Distilling Causal Effect of Data in Continual Few-shot Relation Learning

Weihang Ye, Peng Zhang* , Jing Zhang, Hui Gao, Moyao Wang

Tianjin University

COLING 2024

Outline

Background and Motivation

Background

Continual Few-Shot Relation Learning (CFRL) aims to learn an increasing number of new relational patterns from a data stream. However, due to the limited number of samples and the continual training mode, this method frequently encounters the catastrophic forgetting issues.

Causal Inference recently has been applied to aid NLP tasks, addressing issues such as spurious correlations, biases in textual data , and model interpretability.

[1]Chengwei Q. et al.. 2022. Continual few-shot relation learning via embedding space regularization and data augmentation. arXiv preprint arXiv:2203.02135.

[2] Xinting H. et al. 2021. Distilling causal effect of data in class-incremental learning. In Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, pages 3957–3966.

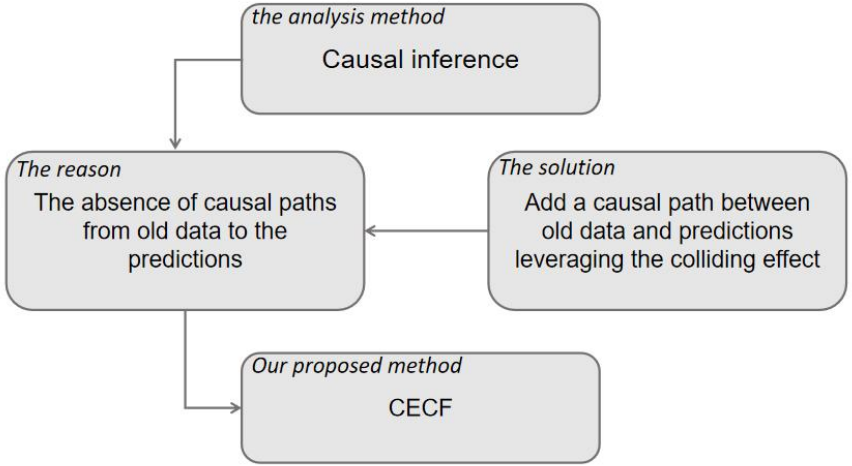
Background and Motivation

Research Idea

Research questions

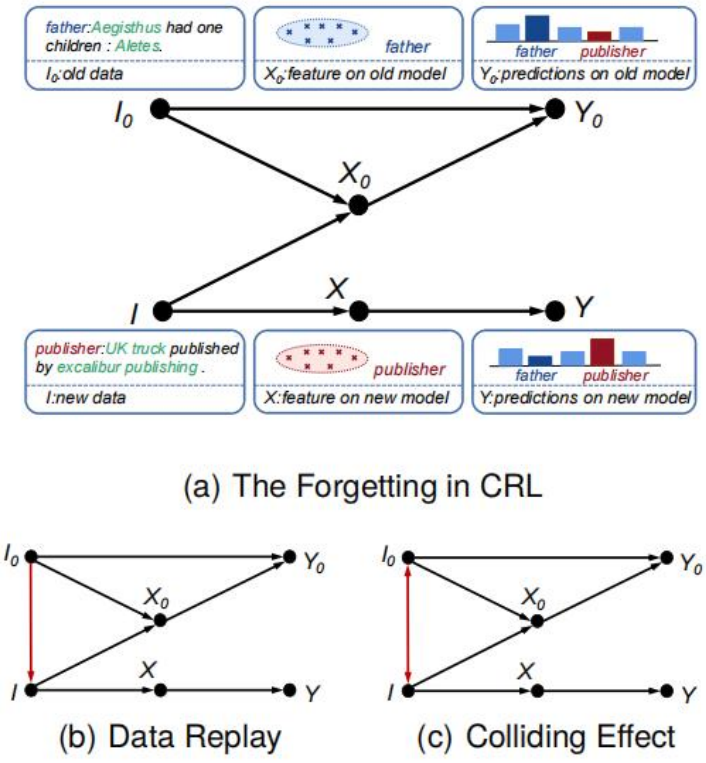
Catastrophic forgetting

Research methods



Methods

(Anti-) Forgetting in Causal Views

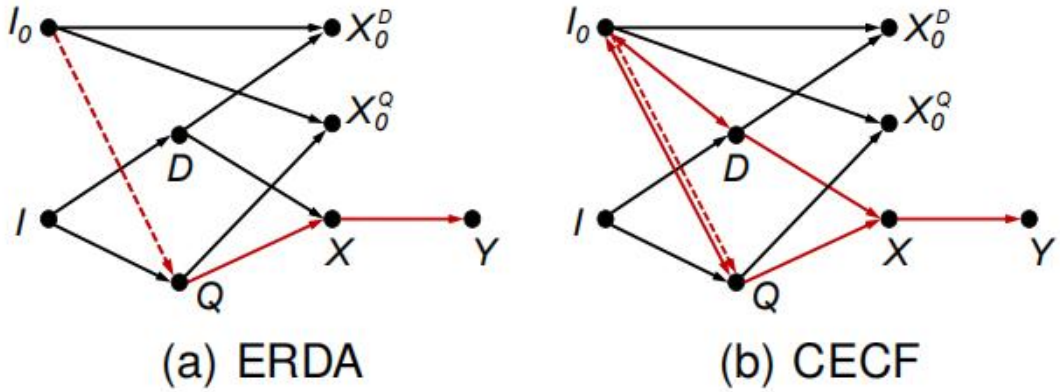


- **I** → **X** signifies the utilization of the new model to extract features **X** from the input sentence **I**.
- **X** → **Y** indicates using the extracted features X to predict the results **Y**
- **(I_0, I)** → **X_0** denotes that for a given new input sentence **I**, we can obtain the feature representation **X_0** in the old feature space by using the old model trained on old data **I_0**.

Methods

Distilling Colliding Effect in CFRL

Based on the distinctive attributes of CFRL, we extended the causal graph from Figure a to Figure b. The key modification is that the new relation data D and the memory data Q collide separately with old data I_0 on nodes X_0^D and X_0^Q in the old feature space. In this way, we introduce two causal paths from old data I^0 to predictions Y .



Methods

Distilling Colliding Effect in CFRL

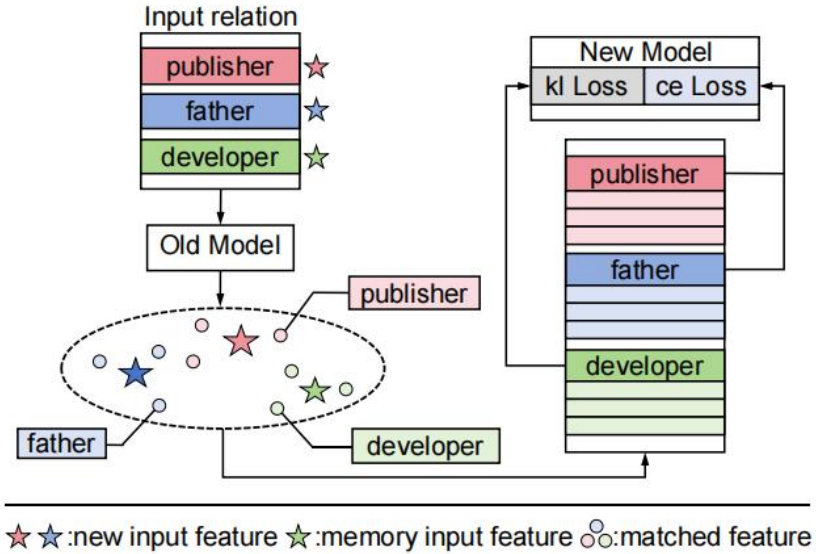
To establish the new causal path using the colliding effect, we identify sentences in new data that have the similar feature representations X_0 to the input in the old feature space. Initially, we compute the old feature representations of the input and other samples. Then, we employ the K − Nearest − Neighbors (KNN) strategy to search the K sentences whose features bear a greater similarity to the feature of input. Next, during the prediction phase for the input sentence, we leverage these matched sentences for joint prediction.

$$dist = \sqrt{\sum_{m=1}^{d} (\mathbf{p}_m - \mathbf{q}_m)^2}$$

$$\begin{aligned} \bar{Y}_k =& W_k Y_k + \textstyle\sum_{j=1}^{K} W_{kj} Y_{kj} \\ \text{s.t.} \quad & W_k \geq W_{k1} \geq W_{k2} \geq \cdots \geq W_{kK} \\ & W_k + \textstyle\sum_{j=1}^{K} W_{kj} = 1 \end{aligned}$$

Methods

Our causal framework for CFRL



The CFRL learning process typically encompasses three steps: learning with new data (simple training), selecting samples for memory, and alleviating forgetting through memory (anti-forgetting training). Our proposed improvements primarily concentrate on the first step.

Methods

Our causal framework for CFRL

$$\mathcal{L} = \lambda_{ce}\mathcal{L}_{ce} + \lambda_{kl}\mathcal{L}_{kl} + \lambda_{mm}\mathcal{L}_{mm} + \lambda_{pm}\mathcal{L}_{pm}$$

$$\mathcal{L}_{ce} = \sum_{(x_k, y_k) \in \widetilde{D_i}} \sum_{n=1}^{|\hat{R}_i|} \delta_{y_k, r_n} \times log(\bar{Y_k})$$

$$\bar{Y_k} = \frac{1}{2}\mathrm{S}(M_i(x_k)) + \frac{1}{2K}\sum_{j=1}^{K}\mathrm{S}(M_i(x_{kj}))$$

$$\mathcal{L}_{kl} = \sum_{(x_l, y_l) \in \hat{\mathcal{Q}}_{i-1}} \sum_{n=1}^{|\hat{R}_{i-1}|} \delta_{y_l, r_n} \times log\frac{\bar{Y_s}}{Y_t}$$

$$\bar{Y_s} = \frac{1}{2}\mathrm{S}(\frac{M_i(x_l)}{T_s}) + \frac{1}{2K}\sum_{j=1}^{K}\mathrm{S}(\frac{M_i(x_{lj})}{T_s})$$

$$Y_t = \mathrm{S}(\frac{M_{i-1}(x_l)}{T_t})$$

Experimental results

Tianjin University

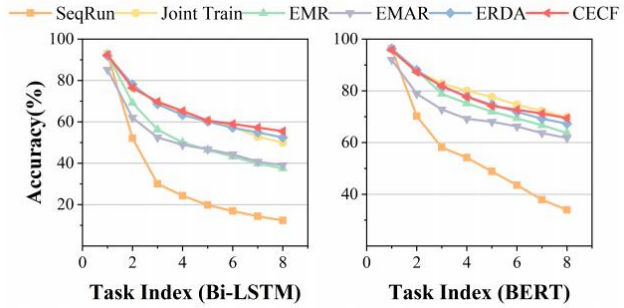
CECF Outperforms ERDA and Other Baselines

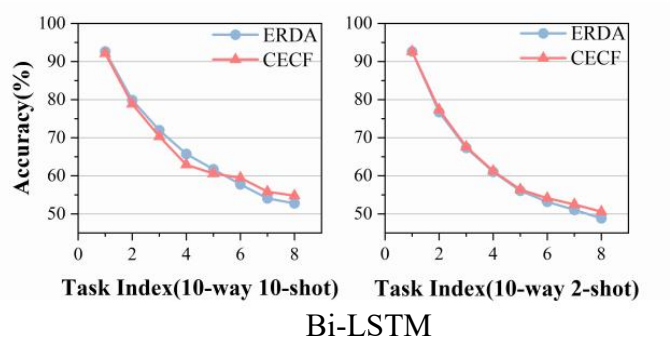
Method	Task index							
	1	2	3	4	5	6	7	8
SeqRun	**92.78**	52.11	30.08	24.33	19.83	16.90	14.36	12.34
Joint Train	**92.78**	76.29	69.39	64.75	60.45	57.64	52.80	50.03
EMR	**92.78**	69.14	56.24	50.03	46.50	43.21	39.88	37.51
EMAR	85.20	62.02	52.45	48.95	46.77	44.33	40.75	39.04
ERDA	91.98	**78.09**	68.59	63.32	60.2	57.13	54.91	52.45
CECF	92.32	76.48	**69.73**	**65.24**	**60.65**	**58.99**	**57.26**	**55.49**

Accuracy (%) of different methods at every time step on FewRel benchmark for 10-way 5-shot CFRL. CECF is better than ERDA with a Bi-LSTM encoder.

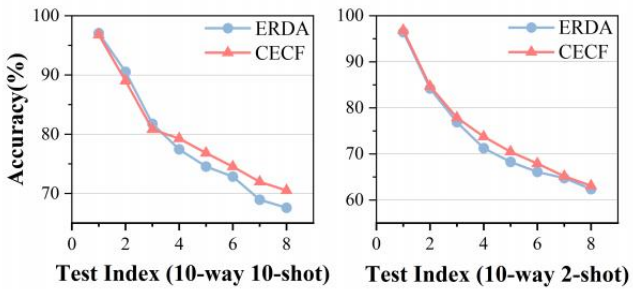
Experimental results

CECF Outperforms ERDA and Other Baselines

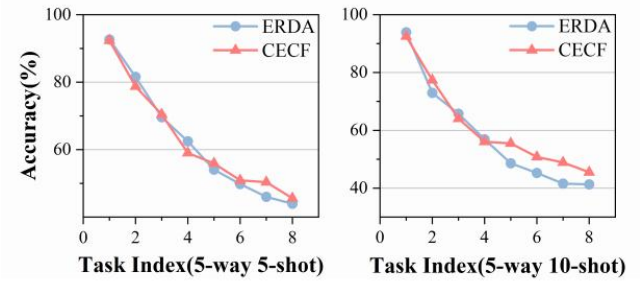




Bi-LSTM



BERT



BERT

From the graph, it can be seen that our performance is better than the baseline under different settings, proving its good generalization ability.

Experimental results

Ablation Study

Method	Task index							
	1	2	3	4	5	6	7	8
CECF	92.32	76.48	**69.73**	**65.24**	60.65	**58.99**	**57.26**	**55.49**
$w.o.$ AW	92.4	77.49	69.08	65.03	60.89	58.42	56.33	54.16
$w.o.E_T$	**92.42**	**78.4**	69.68	64.32	**61.31**	58.23	56.86	53.84
$w.o.E_M$	92.25	77.77	68.93	64.04	59.91	57.34	55.84	54.42
$w.o.E_{M\&T}$	91.98	78.09	68.59	63.32	60.2	57.13	54.91	52.45

We remove one component or combine the removal of two components:

(a) The adaptive weight module AW

(b) The colliding effect E_T in training data, where we calculate the regular cross-entropy loss for classification

(c) The colliding effect E_M in memory data

(d) Both the colliding effects $E_{M\&T}$ in training data and memory data

Experimental results

Hyper-Parameter Analysis

K	1	2	3	5	10
Accuracy(%)	54.75	53.91	55.49	54.35	54.56

λ	0.6	0.9	1.2	1.5	1.8
Accuracy(%)	53.97	53.64	55.49	53.61	53.90

We consider two hyper-parameters: the number of matched sentences K and the initial value of the adaptive weight λ. The results indicate that the best performance is achieved when K = 3. Additionally, the table demonstrates that the model achieves the best accuracy with λ = 1.2.

Thanks