LOIRE: **L**ifel**O**ng learning on **I**ncremental data via pre-trained language model g**R**owth **E**fficiently

Xue Han, Yitong Wang, Junlan Feng, Wenchun Gao, Qian Hu, Chao Deng
JIUTIAN Team, China Mobile Research Institute

ICLR

1. Introduction

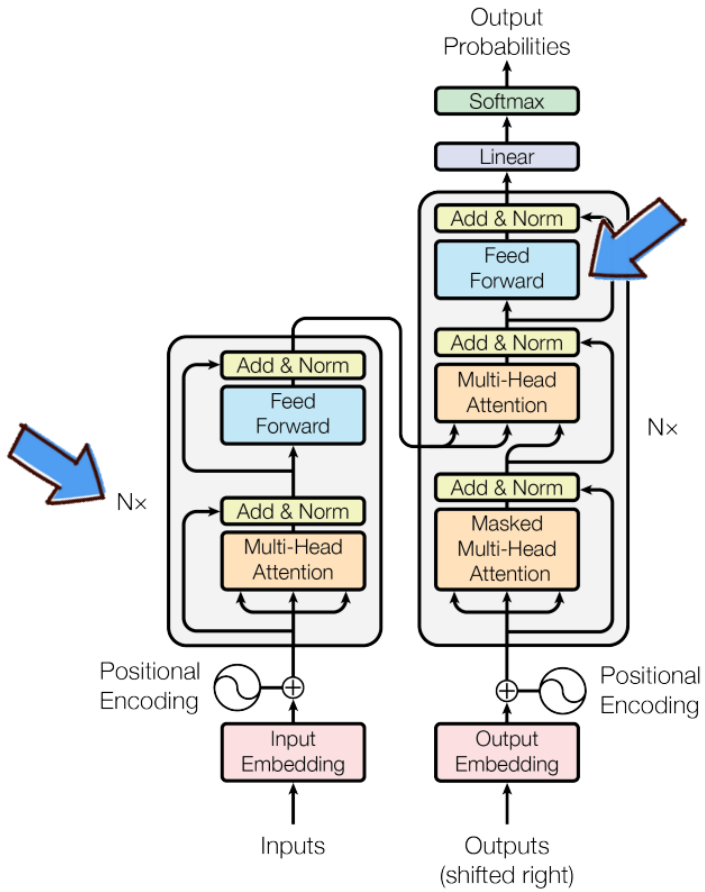
Background:

New data from other sources → PLM's performance may decline → training from scratch: time and computation intensive

Motivation:

Lifelong learning pipelines by employing **model growing approaches**

- **Growth schedules** (when and where) only consider layers and width of FFN
- **Growth operators** (inherit previous knowledge) non-strict function preservation
- **Data distribution** is shifting, causing catastrophic forgetting



1. Introduction

Contributions:

We propose the **LOIRE** (**Lifelong learning framewOrk on Incremental data via PLMs gRowth Efficiently**)

- a novel plug-in layer growth operator that replicates the selected layers and inserts them between the original layer and the subsequent layer, leveraging concept of residual connection.
- a systematic definition for multi-dimensional operators and schedules that includes the layers, the hidden dimension, the FeedForward Network dimension, and the number of heads in the multi-head attention.
- an iterative distillation warmup strategy to accommodate the new data distribution without forgetting earlier distributions during model growth

2. Methodology

Definition of lifelong learning with an **efficient model growth strategy:**

K stages model growth: $M^{(t)}: M_1^{(t)} \Rightarrow M_2^{(t)} ... \Rightarrow M_K^{(t)}$

Optimal Objective:

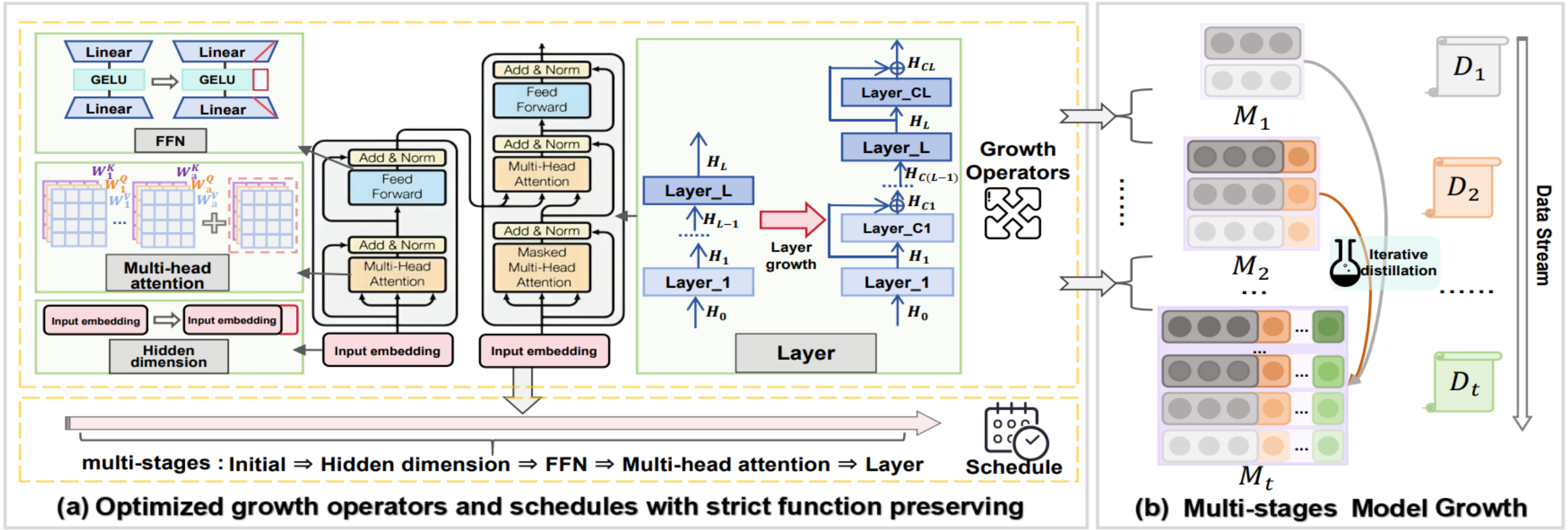
$$arg\underset{\varepsilon}{Min}\{\mathcal{L}(\overline{\varepsilon}), \mathcal{T}(\overline{\varepsilon})\}$$

$$s.t. \quad \begin{cases} \overline{\varepsilon} = \{M_k(x;\theta_k)\}_{k=1}^{K} \\ \theta_k = \varphi(\theta_k^+) + \theta_{k-1}, \varphi \in \overline{\psi} \end{cases}$$

Strict function preserving: $\forall x, M_k(x;\theta_k) = M_k(x;\varphi(\theta_k^+) + \theta_{k-1}) = M_{k-1}(x;\theta_{k-1})$

2. Methodology

The overall framework of the proposed LOIRE

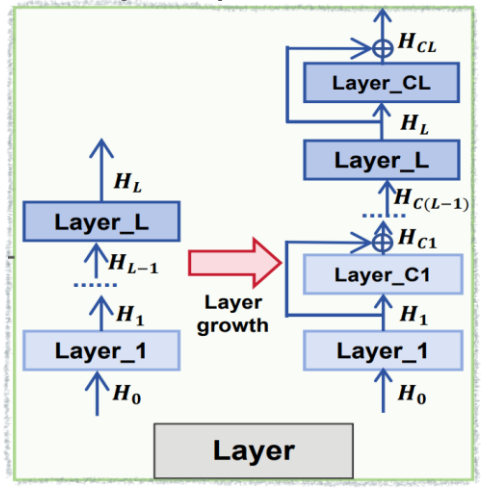


(a) the growth operators and schedules over all four possible Transformer dimensions
(b) multi-stage model growth procedure with an interactive distillation approach

2. Methodology

Growth operator with strict function preservation

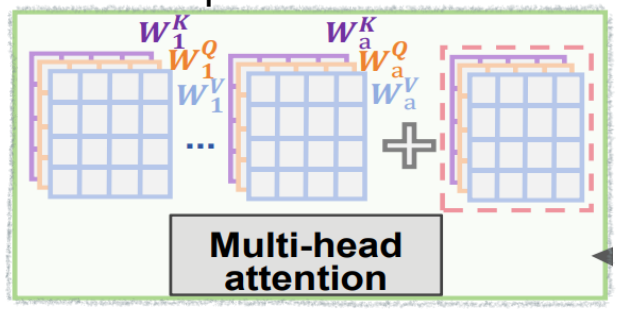
■ Layer operator



$$H^{Cl} = \lambda_l H^l + (1-\lambda_l) Trans_l(H^l), l \in [1, L]$$

$$\begin{aligned} H^{CL} &= Trans(H^{C(L-1)}) = Trans(H^{L-1}) \\ &= Trans^{[2]}(H^{C(L-2)}) = Trans^{[2]}(H^{L-2}) \\ &\cdots \\ &= Trans^{[L-1]}(H^{C1}) = Trans^{[L-1]}(H^1) \end{aligned}$$

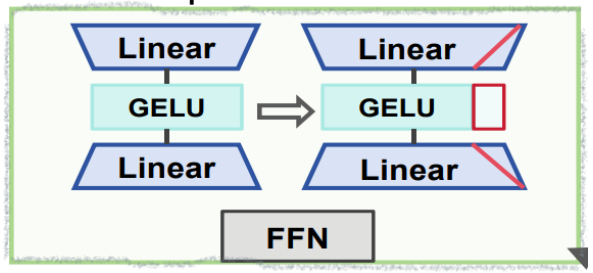
■ MHA operator



$$W_i^{K/Q/V} = \begin{cases} W_i^{K/Q/V} & i \le a_1 \\ any\_value & a_i < 1 \le a_2 \end{cases}$$

$$W^O_{(a_1\times d)\times h} \Rightarrow (W^O)'_{(a_2\times d)\times h} = \begin{bmatrix} W^O_{(a_1\times d)\times h} \\ N_{((a_2-a_1)\times d)\times h} \end{bmatrix}$$

■ FFN operator

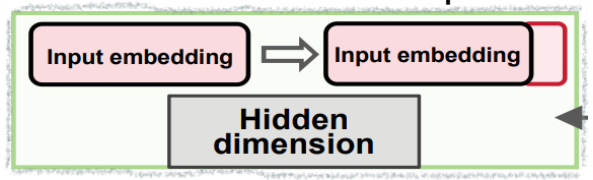


$$W^{l_1}_{h\times f_1} \Rightarrow (W^{l_1}_{h\times f_2})' = \begin{bmatrix} W^{l_1}_{h\times f_1} & R^{W_{l_1}}_{h\times(f_2-f_1)} \end{bmatrix}$$

$$b^{l_1}_{s\times f_1} \Rightarrow (b^{l_1}_{s\times f_2})' = \begin{bmatrix} b^{l_1}_{s\times f_1} & R^{W_{l_1}}_{s\times(f_2-f_1)} \end{bmatrix}$$

$$W^{l_2}_{f_1\times h} \Rightarrow (W^{l_2}_{f_2\times h})' = \begin{bmatrix} W^{l_2}_{f_1\times h} \\ N_{(f_2-f_1)\times h} \end{bmatrix}$$

■ Hidden dimension operator



Hidden states:

$$H_{s\times h_1} \Rightarrow H'_{s\times h_2} = \begin{bmatrix} H_{s\times h_1} & N_{s\times(h_2-h_1)} \end{bmatrix}$$

MHA:

$$W^{l_1}_{h_1\times f} \Rightarrow (W^{l_1}_{h_2\times f})' = \begin{bmatrix} W^{l_1}_{h_1\times f} \\ R_{(h_2-h_1)\times f} \end{bmatrix}$$

$$W^{l_2}_{f\times h_1} \Rightarrow (W^{l_2}_{f\times h_2})' = \begin{bmatrix} W^{l_2}_{f\times h_1} & N_{f\times(h_2-h_1)} \end{bmatrix}$$

$$b^{l_2}_{s\times h_1} \Rightarrow (b^{l_2}_{s\times h_2})' = \begin{bmatrix} b^{l_2}_{s\times h_1} & N_{s\times(h_2-h_1)} \end{bmatrix}$$

FFN:

$$W^{K/Q/V}_{h_1\times d} \Rightarrow (W^{K/Q/V})'_{h_2\times d} = \begin{bmatrix} W^{K/Q/V}_{h_1\times d} \\ R_{(h_2-h_1)\times d} \end{bmatrix}$$

$$W^O_{(a\times d)\times h_1} \Rightarrow (W^O_{(a\times d)\times h_2})' = \begin{bmatrix} W^O_{(a\times d)\times h_1} & N_{(a\times d)\times(h_2-h_1)} \end{bmatrix}$$

2. Methodology

Optimized multi-stage growth schedule

growing the layers and heads in later stages and having a larger hidden dimension in earlier stages can lead to better model performance

$$\bar{\varepsilon} = \{M_1(x;\theta_1), M_2(x;\theta_2), M_3(x;\theta_3), M_4(x;\theta_4), M_5(x;\theta_5)\}$$

$$\theta_1 \Rightarrow \theta_2[\varphi_{hidden}(\theta_2^+) + \theta_1] \Rightarrow \theta_3[\varphi_{ffn}(\theta_3^+) + \theta_2] \Rightarrow \theta_4[\varphi_{mha}(\theta_4^+) + \theta_3] \Rightarrow \theta_5[\varphi_{layer}(\theta_5^+) + \theta_4]$$

Schedule

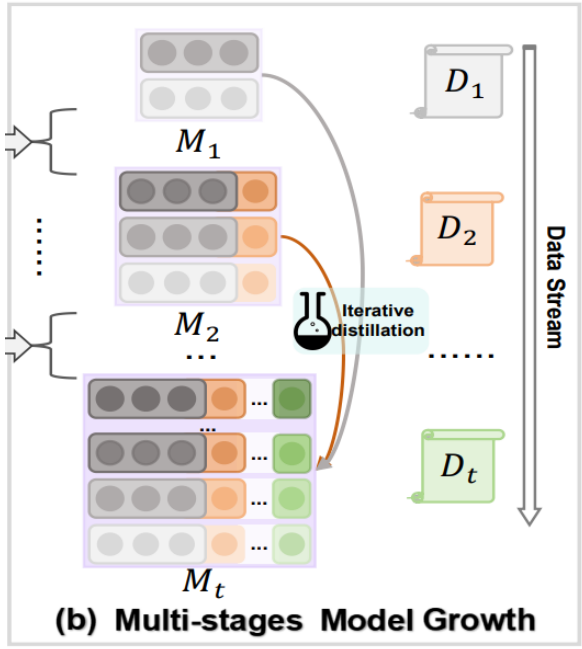
Hidden→FFN→MHA→LAYER

Iterative Distillation warmup

minimize the difference between the distributions of previous models and the current one.

$$L_{distill} = -\sum_{i}^{m} \sum_{v_n \in V} P(v_n|x_{i<m}, \theta_j) log \frac{P(v_n|x_{i<m}, \theta_j)}{P(v_n|x_{i<m}, \theta_t)}, j \leq t-1$$

$$L_{final} = L_{LM} + \sum_{j \leq t-1} \beta_j L_{distill\_j}$$



3. Experiments

How does LOIRE perform during the pre-training stage?

Schedule	M_1		M_2		M_3		M_4		M_5	
Metrics	**AP**	**AP+**	**AP**	AP+	**AP**	**AP+**	**AP**	**AP+**	**AP**	**AP+**
Growing M_1 to M_5 from scratch										
GPT_S	38.69	-	48.83	23.44	82.15	96.61	82.30	117.04	59.59	83.25
GPT_L	22.43	-	27.69	12.86	48.87	55.50	47.75	64.3	36.01	48.88
GPT_R	22.43	-	23.78	3.42	25.92	6.37	22.61	9.14	20.66	8.58
Token KD	38.69	-	48.48	22.93	56.94	90.15	80.53	112.19	56.77	78.35
ER	38.69	-	42.28	9.87	45.30	14.45	40.24	21.67	35.94	19.22
ELLE	38.69	-	34.33	-0.79	31.72	1.21	25.595	4.3	21.75	3.13
LOIRE-GPT1	38.69	-	31.72	**-5.32**	29.18	**-2.68**	24.63	**-0.94**	**19.19**	**-3.84**
Growing M_1 to M_5 from loading an PLM										
LOIRE-GPT2	32.39	-	26.60	-5.79	25.55	-3.95	25.28	-2.90	23.24	-4.22
Growing M_1 to M_5 from loading an PLM										
PPL	Init	Final	Init	Final	Init	Final	Init	Final	Init	Final
LOIRE-Bert	457.41	6.72	7.48	5.68	5.9	4.98	5.18	4.7	4.79	3.91

LOIRE works both on pre-training from scratch and continually training from loading the checkpoint

Still deviations in the function-preserving

LOIRE shows that it is better at keeping knowledge as the training data grows

3. Experiments

How does LOIRE perform in terms of training efficiency?

Schedule	GPT base FLOPs(%)					BERT base Method	BERT base Wall Time
	M_1	M_2	M_3	M_4	AVG		
$\frac{LOIRE}{GPT\_L}$	21.20	78.42	85.69	85.69	76.20	**LIGO**	48h,25min
$\frac{LOIRE}{GPT\_R}$	19.08	70.58	77.12	77.12	70.78	**LOIRE-Bert**	28h,48min

LOIRE can effectively save training time and improve training efficiency

3. Experiments

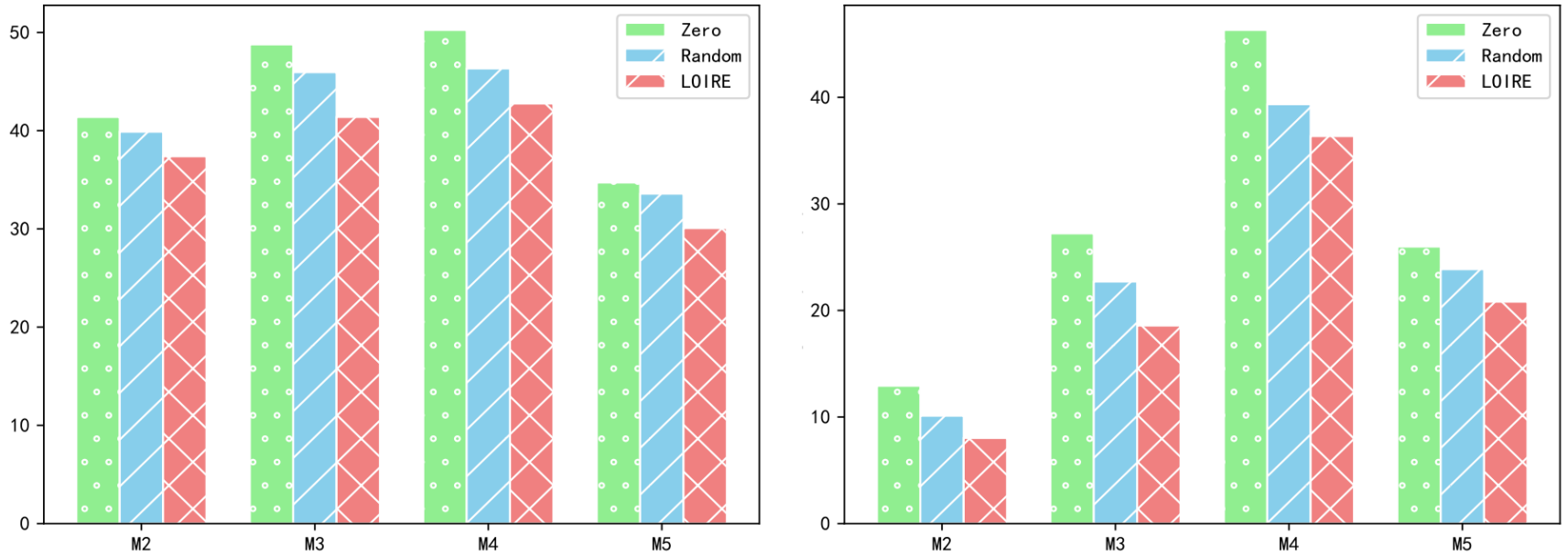
How does LOIRE perform on the multi-domain downstream tasks?

Domain	WB		NEWS		REV		BIO		CS		Avg
Task	MNLI	QNLI	Hyper	Ag news	HELPNESS	IMDB	CHEM	RCT	ACL-ARC	SCIERC	
GPT_S	73.39	80.79	77.08	92.51	86.16	91.58	78.01	86.99	69.53	80.00	81.61
GPT_L	78.69	81.23	75.76	93.02	86.41	92.09	79.74	87.23	70.31	**82.91**	82.74
GPT_R	79.53	82.35	76.38	**93.33**	86.93	93.08	80.57	87.36	69.53	82.70	83.18
Token KD	75.58	80.61	75.16	92.67	86.32	91.55	76.12	86.91	69.53	78.54	81.30
ER	77.05	80.94	78.24	92.61	**87.57**	91.52	77.98	87.13	70.31	82.70	82.61
ELLE	78.12	83.77	78.75	93.21	86.59	92.81	79.98	87.00	73.43	79.79	83.35
LOIRE-GPT1	**79.60**	**84.34**	**81.68**	93.12	87.16	**93.57**	**81.27**	**87.40**	**78.13**	82.08	**84.84**

LOIRE achieves comparable performance while saving computational costs.

3. Experiments

Are our function-preserving operators effective?



(a) AP

(b) AP+

LOIRE's AP and AP+ are significantly lower than those of zero and random

4. Future works

- A gap with existing LLMs with parameters up to 65 billion.
- There are still some deviations from the function-preserving theory in the actual experimental results.
- LOIRE performs worse on a few downstream tasks than other baselines.
- Other data-driven modalities, such as images and video.

Thanks for your listening!

hanxueai@chinamobile.com