ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230584

Identical and Fraternal Twins: Fine-Grained Semantic Contrastive Learning of Sentence Representations

Qingfa Xiao[a,c], Shuangyin Li[a;*] and Lei Chen[b,c]

[a]South China Normal University
[b]The Hong Kong University of Science and Technology
[c]The Hong Kong University of Science and Technology (Guangzhou)
qingfaxiao@m.scnu.edu.cn, shuangyinli@scnu.edu.cn, leichen@cse.ust.hk

Abstract.

The enhancement of unsupervised learning of sentence representations has been significantly achieved by the utility of contrastive learning. This approach clusters the augmented positive instance with the anchor instance to create a desired embedding space. However, relying solely on the contrastive objective can result in sub-optimal outcomes due to its inability to differentiate subtle semantic variations between positive pairs. Specifically, common data augmentation techniques frequently introduce semantic distortion, leading to a semantic margin between the positive pair. While the InfoNCE loss function overlooks the semantic margin and prioritizes similarity maximization between positive pairs during training, leading to the insensitive semantic comprehension ability of the trained model. In this paper, we introduce a novel Identical and Fraternal Twins of Contrastive Learning (named IFTCL) framework, capable of simultaneously adapting to various positive pairs generated by different augmentation techniques. We propose a *Twins Loss* to preserve the innate margin during training and promote the potential of data enhancement in order to overcome the sub-optimal issue. We also present proof-of-concept experiments combined with the contrastive objective to prove the validity of the proposed Twins Loss. Furthermore, we propose a hippocampus queue mechanism to restore and reuse the negative instances without additional calculation, which further enhances the efficiency and performance of the IFCL. We verify the IFCL framework on nine semantic textual similarity tasks with both English and Chinese datasets, and the experimental results show that IFCL outperforms state-of-the-art methods.

1 Introduction

Recent advances in neural network architecture, including the development of novel algorithms and computational techniques, have led to the emergence of universal sentence representations as a promising tool for various downstream tasks in natural language processing. These tasks encompass a wide range of applications such as information retrieval, semantic matching, and machine translation, which are crucial for enhancing the performance of intelligent systems [16, 17, 30, 8]. However, despite their potential, native BERT representations exhibit certain limitations in their ability to accurately capture semantic similarity tasks, which are essential for understanding and processing complex language structures [12, 23].

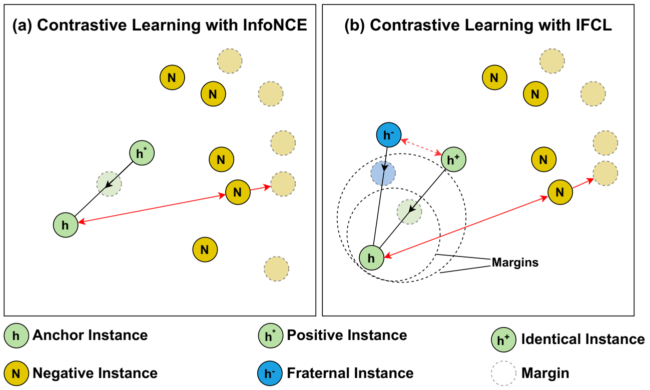

Figure 1. The optimization process in contrastive learning with positive and negative instances. (a) shows the proceeding of maximizing the similarities between the positive pairs in InfoNCE. (b) shows the proceeding of converging the identical and fraternal twins to each margin with the proposed Twins Loss.

To overcome these limitations and improve the effectiveness of sentence representations, researchers have introduced contrastive learning to the field of natural language processing, drawing inspiration from established techniques in the domain of computer vision. This innovative method offers a powerful solution for disentangling overlapping sentence representations and addressing the collapse of embedding space, which is a critical issue in representation learning. By doing so, contrastive learning enables more accurate and robust models for natural language processing tasks.

A key challenge persisting in contrastive learning is the design of high-quality positive pairs for training, an essential aspect that significantly influences the overall performance of the learning process. To generate these positive pairs, data augmentation techniques are employed, which involve creating new, diverse, and semantically-rich training samples from existing data. Data augmentation not only expands the training dataset but also enables the model to learn more effective representations by exposing it to a variety of linguistic structures and nuances. In recent years, an increasing number of studies

[*] Corresponding Author. Email: shuangyinli@scnu.edu.cn
[**] Both authors Qingfa Xiao and Shuangyin Li contributed equally to this research. The partial modification and improvement work of this research was completed by Qingfa Xiao as a visiting student at HKUST (GZ).

on data augmentation have been employed to generate high-quality augmented sentences for contrastive learning [13, 14, 32]. Explicit data augmentation techniques involve randomly inserting, substituting synonyms, and deleting words in a sentence. Other techniques focus on generating positive instances in the sentence embedding space, which is called implicit data augmentation. For instance, SimCSE [14] uses a simple dropout method to generate positive instances by slightly modifying the sentence representations, which is commonly used in current research on contrastive learning.

Despite the potential benefits of data augmentation techniques, they also present several challenges that must be addressed to ensure their effectiveness in contrastive learning of sentence representations. Two key issues that warrant further attention are semantic distortions of augmented sentences and limitations of the InfoNCE loss function. First, data augmentation can unintentionally lead to semantic distortions in the augmented sentences, causing their meanings to diverge from those of the original sentences. This may occur when words are inserted, deleted, or replaced without taking into account the broader context or the subtle nuances of the sentence structure. As a result, the augmented sentences may not accurately represent the intended meaning, which in turn hampers the model's ability to learn effective representations via contrastive learning. Second, the InfoNCE loss function, which is frequently employed in contrastive learning objectives, is unable to differentiate between semantically accurate and distorted samples. Consequently, the learning process may be misguided, and the model may end up with a limited understanding of the semantics. A notable example of this problem can be observed when the contrastive learning method treats the sentence "I do not like apples" and its augmented version "I do like apples" as a positive pair. By attempting to maximize the semantic distance between these semantically dissimilar sentences, the model's performance is adversely affected.

To address this issue, we firstly propose the Identical and Fraternal Twins data augmentations method, which generates two types of high-quality positive instances. Identical twins consist of an anchor instance and an identical instance, both created using dropout augmentation. Fraternal twins also have an anchor instance, but the fraternal instance includes features from a different language family, such as English and German, or Mandarin and Cantonese. These languages are chosen based on their proximity to each other in the embedding space and their high relevance [19]. Compared to the anchor instance, the identical instance has the most similar semantics due to the same augmentation method, while the fraternal instance introduces more diversity as the diversity of language expression. We further introduce the Identical and Fraternal Twins of Contrastive Learning (IFCL) framework based on proposed augmentation technique. Our framework employs a novel and crucial training objective *Twins Loss* to capture the fine-grained semantics and diverse expressions of twins and avoid the local optimum problem. Figure 1(a) illustrates the traditional contrastive learning process, where the positive instance is placed as close as possible to the anchor instance, and the negative instance is scattered in the embedding space. In contrast, Figure 1(b) demonstrates our IFCL optimization process, where the distances of identical and fraternal twins converge to each margin. Our motivation is that the identical instance is naturally closer to the anchor instance than the fraternal instance, as are their margins. Additionally, our framework contains a hippocampus queue mechanism that stores the negative instances in the queue with corresponding forgetting coefficients as short memory. This mechanism improves the efficiency and performance of the IFCL.

Our contributions of the IFCL framework can be synthesized as follows:

1. A novel data augmentation technique for contrastive learning is proposed to generate high-quality positive instances, named Identical and Fraternal Twins, where the identical twins retain the most similar semantic, and the fraternal twins exhibit greater diversity.
2. A novel hippocampus queue mechanism is presented to fully utilize the negative instances by storing the previous mini-batches into a short-term memory, improving the efficiency and performance of the IFCL.
3. Within the IFCL framework, we propose a novel and core training loss function named *Twins Loss* to optimize the identical and fraternal twins according to their margins, capturing fine-grained semantics for sentence representation learning and alleviating the sub-optimal issue.

In the experiments, our method has been validated on the benchmark datasets in both English and Chinese languages, and the results demonstrate our method significantly outperforms other strong baselines, achieving the state-of-the-art average performance. Furthermore, the core innovation of IFCL, the Twins loss function, has been shown to be effective through ablation experiments. And the proof-of-concept experiment has also confirmed that the use of Twins loss can optimize the upper and lower bounds of the contrastive learning objective, which is theoretical analysis about why and how such training objective works.

2 Related Works

Starting from research on word-level tasks [21, 22], sentence representations have garnered significant attention. Several works utilize various weighted combinations of word representations to form sentence representations [24]. However, these methods only consider the composition of word features, disregarding the order of words. Skip-Thought [17] uses the distributed assumption and RNNs and their variants [16] to predict contextual sentence representations. Infersent [11] constructs the SNLI dataset and trains sentence representations in a supervised manner, improving their quality. Furthermore, Universal Sentence Encoder [8] adopts the transformer architecture and multi-task joint training to extract semantic information.

In recent years, pre-trained BERT has become increasingly popular for sentence-level tasks. However, the native sentence representations often perform poorly. To address this, Sentence-BERT [23] leverages sentence similarity tasks with the SNLI and MNLI datasets and employs a siamese structure to generate semantically meaningful representations. Meanwhile, both BERT-flow [18] and BERT-whitening [26] employ a regular mapping of sentence representations to achieve isotropy in the embedding space, significantly outperforming previous unsupervised methods on semantic textual similarity tasks.

Contrastive learning is widely used in computer vision, where models learn the prior knowledge distribution of images in a self-supervised manner [9, 15]. In natural language processing (NLP), pre-trained BERT often performs poorly in sentence similarity tasks due to the inhomogeneous distribution of native representations. To address this, unsupervised contrastive learning has been employed as a training objective to alleviate the semantic collapse issue in NLP. ConSERT [32] inserts adversarial attack, token shuffling, cutoff, and dropout on the embedding layer as data augmentation methods. SimCSE [14] uses inherent dropout masks in transformers architecture to maintain semantic information as much as possible, which significantly improves semantic textual similarity tasks. Both methods gen-

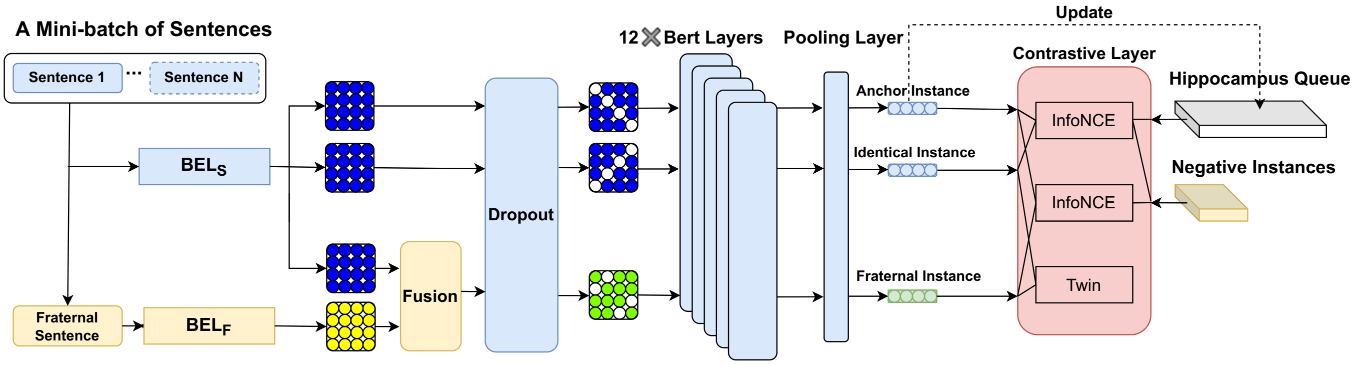

Figure 2. The general framework of the proposed IFCL.

erate positive instances through data augmentation and use other instances in the mini-batch as negative. SNCSE [29] proposes the bidirectional margin loss to distinguish hard negative instances, while ArcCSE [34] proposes a new optimizing objective loss to model pairwise sentence relations. DiffCSE [10] utilizes the ELECTRA model and replaces the token detection task to learn the differences between original and forged sentences. VaSCL [33] generates effective data augmentations using neighborhood methods, while DCLR [35] punishes false negatives and generates noise-based negatives to ensure the uniformity of the representation space. PT-BERT [27] constructs same-length positive and negative pairs using pseudo sentence representations to remove superficial features. MoCoSE [6] builds a two-branch structure framework with a prediction layer for the online branch to create asymmetry between the online and target branches, using a similar momentum encoder as MOCO [15]. However, most previous works focus on designing data augmentations to generate the most similar positive instances for contrastive learning, without considering the relationship of positive pairs in a fine-grained semantic way.

3 The IFCL Framework

The proposed IFCL adapts the pre-trained language model into a triplet network for contrastive learning of sentence representations, where the BERT model shares the same parameters. There are three major components of the IFCL framework as shown in Fig. 2:

1. A novel data augmentation module that includes dropout augmentation and fusion augmentation to generate two types of positive pairs for contrastive learning: identical twins $\left\{h_i, h_i^+\right\}$ and fraternal twins $\left\{h_i, h_i^-\right\}$.
2. A novel hippocampus queue mechanism is presented to fully utilize the negative instances by storing the previous instances into a short-term query, improving the efficiency and performance of the IFCL.
3. The InfoNCE loss is the first training objective function for positive pairs, which minimizes the distance between positive pairs and amplifies inconsistency between negative pairs. Additionally, we propose the *Twins Loss* as the second training objective function to constrain distance margins between identical and fraternal instances in the embedding space.

3.1 Identical and Fraternal Twins

In this section, we present the data augmentation techniques for generating identical and fraternal twins, respectively, and discuss how they can be used in IFCL.

3.1.1 Identical Twins

The first type of positive pair in the proposed approach consists of an anchor instance and an identical instance that share the same semantics, which we call identical twins. To generate these pairs, we adopt the method of dropout augmentation, as used in SimCSE. In the standard transformer architecture, random dropout masks are inherently placed on fully connected layers. We define a set of sentences as $\{x_i\}_{i=1}^N$. During the fine-tuning process, each sentence x_i is fed to the pre-trained BERT model twice. With different dropout masks, the BERT model outputs two different sentence representations, which are used to build the identical twins $\left\{hi, h_i^+\right\}$ as follows:

$$\begin{aligned} h_i &= f_\theta\left(BEL_S\left(x_i\right), z_i\right), \\ h_i^+ &= f_\theta\left(BEL_S\left(x_i\right), z_i^+\right), \end{aligned} \tag{1}$$

where BEL_S is the embedding layer without trainable parameters for x_i, the encoder $f_\theta\left(\odot\right)$ is the pre-trained BERT model, and z_i, z_i^+ are dropout masks representing Bernoulli distribution with probability ρ. The use of dropout masks is necessary for BERT model, so we consider that this simple approach is the only way to obtain h_i and h_i^+ with minimal semantic loss.

3.1.2 Fraternal Twins

The second type of positive pair consists of anchor instances and fraternal instances, where the fraternal instances are more diverse than the identical instances. To generate the fraternal instances, we propose an embedding fusion augmentation method that fuses additional semantic information from other languages into the anchor instances. Specifically, we could choose German as the additional semantics for English and Cantonese for Mandarin. Thus, the fraternal instance fuses the representation of the anchor in English and its diversified expression in German.

To accomplish this, we first translate the sentences from the source language $\{x_i\}_{i=1}^m$ into the fraternal language $\left\{x_i^-\right\}_{i=1}^m$. Next, the

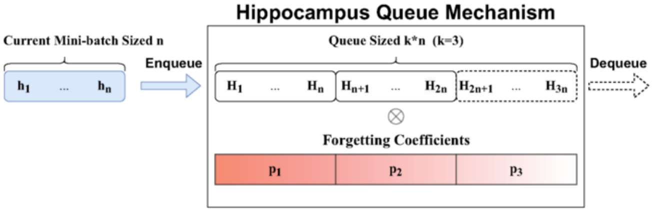

Figure 3. The hippocampus queue mechanism. For each iteration, the queue enqueue n sentences and dequeue n sentences. And the forgetting coefficients represents the weight of the negative instances stored in the queue, where the later updated instances have the larger coefficients.

two sets of sentences are fed into two different embedding layers, respectively, to obtain the embeddings of x_i and x_i^- as follows:

$$y_i = BEL_S\left(x_i\right), y_i^- = BEL_F\left(x_i^-\right), \tag{2}$$

where BEL_F is the embedding layer for x_i^-, extracted from the corresponding linguistic BERT model. Next, we fuse y_i and y_i^- into f_θ as follows:

$$h^- = f_\theta\left(\varepsilon * y_i + (1-\varepsilon) * y_i^-, z_i^-\right), \tag{3}$$

where ε is the fusion rate for $\left\{y_i, y_i^-\right\}$ and z_i^- denotes Bernoulli distribution with probability ρ. Together with the anchor instance h_i, we can obtain the fraternal twins $\left\{h_i, h_i^-\right\}$.

3.1.3 Discussion

Why do we employ two types of positive pairs in contrastive learning? The method of generating positive instances with dropout masks is a fundamental way to generate positive pairs with semantic fidelity, which is also the main characteristic of identical instances. On the other hand, the fraternal instance h^- incorporates features of the most relevant languages from the embedding space [19], such as expression, logical properties, and translational invariance of words, making it superior to identical instances. And Eq. (3) illustrates the subtle differences between identical and fraternal instances. In comparison with randomly changing features in dropout augmentation, our proposed fusion augmentation aims to inject more regular features into the fraternal instance. Therefore, the margins between these positive pairs are the precondition and intuition for the proposed *Twins Loss*. From a contrastive perspective, we apply constraints to preserve their margins and add comparisons between these positive pairs.

3.2 Hippocampus Queue Mechanism

Unsupervised contrastive learning yields sentence representations whose performance on downstream tasks is highly correlated with the number of negative instances. To increase the number of negative instances, one could extend the batch size. However, this operation is limited by finite GPU memory. To address this issue, we propose the hippocampus queue mechanism, which stores short-term memory, including a queue to store negative instances and forgetting coefficients to eliminate the impact of inconsistent instances.

Initially, we divide the datasets into multiple mini-batches of size N and encode N sentences in each iteration. In each mini-batch with batch size N, the other $N-1$ instances were chosen as negative instances. In our proposed strategy, we store $k * N$ anchor instances from previous k mini-batches in the queue and reuse them as negative instances. Consequently, each anchor instance has a total of $(k+1)N-1$ negative instances. Importantly, negative instances in the queue are only used to calculate loss, not to generate gradients for back-propagation, which reduces memory costs and differs from enlarging the batch size. To increase the diversity of negative instances, the queue is updated for each training step, with the current instances enqueued and the oldest instances dequeued.

As the encoder's parameters are continuously updated, the negative instances stored in the queue are progressively denatured. We hypothesize that previous instances in the queue are more inconsistent with the current mini-batch's instances, leading to deceptive similarities between anchor instances and denatured instances. To address this problem, we introduce a forgetting coefficient for each instance in the queue to calculate loss. The forgetting coefficients $\{p_m\}_{m=1}^{k*N}$ for all instances in the queue are as follows:

$$p_m = 1 - \lambda\left\lceil m/N \right\rceil, \tag{4}$$

where λ denotes the forgetting rate indicating that the weight of negative instances decreases progressively and becomes more harmonious.

3.3 Constrastive Learning with Twins Loss

In this section, we introduce how to optimize the sentence representations in the contrastive layer of IFCL framework. We present the unsupervised contrastive learning processing based on the InfoNCE function and the proposed *Twins Loss* function. Initially, we divide the datasets into multiple mini-batches of size N. During training, we encode all sentences three times to obtain sentence representations $\left\{h_i, h_i^+, h_i^-\right\}_{i=1}^N$, which serve as anchor instances, identical instances, and fraternal instances. Based on above two types of positive pairs, contrastive learning aims to learn effective representation by pulling semantically close neighbors together and pushing apart non-neighbors. And the *Twins Loss* aims to maintain the margins between the different positive pairs. In this paper, we use the cosine distance to express the semantic similarity $\operatorname{sim}(\odot)$ between sentences $\mathbf{h_i}, \mathbf{h_j}$ as follows:

$$\operatorname{sim}\left(\mathbf{h_i}, \mathbf{h_j}\right) = \frac{\mathbf{h}_i^\top \mathbf{h}_j}{\left\|\mathbf{h}_i\right\| \cdot \left\|\mathbf{h}_j\right\|}, \tag{5}$$

For the set of identical instances $\left\{h_i, h_i^+\right\}_{i=1}^N$, we define the first loss function by using negative instances $\{\mathbf{H}_m\}_{m=1}^{k*N}$ stored in the hippocampus queue. The loss function is defined as follows:

$$\begin{aligned} \ell_i^I &= -\log \frac{e^{\operatorname{sim}\left(\mathbf{h}_i, \mathbf{h}_i^+\right)/\tau}}{\sum_{j=1}^N e^{\operatorname{sim}\left(\mathbf{h}_i, \mathbf{h}_j^+\right)/\tau} + \varphi}, \\ \varphi &= \sum_{m=1}^{k*N} p_m * e^{\operatorname{sim}\left(\mathbf{h}_i, \mathbf{H}_m\right)/\tau}, \end{aligned} \tag{6}$$

Here, $\operatorname{sim}(\odot)$ refers to cosine similarity, τ represents the temperature that controls the perception of negative instances, and p_m is the forgetting coefficient shown in Section 3.2. It should be noted that $\mathbf{H}_m$ stored in the hippocampus queue is detached from the current graph while back-propagating the gradients to update the parameters of the IFCL.

The InfoNCE loss is used as a second loss function for the fraternal twins $\left\{h_i, h_i^-\right\}_{i=1}^N$ and is formulated as follows:

$$\ell_i^F = -\log \frac{e^{\operatorname{sim}\left(\mathbf{h}_i, \mathbf{h}_i^-\right)/\tau}}{\sum_{j=1}^N e^{\operatorname{sim}\left(\mathbf{h}_i, \mathbf{h}_j^-\right)/\tau}}, \tag{7}$$

This loss function serves as an auxiliary training objective to add diverse features in the embedding space and enhance the robustness of sentence representations. Thus it does not use the hippocampus queue mechanism.

We propose a novel training objective, called *Twins Loss*, to alleviate model's sub-optimal issue by preserving the innate margins between identical and fraternal twins. The objective is formulated as follows:

$$\begin{aligned} \ell_i^T &= \left| e^{\operatorname{sim}\left(\mathbf{h}_i, \mathbf{h}_i^+\right)} - e^{\operatorname{sim}\left(\mathbf{h}_i, \mathbf{h}_i^-\right)} - \mathbf{M}_i \right|, \\ \mathbf{M}_i &= e^{\operatorname{sim}\left(\mathbf{emb}_i, \mathbf{emb}_i^+\right)} - e^{\operatorname{sim}\left(\mathbf{emb}_i, \mathbf{emb}_i^-\right)} \end{aligned} \tag{8}$$

Here, $\mathbf{M}$ represents the innate margins between identical and fraternal twins, and $\mathbf{emb}$ represents the embeddings y after data augmentations. This $\mathbf{M}$ is determined by the similarity of initial embedding with different augmentation approaches, indicating the suitable intervals between the identical and fraternal instances in embedding space as Fig. 1

As mentioned earlier, the semantic distortion in the positive pair occurs naturally due to the dropout mask and embeddings fusion method. However, this distortion must be considered in the loss function InfoNCE, as it is not reasonable to maximize the similarity of the anchor instance and positive instance to 100%. Our IFCL introduces a novel loss function, presented as Eq. (8), which ensures that positive pairs' similarities (the anchor instance and the identical instance, the anchor instance and the fraternal instance) can be converted to optimal margins. For the *Twins Loss* approach, our objective is to minimize the distance between the optimized positive pairs and their distance before optimization, where $\mathbf{M}$ is the newly added comparator in our contrastive learning method.

Overall, we combine the loss functions Eq.6, Eq.7, and Eq. 8 together to form the total loss of our IFCL in each iteration.Thus, our IFCL model $f_\theta\left(\odot\right)$ can learn from both identical twins and fraternal twins, with the predicted representations distribution of the model parameterized by θ:

$$\theta^* = \arg\min_\theta \sum_{i=1}^N (\ell_i^I + \ell_i^F + \ell_i^T) \tag{9}$$

And we present the algorithm of the training process of the IFCL framework in supplementary materials.

4 Experiments

4.1 Setups

Datasets. To fine-tune the pre-trained BERT models, we randomly selected two subsets from Wikipedia as unlabeled training datasets for both English and Chinese languages. The English dataset (EnData) contains 10^6 sentences released by SimCSE [14], and the Chinese dataset (CnData) contains 10^5 sentences.

For the EnData dataset, we evaluate the performance of our models on 7 semantic textual similarity (STS) tasks, namely STS 2012–2016 [1, 2, 3, 4, 5], STS-Benchmark [7], and SICK-Relatedness [20]. On the other hand, for the CnData dataset, we evaluated the models on the Chinese STS-Benchmark (C-STS-B)[25] and SimCLUE[31]. SimCLUE includes most of the available open-source datasets of semantic similarity and natural language inference in the Chinese domain. Each STS task sample consists of a sentence pair and a golden score ranging from 1.0 to 5.0, indicating the degree of semantic similarity between the sentence pair. In contrast, each SimCLUE sample comprises a sentence pair and a binary golden score (0 or 1) indicating whether the sentence pair is similar.

To translate the fraternal sentences, we utilize the T5 pre-trained model from Huggingface to convert the EnData into the German dataset. Additionally, we rely on the Baidu Translator API to translate CnData into the Cantonese dataset. This decision is made due to the paucity of Cantonese translation resources available.

Implemention Details. We implement the IFCL based on Sentence-BERT [23] and initialize models with three different versions of pre-trained BERT [12] and the embedding layer of the BERT model for the fraternal sentences. These versions include bert-base-uncased, bert-base-multilingual-uncased, bert-large-uncased as well as stefan-it/albert-large-german-cased and denpa92/bert-base-cantonese for German and Cantonese datasets respectively. During the experiments, we fine-tune the IFCL for one epoch and evaluate the models with the verification sets of STS-B or C-STS-B [7, 25] after every 151 steps. Our evaluation metric of choice is Spearman's correlation, consistent with previous works. And we conduct some experiments to explore the influence of hyperparameters including the fusion rate and hippocampus queue length in supplementary materials.

4.2 Results on STS Tasks with EnData

On this EnData datasets, we compare the effectiveness of IFCL with other state-of-the-art unsupervised methods, such as BERT-flow, BERT-whitening, unsupervised ConSERT, unsupervised SimCSE, and DCLR. To evaluate the IFCL's effectiveness, we have selected MoCoSE and PT-BERT, as they are also based on contrastive learning and have the strategy to expand the amount of negative instances. We have initialized our models with pre-trained BERT$_{\mathrm{base}}$ and BERT$_{\mathrm{large}}$, called IFCL-BERT$_{\mathrm{base}}$ and IFCL-BERT$_{\mathrm{large}}$, respectively.

As indicated in Table 1, IFCL-BERT$_{\mathrm{base}}$ has an average Spearman's correlation score of 77.80% on the STS tasks, which is the highest average score compared to the other methods. Specifically, when compared to the previous state-of-the-art method DCLR, IFCL-BERT outperforms significantly on four STS tasks. Notably, our model's performance is also exceptional when compared to recent works, including MoCoSE-BERT and PT-BERT, which enhances the STS12, STS16, and STS-B tasks score to 71.57% (+0.37%), 80.17% (+1.27%), and 80.27% (+1.59%), respectively. Although IFCL-BERT does not perform the best on all STS tasks, it ranks in the top 3 and outperforms most of the baselines. In the case of IFCL-BERT$_{\mathrm{large}}$, it shows significant improvement, particularly on STS-B and SICK-R tasks, and achieves the best average result.

We conducte an ablation study using BERT$_{\mathrm{base}}$ on the STS-Benchmark task to examine the contribution of each component. In comparison with SimCSE, we propose three additional components: fraternal instances, Twins Loss, and the hippocampus queue mechanism. As shown in Table 2, we find that our IFCL$_{\mathrm{base}}$ model achieve a substantial improvement when using fraternal instances with Twins Loss. This finding demonstrates that the differences between positive pairs should not be neglected, and that these fine-grained semantics can significantly improve model performance. Moreover, the pro-

Table 1. The performances of IFCL on 7 English STS tasks. We choose [cls] representations for training and the mean of the representations in the last layer for verification. Noted that ◇ indicates they are based on contrastive learning and use the same Wikipedia datasets (EnData).

Results of English tasks								
Method	**STS12**	**STS13**	**STS14**	**STS15**	**STS16**	**STS-B**	**SICK-R**	**Avg.**
BERT$_{\text{base}}$	39.70	59.38	49.67	66.03	66.19	53.87	62.06	56.70
BERT-flow$_{\text{base}}$	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
BERT-whitening$_{\text{base}}$	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
ConSERT$_{\text{base}}$	64.64	78.49	69.07	79.72	75.95	73.97	67.31	72.74
SimCSE-BERT$_{\text{base}}^{\diamond}$	68.40	82.41	74.38	80.91	78.56	76.85	72.23	76.25
VaSCL-BERT$_{\text{base}}^{\diamond}$	69.08	81.95	74.64	82.64	80.57	80.23	71.23	77.19
DCLR-BERT$_{\text{base}}^{\diamond}$	70.81	83.73	75.11	82.56	78.44	78.31	71.59	77.22
MoCoSE-BERT$_{\text{base}}^{\diamond}$	71.48	81.40	74.47	**83.45**	78.99	78.68	**72.44**	77.27
PT-BERT$_{\text{base}}^{\diamond}$	71.20	**83.76**	**76.34**	82.63	78.90	79.42	71.94	77.74
IFCL-BERT$_{\text{base}}^{\diamond}$	**71.57**	82.35	75.08	83.03	**80.17**	**80.27**	72.16	**77.80**
BERT$_{\text{large}}$	57.73	61.17	61.18	68.07	70.25	59.59	60.34	62.62
ConSERT$_{\text{large}}$	70.69	82.96	74.13	82.78	76.66	77.53	70.37	76.45
SimCSE$_{\text{large}}^{\diamond}$	70.88	84.16	76.43	84.50	79.76	79.26	73.88	78.41
DCLR-BERT$_{\text{large}}^{\diamond}$	71.87	**84.83**	**77.37**	**84.70**	**79.81**	79.55	74.19	78.90
MoCoSE-BERT$_{\text{large}}^{\diamond}$	**74.50**	84.54	77.32	84.11	79.67	80.53	73.26	79.13
IFCL-BERT$_{\text{large}}^{\diamond}$	73.88	84.31	76.64	84.01	79.56	**81.37**	**76.30**	**79.44**

Table 2. The ablation study of the IFCL$_{\text{base}}$, where FI denotes fraternal instances, TL denotes Twins Loss, and HQ denotes hippocampus queue mechanism.

Model	STS-B
IFCL$_{\text{base}}$	**80.27**
w/o FI, TL	78.10
w/o HQ	78.81
w/o TL, HQ	77.26
w/o FI, TL, HQ (SimCSE)	76.83

Table 3. The performances on the C-STS-B and SimCLUE tasks. Noted that • indicates that the models are fine-tuned with the data in the C-STS-B task, which can be considered as a weakly supervised training. And ◇ indicates that the models are fine-tuned with unsupervised CnData datasets.

Results of Chinese tasks		
Method	**Chinese STS-B**	**SimCLUE**
BERT	55.52	29.89
BERT-whitening•	68.27	-
SimCSE-BERT•	68.91	40.74
SimCSE-BERT◇	60.41	40.54
IFCL-BERT◇	**71.41**	**44.42**

posed hippocampus queue mechanism also demonstrated a crucial role in IFCL.

4.3 Results on STS Tasks with CnData

We conduct experiments on the CnData datasets, comparing our results with native BERT and unsupervised SimCSE [14] as baselines. Table 3 presents the results of the experiments. Our fine-tuned IFCL-BERT◇ on the CnData datasets outperform other methods on both C-STS-B and SimCLUE, achieving accuracy scores of 71.41% and 44.42%, respectively. However, we notice that SimCSE-BERT◇ performed worse on CnData than on C-STS-B datasets. The results, presented as SimCSE-BERT• in Table 3, showed significant improvement and performed better than SimCSE-BERT◇. This is because the data in the C-STS-B task is related to semantic understanding, which is helpful for our task. Nevertheless, our IFCL-BERT◇ still outperform both on C-STS-B and SimCLUE tasks. In conclusion, the experiments demonstrate that IFCL is more efficient than SimCSE in the Chinese domain.

4.4 Analysis

Why do we use Identical and Fraternal Twins? In the semantic textual similarity tasks, excellent data augmentation methods could avoid semantic distortions and improve the expression diversity, which helps to generate high-quality positive instances for contrastive learning. To verify identical and fraternal twins and prove the effectiveness, we augment the test data of the STS-B and C-STS-B tasks with different data augmentation methods and show the performances on Spearman's correlation.

In detail, Stsb-Bert-Base is used for the STS-B task and Simbert-Chinese-Base is used for C-STS-B task, the two pre-trained BERT are designed for the task of semantic textual similarity. We test five methods of data augmentations, including random insertion, synonymous substitution, random deletion, and our method of identical and fraternal twins. As shown in Fig. 4a, we can observe that the traditional data augmentations, such as random insertion, synonymous substitution, and random deletion perform worse than our method of identical and fraternal twins. Meanwhile, the method of identical twins performs better on the STS-B task, while the method of fraternal twins outperforms on the C-STS-B task, which is proved that we need to adjust the participation of the two methods in different scenarios. Thus in our paper, we specifically propose the *Twins Loss* to optimize them, which is the fine-grained comparison between the positive instances in the IFCL.

Effect of Twins Loss Function. To test the effectiveness of *Twins Loss* function, we design the controlled experiments, where the IFCL$_{\text{w/o TL}}$ indicates that only the ℓ_i^I and ℓ_i^F is used and the *Twins Loss* ℓ_i^T is knocked off. The 1.3k sentence pairs with gold scores (0-5 score) are from the STS-Benchmark dataset, and the models need to predict the cosine similarity of each sentence. As shown in Fig. 4b, the 45-degree line denotes the ground truth, and the line closer to it indicates better performance. From Fig. 4b, the performance of IFCL (80.27%) is better than that of IFCL$_{\text{w/o TL}}$(78.65%), which means the effectiveness of *Twins Loss* function. Moreover, the range of predicted values of the BERT is from 0.75 to 0.9, while, the range of the IFCL is from 0.35 to 0.9. The range shows the capacity of the methods on capturing the fine-grained similarity of the sentences. Thus, from this part of the experiments, we can conclude that the IFCL can learn more diverse semantics from the identical and fraternal instances for the sentence representations in a more reasonable way.

Advantages of Hippocampus Queue Mechanism. To verify the

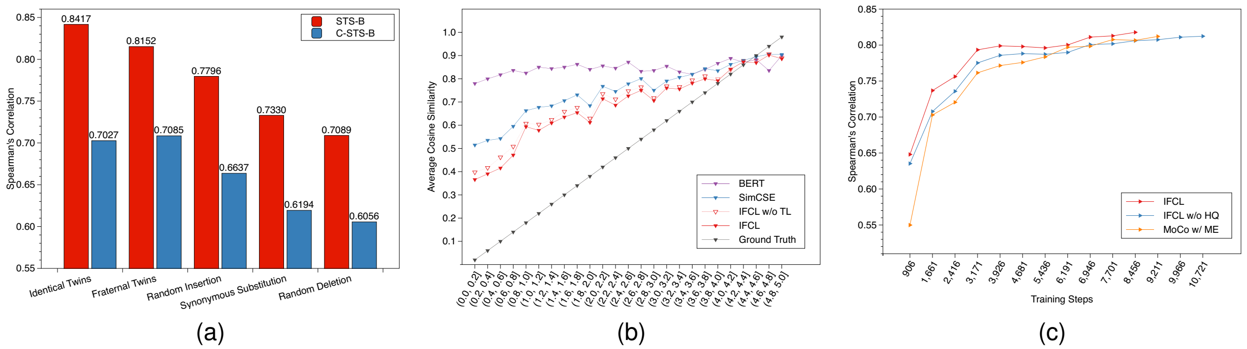

Figure 4. (a) The performances of different data augmentation methods on the STS-B and C-STS-B tasks. Noted that, we choose two types of supervised pre-trained BERT for this experiment, including Stsb-Bert-Base and Simbert-Chinese-Base. (b)The predicted cosine similarity of sentence pairs with different base models. Noted that, the gray line denotes the ground truth. (c) The advantages of the hippocampus queue mechanism, by demonstrating the training steps to the optimal IFCL$_{\mathrm{base}}$ on the STS-B task.

advantages of the hippocampus queue mechanism, we explore the amount of data required for the optimal performance of the IFCL, where the IFCL$_{\mathrm{w/o\ HQ}}$ indicates that the hippocampus queue mechanism is knocked off in the IFCL, and the MoCo$_{\mathrm{w/\ ME}}$ indicates that we replace the hippocampus queue mechanism with momentum encoder mechanism in MoCo [15] for the IFCL. As shown in Fig. 4c, the IFCL can be optimized to the optimal only with 8456 steps of about 540,000 training data, which outperforms the IFCl$_{\mathrm{w/o\ HQ}}$ and the MoCo$_{\mathrm{w/\ ME}}$ in both efficiency and performance. Though the momentum encoder mechanism originated from computer vision is effective, the proposed hippocampus queue mechanism is more available for the contrastive learning of natural language processing.

4.5 Proof of Concept Experiments

In this section, we discuss the rationale behind designing margins for the two types of augmented positive instances in contrastive learning, from the perspective of mutual information. In the findings of the InfoMin [28] study, they proposed a unified perspective on positive instances for contrastive learning, asserting that reducing the mutual information between positive pairs while preserving task-relevant information is optimal for the task. In the ideal optimization process, the mutual information of positive pairs share only task-relevant information, with no irrelevant noise. Consequently, we build upon their study and demonstrate, both experimentally and theoretically, that our designed margins can guide our positive instances to achieve this desired balance. In our IFCL framework, there is a two-stage exploration of the mutual information of instances and the task-relevant information. We firstly examine the mutual information between the dropout augmented instances h and the fusion augmented instances h^*, and investigate whether the margins can effectively reduce their mutual information. And then, we explore the task-relevant information of the textual semantic task, and assess whether the shared mutual information predominantly encompasses task-relevant information without extraneous noise.

For the first stage discussion, we follow the study of InfoMin and define the mutual information $\mathbb{MI}$ between the set of two types of instances $\{h_i, h_i^*\}_{i=1}^N$ as follows:

$$\mathbb{MI}(h, h^*) = \log(N) + \frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{\mathrm{sim}\left(h_i, h_i^*\right)/\tau}}{\sum_{j=1}^{N} e^{\mathrm{sim}\left(h_i, h_j^*\right)/\tau}}, \quad (10)$$

Table 4. Mutual information and task-relevant information. The IFCL-BERT w/o TL means training IFCL-BERT without Twins Loss. The experiments are conducted with EnData and STS-B datasets on Bert-base.

Method	$\mathbb{MI}(h, h^+)$	$\mathbb{MI}(h, h^-)$	$\mathbb{MI}_{task}$
IFCL-BERT	4.15	4.17	4.31
IFCL-BERT w/o TL	4.23	4.20	4.58
SimCSE	4.24	-	4.52

In the first stage, we calculate the mutual information of our identical twins $\mathbb{MI}(h, h^+)$ and fraternal twins $\mathbb{MI}(h, h^-)$ after training as shown in Table 4. We find that both mutual information of our two positive pairs decline in our IFCL, which proves that the margins in *Twins Loss* can reduce the mutual information. Comparing between $\mathbb{MI}(h, h^+)$ and $\mathbb{MI}(h, h^-)$, the designed margins can keep more diverse semantic information from our fusion augmentation in fraternal twins, where the $\mathbb{MI}(h, h^-)$ is higher than $\mathbb{MI}(h, h^+)$ in IFCL. In the second stage, we explore why such reduced mutual information is effective for the task in terms of task-relevant information. In our semantic textual tasks, we evaluate the semantic information between two sentences. Therefore, we define the mutual of sentence pairs which have the highest similarity score 5.0 in STS-Benchmark dataset as task-relevant information. As shown in Table 4, the $\mathbb{MI}_{task}$ in our IFCL is the lowest compared with the others and the most closest to train data $\mathbb{MI}(h, h^+)$ and $\mathbb{MI}(h, h^-)$. Refer to InfoMin, the mutual information of positive pairs in our IFCL contains more task-relevant information. It is the core influence of our designed margins for two types of positive pairs, leading to stronger model semantic recognition.

5 Conclusion

In this paper, we propose the IFCL, an identical and fraternal twins of contrastive learning, which learns the sentence representations in a more fine-grained semantic manner in the embedding space, and which also provides a new perspective on the contrastive learning between positive instances. With a designed *Twins Loss*, our IFCL significantly improves the efficiency and performance of unsupervised sentence representation on semantic textual similarity tasks in both English and Chinese domains. In the future, we will optimize our method for supervised learning, where the labels are fully used for more diverse positive pairs.

ACKNOWLEDGMENTS

This work was supported by Natural Science Foundation of Guangdong (2023A1515012073), National Natural Science Foundation of China(No. 62006083), and National Key Research and Development Program of China (2020YFA0712500).

References

[1] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al., '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)*, pp. 252–263, (2015).

[2] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel M Cer, Mona T Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe, 'Semeval-2014 task 10: Multilingual semantic textual similarity.', in *SemEval@ COLING*, pp. 81–91, (2014).

[3] Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe, '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), (2016).

[4] Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre, '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)*, pp. 385–393, (2012).

[5] Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo, '* 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*, pp. 32–43, (2013).

[6] Rui Cao, Yihao Wang, Yuxin Liang, Ling Gao, Jie Zheng, Jie Ren, and Zheng Wang, 'Exploring the impact of negative samples of contrastive learning: A case study of sentence embeddin', *arXiv preprint arXiv:2202.13093*, (2022).

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

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

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, 'A simple framework for contrastive learning of visual representations', in *International conference on machine learning*, pp. 1597–1607. PMLR, (2020).

[10] Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljačić, Shang-Wen Li, Wen-tau Yih, Yoon Kim, and James Glass, 'Diffcse: Difference-based contrastive learning for sentence embeddings', *arXiv preprint arXiv:2204.10298*, (2022).

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

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 'Bert: Pre-training of deep bidirectional transformers for language understanding', *arXiv preprint arXiv:1810.04805*, (2018).

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

[14] Tianyu Gao, Xingcheng Yao, and Danqi Chen, 'Simcse: Simple contrastive learning of sentence embeddings', *arXiv preprint arXiv:2104.08821*, (2021).

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick, 'Momentum contrast for unsupervised visual representation learning', in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, (2020).

[16] Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen, 'Convolutional neural network architectures for matching natural language sentences', *Advances in neural information processing systems*, **27**, (2014).

[17] Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler, 'Skip-thought vectors', *Advances in neural information processing systems*, **28**, (2015).

[18] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li, 'On the sentence embeddings from pre-trained language models', *arXiv preprint arXiv:2011.05864*, (2020).

[19] Jindřich Libovickỳ, Rudolf Rosa, and Alexander Fraser, 'How language-neutral is multilingual bert?', *arXiv preprint arXiv:1911.03310*, (2019).

[20] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli, 'A sick cure for the evaluation of compositional distributional semantic models', in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 216–223, (2014).

[21] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, 'Distributed representations of words and phrases and their compositionality', *Advances in neural information processing systems*, **26**, (2013).

[22] Jeffrey Pennington, Richard Socher, and Christopher D Manning, 'Glove: Global vectors for word representation', in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, (2014).

[23] Nils Reimers and Iryna Gurevych, 'Sentence-bert: Sentence embeddings using siamese bert-networks', *arXiv preprint arXiv:1908.10084*, (2019).

[24] Stephen Robertson, 'Understanding inverse document frequency: on theoretical arguments for idf', *Journal of documentation*, (2004).

[25] Tang Shancheng, Bai Yunyue, and Ma Fuyu, 'A semantic text similarity model for double short chinese sequences', in *2018 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*, pp. 736–739. IEEE, (2018).

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

[27] Haochen Tan, Wei Shao, Han Wu, Ke Yang, and Linqi Song, 'A sentence is worth 128 pseudo tokens: A semantic-aware contrastive learning framework for sentence embeddings', *arXiv preprint arXiv:2203.05877*, (2022).

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

[29] Hao Wang, Yangguang Li, Zhen Huang, Yong Dou, Lingpeng Kong, and Jing Shao, 'Sncse: Contrastive learning for unsupervised sentence embedding with soft negative samples', *arXiv preprint arXiv:2201.05979*, (2022).

[30] John Wieting and Kevin Gimpel, 'Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations', *arXiv preprint arXiv:1711.05732*, (2017).

[31] Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al., 'Clue: A chinese language understanding evaluation benchmark', *arXiv preprint arXiv:2004.05986*, (2020).

[32] Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu, 'Consert: A contrastive framework for self-supervised sentence representation transfer', *arXiv preprint arXiv:2105.11741*, (2021).

[33] Dejiao Zhang, Wei Xiao, Henghui Zhu, Xiaofei Ma, and Andrew O Arnold, 'Virtual augmentation supported contrastive learning of sentence representations', *arXiv preprint arXiv:2110.08552*, (2021).

[34] Yuhao Zhang, Hongji Zhu, Yongliang Wang, Nan Xu, Xiaobo Li, and Binqiang Zhao, '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)*, pp. 4892–4903, (2022).

[35] Kun Zhou, Beichen Zhang, Wayne Xin Zhao, and Ji-Rong Wen, 'Debiased contrastive learning of unsupervised sentence representations', *arXiv preprint arXiv:2205.00656*, (2022).