VNU Journal of Science: Comp. Science & Com. Eng., Vol. 40, No. 2 (2024) 22–31

VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

Continual Extraction of Semantic Relations using Augmented Prototypes with Energy-based Model Alignment

Thanh Hai Dang, Quynh-Trang Pham Thi, Duc-Hung Nguyen, Tri-Thanh Nguyen, Duc-Trong Le*

VNU University of Engineering and Technology, 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam

Received 02 April 2024;
Revised 24 June 2024; Accepted 16 Dec 2024

Abstract: Continual relation extraction (CRE) is a critical task in natural language processing that aims to learn new relation types incrementally while preserving knowledge of previously learned relations. However, existing CRE models often struggle with catastrophic forgetting and inefficient utilization of memory. In this paper, we propose a CRE model that leverages class-specific prototypes and energy-based latent alignment to address these challenges. Our approach stores relation prototypes instead of real data points, enriching them with Gaussian noise during training. We incorporate contrastive learning to obtain effective representations for memory prototype data and introduce an Energy-based Latent feature space Alignment (ELI) module to mitigate representational shift across tasks. We evaluate our model on two benchmark datasets: FewRel, a balanced few-shot relation classification dataset, and TACRED, a large-scale imbalanced relation extraction dataset. Extensive experiments demonstrate that our proposed method consistently outperforms state-of-the-art CRE models across multiple tasks, with improvements of up to 4% over existing methods. This consistent superior performance highlights our model effectiveness in addressing the challenges of continual relation extraction, particularly in maintaining performance across a sequence of tasks while mitigating catastrophic forgetting.

Keywords: continual learning, contrastive learning, energy-based model

1. Introduction

Relation extraction (RE) is a fundamental task in natural language processing that aims to identify semantic relations between entities mentioned in text [1, 2]. For example, given the ”Location-of” relation and the sentence ”Hanoi, the capital of Vietnam, is known for its iconic landmark, the Temple of Literature.” with the entity pair [Hanoi, Temple of Literature] as an input, an RE model should recognize the given relation between these two entities.

*Corresponding author
E-mail address: trongld@vnu.edu.vn
https://doi.org/10.25073/2588-1086/vnucsce.2476

Traditional RE models are trained on fixed datasets with predefined relation types, limiting their ability to adapt to new relations that may emerge over time [3]. This poses a significant challenge in real-world applications where the number and type of relations can continuously evolve. To address this issue, the paradigm of continual relation extraction (CRE) has been proposed, enabling models to learn new relation types incrementally while preserving knowledge of previously learned relations [4, 5].

Continual learning, also known as lifelong machine learning or never-ending learning, is a novel machine learning paradigm that involves the continuous acquisition and execution of learning tasks [6]. It entails retaining acquired knowledge and selectively leveraging previously stored knowledge to adeptly address new learning tasks. This learning paradigm aims to propel machine learning into a new era, aspiring to emulate human-like adaptability while surmounting the limitations inherent in the isolated learning approach of traditional machine learning. However, catastrophic forgetting (CF), initially identified by McCloskey and Cohen (1989) [7], stands as a formidable challenge when tackling continual learning problems, especially with deep learning models. This phenomenon occurs when a model sequentially learns tasks, wherein learning a new task may significantly degrade the model's predictive performance on previously learned tasks.

The catastrophic forgetting in continual learning emerges due to two main reasons: (1) the parameter updates optimized for new tasks may not align well with older tasks, and (2) the resulting latent representation shift occurs when the model's latent feature space evolves during the learning of new tasks, leading to a discrepancy between the representations of old and new tasks. This shift in the latent space is a critical factor contributing to the CF in continual learning.

Recent advancements in CL have paved the way for the development of CRE models [8, 9]. Existing CL methods can be broadly categorized into three main types: (1) regularization methods [10], (2) dynamic architecture methods [11], and (3) memory-based methods [12]. Among these, the memory-based approach has shown promising results by storing a subset of representative samples from previous tasks and replaying them during the learning of new tasks to mitigate CF[13-16]. However, existing memory-based CRE models often struggle to effectively utilize the limited memory capacity and suffer from embedding space shift when incorporating new relation types [13, 14].

In this paper, we propose a memory-based CRE model that leverages relation prototypes to enhance the learning of consistent and informative representations across tasks. Instead of storing real data points in the memory buffer, our model replays them with their prototypes, which are constructed from typical data samples selected through clustering techniques [5, 14]. These semantic relation prototypes serve as anchors for refining the embedding space and maintaining a stable understanding of both old and new relations [15]. Further, we utilize contrastive learning during the model training to obtain highly effective representations for memory data. Moreover, to address the representational shift problem in CRE, we utilize an Energy-based Latent feature space Alignment (ELI) module [16] for aligning the latent representations of old and new tasks.

We conduct extensive experiments on two benchmark datasets for CRE, namely FewRel [17] and TACRED [2], demonstrating the superior performance of our model compared to state-of-the-art CRE models, particularly in terms of long-term retention of previously learned tasks and robustness to the embedding space shift.

The remainder of this paper is organized as follows: Section 2 provides an overview of related work in contrastive learning and energy-based model. Section 3 describes our proposed CRE model in detail. Section 4 presents the

experimental setup and results. Finally, Section 5 is about conclusion and discussion of future research directions.

2. Related Work

2.1. Contrastive Learning

Contrastive learning is designed to bring similar samples together in the embedded space while pushing dissimilar samples farther apart [18]. In recent years, the surge in popularity of CL has marked substantial progress in self-supervised representation learning[19-22]. What sets these works apart is their shared characteristic of operating without labels, relying instead on forming positive and negative pairs through data augmentations.

Drawing inspiration from studies on supervised contrastive learning, Zhao et al. 2022 introduced an improved approach to Learning through the application of contrastive learning [19].

2.2. Energy-based Model (EBM)

In 2021, Liu et al., proposed using an "energy score" metric to detect out-of-distribution data (OOD), which are outside the training data distribution [20]. Previous methods often relied on the confidence scores from the softmax function for OOD [21]. However, neural networks can generate high softmax confidence scores even for data points significantly far from the training data distribution.

The method by Liu et al. employs an Energy-Based Model (EBM) that maps each input data point to a single scalar as an energy score, which is lower for trained data and higher for untrained data. The critical point is that the energy score can be computed from an existing classification model without relying on a generative model, thus avoiding the challenging optimization process in training generative models. This contrasts with other methods like JEM [22], which generates probability scores from a generative model perspective. JEM can be difficult to optimize and unstable in practice as it requires estimating the normalized density over the entire input space to maximize probability. Additionally, while JEM only utilizes data within the distribution, the approach by Liu et al., [20] allows for the flexible adjustment of energy distance between training and out-of-distribution data, incorporating both in-distribution and out-of-distribution data.

Liu et al. [20] demonstrated that the energy score can replace softmax confidence in pre-trained neural networks during predictions. They also introduced a constrained learning objective regarding energy during training, aiding in model adjustments. This learning process generates an energy surface, assigning low energy values to in-distribution data and high energy values to out-of-distribution training data.

The Energy-Based Model (EBM) has profound connections with modern machine learning models, especially discriminative models. As illustrated in Figure 1, consider a discriminative neural classifier $f(x) : R^D \rightarrow R^K$, mapping an input $x \in R^D$ to a number K, known as logits. These logits are used to generate a classification distribution using the softmax function:

$$p(y|x) = \frac{e^{f_y(x)/T}}{\sum_i e^{f_i(x)/T}} \tag{1}$$

The authors proposed using an energy function with input (x,y) is $E(x, y) = -f_{y(x)}$. The energy function does not alter the parameters of the neural network $f(x)$, and the free energy $E(x; f)$ can be expressed over $x \in R^D$ based on the denominator of the softmax activation function:

$$E(x, f) = -T. \log \sum_i^K e^{f_i(x)/T} \tag{2}$$

Energy can be used as a scoring function for any pre-trained neural network (without retraining). During inference, for an input x, the energy score $E(x; f)$ is computed for a neural

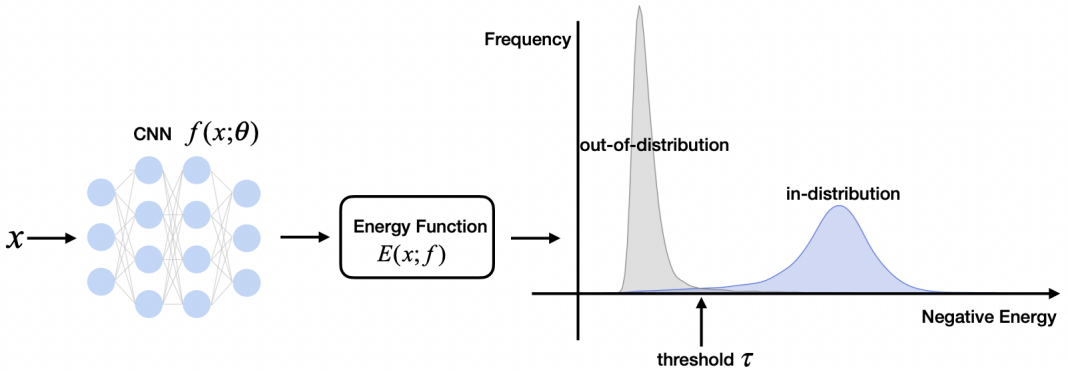

Figure 1. Energy-based Model (EBM) for Out-of-distribution detection [20].

network $f(x)$. An out-of-distribution (OOD) detector will classify the input as OOD if the energy score exceeds a specified threshold.

3. Methodology

Formally, a continual relation extraction (CRE) model learns a sequence of K tasks denoted as $T_1, T_2, \ldots, T_K$. Each task T_k has its dedicated training set $D_k = (x_i^k, y_i^k)_{i=1}^{N_k}$, where x_i^k represents input data, encompassing a sentence together with an entity pair, while $y_i^k \in R_k$ denotes the relation label between the two given entities in the input sentence (an example for x_i^k is given in the introduction section). R_k is predefined for task T_k. The primary objective of continual relation learning is to train a model capable of acquiring new tasks while preventing the occurrence of catastrophic forgetting in previous tasks. In simpler words, after learning the k^{th} task, the model should be able to determine the relation of a given entity pair within $\hat{R}_k$, where $\hat{R}_k = \bigcup_{i=1}^{k} R_i$ represents the set of all relations observed up to the k^{th} task.

For each relation, we collect a set of its exemplars, from which their centroid (aka. prototype) for this relation is calculated. This prototype is then used in the memory augmentation phase. The episodic memory of prototypes for the observed relations in tasks T_1 to T_k is denoted as $\hat{M}_k = \bigcup_{r \in \hat{R}_k} M_r$, where $M_r = \{(x_i, y_i = r)\}_{i=1}^{O}$, with r representing a specific relation, and O indicating the number of prototypes (memory size). The overall architecture of our proposed model is illustrated in Figure 2.

3.1. Model pipeline

The learning process of our model for the current task T_k comprises five key phases, as follows:

Initial training phase: wherein the input is encoded using the encoder **E**. Subsequently, the parameters of the encoder **E** and the projector Pro_j are adjusted based on the current training samples in D_k through supervised contrastive learning.

Exemplar selection phase: For each class $r \in R_k$ that wasn't observed in old tasks, we gather all the samples labelled r from D_k. Next, the k-means algorithm is employed to group these samples into clusters. Within each cluster, the closest sample to the centroid is selected as the representative for that cluster, denoted as M_r. Subsequently, a prototype p_r for r is generated based on M_r to extend the prototype set $\mathbf{P}_r$.

Memory augmentation: Wherein, we introduce a straightforward approach that employs prototype-based memory augmentation to address the issue of catastrophic forgetting

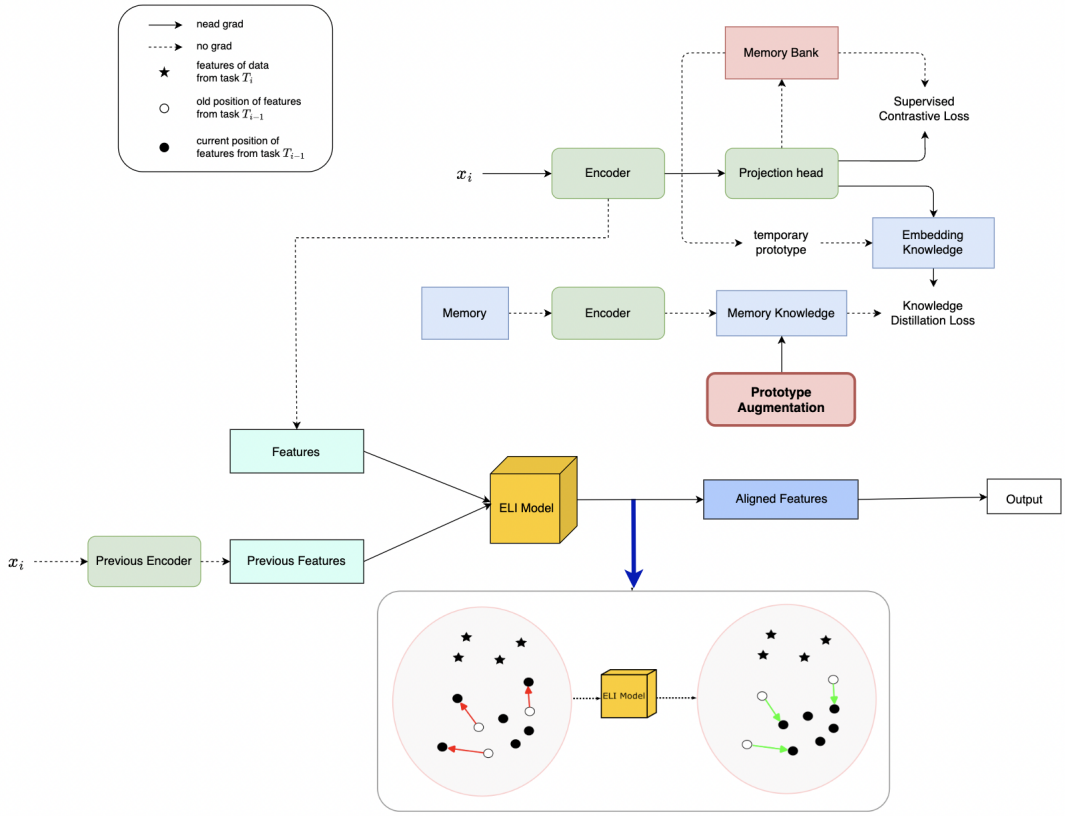

Figure 2. Overall architecture of our proposed continual relation extraction model. Here, x_i denotes the input (including a sentence and an entity pair appearing in it) of our model.

in continual learning. This phase retains a representative prototype for each previous relation and enhances the memorized prototypes by introducing Gaussian noises when learning new relations. Subsequently, the augmented prototypes and stored memory are combined to uphold discrimination and balance between old and new classes.

Memory replay: Following the augmentation phase, newly generated pseudo data are merged with existing data in the memory. During this phase, memory replay is consistently applied to acquire new relation prototypes iteratively while enhancing the distinctiveness of existing relational prototypes.

ELI model training: When continuously learning a sequence of tasks, the optimized hidden feature space of previous tasks is changed over subsequent tasks, reducing the model's performance on earlier tasks. To address this issue, we train a model based on the energy score and then feed it into the ELI algorithm to perform realignment of the skewed hidden spaces.

3.2. Energy-based Latent Feature Alignment (ELI) Model

As depicted in Figure 2, after training each task, we proceed to train the ELI using three components: the data from the current task $\mathbf{x}$, the latent representations of $\mathbf{x}$ from the model trained up to the last task: $\mathbf{z}_{t-1} = \mathbf{E}_{t-1}(\mathbf{x})$, and the latent representations of $\mathbf{x}$ from the model trained up to the current task: $\mathbf{z}_t = \mathbf{E}_t(\mathbf{x})$. The Energy-based Model (EBM) $\mathcal{E}_\psi$ undergoes training to

assign low energy to $\mathbf{z}_{t-1}$ and high energy to $\mathbf{z}_t$, as depicted in Algorithm 1.

In the subsequent step, the trained EBM $\mathcal{E}_\psi$ is utilized to counteract the representational shift occurring in the latent representations of previous task instances when passed through the current model: $\mathbf{z}_t = \mathbf{E}_t(\mathbf{x})$. Due to the representational shift in the latent space, $\mathbf{z}_t$ will exhibit higher energy values in the energy manifold. We aim to align $\mathbf{z}_t$ to alternative locations in the latent space, minimizing their energy on the manifold.

Algorithm 1 EBM training

Require: Feature extractor for the current task: $\mathbf{E}_t$; Feature extractor for the previous task: $\mathbf{E}_{t-1}$; Data distribution of the current task: $p_{data}^{T_t}$
Ensure: $\mathcal{E}_\psi$ is optimized
1: $\mathcal{E}_\psi \leftarrow$ Initialize
2: **while** until required iterations **do**
3: $\mathbf{x} \sim p_{data}^{T_t}$ ▷ Sample a mini-batch
4: $\mathbf{z}^{T_{t-1}} \leftarrow \mathbf{E}_{t-1}(\mathbf{x})$
5: $\mathbf{z}^{T_t} \leftarrow \mathbf{E}_t(\mathbf{x})$
6: $\mathbf{z}_{sampled}^{T_t} \leftarrow$ Sample from EBM with $\mathbf{z}^{T_t}$ as starting point ▷ Refer Equation (6)
7: in_dist_energy $\leftarrow E_\psi(\mathbf{z}^{T_{t-1}})$
8: out_dist_energy $\leftarrow E_\psi(\mathbf{z}_{sampled}^{T_t})$
9: Loss $\leftarrow$ (-in_dist_energy + out_dist_energy) ▷ Refer Equation (5)
10: Optimize $\mathcal{E}_\psi$ with Loss.
11: **end while**

3.2.1. EBM Training

Inspired by [16], we train the EBM model, which is constructed using a neural network that can map hidden representations to energy values (constants). Specifically, for a hidden feature vector $\mathbf{z} \in \mathbb{R}^D$ in the latent space, an energy function $\mathcal{E}_\psi(\mathbf{z}) : \mathbb{R}^D \to \mathbb{R}$ is trained to map it to a scalar energy value. An EBM is defined as the Gibbs distribution $p_\psi(z)$ over $\mathcal{E}_\psi(\mathbf{z})$:

$$p_\psi(\mathbf{z}) = \frac{\exp(-\mathcal{E}_\psi(\mathbf{z}))}{\int_{\mathbf{z}} \exp(-\mathcal{E}_\psi(\mathbf{z}))\, d\mathbf{z}}, \tag{3}$$

where $\int_{\mathbf{z}} \exp(-\mathcal{E}_\psi(\mathbf{z}))d\mathbf{z}$ represents a partition function that is challenging to compute. EBM is trained by maximizing the log-likelihood function over a set of samples drawn from the true distribution $p_{true}(\mathbf{z})$:

$$L(\psi) = \mathbb{E}_{\mathbf{z}\sim p_{true}}[\log p_\psi(\mathbf{z})]. \tag{4}$$

The derivative of the function $L(\psi)$ is as follows [23]:

$$\partial_\psi L(\psi) = \mathbb{E}_{\mathbf{z}\sim p_{true}}[-\partial_\psi \mathcal{E}_\psi(\mathbf{z})] + \mathbb{E}_{\mathbf{z}\sim p_\psi}[\partial_\psi \mathcal{E}_\psi(\mathbf{z})]. \tag{5}$$

The first component in (5) ensures the reduction of energy for samples $\mathbf{z}$ drawn from the genuine data distribution p_{true}, while the subsequent component guarantees that the samples generated from the model itself will possess elevated energy levels. In the ELI context, p_{true} corresponds to the distribution of latent representations from the model trained on the preceding task. Obtaining samples from $p_\psi(\mathbf{x})$ is challenging due to the normalization constant in Eq. (3). Approximate samples are iteratively generated using Langevin dynamics [24], a widely utilized Markov Chain Monte Carlo (MCMC) algorithm.

$$\mathbf{z}_{i+1} = \mathbf{z}_i - \frac{\lambda}{2}\partial_{\mathbf{z}}\mathcal{E}_\psi(\mathbf{z}) + \sqrt{\lambda}\omega_i, \omega_i \sim \mathcal{N}(0, \mathbf{I}) \tag{6}$$

where λ is the step size, and ω represents data uncertainty. In our experiments, we set λ to 0.1, as suggested by [16]. Equation (6) results in a Markov chain that converges to a stationary distribution after a few iterations, starting from an initial value $\mathbf{z}_i$.

We will feed the features through ELI (as depicted in Algorithm 2) for alignment during the testing phase.

Table 1. Experimental accuracy of our proposed CRE model and existing state-of-the-art models across all tasks on the FewRel dataset. T is an abbreviation for Task

FewRel										
Model	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
EA-EMR [4]	89	69	59.1	54.2	47.8	46.1	43.1	40.7	38.6	35.2
EMAR [5]	88.5	73.2	66.6	63.8	55.8	54.3	52.9	50.9	48.8	46.3
CML [13]	91.2	74.8	68.2	58.2	53.7	50.4	47.8	44.4	43.1	39.7
EMAR+BERT	**98.8**	89.1	89.5	85.7	83.6	84.8	79.3	80	77.1	73.8
RP-CRE [14]	97.9	92.7	91.6	89.2	88.4	86.8	85.1	84.1	82.2	81.5
CRL [19]	98.2	94.6	92.5	90.5	89.4	87.9	86.9	85.6	84.5	83.1
Ours	98.2	**94.7**	**93.9**	**92.2**	**90.1**	**89.6**	**88.6**	**87.1**	**86.1**	**84.6**

Table 2. Experimental accuracy of our proposed CRE model and existing state-of-the-art models across all tasks on the TACRED dataset. T is an abbreviation for Task

TACRED										
Model	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
EA-EMR[4]	47.5	40.1	38.3	29.9	24	27.3	26.9	25.8	22.9	19.8
EMAR[5]	73.6	57	48.3	42.3	37.7	34	32.6	30	27.6	25.1
CML[13]	57.2	51.4	41.3	39.3	35.9	28.9	27.3	26.9	24.8	23.4
EMAR+BERT	96.6	85.7	81	78.6	73.9	72.3	71.7	72.2	72.6	71
RP-CRE [14]	97.6	90.6	86.1	82.4	79.8	77.2	75.1	73.7	72.4	72.4
CRL[19]	97.7	93.2	89.8	84.7	84.1	81.3	80.2	79.1	79	78
Ours	**97.8**	**95.2**	**90.6**	**88.2**	**84.6**	**81.7**	**82.8**	**80.5**	**80**	**79.6**

Algorithm 2 Latent space alignment (ELI)

Require: Latent vector: $\mathbf{z}$; EBM: $\mathcal{E}_\psi$; Langevin iterations: L_{steps}; Learning rate: λ

Ensure: $\mathbf{z}$

1: **while** until L_{steps} iterations **do**
2: $\quad grad = \nabla_{\mathbf{z}} \mathcal{E}_\psi(\mathbf{z})$
3: $\quad \mathbf{z} \leftarrow \mathbf{z} - \lambda * grad$
4: **end while**

4. Experiments and Results

4.1. Datasets

We conduct extensive experiments with our proposed model on two continual semantic relation extraction benchmark datasets, namely FewRel and TACRED, as follows. Following the experimental settings of state-of-the-art CRE models, such as [14, 19], these two datasets are divided into three subsets: training set, test set and validation set with a ratio of 3:1:1.

- **FewRel** [17, 25]: an extensive well-balanced few-shot relation classification dataset. It consists 100 relations, each of 700 annotated sentences, resulting in 70,000 sentences in total.

- **TACRED** [2]: a large-scale imbalanced relation extraction dataset comprising 106,264 examples built over newswire and web documents from the yearly TAC Knowledge Base Population (TAC KBP) challenges. Human annotations for these examples are created based on the TAC KBP challenges and crowd-sourcing. As a result, the TACRED database encompasses 42 relations (including *no relation*, which

is removed in our experiments as it does not meet the open relation assumption of CRE). Due to the imbalance of TACRED, the number of training samples and test samples of each relation in our experiments is set to 320 and 40, respectively, which is in accordance with state-of-the-art CRE models, such as [14, 19].

4.2. Results

The experimental results of our CRE model reported on each dataset were averaged over five runs. We compare our model with six state-of-the-art existing methods for continual semantic relation extraction. Tables 1 and 2 present the accuracy scores for each task on the FewRel and TACRED datasets, respectively. Our model demonstrates superior performance on the FewRel dataset, achieving the highest accuracy on nine out of ten tasks on the FewRel dataset. Notably, it outperforms the recent CRL model [19] by 1-2% on 8 out of 10 tasks, as illustrated in Table 1. Our model is onpar with CRL on two remaining tasks, i.e. Task 1 and 2. However, our method performs less effectively than EMAR+BERT [5]. It's worth noting that EMAR+BERT employs BERT, a powerful pre-trained model, which provides a highly effective initialization for Task 1. Without using BERT, EMAR [5] underperforms our model by significantly large margins on all ten tasks, highlighting the effectiveness of our approach in leveraging memory and energy-based latent alignment for continual learning. We note that when learning only on Task 1, our proposed model does not suffer from catastrophic forgetting, so the effectiveness of our approach was not demonstrated in this specific scenario. However, as the number of tasks increases, our model demonstrates superior ability in mitigating catastrophic forgetting, consistently outperforming EMAR+BERT in subsequent tasks.

Similarly, the results on the TACRED dataset, presented in Table 2, further validate the effectiveness of our model. Our model achieves the highest accuracy scores across all tasks, demonstrating its robustness on this more challenging, imbalanced dataset. Especially on Task 4, our model can surpass CRL [5] by nearly 4%, the highest accuracy margin observed across all tasks. Unlike the results on the FewRel dataset, our model outperforms EMR+BERT on all tasks with significantly large accuracy margins, ranging from 1.1% (on Task 1) to 11.1% on Task 7.

The consistent superior performance of our model across both datasets can be attributed to several factors:

- **Effective memory utilization**: Our memory-prototype approach allows for efficient use of stored information, enabling better retention of knowledge from previous tasks.
- **Energy-based Latent space Alignment (ELI)**: The ELI module helps maintain consistency in the latent space across tasks, mitigating the representational shift problem common in continual learning scenarios.
- **Robustness to dataset characteristics**: The model's ability to outperform existing methods on both the balanced FewRel and imbalanced TACRED datasets demonstrates its versatility and robustness to different data distributions.
- **Long-term stability**: The performance improvement becomes more pronounced in later tasks, indicating our model's superior ability to mitigate catastrophic forgetting over extended sequences of tasks.

These results underscore the effectiveness of our proposed approach in addressing the challenges of continual relation extraction, particularly in maintaining performance across a sequence of tasks while mitigating catastrophic forgetting.

5. Conclusion

In this paper, we propose a continual learning model for semantic relation extraction, which is based on storing class-specific prototypes instead of real data points in a memory buffer. The reason for this alternative storage is that the use of real data points raises critical concerns about memory sizes increased over time and data security. We enrich these prototypes during training by adding Gaussian noises. Further, our model incorporates contrastive learning during the model training to obtain highly effective representations for the memory prototype data. Finally, we propose to train an Energy-based Latent feature space Alignment module (ELI) to adjust the hidden feature space before passing it through the classification phase during the model testing. ELI is expected to help our model avoid its performance decreases for old tasks when learning new coming tasks as a consequence of the hidden feature space of old tasks may deviate from the learned optimal space. Extensive experimental results on the FewRel and TACRED benchmark datasets for continual relation extraction show that our proposed method outperforms several baseline methods and state-of-the-art ones in continual semantic relation extraction.

References

[1] D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao. "Relation Classification via Convolutional Deep Neural Network". In: *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*. 2014, pp. 2335–2344. URL: https://aclanthology.org/C14-1220.

[2] Y. Zhang, V. Zhong, D. Chen, G. Angelir, and C. Manning. "Position-aware Attention and Supervised Data Improve Slot Filling". In: *Conference on Empirical Methods in Natural Language Processing*. 2017. DOI: 10.18653/v1/D17-1004.

[3] L. Soares, N. FitzGerald, J. Ling, and T. Kwiatkowski. "Matching the Blanks: Distributional Similarity for Relation Learning". In: 2019. DOI: 10.18653/v1/P19-1279.

[4] H. Wang, W. Xiong, M. Yu, X. Guo, S. Chang, and W. Wang. "Sentence Embedding Alignment for Lifelong Relation Extraction". In: *arXiv preprint arXiv:1903.02588* (2019). DOI: 10.18653/v1/N19-1086.

[5] X. Han, Y. Dai, T. Gao, Y. Lin, Z. Liu, P. Li, M.Sun, and J. Zhou. "Continual Relation Learning via Episodic Memory Activation and Reconsolidation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 6429–6440. DOI: 10.18653/v1/2020.acl-main.573.

[6] Z. Chen and B. Liu. *Lifelong Machine Learning*. Vol. 1. Springer, 2018. URL: https://www.cs.uic.edu/~liub/lifelong-machine-learning.html.

[7] M. McCloskey and NJ. Cohen. "Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem". In: *Psychology of learning and motivation*. Vol. 24. Elsevier, 1989, pp. 109–165. DOI: 10.1016/s0079-7421(08)60536-8.

[8] GI. Parisi, R. Kemker, JL. Part, C. Kanan, and S. Wermter. "Continual Lifelong Learning with Neural Networks: A Review". In: *Neural Networks* 113 (2019), pp. 54–71. DOI: 10.1016/j.neunet.2019.01.012.

[9] R. Hadsell, D. Rao, AA. Rusu, and R. Pascanu. "Embracing Change: Continual Learning in Deep Neural Networks". In: *Trends in Cognitive Sciences* 24.12 (2020), pp. 1028–1040. DOI: https://doi.org/10.1016/j.tics.2020.09.004.

[10] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, AA. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al. "Overcoming Catastrophic Forgetting in Neural Networks". In: *Proceedings of the national academy of sciences* 114.13 (2017), pp. 3521–3526. DOI: https://doi.org/10.1073/pnas.161183511.

[11] C. Fernando, D. Banarse, C. Blundell, Y. Zwols, D. Ha, AA. Rusu, A. Pritzel, and D. Wierstra. "Pathnet: Evolution Channels Gradient Descent in Super Neural Networks". In: *arXiv preprint arXiv:1701.08734* (2017). DOI: https://doi.org/10.48550/arXiv.1701.08734.

[12] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars. "Memory Aware Synapses: Learning What (not) to Forget". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 139–154. DOI: https://doi.org/10.1007/978-3-030-01219-9_9.

[13] T. Wu, X. Li, Y. Li, G. Haffari, G. Qi, Y. Zhu, and G. Xu. "Curriculum-Meta Learning for Order-Robust Continual Relation Extraction". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 12. 2021, pp. 10363–10369. DOI: https://doi.org/10.1609/aaai.v35i12.17241.

[14] L. Cui, D. Yang, J. Yu, C. Hu, J. Cheng, J. Yi, and Y. Xiao. "Refining Sample Embeddings with Relation Prototypes to Enhance Continual 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 1: Long Papers)*. 2021, pp. 232–243. DOI: 10.18653/v1/2021.acl-long.20.

[15] J. Snell, K. Swersky, and R. Zemel. "Prototypical Networks for Few-shot Learning". In: *Advances in Neural Information Processing Systems*. 2017, pp. 4077–4087. URL: https://dl.acm.org/doi/pdf/10.5555/3294996.3295163.

[16] K. Joseph, S. Khan, F. Khan, RM. Anwer, and VN. Balasubramanian. "Energy-based Latent Aligner for Incremental Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 7452–7461. DOI: 10.1109/CVPR52688.2022.00730.

[17] X.Han, H. Zhu, P. Yu, Z. Wang, Y. Yao, Z. Liu, and M. Sun. "FewRel: A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation". In: *arXiv preprint arXiv:1810.10147* (2018). DOI: 10.18653/v1/D18-1514.

[18] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. "A Survey on Contrastive Self-Supervised Learning". In: *Technologies* 9.1 (2020), p. 2. DOI: https://doi.org/10.3390/technologies9010002.

[19] Kang Zhao, Hua Xu, Jiangong Yang, and Kai Gao. "Consistent Representation Learning for Continual Relation Extraction". In: *arXiv preprint arXiv:2203.02721* (2022). DOI: 10.18653/v1/2022.findings-acl.268.

[20] W. Liu, X. Wang, J. Owens, and Y. Li. "Energy-based Out-of-distribution Detection". In: *NeurIPS*. 2020. URL: https://dl.acm.org/doi/pdf/10.5555/3495724.3497526.

[21] D. Hendrycks and K. Gimpel. "A Baseline for Detecting Misclassified and Out-of-distribution Examples in Neural Networks". In: *arXiv preprint arXiv:1610.02136* (2016). DOI: https://doi.org/10.1145/3503161.3548340.

[22] Will Grathwohl. *JEM: Official Code for the paper "Your Classifier is Secretly an Energy Based Model and You Should Treat it Like One"*. https://github.com/wgrathwohl/JEM. 2020.

[23] Oliver Woodford. *Notes on Contrastive Divergence*. Tech. Rep. Department of Engineering Science, University of Oxford, 2006. URL: https://ai.stanford.edu/~gaheitz/Research/ContrastiveDivergence.pdf.

[24] M. Welling and Y. Teh. "Bayesian Learning via Stochastic Gradient Langevin Dynamics". In: *Proceedings of the 28th International Conference on Machine Learning*. 2011, pp. 681–688. URL: http://www.icml-2011.org/papers/398_icmlpaper.pdf.

[25] T. Gao, X. Han, H. Zhu, Z. Liu, P. Li, M. Sun, and J. Zhou. "FewRel 2.0: Towards More Challenging Few-Shot Relation Classification". In: *arXiv preprint arXiv:1910.07124* (2019). DOI: 10.18653/v1/D19-1649.