On Zero-Initialized Attention: Optimal Prompt and Gating Factor Estimation

Nghiem Diep, Huy Nguyen*, Chau Nguyen*, Minh Le, Duy M. H. Nguyen, Daniel Sonntag, Mathias Niepert, Nhat Ho*

DFKI, HCMUS, UT Austin, Qualcomm AI, IMPRS-IS, Stuttgart, Oldenburg

❑ Motivation:

- **Challenge**: Fine-tune LLMs is expensive, make adaptation to new tasks difficult.
- **Solution**: LLaMA-Adapter [1] is proposed as a (PEFT) method for LLaMA models.
- Zero-initialized attention mitigate noise effect to the word tokens at the beginning of training
- **However,** theoretical foundations of zero-initialized attention remain largely unexplored.

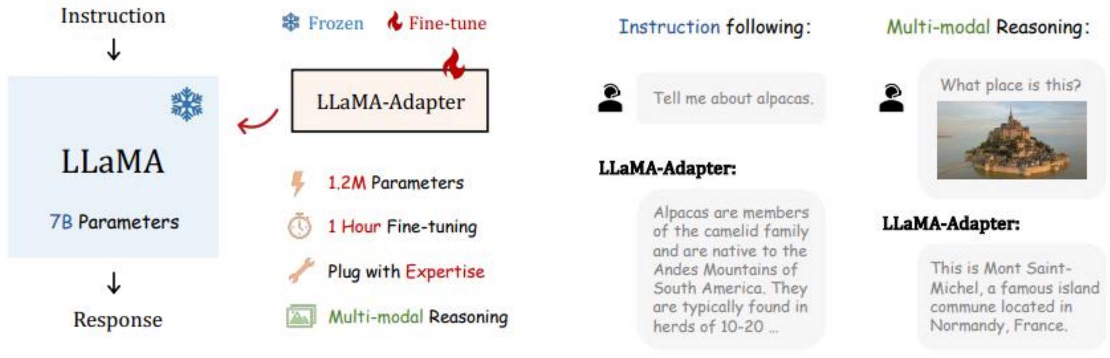

Figure 1: **Characteristics of LLaMA-Adapter.** Our lightweight adaption method efficiently fine-tunes LLaMA (Touvron et al., 2023) 7B model with only 1.2M learnable parameters within one hour, which exhibits superior instruction-following and multi-modal reasoning capacity.

[1] Zhang, Renrui, et al. "Llama-adapter: Efficient fine-tuning of language models with zero-init attention." ICLR 2023

❑ Motivation:

⇒ **Key Innovation**: Zero-Initialized Mechanism.

- Conduct theoretical and empirical investigation into zero-initialized attention.
- This method theoretically linked to Mixture-of-Experts (MoE) models.
- Non-linear prompts further enhance performance, flexibility, and adaptability.

- **LLaMA-Adapter:**

- Attention score: $S = QK^T/\sqrt{C}$, which $S = [S^K, S^{M+1}]^T$, $S^K \in R^{K\times 1}$ and $S^{M+1} \in R^{(M+1)\times 1}$.

- Use zero-initialized, softmax function σ is applied as:

$$S^g = [\sigma(S^K)\cdot \tanh(g)\,;\sigma(S^{M+1})]^T$$

- Finally, output of attention:

$$t^o = Linear_o(S^g V) \in R^{1\times C}$$

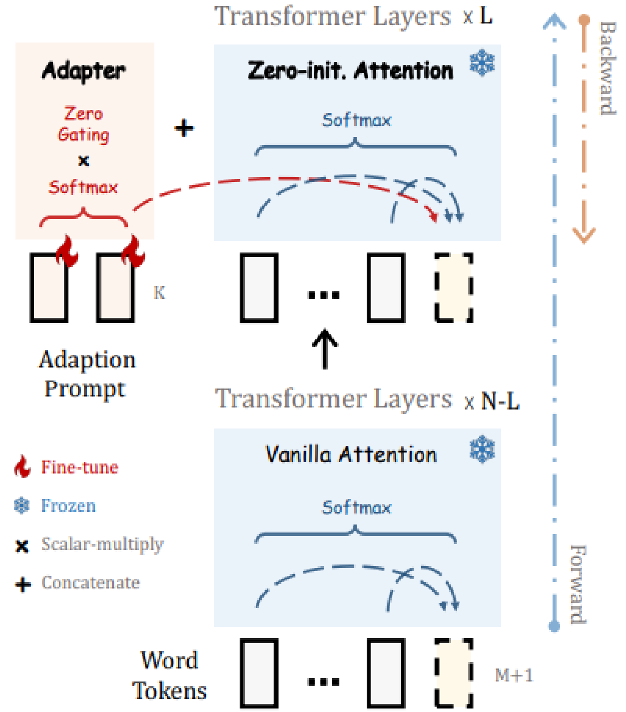

Figure 2: **Details of Zero-initialized Attention.** We insert learnable adaption prompts into the last L out of N transformer layers of LLaMA. To progressively learn the instructional knowledge, we adopt a zero gating factor within the attention for stable training in the early training stages.

❑ Zero-initialized Attention as MoE:

- Analyzing zero-initialized attention by viewing its components as gates and expert responses.
- Value matrix computed in attention is re–formularized as experts $f_i(.)$ and attention weights work as gating functions $G_i(.)$ over token interactions in MoE setting after rewriting softmax attention score matrix.
- Output of zero-initialized attention (having the MoE structure):

$$y = \sum_{j=1}^{M+1} G_j(X) \cdot f_j(X) + \tanh(g) \times \left(\sum_{j'=1}^{K} G_{M+1+j'}(X) \cdot f_{M+1+j'}(X) \right).$$

❑ Linear Prompt:

- **Problem settings:** Assume $\{(X_i, Y_i)\}_{i=1}^N$ are i.i.d samples from the following regression model:

$$Y_i = f_{G_*,\alpha_*}(X_i) + \epsilon_i, \qquad i \in [N]$$

$$f_{G_*,\alpha_*}(X) = \sum_{j=1}^{N} \frac{\exp\left(X^T \bar{A}_j^0 X + \bar{a}_j^0\right)}{\sum_{k=1}^N \exp(X^T \bar{A}_k^0 X + \bar{a}_k^0)} h\left(X, \bar{\eta}_j^0\right) + \tanh(\alpha_*) \cdot \sum_{j=1}^{L} \frac{\exp\left(\left(\bar{B} p_{*,j}\right)^T X + \bar{b}_{*,j}\right)}{\sum_{k=1}^L \exp\left(\left(\bar{B} p_{*,k}\right)^T X + \bar{b}_{*,k}\right)} \bar{C} p_{*,j}$$

- $G_* := \sum_{j=1}^L \exp\left(\bar{b}_{*,j}\right) \delta_{p_{*,j}}$ denote true but unknown measure.
- $\{\epsilon_i\}_{i=1}^N$ are independent Gaussian noise, $E(\epsilon_i | X_i) = 0$ and $Var(\epsilon_i | X_i) = \sigma^2 I$.

❑ Linear Prompt:

- Convergence rates of prompt estimation in original attention are significantly slow, standing at the order of $O_P(1/\log^{\tau}(n))$ for some constant $\tau > 0$, where n is the sample size.
- Convergence rates of linear prompt estimations are of polynomial orders, ranging from $O_P([\log(n)/n]^{\frac{1}{2}})$ to $O_P([\log(n)/n]^{\frac{1}{4}})$

➢ Faster than those under the original attention.

❑ Non-Linear Prompt:

$$f_{G_*,\alpha_*}(X) = \sum_{j=1}^{N} \frac{\exp\left(X^T \bar{A}_j^0 X + \bar{a}_j^0\right)}{\sum_{k=1}^{N} \exp\left(X^T \bar{A}_k^0 X + \bar{a}_k^0\right)} h\left(X, \bar{\eta}_j^0\right) + \tanh(\alpha_*) \cdot \sum_{j=1}^{L} \frac{\exp\left(\left(\bar{B}\sigma(p_{*,j})\right)^T X + \bar{b}_{*,j}\right)}{\sum_{k=1}^{L} \exp\left(\left(\bar{B}\sigma(p_{*,k})\right)^T X + \bar{b}_{*,k}\right)} \bar{C}\sigma(p_{*,j})$$

- Apply the same theoretical framework into non-linear prompt, the convergence rate also range from $O_P([\log(n)/n]^{\frac{1}{2}})$ to $O_P([\log(n)/n]^{\frac{1}{4}})$.

➢ Zero-initialized attention with non-linear prompts is also more sample-efficient than the random-initialized attention in terms of prompt convergence.

➢ Sharing the same sample complexity as when using linear prompts, zero-initialized attention with non-linear prompts will be shown to offer greater flexibility in practical applications.

❑ Non-Linear Prompt:

- Replace linear prompt P with non-linear prompt $\tilde{P} = \sigma(P) \in R^{K \times d}$, where:

$$\sigma(P) = f_2\left(\phi\left(f_1(P)\right)\right)$$

- Where $f_1(.), f_2(.)$ are separate linear layers, $\phi(.)$ is an activation (e.i. ReLU), and P is layer embedding.

- Ensure parameter efficiency and facilitate knowledge sharing across layers, this MLP is shared among the layers that utilize the prompts.

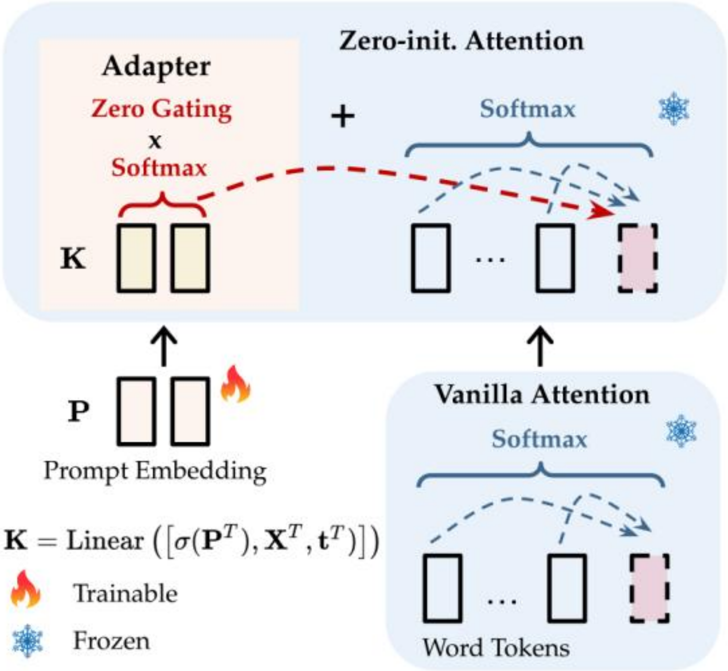

Figure 1. LLaMA-Adapter with non-linear prompt structures. Trainable prompts are integrated into the final layers of the LLaMA model, where a zero-gating mechanism modulates the added prompts. This approach enables progressive learning of instructional knowledge while keeping the remaining model parameters frozen.

❑ Linear Prompt vs Random-Init Prompt:

- Note that Random-Init Prompt is conventional attention combine with PEFT. Linear Prompt is zero-initialized attention combine with PEFT.

Table 1: Commparison between *Linear prompt* (zero-initialized mechanism) and *Random-Init* prompt on 4 LLM tasks using LLaMA-7B and LLaMA-13B models.

Method	ARC			MMLU	Hellaswag	TruthfullQA	Average
	Acc (eas)	*Acc (cha)*	*Acc (aver)*	*Acc*	*Acc*	*Acc*	
LLaMA-7B + zero-init	62.29 ↑1.64	43.17 ↑2.47	52.73 ↑2.06	36.28 ↑1.16	76.79 ↑4.17	45.53 ↑7.71	52.83 ↑3.77
LLaMA-7B + rand-init	60.65	40.7	50.67	35.12	72.62	37.82	49.06
LLaMA-13B + zero-init	81.78 ↑0.17	64.33 ↑0.42	73.06 ↑0.3	49.64 ↑1.62	81.21 ↑0.05	34.88 ↑0.36	59.70 ↑0.58
LLaMA-13B + rand-init	81.61	63.91	72.76	48.02	81.16	34.52	59.12

❑ Linear Prompt vs Non-Linear Prompt:

- Note that Non-Linear Prompt is zero-initialized attention combine with PEFT, and prompt is applied with non-linear mlp. Linear Prompt is zero-initialized attention combine with PEFT.

Table 2: Comparison of `Non-Linear prompt`, `Linear prompt`, and various fine-tuning methods. **Params** denote the total number of parameters updated during the fine-tuning process. **Bold** values indicate better scores between linear and non-linear settings.

Method	Params	ARC			MMLU	Hellaswag	TruthfullQA	Average
		Acc (eas)	*Acc (cha)*	*Acc (aver)*	*Acc*	*Acc*	*Acc*	
LLaMA-7B, Fully Fine-tuning Alpaca	7B	67.47	46.25	56.86	37.25	77.09	42.35	53.39
LLaMA-7B, LoRA Alpaca	4.2M	61.91	42.15	52.03	34.87	77.53	46.14	52.64
LLaMA-7B + zero-init + linear	1.2M	62.29	43.17	52.73	36.28	**76.79**	**45.53**	52.83
LLaMA-7B + zero-init + non-linear	2.6M	**63.51**	**45.39**	**54.45**	**36.95**	76.67	45.04	**53.28**
LLaMA-13B + zero-init + linear	1.9M	81.78	64.33	73.06	49.64	81.21	34.88	59.70
LLaMA-13B + zero-init + non-linear	3.3M	**82.87**	**66.55**	**74.71**	**51.32**	**81.72**	**38.92**	**61.67**

❑ Sample Efficiency:

- Note that Non-Linear Prompt is zero-initialized attention combine with PEFT, and prompt is applied with non-linear mlp. Linear Prompt is zero-initialized attention combine with PEFT.

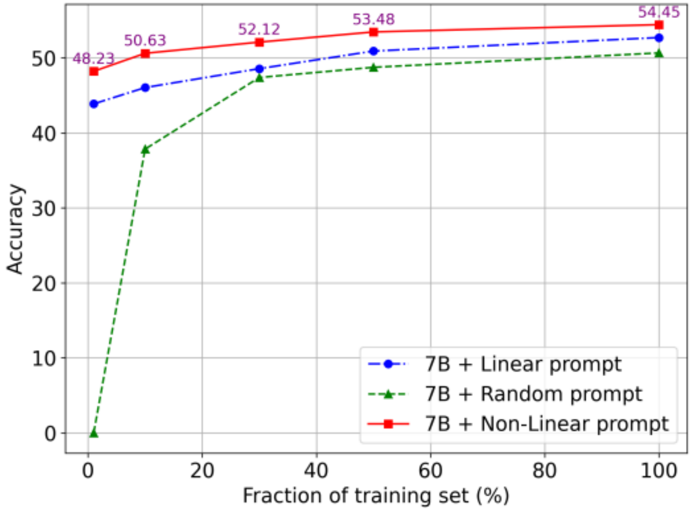

Figure 2. Sample efficiency comparison of three prompt-tuning initialization strategies on the ARC Dataset with LLaMA-7B.

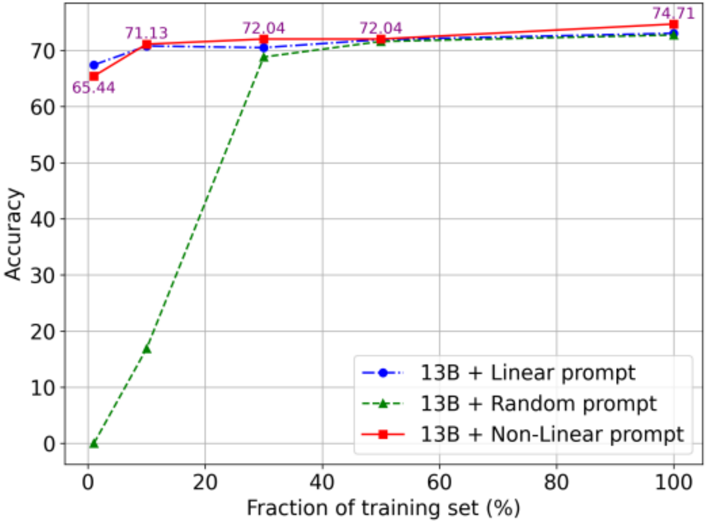

Figure 3. Sample efficiency comparison of three prompt-tuning initialization strategies on the ARC Dataset with LLaMA-13B.

Thank you for listening