Pairwise Supervised Contrastive Learning of Sentence Representations

Dejiao Zhang   **Shang-Wen Li**[*]   **Wei Xiao**[*]   **Henghui Zhu**[*]
**Ramesh Nallapati**   **Andrew O. Arnold**   **Bing Xiang**
AWS AI

Abstract

Many recent successes in sentence representation learning have been achieved by simply fine-tuning on the Natural Language Inference (NLI) datasets with triplet loss or siamese loss. Nevertheless, they share a common weakness: sentences in a contradiction pair are not necessarily from different semantic categories. Therefore, optimizing the semantic entailment and contradiction reasoning objective alone is inadequate to capture the high-level semantic structure. The drawback is compounded by the fact that the vanilla siamese or triplet losses only learn from individual sentence pairs or triplets, which often suffer from bad local optima. In this paper, we propose PairSupCon, an instance discrimination based approach aiming to bridge semantic entailment and contradiction understanding with high-level categorical concept encoding. We evaluate PairSupCon on various downstream tasks that involve understanding sentence semantics at different granularities. We outperform the previous state-of-the-art method with 10%–13% averaged improvement on eight clustering tasks, and 5%–6% averaged improvement on seven semantic textual similarity (STS) tasks.

1 Introduction

Learning high-quality sentence embeddings is a fundamental task in Natural Language Processing. The goal is to map semantically similar sentences close together and dissimilar sentences farther apart in the representation space. Many recent successes have been achieved by training on the NLI datasets (Bowman et al., 2015; Williams et al., 2017; Wang et al., 2018), where the task is often to classify each sentence pair into one of three categories: entailment, contradiction, or neutral.
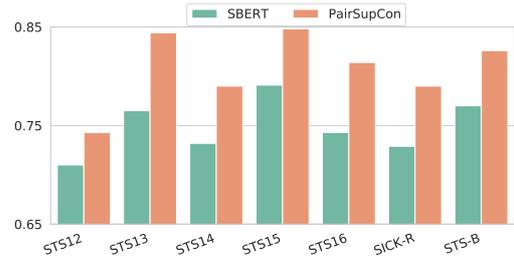
(a) Spearman Correlation Evaluation on STS.
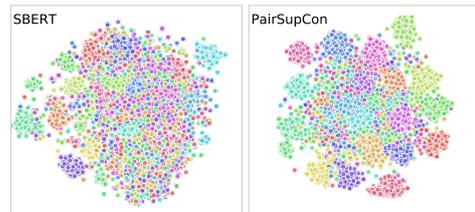
(b) TSNE visualization of the StackOverflow embeddings.

Figure 1: Simultaneoulsy encoding semantic categorical structure and pairwise entailment and contradiction understanding into embeddings.

Despite promising results, prior work (Conneau et al., 2017; Cer et al., 2018; Reimers and Gurevych, 2019a) share a common weakness: the sentences forming a contradiction pair may not necessarily belong to different semantic categories. Consequently, optimizing the model for semantic entailment and contradiction understanding alone is inadequate to encode the high-level categorical concepts into the representations. Moreover, the vanilla siamese (triplet) loss only learns from the individual sentence pairs (triplets), which often requires substantial training examples to achieve competitive performance (Oh Song et al., 2016; Thakur et al., 2020). As shown in Section 4.1, the siamese loss can sometimes drive a model to bad local optima where the performance of high-level semantic concept encoding is degraded when compared with its baseline counterpart.

In this paper, we take inspiration from self-supervised contrastive learning (Bachman et al.,

2019; He et al., 2020; Chen et al., 2020) and propose jointly optimizing the pairwise semantic reasoning objective with an *instance discrimination loss*. We name our approach Pairwise Supervised Contrastive Learning (PairSupCon). As noticed by the recent work (Wu et al., 2018; Zhang et al., 2021), instance discrimination learning can implicitly group similar instances together in the representation space without any explicit learning force directs to do so. PairSupCon leverages this implicit grouping effect to bring together representations from the same semantic category while, simultaneously enhancing the semantic entailment and contradiction reasoning capability of the model.

Although the prior work mainly focuses on pairwise semantic similarity related evaluations, we argue in this paper that the capability of encoding the high-level categorical semantic concept into the representations is an equally important aspect for evaluations. As shown in Section 4, the previous state-of-the-art model that performs best on the semantic textual similarity (STS) tasks results in degenerated embeddings of the categorical semantic structure. On the other hand, better capturing the high-level semantic concepts can in turn promote better performance of the low-level semantic entailment and contradiction reasoning. This assumption is consistent with how human categorize objects in a top-down reasoning manner. We further validate our assumption in Section 4, where PairSupCon achieves an averaged improvement of $10\% - 13\%$ over the prior work when evaluated on eight short text clustering tasks, and yields $5\% - 6\%$ averaged improvement on seven STS tasks.

2 Related Work

Sentence Representation Learning with NLI The suitability of leveraging NLI to promote better sentence representation learning is first observed by InferSent (Conneau et al., 2017), where a siamese BiLSTM network is optimized in a supervised manner with the semantic entailment and contraction classification objective. Universal Sentence Encoder (Cer et al., 2018) later augments an unsupervised learning objective with the supervised learning on NLI, and shows better transfer performance on various downstream tasks.

More recently, SBERT (Reimers and Gurevych, 2019b) finetunes a siamese BERT (Devlin et al., 2018) model on NLI and sets new state-of-the-art results. However, SBERT as well as the above work

adopt the vanilla siamese or triplet loss, which often suffer from slow convergence and bad local optima (Oh Song et al., 2016; Thakur et al., 2020).

Self-Supervised Instance Discrimination Another relevant line of work is self-supervised contrastive learning, which essentially solves an instance discrimination task that targets at discriminating each positive pair from all negative pairs within each batch of data (Oord et al., 2018; Bachman et al., 2019; He et al., 2020; Chen et al., 2020). Owing to their notable successes, self-supervised instance discrimination has become a prominent pre-training strategy for providing effective representations for a wide range of downstream tasks.

While recent successes are primarily driven by the computer vision domain, there is an increasing interest in leveraging variant instance discrimination tasks to support Pretraining Language Models (PLMs) (Meng et al., 2021; Giorgi et al., 2020; Wu et al., 2020; Rethmeier and Augenstein, 2021). Our proposal can be seen as complementary to this stream of work, considering that the instance discrimination learning based PLMs provide a good foundation for PairSupCon to further enhance the sentence representation quality by further learning on NLI. As demonstrated in Section 4, by training a pre-trained BERT-base model for less than an hour, PairSupCon attains substantial improvement on various downstream tasks that involve sentence semantics understanding at different granularities.

Deep Metric Learning Inspired by the pioneering work of (Hadsell et al., 2006; Weinberger and Saul, 2009), many recent works have shown significant benefit in learning deep representations using either siamese loss or triplet loss. However, both losses learn from individual pairs or triplets, which often require substantial training data to achieve competitive performance. Two different streams of work have been proposed to tackle this issue, with the shared focus on nontrivial pairs or triplets optimization. Wang et al. (2014); Schroff et al. (2015); Wu et al. (2017); Harwood et al. (2017) propose hard negative or hard positive mining that often requires expensive sampling. Oh Song et al. (2016) extends the vanilla triplet loss by contrasting each positive example against multiple negatives.

Our work leverages the strength of both lines, with the key difference being the above work requires categorical level supervision for selecting hard negatives. To be more specific, negative sam-
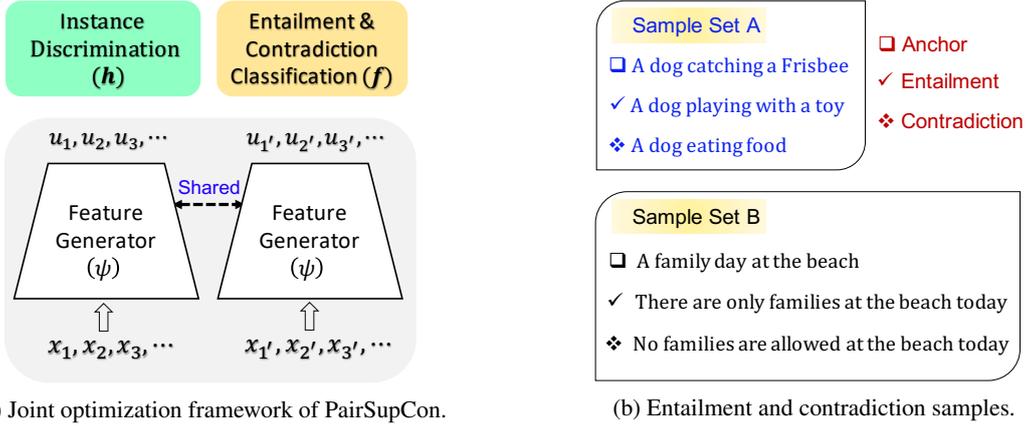
(a) Joint optimization framework of PairSupCon.

(b) Entailment and contradiction samples.

Figure 2: Illustration of PairSupCon. **(a):** Training framework of PairSupCon. **(b):** Despite the entailment sample is more similar to the anchor, both the contradiction and entailment samples are likely from the same semantic category as the anchor.

ples that have different categorical labels from the anchor but are currently mapped close to the anchor in the representation space, are likely to be more useful and hence being sampled. However, there are no categorical labels available in NLI. We thereby contrast each positive pair against multiple negatives collected using an unsupervised importance sampling strategy, for which the hypothesis is that hard negatives are more likely to locate close to the anchor. The effectiveness of this assumption is investigated in Section 4.3.

3 Model

Following SBERT (Reimers and Gurevych, 2019a), we adopt the SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2017) as our training data, and refer the combined data as NLI for convenience. The NLI data consists of labeled sentence pairs and each can be presented in the form: (*premise, hypothesis, label*). The premise sentences are selected from existing text sources and each premise sentence is paired with variant hypothesis sentences composed by human annotators. Each label indicates the hypothesis type and categorizes semantic relation of the associated premise and hypothesis sentence pair into one of three categories: *entailment, contradiction*, and *neural*, correspondingly.

Prior work solely optimizes either siamese loss or triplet loss on NLI. We instead aiming to leverage the implicit grouping effect of instance discrimination learning to better capture the high-level categorical semantic structure of data while, simultaneously promoting better convergence of the low-level semantic textual entailment and contradiction reasoning objective.

3.1 Instance Discrimination

We leverage the *positive (entailment)* pairs of NLI to optimize an instance discrimination objective which tries to separate each positive pair apart from all other sentences. Let $\mathcal{D} = \left\{ (x_j, x_{j'}), y_j \right\}_{j=1}^{M}$ denote a randomly sampled minibatch, with $y_i = \pm 1$ indicating an entailment or contradiction pair. Then for a premise sentence x_i within a *positive* pair $(x_i, x_{i'})$, we aim to separate its hypothesis sentence x_i' from all other $2M$-2 sentences within the same batch $\mathcal{D}$. To be more specific, let $\mathcal{I} = \{i, i'\}_{i=1}^{M}$ denote the corresponded indices of the sentence pairs in $\mathcal{D}$, we then minimize the following for x_i,

$$\ell_{\text{ID}}^i = -\log \frac{\exp(s(z_i, z_{i'})/\tau)}{\sum_{j \in \mathcal{I} \setminus i} \exp(s(z_i, z_j)/\tau)} . \quad (1)$$

In the above equation, $z_j = h(\psi(x_j))$ denotes the output of the instance discrimination head in Figure 2, τ denotes the temperature parameter, and $s(\cdot)$ is chosen as the cosine similarity, *i.e.,* $\text{s}(\cdot) = z_i^T z_{i'} / \|z_i\| \|z_{i'}\|$. Notice that Equation (1) can be interpreted as a $(2M-1)$-way softmax based classification loss of classifying z_i as z_i'.

Similarly, for the hypothesis sentence $(x_{i'})$ we also try to discriminate its premise (x_i) from all the other sentences in $\mathcal{D}$. We denote the corresponding loss as $\ell_{\text{ID}}^{i'}$ that is defined by exchanging the roles of instances $x_{i'}$ and x_i in Equation (1), respectively. In summary, the final instance discrimination loss is averaged over all positive pairs in $\mathcal{D}$,

$$\mathcal{L}_{\text{ID}} = \frac{1}{P_M} \sum_{i=1}^{M} \mathbb{1}_{y_i=1} \cdot \left(\ell_{\text{ID}}^i + \ell_{\text{ID}}^{i'} \right) . \quad (2)$$

Here, $\mathbb{1}_{(\cdot)}$ denotes the indicator function, and P_M is the number of positive pairs in $\mathcal{D}$. As demonstrated in Section 4, optimizing the above loss not only helps implicitly encode categorical semantic structure into representations, but also promotes better pairwise semantic reasoning capability, though no pairwise supervision except the true entailment labels are present to the model.

3.2 Learning from Hard Negatives

Notice that Eq (1) can be rewritten as

$$\ell_{\text{ID}}^i =$$
$$\log \left(1 + \sum_{j \neq i, i'} \exp \left[\frac{\text{s}(z_i, z_j) - \text{s}(z_i, z_{i'})}{\tau} \right] \right) .$$

It can be interpreted as extending the vanilla triplet loss by treating the other $2M{-}2$ samples within the same minibatch as negatives. However, the negatives are uniformly sampled from the training data, regardless of how informative they are. Ideally, we want to focus on hard negatives that are from different semantic groups but are mapped close to the anchor, *i.e.,* z_i, in the representation space. Although the categorical level supervision is not available in NLI, we approximate the importance of the negative samples via the following,

$$\ell_{\text{wID}}^i = \tag{3}$$
$$\log \left(1 + \sum_{j \neq i, i'} \exp \left[\frac{\alpha_j \text{s}(z_i, z_j) - \text{s}(z_i, z_{i'})}{\tau} \right] \right) .$$

Here $\alpha_j = \frac{\exp(\text{s}(z_i, z_j)/\tau)}{\frac{1}{2M-2} \sum_{k \neq i, i'} \exp(\text{s}(z_i, z_k)/\tau)}$, which can be interpreted as the relative importance of z_j among all $2M$-2 negatives of anchor z_i. The assumption is that hard negatives are more likely to be those that are located close to the anchor in the representation space. Although there might exists false negatives, *i.e.,* those located close to the anchor z_i but are from the same category, the probability is low as long as the underlying number of categories of the training data is not too small and each minibatch $\mathcal{D}$ is uniformly sampled.

3.3 Entailment and Contradiction Reasoning

The instance discrimination loss mainly focuses on separating each positive pair apart from the others, whereas there is no explicit force in discriminating contradiction and entailment. To this end, we jointly optimize a pairwise entailment and contradiction reasoning objective. We adopt the softmax-based cross-entropy to form the pairwise classification objective. Let $u_i = \psi(x_i)$ denote the representation of sentence x_i, then for each labeled sentence pair $(x_i, x_{i'}, y_i)$ we minimize the following,

$$\ell_{\text{C}}^i = \text{CE}\left(f(u_i, u_{i'}, |u_i - u_{i'}|), y_i \right) . \tag{4}$$

Here f denotes the linear classification head in Figure 1, and CE is the cross-entropy loss. Different from Reimers and Gurevych (2019b), we exclude the neural pairs from the original training set and focus on the binary classification of semantic entailment and contradiction only. Our motivation is that the concept semantic neural can be well captured by the instance discrimination loss. Therefore, we drop the neural pairs from the training data to reduce the functional redundancy of the two losses in PairSupCon and improve the learning efficiency as well.

Overall loss   In summary, our overall loss is

$$\mathcal{L} = \sum_{i=1}^M \ell_{\text{C}}^i + \beta \mathbb{1}_{y_i=1} \cdot \left(\ell_{\text{wID}}^i + \ell_{\text{wID}}^{i'} \right) , \tag{5}$$

where ℓ_{C}^i and $\ell_{\text{wID}}^i, \ell_{\text{wID}}^{i'}$ are defined in Equations (4) and (3), respectively. In the above equation, β balances between the capability of pairwise semantic entailment and contradiction reasoning, and the capability of the high-level categorical semantic structure encoding. We dive deep into the trade-off between these two aspects by evaluating PairSupCon with different β values in Section 4.3, and show how different β values can benefit different downstream tasks. Unless otherwise specified, we set $\beta = 1$ in this paper for the purpose of providing effective representations to various downstream tasks instead of being tailored to any specific ones.

4 Experiments

Baselines   In this section, we mainly investigate the effective strategies of leveraging the labeled NLI data to enhance the sentence representations of the pre-trained language models (PLMs). We compare PairSupCon against the vanilla BERT (Devlin et al., 2018; Sanh et al., 2019) models and the previous state-of-the-art approach, SBERT (Reimers and Gurevych, 2019a). We noticed a concurrent work SimCSE (Gao et al., 2021) when we were preparing this submission. Although SimCSE also

	AG	SS	SO	Bio	Tweet	G-TS	G-S	G-T	Avg.
BERT$_{distil \cdot base}$	**85.8**	68.4	20.7	32.0	47.5	63.2	56.8	50.0	53.1
SBERT$_{distil \cdot base}^{\diamondsuit}$	61.1	54.9	32.2	32.6	45.9	56.7	50.7	48.2	47.8
SimCSE$^{\spadesuit}$	-	-	-	-	-	-	-	-	-
PairSupCon$_{distil \cdot base}$	82.5	**71.5**	**63.6**	**39.0**	**54.2**	**69.1**	**63.6**	**59.7**	**62.9**
BERT$_{base}$	79.7	64.0	21.8	32.3	45.1	61.0	55.8	46.7	50.8
SBERT$_{base}^{\diamondsuit}$	65.8	62.5	29.6	31.9	45.6	56.5	50.2	49.3	48.9
SimCSE$^{\spadesuit}$	82.4	67.3	49.0	37.5	56.1	68.1	63.2	59.6	60.4
PairSupCon$_{base}$	**83.0**	**73.6**	**63.8**	**38.8**	**55.7**	**69.4**	**64.5**	**60.4**	**63.7**
BERT$_{large}$	83.1	66.5	26.4	31.4	44.3	62.3	55.8	46.4	52.0
SBERT$_{large}^{\diamondsuit}$	66.6	63.7	37.7	34.6	47.7	59.1	53.8	49.5	51.6
SimCSE$^{\spadesuit}$	82.7	66.3	49.0	40.5	**57.9**	68.1	62.4	60.8	61.0
PairSupCon$_{large}$	**84.2**	**75.7**	**63.9**	**41.8**	55.8	**70.4**	**64.8**	**61.3**	**64.7**

Table 1: Clustering accuracy reported on eight shorttext clustering datasets. The results are averaged over 10 clustering runs using KMeans with independent seeds. $\diamondsuit$ and $\spadesuit$: results evaluated on the checkpoints provided by Reimers and Gurevych (2019a) and Gao et al. (2021), respectively.

leverages the instance discrimination learning to improve the sentence embeddings, it shares a different motivation and focus than ours. We compare PairSupCon against SimCSE to show how different instance discrimination based approaches enhance the sentence representations differently, while our contribution is claimed over the previous SOTA models. Please refer to Appendix A for the details of our implementation.

Dataset	N	W	C	L	S
AgNews (AG)	8K	23	4	2K	2K
StackOverflow (SO)	20K	8	20	1K	1K
Biomedical (Bio)	20K	13	20	1K	1K
SearchSnippets (SS)	12K	18	8	2.66K	.37K
GoglenewsTS (G-TS)	11K	28	152	430	3
GoglenewsS (G-S)	11K	22	152	430	3
GoglenewsT (G-T)	11K	6	152	430	3
Tweet (Tweet)	5K	8	89	249	1

Table 2: Statistics of eight short text clustering datasets. N: number of text examples; W: average number of words contained in each text example; C number of clusters; L: the size of the largest cluster; and S: the size of the smallest cluster.

4.1 Clustering

Motivation Existing work mainly focuses on the semantic similarity (a.k.a STS) related tasks. We argue that an equally important aspect of sentence representation evaluation – the capability of encoding the high-level categorical structure into the representations, has so far been neglected. Desirably, a model should map the instances from the same category close together in the representation space while mapping those from different categories farther apart. This expectation aligns well with the underlying assumption of clustering and is consistent with how human categorizes data. We evaluate the capabilities of categorical concept embedding using K-Means (MacQueen et al., 1967; Lloyd, 1982), given its simplicity and the fact that the algorithm itself manifests the above expectation.

We consider eight benchmark datasets for short text clustering. As indicated in Table 2[1], the datasets present the desired diversities of both the size of each cluster and the number of clusters of each dataset. Furthermore, each text instance consists of 6 to 28 words when averaged within each dataset, which well covers the spectrum of NLI where each instance has 12 words on average. Therefore, we believe the proposed datasets can provide an informative evaluation on whether an embedding model is capable of capturing the high-level categorical concept.

Evaluation Results The evaluation results are summarized in Table 1. We run K-Means with the scikit-learn package (Pedregosa et al., 2011) on the representations provided by each model and report the clustering accuracy [2] averaged over 10 independent runs. [3] As Table 1 indicates, in com-

---

[1]Please refer to Appendix B for more details.

[2]We computed the clustering accuracy using the Hungarian algorithm (Munkres, 1957).

[3]We randomly select 10 independent seeds and fix them for all evaluations

	STS12	STS13	STS14	STS15	STS16	SICK-R	STS-B	Avg.
$BERT_{distil\cdot base}$	43.4	64.9	54.1	66.6	68.5	63.5	57.2	59.8
$SBERT^{\diamond}_{distil\cdot base}$	69.8	72.2	70.7	79.9	75.4	74.5	78.4	74.4
$SimCSE^{\spadesuit}$	-	-	-	-	-	-	-	-
PairSupCon$_{distil\cdot base}$	**73.6**	**84.1**	**78.3**	**84.4**	**81.8**	**77.8**	**82.3**	**80.3**
$BERT_{base}$	30.9	59.9	47.7	60.3	63.7	58.2	47.3	52.6
$SBERT^{\diamond}_{base}$	71.0	76.5	73.2	79.1	74.3	72.9	77.0	74.9
$SimCSE^{\spadesuit}$	**75.3**	**84.7**	**80.2**	**85.4**	80.8	**80.4**	**84.3**	**81.6**
PairSupCon$_{base}$	74.3	84.4	79.0	84.8	**81.4**	79.0	82.6	80.8
$BERT_{large}$	27.7	55.8	44.5	51.7	61.9	53.9	47.0	48.9
$SBERT^{\diamond}_{large}$	72.3	78.5	74.9	81.0	76.3	73.8	79.2	76.5
$SimCSE^{\spadesuit}$	**75.8**	**86.3**	**80.4**	**86.1**	80.9	**81.1**	**84.9**	**82.2**
PairSupCon$_{large}$	74.2	85.8	79.5	85.5	**81.7**	80.4	83.7	81.5

Table 3: Spearman rank correlation between the cosine similarity of sentence representations and the ground truth labels on seven Semantic Textual Similarity (STS) tasks. $\diamond$ and $\spadesuit$: results evaluated on the checkpoints provided by Reimers and Gurevych (2019a) and Gao et al. (2021), respectively.

parison with the vanilla BERT models, SBERT results in degenerated embedding of the categorical semantic structure by simply optimizing the pairwise siamese loss. One possible reason is that SBERT uses a large learning rate (2e-05) to optimize the transformer, which can cause catastrophic forgetting of the knowledge acquired in the original BERT models. We find using a smaller learning rate for the backbone can consistently improve the performance of SBERT (see the performance of BERT-base with "Classificaiton" in Table 4).

Nevertheless, PairSupCon leads to an averaged improvement of 10.8% to 15.2% over SBERT, which validates our motivation in leveraging the implicit grouping effect of the instance discrimination learning to better encode the high-level semantic concepts into representations. Moreover, PairSupCon also attains better performance than SimCSE, and we suspect this is because PairSupCon better leverages the training data. Specifically, PairSupCon aims to discriminate an positive (entailment) sentence pair apart from all other sentences, no matter they are premises or hypotheses. In contrast, SimCSE only separates a premise from the other premises through their entailment and contradiction hypotheses, while there is no explicit instance discrimination force within the premises or the hypotheses alone. Considering the statistic data difference (Williams et al., 2017) between premises and hypotheses, PairSupCon can potentially better capture categorical semantic concepts by leverag-

ing additional intrinsic semantic properties of the premises or the hypotheses that are undiscovered by SimCSE.

4.2 Semantic Textual Similarity

Next, we asses the performance of PairSupCon on seven STS tasks, namely STS 2012-2016 (Agirre et al., 2012, 2013, 2014, 2015, 2016), STS Benchmark (Cer et al., 2017), and SICK-Relatedness (Marelli et al., 2014). These datasets include pairs of sentences with a fine-grained gold semantic similarity score ranges from 0 to 5. To enable a fair comparison, we follow the setup in SBERT (Reimers and Gurevych, 2019a), and compute the Spearman correlation[4] between cosine similarity of sentence embeddings and the ground truth similarity scores of each pair.

The evaluation results are reported in Table 3. PairSupCon substantially outperforms both the vanilla BERT and SBERT models. This validates our assumption that, by implicitly encoding the high-level categorical structure into the representations, PairSupCon promotes better convergence of the low-level semantic entailment reasoning objective. This assumption is consistent with the top-down categorization behavior of humans. Although SimCSE leverages STS-Benchmark as the development set while PairSupCon is fully blind

---

[4]Same as SBERT (Reimers and Gurevych, 2019a) and SimCSE (Gao et al., 2021), we concatenate all the topics and report the overall Spearman's correlation.
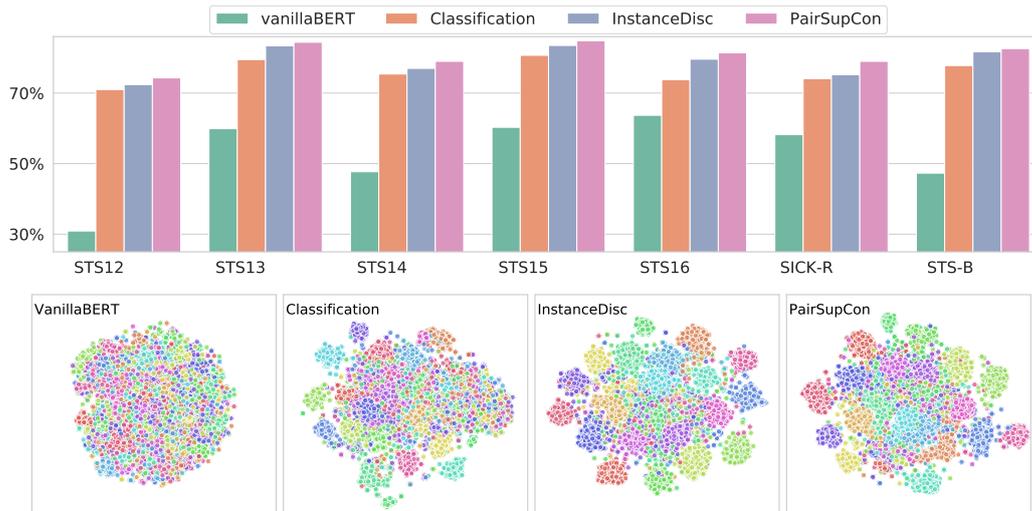
Figure 3: Ablation study of PairSupCon. **First Row**: Spearman correlation evaluated on the STS tasks. **Second Row**: visualization of the StackOverflow (Xu et al., 2017) representations using t-SNE (Maaten and Hinton, 2008).

to the downstream tasks[5], we hypothesize the performance gain of SimCSE on STS is mainly contributed by explicitly merging the entailment and contradiction separation into the instance discrimination loss. On the other hand, as we discussed in Section 4.1, PairSupCon achieves more obvious performance gain on the clustering tasks through a bidirectional instance discrimination loss. Therefore, developing a better instance discrimination based sentence representation learning objective by incorporating the strengths of both SimCSE and PairSupCon could be a promising direction.

Loss Func.	Downstream	
	STS	Clustering
Classification	76.0	52.9
InstanceDisc	79.0	**65.5**
PairSupCon		
$\beta = 0.5$	**81.0**	61.6
$\beta = 1$	80.8	63.7
$\beta = 2$	80.4	64.6
$\beta = 4$	80.0	65.0
$\beta = 4$	79.5	65.4

Table 4: Ablation study of PiarSupCon on BERT-base. We underline the setting used for all our experiments.

4.3 Ablation Study

We run extensive ablations to better understand what enables the good performance of PairSup-

[5]We also evaluate PairSupCon with STS-Benchmark as the development set in Appendix C.

Loss Func.	Downstream	
	STS	Clustering
InstanceDisc		
w/o HardNegSample	78.4	64.3
w/ HardNegSample	**79.0**	**65.5**
PairSupCon		
w/o HardNegSample	80.2	62.1
w/ HardNegSample	**80.8**	**63.7**

Table 5: Ablation of hard negative sampling. We choose BERT-base as backbone.

Con. For notational convenience, we name the pairwise semantic relation classification objective in PairSupCon as **Classification**, and the instance discrimination objective as **InstanceDisc**.

PairSupCon versus. Its Components In Figure 3, we compare PairSupCon against its two components, namely Classification and InstanceDisc. As it shows, InstanceDisc itself outperforms Classification on both STS and categorical concept encoding. The result matches our expectation that contrasting each positive pair against multiple negatives, despite obtained through unsupervised sampling, yields better performance than simply learning from each individual pair. By jointly optimizing both objectives, PairSupCon leverages the implicit grouping effect of InstanceDisc to encode the high-level categorical structure into representations while, simultaneously complementing InstanceDisc with more fine-grained semantic con-

	MR	CR	SUBJ	MPQA	SST	TREC	MRPC
BERT$_{distil\cdot base}$	61.4(4.0)	69.0(2.6)	**86.2**(1.8)	67.4(3.8)	67.3(4.2)	**59.2**(6.3)	59.8(6.9)
SBERT$_{distil\cdot base}^{\diamond}$	**69.8**(1.9)	**83.3**(1.9)	80.4(2.8)	81.9(1.5)	78.8(1.2)	39.8(4.2)	**60.4**(3.4)
SimCSE$^{\spadesuit}$	-	-	-	-	-	-	-
PairSupCon$_{distil\cdot base}$	65.6 (6.4)	80.5(1.5)	82.7 (3.3)	**83.8**(1.1)	**79.9**(1.6)	52.5(3.2)	59.7(2.0)
BERT$_{base}$	61.8(4.3)	68.2(3.1)	**87.4**(1.6)	63.6(5.2)	68.0(1.7)	**58.4**(2.1)	56.5(12.1)
SBERT$_{base}^{\diamond}$	**72.9**(1.7)	81.2(4.9)	81.9(1.1)	80.7(2.4)	78.0 (6.1)	45.8(1.9)	60.0(3.9)
SimCSE$^{\spadesuit}$	71.8(1.8)	**82.9**(1.5)	85.9(2.0)	79.0(4.2)	80.0(3.7)	54.5(4.0)	57.8(5.1)
PairSupCon$_{base}$	72.8(2.0)	81.8(2.6)	85.7(1.9)	**82.7**(1.2)	**80.2**(3.2)	49.3(3.9)	**61.4**(4.4)
BERT$_{large}$	64.4(4.4)	70.9(2.9)	**86.6**(3.0)	62.9(4.6)	73.1(3.0)	**57.2**(2.9)	**63.4**(9.2)
SBERT$_{large}^{\diamond}$	**78.5**(0.9)	**85.1**(6.3)	79.8(3.6)	83.3(1.6)	82.5(4.8)	43.5(4.9)	58.8(4.7)
SimCSE$^{\spadesuit}$	75.4(2.3)	84.5(3.0)	85.5(2.8)	84.0(1.9)	**84.9**(2.2)	52.4(3.4)	59.6(2.4)
PairSupCon$_{large}$	74.8(2.6)	84.8(4.1)	84.6(2.3)	**85.2**(1.3)	83.7(1.7)	50.4(5.7)	62.6(5.3)

Table 6: Few-shot learning evaluation on SentEval. For each task, we randomly sample 16 labeled instances per class and report the mean (standard deviation) performance over 5 different training sets. $\diamond$ and $\spadesuit$: results evaluated on the checkpoints provided by Reimers and Gurevych (2019a) and Gao et al. (2021), respectively.

cept reasoning capability via Classification.

Table 4 indicates a trade-off between the high-level semantic structure encoding and the low-level pairwise entailment and contradiction reasoning capability of PairSupCon. Focusing more on the pairwise classification objective, *i.e.,* using smaller β values, can hurt the embeddings of the high-level semantic structure. This result is not surprising, especially considering that sentences forming a contradiction pair are not necessarily belong to different semantic groups. On the other hand, InstanceDisc only focuses on separating each positive pair from all other samples within the same minibatch, and an explicit force that discriminates semantic entailment from contradiction is necessary for PairSupCon to achieve competitive performance on the more fine-grained pairwise similarity reasoning on STS. As indicated in Table 4, we can tune the β values to attain effective representations for specific downstream tasks according to their semantic granularity focuses. We set $\beta = 1$ for all our experiments with the goal to provide effective universal sentence representations to different downstream tasks.

Hard Negative Sampling Helps In Table 5, we compare both PairSupCon and InstanceDisc against their counterparts where the negatives in the instance discrimination loss are uniformly sampled from data. As it shows, the hard negative sampling approach proposed in Section 3.2 leads to improved performance on both STS and clustering tasks. We associate this performance boost with our assumption that hard negatives are likely located close to the anchor. A properly designed

distance-based sampling approach can drive the model to better focus on hard negative separation and hence lead to better performance.

On the other hand, hard negative sampling without any supervision is a very challenging problem, especially considering that samples within the local region of an anchor are also likely from the same semantic group as the anchor. As a consequence, a solely distance-based sampling approach can induce certain false negatives and hurt the performance. To tackle this issue, leveraging proper structure assumption or domain-specific knowledge could be potential directions, which we leave as future work.

4.4 Transfer Learning

In order to provide a fair and comprehensive comparison with the existing work, we also evaluate PairSupCon on the following seven transferring tasks: MR (Pang and Lee, 2005), CR (Hu and Liu, 2004), SUBJ (Pang and Lee, 2004), MPQA (Wiebe et al., 2005), SST (Socher et al., 2013), TREC (Li and Roth, 2002), and MRPC (Dolan et al., 2004). We follow the widely used evaluation protocol, where a logistic regression classifier is trained on top of the frozen representations, and the testing accuracy is used as a measure of the representation quality. We adopt the default configurations of the SentEval (Conneau and Kiela, 2018) toolkit and report the evaluation results in Table 7 in Appendix D. As we can see, the performance gap between different methods are relatively small.

We suspect the reason is that the transfer learning tasks do not present enough complexities to fully uncover the performance gap between different ap-

proaches, especially considering that most tasks are binary classification with a large amount of labeled training examples. To further examine our hypothesis, we extend the evaluation to the setting of few-shot learning, where we uniformly sample 16 labeled instances per class for each task. We report the mean and standard deviation of the evaluation performance over 5 different sample sets in Table 6. Although we observe more obvious performance gap on each specific task, there is no consistent performance gap between different approaches when evaluated across different tasks. Therefore, to better evaluate the transfer learning performance of sentence representations, more complex and diverse datasets are required.

5 Discussion and Conclusion

In this paper, we present a simple framework for universal sentence representation learning. We leverage the implicit grouping effect of instance discrimination learning to better encoding the high-level semantic structure of data into representations while, simultaneously promoting better convergence of the lower-level semantic entailment and contradiction reasoning objective. We substantially advance the previous state-of-the-art results when evaluated on various downstream tasks that involve understanding semantic concepts at different granularities.

We carefully study the key components of our model and pinpoint the performance gain contributed by each of them. We observe encouraging performance improvement by using the proposed hard negative sampling strategy. On the other hand, hard negative sampling without any supervision is an crucial, yet significantly challenging problem that should motivate further explorations. Possible directions include making proper structure assumption or leveraging domain-specific knowledge. The substantial performance gain attained by our model also suggests developing explicit grouping objectives could be another direction worth investigation.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. 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 Claramunt, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511*. ACL (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, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (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.

Philip Bachman, R Devon Hjelm, and William Buchwalter. 2019. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pages 15535–15545.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

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*.

William Dolan, Chris Quirk, Chris Brockett, and Bill Dolan. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

John M Giorgi, Osvald Nitski, Gary D Bader, and Bo Wang. 2020. Declutr: Deep contrastive learning for unsupervised textual representations. *arXiv preprint arXiv:2006.03659*.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Ben Harwood, Vijay Kumar BG, Gustavo Carneiro, Ian Reid, and Tom Drummond. 2017. Smart mining for deep metric learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2821–2829.

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*, pages 9729–9738.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

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

Yu Meng, Chenyan Xiong, Payal Bajaj, Saurabh Tiwary, Paul Bennett, Jiawei Han, and Xia Song. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *arXiv preprint arXiv:2102.08473*.

James Munkres. 1957. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. 2016. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012.

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

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *arXiv preprint cs/0409058*.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *arXiv preprint cs/0506075*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on World Wide Web*, pages 91–100.

Md Rashadul Hasan Rakib, Norbert Zeh, Magdalena Jankowska, and Evangelos Milios. 2020. Enhancement of short text clustering by iterative classification. *arXiv preprint arXiv:2001.11631*.

Nils Reimers and Iryna Gurevych. 2019a. 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.

Nils Reimers and Iryna Gurevych. 2019b. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Nils Rethmeier and Isabelle Augenstein. 2021. A primer on contrastive pretraining in language processing: Methods, lessons learned and perspectives. *arXiv preprint arXiv:2102.12982*.

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

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2020. Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. *arXiv preprint arXiv:2010.08240*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. 2014. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1386–1393.

Kilian Q Weinberger and Lawrence K Saul. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2).

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

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. 2017. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2840–2848.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742.

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*.

Jiaming Xu, Bo Xu, Peng Wang, Suncong Zheng, Guanhua Tian, and Jun Zhao. 2017. Self-taught convolutional neural networks for short text clustering. *Neural Networks*, 88:22–31.

Jianhua Yin and Jianyong Wang. 2016. A model-based approach for text clustering with outlier detection. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 625–636. IEEE.

Dejiao Zhang, Feng Nan, Xiaokai Wei, Shang-Wen Li, Henghui Zhu, Kathleen McKeown, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2021. Supporting clustering with contrastive learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5419–5430, Online. Association for Computational Linguistics.

Xiang Zhang and Yann LeCun. 2015. Text understanding from scratch. *arXiv preprint arXiv:1502.01710*.

A Implementation

We use the pre-trained BERT models as the backbone. The pairwise entailment and contradiction classification loss is optimized with a linear layer of size $3d \times 2$, where d denotes the dimension of the sentence representations. We choose a two-layer MLP with size ($d \times d$, $d \times 128$) to optimize the instance discrimination loss. We use Adam (Kingma and Ba, 2015) as our optimizer, with a batch size of 1024 and a constant learning rate of $5e$-04. We use a smaller learning rate $5e$-06 for updating the transformers. We train our models for three epochs and evaluate all downstream tasks at the end of training. Same as Reimers and Gurevych (2019a), we use mean pooling to obtain sentence representations for both training and evaluation. For fair comparison, we use the default setting, *i.e.,* [CLS] representation, to evaluate SimCSE (Gao et al., 2021).

	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	Avg.
BERT$_{distil\cdot base}$	80.0	85.7	95.0	88.9	85.4	**90.6**	74.1	85.7
SBERT$^{\diamond}_{distil\cdot base}$	**80.9**	**88.2**	92.7	89.5	**87.4**	83.6	**75.2**	85.4
SimCSE$^{\spadesuit}$	-	-	-	-	-	-	-	-
PairSupCon$_{distil\cdot base}$	80.6	87.8	**94.7**	**89.9**	86.5	89.0	74.4	**86.1**
BERT$_{base}$	81.5	86.7	95.2	88.0	85.9	**90.6**	73.7	86.0
SBERT$^{\diamond}_{base}$	**83.6**	**89.4**	94.4	89.9	**89.0**	89.6	**76.0**	**87.4**
SimCSE$^{\spadesuit}$	82.9	89.2	94.8	89.7	87.3	88.4	73.5	86.5
PairSupCon$_{base}$	82.7	88.8	**95.2**	90.3	87.6	88.8	74.3	86.8
BERT$_{large}$	84.3	89.2	95.6	86.9	89.3	91.4	71.7	86.9
SBERT$^{\diamond}_{large}$	84.6	**90.9**	94.5	90.2	**90.7**	86.8	**76.6**	87.7
SimCSE$^{\spadesuit}$	**85.4**	90.8	95.3	90.3	90.3	90.6	76.3	88.4
PairSupCon$_{large}$	84.5	90.1	**95.7**	**90.6**	**90.7**	92.4	75.3	**88.5**

Table 7: Transfer learning results evaluated on SentEval. $\diamond$ and $\spadesuit$: results evaluated on the checkpoints provided by Reimers and Gurevych (2019a) and Gao et al. (2021), respectively.

B Short Text Clustering Datasets

SearchSnippets is extracted from web search snippets, which contains 12340 snippets associated with 8 groups Phan et al. (2008).

StackOverflow is a subset of the challenge data published by Kaggle[6], where 20000 question titles associated with 20 different categories are selected by Xu et al. (2017).

Biomedical is a subset of PubMed data distributed by BioASQ[7], where 20000 paper titles from 20 groups are randomly selected by Xu et al. (2017).

AgNews is a subset of news titles (Zhang and LeCun, 2015), which contains 4 topics selected by Rakib et al. (2020).

Tweet consists of 89 categories with 2472 tweets in total (Yin and Wang, 2016).

GoogleNews contains titles and snippets of 11109 news articles related to 152 events (Yin and Wang, 2016). We name the full dataset as *GoogleNews-TS* while, *GoogleNews-T* and *GoogleNews-S* are obtained by extracting the titles and the snippets, respectively.

---

[6]https://www.kaggle.com/c/predict-closed-questions-on-stackoverflow/download/train.zip
[7]http://participants-area.bioasq.org

C Leveraging STS-Benchmark as the Development Set

To better understand the underlying causes of the performance gap between PairSupCon and Sim-CSE on STS, we also train PairSupCon by using STS-Benchmark as the development set. We summarize the corresponding evaluation results in Table 8, which indicates that PairSupCon does not benefit from leveraging the STS-Benchmark. We thereby hypothesize the performance gain of Sim-CSE is mainly attributed by merging the entailment and contradiction discrimination into the instance-wise contrastive learning objective.

On the other hand, as discussed in Section 4.1, SimCSE can be roughly interpreted as an uni-directional instance-wise contrastive learning. In contrast, PairSupCon utilizes the training data in a more efficient way through a bidirectional instance discrimination loss, and hence achieves more obvious performance gain on the clustering tasks. Therefore, developing a better instance discrimination based sentence representation learning objective by incorporating the strengths of both SimCSE and PairSupCon could be a promising direction.

D Transfer Learning

To provide a fair and more comprehensive comparison with the existing work, we also evaluate Pair-SupCon on the seven transferring tasks using the SentEval toolkit (Conneau and Kiela, 2018). We follow the widely used evaluation protocol, where a

PairSupCon	STS12	STS13	STS14	STS15	STS16	SICK-R	STS-B	Avg.
DistilBase$_{w/\text{ STS-B}}$	73.4	83.8	78.2	84.2	81.9	77.7	82.2	80.2
DistilBase$_{w/o\text{ STS-B}}$	73.6	84.1	78.3	84.4	81.8	77.8	82.3	80.3
BertBase$_{w/\text{ STS-B}}$	74.2	84.5	79.0	84.7	81.4	78.9	82.6	80.8
BertBase$_{w/o\text{ STS-B}}$	74.3	84.4	79.0	84.8	81.4	79.0	82.6	80.8
BertLarge$_{w/\text{ STS-B}}$	74.3	85.7	79.1	84.8	81.7	80.2	83.3	81.3
BertLarge$_{w/o\text{ STS-B}}$	74.2	85.8	79.5	85.5	81.7	80.4	83.7	81.5

Table 8: Investigation on the effectiveness of leveraging STS-Benchmark as the development set for training PairSupCon. The training of our proposed PairSupCon model is fully blind to the downstream tasks. This table indicates that, different from SimCSE (Gao et al., 2021), no obvious performance gain is attained by PairSupCon when using STS-B as the development set.

logistic regression classifier is trained on top of the frozen representations, and the testing accuracy is used as a measure of the representation quality. We report the evaluation results in Table 7. As we can see, the performance gap between different models are small, yet still not consistent across different tasks. As discussed in Section 4.4, one possible explanation is that the transfer learning tasks do not present enough complexities to discriminate the performance gap between different approaches, since most tasks are binary classification with a large amount of labeled training examples.

Although we observe more obvious performance gap by extending the evaluation to the setting of few-shot learning (Table 6), there is no consistent performance gain across different tasks attained by any specific model investigated in this paper. Moving forward, having more complex and diverse datasets to evaluate the transfer learning performance could better direct the development of universal sentence representation learning.