Learning Monolingual Sentence Embeddings with Large-scale Parallel Translation Datasets

Anonymous ACL submission

Abstract

Although contrastive learning has greatly improved sentence representation, its performance is still limited by the size of monolingual sentence-pair datasets. Meanwhile, there exist large-scale parallel translation pairs (100x larger than monolingual pairs) that are highly correlated in semantic, but have not been utilized for learning sentence representation. Furthermore, given parallel translation pairs, previous contrastive learning frameworks can not well balance the monolingual embeddings' alignment and uniformity which represent the quality of embeddings. In this paper, we build on the top of dual encoder and propose to freeze the source language encoder, utilizing its consistent embeddings to supervise the target language encoder via contrastive learning, where source-target translation pairs are regarded as positives. We provide the first exploration of utilizing parallel translation sentence pairs to learn monolingual sentence embeddings and show superior performance to balance the alignment and uniformity. We achieve a new state-of-the-art performance on the average score of standard semantic textual similarity (STS), outperforming both SimCSE and Sentence-T5, and the best performance in corresponding tracks on transfer tasks.

1 Introduction

It has been a fundamental problem in natural language processing to learn sentence embeddings that provide compact semantic representations (Reimers and Gurevych, 2019; Gao et al., 2021; Ni et al., 2021). Recently, contrastive learning (CL) which aims to learn effective representation by pulling semantically close neighbors together and separating non-neighbors, has widely attracted attention for building representations. Benefited from a powerful contrastive learning framework, scaling up the size of dataset greatly improves robustness and generalization of representations, as suggested by some previous works (Chen et al., 2020; Radford et al., 2021; Jia et al., 2021; Wang et al., 2021).

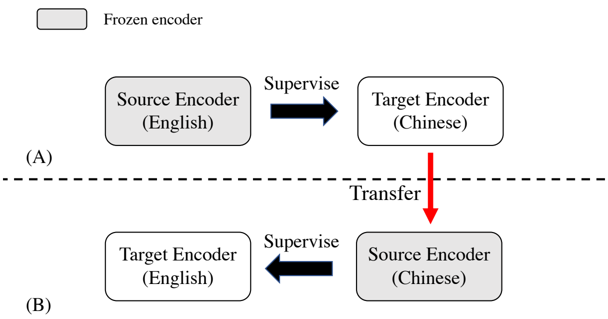

Figure 1: **Training pipeline.** We first obtain a target (Chinese) encoder given a pre-trained SimCSE model as the source encoder. Then, we take the pre-trained Chinese encoder as the source encoder and freeze it to supervise a target (English) encoder. Step (A) and step (B) both follow our proposed framework.

Gao et al. 2021 demonstrates that a contrastive objective can be extremely effective when coupled with pre-trained language models and sentence-pair datasets. However, the generality and capability of the language model are strictly limited by the size of existing sentence-pair datasets (Bowman et al., 2015; Williams et al., 2017). Meanwhile, there have accumulated large-scale parallel translation datasets (100x larger than existing monolingual sentence-pair datasets) in multilingual learning community (Yang et al., 2019a; Feng et al., 2020; Pan et al., 2021), which have not been utilized for learning sentence representations. Furthermore, given parallel translation pairs, previous contrastive learning frameworks (Radford et al., 2021; Gao et al., 2021) cannot well balance[1] the alignment and uniformity (Wang and Isola, 2020) of monolingual sentence embeddings, where alignment calculates the expected distance between positive embeddings and uniformity measures how well the embeddings are uniformly distributed.

Suggested by Frozen (Tsimpoukelli et al., 2021)

[1]The alignment retains steady while uniformity improves.

in multimodal learning, freezing the language model and only updating the vision encoder enables strong generalization. In this paper, we build on the top of dual encoder (Radford et al., 2021; Yang et al., 2019b), and adopt a similar strategy as Frozen, where we freeze the source language encoder and only train the target language encoder for better monolingual sentence embeddings. The source language encoder constructs a large memory queue that stores negative embeddings, and provides consistent embeddings to supervise the target language encoder via contrastive learning, where source-target translation pairs are regarded as positives. Specifically, we utilize available large-scale Chinese-English translation datasets as source-target pairs to learn sentence embeddings in English scenarios. To obtain the source language (Chinese) encoder, instead of adopting a pre-trained model, we conduct the same protocol where a frozen pre-trained English encoder[2] is utilized to supervise our source language (Chinese) encoder, and fine-tune it on Chinese NLI dataset for better performance. We initialize the target language (English) encoder with a pre-trained language model, such as BERT (Devlin et al., 2018) or RoBERTa (Liu et al., 2019). The illustration of training pipeline can be found in Figure 1

We conduct a comprehensive evaluation protocol following SimCSE (Gao et al., 2021) on seven standard semantic textual similarity (STS) tasks (Agirre et al., 2012, 2013; Marelli et al., 2014; Agirre et al., 2014, 2015, 2016; Cer et al., 2017) and seven transfer tasks (Conneau and Kiela, 2018). We achieve a new state-of-the-art on STS tasks, outperforming SimCSE (Gao et al., 2021) and Sentence-T5 (Ni et al., 2021) by a large margin, and also achieve the best performance in corresponding tracks on transfer tasks evaluated by SentEval (Conneau and Kiela, 2018). On the average score of STS tasks, our pre-trained BERT$_{base}$ with or without fine-tuning surpasses SimCSE-BERT$_{base}$ by 4.39% and 3.25% respectively, and RoBERTa$_{large}$ achieves 85.58 on average. Surprisingly, BERT$_{base}$ with fine-tuning achieves better results than Sentence-T5 (11B) with only 1% parameters in comparison.

We summarize our contributions as below:

1. We provide the first exploration of utilizing existing large-scale parallel translation pairs for learning sentence representation.

2. We introduce a new cross-lingual contrastive learning framework to learn sentence embeddings that well balances alignment and uniformity.

3. Our approach achieves a new state-of-the-art on standard semantic textual similarity (STS), and the best performance in corresponding tracks on transfer tasks evaluated by SentEval[3].

2 Related Work

2.1 Sentence Representation

Sentence representation is a well-studied area with many proposed methods (Mikolov et al., 2013; Pennington et al., 2014; Le and Mikolov, 2014). With the progress of pre-training, objectives like BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) are utilized to generate sentence embeddings. To derive semantically meaningful sentence embeddings that can be compared using cosine-similarity from BERT, Sentence-BERT (Reimers and Gurevych, 2019) uses siamese and triplet network structures. SimCSE (Gao et al., 2021) introduces a simple contrastive learning framework, which greatly improves state-of-the-art sentence embeddings on semantic textual similarity tasks both on unsupervised and supervised tracks. Sentence-T5 (Ni et al., 2021) investigates producing sentence embeddings from the pre-trained T5 (Raffel et al., 2019), then fine-tunes the model on natural language inference dataset and achieves the leading results in sentence embeddings benchmark datasets. These works are conducted on monolingual sentence-pair datasets, while not exploring existing large-scale paralllel translation datasets. In this work, we provide an exploration of utilizing available parallel translation pairs for learning sentence embeddings.

2.2 Multilingual Learning

Multilingual learning has attracted increasing interests from the community. Parallel translation datasets have been widely leveraged for Neural Machine Translation (NMT) (Bahdanau et al., 2014; Wu et al., 2016), Semantic Retrieval (SR) (Wagner et al., 2001), Bitext Retrieval (Yang et al., 2019b,a) (BR) and Retrieval Question Answering (ReQA) (Kolomiyets and Moens, 2011), etc. Multilingual Sentence Encoder (Yang et al., 2019b) conducts a multitask trained dual encoder to bridge 16 different languages, and achieves competitive results on SR, BR, ReQA tasks. LaBSE (Feng

[2] We adopt the pre-trained SimRoBERTa$_{large}$ model from https://github.com/princeton-nlp/SimCSE.

[3] https://github.com/facebookresearch/SentEval

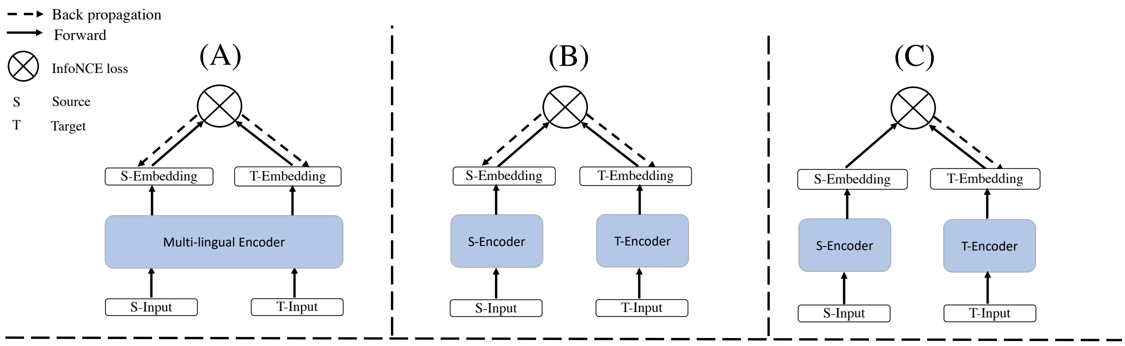

Figure 2: **Comparison of preliminaries and our approach for utilizing parallel translation pairs.** (A), (B) and (C) represent a multilingual encoder, dual encoder and our modified dual encoder, respectively.

et al., 2020) adopts a dual encoder with additive margin softmax combined with masked language model (MLM) (Devlin et al., 2018) and translation language model (TLM) (Lample and Conneau, 2019) to improve multilingual sentence embeddings. mRASP2 (Pan et al., 2021) hypotheses that inner multilingual representations leads to better multilingual translation performance. They regard a corresponding pair as a positive sample, and other in-batch samples including a variety of languages as negative samples, to establish a contrastive learning process. In this way, multiple languages representations are smoothly embedded into the same semantic space. Unlike previous works that focus on embedding text from multiple languages into the same semantic space, we propose utilizing corresponding parallel translation pairs as semantically close neighbors, pulling their embeddings together while pushing apart non-neighbors.

3 Proposed Approach

We start by briefly describing background and preliminaries in 3.1. Then, we introduce the design of our proposed contrastive framework for learning from parallel translation pairs in 3.2. Lastly, we provide analysis for our approach in 3.3.

3.1 Background

Scaling up the size of training dataset (Radford et al., 2021; Jia et al., 2021) has proved to be effective to improve robustness and generalization of representations in contrastive learning framework. However, previous works (Reimers and Gurevych, 2019; Gao et al., 2021) only utilize limited size[4] of monolingual sentence pairs to learn sentence embeddings, such as MNLI datasets (Williams et al., 2017) and SNLI (Bowman et al., 2015). In contrast, there have existed large-scale well-annotated parallel translation pairs (100x larger than monolingual paired datasets) in the community of multilingual learning. Instead of training on limited monolingual sentence pairs, utilizing existing parallel translation datasets shows better flexibility and a potential to further improve the performance of sentence embeddings, where a parallel translation pair that is highly correlated in semantic can be treated as a positive sample.

Preliminaries. To utilize paired inputs, single multilingual encoder (Ma et al., 2020; Pan et al., 2021) and dual encoder (He et al., 2020; Radford et al., 2021; Ni et al., 2021) are the most commonly adopted strategies for learning multilingual representations. Multilingual encoder embeds sentences from different languages into a single semantic space using a unified encoder, based on the hypothesis that multilingual learning leads to better multilingual sentence representation. Its architecture is illustrated in A, Figure 2. Dual encoder, also known as two-tower, models the paired data with two independent encoders, and projects the embeddings of paired inputs into the same semantic space through joint training. Its architecture is illustrated in B, Figure 2.

Alignment and uniformity. Wang and Isola (2020) identifies two key properties related to contrastive learning that measure the quality of representations. The alignment calculates the expected distance between embeddings of the paired positive instances, while the uniformity measures how well the embeddings are uniformly distributed. Following Gao et al. (2021), we also use these metrics to demonstrate the inner workings of our approach.

3.2 Method

Although multilingual encoder and dual encoder can use parallel translation pairs straightforwardly,

[4]SNLI+MNLI only include 314K examples.

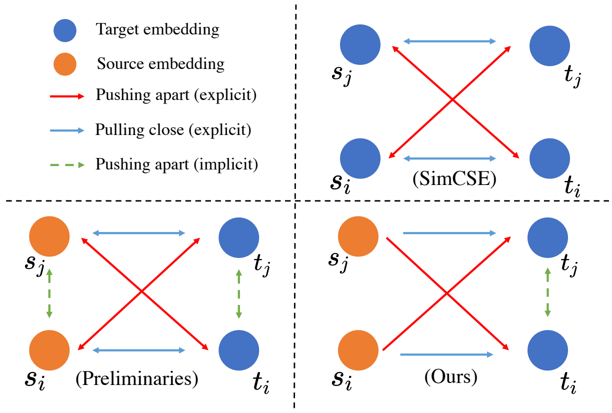

Figure 3: **Illustration of contrastive objectives.** (s_i, t_i) and (s_j, t_j) are two paired samples. In (SimCSE), (s_i, t_i) denotes monolingual pairs, while in (Preliminaries) and (Ours), it denotes parallel translation pairs.

they both suffer from the imbalance between alignment and uniformity, as source language encoder and target language encoder keep updating in the training process. In other words, while they pull the positive samples (source-target translation pairs) closer and the negative samples (source-non target translation pairs) farther away through an explicit contrastive learning objective, the alignment and uniformity of embeddings from monolingual sentence pairs cannot be guaranteed. Specifically, let $(s_i,\ t_i)$ denote the representation of a parallel translation pair generated by the source language encoder and target language encoder, respectively. We simplify the explicit contrastive objective as Eq 1.

$$L_{explicit} = \alpha_1 * L_p - \alpha_2 * L_n \quad (1)$$

Where L_p and L_n represent the distance for positives and negatives of parallel translation pairs as defined in Eq 2 and Eq 3, α denote the linear weights, D is a distance function, and $i \neq j$. The explicit contrastive objective is to minimize the distance between positives and maximize the distance between negatives.

$$L_p = D(s_i, t_i) + D(s_j, t_j) \quad (2)$$

$$L_n = D(s_i, t_j) + D(s_j, t_i) \quad (3)$$

Given parallel translation pairs, we also define the implicit or actual objective that has not been considered into contrastive learning framework in Eq 4, which measures the alignment and uniformity of monolingual sentence embeddings. Although $L_{implicit}$ is not considered in the explicit contrastive objective, we expect to retain good alignment and uniformity of monolingual sentence embeddings from the target encoder, as the actual objective is to learn monolingual sentence embeddings from parallel translation pairs.

$$L_{implicit} = \beta_1 * L_p^{'} - \beta_2 * L_n^{'} \quad (4)$$

Where $L_p^{'}$ and $L_n^{'}$ represent the distance for positives and negatives of monolingual pairs as defined in Eq 5 and Eq 6. s_i^+ and t_i^+ represent the monolingual positive samples for s_i and t_i, respectively. β denote linear weights.

$$L_p^{'} = D(s_i, s_i^+) + D(t_i, t_i^+) \quad (5)$$

$$L_n^{'} = D(s_i, s_j) + D(t_i, t_j) \quad (6)$$

In preliminaries, as shown in (A) and (B), Figure 2, the source language encoder keeps updating in training and can not provide consistent supervision for the target language encoder. The implicit objective for preliminaries is Eq 4, where the alignment and uniformity of source embeddings and target embeddings are both required to be implicitly optimized. However, given two independent implicit objectives, it becomes hard to find a local optimum through Eq 1 without any constraints.

To effectively improve the uniformity and retain the alignment simultaneously, and optimize the implicit objective (4) through an explicit objective (1), we propose to soften the implicit objective for better optimization with our modified architecture, built on the top of regular dual encoder. To be clear, we freeze the side of the source language encoder, so that the alignment and uniformity of source embeddings are frozen in the training. In this case, the implicit objective degrades to Eq 7.

$$L_{implicit} = \beta_1 * D(t_i, t_i^+) - \beta_2 * D(t_i, t_j) \quad (7)$$

As the optimization space shrinks and the implicit objective relaxed, finding the local optimal solution becomes easier and more efficient. We show the differences between our approach (C) and preliminaries (A, B) in Figure 2.

3.3 Analysis

We first analyze the connection between our approach and SimCSE (Gao et al., 2021) and claim that the modified dual architecture with parallel translation pairs as input shares the same implicit

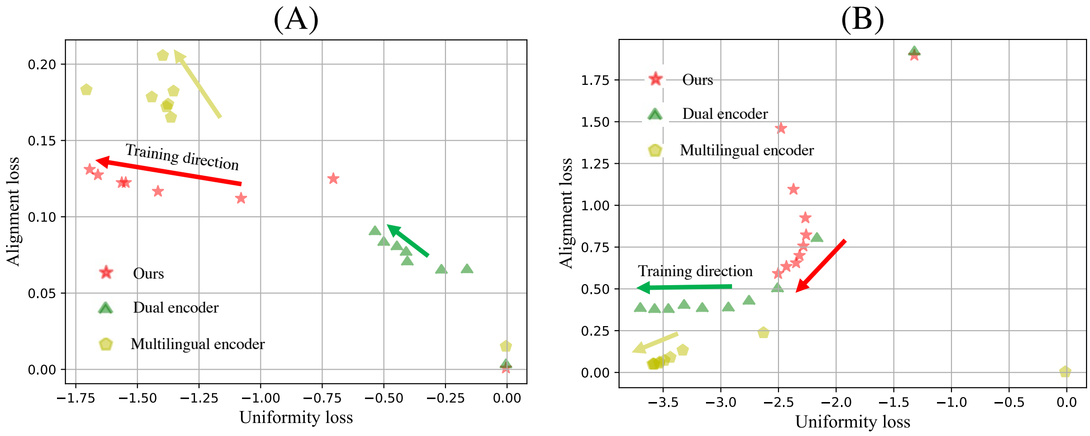

Figure 4: $Loss_{align}$-$Loss_{uniform}$ Plot. We visualize checkpoints every 100 training steps, and the arrows indicate the training direction. (A) shows the results of target encoder given monolingual sentence pairs as input, (B) shows the uniformity and alignment loss of multilingual embeddings given parallel translation pairs as input. Training details refer to 4.4.2. For both $Loss_{align}$ and $Loss_{uniform}$, lower values are better.

contrastive objective as SimCSE with monolingual pairs as input. Then, we provide the visualization results of alignment and uniformity that show superior performance compared to preliminaries.

Connection to SimCSE. As shown in Figure 3, the explicit objective of SimCSE is defined in Eq 1. However, as SimCSE adopts a single monolingual encoder, the source and target language encoder refers to the same model. Given monolingual sentence pairs, $t_i = s_i^+$ is valid, and the implicit objective defined in Eq 4 is identical to its explicit objective. The alignment and uniformity of target language embeddings are optimized in the training. In our approach, as the source encoder is frozen, we soften the implicit objective to the alignment and uniformity of monolingual target embeddings as SimCSE. The only difference is that we optimize the target encoder implicitly with parallel translation pairs, while SimCSE optimizes explicitly with monolingual sentence pairs.

Visualization of alignment and uniformity. To validate the effectiveness of our approach, we take the checkpoint of our model and preliminaries every 100 steps during training and visualize their alignment and uniformity (Wang and Isola, 2020) on a monolingual sentence-pair dataset and parallel translation dataset in Figure 4, training details can be found in 4.4.2 and the data used for visualization is in Appendix A. In A, Figure 4, we show the promising results of implicit objective (the alignment and uniformity of target encoder), given monolingual sentence pairs as input, where we greatly improve uniformity and retain a steady alignment, while others dramatically degrade alignment. In B, Figure 4, We also compare the convergence of explicit objective between three models. Starting from pre-trained checkpoints, all models greatly improve uniformity given parallel translation pairs as input. In contrast, we achieve a better training direction in alignment than other methods, which exhibits a more consistent convergence in cross-lingual training.

4 Experiments

We first describe the datasets in 4.1, and illustrate the training details in 4.2. Then in 4.3, we conduct comprehensive experiments to evaluate the effectiveness of our method. Lastly, we do ablation studies for further analyzing in 4.4.

4.1 Training Datasets

We adopt WMT and source-mixed datasets that have parallel translation pairs for cross-lingual contrastive learning, while the Chinese NLI dataset that has monolingual Chinese sentence pairs is only utilized for fine-tuning.

WMT Dataset[5] is a common-used machine translation dataset composed of various sources. We perform an elaborate cleaning process following (Meng et al., 2020) to filter out low-quality pairs. We get 19,442,200 Chinese-English translation parallel pairs after cleaning.

Source-mixed Dataset collects from more open-sourced translation datasets built on the top of WMT dataset, including AIC (Wu et al., 2017), translation2019zh (Xu, 2019), UN Corpus (Ziemski et al., 2016), etc. Finally, we establish a

[5]http://www.statmt.org/wmt20/

larger-scale dataset including 56,741,808 Chinese-English translation pairs. This dataset is used to show that further scaling up the size of the training set helps improve overall performance.

Chinese NLI Dataset[6] is a Chinese Nature Language Inference dataset which is similar to NLI dataset (Bowman et al., 2015; Williams et al., 2017). We adopt the same method in SimCSE (Gao et al., 2021) to handle the Chinese NLI dataset: given one premise (sentence), we regard the absolutely true (entailment) sentence as the positive, and the definitely false (contradiction) sentence as the hard negative. We establish a dataset containing 315,298 triplets, and each triplet has 3 sentences: premise, positive, hard negative sentences.

4.2 Training Details

We elaborate the training details of our pipeline that is shown in Figure 1. We maintain a consistent memory queue (He et al., 2020) of negative embeddings, where the current mini-batch of the source language encoder's embeddings are enqueued and the oldest are dequeued. The pooling method used in the training is [CLS] with an MLP layer following SimCSE. All experiments are conducted on 8 V100 GPUs. The batch size in experiments represents the batch size on each GPU.

4.2.1 Training a Chinese Encoder

As shown in (A), Figure 1, the first step is to train a target language (Chinese) encoder. Specifically, we adopt the pre-trained SimCSE-RoBERTa$_{large}$ model as the source language (English) encoder, and initialize a Chinese RoBERTa$_{large}$ model[7] with pre-trained weights as the target language (Chinese) encoder. We adopt a series of hyperparameters from 4.2.2: learning rate is 5e-5, batch size is 200, queue size is 200,000, dropout is 0.1, and the input sentence length is 50. In addition, a cosine learning rate scheduler is applied for maintaining the consistency of training. We freeze the source language (English) encoder and only update the target language (Chinese) model. We evaluate every 250 training steps on the development set of Chinese STS-B and save the best checkpoint. The target language (Chinese) model is trained for 2 epochs on WMT or source-mixed dataset. To further boost the performance of the target language (Chinese) model, we fine-tune it on Chinese NLI dataset, with the same settings as described in section 4.2.3.

[6]https://github.com/pluto-junzeng/CNSD

[7]https://huggingface.co/hfl/chinese-RoBERTa-wwm-ext-large

4.2.2 Training an English Encoder

As shown in B, Figure 1, we train a target language (English) encoder that generates sentence embeddings. Specifically, we reuse the pre-trained Chinese encoder from 4.2.1 as the source language (Chinese) encoder and freeze its parameters. We evaluate every 250 training steps on the development set of STS-B and save the best checkpoint.

Effect of Temperature. Temperature is a crucial factor which impacts training convergence and the overall performance in contrastive learning. We evaluate several temperatures recommended by previous works (Gao et al., 2021; Ni et al., 2021; Radford et al., 2021), including 0.05, 0.01, parameter 1 (a learnable parameter in training). As shown in Table 1, a parameter 1 works best.

Temperature	0.01	0.05	Parameter 1
BERT$_{base}$	81.59	86.93	**87.73**

Table 1: Effect of the temperature.

For BERT$_{base}$ (or RoBERTa$_{base}$), the learning rate is we-4, batch size is 400, queue size is 10000, temperature is parameter 1 and the dropout is defaulted set as 0.1. We leverage the cosine learning rate scheduler to adjust the learning rate dynamically. In the term of RoBERTa$_{large}$ (or BERT$_{large}$), we set the learning rate to 5e-5, batch size to 200, queue size to 200,000, all other hyperparameters keep the same as BERT$_{base}$. Refer to appendix B for grid search of hyperparameters.

4.2.3 Fine-tune on NLI Dataset

We investigate the effect of scaling up training dataset by fine-tuning on NLI dataset. The NLI dataset contains 275,602 samples, and each sample consists of a query sentence, a positive sentence, and a hard negative sentence. Following the similar training setting as SimCSE, we set the learning rate to 1e-5, batch size to 128, dropout to 0.1, temperature to 0.05, and input length to 50 for small models (BERT$_{base}$ and RoBERT$_{base}$). While for large models (BERT$_{large}$ and RoBERTa$_{large}$), we set batch size to 96.

4.3 Evaluation Results

Following Gao et al., we evaluate our models on seven transfer and seven STS tasks by SentEval

Model	Fine-tune data	STS12	STS13	STS14	STS15	STS16	STSb	SICK-R	Avg
SBERT$_{base}$	NLI	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
SBERT$_{base}$-flow	NLI	69.78	77.27	74.35	82.01	77.46	79.12	76.21	76.60
SBERT$_{base}$-whitening	NLI	69.65	77.57	74.66	82.27	78.39	79.52	76.91	77.00
CT-SBERT$_{base}$	NLI	74.84	83.20	78.07	83.84	77.93	81.46	76.42	79.39
SimCSE-BERT$_{base}$	NLI	75.30	84.67	80.19	85.40	80.82	84.25	80.39	81.57
Ours-BERT$_{base}$(WMT)	-	80.73	85.82	83.20	88.57	82.50	86.60	80.64	84.01
Ours-BERT$_{base}$(SMD)	-	79.21	87.84	83.24	88.64	82.42	86.87	81.31	84.22
Ours-BERT$_{base}$(WMT)	NLI	**80.85**	87.30	83.42	87.81	83.74	87.42	81.52	84.58
Ours-BERT$_{base}$(SMD)	NLI	80.26	**88.70**	**84.05**	**88.62**	**84.57**	**87.95**	**81.87**	**85.15**
SBERT$_{large}$	NLI	72.27	78.46	74.90	80.90	76.25	79.23	73.75	76.55
SimCSE-BERT$_{large}$	NLI	75.78	86.33	80.44	86.60	80.86	84.87	81.14	82.21
Ours-BERT$_{large}$(WMT)	-	80.71	86.10	83.18	89.13	83.25	86.75	81.43	84.36
Ours-BERT$_{large}$(SMD)	-	79.18	87.75	82.85	88.53	82.60	86.85	81.51	84.18
Ours-BERT$_{large}$(WMT)	NLI	**81.88**	88.78	84.04	88.42	84.94	88.08	81.38	85.36
Ours-BERT$_{large}$(SMD)	NLI	80.86	**89.47**	**84.35**	**88.97**	**85.04**	**88.58**	**81.63**	**85.56**
SRoBERTa$_{base}$-whitening	NLI	70.46	77.07	74.46	81.64	76.43	79.49	76.65	76.60
SimCSE-RoBERTa$_{base}$	NLI	76.53	85.21	80.95	86.03	82.57	85.83	80.50	82.52
Ours-RoBERTa$_{base}$(WMT)	-	**80.59**	85.36	82.16	87.84	82.30	85.96	80.90	83.59
Ours-RoBERTa$_{base}$(SMD)	-	78.60	87.33	83.22	88.64	83.04	86.59	81.15	84.08
Ours-BRoBERTa$_{base}$(WMT)	NLI	80.25	86.97	82.92	87.97	83.78	87.10	81.06	84.29
Ours-RoBERTa$_{base}$(SMD)	NLI	80.02	**87.90**	**83.64**	**88.59**	**85.26**	**87.59**	**81.32**	**84.90**
SRoBERTa$_{large}$	NLI	74.53	77.00	73.18	81.85	76.82	79.10	74.29	76.68
SimCSE-RoBERTa$_{large}$	NLI	77.46	87.27	82.36	86.66	83.93	86.70	81.95	83.76
Ours-RoBERTa$_{large}$(WMT)	-	79.26	87.80	83.76	88.51	83.76	86.94	81.86	84.56
Ours-RoBERTa$_{large}$(SMD)	-	80.86	88.19	84.34	89.20	83.90	87.47	81.26	85.03
Ours-RoBERTa$_{large}$(WMT)	NLI	**81.24**	88.69	84.58	88.59	**85.55**	88.05	82.00	85.53
Ours-RoBERTa$_{large}$(SMD)	NLI	80.07	**89.45**	**84.64**	**88.85**	85.14	**88.60**	**82.28**	**85.58**
ST5-Enc mean (11B)	NLI	77.42	87.50	82.51	87.47	84.88	85.61	80.77	83.74
ST5-EncDec first (11B)	NLI	80.11	88.78	84.33	88.36	85.55	86.82	80.60	84.94
Ours-BERT$_{base}$(SMD)	NLI	80.26	88.70	84.05	88.62	84.57	87.95	81.87	85.15
Ours-BERT$_{large}$(SMD)	NLI	**80.86**	**89.47**	84.35	**88.97**	85.04	88.58	81.63	85.56
Ours-RoBERTa$_{large}$(SMD)	NLI	80.07	89.45	**84.64**	88.85	**85.14**	**88.60**	**82.28**	**85.58**

Table 2: **Comparison with previous state-of-the-art works in STS tasks.** All results are from Gao et al., 2021; Ni et al., 2021; Reimers and Gurevych, 2019; WMT and SMD represent the model is trained on WMT dataset and source-mixed dataset, respectively. The pooling methods used for comparison can be found in Appendix C, and the Ours-RoBERTa$_{large}$(WMT)'s pooling method is [CLS] with MLP.

tools. As the main goal of learning sentence embeddings is to cluster semantically similar sentences, we also take STS result as the main metric.

Semantic textual similarity tasks. We evaluate our approach under zero-shot and fine-tuned settings, respectively. To fairly compare with previous works (Gao et al., 2021; Ni et al., 2021), we adopt seven STS tasks including STS 2012–2016 (Agirre et al., 2012, 2013, 2014, 2015, 2016), STS Benchmark (Cer et al., 2017) and SICK-Relatedness (Marelli et al., 2014). STS tasks are widely used in measuring the discriminative power of sentence embeddings. In STS, sentence embeddings are evaluated by how well their cosine similarities correlate with human-annotated similarity scores. Suggested by Reimers et al., 2016; Gao et al., 2021, we also report Spearman's correlation coefficients to evaluate the performance.

We start from pre-trained checkpoints of BERT or RoBERTa as the backbone. We divide the comparison into 3 tracks for a comprehensive comparison: BERT track, RoBERTa track, and state-of-the-art track. Specifically, BERT track includes Sentence-BERT (Reimers and Gurevych, 2019), CT-BERT (Carlsson et al., 2020), and SimBERT. RoBERTa track includes SimRoBERTa and Sentence-RoBERTa. In the term of the state-of-the-art track, we compare with Sentence-T5 (Ni et al., 2021) 11B model, which contains 11 billion parameters. Table 2 reports the evaluation results on seven STS tasks. Our approach can substantially improve results on all the datasets with or without extra NLI supervision, greatly outperforming the previous state-of-the-art models. Specifically, our approach outperforms the averaged Spearman's correlation of SimCSE by 1.27-2.65 under a zero-shot setting in all tracks. When using NLI datasets, Ours-BERT$_{base}$ further pushes the state-of-the-art results from 84.94 to 85.15. The gains are more pronounced on RoBERTa encoders, and our method achieves 85.58 with RoBERT$_{large}$.

Transfer Tasks. We evaluate on the following transfer tasks: MR (Pang and Lee, 2005), CR (Hu and Liu, 2004), SUBJ (Pang and Lee, 2004), MPQA (Wiebe et al., 2005), SST-2 (Socher et al., 2013), TREC (Voorhees and Tice, 2000) and MRPC (Dolan and Brockett, 2005). We employ the default configurations from SentEval. Results on transfer tasks are shown in Appendix Table 7.

Benefited from the large scale of parallel translation datasets that boosts the power of contrastive learning, our method learns more generalized sentence representations than previous approaches, and improves performance on transfer tasks.

4.4 Ablation Studies

We investigate the impact of source language encoder and contrastive objectives. We use BERT$_{base}$ (WMT) without fine-tuning as our benchmark.

4.4.1 The effect of source language encoder

To analyze the role of source language encoder, we train a SimCSE-RoBERTa$_{large}$ model on the Chinese NLI dataset directly and use it as the source language (Chinese) encoder. For comparison, we train two RoBERTa$_{large}$ models on the WMT dataset following the steps in 4.2.1 with and without fine-tuning. Then, we train three target language (English) encoders as 4.2.2 given different source language models and evaluate them on the SST-B development set. We report the results in table 3. We also directly evaluate the source language (Chinese) encoder on the Chinese STS-B test dataset. The results are in Table 4. All results reveal the superior performance of our approach.

Source Encoder	SimCSE$_{CN}$	Ours	Ours+F
STS-B	86.58	86.91	**88.06**

Table 3: Performance of target language encoders given different source language encoders on STS-B development dataset. SimCSE$_{CN}$ represents the Chinese SimCSE-RoBERTa$_{large}$. Ours+F and Ours are RoBERTa$_{large}$ that trained by our strategy with and without fine-tuning, respectively.

4.4.2 The effect of contrastive objectives

In 3.1, we describe preliminaries in contrastive learning for handling paired data. Figure 2 shows the differences. To show the effectiveness of our cross-lingual contrastive learning scheme, we train models with multilingual encoder, dual encoder and our modified dual architecture, respectively, and evaluate their performance on STS-B development set. For dual encoder, we adopt the pre-trained source language (Chinese) encoder from 4.2.1 and a pre-trained RoBERTa$_{base}$, then train it via contrastive learning. For multilingual encoder, we adopt a RoBERTa$_{base}$-xlm (Lample and Conneau, 2019) model that accepts multilingual input. For our modified dual architecture, we use the same source and target encoder as dual encoder, while keeping the source encoder frozen. All models are trained on WMT dataset.

Model	SimCSE$_{CN}$	Ours	Ours+F
STS-B$_{CN}$	81.13	81.13	**83.37**

Table 4: Performance of source language encoders on Chinese STS-B test dataset. SimCSE$_{CN}$ represents the Chinese SimCSE-RoBERTa$_{large}$. Ours+F and Ours are RoBERTa$_{large}$ that trained by our strategy with and without fine-tuning, respectively.

Models	Multilingual	Dual	Ours
STS-B	71.02	73.13	**86.82**

Table 5: **The effect of contrastive objectives.** Dual, Multilingual and Ours represent dual encoder, multilingual encoder and our modified dual encoder.

For a fair comparison, we unify the hyperparameters of different objectives: batch size is 128, learning rate is 2e-4, queue size[8] is 0, temperature is parameter 1. The only difference between dual encoder and ours is whether the source language encoder is frozen in the training. Table 5 shows the effectiveness of our approach.

5 Conclusion

In this work, we provide the first exploration of utilizing existing large-scale parallel translation pairs for learning sentence representation, propose a modified dual architecture that well balances the alignment and uniformity of embeddings. We demonstrated that our method achieves a new state-of-the-art on standard semantic textual similarity (STS), and the best performance on corresponding tracks on transfer tasks, outperforming both SimCSE and Sentence-T5.

[8] We gather the samples from other GPUs, so the comparative samples in contrastive learning are 128×8=1024.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 252–263.

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

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511.* ACL (Association for Computational Linguistics).

Eneko Agirre, Johan Bos, Mona Diab, Suresh Manandhar, Yuval Marton, and Deniz Yuret. 2012. * 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). 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).*

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.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473.*

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

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

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

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

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

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

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

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852.*

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

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

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

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918.*

Oleksandr Kolomiyets and Marie-Francine Moens. 2011. A survey on question answering technology from an information retrieval perspective. *Information Sciences*, 181(24):5412–5434.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291.*

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

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, Online. Association for Computational Linguistics.

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*.

Shuming Ma, Jian Yang, Haoyang Huang, Zewen Chi, Li Dong, Dongdong Zhang, Hany Hassan Awadalla, Alexandre Muzio, Akiko Eriguchi, Saksham Singhal, et al. 2020. Xlm-t: Scaling up multilingual machine translation with pretrained cross-lingual transformer encoders. *arXiv preprint arXiv:2012.15547*.

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

Fandong Meng, Jianhao Yan, Yijin Liu, Yuan Gao, Xianfeng Zeng, Qinsong Zeng, Peng Li, Ming Chen, Jie Zhou, Sifan Liu, et al. 2020. Wechat neural machine translation systems for wmt20. *arXiv preprint arXiv:2010.00247*.

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

Jianmo Ni, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, Yinfei Yang, et al. 2021. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. *arXiv preprint arXiv:2105.09501*.

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

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

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Nils Reimers, Philip Beyer, and Iryna Gurevych. 2016. Task-oriented intrinsic evaluation of semantic textual similarity. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 87–96.

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

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

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *arXiv preprint arXiv:2106.13884*.

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.

Anthony D Wagner, E Juliana Paré-Blagoev, Jill Clark, and Russell A Poldrack. 2001. Recovering meaning: left prefrontal cortex guides controlled semantic retrieval. *Neuron*, 31(2):329–338.

Jue Wang, Haofan Wang, Jincan Deng, Weijia Wu, and Debing Zhang. 2021. Efficientclip: Efficient cross-modal pre-training by ensemble confident learning and language modeling. *arXiv preprint arXiv:2109.04699*.

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

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

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

Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. 2017. Ai challenger: A large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Bright Xu. 2019. Nlp chinese corpus: Large scale chinese corpus for nlp.

Yinfei Yang, Gustavo Hernández Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019a. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. *arXiv preprint arXiv:1902.08564*.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2019b. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534.

A Validation Set for Visualization

For monolingual sentence-pair dataset, we adopt the STS-B development set and the same settings as the SimCSE(Gao et al., 2021). For parallel translation dataset, UN Corpus development set is used for our visualization. We take out the first 1000 data of the UN Corpus development set. Then, we use the first 250 as positive samples, and replace the Chinese sentence in the last 750 pairs with other Chinese sentences (randomly selected in remaining data in the UN Corpus development set) as negative samples to build a visual validation set of parallel translation data.

B Hyperparameters

We also provide comprehensive analysis of hyperparameters on cross-lingual contrastive learning, including the size of memory queue, learning rate and batch size. We perform grid-search of batch size $\in \{128, 256, 400, 512\}$, learning rate $\in \{5e-5, 1e-4, 2e-4, 5e-4\}$ and queue size $\in \{1024, 4096, 10000, 50000\}$ for BERT$_{base}$, and batch size $\in \{64, 128, 200\}$, learning rate $\in \{1e-5, 2e-5, 5e-5, 1e-4\}$ and queue size $\in \{10000, 50000, 200000, 300000\}$ for RoBERTa$_{large}$. We evaluate on STS-B development set. The results are shown in Table 6.

	BERT		RoBERTa	
	base	large	base	large
Batch size	400	200	400	200
Learning rate	2e-4	5e-5	2e-4	5e-5
Queue size	10 T	200 T	10 T	200 T

Table 6: Our setting of batch sizes, queue size and learning rates for different models. T represents a thousand.

C The Effect of Pooling

Suggested by Gao et al. (2021), pooling strategies make differences in the performance. Li et al. (2020) shows that taking the average embeddings of the pre-trained model leads to better performance than [CLS]. Here, we consider three different pooling settings: (1) Average Pooling, (2) [CLS] with MLP, (3) [CLS] without MLP. Table 8 shows the comparison between different pooling methods. We evaluate on STS-B development set. As shown, we find that CLS without MLP method

Model	**MR**	**CR**	**SUBJ**	**MPQA**	**SST**	**TREC**	**MRPC**	**Avg**
InferSent-GloVe	81.57	86.54	92.50	90.38	84.18	88.20	75.77	85.59
Sentence Encoder	80.09	85.19	93.98	86.70	86.38	93.20	70.14	85.10
SBERT$_{base}$	83.64	89.43	94.39	89.86	88.96	89.60	76.00	87.41
SimCSE-BERT$_{base}$	82.69	89.25	94.81	89.59	87.31	88.40	73.51	86.51
Ours-BERT$_{base}$(SMD)	**85.78**	**91.26**	**94.90**	**91.41**	**90.77**	**91.40**	**77.74**	**89.04**
SRoBERTa$_{base}$	84.91	90.83	92.56	88.75	90.50	88.60	**78.14**	87.76
SimCSE-RoBERTa$_{base}$	84.92	92.00	94.11	89.82	91.27	88.80	75.65	88.08
SimCSE-RoBERTa$_{large}$	**88.12**	92.37	95.11	90.49	**92.75**	91.80	76.64	89.61
Ours-RoBERTa$_{base}$(SMD)	87.02	92.32	95.21	90.92	**92.75**	92.40	77.91	89.79
Ours-RoBERTa$_{large}$(SMD)	88.02	**92.45**	**95.45**	**91.23**	92.70	**94.80**	76.17	**90.12**

Table 7: Performance on transfer tasks. Results are from Gao et al.; Ni et al.; Reimers and Gurevych. SMD represents the model is pre-trained on source-mixed dataset. The models in comparison are both fine-tuned.

Models	[CLS] w/M	AVG	[CLS] wo/M
BERT$_{base}$	85.19	87.28	**88.08**

Table 8: **The effect of different pooling methods.** [CLS] w/M and [CLS] wo/M represent [CLS] with or without an MLP layer, respectively.

works the best for our models. In addition, we adopt the [CLS] with MLP as the fine-tuned models pooling method, as suggested by SimCSE (because we fine-tune our models by SimCSE method).