objectives. Our method has achieved peak performance in non-contrastive sentence embeddings through meticulous fine-tuning and optimization. This comprehensive effort has yielded superior sentence representation models, showcasing the effectiveness of our approach.

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 hace achieved remarkable results on MTEB benchmark (Muennighoff et al., 2023). The fundamental role of the contrastive objective is to regularize the anisotropic embedding space of language models, ultimately enabling them to function effectively as embedding models (Li et al., 2020).

On the contrary, non-contrastive methods have not gained widespread popularity as the primary objective for training sentence embedding models, despite demonstrating regularization efficacy in vision (Bardes et al., 2022; Zbontar et al., 2021; Chen and He, 2020; Grill et al., 2020). This reluctance stems from the fact that non-contrastive objectives tend to perform suboptimally in comparison to contrastive objectives, particularly in the SimCSE (Gao et al., 2021b) setting. For example, SCD (Klein and Nabi, 2022) showcased that Barlow Twins (Zbontar et al., 2021) achieves only 67.57 on the STSBenchmark (Cer et al., 2017) test set, while SimCSE (Gao et al., 2021b) accomplishes 76.85.

Additionally, we demonstrate that the observed performance drawback is not confined to Barlow Twins exclusively. Other well-known non-contrastive methods (Bardes et al., 2022; Ozsoy et al., 2022) also suffer from inferior performance. Specifically, when examining the top evaluation scores in Figure 2 for the STSBenchmark development set, these non-contrastive methods consistently fall short compared to SimCSE, which achieves an impressive score of 82.5.

Despite the comparatively lower performance observed when non-contrastive objectives are employed in a sentence embedding framework, their inherent characteristics, such as the lack of negative samples and the ability to prevent dimensional collapse, as demonstrated in Ozsoy et al. (2022), inspire us to delve deeper into investigating and improving the effectiveness of non-contrastive objectives.

Hence, we begin by presenting empirical evidence of representation collapse observed during training with non-contrastive objectives. This includes instances utilizing siamese networks, dropout as augmentation, and even those incorporating additional parametrization with MLP layers. We delve into the potential reasons behind the suboptimal performance in Section 4.1.

Furthermore, we introduce the target network as a novel augmentation method, which empirically enhances the diversity of embeddings and effectively mitigates the collapse associated with non-contrastive objectives. Subsequently, through additional finetuning and architectural refinements, detailed in Section 4.2 and Section 4.3, we achieve the

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 384–393

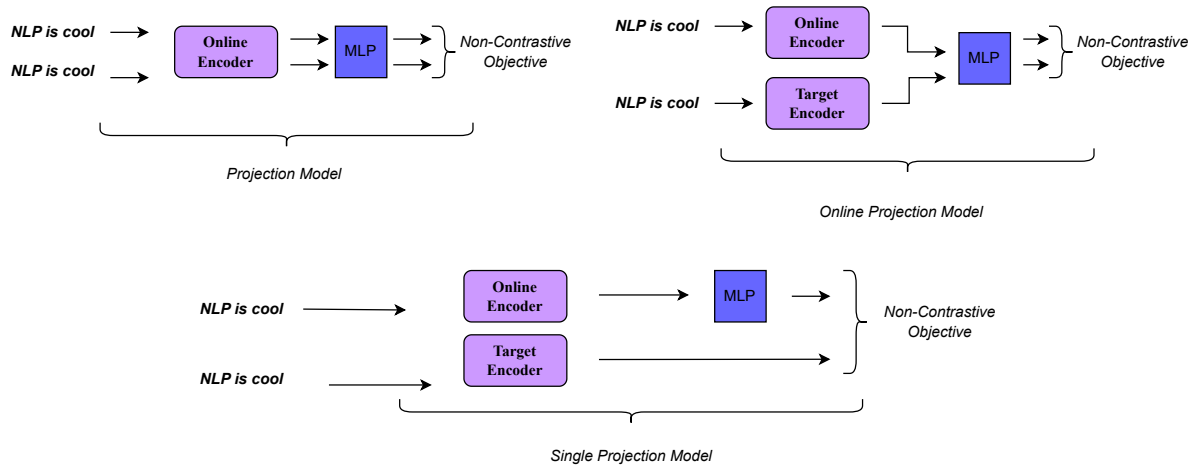

Figure 1: *Projection Model* is the same as SimCSE (Gao et al., 2021b). The Online keyword is to emphasize that the model gets gradient updates. The *Online Projection Model* is similar to the *Projection Model* except for the Target Encoder. The Target Encoder is an exponentially moving average of the 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.

absolute best performance among non-contrastive objectives. In summary, we present a series of non-contrastive models collectively named UNSEE, surpassing SimCSE in the MTEB benchmark. This underscores the potential of non-contrastive objectives as fundamental components for training state-of-the-art embedding models.

2 Related Work

Competitive sentence embedding models are typically built by modifying BERT (Devlin et al., 2018) with diverse configurations. In the early stages of sentence embedding development, models like InferSent (Conneau et al., 2017) and the Universal Sentence Encoder (Cer et al., 2018) predominantly relied on LSTM (Hochreiter and Schmidhuber, 1997) or the Transformer (Vaswani et al., 2017) architecture.

The conventional BERT model (Devlin et al., 2018) exhibits suboptimal performance and operates at a slower pace. Sentence BERT, abbreviated as SBERT (Reimers and Gurevych, 2019), represents a modified version of BERT that utilizes siamese or triplet networks to generate meaningful and accurate sentence embeddings. SBERT improves accuracy and significantly reduces the time required 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 integration of these enhancements into BERT, a fundamental question arises: why are these modifications necessary in the first place?

Li et al. (2020) brings attention to a concern related to BERT's sentence embeddings, specifically highlighting the presence of anisotropy in the embedding space. Their empirical observations reveal that the sentence embedding space lacks smoothness and is poorly defined in certain regions, posing challenges when applying cosine similarity directly. To address this issue, they propose a solution that involves transforming sentence embeddings into a Gaussian distribution that is both smooth and isotropic. This transformation is achieved through the utilization of normalizing flows. The proposed flow-based generative model is trained in an unsupervised manner with the objective of maximizing the likelihood of generating BERT sentence embeddings from a standard Gaussian latent variable.

Liu et al. (2021) present MirrorBERT, a method that improves sentence representations through a straightforward approach of duplicating or slightly augmenting the text input, all without external supervision. These augmentations can take place either within the input space, involving actions like random span masking, or within the feature space, using techniques such as dropout. Notably, dropout is not only implemented within the MLP but also leads to the deactivation of attention heads, all

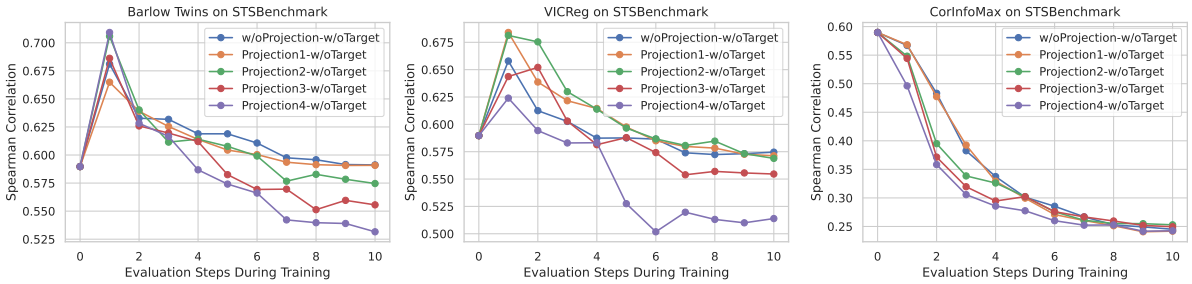

Figure 2: The performance of various non-contrastive objectives on STSBenchmark evaluation dataset (Cer et al., 2017) in the Projection Model or SimCSE setting. The difference between models is the number of MLP layers. MLP layer is adopted from BSL (Zhang et al., 2021b).

while preserving the model's performance across various tasks. Furthermore, it has been demonstrated that MirrorBERT also enhances isotropy.

Gao et al. (2021b) introduce SimCSE, which employs conventional dropout as a means of input augmentation. By feeding a single sentence through two passes, this approach generates two distinct feature embeddings, which can be treated as similar to positive 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 state-of-the-art embedding models (Xiao et al., 2023; Li et al., 2023; Su et al., 2023; Wang et al., 2022) distinguish themselves by their training on exceptionally large and extensive corpora. These corpora encompass a vast amount of both unlabeled and labeled text data. The utilization of such extensive and diverse training data has played a crucial role in the impressive performance exhibited by these models in the MTEB benchmark (Muennighoff et al., 2023), despite their fundamental similarity to SimCSE.

On the contrary, models such as SimCSE follow a significantly different paradigm, undergoing training on a relatively modest dataset consisting of just 1 million sentences. Considering the substantial difference in the scale and diversity of training data, attempting direct comparisons between SimCSE-like models and these state-of-the-art embedding models seems impractical and might not provide meaningful insights into their relative capabilities. Therefore, we exclude them from our analysis.

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) utilizes 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.

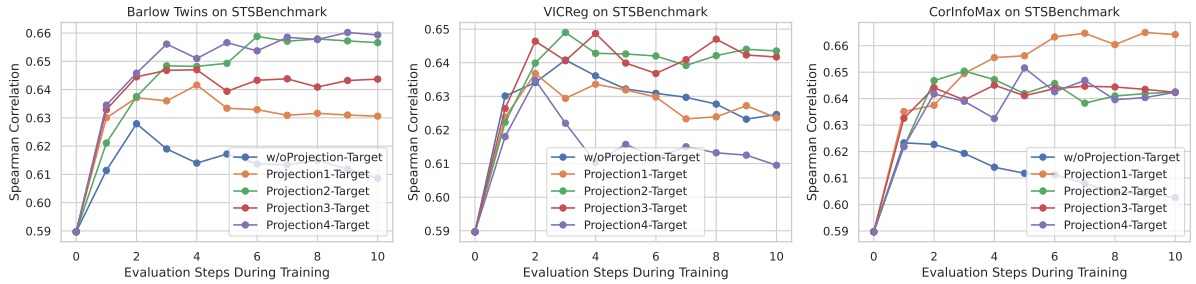

Figure 3: The performance of various non-contrastive objectives on STSBenchmark (Cer et al., 2017) in the Online Projection Model with SimCSE hyperparameters. The difference between models is the number of MLP layers. MLP layer is adopted from BSL (Zhang et al., 2021b).

An approximation of this information maximization objective simplifies into an Euclidean distance-based objective function, which is further regulated by the logarithm of the determinant of the feature 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 naturally. 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 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 straightforwardly. 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, to train 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 labeled data may be limited or unavailable.

4 From SimCSE to the UNSEE

In this section, we detail the methodology employed to derive the final UNSEE models from SimCSE. The STSBenchmark evaluation dataset (Cer et al., 2017) serves as the basis for identifying the optimal configuration. We follow a systematic approach, progressively discussing enhancements and offering justifications for each decision. It's worth noting that SimCSE achieves a score of 82.5 in the STSBenchmark. However, we intentionally ex-

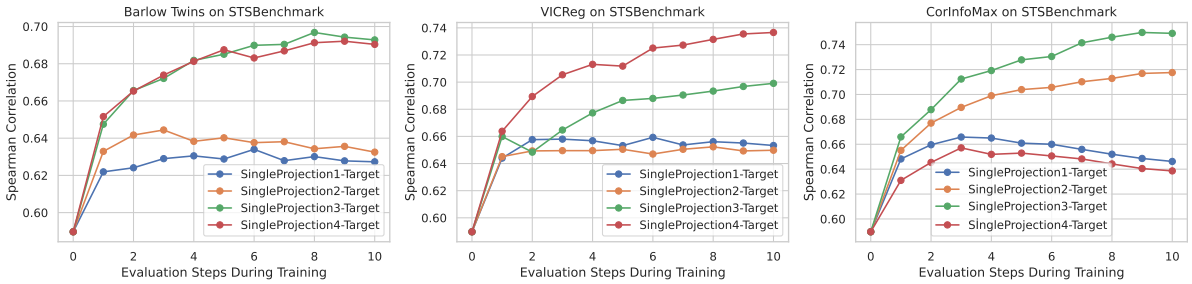

Figure 4: The performance of various non-contrastive objectives on STSBenchmark (Cer et al., 2017) in the Single Projection Model with SimCSE hyperparameters.

clude it from our figures as its high score can distort the visualization in certain experiments.

4.1 Projection Model

In Figure 1, *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 is noteworthy that dropout augmentation plays a pivotal role within the SimCSE paradigm. This realization leads us to consider the prospect of exploring alternative augmentation techniques, aiming to delve deeper into the inherent potential of non-contrastive objectives. This exploration of diverse augmentation strategies has the potential to reveal the true efficacy and versatility of these objectives, providing insights into their capabilities beyond their current limitations.

4.2 Online Projection Model

Considering the notable underperformance of non-contrastive objectives, it becomes imperative to explore novel avenues for their improvement. As highlighted by Gao et al. (2021a), most input space augmentations are not as effective as dropout. This finding casts doubt on the likelihood of discovering an input augmentation method superior to dropout.

This recognition has guided our exploration towards the creation of a new augmentation technique, specifically, the incorporation of a target network. This method constitutes a relatively straightforward feature space augmentation strategy aimed at infusing greater diversity into the embeddings, surpassing the effectiveness of conventional dropout. An analogy can be drawn to *lagged dropout*, where networks undergoing dropout display subtle variations, and the target network functions as a slow-moving average of the online network, actively contributing to the diversification of embeddings.

Figure 3 demonstrates that the utilization of a target network effectively prevents representation collapse, ensuring a more stable training process. However, it is noteworthy that, even in situations where representation collapse is avoided, the overall performance remains suboptimal. The introduction of additional parametrization through MLP layers has only yielded a marginal impact on improving performance.

An argument can be made that creating effective sentence embeddings presents a more formidable challenge when non-contrastive objectives are utilized, especially in comparison to tasks related to vision. In contrastive learning, the approach involves actively pushing data samples apart to improve discrimination. However, in sentence embeddings with non-contrastive objectives, this process becomes implicit.

To draw a parallel, envision a scenario where each sample is assigned a distinct label, yet some labels are shared among the samples. Similarly, when training a sentence embedding model with non-contrastive objectives, it reflects this intricate situation. We utilize a dataset consisting of randomly sampled Wikipedia sentences collected in SimCSE (Gao et al., 2021b). While each sentence in the dataset may possess unique content, there exist underlying semantic or syntactic relationships among them, akin to the shared labels in the prob-

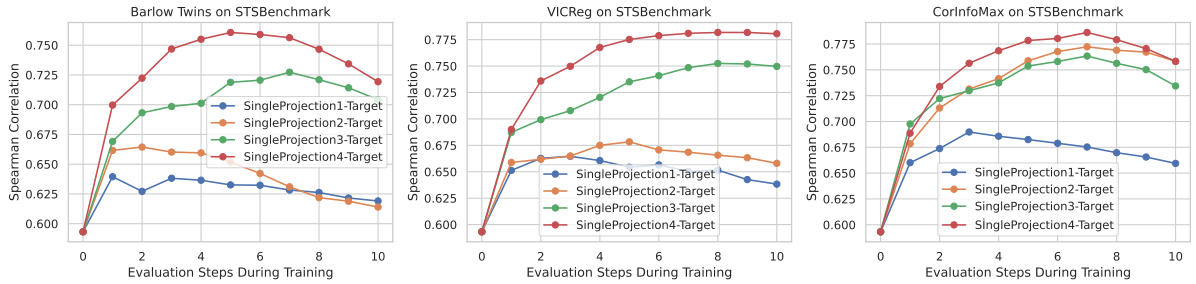

Figure 5: The performance of various non-contrastive objectives on STSBenchmark (Cer et al., 2017) in the Single Projection Model with slightly optimized hyperparameters. The difference between models is the number of MLP layers. MLP layer is adopted from BSL (Zhang et al., 2021b).

lem we are considering. The inherent complexity and the necessity to implicitly capture these relationships contribute to the intricacy of the sentence embedding task when utilizing non-contrastive objectives.

4.3 Single Projection Model

In our *Online Projection Model*, it is crucial to emphasize the significant contribution of MLP layers for both target and online embeddings. Importantly, the sentence embeddings themselves are initially obtained from the BERT model.

The MLP layers should not be viewed as static components in our model architecture; instead, they play a dynamic and transient role during the training phase. Their function is crucial in continually shaping the embeddings for effective loss minimization. However, it is important to emphasize that the outputs produced by these MLP layers do not represent the definitive embeddings used for subsequent evaluation.

This leads us to an intriguing hypothesis: What if we were to consider avoiding the involvement of 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 aim to explore whether such architectural simplification could yield substantial advantages. This modification holds the potential to provide insights into whether a more simplified approach might enhance both the efficiency of loss minimization and the quality of the resultant embeddings, thereby refining the overall training process.

The outcomes presented in Figure 4 closely align with our hypothesis. Throughout the training process, the models consistently showcased incremental performance improvements, surpassing the accomplishments of the preceding model while maintaining identical complexities and hyperparameters. While these results are undeniably promising, it is crucial to acknowledge that they have not yet reached the performance level observed in SimCSE. This suggests that additional optimization endeavors are necessary to narrow the gap and enable our models to attain the performance parity with their SimCSE counterparts. Hence, there is ample room for refinement and enhancement in our pursuit of achieving comparable or even superior performance.

We have significantly improved our model's performance by making relatively minor adjustments to specific hyperparameters, with a particular focus on the learning rate, batch size, and sequence length. The optimal hyperparameters are set to 1e-4, 32, and 64, respectively. The decay rate is maintained at 0.999 consistently across all experiments. Remarkably, these subtle modifications have enabled us to achieve the highest attainable scores among non-contrastive objectives, all without delving into the optimization of hyperparameters within the loss objective. It's important to note that we intentionally adhered 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, demonstrating their effective performance even when applied in contexts beyond their original domain.

The outcomes shown in Figure 5 do not signify the peak of our accomplishments. We have obtained superior results by increasing the frequency of evaluations (20 evaluations per run) throughout the training process and introducing a checkpointing system to preserve the best-performing model. These particular runs were crafted to be consistent

Num. Datasets (→)	Class. 12	Clust. 11	PairClass. 3	Rerank. 4	Retr. 15	STS 10	Summ. 1	Avg. 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)

with our earlier experiments, intending to showcase the effectiveness of the implemented adjustments.

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.

While MTEB aims to encompass a wide range of applications for sentence embeddings, there are noticeable score discrepancies within UNSEE models. Despite their shared objective of optimizing feature decorrelation, implicit in the case of BYOL, differences in their problem formulations lead to variations in scores across different subtasks. For instance, UNSEE-Barlow excels significantly in clustering compared to other objectives. One could argue that the exclusive focus of Barlow Twins on minimizing feature decorrelation might make it more effective in information dissemination, resulting in superior clustering. However, VICReg's incorporation of variance and invariance aspects may pose challenges in achieving the same level of clustering performance. Another question arises regarding why this performance difference doesn't extend to retrieval. One possible explanation is that retrieval requires a finer-grained spread within a subspace, a quality that other objectives (excluding Barlow Twins) may achieve due to their invariance objective.

Nonetheless, 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.

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.

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.

Training Details

We implement UNSEE with *SentenceTransformers* from (Reimers and Gurevych, 2019). Our code is available at GitHub. 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.

Computational Requirements

We only use Tesla T4 GPUs for our experiments.

Acknowledgements

We are grateful to Alper Erdogan, and Deniz Yuret for advising the project initially and hereby thank KUIS AI for providing computing resources for our project.

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.