TransAug: Translate as Augmentation for Sentence Embeddings

Jue Wang
Zhongnan University of Economics and Law
201821090281@stu.zuel.edu.cn

Haofan Wang
Kuaishou Technology
wanghaofan@kuaishou.com

Xing Wu
Kuaishou Technology

Chaochen Gao
Kuaishou Technology

Debing Zhang
Kuaishou Technology

Abstract

While contrastive learning greatly advances the representation of sentence embeddings, it is still limited by the size of the existing sentence datasets. In this paper, we present TransAug (Translate as Augmentation), which provide the first exploration of utilizing translated sentence pairs as data augmentation for text, and introduce a two-stage paradigm to advances the state-of-the-art sentence embeddings. Instead of adopting an encoder trained in other languages setting, we first distill a Chinese encoder from a SimCSE encoder (pretrained in English), so that their embeddings are close in semantic space, which can be regarded as implicit data augmentation. Then, we only update the English encoder via cross-lingual contrastive learning and frozen the distilled Chinese encoder. Our approach achieves a new state-of-art on standard semantic textual similarity (STS), outperforming both SimCSE and Sentence-T5, and the best performance in corresponding tracks on transfer tasks evaluated by SentEval.

1 Introduction

It has been a fundamental problem in natural language processing to learn sentence embeddings that provide compact semantic representations (Le and Mikolov, 2014; Gao et al., 2021; Ni et al., 2021; Reimers and Gurevych, 2019). Recently, contrastive learning (CL) which aims to learn effective representation by pulling semantically close neighbors together and separate non-neighbors, has widely attracted attention for building universal representations. It is noteworthy that benefit from powerful contrastive learning framework, scaling up the size of dataset greatly improve robustness and generalization of representations as suggested by some previous works (Radford et al., 2021; Chen et al., 2020; Jia et al., 2021; Wang et al., 2021).

SimCSE (Gao et al., 2021) demonstrates that a contrastive objective can be extremely effective when coupled with pre-trained language models. However, the generality and capability of the language model is strictly limited by the size of parallel sentence pairs (less than 1 million). To alleviate this issue, it is sensible and practical to construct a comparably large-scale paired sentence dataset through translation, inspired by previous works (Pan et al., 2021; Feng et al., 2020; Yang et al., 2019a) in multilingual filed, yet there is no efficient way to utilize the translated pairs for sentence representation learning.

In this paper, we provide the first exploration of using translated sentence pairs as data augmentation (TransAug) for text, and introduce a two-stage paradigm to learn superior sentence embeddings. To construct positive embedding pairs for contrastive learning, the most naive idea is to employ two independent encoders trained on different language datasets (Chinese and English in our case) to produce sentence embeddings given the translated pair as input, or adopt a single encoder which is able to accept multi-language input. However, due to the distribution deviation of different language inputs, the generated two embeddings usually can not smoothly lie in the same representation space, which degrades the power of contrastive learning. Thus, instead of directly adopting an existed SimCSE (trained in Chinese) model, we first conduct multilingual semantic distillation (MSD) to obtain a Chinese encoder from a pretrained SimCSE model (trained in English), so that their embeddings are close in semantic space and can be regarded as implicit data augmentation. In stage two, we propose a Cross-lingual contrastive method and a multilingual teacher-student contrastive architecture, where the distilled Chinese encoder (as teacher) is frozen and supervise the English encoder (as student) through contrastive loss (Hadsell et al., 2006). Specifically, the student encoder produces query embeddings and the teacher encoder generates key embeddings and negative em-
Figure 1: The pipeline of two-stage TransAug. Stage 1 and stage 2 describe the distillation and contrastive learning process, respectively. $F^{E-T}$, $F^{C-S}$, $F^{E-S}$, $F^{C-T}$ represent English-teacher, Chinese-student, English-student, Chinese-teacher, respectively. Stage 1.2 and stage 2.2 represent the target of optimization in different stage, the goal in stage 1 is to minimize the distance, while the goal in stage 2 is to discriminate.

beddings, the objective is to distinguish whether the query embeddings match the corresponding key embeddings. The pipeline is illustrated in Figure 1.

To better validate the predominant performance of TransAug, we conduct a comprehensive evaluation protocol following the same setting as SimCSE on seven standard semantic textual similarity (STS) tasks (Agirre et al., 2012, 2013, 2014, 2015, 2016; Cer et al., 2017; Marelli et al., 2014) and seven transfer tasks (Conneau and Kiela, 2018). TransAug achieves a new state-of-art on STS tasks, outperforming SimCSE and Sentence-T5 (Ni et al., 2021) by margin, and also achieves state-of-art performance in corresponding tracks on transfer tasks evaluated by SentEval (Conneau and Kiela, 2018). On the average score of STS tasks, our pre-trained TranAug-bert$_{base}$ with or without fine-tuning surpass SimBert$_{base}$ by 3.58% and 2.65% respectively, and TranAug-RoBERTa$_{large}$ achieves 85.60 on average. Surprisingly, TranAug-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 using translated sentence pairs as data augmentation for text.
2. A two-stage paradigm is introduced to utilize translated sentence pairs and improve the representation of sentence embeddings.
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.

2 Related Work

2.1 Universal 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). About 7 years ago, researchers just represent a sentence by the average embeddings of words in sentence through Word2vec (Mikolov et al., 2013), Glove (Pennington et al., 2014), etc, which results in a poor loss of sentence’s information. With the progress of pre-training, some pre-training objectives like BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019) are utilized to generate the sentence embeddings. However, the performance of embeddings produced by pre-trained models is not ideal enough, even worse than the averaging word embeddings like Glove and Word2vec. To derive sentence embeddings from BERT, Sentence-BERT (Reimers and Gurevych, 2019) use siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity. SimCSE (Gao et al., 2021) introduce a simple contrastive learning framework, which greatly improves state-of-the-art sentence embeddings on semantic textual similarity tasks both on unsuper-
vised and supervised tracks. Sentence-T5 (Ni et al., 2021) investigates producing sentence embeddings from the pre-trained T5 (Raffel et al., 2019) models and fine-tune them on downstream datasets that achieve the leading results in sentence embedding benchmark datasets.

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), Semantic Retrieval (SR), Bitext Retrieval (BR) and Retrieval Question Answering (ReQA), etc. Multilingual Universal Sentence Encoder (Yang et al., 2019b) conduct a multitask trained dual encoder to bridge 16 different languages, and achieves competitive results on SR, BR, ReQA tasks. LaBSE (Feng et al., 2020) adopt a dual encoder with additive margin softmax combined masked language model (MLM) and translation language model (TLM) to improve multilingual sentence embeddings. mRASP2 (Pan et al., 2021) hypothesis that a universal cross-language representation 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 a close semantic space. Different from previous works that focus on embedding text from multiple languages into a close semantic space, we propose to utilize translation datasets as data augmentation or amplification for learning robust universal sentence embedding.

3 Methods

In this section, we first compare two practical strategies for translated sentence pairs, then illustrate the two-stage paradigm of our proposed TransAug and show the necessity of each stage. The pipeline of TransAug is shown in Figure 1.

3.1 Preliminary

In this subsection, we briefly introduce two preliminaries for utilizing paired, which have been commonly used in contrastive learning approaches.

Multilingual Single Encoder (Yang et al., 2019b; Pan et al., 2021) embeds sentence from different languages into a single semantic space using a unified encoder, based on the hypothesis that a universal cross-language learning leads to better sentence representation. Its architecture is illustrated in A, Figure 2.

Dual Encoder (He et al., 2020; Radford et al., 2021; Ni et al., 2021), also known as two-tower, models the paired data with two separate encoders, and project the representation of paired input into the same embedding space through jointly training. Its architecture is illustrated in B, Figure 2.

However, these methods are not designed to learn universal representation of sentence and lead to poor generalization. Thus, to make up this gap, we slightly modify the dual encoder architecture and propose a two-stage paradigm (TransAug) to advance the representation of sentence. We introduce the two-stage in Section 3.2 and Section 3.3 respectively, and conduct comprehensive experiments to verify the effectiveness of TransAug compared with two preliminaries in Section 4.4.2. The simplified comparison is shown in Figure 2.

3.2 Multilingual Semantic Distillation

In the first stage, we conduct multilingual semantic distillation (MSD) to obtain a Chinese sentence encoder that has similar semantic space as the English sentence encoder. Specifically, we freeze the pre-trained SimCSE model (trained in English) and use its embeddings to supervise a BERT or RoBERTa (pre-trained in Chinese), and minimize the L2 distance between the English embeddings and Chinese
3.3 Cross-Lingual Contrastive Learning

After we obtain a distilled Chinese encoder that generate embeddings closing to English sentence encoder from stage one, the next key step is to apply contrastive objectives on translation datasets. Different from previous works that adopt a single (Gao et al., 2021; Pan et al., 2021) or dual (He et al., 2020; Radford et al., 2021; Ni et al., 2021) encoder as backbone, TransAug introduces a new-fashioned multilingual teacher-student architecture to conduct contrastive learning effectively.

Why not use a pre-trained encoder? To encode the translated sentence pair, the most direct way is adopting an existed pre-trained encoder (trained in Chinese). However, as the distribution deviation of language datasets, the generated two embeddings usually do not lie in the similar representation space, which degrades the power of contrastive learning. Thus, instead of directly adopting an existed SimCSE (trained in Chinese) model, we first conduct multilingual semantic distillation to obtain a Chinese encoder from a pretrained SimCSE model (trained in English), so that their embeddings are close in semantic space and can be regarded as implicit data augmentation. To validate the hypothesis that distilled Chinese embedding is closer to English embedding, we randomly sample 5 translated pairs (Chinese-English sentences) from the WMT\(^1\) validation set, and visualize the generated embeddings through PCA (Shlens, 2014) method. As shown in Figure 3, the distilled Chinese embedding is indeed closer to the corresponding English embedding than its counterpart from a pre-trained Chinese encoder. To better show the effectiveness of multilingual semantic distillation, we also conduct an ablation study in Section 4.4.1 to confirm that the distilled Chinese encoder is superior to pre-trained Chinese encoder.

---

\(^1\)http://www.statmt.org/wmt20/

Figure 4: Embedding distribution of the teacher and student model in CCL. The dots in the same color are the representations of the corresponding pair. The dots connected by the solid red line are in Chinese, and the solid blue line is in English.

We assign the distilled Chinese encoder from stage one as teacher model and a pre-trained BERT or RoBERTa model as student model. Then, we freeze the teacher model and only use its consistent embeddings to build a large memory queue, and produce the key embeddings. For student model, it supplies the query embedding to match the key embedding via contrastive learning framework.

Notably, we notice that using the frozen robust embeddings from teacher model as the contrastive labels to separate the student model’s embeddings greatly encourages the training efficiency and makes embedding from student model more discriminative. As visualized in Figure 4, where we randomly sample 10 pairs from WMT evaluation set and visualize the embedding distribution, the semantic space of the student model is larger than its teacher counterpart. We provide more analysis in Section 4.4.2.
4 Experiments

We first introduce the datasets adopted in our work, and illustrate details of each module in our proposed framework. Then, we conduct experiments on 7 semantic textual similarity (STS) tasks following previous work (Gao et al., 2021; Ni et al., 2021). We also evaluate 7 transfer learning tasks and provide detailed results. Finally, we do ablation studies to validate the effectiveness of proposed modules.

4.1 Datasets

For the two-stage training process, we use two datasets in our work: WMT dataset and a large-scale dataset collecting from the internet.

4.1.1 WMT Dataset

This is a common-used machine translation dataset composed of a collection of various sources. We translate the original sentence in English to Chinese. The corpus file has 19,442,200 Chinese-English parallel sentence pairs.

4.1.2 Source-mixed Dataset

To further scale up the size of the training dataset, we extra collect open-sourced translation datasets from the internet on the top of WMT dataset, including AIC (Wu et al., 2017), translation2019zh (Xu, 2019), etc. Finally, we establish a larger-scale dataset including 67,307,798 Chinese-English parallel pairs.

4.2 Training Details

We elaborate the training details of Multilingual Semantic Distillation (MSD) and Cross-Lingual Contrastive Learning (CCL), respectively. All experiments are conducted on 8 NVIDIA V100 GPUs.

4.2.1 Multilingual Semantic Distillation

In the stage one of TransAug, the main goal is to obtain a Chinese encoder that generate embeddings closed to the original English embeddings in semantic space. Specifically, we adopt the pre-trained SimCSE-RoBERTa large model as the English-teacher encoder, and establish a RoBERTa large model as the Chinese-student encoder. We set learning rate to 0.00005, batch size to 160, dropout to 0.1, and the input sentence length to 50. In addition, a cosine learning rate scheduler is applied for maintaining the consistency of training. We freeze the teacher encoder and only update the student model as regular multilingual semantic distillation setting, which minimize the distance between English and Chinese embeddings. The student model is trained for 2 epochs with source-mixed dataset.

4.2.2 Cross-Lingual Contrastive Learning

After obtain a distilled Chinese encoder from stage one, the next step is to conduct efficient contrastive learning for utilizing cross-lingual embeddings. To be more specific, we froze the parameters of the distilled encoder, and align the same training setting as SimCSE. We evaluate every 250 training steps on the dev set of STS-B and keep the best checkpoint for the final evaluation on test sets.

We also provide comprehensive analysis of hyperparameters on cross-lingual contrastive learning, including the size of negative sample queue, learning rate and batch size. All experiments are conducted on STS-B development set.

Size of Memory Queue. The negative sample queue is a critical component in the contrastive learning framework. We analyze the effect of queue size for different student backbones (BERT\textsubscript{base} and RoBERTa\textsubscript{large}) on cross-lingual learning process.

| Queue size | 1024 | 4096 | 10T | 50T |
|------------|------|------|-----|-----|
| BERT\textsubscript{base} | 87.82 | 88.08 | 87.79 | 87.92 |

Table 2: Effect of the queue size on BERT\textsubscript{base}. T is the abbreviation for Thousands.

| Queue size | 10T | 50T | 200T | 300T |
|------------|-----|-----|------|------|
| RoBERTa\textsubscript{large} | 87.95 | 88.04 | 88.36 | 87.36 |

Table 3: Effect of the queue size on RoBERTa\textsubscript{large}. T is the abbreviation for Thousands.

As shown in Table 2 and Table 3, BERT\textsubscript{base} achieves the best result with a small queue size, while RoBERTa\textsubscript{large} requires a large queue size for better performance.

Effect of Learning Rate. We perform grid searching for finding a suitable learning rate both for BERT\textsubscript{base} and RoBERTa\textsubscript{large}. The results are reported in 4 and 5 respectively.

| LR | 5e-5 | 1e-4 | 2e-4 | 5e-4 |
|----|------|------|------|------|
| BERT\textsubscript{base} | 86.16 | 87.95 | 88.08 | 84.68 |

Table 4: Effect of the learning rate on BERT\textsubscript{base}.

| LR | 1e-5 | 2e-5 | 5e-5 | 1e-4 |
|----|------|------|------|------|
| RoBERTa\textsubscript{large} | 86.47 | 86.86 | 88.36 | 87.76 |

Table 5: Effect of the learning rate on RoBERTa\textsubscript{large}.
Table 8: Effect of the temperature. L1 represents the learnable temperature 1.

| Model                  | Fine-tune data | STS12 | STS13 | STS14 | STS15 | STS16 | STSb | SICK-R | Avg |
|------------------------|----------------|-------|-------|-------|-------|-------|------|--------|-----|
| SBERTbase              | NLI            | 70.97 | 76.53 | 73.19 | 79.09 | 74.30 | 77.03 | 72.91  | 74.89|
| SBERTbase-flow         | NLI            | 69.78 | 77.27 | 74.35 | 82.01 | 77.46 | 79.12 | 76.21  | 76.60|
| SBERTbase-whitening    | NLI            | 69.65 | 77.57 | 74.66 | 82.27 | 78.39 | 79.52 | 76.91  | 77.00|
| CT-SBERTbase           | NLI            | 74.84 | 83.20 | 78.07 | 83.84 | 77.93 | 81.46 | 76.42  | 79.39|
| SimCSE-BERTbase        | NLI            | 75.30 | 84.67 | 80.19 | 85.40 | 80.82 | 84.25 | 80.39  | 81.57|
| TransAug-BERTbase(WMT) | N/A            | 80.13 | 86.80 | 83.22 | 88.72 | 82.42 | 86.73 | 81.15  | 84.17|
| TransAug-BERTbase(SMD) | N/A            | 79.21 | 87.84 | 83.24 | 88.64 | 82.42 | 86.87 | 81.31  | 84.22|
| TransAug-BERTbase(SMD) | NLI            | 81.08 | 88.19 | 84.07 | 88.28 | 84.48 | 87.14 | 81.36  | 84.94|
| TransAug-BERTbase(SMD) | NLI            | 80.26 | 88.70 | 84.05 | 88.62 | 84.57 | 87.95 | 81.87  | 85.15|
| SBERTlarge             | NLI            | 72.27 | 78.46 | 74.90 | 80.90 | 76.25 | 79.23 | 73.75  | 76.55|
| SimCSE-BERTlarge       | NLI            | 75.78 | 86.33 | 80.44 | 86.60 | 80.86 | 84.87 | 81.14  | 82.21|
| TransAug-BERTlarge(WMT)| N/A            | 78.41 | 87.23 | 83.21 | 89.08 | 82.97 | 87.00 | 81.65  | 84.22|
| TransAug-BERTlarge(SMD)| N/A            | 79.18 | 87.75 | 82.85 | 88.53 | 82.60 | 86.85 | 81.51  | 84.18|
| TransAug-BERTlarge(WMT)| NLI            | 80.86 | 88.93 | 84.01 | 88.81 | 84.71 | 87.96 | 81.03  | 85.19|
| TransAug-BERTlarge(SMD)| NLI            | 80.86 | 89.47 | 84.35 | 88.97 | 85.04 | 88.58 | 81.63  | 85.56|

Table 1: Comparison with previous state-of-the-art works in STS task. 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.

Effect of Batch Size. As shown in Table 6, we set the batch size of BERTbase to 400 for the best result. Restricted by the computing resource, 160 is the largest batch size we can set for RoBERTalarge.

| Batch size | 128 | 256 | 400 | 480 |
|------------|-----|-----|-----|-----|
| BERT-base  | 85.33 | 87.87 | **88.08** | 88.00 |

Table 6: Effect of batch size on BERTbase.

| Batch size | 64 | 128 | 160 | /   |
|------------|----|-----|-----|-----|
| RoBERTalarge | 87.75 | 87.82 | **88.36** | /   |

Table 7: Effect of batch size on RoBERTalarge.

Effect of Temperature. Temperature is a crucial factor which impact the coverage of training and the model’s performance in contrastive learning. We evaluate a number of temperatures recommended by previous works (Gao et al., 2021; Ni et al., 2021; Radford et al., 2021), including 0.05, 0.01, learnable temperature 1 (a learnable parameter in training). As shown in Table 8, a learnable temperature 1 works best.

| Temperature | 0.01 | 0.05 | L1  |
|-------------|------|------|-----|
| BERTbase    | 82.21 | 86.80 | **88.08** |

Table 8: Effect of the temperature. L1 represents the learnable temperature 1.

For BERTbase, the learning rate is 0.0002, batch size is 400, queue size is 4096, 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 RoBERTalarge, we set the learning rate to 0.00005, batch size to 160, queue size to 200,000, all other hyperparameters keep the same as BERTbase-uncased.

4.2.3 Finetune on NLI Dataset

We investigate whether more training data are additive for better sentence representations by fine-tuning on NLI dataset (SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2017)). The NLI dataset contains 275,602 samples and each sample is consisted of a query sentence, a positive sentence, and a hard negative sentence. Following a similar training setting of SimCSE, we set the learning rate to 0.00001, batch size to 128, dropout to 0.1, temperature to 0.05 and input length to 50 for small models (BERTbase and RoBERTalarge). While for large models (BERTlarge-uncased and RoBERTalarge), we set learning rate to 0.00001, batch size to 96.

4.3 Evaluation Results

Following the same setting as previous works (Gao et al., 2021; Ni et al., 2021), we evaluate using
Table 9: Performance on transfer tasks on the SentEval benchmark. All results are from Gao et al., 2021; Ni et al., 2021; Reimers and Gurevych, 2019. SMD represents the model is pre-trained on source-mixed dataset.

| 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|
| Universal Sentence Encoder   | 80.09| 85.19| 93.98| 86.70 | 86.38 | 93.20 | 70.14 | 85.10|
| SimCSE-BERT_{base}           | 83.64| 89.43| 94.39| 89.86 | 88.96 | 91.29 | 88.91 | 87.41|
| SimCSE-ROBERTa_{base} (SMD)  | 84.91| 90.83| 92.56| 87.50 | 89.50 | 92.20 | 88.57 | 78.14|
| TransAug-BERT_{base} (SMD)   | 85.08| 91.68| 94.61| 90.68 | 91.32 | 94.63 | 90.20 | 87.76|
| SimCSE-ROBERTa_{large}       | 88.12| 92.37| 95.11| 90.49 | 92.75 | 91.80 | 76.64 | 89.61|
| TransAug-ROBERTa_{large} (SMD)| 87.22| 92.66| 95.22| 91.34 | 92.59 | 93.40 | 77.62 | 90.01|

4.3.1 Semantic textual similarity tasks

We evaluate TransAug under zero-shot and fine-tuned settings. To fairly compare with previous works (Gao et al., 2021; Ni et al., 2021), we adopt 7 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). For STS, sentence embeddings are evaluated by how well their cosine similarities correlate with human annotated similarity scores, which has been widely used in measuring the discriminative power of sentence embeddings. Suggested by Gao et al., 2021, we also report Spearman’s correlation coefficients.

We start from pre-trained checkpoints of BERT or RoBERTa as backbone. For comprehensive comparison, we divide the comparison into 3 track: 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 State-of-the-art track, we compare with Sentence-T5 (Ni et al., 2021) 11B model, which contains 11 billion parameters. Table 1 reports the evaluation results on 7 STS tasks. TransAug can substantially improve results on all the datasets with or without extra NLI supervision, greatly outperforming the previous state-of-the-art models.

Specifically, TransAug outperforms the averaged Spearman’s correlation of SimCSE by 0.89-2.65 under zero-shot setting. When using NLI datasets, TransAug-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 TransAug achieves 85.60 with RoBERTa_{large}.

4.3.2 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 Table 9.

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

4.4 Ablation Studies

We investigate the impact of multilingual semantic distillation, the multilingual teacher-student architecture and different pooling methods. Our benchmark used in this section is the TransAug-BERT_{base} (WMT) without any fine-tuning.

4.4.1 Choices of Chinese Encoder

In Section 3.2, we have briefly provided the reason why we do not use a pre-trained encoder in stage one. To further support our claim, in stage one, instead of distillation, we train two pre-trained Chinese sentence encoders. One is a RoBERTa_{large} model trained with CCL, the other is a SimCSE-RoBERTa_{large} model. Both are trained on Chinese

https://github.com/facebookresearch/SentEval
NLI dataset\textsuperscript{3}. In stage two, the pre-trained encoders and distilled encoder follow the same setting. We evaluate on SST-B development set and report the result in table 10. As shown, distilled model improves from 86.57 to 88.08 than pretrained model.

| Models | PT-SimCSE | PT-CCL | DT |
|--------|-----------|--------|----|
| STS-B  | 86.06     | 86.57  | **88.08** |

Table 10: Comparison of distilled and pretrained encoders. PT represents ‘pre-trained’ while DT represents ‘distilled’. DT is TransAug-BERT\textsubscript{base} (WMT), PT-SimCSE and PT-CCL are RoBERT\textsubscript{a large} and SimRoberta\textsubscript{large} that trained with SimCSE and CCL strategies, respectively.

### 4.4.2 Choices of Training Strategies

In Section 3.1, we introduce two common strategies in machine translation approaches for handling translated sentence pairs. Figure 2 shows the difference between TransAug and these works. To show the effectiveness of our cross-lingual contrastive learning scheme, we train models with single multilingual encoder, regular dual encoder and our TransAug architecture, respectively, and evaluate their performance on STS-B development set.

For dual encoder, we use the same distilled Chinese encoder from stage one and a BERT\textsubscript{base}, then train via contrastive learning, instead of freezing the parameters of distilled Chinese encoder. In the term of single encoder, we adopt a RoBERT\textsubscript{a base}-xlm (Lample and Conneau, 2019) model that accept multilingual input, and train this model following the same method as SimCSE for RoBERT\textsubscript{a base}. Both are trained on WMT dataset.

| Models | DE | XLM | TBW |
|--------|----|-----|-----|
| STS-B  | 68.10 | 72.71 | **88.08** |

Table 11: Comparison of different strategies for translated sentence pairs. DE, XLM and TBW represent dual encoder, single multilingual encoder and our TransAug-BERT\textsubscript{base}(WMT).

The result of the correlation analyses is shown in 11. The multilingual teacher-student architecture exhibits the best result, showing its great advantages for cross-lingual contrastive learning.

To further analyze its effectiveness, we evaluate the training process with alignment and uniformity (Wang and Isola, 2020) suggested by SimCSE (Gao et al., 2021). We take the checkpoint every 100 steps during training and calculate the alignment and uniformity loss. As clearly shown in Figure 5, all models greatly improve uniformity, especially TransAug-BERT\textsubscript{base}. However, the alignment of the two counterparts degrades drastically, while our TransAug-BERT\textsubscript{base} keeps a steady alignment thanks to a frozen operation.

### 4.4.3 Pooling Methods

As suggested by previous works (Gao et al., 2021), pooling strategies make difference on the performance. Li et al (Li et al., 2020) shows that taking the average embeddings of pre-trained model leads to better performance than [CLS]. Here, we consider three different pooling settings: (1) Average Pooling, (2) CLS, (3) CLS before pooler. Table 12 shows the comparison between different pooling methods in TransAug. We evaluate on STS-B development set. As shown, we find that CLS before pooler method works the best for TransAug.

| Models | CLS | AVG | CBP |
|--------|-----|-----|-----|
| STS-B  | 85.19 | 87.28 | **88.08** |

Table 12: Performance of different pooling methods. CBP represent [CLS] before pooler method.

### 5 Conclusion

In this work, we propose TransAug, a simple but effective data augmentation method for sentence embeddings via translation. To utilize the translated pairs, we introduce a two-stage paradigm to advances the state-of-the-art sentence embeddings.
We demonstrated that TransAug achieves a new state-of-art on both downstream transfer tasks and standard semantic textual similarity (STS), outperforming both SimCSE and Sentence-T5.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Marítxalar, 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.

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 Gylensten, Evangelia Gogouliou, Erik Ylitalo 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.

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

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.

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*.

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-5: 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 and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Jonathon Shlens. 2014. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*.

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.

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.

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*.

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*. 