Synergistic Anchored Contrastive Pre-training for Few-Shot Relation Extraction

Da Luo, Yanglei Gan, Rui Hou , Run Lin , Qiao Liu *, Yuxiang Cai , Wannian Gao

University of Electronic Science and Technology of China
{luoda, yangleigan, hour, runlin, yuxiangcai, wanniangao}@std.uestc.edu.cn, qliu@uestc.edu.cn

Abstract

Few-shot Relation Extraction (FSRE) aims to extract relational facts from a sparse set of labeled corpora. Recent studies have shown promising results in FSRE by employing Pre-trained Language Models (PLMs) within the framework of supervised contrastive learning, which considers both instances and label facts. However, how to effectively harness massive instance-label pairs to encompass the learned representation with semantic richness in this learning paradigm is not fully explored. To address this gap, we introduce a novel synergistic anchored contrastive pre-training framework. This framework is motivated by the insight that the diverse viewpoints conveyed through instance-label pairs capture incomplete yet complementary intrinsic textual semantics. Specifically, our framework involves a symmetrical contrastive objective that encompasses both sentence-anchored and label-anchored contrastive losses. By combining these two losses, the model establishes a robust and uniform representation space. This space effectively captures the reciprocal alignment of feature distributions among instances and relational facts, simultaneously enhancing the maximization of mutual information across diverse perspectives within the same relation. Experimental results demonstrate that our framework achieves significant performance enhancements compared to baseline models in downstream FSRE tasks. Furthermore, our approach exhibits superior adaptability to handle the challenges of domain shift and zero-shot relation extraction. Our code is available online at https://github.com/AONE-NLP/FSRE-SaCon.

Introduction

Relation extraction (RE) is a crucial task in Natural Language Processing (NLP), aiming to extract relationships between entities in a given sentence (Zhou et al. 2005). It has broad applications in question answering (Bordes, Chopra, and Weston 2014) and knowledge base construction (Luan et al. 2018). However, the scarcity of annotated data poses a significant challenge in developing RE models. Therefore, recent studies have introduced the few-shot relation extraction (FSRE) task as an effective solution.

Inspired by the success of pre-trained language models (PLMs) in NLP (Dong et al. 2019; Qin et al. 2021), various pre-training frameworks have been introduced into FSRE (Peng et al. 2020; Liu et al. 2022a; Zhang and Lu 2022). These frameworks aim to learn transferable knowledge in the form of initial embeddings, which are then fine-tuned with a learned optimization strategy to enhance performance with a limited amount of labeled corpora for training.

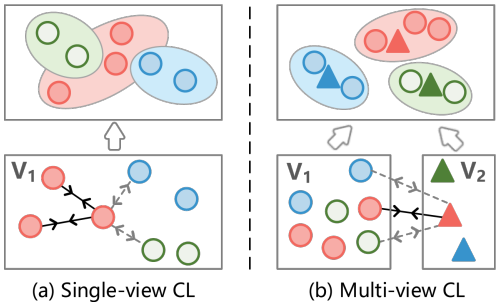


Figure 1: The concepts of single-view CL and multi-view CL. Each color corresponds to a distinct relation, with circles and triangles symbolizing instances and labels. Solid lines denote proximity between instances (a) or instance-label pair (b), while dashed lines indicate instances moving apart (a) or mismatched instance-label pairs (b).

Among these pre-training frameworks, contrastive learning (CL) has emerged as a popular paradigm. The key idea is to *bring together* samples from the same class in semantic embedding space, while *pushing apart* samples from different classes (Chen et al. 2020). In this way, the target for these frameworks in FSRE changes from learning generic representations that capture meaningful patterns and structures from large external corpus to creating discriminative embeddings by pushing apart positive and negative samples.

Recent CL-based pre-training frameworks have achieved state-of-the-art performance on FSRE benchmarks. However, existing studies (Soares et al. 2019; Peng et al. 2020; Zhang and Lu 2022) have predominantly focused on establishing instance-label alignment through single-view contrastive learning, where the correlation between instances and labels is dependent on a single view per instance. In this regard, we argue that the adoption of mutual contrastive learning among multiple views can yield representations that exhibit both robustness to inconsequential variations and re-

*Corresponding author

tention of necessary task-relevant information (Tian, Krishnan, and Isola 2020). The primary goal in this context is to maximize the lower bound on mutual information between the instance (V_1) and its label (V_2), as illustrated in Figure 1. Despite recent attempts to introduce a multi-view contrastive learning framework (Dong, Pan, and Luo 2021), wherein view V_1 is considered as an anchor and iteratively compared with V_2, this approach may introduce bias in relation representation and hinder the model's capacity to generalize to previously unseen relations.

In light of these considerations, we propose a novel pre-training framework based on **S**ynergistic **A**nchored **Con**trastive (SaCon) learning. Specifically, from the point of multi-view coding, we first employ two separate encoders to map sentences and labels into the same vector space and obtain instance-level and label-level representations, respectively. Then these representations serve as anchors for each other, enabling contrastive similarity distributions between diverse embedding spaces derived from two views. To this end, we introduce a symmetrical contrastive loss to enforce the consistency between view V_1 and V_2 for mutual calibration. This symmetrical loss can help SaCon learn robust semantic representations within the same relation space, thus improving the generalization capability of the model. Extensive experiments demonstrate that our proposed SaCon indeed enhances the performance of various downstream FSRE baselines and achieves state-of-the-art results comparing to other pre-training framework for FSRE. Our main contributions are as follows:

- We propose a novel pre-training framework for few-shot relation extraction based on the concept of multi-view contrastive learning, aiming to learn robust representations via modeling instance-label correlations between diverse views in a synergistic manner.
- We present a novel symmetrical loss function that incorporates a consistency cost to facilitate the learning of representations invariant to certain relation classes of variations, thus enhancing the generalizability of the model.
- Extensive experiments on two FSRE benchmarks demonstrate the effectiveness of our proposed framework, even in challenging scenarios involving domain shift and zero-shot relation extraction tasks settings.

Related Work

Few-Shot Relation Extraction

Few-shot relation extraction (FSRE) aims to classify relations between entities with a limited quantity of labeled data. One popular method is the prototypical network (Snell, Swersky, and Zemel 2017). For instance, prior research (Gao et al. 2019a; Sun et al. 2019; Ye and Ling 2019) endeavors to improve prototypical network's performance by incorporating attention mechanisms. However, these approaches only utilize information from sentences, which have shown limited improvement. Therefore, several algorithms (Qu et al. 2020; Han, Cheng, and Lu 2021; Liu et al. 2022b) have been proposed to compensate the limited information in FSRE by introducing external relation label information, achieving comparable results to the state-of-the-art approaches.

Another line of work focuses on further training PLMs, coupled with a new architecture for fine-tuning relation representations in BERT (Devlin et al. 2019). Based on the objective that the relation representations should be similar if they range over the same pairs of entities, (Soares et al. 2019) proposed a new method, matching the blanks (MTB), to learn relation representations directly from text. To sample data with more diversity than MTB, (Peng et al. 2020) adopted a less strict rule, assuming that sentences with the same relation should have similar representations, and proposed an entity-masked contrastive pre-training framework for FSRE. More recently, (Zhang and Lu 2022) have introduced label prompts to enhance instance representations and proposed a contrastive pre-training approach with label prompt drop out to create a more challenging learning setup.

All of the aforementioned approaches can be regarded as single-view learning, which only use instance-view or concatenate all instance and label views into one single view. These learning methods would be prone to the issue of overfitting, wherein the model becomes excessively tailored to the training data, hindering its ability to generalize to new and unseen examples (Xu, Tao, and Xu 2013). To overcome this issue, (Dong, Pan, and Luo 2021) have focusd on multi-view learning, proposed a framework considering both instance-view and label-view semantic mapping information. However, it ignored the interaction between this two views, leading to incomplete and biased relation representations. To fill this gap, we propose a novel pre-training framework, named SaCon, to learn both instance-view and label-view representations in a synergistic way.

Contrastive Learning

Contrastive Learning has recently received interest due to its success in self-supervised representation learning in Computer Vision (He et al. 2020; Chen et al. 2020). The common idea of these works is pulling together an anchor and a "positive" sample in the embedding space, and pushing apart the anchor from many "negative" samples. Recently, from the classic hypothesis that a powerful representation is one that models view-invariant factors, (Tian, Krishnan, and Isola 2020) extended the contrastive learning to the multiview settings, attempting to maximize mutual information between representations of different views of the same scene. Based on the idea of multiview contrastive learning, (Zhang et al. 2022) proposed ConVIRT, an supervised strategy to learn medical visual representations by exploiting naturally occurring paired descriptive text. And (Radford et al. 2021) presented a more universal pre-training paradigm to enable zero-shot transfer to many existing image classification task via natural language prompting. Inspired by this work, we propose a synergistic anchored contrastive learning method for our SaCon, equipped with a symmetrical loss function to obtain more robust and invariant relation representations.

Zero-Shot Relation Extraction

Zero-shot relation extraction (ZSRE) aims to predict semantic relationships between entities without requiring any labeled training instances specific to the target relations. (Levy et al. 2017; Cetoli 2020; Bragg et al. 2021) have learned

this task by extending the machine reading-comprehension techniques. Another direction is matching-based methods including text-entailment-based methods and representation matching-based methods. Text-entailment-based methods (Sainz et al. 2021) require the model to predict whether the input sentence containing two entities matches the description of a given relation; Representation matching-based methods (Qu et al. 2020; Chen and Li 2021; Dong, Pan, and Luo 2021) separately encode the relations and instances into the same semantic space and then take these two embeddings for comparison. In our work, we will undertake the pre-training of two distinct semantic encoders, one specialized for instances and the other for labels, thus facilitating the inherent implementation of representation matching in the ZSRE paradigm.

Task Definition

Few-Shot RE We follow the typical N-way-K-shot settings for FSRE, which contains a support set S and a query set Q in each episode. Specifically, for a FSRE task, we randomly select N relation classes from base classes, each with K instances to form a support set $S = \{s_k^i; 1 \leq i \leq N, 1 \leq k \leq K\}$. Meanwhile, a query set Q is sampled from the remaining data of the N classes, denoted as $Q = \{q_j; 1 \leq j \leq T\}$, with T representing the number of samples. The model is trained on support set S and then used to predict the correct relations for the query instance in Q.

Zero-Shot RE Different from FSRE, there is no support instances in ZSRE. That is, the K in N-way-K-shot FSRE settings is 0. Given a sentence with a relation $\mathcal{R}$ targeted for extraction, zero-shot relation extraction is the learning of a model without any use of instances from $\mathcal{R}$ during training.

Approach

Knowledge Base Construction

In this work, we aim to construct a comprehensive knowledge base using the English Wikipedia corpus to facilitate the extraction of prior knowledge for RE. The knowledge base consists of a label dictionary represented as $\mathcal{L} = \{L_1, L_2, ..., L_M\}$ and a set of sentences denoted as $\mathcal{S} = \{S_1, S_2, ..., S_N\}$. Here, M and N correspond to the number of relation classes and sentences, respectively.

The label dictionary is created by mapping each sentence's relation identifier to its corresponding relation label and description. As an example, considering the following sample sentence "*The Pythagorean theorem states how to determine when a triangle is a right triangle.*", which entails the relational fact "`statement describes`", accompanied by the relational description of "*formalization of the statement contains a bound variable in this class*". By combining the relation label and description, the entry "`statement describes`: *formalization of the statement contains a bound variable in this class*" is added to the label dictionary.

Encoder

The overall architecture of SaCon is shown in Figure 2. Given a mini-batch of training examples denoted as $\{s_n, l_m\}$, where $s_n \in \mathcal{S}$ and $l_m \in \mathcal{L}$, our SaCon is built upon a bi-encoder architecture consisting of two isomorphic and fully decoupled BERT models (Devlin et al. 2019), i.e. a label encoder Φ_l and a sentence encoder Φ_s.

Label Embeddings The objective of the label encoder is to generate label embeddings $\mathbf{l}_i^e$ for each label l_i in the label dictionary, denoted as:

$$\begin{aligned} \mathbf{l}_i^f = \Phi_l(l_i) = \{\mathbf{h}_{cls}^L, \mathbf{h}_1^L, \mathbf{h}_2^L, ..., \mathbf{h}_{l_c}^L | \mathbf{h}_j^L \in \mathbb{R}^d\}, \\ \mathbf{l}_i^e = \mathbf{h}_{cls}^L \oplus Meanpool(\mathbf{l}_i^f), i \in m \end{aligned} \tag{1}$$

where l_c represents the length of relation class, $\mathbf{h}_j^L$ denotes the hidden state of the j-th token in the input sequence, and $\mathbf{l}_i^e \in \mathbb{R}^{2\times d}$, where d denotes the dimension of the hidden state. $Meanpool(\mathbf{l}_i^f)$ is the average of embeddings of all tokens in l_i, while $\oplus$ signifies the concatenation operation.

Sentence Embeddings Inspired by the context representations architecture in (Soares et al. 2019), we adopt special markers ($[E1_s, E1_e, E2_s, E2_e]$) to highlight the positions of entity mentions within the input sentence. To illustrate, considering the sentence "**Entity1** was founded by **Entity2**.", the input sequence is "$[CLS]$ $[E1_s]$ Entity1 $[E1_e]$ was founded by $[E2_s]$ Entity2 $[E2_e]$. $[SEP]$". Each embedding of the input sequence $\mathbf{s}_i^e$ is denoted as:

$$\begin{aligned} \mathbf{s}_i^f = \Phi_s(s_i) = \{\mathbf{h}_{cls}^S, \mathbf{h}_1^S, \mathbf{h}_2^S, ..., \mathbf{h}_{l_s}^S | \mathbf{h}_j^S \in \mathbb{R}^d\}, \\ \mathbf{s}_i^e = \mathbf{h}_b^S \oplus \mathbf{h}_e^S, i \in n \end{aligned} \tag{2}$$

Where l_s represents the length of input sequence, $\mathbf{s}_i^e \in \mathbb{R}^{2\times d}$, b and e corresponds to the positions of special tokens $[E1_s]$ and $[E2_s]$, respectively. $\oplus$ stands for concatenation. To mitigate the model's reliance on shallow cues of entity mentions during pre-training (Peng et al. 2020), we randomly mask entity spans by replacing them with the special $[BLANK]$ token. The probability of masking an entity span is denoted as $\rho_{blank} = 0.7$.

Pre-training Stage

Our proposed SaCon incorporates two pre-training tasks: symmetrical contrastive learning (SCL) and masked language modeling (MLM).

SCL In our training setup, we work with a mini-batch consisting of n sentences and m labels. The objective is to predict the actual (sentence, label) pairs out of the $n \times m$ possible combinations. To achieve this, we employ contrastive learning as part of our SaCon model. The main goal of contrastive learning is to establish a robust and uniform representation space by training a sentence encoder and a label encoder in a synergistic way. This training process involves maximizing the cosine similarity between embeddings of n correct pairs, while minimizing the cosine similarity of embeddings belonging to the $n \times m - n$ incorrect pairs.

To achieve effective contrast, we incorporate a symmetrical contrastive objective, which encompasses two components: sentence-anchored contrastive learning (SCL_s) and label-anchored contrastive learning (SCL_l). This symmetrical contrastive objective ensures that our framework captures the inherent relationships between sentences and labels, leading to improved performance in the downstream few-shot relation extraction tasks.

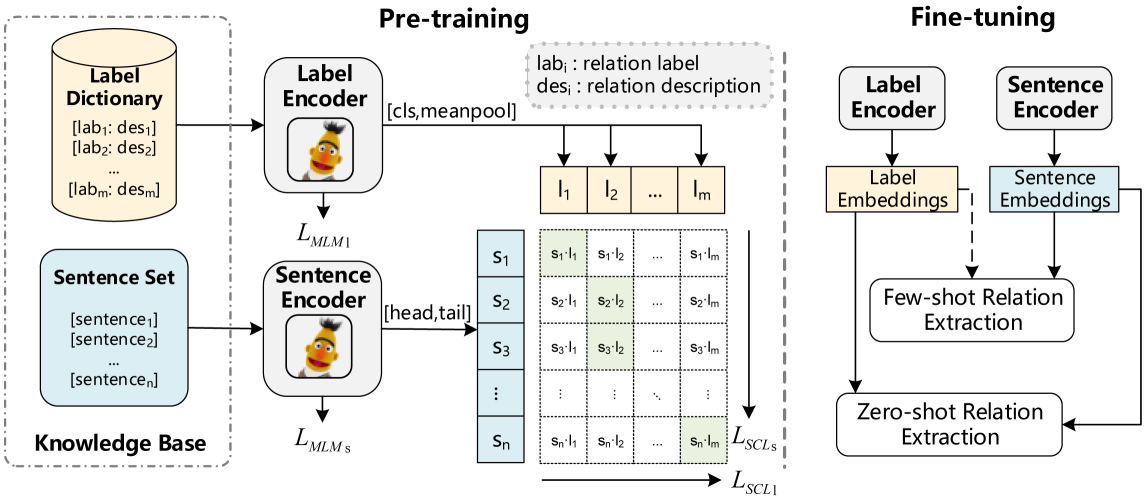


Figure 2: The model overview of SaCon. The SaCon framework involves simultaneous training of a label encoder and a sentence encoder during the pre-training stage to predict accurate pairings (depicted in green) for a batch of [label, sentence] training examples. In the subsequent fine-tuning stage, the adeptly trained label encoder and sentence encoder handle tasks like few-shot relation extraction or zero-shot relation extraction by effectively incorporating information from both labels and sentences.

$\mathbf{SCL_s}$ The sentence-anchored contrastive learning focuses on each individual sentence s_i as the anchor and extracts positive and negative samples from the set of relation labels. Specifically, for a given sentence s_i in the positive pair (s_i, l_i), any label in the remaining pairs forms a negative pair with s_i, denoted as (s_i, l_j), where $1 \leq j \leq m-1$. The training loss is denoted as:

$$L_{SCL_s} = -\frac{1}{n}\sum_{i=1}^{n} log\frac{exp(< \mathbf{s}_i^e, \mathbf{l}_i^e > /\tau)}{\sum_{k=1}^{m} exp(< \mathbf{s}_i^e, \mathbf{l}_k^e > /\tau)}, \quad (3)$$

where $< \mathbf{s}_i^e, \mathbf{l}_i^e >$ represents the cosine similarity between $\mathbf{s}_i^e$ and $\mathbf{l}_i^e$, and τ represents a temperature parameter.

$\mathbf{SCL_l}$ Symmetrically, we consider the relation labels in a mini-batch as anchors and extract positive and negative samples from the corresponding sentences. For a given label l_i, the set A includes all positive sentences whose labels are l_i, while the negative pairs consist of sentences whose labels are not l_i. The contrastive loss for SCL_l is calculated as:

$$L_{SCL_l} = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{A} log\frac{exp(< \mathbf{l}_i^e, \mathbf{s}_j^e > /\tau)}{\sum_{k=1}^{n} exp(< \mathbf{l}_i^e, \mathbf{s}_k^e > /\tau)}. \quad (4)$$

For any label anchor, all positive sentences in a mini-batch contribute to the numerator. Therefore, the L_{SCL_l} loss encourages the encoder to give closely aligned representations to all positive pairs, resulting in a more robust clustering of the representation space and it will update more stably.

By minimizing these losses, SaCon aims to preserve the maximum mutual information between the true pairs in the latent space under respective representation functions. Our final contrastive training loss is then computed as the combination of the two losses:

$$L_{SCL} = L_{SCL_l} + L_{SCL_s}, \quad (5)$$

where the two contrastive learning tasks are complementary to each other.

MLM To retain the language understanding capabilities of BERT and mitiagte the issue of catastrophic forgetting (Peng et al. 2020), we incorporate the masked language modeling (MLM) objective into our model. The MLM objective involves masking certain tokens or words in a sentence and then predicting those masked tokens based on the surrounding context. For the bi-encoder architecture of SaCon, we have two training objectives for the MLM task, namely L_{MLM_s} and L_{MLM_l}. These objectives are designed to optimize the MLM task for the sentence encoder and label encoder, respectively. The loss function for the masked language model is defined as:

$$L_{MLM} = L_{MLM_s} + L_{MLM_l}, \quad (6)$$

and we have noticed that the loss of SCL is approximately 1.5 to 2 times larger than the MLM loss, which operates on a larger scale. To stike this balance between the two losses, we have set the weight for SCL to 1/2. So the overall pre-training loss is denoted as:

$$L = \frac{1}{2}L_{SCL} + L_{MLM}. \quad (7)$$

Fine-tuning Stage

To facilitate the fine-tuning of downstream models for FSRE, we classify these models into two groups. The first group comprises methods that do not use additional label information. In this case, we initialize the BERT encoder with the pre-trained parameters obtained from our sentence encoder. The second group includes methods that incorporate label information. For these methods, we replace the original encoders with our pre-trained label encoder and sentence encoder, respectively. In the case of zero-shot relation extraction, both the sentence and label encoders are utilized.

Model	5-way-1-shot	5-way-5-shot	10-way-1-shot	10-way-5-shot
FewRel 1.0				
Proto-BERT *	82.92 / 80.68	91.32 / 89.60	73.24 / 71.48	83.68 / 82.89
Proto-BERT+**SaCon**	92.38 / 95.35	96.62 / 97.71	86.78 / 91.02	93.29 / 95.32
BERT-PAIR ♣	85.66 / 88.32	89.48 / 93.22	76.84 / 80.63	81.76 / 87.02
BERT-PAIR + **SaCon**	88.88 / 90.99	92.66 / 95.70	78.65 / 84.24	84.88 / 91.85
REGRAB	87.95 / 90.30	92.54 / 94.25	80.26 / 84.09	86.72 / 89.93
REGRAB + **SaCon**	94.21 / 95.15	97.07 / 97.72	90.80 / 92.35	95.45 / 96.11
HCRP	90.90 / 93.76	93.22 / 95.66	84.11 / 89.95	87.79 / 92.10
HCRP + **SaCon**	96.16 / 96.90	97.57 / 97.81	91.74 / 93.99	95.40 / 96.21
SimpleFSRE	91.29 / 94.42	94.05 / 96.37	86.09 / 90.73	89.68 / 93.47
SimpleFSRE + **SaCon**	**98.17 / 97.88**	**97.98 / 98.12**	**96.21 / 96.65**	**96.46 / 96.50**
FewRel 2.0 Domain Adaptation				
Proto-BERT ♣	40.12	51.50	26.45	36.93
Proto-BERT + **SaCon**	78.30	88.84	**66.81**	78.30
BERT-PAIR ♣	67.41	78.57	54.89	66.85
BERT-PAIR + **SaCon**	70.89	82.49	59.14	69.49
HCRP	76.34	83.03	63.77	72.94
HCRP + **SaCon**	**80.23**	89.57	66.55	80.14
SimpleFSRE ⋆	72.42	89.99	56.03	78.90
SimpleFSRE + **SaCon**	76.41	**90.32**	59.33	**81.12**

Table 1: Accuracy (%) of few-shot classification on the FewRel 1.0 validation / test set and FewRel 2.0 Domain Aadptative test set. FewRel 1.0 is trained and tested on Wikipedia domain. FewRel 2.0 is trained on Wikipedia domain but tested on biomedical domain. ♣ are from FewRel public leaderboard, $*$ is reported by (Peng et al. 2020), and $\star$ represents the results of our implementation. Others are obtained from results reported by papers. The best results are marked in bold.

Experiment

Datasets and Implementation

The pre-training dataset (Peng et al. 2020) is constructed by using Wikipedia articles as the corpus and Wikidata (Vrandečić and Krötzsch 2014) as the knowledge graph, which consists of 744 relations and 867,278 sentences. The fine-tuning process is conducted on FewRel 1.0 (Han et al. 2018) and FewRel 2.0 (Gao et al. 2019b) datasets. FewRel 1.0 is a large-scale few-shot relation classification dataset, consisting of 70,000 sentences on 100 relations derived from Wikipedia and annotated by crowdworkers. FewRel 2.0 is built upon the FewRel 1.0 dataset by adding an additional test set in a quite different domain. Our experiments follow the splits used in official benchmarks, which split the dataset into 64 base classes for training, 16 classes for validation, and 20 novel classes for testing.

Our SaCon is trained on the NVIDIA A100 Tensor Core GPU. It undergoes pre-training on the BERT-base model from the Huggingface Transformer library [1] and uses AdamW (Loshchilov and Hutter 2018) for optimization with the learning rate as 3e-5. The temperature parameter τ is learnable and initialized to 0.07 from (Wu et al. 2018) and clipped to prevent scaling the logits by more than 100. The epoch and batch size are set to 20 and 128. As for fine-tuning, we select the same values as those reported in the baseline methods. The batch size and training iteration are set to 4 and 10,000 with learning rate as {1e-5, 2e-5, 1e-6}.

[1] https://github.com/huggingface/transformers

Evaluation

To evaluate the efficacy of our proposed SaCon, we conduct evaluations based on the quality of the learned representations on downstream baselines. The performance of the downstream models is measured in terms of accuracy on the query set of N-way-K-shot tasks. Following previous studies (Gao et al. 2019b; Qu et al. 2020; Zhang and Lu 2022), we consider N as either 5 or 10, and K as either 1 or 5. We follow the official evaluation settings by randomly selecting 10,000 tasks from the validation data. To determine the final test accuracy, we submit the model predictions to the FewRel 1.0 leaderboard [2] and the FewRel 2.0 leaderboard [3].

Results

Main Results

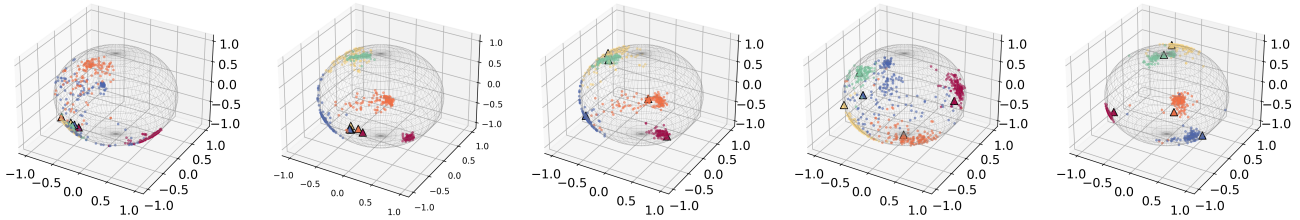
Our pre-trained encoders in SaCon are applied to following baseline approaches: 1) **Proto-BERT** (Snell, Swersky, and Zemel 2017), an original prototypical network with Bert-base (Devlin et al. 2019) as the encoder. 2) **BERT-PAIR** (Gao et al. 2019b), a method measuring the score of each support-query pair corresponding to the same relation. 3) **REGRAB** (Qu et al. 2020), a Bayesian meta-learning approach via a global relation graph. 4) **HCRP** (Han, Cheng, and Lu 2021), an approach based on meta-learning and contrastive learning that learns better representations by utilizing relation label. 5) **SimpleFSRE** (Liu et al. 2022b), a sim-

[2] https://codalab.lisn.upsaclay.fr/competitions/7395

[3] https://codalab.lisn.upsaclay.fr/competitions/7397

Model	5-way-1-shot		5-way-5-shot		10-way-1-shot		10-way-5-shot	
	Proto-BERT	HCRP	Proto-BERT	HCRP	Proto-BERT	HCRP	Proto-BERT	HCRP
FewRel 1.0								
MapRE	76.48	91.77	81.83	94.24	66.13	82.31	71.85	88.89
LPD	92.01	94.97	96.23	96.39	86.61	90.82	92.97	93.49
SaCon	**92.38**	**96.16**	**96.62**	**97.57**	**86.78**	**91.74**	**93.29**	**95.40**
FewRel 2.0 Domain Adaptation								
MapRE	64.62	74.12	74.96	87.76	48.96	57.90	62.88	**80.30**
LPD	63.77	**81.22**	76.05	88.13	52.15	67.57	61.86	80.06
SaCon	**78.40**	81.18	**88.16**	**88.58**	**68.29**	**69.47**	**79.21**	80.07

Table 2: Accuracy of different pre-training methods applied to Proto-BERT and HCRP baselines on FewRel 1.0 and FewRel 2.0 validation datasets. The best results are marked in bold.



(a) MTB (b) CP (c) LPD (d) MapRE (e) SaCon

Figure 3: Instance-level (●) and Label-level (▲) feature distribution plots of five sampled relations on unit hypersphere with five pre-training frameworks.

ple approach using relation information by direct additional way to assist prototypical network.

Table 1 presents comparative results for the two few-shot learning datasets. Our SaCon significantly improves the performance of all FSRE baselines on the FewRel 1.0 dataset by 2% to 10%, with the best results achieved by fine-tuning on the SimpleFSRE base model, demonstrating SaCon's effectiveness. Additionally, we evaluate SaCon's cross-domain transferability by assessing its performance on the FewRel 2.0 domain-adaptive dataset, resulting in considerable improvements of about 3% to 35% that confirm SaCon's robustness. Notably, these enhancements are more pronounced with fewer support instances such as 5-way-1-shot and 10-way-1-shot settings, indicating SaCon's ability under extremely limited resources.

Comparison with Other Contrastive Pre-training Methods

To validate the performance of our SaCon, we visualize the clustering plots of various pre-training methods for FSRE on the unit hypersphere. These methods can be classified into two distinct categories: **Single-view**: MTB (Soares et al. 2019), CP (Peng et al. 2020), LPD (Zhang and Lu 2022). **Multi-view**: MapRE (Dong, Pan, and Luo 2021).

We randomly select five relation classes and their associated instances to illustrate SaCon's representation effectiveness. Using the encoded representations from various FSRE pre-trained models, we visualize their clustering on the unit hypersphere. Each relation category is represented by a unique color. As depicted in Figure 3, when using single-view CL for pre-training (subfigures (a), (b), (c)), the clustering is relatively compact but still exhibits numerous outliers. Notably, the positions of orange instances are relatively sparse, with partial overlap with blue instances, making it challenging to distinguish between them. Furthermore, these representations of relation classes do not form distinct clusters closely associated with their respective instance representations, resulting in inconsistency between instances and their labels. On the other hand, MapRE, which employs multi-view CL, produces label-instance aligned representations but leads to significant dispersion among instances within the same relation class. This could be attributed to its learning pattern, utilizing sentence-anchored CL in both different views, resulting in a biased learning mode that fails to achieve complementarity. In contrast, our proposed SaCon achieves complete alignment and consistency in instance and relation representations.

Moreover, we also compare our SaCon with the most recent pre-training methods through the performance of fine-tuning tasks: **MapRE** and **LPD**. We employ two baselines for fine-tuning: Proto-BERT (Snell, Swersky, and Zemel 2017) and HCRP (Han, Cheng, and Lu 2021). Proto-BERT is a simple approach solely relying on sentence information, whereas HCRP incorporates both sentence and label information. The results in Table 2 indicate that SaCon outperforms MapRE and LPD on almost all FSRE settings.

Ablation Studies

We investigate two variants of SaCon, namely, sentence-anchored contrastive learning (SCL_s) and label-anchored

Method	5-way 1-shot	5-way 5-shot	10-way 1-shot	10-way 5-shot
sentence-anchored	97.65	97.76	95.59	95.91
label-anchored	95.19	97.84	93.33	96.45
SaCon	**98.17**	**97.98**	**96.21**	**96.46**

Table 3: Accuracy of different variants of SaCon applied to SimpleFSRE on the FewRel 1.0 validation set.

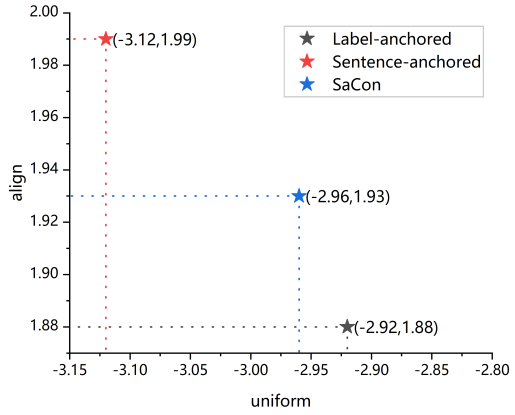


Figure 4: Mean statistics of alignment and uniformity. Lower values indicate better alignment and uniformity.

contrastive learning (SCL_l), to examine the impact of different pre-training options on the fine-tuning performance of the framework. The results of these variants are presented in Table 3. This suggests that both SCL_s and SCL_l contribute synergistically to the overall performance, demonstrating the complementary nature of the multi-view contrastive learning approach.

To further support the notion of complementarity between these contrastive objectives, we examined two key properties related to contrastive learning: alignment and uniformity. Alignment reflects the ability of encoders to assign similar features to similar samples, while uniformity promotes a uniform distribution on the unit hypersphere (Wang and Isola 2020). Followed by (Wang et al. 2022), we introduce two metrics, **align** and **uniform**, to quantitatively evaluate the alignment and uniformity of the learned representations. The calculation of the align and uniform metrics are performed according to the following formulation:

$$\begin{aligned} align &= \mathop{\mathbb{E}}_{(s_i,l_i)\sim p_{\text{pos}}} \|f(\mathbf{s}_i^e) - f(\mathbf{l}_i^e)\|^2, \\ uniform &= \log \mathop{\mathbb{E}}_{s_i,s_j\sim p_{\text{instance}}} e^{-2\|f(\mathbf{s}_i^e)-f(\mathbf{s}_j^e)\|^2}/2 + \\ &\quad \log \mathop{\mathbb{E}}_{l_i,l_j\sim p_{\text{label}}} e^{-2\|f(\mathbf{l}_i^e)-f(\mathbf{l}_j^e)\|^2}/2, \end{aligned} \tag{8}$$

where $p_{\text{pos}}(\cdot,\cdot)$ denotes the distribution of positive pairs, $p_{\text{instance}}(\cdot)$ and $p_{\text{label}}(\cdot)$ denote the distribution of sentences and labels. $\mathbf{s}_i^e$ and $\mathbf{l}_i^e$ denote each sentence and label representations, respectively. $f(.)$ indicates L2 normalization.

Figure 4 provides visual evidence of the characteristics of features obtained through different variants of SaCon. The features derived from label-anchored contrastive learning demonstrate a higher degree of clustering for positive pairs, indicating a stronger alignment in terms of similarity between related samples. However, these features exhibit a poorer uniformity, suggesting a less diverse distribution across the feature space. On the other hand, features obtained through sentence-anchored contrastive learning exhibit the most uniform distribution, indicating a better coverage of the feature space. However, they demonstrate a weaker alignment among related samples.

In contrast, our proposed SaCon demonstrates an intermediate behavior with respect to both alignment and uniformity. By leveraging this combination, SaCon achieves a more optimal trade-off between alignment and uniformity, resulting in enhanced representation learning.

Method	5-way-0-shot	10-way-0-shot
Bert + SQUAD	52.50	37.50
REGRAB	86.00	76.20
MapRE	90.65	81.46
SaCon + SimpleFSRE	**97.23**	**95.20**

Table 4: The comparison results of ZSRE task on FewRel 1.0 validation set in accuracy.

Zero-Shot Relation Extraction

We further investigate the extreme condition of FSRE, zero-shot RE, where no support instances are available during prediction. Table 4 presents the results of our proposed framework compared to three recent ZSRE methods. SaCon obtains significantly performance across all zero-shot settings. Particularly, in the 5-way and 10-way settings, Sacon outperforms the state-of-the-art MapRE by 6.58% and 13.74%, respectively, proving the robust representation capabilities of our proposed pre-training framework.

Conclusion

In this paper, we propose a novel synergistic anchored contrastive pre-training framework which enables the learning of consistent semantic representations from multiple views within a large scale dataset. The framework incorporates a symmetrical contrastive objective, comprising a sentence-anchored contrastive loss and a label-anchored contrastive loss, to ensure consistency across different views. Extensive experiments conducted on five advanced baselines for FSRE demonstrate the effectiveness and generalization capabilities of our framework. Ablation studies further confirm the complementary nature of the sentence-anchored and label-anchored contrastive learning in SaCon. Additionally, SaCon showcases strong capacity in domain adaptive and zero-shot settings, highlighting its robustness.

Moving forward, our furture work will focus on advancing universal contrastive pre-training techniques, particularly addressing the many-to-many relations between labels and instances within large-scale pre-training datasets for RE.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable disucssion and constructive feedback. This work was supported by the National Natural Science Foundation of China (U22B2061, U19B2028), the National Key R&D Program of China (2022YFB4300603) and Sichuan Science and Technology Program (2023YFG0151).

References

Bordes, A.; Chopra, S.; and Weston, J. 2014. Question Answering with Subgraph Embeddings. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 615–620.

Bragg, J.; Cohan, A.; Lo, K.; and Beltagy, I. 2021. Flex: Unifying evaluation for few-shot nlp. *Advances in Neural Information Processing Systems*, 34: 15787–15800.

Cetoli, A. 2020. Exploring the zero-shot limit of FewRel. In *Proceedings of the 28th International Conference on Computational Linguistics*, 1447–1451.

Chen, C.-Y.; and Li, C.-T. 2021. ZS-BERT: Towards Zero-Shot Relation Extraction with Attribute Representation Learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3470–3479.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, 1597–1607.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Dong, L.; Yang, N.; Wang, W.; Wei, F.; Liu, X.; Wang, Y.; Gao, J.; Zhou, M.; and Hon, H.-W. 2019. Unified language model pre-training for natural language understanding and generation. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 13063–13075.

Dong, M.; Pan, C.; and Luo, Z. 2021. MapRE: An Effective Semantic Mapping Approach for Low-resource Relation Extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2694–2704.

Gao, T.; Han, X.; Liu, Z.; and Sun, M. 2019a. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 6407–6414.

Gao, T.; Han, X.; Zhu, H.; Liu, Z.; Li, P.; Sun, M.; and Zhou, J. 2019b. FewRel 2.0: Towards More Challenging Few-Shot Relation Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6250–6255.

Han, J.; Cheng, B.; and Lu, W. 2021. Exploring Task Difficulty for Few-Shot Relation Extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2605–2616.

Han, X.; Zhu, H.; Yu, P.; Wang, Z.; Yao, Y.; Liu, Z.; and Sun, M. 2018. FewRel: A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4803–4809.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Levy, O.; Seo, M.; Choi, E.; and Zettlemoyer, L. 2017. Zero-Shot Relation Extraction via Reading Comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, 333–342.

Liu, F.; Lin, H.; Han, X.; Cao, B.; and Sun, L. 2022a. Pre-training to Match for Unified Low-shot Relation Extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5785–5795.

Liu, Y.; Hu, J.; Wan, X.; and Chang, T.-H. 2022b. A Simple yet Effective Relation Information Guided Approach for Few-Shot Relation Extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, 757–763.

Loshchilov, I.; and Hutter, F. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Luan, Y.; He, L.; Ostendorf, M.; and Hajishirzi, H. 2018. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3219–3232.

Peng, H.; Gao, T.; Han, X.; Lin, Y.; Li, P.; Liu, Z.; Sun, M.; and Zhou, J. 2020. Learning from Context or Names? An Empirical Study on Neural Relation Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3661–3672.

Qin, Y.; Lin, Y.; Takanobu, R.; Liu, Z.; Li, P.; Ji, H.; Huang, M.; Sun, M.; and Zhou, J. 2021. ERICA: Improving Entity and Relation Understanding for Pre-trained Language Models via Contrastive Learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3350–3363.

Qu, M.; Gao, T.; Xhonneux, L.-P. A.; and Tang, J. 2020. Few-shot relation extraction via Bayesian meta-learning on relation graphs. In *Proceedings of the 37th International Conference on Machine Learning*, 7867–7876.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Sainz, O.; de Lacalle, O. L.; Labaka, G.; Barrena, A.; and Agirre, E. 2021. Label Verbalization and Entailment for

Effective Zero and Few-Shot Relation Extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1199–1212.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4080–4090.

Soares, L. B.; Fitzgerald, N.; Ling, J.; and Kwiatkowski, T. 2019. Matching the Blanks: Distributional Similarity for Relation Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2895–2905.

Sun, S.; Sun, Q.; Zhou, K.; and Lv, T. 2019. Hierarchical attention prototypical networks for few-shot text classification. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 476–485.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, 776–794. Springer.

Vrandečić, D.; and Krötzsch, M. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10): 78–85.

Wang, C.; Yu, Y.; Ma, W.; Zhang, M.; Chen, C.; Liu, Y.; and Ma, S. 2022. Towards representation alignment and uniformity in collaborative filtering. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1816–1825.

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

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3733–3742.

Xu, C.; Tao, D.; and Xu, C. 2013. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*.

Ye, Z.-X.; and Ling, Z.-H. 2019. Multi-Level Matching and Aggregation Network for Few-Shot Relation Classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2872–2881.

Zhang, P.; and Lu, W. 2022. Better Few-Shot Relation Extraction with Label Prompt Dropout. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 6996–7006. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Zhang, Y.; Jiang, H.; Miura, Y.; Manning, C. D.; and Langlotz, C. P. 2022. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, 2–25. PMLR.

Zhou, G.; Su, J.; Zhang, J.; and Zhang, M. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd annual meeting of the association for computational linguistics*, 427–434.