北京大學
PEKING UNIVERSITY

Mixture of Experts Meets Prompt-Based Continual Learning

Jiayu Yao

Multi-head Self Attention & MoE

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_k}})\boldsymbol{V}$$

$$\mathrm{MSA}(\boldsymbol{X}^Q,\boldsymbol{X}^K,\boldsymbol{X}^V) := \mathrm{Concat}(\boldsymbol{h}_1,\ldots,\boldsymbol{h}_m)W^O \in \mathbb{R}^{N\times d},$$

$$\boldsymbol{h}_i := \mathrm{Attention}(\boldsymbol{X}^Q W_i^Q, \boldsymbol{X}^K W_i^K, \boldsymbol{X}^V W_i^V),\ i\in[m].$$

$$\mathbf{y} := \sum_{j=1}^{N} G(\boldsymbol{h})_j \cdot f_j(\boldsymbol{h}) := \sum_{j=1}^{N} \frac{\exp\left(s_j(\boldsymbol{h})\right)}{\sum_{\ell=1}^{N}\exp\left(s_\ell(\boldsymbol{h})\right)} \cdot f_j(\boldsymbol{h}),$$

Prefix Tuning

$$\boldsymbol{h}_{l,i} = \sum_{j=1}^{N} \frac{\exp\left(\frac{\boldsymbol{x}_i^\top W_l^Q {W_l^K}^\top \boldsymbol{x}_j}{\sqrt{d_v}}\right)}{\sum_{k=1}^{N} \exp\left(\frac{\boldsymbol{x}_i^\top W_l^Q {W_l^K}^\top \boldsymbol{x}_k}{\sqrt{d_v}}\right)} {W_l^V}^\top \boldsymbol{x}_j = \sum_{j=1}^{N} \frac{\exp(s_{i,j}(\boldsymbol{X}))}{\sum_{k=1}^{N} \exp(s_{i,k}(\boldsymbol{X}))} f_j(\boldsymbol{X}),$$

$$\tilde{\boldsymbol{h}}_l = \mathrm{Attention}\left(\boldsymbol{X}^Q W_l^Q, \begin{bmatrix} \boldsymbol{p}^K \\ \boldsymbol{X}^K \end{bmatrix} W_l^K, \begin{bmatrix} \boldsymbol{p}^V \\ \boldsymbol{X}^V \end{bmatrix} W_l^V\right) = \left[\tilde{\boldsymbol{h}}_{l,1}, \ldots, \tilde{\boldsymbol{h}}_{l,N}\right]^\top \in \mathbb{R}^{N \times d_v},$$

$$\tilde{\boldsymbol{h}}_{l,i} = \sum_{j=1}^{N} \frac{\exp(s_{i,j}(\boldsymbol{X}))}{\sum_{k=1}^{N} \exp(s_{i,k}(\boldsymbol{X})) + \sum_{k'=1}^{L} \exp(s_{i,N+k'}(\boldsymbol{X}))} f_j(\boldsymbol{X})$$

$$+ \sum_{j'=1}^{L} \frac{\exp(s_{i,N+j'}(\boldsymbol{X}))}{\sum_{k=1}^{N} \exp(s_{i,k}(\boldsymbol{X})) + \sum_{k'=1}^{L} \exp(s_{i,N+k'}(\boldsymbol{X}))} f_{N+j'}(\boldsymbol{X})$$

$$\mathcal{O}\left(\frac{1}{\log^\tau n}\right)$$

Non-linear Residual Gate Meet Prefix Tuning

$$\hat{s}_{i,N+j}(\boldsymbol{X}) := \frac{\boldsymbol{X}^\top E_i^\top W_l^Q {W_l^K}^\top \boldsymbol{p}_j^K}{\sqrt{d_v}} + \alpha \cdot \sigma\left(\tau \cdot \frac{\boldsymbol{X}^\top E_i^\top W_l^Q {W_l^K}^\top \boldsymbol{p}_j^K}{\sqrt{d_v}}\right)$$
$$= s_{i,N+j}(\boldsymbol{X}) + \alpha \cdot \sigma(\tau \cdot s_{i,N+j}(\boldsymbol{X})),\ i \in [N],\ j \in [L],$$

$$g_{G_*}(\boldsymbol{X}) := \sum_{j=1}^{N} \frac{\exp(\boldsymbol{X}^\top B_j^0 \boldsymbol{X} + c_j^0)}{T(\boldsymbol{X})} \cdot h(\boldsymbol{X}, \eta_j^0)$$
$$+ \sum_{j'=1}^{L} \frac{\exp((\beta_{1j'}^*)^\top \boldsymbol{X} + \alpha\sigma(\tau(\beta_{1j'}^*)^\top \boldsymbol{X}) + \beta_{0j'}^*)}{T(\boldsymbol{X})} \cdot h(\boldsymbol{X}, \eta_{j'}^*),$$

Non-linear Residual Gate Meet Prefix Tuning

Theorem 4.1 (Regression Estimation Rate). *Equipped with a least squares estimator $\widehat{G}_n$ given in equation (15), the model estimation $g_{\widehat{G}_n}(\cdot)$ converges to the true model $g_{G_*}(\cdot)$ at the following rate:*

$$\|g_{\widehat{G}_n} - g_{G_*}\|_{L_2(\mu)} = \mathcal{O}_P(\sqrt{\log(n)/n}). \tag{16}$$

Theorem 4.3. *Assume that the expert function $h(x, \eta)$ and the activation $\sigma(\cdot)$ are algebraically independent, then we achieve the following lower bound for any $G \in \mathcal{G}_{L'}(\Theta)$:*

$$\|g_G - g_{G_*}\|_{L_2(\mu)} \gtrsim \mathcal{L}_1(G, G_*),$$

which together with Theorem 4.1 indicates that $\mathcal{L}_1(\widehat{G}_n, G_*) = \widetilde{\mathcal{O}}_P(n^{-1/2})$.

Non-linear Residual Gate Meet Prefix Tuning

What do you see?

Experiment

Table 1: Overall performance comparison on Split CIFAR-100 and Split ImageNet-R. We present Final Average Accuracy (FA), Cumulative Average Accuracy (CA), and Average Forgetting Measure (FM) of all methods under different pre-trained models.

PTM	Method	Split CIFAR-100			Split Imagenet-R		
		FA (↑)	**CA**(↑)	FM(↓)	**FA** (↑)	**CA**(↑)	FM(↓)
Sup-21K	L2P	83.06 ± 0.17	88.27 ± 0.71	5.61 ± 0.32	67.53 ± 0.44	71.98 ± 0.52	5.84 ± 0.38
	DualPrompt	87.30 ± 0.27	91.23 ± 0.65	3.87 ± 0.43	70.93 ± 0.08	75.67 ± 0.52	5.47 ± 0.19
	S-Prompt	87.57 ± 0.42	91.38 ± 0.69	3.63 ± 0.41	69.88 ± 0.51	74.25 ± 0.55	4.73 ± 0.47
	CODA-Prompt	86.94 ± 0.63	91.57 ± 0.75	4.04 ± 0.18	70.03 ± 0.47	74.26 ± 0.24	5.17 ± 0.22
	HiDe-Prompt	92.61 ± 0.28	94.03 ± 0.01	1.50 ± 0.28	75.06 ± 0.12	76.60 ± 0.01	$\mathbf{4.09} \pm 0.13$
	NoRGa (Ours)	$\mathbf{94.48} \pm 0.13$	$\mathbf{95.83} \pm 0.37$	$\mathbf{1.44} \pm 0.27$	$\mathbf{75.40} \pm 0.39$	$\mathbf{79.52} \pm 0.07$	4.59 ± 0.07
iBOT-21K	L2P	79.13 ± 1.25	85.13 ± 0.05	7.50 ± 1.21	61.31 ± 0.50	68.81 ± 0.52	10.72 ± 0.40
	DualPrompt	78.84 ± 0.47	86.16 ± 0.02	8.84 ± 0.67	58.69 ± 0.61	66.61 ± 0.67	11.75 ± 0.92
	S-Prompt	79.14 ± 0.65	85.85 ± 0.17	8.23 ± 1.15	57.96 ± 1.10	66.42 ± 0.71	11.27 ± 0.72
	CODA-Prompt	80.83 ± 0.27	87.02 ± 0.20	7.50 ± 0.25	61.22 ± 0.35	66.76 ± 0.37	9.66 ± 0.20
	HiDe-Prompt	93.02 ± 0.15	94.56 ± 0.05	$\mathbf{1.26} \pm 0.13$	70.83 ± 0.17	73.23 ± 0.08	$\mathbf{6.77} \pm 0.23$
	NoRGa (Ours)	$\mathbf{94.76} \pm 0.15$	$\mathbf{95.86} \pm 0.31$	1.34 ± 0.14	$\mathbf{73.06} \pm 0.26$	$\mathbf{77.46} \pm 0.42$	6.88 ± 0.49
iBOT-1K	L2P	75.51 ± 0.88	82.53 ± 1.10	6.80 ± 1.70	59.43 ± 0.28	66.83 ± 0.92	11.33 ± 1.25
	DualPrompt	76.21 ± 1.00	83.54 ± 1.23	9.89 ± 1.81	60.41 ± 0.76	66.87 ± 0.41	9.21 ± 0.43
	S-Prompt	76.60 ± 0.61	82.89 ± 0.89	8.60 ± 1.36	59.56 ± 0.60	66.60 ± 0.13	8.83 ± 0.81
	CODA-Prompt	79.11 ± 1.02	86.21 ± 0.49	7.69 ± 1.57	66.56 ± 0.68	73.14 ± 0.57	7.22 ± 0.38
	HiDe-Prompt	93.48 ± 0.11	95.02 ± 0.01	1.63 ± 0.10	71.33 ± 0.21	73.62 ± 0.13	7.11 ± 0.02
	NoRGa (Ours)	$\mathbf{94.01} \pm 0.04$	$\mathbf{95.11} \pm 0.35$	$\mathbf{1.61} \pm 0.30$	$\mathbf{72.77} \pm 0.20$	$\mathbf{76.55} \pm 0.46$	$\mathbf{7.10} \pm 0.39$
DINO-1K	L2P	72.23 ± 0.35	79.71 ± 1.26	8.37 ± 2.30	57.21 ± 0.69	64.09 ± 0.74	7.47 ± 0.96
	DualPrompt	73.95 ± 0.49	81.85 ± 0.59	9.32 ± 1.42	57.98 ± 0.71	65.39 ± 0.27	9.32 ± 0.69
	S-Prompt	74.39 ± 0.17	81.60 ± 0.74	9.07 ± 1.13	57.55 ± 0.72	64.90 ± 0.13	8.73 ± 0.56
	CODA-Prompt	77.50 ± 0.64	84.81 ± 0.30	8.10 ± 0.01	63.15 ± 0.39	69.73 ± 0.25	6.86 ± 0.11
	HiDe-Prompt	92.51 ± 0.11	94.25 ± 0.01	1.67 ± 0.20	68.11 ± 0.18	71.70 ± 0.01	6.45 ± 0.58
	NoRGa (Ours)	$\mathbf{93.43} \pm 0.33$	$\mathbf{94.65} \pm 0.62$	$\mathbf{1.65} \pm 0.25$	$\mathbf{71.77} \pm 0.44$	$\mathbf{75.76} \pm 0.49$	$\mathbf{6.42} \pm 0.68$
MoCo-1K	L2P	77.24 ± 0.69	83.73 ± 0.70	5.57 ± 0.75	54.13 ± 0.67	62.09 ± 0.76	$\mathbf{4.88} \pm 0.42$
	DualPrompt	77.56 ± 0.63	84.37 ± 0.51	6.54 ± 0.50	54.45 ± 0.30	62.92 ± 0.41	5.34 ± 0.41
	S-Prompt	77.20 ± 0.39	84.47 ± 0.37	7.00 ± 0.62	53.94 ± 0.32	62.42 ± 0.51	5.16 ± 0.48
	CODA-Prompt	77.83 ± 0.34	84.97 ± 0.23	12.60 ± 0.02	55.75 ± 0.26	65.49 ± 0.36	10.46 ± 0.04
	HiDe-Prompt	91.57 ± 0.20	93.70 ± 0.01	$\mathbf{1.51} \pm 0.17$	63.77 ± 0.49	68.26 ± 0.01	9.37 ± 0.71
	NoRGa (Ours)	$\mathbf{93.52} \pm 0.06$	$\mathbf{94.94} \pm 0.29$	1.63 ± 0.13	$\mathbf{64.52} \pm 0.16$	$\mathbf{70.21} \pm 0.64$	9.06 ± 0.19

Experiment

Table 2: Final average accuracy (FA) on Split CUB-200 and 5-Datasets.

Method	Split CUB-200		5-Datasets	
	Sup-21K	iBOT-21K	Sup-21K	iBOT-21K
L2P	75.46	46.60	81.84	82.25
DualPrompt	77.56	45.93	77.91	68.03
S-Prompt	77.13	44.22	86.06	77.20
CODA-Prompt	74.34	47.79	64.18	51.65
HiDe-Prompt	86.56	78.23	93.83	94.88
NoRGa (Ours)	**90.90**	**80.69**	**94.16**	**94.92**

Table 3: Ablation study of different activation functions, measured by final average accuracy (FA).

Method	Split CIFAR-100		Split CUB-200	
	Sup-21K	iBOT-21K	Sup-21K	iBOT-21K
HiDe-Prompt	92.61	93.02	86.56	78.23
NoRGa tanh	94.36	**94.76**	90.87	**80.69**
NoRGa sigmoid	**94.48**	94.69	**90.90**	80.18
NoRGa GELU	94.05	94.63	90.74	80.54

Thanks