Efficient Information Extraction in Few-Shot Relation Classification through Contrastive Representation Learning

Anonymous ACL submission

Abstract

Differentiating relationships between entity pairs with limited labeled instances poses a significant challenge in few-shot relation classification. Representations of textual data extract rich information spanning the domain, entities, and relations. In this paper, we introduce a novel approach to enhance information extraction using multiple noisy representations and contrastive learning. While sentence representations in relation classification commonly combine information from entity marker tokens, we argue that substantial information within the internal model representations remains untapped. To address this, we propose aligning multiple noisy sentence representations, such as the [CLS] token, the [MASK] token used in prompting, and entity marker tokens. We employ contrastive learning to reduce the noise contained in the individual representations. We demonstrate the adaptability of our representation contrastive learning approach, showcasing its effectiveness for both sentence representations and additional data sources, such as relation description representations. Our evaluation underscores the efficacy of incorporating multiple noisy representations through contrastive learning, enhancing information extraction in settings where available data is limited.[1]

1 Introduction

Relation classification (RC) is an important subtask in the relation extraction framework. It entails identifying relation types that correspond to a pair of entities within a given textual context. Extracting relevant information is central to this task. To achieve this, RC models must distill rich information from sentences, including contextual cues, entity attributes, and relation characteristics. While language models are essential to extract representations from text, it is noteworthy that previous research has highlighted the suboptimal use of vector space in sentence representations (Ethayarajh, 2019). Recent advances have addressed this limitation by improving sentence representations through various techniques, including flow-based approaches (Li et al., 2020), whitening operations (Huang et al., 2021), prompting (Jiang et al., 2022), and contrastive learning (Gao et al., 2021; Kim et al., 2021; Zhou et al., 2022).

Relation extraction applications suffer from a long-tail of relation types characterized by limited data availability and disproportional data acquisition costs (Yang et al., 2021). To address this challenge, few-shot RC tasks models with quickly adapting to unseen relation types using only few labeled examples. Common approaches to this task include meta-learning and prototypical networks which leverage representation similarity to match unseen query instances with few labeled support instances (Snell et al., 2017). Recent research incorporated supplementary data to enrich model representations. Yang et al. (2021) and Qu et al. (2020) incorporate information from external knowledge bases, augmenting entity-related knowledge. Wang et al. (2020b) and Yu et al. (2022) utilize linguistic dependencies to integrate structural sentence information into the model. Textual relation descriptions provide an additional perspective on relation types, thereby enhancing the performance of prototypical networks (Han et al., 2021; Dong et al., 2021; Liu et al., 2022).

To capture contextual information in sentences, language models create representations of the textual data. Given the inherent complexity of distinguishing between various relation types, RC applications commonly combine representations of entity marker tokens as sentence representations (Baldini Soares et al., 2019; Dong et al., 2021). Additionally, recent work uses contrastive learning to create more discriminative representations in few-shot RC (Han et al., 2021; Zhang and Lu, 2022; Dong et al., 2021). Other studies suggest that rep-

[1] Our model is available at https://anonymous.4open.science/r/MultiRep-6E39.

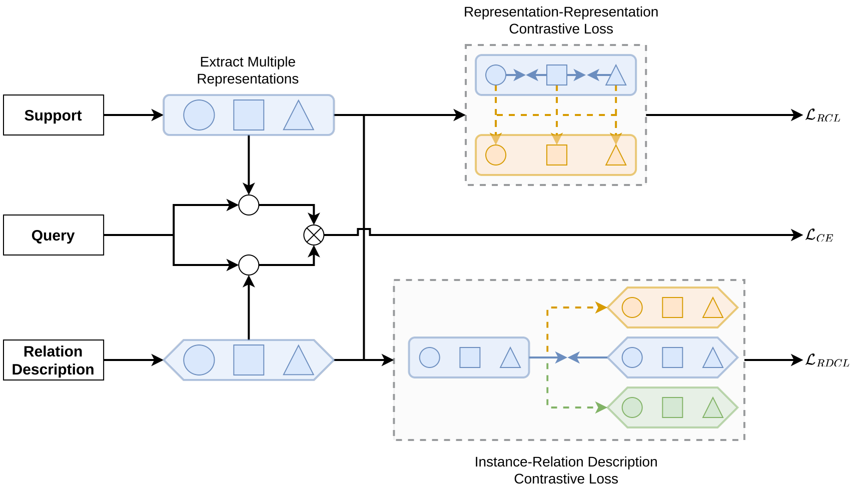
Representation-Representation Contrastive Loss
Extract Multiple Representations
Support
Query
Relation Description
$\mathcal{L}_{RCL}$
$\mathcal{L}_{CE}$
$\mathcal{L}_{RDCL}$
Instance-Relation Description Contrastive Loss

Figure 1: Overview of the MultiRep model, which integrates relation description information. The ∘ represents the vector dot product between the instance or relation description and the query samples, while the addition operation is denoted by ⊗. Attracting and repelling forces in contrastive learning are represented by →← and - - →, respectively.

resenting sentences with the [MASK] token through prompting improves sentence representations as they avoid embedding biases (Jiang et al., 2022).

In this study, we propose aligning multiple noisy sentence representations, such as the [CLS] token, the [MASK] token used in prompting, and entity marker tokens. Recognizing that encoder representations are compressed perspectives of the model's internal representations and consequently contain noise, we combine multiple noisy representations to construct richer sentence embeddings. To achieve alignment among these noisy representations, we employ contrastive learning, which aims to maximize the intra-sentence representation similarity. We concatenate the different representations to form the instance representation. This approach enriches sentence embeddings in two fundamental ways: (i) by merging multiple perspectives, it encapsulates more information obtained from the model's internal representations, and (ii) through the contrastive learning objective, it distills commonalities among the embeddings while reducing the impact of noise. A key advantage in our approach is the efficient utilization of resources, since all representations are derived from a single forward pass. We demonstrate that this approach can be extended to additional information sources, particularly relation descriptions. In summary, our contributions are:

- We introduce a novel methodology for information extraction in few-shot relation classification, which demonstrates how to align multiple noisy representations through contrastive learning.
- Our approach extends its utility to diverse information sources, including relation descriptions, showcasing its adaptability.
- We emphasize the resource-efficiency of our approach, streamlining the information extraction process while maintaining performance.

2 Approach

This section provides a detailed overview of our approach, as depicted in Figure 1.

2.1 Task Definition

In the N-way K-shot evaluation setting, episodes are randomly sampled from the training set. An episode consists of $N \times K$ input sentences x in the support set $S = \{(x_i, rel_i)\}_{i=1}^{N \times K}$ and $N \times K$ inputs from the query set $Q = \{x_i\}_{i=1}^{N \times K}$. The relations are randomly sampled from the relation types included in the training dataset. Importantly, the relation types in the training set are not overlapping with the test set (and validation set) $rel_{\text{train}} \cap rel_{\text{test}} = \varnothing$ (Gao et al., 2019).

2.2 Sentence Representations

In line with related work, we utilize the BERT-Base model (Devlin et al., 2019) to encode textual inputs. This model creates representations of $h = 768$ dimensions for each input token. Below, we elaborate on the methods used to create multiple sentence representations from the BERT encoder.

Average Pooling is a simple technique that involves computing sentence representations by averaging the token representations. Devlin et al. (2019) append the **[CLS]** token to all model inputs and employ its representation for next sentence prediction. The **entity marker** approach consists of augmenting the input sentence x with position markers that identify the tokens corresponding to the entities (Baldini Soares et al., 2019). This results in a modified input $\bar{x} = [x_0, ..., \texttt{[E1\_Start]}, x_i, \texttt{[E1\_End]}, ..., x_n]$. The sentence representation is constructed by concatenating the entity start marker representations `[E1_Start]` and `[E2_Start]` (Baldini Soares et al., 2019). In the prompting approach, the RC task is reformulated as a masked language modeling problem. With a template $\mathcal{T}$, each input is transformed into $x_{\text{prompt}} = \mathcal{T}(x)$ containing at least one **`[MASK]`** token. The masked token is interpreted as the relation label and predicted based on the context, i.e. $\bar{x}$ = `[MASK]`: x. (Schick and Schütze, 2021). Gao et al. (2021) use various **dropout** masks to create augmented representations with varying levels of noise. As entity marker representations are not available for relation descriptions, we instead use the prompting and `[CLS]` representations with different dropout masks.

2.3 Contrastive Representation Learning

The objective of our **representation-representation contrastive loss** term is reducing noise within in the sentence representations obtained from the encoder. A key difference to contrastive learning objectives in related work lies in our method of constructing positive instance pairs. In a single forward pass, we derive M different representations from each sentence, and consider these representations as positive pairs. Consequently, representations from other sentences in the training set serve as negative instance pairs. For a given representation r_i^m (where $m \in M$, $i \in N \times K$), we define positive instances r_i^+ and negative instances r_i^- as follows:

$$r_i^+ = \{r_i^{k \neq m} \,|\, k \in M\}$$

$$r_i^- = \{r_{j \neq i}^m \,|\, j \in N \times K\}$$

This aims to maximize the similarity between different representations of the same sentence and minimize the similarity to representations obtained from other sentences (van den Oord et al., 2019; Gao et al., 2021). It ensures that the differentiating factors encoded in the embeddings primarily reflect the underlying sentences, regardless of how these representations are derived from the internal model representations. The representation-representation contrastive loss is computed as follows:

$$\mathcal{L}_{RCL} = \sum_{i=1}^{N \times K} \sum_{m=1}^{M} -log \frac{exp\left(\phi(r_i^m, r_i^+)/\tau\right)}{exp\left(\phi(r_i^m, r_j^-)/\tau\right)},$$

where τ is a temperature scaling parameter, and $\phi(r_i^m, r_i^+)$ represents the element-wise cosine similarity $\sum_{k=1}^{M-1} r_i^m \cdot r_i^k / \|r_i^m\| \|r_i^k\|$ between representation r_i^m and each representation in r_i^+.

In the **instance-relation description contrastive loss**, we leverage the relation descriptions to maximize the similarity between instance representations and corresponding relation description representations. To construct the instance representations R_i and the relation description representations D_i, we concatenate all representations extracted from the encoder $R_i = [r_i^1; r_i^2; ...; r_i^M]$ and $D_i = [d_i^1; d_i^2; ...; d_i^M]$. For instance representation R_i, we select the corresponding relation description D^+ based on the label information in the support set. Non-corresponding relation descriptions D^- form negative pairs. The instance-relation description contrastive loss is computed as follows:

$$\mathcal{L}_{RDCL} = \sum_{i=1}^{N \times K} -log \frac{exp\left(\phi(R_i, D^+)/\tau\right)}{exp\left(\phi(R_i, D^-)/\tau\right)}$$

2.4 Relation Classification

We obtain N class prototypes by averaging the K instance representations in the support set. We compute the similarity between query instances and support prototypes using the vector dot product and selecting the most similar class prototype. For the relation description, we compute the similarity between query instances and relation description representations D. We add the similarities obtained from the relation descriptions with the similarities obtained from the class prototypes and select the most similar prototype and relation description. This is in line with Liu et al. (2022), who instead directly add the prototype and relation description representations. We compute the cross-entropy loss $\mathcal{L}_{CE} = -log\,(z_y)$, where z_y is the probability for class y. The total loss is defined as the sum of the individual loss terms $\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{RCL} + \mathcal{L}_{RDCL}$.

3 Experiments

3.1 Dataset and Evaluation

We conducted our experiments on the FewRel dataset, which consists of 700 instances for each of the 100 different relation types (Han et al., 2018). This dataset is derived from Wikipedia and is divided into training, validation, and test sets, encompassing 64, 16, and 20 relation types, respectively. The training process involves exposing MultiRep to a large number of episodes sampled from the

Model	**Relation Descriptions**	**5-1**	**5-5**	**10-1**	**10-5**	**Avg.**
Proto	-	- / 80.68	- / 89.60	- / 71.48	- / 82.89	- / 81.16
BERT-Pair	-	85.66 / 88.32	89.48 / 93.22	76.84 / 80.63	81.76 / 87.02	83.44 / 87.30
CTEG	-	84.72 / 88.11	92.52 / **95.25**	76.01 / 81.29	84.89 / 91.33	84.54 / 89.00
DAPL	-	- / 85.94	- / 94.28	- / 77.59	- / 89.26	- / 86.77
SimpleFSRE	-	84.77 / **89.33**	89.54 / 94.13	76.85 / 83.41	83.42 / 90.25	83.64 / 89.28
MultiRep (Ours)	-	**87.13** / 89.20	**92.93** / 95.09	**78.42 / 84.18**	**87.29 / 91.65**	**86.44 / 90.03**
TD-Proto	✓	- / 84.76	- / 92.38	- / 74.32	- / 85.92	- / 84.34
HCRP	✓	90.90 / 93.76	93.22 / 95.66	84.11 / 89.95	87.79 / 92.10	89.01 / 92.87
SimpleFSRE	✓	91.29 / **94.42**	**94.05 / 96.37**	86.09 / 90.73	**89.68 / 93.47**	90.28 / **93.75**
MultiRep (Ours)	✓	**92.73** / 94.18	93.79 / 96.29	**86.12 / 91.07**	88.80 / 91.98	**90.36** / 93.38

Table 1: Accuracy on the FewRel validation / test set.

Model	**5-1**	**10-1**
MultiRep	92.73	86.12
w/o $\mathcal{L}_{RCL}$	92.08	85.95
w/o $\mathcal{L}_{RDCL}$	90.14	84.16
w/o Avg. Pooling	92.26	85.82
w/o Entity Marker	91.90	84.83
w/o [CLS]	91.35	85.51
w/o [MASK]	91.87	85.80
w/ prototype addition	91.75	85.82

Table 2: Model variants with (w/) or without (w/o) indicated representations and architectural changes evaluated on the FewRel validation set.

training set. Model performance is subsequently evaluated on previously unseen data from the validation and test sets. MultiRep was trained for 30,000 iterations on the FewRel training set with a batch size of 4 and a learning rate of 2e-5.

3.2 Results

We present the results of our MultiRep approach and compare them to relevant benchmark models designed for few-shot RC, some of which incorporate relation descriptions as additional information. For consistency, all benchmarked models use BERT-Base (Devlin et al., 2019) as the sentence encoder. The benchmark models include Proto (Gao et al., 2019), BERT-Pair (Gao et al., 2019), TD-Proto (Yang et al., 2020), CTEG (Wang et al., 2020a), DAPL (Yu et al., 2022), HCRP (Han et al., 2021), and SimpleFSRE (Liu et al., 2022).

Our model evaluation results are summarized in Table 1. We analyze these results for two distinct scenarios: (i) models that do not incorporate additional information, and (ii) models that incorporate relation description information. We observe that MultiRep outperforms existing models, particularly in settings where information is limited. Specifically, this includes scenarios where relation description information is unavailable, as well as 1-Shot settings in the presence of relation description information. To validate the importance of individual components in the MultiRep model, we conducted ablation studies and present the results in Table 2. These results are based on the MultiRep model that incorporates relation description information, evaluated on the FewRel validation set in the 5-Way 1-Shot and 10-Way 1-Shot settings. Our findings indicate that removing the contrastive learning loss terms, $\mathcal{L}_{RCL}$ and $\mathcal{L}_{RDCL}$, substantially reduces model performance. Furthermore, removing individual representations from the MultiRep model has a negative impact on performance, and there are no specific representations that disproportionately affect the model's performance. Additionally, we validate our approach of computing separate instance prototypes and relation description prototypes, as compared to the direct prototype addition method introduced by Liu et al. (2022). Our results demonstrate that our approach yields the best model performance for MultiRep.

4 Conclusion

In this study, we propose aligning multiple noisy sentence representations for few-shot RC using contrastive learning to efficiently extract discriminative sentence representations. We demonstrate the adaptability of our representation contrastive learning approach, showcasing its effectiveness for both sentence representations and relation description representations. We demonstrate that our approach efficiently extracts relevant information from multiple sentence representations. It is particularly performant in low-resource settings, such as few-shot RC not including any additional data sources and 1-Shot scenarios. A key advantage of our approach lies in its efficient use of resources, achieved by obtaining all sentence representations from a single forward pass.

5 Limitations

Although our approach efficiently utilizes multiple sentence representations within a single forward pass, it is important to note that this involves combining these representations into larger vectors. This aggregation process may require additional memory and computational resources. Moreover, the application of contrastive learning comes with additional computational requirements. Our method is specifically designed for few-shot RC tasks, and its performance might vary when applied to different types of NLP tasks.

References

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Manqing Dong, Chunguang Pan, and Zhipeng Luo. 2021. MapRE: An effective semantic mapping approach for low-resource relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2694–2704, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. 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)*, pages 6250–6255, Hong Kong, China. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiale Han, Bo Cheng, and Wei Lu. 2021. Exploring task difficulty for few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2605–2616, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. 2021. WhiteningBERT: An easy unsupervised sentence embedding approach. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 238–244, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ting Jiang, Jian Jiao, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Denvy Deng, and Qi Zhang. 2022. PromptBERT: Improving BERT sentence embeddings with prompts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8826–8837, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-guided contrastive learning for BERT sentence representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2528–2540, Online. Association for Computational Linguistics.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Yang Liu, Jinpeng Hu, Xiang Wan, and Tsung-Hui Chang. 2022. A simple yet effective relation information guided approach for few-shot relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 757–763, Dublin, Ireland. Association for Computational Linguistics.

Meng Qu, Tianyu Gao, Louis-Pascal A. C. Xhonneux, and Jian Tang. 2020. Few-shot relation extraction via bayesian meta-learning on relation graphs. *International Conference on Machine Learning (ICML)*, pages 7867–7876.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical Networks for Few-shot Learning. *Advances in Neural Information Processing Systems*, page 11.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding.

Yingyao Wang, Junwei Bao, Guangyi Liu, Youzheng Wu, Xiaodong He, Bowen Zhou, and Tiejun Zhao. 2020a. Learning to decouple relations: Few-shot relation classification with entity-guided attention and confusion-aware training. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5799–5809, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yuxia Wang, Karin Verspoor, and Timothy Baldwin. 2020b. Learning from unlabelled data for clinical semantic textual similarity. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 227–233, Online. Association for Computational Linguistics.

Kaijia Yang, Nantao Zheng, Xinyu Dai, Liang He, Shujian Huang, and Jiajun Chen. 2020. Enhance prototypical network with text descriptions for few-shot relation classification. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 2273–2276, New York, NY, USA. Association for Computing Machinery.

Shan Yang, Yongfei Zhang, Guanglin Niu, Qinghua Zhao, and Shiliang Pu. 2021. 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)*, pages 987–991, Online. Association for Computational Linguistics.

Tianshu Yu, Min Yang, and Xiaoyan Zhao. 2022. Dependency-aware prototype learning for few-shot relation classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2339–2345, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Peiyuan Zhang and Wei Lu. 2022. Better few-shot relation extraction with label prompt dropout. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6996–7006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kun Zhou, Beichen Zhang, Xin Zhao, and Ji-Rong Wen. 2022. Debiased contrastive learning of unsupervised sentence representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6120–6130, Dublin, Ireland. Association for Computational Linguistics.

A Model Training

The MultiRep model consists of 109.48 million parameters and was trained on a single NVIDIA A6000 48GB GPU. The combined training and evaluation time for the 5-Way 1-Shot and 10-Way 5-Shot models, incorporating relation descriptions, was 16 hours and 25 hours, respectively.