Paper ID: Srp3559

DECOMPOSING LOGITS DISTILLATION FOR INCREMENTAL NAMED ENTITY RECOGNITION

Duzhen Zhang, Yahan Yu, Feilong Chen, Xiuyi Chen
Baidu Inc & Huawei Inc, Beijing, P.R.China
zhangduzhen@baidu.com, yuyahan@baidu.com, chenfeilong10@huawei.com, chenxiuyi01@baidu.com

Abstraction

Incremental Named Entity Recognition (INER) aims to continually train a model with new data, recognizing emerging entity types without forgetting previously learned ones. Prior INER methods have shown that Logits Distillation (LD), which involves preserving predicted logits via knowledge distillation, effectively alleviates this challenging issue. In this paper, we discover that a predicted logit can be decomposed into two terms that measure the likelihood of an input token belonging to a specific entity type or not. However, the traditional LD only preserves the sum of these two terms without considering the change in each component. To explicitly constrain each term, we propose a novel Decomposing Logits Distillation (DLD) method, enhancing the model's ability to retain old knowledge and mitigate catastrophic forgetting. Moreover, DLD is model-agnostic and easy to implement. Extensive experiments show that DLD consistently improves the performance of state-of-the-art INER methods across ten INER settings in three datasets.

Index Terms— Named Entity Recognition, Incremental Learning

Method

In this paper, we discover that predicted logits can be expressed as the sum of positive and negative terms, representing the likelihood and unlikelihood of an input token belonging to a specific entity type, respectively. However, the previous LD approach only focuses on maintaining the relative difference between positive and negative terms, neglecting the change of each term. To overcome this limitation and explicitly impose constraints on each term, we propose a simple yet effective method called Decomposing Logits Distillation (DLD), which improves the model's ability to retain old knowledge. We provide a comparison between LD and DLD in Figure 1. Furthermore, DLD is model-agnostic and can be combined with existing LD-based INER methods to effectively alleviate catastrophic forgetting.

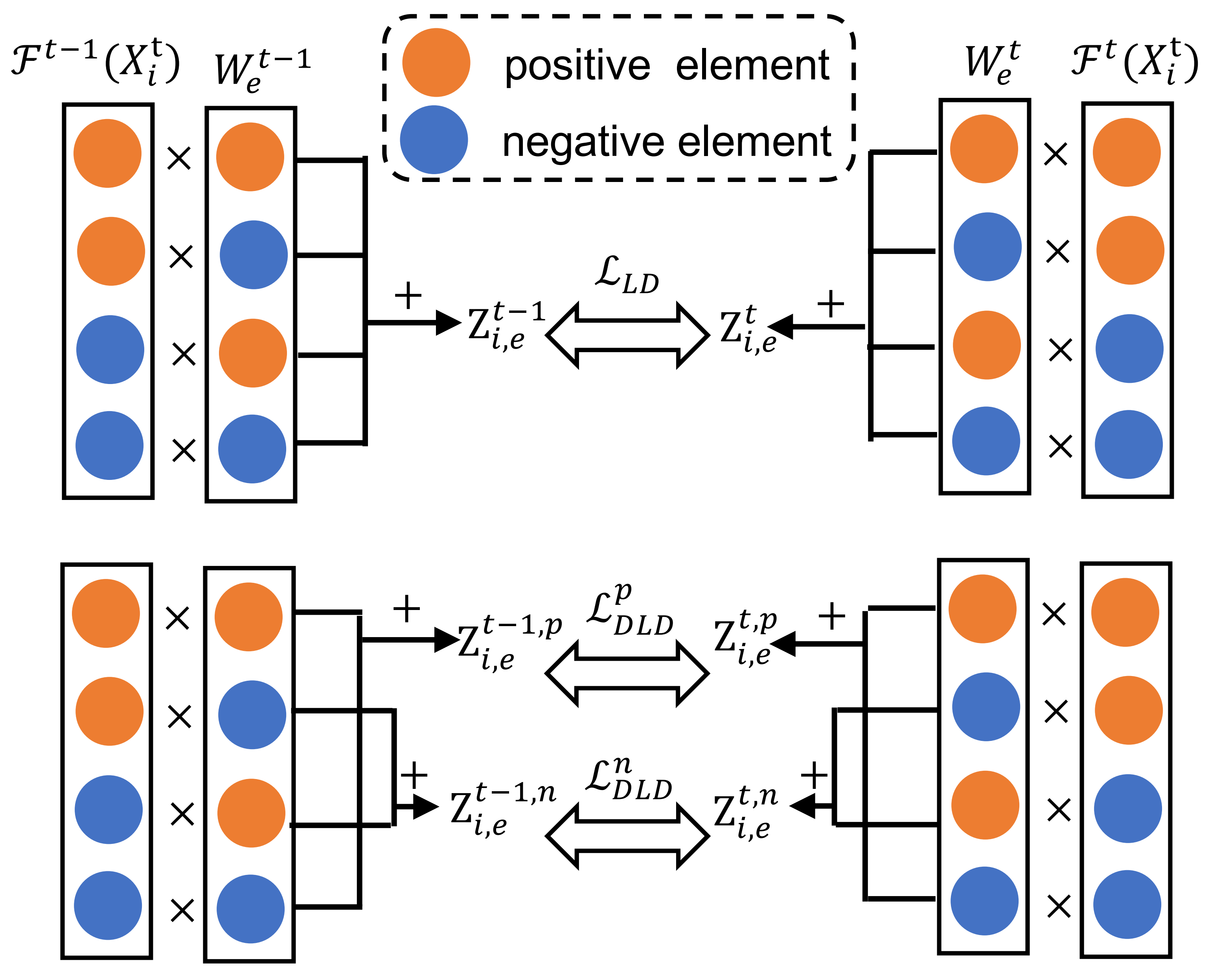

Figura 1: Comparison of LD and DLD. X_i^t denotes the i-th token of an input sequence in incremental step t. W_e^t denotes the weights of a classifier for an entity type e and $\mathcal{F}^t$ denotes a feature extractor in incremental step t. LD loss $\mathcal{L}_{LD}$ encourages the new model to predict logits $\mathbf{Z}^t$ similar to the ones $\mathbf{Z}^{t-1}$ obtained by the old model. DLD loss $\mathcal{L}_{DLD}^p$ encourages the new model to predict positive term $\mathbf{Z}^{t,p}$ similar to the ones $\mathbf{Z}^{t-1,p}$ obtained by the old model, and another DLD loss $\mathcal{L}_{DLD}^n$ constrains the negative terms $\mathbf{Z}^{t,n}$ and $\mathbf{Z}^{t-1,n}$.

Our main contributions can be summarized as follows:

- We introduce a simple yet effective DLD method, which decomposes predicted logits into two terms and imposes explicit constraints on them, improving the discriminative ability for old entity types together with the LD effectively.
- We conduct extensive experiments on ten INER settings of three datasets (CoNLL2003, I2B2, and OntoNotes5). The results show the effectiveness of DLD, consistently improving the performance of SOTA INER methods.

Datasets

We evaluate our DLD on three widely used NER datasets: CoNLL2003, I2B2, and OntoNotes5. The dataset statistics are summarized in Table 1.

Tabela 1: The statistics for each NER dataset.

Dataset	# Entity Type	# Sample	Entity Type Sequence (Alphabetical Order)
CoNLL2003	4	21k	LOCATION, MISC, ORGANISATION, PERSON
I2B2	16	141k	AGE, CITY, COUNTRY, DATE, DOCTOR, HOSPITAL, IDNUM, MEDICALRECORD, ORGANIZATION, PATIENT, PHONE, PROFESSION, STATE, STREET, USERNAME, ZIP
OntoNotes5	18	77k	CARDINAL, DATE, EVENT, FAC, GPE, LANGUAGE, LAW, LOC, MONEY, NORP, ORDINAL, ORG, PERCENT, PERSON, PRODUCT, QUANTITY, TIME, WORK OF ART

Experimental results

In this section, we present the results of extensive experiments on ten INER settings of three datasets: I2B2, OntoNotes5, and CoNLL2003. Tables 2 and 3 display the performances of our DLD and baselines. Our results demonstrate that, after applying DLD, the Micro-F1 and Macro-F1 scores of ExtendNER and CFNER consistently improve by approximately 2 points across nearly all INER settings of the three datasets. Furthermore, CFNER+DLD achieves new SOTA performance on ten INER settings of the three datasets. These findings highlight the universal and effective nature of our proposed DLD method.

Tabela 2: Comparisons with baselines on the CoNLL2003 dataset. The **bold** denotes the highest result. ‡ represents our reproduced results with the open codebases. Other baseline results are directly cited from CFNER. ExtendNER and CFNER with **DLD** significantly outperform their corresponding vanilla methods (with $p < 0.05$).

Baseline	FG-1-PG-1		FG-2-PG-1	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1
Finetuning	50.84±0.10	40.64±0.16	57.45±0.05	43.58±0.18
PODNet	36.74±0.52	29.43±0.28	59.12±0.54	58.39±0.99
LUCIR	74.15±0.43	70.48±0.66	80.53±0.31	77.33±0.31
Self-Training	76.17±0.91	72.88±1.12	76.65±0.24	66.72±0.11
ExtendNER	76.36±0.98	73.04±1.80	76.66±0.66	66.36±0.64
ExtendNER‡	76.07±0.35	73.06±0.29	77.89±0.42	69.92±1.02
ExtendNER‡ + **DLD**	78.56±0.59	75.95±0.83	78.97±0.69	72.60±1.07
CFNER	80.91±0.29	79.11±0.50	80.83±0.36	75.20±0.32
CFNER‡	80.29±0.21	78.44±0.24	81.52±0.43	77.20±0.82
CFNER‡ + **DLD**	**81.81±0.91**	**80.32±0.31**	**83.40±0.87**	**79.54±0.71**

Tabela 3: Comparisons with baselines on the I2B2 and OntoNotes5 datasets. The **bold** denotes the highest result. ‡ represents our reproduced results with the open codebases. Other baseline results are directly cited from CFNER. ExtendNER and CFNER with **DLD** significantly outperform their corresponding vanilla methods (with $p < 0.05$).

Dataset	Baseline	FG-1-PG-1		FG-2-PG-2		FG-8-PG-1		FG-8-PG-2	
		Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
I2B2	Finetuning	17.43±0.54	13.81±1.14	28.57±0.26	21.43±0.41	20.83±1.78	18.11±1.66	23.60±0.15	23.54±0.38
	PODNet	12.31±0.35	17.14±1.03	34.67±2.65	24.62±1.76	39.26±1.38	27.23±0.93	36.22±12.9	26.08±7.42
	LUCIR	43.86±2.43	31.31±1.62	64.32±0.76	43.53±0.59	57.86±0.87	33.04±0.39	68.54±0.27	46.94±0.63
	Self-Training	31.98±2.12	14.76±1.31	55.44±4.78	33.38±3.13	49.51±1.35	23.77±1.01	48.94±6.78	29.00±3.04
	ExtendNER	42.85±2.86	24.05±1.35	57.01±4.14	35.29±3.38	43.95±2.01	23.12±1.79	52.25±5.36	30.93±2.77
	ExtendNER‡	41.65±10.11	23.11±2.70	67.60±1.15	42.58±1.59	45.14±2.91	27.41±0.88	56.48±2.41	38.88±1.38
	ExtendNER‡ + **DLD**	43.96±3.19	25.07±2.22	69.04±1.83	44.02±2.03	47.84±1.75	28.92±1.55	57.97±1.12	40.81±2.19
	CFNER	62.73±3.62	36.26±2.24	71.98±0.50	49.09±1.38	59.79±1.70	37.30±1.15	69.07±0.89	51.09±1.05
	CFNER‡	64.79±0.26	37.79±0.65	72.58±0.59	51.71±0.84	56.66±3.22	36.84±1.35	69.12±0.94	51.61±0.87
	CFNER‡ + **DLD**	**66.48±0.78**	**39.53±1.83**	**75.77±0.66**	**52.83±0.97**	**63.64±1.55**	**40.44±2.26**	**70.88±0.70**	**53.21±0.92**
OntoNotes5	Finetuning	15.27±0.26	10.85±1.11	25.85±0.11	20.55±0.24	17.63±0.57	12.23±1.08	29.81±0.12	20.05±0.16
	PODNet	9.06±0.56	8.36±0.57	34.67±1.08	24.62±0.85	29.00±0.86	20.54±0.91	37.38±0.26	25.85±0.29
	LUCIR	28.18±1.15	21.11±0.84	64.32±1.79	43.53±1.11	66.46±0.46	46.29±0.38	76.17±0.09	55.58±0.55
	Self-Training	50.71±0.79	33.24±1.06	68.93±1.67	50.63±1.66	73.59±0.66	49.41±0.77	77.07±0.62	53.32±0.63
	ExtendNER	50.53±0.86	32.84±0.84	67.61±1.53	49.26±1.49	73.12±0.93	49.55±0.90	76.85±0.77	54.37±0.57
	ExtendNER‡	51.36±0.77	33.38±0.98	63.03±9.39	47.64±5.15	73.65±0.19	50.55±0.56	77.86±0.10	55.21±0.51
	ExtendNER‡ + **DLD**	54.04±1.40	35.56±1.14	68.65±1.31	51.17±2.07	75.38±0.89	52.57±0.72	78.93±0.46	56.78±0.55
	CFNER	58.94±0.57	42.22±1.10	72.59±0.48	55.96±0.69	78.92±0.58	57.51±1.32	80.68±0.25	60.52±0.84
	CFNER‡	58.44±0.71	41.75±1.51	72.10±0.31	55.02±0.35	78.25±0.33	58.64±0.42	80.09±0.37	61.06±0.37
	CFNER‡ + **DLD**	**62.75±1.22**	**45.62±1.38**	**73.44±0.26**	**57.33±0.65**	**79.86±0.36**	**60.44±0.70**	**82.26±0.55**	**63.07±1.13**