Supervised Contrastive Learning for Cross-lingual Transfer Learning

Shuaibo Wang[1], **Hui Di**[2], **Hui Huang**[3], **Siyu Lai**[1]
Kazushige Ouchi[2], **Yufeng Chen**[1]*, **Jinan Xu**[1]
[1] School of Computer and Information Technology, Beijing Jiaotong University,
Beijing 100044, China
[2] Toshiba (China) Co., Ltd. [3] Harbin Institute of Technology
{wangshuaibo,20120374,chenyf,jaxu}@bjtu.edu.cn;
dihui@toshiba.com.cn; huanghui_hit@126.com;
kazushige.ouchi@toshiba.co.jp

Abstract

Multilingual pre-trained representations are not well-aligned by nature, which harms their performance on cross-lingual tasks. Previous methods propose to post-align the multilingual pre-trained representations by multi-view alignment or contrastive learning. However, we argue that both methods are not suitable for the cross-lingual classification objective, and in this paper we propose a simple yet effective method to better align the pre-trained representations. On the basis of cross-lingual data augmentations, we make a minor modification to the canonical contrastive loss, to remove false-negative examples which should not be contrasted. Augmentations with the same class are brought close to the anchor sample, and augmentations with different class are pushed apart. Experiment results on three cross-lingual tasks from XTREME benchmark show our method could improve the transfer performance by a large margin with no additional resource needed. We also provide in-detail analysis and comparison between different post-alignment strategies.

1 Introduction

Cross-lingual transfer learning aims to transfer the learned knowledge from a resource-rich language to a resource-lean language. The main idea of crosss-lingual transfer is to learn a shared language-invariant feature space for both languages, so that a model trained on the source language could be applied to the target language directly. Such generalization ability greatly reduces the required annotation efforts, and has urgent demand in real-world applications.

Recent multilingual pre-trained models, such as XLM-RoBERTa(XLM-R) (Conneau et al., 2020), have been demonstrated surprisingly effective in the cross-lingual scenario. By fine-tuning on labeled data in a source language, such models can generalize to other target languages even without any additional training. This has become a de-facto paradigm for cross-lingual language understanding tasks.

Despite their success in cross-lingual transfer tasks, multilingual pre-training commonly lacks explicit cross-lingual supervision, and the representations for different languages are not inherently aligned. To further improve the transferability of multilingual pre-trained representations, previous works propose different methods for cross-lingual alignment. Zheng et al. (2021) and Lai et al. (2021) propose to augment the training set with different views, and align the pre-trained representations of different languages by dragging two views closer. However, simply bringing different views closer would easily lead to representation collapse and performance degradation (Tao et al., 2021). Meanwhile, Pan et al. (2021) and Wei et al. (2021) propose to incorporate additional parallel data, and align the pre-trained representations by contrasting positive and negative samples. However, monotonously treating all random samples equally negative is inconsistent with the classification objective.

In this work, we propose a simple yet effective method to better post-align the multilingual representations on downstream tasks, which can both avoid representation collapse and meanwhile induce classification bias. With only training data for the source language available, our method performs cross-lingual fine-tuning by two steps. 1) Firstly, the original training data is augmented with different views,

* Corresponding author.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 884-895, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

including code-switching, full-translation and partial-translation. All views could provide cross-lingual supervision for post-alignment. 2) Given one training sentence as the anchor point, the corresponding augmented view serves as the positive sample, and other augmented views with different labels serve as the negative samples, contrastive learning is performed by pulling positive samples together and pushing apart negative samples. This is called Supervised Contrastive Learning (SCL), and can be deemed as a cross-lingual regularizer to be combined with conventional fine-tuning.

We perform experiments on two cross-lingual classification tasks, namely XNLI (cross-lingual inference) and PAWS-X (cross-lingual paraphrase identification) (Conneau et al., 2018; Yang et al., 2019a). We compare different alignment methods, and our proposed method outperforms previous methods by a large margin, proving its effectiveness. Besides, we also apply our method on the cross-lingual retrieval task of BUCC[0] and tatoeba (Artetxe and Schwenk, 2019). We use the data from PAWS-X as supervision, and fine-tune the pretrained model by contrasting samples with their machine translation. Our proposed method again outperforms other methods by a large margin.

Detailed analysis and discussion are provided to compare different post-alignment methods for pre-trained representations, and to prove the necessity of label-supervision when performing cross-lingual contrastive learning.

2 Background

2.1 Contrastive Learning

Contrastive learning aims at maximizing the similarity between the encoded query q and its matched positive samples k^+ while keeping randomly sampled keys $\{k_0, k_1, k_2, ...\}$ far away from it. With similarity measured by a score function $s(q, k)$, InfoNCE (van den Oord et al., 2018) loss is commonly used to this end:

$$L_{ctl} = \frac{\exp(s(q, k^+))}{\exp(s(q, k^+)) + \sum_{i=1}^{n} \exp(s(q, k_i^-))}$$

Contrastive learning has led to significant improvements in various domains (He et al., 2020; Gao et al., 2021). Recently, Khosla et al. (2020) propose to incorporate label-supervision to the fine-tuning of pre-trained models, and obtain improvement on multiple datasets of the GLUE benchmark, and our work is inspired by them. However, their method is only targeted at monolingual tasks.

2.2 Cross-lingual Transfer

Cross-lingual transfer learning aims to transfer the learned knowledge from a resource-rich language to a resource-lean language. Despite recent success in large-scale language models, how to adapt models trained in high-resource languages (e.g., English) to low-resource ones still remains challenging. Several benchmarks are proposed to facilitate the progress of cross-lingual transfer learning (Hu et al., 2020; Liang et al., 2020), where models are fine-tuned on English training set and directly evaluated on other languages.

Recently, several pre-trained multilingual language models are proposed for cross-lingual transfer, including multilingual BERT (Devlin et al., 2019), XLM (Lample and Conneau, 2019), and XLM-R (Conneau et al., 2020). The models work by pre-training multilingual representations using some form of language modeling, and have made outstanding progress in cross-lingual tasks. However, most existing models use only single-language input for language model finetuning, without leveraging the intrinsic cross-lingual alignment. Therefore, several methods have been proposed to post-align the pre-trained representations, by introducing some form of cross-lingual supervision. Cao et al. (2020) and Dou et al. (2021) propose to generate word alignment information from parallel data, and push the aligned words in parallel data to have similar representations. Pan et al. (2021), Wang et al. (2021) and Wei et al. (2021) propose to utilize contrastive learning for post-alignment by contrasting positive and negative samples, where positive samples are parallel to each other while negative samples are randomly picked.

[0]https://comparable.limsi.fr/bucc2017/

Proceedings of the 21st China National Conference on Computational Linguistics, pages 884-895, Nanchang, China, October 14 - 16, 2022.

Zheng et al. (2021) and Lai et al. (2021) propose to augment the training set with different views, and align the representations by dragging two views close to each other. In a nutshell, despite all variations of supervision in both sentence or word-level, from both parallel data or automatically crafted data, the alignment must be performed by inter-lingual comparing, either by bringing two representations closer or contrasting a representation with random sampled representations. However, we argue that both methods are in contradiction with the cross-lingual classification objective, for which we will give detailed analysis in Section 3.2.

3 Approach

In this section, we first introduce the three cross-lingual data augmentation methods. Based on that, we propose three paradigms to post-align the multilingual representations, and provide theoretical analysis and comparison for them.

3.1 Cross-lingual Data Augmentation

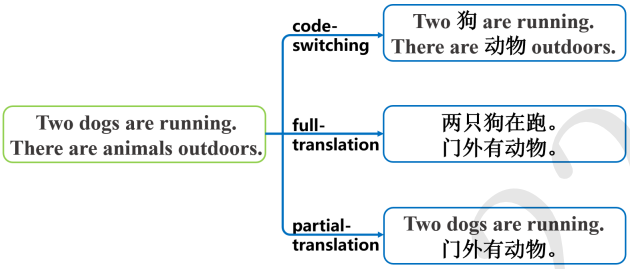

Figure 1: Different cross-lingual data-augmentation methods. Here we use sentence-pair classification as an example, therefore each sample contains two sentences.

In this work, we do not want to incorporate any parallel data (which is inaccessible in a lot of scenarios, especially for a resource-lean language that we want to transfer to). Therefore, to provide cross-lingual supervision for post-alignment, we propose three data augmentation methods:

1. Code-switching: Following Qin et al. (2020), we randomly select words in the original text in the source language and replace them with target language words in the bilingual dictionaries, to generate code-switched data. The intuition is to help the model automatically and implicitly align the replaced word vectors in the source and all target languages by mixing their context information, and the switched words can serve as anchor point for aligning two representation space.

2. Full-translation: Machine translation has been proved to be an effective data augmentation strategy under the cross-lingual scenario. It can provide translations almost in-line-with human performance, and therefore serves as a strong baseline for cross-lingual tasks.

3. Partial-translation: This method simply takes a portion of input and replace it with its translation in another language. According to Singh et al. (2019), partial-translation could provide inter-language information, where the non-translated portion serves as the anchor point. This is somehow akin to code-switching, and can be deemed as code-switching in segment-level.

The three methods can provide cross-lingual supervision in a coarse-to-fine manner (sentence-level, segment-level, word-level). We perform all the three methods to the whole training set. Each training sample could be code-switched multiple times with different results, and each task contains translation into multiple languages, leading to multiple views from a cross-lingual perspective.

3.2 Cross-lingual Alignment: What do we want?

Many experiments (Cao et al., 2020; Kulshreshtha et al., 2020) suggest that to achieve reasonable performance in the cross-lingual setup, the source and the target languages need to share similar representations. However, current multilingual pre-trained models are commonly pre-trained without explicit cross-lingual supervision. Therefore, the cross-lingual transfer performance can be further improved by additional cross-lingual alignment.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 884-895, Nanchang, China, October 14 - 16, 2022.

Given the training sample in source language and its cross-lingual augmentations, previous methods perform cross-lingual alignment in two different trends: Multi-view Alignment (Zheng et al., 2021; Lai et al., 2021) or Contrastive Learning (Wei et al., 2021; Pan et al., 2021). The multi-view alignment is to bring the sample and the corresponding augmentation together, while the contrastive learning is to bring these two together while pushing apart other random sampled augmentations. Suppose we are working with a batch of training examples of size N, $\{x_i, y_i\}, i = 1, ...N$, x_i denotes the training sample, while y_i is the label, the two different objectives can be denoted as follows:

$$L_{MVA} = -s(\Phi(x_i), \Phi(\hat{x}_i))$$

$$L_{CL} = -\log \frac{s(\Phi(x_i), \Phi(\hat{x}_i))}{s(\Phi(x_i), \Phi(\hat{x}_i)) + \sum_{j=1}^{N} \mathbb{I}_{j \neq i} s(\Phi(x_i), \Phi(\hat{x}_j))}$$

where $\Phi(\cdot) \in R_d$ denotes the $L2$-normalized embedding of the final encoder hidden layer before the softmax projection, and $\hat{x}_i$ denotes the augmented view (code-switching, full-translation, partial-translation, etc.), and s(q, k) denotes the similarity measure (cosine similarity, KL divergence, etc.). MVA is short for multi-view alignment, and CL is short for contrastive learning.

Since in vanilla contrastive learning, the similarity function is normally in the form of exponential, therefore L_{CL} can be detached into two terms:

$$L_{CL} = \underbrace{-s(\Phi(x_i), \Phi(\hat{x}_i))}_{alignment} + \underbrace{\log(e^{s(\Phi(x_i), \Phi(\hat{x}_i))} + e^{\sum_{j=1}^{N} \mathbb{I}_{j \neq i} s(\Phi(x_i), \Phi(\hat{x}_j))})}_{uniformity}$$

where the first term optimize the alignment of representation space, and the second term optimize the uniformity, as discussed in Wang et al. (2020). According to Gao et al. (2021), let W be the sentence embedding matrix corresponding to x_i, i.e., the i-th row of W is $\Phi(x_i)$, optimizing the $uniformity$ term essentially minimizes an upper bound of the summation of all elements in $WW^\top$, and inherently "flatten" the singular spectrum of the embedding space.

However, the $uniformity$ term in L_{CL} is in contradiction with the classification objective. In classification task, we want the representations to be clustered in several bunches, each bunch corresponds to a class. Or else to say, we want the representations to be inductively biased, rather than uniformly distributed.

On the other hand, it is obvious that the multi-view alignment objective L_{MVA} is to solely maximize the alignment. This would easily lead to representation collapse, since simply projecting all representations to one data point could easily reduce the $alignment$ term to zero. Contrast between samples is necessary to avoid collapse, and simply removing the $uniformity$ term is also not what we want.

3.3 Better Alignment with SCL

To better perform cross-lingual alignment, we propose to introduce label information to the vanilla contrastive learning, named as Supervised Contrastive Learning (SCL):

$$L_{SCL} = -\log \frac{s(\Phi(x_i), \Phi(\hat{x}_i))}{s(\Phi(x_i), \Phi(\hat{x}_i)) + \sum_{j=1}^{N} \mathbb{I}_{y_j \neq y_i} s(\Phi(x_i), \Phi(\hat{x}_j))}$$

More concretely, our modification is based on InfoNCE loss (van den Oord et al., 2018), therefore the similarity function is written as:

$$s(\Phi(x_i), \Phi(\hat{x}_i)) = e^{cos(\Phi(x_i), \Phi(\hat{x}_i))/\tau}$$

where $\tau > 0$ is an adjustable scalar temperature parameter that controls the separation of classes. Empirical observations show that both $L2$-normalization of the encoded embedding representations (which is incorporated in the calculation of cosine similarity) and an adjustable scalar temperature parameter τ

Proceedings of the 21st China National Conference on Computational Linguistics, pages 884-895, Nanchang, China, October 14 - 16, 2022.

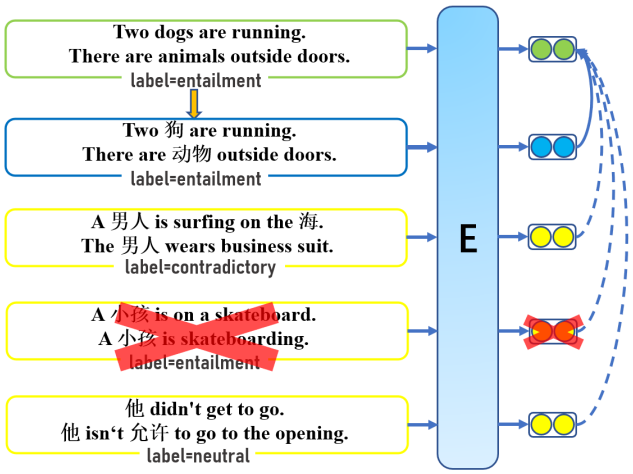

Figure 2: Our proposed supervised contrastive learning. Solid line connects positive pairs while dashed line connects negative pairs. Notice the false negative sample is removed.

improve performance. This can serve as a cross-lingual regularization term and be combined with the canonical classification loss:

$$L_{CE} = y_i \cdot \log(1 - \hat{y}_i) + \hat{y}_i \cdot \log(1 - y_i)$$

$$L_{total} = L_{CE} + \lambda L_{SCL}$$

where λ is a scalar weighting hyperparameter that we tune for each downstream task.

The core idea is simple, just to remove the negative samples which belong to the same class with the anchor point. Therefore, only samples from different classes would be pulled apart. The modified $uniformity$ term is not to unify the representations any more, but to push the multilingual decision clusters apart from each other.

This loss can be applied to a variety of encoders, not just limited to multilingual pre-trained transformer-like models. The loss is meant to capture similarities between examples of the same class and contrast them with examples from other classes. This is in line with the objective of cross-lingual alignment. When we are doing cross-lingual alignment, what we really want to do is to transfer the representation for a certain class to another language, rather than to learn a unified multilingual representation space.

4 Experiments

4.1 Data Preparation

In this work, we mainly focus on sentence-level tasks, for which the aggregated representation is easily accessible. We conduct experiments on two cross-lingual sentence-pair classification tasks: natural language inference and paraphrase identification. The Cross-lingual Natural Language Inference corpus (XNLI) (Conneau et al., 2018) asks whether a premise sentence entails, contradicts, or is neutral toward a hypothesis sentence. The Cross-lingual Paraphrase Adversaries from Word Scrambling (PAWS-X) (Yang et al., 2019a) dataset requires to determine whether two sentences are paraphrases. Both tasks are from XTREME benchmark (Hu et al., 2020). Despite their intrinsic different objective, both tasks can be formalized as sentence-pair classification tasks. For both tasks, the training set is in English, while human annotated development and test sets are available for a bunch of different languages. The model is evaluated on the test data of the task in the target languages.

For cross-lingual data augmentation, we first randomly sample a target language and then adapt the generating method for each data augmentation method. Since XNLI covers more target languages than PAWS-X, we set $t_f = 2, t_p = 2, t_c = 1$ in XNLI, and $t_f = 1, t_p = 1, t_c = 1$ in PAWS-X, where t_f, t_p

Proceedings of the 21st China National Conference on Computational Linguistics, pages 884-895, Nanchang, China, October 14 - 16, 2022.

Method	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	Avg
cross-lingual transfer (Models are fine-tuned on English training data only.)																
InfoXLM	86.4	74.2	79.3	79.3	77.8	79.3	80.3	72.2	77.6	67.5	74.6	75.6	67.3	77.1	77.0	76.5
HITCL	86.3	74.8	80.6	79.5	78.9	81.3	80.5	73.1	79.0	69.9	75.7	75.4	69.7	77.4	77.6	77.3
xTune*	84.7	76.7	81.0	79.9	79.4	81.6	80.5	75.6	77.9	68.4	75.4	77.2	72.2	78.1	77.4	77.7
XLMR-base	84.8	72.7	78.8	77.9	76.5	79.8	78.9	72.2	76.5	66.8	73.9	73.7	68.0	76.8	75.4	75.5
MVA	85.0	75.0	79.1	78.2	78.1	79.7	79.1	72.5	76.8	68.9	75.5	74.5	70.0	76.9	77.4	76.5
CL	84.4	75.5	80.0	79.3	78.7	80.4	79.8	74.1	78.3	71.5	76.1	76.0	71.0	78.2	77.8	77.4
SCL	**86.3**	**77.8**	**81.7**	**81.3**	**80.6**	**82.7**	**81.8**	**76.3**	**80.4**	**73.8**	**78.9**	**78.1**	**73.1**	**80.5**	**80.2**	**79.6**
translate-train (Models are fine-tuned on both English data and its translations.)																
InfoXLM	86.5	78.9	82.4	82.3	81.3	83.0	82.6	77.8	80.6	73.3	78.9	79.5	71.6	81.0	80.7	80.0
HITCL	86.5	78.1	82.2	80.8	81.6	83.2	82.3	76.7	81.3	73.8	78.6	80.5	73.9	80.4	80.7	80.0
xTune*	86.6	79.7	82.7	82.2	81.9	83.1	82.3	78.9	80.9	75.7	78.4	79.8	75.3	80.5	80.0	80.5
XLMR-base	84.3	76.9	80.3	79.8	79.1	81.5	80.3	75.3	78.1	72.9	77.1	77.4	70.8	79.8	79.7	78.2
MVA	85.4	78.5	81.5	81.8	**80.6**	82.3	81.0	77.3	79.9	74.1	**78.8**	78.2	73.5	80.2	80.2	79.6
CL	85.9	77.2	81.6	80.5	80.0	81.7	81.5	76.5	80.3	73.5	77.8	78.2	72.5	79.9	79.9	79.1
SCL	**86.4**	**78.8**	**82.0**	**82.0**	80.5	**82.9**	**82.3**	**77.3**	**80.5**	**74.5**	78.6	**79.7**	**74.2**	**80.9**	**80.3**	**80.1**

Table 1: Experiment results on XNLI. Results with * are reimplemented by us with their released codes. InfoXLM (Chi et al., 2021a) and HITCL (Wei et al., 2021) use contrastive learning while xTune (Zheng et al., 2021) uses multi-view alignment. Notice xTune uses more augmentation data and model ensemble compared to us.

and t_c respectively represent the number of samples generated by full-translation, partial translation and code-switching for each training data. Therefore, each training batch contains $6 \times batch\_size$ sentence pairs in XNLI and $4 \times batch\_size$ sentence pairs in PAWS-X. The code-switching ratio r_c is set as 0.75 in XNLI and 0.5 in PAWS-X. For cross-lingual retrieval tasks mentioned below, each training pair from PAWS-X is detached into two sentences when feeding to the model, and we do not incorporate code-switching as data augmentation.

4.2 Setup

For sentence pair classification tasks of XNLI and PAWS-X, we concatenate the input as the formation defined by XLM-R:

```
[s] input1 [\s] input2 [\s]
```

and we use the final hidden layer corresponding to `[s]` as aggregated representation. For retrieval tasks of BUCC and tatoeba, we perform alignment on the same aggregated representation, but the retrieval is performed on the averaged pooled eighth layer, following the related works (Chi et al., 2021b; Chi et al., 2021a). Adam optimizer is applied with a learning rate of 5e-6. Batch size is set as 24 for XNLI, 36 for PAWS-X and 48 for retrieval.

We evaluate a number of strong baselines and the three post-align strategies discussed in the former section. The baseline is trained with cross-entropy loss with no alignment term serving as cross-lingual regularizer. Then we create cross-lingual augmentations with different methods, and apply different alignment strategies. Three groups of augmentations (full-translation, partial translation, code-switching) are mixed together. The bilingual dictionaries we used for code-switch substitution are from MUSE (Lample et al., 2018). For languages that cannot be found in MUSE, we ignore these languages since other bilingual dictionaries might be of poorer quality. The machine translated training set is taken from the XTREME repository, which is obtained by an in-house translation model from Google.

We mainly compare with models that learn multilingual contextual representations as they have achieved state-of-the-art results on cross-lingual tasks. All cross-lingual alignment strategies are applied to pre-trained XLM-R-base. Following the trend of Hu et al. (2020), we mainly consider the following two scenarios:

Cross-lingual Transfer: the models are fine-tuned on English training data, and directly evaluated on different target languages.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 884-895, Nanchang, China, October 14 - 16, 2022.

Method	en	de	es	fr	ja	ko	zh	avg
cross-lingual transfer (Models are fine-tuned on English training data only.)								
InfoXLM*	94.7	89.7	90.1	90.4	78.7	79.0	82.3	86.4
xTune*	93.7	90.2	89.9	90.4	82.6	81.9	84.3	87.6
XLMR-base	94.5	88.4	89.4	89.3	76.0	77.2	82.6	85.3
MVA	95.0	89.1	90.9	90.6	79.5	81.1	83.7	87.1
CL	94.6	89.8	91.3	90.9	78.9	80.0	82.8	86.9
SCL	**95.3**	**91.3**	**91.8**	**91.7**	**83.2**	**84.5**	**85.7**	**89.0**
translate-train (Models are fine-tuned on both English data and its translations.)								
InfoXLM*	94.5	90.5	91.6	91.7	84.4	83.9	85.8	88.9
xTune*	93.9	90.4	90.9	91.7	85.6	86.8	86.6	89.4
XLMR-base	95.0	89.8	91.8	91.6	81.2	84.3	84.4	88.3
MVA	95.3	90.9	92.0	91.8	83.1	83.6	85.3	88.8
CL	95.4	90.2	92.1	91.4	81.7	84.0	85.3	88.6
SCL	**95.5**	**91.4**	**92.3**	**92.3**	**83.2**	**85.0**	**87.2**	**89.5**

Table 2: Experiment results on PAWS-X. Results with * are reimplemented by us with their released codes.

Translate-train: the models are fine-tuned on the concatenation of English training data and its translation to all target languages. Translate-train is normally a strong baseline for cross-lingual transfer tasks. For classification tasks, it is straightforward that the translation should be assigned with the same label.

In both settings, the alignment term is combined with the canonical cross-entropy loss to be back-propagated together. We use KL Divergence as the similarity measure for multi-view alignment. For contrastive learning, we only consider in-batch negative samples, leaving more complicated methods (e.g. to maintain a memory bank for negative samples (He et al., 2020)) to the future.

4.3 Main Results

As shown in Table 1 and Table 2, we can see that our proposed method could improve the cross-lingual transfer results of pre-trained XLM-R by a large margin. Our method is especially effective in zero-shot setting, where the accuracy is improved by 4.1 points on XNLI and 3.7 points on PAWS-X. Our method can also achieve significant improvement in translate-train setting, where the accuracy is improved by 1.9 points on XNLI and 1.2 points on PAWS-X. Results are consistently improved among all languages, despite their relation with English close or not.

The results of multi-view alignment and vanilla contrastive learning, despite using the same augmentation data, underperform our method on both datasets. This proves the pre-trained representations are better aligned according to the label information after SCL. Different representations, despite belonging to different languages, are projected to the same cluster if they belong to the same class.

SCL is a simple yet effective framework to align the pre-trained multilingual representations on downstream tasks. Cross-lingual signals can be obtained by machine translation or bilingual dictionary, therefore no extra human annotation is needed. While previous works also propose other methods to align the pre-trained representations, the results in Table 1 and 2 prove the superiority of our method.

5 Analysis and Discussion

5.1 Different Augmentations

In this section, we want to explore the influence of different cross-lingual augmentations. We apply different groups of augmentations under the zero-shot setting, and compare the results on different tasks.

As shown in Table 3, we can see that the results of full translation and partial translation are better than code-switching. We think it is because the information provided by code-switching is comparably sparse, only a few anchor words covered by the bilingual dictionary. On the other side, well-trained machine translation system can provide fluent and accurate translation, therefore the multilingual representation can be better aligned. We can also tell that the results of our proposed method outperform the counterparts again on both datasets, proving its superiority.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 884-895, Nanchang, China, October 14 - 16, 2022.

AugData	Method	XNLI en	XNLI avg	PAWS-X en	PAWS-X avg
None	XLMR	84.9	75.5	94.5	85.3
full-trans	MVA	85.2	76.6	94.9	87.1
	CL	85.0	77.9	94.9	87.2
	SCL	85.6	**79.2**	95.3	**88.7**
partial-trans	MVA	83.7	75.7	95.2	86.5
	CL	84.5	76.9	94.9	86.6
	SCL	85.3	78.4	95.3	88.1
code-switch	MVA	85.3	76.4	94.7	86.1
	CL	84.5	76.1	95.2	86.5
	SCL	84.8	76.2	95.1	87.2

Table 3: Experiment results on XNLI and PAWS-X based on different cross-lingual data augmentations, including full-translation, partial translation, and code-switching. For each group of data, we apply all three post-align methods.

similarity measure	lambda	XNLI en	XNLI avg
KLDiv	1	85.19	76.64
	10	85.05	76.71
Symmetric KLDiv	1	84.67	76.17
	10	83.85	76.20
Cosine Similarity	1	83.03	75.16
	10	84.05	76.38
Mean-Square Error	1	83.95	75.37
	10	84.35	76.58

Table 4: Experiment results of different similarity measures and loss weight λ on XNLI. Here we only use the augmentation of full-translation, and the results is in cross-lingual setting. We do not experiment on PAWS-X due to resource limitation.

5.2 Similarity Measure

The similarity measure in L_{MVA} has many alternatives. Previous studies on multi-view learning propose all kinds of measures (Yang et al., 2019b), such as Cosine-Similarity, Mean-Square Error, Kullback-Leibler Divergence and Symmetric Kullback-Leibler Divergence. Suppose we are dealing with an input x and its augmentation $\hat{x}$, different similarity measures can be denoted as:

$$L_{KLDiv} = \Phi(x) log \frac{\Phi(\hat{x})}{\Phi(x)}$$

$$L_{SymKLdiv} = \Phi(x) log \frac{\Phi(\hat{x})}{\Phi(x)} + \Phi(\hat{x}) log \frac{\Phi(x)}{\Phi(\hat{x})}$$

$$L_{cosine} = \frac{\Phi(x) \cdot \Phi(\hat{x})}{||\Phi(x)|| \, ||\Phi(\hat{x})||}$$

$$L_{MSE} = ||\Phi(x) - \Phi(\hat{x})||^2$$

where $\Phi(\cdot)$ denotes the $L2$-normalized aggregated representation. We experiment different similarity measures on the multi-view alignment objective, in combination with different loss weight λ, and the results are shown in Table 4. Surprisingly, we do not see a clear difference between different measures, and in the end we decide to use cosine similarity with $\lambda = 10$ in all experiments. On the other hand, λ is set as 1 for contrastive learning.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 884-895, Nanchang, China, October 14 - 16, 2022.

setting	temp	XNLI en	XNLI avg	PAWS-X en	PAWS-X avg
cross-transfer	1.0	85.6	79.2	95.3	88.7
	0.3	85.2	79.1	94.8	88.7
	0.1	85.8	79.2	95.3	88.2
translate-train	1.0	86.4	79.8	95.4	89.0
	0.3	85.8	79.8	95.4	89.1
	0.1	85.9	79.5	95.3	89.2

Table 5: Experiment results of different contrast temperatures on XNLI and PAWS-X. Here we only use the augmentation of full-translation, and the results are based on supervised contrastive learning.

5.3 Contrast Temperature

Previous empirical observations show that an adjustable scalar temperature parameter τ can improve the performance of contrastive learning (Wang and Isola, 2020; He et al., 2020). Lower temperature increases the influence of examples that are harder to separate, effectively creating harder negatives. However, we do not find such a pattern in our experiments, as shown in Table 5, and finally we decide to set the temperature τ as 1.0 in all experiments.

5.4 SCL for Cross-lingual Retrieval

To further prove the importance of label information in cross-lingual fine-tuning, we also apply the alignment methods on cross-lingual sentence retrieval tasks. We experiment on two datasets, BUCC[1] and tatoeba (Artetxe and Schwenk, 2019). Both datasets aim at extracting parallel sentences from a comparable corpus between English and other languages, with BUCC covering 4 languages and tatoeba covering more than 100 languages. To compare with previous works, we only use a subset of tatoeba (33 languages) in this work.

The pre-trained multilingual models are able to provide language-deterministic representations by nature. Previous works directly calculate the similarity of different sentences by representations from the pre-trained model, to determine whether two sentences are parallel or not (Hu et al., 2020; Chi et al., 2021b; Chi et al., 2021a). In this work, we propose to use the data of paraphrase identification, including the original training sentence pairs and their translations to six languages, to post-align the pre-trained representations.

We compare the previously proposed three strategies to post-align the pre-trained representations. Since we are dealing with retrieval task, the sentence pair from two different languages are encoded separately by the pre-trained XLM-R. We apply the alignment training methods on the aggregated representation. For multi-view alignment, only two translation pairs are pulled closer to each other. For vanilla contrastive learning, we treat all translation pairs as positive while the others as negative. For our proposed SCL, both translation pairs and translation with paraphrasing pairs are deemed as positive, while the others are deemed as negative, as denoted by the following formula:

$$L_{SCL} = -\sum_{j=1}^{N} \mathbb{I}_{y_{ij}=1} \log \frac{s(\Phi(x_i), \Phi(\hat{x}_j))}{s(\Phi(x_i), \Phi(\hat{x}_j)) + \sum_{k=1}^{N} \mathbb{I}_{y_{ik}\neq 1} s(\Phi(x_i), \Phi(\hat{x}_k))}$$

where x_i is a training sample and $\hat{x_i}$ is its translation, and $y_{ij} = 1$ denotes x_i and x_j are a paraphrase pair. After the fine-tuning stage, following previous work, we utilize the average pooled hidden representation of the eighth layer of the pre-trained model as the sentence representation.

As shown in Table 6 and Table 7, paraphrase identification dataset with translated augmentation, despite containing noise generated by the MT model, can provide cross-lingual signal to post-align the multilingual representations. Vanilla contrastive learning can perform alignment space by pulling translation pairs together and pushing translation pairs apart, but paraphrase pairs also possess the same semantics, and should not be contrasted as negative samples. After introducing label information into contrast, the

[1] https://comparable.limsi.fr/bucc2017

Proceedings of the 21st China National Conference on Computational Linguistics, pages 884-895, Nanchang, China, October 14 - 16, 2022.

Method	en-de	en-fr	en-ru	en-zh	avg
mBERT*	62.5	62.6	51.8	50.0	56.7
XLM*	56.3	63.9	60.6	46.6	56.8
XLMR-large*	67.6	66.5	73.5	56.7	66.0
XLMR-base	82.68	74.85	82.08	64.09	75.93
MVA	43.92	26.24	38.71	7.58	29.11
CL	87.22	79.93	86.88	78.83	83.21
SCL	**88.82**	**81.88**	**88.01**	**82.47**	**85.29**

Table 6: Experiment results on BUCC2018 test set. Results with * are released by XTREME(Hu et al., 2020). We apply different post-align strategies on pre-trained XLM-RoBERTa-base model using the training set of PAWS-X with translation augmentation.

Method	en-xx	xx-en
XLMR-base*	55.50	53.40
XLM-E(Chi et al., 2021b)	65.00	62.30
InfoXLM(Chi et al., 2021a)	68.62	67.29
XLMR-base	55.60	53.49
MVA	28.00	27.79
CL	78.80	77.87
SCL	**80.41**	**80.84**

Table 7: Experiment results on tatoeba. Result with $*$ is released by (Chi et al., 2021b). xx denotes the 33 languages as experimented in (Chi et al., 2021a) and (Chi et al., 2021b), and we release the averaged accuracy in both directions.

retrieval accuracy is further improved by 2-3 points. On the contrary, multi-view alignment would lead to representation collapse and cannot converge at all. This is in line with our previous analysis.

6 Conclusion

In this paper, we propose to improve cross-lingual fine-tuning with supervised contrastive learning. Cross-lingual supervision is created by augmenting the training set, and different methods to post-align the multilingual pre-trained representation are compared. We propose to incorporate label-information when performing cross-lingual contrastive fine-tuning, and outperforms previous methods by a large margin on four cross-lingual transfer benchmark datasets.

Canonical cross-entropy has many intrinsic problems, especial when performing transfer learning tasks, and contrastive learning can be a decent supplementary. By alleviating the commonality and differences between different examples, representations are efficiently transferred from one domain or language to another. In the future, we would explore the application of supervised contrastive learning on other transfer learning tasks, including token-level classification, language generation, cross-domain transfer, etc.

Acknowledgement

This research work is supported by the National Key R&D Program of China (2019YFB1405200), the National Nature Science Foundation of China (No. 61976016, 61976015 and 61876198) and Toshiba (China) Co.,Ltd. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *Proc. of ICLR*.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021a. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proc. of ACL*.

Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Saksham Singhal, Payal Bajaj, Xia Song, and Furu Wei. 2021b. XLM-E: cross-lingual language model pre-training via ELECTRA. *CoRR*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proc. of EMNLP*.

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 *Proc. of ACL*.

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

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proc. of ACL*.

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

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

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Proc. of NeurIPS*.

Saurabh Kulshreshtha, Jose Luis Redondo Garcia, and Ching-Yun Chang. 2020. Cross-lingual alignment methods for multilingual BERT: A comparative study. In *Proc. of ACL*.

Siyu Lai, Hui Huang, Dong Jing, Yufeng Chen, Jinan Xu, and Jian Liu. 2021. Saliency-based multi-view mixed language training for zero-shot cross-lingual classification. In *Proc. of ACL*.

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

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *Proc. of ICLR*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv*.

Lin Pan, Chung-Wei Hang, Haode Qi, Abhishek Shah, Saloni Potdar, and Mo Yu. 2021. Multilingual BERT post-pretraining alignment. In *Proc. of ACL*.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP. In *Proc. of IJCAI*.

Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. XLDA: cross-lingual data augmentation for natural language inference and question answering. *CoRR*.

Chenxin Tao, Honghui Wang, Xizhou Zhu, Jiahua Dong, Shiji Song, Gao Huang, and Jifeng Dai. 2021. Exploring the equivalence of siamese self-supervised learning via A unified gradient framework. *CoRR*.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*.

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

Proceedings of the 21st China National Conference on Computational Linguistics, pages 884-895, Nanchang, China, October 14 - 16, 2022.

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

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2021. On learning universal representations across languages. In *Proc. of ICLR*.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019a. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In *Proc. of EMNLP*.

Zonghan Yang, Yong Cheng, Yang Liu, and Maosong Sun. 2019b. Reducing word omission errors in neural machine translation: A contrastive learning approach. In *Proc. of ACL*.

Bo Zheng, Li Dong, Shaohan Huang, Wenhui Wang, Zewen Chi, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. 2021. Consistency regularization for cross-lingual fine-tuning. In *Proc. of ACL*.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 884-895, Nanchang, China, October 14 - 16, 2022.