Statistical Advantages of Perturbing Cosine Router in Mixture of Experts

Huy Nguyen, Pedram Akbarian, Trang Pham, Trang Nguyen, Shuijan Zhang, Nhat Ho

Presenter: Huy Nguyen
The University of Texas at Austin

TEXAS
The University of Texas at Austin

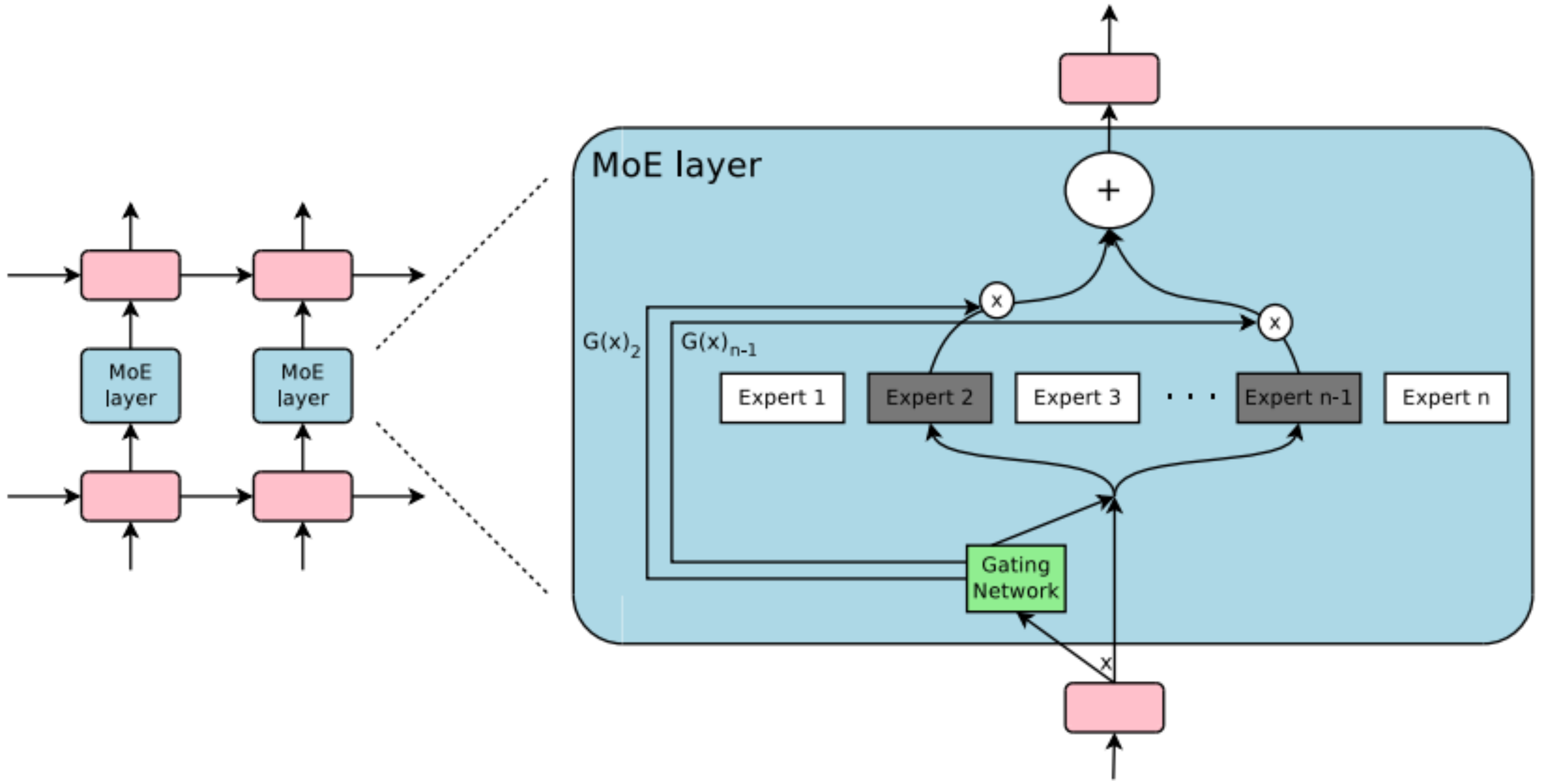
March 2025

ICLR
International Conference On Learning Representations

Sparse Mixture of Experts

- **Sparse mixture of experts** (MoE) [1] employs an **adaptive router** to activate only **a few experts** per input.
- $\longrightarrow$ **Increase the model capacity** while **remaining the computation overhead.**
- **Formulation:**

$$y = \sum_{i=1}^{k} \mathrm{softmax}(\mathrm{TopK}(\underbrace{\mathscr{R}(x)}_{\text{Router}}))_i \cdot \underbrace{h(x, \eta_i)}_{\text{Experts}}$$



[1] N. Shazeer et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In ICLR, 2017.

Router Choices

$$y = \sum_{i=1}^{k} \mathrm{softmax}(\mathrm{TopK}(\mathscr{R}(x)))_i \cdot h(x, \eta_i)$$

- **Linear router [1]:** $\mathscr{R}(x) := \left(\beta_{1i}^{\top} x + \beta_{0i}\right)_{i=1}^{k}$ → **Representation collapse issue [2].**

- **Cosine router [2]:** $\mathscr{R}(x) := \left(\frac{\beta_{1i}^{\top} x}{\|\beta_{1i}\| \cdot \|x\|} + \beta_{0i}\right)_{i=1}^{k}$ → **Alleviate representation collapse** but **slow expert convergence**.

- **Perturbed cosine Router (Ours):** $\mathscr{R}(x) := \left(\frac{\beta_{1i}^{\top} x}{(\|\beta_{1i}\| + \tau_1) \cdot (\|x\| + \tau_2)} + \beta_{0i}\right)_{i=1}^{k}$ → **Alleviate representation collapse** and **improve expert convergence**.

[1] N. Shazeer et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In ICLR, 2017.

[2] Z. Chi et al. On the Representation Collapse of Sparse Mixture of Experts. Advances in NeurIPS, 2022.

Expert Convergence Analysis

- **Setup:** Suppose that the data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ in $\mathbb{R}^d \times \mathbb{R}$ are sampled from the regression model:

$$Y_i = f_{G_*}(X_i) + \varepsilon_i, \quad i = 1,2,\ldots,n,$$

  - IID input: $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mu$
  - Independent Gaussian noise variables: $\varepsilon_i | X_i \sim \mathcal{N}(0,\nu)$
  - The regression function: $f_{G_*}(x) := \sum_{i=1}^{k_*} \mathrm{softmax}(\mathscr{R}(x; \beta_1^*, \beta_0^*))_i \cdot h(x, \eta_i^*)$.

Least Squares Estimation

- **Least squares estimation:** We estimate parameters via estimating mixing measure $G_* = \sum_{i=1}^{k_*} \exp(\beta_{0i}^*)\delta_{(\beta_{1i}^*,\eta_i^*)}$:

$$\widehat{G}_n := \arg\min_{G} \sum_{i=1}^{n} \Big(Y_i - f_G(X_i)\Big)^2.$$

- **Goals:** Determine the convergence rate of expert estimation $h(x, \hat{\eta}_i)$ to $h(x, \eta_i^*)$.

Practical Implications

Table 1: Summary of expert convergence rates.

Routers/ Experts	**Linear:** $a^\top x + b$	**Polynomial:** $(a^\top x + b)^p, p \geq 2$	**ReLU FFN**
Linear	$1/\log^\tau(n)$	$1/\log^\tau(n)$	$n^{-1/4}$
Cosine	$1/\log^\tau(n)$	$1/\log^\tau(n)$	$1/\log^\tau(n)$
Perturbed cosine	$1/\log^\tau(n)$	$n^{-1/4}$	$n^{-1/4}$

- **(P.1)** **Expert convergence rates are faster** when using the perturbed cosine router than than those when using the cosine/linear router.
- **(P.2)** The perturbed cosine router is **compatible with a broader range of experts** (polynomial and ReLU FFN experts) than the cosine/linear router.

Experiments: Language Modeling

- **Language modeling tasks.** We evaluate the model's pre-training capabilities on character-level language modeling using Enwik8 and Text8 datasets [3], and assess its word-level language modeling performance on Wikitext-103 [4].

Table 2: Performance of vanilla and perturbed cosine routers on language modeling tasks.

Router/Experts	**Enwik8 (BPC ↓)**		**Text8 (BPC ↓)**		**Wikitext-103 (PPL ↓)**	
	Small	**Medium**	**Small**	**Medium**	**Small**	**Medium**
Cosine	1.213	1.161	1.310	1.271	90.070	38.018
Perturbed cosine	**1.197**	**1.147**	**1.303**	**1.251**	**89.910**	**37.859**

[3] N. Mahoney. Large text compression benchmark, 2011.

[4] S. Merity et al. Pointer sentinel mixture models, 2016.

Experiments: Domain Generalization

- **Domain generalization tasks:** Generalizing a model's performance to unseen test domains with distributions different from those encountered during training.

Table 3: Average out-of-distribution test accuracies.

Router/Experts	PACS	VLCS	OfficeHome	TerraIncognita	DomainNet	Avg.
Linear	86.33	78.15	73.02	41.30	48.19	65.40
Cosine	87.22	78.99	73.27	45.55	48.45	66.70
Perturbed cosine	**89.36**	**80.01**	**74.09**	**49.87**	**48.51**	**68.37**

Table 4: Per-domain performance of PACS, VLCS, OfficeHome, TerraIncognita.

	Router/Experts	clipart	infograph	painting	quickdraw	real	sketch
DomainNet	Linear	**69.11**	**24.95**	54.81	16.88	68.95	54.41
	Cosine	68.05	24.48	**55.75**	17.39	69.41	55.59
	Perturbed	68.31	24.52	55.03	**17.90**	**69.46**	**55.83**

Table 5: Per-domain performance of DomainNet.

	Router/Experts	A	C	P	S
PACS	Linear	87.29	81.20	**98.50**	78.34
	Cosine	89.24	86.11	97.60	75.92
	Perturbed cosine	**89.87**	**86.97**	97.90	**82.68**
	Router/Experts	**C**	**L**	**S**	**V**
VLCS	Linear	97.53	63.65	74.09	77.33
	Cosine	98.59	67.42	70.88	**79.07**
	Perturbed cosine	98.59	**67.80**	**74.70**	78.95
	Router/Experts	**A**	**C**	**P**	**R**
OfficeHome	Linear	72.99	57.27	79.03	82.78
	Cosine	73.40	57.27	78.69	83.70
	Perturbed cosine	**74.64**	**57.85**	**79.59**	**84.27**
	Router/Experts	**L100**	**L30**	**L43**	**L46**
TerraIncognita	Linear	45.99	28.51	54.66	36.05
	Cosine	50.00	37.49	53.02	41.67
	Perturbed cosine	**57.59**	**43.30**	**56.93**	41.67

Thank You!