Multi-stage Distillation Framework for Cross-Lingual Semantic Similarity Matching

Anonymous ACL submission

Abstract

Previous studies have proved that cross-lingual knowledge distillation can significantly improve the performance of pre-trained models for cross-lingual similarity matching tasks. However, the student model needs to be large in this operation. Otherwise, its performance will drop sharply, thus making it impractical to be deployed to memory-limited devices. To address this issue, we delve into cross-lingual knowledge distillation and propose a multi-stage distillation framework for constructing a small-size but high-performance cross-lingual model. In our framework, contrastive learning, bottleneck, and parameter recurrent strategies are delicately combined to prevent performance from being compromised during the compression process. The experimental results demonstrate that our method can compress the size of XLM-R and MiniLM by more than 50%, while the performance is only reduced by about 1%.

1 Introduction

On the internet, it is widespread to store texts in dozens of languages in one system. Cross-lingual similar text matching in multilingual systems is a great challenge for many scenarios, e.g., search engines, recommendation systems, question-answer robots, etc. (Cer et al., 2017; Hardalov et al., 2020; Asai et al., 2021).

In the monolingual scenario, benefiting from the robust performance of the pre-trained language models (PLMs) (e.g., BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), T5 (Raffel et al., 2020), etc.), significant success has been achieved in text-similarity matching tasks. For example, Reimers and Gurevych (2019) proposed the SBERT model trained with similar text pairs and achieved the state-of-the-art performance in the supervised similarity matching. In unsupervised scenarios, Gao et al. (2021) proposed the SimCSE model, which was trained on Wiki corpus through contrastive learning task.

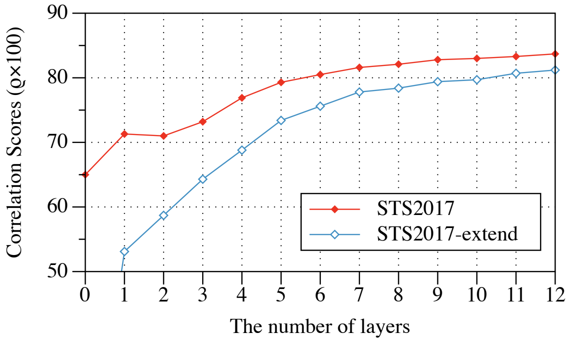

Figure 1: Evaluation results of XLM-R with different number of encoder layers on the STS2017 **monolingual** task and the STS2017-extend **cross-lingual** task, using SBERT-paraphrases for knowledge distillation.

Drawing on the success in the monolingual scenario, researchers began to introduce pre-training technology into cross-lingual scenarios and proposed a series of multilingual pre-trained models, e.g., mBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019), XLM-R (Conneau et al., 2020), etc. Due to the vector collapse issue (Li et al., 2020), the performances of these cross-lingual models on similarity matching tasks are still not satisfactory. Reimers and Gurevych (2020) injected the similarity matching ability of SBERT into the cross-lingual model through knowledge distillation, which alleviated the collapse issue and improved the performance of cross-lingual matching tasks.

Although the cross-lingual matching tasks have achieved positive results, the existing cross-lingual models are huge and challenging to be deployed in devices with limited memory. We try to distill the SBERT model into an XLM-R with fewer layers following Reimers and Gurevych (2020). However, as shown in Figure 1, the performance will be significantly reduced as the number of layers decreases. This phenomenon indicates that cross-lingual capabilities are highly dependent on the model size, and simply compressing the number of layers will bring a serious performance loss.

In this work, we propose a multi-stage distillation compression framework to build a small-size but high-performance model for cross-lingual similarity matching tasks. In this framework, we design three strategies to avoid semantic loss during compression, i.e., multilingual contrastive learning, parameter recurrent, and embedding bottleneck. Besides, we respectively explore the performance impact of reducing the embedding size and encoder size. Experimental results demonstrate that our method effectively reduces the size of the multilingual model with minimal semantic loss. We further investigate the effectiveness of the three strategies through ablation studies. Finally, a series of small-size cross-lingual models are released on Github[1] along with their code.

The main contributions of this paper can be summarized as follows:

- We validate that cross-lingual capability requires a larger model size and explore the semantic performance impact of shrinking the embedding or encoder size.
- A multi-stage distillation framework is proposed to compress the size of cross-lingual models, where three strategies are combined to reduce semantic loss.
- Extensive experiments examine the effectiveness of these three strategies and multi-stages used in our framework.

2 Related work

2.1 Multilingual models

Existing multilingual models can be divided into two categories, namely Multilingual general model and Cross-lingual representation model.

In the first category, transformer-based pretrained models have been massively adopted in multilingual NLP tasks (Huang et al., 2019; Chi et al., 2021; Luo et al., 2021; Ouyang et al., 2021). mBERT (Devlin et al., 2019) was pre-trained on Wikipedia corpus in 104 languages, achieved significant performance in the downstream task. XLM (Conneau and Lample, 2019) presented the translation language modeling (TLM) objective to improve the cross-lingual transferability by leveraging parallel data. XLM-R (Conneau et al., 2020) was built on RoBERTa (Liu et al., 2019) using CommonCrawl Corpus.

In the second category, LASER (Artetxe and Schwenk, 2019a) used an encoder-decoder architecture based on a Bi-LSTM network and was trained on the parallel corpus obtained by neural machine translation. Multilingual Universal Sentence Encoder (mUSE) (Chidambaram et al., 2019; Yang et al., 2020) adopted a bi-encoder architecture and was trained with an additional translation ranking task. LaBSE (Feng et al., 2020) turned the pretrained BERT into a bi-encoder mode and was optimized with the objectives of mask language model (MLM) and TLM. Recently, Mao et al. (2021) presented a lightweight bilingual sentence representation method based on the dual-transformer architecture.

2.2 Knowledge distillation

However, Multilingual models do not necessarily have cross-lingual capabilities, especially in the first category, in which vector spaces of different languages are not aligned. Knowledge distillation (Hinton et al., 2015) used knowledge from a teacher model to guide the training of a student model, which can be used to compress the model and align its vector space at the same time.

For model compression, knowledge distillation aimed to transfer knowledge from a large model to a small model. BERT-PKD (Sun et al., 2019) extracted knowledge from both last layer and intermediate layers at fine-tuning stage. DistilBERT (Sanh et al., 2019) performed distillation at pre-training stage to halve the depth of BERT. TinyBERT (Jiao et al., 2020) distilled knowledge from BERT at both pre-training and fine-tuning stages. MobileBERT (Sun et al., 2020) distilled bert into a model with smaller dimensions at each layer. MiniLM (Wang et al., 2021) conducted deep self-attention distillation.

Unlike previous works presenting general distillation frameworks, we focus on compressing multilingual pre-trained models while aligning their cross-lingual vector spaces. In addition, we take inspiration from Reimers and Gurevych (2020), which successfully aligned the vector space of the multilingual model through cross-lingual knowledge distillation (X-KD). Our framework combines the advantages of X-KD for aligning vectors and introduces three strategies and an assistant model to prevent performance from being compromised during compression.

[1] Will be publicly available once accepted.

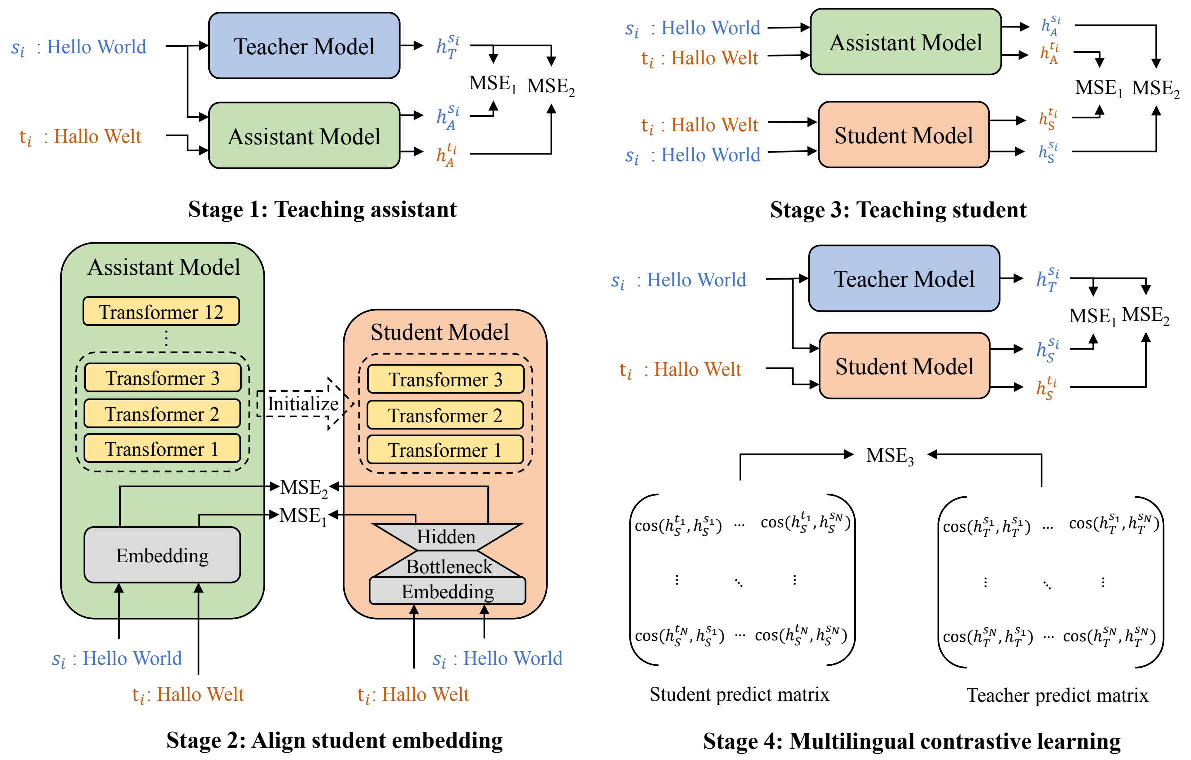

Figure 2: The overview of the model architecture and the multi-stage distillation. It consists of four stages and aims to obtain a small multilingual student model. For convenience, we take the English SBERT as the teacher model, XLM-R as the assistant model. $< s_i, t_i >$ is a pair of parallel sentences in two different language. N is the batch size. MSE is the mean squared error loss function.

3 Method

In this section, we will introduce our method in detail. First, we exhibit the model architecture, and then introduce the multi-stage distillation strategy for the model training. An overview of our approach is shown in Figure 2.

3.1 Model architecture

Given a large-size monolingual model as teacher T and a small-size multilingual model as student S, our goal is to transfer semantic similarity knowledge from T to S and simultaneously compress the size of S with m parallel sentences $P = \{< s_1, t_1 >, < s_2, t_2 >, \cdots < s_m, t_m >\}$.

3.1.1 Teacher model

In this work, we use SBERT (Reimers and Gurevych, 2019) as the teacher model, which has been proven to perform well on monolingual semantic similarity tasks. SBERT adopts a siamese network structure to fine-tune a BERT (Devlin et al., 2019) encoder, and applies a mean pooling operation to its output to derive sentence embedding.

3.1.2 Assistant model

Mirzadeh et al. (2020) proved that when the gap between the student and teacher is large, the performance of the student model will decrease. We hope to get a small student model with cross-lingual capabilities, while the teacher is a large monolingual model. To alleviate the gaps, we introduce an assistant model A (Mirzadeh et al., 2020), which is a large multilingual model with cross-lingual ability.

3.1.3 Student model

Inspired by ALBERT (Lan et al., 2020), we design the student model with Parameter Recurrent and Embedding Bottleneck strategy. Since there is no available multilingual ALBERT, we need to design from scratch.

Parameter Recurrent. We choose the first M layers of the assistant model as a recurring unit (RU). The role of RU is to initialize the student model with layers from the assistant model. Concretely, the RU is defined as,

$$RU = \{L_i | i \in [1, M]\}, \qquad (1)$$

where L_i is the i^{th} transformer layer.

Embedding Bottleneck. Multilingual pre-trained models usually require a large vocabulary V to support more languages, which leads to large embedding layer parameters. We add a bottleneck layer (He et al., 2016; Lan et al., 2020; Sun et al., 2020) of size B between embedding layer and hidden layer H. In this way, the embedding layer is reduced from $O(V \times H)$ to $O(V \times B + B \times H)$.

3.2 Multi-stage distillation

Multi-stage Distillation is the key for enabling the small-size student model with cross-lingual matching ability.

Stage 1. Teaching assistant

As the **Stage 1** in Figure 2, we use the teacher model and parallel corpus to align vector space between different languages through the loss function in (2), enabling its cross-lingual ability (Reimers and Gurevych, 2020).

$$\ell_{stage1} = \frac{1}{|N|} \sum_{i}^{N} \left[(h_T^{si} - h_A^{si})^2 + (h_T^{si} - h_A^{ti})^2 \right], \tag{2}$$

where N is the batch size, and s_i and t_i denotes the parallel sentences in a mini batch.

Stage 2. Align student embedding

As the **Stage 2** in Figure 2, we align the embedding bottleneck layer with the assistant embedding space through the loss function in (3),

$$\ell_{stage2} = \frac{1}{|N|} \sum_{i}^{N} \left[(h_{Ae}^{si} - h_{Be}^{si})^2 + (h_{Be}^{ti} - h_{Ae}^{ti})^2 \right], \tag{3}$$

where h_{Ae}^{si}, h_{Ae}^{ti} denotes the output of assistant embedding layer, h_{Be}^{si}, h_{Be}^{ti} denotes the output of embedding bottleneck layer.

Stage 3. Teaching student

In the **Stage 3**, the student model is trained to imitate the output of the assistant model with loss function in (4),

$$\ell_{stage3} = \frac{1}{|N|} \sum_{i}^{N} \left[(h_A^{si} - h_S^{si})^2 + (h_S^{ti} - h_A^{ti})^2 \right], \tag{4}$$

where h_A^{si}, h_A^{ti} denotes the output of assistant model, h_S^{si}, h_S^{ti} denotes the output of student model.

Stage 4. Multilingual contrastive learning

After the above three stages, we can get a small multilingual sentence embedding model. However, as shown in Figure 1, when the model size decrease, its cross-lingual performance decreases sharply. Therefore, in this stage, we propose multilingual contrastive learning (MCL) task further to improve the performance of the small student model.

Assuming the batch size is N, for a specific translation sentence pair (s_i, t_i) in one batch, the mean-pooled sentence embedding of the student model is (h_S^{si}, h_S^{ti}). The MCL task takes parallel sentence pair (h_S^{si}, h_S^{ti}) as positive one, and other sentences in the same batch $\left\{ (h_S^{si}, h_S^{tj}) | j \in [1, N], j \neq i \right\}$as negative samples. Considering that the MCL task needs to be combined with knowledge distillation. Unlike the previous work (Yang et al., 2019; Feng et al., 2020; Mao et al., 2021), the MCL task does not directly apply the temperature-scaled cross-entropy loss function.

Here, we introduce the implementation of the MCL task. For each pair of negative examples (s_i, t_j) in the parallel corpus, the MCL task first unifies (s_i, t_j) into the source language (s_i, s_j), then uses the fine-grained distance between h_T^{si} and h_T^{sj} in the teacher model to push away the semantic different pair (h_S^{si}, h_S^{tj}) in the student model. For positive examples, the MCL task pull semantically similar pair (h_S^{si}, h_S^{ti}) together. The MCL task loss is (5),

$$\ell_{MCL} = \frac{1}{N^2} \sum_{i}^{N} \sum_{j}^{N} \left(\phi(h_T^{si}, h_T^{sj}) - \phi(h_S^{si}, h_S^{tj}) \right)^2, \tag{5}$$

where ϕ is the distance function. Following prior work (Yang et al., 2019; Feng et al., 2020), we set $\phi(x, y) = cosine(x, y)$. we also add the knowledge distillation task for multilingual sentence representation learning. The knowledge distillation loss is defined as,

$$\ell_{KD} = \frac{1}{|N|} \sum_{i}^{N} \left[(h_T^{si} - h_S^{si})^2 + (h_T^{si} - h_S^{ti})^2 \right]. \tag{6}$$

In stage 4, the total loss function is added by ℓ_{MCL} and ℓ_{KD}.

$$\ell_{stage4} = \ell_{MCL} + \ell_{KD}. \tag{7}$$

Model	AR-AR	ES-ES	EN-EN	Avg.	Embedding size	Encoder size
Pre-trained Model						
mBERT(mean)	50.9	56.7	54.4	54.0	92.20M	85.05M
XLM-R(mean)	25.7	51.8	50.7	42.7	192.40M	85.05M
mBERT-nli-stsb	65.3	83.9	80.2	76.5	92.20M	85.05M
XLM-R-nli-stsb	64.4	83.1	78.2	75.3	192.40M	85.05M
LASER	68.9	79.7	77.6	75.4	23.56M	17.06M
LaBSE	69.1	80.8	79.4	76.4	385.28M	85.05M
Knowledge Distillation						
mBERT← SBERT-nli-stsb	78.8	83.0	82.5	81.4	92.20M	85.05M
XLM-R← SBERT-nli-stsb	79.9	83.5	82.5	82.0	192.40M	85.05M
mBERT← SBERT-paraphrases	79.1	86.5	88.2	84.6	92.20M	85.05M
DistilmBERT← SBERT-paraphrases	77.7	85.8	88.5	84.0	92.20M	46.10M
XLM-R← SBERT-paraphrases	79.6	86.3	88.8	**84.6**	192.40M	85.05M
MiniLM← SBERT-paraphrases	80.3	84.9	85.4	**83.5**	96.21M	21.29M
Ours(Teacher model=SBERT-paraphrases)						
XLM-R($b = True$, bs = 128, $\lvert RU\rvert = 3$)	76.7	84.5	86.6	82.6	32.49M	21.26M
XLM-R($b = True$, bs = 128, $\lvert RU\rvert = 12$)	79.0	85.5	88.4	**84.3**	**32.49M**	85.05M
XLM-R($b = False$, $\lvert RU\rvert = 3$)	79.9	86.8	88.4	**85.0**	192.40M	**21.26M**
Ours(Teacher model=SBERT-paraphrases)						
MiniLM($b = True$, bs = 128, $\lvert RU\rvert = 3$)	72.8	79.3	84.4	78.8	32.05M	5.32M
MiniLM($b = True$, bs = 128,$\lvert RU\rvert = 12$)	79.0	84.4	85.2	**82.9**	**32.05M**	21.29M
MiniLM($b = False$, $\lvert RU\rvert = 3$)	79.9	85.3	85.6	**83.6**	96.21M	**5.32M**

Table 1: Spearman rank correlation ($\rho \times 100$) between the cosine similarity of sentence representations and the gold labels for STS 2017 **monolingual** dataset. b indicates whether to use the Embedding Bottleneck strategy, bs indicates the hidden size of Bottleneck layer. $|RU|$ indicates the first $|RU|$ layers from the basic model are taken as Recurrent Unit, the recurrent times = basic model layers$/|RU|$.

4 Experimental results

4.1 Evaluation setup

Dataset. The semantic text similarity (STS) task requires models to assign a semantic similarity score between 0 and 5 to a pair of sentences. Following Reimers and Gurevych (2020), we evaluate our method on two multilingual STS tasks, i.e., STS2017 (Cer et al., 2017) and STS2017-extend (Reimers and Gurevych, 2020), which contain three monolingual tasks (EN-EN, AR-AR, ES-ES) and six cross-lingual tasks (EN-AR, EN-ES, EN-TR, EN-FR, EN-IT, EN-NL).

Parallel corpus. In stage 1, stage 2 and stage 3, we use TED2020 (Reimers and Gurevych, 2020), WikiMatrix (Schwenk et al., 2021), Europarl (Koehn, 2005) and NewsCommentary (Tiedemann, 2012) as parallel corpus for training. In stage 4, TED2020 is enough for contrastive learning. In this way, the student model first learns generalized multilingual knowledge and then possesses semantic similarity capabilities.

Metric. Spearman's rank correlation ρ is reported in our experiments. Specifically, we first compute the cosine similarity score between two sentence embeddings, then calculate the Spearman rank correlation ρ between the cosine score and the golden score.

4.2 Implementation details

Mean pooling is applied to obtain sentence embeddings, and the max sequence length is set to 128. We use AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate of 2e-5 and a warm-up of 0.1. In stage1, stage2, and stage3, the models are trained for 20 epochs with a batch size of 64, while in stage 4, the student model is trained for 60 epochs. The mBERT, XLM-R used in this work are base-size model obtained from Huggingface's *transformers* package (Wolf et al., 2020), and the MiniLM refers to *MiniLM-L12-H384*[2]

4.3 Performance comparison

We compare the model obtained from our multi-stage distillation with the previous state-of-the-art models, and results are shown in Table 1 and Table 2. In *Pre-trained Model*, mBERT(mean) and XLM-R(mean) are mean pooled mBERT and XLM-R models. mBERT-nli-stsb and XLM-R-nli-stsb are mBERT and XLM-R fine-tuned on the NLI and STS training sets. LASER and LaBSE are obtained from Artetxe and Schwenk (2019b) and Feng et al. (2020). In *Knowledge Distillation*, we use the paradigm of Student←Teacher to represent the Student model distilled from the Teacher model. There are two teacher models, i.e., SBERT-nli-stsb

[2] https://huggingface.co/microsoft/Multilingual-MiniLM-L12-H384

Model	EN-AR	EN-DE	EN-TR	EN-ES	EN-FR	EN-IT	EN-NL	Avg.	Embedding size	Encoder size
Pre-trained Model										
mBERT(mean)	16.7	33.9	16.0	21.5	33.0	34.0	35.6	27.2	92.20M	85.05M
XLM-R(mean)	17.4	21.3	9.2	10.9	16.6	22.9	26.0	17.8	192.40M	85.05M
mBERT-nli-stsb	30.9	62.2	23.9	45.4	57.8	54.3	54.1	46.9	92.20M	85.05M
XLM-R-nli-stsb	44.0	59.5	42.4	54.7	63.4	59.4	66.0	55.6	192.40M	85.05M
LASER	66.5	64.2	72.0	57.9	69.1	70.8	68.5	67.0	23.56M	17.06M
LaBSE	74.5	73.8	72.0	65.5	77.0	76.9	75.1	73.5	385.28M	85.05M
Knowledge Distillation										
mBERT← SBERT-nli-stsb	77.2	78.9	73.2	79.2	78.8	78.9	77.3	77.6	92.20M	85.05M
DistilmBERT← SBERT-nli-stsb	76.1	77.7	71.8	77.6	77.4	76.5	74.7	76.0	92.20M	46.10M
XLM-R← SBERT-nli-stsb	77.8	78.9	74.0	79.7	78.5	78.9	77.7	77.9	192.40M	85.05M
mBERT← SBERT-paraphrases	80.8	83.6	77.9	83.6	84.6	84.6	84.2	82.7	92.20M	85.05M
DistilmBERT← SBERT-paraphrases	79.7	81.7	76.4	82.3	83.2	84.3	83.0	81.5	92.20M	46.10M
XLM-R← SBERT-paraphrases	82.3	84.0	80.9	83.1	84.9	86.3	84.5	**83.7**	192.40M	85.05M
MiniLM← SBERT-paraphrases	81.3	82.7	74.8	83.2	80.3	82.4	82.2	**80.9**	96.21M	21.29M
Ours(Teacher model=SBERT-paraphrases)										
XLM-R($b = True, bs = 128, \|RU\| = 3$)	78.0	79.8	73.9	80.5	82.1	80.3	81.2	79.4	32.49M	21.26M
XLM-R($b = True, bs = 128, \|RU\| = 12$)	79.4	83.6	78.7	83.3	84.2	85.6	84.8	**82.8**	**32.49M**	85.05M
XLM-R($b = False, \|RU\| = 3$)	81.1	84.3	79.8	82.6	84.5	84.8	85.4	**83.2**	192.40M	**21.26M**
Ours(Teacher model=SBERT-paraphrases)										
MiniLM($b = True, bs = 128, \|RU\| = 3$)	73.0	76.0	63.7	71.4	71.8	72.1	74.7	71.8	32.05M	5.32M
MiniLM($b = True, bs = 128, \|RU\| = 12$)	79.7	81.0	74.1	81.9	80.1	80.8	80.7	**79.8**	**32.05M**	21.29M
MiniLM($b = False, \|RU\| = 3$)	82.3	82.8	76.9	82.1	80.5	82.3	82.4	**81.3**	96.21M	**5.32M**

Table 2: Spearman rank correlation ($\rho \times 100$) between the cosine similarity of sentence representations and the gold labels for STS 2017-extend **cross-lingual** dataset. b indicates whether to use Embedding Bottleneck strategy, bs indicates the hidden size of Bottleneck layer. $|RU|$ indicates the first $|RU|$ layers from the basic model are taken as Recurrent Unit, the recurrent times = basic model layers/$|RU|$.

and SBERT-paraphrases, which are released by UKPLab[3]. The former is fine-tuned on the English NLI and STS training sets, and the latter is trained on more than 50 million English paraphrase pairs. The student models include mBERT, XLM-R, DistilmBERT (Sanh et al., 2019) and MiniLM (Wang et al., 2021).

Table 1 and Table 2 show the evaluation results on monolingual and multilingual STS task, respectively. For the XLM-R, our method compresses the embedding size by 83.2% with 0.3% worse monolingual performance and 0.9% worse cross-lingual performance, compresses the encoder size by 75% with slightly higher (0.4%) monolingual performance and 0.5% worse cross-lingual performance. When compressing the embedding layer and the encoder simultaneously, the model size is reduced by 80.6%, its monolingual performance drop by 2% and cross-lingual performance drop by 4%, but it still outperforms the pre-trained models.

For comparison with other distillation methods, MiniLM← SBERT-paraphrases is taken as a strong baseline. Our framework can further compress its embedding size by 66.7% with 0.6% worse in monolingual performance and 1.1% worse in cross-lingual performance. Its encoder size is further compressed by 75% with slightly higher monolingual (0.1%) and cross-lingual (0.4%) performance.

Model	AR-AR	ES-ES	EN-EN	Avg.
ours	76.7	84.5	86.6	82.6
w/o MCL	76.4	83.9	86.8	82.3
w/o Rec.	67.4	80.1	86.6	78.0
w/o MCL+Rec.	67.9	79.3	86.6	77.9

Table 3: Results of ablation studies on STS-2017 monolingual task

Model	EN-AR	EN-DE	EN-TR	EN-ES	EN-FR	EN-IT	EN-NL	Avg.
ours	78.0	79.8	73.9	80.5	82.1	80.3	81.2	79.4
w/o MCL	75.9	79.7	73.2	79.9	80.4	80.4	80.5	78.5
w/o Rec.	69.1	73.4	66.5	70.2	73.7	73.0	75.9	71.7
w/o MCL+Rec.	67.8	73.6	66.4	68.5	72.8	71.8	75.2	70.9

Table 4: Results of ablation studies on STS2017-extend cross-lingual task

In addition, our compressed XLM-R($b = True$, $bs = 128$, $|RU| = 12$) achieves higher monolingual(0.8%) and cross-lingual(1.9%) performance with the same model size.

4.4 Ablation study

Among the three key strategies, multilingual contrastive learning (**MCL**) and parameter recurrent (**Rec.**) are two crucial mechanisms to improve model performance. The bottleneck is used to compress the model. In this section, ablation studies is performed to investigate the effects of **MCL** and **Rec.**. The effects of the bottleneck will be discussed in section 4.7.

[3] https://github.com/UKPLab/sentence-transformers

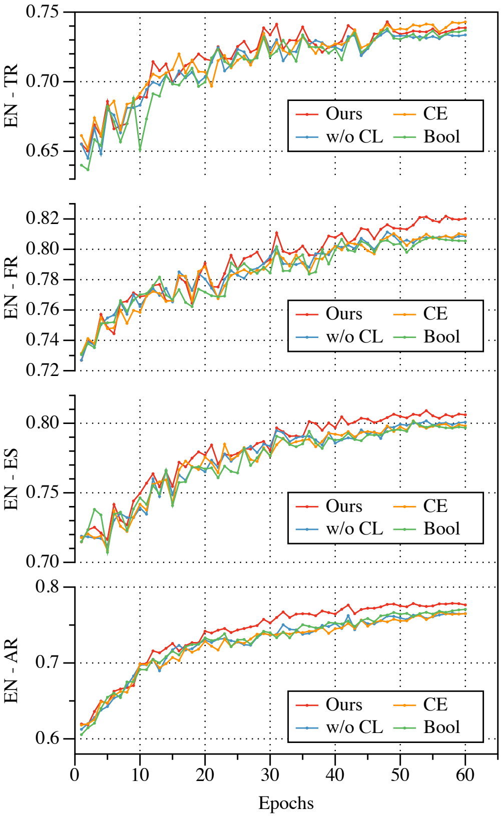

Figure 3: Performance of XLM-R (b=True, bs=128, $|RU|$ = 3) after each training epoch on EN-AR, EN-ES, EN-FR, EN-TR tasks with different contrastive learning settings.

Settings	EN-AR	EN-DE	EN-TR	EN-ES	EN-FR	EN-IT	EN-NL	Avg.
Ours	78.0	79.8	73.9	80.5	82.1	80.3	81.2	79.4
Bool	77.0↓	80.5↑	73.5↓	79.8↓	80.3↓	80.7↑	81.2	79.0↓
CE	76.6↓	79.9↑	74.3↑	80.0↓	80.8↓	80.6↑	80.7↓	78.9↓
w/o CL	75.9↓	79.7↓	73.2↓	79.9↓	80.4↓	80.4↑	80.5↓	78.5↓

Table 5: Evaluation results of XLM-R ($b = True, bs = 128$, $|RU| = 3$) on the STS2017-extend cross-lingual task with different contrastive learning settings.

Settings	Avg. (Monolingual)	Avg. (Cross-lingual)
Single-stage		
Random Initialize	78.1	71.1
+ Pre-Distillation	79.0	73.8
Multi-stage		
stage 1 + 2	48.4	20.8
stage 1 + 2 + 3	75.2	70.6
stage 1 + 2 + 3 + 4	**82.6**	**79.4**

Table 6: Comparison of using different stage settings on monolingual and multilingual STS task. XLM-R is the basic model. The first three layers from XLM-R are taken as a Recurrent Unit, bottleneck hidden size is 128.

XLM-R(b=True, bs=128, $|RU|$ = 3) is selected as the basic model. We consider three different settings: 1) training without MCL task. 2) training without parameter recurrent. 3) training without both. The monolingual results and multilingual results are presented in Table 3 and Table 4..

It can be observed that: 1) without MCL task, the model performs poorer on the cross-lingual tasks. 2) without parameter sharing, the model performs poorer on all datasets. 3) MCL task can significantly improve the cross-lingual performance on EN-AR, EN-ES, EN-FR, EN-NL. It can be concluded that both MCL task and parameter recurrent play a key role in our method.

4.5 Effect of contrastive learning

To investigate the effects of contrastive learning in stage 4, we select XLM-R(b=True, bs=128, $|RU|$ = 3), modify the original objective in (5) into three different settings, namely, *Bool*, *CE* and *w/o CL*.

In the *Bool* setting, the soft label in (5) is replaced with hard label (0 or 1), as (8),

$$\ell_{Bool} = \frac{1}{N^2} \sum_i^N \sum_j^N \left(\delta(h_T^{si}, h_T^{sj}) - \phi(h_S^{si}, h_S^{tj}) \right)^2, \tag{8}$$

where $\delta(x, y) = 1$, if $x = y$, otherwise 0.

In the *CE* setting, the objective in (5) is replaced with temperature-scaled cross-entropy, as (9),

$$\ell_{CE} = -\sum_i^N \sum_j^N \phi_T \log \frac{e^{\phi_S/\tau}}{\sum_{k=1}^N e^{\phi_S/\tau}}, \tag{9}$$

where $\phi_T = cos(h_T^{si}, h_T^{sj})$, $\phi_S = cos(h_S^{si}, h_S^{tj})$, $\tau = 0.05$ is a hyperparameter called temperature.

In the *w/o CL* setting, the contrastive learning is removed in Stage 4.

Table 5 presents the model performance of cross-lingual semantic similarity task with different settings. It can be observed that all the above training objectives can improve the model performance on the cross-lingual task, compared with the *w/o CL* settings. Model trained with (8) and (9) underperform that trained with (5), especially on EN-AR, EN-ES, EN-FR, EN-NL task.

We plot the convergence process of different settings in Figure 3. On EN-AR, EN-ES, EN-FR tasks, our setting outperform other settings. It is worth mentioning that on the EN-TR task, our setting underperform the *CE* setting according to Table 5. However, our setting reaches the same level as *CE* setting during the 30 to 40 epoch.

Model	Monolingual Avg.	Cross-lingual Avg.	Embedding size	Encoder size
Teacher model=SBERT-paraphrases, Student model=XLM-R, $\lvert RU\rvert = 3$				
$b = True, bs = 128$	82.6	79.4	32.49M	21.26M
$b = True, bs = 256$	82.8	80.6	64.59M	21.26M
$b = False$	85.0	83.2	192.40M	21.26M
Teacher model=SBERT-paraphrases, Student model=XLM-R, $b = True, bs = 128$				
$\lvert RU\rvert = 3$	82.6	79.4	32.49M	21.26M
$\lvert RU\rvert = 6$	83.4	81.1	32.49M	42.52M
$\lvert RU\rvert = 12$	84.3	82.8	32.49M	85.05M
Teacher model=SBERT-paraphrases, Student model=MiniLM, $\lvert RU\rvert = 3$				
$b = True, bs = 128$	78.8	71.8	32.05M	5.32M
$b = True, bs = 256$	79.5	72.8	64.10M	5.32M
$b = False$	83.6	81.3	96.21M	5.32M
Teacher model=SBERT-paraphrases, Student model=MiniLM, $b = True, bs = 128$				
$\lvert RU\rvert = 3$	78.8	71.8	32.05M	5.32M
$\lvert RU\rvert = 6$	81.5	76.1	32.05M	10.64M
$\lvert RU\rvert = 12$	82.9	79.8	32.05M	21.29M

Table 7: The performance of STS monolingual and cross-lingual task based on XLM-R(b=True, bs=128, $\lvert RU\rvert$ = 3) and MiniLM(b=True, bs=128, $\lvert RU\rvert$ = 3), We evaluated the effect of increasing bs or $\lvert RU\rvert$.

4.6 Effect of multi-stages

To verify the effectiveness of multi-stages, we shows the performance comparison of using different stage settings in Table 6. In the *Single-stage* setting, we first initialize the shrunk student model in two ways: (1) Random Initialize: Adding the untrained embedding bottleneck layers to the student model. (2) Pre-Distillation: The student model with bottleneck layer is initialized by distillation using XLM-R and the same corpus as section 4.1. Then we follow Reimers and Gurevych (2020) to align vector space between different languages. In the *Multi-stage* setting, the performance of the student model is reported after each stage.

As shown in Table 6, the *Multi-stage* setting outperforms the single-stage one, indicating that our multi-stage framework with an assistant model is effective. Adding stage3 and stage4 further improves the student model performance, suggesting that multi-stage training are necessary.

4.7 Effect of bottleneck and recurrent unit

In this section, we study the impact of embedding bottleneck and recurrent unit strategies on multilingual semantic learning. We consider three settings for each strategy, as shown in Table 7.

First, we found that both XLM-R and MiniLM perform better as the bottleneck hidden size bs increases. The performance is best when the entire embedding layer is retained, The MiniLM(b=False) can outperform its original model in Table 1 and Table 2. But the benefit of increasing bs is not obvious unless the entire embedding layer is retained.

Second, by increasing the number of recurrent unit layers $\lvert RU\rvert$, XLM-R and MiniLM have been steadily improved on these two tasks. The increase in model size caused by the $\lvert RU\rvert$ is less than the bs. For example, the performance of MiniLM on cross-lingual tasks increased by 8%, while its size only increased by 15.9M.

Finally, it can be observed that when using the bottleneck layer (b=True), the model performance will increase steadily as $\lvert RU\rvert$ increases. The smaller the encoder hidden size, the more significant effect caused by $\lvert RU\rvert$ increasing (ΔMiniLM>ΔXLM-R). However, the increase of bs can not improve performance significantly but make the embedding size larger. Therefore, an effective way to compress the multilingual model is reducing bs while increasing $\lvert RU\rvert$. In this way, we shrink XLM-R by 58%, MiniLM by 55%, with less than 1.1% performance degradation.

5 Conclusion

In this work, we realize that the cross-lingual similarity matching task requires a large model size. To obtain a small-size model with cross-lingual matching ability, we propose a multi-stage distillation framework. Knowledge distillation and contrastive learning are combined in order to compress model with less semantic performance loss.

Our experiments demonstrate promising STS results with three monolingual and six cross-lingual pairs, covering eight languages. The empirical results show that our framework can shrink XLM-R or MiniLM by more than 50%. In contrast, the performance is only reduced by less than 0.6% on monolingual and 1.1% on cross-lingual tasks. If we slack the tolerated loss performance in 4%, the size of XLM-R can be reduced by 80%.

References

Mikel Artetxe and Holger Schwenk. 2019a. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics (TACL)*, 7:597–610.

Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 547–564.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 3576–3588.

Muthu Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yunhsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Learning cross-lingual sentence representations via a multi-task dual-encoder model. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 250–259.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186.

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. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6894–6910.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5427–5444.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2485–2494.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*, pages 79–86, Phuket, Thailand.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of International Conference on Learning Representations (ICLR)*.

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.

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*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 2019 International Conference on Learning Representations (ICLR)*.

Fuli Luo, Wei Wang, Jiahao Liu, Yijia Liu, Bin Bi, Songfang Huang, Fei Huang, and Luo Si. 2021. VECO: Variable and flexible cross-lingual pre-training for language understanding and generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 3980–3994.

Zhuoyuan Mao, Prakhar Gupta, Chenhui Chu, Martin Jaggi, and Sadao Kurohashi. 2021. Lightweight cross-lingual sentence representation learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 2902–2913.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5191–5198.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 27–38.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. In *Journal of Machine Learning Research*, pages 1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525.

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

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (ECACL)*, pages 1351–1361.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for BERT model compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4323–4332.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. MobileBERT: a compact task-agnostic BERT for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2158–2170.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, pages 2214–2218.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-Art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 38–45.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 87–94.

Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Improving multilingual sentence embedding using bidirectional dual encoder with additive margin softmax. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5370–5378.