A Multi-Level Supervised Contrastive Learning Framework for Low-Resource Natural Language Inference

Shu'ang Li, Xuming Hu, Li Lin, Aiwei Liu, Lijie Wen, and Philip S. Yu, *Life Fellow, IEEE*

***Abstract*—Natural Language Inference (NLI) is a growingly essential task in natural language understanding, which requires inferring the relationship between the sentence pairs (premise and hypothesis). Recently, low-resource natural language inference has gained increasing attention, due to significant savings in manual annotation costs and a better fit with real-world scenarios. Existing works fail to characterize discriminative representations between different classes with limited training data, which may cause faults in label prediction. Here we propose a multi-level supervised contrastive learning framework named MultiSCL for low-resource natural language inference. MultiSCL leverages a sentence-level and pair-level contrastive learning objective to discriminate between different classes of sentence pairs by bringing those in one class together and pushing away those in different classes. MultiSCL adopts a data augmentation module that generates different views for input samples to better learn the latent representation. The pair-level representation is obtained from a cross attention module. We conduct extensive experiments on two public NLI datasets in low-resource settings, and the accuracy of MultiSCL exceeds other models by 1.8%, 3.1% and 4.1% on SNLI, MNLI and Sick with 5 instances per label respectively. Moreover, our method outperforms the previous state-of-the-art method on cross-domain tasks of text classification.**

***Index Terms*—Natural language inference, contrastive learning, low-resource, multi-level.**

I. Introduction

NATURAL Language Inference (NLI), also known as Recognizing Textual Entailment (RTE), is a key topic in the research field of natural language understanding [1], [2], and could support tasks such as question answering, reading comprehension, document summarization and relation extraction [3], [4], [5], [6]. In NLI scenarios, the model is given a pair of sentences, namely **premise** and **hypothesis**, and asked to infer the relationship between them from a set of relationships, including `entailment`, `contradiction` and `neutral`. Several concrete examples are illustrated in Table I.

TABLE I
EXAMPLES OF THE RELATIONSHIP BETWEEN **PREMISE** AND **HYPOTHESIS**: **E**NTAILMENT, **C**ONTRADICTION, AND **N**EUTRAL

Premise	Two men on bicycles competing in a race.	
Hypothesis	People are riding bikes.	E
	Men are riding bicycles on the streets.	C
	A few people are catching fish.	N

Large annotated datasets, such as SNLI [7] and MultiNLI datasets [8] have been available in recent years, making it possible to apply sophisticated deep learning models. These neural network models require a large number of training parameters to achieve good results in NLI [9], [10]. However, large-scale datasets are obtained from a large number of manual annotations and have a high annotation cost. Therefore, natural language inference for low-resource scenarios has gained more widespread attention in recent years. Compared with traditional task scenarios, NLI in low-resource scenarios focuses on using a small amount of manually annotated data to achieve similar results as the full amount of data. This can save a large number of manual annotation costs and is more in line with realistic application scenarios, thus with high research value and practical application value [11].

Recent work has shown advantages of generative classifiers in term of low-resource and robustness. Ding et al. [12] propose a generative classifier that defines the conditional probabilities assumed given the premises and labels and has better performance in very few labeled data settings. Liu et al. [13] propose a multi-task deep neural network for learning semantic representations across multiple natural language understanding tasks to enhance the semantic representation in low-resource scenarios. The network not only utilizes a large amount of cross-task data but also benefits from regularization effects to learn more general representations that can be adapted to NLI in fewer sample scenarios. However, these methods only use the feature of the sentence pair itself to predict the class, without considering the comparison between the sentence pairs in different classes. When the training instances per class are small, it is difficult for these methods to distinguish the true class of similar sentences.

Manuscript received 17 May 2022; revised 17 December 2022; accepted 24 April 2023. Date of publication 27 April 2023; date of current version 8 May 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFB1704003, in part by the National Nature Science Foundation of China under Grants 71690231 and 62021002, in part by the N S F under Grants III-1763325, III-1909323, III-2106758, and SaTC-1930941, and in part by the Tsinghua BNRist and Beijing Key Laboratory of Industrial Bigdata System and Application. The associate editor coordinating the review of this manuscript and approving it for publication was Jing Huang. *(Corresponding author: Lijie Wen.)*

Shu'ang Li, Xuming Hu, Li Lin, Aiwei Liu, and Lijie Wen are with the School of Software, Tsinghua University, Beijing 100190, China (e-mail: lisa18@mails.tsinghua.edu.cn; hxm19@mails.tsinghua.edu.cn; lin-l16@mails.tsinghua.edu.cn; liuaw20@mails.tsinghua.edu.cn; wenlj@tsinghua.edu.cn).

Philip S. Yu is with the Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607 USA (e-mail: psyu@uic.edu).

Digital Object Identifier 10.1109/TASLP.2023.3270771

2329-9290 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: University of Illinois at Chicago Library. Downloaded on May 13,2023 at 15:49:11 UTC from IEEE Xplore. Restrictions apply.

Because these methods can't well characterize discriminative representations between different classes with limited training data, which may cause faults in label prediction.

Many recent works explored using contrastive learning to tackle this problem. Contrastive learning is a popular technique in computer vision area [14], [15], [16] and the core idea is to learn a function that maps positive pairs closer together in the embedding space, while pushing apart negative pairs. A contrastive objective is used by [17] to fine-tune pre-trained language models to obtain sentence embeddings with the relationship of sentences in NLI. The model achieved state-of-the-art performance in sentence similarity tasks. Yan et al. [18] propose a simple but effective training objective based on contrastive learning. It mitigates the collapse of BERT-derived representations and transfers them to downstream tasks. However, these approaches can't distinguish well between the representation of sentence pairs in different classes.

In our previous work [19], we propose a pair-level supervised contrastive learning approach (PairSCL), which obtains new state-of-the-art performance in NLI. However, this method cannot be adapted to low-resource settings for the reason of limited discriminative ability of sentence pairs in very few sample scenarios. Therefore, in this paper, we comprehensively investigate the potential of contrastive learning in low-resource NLI. Based on our analysis of PairSCL that contrastive learning can help discriminate the class of sentence pairs, we propose a multi-level supervised contrastive learning framework named MultiSCL for low-resource NLI. In addition to pair-level contrastive learning, MultiSCL leverages the sentence-level contrastive learning objective to characterize the latent embeddings of sentences in semantic space. Furthermore, to better learn the semantic representation, we adopt a data augmentation module that generates different views for input sentences with sentence-level supervised contrastive learning by regarding the contradiction pairs as negatives, and entailment pairs as positives. The pair-level representation can perceive the class information of sentence pairs and is obtained from the Cross Attention module which captures the relevance and characterizes the relationship between the sentence pair. Then we adopt contrastive learning to differentiate the pair-level representation by capturing the similarity between pairs in one class and contrasting them with pairs in other classes.

For example, the entailment pair (P_1, H_1) and contradiction pair (P_2, H_2) are from Table I (P_1: Two men on bicycles competing in a race. H_1: People are riding bikes. P_2: Two men on bicycles competing in a race. H_2: Men are riding bicycles on the streets.). For sentence-level contrastive learning, we take advantage of the fact that entailment pairs can be naturally used as positives and the contradiction pairs can be regarded as negatives. We consider H_1 as the positive set for P_1 and H_2 as the negative set for P_2. In this way, the encoder can capture the semantic representation of the sentences more accurately. For pair-level contrastive learning, our model regards the pair (P_2, H_2) as the negative set for the pair (P_1, H_1) with the representation obtained from Cross Attention module to distinguish the pairs from different classes.

Our contributions can be summarized as follows:

- We propose a novel multi-level supervised contrastive learning framework named MultiSCL for low-resource NLI. It applies the sentence-level and pair-level contrastive learning to learn the discriminative representation with limited labeled training data.
- We adopt a data augmentation module to generate the different views for input samples. We explore various effective text augmentation strategies for contrastive learning and analyze their effects on low-resource NLI.
- We conduct extensive experiments on two public NLI datasets in low-resource settings, and the accuracy of MultiSCL exceeds other models by 1.8%, 3.1% and 4.1% on SNLI, MultiNLI and Sick with 5 instances per label respectively. Moreover, our method outperforms the previous state-of-the-art method on cross-domain tasks of text classification.

This paper is substantially an extended version of our previous paper [19] that is published at ICASSP 2022. Compared to the previous version, we make heavy extensions as follows: (1) By adding the sentence-level contrastive learning objective, we propose a new multi-level supervised contrastive learning framework called MultiSCL for low-resource NLI. (2) We adopt a data augmentation module to generate the views for input sentences and explore various effective text augmentation strategies. (3) We conduct extensive experiments on NLI datasets in low-resource scenarios. We conduct experiments on cross-domain datasets to validate the transfer capability of our model.

The structure of this paper is as follows. In Section II, we review the related work to natural language inference and contrastive learning. Section III introduces the architecture of our framework. Section IV presents experimental design details and Section V reports our experimental results and analysis. Finally, in Section VI, we conclude this paper and present some future work.

II. Related Works

A. Natural Language Inference

Early methods for NLI mainly relied on conventional, feature-based methods trained from small-scale datasets [3], [20]. The release of large datasets, such as SNLI [7] and MultiNLI [8], made neural network methods feasible. Such methods can be roughly categorized into two classes: sentence embedding bottleneck methods which first encode the two sentences as vectors and then feed them into a classifier for classification [21], [22], [23], and more general methods which usually involve interactions while encoding the two sentences in the pair [10], [24], [25]. [26] enables the use of various kinds of external knowledge bases to retrieve information related to **premise** and **hypothesis**. Wang et al. [27] propose a novel Knowledge Graph-enhanced NLI (KGNLI) model to leverage the usage of background knowledge stored in knowledge graphs in the field of NLI.

Recently, large-scale pre-trained language representation models such as BERT [28], GPT [29], BART [30], etc., have

Authorized licensed use limited to: University of Illinois at Chicago Library. Downloaded on May 13,2023 at 15:49:11 UTC from IEEE Xplore. Restrictions apply.

achieved dominating performance in NLI. These neural network models have a large number of training parameters to achieve good results in NLI. However, large-scale datasets are obtained from a large number of manual annotations and have a high annotation cost. Therefore, NLI for low-resource scenarios has gained more widespread attention in recent years.

Ding et al. [12] propose GenNLI, a generative classifier for NLI tasks. The model defines conditional probabilities assumed given premises and labels, parameterizing the distribution using a sequence-to-sequence model with attention [31] and a replication mechanism [32]. They explore training objectives for discriminative fine-tuning of the generative classifier, comparing several classical discriminative criteria. Liu et al. [13] propose a multi-task deep neural network (MT-DNN) for learning semantic representations across multiple natural language understanding tasks. MT-DNN not only utilizes a large amount of cross-task data but also benefits from regularization effects to learn more general representations that can be adapted to natural language reasoning in very few sample scenarios. Schick et al. [33] introduce Pattern Exploiting Training concerning a partially pre-trained language model using task descriptions in natural language [34]. They reformulate a small amount of labeled data into fill-in-the-blank phrases to help the language model understand the given task.

The above methods only use the feature of the sentence pair itself to predict the class, without considering the comparison between the sentence pairs in different classes. In our work, we propose a multi-level contrastive learning framework named MultiSCL for low-resource NLI. MultiSCL leverages a sentence-level and pair-level contrastive learning objective to learn discriminative representations between different classes.

B. Contrastive Learning

Contrastive learning has shown promising results in computer vision area in an unsupervised/self-supervised way [14], [15]. The key idea of contrastive learning is: first create augmentations of original examples, then learn representations by predicting whether two augmented examples are from the same original data example or not. Dating back to [35], these approaches learn representations by contrasting positive pairs against negative pairs. Along these lines, Dosovitskiy et al. [36] propose to treat each instance as a class represented by a feature vector (in a parametric form). Wu et al. [37] propose to use a memory bank to store the instance class representation vector, which was an approach adopted and extended in several recent papers [38], [39]. He et al. [14] propose Momentum Contrast (MoCo) by building a dynamic dictionary with a queue and a moving-averaged encoder and showed state-of-the-art results. Chen et al. [15] propose a simple framework for contrastive learning to learn visual representations without specialized architectures or a memory bank.

Some works use contrastive learning to solve Natural Language Processing (NLP) tasks. Qin et al. [40] propose a novel contrastive learning framework in the pre-training phase to obtain a deeper understanding of the entities and their relations in text. Giorgi et al. [41] propose a self-supervised method for minimizing sentence embeddings of textual segments randomly sampled from nearby in the same document and obtained state-of-the-art performance on SentEval [42]. Gunel et al. [43] propose a supervised contrastive learning (SCL) objective which uses SCL loss combined with cross-entropy loss for the fine-tuning stage. The proposed model shows improved performance on multiple datasets of the GLUE benchmark [44] in both the high-data and low-data regimes. Yan et al. [18] explore a simple but effective sentence-level training objective with various effective text augmentation strategies to generate views for contrastive learning. Suresh et al. [45] incorporate inter-class relationships into a supervised contrastive loss by differentiating the weights between different negative samples for fine-grained text classification. Li et al. [46] adopt supervised contrastive pre-training to capture both implicit and explicit sentiment orientation towards aspects by aligning the representation of implicit sentiment expressions to those with the same label for aspect-based sentiment analysis. Wu et al. [47] propose a new framework, combining word-level masked language modeling objectives with sentence-level contrastive learning objective to pre-train a language model. Zhang et al. [48] propose an instance discrimination-based approach aiming to bridge semantic entailment and contradiction understanding with high-level categorical concept encoding. Wang et al. [49] propose Contrastive Learning with semantIc Negative Examples (CLINE), which constructs semantic negative examples unsupervised to improve the robustness under semantically adversarial attacking. By comparing with similar and opposite semantic examples, the model can effectively perceive the semantic changes caused by small perturbations. Li et al. [19] propose a pair-level supervised contrastive learning approach. The pair-level representation is obtained by Cross Attention module which can capture the relevance and well characterize the relationship between the sentence pair.

However, the above methods can't learn effective semantic representations in low-resource scenarios. In our work, we will focus on the use of multi-level contrastive learning for low-resource NLI.

III. Approach

In this section, we describe our approach MultiSCL. The overall architecture of the model is illustrated in the left part of Fig. 1. MultiSCL comprises the following four major components: a data augmentation module that generates different views for input samples; an encoder that computes sentence-level representations of **premise** and **hypothesis**; a cross attention module to obtain the pair-level representation of the sentence pair and a joint-training layer including the sentence-level and pair-level contrastive learning term and the cross-entropy term.

A. Data Augmentation Module

In low-resource NLI scenarios, it's a challenge for a model to learn discriminative representations and infer the relationship between sentences. Therefore, we adopt a data augmentation module to generate different views of sentences to enhance the semantic understanding and inference capability of the model.

Authorized licensed use limited to: University of Illinois at Chicago Library. Downloaded on May 13,2023 at 15:49:11 UTC from IEEE Xplore. Restrictions apply.

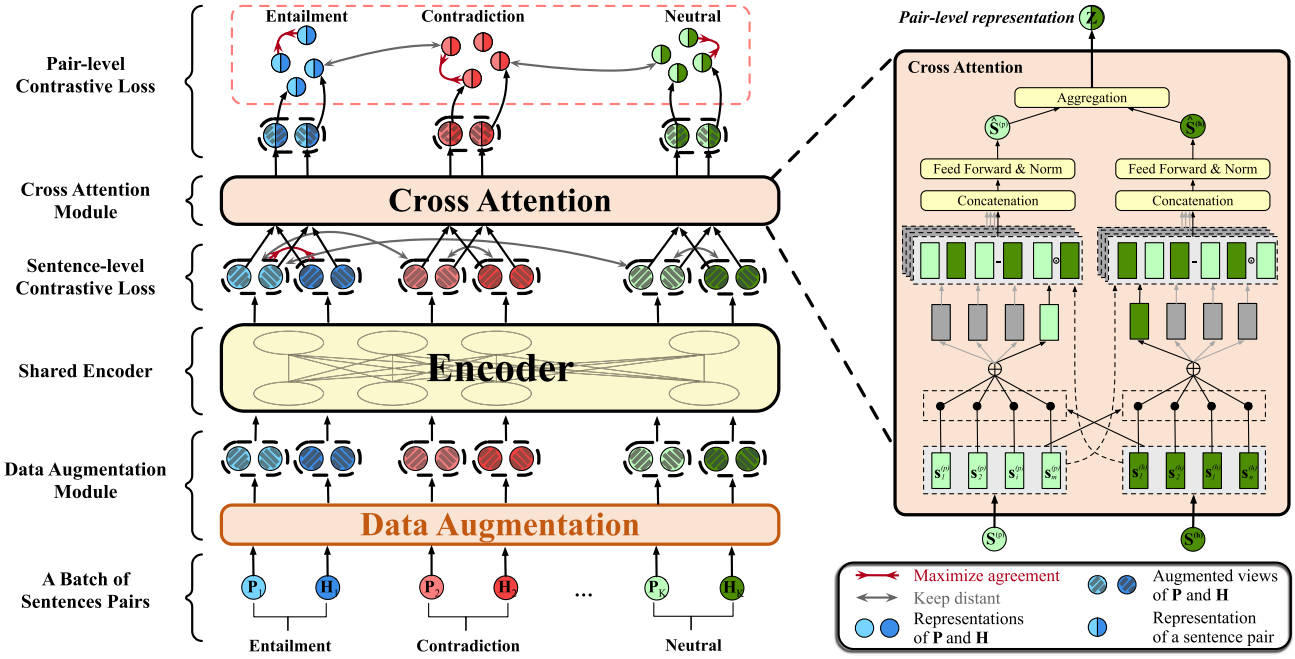


Fig. 1. The overall view of MultiSCL. The left part is the main framework of our model. The right part is the detailed structure of Cross Attention module.

We explore and test six data augmentation strategies to generate views, including synonym replacement [50], reordering [47], word insertion, word deletion, dropout [17], and back translation [51].

Synonym Replacement randomly chooses n words from the sentence that are not stop words. Replace each of these words with one of its synonyms chosen at random, so that the augmented data fit the original semantics as closely as possible.

Reordering is another widely-studied augmentation method that can keep the original sentence's features. We randomly select n pairs of spans and switch them to construct the reordering augmentation in our implementation.

Word Insertion finds the random synonym of n random word in the sentence that is not a stop word. Insert that synonym into a random position in the sentence.

Word Deletion randomly selects n tokens in the sentence and replaces them with a special token `[DEL]`, which is similar to the token `[MASK]` in BERT [28].

Dropout has been proven an effective augmentation strategy for contrastive learning [17], [18]. For this setting, we randomly drop elements in the token embedding layer by a specific number n and set their values to zero.

Back Translation first translates the sentence into another language and translates it back to the original language. Then the new sentence is regarded as an augmented sentence of the original.

In our model, we select two augmentation strategies to generate two different views of input texts. We will explore and test the effect of different combinations in our experiment. The number of changed words or tokens for augmentation methods n is based on the length of the sentence l with the formula $n = \eta l$, where η is a hyper-parameter that indicates the percent of the changed words in a sentence.

B. Text Encoder

After we get the different views of the sentence, we need to get the context-based semantic information. We give the formal definition of NLI as follows. Each instance in a NLI dataset consists of two sentences and a label indicating the relation between them. Formally, we denote **premise** as $X^{(p)} = \{x_1^{(p)}, x_2^{(p)}, \ldots, x_m^{(p)}\}$ and **hypothesis** as $X^{(h)} = \{x_1^{(h)}, x_2^{(h)}, \ldots, x_n^{(h)}\}$, where m and n are length of the sentences respectively. The instance in the batch $\mathcal{I}$ is denoted as $(X^{(p)}, X^{(h)}, y)_i$, where $i = \{1, \ldots, K\}$ is the indices of the samples and K is the batch-size. After passing the input samples to data augmentation module, we construct the new batch $\tilde{\mathcal{I}}$ with size $2K$ by randomly augmenting twice for all the sentences. The encoder (e.g., BERT) takes $X^{(p)}, X^{(h)}$ as inputs and computes the semantic representations, denoted as $\mathbf{S}^{(p)} = \{\mathbf{s}_i^{(p)} | \mathbf{s}_i^{(p)} \in \mathbb{R}^k, i = 1, 2, \ldots, m\}$ and $\mathbf{S}^{(h)} = \{\mathbf{s}_j^{(h)} | \mathbf{s}_j^{(h)} \in \mathbb{R}^k, j = 1, 2, \ldots, n\}$, where k is the dimension of the encoder's hidden state.

C. Cross Attention Module

Different from single sentence classification, we need a proper interaction module to better clarify the relationship of sentences pair for NLI task. In practice, we need to compute token-level weights between words in **premise** and **hypothesis** to obtain information about their interaction. Therefore, we introduce Cross Attention module to calculate the co-attention matrix $\mathbf{C} \in \mathbb{R}^{m \times n}$ of the token level. If the value of attention weight is relatively large, the correlation between the words is stronger. Each element $\mathbf{C}_{i,j} \in \mathbb{R}$ indicates the relevance between the i-th

Authorized licensed use limited to: University of Illinois at Chicago Library. Downloaded on May 13,2023 at 15:49:11 UTC from IEEE Xplore. Restrictions apply.

word of **premise** and the j-th word of **hypothesis**:

$$\mathbf{C}_{i,j} = \mathbf{P}^T tanh(\mathbf{W}(\mathbf{s}_i^{(p)} \odot \mathbf{s}_j^{(h)})), \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{d\times k}$, $\mathbf{P} \in \mathbb{R}^d$, and $\odot$ denotes the element-wise production operation. $\mathbf{W}$ and $\mathbf{P}$ are trainable parameters to map the feature in the semantic space, where k is the dimension of the encoder's hidden state and d is the dimension for nonlinear mapping to a latent representation. Then the attentive matrix could be formalized as:

$$\mathbf{c}_i^{(p)} = softmax(\mathbf{C}_{i,:}), \quad \mathbf{c}_j^{(h)} = softmax(\mathbf{C}_{:,j}), \tag{2}$$

$$\mathbf{s}_i^{(p)'} = \mathbf{S}^{(h)} \cdot \mathbf{c}_i^{(p)}, \quad \mathbf{s}_i^{(h)'} = \mathbf{S}^{(p)} \cdot \mathbf{c}_j^{(h)}. \tag{3}$$

We further enhance the collected local semantic information:

$$\mathbf{s}_i^{(p)''} = [\mathbf{s}_i^{(p)}; \mathbf{s}_i^{(p)'}; \mathbf{s}_i^{(p)} - \mathbf{s}_i^{(p)'}; \mathbf{s}_i^{(p)} \odot \mathbf{s}_i^{(p)'}], \tag{4}$$

$$\tilde{\mathbf{s}}_i^{(p)} = ReLU(\mathbf{W}_i^{(p)} \mathbf{s}_i^{(p)''} + \mathbf{b}_i^{(p)}), \tag{5}$$

where $[\cdot;\cdot;\cdot;\cdot]$ refers to the concatenation operation. $\mathbf{s}_i^{(p)} - \mathbf{s}_i^{(p)'}$ indicates the difference between the original representation and the **hypothesis**-information enhanced representation of **premise**, and $\mathbf{s}_i^{(p)} \odot \mathbf{s}_i^{(p)'}$ represents their semantic similarity. Both values are designed to measure the degree of semantic relevance between the sentence pair. The smaller the difference and the larger the semantic similarity, the sentences pair are more likely to be classified into the Entailment category. The difference and element-wise product are then concatenated with the original vectors ($\mathbf{S}^{(p)}, \mathbf{S}^{(p)'}$). We expect that such operations could help enhance the pair-level information and capture the inference relationships of **premise** and **hypothesis**. We get the new representation containing **hypothesis**-guided inferential information for **premise**:

$$\tilde{\mathbf{S}}^{(p)} = (\tilde{\mathbf{s}}_1^{(p)}, \tilde{\mathbf{s}}_2^{(p)}, \ldots, \tilde{\mathbf{s}}_m^{(p)}), \tag{6}$$

$$\hat{\mathbf{S}}^{(p)} = LayerNorm(\tilde{\mathbf{S}}^{(p)}), \tag{7}$$

where $LayerNorm(.)$ is a layer normalization. The result $\hat{\mathbf{S}}^{(p)}$ is a 2D-tensor that has the same shape as $\mathbf{S}^{(p)}$. The representation of **hypothesis** $\hat{\mathbf{S}}^{(h)}$ is calculated in the same way.

Then we convert these representations obtained above to a fixed-length vector with pooling. More specifically, we compute max pooling and mean pooling for $\hat{\mathbf{S}}^{(p)}$ and $\hat{\mathbf{S}}^{(h)}$. where $\hat{\mathbf{S}}^{(p)} = \{\hat{\mathbf{s}}_1^{(p)}, \hat{\mathbf{s}}_2^{(p)}, \ldots, \hat{\mathbf{s}}_m^{(p)}\}$ and $\hat{\mathbf{S}}^{(h)} = \{\hat{\mathbf{s}}_1^{(h)}, \hat{\mathbf{s}}_2^{(h)}, \ldots, \hat{\mathbf{s}}_n^{(h)}\}$. Formally:

$$\hat{\mathbf{S}}_{\text{mean}}^{(p)} = \sum_{i=1}^{m} \frac{\hat{\mathbf{s}}_i^{(p)}}{m}, \quad \hat{\mathbf{S}}_{\text{max}}^{(p)} = \max_{i=1}^{m} \hat{\mathbf{s}}_i^{(p)}, \tag{8}$$

$$\hat{\mathbf{S}}_{\text{mean}}^{(h)} = \sum_{j=1}^{n} \frac{\hat{\mathbf{s}}_j^{(h)}}{n}, \quad \hat{\mathbf{S}}_{\text{max}}^{(h)} = \max_{j=1}^{n} \hat{\mathbf{s}}_j^{(h)}. \tag{9}$$

We aggregate these representations and the pair-level representation $\mathbf{Z}$ for the sentence pair is obtained as follows:

$$\mathbf{Z} = [\hat{\mathbf{S}}_{\text{mean}}^{(p)}; \hat{\mathbf{S}}_{\text{max}}^{(p)}; \hat{\mathbf{S}}_{\text{mean}}^{(h)}; \hat{\mathbf{S}}_{\text{max}}^{(h)}]. \tag{10}$$

As described, Cross Attention module can capture the relevance of the sentence pair and well characterize the relationship. Therefore, the pair-level representation can perceive the class information of sentence pairs.

D. Training Objective

1) Text-Level Supervised Contrastive Loss: The core idea of contrastive learning is to learn a function that maps positive pairs closer together in the embedding space while pushing apart negative pairs. In general, two variants augmented from the same original sentence form the positive pair, while all other instances from the same batch are regarded as negative samples for them. Especially, the NLI datasets consist of high-quality and crowd-sourced labeled sentence pairs and each can be presented in the form: (**premise**, **hypothesis**, label) as described in Section I. Given one premise, human annotators are required to manually write one sentence that is absolutely true (entailment), one that might be true (neutral), and one that is definitely false (contradiction). Thus for each **premise** and its entailment **hypothesis**, there is an accompanying contradiction **hypothesis** and neutral **hypothesis** (see Table I for an example). Therefore, it is natural to take the entailment **hypothesis** for **premise** as its positive set and the contradiction **hypothesis** as negative set.

In the training stage, we randomly sample a batch $\mathcal{I}$ of K examples $(X^{(p)}, X^{(h)}, y)_{i\in\mathcal{I}=\{1,\ldots,K\}}$ as denoted in Section III-B. After passing the input samples to data augmentation module, we construct the new batch $\tilde{\mathcal{I}}$ with size $2K$ by randomly augmenting twice for all the sentences. We denote the representation obtained from the text encoder of the new batch as $(\mathbf{S}^{(p)}, \mathbf{S}^{(h)})_{i\in\tilde{\mathcal{I}}=\{1,\ldots,2K\}}$. For the premise $\mathbf{S}_i^{(p)}$ in the augmented batch, we denote the set of positives as $\mathbf{S}_i^{(p)+}$, including the augmented views of the entailment **hypothesis** and the augmented views of the same original **premise**. The negative set is denoted as $\mathbf{S}_i^{(p)-}$, including the augmented views of the contradiction **hypothesis**. The sentence-level supervised contrastive loss on the batch $\tilde{\mathcal{I}}$ is defined as:

$$\ell_i = \frac{e^{\text{sim}(\mathbf{S}_i^{(p)}, \mathbf{S}_i^{(p)+})/\tau}}{\sum_{k=1}^{2K} \mathbb{1}_{[k\neq i]} (e^{\text{sim}(\mathbf{S}_i^{(p)}, \mathbf{S}_k^{(p)+})/\tau} + e^{\text{sim}(\mathbf{S}_i^{(p)}, \mathbf{S}_k^{(p)-})/\tau})}, \tag{11}$$

where $\text{sim}(\cdot)$ indicates the cosine similarity function, τ controls the temperature.

Finally, we average all $2K$ in-batch losses ℓ_i to obtain the final sentence-level contrastive loss $\mathcal{L}_{SCL(sent)}$:

$$\mathcal{L}_{SCL(sent)} = \sum_{i\in\tilde{\mathcal{I}}} -\log \ell_i, \tag{12}$$

In this way, we can map the representations from the encoder of the semantically similar sentences closer together in the embedding space, while pushing apart irrelevant sentences. Thus, the pair-level representations based on the output of the encoder can better capture the relationships between **premise** and **hypothesis**.

2) Pair-Level Supervised Contrastive Loss: In [16], the authors extended the above loss to a supervised contrastive loss by regarding the samples belonging to the same class as the positive set. Inspired by this, we adopt a supervised contrastive learning objective to align the pair-level representation obtained from

Authorized licensed use limited to: University of Illinois at Chicago Library. Downloaded on May 13,2023 at 15:49:11 UTC from IEEE Xplore. Restrictions apply.

Cross Attention module to distinguish sentence pairs from different classes. The pair-level supervised contrastive loss brings the latent representations of pairs belonging to the same class closer together.

In the training stage, we take the augmented batch $\tilde{\mathcal{I}}$ of 2 K examples $(X^{(p)}, X^{(h)}, y)_{i=\{1,\ldots,2K\}}$ as denoted in Section III-D1. For the pair $(X^{(p)}, X^{(h)}, y)_i$, we denote the set of positives as $\mathcal{G} = \{g : g \in \mathcal{I}, y_g = y_i \wedge g \neq i\}$, with size $|\mathcal{G}|$. The supervised contrastive loss on the batch $\tilde{\mathcal{I}}$ is defined as:

$$\ell_{i,g} = \frac{e^{\mathrm{sim}(\mathbf{Z}_i \cdot \mathbf{Z}_g/\tau)}}{\sum_{k \in \tilde{\mathcal{I}}/i} e^{\mathrm{sim}(\mathbf{Z}_i \cdot \mathbf{Z}_k/\tau)}}, \tag{13}$$

$$\mathcal{L}_{SCL(pair)} = \sum_{i \in \tilde{\mathcal{I}}} -\log \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \ell_{i,g}, \tag{14}$$

where $\ell_{i,g}$ indicates the likelihood that pair i is most similar to pair g and τ is the temperature hyper-parameter. Larger values of τ scale down the dot-products, creating more difficult comparisons. $\mathbf{Z}_i$ is the pair-level representation of pair $(X^{(p)}, X^{(h)})_i$ from Cross Attention module. Supervised contrastive loss $\mathcal{L}_{SCL(pair)}$ is calculated for every sentence pair among the batch $\mathcal{I}$. To minimize contrastive loss $\mathcal{L}_{SCL(pair)}$, the similarity of pairs in the same class should be as large as possible, and the similarity of negative examples should be as small as possible.

3) Cross-Entropy Loss: Supervised contrastive loss mainly focuses on separating each pair apart from the others of different classes, whereas there is no explicit force in discriminating contradiction, neutral, and entailment. Therefore, we adopt the softmax-based cross-entropy to form the classification objective:

$$\mathcal{L}_{CE} = CrossEntropy(\mathbf{WZ} + \mathbf{b}, y), \tag{15}$$

where $\mathbf{W}$ and $\mathbf{b}$ are trainable parameters. $\mathbf{Z}$ is the pair-level representation from Cross Attention module and y is the corresponding label of the pair.

4) Overall Loss: The overall loss is a weighted average of CE and the multi-level SCL loss, denoted as:

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha \mathcal{L}_{SCL(sent)} + \beta \mathcal{L}_{SCL(pair)}, \tag{16}$$

where α, β is a hyper-parameter to balance the objectives.

IV. Experimental Setup

A. Benchmark Dataset

We evaluate our model on four popular datasets: the Stanford Natural Language Inference (SNLI), the MultiGenre NLI Corpus (MultiNLI), Sick and SciTail. We also conduct cross-domain experiment while trained with all source domain data of one dataset and zero-shot transferred to the target domain of another dataset to evaluate the domain adaptation capability of the model. Detailed statistical information of these datasets is shown in Table II. Len(P) and Len(H) refer to the average length of **premise** and **hypothesis** respectively. MultiNLI(m) and MultiNLI(mm) indicate the matched and mismatched datasets respectively. We use classification accuracy as the evaluation metric.

TABLE II
Statistics of Datasets

Dataset	Train	Dev	Test	Len(P)	Len(H)
SNLI	549k	9.8k	9.8k	14	8
MultiNLI(m)	392k	9.8k	9.8k	22	11
MultiNLI(mm)		9.8k	9.8k	22	11
Sick	4.5k	0.5k	4.9k	12	10
SciTail	23.5k	1.3k	2.1k	20	12

SNLI The Stanford Natural Language Inference (SNLI) dataset contains 570 k human-annotated sentence pairs, in which the premises are drawn from the captions of the Flickr30 corpus and hypotheses are manually annotated [7]. This is the most widely used entailment dataset for NLI.

MultiNLI The corpus [8] is a new dataset for NLI, which contains 433 k sentence pairs. Similar to SNLI, each pair is labeled with one of the following relationships: entailment, contradiction, or neutral. We use the matched dev set and mismatched dev set as our validation and test sets, respectively.

Sick This is a large data set on compositional meaning, annotated with subject ratings for both relatedness and entailment relation between sentences [20]. The SICK data set consists of around 10000 English sentence pairs, each annotated for relatedness in meaning.

SciTail This is a textual entailment dataset derived from a science question answering (SciQ) dataset [52]. The task involves assessing whether a given **premise** entails a given **hypothesis**. In contrast to other entailment datasets mentioned previously, the hypotheses in SciTail are created from science questions while the corresponding answer candidates and premises come from relevant web sentences retrieved from a large corpus. The dataset is only used for domain adaptation.

B. Implementation Details

We start from pre-trained checkpoints of BERT [28] (uncased). We implement MultiSCL based on Huggingface's `transformers` package [53]. All experiments are conducted on 1 Nvidia GTX 3090 GPU.

We train our models for 10 epochs with a batch size of 512 and temperature $\tau = 0.08$ using an Adam optimizer [54]. The hyper-parameter α and β are set as 1 for combining objectives. The learning rate is set as 5e-5 for base models. Weight decay is used with a coefficient of 1e-5. The maximum sequence length is set to 128. All the experiments are conducted 5 times with different random seeds and we report the average scores. The hyperparameter η is set to 0.1, which indicates the percent of the changed words in a sentence during the data augmentation module. We select Reordering and Dropout as augmentation strategies in the main experiments.

C. Baseline Models

To analyze the effectiveness of our model, we evaluate several approaches for traditional NLI scenarios and state-of-the-art methods for low-resource NLI scenarios as baselines as follows on the above datasets.

Authorized licensed use limited to: University of Illinois at Chicago Library. Downloaded on May 13,2023 at 15:49:11 UTC from IEEE Xplore. Restrictions apply.

TABLE III
PERFORMANCE ON TEST DATASETS WITH VARIOUS AMOUNTS OF TRAINING DATA*

Dataset	Model	5	20	100	500	1000	all
SNLI	InferSent	37.5	39.6	44.1	56.0	63.9	84.5
	ESIM	38.4	38.6	46.7	58.2	65.4	87.6
	BERT	33.4	37.3	47.4	70.1	78.7	90.6
	PairSCL	35.7	38.4	48.3	71.0	79.4	91.9
	GenNLI	43.5	45.6	50.6	60.6	64.2	82.2
	MT-DNN	40.2	41.4	47.5	72.4	77.3	91.0
	MultiSCL	**45.3**	**48.2**	**55.4**	**73.7**	**80.1**	**92.2**
MNLI	InferSent	34.1	33.7	35.2	44.9	47.9	70.4
	ESIM	36.9	35.4	40.5	49.8	54.2	76.7
	BERT	33.0	34.9	41.6	63.6	68.5	83.3
	PairSCL	32.1	34.4	40.8	64.1	68.9	84.6
	GenNLI	44.1	47.1	49.0	60.6	63.4	67.5
	MT-DNN	39.8	42.5	45.6	65.2	69.1	81.4
	MultiSCL	**47.2**	**49.3**	**52.7**	**66.4**	**70.6**	**85.4**
SICK	InferSent	35.5	46.3	60.2	73.2	-	83.6
	ESIM	34.5	48.4	62.9	75.4	-	84.6
	BERT	36.7	56.7	63.6	78.6	-	86.0
	PairSCL	35.8	55.2	63.1	79.2	-	86.5
	GenNLI	50.6	64.7	68.7	75.2	-	80.4
	MT-DNN	46.4	61.9	64.2	78.3	-	85.8
	MultiSCL	**54.7**	**67.5**	**71.7**	**81.4**	**-**	**87.1**

1) Traditional NLI baselines:
   - InferSent [21] uses a BiLSTM network with max-pooling to learn generic sentence embeddings that perform well on several NLI tasks.
   - ESIM [24] is a previous state-of-the-art model for the natural language inference (NLI) task. It is a sequential model that incorporates the chain LSTM and the tree LSTM to infer local information between two sentences.
   - BERT [28] is the naturally bidirectional masked language model, configured with 'bert-base-uncased'.
   - PairSCL [19] is a pair-level supervised contrastive learning approach with BERT as encoder. It adopts a Cross Attention module to learn the joint representations of the sentence pairs.
2) Low-resource NLI baselines:
   - Gen-NLI [12] is a generative classifier for NLI tasks. The model defines conditional probabilities assumed given premises and labels, parameterizing the distribution using a sequence-to-sequence model with attention and a replication mechanism.
   - MT-DNN [13] is a multi-task deep neural network for learning semantic representations across multiple natural language understanding tasks.

V. EXPERIMENT RESULTS

A. Main Results

We first empirically compare MultiSCL with baselines for three NLI datasets in low-resource scenarios. We construct smaller training sets by randomly selecting 5, 20, 100, 500, and 1000 instances per class, and then train separate models across these different-sized training sets. Table III shows the average performance and standard deviation of the three runs of our model in comparison with the baselines on SNLI, MNLI, and SICK.[1] The results of Gen-NLI and MT-DNN are obtained from the original papers, and the results of other baselines are obtained from our re-running of source code released by the authors. The best result for each dataset and data amount is shown in bold. We also conduct the students paired t-test and the p-value of the significance test between the results of MultiSCL and GenNLI is less than 0.05 and 0.01, respectively. The results show that MultiSCL outperforms baselines on all data volume settings for all three datasets. When using training sets with 5/20/100 instances per class on three datasets, MultiSCL outperforms the state-of-the-art model by 3%, 2.5%, and 3.8% respectively, which proves that MultiSCL can better capture the latent semantic representations by multi-level contrastive learning in low-resource scenarios. We can observe that transformer-based models such as BERT and PairSCL have poor performance results when the training data is less than 100 instances, even lower than LSTM-based models such as ESIM and InferSent. This shows that large-scale pre-trained models require a large amount of supervised data to be finetuned.

When the training set gets larger, the performance gap between MultiSCL and baselines does shrink. When trained with 500/1000/all instances per label, the accuracy exceeds the state-of-the-art model by 1.6%, 1.1%, and 0.6%, which shows that MultiSCL has a more significant advantage compared with other models when trained with a smaller amount of training data. Furthermore, MultiSCL outperforms our previous work PairSCL by 10.9% on average with trained less than 100 instances per class. The performance gains are due to the data augmentation module and the stronger ability to learn sentence-level latent embeddings in semantic space. The encoder of MultiSCL can capture sentence-level semantics effectively by the specifically-designed contrastive signal – regarding the entailment pairs as positives and the contradiction pairs as negatives with limited training data. When trained with full training data, MultiSCL exceeds PairSCL by 0.6%. We will further analyze the role of each module of MultiSCL in more detail in Section V-B.

TABLE IV
ABLATION STUDY ON SNLI

Model	5	20	100	500
MultiSCL (-Data Augmentation)	39.5	44.8	52.1	71.5
MultiSCL (-Cross Attention)	42.3	46.0	52.8	69.5
MultiSCL (-SCL(sent) loss)	41.0	45.1	51.7	71.2
MultiSCL (-SCL(pair) loss)	42.7	44.3	52.5	70.6
MultiSCL (-CE loss)	44.2	47.0	54.8	73.7
MultiSCL	**45.3**	**48.2**	**55.4**	**73.7**

B. Ablation Study

We run extensive ablations to better understand the contribution of each key component of MultiSCL. We conduct experiments with the training set of 5/20/100/500 instances per class on SNLI and select Reordering and Dropout as augmentation strategies. The results are shown in Table IV.

[1]SICK does not have results in the 1000 column because the 'contradiction' label has only 665 instances.

Authorized licensed use limited to: University of Illinois at Chicago Library. Downloaded on May 13,2023 at 15:49:11 UTC from IEEE Xplore. Restrictions apply.

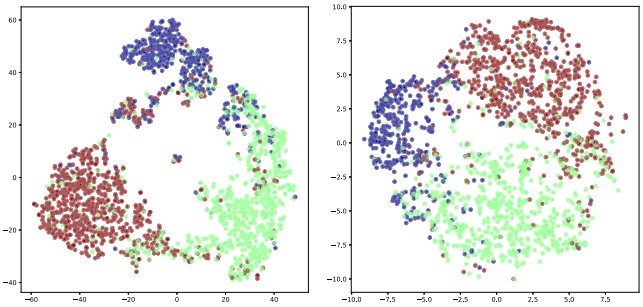


Fig. 2. Visualization of the SNLI representations using t-SNE for MultiSCL (left) and MultiSCL w/o Cross Attention Module (right).

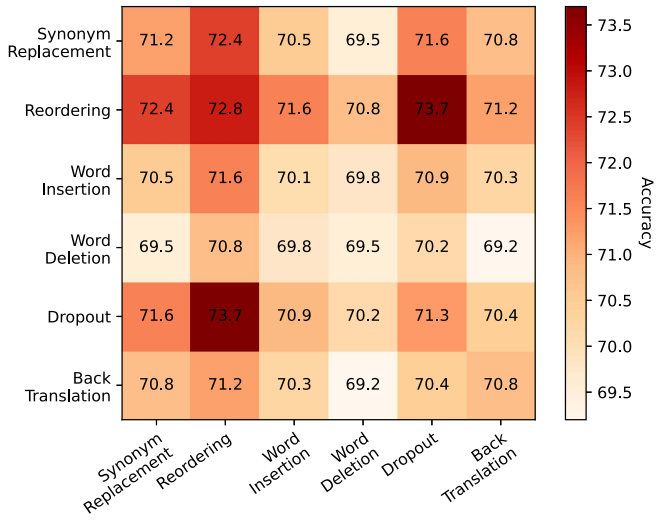


Fig. 3. The performance visualization with different combinations of data augmentation strategies. The row indicates the 1st data augmentation strategy while the column indicates the 2nd data augmentation strategy.

After removing Data Augmentation module, the performance of MultiSCL is reduced by an average of 2.94%. Moreover, we can observe that the importance of Data Augmentation module gradually decreases as the number of training instances increases. When training with 5 instances per class, the accuracy decreases by 5.8% on the test set. This indicates that when the training data is small, augmenting the sentences can bring larger enhancements to the model. The reason is that Data Augmentation module can create challenging views which allow the encoder to learn the semantic representation better with multi-level contrastive learning. After removing Cross Attention mechanism, the model simply concatenates the representation of two sentences to obtain the representation of the sentence pair. The performance decreases by 3.0%, 2.2%, 2.6%, and 3.2% respectively with different sizes of training data, which shows the joint representation obtained by cross attention can well characterize the relationship between the sentence pair. Without the sentence-level supervised contrastive learning loss, the accuracy of our model is decreased by 4.3%, 3.1%, 3.7%, and 2.5% with 5/20/100/500 training data per class. This demonstrates that by regarding entailment sentence pairs as positive samples and contradiction pairs as negative samples allows the encoder to discriminate the semantic difference between sentences. After removing pair-level supervised contrastive learning loss, the performance decreases by 2.5% on average. The reason is that the contrastive learning objective can learn the discrepancy between the sentence pairs of different classes by pulling the sentence pairs from the same class together and pushing the pairs of different classes further apart. The test accuracy decreases by 0.7% on average without the cross-entropy loss.

VI. Qualitative Analysis

In this section, we further conduct extensive experiments to understand the inner workings of MultiSCL.

A. Analysis of Cross Attention Module

To further investigate the effect of Cross Attention module, we conduct the t-SNE visualization experiments of the representations $\mathbf{Z}$ for the sentence pairs on SNLI test set. Fig. 2 illustrates that the contrastive loss can map the pairs of the same category closer together in the embedding space while pushing apart negative pairs in different classes. The right part of Fig. 2 shows the results after removing Cross Attention module. We observe that there is very little overlap between different classes in left part and the distance of the representations of different classes are distant. In contrast, the representations without Cross Attention module in different classes are closer and overlap with each other. The representations of the same category with Cross Attention module in left part of Fig. 2 are better grouped together compared to the representations in right part of Fig. 2. That indicates Cross Attention module can learn the joint representation between **premise** and **hypothesis** very well. The representations of the positive pairs obtained from Cross Attention module can be mapped together in the semantic space by contrastive learning. This could well explain why removing Cross Attention module would give a high accuracy drop.

B. Analysis of Data Augmentation Module

In this section, we first analyze the effect of different combinations of data augmentation strategies on SNLI with 500 training instances per class. As described in Section III-A, we consider six options for each augmentation, including Synonym Replacement, Reordering, Word Insertion, Word Deletion, Dropout, and Back Translation, resulting in 6×6 combinations.

The results are shown in Fig. 3. When using Word Deletion, MultiSCL has poorer performance than other strategies on average. The performance decreases to the lowest with an accuracy of 69.2% with Word Deletion and Back Translation. We argue that Word Deletion and Back Translation may change the meaning compared with the original sentences, resulting in an unexpected change in the relationship between **premise** and **hypothesis**. In this situation, MultiSCL may learn to misunderstand semantic representation with multi-level contrastive learning by incorrectly setting the positive and negative set. Compared to Word Deletion and Back Translation, Word Insertion and Synonym

Authorized licensed use limited to: University of Illinois at Chicago Library. Downloaded on May 13,2023 at 15:49:11 UTC from IEEE Xplore. Restrictions apply.

TABLE V
INFLUENCE OF η IN DATA AGUMENTATION MODULE

η	6%	7%	8%	9%	10%	20%	40%	60%	80%
MultiSCL	71.8	72.1	72.8	73.4	**73.7**	70.1	63.4	54.6	41.8

Replacement improves the performance but does not achieve the best result. These two strategies can create views that have the same meaning as the original sentence but don't introduce meaningful changes. Therefore, the model cannot construct effective positive/negative sets with these augmented views in contrastive learning.

We can observe that Reordering and Dropout are the two most effective strategies with an accuracy of 73.7% (where Reordering is slightly better than Dropout). We argue that Reordering and Dropout can create challenging sentence pairs for contrastive learning without changing the semantic information. The augmented views are useful for contrastive learning without confusing the model, and thus improve the model robustness in low-resource scenarios. We adopt Reordering and Dropout as augmentation strategies in most experiments.

Furthermore, we explore the effect of hyperparameter η on SNLI with 500 training instances per class. η is a hyperparameter that indicates the percent of the changed words in a sentence. The number of words changed n is calculated with the formula $n = \eta$l, where l is the length of the sentence. Table V shows that the accuracy first increases then decreases dramatically when the value of η becomes larger. The performance is highest when η is set to 10%. The accuracy decreases 3.6% when η is set to 20%. When the value of η is 80% in the extreme case, the accuracy of MultiSCL drops to a minimum of 41.8%. The results are not surprising. When η's value increases, the number of changed words increases which makes the augmented sentence and the original sentence more likely to have different meanings. In this situation, Data Augmentation module can introduce very serious noise to the model, so that the positive/negative pairs that we regard in multi-level contrastive learning do not work actually.

C. Influence of Temperature

The temperature in sentence-level (11) and pair-level (12) contrastive loss is used to control the smoothness of the distribution normalized by softmax operation and thus influences the gradients when backpropagation. A higher temperature smooths the distribution while a low temperature scales up the dot-products and sharpens the distribution. In our experiments, we explore the influence of temperature τ on SNLI dataset with 500 training instances per class. The result is illustrated in Fig. 4.

As shown in the figure, we can observe that the performance of MultiSCL is very sensitive to the value of temperature τ. As the temperature becomes higher, the performance of the model first improves and then decreases. Either too low or too high temperature will make our model perform badly. The optimal temperature value is 0.08 when MultiSCL has the highest accuracy of 73.7%. This phenomenon again demonstrates that the temperature determines how much attention is paid to difficult negative samples in contrastive loss. The higher the temperature, the less attention is paid to difficult negative samples, while the lower the temperature, the model focuses more on negative samples that are very different from the anchor. We select 0.08 as the temperature in most of our experiments.

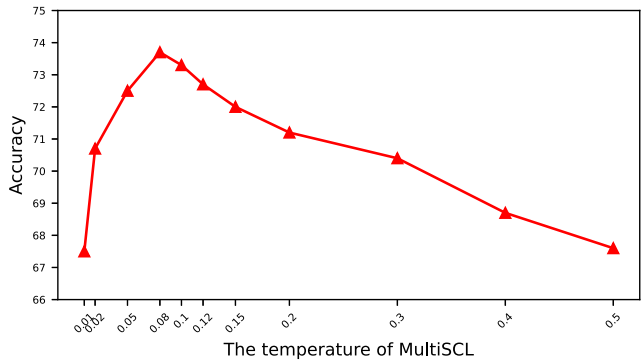


Fig. 4. The influence of different temperatures τ in MultiSCL. The best performance is achieved when the temperature is set to 0.08.

TABLE VI
INFLUENCE OF α AND β IN TRAINING OBJECTIVE

α \ β	0.2	0.5	1.0	2.0
0.2	69.5	70.2	70.6	71.6
0.5	70.4	71.3	72	72.1
1.0	71.4	72.5	**73.7**	72.8
2.0	71.2	72.7	73	71.5

D. Influence of Hyper-Parameters in Training Objective

To investigate the effect of hyper-parameters α and β in 15, we conduct the experiments with 500 instances per class on SNLI by setting different values. The results are illustrated in Table VI and we underline the setting used for all our experiments. The hyper-parameters α and β are used to balance sentence-level contrastive loss and pair-level contrastive loss in the training objective. MultiSCL focuses more on semantic similarity discrimination of sentences with a larger value of α, while the model is paying more attention to learning the discrepancy of sentence pairs in different categories with a larger value of β.

Table VI indicates a trade-off between the sentence-level semantic encoding capability and the pair-level reasoning capability of MultiSCL. When the values of both α and β are 0.2, MultiSCL almost removes contrastive learning and uses only cross-entropy loss, and the accuracy decreases by 4.2%. The performance of MultiSCL keeps improving as the value of α and β increases until the highest accuracy of 73.7% with $\alpha = \beta = 1.0$. However, as the values of α and β continue to increase, the accuracy begins to decrease. This result is not surprising, especially considering that the joint representation of a contradiction pair in pair-level contrastive learning is obtained from two sentences that are regarded as negative sets in the sentence-level contrastive learning. Focusing too much on the pair-level classification objective, i.e., using larger β values, can hurt the embeddings of the sentence from the encoder.

Authorized licensed use limited to: University of Illinois at Chicago Library. Downloaded on May 13,2023 at 15:49:11 UTC from IEEE Xplore. Restrictions apply.

TABLE VII
SELECTED EXAMPLES FROM SNLI

ID	**Premise**	**Hypothesis**	GenNLI	MultiSCL	Gold
A	Two kids are standing in the ocean hugging each other.	Two kids enjoy their day at the beach.	E	N	N
B	They are sitting on the edge of a fountain.	The fountain is splashing the persons seated.	E	N	N
C	An old man with a package poses in front of an advertisement.	A man poses in front of an ad.	N	E	E
D	A young family enjoys feeling ocean waves lap at their feet.	A family is at the beach.	N	E	E
E	A man and woman sit at a cluttered table.	The table is neat and clean.	E	C	C
F	A race car sits in the pits.	The car is going fast.	N	C	C

On the other hand, focusing overly on separating semantically dissimilar sentences also affects the discrimination of sentence pairs in different categories. We set α and β as 1.0 for all our experiments with effective multi-level supervised contrastive learning.

E. Case Study

To illustrate the advantages of our model in more detail, we conduct a case study. Table VII includes some examples from the SNLI test set, where MultiSCL successfully predicts the relation and GenNLI fails. In examples A and B, both sentences contain phrases that are either identical or highly lexically related (e.g. "Two kids", "ocean/beach", and "fountain"), which confuses GenNLI's judgment as entailment category. MultiSCL can infer the correct relationship of Neutral by capturing the difference between words in **premise** and **hypothesis** with Cross Attention module. For examples C and D, GenNLI regarded their relationships as Neutral but the gold labels are Entailment. The reason may be that the two sentences do not have some identical words, so GenNLI cannot easily recognize their semantic similarity. However, MultiSCL can correctly capture the semantics of the sentences through multi-level contrastive learning. For example E, GenNLI predicts the relationship as Entailment while MultiSCL can infer the correct relationship as Contradiction from "cluttered" and "neat and clean". For example F, MultiSCL can predict the relationship as Contradiction from "sits" and "going fast" while GenNLI considers their relationship to be Neutral. These results show that MultiSCL can understand the semantic information by capturing the interaction of words in **premise** and **hypothesis**. Furthermore, MultiSCL can better infer the relationship of the challenging pairs with the multi-level contrastive learning.

F. Error Analysis

To analyze the limitation of MultiSCL, we select some very challenging cases in which both GenNLI and MultiSCL cannot infer the relationships correctly in low-resourse scenarios. For example, MultiSCL predicts the relationship as entailment for the contradiction pair (**premise**: A person wearing a straw hat, standing outside working a steel apparatus. **hypothesis**: A person is burning a straw hat.). The most likely reason is that all the words in **hypothesis** except "burning" are included in **premise**. In this situation, MultiSCL ignores the different words ("wearing" and "burning") and simply assumes that **premise** and **hypothesis** are nearly identical. Therefore, the model predicts the relationship as entailment. Another example of model misclassification is the pair (**premise**: Two women having drinks and smoking cigarettes at the bar. **hypothesis**: Three women are at a bar.). The gold label is contradiction but MultiSCL believes their relationship is entailment. We can observe that the two sentences describe basically the same scenario, only the number of people mentioned is different. This indicates that our model ignores differences in count words when the semantics of the sentences are similar.

VII. DOMAIN ADAPTATION RESULTS

To investigate the performance of the model in low-resource scenarios more deeply, we conduct the domain adaptation experiments. We evaluate MultiSCL with BERT, GenNLI, and MT-DNN by domain adaptation between SNLI, MNLI, and Scitail. The models are trained to convergence using all the source domain data and zero-shot transferred to the target domain, which can evaluate the ability for domain-independent reasoning. The domain adaptation results are illustrated in Table VIII. MultiSCL outperforms all baselines by 8.5% on average for all six domain adaptation settings. Furthermore, we can find that all models perform the best for domain adaptation of SNLI and MNLI in various combinations of source and target datasets. The reason is that MNLI is modeled on SNLI but differs in that it covers a range of genres including transcribed speech, fiction, and government reports. Therefore, the domain similarity between SNLI and MNLI is higher and the models are more likely to transfer semantic knowledge. Moreover, the model is easier to transfer from MNLI to SNLI, compared with SNLI to MNLI.

For the domain adaptation results between SNLI/MNLI and SciTail, the accuracy is lower compared to other cases. The reason is that SciTail is a textual entailment dataset from the science domain and domain information differs significantly from other datasets. In addition, SciTail is a smaller dataset that only contains 23.5 k training data as introduced in IV-A. Therefore, it is a challenge to transfer between dissimilar domains such as SciTail to SNLI/MNLI. Our model, Multi-SCL, outperforms the state-of-the-art model by 8.8% on average in six domain adaptation scenarios. We can observe that MulitSCL achieves an outstanding performance of 49.2% and 51.4% adapted from SciTail to SNLI/MNL, exceeding the state-of-the-art model by 10.1% and 10.4%, respectively. These results indicate that MultiSCL has a stronger ability to learn domain invariant latent representations through multi-level contrastive learning.

Authorized licensed use limited to: University of Illinois at Chicago Library. Downloaded on May 13,2023 at 15:49:11 UTC from IEEE Xplore. Restrictions apply.

TABLE VIII
DOMAIN ADAPTATION PERFORMANCE

Source Dataset	Target Dataset	Model	Accuracy
SNLI	MNLI	BERT	67.0
		GenNLI	61.4
		MT-DNN	69.3
		Multi-SCL	**77.5**
SNLI	SciTail	BERT	49.0
		GenNLI	42.7
		MT-DNN	49.3
		Multi-SCL	**60.2**
MNLI	SNLI	BERT	77.4
		GenNLI	71.8
		MT-DNN	78.5
		Multi-SCL	**85.0**
MNLI	SciTail	BERT	49.8
		GenNLI	47.6
		MT-DNN	50.3
		Multi-SCL	**58.2**
SciTail	SNLI	BERT	39.1
		GenNLI	35.9
		MT-DNN	40.1
		Multi-SCL	**49.2**
SciTail	MNLI	BERT	41.0
		GenNLI	37.8
		MT-DNN	42.5
		Multi-SCL	**51.4**

MultiSCL can accurately characterize the sentence pairs in the semantic space by the specifically-designed contrastive signal and zero-shot transferred to different domains.

VIII. CONCLUSION AND FUTURE WORK

In this paper, we propose a multi-level supervised contrastive learning approach (MultiSCL) for low-resource NLI. We adopt a data augmentation module to create different views for input sentences. MultiSCL leverages sentence-level contrastive learning to enhance the capability of semantic modeling by naturally taking the entailment **hypothesis** as the corresponding **premise**'s positive set and the contradiction **hypothesis** as the negative set. The cross attention module is designed to learn the joint representations of the sentence pairs. The pair-level contrastive learning objective is aimed to distinguish the varied classes of sentence pairs by pulling those in one class together and pushing apart the pairs in other classes. We evaluate MultiSCL on three popular NLI datasets in low-resource settings. The experiment results show that MultiSCL outperforms the previous state-of-the-art method performance by 1.8%, 3.1% and 4.1% on SNLI, MNLI and Sick with 5 instances per label respectively. For the domain adaptation tasks, the accuracy of MultiSCL exceeds existing models by 8.5% on average. We carefully study the components of MutliSCL and show the effects of different parts. We also compare multiple combinations of data augmentation strategies and provide fine-grained analysis of several hyperparameters to interpret how our approach works.

In future work, we intend to exploit using contrastive learning to obtain representations that can more accurately express the relationship between sentences in low-resource settings. Furthermore, we will investigate more effective data augmentation methods for texts. Other future work will be to measure the performance of MultiSCL on adversarial and similarly challenging NLI datasets. We hope our work will provide a new perspective for future research on contrastive learning.

REFERENCES

[1] B. MacCartney and C. D. Manning, "Modeling semantic containment and exclusion in natural language inference," in *Proc. 22nd Int. Conf. Comput. Linguistics*, 2008, pp. 521–528. [Online]. Available: https://www.aclweb.org/anthology/C08-1066

[2] L. Shen et al., "Towards natural language interfaces for data visualization: A survey," 2021, *arXiv:2109.03506*.

[3] I. Dagan, D. Roth, M. Sammons, and F. M. Zanzotto, "Recognizing textual entailment: Models and applications," *Synth. Lectures Hum. Lang. Technol.*, vol. 6, no. 4, pp. 1–220, 2013.

[4] X. Hu, L. Wen, Y. Xu, C. Zhang, and S. Y. Philip, "SelfORE: Self-supervised relational feature learning for open relation extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 3673–3682.

[5] X. Hu et al., "Gradient imitation reinforcement learning for low resource relation extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 2737–2746.

[6] X. Hu, C. Zhang, F. Ma, C. Liu, L. Wen, and P. S. Yu, "Semi-supervised relation extraction via incremental meta self-training," in *Proc. Conf. Empirical Methods Natural Lang. Process.: Findings*, 2021, pp. 487–496.

[7] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 632–642. [Online]. Available: https://www.aclweb.org/anthology/D15-1075

[8] A. Williams, N. Nangia, and S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2018, pp. 1112–1122.

[9] Q. Chen, X. Zhu, Z.-H. Ling, D. Inkpen, and S. Wei, "Neural natural language inference models enhanced with external knowledge," in *Proc. Assoc. Comput. Linguistics*, 2018, pp. 2406–2417.

[10] Y. Gong, H. Luo, and J. Zhang, "Natural language inference over interaction space," in *Proc. Int. Conf. Learn. Representations*, 2018.

[11] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow, "A survey on recent approaches for natural language processing in low-resource scenarios," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2021, pp. 2545–2568.

[12] X. Ding, T. Liu, B. Chang, Z. Sui, and K. Gimpel, "Discriminatively-tuned generative classifiers for robust natural language inference," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 8189–8202.

[13] X. Liu, P. He, W. Chen, and J. Gao, "Multi-task deep neural networks for natural language understanding," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 4487–4496.

[14] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.

[15] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[16] P. Khosla et al., "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 18661–18673.

[17] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 6894–6910.

[18] Y. Yan, R. Li, S. Wang, F. Zhang, W. Wu, and W. Xu, "Consert: A contrastive framework for self-supervised sentence representation transfer," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 5065–5075.

[19] S. Li, X. Hu, L. Lin, and L. Wen, "Pair-level supervised contrastive learning for natural language inference," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 8237–8241.

Authorized licensed use limited to: University of Illinois at Chicago Library. Downloaded on May 13,2023 at 15:49:11 UTC from IEEE Xplore. Restrictions apply.

[20] M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, and R. Zamparelli, "A sick cure for the evaluation of compositional distributional semantic models," in *Proc. 9th Int. Conf. Lang. Resour. Eval.*, 2014, pp. 216–223.
[21] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 670–680.
[22] Q. Chen, X. Zhu, Z.-H. Ling, S. Wei, H. Jiang, and D. Inkpen, "Recurrent neural network-based sentence encoder with gated attention for natural language inference," in *Proc. 2nd Workshop Evaluating Vector Space Representations NLP*, 2017, pp. 36–40. [Online]. Available: https://www.aclweb.org/anthology/W17-5307
[23] W. Wu, H. Wang, T. Liu, and S. Ma, "Phrase-level self-attention networks for universal sentence encoding," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 3729–3738.
[24] Q. Chen, X. Zhu, Z.-H. Ling, S. Wei, H. Jiang, and D. Inkpen, "Enhanced LSTM for natural language inference," in *Proc. Assoc. Comput. Linguistics*, 2017, pp. 1657–1668. [Online]. Available: https://www.aclweb.org/anthology/P17-1152
[25] D. Liang, F. Zhang, Q. Zhang, and X.-J. Huang, "Asynchronous deep interaction network for natural language inference," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2019, pp. 2692–2700.
[26] X. Wanget al., "Improving natural language inference using external knowledge in the science questions domain," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 7208–7215.
[27] Z. Wang, L. Li, and D. Zeng, "Knowledge-enhanced natural language inference based on knowledge graphs," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 6498–6508.
[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.
[29] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.
[30] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7871–7880.
[31] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1412–1421.
[32] J. Gu, Z. Lu, H. Li, and V. O. Li, "Incorporating copying mechanism in sequence-to-sequence learning," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 1631–1640.
[33] T. Schick and H. Schütze, "Exploiting cloze-questions for few-shot text classification and natural language inference," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics: Main Volume*, 2021, pp. 255–269.
[34] A. Radford et al., "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.
[35] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 1735–1742.
[36] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with exemplar convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1734–1747, Sep. 2016.
[37] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3733–3742.
[38] C. Zhuang, A. L. Zhai, and D. Yamins, "Local aggregation for unsupervised learning of visual embeddings," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6002–6012.
[39] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," 2019, *arXiv:1906.05849*.
[40] Y. Qin et al., "Erica: Improving entity and relation understanding for pre-trained language models via contrastive learning," 2020, *arXiv:2012.15022*.
[41] J. M. Giorgi, O. Nitski, G. D. Bader, and B. Wang, "Declutr: Deep contrastive learning for unsupervised textual representations," 2020, *arXiv:2006.03659*.
[42] A. Conneau and D. Kiela, "Senteval: An evaluation toolkit for universal sentence representations," in *Proc. 11th Int. Conf. Lang. Resour. Eval.*, 2018.
[43] B. Gunel, J. Du, A. Conneau, and V. Stoyanov, "Supervised contrastive learning for pre-trained language model fine-tuning," 2020, *arXiv:2011.01403*.
[44] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proc. EMNLP Workshop BlackboxNLP: Analyzing Interpreting Neural Netw. NLP*, 2018, pp. 353–355.
[45] V. Suresh and D. Ong, "Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 4381–4394.
[46] Z. Li, Y. Zou, C. Zhang, Q. Zhang, and Z. Wei, "Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 246–256.
[47] Z. Wu, S. Wang, J. Gu, M. Khabsa, F. Sun, and H. Ma, "Clear: Contrastive learning for sentence representation," 2020, *arXiv:2012.15466*.
[48] D. Zhang et al., "Pairwise supervised contrastive learning of sentence representations," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 5786–5798.
[49] D. Wang, N. Ding, P. Li, and H. Zheng, "Cline: Contrastive learning with semantic negative examples for natural language understanding," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 2332–2342.
[50] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 6382–6388. [Online]. Available: https://www.aclweb.org/anthology/D19-1670
[51] S. Edunov, M. Ott, M. Auli, and D. Grangier, "Understanding back-translation at scale," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 489–500.
[52] T. Khot, A. Sabharwal, and P. Clark, "Scitail: A textual entailment dataset from science question answering," in *Proc. AAAI Conf. Artif. Intell.*, 2018.
[53] T. Wolf et al., "Huggingface's transformers: State-of-the-art natural language processing," 2019, *arXiv:1910.03771*.
[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

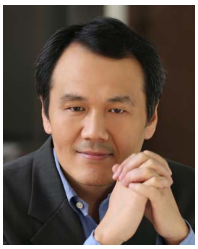
Shu'ang Li received the B.S. degree from Tsinghua University, Beijing, China, in 2018. He is currently working toward the Ph.D. degree with the School of Software, Tsinghua University. His research interests include text classification and deep learning.

Xuming Hu received the B.E. degree in computer science and technology from the Dalian University of Technology, Dalian, China. He is currently working toward the Ph.D. degree with Tsinghua University, Beijing, China. His research interests include natural language processing and information extraction.

Li Lin received the B.S. degree from Xi'an Jiaotong University, Shaanxi, China, in 2016 and the Ph.D. degree with the School of Software, Tsinghua University, Beijing, China, in 2022. Her research interests include sequence modeling and deep learning.

Authorized licensed use limited to: University of Illinois at Chicago Library. Downloaded on May 13,2023 at 15:49:11 UTC from IEEE Xplore. Restrictions apply.

Aiwei Liu received the B.E. degree in software engineering from Nanjing University, Nanjing, China. He is currently working toward the Ph.D. degree with Tsinghua University, Beijing, China. His research interests include natural language processing and semantic parsing.



Lijie Wen received the B.S., M.S., and Ph.D. degrees with the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2000, 2003, and 2007 respectively. He is currently an Associate Professor with the School of Software, Tsinghua University. He has authored or coauthored more than 150 academic papers on conferences and journals, which are cited more than 4100 times by Google Scholar. His research interests include process data management, lifecycle management of computational workflow, and natural language processing.

Philip S. Yu (Life Fellow, IEEE) is currently a Distinguished Professor and the Wexler Chair of information technology with the Department of Computer Science, University of Illinois Chicago (UIC), Chicago, IL, USA. Before joining UIC, he was with IBM Watson Research Center, where he built a World-Renowned Data Mining and Database Department. He has authored or coauthored more than 780 papers in refereed journals and conferences. He holds or has applied for more than 250 U.S. Patents. His research interest include Big Data, data mining, data stream, database, and privacy. He is a Fellow of ACM. Dr. Yu was the Editor-in-Chief of *ACM Transactions on Knowledge Discovery from Data* during 2011–2017 and IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING during 2001–2004. He was the recipient of several IBM honors, including the two IBM Outstanding Innovation Awards, Outstanding Technical Achievement Award, two Research Division Awards, and 94th Plateau of Invention Achievement Awards.

Authorized licensed use limited to: University of Illinois at Chicago Library. Downloaded on May 13,2023 at 15:49:11 UTC from IEEE Xplore. Restrictions apply.