applied sciences

MDPI

Article

Multi-Head Self-Attention-Enhanced Prototype Network with Contrastive–Center Loss for Few-Shot Relation Extraction

Jiangtao Ma [1,2], Jia Cheng [1], Yonggang Chen [3], Kunlin Li [1], Fan Zhang [4] and Zhanlei Shang [1,*]

[1] College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China; majiangt@139.com (J.M.); 332107040586@email.zzuli.edu.cn (J.C.); 332207050714@email.zzuli.edu.cn (K.L.)
[2] Songshan Laboratory, Zhengzhou 450046, China
[3] The State Information Center, Beijing 100045, China; yonggang@sic.gov.cn
[4] China National Digital Switching System Engineering and Technology R&D Center, Zhengzhou 450001, China; ffzhang@fudan.edu.cn
* Correspondence: shangzl@zzu.edu.cn

Abstract: Few-shot relation extraction (FSRE) constitutes a critical task in natural language processing (NLP), involving learning relationship characteristics from limited instances to enable the accurate classification of new relations. The existing research primarily concentrates on using prototype networks for FSRE and enhancing their performance by incorporating external knowledge. However, these methods disregard the potential interactions among different prototype networks, and each prototype network can only learn and infer from its limited instances, which may limit the robustness and reliability of the prototype representations. To tackle the concerns outlined above, this paper introduces a novel prototype network called SACT (multi-head **s**elf-**a**ttention and **c**ontrastive-cen**t**er loss), aimed at obtaining more comprehensive and precise interaction information from other prototype networks to bolster the reliability of the prototype network. Firstly, SACT employs a multi-head self-attention mechanism for capturing interaction information among different prototypes from traditional prototype networks, reducing the noise introduced by unknown categories with a small sample through information aggregation. Furthermore, SACT introduces a new loss function, the contrastive–center loss function, aimed at tightly clustering samples from a similar relationship category in the center of the feature space while dispersing samples from different relationship categories. Through extensive experiments on FSRE datasets, this paper demonstrates the outstanding performance of SACT, providing strong evidence for the effectiveness and practicality of SACT.

Keywords: few-shot; relation extraction; prototype network; multi-head self-attention; contrastive–center loss

Citation: Ma, J.; Cheng, J.; Chen, Y.; Li, K.; Zhang, F.; Shang, Z. Multi-Head Self-Attention-Enhanced Prototype Network with Contrastive–Center Loss for Few-Shot Relation Extraction. *Appl. Sci.* **2024**, *14*, 103. https://doi.org/10.3390/app14010103

Academic Editors: Gloria Bordogna and Fugazza Cristiano

Received: 30 October 2023
Revised: 3 December 2023
Accepted: 18 December 2023
Published: 21 December 2023

1. Introduction

Relation extraction (RE) serves as a crucial subtask in natural language processing (NLP) [1] and a key step in constructing Knowledge Graphs (KGs). Its objective is to extract semantic relationships between entities from unstructured text. For example, from the sentence "The hacker gained access to sensitive information by executing a media-less attack on Android", the relationship "gained access to" can be extracted, resulting in the triple <Hacker, gained access to, sensitive information>. RE can extract numerous relationship instances, which can be applied in downstream tasks, such as intelligent question-answering systems [2], machine translation [3], information retrieval [4], recommendation systems [5], and so on.

The existing RE methods commonly employ deep learning techniques and can be categorized into supervised, semi-supervised, and unsupervised approaches. Extensive research has shown that supervised RE methods achieve excellent results. However, supervised methods need to utilize a substantial mass of high-quality manually labeled training data during the training phase, which entails significant human and time costs. Meanwhile, supervised RE is hindered by the presence of annotation errors in the manually labeled data, limiting its performance. To address the limitations of manually annotated data, Mintz [6] proposed the use of a distant supervision algorithm to automatically generate large-scale labeled data. However, utilizing labeled data generated by distant supervision algorithms has certain drawbacks. It introduces a substantial amount of label noise [7] and also results in a long-tail distribution of data [8], where only a few categories have extremely limited labeled data available. While some established methods [9–11] have effectively mitigated the issue of noisy data, they encounter a substantial decline in performance when confronted with a limited quantity of training data.

To address the issue of data sparsity [12] caused by distant supervision for RE, researchers have advocated for the adoption of few-shot learning (FSL) [13]. This approach typically applies meta-learning [14] to tackle this task. Few-shot relation extraction (FSRE) tasks involve the generalization of a limited number of labeled examples for the extraction of new relations. These tasks often employ the N-way-K-shot setup, as illustrated in Figure 1. Recently, many studies have utilized metric learning within a meta-learning framework to address FSRE, and prototype networks [15] have become a hot research topic due to their simplicity and efficiency. While prototype networks have made significant advancements, many studies have aimed to enhance the model's performance by incorporating external information for better prototype representations. For instance, TDproto [16] leverages relation and entity description information to enhance prototype networks. HCRP [17] employs relation–prototype contrastive learning to better leverage relationship information and obtain diverse and discriminative prototype representations. SimpleFSRE [18] improves model performance by concatenating two representations of relationship information and directly incorporating them into the prototype representation. PRM [19] combines a gating mechanism to utilize relationship description information, determining the degree of preservation and an update of both the prototype and relationship information. CBPM [20] corrects the prototype network by utilizing category information from the query set and hierarchical information from relationship synonyms.

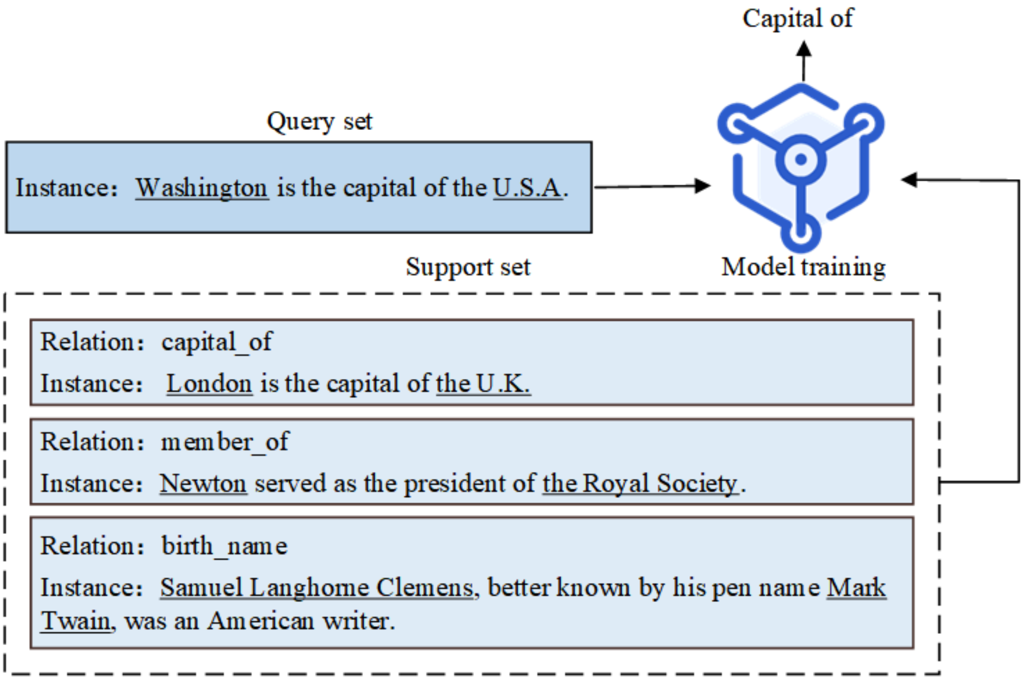
Capital of
Query set
Instance: Washington is the capital of the U.S.A.
Model training
Support set
Relation: capital_of
Instance: London is the capital of the U.K.
Relation: member_of
Instance: Newton served as the president of the Royal Society.
Relation: birth_name
Instance: Samuel Langhome Clemens, better known by his pen name Mark Twain, was an American writer.

Figure 1. A demonstration of 3-way 1-shot scenario. Words with underscores signify entity mentions. The model is trained on support set instances to predict the relationship between the two known entities in the query set.

However, these investigations often employ a simplistic strategy, which entails averaging the sentence representations associated with each relation class in the support set to obtain the prototype representations. Despite introducing relationship information to constrain prototypes for better representations, there are limitations to consider: Firstly, they overlook the interactivity among prototypes, which would constrain the model's access to global information within the data. Each prototype network merely learns from its own limited instances, lacking a comprehensive perspective from other prototype networks. This limitation may result in the model's inability to capture potential correlations and shared features among distinct prototypes, leading to a diminished expressive capability of the prototype network. In practical applications, variations in unknown categories or samples may not be confined to a single prototype network but rather involve interactions among multiple prototype networks. Neglecting such interactions could render the model less robust when confronted with unfamiliar situations, making generalization to broader data distributions challenging. As shown in Figure 2, the interaction information between different prototypes can provide strong supporting evidence for RE. Additionally, when employing contrastive learning or graphs to constrain prototypes, there is limited utilization of global information. Consequently, this model may exhibit poor classification performance for outlier samples with low semantic similarity due to substantial differences between their prototypes.

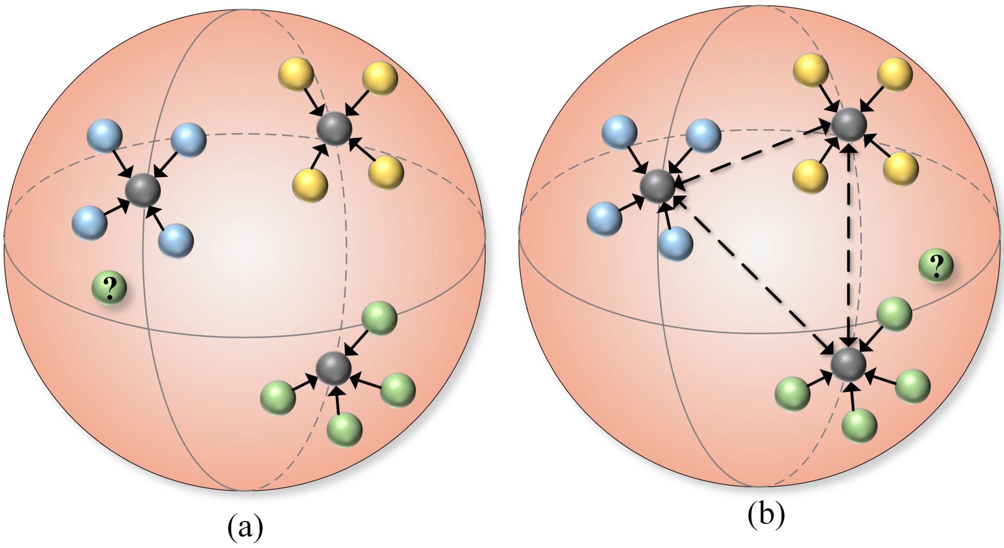

Figure 2. An illustration of the impact of prototype interaction information on query instances. Gray spheres represent prototype networks, while spheres of other colors represent representations of support instances. The green spheres with question marks represent representations of query instances. (**a**) Originally, the representation of the query instance closely resembles the blue prototype. (**b**) After interacting with information from different prototypes, the position of the query instance representation changes, thereby modifying the prototypes.

To tackle the previously mentioned concerns, this paper presents a relationship–prototype fusion method based on a multi-head self-attention network (SACT). SACT leverages prototype interrelationships more effectively and combines them with relationship description information to generate improved prototype representations. Specifically, SACT introduces a prototype enhancement module that enhances prototype representations by adding a multi-head self-attention mechanism based on the relationships between prototypes. This results in enriched prototype representations that are integrated with the basic prototypes. Furthermore, SACT employs an adaptive space fusion technique to merge relationship information with prototype representations, ultimately obtaining the final prototype representations. This approach facilitates the acquisition of more comprehensive and effective prototypes by the model. Additionally, SACT introduces a contrastive–center loss function, which simultaneously minimizes distances among the same class while increasing distances between different classes, thereby improving the handling of outlier samples.

The following outline the principal contributions of this paper:

(1) This paper presents a novel prototype network with multi-head self-attention enhancement with contrast–center loss, named SACT. This model leverages interaction information among sufficient prototypes to enhance them. It employs an adaptive fusion mechanism to integrate relationship information with the improved prototypes, thereby enhancing the classification accuracy.

(2) This paper introduces a contrastive–center loss function that enhances intra-class cohesion and inter-class separability by comparing the distances between query samples and their respective class centers with the distances to non-corresponding class centers.

(3) Extensive experiments were conducted using two extensive FSRE datasets, FewRel 1.0 and FewRel 2.0, which yielded results surpassing those of other SOTA models. Furthermore, ablation experiments were carried out to showcase the efficacy of SACT.

2. Related Work

2.1. Relation Extraction

While traditional RE methods [21,22] have achieved notable results, they entail substantial investments in human and material resources and exhibit limited robustness. As deep learning [23] is rapidly evolving, many researchers have made significant strides by leveraging it for RE. Currently, supervised RE research heavily relies on various neural network models, including CNNs [24,25], RNNs [26,27], LSTMs [28,29], and more. The introduction and application of these methods have contributed to the advancement of the field of RE. However, supervised RE methods necessitate a substantial deal of annotated data, a resource-intensive and time-consuming endeavor that places a significant strain on human resources. Building upon this, in 2009, Mintz was the pioneer in proposing the utilization of distant supervision to address the aforementioned issue. Huang et al. [30] first proposed the use of residual learning in conjunction with a multilayer CNN to solve the RE problem. Zeng and Qin used adversarial learning [31], deep reinforcement learning [32], and generative adversarial learning [33], respectively, to solve the remotely supervised noise problem. While the aforementioned methods provide improved solutions to the noise problem associated with distant supervision, addressing the long-tailed distribution of data resulting from remote supervision algorithms remains an ongoing challenge.

2.2. Few-Shot Learning

FSL endeavors to explore how to train models using a limited quantity of data. Research on FSL is generally categorized into three main approaches: FSL methods that focus on enhancing the model structure, FSL methods that utilize metrics, and FSL methods that employ optimization techniques.

Methods based on improving model structure do not require the use of meta-learning and directly strengthen the model to address FSL problems. Santoro et al. [34] eliminated the drawbacks of traditional models by improving the model's memory mechanism, using Neural Turing Machines (NTMs) to perform short-term memory and update long-term memory, enabling rapid and accurate predictions for data that only appear once. Employing a meta-learning framework grounded in temporal convolution and soft attention, Mishra et al. [35] facilitated the utilization of historical information and the precise localization of required information segments. Ren et al. [36] combined incremental learning based on attention guidance mechanisms with FSL, introducing a method called the attraction mechanism that directs the model's attention to features related to new classes, enabling rapid adaptation to new categories.

In optimization-based FSL methods, Finn et al. [37] utilized a straightforward and efficient task-agnostic algorithmic model. Parameters are acquired by executing gradient descent on a collection of small tasks, allowing the current task to quickly converge with only a few iterations on the training set. Elsken et al. [38] proposed a method that fully integrates the Neural Architecture Search (NAS) with gradient-based meta-learning.

Metric-based FSL methods learn the distances between different classes of samples and have found widespread applications in FSL. Koch [39] and Vinyals [40] introduced Siamese neural networks and matching networks for FSL, respectively. Snell et al. [41] made the first attempt to solve the FSL problem using a prototype network. Prototype networks create a prototype representation for each class, establishing a metric space for classification by measuring the distance between the prototype vectors of all classes and query points in the metric space. However, prototype networks have not been well explored in RE. The research conducted in this study aims to address this gap by focusing on the information interaction between prototypes. The objective of this work is to delve into the interactions between prototypes and present a prototype refinement method that relies on a multi-head self-attention mechanism. The aim is to better capture the interactions and information flow among different prototypes.

2.3. Few-Shot Relation Extraction

FSRE endeavors to extract relationships among entities using a limited amount of labeled data. Han et al. [42] played a pioneering role in creating FewRel 1.0, a comprehensive large-scale dataset for FSRE. This dataset provides various evaluation metrics for comparing different FSRE models. However, both the training and validation sets originate from Wikidata, leading to a lack of domain adaptability and affecting its generalization performance. In response to this issue, Gao et al. [43] re-annotated a cross-domain test set utilizing the FewRel 1.0 dataset as the foundation, making it more challenging for models to transfer knowledge across different domains. Ye et al. [44] consider the information about the matching of each query and support instances at local and instance levels, facilitating instance classification in the query set based on relationships.

However, in FSRE, issues like dataset noise and feature sparsity can result in diminished model performance. To tackle these issues, Gao et al. [45] introduced a hybrid attention mechanism to enhance the handling of instance tasks and address the issue of feature sparsity, thereby adapting to the challenges of FSRE tasks. Han et al. [17] investigated the impact of task difficulty on model performance in few-shot tasks. Their work introduced a combination of prototype networks and relationship prototype contrastive learning, providing an effective solution to address the difficulty of FSRE tasks. Wang et al. [46] introduced a rule-based discriminative knowledge method to address the prediction confusion issue in FSRE. This approach utilizes a logical perception reasoning module and a distinctiveness discovery module to enhance prediction accuracy.

Furthermore, some researchers have attempted to introduce external knowledge to enhance the accuracy of RE. Yang et al. [47] leveraged external knowledge bases to extract inherent concepts of entities. They used a concept attention module to select the entity concepts that are most semantically similar to the sentence and employed a self-attention fusion module to combine them with entity-embedding vectors, enhancing entity representations and improving the accuracy of RE. Peng et al. [48] introduced an innovative method known as the Entity-Masked RE Contrastive Pre-training Framework. This approach aims to leverage contextual information and entity type information for RE by studying and analyzing the impact of contextual context and entity mentions on RE performance. Dong et al. [49] improved the performance of RE by utilizing context information and label-agnostic and label-aware knowledge provided by relation labels. Although the previous model performed well, its ability to handle outlier samples in FSRE tasks is limited. This paper adopts a contrastive–center loss function to improve the model's capability in managing outlier samples by enhancing both the clustering of similar samples and the separation of dissimilar ones.

3. Problem Formulation

In the context of FSRE, the conventional approach involves meta-learning, encompassing two fundamental phases: meta-training and meta-testing. During the meta-learning phase, the model acquires the ability to learn, and then during the meta-testing stage, it leverages the knowledge acquired from the meta-training set to rapidly generalize to new relation categories. This process is usually conducted on a per-task basis and follows the N-way-K-shot setting. In each task, the training data are classified as two segments, the support set $\mathcal{S}$ and the query set $\mathcal{Q}$, with non-overlapping relation types between $\mathcal{S}$ and $\mathcal{Q}$. The meta-training stage includes an auxiliary dataset called T_{base}, which contains a diverse set of base classes for model training, and these base classes do not overlap with the novel classes. During the meta-training stage, $\mathcal{N}$ relation categories are randomly sampled from T_{base}, with each category containing $\mathcal{K}$ instances, to construct the support set $\mathcal{S} = \left\{ \mathcal{S}_j^i \in R^{2d}, i = 1, 2, \ldots, \mathcal{N}; j = 1, 2, \ldots, \mathcal{K} \right\}$. Then, a random selection of $\mathcal{M}$ instances is drawn from the remaining instances within these $\mathcal{N}$ categories to create the query set $\mathcal{Q} = \left\{ \mathcal{Q}^l \in R^{2d}, l = 1, 2, \ldots, \mathcal{M} \right\}$. Predicting the relations among instances within the query set $\mathcal{Q}$ is the primary objective of this task. The model undergoes iterative training, where the disparity between the anticipated labels within the query set $\mathcal{Q}$ and the actual labels serves as informative feedback signals.

4. Methodology

4.1. Framework

In order to better leverage the interactive information among prototypes, we propose a prototype network model for FSRE, termed multi-head self-attention and contrastive–center loss (SACT). To enhance the representational power of the prototype network in scenarios with only a limited number of support instances, we introduce a multi-head self-attention mechanism, facilitating the utilization of the interaction information in each prototype. This approach combines the relationship information and employs an adaptive prototype space fusion technique to generate more enriched prototypes. Additionally, we design a contrastive–center loss to assist the model in effectively aggregating samples from the same category while distinguishing differences between different categories, enabling the model to learn a more discriminative metric space.

In this section, we will provide a detailed exposition of the primary framework of the proposed model, SACT, as illustrated in Figure 3. The framework comprises the following modules: (1) Input: Initially, the sampled support set, query set, and relation information are input into the sentence encoder. (2) Sentence Encoder: This is responsible for encoding the input support and query sets into corresponding sentence representations. Simultaneously, the relationship information is transformed into relationship representations, and their final representation is obtained through additive operations. (3) Prototype Enhancement: By averaging the sentence representations for each relation class in the query set, basic prototypes are obtained. These prototypes are then enhanced using a multi-head self-attention mechanism to generate more expressive prototype representations. Finally, an adaptive prototype fusion mechanism combines the representations of the relationship information and enhanced prototypes to form the ultimate prototype representation. (4) Contrastive–Center Loss: This module compels the prototype network to learn a more discriminative metric space in the semantic domain. It introduces the contrastive–center loss function, aiding the model in effectively aggregating samples from the same category and distinguishing differences between the different categories. Subsequently, we will delve into detailed explanations of the latter three components in the following sections.

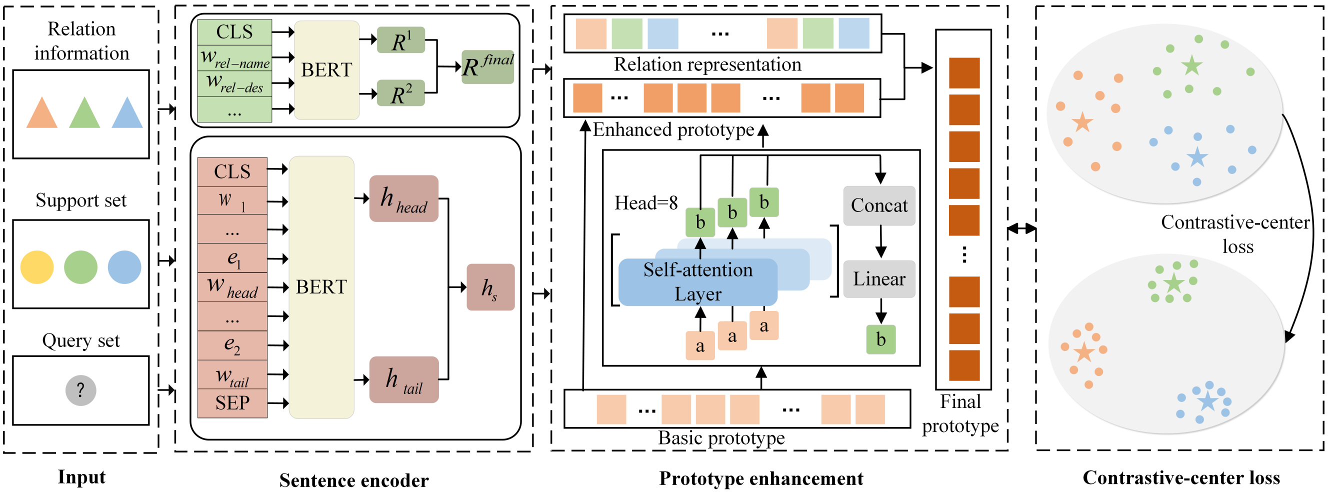

Figure 3. The architecture of our proposed SACT for the FSRE task. SACT first introduces the input relation information, support set, and query set into a BERT encoder to obtain the relationship information representation in the upper part of the sentence encoder module and the sentence representations in the lower part. Subsequently, the prototype network is further enhanced through a multi-head self-attention mechanism and optimized using the contrast–center loss function. In the diagram, relationship information is represented by triangles, the support set is denoted by circles, circles with question marks represent the query set, and pentagrams symbolize prototype representations.

4.2. Sentence Encoder

In this section, we will provide a detailed overview of the structure and functionality of the sentence encoder. By introducing the sentence encoder, we encode the input instances and relationship information into low-dimensional vector representations, laying the foundation for subsequent processing and analysis. The encoder layer comprises two primary components: (1) Sentence Representation: Utilizing a pre-trained language model to encode each word in the input instances, we obtain the sentence representation. (2) Relationship Representation: By employing the encoder to represent relationship information, we obtain two vector representations for the relationship. This process involves encoding the relationship information to acquire a relationship-level representation and directly concatenating them to form the final relationship representation.

4.2.1. Sentence Representations

The existing research utilizes various types of encoders, including CNN, RNN, LSTM, and others, for the extraction of features from sentences, each with its own strengths and limitations. CNNs can capture local features and semantic information in sentence feature extraction to extract useful features from them, but there are limitations in processing long text and entity context information. In contrast, RNNs can capture context information by memorizing previous inputs and handle variable-length input sequences but process long texts only in a forward manner. LSTMs, on the other hand, effectively capture bidirectional dependency information in sequences, which is crucial for RE tasks. However, both RNNs and LSTMs suffer from issues, like gradient vanishing and exploding.

With the advancement of pre-trained language models like Transformer and BERT [50], they acquire richer semantic information through pre-training on extensive data to provide more accurate and comprehensive representations for sentences in the support and query sets. Compared to other encoders, the BERT encoder simultaneously trains bidirectional language representations, extracting features for each word in the input sequence in parallel. By leveraging contextual information from surrounding words during feature extraction, BERT more effectively captures global context information. Therefore, this paper employs BERT as the sentence encoder to encode sentences from the input $\mathcal{S}$ and $\mathcal{Q}$. Specifically, the input sentences are preprocessed by tokenizing the sentences and splitting them into individual tokens or words, each of which usually represents a word called tokens, thus making the sentences easier to process by the model. On the basis of tokenization, [CLS] and [SEP] tokens are appended to the commencement and termination of sentences, respectively, to signify the start and finish of the sentence. Two pairs of special tokens, [e1] and [/e1] for the subject entity, and [e2] and [/e2] for the object entity, are inserted before and after the corresponding entities. For example, in the sentence "The hacker gained access to sensitive information by executing a media-less attack on Android", the given sentence is processed as "[CLS] [e1] The hacker [/e1] gained access to [e2] sensitive information [/e2] by executing a media-less attack on Android. [SEP]". Next, this paper employs BERT to encode the preprocessed sentences, obtaining vector representations for the sentences, as illustrated in Figure 4. Subsequently, these vectors are concatenated to the positions corresponding to "[e1]" and "[e2]" to acquire sentence representations that incorporate entity information.

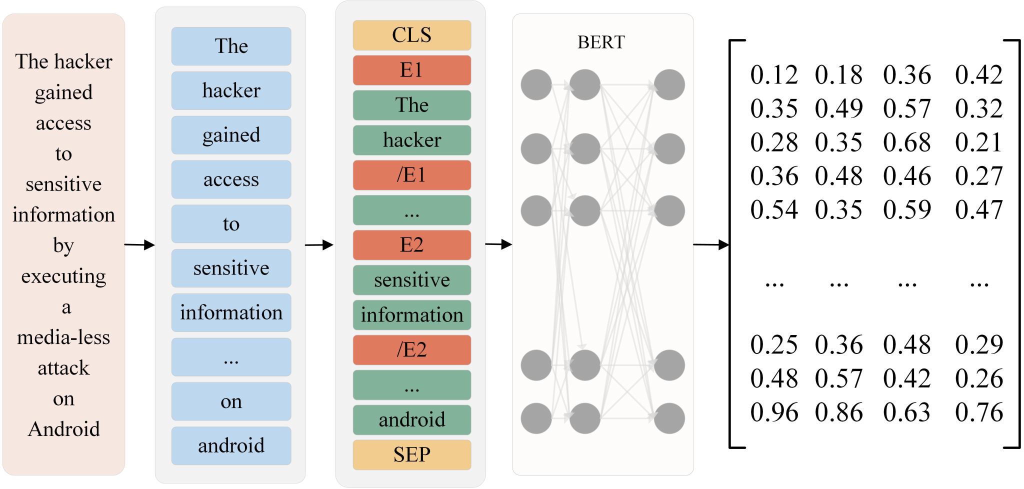

Figure 4. An example of sentence representation generated by the sentence encoder. It illustrates how an input sentence is transformed into a numerical representation that can be utilized for further processing.

4.2.2. Relation Representations

In the case of each piece of relation information, SACT concatenates the relation name and relation description in the format of "Relation Name: Relation Description". This sequence is then input into the BERT encoder, and it is encoded to generate two components of the relation representation: the [CLS] token embedding and the average embedding of all the tokens. These components are denoted as $\left\{\mathcal{R}_i^1 \in \mathcal{R}^{2d}; i = 1, 2, \ldots, N\right\}$, $\left\{\mathcal{R}_i^2 \in \mathcal{R}^{2d}; i = 1, 2, \ldots, N\right\}$, respectively. This paper concatenates the two components of the relation representation directly, represented as $\mathcal{R}_i^1$ and $\mathcal{R}_i^2$, to generate the final relation representation. This concatenation is performed without introducing additional linear layers or parameters, thereby preserving more of the original information, as shown in Equation (1):

$$\mathcal{R}_i^{\text{final}} = \mathcal{R}_i^1 \oplus \mathcal{R}_i^2 \tag{1}$$

4.3. Prototype Enhancement Module

The prototype network creates a prototype representation for each instance and utilizes the distance between the prototype vector and the query instances for classifying the query set. Typically, a straightforward averaging of the representations of the support set instances is used to obtain the prototype vector. However, this method overlooks effective interaction information among the prototypes. Additionally, relationship information is crucial for obtaining improved class prototypes. Therefore, to enhance the representational capacity of the prototypes, we introduce a multi-head self-attention mechanism and relationship information. By merging the relationship information with the enhanced prototype representations, a more comprehensive and expressive prototype representation is formed, providing stronger support for the classification tasks.

4.3.1. Basic Prototype

To obtain sentence representations for the support set, the support set is provided as input to the sentence encoder. Following the conventional prototype network approach, SACT simply averages the sentence representations of the support set to derive the basic prototype p_i^{basic}, calculated as shown in Equation (2):

$$P_i^{\text{basic}} = \frac{1}{|\mathcal{K}|} \sum_{m=1}^{\mathcal{K}} \Gamma(s_i^m) \tag{2}$$

where Γ represents the sentence encoder, $\mathcal{K}$ denotes the total samples for the i-th relation within $\mathcal{M}$, S_i^m represents the embedding of the m-th support set for the i-th relation, and p_i^{basic} represents the basic prototype of relation i.

4.3.2. Enhanced Prototype

In the prototype space, similar data points may be assigned to adjacent prototype vectors, and there might be some correlation between these similar prototypes. However, the basic prototypes obtained from Equation (2) overlook interactions between classes. To explore the inherent relationships between the class prototypes, SACT introduces a multi-head self-attention mechanism to consider the influence of different prototypes on each other. Self-attention allows each prototype to focus on different regions in the embedding space and gather information from other prototype vectors. The key idea behind multi-head self-attention is to utilize multiple self-attention operators to simultaneously process features from different subspaces. Each self-attention operator can capture different focal points in the input sequence and assign distinct weights to each focal point. By employing multiple self-attention operators, the model can concurrently focus on different parts of the input and merge their features together to enhance the contextual information for local features. This allows our model to better handle few-shot unknown categories and achieve accurate classification when facing new query instances.

When the prototype network encounters FSRE, it may produce considerable noise for classes with limited samples due to insufficient training. To mitigate the impact of noise and enhance the expressiveness of the prototype network, SACT employs an information-aggregation approach. Considering each prototype as a head of the self-attention mechanism, we calculate the attention weights between each head and the others. This allows us to obtain interaction information from different prototypes. Such interaction information can be regarded as a form of global contextual information, aiding the model in better understanding the relationships and interactions between various prototypes. Consequently, when dealing with unknown categories, the model can more effectively capture their associations and common features. Finally, by aggregating the features from multiple heads, we synthesize these rich interactive pieces of information to obtain a comprehensive prototype representation. The prototype enhancement process, as shown in Figure 5, takes the basic prototypes obtained from Equation (2) as input to a multi-head self-attention model. By applying a linear transformation to the input prototype vectors, SACT obtains the query Q, key K, and value V matrices as described in Equation (3):

$$\begin{aligned} \text{Query } &= W_Q \times p_i^{\text{basic}} \\ \text{Key } &= W_K \times p_i^{\text{basic}} \\ \text{Value } &= W_V \times p_i^{\text{basic}} \end{aligned} \tag{3}$$

where W_Q, W_K, and W_V are the learnable matrices in different linear transformations. Next, SACT computes the correlation between query and key, which involves calculating the attention scores represented by the sparse matrix $\alpha_{i,j}$ using vector dot products:

$$\alpha_{i,j} = \mathrm{softmax}\left(\frac{[W_Q \times p_i] \times [W_K \times p_j]^T}{\sqrt{d_k}}\right) \tag{4}$$

where $\sqrt{d_k}$ is the dimension of the prototype vectors, and p_i and p_j are the basic prototypes for the i-th and j-th classes, respectively. The basic prototypes are inputted into the self-attention matrix to obtain the enhanced prototypes, and the basic prototypes are integrated with the enhanced prototypes to obtain a more enriched prototype representation. The enhanced prototypes p_i^{enhanced} are denoted as in Equation (5):

$$p_i^{\text{enhanced}} = p_i^{\text{basic}} + W_v \times p_i \times \alpha_{i,j} \tag{5}$$

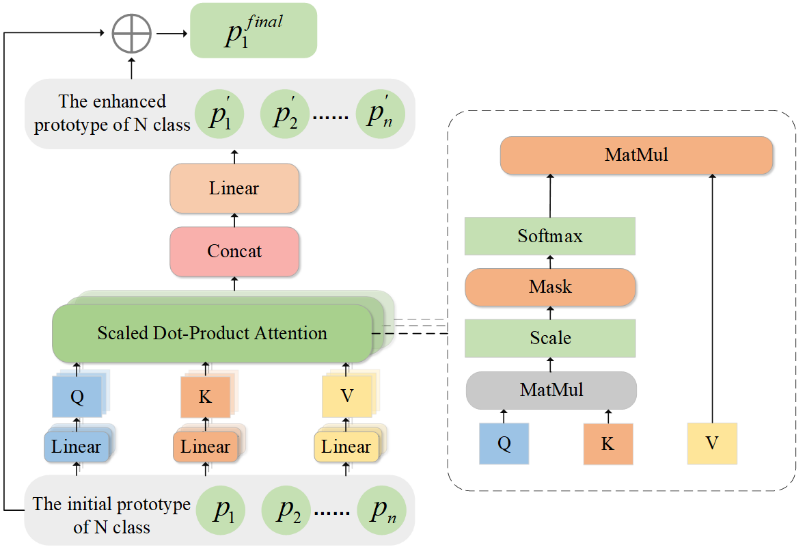
p_1^{final}
The enhanced prototype of N class
p_1' p_2' ...... p_n'
Linear
Concat
Scaled Dot-Product Attention
Q
K
V
Linear
Linear
Linear
The initial prototype of N class
p_1 p_2 ...... p_n
MatMul
Softmax
Mask
Scale
MatMul
Q
K
V

Figure 5. Schematic diagram of the prototype enhancement process. The initial prototypes of N categories are input into a multi-head self-attention mechanism to obtain enhanced prototypes. These enhanced prototypes are then combined with the initial prototypes to form the final prototypes.

4.3.3. Final Prototype

To further enhance the expressiveness and discriminability of prototypes, SACT adjusts the positions of the prototypes by utilizing relation information to better represent the features of the data. Inspired by adaptive prototype space fusion [51], SACT combines the relation information with the enhanced prototypes to obtain the final prototype representation, as shown in Equation (6):

$$p_i^{\text{final}} = \varepsilon \times p_i^{\text{enhanced}} + \beta \times \mathcal{R}_i^{\text{final}} \tag{6}$$

where ε and β are two learnable weights that depend on the importance of the relation information and the prototype representation to the final prototype.

4.4. Contrastive–Center Loss

Neural networks employ a loss function to measure the discrepancy between the model's output and the ground truth. The model's weights are then adjusted based on this disparity to update the network. However, defining an appropriate loss function is a challenging problem when tackling FSRE tasks where the goal is to generate separable representations for new classes. This paper introduces a novel loss function, namely, the contrastive–center loss function, designed to enhance the discriminability of samples from different relationship categories by encouraging both aggregation and dispersion in the feature space. It aims to tightly cluster samples of the same relationship category in the feature space while dispersing samples from different relationship categories, thereby improving the model's distinctiveness and classification performance. This enhancement makes a significant contribution to achieving better relationship classification results with the model.

The center loss function was initially introduced by Wen et al. [52] in 2016 to address face recognition tasks. Unlike the traditional cross-entropy loss function, the center loss function calculates the distance of each sample from its respective class center. It penalizes the distance between the current sample's feature vector and its assigned class center vector during each forward pass, aiming to minimize the distance within the same category. Equation (7) represents the distance in the center loss function:

$$\mathcal{L}_{center} = \frac{1}{2}\sum_{i}^{n}\|s_i - c_{k^i}\|_2^2 \tag{7}$$

where $\mathcal{L}_c$ represents the center loss. S_i denotes the i-th training sample. k^i represents the label of S_i. $c_{k^i} \in R^d$ signifies the feature center of the k^i class's deep feature in each mini-batch, where d denotes the feature dimension. In the training of deep neural networks, using a single loss function might not fully optimize the model. Therefore, Wen et al. [52] combined softmax with center loss to train the network when addressing facial recognition problems. The formula is as described in Equation (8):

$$\mathcal{L} = \mathcal{L}_{softmax} + \lambda\mathcal{L}_{center} \tag{8}$$

where $\mathcal{L}$ represents the final loss function of the network, $\mathcal{L}_s$ stands for softmax loss, and λ is used to adjust the weighting between the two.

However, the center loss is primarily focused on adjusting the distances between samples of the same class to bring them closer together, without considering the distances between different classes. In this paper, using a prototype network to address FSRE, SACT aims for not only the reduction in intra-class distances but also to enhance inter-class distances, making the feature distributions between classes more discriminative. Merely shifting feature vectors toward class centroids proves inadequate for this task. Therefore, SACT combines the advantages of contrastive loss and center loss to propose a new loss function, the contrastive–center loss function, as shown in Figure 6. Specifically, the training samples are compared with their corresponding class centers and non-corresponding class centers, penalizing the contrast between them. This encourages intra-class compactness while promoting inter-class separability, as shown in Equation (9):

$$\mathcal{L}_{ct-center} = \frac{1}{2}\sum_{i=1}^{m}\frac{\left\|s_i - c_{y_i}\right\|_2^2}{\left(\sum_{j=1,j\neq y_i}^{k}\left\|s_i - c_{y_i}\right\|_2^2\right) + \delta} \tag{9}$$

where $\mathcal{L}_{ct-center}$ represents the contrastive–center loss function, and m denotes the amounts of samples in a mini-batch. In the experiments conducted in this paper, a constant δ is introduced to avoid division by zero. Specifically, the default value of δ was set to 1 in this paper.

During the training process, the class centers c_{y_i} are continually updated in each mini-batch. In contrast to the center loss, our approach induces a more discrete distribution of class centers through the contrastive–center loss. When the distances between different categories are too small, it can negatively impact the effectiveness of classification and should be penalized. In this manner, the prototype network can better learn the similarity relationships among the samples, clustering similar samples together to enhance the reliability of the prototype network. By updating the class centers, the network is better able to learn the feature representations for each category, thereby improving the performance of the classification task.

To obtain a more comprehensive and refined model optimization performance, SACT combines these two loss functions, cross-entropy loss and contrastive–center loss, and jointly trains the model to produce more discriminative feature representations and create clearer class boundaries in the feature space, as shown in Equation (10):

$$\mathcal{L}_{final} = -\log(p_i) + \gamma\mathcal{L}_{ct-center} \quad (10)$$

where p_i stands for the probability of the query instance being part of class i, and γ serves as a hyperparameter that regulates the weight allocation between the two loss functions during training.

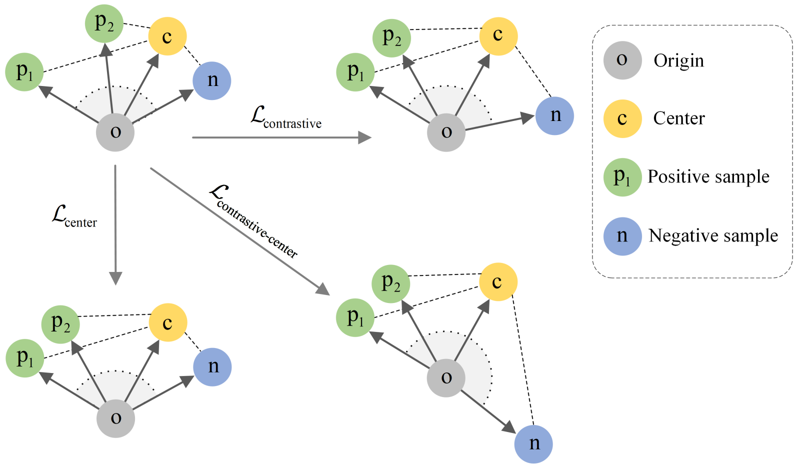

Figure 6. Diagram of the contrastive–center loss. In the upper-left corner, we have the basic prototype representations. Through the influence of the center loss function, you can observe a significant reduction in the distance between positive samples. After being affected by the contrastive loss function, there is some increase in the distance between class centers and negative samples. However, the contrast–center loss function used by SACT not only reduces the distance between positive samples but also increases the distance between centers and negative samples.

5. Experimental Settings

5.1. Dataset

To validate the effectiveness of SACT, the authors of this paper followed previous research and evaluated SACT on two commonly used FSRE datasets, FewRel 1.0 [42] and FewRel 2.0 [43]. FewRel 1.0, made available by Tsinghua University, is an expansive dataset designed for FSRE and is sourced from Wikipedia. It encompasses 100 distinct relations, each with 700 annotated instances. The dataset comprises training, testing, and validation subsets, featuring 65, 16, and 20 unique relations, respectively. Notably, there is no overlap in the relations between the training, testing, and validation sets. Please be aware that the validation set is not publicly accessible, but researchers can obtain the test scores by submitting their models via official testing scripts. However, it is worth noting that all the instances in the FewRel 1.0 dataset originate from a single domain's corpus. In real-world scenarios, data often span various domains such as medicine, education, biology, and cybersecurity, each with distinct syntax and content. These differences can potentially influence the performance of the model. To assess the performance of SACT in cross-domain scenarios, this paper also utilizes the FewRel 2.0 dataset. FewRel 2.0 extends upon FewRel 1.0 by including 25 additional relations from the biomedical domain, with each relation consisting of 100 instances as its validation set. The details of the dataset are provided in Table 1.

The authors of this paper evaluated the model's accuracy in the following four FSL task settings: 5-way-1-shot, 5-way-5-shot, 10-way-1-shot, and 10-way-5-shot.

Table 1. Details of the FewRel 1.0 dataset.

Corpus	Task	#Relation	#Entity	#Sentences	#Test
FewRel 1.0	Train	64	89,600	44,800	-
	Validation	16	22,400	11,200	-
	Test (unpublished)	20	28,000	14,000	10,000
FewRel 2.0	Validation	10	2000	1000	-
	Test (unpublished)	15	3000	1500	10,000

5.2. Baselines

The authors of this paper compared SACT against the following baseline models: (1) Proto-CNN [41]: Utilizes a CNN-based encoder in a prototype network. (2) Proto-HATT [45]: Introduces an attention mechanism at both the instance level and feature level, which is built upon the prototype network. (3) MLMAN [44]: Utilizing an innovative hierarchical network, performs feature matching between different levels and subsequently aggregates the results. Takes into account both local and instance-level matching information simultaneously. (4) Proto-BERT [41]: Employs a prototype net-work with BERT as the sentence encoder. (5) TD-Proto [16]: Incorporates textual descriptions of entities and relationships into the prototype network. (6) ConceptFERE [47]: Introduces the inherent concept of entities as external knowledge and fully leverages their essential attributes. (7) HCPR [17]: Utilizes a contrastive learning framework with relation label information (8) DRK [46]: Adopts a rule-based discriminative knowledge approach to mitigate adverse effects stemming from entity type feature confusion. (9) DAPL [53]: Utilizes the shortest dependency path information between entities in the prototype network. (10) SimpleF-SRE [18]: Concatenates the two representations of relational information and directly incorporates them into the prototype representation. (11) CP [48]: Utilizes the entity-masking contrast pre-training framework by randomly masking entity references. (12) MapRE [49]: Combines label-agnostic semantic information with label-aware information to consider the semantic knowledge of the relationship. (13) LPD [54]: Improves the utilization of textual labels by employing a random deletion approach for relationship label prompts. (14) CBPM [20]: Utilizes adaptive local loss based on relational similarity in a network prototype. (15) BERT-Pair [43]: BERT model based on sequence.

5.3. Implementation Details

The experiments in this paper were conducted in the Python 3.7.13 environment, employing PyTorch 1.9.1. The instance encoder employed in this paper utilized “bert-base-uncased” as the pre-trained parameters for the BERT model and used AdamW as the optimizer for SACT. This paper employed two different backbone models, namely, BERT and CP, for comparison with other baseline models to showcase the efficacy of SACT. The specific hyperparameter settings are detailed in Table 2.

Table 2. List of specific hyperparameter settings.

Parameter	Value
Encoder	**BERT**
Backend model	Bert /cp
Learning_rate	$1 \times 10^{-5}/5 \times 10^{-6}$
Max_length	128
Hidden_size	768
Batch_size	4
Optimizer	AdamW
Validation_step	1000
Max training iterations	30,000

5.4. Main Results

The authors of this paper conducted a comprehensive evaluation of SACT on both the FewRel 1.0 and FewRel 2.0 datasets, utilizing accuracy as the primary performance metric. The authors of this paper employed the traditional FSRE setting, specifically the N-way-K-shot setting, and the results are presented in Tables 3 and 4. Table 3 displays the outcomes of SACT on the validation and test sets of FewRel 1.0. In terms of the encoders, this paper includes models based on both CNN encoders and BERT encoders. For the part based on BERT encoders, the upper section corresponds to the BERT-based backend model, while the lower section corresponds to the CP-based backend model. Based on the data in Table 3, the authors of this paper draw the following conclusions:

Table 3. Accuracy (%) of FSRE on the FewRel 1.0 validation/test set.

Encoder	Model	5-Way-1-Shot	5-Way-5-Shot	10-Way-1-Shot	10-Way-5-Shot
CNN	Proto-CNN	72.65/74.52	86.15/88.40	60.13/62.38	76.20/80.45
	Proto-HATT	75.01/——	87.09/90.12	62.48/——	77.50/83.05
	MLMAN	79.01/——	88.86/92.66	67.37/75.59	80.07/87.29
BERT	Proto-BERT	84.77/89.33	89.54/94.13	76.85/83.41	83.42/90.25
	TD-proto	——/84.76	——/92.38	——/74.32	——/85.92
	ConceptFERE	——/89.21	——/90.34	——/75.72	——/81.82
	DAPL	——/85.94	——/94.28	——/77.59	——/89.26
	HCRP (BERT)	90.90/93.76	93.22/95.66	84.11/89.95	87.79/92.10
	DRK	——/89.94	——/92.42	——/81.94	——/85.23
	SimpleFSRE	91.29/94.42	94.05/96.37	86.09/90.73	89.68/93.47
	Ours (BERT)	92.31/94.83	94.05/97.07	86.92/90.46	89.36/93.65
	CP	——/95.10	——/97.10	——/91.20	——/94.70
	MapRE	——/95.73	——/97.84	——/93.18	——/95.64
	HCRP (CP)	94.10/96.42	96.05/97.96	89.13/93.97	93.10/96.46
	LPD	93.51/95.12	94.33/95.79	87.77/90.73	89.19/92.15
	CBPM	——/90.89	——/94.68	——/82.54	——/89.67
	Ours (CP)	96.48/97.14	97.93/97.98	93.88/95.24	95.61/96.27

Table 4. Accuracy (%) of FSRE on the FewRel 2.0 domain adaptation test set.

Model	5-Way-1-Shot	5-Way-5-Shot	10-Way-1-Shot	10-Way-5-Shot
Proto-CNN *	35.09	49.37	22.98	35.22
Proto-BERT *	40.12	51.50	26.45	36.93
BERT-PAIR *	56.25	67.44	43.64	53.17
Proto-CNN-ADV *	42.21	58.71	28.91	44.35
Proto-BERT-ADV *	41.90	54.74	27.36	37.40
HCRP	76.34	83.03	63.77	72.94
Ours (CP)	81.28	88.92	68.18	79.03

* Representative results from FewRel rankings.

The BERT encoder exhibits greater competitiveness compared to the CNN encoder. Figure 7 illustrates the outcomes of the three CNN-based encoder models compared to SACT on FewRel 1.0. The figure distinctly shows that the precision of the models when using BERT as an encoder significantly outperforms the models using CNN as an encoder. In terms of the average accuracy, BERT-based models achieve an impressive 83.42%, while CNN-based models lag at 71.74%. This demonstrates the superiority of BERT as an instance encoder, as it can more effectively capture and represent the semantic information of instances.

SACT has achieved significant improvements in few-shot relation extraction tasks. Figure 8 illustrates the comparative analysis between SACT and the suboptimal model (HCPR) on the FewRel 1.0 dataset under the 1-shot setting (specifically, 5-way-1-shot and 10-way-1-shot settings). As depicted in the diagram, SACT demonstrates notably high accuracy in the 1-shot setting, surpassing the baseline models. Specifically, in the two few-shot settings, SACT (BERT) and SACT (CP) achieved accuracy improvements of over 0.72% and 1.27%, respectively, compared to the currently best-performing models. This result indicates that SACT performs well in the few-shot scenario. In other words, the SACT (CP) model demonstrates superior performance and generalization when dealing with a limited number of samples.

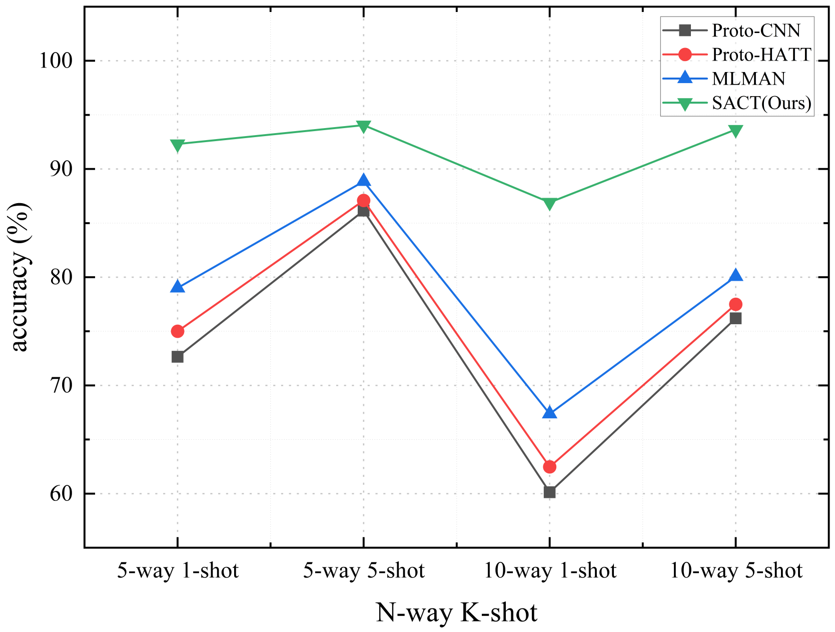

Figure 7. Comparison of three CNN encoder-based models with SACT on the FewRel 1.0 dataset.

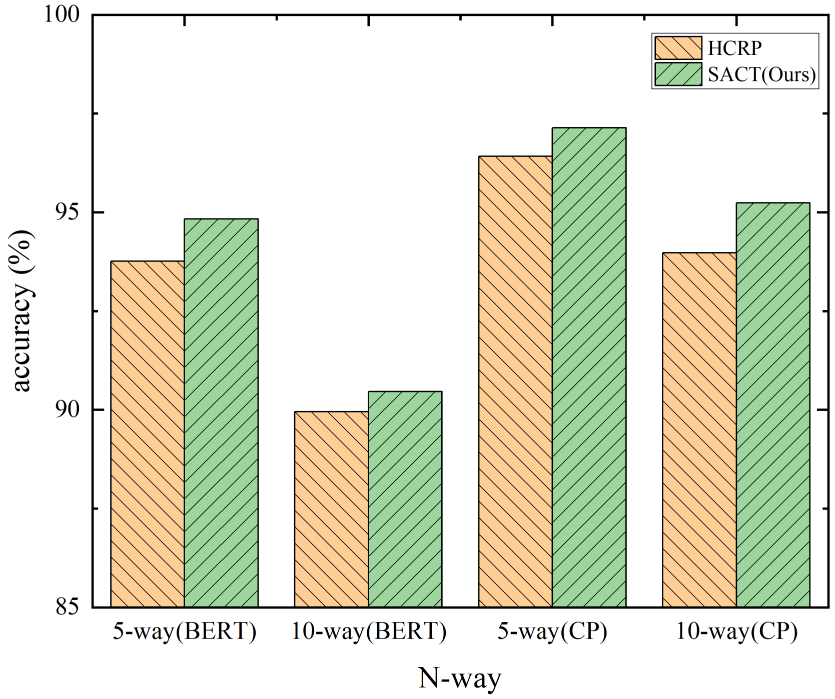

Figure 8. Comparison between HCPR and SACT in 1-shot setting.

SACT demonstrates greater robustness compared to previous methods. Figures 9 and 10 provide a comparative analysis of SACT alongside other BERT-based and CP-based models on the FewRel 1.0 dataset. The SACT model incorporates prototype refinement through a multi-head self-attention mechanism and incorporates a contrast–center loss in its design.

As evident from the figures, this strategy confers substantial advantages to SACT when compared to other BERT-based and CP-based models. In the four meta-tasks based on BERT, SACT achieved accuracies of 94.83%, 97.07%, 90.46%, and 93.65%, respectively. In the four meta-tasks based on CP, SACT achieved accuracies of 97.14%, 97.98%, 95.24%, and 96.27%, respectively. Therefore, the SACT model has attained the SOTA level in the contemporary research field. This result signifies significant accomplishments in the model design and optimization strategies, providing an effective solution for RE tasks.

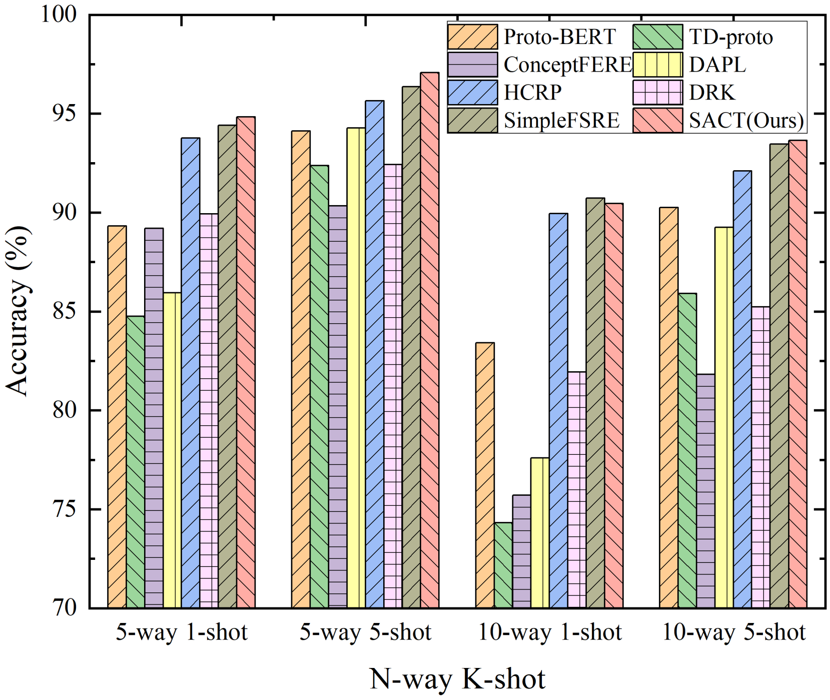

Figure 9. Comparison of SACT with other BERT-based models on the FewRel 1.0 dataset.

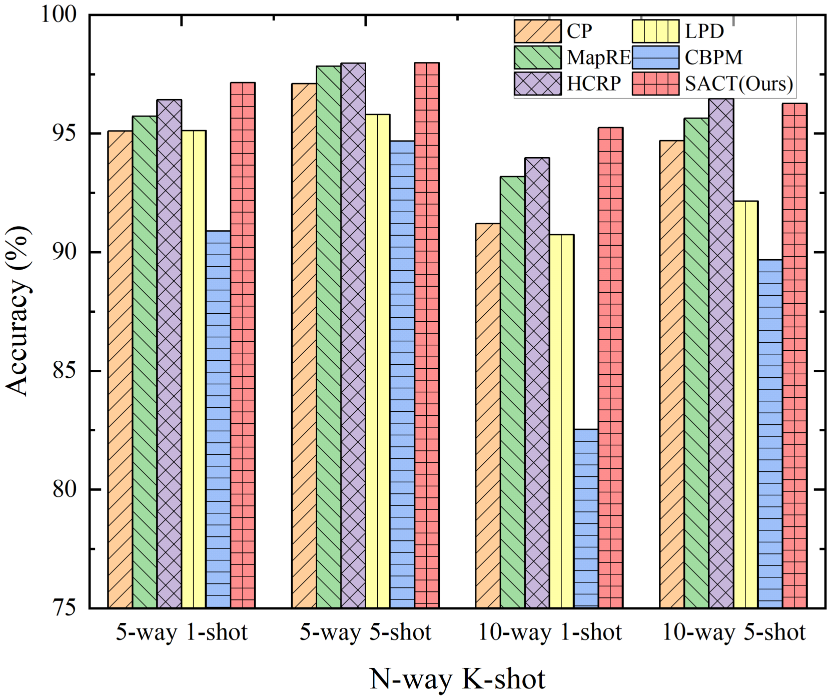

Figure 10. Comparison of SACT with other CP-based models on FewRel 1.0 dataset.

SACT demonstrates outstanding expressiveness compared to other prototype networks. Figure 11 illustrates the comparison between SACT and the existing prototype network models on the FewRel 1.0 dataset. Through this comparison, we observe a significant improvement in SACT's performance across all configurations, particularly excelling in enhancing model accuracy. This outcome strongly validates the effectiveness and superiority of our proposed approach, providing compelling evidence that our method enhances the expressiveness of prototype networks and propels the development of prototype network models in various application domains.

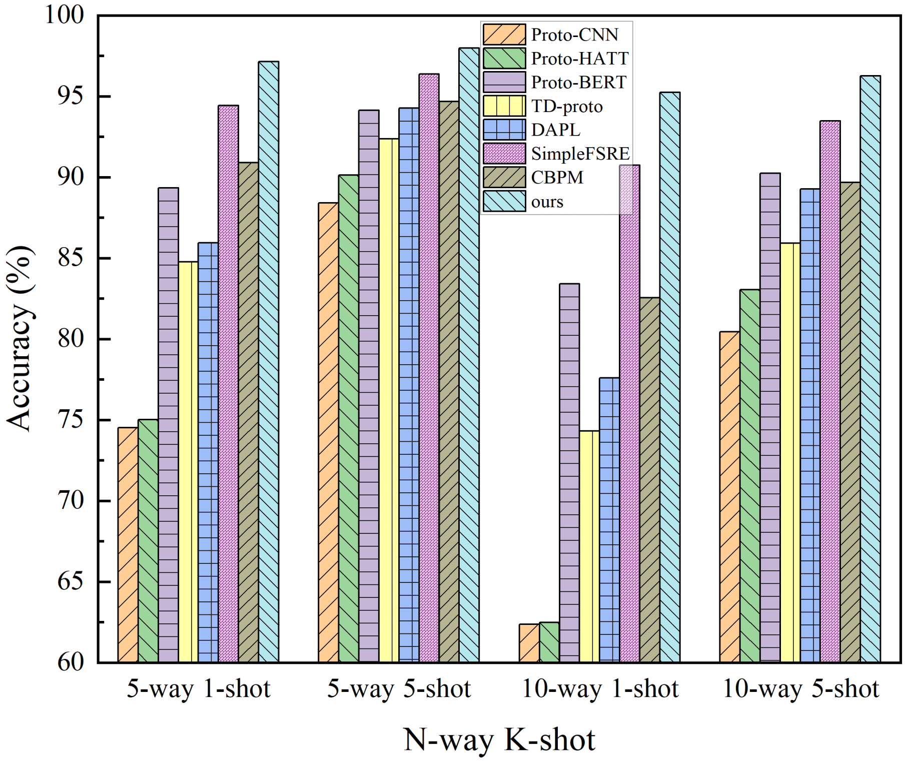

Figure 11. Comparison of SACT with other prototype network models on FewRel 1.0 dataset.

5.5. Domain Adaptation Results

In the general domain, the SACT model achieved excellent performance. To assess the domain adaptability of the SACT model, its performance was tested in the biomedical domain using the FewRel 2.0 dataset as the test set. Figure 12 shows a comparative analysis between SACT and the other models on the FewRel 2.0 dataset. The figure conspicuously illustrates that SACT outperforms the other models on the FewRel 2.0 dataset. The performance of the SACT model exhibits a noteworthy superiority over the other models, which signifies that SACT has a higher accuracy and generalization ability on the RE task. As evident from the outcomes presented in Table 4, the SACT model has demonstrated substantial accuracy enhancements of 4.94%, 5.89%, 4.41%, and 6.09% across four tasks within the biomedical domain when compared to the runner-up model (HCPR). These findings furnish compelling evidence for the effectiveness and robust domain adaptability of SACT.

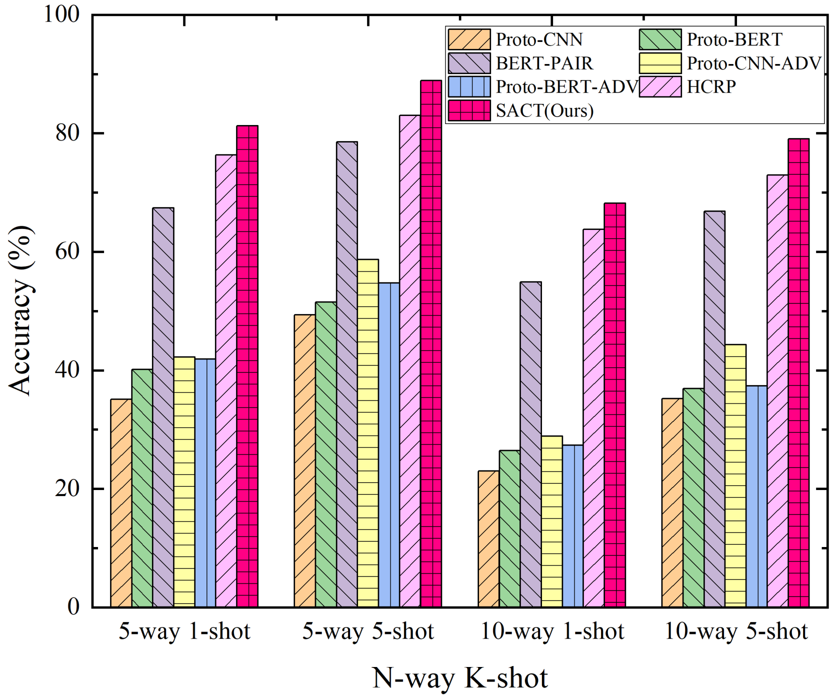

Figure 12. Comparison of SACT with other models on the FewRel 2.0 dataset.

5.6. Ablation Study

To validate the impact of the multi-head self-attention mechanism on prototype network enhancement and the effectiveness of the contrastive–center loss function, a thorough set of ablation experiments was conducted on the FewRel 1.0 dataset. These experiments covered both the 5-way-1-shot and 10-way-1-shot scenarios. The experimental outcomes are presented in Table 5. In these experiments, the authors of this paper designed several experimental variants. Here, 'w/o modification prototype' indicates that only the basic prototype network was used without introducing the multi-head self-attention mechanism to enhance the prototype network. In addition, 'w/o Contrastive-center loss' represents experiments using the loss function opposite to the contrastive–center loss.

Table 5. The outcomes of the ablation analysis for SACT.

Model	5-Way-1-Shot	10-Way-1-Shot
SACT	96.48	93.88
w/o modification prototype	94.89	87.07
w/o Contractive-center loss	94.86	87.47

The multi-head self-attention mechanism and the contrastive–center loss function, as evidenced by the data in Table 5, play crucial roles in improving model accuracy. In both task settings, the simultaneous removal of both the multi-head self-attention mechanism and the contrastive–center loss function resulted in a substantial decline in the model's performance. This further validates the effectiveness of SACT. Specifically, substituting the initial prototype network with the prototype network enhanced by the multi-head self-attention mechanism resulted in a reduction in model accuracy by 1.59% and 6.81% in the 5-way-1-shot and 10-way-1-shot scenarios, respectively. The importance of the multi-head self-attention mechanism is clearly demonstrated. Furthermore, replacing the contrastive–center loss function with the cross-entropy (CE) loss function results in a discernible decrease in model accuracy by 1.62% and 6.41%, respectively, under identical conditions. This further substantiates the efficacy of the contrast–center loss function.

Furthermore, compared to the 5-way-1-shot setting, the decline in model accuracy was more pronounced in the 10-way-1-shot setting, as further observation reveals. This may be attributed to the fact that the 10-way-1-shot task setting involves a greater variety of tasks, providing the model with a broader range of task samples. The importance of the multi-head self-attention mechanism and contrastive–center loss function is further validated by this observation. These mechanisms empower the model to adapt and learn different types of tasks, effectively capturing both the commonalities and differences among them.

6. Conclusions

This paper introduces SACT, a novel FSRE model. SACT incorporates a multi-head self-attention mechanism to capture inherent relationships among different category prototypes, addressing the limited generalization capabilities of traditional prototype networks when faced with new relationships. In addition, SACT employs an adaptive prototype fusion technique that combines relational information with enhanced prototypes, enhancing the overall performance of the prototype network. Moreover, SACT introduces a novel loss function, the contrastive–center loss. It effectively tightens the feature vectors of samples from the same class by maximizing the angle between negative pairs. This enhances sample distribution, improving intra-class closeness and inter-class distinguishability. This innovative loss function boosts the learning capacity of our model, offering a fresh perspective for RE research. The experimental results on the FewRel 1.0 and FewRel 2.0 datasets present the superior performance of SACT. SACT not only holds significant relevance for FSRE but also carries broad potential implications for the fields of NLP and RE. It provides valuable insights for future research endeavors in these domains.

However, there are still some limitations to acknowledge. Firstly, SACT does not consider the “none of the above” category, leading to insufficient classification capabilities for similar relationships. Secondly, the model relies solely on predefined relationship categories, making it inflexible to adapt to new or unknown relationship types. In future research, we plan to integrate the SACT model with other large-scale language models possessing robust text comprehension and representation learning capabilities, such as ChatGPT-4, to address SACT’s limitations and enhance its classification capabilities for similar and unknown relationship categories. Additionally, we will explore how to apply the SACT model to other NLP tasks, such as few-shot named entity recognition and few-shot text classification.

Author Contributions: Conceptualization, J.M. and J.C.; investigation, Z.S. and Y.C.; software, J.C. and K.L.; writing—original draft preparation, J.C.; validation, F.Z. and Z.S.; writing—review and editing, J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partly funded by the Songshan Laboratory Pre-research Project (YYJC0120 -22023), Henan Province Science Foundation (23230 0420150), Henan Province Collaborative Innovation Center for Internet Medical and Health Services (IH2019006), Zhengzhou University of Light Industry (2021ZCKJ214), Open Foundation of Henan Key Laboratory of Cyberspace Situation Awareness (HNTS2022005), and the Science and Technology Innovation Project of Zhengzhou University of Light Industry (23XNKJTD0205)

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/thunlp/FewRel.

Acknowledgments: We thank the reviewers for their valuable comments on this paper, which enhanced the quality of the manuscript. At the same time, we sincerely appreciate the guidance and suggestions provided by Yaqiong Qiao during the process of revising our manuscript. Her contributions have greatly enhanced the quality of our work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lauriola, I.; Lavelli, A.; Aiolli, F. An introduction to deep learning in natural language processing: Models, techniques, and tools. *Neurocomputing* **2022**, *470*, 443–456. [CrossRef]
2. Xiao, G.; Corman, J. Ontology-Mediated SPARQL Query Answering over Knowledge Graphs. *Big Data Res.* **2021**, *23*, 100177. [CrossRef]
3. Garcia, X.; Bansal, Y.; Cherry, C.; Foster, G.; Krikun, M.; Johnson, M.; Firat, O. The Unreasonable Effectiveness of Few-shot Learning for Machine Translation. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; Volume 202, pp. 10867–10878.
4. Lawrie, D.; Yang, E.; Oard, D.W.; Mayfield, J. Neural Approaches to Multilingual Information Retrieval. In *Advances in Information Retrieval*; Kamps, J., Goeuriot, L., Crestani, F., Maistro, M., Joho, H., Davis, B., Gurrin, C., Kruschwitz, U., Caputo, A., Eds.; ECIR: Cham, Switzerland, 2023; pp. 521–536.
5. Wang, Y.; Ma, W.; Zhang, M.; Liu, Y.; Ma, S. A Survey on the Fairness of Recommender Systems. *ACM Trans. Inf. Syst.* **2023**, *41*, 1–43. [CrossRef]
6. Mintz, M.; Bills, S.; Snow, R.; Jurafsky, D. Distant supervision for relation extraction without labeled data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Suntec, Singapore, 2–7 August 2009; pp. 1003–1011.
7. Ye, Q.; Liu, L.; Zhang, M.; Ren, X. Looking Beyond Label Noise: Shifted Label Distribution Matters in Distantly Supervised Relation Extraction. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 3841–3850. [CrossRef]
8. Zhang, N.; Deng, S.; Sun, Z.; Wang, G.; Chen, X.; Zhang, W.; Chen, H. Long-tail Relation Extraction via Knowledge Graph Embeddings and Graph Convolution Networks. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, 2–7 June 2019; pp. 3016–3025. [CrossRef]
9. Luo, X.; Zhou, W.; Wang, W.; Zhu, Y.; Deng, J. Attention-Based Relation Extraction With Bidirectional Gated Recurrent Unit and Highway Network in the Analysis of Geological Data. *IEEE Access* **2018**, *6*, 5705–5715. [CrossRef]
10. Li, Y.; Long, G.; Shen, T.; Zhou, T.; Yao, L.; Huo, H.; Jiang, J. Self-attention enhanced selective gate with entity-aware embedding for distantly supervised relation extraction. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 8269–8276.
11. Lin, X.; Liu, T.; Jia, W.; Gong, Z. Distantly Supervised Relation Extraction using Multi-Layer Revision Network and Confidence-based Multi-Instance Learning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 17 April 2021; pp. 165–174. [CrossRef]
12. Augenstein, I.; Maynard, D.; Ciravegna, F. Relation Extraction from the Web Using Distant Supervision. In *Knowledge Engineering and Knowledge Management*; Janowicz, K., Schlobach, S., Lambrix, P., Hyvönen, E., Eds.; Springer: Cham, Switzerland, 2014; pp. 26–41.
13. Sun, Q.; Liu, Y.; Chua, T.S.; Schiele, B. Meta-Transfer Learning for Few-Shot Learning. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 403–412. [CrossRef]
14. Lee, H.y.; Li, S.W.; Vu, T. Meta Learning for Natural Language Processing: A Survey. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, DC, USA, 10–15 July 2022; pp. 666–684. [CrossRef]
15. Mettes, P.; van der Pol, E.; Snoek, C.G.M., Hyperspherical Prototype Networks. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Curran Associates Inc.: Red Hook, NY, USA, 2019.
16. Yang, K.; Zheng, N.; Dai, X.; He, L.; Huang, S.; Chen, J. Enhance prototypical network with text descriptions for few-shot relation classification. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Shanghai, China, 19–23 October 2020; pp. 2273–2276.
17. Han, J.; Cheng, B.; Lu, W. Exploring Task Difficulty for Few-Shot Relation Extraction. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 17 April 2021; pp. 2605–2616. [CrossRef]
18. Liu, Y.; Hu, J.; Wan, X.; Chang, T.H. A Simple yet Effective Relation Information Guided Approach for Few-Shot Relation Extraction. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, 22–27 May 2022; pp. 757–763. [CrossRef]
19. Liu, Y.; Hu, J.; Wan, X.; Chang, T.H. Learn from Relation Information: Towards Prototype Representation Rectification for Few-Shot Relation Extraction. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, DC, USA, 10–15 July 2022; pp. 1822–1831. [CrossRef]
20. Wen, M.; Xia, T.; Liao, B.; Tian, Y. Few-shot relation classification using clustering-based prototype modification. *Knowl.-Based Syst.* **2023**, *268*, 110477. [CrossRef]
21. Zelenko, D.; Aone, C.; Richardella, A. Kernel Methods for Relation Extraction. *J. Mach. Learn. Res.* **2002**, *3*, 1083–1106.
22. Deng, B.; Fan, X.; Yang, L. Entity relation extraction method using semantic pattern. *Jisuanji Gongcheng/Comput. Eng.* **2007**, *33*, 212–214.

23. Shlezinger, N.; Whang, J.; Eldar, Y.C.; Dimakis, A.G. Model-Based Deep Learning. *Proc. IEEE* **2023**, *111*, 465–499. [CrossRef]
24. Shen, Y.; Huang, X. Attention-Based Convolutional Neural Network for Semantic Relation Extraction. In Proceedings of the Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, 11–16 December 2016; pp. 2526–2536.
25. Wang, L.; Zhu, C.; de Melo, G.; Zhiyuan, L. Relation Classification via Multi-Level Attention CNNs. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016. [CrossRef]
26. Ebrahimi, J.; Dou, D. Chain based RNN for relation classification. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, CO, USA, 31 May–5 June 2015; pp. 1244–1249.
27. Nguyen, T.H.; Grishman, R. Combining Neural Networks and Log-linear Models to Improve Relation Extraction. *arXiv* **2015**, arXiv:1511.05926.
28. Li, F.; Zhang, M.; Fu, G.; Qian, T.; Ji, D.H. A Bi-LSTM-RNN Model for Relation Classification Using Low-Cost Sequence Features. *arXiv* **2016**, arXiv:1608.07720.
29. Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; Xu, B. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016. [CrossRef]
30. Huang, Y.Y.; Wang, W.Y. Deep Residual Learning for Weakly-Supervised Relation Extraction. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 1803–1807. [CrossRef]
31. Zeng, D.; Dai, Y.; Li, F.; Sherratt, R.S.; Wang, J. Adversarial learning for distant supervised relation extraction. *Comput. Mater. Contin.* **2018**, *55*, 121–136.
32. Qin, P.; Xu, W.; Wang, W.Y. Robust Distant Supervision Relation Extraction via Deep Reinforcement Learning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 2137–2147. [CrossRef]
33. Qin, P.; Xu, W.; Wang, W.Y. DSGAN: Generative Adversarial Training for Distant Supervision Relation Extraction. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 496–505. [CrossRef]
34. Santoro, A.; Bartunov, S.; Botvinick, M.; Wierstra, D.; Lillicrap, T. Meta-Learning with Memory-Augmented Neural Networks. In Proceedings of the 33rd International Conference on International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; Volume 48, pp. 1842–1850.
35. Mishra, N.; Rohaninejad, M.; Chen, X.; Abbeel, P. A Simple Neural Attentive Meta-Learner. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
36. Ren, M.; Liao, R.; Fetaya, E.; Zemel, R. Incremental few-shot learning with attention attractor networks. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 5275–5285.
37. Finn, C.; Abbeel, P.; Levine, S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, Sydney, Australia, 6–11 August 2017; pp. 1126–1135.
38. Elsken, T.; Staffler, B.; Metzen, J.; Hutter, F. Meta-Learning of Neural Architectures for Few-Shot Learning. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Los Alamitos, CA, USA, 13–19 June 2020; pp. 12362–12372. [CrossRef]
39. Koch, G.; Zemel, R.; Salakhutdinov, R. Siamese neural networks for one-shot image recognition. In Proceedings of the ICML Deep Learning Workshop, Lille, France, 6–11 July 2015; Volume 2.
40. Vinyals, O.; Blundell, C.; Lillicrap, T.; kavukcuoglu, K.; Wierstra, D. Matching Networks for One Shot Learning. In *Advances in Neural Information Processing Systems*; Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2016; Volume 29.
41. Snell, J.; Swersky, K.; Zemel, R. Prototypical Networks for Few-shot Learning. In *Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H.,Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2017; Volume 30.
42. Han, X.; Zhu, H.; Yu, P.; Wang, Z.; Yao, Y.; Liu, Z.; Sun, M. FewRel: A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018. [CrossRef]
43. Gao, T.; Han, X.; Zhu, H.; Liu, Z.; Li, P.; Sun, M.; Zhou, J. FewRel 2.0: Towards More Challenging Few-Shot Relation Classification. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 7 November 2019; pp. 6250–6255.
44. Ye, Z.X.; Ling, Z.H. Multi-Level Matching and Aggregation Network for Few-Shot Relation Classification. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 2872–2881. [CrossRef]
45. Gao, T.; Han, X.; Liu, Z.; Sun, M. Hybrid Attention-Based Prototypical Networks for Noisy Few-Shot Relation Classification. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'19/IAAI'19/EAAI'19, Honolulu, HI, USA, 27 January–1 February 2019; AAAI Press: Washington, DC, USA, 2019. [CrossRef]

46. Wang, M.; Zheng, J.; Cai, F.; Shao, T.; Chen, H. DRK: Discriminative Rule-based Knowledge for Relieving Prediction Confusions in Few-shot Relation Extraction. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022; pp. 2129–2140.
47. Yang, S.; Zhang, Y.; Niu, G.; Zhao, Q.; Pu, S. Entity Concept-enhanced Few-shot Relation Extraction. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Online, 1–6 August 2021; pp. 987–991. [CrossRef]
48. Peng, H.; Gao, T.; Han, X.; Lin, Y.; Li, P.; Liu, Z.; Sun, M.; Zhou, J. Learning from Context or Names? An Empirical Study on Neural Relation Extraction. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 19–20 November 2020; pp. 3661–3672. [CrossRef]
49. Dong, M.; Pan, C.; Luo, Z. MapRE: An Effective Semantic Mapping Approach for Low-resource Relation Extraction. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 17 April 2021; pp. 2694–2704. [CrossRef]
50. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, 2–7 June 2019; pp. 4171–4186. [CrossRef]
51. Liu, S.; Huang, D.; Wang, Y. Learning Spatial Fusion for Single-Shot Object Detection. *arXiv* **2019**, arXiv:1911.09516.
52. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A Discriminative Feature Learning Approach for Deep Face Recognition. In *Computer Vision–ECCV 2016, Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 499–515.
53. Yu, T.; Yang, M.; Zhao, X. Dependency-aware Prototype Learning for Few-shot Relation Classification. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022; pp. 2339–2345.
54. Zhang, P.; Lu, W. Better Few-Shot Relation Extraction with Label Prompt Dropout. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 6996–7006. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.