

Multi-Level Contrastive Distillation for Semi-Supervised Relation Extraction

Huiming Wang^{*1} Guoshun Nan² Wei Lu³

¹Zhejiang University, China ²Beijing University of Posts and Telecommunications, China

³Singapore University of Technology and Design, Singapore

steve.wang@zju.edu.cn nanguo2021@bupt.edu.cn

luwei@sutd.edu.sg

Abstract


Semi-supervised relation extraction (SSRE) aims to extract structured information from unstructured text where limited labeled data and a large amount of unlabeled data are available during learning. Existing methods mainly focus on refining labels for unlabeled instances with a wrapper method (Van Engelen and Hoos, 2020). However, learning high-quality representations on the unlabeled data for relation classification remains a challenging research question, which is largely under-explored in previous efforts. In this paper, we present a novel model that learns token-level, sentence-level, and relation-level representations in a joint and interactive manner, using a multi-level contrastive learning approach under an iterative knowledge distillation framework. Experiments on two public datasets demonstrate the effectiveness of our method under various SSRE settings.

1 Introduction

Relation extraction (RE) aims to detect semantic relations among entities from plain texts. Neural RE models (Zhang et al., 2018; Guo et al., 2019) learn high-dimensional representations from input textual data for better prediction. However, these models are data-hungry, requiring a large amount of labeled instances. Distantly supervised RE methods (Zeng et al., 2015; Quirk and Poon, 2017) were proposed to automatically align texts with relation triples with the help of external knowledge bases (KB). While effective, these approaches heavily rely on the readiness of the KBs. As a result, growing attention is drawn to semi-supervised RE (SSRE; Agichtein and Gravano 2000), in which a large amount of unlabeled samples are incorporated into the learning process, together with a small number of labeled samples.

Figure 1 shows a scenario of SSRE with three RE instances, where the first instance (color orange)

$$\mathcal{D}_l = \{\{x_i, y_i\}, i \in \{1, \dots, n\}\} \quad \mathcal{D}_u = \{\{x_j\}, j \in \{n+1, \dots, n+m\}\}$$



| Sentence | Relation Label |
|--|------------------|
| ● Famous <u>India</u> classical musician <u>Ali Akbar Khan</u> dies. | per:origin |
| ● <u>Trump</u> believes the <u>United States</u> has potential. | Relation Unknown |
| ● He had lived in <u>Cambridge</u> for more than 10 years. | Relation Unknown |

Figure 1: SSRE learns with limited labeled data \mathcal{D}_l and a large amount of unlabeled data \mathcal{D}_u . We show three samples, where the first has a relation label “per:origin” and the other two are unlabeled (unknown relation).

labeled with a relation “per:origin” sampled from the labeled data set \mathcal{D}_l , and the other two (color purple and blue) are sampled from the unlabeled dataset \mathcal{D}_u without any relation labels. Unlike supervised models, a SSRE model will be trained on both labeled and unlabeled instances.

Existing approaches to learning with unlabeled data for relation extraction largely focused on designing modules that can iteratively assign or refine labels for unlabeled instances (Lin et al., 2019; Zhou et al., 2020), following a *wrapper method* (Van Engelen and Hoos, 2020), with representative approaches such as bootstrapping or self-training (Yarowsky, 1995). While effective, such approaches did not explicitly focus on learning appropriate input data representations. It is believed that, however, central to the success of semi-supervised learning is the *cluster assumption* (Chapelle et al., 2006), which says input data points coming from the same cluster belong to the same class¹. How to design an effective approach to learn suitable data representations so that we can bear out such an assumption for the SSRE task remains an open research question.

In this work, we draw inspirations from InfoGraph (Sun et al., 2019) for semi-supervised graph-

^{*} Work done while Huiming Wang visited SUTD

¹There exist other assumptions for semi-supervised learning, such as smoothness, low-density, and manifold assumptions, which can be regarded as special cases of the cluster assumption (Van Engelen and Hoos, 2020).

level representation learning, and introduce a novel *multi-level contrastive learning* (MCL) approach to learn better data representations for SSRE. Our method can be considered as a novel contrastive learning (CL) approach that is between unsupervised (He et al., 2020) and supervised CL (Khosla et al., 2020). The proposed approach can learn high quality token-level, sentence-level, and relation-level representations by explicitly capturing dependencies between different representation spaces. To effectively inject the supervision signal into the input data representation learning process, we further integrate our MCL into an *iterative knowledge distillation* (IKD) framework, which consists of teacher and student networks. With such a joint and interactive learning approach, the proposed MCL-IKD model can yield more informative token-, sentence-, and relation-level representations, where similar data points are grouped together in the representation space and dis-similar ones are pushed apart from each other. This allows clusters over the input data points to be easily formed in the new representation space, facilitating the SSRE task.

Experiments on two public datasets under various SSRE settings show the effectiveness of our model, yielding better representations for classification. Under the same setting, MCL outperforms baselines adapted from popular CL methods including MoCo (He et al., 2020) and SimCLR (Chen et al., 2020a). Also, the encoder component of our semi-supervised method is compatible with various supervised RE models, including sequence-based models (Zhang et al., 2017), graph-based models (Zhang et al., 2018), MetaSRE (Hu et al., 2020) with BERT encoder (Devlin et al., 2019). Our contributions are summarized as follows²:

- We present a novel multi-level contrastive learning method that is able to learn better token-level, sentence-level, and relation-level representations for SSRE.
- To enhance the model’s capability in representation learning from unlabeled data, we further present a novel iterative knowledge distillation method. Such a joint and interactive learning approach is able to yield more informative representations.
- We conduct extensive quantitative and qualitative experiments on several standard benchmark datasets, confirming the superiority of our proposed multi-level contrastive distilla-

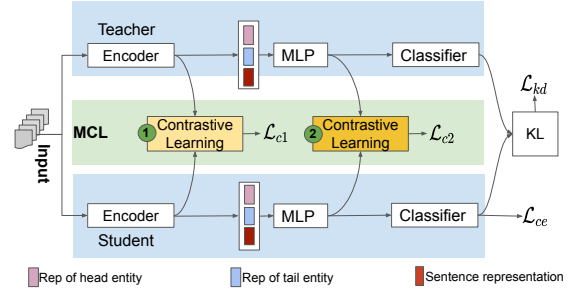


Figure 2: Architecture of our proposed approach, which consists of MCL, IKD, and four losses.

tion method under various SSRE settings.

2 Model Overview

Figure 2 depicts the overall architecture of our MCL-IKD model, which consists of two key ingredients: a multi-level contrastive learning (MCL) module, which is responsible for learning good representations of input data, and an iterative knowledge distillation (IKD) module that facilitates the integration of supervision signal while learning the first module. We first describe components including the encoder and the teacher-student framework.

Encoder: Our method is compatible with various encoders, including Position-aware Recurrent Neural Network (PRNN; Zhang et al. 2017), CGCN (Zhang et al., 2018), and pre-trained BERT.

Teacher-Student: As shown in Figure 2, for PRNN and CGCN encoders³, both teacher and student consist of an encoder and two multi-layer perceptrons (MLPs). We follow Zhang et al. (2018) to construct a relation representation by concatenating the representations of head entity, tail entity, and the sentence. The relation representations of the teacher will be fed into a classifier to generate soft labels (logits), which will be used to guide the student by minimizing the Kullback-Leibler (KL) divergence (Rached et al., 2004) \mathcal{L}_{kd} between the teacher and student:

$$\mathcal{L}_{kd} = \text{KL}(T_o \| S_o) \quad (1)$$

where T_o and S_o refer to the probabilistic output distributions of the teacher and student, respectively. Such a design can be viewed as self-distillation (Yun et al., 2020), e.g., born-again network (Furlanello et al., 2018). Intuitively, the student learns to “follow” the teacher, while learning how to refine the internal data representations. The

³We use MetaSRE as our backbone for BERT encoder, as it is the state-of-the-art approach on BERT. Due to space limitation, implementation details are available in the Appendix.

²Our code is available at *url*.

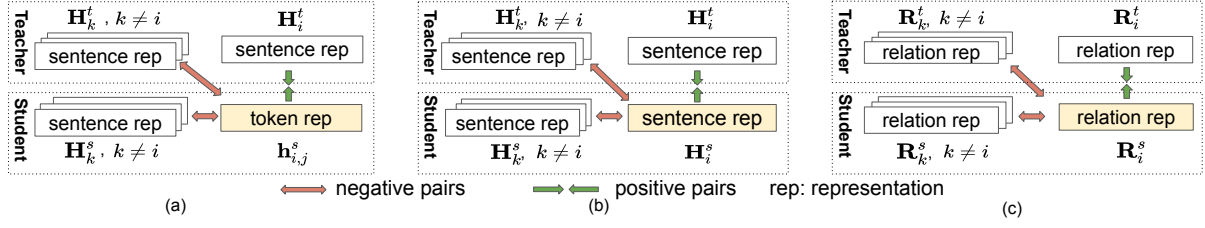


Figure 3: MCL: (a) token-level, (b) sentence-level, and (c) relation-level CL, respectively.

latter is achieved by our introduced MCL objectives, to be discussed next.

3 Multi-Level Contrastive Distillation

The goal of contrastive learning (CL) is to learn representations where a “positive” pair of data points are close to each other in the representation space and points from “negative” pairs are far apart from each other. Our MCL borrows this idea to learn better representations for SSRE, so that the unlabeled data can be better explored without any human interventions. As shown in Figure 3, we contrastively learn token-level, sentence-level, and relation-level representations simultaneously under a KD framework, since a relation representation is constructed by concatenating the representations of head entity, tail entity, and the sentence. During the training process, mutual information (MI; Ghahramani (2006)) is employed to compute the contrastive losses. We maximize the MI of positive pairs and minimize the MI of negative pairs to learn better representations. However, MI estimation is generally intractable for continuous random variables. We instead maximize the lower bound of MI with Jensen-Shannon estimator (Nowozin et al., 2016). Specifically, following InfoGraph, given a pair of representations $\mathbf{u} \in \mathbb{R}^d$ and $\mathbf{v} \in \mathbb{R}^d$, the MI estimator is defined as:

$$\mathcal{T}(\mathbf{u}, \mathbf{v}, x) = \begin{cases} \gamma - \text{softplus}(-\mathbf{u}\mathbf{v}^T) & x = +1 \\ \text{softplus}(-\mathbf{u}\mathbf{v}^T) + \mathbf{u}\mathbf{v}^T - \gamma & x = -1 \end{cases} \quad (2)$$

where $\text{softplus}(x) = \log(1 + e^x)$, x is the indicator for positive (+1) or negative (−1) pairs, and γ is a hyperparameter. Next we detail each level’s CL formulation.

3.1 Token-Level CL

Figure 3 (a) shows the token-level CL procedure, aiming to obtain better token representations generated by the student encoder. We use N to denote the size of a training batch \mathcal{D} . Without loss of generality, we denote all token representations of

the i -th sentence as $\mathbf{h}_i \in \mathbb{R}^{N_i \times d}$ ($i \in [1, N]$), the representation of each token in the i -th sentence as $\mathbf{h}_{i,j} \in \mathbb{R}^d$ ($j \in [1, N_i]$), and the representation of the k -th sentence in the batch as $\mathbf{H}_k \in \mathbb{R}^d$ ($k \in [1, N]$), where N_i is the number of tokens in the sentence. We consider the following quantity which concerns the estimation of the average MI between each token in the i -th sentence and \mathbf{H}_k :

$$\mathcal{I}(\mathbf{h}_i, \mathbf{H}_k, x) = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathcal{T}(\mathbf{h}_{i,j}, \mathbf{H}_k, x) \quad (3)$$

We feed a batch of sequences to the student and teacher networks simultaneously to get contextualized representations. As the representations of the same sentence generated by the two networks are expected to be close to each other, we treat the representation of j -th token $\mathbf{h}_{i,j}^s \in \mathbb{R}^d$ in the i -th sentence in the student networks, and the corresponding sentence representation $\mathbf{H}_i^t \in \mathbb{R}^d$ in the teacher networks as a positive pair. Hence, the aggregated MI score $\mathcal{I}^{(+)}$ for these pairs in the batch can be estimated by:

$$\mathcal{I}^{(+)} = \sum_{i=1}^N \mathcal{I}(\mathbf{h}_i^s, \mathbf{H}_i^t, +1) \quad (4)$$

where $\mathbf{h}_i^s \in \mathbb{R}^{N_i \times d}$ refers to all token representations of i -th sentence in student network.

Meanwhile, we consider $\mathbf{h}_{i,j}^s$, and the other sentence representation in student or teacher networks, as a negative pair, as they are far apart from each other. This leads to the following term for estimating the aggregated MI for these negative pairs in a batch:

$$\mathcal{I}^{(-)} = \sum_{i,k=1; i \neq k}^N \mathcal{I}(\mathbf{h}_i^s, \mathbf{H}_k^s, -1) + \mathcal{I}(\mathbf{h}_i^s, \mathbf{H}_i^t, -1) \quad (5)$$

where $\mathbf{H}_k^s \in \mathbb{R}^d$ denotes the k -th sentence representation in the student networks. The overall MI for the token-level CL can be estimated by:

$$\mathcal{I} = \mathcal{I}^{(+)} - \mathcal{I}^{(-)} \quad (6)$$

3.2 Sentence-Level CL

Figure 3 (b) shows our sentence-level CL procedure, with a goal of improving the sentence representations generated by the student encoder. For a sentence representation \mathbf{H}_i^s ($i \in [1, N]$) in the student networks, we form a positive pair with the corresponding sentence representation \mathbf{H}_i^t ($i \in [1, N]$) in the teacher networks to contrastively distill the knowledge. The MI score $\mathcal{I}'^{(+)}$ for these positive samples in the batch can be estimated by:

$$\mathcal{I}'^{(+)} = \sum_i^N \mathcal{T}(\mathbf{H}_i^s, \mathbf{H}_i^t, +1) \quad (7)$$

For \mathbf{H}_i^s , we construct negative pairs with the representations of other sentences. The MI scores $\mathcal{I}'^{(-)}$ for negative pairs in the batch can be estimated by:

$$\mathcal{I}'^{(-)} = \sum_{i,k=1; i \neq k}^N \mathcal{T}(\mathbf{H}_i^s, \mathbf{H}_k^s, -1) + \mathcal{T}(\mathbf{H}_i^s, \mathbf{H}_k^t, -1) \quad (8)$$

The overall MI for the sentence-level CL in the batch can be computed by:

$$\mathcal{I}' = \mathcal{I}'^{(+)} - \mathcal{I}'^{(-)} \quad (9)$$

Combining such approximations with Equation (6), we can contrastively train the student encoder with the objective \mathcal{L}_{c1} defined as follows:

$$\mathcal{L}_{c1} = -\mathcal{I} - \mathcal{I}' \quad (10)$$

3.3 Relation-Level CL

Figure 3 (c) illustrates the relation-level CL procedure, which aims to improve relation representations outputted by the encoder and the MLP module in the student networks. We denote the relation representation in the i -th sentence in student networks as \mathbf{R}_i^s , where $i \in [1, N]$. For \mathbf{R}_i^s , we construct positive pairs with the relation representation (\mathbf{R}_i^t) of the same sentence in the teacher networks. Such a method allows the relation representations in student to contrastively distill knowledge from the ones in the teacher. The MI $\mathcal{I}_r^{(+)}$ for these positive pairs in the batch \mathcal{D} can be estimated by:

$$\mathcal{I}_r^{(+)} = \sum_{i=1}^N \mathcal{T}(\mathbf{R}_i^s, \mathbf{R}_i^t, +1) \quad (11)$$

Meanwhile, we consider \mathbf{R}_i^s and the relation representations of other sentences in the student and

Algorithm 1: Workflow of our model.

Input: Labeled data $\mathcal{D}_l = \{(x_1, y_1), \dots, (x_n, y_n)\}$, unlabeled data $\mathcal{D}_u = \{x_{n+1}, \dots, x_{n+m}\}$, max iterations I

Output: Student model parameterized by ξ_s^{I-1}

(1) Train a base encoder parameterized by ξ_{base} with the labeled data \mathcal{D}_l ,

(2) Initialize parameters of a teacher model with ξ_{base} ,

(3) Annotate \mathcal{D}_u based on the teacher’s predictions, and construct a new dataset \mathcal{D}^0 by $\mathcal{D}_l + \mathcal{D}_{ul}^0 = \{(x_1, y_1), \dots, (x_n, y_n), (x_{n+1}, y_{n+1}^0), \dots, (x_{m+n}, y_{m+n}^0)\}$

for $i = 0, \dots, I - 1$ **do**

Initialize the parameters (ξ_s^i) of a student S^i with ξ_{base}

Update the parameters ξ_s^i on the dataset \mathcal{D}^i with the teacher \mathcal{M}^i

Re-annotate \mathcal{D}_u as \mathcal{D}_{ul}^{i+1} using the student S^i , and generate a new dataset \mathcal{D}^{i+1} by $\mathcal{D}_l + \mathcal{D}_{ul}^{i+1}$

Make the student S^i the new teacher \mathcal{M}^{i+1}

teacher networks as the negative pairs. The MI $\mathcal{I}_r^{(-)}$ for these negative pairs can be estimated by:

$$\mathcal{I}_r^{(-)} = \sum_{i,k=1; i \neq k}^N \mathcal{T}(\mathbf{R}_i^s, \mathbf{R}_k^s, -1) + \mathcal{T}(\mathbf{R}_i^s, \mathbf{R}_k^t, -1) \quad (12)$$

where \mathbf{R}_k^s and \mathbf{R}_k^t are the relation representations of the k -th sentence in the student and teacher networks, respectively. With the Equation (11) and Equation (12), we can contrastively train the student encoder and MLP layer with the objective \mathcal{L}_{c2} defined as:

$$\mathcal{L}_{c2} = -(\mathcal{I}_r^{(+)} - \mathcal{I}_r^{(-)}) \quad (13)$$

4 Iterative Learning

To effectively inject the supervision signal into the data, we integrate our MCL to an iterative knowledge distillation (IKD) framework, which is presented in Algorithm 1. We refer to KD4RE (Zhang et al., 2020) to train a base encoder for initializations. The unlabeled data \mathcal{D}_u is re-annotated at each iteration to construct a new training set \mathcal{D}^{i+1} , where i indicates the i -th iteration. We put back a student as a new teacher and iterate such a process, allowing the student to better explore the unlabeled data, at the meantime, obtaining higher quality representations by our MCL.

We denote \mathcal{L}_{ce} as the cross-entropy loss defined at the end of the student network. Combining Equation (1) in Section 2, Equation (10), and Equation (13), we arrive at the overall training loss \mathcal{L} :

$$\mathcal{L} = \alpha \mathcal{L}_{kd} + (1 - \alpha) \mathcal{L}_{ce} + \beta \mathcal{L}_{c1} + \lambda \mathcal{L}_{c2} \quad (14)$$

where α , β , and λ are hyperparameters indicating relative significance of each component. The distillation loss \mathcal{L}_{kd} and two contrastive losses \mathcal{L}_{c1} and \mathcal{L}_{c2} facilitate the learning procedure to obtain better representations with our multi-level contrastive distillation mechanism.

5 Experiments

5.1 Datasets and Settings

We conduct experiments on two RE datasets, including the SemEval-2010 Task 8 dataset (SemEval; Hendrickx et al. 2009), and the TAC relation extraction dataset (TACRED; Zhang et al. 2018). SemEval is a standard benchmark dataset for RE systems with 19 relation types. TACRED is a large-scale crowd-sourced RE dataset which is collected from all the prior TAC KBP shared tasks, involving 42 relation types. We follow InfoGraph (Sun et al., 2019) to set γ as $\log(2.0)$. We refer to a previous study (Furlanello et al., 2018) to decrease the loss weight α from 1 to 0 linearly throughout the training stage of a student. The loss weights β and λ are configured as 0.05 and 1.8, respectively⁴.

5.2 Baselines

As our framework is compatible with different encoders, we conduct experiments based on three encoders including PRNN (Zhang et al., 2017), CGCN (Zhang et al., 2018), and BERT (Devlin et al., 2019). We further provide a comprehensive comparison among the following three type of baselines equipped with the same encoders.

Sequence-based Models: To fairly compare with the results reported in the previous work (Lin et al., 2019), all these sequence-based baselines use PRNN (Zhang et al., 2017) as the encoder. 1) KD4RE (Zhang et al., 2020) is also a KD-based learning framework which is originally proposed for supervised RE. 2) Mean-Teacher (Tarvainen and Valpola, 2017) is a consistency training method which was first proposed for semi-supervised image classification. 3) Self-Training (Rosenberg et al., 2005) trains a model on labeled data and then relies on this model to generate pseudo labels for the unlabeled data. 4) RE-Ensemble (Lin et al., 2019) is similar to Self-Training and the main difference is that it generates pseudo labels with outputs from two models. 5) DualRE-Pairwise and DualRE-Pointwise are two variants of DualRE (Lin

et al., 2019), which include a prediction module and a retrieval module.

Graph-based Models: These baselines involve GCN (Kipf and Welling, 2017) that encodes a dependency tree, CGCN (Zhang et al., 2018) that constructs a graph by pruning a dependency tree, and AGGCN (Guo et al., 2019) which uses multi-head attention to build a graph for information aggregation. These baselines also use PRNN as the sequence encoder for fair comparisons.

BERT-based Models: These baselines are equipped with the BERT-Base encoder (Devlin et al., 2019), including Mean-Teacher_{BERT} (Tarvainen and Valpola, 2017), Self-Training_{BERT} (Rosenberg et al., 2005), DuralRE_{BERT} (Lin et al., 2019), MRefG_{BERT} (Li and Qian, 2020), and MetaSRE_{BERT} (Hu et al., 2020). For the first three baselines, we have outlined them in the introduction of Sequence-based models. The fourth baseline MRefG_{BERT} relies on multiple reference-based graphs for information aggregation. MetaSRE_{BERT} is the previous state-of-the-art SSRE model that uses meta learning to improve the quality of pseudo labels.

5.3 Results

We follow the same settings as DualRE (Lin et al., 2019) to split labeled and unlabeled data, i.e., 3%, 10%, and 15% labeled training on TACRED, and 5%, 10%, and 30% on SemEval, and report the average F1 scores with standard deviations.

Comparisons against sequence-based models:

We only use PRNN as an encoder for both student and teacher networks for fair comparisons with these baselines. Table 1 show that our model consistently performs the best under most of the settings on the two datasets, demonstrating the effectiveness of the proposed MCL-IKD for SSRE. We believe the gains mainly stem from the MCL and IKD, which explicitly capture the correlations between difference level of representations in an iterative learning manner. Table 1 shows that our model outperforms KD4RE under various settings on SemEval and TACRED. Compared with another KD-based approach Mean-Teacher, we also achieve 4.10 and 3.82 points improvements on these two datasets. These results confirm our hypothesis that directly applying a KD-based model for SSRE may not achieve good results. We also show comparisons with the iterative learning methods under var-

⁴Detailed hyper-parameters are attached in the Appendix.

| Encoder | Methods / % Labeled Data | SemEval | | | TACRED | | |
|---------|---------------------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| | | 5% | 10% | 30% | 3% | 10% | 15% |
| PRNN | PRNN* | 55.34±1.08 | 62.63±1.42 | 69.02±1.01 | 39.11±1.92 | 52.23±1.20 | 54.55±1.92 |
| | KD4RE* | 56.51±1.20 | 64.60±0.67 | 71.62±0.21 | 46.89±1.31 | 55.13±0.87 | 56.58±0.42 |
| | Mean-Teacher* | 51.51±3.58 | 61.36±0.75 | 69.24±0.56 | 46.74±1.70 | 53.94±0.91 | 55.13±0.05 |
| | Self-Training* | 56.30±0.96 | 63.79±0.28 | 70.74±0.58 | 43.86±1.26 | 54.71±0.09 | 57.31±0.47 |
| | RE-Ensemble* | 58.63±0.62 | 64.83±0.61 | 71.69±0.47 | 44.62±0.39 | 55.54±0.29 | 57.72±0.38 |
| | DualRE-Pairwise* | 61.51±0.56 | 66.00±0.48 | 72.36±0.60 | 43.55±0.67 | 55.09±0.25 | 57.30±0.81 |
| | DualRE-Pointwise* | 60.43±1.67 | 66.03±1.00 | 72.36±0.35 | 44.73±0.66 | 56.65±0.42 | 58.58±0.69 |
| | Ours (PRNN) | 63.13±0.52 | 68.12±0.32 | 73.34±0.27 | 46.92±0.90 | 57.20±0.76 | 58.95±0.35 |
| CGCN | GCN [†] | N/A | N/A | 49.04±2.01 | 33.39±2.67 | 54.02±1.50 | 57.11±0.73 |
| | CGCN [†] | N/A | N/A | 39.45±2.52 | 22.01±2.89 | 54.08±1.39 | 55.89±1.15 |
| | AGGCN [†] | N/A | N/A | 68.92±0.89 | 47.80±0.73 | 55.88±0.59 | 56.41±0.42 |
| | Ours (CGCN) | N/A | N/A | 74.11±0.45 | 48.20±0.72 | 56.49±0.57 | 59.37±0.39 |
| BERT | BERT* | 72.71±1.24 | 73.93±0.99 | 80.55±0.87 | 41.11±3.88 | 54.23±1.67 | 56.55±0.82 |
| | Mean-Teacher _{BERT} * | 69.05±3.89 | 73.37±1.42 | 80.61±0.81 | 44.34±1.78 | 53.08±1.01 | 53.79±1.38 |
| | Self-Training _{BERT} * | 71.34±1.68 | 74.25±1.10 | 81.71±0.79 | 42.11±1.04 | 54.17±0.53 | 56.52±0.40 |
| | DualRE _{BERT} * | 74.35±1.76 | 77.13±1.10 | 82.88±0.67 | 43.06±1.73 | 56.03±0.55 | 57.99±0.67 |
| | MRefG _{BERT} * | 75.48±1.34 | 77.96±0.90 | 83.24±0.71 | 43.81±1.44 | 55.42±1.40 | 58.21±0.71 |
| | MetaSRE _{BERT} * | 78.33±0.92 | 80.09±0.78 | 84.81±0.44 | 46.16±1.02 | 56.95±0.34 | 58.94±0.36 |
| | Ours (BERT) | 80.02±0.41 | 82.13±0.39 | 86.42±0.27 | 49.00±0.76 | 57.39±0.45 | 60.12±0.32 |

Table 1: F1 (%) comparisons on the SemEval and TACRED datasets under various semi-supervised settings, e.g., with only 5% of the training instances used as labeled data. The results with * are directly copied from the existing works DualRE (Lin et al., 2019) or MetaSRE (Hu et al., 2020), and the ones with † are produced by us as there are no previous results reported for such settings. All our models are significant over the baselines at $p < 0.01$.

ious settings on the two datasets. These methods are similar to our IKD, which attempts to refine the model by iteratively annotating the unlabeled data. The comparisons show that our MCL-IKD can better explore knowledge from unlabeled data, and justify our claim that the proposed MCL benefits the IKD for SSRE.

Comparisons against graph-based models: To align with these graph-based baselines, we employ PRNN as a sequence encoder and CGCN as a graph encoder on the two datasets. With very limited number of training samples, e.g., 3% on TACRED, our model outperforms GCN by 14.81 points in terms of F1 score. These results further confirm the superiority of our model in handling SSRE when there are very few labeled data available. We observe that our model and all graph-based baselines can hardly converge with a GCN graph encoder under the setting of 5% and 10% labeled training data on SemEval, and hence the results are indicated as “NA”. We hypothesis that the graph-based encoders required more data to train parameters. Under the 3% setting on TACRED, our performance is better than AGGCN, suggesting that MCL under IKD can also benefit a graph encoder for the SSRE.

Comparisons against BERT-based models: As shown in Table 1, our model still consistently outperforms all baselines on the two datasets. For example, compared with MetaSRE_{BERT}, our method achieves 2.04 and 1.61 points gains on 10% and

30% labeled data respectively on the SemEval dataset. Note that we integrate our MCL in MetaSRE_{BERT} and train the model under our IKD framework⁵, as MetaSRE_{BERT} is the previous state-of-the-art method. These results suggest that distilling from BERT representations with our MCL and IKD is still able to yield better relation representations for SSRE.

Comparisons against NERO: As shown in Table 2, we also compare our MCL-IKD against the state-of-the-art weakly-supervised model NERO. NERO requires first mine rules from the data (soft) and then manually refine them by annotators (hard). However, refining rules on a large dataset (TACRED) is non-trivial and mining informative rules on a small dataset (SemEval) is challenging. Our model outperforms NERO on SemEval and achieves comparative results on TACRED without using any pre-defined rules. These results further confirm our hypothesis that the MCL-IKD is able to learn high-quality representations solely from data even the training resources are limited.

5.4 Ablation Study

Table 3 shows the ablation study on SemEval, with the PRNN encoder and 10% of labeled data.

Effect of MCL: We turn off one of the CL in MCL each time and observe that removing each component leads to a performance drop by 1.93

⁵Details are available in Appendix due to space limitation.

| Methods / Dataset | TACRED | SemEval |
|--|-----------------|-----------------|
| NERO _{hard} (Zhou et al., 2020) | 42.9±1.4 | 58.6±0.6 |
| NERO _{soft} (Zhou et al., 2020) | 45.3±1.0 | 54.9±0.6 |
| NERO (Zhou et al., 2020) | 51.3±0.6 | 60.5±0.7 |
| Ours (PRNN) | 50.1±0.4 | 61.3±0.4 |

Table 2: F1 (%) comparisons with NERO variants using the exact same data splits as used by NERO. NERO_{hard} only performs hard-matching on rules, while NERO_{soft} uses the soft rule matcher.

| Model | F1 |
|--|-------|
| Full model | 69.10 |
| w/o Token-level and Sentence-level CL | 67.17 |
| w/o Relation-level CL | 67.59 |
| w/o Multi-level Contrastive Learning (MCL) | 66.14 |
| w/o Iterative Knowledge Distillation (IKD) | 66.92 |
| w/o MCL and IKD | 64.57 |
| w 1 iteration | 68.06 |
| w/o Unlabeled Data | 63.74 |

Table 3: Ablation study on dev set of SemEval with 10% labeled data and the PRNN encoder.

and 1.51 F1 points, respectively. Removal of MCL leads to a 2.96 points F1 drop. These prove that each CL module contributes to the performance. Combining them together obtains more gains.

Effect of IKD: We wipe off the IKD module and observe that the F1 drops from 66.14 to 64.57. This validates our hypothesis that a well-designed KD approach is able to better utilize the soft-labels that contain rich semantic relation information, allowing the student network learns better representations by contrastive knowledge distillation.

Effect of iterative learning: Our model obtains best performance with 2 iterations. Using only 1 iteration leads to a 1.04 points F1 drop. This shows that our IKD can learn better representation by incrementally exploring unlabeled data. However, F1 will drop 0.31 when the number of iterations increases from 2 to 3. We hypothesized that more iterations may lead to overfitting. Similar observations were discussed in (Furlanello et al., 2018).

Effect of the unlabeled data: We observe that F1 drops significantly from 69.10 to 63.47 without using unlabeled data. This shows that unlabeled data plays an important role for SSRE and exploring high quality representations from unlabeled corpus is able to benefit SSRE. Such an observation supports our motivation raised at the beginning.

5.5 Discussion

Can MCL-IKD outperform existing CL methods? To answer this question, we adapt three very recent CL methods from visual tasks to SSRE,

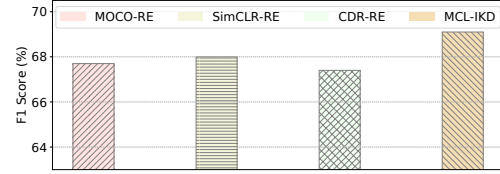


Figure 4: Comparisons with recent CL methods (SemEval, 10% training instances).

| Model | Time (H) | F1 | Model | Time (H) | F1 |
|-------|----------|-------|-----------|----------|-------|
| MCL | 2.08 | 72.91 | MCL-IDK | 2.10 | 74.11 |
| KD4RE | 2.07 | 71.62 | MCL-AGGCN | 0.83 | 69.91 |

Table 4: Training time comparisons measured in hours.

including 1) MoCo-RE adapted from MoCo (He et al., 2020), which is originally proposed for image classification by building a large dictionary for CL. MoCo-RE refers to the original MoCo to build a dictionary for entities from the overall training corpus. 2) SimCLR-RE adapted from SimCLR (Chen et al., 2020a), which uses a learnable non-linear transformation between two representations. SimCLR-RE follows a similar idea to introduce a network between relation representations. For fair comparisons, SimCLR-RE does not adapt the data augmentation method used in SimCLR, as it introduces additional instances. 3) CRD-RE adapted from CRD (Tian et al., 2019) that proposes a contrastive loss based on KD for image classification. CRD-RE uses the same loss to CRD for the relation representations. Figure 4 shows the comparison results on SemEval with 10% training instances. Our MCL-IKD outperforms the above three baselines by average 1.72, 1.09, and 1.35 points respectively based on 5 runs of experiments, showing the superiority of our method for SSRE.

Quantitative analysis to gains and computation cost:

We compare our method with two baselines for this purpose. Table 4 shows the comparisons of training time for MCL and IKD on SemEval. To evaluate IKD, we compare our model with a baseline that integrates a variant of MCL to AGGCN, termed as MCL-AGGCN. Our model requires 2.5x training time, with a 4.20 points gain under 30% labeled data. Under the same KD framework, our MCL requires very little extra time compared with KD4RE, with 1.29 points improvement with 30% labeled data. We conclude that MCL can bring large gain with almost negligible additional running time, and the IKD framework requires more computational costs. The reason why we use the

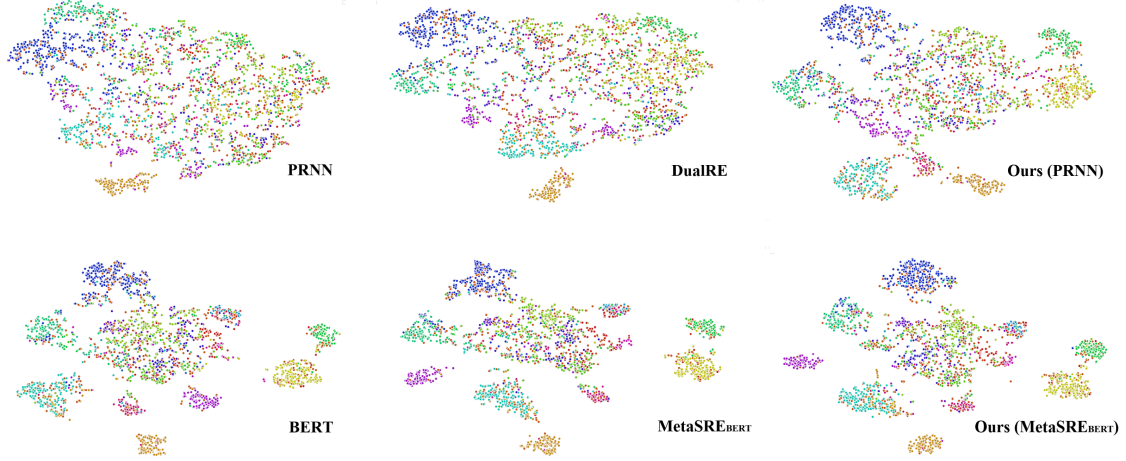


Figure 5: The t -SNE visualization of relation representations with 10% training data on SemEval. Our MCL-IKD models, which are the rightmost two, yield better representation clustering than corresponding baselines.

IKD is that it can effectively inject the supervision signal into the data while learning with MCL, yielding better representations on the unlabeled data.

5.6 Visualization

To illustrate why our proposed MCL-IKD model performs better with a multi-level contrastive learning method, we leverage t -SNE (Maaten and Hinton, 2008) to project the relation representations to the 2D space. Figure 5 depicts the visualizations of relation representations generated by 6 models trained on the development set of the SemEval with 19 relation types and 10% labeled data. The comparisons show that the ones outputted by the MCL-IKD model are better clustered. This indicates that our method can indeed learn better data representations that form good clusters for SSRE, and also explains the success of our approach under the *cluster assumption* used for semi-supervised learning, as discussed at the beginning.

6 Related Work

Semi-supervised RE: There are plenty of studies for supervised RE (Zeng et al., 2014; Miwa and Bansal, 2016; Han et al., 2018; Guo et al., 2019) and semi-supervised learning (Qiao et al., 2018; Li et al., 2019; Ouali et al., 2020; Chen et al., 2020b). Earlier efforts (Brin, 1998; Agichtein and Gravano, 2000) of SSRE relied on pattern-relation duality to iteratively exploit the labeling pattern from a small corpus. Recently, DualRE and NERO (Lin et al., 2019; Zhou et al., 2020) advance these two methods by iterative self-training and refined labeling rules with human interventions. Recent MetaSRE (Hu et al., 2020) employs meta learning for SSRE.

Unlike these studies, our MCL-IKD learn representations based on a joint and interactive learning method without relying on any human efforts.

Knowledge Distillation: KD enables a student to learn how to refine the representations from a teacher (Hinton et al., 2015; Frosst and Hinton, 2017; Tarvainen and Valpola, 2017). For example, KD4RE (Zhang et al., 2020) exploits well-informed soft labels for supervised RE. Born-Again Networks (Furlanello et al., 2018) stacks multiple students to achieve better results. Closest to our work, CRD (Tian et al., 2019) designs a contrastive loss for image classification. Different from CRD, we further develop an MCL method for SSRE.

Contrastive Learning: CL method learns representations by contrasting positive pairs and negative pairs (Tian et al., 2019; He et al., 2020; Chen et al., 2020a; Klein and Nabi, 2020; Khosla et al., 2020). Some prior works (Hjelm et al., 2019; Velickovic et al., 2019; Sun et al., 2019; Wei et al., 2021) use mutual information (Bell and Sejnowski, 1995) to estimate the scores between latent representations. Inspired by InfoGraph (Sun et al., 2019) for learning graph representation, our MCL learns token-level, sentence-level, and relation-level representations simultaneously for SSRE under an iterative knowledge distillation framework.

7 Conclusion

This paper presents MCL-IKD, a novel model for SSRE, which learns token-level, sentence-level, and relation-level representations in a joint and interactive manner based on multi-level contrastive learning within an iterative knowledge distillation

framework. Experiments on three benchmarks validate the effectiveness of our method. While the model is specifically designed for the SSRE task, our conceptual ideas underlying the design of the approach are general. We believe such ideas can also be applied to other semi-supervised tasks within NLP, such as sentiment analysis, text classification, and so on. We leave such directions for future research.

References

- Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of Digital Library*.
- Anthony J Bell and Terrence J Sejnowski. 1995. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159.
- Sergey Brin. 1998. Extracting patterns and relations from the world wide web. In *Proceedings of Web and Databases*.
- Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. 2006. *Semi-Supervised Learning*. The MIT Press.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *Proceedings of ICML*.
- Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. 2020b. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.
- Nicholas Frosst and Geoffrey Hinton. 2017. Distilling a neural network into a soft decision tree. *arXiv preprint arXiv:1711.09784*.
- Tommaso Furlanello, Zachary Lipton, Michael Tschanen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks. In *Proceedings of ICML*.
- Zoubin Ghahramani. 2006. Information theory. *Encyclopedia of Cognitive Science*.
- Zhijiang Guo, Yan Zhang, and Wei Lu. 2019. Attention guided graph convolutional networks for relation extraction. In *Proceedings of ACL*.
- Xu Han, Pengfei Yu, Zhiyuan Liu, Maosong Sun, and Peng Li. 2018. Hierarchical relation extraction with coarse-to-fine grained attention. In *Proceedings of EMNLP*.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of CVPR*.
- Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. *SEW-2009 Semantic Evaluations: Recent Achievements and Future Directions*, page 94.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2019. Learning deep representations by mutual information estimation and maximization. In *Proceedings of ICLR*.
- Xuming Hu, Fukun Ma, Chenyao Liu, Chenwei Zhang, Lijie Wen, and Philip S Yu. 2020. Semi-supervised relation extraction via incremental meta self-training. *arXiv preprint arXiv:2010.16410*.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Proceedings of NeurIPS*, 33.
- Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of ICLR*.
- Tassilo Klein and Moin Nabi. 2020. Contrastive self-supervised learning for commonsense reasoning. In *Proceedings of ACL*.
- Wanli Li and Tieyun Qian. 2020. Exploit multiple reference graphs for semi-supervised relation extraction. *arXiv preprint arXiv:2010.11383*.
- Xinzhe Li, Qianru Sun, Yaoyao Liu, Qin Zhou, Shibao Zheng, Tat-Seng Chua, and Bernt Schiele. 2019. Learning to self-train for semi-supervised few-shot classification. In *Proceedings of NeurIPS*.
- Hongtao Lin, Jun Yan, Meng Qu, and Xiang Ren. 2019. Learning dual retrieval module for semi-supervised relation extraction. In *Proceedings of WWW*.
- Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.
- Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of ACL*.
- Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. 2016. f-gan: Training generative neural samplers using variational divergence minimization. In *Proceedings of NeurIPS*.

- Yassine Ouali, Céline Hudelot, and Myriam Tami. 2020. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings. of CVPR*.
- Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. 2018. Deep co-training for semi-supervised image recognition. In *Proceedings. of ECCV*.
- Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings. of EACL*.
- Ziad Rached, Fady Alajaji, and L Lorne Campbell. 2004. The kullback-leibler divergence rate between markov sources. *IEEE Transactions on Information Theory*, 50(5):917–921.
- Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. 2005. Semi-supervised self-training of object detection models. In *Proceedings of the Seventh IEEE Workshops on Application of Computer Vision (WACV/MOTION’05)-Volume 1-Volume 01*, pages 29–36.
- Fan-Yun Sun, Jordan Hoffman, Vikas Verma, and Jian Tang. 2019. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In *Proceedings. of ICLR*.
- Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings. of NeurIPS*.
- Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2019. Contrastive representation distillation. In *Proceedings. of ICLR*.
- Jesper E Van Engelen and Holger H Hoos. 2020. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440.
- Petar Velickovic, William Fedus, William L Hamilton, Pietro Lio, Yoshua Bengio, and R Devon Hjelm. 2019. Deep graph infomax. In *Proceedings. of ICLR*.
- Xiangpeng Wei, Yue Hu, Rongxiang Weng, Luxi Xing, Heng Yu, and Weihua Luo. 2021. On learning universal representations across languages. In *Proceedings. of ICLR*.
- David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings. of ACL*.
- Sukmin Yun, Jongjin Park, Kimin Lee, and Jinwoo Shin. 2020. Regularizing class-wise predictions via self-knowledge distillation. In *Proceedings. of CVPR*.
- Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings. of EMNLP*.
- Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jian Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings. of COLING*.
- Yuhao Zhang, Peng Qi, and Christopher D Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings. of EMNLP*.
- Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings. of EMNLP*.
- Zhenyu Zhang, Xiaobo Shu, Bowen Yu, Tingwen Liu, Jiapeng Zhao, Quangan Li, and Li Guo. 2020. Distilling knowledge from well-informed soft labels for neural relation extraction. In *Proceedings. of AAAI*.
- Wenxuan Zhou, Hongtao Lin, Bill Yuchen Lin, Ziqi Wang, Junyi Du, Leonardo Neves, and Xiang Ren. 2020. Nero: A neural rule grounding framework for label-efficient relation extraction. In *Proceedings. of WWW*.