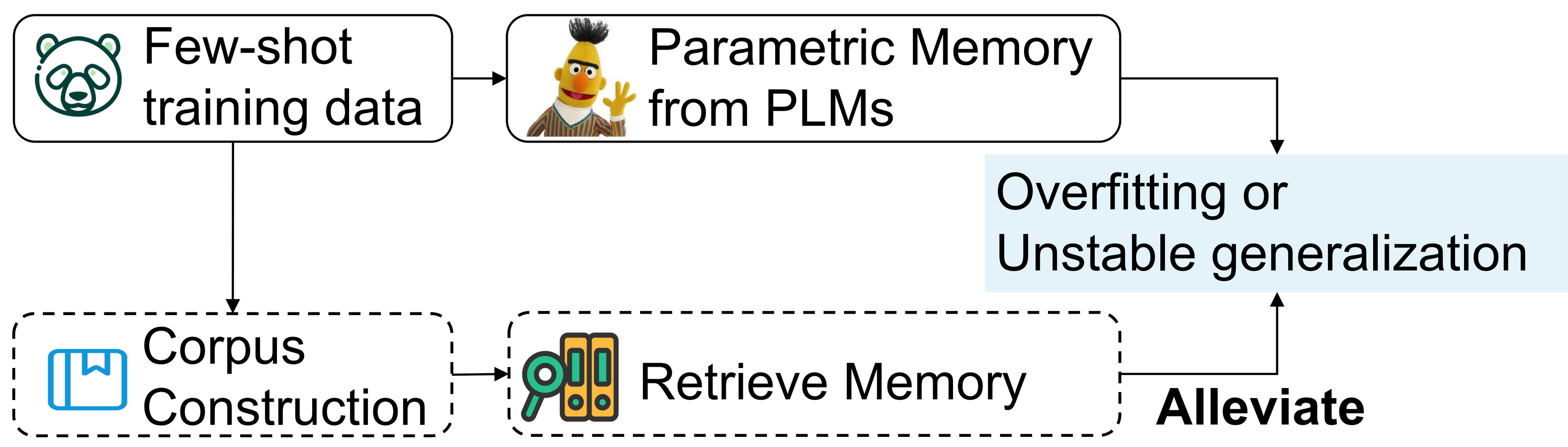
Retrieval-Augmented Few-shot Text Classification

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

1 Key Lab of Intelligent Information Processing of Chinese Academy of Sciences, Institute of Computing Technology. **2** University of Chinese Academy of Sciences. **3** Tencent AI Lab, China.

1 Motivation

- For few-shot text classification, 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.



- Compared with conventional methods, retrieval methods in text classification could comprise an example retriever $f_{retr}(x, z_j)$ and a text classifier $f_{clf}(x \oplus z_j)$.

$$P(y|x) = \text{softmax}(f_{clf}(x)) \quad \Rightarrow \quad P_{\theta,\varphi}(y|x) = \sum_{j=1}^{m} P_\theta(y|x, z_j) P_\varphi(z_j|x)$$

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

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

2 Challenges

- Retrieving examples from a narrow space to improve few-shot learning is still challenging due to limited training data.
  - Static retrieval whose metric is not task-specific (BM25/TF-IDF) cannot be reliable for retrieving helpful samples for target prediction.

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

  - Joint learning-based retrieval suffers from the gradient vanishing problem during the optimizing process, since its retrieval metric is updated towards the downstream task by minimizing the standard cross-entropy loss.

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

- The gradient norm of the joint learning-based retriever exceeds the threshold of 1e−6 for only about 40% of the steps.
- We aims to meet the challenge of weak supervision signals for the retriever and insufficient data.

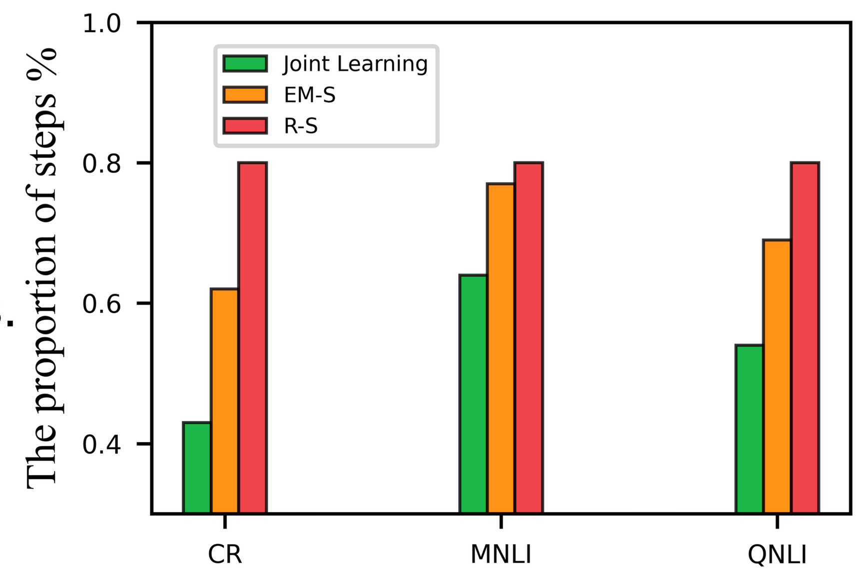


Figure 2: The proportion of steps in which the average gradient of retriever's all parameters is more than $1e-6$.

3 Method: Retrieval with EM-L and R-L

- EM-based Loss (EM-L) considers z_j as a latent variable and alternates between an E-step and a M-step until convergence.
  - The Expectation-step computes the conditional probabilities:

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

  - The Maximization-step updates the parameters by maximizing the expected log-likelihood:

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

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

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

- Ranking-based Loss (R-L) considers the process of retrieving z_j as a ranking task.
  - R-L employs a ranking loss to enhance the consistency between $P_\theta(y|x, z_i)[y_i]$ and $P_\varphi(z_i|x)$ and provide more direct signals to the retriever.

$$\mathcal{L}oss_{rank} = \sum_{i}^{n} \sum_{j}^{m} \max(P_\theta(y|x_i, z_j)[y_i] - P_\varphi(z_j, x_i) + \delta, 0)$$

$$\mathcal{L}oss_{cls} = -\sum_{i}^{n} \sum_{j}^{m} \log[P(z_j|x_i) \cdot P(y|x_i, z_j)][y_i]$$

$$\mathcal{L}oss = \lambda \cdot \mathcal{L}oss_{rank} + \mathcal{L}oss_{cls}$$

- Both of EM-L and R-L aim to retrieve examples from a limited space more effectively and prioritize more beneficial examples for downstream tasks.

4 Experiment

- Retrieving examples from the training set is effective in few-shot scenarios.
- 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, such as QQP, QNLI, and LAP.

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	$\underline{91.31}_{(1.30)}$	$\underline{87.58}_{(1.40)}$	$\mathbf{90.00}_{(0.90)}$	$\underline{92.13}_{(1.41)}$	$\mathbf{74.41}_{(0.74)}$	$\mathbf{67.66}_{(3.77)}$	$\underline{64.85}_{(3.21)}$	$\underline{69.52}_{(3.69)}$	$\underline{73.74}_{(3.46)}$	$\mathbf{76.02}_{(1.90)}$	$\underline{78.72}$
R-L	$\mathbf{91.58}_{(1.30)}$	$\mathbf{87.47}_{(0.09)}$	$\underline{89.93}_{(1.70)}$	$\mathbf{92.86}_{(1.21)}$	$\underline{73.79}_{(2.28)}$	$\underline{67.62}_{(5.79)}$	$\mathbf{66.04}_{(3.18)}$	$\mathbf{73.08}_{(4.59)}$	$\mathbf{76.79}_{(2.60)}$	$\underline{75.59}_{(1.51)}$	$\mathbf{79.46}$
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	$\mathbf{85.38}_{(1.30)}$	$\mathbf{75.80}_{(2.20)}$	$\mathbf{83.81}_{(5.36)}$	$\mathbf{89.36}_{(2.64)}$	$\underline{65.70}_{(8.17)}$	$\underline{60.93}_{(1.56)}$	$\mathbf{62.24}_{(3.12)}$	$\underline{65.25}_{(3.20)}$	$\underline{71.64}_{(3.36)}$	$\mathbf{72.69}_{(3.18)}$	$\underline{73.27}$
R-L	$\underline{84.69}_{(2.29)}$	$\underline{75.35}_{(2.20)}$	$\underline{83.17}_{(3.22)}$	$\underline{88.92}_{(3.81)}$	$\mathbf{70.53}_{(2.68)}$	$\mathbf{61.37}_{(0.12)}$	$\underline{62.18}_{(1.72)}$	$\mathbf{66.31}_{(3.30)}$	$\mathbf{73.28}_{(3.13)}$	$\mathbf{72.69}_{(3.01)}$	$\mathbf{73.85}$

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.

5 Analysis

- Higher Kendall's τ' of EM-L and R-L in 16-shot and 8-shot text classification indicates that they could prioritize more helpful examples according to their corresponding metrics.

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	$\underline{0.6853}$	$\underline{0.6451}$	**0.6265**	**0.7500**	**0.6598**
R-L	**0.7442**	**0.6562**	$\underline{0.6057}$	$\underline{0.7185}$	$\underline{0.6125}$

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

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	$\underline{87.30}$	**78.75**	$\underline{87.52}$	**67.90**
R-L	**89.79**	$\underline{77.38}$	**88.78**	$\underline{66.77}$

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

- EM-L and R-L maintain sustaining advantages and stability as the number of retrieval examples varies, which verifies their stronger supervision signals.

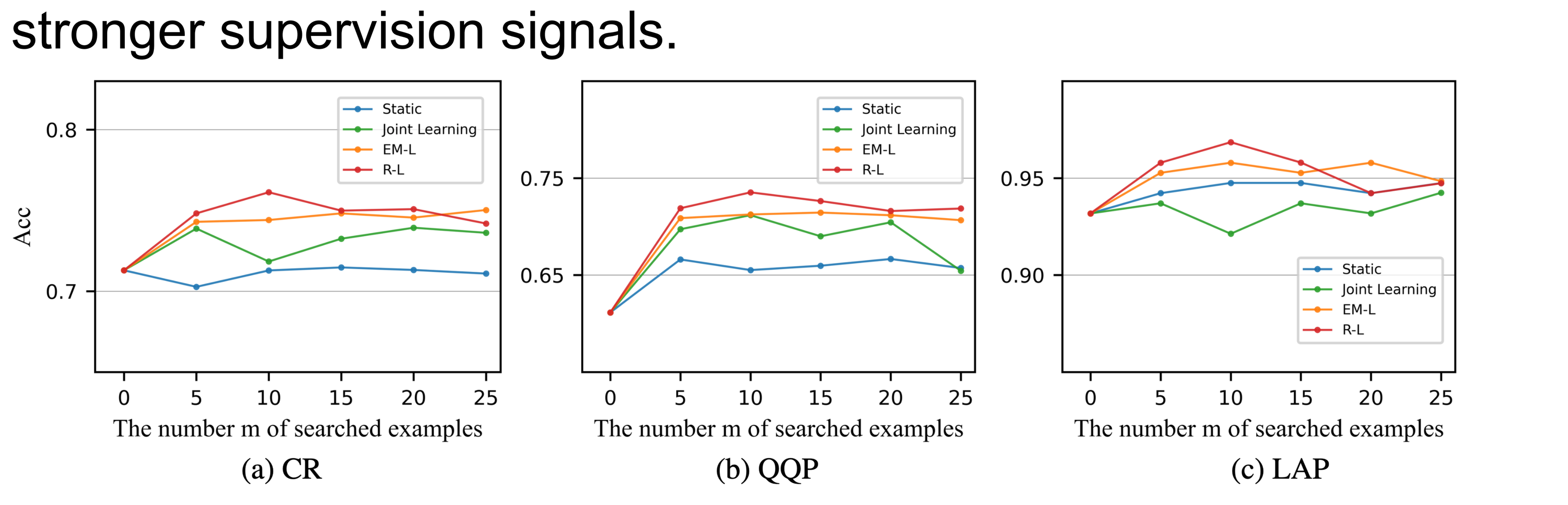


Figure 1: Effects of the number m of retrieved examples. The results are average *Accuracy* on the validation set.