SCD: Self-Contrastive Decorrelation for Sentence Embeddings

Tassilo Klein
SAP AI Research
tassilo.klein@sap.com

Moin Nabi
SAP AI Research
m.nabi@sap.com

Abstract

In this paper, we propose Self-Contrastive Decorrelation (SCD), a self-supervised approach. Given an input sentence, it optimizes a joint self-contrastive and decorrelation objective. Learning a representation is facilitated by leveraging the contrast arising from the instantiation of standard dropout at different rates. The proposed method is conceptually simple yet empirically powerful. It achieves comparable results with state-of-the-art methods on multiple benchmarks without using contrastive pairs. This study opens up avenues for efficient self-supervised learning methods that are more robust than current contrastive methods.[1]

1 Introduction

Unsupervised learning of representation (a.k.a. embedding) is a fundamental problem in NLP and has been studied extensively in the literature (Mikolov et al., 2013; Pennington et al., 2014; McCann et al., 2017; Peters et al., 2018). Sentence embeddings are essential for numerous language processing applications, such as machine translation, sentiment analysis, information retrieval, and semantic search. Recently, self-supervised pre-training schemes have been successfully used in the context of transformer architectures, leading to a paradigm shift in natural language processing and understanding (Devlin et al., 2018; Liu et al., 2019; Radford et al., 2018) The idea here is to employ an auxiliary task, which enforces an additional objective during training. Typically, this entails predictions based on a subset of information from the context. Most objectives found effective in practice are quite simple. Some successful examples of such pretext tasks are Masked Language Model (MLM), Next Sentence Prediction (NSP), Sentence Order Prediction (SOP), etc. (Devlin et al., 2019; Liu et al., 2019; Lan et al., 2019). When working with unlabeled data, contrastive learning is among the most powerful approaches in self-supervised learning. The goal of contrastive representation learning is to learn an embedding space in such a manner that similar sample pairs (i.e., *positive pairs*) stay close to each other. Simultaneously, dissimilar sample pairs (i.e., *negative pairs*) are far pushed apart. To this end, different augmented views of the same sample and the augmented views from different samples are used as positive and negative pairs. These methods have shown impressive results over a wide variety of tasks from visual to textual representation learning (Chen et al., 2020a,b; Gao et al., 2021; Grill et al., 2020; Chen and He, 2021).

Different techniques have been proposed for the augmentation and selection of positive and negative pairs. For example, DeCLUTR (Giorgi et al., 2021) proposes to take different spans from the same document as positive pairs, while CT (Carlsson et al., 2020) aligns embeddings of the same sentence from two different encoders. CERT (Fang et al., 2020) applies the back-translation to create augmentations of original sentences, and IS-BERT (Zhang et al., 2020) maximizes the agreement between global and local features. Finally, CLEAR (Wu et al., 2020) employs multiple sentence-level augmentation strategies to learn a sentence representation. Despite the simplicity of these methods, they require careful treatment of negative pairs, relying on large batch sizes (Chen et al., 2020a) or sophisticated memory strategies. These include memory banks (Chen et al., 2020b; He et al., 2020) or customized mining strategies (Klein and Nabi, 2020) to retrieve negative pairs efficiently. In NLP specifically, the endeavor of "hard negative mining" becomes particularly challenging in the unsupervised scenario. Increasing training batch size or the memory bank size implicitly introduces more hard negative samples, coming along with the heavy burden of large memory requirements.

[1] Source code and pre-trained models are available at: https://github.com/SAP-samples/acl2022-self-contrastive-decorrelation/

In this paper, we introduce SCD, a novel algorithm for self-supervised learning of sentence embedding. SCD achieves comparable performance in terms of sentence similarity-based tasks compared with state-of-the-art contrastive methods *without*, e.g., employing explicit contrastive pairs. Rather, in order to learn sentence representations, the proposed approach leverages the *self*-contrast imposed on the augmentations of a *single* sample. In this regard, the approach builds upon the idea that sufficiently strong perturbation of the sentence embedding reflects the semantic variations of the sentence. However, it is unclear which perturbation is simply a slight variation of the sentence without changing the semantic (positive pair) and which perturbation sufficiently modifies the semantic to create a negative sample. Such ambiguity manifests itself in the augmented sample sharing the characteristics of both negative and positive samples. To accommodate this, we propose an objective function consisting of two opposing terms, which acts on augmentations pairs of a sample: **i)** self-contrastive divergence (*repulsion*), and **ii)** feature decorrelation (*attraction*). The first term treats the two augmentations as a negative pair pushing apart the different views. In contrast to that, the second term attends to the augmentations as a positive pair. Thus, it maximizes the correlation of the same feature across the views, learning invariance w.r.t. the augmentation. Given the opposing nature of the objectives, integrating them in a joint loss yields a min-max optimization scheme. The proposed approach avoids degenerated embeddings by framing the representation learning objective as an attraction-repulsion trade-off. Simultaneously, it learns to improve the semantic expressiveness of the representation. Due to the difficulty of augmentation in NLP, the proposed approach generates augmentation "on-the-fly" for each sample in the batch. To this end, multiple augmentations are produced by *varying* dropout rates for each sample. We empirically observed that SCD is more robust to the choice of augmentations than pairwise contrastive methods; we believe that not relying on contrastive pairs is one of the main reasons for this, an observation also made in self-supervised learning literature such as BYOL (Grill et al., 2020). While other methods take different augmentation or different copies of models, we utilized the different outputs of the same sentence from standard dropout.

Most related to our paper is (Gao et al., 2021), which considers using dropout as data augmentation in the context of contrastive learning. A key novelty of our approach is that we use the dropout for creating the *self*-contrastive pairs, which can be utilized as *both* positive and negative. At last, we note that our model is different from the *pairwise* feature decorrelation or whitening in (Zbontar et al., 2021; Su et al., 2021; Ermolov et al., 2021), which encourage similar representations between augmented views of a sample while minimizing the redundancy within the representation vector. A key difference compared to these methods is that they ignore the contrastive objective completely. In contrast, our method takes it into account and provides the means to treat self-contrastive views as positive and negative pairs simultaneously.

Our contribution: i) generation of sentence embeddings by leverage multi-dropout **ii)** elimination of reliance on negative pairs using self-contrast, **iii)** proposing feature decorrelation objective for non-contrastive self-supervised learning in NLP.

2 Method

Our approach relies on the generation of two views A and B of samples. To this end, augmentations are generated in embedding space for each sample x_i in batch X. Batches are created from samples of set $\mathcal{D} = \{(x_i)\}_{i=1}^{N}$, where N denotes the number of sample (sentences). Augmentations are produced by an encoder f_θ, parametrized by θ. The output of the encoder is the embeddings of samples in X denoted as $H^A \in \mathcal{T}$ and $H^B \in \mathcal{T}$. Here $\mathcal{T}$ denotes the embedding space. Next, we let, $\boldsymbol{h}_i \in \mathcal{T}$ denote the associated representation of the sentence. The augmentation embeddings produced per sample are then denoted $\boldsymbol{h}_i^A$ and $\boldsymbol{h}_i^B$. To obtain the different embedding, we leverage a transformer language model as an encoder in combination with *varying* dropout rates. Specifically, one augmentation is generated with *high* dropout and one with *low* dropout. This entails employing different random masks during the encoding phase. The random masks are associated with *different* ratios, r_A and r_B, with $r_A < r_B$. Integrating the distinct dropout rates into the encoder, we yield $\boldsymbol{h}_i^A = f_\theta(x_i, r_A)$ and $\boldsymbol{h}_i^B = f_\theta(x_i, r_B)$. Given the embeddings, we leverage a joint loss, consisting of two objectives:

$$\min_{\theta_1,\theta_2} \mathcal{L}_S(f_{\theta_1}) + \alpha\mathcal{L}_C(f_{\theta_1}, p_{\theta_2}) \qquad (1)$$

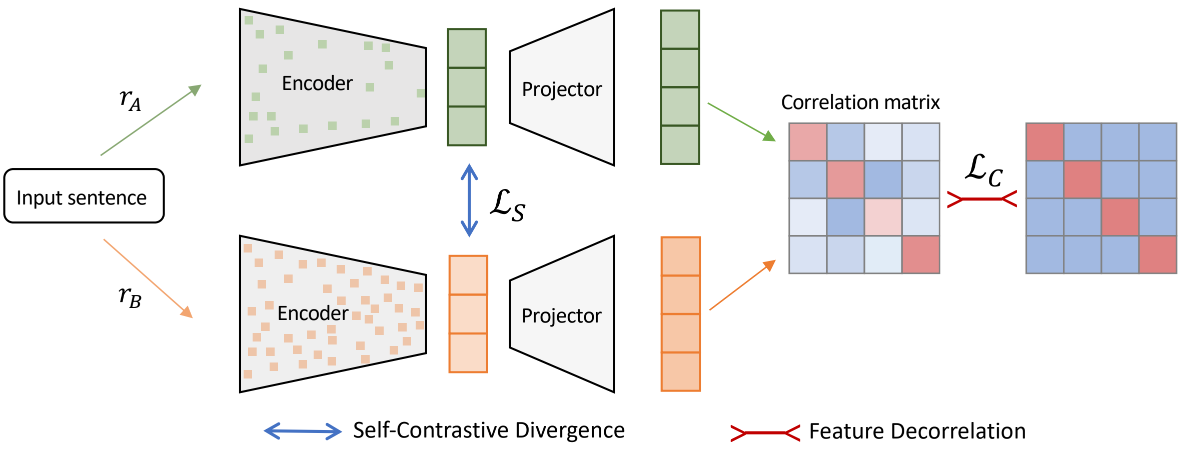

FIGURE 1. Schematic illustration of the proposed approach (best shown in color). Starting from an input sentence **(left)**, two embeddings are produced by varying the dropout-rate in the encoder. Patches within the encoder indicate masking due to dropout. Different dropout rates and resulting embeddings color-coded: **low dropout**, **high dropout**. Self-contrastive loss is imposed on the embeddings **(center)**. A projector maps embeddings to to a high-dimensional feature space, where the features are decorrelated **(right)**.

Here $\alpha \in \mathbb{R}$ denotes a hyperparameter and $p : \mathcal{T} \to \mathcal{P}$ is a projector (MLP) parameterized by θ_2, which maps the embedding to $\mathcal{P}$, with $|\mathcal{P}| \gg |\mathcal{T}|$. The objective of $\mathcal{L}_S$ is to increase the contrast of the augmented embedding, pushing apart the embeddings $\boldsymbol{h}_i^A$ and $\boldsymbol{h}_i^B$. The objective of $\mathcal{L}_C$ is to reduce the redundancy and promote invariance w.r.t. augmentation in a high-dimensional space $\mathcal{P}$. See Fig. 1 for a schematic illustration of the method.

2.1 Self-Contrastive Divergence:

Self-contrast seeks to create a contrast between the embeddings arising from different dropouts. Hence, $\mathcal{L}_S$ consists of the cosine similarity of the samples in the batch as:

$$\mathcal{L}_S = \frac{1}{N}\sum_i^N \boldsymbol{h}_i^A \cdot (\boldsymbol{h}_i^B)^T \left(\|\boldsymbol{h}_i^A\|\|\boldsymbol{h}_i^B\|\right)^{-1} \quad (2)$$

2.2 Feature Decorrelation:

$\mathcal{L}_C$ seeks to make the embeddings invariant to augmentation while at the same time reducing the redundancy in feature representation. To this end, the embedding $\boldsymbol{h}_i$ is projected up from $\mathcal{T}$ to a high-dimensional space $\mathcal{P}$, where decorrelation is performed. To avoid clutter in notation, we let $p_i^* = p(h_i^*)$ and $* \in \{A, B\}$, denote the augmented embedding vectors of sample x_i after applying a projection with $p(.)$. Then, a correlation matrix is computed from the projected embeddings. Its entries $C_{j,k}$ are:

$$C_{jk} = \sum_i p_{i,j}^A \cdot p_{i,k}^B \left(\sum_i (p_{i,j}^A)^2 (p_{i,k}^B)^2\right)^{-\frac{1}{2}} \quad (3)$$

Here, $p_{i,j}^* \in \mathbb{R}$ denotes the j^{th} component in the projected embedding vector. Then the loss objective for feature decorrelation is defined as:

$$\mathcal{L}_C = -\sum_j (1 - C_{jj})^2 + \lambda \sum_j \sum_{j \neq k} C_{jk}^2 \quad (4)$$

The first term seeks to achieve augmentation invariance by maximization of the cross-correlation along the diagonal. The second term seeks to reduce redundancy in feature representation by minimizing correlation beyond the diagonal. Given that these objectives are opposing, $\lambda \in \mathbb{R}$ is a hyperparameter, controlling the trade-off.

3 Experiments & Results

3.1 Training Setup:

Training is started from a pre-trained transformer LM. Specifically, we employ the Hugging Face (Wolf et al., 2020) implementation of BERT and RoBERTa. For sentence representation, we take the embedding of the `[CLS]` token. Then similar to (Gao et al., 2021), we train the model in an unsupervised fashion on 10^6 randomly samples sentences from Wikipedia. The LM is trained with a learning rate of $3.0e{-}5$ for 1 epoch at batch-size of 192. The projector MLP q has three linear layers, each with 4096 output units in conjunction with ReLU and BatchNorm in between. For BERT hyperparameters are $\alpha = 0.005$, $\lambda = 0.013$, and dropout rates are $r_A = 5.0\%$ and $r_B = 15.0\%$. For RoBERTa hyperparameters are $\alpha = 0.0033$, $\lambda = 0.028$, and dropout rates are $r_A = 6.5\%$ and

Semantic Textual Similarity (STS) Benchmark

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
GloVe embeddings (avg.)♣	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
BERT$_{\texttt{base}}$ [CLS]-embedding	21.54	32.11	21.28	37.89	44.24	20.29	42.42	31.40
BERT$_{\texttt{base}}$(first-last avg)♢	39.70	59.38	49.67	66.03	66.19	53.87	62.06	56.70
BERT$_{\texttt{base}}$-flow♢	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
BERT$_{\texttt{base}}$-whitening♢	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
IS-BERT$_{\texttt{base}}$ ♡	56.77	69.24	61.21	75.23	70.16	69.21	64.25	66.58
CT-BERT$_{\texttt{base}}$ ♢	61.63	76.80	68.47	77.50	76.48	74.31	69.19	72.05
SimCSE-BERT$_{\texttt{base}}$	**68.05**	**80.38**	**72.62**	**78.96**	**76.90**	75.11	69.37	**74.48**
$*$ SCD-BERT$_{\texttt{base}}$	66.94	78.03	69.89	78.73	76.23	**76.30**	**73.18**	74.19
RoBERTa$_{\texttt{base}}$ [CLS]-embedding	16.67	45.56	30.36	55.08	56.99	38.82	61.89	43.62
RoBERTa$_{\texttt{base}}$(first-last avg)♢	40.88	58.74	49.07	65.63	61.48	58.55	61.63	56.57
SimCSE-RoBERTa$_{\texttt{base}}$	**67.05**	**80.01**	**70.93**	79.66	**80.06**	**78.38**	68.30	**74.91**
$*$ SCD-RoBERTa$_{\texttt{base}}$	63.53	77.79	69.79	**80.21**	77.29	76.55	**72.10**	73.89

TABLE 1. Sentence embedding performance on STS tasks measured as Spearman's correlation. ♣: results from Reimers and Gurevych (2019); ♡: results from Zhang et al. (2020); ♢ results from Gao et al. (2021); other results are by ourselves. Dashed line (- -), separates BERT (upper part) and RoBERTa (lower part) language models.

$r_B = 24.0\%$. The values were obtained by grid-search. First a coarse-grid was put in place with a step-size of 0.1 for α, 10% for the dropout rates r_A, r_B. For λ the coarse-grid consisted of different magnitudes $\{0.1, 0.01, 0.001\}$. Second, on a fine-grid with step-size of 0.01 and 1%, respectively.

3.2 Evaluation Setup:

Experiments are conducted on 7 standard semantic textual similarity (STS) tasks. In addition to that, we also evaluate on 7 transfer tasks. Specifically, we employ the SentEval toolkit (Conneau and Kiela, 2018) for evaluation. As proposed by (Reimers and Gurevych, 2019; Gao et al., 2021), we take STS results as the main comparison of sentence embedding methods and transfer task results for reference. For the sake of comparability, we follow the evaluation protocol of (Gao et al., 2021), employing Spearman's rank correlation and aggregation on all the topic subsets.

3.3 Main Results

3.3.1 Semantic Textual Similarity:

We evaluate on 7 STS tasks: (Agirre et al., 2012, 2013, 2014, 2015, 2016), STS Benchmark (Cer et al., 2017) and SICK-Relatedness (Marelli et al., 2014). These datasets come in sentence pairs together with correlation labels in the range of 0 and 5, indicating the semantic relatedness of the pairs. Results for the sentence similarity experiment can be seen in Tab. 1. The proposed approach is on-par with state-of-the-art approaches. Using BERT-LM, we outperform the next-best approach on STS-B (**+1.19**) and on SICK-R (**+3.81**) points. Using RoBERTa-LM, we outperform the next best comparable approach (SimCSE-RoBERTA$_{\texttt{base}}$) on STS-15 (**+0.55%**) and SICK-R (**+3.8%**).

3.3.2 Transfer task:

We evaluate our models on the following transfer tasks: MR (Pang and Lee, 2005), CR (Hu and Liu, 2004), SUBJ (Pang and Lee, 2004), MPQA (Wiebe et al., 2005), SST-2 (Socher et al., 2013), TREC (Voorhees and Tice, 2000) and MRPC (Dolan and Brockett, 2005). To this end, a logistic regression classifier is trained on top of (frozen) sentence embeddings produced by different methods. We follow default configurations from SentEval. Results for the transfer task experiment can be seen in Tab. 2. SCD is on-par

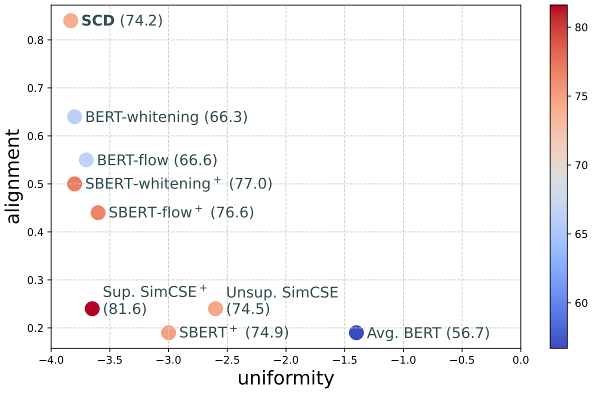

FIGURE 2. Quantitative analysis of embeddings - *alignment* vs. *uniformity* (the smaller, the better). Points represent average STS performance using BERT$_{\texttt{base}}$, with Spearman's correlation color coded (+ corresponds to supervised methods).

Transfer Benchmark								
Model	**MR**	**CR**	**SUBJ**	**MPQA**	**SST**	**TREC**	**MRPC**	**Avg.**
GloVe embeddings (avg.)$^{\clubsuit}$	77.25	78.30	91.17	87.85	80.18	83.00	72.87	81.52
Skip-thought$^{\heartsuit}$	76.50	80.10	93.60	87.10	82.00	92.20	73.00	83.50
Avg. BERT embeddings$^{\clubsuit}$	78.66	86.25	94.37	88.66	84.40	**92.80**	69.54	84.94
BERT `[CLS]`-embedding	**81.83**	**87.39**	95.48	88.21	**86.49**	91.00	72.29	**86.10**
IS-BERT$_{\texttt{base}}^{\heartsuit}$	81.09	87.18	94.96	**88.75**	85.96	88.64	74.24	85.83
SimCSE-BERT$_{\texttt{base}}$	80.74	85.75	93.96	88.60	84.57	86.20	73.51	84.76
$*$ SCD-BERT$_{\texttt{base}}$	73.21	85.80	**99.56**	88.67	85.89	89.80	**75.71**	85.52
RoBERTa `[CLS]`-embedding	81.27	84.77	94.15	84.18	86.71	81.20	72.17	83.49
SimCSE-RoBERTa$_{\texttt{base}}$	65.00	87.28	**99.60**	**86.63**	87.26	80.80	72.23	82.69
$*$ SCD-RoBERTa$_{\texttt{base}}$	**82.17**	**87.76**	93.67	85.69	**88.19**	**83.40**	**76.23**	**85.30**

TABLE 2. Transfer task result measured as accuracy. ♣: results from Reimers and Gurevych (2019); ♡: results from Zhang et al. (2020); ◇ results from Gao et al. (2021); other results are by ourselves. Dashed line (- -), separates BERT (upper part) and RoBERTa (lower part) language models.

with state-of-the-art approaches. Using BERT-LM, we outperform the next best approach on SUBJ (**+4.6%**) and MRPC (**+2.2%**). Using RoBERTa-LM, we outperform the next best comparable approach (SimCSE-RoBERTA$_{\texttt{base}}$) on almost all benchmarks, with an average margin of (**+2.61%**).

3.4 Analysis

3.4.1 Ablation Study:

We evaluated each component's performance by removing them individually from the loss to assess both loss terms' contributions. It should be noted that $\mathcal{L}_S$ of Eq. 2 and $\mathcal{L}_C$ of Eq. 4 both interact in a competitive fashion. Hence, only the equilibrium of these terms yields an optimal solution. Changes - such as eliminating a term - have detrimental effects, as they prevent achieving such an equilibrium, resulting in a significant drop in performance. See Tab. 3 for the ablation study on multiple benchmarks. Best performance is achieved in the presence of all loss terms.

3.4.2 Uniformity and Alignment Analysis:

To better understand the strong performance of SCD, we borrow the analysis tool from (Wang and Isola, 2020), which takes *alignment* between semantically-related positive pairs and *uniformity* of the whole representation space to measure the quality of learned embeddings. Figure 2 shows *uniformity* and *alignment* of different methods and their results on the STS. SCD achieves the best in terms of *uniformity*, reaching to the supervised counterparts (**-3.83**), which can be related to the strong effect of the self-contrastive divergence objective. It shows the *self*-contrastive pairs can effectively compensate for the absence of contrastive pairs. In terms of *alignment*, SCD is inferior to other counterparts (**0.84**), which can be attributed to the fact that our repulsion objective mainly focuses on the feature decorrelation aiming to learn a more effective and efficient representation. This is reflected in the final results on the STS where SCD obtains significantly higher correlation even compared to the method with lower *alignment* such as BERT-whitening or BERT-flow.

Method	**STS**	**STS-B**	**SICK-R**
BERT-$_{\texttt{base}}$	31.41	20.29	42.42
SCD ($\mathcal{L}_S$)	35.70	23.59	49.88
SCD ($\mathcal{L}_C$)	66.48	67.57	67.97
SCD ($\mathcal{L}_S$+$\mathcal{L}_C$)	**73.96**	**76.30**	**73.18**

TABLE 3. Ablation study, performance in average Spearman correlation on Semantic Texual Similarity task. **STS** denotes the average of STS12 to STS16.

4 Conclusion & Future Work

We proposed a self-supervised representation learning approach, which leverages the self-contrast of augmented samples obtained by dropout. Despite its simplicity, it achieves comparable results with state-of-the-arts on multiple benchmarks. Future work will deal with sample-specific augmentation to improve the embeddings and, particularly, the representation alignment.

Acknowledgement: We would like to thank Mahdyar Ravanbakhsh for valuable feedback on the manuscript.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task (SemEval 2012)*, pages 385–393.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2020. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations*.

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.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. pages 1597–1607.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations.

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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. 2021. Whitening for self-supervised representation learning. In *International Conference on Machine Learning*, pages 3015–3024. PMLR.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*.

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

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep contrastive learning for unsupervised textual representations. 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 879–895, Online. Association for Computational Linguistics.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *ACM SIGKDD international conference on Knowledge discovery and data mining*.

Tassilo Klein and Moin Nabi. 2020. Contrastive self-supervised learning for commonsense reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7517–7523, Online. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

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 preprint arXiv:1907.11692*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. pages 216–223.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. *arXiv preprint arXiv:1708.00107*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. pages 271–278.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. pages 115–124.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. pages 1631–1642.

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

Ellen M Voorhees and Dawn M Tice. 2000. Building a question answering test collection. In *the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 200–207.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. 2021. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. pages 1601–1610.

SCD: Self-Contrastive Decorrelation for Sentence Embeddings

Klein, Tassilo; Nabi, Moin

01	guimin chen 24/4/2022 17:06	Page 1
02	guimin chen 24/4/2022 17:06	Page 1
03	guimin chen 24/4/2022 17:06	Page 1
04	guimin chen 24/4/2022 17:06	Page 1
05	guimin chen 24/4/2022 17:06	Page 1
06	guimin chen 22/6/2022 2:25	Page 2
07	guimin chen 22/6/2022 2:25	Page 2