Proposal: Bayesian Neural Network for MoE

Daoyuan GUO
Student id: 50026444

Abstract. We propose a *lightweight Bayesian router* for sparse Mixture-of-Experts (MoE). A single Bayesian logistic layer produces per-expert probabilities. At inference, we choose the smallest k whose top-k mass covers at least $(1-\varepsilon)$ with confidence $(1-\delta)$ ("missing-mass coverage"). Training avoids hard thresholds via a soft budget on the effective number of experts plus diversity and load-balancing regularizers. We expect improved utilization, calibration, and robustness at negligible end-to-end overhead compared to fixed top-k [3–5].

Background & Motivation. Sparse MoE accelerates large models by activating few experts per input [3, 5], yet fixed top-k induces imbalance and does not guarantee that useful experts are included; deterministic gates are often miscalibrated under shift [4]. Lightweight Bayesian approximations add epistemic awareness with small cost [2]. We turn routing into a coverage-controlled statistical decision: input-adaptive k with predictable average compute.

Related Work. Sparsely-gated MoE and Switch Transformers scale via conditional computation [3, 5]. Calibration and uncertainty are addressed by temperature scaling and Bayesian approximations [2, 4]. A Gaussian logit with softmax induces a logistic-normal on probabilities, enabling Delta-method variance propagation [1].

Methods & Approach. **Bayesian router.** Given trunk features $h = f_\theta(x) \in \mathbb{R}^d$ and M experts, $\eta_i = w_i^\top h + b_i$ with priors $w_i \sim \mathcal{N}(0, \sigma_w^2 I)$, $b_i \sim \mathcal{N}(0, \sigma_b^2)$. Maintain means (μ_{w_i}, μ_{b_i}) and diagonal variances via diagonal-Fisher EMA (Laplace-Lite): $F_i \leftarrow \rho F_i + (1-\rho)p_i(1-p_i)(h \odot h)$, $s_i^2 \approx (F_i + \lambda I)^{-1}$. At inference, $m_i = \mu_{w_i}^\top h + \mu_{b_i}$, $v_i = h^\top \Sigma_{w_i} h + s_{b_i}^2$, $\tilde{m}_i = \frac{m_i}{\sqrt{1+\frac{\pi}{8}v_i}} \cdot \frac{1}{T}$, $p = \mathrm{softmax}(\tilde{m})$. **Inference.** Sort p to $p_{(1)} \geq \cdots \geq p_{(M)}$; let $S_k = 1 - \sum_{j=1}^{k} p_{(j)}$. Using a first-order Delta approximation for the logistic-normal, the tail variance is $\mathrm{Var}(S_k) \approx (1-S_k)^2 \sum_{j>k} v_{(j)} p_{(j)}^2 + S_k^2 \sum_{j\leq k} v_{(j)} p_{(j)}^2$. Choose the smallest k with $S_k + z_{1-\delta}\sqrt{\mathrm{Var}(S_k)} \leq \varepsilon$, then re-normalize weights within the selected set. **Training.** Per batch, set loss $\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{budget}}\big(N_{\text{eff}}(p) - k_{\text{target}}\big)^2 + \lambda_{\text{div}} \sum_i p_i^2 + \lambda_{\text{lb}} \mathrm{KL}(\bar{p}\|u) + \lambda_{\text{temp}}(T-1)^2$, with $N_{\text{eff}}(p) = 1/\sum_i p_i^2$. Warm-up: higher T and larger k_{target}, then anneal to run-time values. Forward uses hard selection (top-$\hat{k}(x)$) with standard gather/scatter; backward uses a straight-through estimator restricted to the selected subset.

Experiment Plan. Datasets like CIFAR-100, CIFAR-10C/SVHN. Baselines: dense; fixed top-k MoE; Bayesian-router+fixed-k. Metrics: Acc/Top-1, NLL, ECE, etc. Ablations: (ε, δ), λ_{budget}, temperature strategy.

Expected Results, Impact, and Complexity Estimation. (i) Input-adaptive k with explicit (ε, δ) guarantees; (ii) better robustness from Bayesian gating; (iii) improved utilization and fewer collapses with same budget; (iv) drop-in compatibility with existing MoE. Over a standard linear gate+softmax+top-k, we add one diagonal quadratic form for v (cost $\mathcal{O}(Md)$, $\sim$ one extra linear), existing sort/partial-topk, and an $\mathcal{O}(M)$ scan for coverage. Gate-side FLOPs $\approx 2\times$ the conventional gate, but gates are typically $< 5\%$ of total compute, yielding $\sim 1\% - 3\%$ overall overhead.

References

[1] J. Aitchison. *The Statistical Analysis of Compositional Data*. Chapman and Hall, 1986.
[2] E. Daxberger, A. Kristiadi, A. Fischer, P. Hennig, and D. Barber. Laplace redux: Effortless bayesian deep learning. In *NeurIPS*, 2021.
[3] W. Fedus, B. Zoph, and N. Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv:2101.03961*, 2021.
[4] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *ICML*, pages 1321–1330, 2017.
[5] N. Shazeer, A. Mirhoseini, K. Maziarz, et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv:1701.06538*, 2017.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).