Decoupling Knowledge from Memorization: Retrieval-augmented Prompt Learning

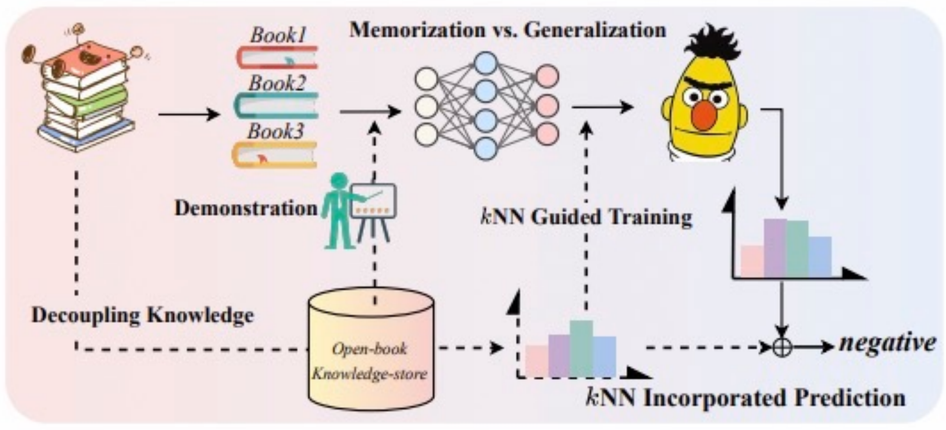
Xiang Chen [1,2], Lei Li [1,2], Ningyu Zhang [1,2]*, Xiaozhuan Liang [1,2], Shumin Deng [1,2]

Chuanqi Tan [3], Fei Huang [3], Luo Si [3], Huajun Chen [1,2]*

[1] Zhejiang University & AZFT Joint Lab for Knowledge Engine,

[2] Hangzhou Innovation Center, Zhejiang University, [3] Alibaba Group

Introduction



- **Limitations of Prompt Learning:**

different Prompt learning with PLMs usually generalizes unstably in an extremely low-resource setting or emerging domains. One potential reason is that, it is non-trivial for parametric models to learn rare or hard patterns well with rote memorization, thus, resulting in inefficient generalizable performance.

- **Decoupling knowledge from memorization :**

with the motivation of decoupling knowledge from memorization to help the model strike a balance between generalization and memorization, we constructs an open-book knowledge-store from training instances and implements a retrieval mechanism during the process of input, training and inference, thus equipping the model with the ability to retrieve related contexts from the training corpus as cues for enhancement.

Retrieval-augmented Prompt Learning

Open-book Knowledge-store

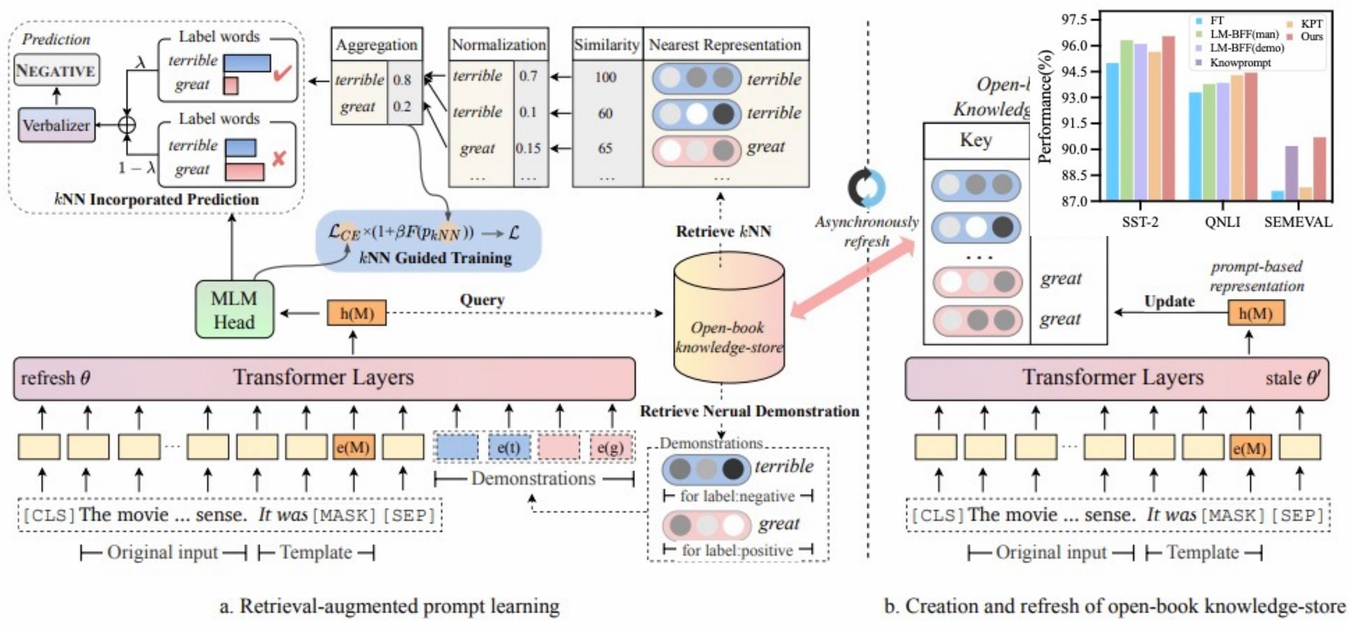
Given the i-th example $(\boldsymbol{c}_i, y_i)$ in the training data $\mathcal{C}$, we obtain the key-value pair $(\boldsymbol{h}_{\hat{c}_i}, v_i)$, in which $\hat{\boldsymbol{c}}_i = \mathcal{T}(\boldsymbol{c}_i), \boldsymbol{h}_{\hat{c}_i} \in \mathbb{R}^d$ is the embedding of the [MASK] token in the last layer of the PLM, and $v_i = f(y_i)$ denotes the label word of the i-th example.

$$(\mathcal{K}, \mathcal{V}) = \left\{ \left(\boldsymbol{h}_{\hat{c}_i}, v_i\right) \mid (\boldsymbol{c}_i, y_i) \in \mathcal{C} \right\}$$

Retrieval of Neural Demonstration

We intuitively aggregate the m neighbor vectors for each class according to their similarity and incorporate the demonstration into the input representation of $\hat{\boldsymbol{x}}$ after the word embedding layer of the $\mathcal{M}$ as follows:

$$\mathcal{I} = e(\hat{\boldsymbol{x}}) \oplus \left[\sum_{i \in [1:m]} \alpha_i^{(1)} \boldsymbol{h}_{\hat{c}_i}^{(1)}, e\left(v^{(1)}\right) \right] \oplus \ldots \oplus \left[\sum_{i \in [1:m]} \alpha_i^{(L)} \boldsymbol{h}_{\hat{c}_i}^{(L)}, e\left(v^{(L)}\right) \right]; \alpha_i^{(l)} = \frac{e^{\boldsymbol{h}_{\hat{q}} \cdot \boldsymbol{h}_{\hat{c}_i}^{(l)}}}{\sum_{i \in [1:m]} e^{\boldsymbol{h}_{\hat{q}} \cdot \boldsymbol{h}_{\hat{c}_i}^{(l)}}}$$



a. Retrieval-augmented prompt learning b. Creation and refresh of open-book knowledge-store

Figure 2: Overview of RETROPROMPT. Note that $e(\cdot)$ denotes word embedding function in the PLM $\mathcal{M}$, while "M","t" and "g" in $e(\cdot)$ specifically refers to "[MASK]", "terrible" and "great".

Retrieve kNN for Guiding Training

Our intuition is to differentiate between easy and hard examples according to the prediction of kNN.

$$P_{k\text{NN}}(y \mid \boldsymbol{q}_t) \propto \sum_{(\boldsymbol{c}_i, y_i) \in \mathcal{N}} \mathbf{1}_{y=y_i} \exp\left(d\left(\boldsymbol{h}_{\hat{\boldsymbol{q}}_t}, \boldsymbol{h}_{\hat{\boldsymbol{c}}_i}\right)\right)$$

$$F(p_{k\text{NN}}) = -\log(p_{k\text{NN}}), \quad \mathcal{L} = (1 + \beta F(p_{k\text{NN}}))\mathcal{L}_{CE}$$

kNN based probability for Cloze-style Prediction

we reformulate the $P(y \mid \boldsymbol{q}_t)$ by interpolating the $P_{k\text{NN}}$ with the already-trained base PLM's MLM prediction $P_{\mathcal{M}}$ using parameter λ to produce the final probability of the label:

$$P(y \mid \boldsymbol{q}_t) = \lambda P_{k\text{NN}}(y \mid \boldsymbol{q}_t) + (1 - \lambda) g(P_{\mathcal{M}}([\text{MASK}] = v \mid \mathcal{T}(\boldsymbol{q}_t)))$$

Experiments

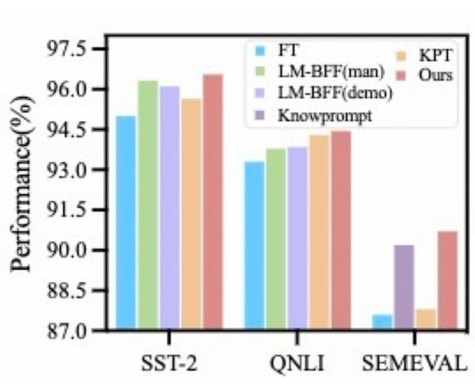
Few-shot/Zero-shot Results

St.	Model	Single Sentence			Sentence Pair			Model	Information Extraction			Avg.
		SST-2 (acc)	MR (acc)	CR (acc)	MNLI (acc)	QNLI (acc)	QQP (F1)		FewN (acc)	SemEval (acc)	TACRED (F1)	
16	FT	81.4 (3.8)	76.9 (5.9)	75.8 (3.2)	45.8 (6.4)	60.2 (6.5)	60.7 (4.3)	FT	52.7 (2.2)	66.1 (1.2)	25.8 (2.8)	60.6
	LM-BFF (man)	91.6 (1.2)	87.0 (2.0)	90.3 (1.6)	64.3 (2.5)	64.6 (5.4)	65.4 (5.3)	KnPr	65.3 (1.1)	80.9 (2.5)	33.2 (2.0)	71.4
	LM-BFF (D-demo)	91.8 (1.2)	86.6 (1.8)	90.2 (1.4)	64.8 (2.3)	69.2 (5.4)	68.2 (3.2)	KnPr (D-demo)	—	—	—	72.2*
	KPT †	90.3 (1.6)	86.8 (1.8)	88.8 (3.7)	61.4 (2.1)	61.5 (2.8)	71.6 (2.7)	KPT †	65.9 (1.5)	78.8 (2.1)	32.8 (1.7)	70.9
	Ours	**93.9** (0.4)	**88.0** (0.8)	**91.9** (0.7)	**71.1** (1.8)	**71.6** (1.8)	**74.0** (2.0)	**Ours**	**67.3** (0.9)	**81.5** (1.3)	**40.7** (0.7)	**75.6**
4	FT	60.2 (2.8)	57.6 (1.4)	66.4 (5.5)	35.0 (0.3)	54.2 (3.9)	52.8 (4.7)	FT	32.7 (2.9)	38.8 (2.0)	14.7 (2.8)	45.8
	LM-BFF (man)	90.7 (0.8)	85.2 (2.8)	89.9 (1.8)	51.0 (2.5)	61.1 (6.1)	48.0 (4.9)	KnPr	52.5 (1.5)	58.4 (3.7)	28.8 (2.5)	62.8
	LM-BFF (D-demo)	90.2 (1.5)	85.5 (2.1)	89.7 (0.6)	56.1 (1.0)	61.7 (7.6)	63.2 (5.6)	KnPr (D-demo)	—	—	—	65.1*
	KPT †	88.2 (5.7)	83.4 (1.5)	87.2 (2.5)	53.7 (2.7)	59.2 (2.8)	54.9 (7.9)	KPT †	58.8 (2.2)	57.2 (3.2)	27.5 (2.2)	63.3
	Ours	**91.5** (1.8)	**87.4** (0.5)	**91.4** (0.6)	**57.6** (5.5)	**62.2** (6.0)	**66.1** (4.1)	**Ours**	**60.9** (1.9)	**59.2** (3.0)	**32.1** (2.0)	**67.6**
0	LOTClass♣	71.8	81.7	50.1	50.4	36.5	55.9	LOTClass♣	11.5	9.8	2.5	41.1
	FT	49.1	50.0	49.8	34.4	49.5	31.6	FT	10.0	6.2	0.5	31.2
	LM-BFF (man)	83.5	80.3	78.4	49.7	50.5	49.7	KnPr	15.9	10.3	2.3	46.7
	LM-BFF (D-demo)	82.9	80.7	**81.4**	52.2	53.5	44.0	KnPr (D-demo)	—	—	—	47.0*
	KPT †	78.4	81.9	71.4	37.1	55.3	47.5	KPT †	24.6	11.6	0.8	45.7
	Ours	**86.8**	**83.5**	79.7	**53.7**	**56.2**	**56.7**	**Ours**	**41.3**	**12.2**	**2.8**	**52.5**

Cross-domain Results

Model	Source	Target Domain	
	16-shot MR	SST-2	CR
FT	76.9	71.4	64.7
LM-BFF (man)	87.0	88.9	86.9
LM-BFF (D-demo)	86.6	89.3	87.5
KPT	86.8	86.8	86.7
RETROPROMPT	**88.0**	**91.4**	**88.8**
	16-shot QQP	MRPC	RTE
FT	60.7	43.7	48.0
LM-BFF (man)	65.4	20.9	65.5
LM-BFF (D-demo)	68.2	38.8	66.2
KPT	71.6	42.3	65.8
RETROPROMPT	**74.0**	**49.4**	**67.3**

Full-data Results



Analysis of Memorization

Definition of Memorization Measurement

we define memorization measures as to how the classification varies when a training instance z is deleted from the trainset. We define and derive the memorization score for a training instance z as follows：

$$\mathcal{S}_{\text{delate}}(\boldsymbol{z}) \overset{\text{def}}{=} -\left.\frac{dP\left(y \mid \boldsymbol{x}; \hat{\theta}_{\xi,-\boldsymbol{z}}\right)}{d\xi}\right|_{\xi=0} = -\nabla_\theta P(y \mid \boldsymbol{x}; \hat{\theta})^\top \left.\frac{d\hat{\theta}_{\xi,-\boldsymbol{z}}}{d\xi}\right|_{\xi=0} = -\nabla_\theta P(y \mid \boldsymbol{x}; \hat{\theta})^\top H_{\hat{\theta}}^{-1} \nabla_\theta \mathcal{L}(\boldsymbol{z}, \hat{\theta})$$

Top-memorized Instances: Typical or Atypical?

we adopt SST-2 to analyze the memorization by judging the atypical of an instance by checking the percentage of positive phrases.

Mem Group	Negative			Postive		
	FT	LM-BFF	OURS	FT	LM-BFF	OURS
Top-10%	34.29	32.78	30.23	68.75	69.71	75.67
ALL		23.40			86.39	
Bottom-10%	17.63	16.25	14.42	95.92	95.08	94.53

	FT	LM-BFF	OURS
MEM SCORE	4.597	0.121	0.032