Rethinking Gating Mechanism in Sparse MoE: Handling Arbitrary Modality Inputs with Confidence-Guided Gate

Liangwei Nathan Zheng, Wei Emma Zhang, Mingyu Guo, Olaf Maennel and Weitong Chen

Adelaide University

AUSTRALIAN INSTITUTE FOR MACHINE LEARNING

ICML International Conference On Machine Learning

Introduction

Effectively managing missing modalities is a fundamental challenge in real-world multimodal learning scenarios, where data incompleteness often results from systematic collection errors or sensor failures. We propose **ConfSMoE** to introduce a two-stage imputation module to handle the missing modality problem for the SMoE architecture by taking the opinion of experts and revealing the insight of expert collapse from gradient analysis with strong empirical evidence. Inspired by our gradient analysis, ConfSMoE proposed a novel expert gating mechanism by detaching the softmax routing score to task confidence score w.r.t ground truth signal. This naturally relieves expert collapse without introducing additional load balance loss function. We show that the insights of expert collapse empirically align with other gating mechanism such as Gaussian and Laplacian gate.

Highlights

- We theoretically identified the optimization of SMoE token selection will drive a rich-get-richer expert due to the extreme distribution of softmax in router. This can be worse when the modality is missing.
- Typical Load balance loss generate a conflict gradient to SMoE architecture during the optimization, resulting an ambiguous expert selection.
- We address this by replacing the routing score to downstream task confidence score that guide the routing behavior specifically by the task-preference instead of token-expert preference.
- We also address missing modality problem in multimodal learning task by a two-stage imputation strategy: Impute the modality from its own distribution and Further refine the missing modality by cross-modal interaction.

Gradient Conflict of Load Balance Load

1. We assume the all load balance load is trying to make the routing distribution evenly, which the behaviour is equivalent to minise inverse of entropy of routing score. $\mathcal{L}_{Load} = \frac{1}{H(g)}$, where $H(g) = \sum_{i=1}^{N} g_i log(g_i)$ is the entropy of gating score. We can then have the gradient of $\mathcal{L}_{Load}$ in the form of Jacobian.

$$\mathbf{J}_{load} = \frac{\partial \mathcal{L}_{load}}{\partial \mathbf{g}} = \left[\frac{1}{\mathcal{H}(\mathbf{g})^2}(log\mathbf{g} + 1)^\top\right] \cdot (\mathbf{diag(g)} - \mathbf{g}\mathbf{g}^\top)$$

Will be negative when routing distribution is sharp

2. The set-up of Sparse MoE is as follow, where g_i is the gating score of expert E_i with input data h. m_i represents the TopK operation, which is a binary masking operation.

$$f(\mathbf{h}) = \mathbf{h} + \sum_{i=1}^{N} m_i g_i E_i(\mathbf{h})$$

Calculate the Jacobian of Spase MoE:

Always positive

$$\mathbf{J}_{MoE} = \underbrace{\mathbf{E(h)m(diag(g) - gg^\top)}}_{\text{Learning better routing score}} + \underbrace{\sum_{i=1}^{N} m_i g_i E_i'(\mathbf{h})}_{\text{Learning better representation}}$$

Gradient Conflict

Proposed Method

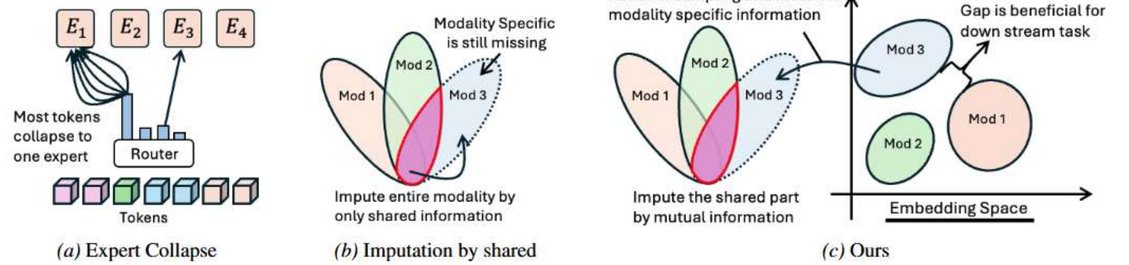


Figure 1: Expert Collapse of MoE and Comparison with Two Different Modality Imputation

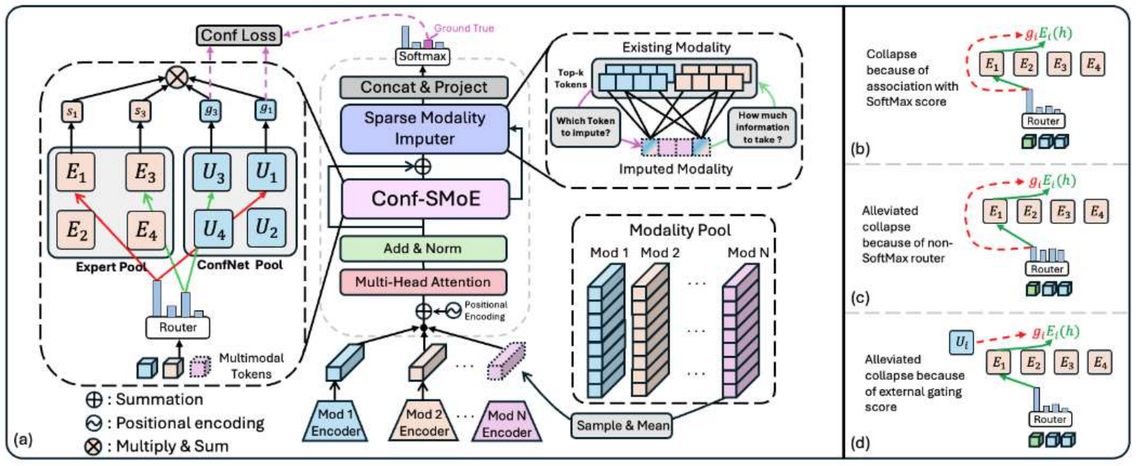
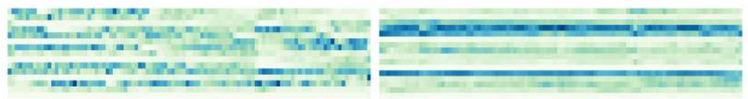


Figure 2: Proposed Method

Results



(a) Expert selection of softmax w/ $\mathcal{L}_{load}$ (b) Expert selection of ConfNet

Figure 3. Expert selection: X-axis represents epoch and Y-axis represents expert ID. The darker the color, the more selection will be for a particular expert. See detailed and more plots in Appendix H

Table 1. Experiment setting I: Main results on MIMIC-IV

Task	Metric	SMIL	ShaSpec	mmFormer	TF	LIMoE	FuseMoE-S	FuseMoE-L	FlexMoE	ConfSMoE-T	ConfSMoE-E
48-IHM	F1	39.58 ± 1.12	32.97 ± 1.08	45.39 ± 1.60	11.60 ± 0.54	43.76 ± 0.44	30.14 ± 1.15	40.21 ± 1.29	35.29 ± 0.30	**49.18 ± 0.22**	48.32 ± 0.30
	AUC	76.60 ± 0.98	78.29 ± 0.25	80.43 ± 0.76	67.08 ± 0.77	82.97 ± 0.61	71.34 ± 0.88	78.05 ± 0.72	80.45 ± 0.44	**85.24 ± 0.10**	85.09 ± 0.17
LOS	F1	58.52 ± 0.44	56.22 ± 0.28	57.06 ± 0.32	42.01 ± 1.12	59.03 ± 0.39	57.59 ± 0.76	58.31 ± 0.46	56.96 ± 0.53	61.33 ± 0.39	**61.35 ± 0.41**
	AUC	75.86 ± 0.66	72.22 ± 0.18	74.65 ± 0.32	64.11 ± 0.40	76.17 ± 0.17	72.59 ± 1.07	72.59 ± 0.61	74.81 ± 0.37	**78.22 ± 0.15**	77.85 ± 0.12
25-PHE	F1	27.77 ± 0.91	25.53 ± 0.11	26.15 ± 0.12	26.52 ± 0.29	25.38 ± 0.52	12.45 ± 0.54	12.25 ± 0.55	24.61 ± 0.71	**28.67 ± 0.33**	28.54 ± 0.25
	AUC	63.90 ± 1.31	62.57 ± 0.23	70.33 ± 0.12	56.28 ± 0.16	72.50 ± 0.67	55.48 ± 0.39	58.52 ± 0.27	71.57 ± 0.47	**74.56 ± 0.19**	74.42 ± 0.34

Table 2. Ablation I: Different router. CMU-MOSI adopt 50% missing setting in Experiment III

Task	Metric	Mean	Softmax	Softmax w/ $\mathcal{L}_{load}$	LFB	Gaussian	Laplacian	ConfNet
MIMIC-III	F1	48.34 ± 0.22	48.59 ± 1.97	51.67 ± 0.62	48.36 ± 0.88	46.50 ± 1.62	50.10 ± 1.17	**53.44 ± 0.27**
	AUC	85.13 ± 0.16	85.91 ± 0.84	85.97 ± 0.52	85.05 ± 0.67	86.70 ± 0.33	85.47 ± 1.05	**87.05 ± 0.58**
CMU-MOSI	F1	42.29 ± 0.87	41.47 ± 0.94	42.43 ± 1.34	43.86 ± 0.56	42.57 ± 1.81	43.15 ± 1.53	**44.34 ± 0.77**
	AUC	68.01 ± 1.68	68.77 ± 1.66	67.40 ± 0.54	66.60 ± 2.24	67.74 ± 0.45	67.60 ± 1.54	**70.41 ± 1.31**

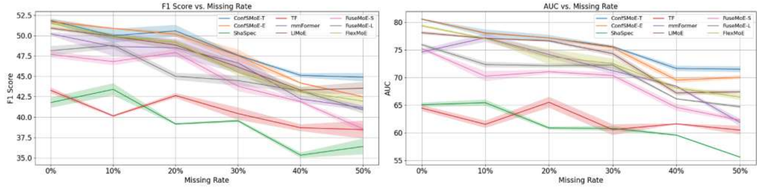


Figure 5: Experiment III random dropping modality in X%

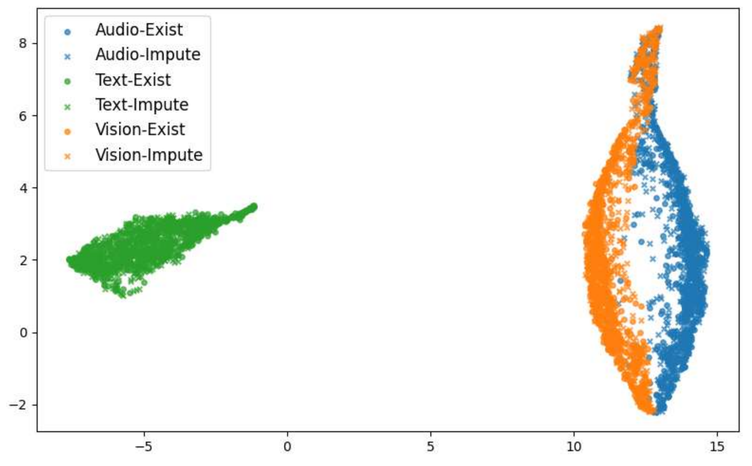


Figure 5: Imputation in Embedding Space