Harnessing Dimension-level Contrastive Learning and Information Compensation Mechanism for Sentence Embedding Enhancement

Anonymous ACL submission

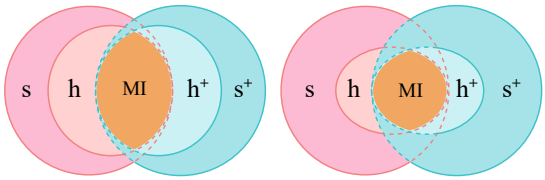
Abstract

Although unsupervised sentence embedding learning has achieved great success through the construction of positive samples and instance-level contrastive learning (ICL), the learned sentence embeddings can be over-compressed or suffer from dimensional pollution due to noisy data augmentation and unconstrained ICL learning processes. To address the above issues, we design a novel sentence embedding enhancement method, namely MSSE, where an information compensation mechanism (ICM) and a dimensional-level contrastive learning mechanism (DCM) are proposed. ICM is motivated by the information bottleneck principle and can prevent excessive compression of representation learning. DCM constrains the learning process of ICL and reduces information contamination across different dimensions. Experimental results demonstrate that our method outperforms the current competitive baselines for 7 STS tasks across unsupervised, few-shot, and supervised learning of sentence embeddings. The source code is available at https://anonymous.4open.science/r/MSSE-main.

1 Introduction

Learning sentence embeddings is a fundamental problem in natural language processing, aiming to map sentences into a unified representation space. It can improve downstream tasks, such as semantic textual similarity (STS) (Agirre et al., 2012), sentiment analysis (Tang et al., 2014; Yu et al., 2017) and information retrieval (Ma et al., 2016; Xiong et al., 2020), etc.

With the success of contrastive learning (Ye et al., 2019; He et al., 2020) and pre-trained language models (PLMs), such as BERT/RoBERTa (Devlin et al., 2019; Liu et al., 2019), many studies focus on learning sentence mutual information by constructing positive and negative samples (Gao et al., 2021; Zhou et al., 2022; Zhuo et al., 2023; Chen et al., 2023). They use *instance-level contrastive learning* (ICL) losses (e.g., InfoNCE (Oord et al., 2018)) to make the representations of positive samples similar and those of negative samples dissimilar, aiming to capture the *mutual information* (MI) present in sentences. In essence, this mechanism guides the model to compress input data, retaining key information while discarding irrelevant information. This aligns with the goal of the *information bottleneck* (IB) principle (Tishby et al., 2000) to find an optimal representation h that maximizes the following term:



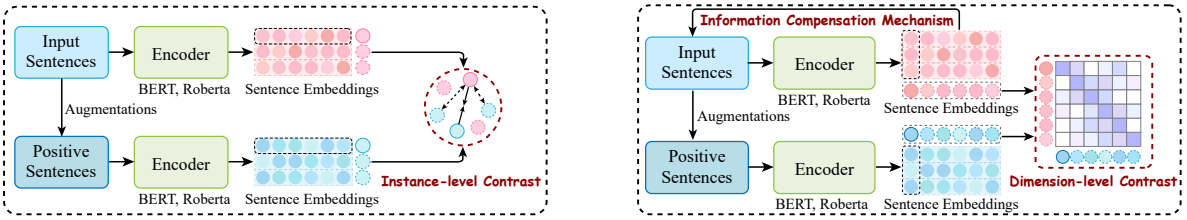
(a) Minimal sufficient MI (b) Over-compressed MI

Figure 1: (a) Instance-level contrastive learning aims to compress positive sentence pairs (s, s^+) into their representation spaces (h, h^+) while maintaining the effectiveness for predicting y. (b) The noise introduced by bad augmentation strategies leads to over-compressed representation spaces of (h, h^+) as well as their mutual information (MI).

$$\mathcal{L}_{\text{IB}} = I(h, y) - \beta I(h, s), \beta > 0 \quad (1)$$

where s and y respectively indicate the input (e.g., a sentence) and output (e.g., the label related to a downstream task such as STS). Maximizing $\mathcal{L}_{\text{IB}}$ is equivalent to maximizing the mutual information between h and y while minimizing the MI between h and s. As a result, h provides the most useful information for predicting y while discarding irrelevant information as much as possible.

Instance-level contrastive learning (ICL) utilizes the InfoNCE loss as an optimization objective to learn mutual information between different views of sentences. It is noteworthy that the InfoNCE loss has been proven to maximize mutual information between in-batch samples (Hjelm et al., 2018; Belghazi et al., 2018), i.e., maximizing the mutual infor-



(a) Previous Methods (b) Our Method

Figure 2: (a) Previous methods use various augmentations and instance-level contrastive learning to pull similar representations closer and push dissimilar representations apart. (b) Our method utilizes a dimension-level contrastive mechanism to alleviate the dimensional pollution and an information compensation mechanism to loosen the constraint of information bottleneck.

mation between positive pairs (h, h^+) and minimizing the mutual information between negative pairs (h, h^-), which is denoted as $I(h, h^+) - I(h, h^-)$. Tsai et al. (2020) have shown that optimizing the above term is equivalent to optimizing Eq. 1, indicating that ICL is closely related to mutual information optimization and information bottleneck principle.

Tian et al. (2020) have found that the mutual information learned by ICL is minimally sufficient for the output y, as shown in the orange part of Figure 1(a). However, some augmentation strategies used in ICL can introduce irrelevant noise and even disrupt the semantic structure of sentences (Wang and Isola, 2020; Gao et al., 2021), leading to the reduction of the representation spaces of h and h^+, as shown in Figure 1(b). This causes *over-compression* of the learned mutual information, which is detrimental to sentence embedding learning and decreases the model's generalization ability. When the task changes, the performance of the learned representation h can be adversely affected in downstream tasks (Tsai et al., 2020; Wang et al., 2021).

Besides *over-compression*, another shortage in prior studies is that they mainly focus on how to generate better positive or negative samples for ICL, thus mapping each sentence into a multi-dimension vector. The ideal situation is that each dimension presents unique information, but in reality, irrelevant information may propagate to other dimensions (Locatello et al., 2019; Chen and He, 2021). We call this phenomenon *dimensional pollution*, which is neglected in ICL and may weaken the quality of sentence embeddings.

In this paper, we propose an intuitively simple yet powerfully effective enhancement method in Multi-Scenario settings for Sentence Embeddings, termed **MSSE**. First, as depicted in Figure 2(b), to tackle the issue of *dimensional pollution*, we propose a **Dimensional-level Contrastive Learning Mechanism** (DCM). Specifically, DCM reduces information contamination between dimensions and enhances the quality of sentence embeddings by pulling the same dimensions (positives) closer and pushing different dimensions (negatives) farther apart. Second, to prevent the *over-compression* of mutual information and enhance generalization, we devise an **Information Compensation Mechanism** (ICM). ICM loosens the constraint of the IB principle by increasing the mutual information $I(h, s)$ between the original input s and the representation h. This compensates for the over-compressed mutual information, enhances the generalization ability, and further improves the performance of sentence embeddings.

We conduct extensive experiments on the STS tasks, the results demonstrate that our method significantly improves the state-of-the-art (SOTA) approaches in unsupervised, few-shot, and supervised settings. In the unsupervised setting, compared to SimCSE, our method using $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{large}}$ achieves absolute gains of 3.90% and 2.95% on average Spearman's correlation, respectively, significantly outperforming competitive baselines. In the few-shot setting, with only 0.1% of the data, MSSE-$\text{BERT}_{\text{base}}$ outperforms SimCSE-$\text{BERT}_{\text{base}}$ by 10.09%. Moreover, we conduct a set of experimental analyses to demonstrate the effectiveness of our approach. The contributions of our work are as follows:

- We propose a dimension-level contrastive learning mechanism to constrain the problem of *dimensional pollution*.
- We devise an information compensation mechanism to loosen the constraint of the information bottleneck, ensuring that mutual information is not over-compressed.
- Our approach achieves the SOTA performance in the widely-used STS task.

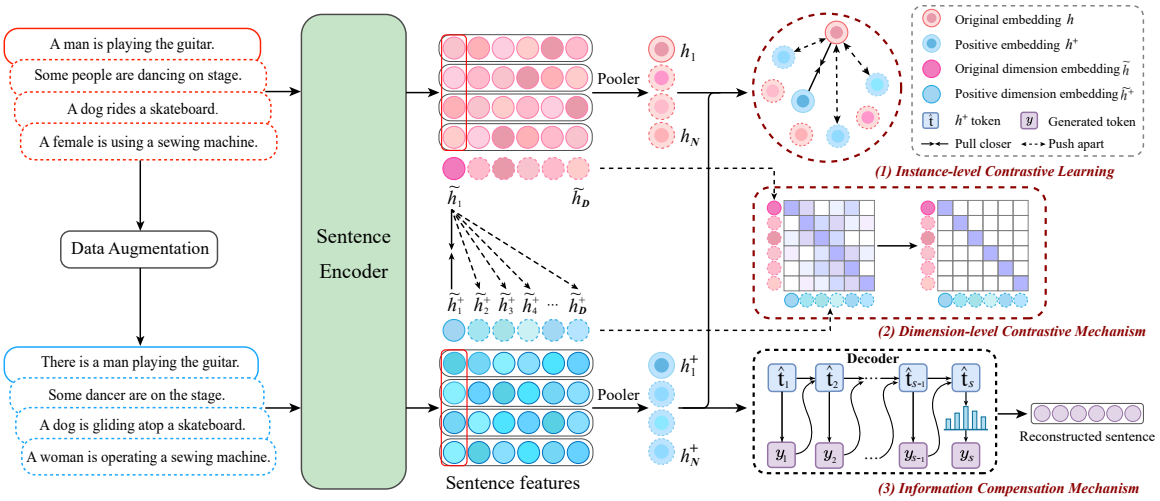


Figure 3: The overall framework of MSSE consists of three components: (1) Instance-level contrastive learning based on InfoNCE loss; (2) Dimension-level contrastive learning mechanism mitigates the impact of dimensional pollution; (3) Information compensation mechanism loosens the constraints and compensates for the over-compressed mutual information. Solid circles represent positive samples and dashed circles represent negative samples within the same batch.

2 Related Work

2.1 Contrastive Learning for Unsupervised Sentence Embeddings (CSE)

Unsupervised sentence embedding learning is a fundamental task in natural language processing. With the development of PLMs (e.g., BERT (Devlin et al., 2019), Roberta (Liu et al., 2019)) and the emergence of SimCSE (Gao et al., 2021), the paradigm of "PLMs + contrastive learning" has been widely used in sentence embedding learning. In practice, these works (Gao et al., 2021; Wu et al., 2022a; Jiang et al., 2022; Zhang et al., 2022a,b; Zhuo et al., 2023) propose various data augmentation methods to generate positive pairs and use other in-batch sentences as negatives. Meanwhile, the selection of negative sentences is also crucial. DCLR (Zhou et al., 2022), ClusterNS (Deng et al., 2023), CLSEP (Wang et al., 2023), SSCL (Chen et al., 2023) are focused on utilizing different approaches to provide negative samples. Compared to ICL, we propose a dimension-level contrastive learning mechanism to reduce *dimensional pollution*. Moreover, we devise an information compensation mechanism to compensate for the learned mutual information. This approach is novel and has not been considered in previous studies.

2.2 Information Bottleneck Principle

Information bottleneck (Tishby and Zaslavsky, 2015; Shwartz-Ziv and Tishby, 2017) divides deep learning into two steps: the first step aims to maximize the mutual information between representations and tasks, while the second step aims to compress the input and representation information as much as possible. InforMin-CL (Chen et al., 2022) and InfoCSE (Wu et al., 2022b) propose contrastive learning based on information minimization, minimizing the information entropy between positive pairs. *mi*CSE (Klein and Nabi, 2023) integrates mutual information and incorporates the consistency between attention from different views. However, the above works only consider inter-sentence mutual information, which results in the mutual information being over-compressed. To address this gap, we devise an information compensation mechanism to loosen the IB principle's constraint.

3 Methodology

The sentence embedding task aims to learn high-quality sentence vectors to achieve isotropy in the representation space (Li et al., 2020; Su et al., 2021). To address this task, we propose a novel method, called MSSE, illustrated in Figure 3. Subsequently, we elaborate on the specific techniques in our framework.

3.1 Data Augmentation

Traditional data augmentation operations disrupt key information and introduce a large amount of noise, which weakens the model's learning capability (Gao et al., 2021). Considering that the key information of sentences lies in the semantic information, we adopt an augmentation strategy (Xie et al., 2020) that ensures consistency of information by using back-translation to rewrite sentences to obtain positive samples. In practice, we use pre-

trained translation models[1] to first translate English into German, and then translate the German back into English.

3.2 Dimension-level Contrastive Learning Mechanism

Effective representation learning aims to capture the semantic content of sentences (Schölkopf et al., 2021). When the model maps sentences to vectors, certain dimensions of these vectors consistently represent specific types of information, such as some dimensions encode tense, some encode semantic information, etc (Tenney et al., 2019). However, not all dimensions can perfectly separate this information (Locatello et al., 2019; Wang et al., 2022b). Key information and irrelevant information often spread across multiple dimensions, leading to their confusion and resulting in *dimensional pollution* . When irrelevant information contaminates the dimensions that are supposed to represent key information, the key information gets diluted (Chen and He, 2021), thereby reducing the quality of the sentence embeddings and undermining downstream performance.

To address the issue of *dimensional pollution*, we propose a novel dimension-level contrastive learning mechanism (DCM). We denote the PLMs as $f_{\text{Enc}}()$, and the representation obtained by feeding into the model as $f_{\text{Enc}}(s) \in \mathbb{R}^{1\times D}$, where D is the representation vector's dimension. We have $h_i = f_{\text{Enc}}(s_i)$ and $h_i^+ = f_{\text{Enc}}(s_i^+)$ for $i \in \{1, 2, ..., N\}$. To keep the key information invariant, we optimize $f_{\text{Enc}}()$:

$$\max_{f_{\text{Enc}}} \frac{1}{D} \sum_{i=1}^{D} f_{\text{Rela}} \left(\tilde{h}_i, \tilde{h}_i^+ \right), \tag{2}$$

where $\tilde{h}_i$ and $\tilde{h}_i^+$ denote the normalized i-th column of $H = \left[(h_1)^T, (h_2)^T, \ldots, (h_N)^T \right]^T \in \mathbb{R}^{N\times D}$ and $H^+ = \left[(h_1^+)^T, (h_2^+)^T, \ldots, (h_N^+)^T \right]^T$, respectively. $f_{\text{Rela}}(\cdot)$ represents a function to measure the correlation between representations after data augmentation. And then, we use $f_{\text{Rela}}(\tilde{h}_i, \tilde{h}_j^+)$ to obtain the final matrix factor C_{ij}:

$$C_{ij} = \frac{\tilde{h}_i \cdot \tilde{h}_j^+}{\| \tilde{h}_i \| \| \tilde{h}_j^+ \|}, i, j \in \{1, 2, \ldots, D\}, \tag{3}$$

where $\tilde{h}_j^+$ represent the normalized j-th column of H^+. We consider the same dimensions in H and H^+ as positive pairs ($\tilde{h}_i$ and $\tilde{h}_i^+$), while different dimensions regard as negative pairs ($\tilde{h}_i$ and $\tilde{h}_j^+$).

[1]We use En-De and De-En model to guild the process.

The dimension-level contrastive loss $\mathcal{L}_{\text{DCM}}$ needs to maximize the positive pairs of correlation and minimize the negative pairs of correlation. This selective optimization helps constrain the propagation of information across dimensions and reduces dimensional contamination. In the matrix C, since the same dimensions are treated as positive pairs, we need to maximize the values on the diagonal of the matrix and minimize the off-diagonal values:

$$\mathcal{L}_{\text{DCM}} = \sum_{i,j} (C_{ij} - \delta_{ij})^2, \quad i, j \in \{1, 2, ..., D\}$$
$$\delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \tag{4}$$

where δ_{ij} represents the Kronecker delta function, which equals 1 when $i = j$ and 0 otherwise. C_{ij} denotes the entry in the i-th row and j-th column of the matrix C.

3.3 Information Compensation Mechanism

Some augmentation strategies introduce irrelevant noise and even disrupt the semantic information of sentences (Gao et al., 2021). According to Eq. 1, during the information compression process in contrastive learning, key information may be inappropriately compressed along with the noise, affecting the quality of sentence embeddings. We describe this as *over-compression*, which means that the model accidentally loses some key information while removing noise (Tian et al., 2020; Wang et al., 2022a). Therefore, we devise an information compensation mechanism (ICM) to loosen the information bottleneck. This allows the model to retain more original information while learning representations. Hence, our optimization objective is formulated as below:

$$\max I(h, h^+) + \lambda I(h, s), \tag{5}$$

where $h = f_{\text{Enc}}(s)$ and $h^+ = f_{\text{Enc}}(s^+)$. λ is used to control the increase in $I(h, s)$. To maximize $I(h, h^+)$, we utilize a commonly used form in ICL, the InfoNCE loss (Oord et al., 2018). It has been proven to be a lower bound for mutual information, equivalent to maximizing $\mathcal{L}_{\text{IB}}$ of the information bottleneck principle in Eq. 1 (detailed in Appendix F). The InfoNCE loss is defined as:

$$\mathcal{L}_{\text{NCE}} = -\log \frac{e^{sim(h_i, h_i^+)/\tau}}{\sum_{j=1}^{N} e^{sim(h_i, h_j^+)/\tau}}, \tag{6}$$

where $sim(\cdot, \cdot)$ is the similarity metric, N is the batch size, τ is a temperature hyper-parameter. In

order to maximize the mutual information $I(h, s)$, it can be represented as:

$$\begin{aligned} I(h,s) &= D_{\mathrm{KL}}(p(h,s)||p(h)p(s)) \\ &= \mathbb{E}_{p(h,s)}\left[\log \frac{p(h,s)}{p(h)p(s)}\right] \\ &= \mathbb{E}_{p(h,s)}\left[\log \frac{p(s|h)}{p(s)}\right], \end{aligned} \tag{7}$$

where $D_{\mathrm{KL}}(\cdot)$ represents the KL divergence function, used to measure the distance between distributions. $p(h)$ and $p(s)$ respectively represent the marginal distributions of h and s. $p(s|h)$ is the conditional distribution, which is difficult to estimate. Instead of directly estimating it, we utilize variational approximation to train $q(s|h)$ to approximate the true probability distribution $p(s|h)$. Thus, maximizing the lower bound on mutual information leads to the maximization of $I(h, s)$. We use the BA bound (Barber and Agakov, 2004) to estimate the lower bound on mutual information:

$$\begin{aligned} I(h,s) &= \mathbb{E}_{p(h,s)}\left[\log \frac{q(s|h)}{p(s)}\right] \\ &\quad + \mathbb{E}_{p(h)}[D_{\mathrm{KL}}(p(s|h)||q(s|h))] \\ &\geq \mathbb{E}_{p(h,s)}[\log q(s|h)] + H(s), \end{aligned} \tag{8}$$

where $H(s)$ is the entropy of the original input s, $\mathbb{E}_{s\sim p(s)} \log p(s)$ is only related to the data generation process and is independent of the representation h obtained through model learning. Therefore, we only need to maximize $\mathbb{E}_{p(h,s)}[\log q(s|h)]$, encouraging the model to learn more information containing the original input s, i.e., learning a decoder $q(s|h)$ to maximize the increased mutual information. Thus, we utilize the representation h to reconstruct the original input s, by comparing the difference between the original input and the reconstructed sentence $\hat{s}$ generated by the decoder $q(s|h)$, This approach loosens the constraint of the information bottleneck and compensates for the mutual information. Our compensatory mutual information loss is as follows:

$$\mathcal{L}_{\mathrm{ICM}} = -\sum_{s_i}^{N}\sum_{j=1}^{s_i} \log P(x_j|\hat{x}_1, \hat{x}_2, ..., \hat{x}_{s_i}), \tag{9}$$

where x_j is the j-th word of the s_i-th sentence in the original inputs, and $\hat{x}_j$ is the corresponding part in the reconstructed sentence $\hat{s}_i$.

According to Eq. 4, Eq. 6 and Eq. 9, the above losses can be simply added to form the final loss:

$$\mathcal{L} = \mathcal{L}_{\mathrm{NCE}} + \mu\mathcal{L}_{\mathrm{DCM}} + \gamma\mathcal{L}_{\mathrm{ICM}}, \tag{10}$$

where μ and γ are the hyper-parameters for weights balance.

4 Experiments

4.1 Experiment Setup

We conduct a set of experiments to evaluate our method on seven semantic textual similarity (STS) tasks: STS 2012-2016 (Agirre et al., 2012, 2013, 2014, 2015, 2016), STS Benchmark (Cer et al., 2017) and SICK-Relatedness (Marelli et al., 2014). We preform experiments with backbones of $\mathrm{BERT}_{\mathrm{base}}$ and $\mathrm{BERT}_{\mathrm{large}}$ (Devlin et al., 2019). We also evaluate 7 transfer learning tasks and provide detailed results in Appendix E. Reimers and Gurevych (2019) argue that the primary goal of sentence embedding learning is to cluster semantically similar sentences. Therefore, we use the STS tasks as the main results.

Training Details We evaluate the model every 125 training steps on the development set of STS-B and keep the best checkpoint for the evaluation on test sets of all STS tasks in unsupervised, few-shot, and supervised scenarios. All the experiments are conducted on 2 NVIDIA Tesla A100 GPUs (80GB memory). More training details are in Appendix A.

Datasets Following SimCSE (Gao et al., 2021), we use 1,000,000 (10^6) sentences randomly sampled from Wikipedia as our training corpus. Additionally, we randomly sample fixed proportions $\{0.1\%, 0.5\%, 1\%, 10\%, 100\%\}$ from the dataset to train our model in the few-shot learning setting. Importantly, to eliminate variance in data sampling during few-shot training, we repeat training for each proportion of the dataset five times with different random seeds and report the average results of the final results. For the supervised learning, we use natural language inference (NLI) datasets (Conneau et al., 2017; Reimers and Gurevych, 2019) to train our model. Following Gao et al. (2021), we use the entailment as positive pairs and the contradiction as hard negative pairs.

Baselines To validate the effectiveness of our approach in different settings, we compare MSSE with a range of competitive sentence embedding learning methods.

In unsupervised setting, we compare MSSE with various competitive methods, which include: SimCSE (Gao et al., 2021), DiffCSE (Chuang et al., 2022), DCLR (Zhou et al., 2022), ArcCSE (Zhang et al., 2022b), PCL (Wu et al., 2022a), CompCSE (Chanchani and Huang, 2023), WhitenedCSE (Zhuo et al., 2023), miCSE (Klein and Nabi,

PLMs	Methods	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
$\text{BERT}_{\text{base}}$	SimCSE	68.40	82.41	74.38	80.91	78.56	76.85	72.23	76.25
	DCLR	70.81	83.73	75.11	82.56	78.44	78.31	71.59	77.22
	ArcCSE	72.08	84.27	76.25	82.32	79.54	79.92	72.39	78.11
	*mi*CSE	71.71	83.09	75.46	83.13	80.22	79.70	73.62	78.13
	CompCSE	72.14	84.06	75.38	83.82	80.43	80.29	71.12	78.18
	PCL	72.84	83.81	76.52	83.06	79.32	80.01	73.38	78.42
	DiffCSE	72.28	84.43	76.47	83.90	80.54	80.59	71.23	78.49
	WhitenedCSE	74.03	84.90	76.40	83.40	80.23	81.14	71.33	78.78
	OssCSE	71.78	84.40	77.71	83.95	79.92	80.57	75.25	79.08
	DistillCSE	74.54	84.51	77.67	84.87	80.70	81.48	72.16	79.42
	RankCSE †	75.21	**85.80**	77.45	84.17	**80.77**	81.21	74.81	79.92
	RankEncoder	74.88	85.59	**78.61**	83.50	80.56	**81.55**	**75.78**	80.07
	MSSE	**76.43**	84.92	78.49	**85.47**	80.11	81.32	74.30	**80.15**
$\text{BERT}_{\text{large}}$	SimCSE	70.88	84.16	76.43	84.50	79.76	79.26	73.88	78.41
	DCLR	71.87	84.83	77.37	84.70	79.81	79.55	74.19	78.90
	ArcCSE	73.17	86.19	77.90	84.97	79.43	80.45	73.50	79.37
	CompCSE	75.10	86.57	77.70	84.72	80.25	80.17	73.21	79.67
	WhitenedCSE	74.65	85.79	77.49	84.71	80.33	81.48	75.34	79.97
	PCL	74.87	86.11	78.29	85.65	80.52	81.62	73.94	80.14
	RankCSE †	75.24	86.17	78.67	85.11	81.12	81.30	75.27	80.41
	OssCSE	72.64	86.36	79.16	85.04	80.80	82.61	**76.65**	80.47
	DistillCSE	75.08	**86.64**	79.53	86.45	81.29	82.72	76.17	81.13
	MSSE	**76.95**	85.58	**79.71**	**86.59**	**81.33**	**82.96**	76.50	**81.36**

Table 1: Sentence representations performance on STS tasks in unsupervised setting. We directly use the results from the original papers except for †. †: reproduce the results using publicly available code without two teacher models to ensure a fair comparison. We mark the best (bold) and second-best (underlined) results among methods with the same PLMs.

2023), OssCSE (Shi et al., 2023), DistillCSE (Xu et al., 2023), RankEncoder (Seonwoo et al., 2023) and RankCSE (Seonwoo et al., 2023).

In the few-shot setting, only *mi*CSE has the same experimental settings as ours. Therefore, we select some methods from the past three years' works and reproduce them for evaluation: SimCSE, PCL, ISCSE (He et al., 2023), RankEncoder, and RankCSE.

In the supervised setting, due to the limited previous works considering this setting, we select some models which have this experimental setting as baselines, such as SimCSE, PromCSE (Jiang et al., 2022) and CLSEP (Wang et al., 2023). Additionally, we reproduce other works, such as PCL and ISCSE, for comparison.

By comparing MSSE with these baselines, we can more accurately assess its performance and ensure its superiority in unsupervised, few-shot, and supervised scenarios. We provide more details of these baselines in Appendix B.

4.2 Main Results

Unsupervised Sentence Embeddings We conduct our experiments on 7 STS tasks and evaluate them using the SentEval toolkit (Conneau and Kiela, 2018). To ensure fairness, we follow the standards of Gao et al. (2021), using Spearman's correlation coefficient as the evaluation metric. Table 1 shows different methods' performances, it is clear that MSSE significantly outperforms the previous SOTA methods on all PLMs, which demonstrates the effectiveness of our method. What's more, compared to SimCSE-$\text{BERT}_{\text{base}}$, MSSE-$\text{BERT}_{\text{base}}$ increases the performance from 76.25% to 80.15% (+3.90%), and compared to SimCSE-$\text{BERT}_{\text{large}}$, MSSE-$\text{BERT}_{\text{large}}$ increases the performance from 78.41% to 81.36% (+2.95%). Specifically, MSSE-$\text{BERT}_{\text{base}}$ achieves on average 1.74% absolute improvements in terms of Spearman's correlation on SimCSE-$\text{BERT}_{\text{large}}$.

Few-shot Sentence Embeddings In the few-shot setting, we utilize $\text{BERT}_{\text{base}}$ to retrain methods and meticulously evaluate their performance. With the increase in dataset size, the model's average performance consistently improves and MSSE consistently outperforms other methods. As shown in Table 2, on a dataset containing only 0.1% of the data volume, MSSE achieves outstanding performance, with an average performance of 75.81%. Compared to the SimCSE (65.72%), MSSE achieves an absolute performance gain of 10.09%. This precisely reflects how MSSE efficiently learns mutual information during training. It utilizes a dimension-

Methods	0.1%	0.5%	1%	10%	100%
SimCSE†	65.72±0.25	73.67±1.99	74.72±1.41	75.08±0.40	76.13
ISCSE†	67.30±1.15	75.69±0.56	76.27±0.37	76.46±0.52	78.07
PCL†	70.33±0.78	74.08±0.53	75.46±1.17	76.53±0.92	78.21
*mi*CSE♣	73.68±0.89	75.15±0.63	76.40±0.48	76.38±0.35	78.13
RankEncoder†	72.61±0.72	75.66±0.90	76.94±0.39	77.84±0.69	79.81
RankCSE†	74.39±1.09	76.02±0.46	77.91±0.86	78.67±0.27	79.92
MSSE	**75.81±0.41**	**77.89±0.37**	**78.78±0.34**	**79.11±0.22**	**80.15**

Table 2: Few-shot sentence representations average performance on 7 STS tasks (Spearman's correlation) based on $\text{BERT}_{\text{base}}$. ♣: results from the original papers; †: reproduce the results based on publicly available code. In addition, the results provided by *mi*CSE lack the result of 0.5%, we only reproduce the result of 0.5%.

Methods	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
CT-SBERT♡	74.84	83.20	78.07	83.84	77.93	81.46	76.42	79.39
SimCSE♡	75.30	84.67	80.19	85.40	80.82	84.25	80.39	81.57
PromCSE♣	75.58	84.33	79.67	85.79	81.24	84.25	80.79	81.81
PCL†	76.21	84.38	79.64	85.98	81.18	84.97	81.00	81.91
ISCSE†	76.22	83.97	79.82	86.18	81.67	85.43	81.05	82.05
CLSEP♣	75.76	84.82	80.30	86.29	81.43	**85.58**	80.99	82.17
MSSE	**76.82**	**85.32**	**80.61**	**87.33**	**82.45**	85.50	**81.93**	**82.85**

Table 3: Supervised sentence representations performance on STS tasks of different methods. ♡: results from Gao et al. (2021); ♣: results from the original papers; †: reproduced by ourselves. All methods are based on $\text{BERT}_{\text{base}}$.

level contrastive learning mechanism to reduce *dimensional pollution* and compensates for the over-compressed mutual information learned by contrastive learning. This also demonstrates the excellent performance of our approach.

Supervised Sentence Embeddings As shown in Table 3, compared to the previous methods, our supervised MSSE-$\text{BERT}_{\text{base}}$ further improves the SOTA results from 82.17% to 82.85% (+0.68%).

4.3 Analysis and Discussion

Ablation Study We conduct a set of ablation studies to investigate the impact of $\mathcal{L}_{\text{NCE}}$, $\mathcal{L}_{\text{DCM}}$ and $\mathcal{L}_{\text{ICM}}$ from Eq. 10. Table 4 reports the average results of the STS tasks. After only removing $\mathcal{L}_{\text{NCE}}$, $\mathcal{L}_{\text{DCM}}$ or $\mathcal{L}_{\text{ICM}}$, the average performance of MSSE decreases by 4.01%, 1.43%, or 2.05%, respectively. This indicates that the proposed two novel mechanisms both contribute to learning sentence embeddings, while traditional instance-level contrastive learning can capture mutual information between different views. If both $\mathcal{L}_{\text{DCM}}$ and $\mathcal{L}_{\text{ICM}}$ are removed simultaneously, the average performance drops by 3.27%, demonstrating the complementary nature of the two modules in advancing the learning of sentence embeddings. More ablation studies(Pooler methods, Augmentation strategies, Hyper-parameters) are provided in Appendix D due to the page limit.

	STS(Avg.)
MSSE	80.15
w/o $\mathcal{L}_{\text{DCM}}$	78.72 (-1.43)
w/o $\mathcal{L}_{\text{ICM}}$	78.10 (-2.05)
w/o $\mathcal{L}_{\text{NCE}}$	76.14 (-4.01)
w/o $\mathcal{L}_{\text{DCM}}+\mathcal{L}_{\text{ICM}}$	76.88 (-3.27)

Table 4: Ablation studies of different loss functions using $\text{BERT}_{\text{base}}$. NCE, DCM and ICM denote the InfoNCE loss used in the instance-level contrastive learning, the Dimension-level Contrastive Learning Mechanism and the Information Compensation Mechanism, respectively.

Impact of the DCM The t-SNE (Reif et al., 2019) plot in Figure 4 demonstrates the advantages of the dimension-level contrastive learning mechanism. We evaluate sentence embeddings using the original $\text{BERT}_{\text{base}}$, RankCSE, and MSSE on 20,000 sentences from the Stackoverflow (Xu et al., 2017) dataset. We apply K-Means clustering to group similar sentence embeddings. The results in Figure 4 show that when we remove the DCM module, the resulting similar sentence pairs (marked with the same color) do not cluster, reflecting that *dimensional pollution* indeed affects the performance of sentence embeddings. However, when we add the DCM module, the resulting similar sentence pairs are better aligned and more clustered.

Impact of the ICM To validate the effectiveness of the information compensation mechanism, we

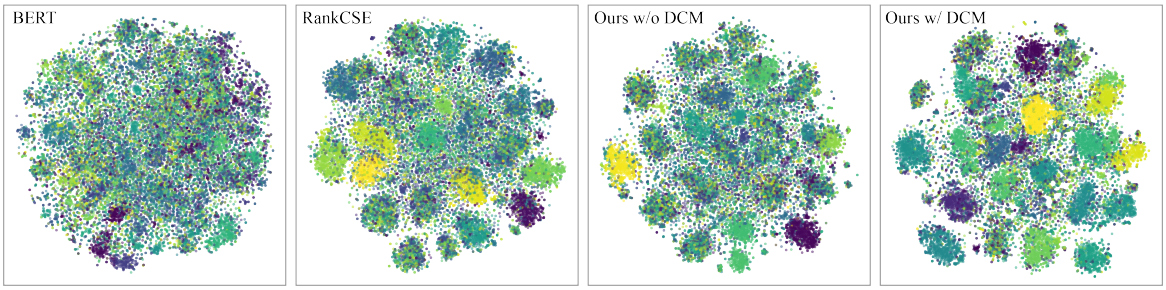


Figure 4: The t-SNE of sentence representations learned with BERT, RankCSE and our method using $\text{BERT}_{\text{base}}$. The points are embeddings of sentences sampled from the Stackoverflow (Xu et al., 2017) dataset. We use K-Means clustering to group similar sentence embeddings and form 30 clusters. (Best viewed in color)

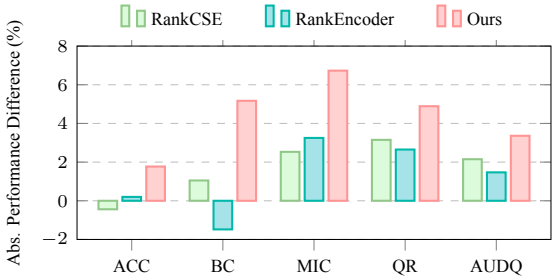
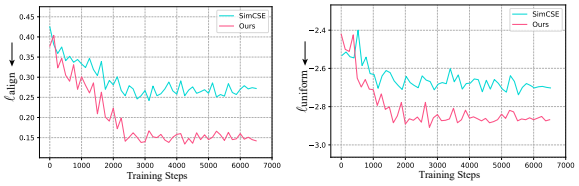


Figure 5: Absolute performance difference on classification, reranking and retrieval tasks compared to SimCSE based on $\text{BERT}_{\text{base}}$. ACC, BC, MIC, QR and AUDQ denote Amazon Counterfactual Classification (O'Neill et al., 2021), Banking77 Classification (Casanueva et al., 2020), Massive Intent Classification (FitzGerald et al., 2023), Quora Retrieval (Sharma et al., 2019), AskUbuntu DupQuestions (Lei et al., 2016).



(a) alignment (b) uniformity

Figure 6: ℓ_{align} and ℓ_{uniform} of our method and SimCSE based on $\text{BERT}_{\text{base}}$. For both measures, lower number are better.

conduct a comprehensive set of experiments on classification, retrieval and reranking tasks from Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2023). As shown in Figure 5, our approach consistently outperforms SimCSE across the five tasks. Compared to the current SOTA methods, such as RankEncoder (Seonwoo et al., 2023) and RankCSE (Liu et al., 2023), our method demonstrates superior performance and robustness across various tasks and domains, further validating the effectiveness of the ICM in loosening the constraint of the information bottleneck and introducing more information. Additionally, we also conduct experiments on retrieval tasks for qualitative analysis and the results are provided in Appendix G.

Alignment and Uniformity Prior work (Wang and Isola, 2020) has demonstrated that models with better *alignment* and *uniformity* can achieve better performance (detailed in Appendix C). We calculate the alignment and uniformity loss on the STS-B development set every 125 training steps. Figure 6 shows that compared with SimCSE, our approach performs better both on the *alignment* measure and the *uniformity* measure. This confirms that our method can improve the quality of sentence representation more effectively. MSSE promotes *alignment* and *uniformity* of information through the dimension-level contrastive learning mechanism to alleviate the *dimensional pollution*, while the information compensation mechanism utilizes the final representations to guide the reconstruction of inputs, loosening the IB principle and compensating for the over-compressed mutual information learned by sentence embeddings.

5 Conclusion

In this work, we propose MSSE, a novel sentence embedding approach, which is applicable to unsupervised, few-shot, and supervised learning settings. MSSE enhances the model's ability to learn mutual information by utilizing the dimension-level contrastive learning mechanism and the information compensation mechanism based on information bottleneck. Experimental results demonstrate that MSSE outperforms previous SOTA methods in all settings. Additionally, we conduct comprehensive ablation experiments and analyses to demonstrate the effectiveness of each component and the rationale behind our approach.

Limitations

In this paper, the limitations of our work are as follows. Firstly, this work follows the standard experimental settings used in previous unsupervised

sentence representation learning works (Gao et al., 2021), but it does not consider the multi-modal scenario. We plan to extend MSSE to multi-modal datasets, such as Flickr30k (Young et al., 2014a), and MS-COCO (Lin et al., 2014), to obtain more new discoveries in the future. Secondly, the performance of sentence embeddings needs to be evaluated through downstream tasks such as STS, which lacks a certain degree of interpretability. Providing interpretability for sentence embeddings is also our next research direction.

Ethics Statement

We focuse on sentence embedding learning and propose a novel multi-scenario sentence embedding enhancement method. What's more, the training corpus and benchmark datasets are open-source, containing no personal sensitive information and no potential malicious content. In practice, we use back-translation for augmentation to obtain positive samples, which has no impact on social and does not involve any ethical issues. Furthermore, we are willing to open-source our code and data to promote better research in this field.

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.

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.

David Barber and Felix Agakov. 2004. The im algorithm: a variational approach to information maximization. *Advances in neural information processing systems*, 16(320):201.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual information neural estimation. In *Proceedings of International conference on machine learning*, pages 531–540.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45.

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.

Sachin Chanchani and Ruihong Huang. 2023. Composition-contrastive learning for sentence embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15836–15848.

Nuo Chen, Linjun Shou, Jian Pei, Ming Gong, Bowen Cao, Jianhui Chang, Jia Li, and Daxin Jiang. 2023. Alleviating over-smoothing for unsupervised sentence representation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3552–3566.

Shaobin Chen, Jie Zhou, Yuling Sun, and Liang He. 2022. An information minimization based contrastive learning model for unsupervised sentence embeddings learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4821–4831.

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.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218.

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

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

Jinghao Deng, Fanqi Wan, Tao Yang, Xiaojun Quan, and Rui Wang. 2023. Clustering-aware negative sampling for unsupervised sentence representation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8713–8729.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 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)*.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302.

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.

Hongliang He, Junlei Zhang, Zhenzhong Lan, and Yue Zhang. 2023. Instance smoothed contrastive learning for unsupervised sentence embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12863–12871.

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.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

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.

Yuxin Jiang, Linhan Zhang, and Wei Wang. 2022. Improved universal sentence embeddings with prompt-based contrastive learning and energy-based learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3021–3035.

Tassilo Klein and Moin Nabi. 2023. miCSE: Mutual information contrastive learning for low-shot sentence embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6159–6177.

Tao Lei, Hrishikesh Joshi, Regina Barzilay, Tommi Jaakkola, Kateryna Tymoshenko, Alessandro Moschitti, and Lluís Màrquez. 2016. Semi-supervised question retrieval with gated convolutions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1279–1289.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755.

Jiduan Liu, Jiahao Liu, Qifan Wang, Jingang Wang, Wei Wu, Yunsen Xian, Dongyan Zhao, Kai Chen, and Rui Yan. 2023. RankCSE: Unsupervised sentence representations learning via learning to rank. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13785–13802.

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*.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124.

Shutian Ma, Chengzhi Zhang, and Daqing He. 2016. Document representation methods for clustering bilingual documents. *Proceedings of the Association for Information Science and Technology*, 53(1):1–10.

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. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223.

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

James O'Neill, Polina Rozenshtein, Ryuichi Kiryo, Motoko Kubota, and Danushka Bollegala. 2021. I wish I would have loved this one, but I didn't – a multilingual dataset for counterfactual detection in product review. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7092–7108.

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. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278.

Bo Pang and Lillian Lee. 2005. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, page 115–124.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. *Advances in Neural Information Processing Systems*, 32.

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.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634.

Yeon Seonwoo, Guoyin Wang, Changmin Seo, Sajal Choudhary, Jiwei Li, Xiang Li, Puyang Xu, Sunghyun Park, and Alice Oh. 2023. Ranking-enhanced unsupervised sentence representation learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15783–15798.

Lakshay Sharma, Laura Graesser, Nikita Nangia, and Utku Evci. 2019. Natural language understanding with the quora question pairs dataset. *arXiv preprint arXiv:1907.01041*.

Zhan Shi, Guoyin Wang, Ke Bai, Jiwei Li, Xiang Li, Qingjun Cui, Belinda Zeng, Trishul Chilimbi, and Xiaodan Zhu. 2023. Osscse: Overcoming surface structure bias in contrastive learning for unsupervised sentence embedding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7242–7254.

Ravid Shwartz-Ziv and Naftali Tishby. 2017. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*.

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. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 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*.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1555–1565.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839.

Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *Proceedings of 2015 ieee information theory workshop (itw)*, pages 1–5.

Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Self-supervised learning from a multi-view perspective. *arXiv preprint arXiv:2006.05576*.

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

Haoqing Wang, Xun Guo, Zhi-Hong Deng, and Yan Lu. 2021. What makes for good representations for contrastive learning.

Haoqing Wang, Xun Guo, Zhi-Hong Deng, and Yan Lu. 2022a. Rethinking minimal sufficient representation in contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16041–16050.

Qian Wang, Weiqi Zhang, Tianyi Lei, Yu Cao, Dezhong Peng, and Xu Wang. 2023. Clsep: Contrastive learning of sentence embedding with prompt. *Knowledge-Based Systems*, 266:110381.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of International conference on machine learning*, pages 9929–9939.

Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. 2022b. Disentangled representation learning. *arXiv preprint arXiv:2211.11695*.

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

Qiyu Wu, Chongyang Tao, Tao Shen, Can Xu, Xiubo Geng, and Daxin Jiang. 2022a. PCL: Peer-contrastive learning with diverse augmentations for unsupervised sentence embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 12052–12066.

Xing Wu, Chaochen Gao, Zijia Lin, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2022b. InfoCSE: Information-aggregated contrastive learning of sentence embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3060–3070.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

Jiahao Xu, Wei Shao, Lihui Chen, and Lemao Liu. 2023. DistillCSE: Distilled contrastive learning for sentence embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8153–8165.

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.

Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. 2019. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6210–6219.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014a. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014b. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xuejie Zhang. 2017. Refining word embeddings for sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 534–539.

Yanzhao Zhang, Richong Zhang, Samuel Mensah, Xudong Liu, and Yongyi Mao. 2022a. Unsupervised sentence representation via contrastive learning with mixing negatives. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11730–11738.

Yuhao Zhang, Hongji Zhu, Yongliang Wang, Nan Xu, Xiaobo Li, and Binqiang Zhao. 2022b. A contrastive framework for learning sentence representations from pairwise and triple-wise perspective in angular space. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4892–4903.

Kun Zhou, Beichen Zhang, Xin Zhao, and Ji-Rong Wen. 2022. Debiased contrastive learning of unsupervised sentence representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6120–6130.

Wenjie Zhuo, Yifan Sun, Xiaohan Wang, Linchao Zhu, and Yi Yang. 2023. WhitenedCSE: Whitening-based

contrastive learning of sentence embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12135–12148.

A Training Details

We preform experiments with backbones of $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{large}}$. We implement all experiments with the deep learning framework PyTorch on 2 NVIDIA Tesla A100 GPUs (80GB memory). In the unsupervised scenario, as shown in Table 5, the learning rate for $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{large}}$ are set to 3e-5 and 1e-5. The batch size for $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{large}}$ are both set to 256. We use AdamW as the optimizer with a warm-up step of 500. τ is set to 0.05, μ and γ are set to 0.8 and 0.2, respectively. We train our model for 1 epoch and evaluate the model every 125 steps. In the few-shot scenario, we adopt the same training method as miCSE (Klein and Nabi, 2023), keeping the total number of optimization steps unchanged for training different dataset sizes. For the training set of size 10^6 (100%), we train for 1 epoch; for the size 10^5 (10%), we train for 10 epochs, etc. In the supervised learning scenario, we follow the setup of SimCSE (Gao et al., 2021) and utilize natural language inference (NLI) dataset (Conneau et al., 2017; Reimers and Gurevych, 2019) as the training corpus, with the batch size set to 512 and the learning rate set to 5e-5.

B Baselines Details

We compare MSSE with the following SOTA sentence embedding methods:

- SimCSE (Gao et al., 2021) conducts thorough experiments in both unsupervised,few-shot, and supervised settings using different dropout encodings to obtain positive pairs. The results are from the original paper. We rerun SimCSE with the same settings, but its performance (76.13) is worse than the reported number in the original paper.

- DiffCSE (Chuang et al., 2022) learns the differences between original and fake sentences by generating fake samples using the ELECTRA model and Replaced Token Detection (RTD) task to enhance the effectiveness of sentence vector representation models.

- DCLR (Zhou et al., 2022) designs an instance-weighting method to penalize false negatives and generate noise-based negatives to ensure the uniformity of the representation space.

PLMs	$\text{BERT}_{\text{base}}$		$\text{BERT}_{\text{large}}$	
	Unsup.	Sup.	Unsup.	Sup.
Batch size	256	512	256	512
Learning rate	3e-5	5e-5	1e-5	5e-5

Table 5: Batch sizes and learning rates for DEMI.

- ArcCSE (Zhang et al., 2022b) enhances the discriminability of positive and negative samples by maximizing the decision margin in the angular space. It also models the semantic partial order between sentences by automatically constructing ternary sentences and their entailment relationships.

- PCL (Wu et al., 2022a) introduces a novel companion contrastive learning with various enhancement functions to construct different positive and negative pairs for unsupervised sentence embeddings.

- ISCSE (He et al., 2023) retrieves embeddings from a dynamic memory buffer based on semantic similarity to obtain positive embedding groups, then aggregates embeddings in the group through self-attention operations to generate smooth instance embeddings.

- CompCSE (Chanchani and Huang, 2023) extracts atomic semantic units using a discourse parser, then maximizes the alignment between text and its phrase components to enhance performance.

- miCSE (Klein and Nabi, 2023) proposes a contrastive learning framework based on mutual information to improve the efficiency of unsupervised learning by enhancing the consistency between attention across different views.

- OssCSE (Shi et al., 2023) considers surface structural deviations and balances learning objectives and word semantics by using a data offsetting bias and recall loss.

- DistillCSE (Xu et al., 2023) provides additional supervised signals using the base model and proposes two knowledge distillation solutions to learn stronger representations.

- CLSEP (Wang et al., 2023) employs a prompting mechanism to provide effective sentence embeddings and introduces Partial Word Vector Augmentation (PWVA), a text data augmentation strategy. This strategy enhances the data in the word embedding space, preserving more semantic information.

- RankEncoder (Seonwoo et al., 2023) introduces a novel unsupervised sentence encoder that predicts the semantic vectors of input sentences based on their relationships with other sentences in an external corpus and the input sentences themselves.

- RankCSE (Liu et al., 2023) addresses the inability of previous works to obtain fine-grained ranking information, proposes ranking consistency and ranking distillation methods, and integrates them with contrastive learning into one framework. RankCSE utilizes pre-trained ranking models, SimCSE and DiffCSE, as teacher models during training, providing a certain level of supervision. To ensure a fair comparison, we reproduce the results by removing the two pre-trained teacher models.

C Alignment and Uniformity

Contrastive representation learning has two key properties: (1) *alignment* of positive pairs; (2) *uniformity* on the hypersphere. Wang and Isola (2020) argues that directly optimizing these two metrics can lead to representations with performance comparable to or better than contrastive learning in downstream tasks. *Alignment* measures the expected distance between normalized representations of positive pairs p_{pos}:

$$\ell_{\text{align}} \triangleq \mathop{\mathbb{E}}_{(x,x^+)\sim p_{\text{pos}}} \|f(x) - f(x^+)\|^2, \qquad (11)$$

while *uniformity* measures the uniform distribution of normalized representations:

$$\ell_{\text{uniform}} \triangleq \log \mathop{\mathbb{E}}_{x,y \overset{i.i.d.}{\sim} p_{\text{data}}} e^{-2\|f(x)-f(y)\|^2}, \qquad (12)$$

where p_{data} represents the distribution of sentence pairs. Smaller values for both metrics are better, which aligns closely with the objectives of contrastive learning: positive instances should be as close as possible, indicating smaller alignment, while random instances should be scattered on the hypersphere, indicating smaller uniformity.

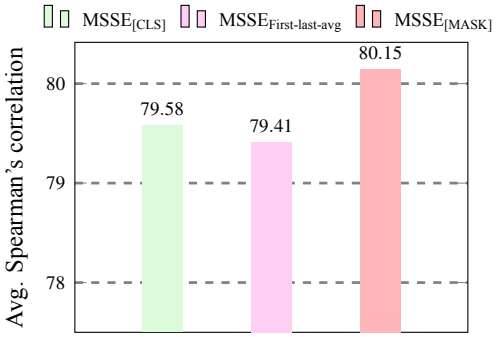


Figure 7: Ablation studies of different pooling methods in unsupervised MSSE based on $\text{BERT}_{\text{base}}$.

D Ablation Studies

We also investigate the impact of different pooling methods, data augmentation strategies, and hyperparameters. We use the average results of the 7 STS tasks as our final report results.

Pooling Methods Most previous works use the [CLS] representation as the final sentence embedding. However, Reimers and Gurevych (2019); Gao et al. (2021) demonstrate that using the first-last-average embedding of a pre-trained model (particularly the first and last layers) can yield better performance than [CLS]. To evaluate the impact of different pooling strategies on performance, we conduct comparative experiments with various pooling methods under unsupervised settings.

As shown in Figure 7, using the [MASK] embeddings, our approach outperforms both [CLS] and first-last-average embeddings. Considering the characteristics of our approach, the information compensation mechanism loosens the IB principle and compensates for the over-compressed mutual information by reconstructing the original sentence through the final representation. Therefore, the [MASK] representation can obtain better performance than [CLS] and first-last-average embeddings.

Augmentation Strategies To investigate the impact of different augmentation strategies on the performance of MSSE in generating positive samples, we also conduct a set of comparative experiments using some traditional augmentation strategies according to PCL (Wu et al., 2022a): shuffled Sentence (SS), word deletion (WD), word repetition (WR), dropout (DP) and back-translation (BT).

Figure 8 shows the average Spearman's correlation performance for the 7 STS tasks using different augmentation strategies. The experimental results indicate that back-translation indeed outperforms

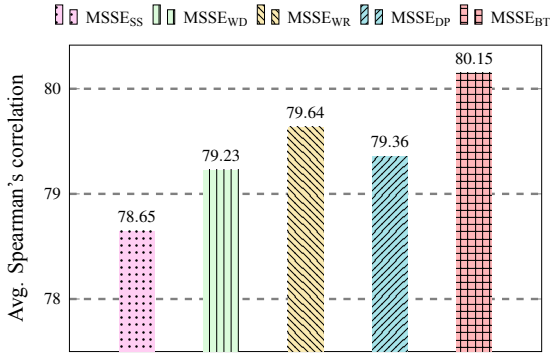


Figure 8: The average performance of Spearman's correlation on 7 STS tasks obtained using different augmentation strategies based on BERT$_{\text{base}}$. MSSE$_{\text{BT}}$ is the augmentation strategy used in our approach.

Batch Size	**32**	**64**	**128**	**256**	**512**
STS(Avg.)	78.80	79.37	79.96	80.15	80.01
τ	**0.01**	**0.02**	**0.05**	**0.10**	**0.20**
STS(Avg.)	79.62	79.94	80.15	79.97	79.43

Table 6: Comparisons of different batch sizes and temperature parameters. Results of MSSE are average STS performance based on BERT$_{\text{base}}$.

traditional augmentation strategies. As mentioned in the main text, the aforementioned strategies introduce noise and irrelevant information, disrupting the sentence structure and semantic information, thereby limiting the model's performance. What's more, back-translation not only avoids introducing more noise but also enriches sentence information, which complements our proposed dimension-level contrastive learning mechanism.

Hyper-parameters To study the influence of hyper-parameters on STS average performance, we conduct experiments by setting different batch sizes and different temperature hyper-parameters. As shown in Table 6, the optimal batch size is 256. With the increase in batch size, the average performance of the model improves, but when the batch size exceeds 256, the average performance of the model significantly decreases. It also shows that the temperature setting for MSSE should be moderate, with the optimal temperature for BERT$_{\text{base}}$ being 0.05.

E Transfer Tasks

For the transfer learning (TR) task, we evaluate 7 datasets using SentEval's default configuration: 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). For each task, we train a logistic regression classifier on the frozen sentence embeddings and test the classification accuracy.

As shown in Table 7, the results demonstrate that MSSE outperforms other competitive SOTA baselines, both on BERT$_{\text{base}}$ and BERT$_{\text{large}}$. Compared to SimCSE-BERT$_{\text{base}}$ (85.81%), MSSE-BERT$_{\text{base}}$ (87.43%) achieves an absolute improvement of 1.62%. On BERT$_{\text{large}}$, MSSE achieves the best performance across five transfer tasks and also significantly outperforms the previous SOTA methods.

F Estimating the Mutual Information with InfoNCE

For the expression $I(h, h^+)$, we have:

$$I(h, h^+) = \sum_{h,h^+} p(h, h^+) \log \frac{p(h^+|h)}{p(h^+)}, \tag{13}$$

where $\frac{p(h^+|h)}{p(h^+)}$ is the density ratio defined by Oord et al. (2018). The definition of the InfoNCE loss is:

$$\begin{aligned} \mathcal{L}_{\text{NCE}} &= -\sum_{H} \left[p(h, h^+) \log \frac{f_k(h^+, h)}{\sum_{h_j^+ \in H} f_k(h_j^+, h)} \right] \\ &= -\mathbb{E}_H \left[log \frac{f_k(h^+, h)}{\sum_{h_j^+ \in H} f_k(h_j^+, h)} \right], \end{aligned} \tag{14}$$

where $f_k(h^+, h)$ represents the quantification of the similarity between the predicted result h^+ and the ground truth h. Oord et al. (2018) demonstrated that $f_k(h^+, h)$ is positively correlated with the density ratio. Therefore, we have:

$$f_k(h^+, h) \propto \frac{p(h^+|h)}{p(h^+)}. \tag{15}$$

Based on the above Eq.13, Eq.14, Eq.15, we partition the data into $H = \{H_{\text{pos}} + H_{\text{neg}}\}$, H_{neg} includes $N - 1$ negative samples from the same

PLMs	Methods	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	Avg.
$\text{BERT}_{\text{base}}$	SimCSE	81.18	86.46	94.45	88.88	85.50	89.80	74.43	85.81
	ArcCSE	79.91	85.25	**99.58**	89.21	84.90	89.20	74.78	86.12
	PCL	80.11	85.25	94.22	89.15	85.12	87.40	**76.12**	85.34
	RankCSE†	83.07	87.91	94.98	**89.65**	**88.91**	89.60	76.02	87.16
	MSSE	**84.10**	**88.54**	94.17	89.07	88.67	**91.60**	75.81	**87.43**
$\text{BERT}_{\text{large}}$	SimCSE	85.36	89.38	95.39	89.63	90.44	91.80	76.41	88.34
	ArcCSE	84.34	88.82	**99.58**	89.79	90.50	92.00	74.78	88.54
	PCL	82.47	87.87	95.04	89.59	87.75	**93.00**	76.00	87.39
	RankCSE †	85.47	89.15	94.93	90.42	90.56	**93.00**	76.81	88.62
	MSSE	**85.84**	**90.10**	94.73	**90.69**	**91.05**	92.30	**77.04**	**88.82**

Table 7: Sentence representations performance on seven transfer tasks. We report the accuracy results based on $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{large}}$. The results are imported from the original papers except for †. We also mark the best (bold) and second-best (underlined) results among methods with the same PLMs.

batch. We reformulate $\mathcal{L}_{\text{NCE}}$ as follows:

$$\begin{aligned}
\mathcal{L}_{\text{NCE}} &= -\mathbb{E}_H \log \left[\frac{\frac{p(h^+|h)}{p(h^+)}}{\frac{p(h^+|h)}{p(h^+)} + \sum_{h_j^+ \in H_{\text{neg}}} \frac{p(h_j^+|h)}{p(h_j^+)}} \right] \\
&= \mathbb{E}_H \log \left[1 + \frac{p(h^+)}{p(h^+|h)} \sum_{h_j^+ \in H_{\text{neg}}} \frac{p(h_j^+|h)}{p(h_j^+)} \right] \\
&\approx \mathbb{E}_H \log \left[1 + \frac{p(h^+)}{p(h^+|h)} (N-1) \mathbb{E} \frac{p(h_j^+|h)}{p(h_j^+)} \right] \\
&= \mathbb{E}_H \log \left[1 + \frac{p(h^+)}{p(h^+|h)} (N-1) \right] \\
&\geq \mathbb{E}_H \log \left[\frac{p(h^+)}{p(h^+|h)} N \right] \\
&= \log(N) - I(h^+, h),
\end{aligned} \tag{16}$$

where $p(h^+)$ represents the marginal distribution of h^+, and $p(h^+|h)$ is the conditional distribution of h^+ given h. N reprensentes the batch size. Based on the above Eq. 16 , we can infer:

$$I(h^+, h) = \log(N) - \mathcal{L}_{\text{NCE}}. \tag{17}$$

Thus, $\mathcal{L}_{\text{NCE}}$ can be regarded as a lower bound of $I(h^+, h)$, and its tightness increases with the growth of N.

In practice, we establish a connection between $f_k(h^+, h)$ and adopot cosine similarity $sim(\cdot, \cdot)$ for measure metric with a temperature hyper-parameter. Hence, we have:

$$\begin{aligned}
\mathcal{L}_{\text{NCE}} &= -\mathbb{E}_H \left[log \frac{f_k(h^+, h)}{\sum_{h_j^+ \in H} f_k(h_j^+, h)} \right] \\
&= -\mathbb{E}_{h_i} \left[log \frac{e^{sim(h_i, h_i^+)/\tau}}{\sum_{j=1}^N e^{sim(h_i, h_j^+)/\tau}} \right] \\
&= -\frac{1}{n} \sum_{h_i}^{H} \log \frac{e^{sim(h_i, h_i^+)/\tau}}{\sum_{j=1}^N e^{sim(h_i, h_j^+)/\tau}} \\
&= -\log \frac{e^{sim(h_i, h_i^+)/\tau}}{\sum_{j=1}^N e^{sim(h_i, h_j^+)/\tau}}.
\end{aligned} \tag{18}$$

G Qualitative Analysis

We conduct small-scale retrieval experiments using RankCSE and MSSE based on $\text{BERT}_{\text{base}}$. We use 30k captions from the Flickr30k (Young et al., 2014b) dataset as the retrieval data and randomly select any sentence from them as a query to retrieve the Top-3 similar sentences (based on cosine similarity). As shown in Table 8, the retrieval results demonstrate that sentences retrieved by MSSE are semantically closer to the query sentences and of higher quality compared to those retrieved by RankCSE, further demonstrating the effectiveness of MSSE.

	RankCSE-BERT$_{\text{base}}$	**MSSE-BERT$_{\text{base}}$**
Query: A group of men climb ladders outdoors.		
#1	Two people standing on a roof while another climbs a ladder.	Two people standing on a roof while another climbs a ladder.
#2	A firefighter climbs a ladder towards the fire above him.	Two men sitting on the roof of a house while another one stands on a ladder.
#3	A person is climbing a wooden ladder up a rocky ledge.	Three people in t-shirt, yellow helmets and harnesses begin to climb ladder.
Query: A man in a white cap and shirt plays the violin with other street performers.		
#1	A man in a white shirt is playing the flute to someone in a red skirt.	A man in a white shirt is playing the flute to someone in a red skirt.
#2	A man in a white shirt plays an electric violin.	A man in a white shirt plays an electric violin.
#3	A man in a red shirt plays the guitar.	A man with glasses wearing a tie plays the violin.
Query: A man in a black outfit poses in front of the eiffel tower.		
#1	A man carrying trinkets with the Eiffel tower in the background.	A man carrying trinkets with the Eiffel tower in the background.
#2	A man wearing black jacket poses with a smile.	A man in formal wear is posing in front of a building.
#3	A man wearing a black long-sleeved shirt is taking a photo of a building.	A man wearing a black long-sleeved shirt is taking a photo of a building.
Query: Two women wearing ceremonial costumes are walking outside a white building.		
#1	Two women wearing blue jeans are walking outside.	Two women wearing dresses are walking by a building.
#2	Two women wearing dresses are walking by a building.	Two people are wearing flower costumes and walking down a street.
#3	Men in traditional dress stand outside .	Two women wearing skirts and heels walking down a sidewalk.

Table 8: Retrieval examples of retrieved Top-3 sentences from queries by RankCSE and MSSE from Flickr30k dataset (30k sentences).