SNCSE: Contrastive Learning for Unsupervised Sentence Embedding with Soft Negative Samples

Hao Wang[†,‡], Yangguang Li[‡], Zhen Huang[†], Yong Dou[†], Lingpeng Kong[§], Jing Shao[‡]
[†] National University of Defense Technology
[‡] SenseTime
[§] The University of Hong Kong
{wanghao2, liyangguang}@sensetime.com

Abstract

Unsupervised sentence embedding aims to obtain the most appropriate embedding for a sentence to reflect its semantic. Contrastive learning has been attracting developing attention. For a sentence, current models utilize diverse data augmentation methods to generate positive samples, while consider other independent sentences as negative samples. Then they adopt InfoNCE loss to pull the embeddings of positive pairs gathered, and push those of negative pairs scattered. Although these models have made great progress on sentence embedding, we argue that they may suffer from *feature suppression*. The models fail to distinguish and decouple textual similarity and semantic similarity. And they may overestimate the semantic similarity of any pairs with similar textual regardless of the actual semantic difference between them. This is because positive pairs in unsupervised contrastive learning come with similar and even the same textual through data augmentation. To alleviate feature suppression, we propose contrastive learning for unsupervised sentence embedding with soft negative samples (SNCSE). Soft negative samples share highly similar textual but have surely and apparently different semantic with the original samples. Specifically, we take the negation of original sentences as soft negative samples, and propose Bidirectional Margin Loss (BML) to introduce them into traditional contrastive learning framework, which merely involves positive and negative samples. Our experimental results show that SNCSE can obtain state-of-the-art performance on semantic textual similarity (STS) task with average Spearman's correlation coefficient of 78.97% on $BERT_{base}$ and 79.23% on RoBERTa$_{base}$. Besides, we adopt rank-based error analysis method to detect the weakness of SNCSE for future study.

1 Introduction

Pretrained language models, e.g. BERT [Devlin *et al.*, 2019] and RoBERTa [Liu *et al.*, 2019], have shown excellent performance on several natural language processing tasks with downstream fine-tuning. However, it seems that they may fail to product appropriate sentence embedding to reflect the semantic [Reimers and Gurevych, 2019; Li *et al.*, 2020; Yan *et al.*, 2021]. Recently, contrastive learning has been attracting developing attention for unsupervised sentence embedding [Yan *et al.*, 2021; Gao *et al.*, 2021; Wu *et al.*, 2021]. Contrastive learning firstly comes from computer vision field [He *et al.*, 2020; Chen *et al.*, 2020a; Chen *et al.*, 2020b]. The main idea is to utilize diverse data augmentation methods to generate positive pairs, while take two independent samples as negative pairs, and then adopt InfoNCE loss [Oord *et al.*, 2018] to pull the embeddings of positive pairs gathered, and push those of negative pairs scattered.
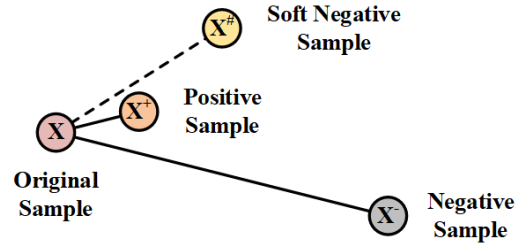
(a) Spatial positions of different samples.

X : Tom and Jerry became good friends.
X⁺: Tom and Jerry became good friends.
X⁻: It has cast stone lintels and window sills.
X#: Tom and Jerry did not become good friends.

(b) Textual sentences of different samples.

Figure 1: Spatial positions and textual sentences of different samples. Previous unsupervised models aim to pull positive pairs gathered, and push negative pairs scattered. The positive sample may share the same textual with the original one with different embedding through cutoff [Yan *et al.*, 2021] or dropout [Gao *et al.*, 2021], while the negative one is totally another independent sentence. SNCSE tries to take the negation of original sentences as soft negative samples, and introduces them into traditional contrastive learning framework through BML loss.

Unsupervised sentence embedding takes similar strategy. As shown in Figure 1, positive pairs may share the same tex-
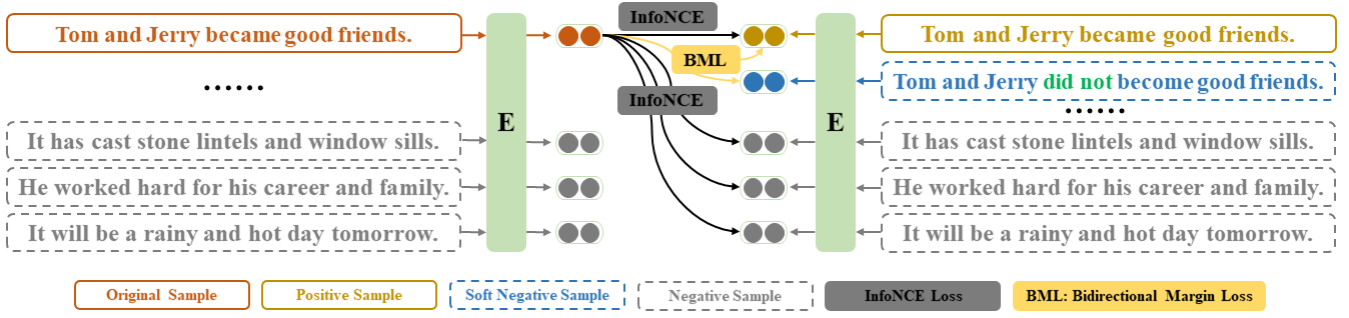
Figure 2: SNCSE framework. InfoNCE loss is utilized to distinguish positive samples from negative samples, while bidirectional margin loss is utilized to distinguish positive samples from soft negative samples. SNCSE does not contrast soft negative samples and negative samples due to the unsure relationship between these two.

tual but have different embeddings through cutoff [Yan *et al.*, 2021] or dropout [Gao *et al.*, 2021], while negative pairs will come from two independent sentences which share less textual. In this condition, we argue that this will lead to **feature suppression**, which has been discussed deeply in vision field [Robinson *et al.*, 2021; Chen *et al.*, 2021]. That is, the models fail to distinguish and decouple textual similarity and semantic similarity. As a result, they may overestimate the semantic similarity of any pairs with similar textual regardless of the actual semantic difference between them. And the models may underestimate the semantic similarity of pairs with less words in common. (Please refer to Section 5 for several instances and detailed analysis.) To **alleviate** feature suppression, we take following issues into consideration:

• whether there exist some **easily available** samples, which have highly similar textual but **surely and apparently** different semantic with the original samples

• how to introduce these samples into the traditional contrastive learning framework to enhance unsupervised sentence embedding, where the traditional framework only involves pure positive and negative samples

We call these samples **soft negative sample**. Here *negative* means they have different semantics with the original samples, while *soft* means they have high textual similarity with the original samples and cannot be merely considered as pure negative samples like other independent sentences. Moreover, we propose contrastive learning for unsupervised sentence embedding with soft negative samples (SNCSE). The framework takes the negation of original sentences as soft negative samples, and proposes bidirectional margin loss (BML) to model the semantic difference between positive samples and soft negative samples, which both have highly similar textual with the original samples. In this way, SNCSE can learn to distinguish and decouple textual similarity and semantic similarity, and thus alleviate feature suppression.

We conduct our experiments on semantic textual similarity (STS) task. Our experimental results show that our SNCSE framework can achieve state-of-the-art performance on the task with diverse pretrained encoders, where average Spearman's correlation coefficient are respectively 78.97% for BERT_{base}, 80.19% for BERT_{large}, 79.23% for RoBERTa_{base} and 81.77% for RoBERTa_{large}. Moreover, our ablation study

results indicate that SNCSE has learned to distinguish a sentence and its negation for the alleviation of feature suppression. Besides, we adopt rank-based error analysis method to detect the weakness of SNCSE for future study. Our code and models[1] have been released for the development of the field.

2 SNCSE Framework

As shown in Figure 2, in SNCSE framework, InfoNCE loss is utilized to distinguish positive samples from negative samples, while bidirectional margin loss is utilized to distinguish positive samples from soft negative samples. SNCSE does not contrast soft negative samples and negative samples, since it is difficult to clarify the relationship between them.

In SNCSE, we take the negation of the original sentences as soft negative samples. Specifically, the negated sentences are generated through a rule-based method. Specifically, we utilize Spacy tools[2] to parse the sentence to obtain the dependency syntax tree of the sentence, and the pos tag as well as the stem of each token. Then according to this information, we convert the sentences to its negation with correct syntax and clear semantic. We mainly focus on explicit negation with negation words, which converts the original sample to *"Tom and Jerry did not become good friends"*. However, we do not consider implicit negation based on antonym, which may convert the original sample to *"Tom and Jerry became bad friends"*.

Formally, in SNCSE, for a sentence, let X, X^+, $X^\#$ respectively denote the original sample, positive sample and soft negative sample. Here, X and X^+ share the same textual. Inspired by recent studies on prompt learning [Ding *et al.*, 2021; Yao *et al.*, 2021; Liu *et al.*, 2021], the three samples are encoded with pretrained language models, e.g. BERT and RoBERTa, within different prompts[3]:

$$
\begin{aligned}
&The\ sentence\ :\ "\ X\ "\ means\ [MASK].\\
&The\ sentence\ of\ "\ X^+\ "\ means\ [MASK].\qquad(1)\\
&The\ sentence\ of\ "\ X^\#\ "\ means\ [MASK].
\end{aligned}
$$

---

[1]https://github.com/Sense-GVT/SNCSE

[2]https://spacy.io/

[3]We are also inspired by PromptBERT, which has been under review but withdrawn on OpenReview.net.

We take the hidden state of the special [MASK] token $h_{[MASK]}$ as the embedding h of the corresponding sample. During training, like SimCSE [Gao *et al.*, 2021], we add an MLP layer with nonlinear activation function on $h_{[MASK]}$ to obtain h:

$$h = Tanh(MLP(h_{[MASK]})) \quad (2)$$

InfoNCE loss is utilized to distinguish positive samples from negative samples:

$$\mathcal{L}_{InfoNCE} = -log \frac{e^{cos\_sim(h_i,h_i^+)/\tau}}{\sum\limits_{j=1}^{N} e^{cos\_sim(h_i,h_j^+)/\tau}} \quad (3)$$

where *cos_sim* means cosine similarity, τ means temperature factor, while N means batch size.

Then we calculate the cosine similarity difference Δ between positive pairs and soft negative pairs:

$$\Delta = cos\_sim(h_i, h_i^{\#}) - cos\_sim(h_i, h_i^+) \quad (4)$$

and propose bidirectional margin loss (BML) to model the semantic similarity difference:

$$\mathcal{L}_{BML} = ReLU(\Delta + \alpha) + ReLU(-\Delta - \beta) \quad (5)$$

BML loss aims to constraint Δ within this interval:

$$\Delta \in [-\beta, -\alpha] \quad (6)$$

where α and β respectively means the bottom and upper semantic similarity difference between positive pairs and soft negative pairs.

Finally, the total loss is the weighted summary of InfoNCE loss and bidirectional margin loss:

$$\mathcal{L}_{SNCSE} = \mathcal{L}_{InfoNCE} + \lambda \mathcal{L}_{BML} \quad (7)$$

where the weight λ reflects the importance of soft negative samples, and also adjusts the form difference of these two loss functions.

For comparison, we also consider to ignore soft negative samples, and merely take the negations as pure positive samples, or pure negative samples. The corresponding positive loss *PL* and negative loss *NL* are respectively:

$$\mathcal{L}_{PL} = -log \frac{e^{cos\_sim(h_i,h_i^+)/\tau}}{\sum\limits_{j=1}^{N} e^{cos\_sim(h_i,h_j^+)/\tau}} - log \frac{e^{cos\_sim(h_i,h_i^{\#})/\tau}}{\sum\limits_{j=1}^{N} e^{cos\_sim(h_i,h_j^{\#})/\tau}} \quad (8)$$

$$\mathcal{L}_{NL} = -log \frac{e^{cos\_sim(h_i,h_i^+)/\tau}}{e^{cos\_sim(h_i,h_i^{\#})/\tau} + \sum\limits_{j=1}^{N} e^{cos\_sim(h_i,h_j^+)/\tau}} \quad (9)$$

3 Experiments

3.1 Evaluation Metric

Following previous works [Yan *et al.*, 2021; Gao *et al.*, 2021; Wu *et al.*, 2021], we evaluate SNCSE on semantic textual

similarity (STS) task. STS task aims to measure the semantic similarity of sentence pairs, and is one of the most widely used benchmark datasets for evaluation of unsupervised sentence embedding. The task include 7 subtasks in total, which are respectively STS 12-16, STS-B, and SICK-R [Agirre *et al.*, 2012; Agirre *et al.*, 2013; Agirre *et al.*, 2014; Agirre *et al.*, 2015; Agirre *et al.*, 2016; Cer *et al.*, 2017; Marelli *et al.*, 2014]. Each subtask contains several sentence pairs, and the semantic similarity of a sentence pair is labeled as a number between 0.0 and 5.0, where higher score means higher semantic similarity. The detailed information of these subtasks are summarized in Table 1. During evaluation, we calculate cosine similarity of each pair, and focus on the test sets. For a pair, golden semantic similarity ranges from 0.0 to 5.0, while cosine similarity ranges from -1.0 to 1.0. To deal with the range difference, rank-based Spearman's correlation coefficient is adopted.

Task	STS12	STS13	STS14	STS15
train	0	0	0	0
dev	0	0	0	0
test	3108	1500	3750	3000
Task	STS16	STS-B	SICK-R	
train	0	5749	4500	
dev	0	1500	500	
test	1186	1379	4927	

Table 1: Statistic information of STS task.

Specifically, for a subtask with n sentence pairs, we first sort the golden similarity to obtain the golden rank R_G for each sentence pair. And we also calculate and sort estimated cosine similarity to obtain estimated rank R_E for each pairs. Then, rank-based Spearman's correlation coefficient ρ for the subtask is defined as:

$$\rho = 1 - \frac{\sum\limits_{i=1}^{n} (R_E^i - R_G^i)^2}{\frac{1}{6}(n^3 - n)} \quad (10)$$

Here ρ ranges from -1.0 to 1.0, where higher value means more advanced sentence embedding and semantic comprehension ability of the model. We calculate Spearman's correlation coefficient ρ for each subtask independently and take the average of ρ on all subtasks as main evaluation metric.

Moreover, we conduct rank-based error analysis to detect the weakness of our SNCSE. For each sentence pair, we roughly[4] take the rank difference between R_G and R_E:

$$(R_E^i - R_G^i)^2 \quad (11)$$

as the semantic estimation error of the pair. Then we focus on sentences pairs with the most severe semantic estimation error for further analysis.

3.2 Experiment Setup

Our framework is implemented through Pytorch 1.8.0 and hugging face transformers[5]. All our experiments are conducted on a computation node with Nvidia A100 GPUs. For

---

[4]Here the rank difference of a pair is related to not only the pair itself, but also other pairs of the same subtask.

[5]https://huggingface.co

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
ConSERT-BERT$_{base}$♡	64.64	78.49	69.07	79.72	75.95	73.97	67.31	72.74
SimCSE-BERTe$_{base}$♣	68.40	82.41	74.38	80.91	78.56	76.85	72.23	76.25
ESimCSE-BERT$_{base}$♠	**73.40**	83.27	**77.25**	82.66	78.81	80.17	72.30	78.27
SNCSE-BERT$_{base}$	70.67	**84.79**	76.99	83.69	**80.51**	**81.35**	**74.77**	**78.97**
w.o. BML Loss	71.58	84.68	76.96	**83.78**	79.40	80.46	70.64	78.21
ConSERT-BERT$_{large}$♡	70.69	82.96	74.13	82.78	76.66	77.53	70.37	76.45
SimCSE-BERT$_{large}$♣	70.88	84.16	76.43	84.50	79.76	79.26	73.88	78.41
ESimCSE-BERT$_{large}$♠	**73.21**	85.37	77.73	84.30	78.92	80.73	74.89	79.31
SNCSE-BERT$_{large}$	71.94	86.66	**78.84**	**85.74**	**80.72**	**82.29**	**75.11**	**80.19**
w.o. BML Loss	73.18	**86.74**	78.44	85.66	79.69	81.98	71.69	79.62
SimCSE-RoBERTa$_{base}$♣	70.16	81.77	73.24	81.36	80.65	80.22	68.56	76.57
ESimCSE-RoBERT$_{base}$♠	69.90	82.50	74.68	83.19	80.30	80.99	70.54	77.44
SNCSE-RoBERTa$_{base}$	70.62	**84.42**	77.24	**84.85**	**81.49**	**83.07**	**72.92**	**79.23**
w.o. BML Loss	**71.22**	84.10	76.44	84.35	80.80	81.77	69.82	78.36
SimCSE-RoBERTa$_{large}$♣	72.86	83.99	75.62	84.77	81.80	81.98	71.26	78.90
ESimCSE-RoBERTa$_{large}$♠	73.20	84.93	76.88	84.86	81.21	82.79	72.27	79.45
SNCSE-RoBERTa$_{large}$	73.71	**86.73**	**80.35**	**86.80**	**83.06**	**84.31**	**77.43**	**81.77**
w.o. BML Loss	**73.73**	85.66	79.00	86.34	81.78	83.64	72.10	80.32

Table 2: Sentence embedding performance on semantic textual similarity (STS) test sets, in terms of Spearman's correlation coefficient. ♡ denotes the results from Yan *et al.* [2021], ♣ denotes the results from Gao *et al.* [2021], while ♠ denotes the results from Wu *et al.* [2021]. We also conduct ablation study to verify the influence of BML loss, and *w.o.* means *without*.

all experiments, we adopt AdamW optimizer. For all experiments, we keep a random number seed of 42, and set dropout rate as 0.1, temperature factor τ as 0.05, α as 0.1 and β as 0.3. Besides, we set λ as 1e-3 for BERT encoder and 5e-4 for RoBERTa encoder. For base models, we adopt batch size of 256 and initial learning rate of 1e-5, while for large models, we adopt batch size of 128 and learning rate of 5e-6 to avoid out of memory problem. Following SimCSE [Gao *et al.*, 2021] and ESimCSE [Wu *et al.*, 2021], we adopt 1 million sentences[6] randomly drawn from English Wikipedia as train sentences. And we also train 1 epoch, evaluate every 125 steps and choose model parameters with the highest performance on STSB development set.

3.3 Main Results

Our main experimental results are shown in Table 2. Our SNCSE framework obtains the best performance with the highest average Spearman's correlation coefficient on all encoders, which are respectively 78.97% for BERT$_{base}$, 80.19% for BERT$_{large}$, 79.23% for RoBERTa$_{base}$ and 81.77% for RoBERTa$_{large}$. Compared with ESimCSE [Wu *et al.*, 2021], which is previous state-of-the-art model, SNCSE respectively obtain 0.7%, 0.87%, 1.79% and 2.32% absolute increase on average Spearman's correlation coefficient. Moreover, we also conduct ablation study to detect the influence of BML loss. For the four encoders, with BML loss, our SNCSE respectively obtain 0.76%, 0.57%, 0.87% and 1.45% absolute increase on average Spearman's correlation coefficient. These indicate the strength of SNCSE and BML loss.

---

[6]https://huggingface.co/datasets/princeton-nlp/
datasets-for-simcse/blob/main/wiki1m_for_simcse.txt

4 Ablation Study

4.1 Role of the Negation

For comparison, we try to change the role of the negation with different loss functions. The experimental results with BERT$_{base}$ encoder are shown in Table 3. Traditional InfoNCE loss merely takes positive and negative pairs into consideration and ignore the negation, and obtains average Spearman's correlation coefficient of 78.21%. Positive loss in Equation 8 takes the negation as pure positive samples, and obtains 77.11%, while negative loss in Equation 9 takes the negation as pure negative samples and obtains 74.37%. SNCSE takes the negation as soft negative samples, contrasts them with the positive samples through BML loss, and obtains the highest Spearman's correlation coefficient of 78.97%. Therefore, it can be concluded that SNCSE considers the negation in a more appropriate way.

Loss	PL	NL	InfoNCE	SNCSE
Avg.	77.11	74.37	78.21	**78.97**

Table 3: Influence of the role of the negation. *PL* takes the negation as pure positive samples, *NL* take as pure negative sample, *InfoNCE* ignores the negation, while *SNCSE* takes as soft negative samples.

4.2 Influence of Hyperparameters in BML Loss

In SNCSE, we introduce BML loss with 3 main hyperparameters to constraint the semantic similarity difference between positive pairs and soft negative pairs. λ denotes the importance soft negative samples, and adjusts the form difference between InfoNCE loss and margin loss. α denotes the bottom difference between positive pairs and soft negative pairs, and a higher value requires more apparent difference. β denotes
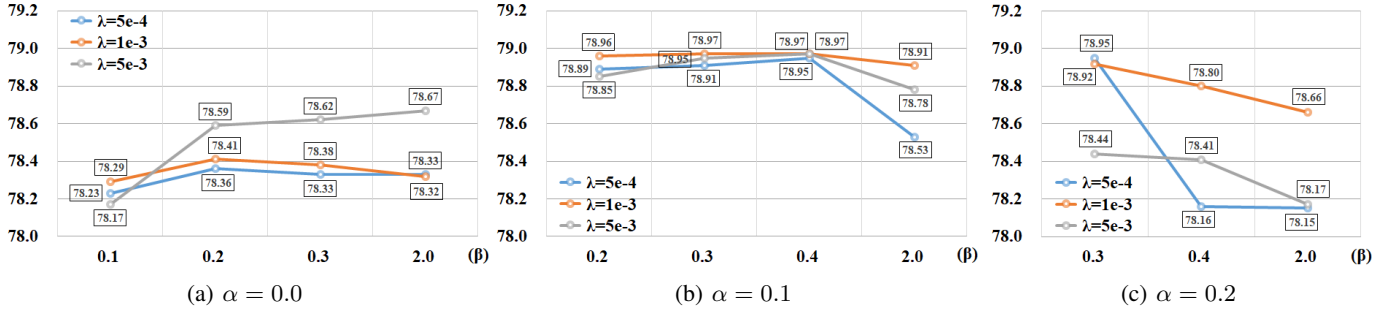
(a) $\alpha = 0.0$ (b) $\alpha = 0.1$ (c) $\alpha = 0.2$

Figure 3: Influence of hyperparameters of BML loss. λ denotes the importance of soft negative sample, and adjusts the form difference between InfoNCE loss and margin loss. α and β denotes the bottom and upper difference between positive pairs and soft negative pairs. The upper difference is ignored with β as 2.0.

the upper difference, and upper difference is ignored with β as 2.0, since cosine similarity ranges from -1.0 to 1.0.

As shown in Figure 3, it seems that α is of the most importance. With different λ and β, setting α as 0.1 shows the best performance. This means α as 0.1 is more appropriate to reflect the difference between positive pairs and soft negative pairs. Besides, with α as 0.1, the performance gets slight increase with β from 0.2 to 0.4, but gets apparent decrease with β as 2.0. With α as 0.2, the performance also gets apparent decrease with the increase of β. This may indicate it is better to set meaningful upper difference between positive pairs and soft negative pairs. Moreover, λ also shows great influence on the performance.

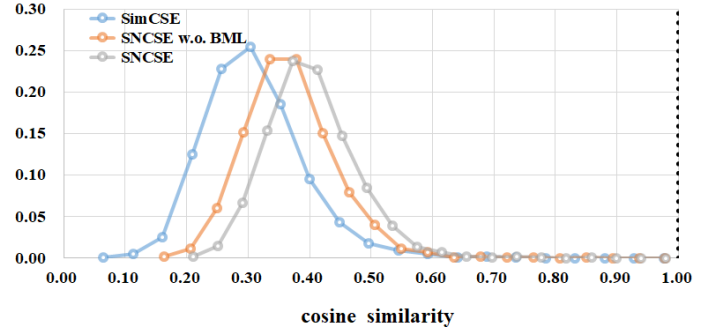### 4.3 Alleviation of Feature Suppression

To further detect the influence of SNCSE on the alleviation of feature suppression, we take soft negative samples for instance. Specifically, we randomly select 5000 sentences from STS task, and visualize the distribution of cosine similarity of different models on positive pairs, negative pairs, and soft negative pairs. For each sentence, we take itself as positive pair, and set default cosine similarity as 1. Then we take the negation of the sentence to construct soft negative pair, and take another randomly selected sentence as negative pair[7]. We also take SimCSE [Gao *et al.*, 2021] for comparison.

The results are shown in Figure 4. For positive pairs, the default similarity is 1.0. For negative pairs, the similarity peak of SimCSE, SNCSE without BML loss and SNCSE respectively appear at about 0.30, 0.38 and 0.37, which are far lower than 1.0. This indicates that all models can clearly distinguish positive pairs from negative pairs. For soft negative pairs, SimCSE, with similarity peak at 0.93, can somewhat distinguish soft negative pairs from positive pairs. However, SNCSE without BML loss, with the peak at 0.98, can hardly distinguish these two, although it highly outperforms SimCSE. It seems that the model simply takes soft negative pairs as pure positive pairs because of the high textual similarity between them, and can hardly recognize the semantic difference between them with severe feature suppression. Never-

---

[7]We do not take the original sentence pairs in STS task, since several pairs have high semantic similarity and should not be considered as negative pairs.
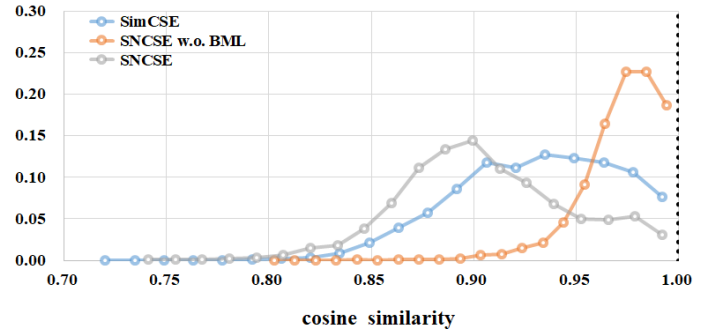
(a) Cosine similarity distribution of *soft negative* pairs.

(b) Cosine similarity distribution of *negative* pairs.

Figure 4: Cosine similarity distribution on different pairs. The vertical black point line means the default similarity of positive pairs is 1.0. We take soft negative samples for instance to detect the influence of SNCSE on the alleviation of feature suppression.

theless, SNCSE, with the peak at about 0.90, is able to clearly model the difference between these two , thus alleviate feature suppression and obtain better performance.

5 Discussion

To further detect the weakness of SNCSE, we propose rank-based error analysis. Specifically, we focus on top 5 error pairs, which have the highest rank estimation error as defined in Equation 11, on STS14 and SICK-R, since these two have

ID	Sentence1	Sentence2	R_G	R_E
		STS14		
1	I wood have asked her if she is up for a treesome.	I wood have at least asked if a threesome was out of the question.	3564	584
2	"being against nukes" does not mean "not wanting to use nukes"	"being against using nukes" means "not wanting to use nukes".	799	3695
3	I oppose the death penalty.	This is why I support the death penalty .	636	3475
4	fundimental difference, please?	fundamental difference?	3324	535
5	what conviction would that be?	what conviction where you talking about before?	3164	495
		SICK-R		
1	A man is doing pull-ups	The man is exercising	4187	304
2	A person is kicking a monkey	A monkey is kicking a person	1140	4869
3	A gloved person is kicking at the hand of a monkey	A monkey is kicking at the gloved hand of a person	1328	4859
4	Ferrets are climbing up a shelf	Ferrets are climbing down from a shelf	1129	4615
5	A small girl is riding in a toy car	A small toy girl is in a riding car	871	4354

Table 4: Top 5 error pairs of STS14 and SICK-R. The similarity scores are sorted from low to high and the error is defined as Equation 11. For a sentence pair, the similarity score is overestimated with R_E higher than R_G, and vice versa. We mark possible key issue with different colors, where red involves typos, green involves negation logic, blue involves textual independency, while cyan involves word order.

the most sentence pairs. As shown in Table 4, the main errors can be concluded as 4 kinds:

• **Typos**. Like the No.1 and No.4 pairs of STS14, the pairs themselves are annotated as high similarity. However, SNCSE underestimates the semantic similarity, since the model cannot understand the relationship between the word *"threesome"* and the typo *"treesome"*, and the word *"fundamental"* and the typo *"fundimental"*. This can be considered as annotation error.

• **Negation logic**. Like No.2, No.3 of STS14, and No.4 of SICK-R, one sentence of these pairs is the negation of another one. Differently, No.2 of STS14 involves explicit negation with the word "not", while the other two involve implicit negation based on antonyms. It is because of feature suppression that SNCSE overestimates the similarity.

• **Textual Independency**. Like No.5 of STS14 and No.1 of STS, the pairs are of high semantic similarity. However, the pairs share less textual and look like totally independent sentences. Therefore, SNCSE, which takes two independent sentences as negative pairs, tends to underestimate the semantic similarity of these pairs due to feature suppression.

• **Word Order**. Like No.2, No.3 and No.5 of SICK-R, the pairs share the same words but come with different word order. Actually, they have great semantic difference. However, SNCSE ignores the semantic difference due to the extremely similar textual. This is also caused by feature suppression.

In summary, it is far to totally handle feature suppression, although SNCSE has made some progress. The main problem is to distinguish and decouple textual similarity and semantic similarity. However, this will be extremely difficult for **self-attention based** transformer encoders, e.g. BERT and RoBERTa, where the embedding of a sentence comes from the weighted summary of the embedding of the words. Especially, it is far more difficult to accurately model the relationship between word order and sentence semantic with the position embedding. For instance, "Tom and Jerry became good friends" and "Jerry and Tom became good friends" share sim-

ilar semantic, while the pairs involving *word order* have apparently different semantics. Therefore, it is necessary to develop stronger models to deal with feature suppression.

6 Related Work

Reimers and Gurevych [2019], Li *et al.* [2020], and Yan *et al.* [2021] find that vanilla sentence embeddings from pretrained language models are of low quality to reflect the semantic, where the embeddings of any sentence pairs are of high similarity. To handle the problem, Reimers and Gurevych [2019] propose Sentence-BERT, which trains the encoders through textual entailment tasks in a supervised way. Early unsupervised sentence embedding models include BERT-flow [Li *et al.*, 2020] and BERT-Whitening [Su *et al.*, 2021], which are respectively based on reversible transformation and whitening operation. Yan *et al.* [2021] fistly introduce contrastive learning for unsupervised sentence embedding, and propose several data augmentation methods. Gao *et al.* [2021] propose SimCSE, which considers dropout operation as a kind of data augmentation method. Wu *et al.* [2021] point out that SimCSE has the bias to overestimate the semantic similarity of sentence pairs with the same length. To eliminate the bias, they propose ESimCSE with word repetition as data augmentation. They also introduce momentum contrastive [He *et al.*, 2020; Chen *et al.*, 2020b] for sentence embedding. Differently, in this paper, we mainly focus on feature suppression.

7 Conclusion

In this paper, we propose SNCSE with bidirectional margin loss (BML) to introduce soft negative samples to enhance sentence embedding. Our model makes significant progress and obtain state-of-the-art performance on STS task with all encoders. Our ablation study results show the importance of BML loss and the improvement on alleviating feature suppression. We also propose rank-based error analysis method to detect the weakness of SNCSE. In the future, we will try to study much stronger models to deal with feature suppression.

References

[Agirre *et al.*, 2012] Eneko Agirre, Daniel M. Cer, Mona T. Diab, and Aitor Gonzalez Agirre. Semeval-2012 task 6: A pilot on semantic textual similarity. In *SemEval 2012*, pages 385–393. ACL, 2012.

[Agirre *et al.*, 2013] Eneko Agirre, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez Agirre, and Weiwei Guo. *sem 2013 shared task: Semantic textual similarity. In **SEM 2013*, pages 32–43. ACL, 2013.

[Agirre *et al.*, 2014] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. Semeval-2014 task 10: Multilingual semantic textual similarity. In *SemEval 2014*, pages 81–91. ACL, 2014.

[Agirre *et al.*, 2015] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez Agirre, Weiwei Guo, Inigo Lopez Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *SemEval 2015*, pages 252–263. ACL, 2015.

[Agirre *et al.*, 2016] Eneko Agirre, Carmen Banea, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval 2016*, pages 497–511. ACL, 2016.

[Cer *et al.*, 2017] Daniel M. Cer, Mona T. Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *SemEval 2017*, pages 1–14. ACL, 2017.

[Chen *et al.*, 2020a] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML 2020*, pages 1597–1607. PMLR, 2020.

[Chen *et al.*, 2020b] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020.

[Chen *et al.*, 2021] Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. In *NIPS 2021*. NeurIPS, 2021.

[Devlin *et al.*, 2019] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL 2019*, pages 4171–4186. ACL, 2019.

[Ding *et al.*, 2021] Ning Ding, Yulin Chen, Xu Han, Guangwei Xu, Pengjun Xie, Hai Tao Zheng, Zhiyuan Liu, Juanzi Li, and Hong Gee Kim. Prompt-learning for fine-grained entity typing. *CoRR*, abs/2108.10604, 2021.

[Gao *et al.*, 2021] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *EMNLP 2021*, pages 6894–6910. ACL, 2021.

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR 2020*, pages 9726–9735. IEEE/CVF, 2020.

[Li *et al.*, 2020] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. In *EMNLP 2020*, pages 9119–9130. ACL, 2020.

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

[Liu *et al.*, 2021] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586, 2021.

[Marelli *et al.*, 2014] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A sick cure for the evaluation of compositional distributional semantic models. In *LREC 2014*, pages 216–223. ELRA, 2014.

[Oord *et al.*, 2018] Aaron Van Den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP 2019*, pages 3980–3990. ACL, 2019.

[Robinson *et al.*, 2021] Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? In *NIPS 2021*. NeurIPS, 2021.

[Su *et al.*, 2021] Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. Whitening sentence representations for better semantics and faster retrieval. *CoRR*, abs/2103.15316, 2021.

[Wu *et al.*, 2021] Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. Esimcse: Enhanced sample building method for contrastive learning of unsupervised sentence embedding. *CoRR*, abs/2109.04380, 2021.

[Yan *et al.*, 2021] Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. Consert: A contrastive framework for self-supervised sentence representation transfer. In *ACL 2021*, pages 5065–5075. ACL, 2021.

[Yao *et al.*, 2021] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *CoRR*, abs/2109.11797, 2021.