ConGen: Unsupervised Control and Generalization Distillation For Sentence Representation

Peerat Limkonchotiwat[†], **Wuttikorn Ponwitayarat**[†], **Lalita Lowphansirikul**[†],
Can Udomcharoenchaikit[†], **Ekapol Chuangsuwanich**[‡], **Sarana Nutanong**[†]
[†]School of Information Science and Technology, VISTEC, Thailand
[‡]Department of Computer Engineering, Chulalongkorn University, Thailand
{peerat.l_s19,wuttikornp.p_s22,lalita.l_s22
,canu_pro,snutanon}@vistec.ac.th,
ekapolc@cp.eng.chula.ac.th

Abstract

Sentence representations are essential in many NLP applications operating at the sentence level. Recently, research attention has shifted towards learning how to represent sentences without any annotations, i.e., unsupervised representation learning. Despite the benefit of training without supervised data, there is still a performance penalty compared to supervised methods. Furthermore, the supervised-unsupervised performance gap widens as we reduce the model size. In this paper, we propose an unsupervised sentence representation method to reduce the supervised-unsupervised performance gap, especially for smaller models. Utilizing the concept for knowledge distillation, we derive a distillation framework comprising two training objectives, *control* and *generalize*, called ConGen. Experiments on semantic textual similarity (STS), text classification (transfer), and natural language inference (NLI) tasks show that ConGen is on par with supervised training even on smaller models. Furthermore, our method consistently outperformed competitors on multilingual STS.

1 Introduction

In recent years, sentence representation has played a crucial role in various NLP tasks operating at the sentence level (Reimers and Gurevych, 2019; Yang et al., 2020; Zhang et al., 2021; Yang et al., 2021). Many researchers use a transformer language model (LM) (Devlin et al., 2019; Liu et al., 2019), as a backbone of sentence representation by finetuning LM on natural language inference (NLI) and semantic textual similarity (STS) labeled data, which yields promising results (Reimers and Gurevych, 2019; Li et al., 2020). However, these techniques require labeled data during the finetuning process, which can be a limiting factor in low-resource settings.

In order to incorporate unlabeled data into the training process, unsupervised learning paradigms have gained popularity. The contrastive learning paradigm has recently led to significant advancement in unsupervised learning. The main idea of contrastive learning in sentence representation is learning a meaningful representation by maximizing the similarity between differently augmented views (Kim et al., 2021; Yan et al., 2021; Gao et al., 2021; Carlsson et al., 2021; Kim et al., 2021; Giorgi et al., 2021; Liu et al., 2021a; Fang et al., 2020). For example, Gao et al. (2021) proposed a state-of-the-art contrastive framework called SimCSE, the learning framework that benefits from the dropout normalization to produce differently augmented views and works well with unsupervised and supervised learning. In particular, SimCSE is SOTA on the STS benchmark, and the performance gap between unsupervised and supervised settings in SimCSE on large networks (e.g., BERT-base) is only five points when evaluated on the STS benchmark.

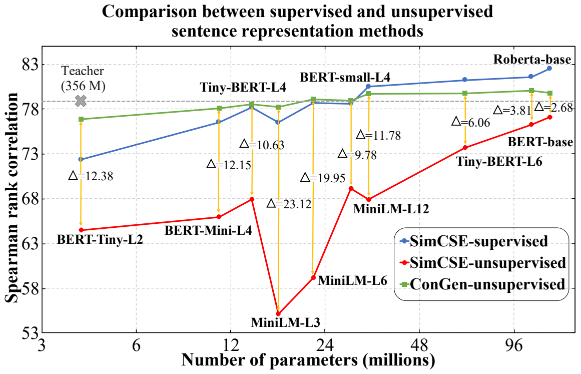


Figure 1: Comparison between finetuning LMs (SimCSE) vs. knowledge distillation (ConGen) on the average of 7 semantic textual similarity (STS) benchmark datasets and Δ is the improvement of ConGen from SimCSE.

However, the performance of SimCSE rapidly degrades as we decrease the model size, which is undesirable when the computational resource is limited, e.g., edge computing (Jiao et al., 2020; Sun et al., 2020b). For instance, when we use the

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 6467–6480
December 7-11, 2022 ©2022 Association for Computational Linguistics

MiniLM-L3 (#parameters: 17M) instead of the BERT-base model (#parameters: 109M), the Spearman rank correlation of SimCSE-unsupervised drops from 76.25 to 55.10 (averaged across 7 STS corpora). The gap between compressed and base LMs is 21.15, as shown in Figure 1. In addition, the gap between unsupervised and supervised learning in MiniLM-L3 is 21.56, while the gap of larger models like BERT-base is only 5.32 points. Maintaining a high performance for supervised and unsupervised learning is challenging for smaller LMs.

In this paper, we aim to retain the advantage of unsupervised learning while mitigating the performance penalty from model compression at the same time. We propose an unsupervised *control* and *generalization* distillation, *ConGen*, a distillation framework that transfers knowledge from a large model to any model regardless of its architecture and size. Not only does ConGen outperform state-of-the-art unsupervised sentence representation, its performance is also similar to supervised learning (Figure 1).

The crux of ConGen lies in the distillation mechanism, which handles two different data augmentation views. In particular, we employ inputs derived from two data augmentation operations, which we refer to as *control* and *generalization*. The student observes both control and generalization inputs, whereas the teacher observes only control input, which we refer to as a *reference* input. We derive a similarity distribution between the student inputs (control and generalization) and the instance queue (He et al., 2020; Fang et al., 2021), and we also do the same to the teacher input (reference). To compare similarity distributions from the teacher to student models, we minimize the discrepancy between the student and teacher distributions in the two following control & generalization learning objectives. First, we *control* the similarity distribution of the control distribution and the reference distribution to be the same. Second, to increase the model's *generalizability*, we enforce the similarity distribution of the generalization distribution to be the same as the reference distribution.

To demonstrate our method's effectiveness, we compare it to other distillation methods in three tasks: semantic textual similarity (STS), text classification (transfer), and natural language inference (NLI). The experimental results from STS demonstrate that our method significantly improves the performance of compression models and consistently outperforms competitors. In addition, when the model parameters are less than 33 Million, ConGen outperforms or matches the supervised baseline (Figure 1). Moreover, in transfer and NLI, ConGen outperforms unsupervised learning, i.e., SimCSE, and other distillation methods in 11 of 12 cases. Additionally, we extend our method to multilingual sentence representation; experimental results from multilingual STS demonstrate that ConGen outperforms competitors in all cases.

The contributions of our work are as follows:

- We extend the relational learning approach often used in computer vision to sentence representation learning by designing a novel training process for textual learning.
- We propose a knowledge distillation technique called ConGen that works regardless of architecture or model size. ConGen distills knowledge from a large model to a small model without using labeled data.
- We design two learning objectives to minimize teacher-student discrepancies, namely, control and generalization.
- We demonstrate the models' performance and efficiency using an extensive set of experimental studies, including three monolingual tasks, one multilingual task, and eight competitors.

2 Related Work

2.1 Unsupervised Sentence Representation

Transformer-based language models (LM), i.e., BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), have shown outstanding performance in many downstream tasks including sentence representation. Contrastive learning is often utilized for training an unsupervised sentence encoder based on a pretrained LM. The main idea behind contrastive loss in unsupervised learning is to enforce similarity between the representations of anchor and positive samples. Anchor samples can be randomly sampled from the training data. In contrast, the positive samples can be obtained from various techniques, e.g., generating from another LM (Carlsson et al., 2021; Kim et al., 2021), sampling sentence from the same document or dialogue (Giorgi et al., 2021; Liu et al., 2021a), generating similar sentences from back-translation operations (Fang et al., 2020), and randomly dropped some features of a vector (Yan et al., 2021; Gao et al., 2021; Liu et al., 2021b), the performance from these techniques outperformed previous unsupervised methods. However, those frameworks only focus on large models (BERT/RoBERTa-base and BERT/RoBERTa-large), without any consideration for smaller models. The experimental results from Wu et al. (2021) have demonstrated that the current SOTA unsupervised learning technique, SimCSE (Gao et al., 2021), fails to produce meaningful sentence representation when SimCSE is trained on compressed LMs.

2.2 Sentence Representational Knowledge Transfer

Knowledge distillation (KD) is a technique for transferring knowledge from a source model (teacher) to a target model (student), where the learning objective is minimizing the discrepancy between the two models. In particular, directly transferring knowledge from a teacher vector to a student model by using prediction outputs (Turc et al., 2019; Sanh et al., 2019) or transformer probabilities (Jiao et al., 2020; Sun et al., 2020b; Wang et al., 2020, 2021c) to create soft labels for compressed student models.

Labeled Sentence-Pair Knowledge Distillation. Reimers and Gurevych (2020) propose an LM finetuning method that minimizes the discrepancy between English and other languages vector representations using the L2 loss. Cheng (2021) propose a dual-view distillation method called DvBERT, which minimizes the discrepancy of a student NLI output with respect to outputs from two teachers using KL divergence. Notably, a recent concurrent work, namely DisCo (Wu et al., 2021) also points out the importance of developing better sentence representation for compressed models. DisCo is based on contrastive distillation (Sun et al., 2020a), where the positive and negative samples in contrastive learning are obtained from a memory bank (Liu and Mukhopadhyay, 2018) produced by a supervised teacher model.

Unlabeled Sentence-Pair Knowledge Distillation. Liu et al. (2022) propose a binary cross-entropy self-distillation called Trans-Encoder. Trans-Encoder imitates the similarity of pair-wise datasets from the teacher to student models by using binary cross-entropy loss. However, the techniques mentioned above require sentence-pair for training, thereby not entirely unsupervised.

Knowledge Distillation without Sentence-Pair. Previous KD literature on sentence representation focuses on weakly-supervised and supervised settings but remains unexplored for unsupervised knowledge distillation. In contrast to previous literature, in the computer vision community, Fang et al. (2021) propose a self-supervised knowledge distillation (SEED) for visual representation learning. SEED is based on two components: (i) large-scale negative samples and (ii) similarity distribution to transfer the knowledge from large to small models without pair-wise or labeled datasets. This allows us to perform unsupervised distillation. We apply these components from computer vision to sentence representation models by designing a new training process, new generalization technique, loss function, and data augmentation methods.

3 Proposed Method

In this section, we describe our *Control* and *Generalization* distillation (ConGen) method. ConGen is a knowledge distillation technique comprising two objectives: (i) transferring the knowledge from large to small models and (ii) improving the model's generalizability. As illustrated in Figure 2, we describe our framework's training process, including how we organize the inputs and outputs, compare the outputs, and train the model.

3.1 How We Organize the Inputs and Outputs

As shown in Figure 2, given a new batch sample x, we first obtain two differently augmented samples $x_1 = \mathcal{T}(x)$ and $x_2 = \mathcal{T}'(x)$, where $\mathcal{T}$ and $\mathcal{T}'$ are back-translation from English-to-German-to-English and English-to-French-to-English (Zhang et al., 2021), respectively. Unlike SEED which uses single-view distillation, we use two augmented methods $(\mathcal{T}, \mathcal{T}')$ to achieve the *control* and *generalize* objectives.

Let f_θ^T and f_θ^S denote the teacher and student encoders, respectively. Sentence representations are extracted from the student model (S) for different augmented views: $\mathbf{z}_{\text{con}}^S = f_\theta^S(x_1)/||f_\theta^S(x_1)||$ and $\mathbf{z}_{\text{gen}}^S = f_\theta^S(x_2)/||f_\theta^S(x_2)||$. On the other hand, the teacher model (T) observes only one augmentation view, $\mathbf{z}_{\text{ref}}^T = f_\theta^T(x_1)/||f_\theta^T(x_1)||$.

3.2 How We Compare the Outputs

A simple method to assess the discrepancy between the teacher and student outputs is to directly compare the two vectors using a function such as L2 (Reimers and Gurevych, 2020) or cosine similarity (Sun et al., 2020a; Wu et al., 2021). In this work, however, we adopt a more robust alternative

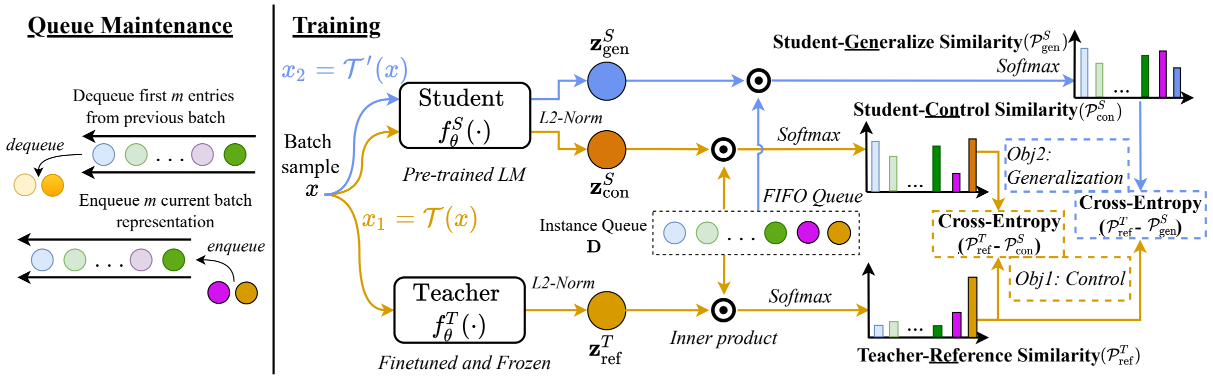

Queue Maintenance
Dequeue first m entries from previous batch
dequeue
Enqueue m current batch representation
enqueue
Training
$x_2 = \mathcal{T}'(x)$
Batch sample x
$x_1 = \mathcal{T}(x)$
Student $f_\theta^S(\cdot)$
Pre-trained LM
Teacher $f_\theta^T(\cdot)$
Finetuned and Frozen
L2-Norm
L2-Norm
$\mathbf{z}_{\text{gen}}^S$
$\mathbf{z}_{\text{con}}^S$
$\mathbf{z}_{\text{ref}}^T$
Instance Queue D
FIFO Queue
Inner product
Softmax
Softmax
Softmax
Student-Generalize Similarity($\mathcal{P}_{\text{gen}}^S$)
Student-Control Similarity($\mathcal{P}_{\text{con}}^S$)
Teacher-Reference Similarity($\mathcal{P}_{\text{ref}}^T$)
Obj2: Generalization
Cross-Entropy ($\mathcal{P}_{\text{ref}}^T$- $\mathcal{P}_{\text{con}}^S$)
Cross-Entropy ($\mathcal{P}_{\text{ref}}^T$- $\mathcal{P}_{\text{gen}}^S$)
Obj1: Control

Figure 2: Illustration of *Control and Generalization Distillation (ConGen)* training pipeline. For the teacher model, we freeze the weights during the distillation. We train student model by minimizing the cross-entropy of teacher & student similarity distributions computed over an instance queue.

which uses a large set of negative samples (Fang et al., 2021) to compare the outputs $\mathbf{z}_{\text{con}}^S$, $\mathbf{z}_{\text{gen}}^S$, and $\mathbf{z}_{\text{ref}}^T$. In particular, we represent the teacher and student outputs as *similarity distributions* computed from an *instance queue of negative samples* used in the loss calculations.

Instance Queue of Negative Samples. Since we use negative samples to describe the teacher and student outputs, we want them to provide sufficient coverage of the entire dataset. However, we want to keep the number of samples small due to the computational cost. Consequently, we adopt the instance queue approach (He et al., 2020) to achieve these goals. Let $\mathbf{D} = [d_1....d_K]$ denote the instance queue where d is a sentence representation obtained from the teacher, and K is the queue length. To cover the entire dataset, our framework progressively updates the instance queue $\mathbf{D}$ using the "first-in-first-out" (FIFO) strategy (He et al., 2020; Fang et al., 2021). At the beginning of each minibatch, we dequeue the first m entries where m is the minibatch size. We then enqueue the representation $\mathbf{z}_{\text{ref}}^T$ of each batch sample x bringing the total queue length back to K. The queue contains reference points for distillation and keeps rotating representations of the entire dataset for coverage. This practice reduces the overhead cost of computing negative samples. There are many ways to initialize $\mathbf{D}$, i.e., random vector (He et al., 2020; Fang et al., 2021). We found that random initialization by sampling from the training data produces acceptable results since we used a pre-trained LM as the student model.

Similarity Distribution. We use discrepancies between similarity distributions to help transfer the knowledge from the teacher to student models. Equation 1 describes how we compute the similarity distribution from a given representation $\mathbf{z}$ and an instance queue $\mathbf{D}$'s $(j = 1, ..., K)$.

$$\mathbf{p}\left(\mathbf{z}, \mathbf{D}, \tau\right) = \left[p_1 \ldots p_K\right], \quad \text{where } p_j = \frac{e^{\text{sim}(\mathbf{z},\mathbf{d}_j)/\tau}}{\sum_{\mathbf{d}\sim\mathbf{D}} e^{\text{sim}(\mathbf{z},\mathbf{d})/\tau}}, \tag{1}$$

where τ denotes the temperature parameter, and sim$(\cdot)$ denotes the cosine similarity between two feature vectors. As shown in Figure 2, we created three distributions: (i) Student-**Con**trol similarity distribution: $\mathcal{P}_{\text{con}}^S = \mathbf{p}(\mathbf{z}_{\text{con}}^S, \mathbf{D}, \tau^S)$; (ii) Student-**Gen**eralize similarity distribution: $\mathcal{P}_{\text{gen}}^S = \mathbf{p}(\mathbf{z}_{\text{gen}}^S, \mathbf{D}, \tau^S)$; and (iii) Teacher-**Ref**erence similarity distribution: $\mathcal{P}_{\text{ref}}^T = \mathbf{p}(\mathbf{z}_{\text{ref}}^T, \mathbf{D}, \tau^T)$. We found that using different temperature scaling for the teacher and student models (τ^T, τ^S) yields better performance than using the same value for both models (see Appendix A.2).

Distilling knowledge via the similarity distribution from the instance queue achieves three objectives: (i) the student learns to match the positive examples via the reference; (ii) the student learns to contrast the positive sample against a large number of negative samples efficiently via the instance queue; and (iii) the student model learns the difference between each negative samples within each distribution.

3.3 How We Train the Model

The training objective of ConGen facilitates the knowledge transfer from the teacher to student models (Figure 2). Specifically, we use the reference distribution to compute the control and generalization discrepancies.

ConGen transfers the knowledge using our novel loss function $\mathcal{L}_{\text{ConGen}}$:

$$\alpha \underbrace{\text{CE}(\mathcal{P}_{\text{ref}}^T, \mathcal{P}_{\text{con}}^S)}_{\text{control}} + (1-\alpha) \underbrace{\text{CE}(\mathcal{P}_{\text{ref}}^T, \mathcal{P}_{\text{gen}}^S)}_{\text{generalize}}, \tag{2}$$

where α represents the control-generalize trade-off and $\mathrm{CE}(\cdot)$ is the cross-entropy function computed between the teacher and student distributions.

The intuition behind the objectives are:

- *Obj 1: Control.* The first objective is to minimize the discrepancy between the student distribution $\mathcal{P}^S_{\text{con}}$ and the teacher distribution $\mathcal{P}^T_{\text{ref}}$ when the inputs are identical, i.e., both are compute from the same augmentation method $\mathcal{T}(\cdot)$.
- *Obj 2: Generalize.* The second objective is to improve the generalizability by minimizing the discrepancy between the student distribution $\mathcal{P}^S_{\text{gen}}$ and the teacher distribution $\mathcal{P}^T_{\text{ref}}$ when the inputs are slightly different, i.e., $\mathcal{T}(\cdot)$ and $\mathcal{T}'(\cdot)$.

With the two objectives, we are ensuring that the student behaves similarly to the teacher with the added robustness from the multiple views.

4 Experiment Setting

4.1 Pre-training

Teacher Model. By default, we use the current state-of-the-art unsupervised sentence representation, SimCSE-RoBERTa-large (Gao et al., 2021) (#parameters: 356M), as the teacher model. Note that our distillation framework is compatible with any teacher method and model. We also compare SimCSE to other recent unsupervised finetuning methods, i.e., BSL (Zhang et al., 2021) and ConSERT (Yan et al., 2021). For the teacher model, we also consider BERT-large and BERT-base. See Table 4.

Student Model. We experiment with multiple pretrained language models from small compressed models to large state-of-the-art models: BERT-Tiny, -Mini, and -Small (Turc et al., 2019), MiniLM (Wang et al., 2020), TinyBERT (Jiao et al., 2020), BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019). To produce sentence representations from these LMs, we use mean pooling (Reimers and Gurevych, 2019). In addition, we add one additional linear layer with the TanH activation function, where the number of hidden dimensions of the linear layer is equal to the teacher's.

Training Setup. For the training data, we use unlabeled texts from two NLI datasets, SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) datasets, to make it comparable with the previous works (Li et al., 2020; Zhang et al., 2020, 2021). We train the student model with the AdamW optimizer, a linear learning rate warm-up over 10% of the training data, and a batch size of 128 for 20 epochs. For hyperparameter settings, we use grid search to find the best parameter settings for the learning rate, teacher temperature (τ^T), student temperature (τ^S), and instance queue size (K). The full hyper-parameter configurations are given in Appendix A.1. Lastly, we randomly pick sentences from the training data to initialize the instance queue, which is more efficient than the random vector.

4.2 Competitive Methods

To show the effectiveness of our method, we compare our work to six competitors as follows.

Finetune-based. We use the state-of-the-art sentence representation, SimCSE (Gao et al., 2021), with *unsupervised* settings as the baseline results and *supervised* settings as an supervised baseline. Unsupervised settings are trained with contrastive loss and dropout as the data augmentation method, while supervised settings are trained with contrastive loss and NLI labeled datasets.

Distillation-based. We also compare our work with other distillation techniques:

- *L2*: A L2 minimization between the teacher and student representations ($\mathbf{z}^T_{\text{ref}}, \mathbf{z}^S_{\text{con}}$) (Romero et al., 2015).
- *Dual-L2*: Two terms L2 minimization where the first term is $\mathrm{L2}(\mathbf{z}^T_{\text{ref}}, \mathbf{z}^S_{\text{con}})$ and the second term is $\mathrm{L2}(\mathbf{z}^T_{\text{ref}}, \mathbf{z}^S_{\text{gen}})$ (Reimers and Gurevych, 2020).
- *SKD*: A self-knowledge distillation method that uses the same two terms L2 minimization as *Dual-L2*, with additional term $\mathrm{L2}(\mathbf{z}^S_{\text{con}}, \mathbf{z}^S_{\text{gen}})$ (Limkonchotiwat et al., 2022).
- *CKD*: An adaptation of contrastive knowledge distillation, where the positive and negative samples obtained from the teacher model (Wu et al., 2021). In this paper, however, we change from a supervised teacher to an unsupervised teacher.

We retrained all models with our training data.

4.3 Evaluation Setup

Semantic Textual Similarity (STS). Following previous works (Zhang et al., 2020, 2021; Yan et al., 2021; Gao et al., 2021), we report the average Spearman rank correlation on seven STS datasets include STS 2012-2016 (Agirre et al., 2012, 2013, 2014, 2015, 2016), STS-B (Cer et al., 2017), and SICK-R (Marelli et al., 2014).

Transfer and NLI. For transfer tasks, we report average scores across seven classification datasets such as MR (Pang and Lee, 2005), CR (Hu and Liu, 2004), SUBJ (Pang and Lee, 2004),

Methods	Semantic Textual Similarity (STS) average scores									
	BERT Tiny	**BERT Mini**	**Tiny BERT-L4**	**MiniLM L3**	**MiniLM L6**	**BERT Small**	**MiniLM L12**	**Tiny BERT-L6**	**BERT Base**	**RoBERTa Base**
#Param (M)	4	11	14	17	22	29	33	67	109	125
Finetune-based										
Teacher	*Unsup-SimCSE-RoBERTa-large: 78.90*† *(#parameters: 356M)*									
Sup-SimCSE$^{\clubsuit}$	*72.35*	*76.52*	*78.19*	*76.49*	*78.86*	*78.59*	*80.48*	*81.23*	*81.57*	*82.52*
Unsup-SimCSE	64.47	65.94	67.91	55.10	59.15	69.13	67.90	73.67	76.25	77.10
Distillation-based										
L2	73.32	76.07	77.03	76.66	77.51	77.30	78.79	78.95	78.97	79.00
Dual-L2	70.76	74.42	76.39	75.34	74.74	76.92	76.91	78.67	78.07	79.06
SKD	68.83	72.02	73.05	72.66	73.59	75.06	74.58	77.62	78.05	77.44
CKD	76.19	76.59	77.48	77.14	77.90	76.97	77.92	78.29	78.54	78.34
Our propose method										
ConGen	**76.85**	**78.09**	**78.54**	**78.22**	**79.10**	**78.91**	**79.68**	**79.73**	**80.06**	**79.78**

Table 1: Sentence embedding performance on STS tasks (Spearman rank correlation). The results of BERT-based, RoBERTa-base, and RoBERTa-large (the teacher model) are from SimCSE (Gao et al., 2021), † is the teacher performance without using any of the task data, and ♣ is the performance of supervised learning.

MPQA (Wiebe et al., 2005), SST (Socher et al., 2013), TREC (Voorhees and Tice, 2000), and MRPC (Dolan and Brockett, 2005) following Gao et al. (2021)'s experimental settings. For NLI tasks, we use two standard NLI datasets, such as SNLI and SICK-E datasets, and we use the same setting as the transfer benchmark.

Multilingual STS. We demonstrate the versatility of our approach on eight multilingual STS-2017 datasets (Cer et al., 2017) including EN-EN, AR-AR, ES-ES, EN-AR, EN-ES, EN-TR, EN-DE, and EN-FR following previous works (Reimers and Gurevych, 2020; Zhang et al., 2021). To extend our work to a multilingual setting, we changed one of the data augmentation operations ($\mathcal{T}'$) from back-translation to machine translation (Google NMT) from English to languages in multilingual STS-2017 following Zhang et al. (2021). For simplicity, we changed the student LMs from monolingual to Multilingual-DistilBert-cased (Sanh et al., 2019) and Multilingual-MiniLM-L12 (Wang et al., 2020), and we use the same teacher model (Unsupervised-SimCSE-RoBERTa-large). For competitors, we compare different unsupervised and multilingual settings with finetune-based, i.e., BSL, and distillation-based, i.e., Dual-L2 and L2. We show the average score from three random seeds for all experimental results.

5 Experimental Results

5.1 Semantic Textual Similarity

Table 1 shows the performance of our distilled models produced by our method compared to those of competitors on STS tasks. As mentioned in Section 1, finetuned-based methods do not perform well for small models. The experimental results demonstrate that distillation from a large model improves the performance of compressed models. For instance, using Unsupervised-SimCSE-RoBERTa-large as the teacher model of MiniLM-L3, the Spearman rank correction of ConGen-MiniLM-L3 can be improved from 55.10 to 78.22. Moreover, when the number of parameters is greater than 22 million, our models perform on par with the teacher model. In addition, when the number of parameters is less than 33 million, our models perform on par with Supervised-SimCSE. ConGen outperforms unsupervised methods in every compression model. For the full results of each dataset, see Appendix A.4.

5.2 Transfer and NLI

This study shows how our proposed models performed on transfer and NLI benchmarks. We continue to use the same baseline as the previous experiment without any modification. The setting of these tasks is described in Section 4.3.

As shown in Table 2, in the transfer learning task, our distillation models improved compression models' performance and were comparable to supervised-SimCSE. Compared to other distillation models, ConGen outperforms competitive models in five out of six models, except the CKD-BERT-base result. Furthermore, in the NLI task, the performance of unsupervised-BERT-Tiny is improved from 68.52 to 78.01 with our distillation method, which is slightly better than supervised-SimCSE and other distillation models. We hypothesize that using the similarity distribution is crucial

Methods	**Transfer Learning (TL) average scores**						**Natural Language Inference (NLI) average scores**					
	BERT Tiny	**Tiny BERT-L4**	**MiniLM L12**	**Tiny BERT-L6**	**BERT Base**	**RoBERTa Base**	**BERT Tiny**	**Tiny BERT-L4**	**MiniLM L12**	**Tiny BERT-L6**	**BERT Base**	**RoBERTa Base**
Finetune-based												
Teacher	*Unsup-SimCSE-RoBERTa-large: 86.81*†						*Unsup-SimCSE-RoBERTa-large: 77.01*†					
Sup-SimCSE♣	*75.30*	*83.01*	*83.86*	*85.55*	*86.98*	*88.85*	*75.91*	*78.64*	*78.86*	*80.63*	*81.84*	*82.00*
Unsup-SimCSE	77.00	82.14	83.38	85.49	85.17	85.10	68.52	74.30	74.25	75.87	76.17	76.51
Distillation-based												
L2	75.69	81.17	81.17	84.96	84.47	86.22	76.85	78.65	78.62	78.85	78.39	76.86
Dual-L2	75.21	80.98	80.98	83.36	83.27	84.46	74.59	77.61	77.65	77.98	78.38	77.84
SKD	75.14	76.98	76.98	82.82	83.21	83.51	75.69	77.25	77.81	79.39	79.14	79.27
CKD	78.62	83.11	83.11	84.87	**85.79**	86.30	77.72	78.56	79.06	79.44	78.95	79.20
Our propose method												
ConGen	**78.87**	**83.45**	**85.08**	**85.57**	85.64	**86.37**	**78.01**	**79.38**	**79.36**	**79.78**	**79.74**	**79.93**

Table 2: Sentence embedding performance on transfer and NLI tasks (accuracy score). Where all settings from SimCSE (Gao et al., 2021), † is the teacher performance, and ♣ is the performance of supervised learning.

for smaller models. For the full results of each dataset, see Table 10 and 11.

5.3 Multilingual STS

This study shows how well our method can be extended for multilingual sentence representation. To extend from monolingual to multilingual settings, we use multilingual compression models, i.e., Multilingual-MiniLM-L12 and Multilingual-DistilBERT-cased, as the student model. For the full setting's detail, see Section 4.3.

The results are illustrated in Table 3. The performance of our distillation method greatly outperforms the finetune-base results, BSL. In comparison to other distillation techniques, L2 has a critical issue in terms of performance, e.g., the Spearman rank correlation of L2-Multilingual-DistilBERT-based on EN-TR is only 10.73. Meanwhile, our method outperforms other distillation methods in all settings.

Model	**AR-AR**	**EN-AR**	**EN-DE**	**ES-ES**	**EN-ES**	**EN-EN**	**EN-FR**	**EN-TR**
Multilingual-MiniLM-L12 (#parameters: 117M)								
BSL	41.66	43.18	45.96	58.23	41.99	52.99	50.87	50.45
L2	60.25	39.50	59.92	74.74	52.14	83.61	57.08	33.27
Dual-L2	77.00	81.90	82.37	86.27	82.56	85.51	83.76	80.77
ConGen	**78.02**	**83.38**	**84.23**	**88.67**	**84.06**	**87.36**	**86.74**	**83.31**
Multilingual-DistilBERT-cased (#parameters: 134M)								
BSL	67.05	69.72	68.23	76.45	63.56	73.31	72.84	70.27
L2	56.41	27.16	53.59	77.10	41.51	83.76	46.48	10.73
Dual-L2	77.21	81.10	82.09	87.21	81.54	86.24	84.97	80.76
ConGen	**80.17**	**83.17**	**83.36**	**89.31**	**83.41**	**87.10**	**86.72**	**83.56**

Table 3: Sentence embedding performance on multilingual STS datasets. Where the teacher model for distillation methods, e.g., L2, Dual-L2, and ConGen is Unsup-SimCSE-RoBERTa-large.

5.4 Ablation Studies

This subsection explores the effect of various design decisions, such as teacher architectures, learning methods, loss functions, data augmentation techniques, anisotropy study, qualitative analysis, and instance queue.

Different Teacher Pretraining. This study shows the performance of our distillation in other teacher models and techniques. For the diversity of teacher models, we use BERT base and large versions trained on BSL (Zhang et al., 2021) and ConSERT (Yan et al., 2021), respectively. As shown in Table 4, our method works well regardless of the teacher model.

Models	**STS average scores**		
	ConSERT BERT-large	**BSL BERT-base**	**SimCSE RoBERTa-large**
Teacher	*76.45*	*73.85*	*78.90*
ConGen models			
BERT-Tiny	74.88	70.61	76.85
TinyBERT-L4	76.28	71.63	78.54
MiniLM-L12	76.93	72.45	79.68
TinyBERT-L6	76.87	**73.44**	79.73
BERT-base	**77.03**	72.87	**80.06**

Table 4: Sentence embedding performance on STS average scores. Where we changed from SimCSE to other finetuning algorithms such as ConSERT and BSL.

Loss function Study. In this study, we show the effectiveness of each objective in ConGen. Since our work is inspired by SEED (Fang et al., 2021) with an additional generalization loss term and data augmentation process designed for NLP tasks, we study how much gain the generalization term gives. In table 5, we show the Spearman rank correlation on average 7 STS corpus. We found that the performance of the original SEED (control only) is similar to our method ConGen when the number of the parameter is less than 14 million, i.e., BERT-Tiny and TinyBERT-L4. Nonetheless, when the number of parameters increases, the gap between SEED and ConGen widens, e.g., the gap between SEED and ConGen on TinyBERT-L4 is only 0.44;

in BERT-base, the gap increased to 0.77. In addition, using only the generalize term (0%) performs slightly better than the original SEED (100%); still, combining the two learning objectives (50%, ConGen) yields the best performance. We investigated more about control and generalization objectives in error analysis.

Model	**alpha's weight (α)**				
	100% (Control only)	**75%**	**50% (ConGen)**	**25%**	**0% (Generalize only)**
BERT-Tiny	76.25	76.72	**76.85**	76.62	76.26
TinyBERT-L4	78.10	78.45	**78.54**	78.43	78.30
MiniLM-L12	79.36	79.56	**79.68**	79.52	79.37
BERT-base	79.29	80.03	**80.06**	80.02	79.72

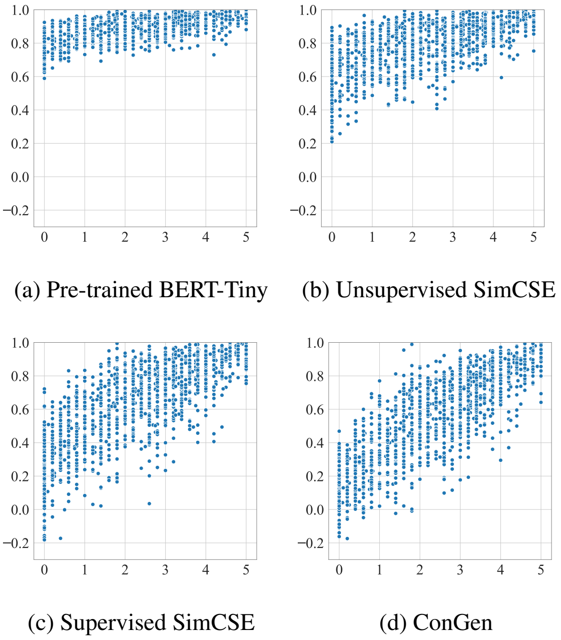
Table 5: Sentence embedding performance on STS average scores. Where we change the portion of α into 0%, 25%, 50% (baseline), 75%, and 100%.

Effect of Data Augmentation Choice. We evaluate the effectiveness of different data augmentation methods for the generalization objective. For simplicity, we use data augmentation methods from Gao et al. (2021). The experimental results show in Table 6. The results showed that using Google NMT for data augmentation yields the best results conforming to previous works (Zhang et al., 2021; Fang et al., 2020). On the other hand, we can also use word deletion and delete one word when Google NMT is unavailable, which is not much different in the larger model.

Model	**STS average scores**	
	BERT-Tiny	**BERT-base**
Baseline		
EN→DE→EN (Google NMT)	**76.85**	**80.06**
Other augmentation methods		
EN→DE→EN (MBart)	71.35	75.37
MLM 15%	74.99	78.44
Synonym replacement	76.01	80.01
Crop 10%	76.14	79.95
Word deletion 10%	76.15	**80.06**
Delete one word	76.14	80.02

Table 6: Comparison between data augmentation operations for the generalize objective.

Anisotropy Study and Qualitative Analysis The anisotropic property of contextualized representations derived from pre-trained BERT is studied in several works (Ethayarajh, 2019; Li et al., 2020). This phenomenon consequently leads to degradation in semantic retrieval performance (Wang and Isola, 2020). Figure 3 shows the correlation between ground-truth similarity scores and model-derived cosine similarity. ConGen shows a better correlation between the gold standard and cosine similarities from sentence pair representations on STS-B than unsupervised-SimCSE and similar to supervised-SimCSE. This result confirms that our work can significantly decrease the unsupervised and supervised learning gap.



(a) Pre-trained BERT-Tiny

(b) Unsupervised SimCSE

(c) Supervised SimCSE

(d) ConGen

Figure 3: Scatter plot of the groundtruth similarity scores (x-axis) and the cosine similarities (y-axis) between sentence pairs in the STS-B (dev set).

We also performed qualitative comparisons following Gao et al. (2021). Using the 150,000 Flickr30k captions dataset, we randomly selected sentences from the dataset to retrieve similar sentences using embeddings from SimCSE, ConGen (only Obj 1), and ConGen (both objectives). Unsupervised simCSE failed to yield good top-5 retrieval results, while ConGen (only Obj 1) started to do badly at the top-10. However, ConGen (two objectives) yielded the most robust results. Example retrievals are available in Table 12.

Instance Queue Study. This study shows the effect of instance queue size on LMs. For simplicity, we select two popular models, BERT-base and RoBERTa-base. For the instance queue sizes, we set the size as follows: 128, 1024, 16384, and 65536. As shown in Figure 4, the instance queue size has affected the performance of LMs. Unlike previous works that also used instance queue (Fang et al., 2021; Wang et al., 2021b), we found that the best instance queue is not always 65536 sentences. The best size for RoBERTa-base is only

1024 sentences. These results show the importance to finetune the instance queue size for distillation since each model has a different instance queue size. However, the difference is relatively small. For the best queue size of each model, see Appendix A.1.

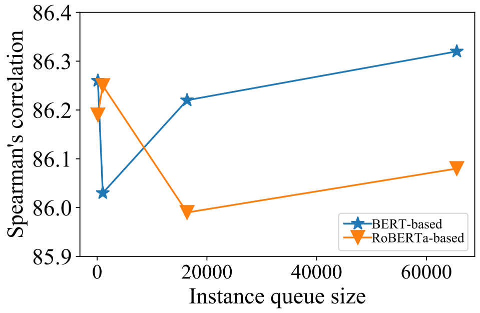


Figure 4: Effect of queue size in BERT-base and RoBERTa-base.

6 Conclusion

In this paper, we propose a novel unsupervised *Control and Generalization Distillation (ConGen)*. ConGen is a distillation framework that transfers knowledge from a large model to any model regardless of its architecture and size by exploiting the concept of control and generalization mechanism. Our method outperforms competitive methods in all cases in monolingual and multilingual STS and five out of six text classification benchmarks. Furthermore, we demonstrate that our distillation framework can reduce the gap between compressed and base LMs. Using ConGen, the performance differences between supervised and unsupervised methods are slim for smaller models.

7 Limitation

Out-of-domain data might pose certain difficulties to our method. We strongly advise against using our model with out-of-domain data i.e., health or legal texts, directly. For example, we measure the cosine similarity between *"The risks and benefits of the procedure were discussed, and the patient consented to this procedure"* and *"The content of this note has been reproduced, signed by an authorized physician in the space above, and mailed to the patient's parents, the patient's home care company."*, the result from ConGen-BERT-Base is 0.3 (indicate that the two sentences are not equivalent, but share some details or are on the same topic) while the answer similarity is 0 (the two sentences are completely dissimilar). Both texts are from the MedSTS corpus (Peng et al., 2019), which is considered out-of-domain. To tackle this problem, we advise detecting out-of-domain samples or incorporating techniques that can help handle out-of-domain samples (Limkonchotiwat et al., 2020, 2021; Trijakwanich et al., 2021; Wang et al., 2021a). In addition, we did not try our method on non-MLM families such as GPT, BART, or CLIP.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *SemEval 2015*.

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 *SemEval 2014*.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016*.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*.

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*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP 2015*.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2021. Semantic re-tuning with contrastive tension. In *ICLR 2021*.

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)*.

Xingyi Cheng. 2021. *Dual-View Distilled BERT for Sentence Embedding*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019*.

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

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *EMNLP-IJCNLP 2019*.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding.

Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. 2021. SEED: self-supervised distillation for visual representation. In *ICLR 2021*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *EMNLP 2021*.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In *ACL/IJCNLP 2021*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR 2020*.

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*, KDD '04.

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*.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-guided contrastive learning for BERT sentence representations. In *ACL/IJCNLP 2021*.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *EMNLP 2020*.

Peerat Limkonchotiwat, Wannaphong Phatthiyaphaibun, Raheem Sarwar, Ekapol Chuangsuwanich, and Sarana Nutanong. 2020. Domain adaptation of Thai word segmentation models using stacked ensemble. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3841–3847, Online. Association for Computational Linguistics.

Peerat Limkonchotiwat, Wannaphong Phatthiyaphaibun, Raheem Sarwar, Ekapol Chuangsuwanich, and Sarana Nutanong. 2021. Handling cross- and out-of-domain samples in Thai word segmentation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1003–1016, Online. Association for Computational Linguistics.

Peerat Limkonchotiwat, Wuttikorn Ponwitayarat, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. 2022. CL-ReLKT: Cross-lingual language knowledge transfer for multilingual retrieval question answering. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2141–2155, Seattle, United States. Association for Computational Linguistics.

Che Liu, Rui Wang, Jinghua Liu, Jian Sun, Fei Huang, and Luo Si. 2021a. DialogueCSE: Dialogue-based contrastive learning of sentence embeddings. In *EMNLP 2021*.

Fangyu Liu, Yunlong Jiao, Jordan Massiah, Emine Yilmaz, and Serhii Havrylov. 2022. Trans-encoder: Unsupervised sentence-pair modelling through self- and mutual-distillations. In *ICLR 2022*.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021b. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In *EMNLP, 2021*.

Qun Liu and Supratik Mukhopadhyay. 2018. Unsupervised learning using pretrained cnn and associative memory bank. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 01–08.

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. *CoRR*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A sick cure for the evaluation of compositional distributional semantic models. In *LREC'14*.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *ACL 2004*.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL 2005)*.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, pages 58–65.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP 2019*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *EMNLP 2020*.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. Fitnets: Hints for thin deep nets. In *ICLR 2015*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP 2013*.

Siqi Sun, Zhe Gan, Yuwei Fang, Yu Cheng, Shuohang Wang, and Jingjing Liu. 2020a. Contrastive distillation on intermediate representations for language model compression. In *EMNLP 2020*.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020b. Mobilebert: a compact task-agnostic BERT for resource-limited devices. In *ACL 2020*.

Nattapol Trijakwanich, Peerat Limkonchotiwat, Raheem Sarwar, Wannaphong Phatthiyaphaibun, Ekapol Chuangsuwanich, and Sarana Nutanong. 2021. Robust fragment-based framework for cross-lingual sentence retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 935–944, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*.

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*, SIGIR '00.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021a. TSDAE: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Liang Wang, Wei Zhao, and Jingming Liu. 2021b. Aligning cross-lingual sentence representations with dual momentum contrast. In *EMNLP 2021*.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML 2020*.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021c. Minilmv2: Multi-head self-attention relation distillation for compressing pre-trained transformers. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *NeurIPS 2020*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Lang. Resour. Evaluation*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL 2018*.

Xing Wu, Chaochen Gao, Jue Wang, Liangjun Zang, Zhongyuan Wang, and Songlin Hu. 2021. Disco: Effective knowledge distillation for contrastive learning of sentence embeddings.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *ACL, 2021*.

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: System Demonstrations*.

Ziyi Yang, Yinfei Yang, Daniel Cer, Jax Law, and Eric Darve. 2021. Universal sentence representation learning with conditional masked language model. In *EMNLP 2021*.

Yan Zhang, Ruidan He, Zuozhu Liu, Lidong Bing, and Haizhou Li. 2021. Bootstrapped unsupervised sentence representation learning. In *ACL/IJCNLP 2021*.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. In *EMNLP 2020*.

A Appendix

A.1 Parameter Settings

To align between the teacher and student similarity distribution, we use grid search on teacher's temperature (τ^T), student's temperature (τ^S), queue size (K), and learning rate (LR). In addition, we evaluate the Spearman rank correlation of STS-B development set in every 512 training step to save the best model. The best parameter is shown in Table 7. We obtained the best parameter of each model by evaluating on STS-dev set.

Model	τ^T	τ^S	K	**LR**
BERT-Tiny	0.05	0.05	16384	$5e^{-4}$
BERT-Mini	0.05	0.07	16384	$3e^{-4}$
Tiny-BERT-L4	0.05	0.05	65536	$1e^{-4}$
MiniLM-L3	0.05	0.07	16384	$5e^{-4}$
MiniLM-L6	0.05	0.07	65536	$3e^{-4}$
BERT-Small	0.05	0.07	65536	$3e^{-4}$
MiniLM-L12	0.05	0.07	16384	$5e^{-5}$
Tiny-BERT-L6	0.05	0.07	65536	$5e^{-5}$
BERT-base	0.05	0.07	65536	$5e^{-5}$
RoBERTa-base	0.1	0.1	1024	$5e^{-5}$
Multilingual-DistilBERT	0.05	0.07	65536	$3e^{-4}$
Multilingual-MiniLM-L12	0.05	0.07	65536	$3e^{-4}$

Table 7: The best parameter of each model in this paper.

A.2 Temperature Scaling

In this study, we show the effect of different temperature settings. As shown in Table 8, the best temperatures for the teacher and student models are different. The results show that, in BERT-Tiny, the best scaling value is 0.05, while the best scaling of BERT-base is 0.07. However, setting both to 0.05 does not have a large effect on performance.

A.3 Effect of Lexical Change on Sentence Pair Similarity.

We compute edit distance at word-level on sentence pairs from STS-B (dev set), Similar to a study of lexical and semantic similarity conducted by (Li et al., 2020). The correlation between model-derived similarity scores and edit distance shows in Figure 5. The results show that the correlation between edit distance and cosine similarity for ConGen ($\rho = -34.73$) is weaker than unsupervised-SimCSE ($\rho = -46.73$) and is similar to that of supervised-SimCSE ($\rho = -35.43$).

A.4 Full results

We show the full results of our work on STS, transfer, and NLI in Table 9, 10, and 11, respectively.

τ^T	τ^S	**STS-B dev set**	
		BERT-Tiny	**BERT-base**
0.1	0.1	83.69	86.02
	0.09	83.71	85.37
	0.07	82.15	83.85
	0.05	79.75	83.58
0.05	0.1	84.02	86.17
	0.09	84.14	86.25
	0.07	84.18	**86.31**
	0.05	**84.83**	85.96
0.01	0.1	84.08	86.22
	0.09	84.18	86.23
	0.07	84.16	86.09
	0.05	83.79	85.19

Table 8: Effect of temperature for the distillation on BERT-Tiny and BERT-base. We evaluated Spearman rank correlation on STS-B dev set.

Model	STS-12	STS-13	STS-14	STS-15	STS-16	STS-B	SICK-R	Avg.
BERT-Tiny	72.18	81.12	75.45	83.22	77.89	79.03 69.05	76.85	
BERT-Mini	74.17	82.69	76.58	84.30	78.23	80.84	69.82	78.09
Tiny-BERT-L4	74.3	83.07	77.37	84.70	79.06	80.99	70.26	78.54
MiniLM-L3	74.00	82.93	76.58	84.35	78.57	81.00	70.09	78.22
MiniLM-L6	75.06	83.86	77.29	85.01	79.67	81.92	70.89	79.10
BERT-Small	74.50	83.58	77.29	84.83	79.72	81.93	70.55	78.91
MiniLM-L12	75.25	84.61	78.27	85.51	80.52	82.32	71.32	79.68
Tiny-BERT-L6	75.53	84.76	78.33	85.72	80.42	82.25	71.12	79.73
BERT-base	75.58	85.13	78.54	85.75	81.12	82.81	71.47	80.06
RoBERTa-base	75.32	84.56	77.26	85.33	81.34	82.67	72.00	79.78

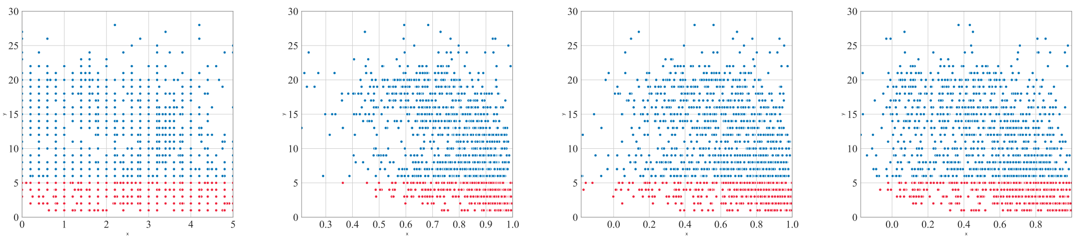
Table 9: The full results of our work on STS tasks (Spearman rank correlation).

Model	MR	CR	MPQA	SUBJ	SST2	TREC	MRPC	avg
BERT-Tiny	72.81	78.76	87.22	87.77	77.30	74.67	73.60	78.87
BERT-Mini	75.70	83.62	88.27	90.06	81.51	83.73	74.11	82.43
Tiny-BERT-L4	76.82	85.18	88.43	90.82	81.82	86.60	74.51	83.45
MiniLM-L3	76.18	83.60	88.47	90.82	82.32	86.13	73.24	82.97
MiniLM-L6	77.99	85.42	88.97	91.66	83.87	86.67	74.90	84.21
BERT-Small	78.09	84.88	89.01	91.89	83.87	87.20	74.55	84.21
MiniLM-L12	79.24	85.84	89.16	92.40	85.52	87.80	75.57	85.08
Tiny-BERT-L6	79.87	86.48	89.28	93.32	85.41	89.13	75.40	85.57
BERT-base	80.44	86.76	89.36	93.46	85.36	88.40	75.67	85.64
RoBERTa-base	81.83	87.74	88.43	93.47	87.11	90.40	75.63	86.37

Table 10: The full results of our work on transfer tasks (Accuracy).

Model	SICKEntailment	SNLI	avg
BERT-Tiny	82.92	73.1	78.01
BERT-Mini	82.57	74.87	78.72
Tiny-BERT-L4	83.73	75.11	79.42
MiniLM-L3	83.19	74.99	79.09
MiniLM-L6	83.18	75.52	79.35
BERT-Small	83.23	75.53	79.38
MiniLM-L12	83.29	75.43	79.36
Tiny-BERT-L6	83.22	76.34	79.78
BERT-base	83.26	76.23	79.74
RoBERTa-base	83.7	76.17	79.93

Table 11: The full results of our work on NLI tasks (Accuracy).



(a) Groundtruth

(b) Unsupervised SimCSE

(c) Supervised SimCSEE

(d) ConGen

Figure 5: Scatter plot of edit distance and sentence similarity. The vertical axis represents the edit distance between sentence pairs in the STS-B (dev set). The horizontal axis represents the groundtruth similarity for (a), and model-derived cosine similarities for (b – d). Red dots indicate sentence pairs with edit distance ≤ 5.

Retrieve @K	**Unsupervised SimCSE-BERT-Tiny**	**ConGen-BERT-Tiny (Only Obj 1)**	**ConGen-BERT-Tiny**
Query: A group of men playing music for people walking down the street .			
K = 1	Many people are watching two men performing in a busy street .	a group of musicians playing music in the street .	a group of musicians playing music in the street .
K = 5	Two women are walking together on a public street .	A group of men walk down a city street .	A man walking on the street listening to music .
K = 10	Two teenage boys are walking together down the sidewalk in a city .	A man walking on the street listening to music .	This men are marching band playing instruments on streets .
Query: Two guys - one with a basketball in his hand - are jumping .			
K = 1	Four guys look up at something underneath a basketball hoop .	two men jump for the basketball .	two men jump for the basketball .
K = 5	Two men are playing basketball , and one man is making a shot .	There are two boys on a basketball court and one is holding a basketball .	A man jumps with a basketball to make a hoop .
K = 10	There are several dogs in the field where the girl with the ponytail is sitting down .	Professional basketball players watching as one man jumps in the air with the basketball in one hand .	Two men jump up for the ball near the hoop in a basketball game .
Query: A brown dog and a gray dog are playing in the snow .			
K = 1	A black dog and a white dog are standing on snow .	A brown dog and a gray dog play in the grass .	A brown dog and a gray dog play in the grass .
K = 5	The boy and a brown dog are standing in the snow .	Two brown dogs playing in snow	A grayish dog running in the snow .
K = 10	A man in a winter hat plays with a small black dog in the snow .	A grayish dog running in the snow .	A black and gray dog
Query: Two young men in uniform involved in a basketball game .			
K = 1	Three men in basketball uniforms looking at something .	Two young males playing basketball on a basketball court .	Young men playing basketball in a competition .
K = 5	Two black men playing basketball in a gym .	Two young men compete in an informal basketball game .	Some young men are playing basketball .
K = 10	Two men in the military are playing baseball .	Two young men in basketball clothes walking together .	Two women in basketball uniforms playing basketball on a court .
Query: A man is using a pro-saw to fix the street .			
K = 1	An older man with a can is standing next to a counter .	Men working on street repair .	Men working on street repair .
K = 5	A young person in a hat is coming through an entrance labeled " no fumar " (no smoking) .	A man helping a biker fix his bike on a road .	A man helping a biker fix his bike on a road .
K = 10	A man in a tie on a street .	A man performing a bicycle repair on the side of the road .	Workman are completing repairs on a city street .

Table 12: Qualitative study on the Flick30kr dataset (150k sentences), we randomly selected one caption as the query and retrieved top-1,5, and 10 highest similarities.