EMNLP 2023

Retrieval-Augmented Few-shot Text Classification

Guoxin Yu[1,2,3], Lemao Liu[4*], Haiyun Jiang[4], Shuming Shi[4], Xiang Ao[1,3*]

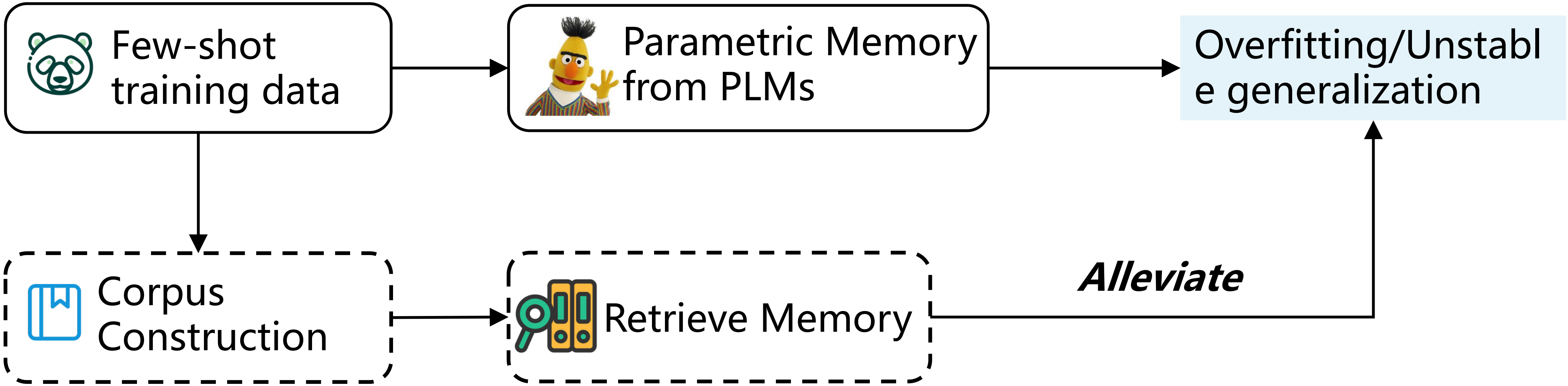
1 Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China.
2 Peng Cheng Laboratory.
3 University of Chinese Academy of Sciences, Beijing 100049, China.
4 Tencent AI Lab, China.

PART .01

Introduction

EMNLP 2023

- Training numerous parameters of PLMs on scarce data is prone to produce *over-fitting* and *unstable generalization*.
- ***Retrieval-based methods*** have shown the capability to incorporate ***retrieved memory*** alongside parameters for *better generalization*.



Introduction

Conventional few shot text classification

$$P(y|x) = \mathrm{softmax}(f_{clf}(x))$$

Retrieval-based few shot text classification

$$P_{\theta,\varphi}(y|x,z_j) = P_{\theta,\varphi}(y|x,z_j)P_{\varphi}(z_j|x)$$

$$P_{\theta}(y|x,z_i) = \mathrm{softmax}(f_{clf}(x \oplus z_j))$$

$$P_{\varphi}(z_j|x) = f_{retr}(x,z_j)$$

PART .01 Challenge

EMNLP 2023

Limited Search Space Z

Retriever $\mathbf{P}_{\theta,\varphi}$

Static Retriever

$$P_\varphi(z_j|x) = f_{retr}(x, z_j) = \mathrm{sim}(x, z_j)$$

sim can be BM25 or TF-IDF

z_j with high BM25/TF-IDF scores are limited

Joint Learning based Retriever

$$P_\varphi(z_j|x) = f_{retr}(x, z_j) = \frac{\exp(x \cdot z_j^{\mathrm{T}})}{\sum_{j=1}^{m} \exp(x \cdot z_j^{\mathrm{T}})}$$

f_{retr} is trainable.

Suffers from gradient vanishing.

<u>The gradient norm of the joint learning-based retriever exceeds the threshold of 1e−6 for only about 40% of the steps. EM-L and R-L are better.</u>

Task Input x

Fixed

Few-shot Classifier $P_\theta(y|x, z_i)$

Parameters Update

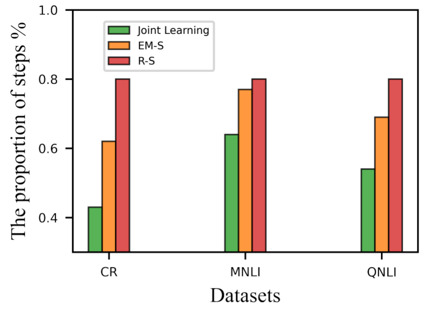


Figure 2: The proportion of steps in which the average gradient of retriever's all parameters is more than $1e-6$.

PART .02

Expectation Maximization-based Loss (EM-L)

EMNLP 2023

1. Expectation-step: z_j is considered as a **latent variable**

Retrieve m z_j

Conditional probabilities

$$P_{\theta,\varphi}(z_j|x,y) = \frac{P_{\theta,\varphi}(y|x,z_j)P_{\varphi}(z_j|x)}{\sum_{j=1}^{m} P_{\theta,\varphi}(y|x,z_j)P_{\varphi}(z_j|x)}$$

Alternates between an E-step and a M-step until convergence

2. Maximization-step: the parameters are updated by **maximizing the expected log-likelihood**

$$P_{\theta,\varphi}(y|x) = P(z_i|x|y|\theta_i) \cdot log\ P(y|x|z_i|\theta)$$

$$\theta_{i+1} = argmax_{\theta} P(z_i|x|y|\theta_i) \cdot log\ P(y|x|z_i|\theta)$$

$$\mathcal{L}oss_{em} = \sum_{i}^{N} \sum_{j}^{M} [P(z_j|x_i,y) \cdot logP(y|x_i,z_j)][c]$$

Ranking-based Loss (R-L)

EMNLP 2023

R-L considers the process of retrieving z_j as a **ranking task**.

Retrieve m z_j

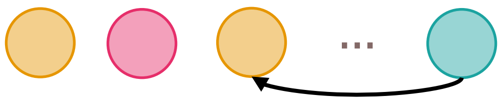
$$\mathcal{L}oss_{cls} = -\sum_{i}^{N}\sum_{j}^{M} log\left[P\left(z_j \middle| x_i\right) \cdot P\left(y \middle| x_i, z_j\right)\right][c]$$

$$\mathcal{L}oss_{rank} = \sum_{i}^{N}\sum_{j}^{M} max\left(0, margin + P\left(z_j \middle| x_i\right) - P\left(y \middle| x_i, z_j\right)[c]\right)$$

$$\mathcal{L}oss = \lambda \cdot \mathcal{L}oss_{rank} + (1-\lambda) \cdot \mathcal{L}oss_{cls}$$

Ranking of m examples

Prioritizing more beneficial examples



The assistance of z_j to x measured by $P\left(y \middle| x_i, z_j\right)[c]$

PART .03

Experiments

- EM-L and R-L approaches train the retriever **more effectively** than static retrieval and joint learning-based retrieval.
- The advantages of EM-L and R-L are **more pronounced on challenging tasks**.

Model	Single Sentence				Sentence Pair				ABSA		Avg.
	SST-2	MR	CR	TREC	QQP	QNLI	MNLI	SNLI	RES	LAP	
Prompt Learning with RoBerta-Large											
Vanilla	84.84$_{(6.80)}$	77.88$_{(7.90)}$	88.36$_{(2.89)}$	87.20$_{(7.70)}$	67.09$_{(6.70)}$	64.25$_{(7.45)}$	60.69$_{(4.08)}$	64.56$_{(4.08)}$	72.05$_{(4.08)}$	71.81$_{(2.88)}$	73.87
Static	88.60$_{(4.10)}$	83.67$_{(6.80)}$	87.06$_{(3.84)}$	90.95$_{(1.36)}$	68.31$_{(7.70)}$	66.27$_{(4.98)}$	60.38$_{(6.70)}$	68.17$_{(5.62)}$	70.95$_{(5.46)}$	73.01$_{(3.03)}$	75.74
Joint	90.71$_{(1.20)}$	85.83$_{(2.40)}$	86.76$_{(6.50)}$	90.57$_{(4.17)}$	67.26$_{(4.40)}$	63.15$_{(7.16)}$	61.95$_{(4.65)}$	67.64$_{(5.80)}$	71.07$_{(2.97)}$	73.32$_{(2.26)}$	75.83
EM-L	*91.31*$_{(1.30)}$	*87.58*$_{(1.40)}$	***90.00***$_{(0.90)}$	*92.13*$_{(1.41)}$	***74.41***$_{(0.74)}$	***67.66***$_{(3.77)}$	*64.85*$_{(3.21)}$	*69.52*$_{(3.69)}$	*73.74*$_{(3.46)}$	***76.02***$_{(1.90)}$	*78.72*
R-L	***91.58***$_{(1.30)}$	***87.47***$_{(0.09)}$	*89.93*$_{(1.70)}$	***92.86***$_{(1.21)}$	*73.79*$_{(2.28)}$	*67.62*$_{(5.79)}$	***66.04***$_{(3.18)}$	***73.08***$_{(4.59)}$	***76.79***$_{(2.60)}$	*75.59*$_{(1.51)}$	***79.46***
Fine-tune RoBerta-Large											
Vanilla	81.59$_{(4.50)}$	73.59$_{(9.90)}$	81.63$_{(4.08)}$	85.95$_{(5.57)}$	61.42$_{(8.19)}$	57.20$_{(2.09)}$	59.90$_{(5.72)}$	59.19$_{(5.58)}$	69.21$_{(4.14)}$	71.06$_{(5.11)}$	70.07
Static	81.99$_{(10.8)}$	72.69$_{(5.05)}$	82.75$_{(5.50)}$	87.02$_{(3.25)}$	60.23$_{(9.60)}$	57.11$_{(3.90)}$	54.69$_{(4.78)}$	62.65$_{(5.10)}$	70.48$_{(8.74)}$	71.37$_{(3.03)}$	70.10
Joint	83.49$_{(3.20)}$	74.89$_{(2.90)}$	80.63$_{(5.42)}$	86.33$_{(3.17)}$	63.50$_{(8.08)}$	57.66$_{(2.69)}$	60.99$_{(4.98)}$	61.01$_{(5.80)}$	70.23$_{(3.57)}$	70.62$_{(4.47)}$	70.94
EM-L	***85.38***$_{(1.30)}$	***75.80***$_{(2.20)}$	***83.81***$_{(5.36)}$	***89.36***$_{(2.64)}$	*65.70*$_{(8.17)}$	*60.93*$_{(1.56)}$	***62.24***$_{(3.12)}$	*65.25*$_{(3.20)}$	*71.64*$_{(3.36)}$	***72.69***$_{(3.18)}$	*73.27*
R-L	*84.69*$_{(2.29)}$	*75.35*$_{(2.20)}$	*83.17*$_{(3.22)}$	*88.92*$_{(3.81)}$	***70.53***$_{(2.68)}$	***61.37***$_{(0.12)}$	*62.18*$_{(1.72)}$	***66.31***$_{(3.30)}$	***73.28***$_{(3.13)}$	***72.69***$_{(3.01)}$	***73.85***

Table 1: Comparison results on 16-*shot* text classification. "Vanilla" denotes methods without retrieval, which only consists of a sentence encoder and a classifier. "Static" and "Joint" are static retrieval and joint learning-based retrieval, which are introduced in §2. "EM-L" and "R-L" are methods implemented with our proposed new objectives. All the reported results are average *Accuracy* and the standard deviation in the subscript.

Experiments

- Higher τ ′ of **EM-L and R-L** indicates that they could **prioritize more helpful examples according to their corresponding metrics** and improve the performance by training more effective retrievers.
- Retrieving examples according to **static metrics and joint learning-based** metrics may result in the inclusion of **harmful examples in the final performance**.

$$\iota = \frac{1}{m(m-1)} \sum_{i<j}^{i,j\epsilon m} sign(p_{i'} p_j) sign(q_{i'} q_j)$$

$$sign(p_i, p_j) = \begin{cases} 1, p_i < p_j \\ 0, p_i = p_j \\ -1, p_i > p_j \end{cases}$$

Kendall's τ'	SST-2	CR	QQP	QNLI	RES
Static	0.5344	0.5837	0.4307	0.5312	0.47857
Joint	0.5413	0.6129	0.4776	0.5937	0.4732
EM-L	<u>*0.6853*</u>	<u>*0.6451*</u>	***0.6265***	***0.7500***	***0.6598***
R-L	***0.7442***	***0.6562***	*0.6057*	*0.7185*	*0.6125*

Table 2: *Kendall's* τ' of $P_\phi(\mathbf{z}_j|\mathbf{x}_i)$ and $P_\theta(y|\mathbf{x}_i, \mathbf{z}_j)[y_i]$.

PART .03

Experiments

EMNLP 2023

- Auxiliary Experiments on different types of training sets proves the effectiveness of EM-L and R-L.

Accuracy	SST-2	MR	TREC	QQP
Vanilla	80.22	60.71	86.05	64.27
Static	76.58	67.51	86.94	60.30
Joint	85.41	71.01	86.57	61.92
EM-L	*87.30*	***78.75***	*87.52*	***67.90***
R-L	***89.79***	*77.38*	***88.78***	*66.77*

Table 3: Comparison results on 8-*shot* text classification. Standard deviations are omitted to save space.

	MR	TREC	RES	LAP
Accuracy				
Vanilla	90.80	96.80	86.53	80.87
Static	91.40	97.60	87.50	81.19
Joint	90.90	97.80	87.58	82.13
EM-L	*91.70*	***98.00***	*88.04*	*82.76*
R-L	***91.45***	***98.00***	***88.48***	***83.22***
Kendall's τ'				
Static	0.4340	0.5280	0.5705	0.4310
Joint	0.5075	0.6580	0.7187	0.7492
EM-L	***0.9195***	***0.7880***	*0.8700*	*0.8564*
R-L	*0.9090*	*0.7160*	***0.8889***	***0.8903***

Table 4: Comparison results with full supervision of the original datasets. Standard deviations are omitted to save space.

PART .03

Experiments

EMNLP 2023

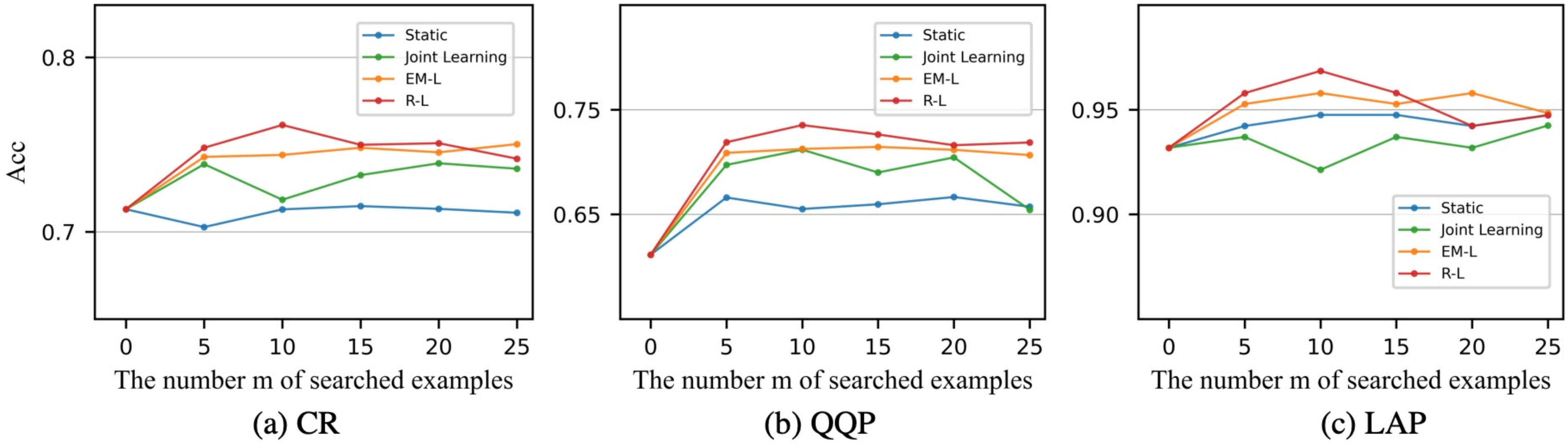


Figure 1: Effects of the number m of retrieved examples. The results are average *Accuracy* on the validation set.

PART .03

Experiments

EMNLP 2023

Input: ***Startup times*** are incredibly long : over two minutes. The sentiment polarity of ***startup times*** was <mask> .

Methods	Predictions	Retrieved Examples	Labels for Retrieved Examples
Static	positive ✗	The ***internet speed*** is spectacular. The sentiment polarity of ***internet speed*** was <mask> .	positive
Joint	positive ✗	That included the extra Sony Sonic Stage software , the speakers and the subwoofer I got -LRB- that WAS worth the money -RRB- , the bluetooth mouse for my supposedly bluetooth enabled computer , the extended life battery and the ***docking port***. The sentiment polarity of ***docking port*** was <mask> .	neutral
EM-L	negative ✓	Its not just slow on the ***internet***, its slow in general. The sentiment polarity of ***internet*** was <mask> .	negative
R-L	negative ✓	Another thing is that after only a month the ***keyboard*** broke and it costed $175 to send it in to fix it . The sentiment polarity of ***keyboard*** was <mask> .	negative

Figure 3: Case Study. “Input” denotes an input sentence from LAP, “Predictions” represents the predicted sentiment polarities of different methods, and “Retrieved Examples” is the fetched examples with the highest metric in the training set. “Labels for Retrieved Example” denotes sentiment labels of the fetched examples.

EMNLP 2023

EMNLP 2023

THANK YOU

Retrieval-Augmented Few-shot Text Classification

Guoxin Yu[1,2,3], Lemao Liu[4*], Haiyun Jiang[4], Shuming Shi[4], Xiang Ao[1,3*]

1 Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China.
2 Peng Cheng Laboratory.
3 University of Chinese Academy of Sciences, Beijing 100049, China.
4 Tencent AI Lab, China.