TRAINING MIXTURE-OF-EXPERTS: A FOCUS ON EXPERT-TOKEN MATCHING

Fateme Vesaghati, Masoumeh Zareapoor
Tehran Azad University
Shanghai Jiao Tong University
mzarea@ieee.com

ABSTRACT

Recent advancements in sparse Mixture-of-Experts (MoE) models, particularly in the Vision MoE (VMoE) framework, have demonstrated promising results in enhancing vision task performance. However, a key challenge persists in optimally routing tokens (such as image patches) to the right experts, without incurring excessive computational costs. Addressing this, we apply the regularized optimal transport, which relies on the Sinkhorn algorithm to the Vision MoE (VMoE) framework, aiming at improving the token-expert matching process. The resulting model, Sinkhorn-VMoE (`SVMoE`), represents a meaningful step in optimizing efficiency and effectiveness of sparsely-gated MoE models.

1 INTRODUCTION

Sparse Mixture-of-Experts (MoEs) (Shazeer et al., 2017) models represent a significant advancement in deep learning. These models are characterized by their unique architecture, where only a small subset of the model, known as 'experts', are activated based on the input they receive. A key feature of MoEs is to learn a gating function to route each input token to the most appropriate experts. These models have shown impressive results in natural language processing (Zhou et al., 2022; Komatsuzaki et al., 2023; Sander et al., 2023). For detailed methodologies, see Fedus et al. (2022) and Puigcerver et al. (2023). Building upon this foundation, the Vision MoE (VMoE) (Riquelme et al., 2021), a novel adaptation of the Vision Transformer (ViT) (Dosovitskiy et al., 2021) for image classification is proposed. The idea of VMoEs is to replace a subset of dense feedforward layers MLP in ViT by MoEs, where each image patch is `routed` to a selected group of `experts` (MLPs). More precisely, the dense feedforward layers MLP: $x \in R^d \rightarrow \text{MLP}(x) \in R^d$ in ViT are replaced with MoEs layers, such that $MoE(x) = \sum_{r=1}^{t} G_r(x) \cdot MLP_r(x)$. This means we have multiple experts $MLP_r(.)$, and for each input x, the model uses a gating function $G_r(x)$ to decide which expert should be activated on that data. For each input only those experts with a nonzero gating function G_r are activated, meaning the input x is 'routed' to these selected experts r. However, we explore a different design methodology for training VMoE, which leads to better performance than conventional dense models. For a batch of image patches, we use the regularized optimal transport which relies on the Sinkhorn algorithm (Cuturi, 2013), to efficiently assign tokens to experts, while limiting the number of tokens assigned to any single expert. Our Sinkhorn VMoE (SVMoE) model refines this matching process, boosting the VMoE method's performance.

2 VMOE WITH SINKHORN ALGORITHM (SVMOE)

We use the VMoE architecture (Riquelme et al., 2021), with a specific focus on its routing mechanism that assigns input tokens to appropriate experts. Consider a set of t tokens $\{x_1, ..., x_t\}$ in R^d and a corresponding matrix $X \in R^{t \times d}$, where each row represents a token. Additionally, we have a matrix $W \in R^{d \times e}$ that contains the weights of the experts, with each column corresponding to the feature vector of an expert. The product of these two matrices $L = XW$, yields a new matrix $L \in R^{t \times e}$ that represents the affinities between tokens and experts, i.e., the entry in the i-th row and j-th column of L is the similarity score between the i-th token and the j-th expert, calculated via inner product. To assign tokens to experts, Riquelme et al. (2021) uses the Top_k router, creating a

gating matrix Π that selects the k largest values from each row of matrix L using the top_k function,

$$\Pi = \text{Top}_k(\texttt{softmax}(L + \alpha\sigma)) \in R^{t \times e} \tag{1}$$

where $L = XW$ denotes the token-to-expert affinity scores, and σ represents noise added into the token-expert matrix XW. The parameter α controls the noise intensity, which has a standard deviation of $1/e$ during the training. To optimize the model, Riquelme et al. (2021) set a buffer capacity for each expert, defining the maximum number of tokens an expert can process. With this capacity and the computed gating matrix Π, the Top_k router assigns each token to the highest-ranked expert, as long as the chosen expert has not reached its full capacity. In our model, the token-expert matching process is formulated as a regularized optimal transport problem, solved using the Sinkhorn algorithm. Our aim is to ensure that the number of tokens assigned to each expert does not exceed a predetermined buffer capacity, k. To achieve this, we define the gating matrix for regularized OT as $\Pi_{OT} = \underset{T \in u(a,b)}{\arg\min} < T, -L > + \frac{1}{2k}\|T\|_2^2$, with $a = 1_e$, $b = (t/e)1_e$. The first term, minimizes the total cost of assigning tokens to experts. The second term, ensures that the solution is sparse (not too many experts assigned per token). The matrix L is computed by $\texttt{softmax}(XW)$, and the transport plan Π_{OT} can be obtained using the Sinkhorn algorithm. With a computed Π_{OT}, the SVMoE router assigns each token to its top-chosen expert in the same way of the Top_k router.

3 EXPERIMENT

We replace the Top_k router with a regularized optimal transport router in standard VMoE architectures (see Figure 1). Specifically, we use the VMoE B/32 and B/16 architectures, which process image patches of sizes 32×32 and 16×16, respectively. In line with the 'Every-2' variant from (Riquelme et al., 2021), we placed SVMoEs in alternate layers of these architectures. For all experiments, we set the total number of experts to 32 ($e = 32$) and assigned 2 experts to each token, $k = 2$, as shown in the Appendix (Figure 3). Consequently, the buffer capacity is fixed to $e/k = 32/2 = 16$, meaning each expert handles a maximum of 16 tokens. We also used the ADAM optimizer (Kingma & Ba, 2015) for 50 steps with a learning rate of 10^{-2}. The results of our experiments are detailed in Table 1, focusing on the accuracy performance of the ImageNet (Deng et al., 2009) and JFT (Sun et al., 2017) datasets. Our findings indicate an improvement with the Sinkhorn VMoE (SVMoE) compared to the baseline VMoE with both B/16 and B/32 architectures on both benchmarks. Additionally, a reduction in training time was observed - the SVMoE enhances performance over VMoE by operating with 15% fewer FLOPs, while also 6% faster. This enhancement in training efficiency and classification accuracy highlights the effectiveness of SVMoE in large-scale vision tasks.

Table 1: Accuracy comparison of ViT, VMoE, and SVMoE on ImageNet-1K and JFT datasets using B/32 and B/16 architectures. Compared to the baseline methods, our SVMoE achieves the highest accuracy across both architectures. The performance in terms of FLOPs is given in the Appendix.

Method	**B/32 architecture**		**B/16 architecture**	
	JFT	ImageNet-1K	JFT	ImageNet-1K
ViT (Dosovitskiy et al., 2021)	39.86	59.30	45.93	66.97
VMoE (Riquelme et al., 2021)	43.51	65.93	49.16	72.34
SVMoE (Sinkhorn VMoE)	**44.32**	**66.47**	**50.29**	**72.85**

4 CONCLUSION

We presented Sinkhorn VMoE (SVMoE), an approach integrating regularized optimal transport that relies on the Sinkhorn algorithm, into the Vision MoE framework. Our key objective was to enhance the token-expert matching process while maintaining strict control over the computational costs. Our experiment demonstrates that SVMoE not only matches or slightly exceeds the performance of the VMoE models but also proves to be a scalable approach for large-scale image classification tasks. While our evaluation currently focuses on ImageNet and JFT, the applicability of SVMoE is not limited to image classification alone. Looking ahead, we plan to extend our research to assess the adaptability and effectiveness of SVMoE across a wider range of vision-based applications.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of the ICLR 2024 Tiny Papers Track.

REFERENCES

Aidan Clark, Diego de Las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, et al. Unified scaling laws for routed language models. In *ICML*, pp. 4057–4086. PMLR, 2022.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *NIPS*, 26, 2013.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

William Fedus, Jeff Dean, and Barret Zoph. A review of sparse expert models in deep learning. *arXiv preprint arXiv:2209.01667*, 2022.

Abdelwahed Khamis, Russell Tsuchida, Mohamed Tarek, Vivien Rolland, and Lars Petersson. Scalable optimal transport methods in machine learning: A contemporary survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. Sparse upcycling: Training mixture-of-experts from dense checkpoints. *ICLR*, 2023.

Wouter Kool, Chris J Maddison, and Andriy Mnih. Unbiased gradient estimation with balanced assignments for mixtures of experts. *NeurIPS I Can't Believe It's Not Better (ICBINB) Workshop*, 2021.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Joan Puigcerver, Carlos Riquelme, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. *arXiv preprint arXiv:2308.00951*, 2023.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *NIPS*, 34:8583–8595, 2021.

Michael Eli Sander, Joan Puigcerver, Josip Djolonga, Gabriel Peyré, and Mathieu Blondel. Fast, differentiable and sparse top-k: a convex analysis perspective. In *ICML*, pp. 29919–29936. PMLR, 2023.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *ICLR*, 2017.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, pp. 843–852, 2017.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. *NIPS*, 35:7103–7114, 2022.

A Related Work

A.1 Sparse mixture of experts (MoE)

In an MoE model, an expert refers to a sub-model designed to handle specific types of input data. A single MoE model contains multiple such experts. The gating mechanism is a crucial component that decides which experts are activated in response to a given input. The gating mechanism routes the inputs to one or more experts based on the input's characteristics. Unlike traditional deep learning models that utilize all parts of the network for processing any input, sparse MoEs activate only a subset of experts for each input. This sparsity is what makes the MoE model computationally efficient. Indeed, by activating only relevant parts of the model, sparse MoEs handle more parameters and computations more efficiently than standard models. Sparse MoE models represent a powerful paradigm for building scalable and efficient deep learning systems that can handle extensive and diverse datasets across different domains like computer vision (Riquelme et al., 2021; Komatsuzaki et al., 2023), natural language processing (Shazeer et al., 2017; Zhou et al., 2022), and more (Sander et al., 2023; Clark et al., 2022). However, a key feature of an MoE model is its routing mechanism, which determines which experts are best suited for a given input. Token-based routing is one of the most common among them. In transformer-based models, the routing mechanism can operate at the token level— each word or image patch (token) is independently routed to the most suitable expert. To balance the assignments of tokens to experts, recent works cast the assignment problem as regularized optimal transport (Clark et al., 2022; Kool et al., 2021) for language modeling tasks. In this paper, we use entropy-regularized optimal transport as a routing mechanism in vision sparse MoEs, to balance the load among experts, ensuring no single expert becomes a bottleneck. It introduces a form of regularization that maximizes entropy, thus spreading out the token assignments in a more balanced and efficient way.

A.2 Optimal Transport (OT)

OT is a mathematical framework used for comparing probability distributions and finding the most efficient way of transforming one distribution into another (Peyré et al., 2019). This framework has applications across various fields, including machine learning and data analysis (Khamis et al., 2024). We now break down the preliminary concepts of OT and its specific formulations. Let us define the probability simplex Σ^r as follows,

$$\Sigma_r := \{x \in \mathbb{R}_+^r | x^\top \mathbf{1}_r = 1\}$$

where $\mathbf{1}_r$ is an r-dimensional vector of ones. A vector x in this simplex are probability distributions because their elements are non-negative and sum to one. Given two distributions $\alpha \in \Sigma_r$ and $\beta \in \Sigma_c$, the transport polytope $U(\alpha, \beta)$ is defined as,

$$U(\alpha, \beta) = \{Q \in \mathbb{R}_+^{r \times c} | Q\mathbf{1}_c = \alpha, Q^\top \mathbf{1}_r = \beta\}$$

Here, Q represents a matrix of transport plans or joint probabilities where Q_{ij} is the amount of mass transported from the i-th element of α to the j-th element of β. Essentially, each matrix Q in this set describes a different way of transporting mass from distribution α to distribution β. However, incorporating an entropy regularization term leads to a more differentiable and often more computationally feasible problem that can be defined as

$$\text{OT}_\epsilon(M, \alpha, \beta) = \arg \max_{Q \in U(\alpha,\beta)} \text{Tr}(Q^\top M) + \epsilon H(Q)$$

where, M is a similarity (or cost) matrix between the elements of α and β. ϵ is a regularization parameter, and $H(Q) = -\sum_{ij} Q_{ij} \log Q_{ij}$ is the entropy of Q, which discourages extreme transport plans and promotes plans that spread out the mass more evenly. The matrix Q^*, which is the solution to the entropy-regularized OT problem, has a specific form given by, $Q^* = \text{Diag}(u) \exp(M/\epsilon) \text{Diag}(v)$, where u and v are vectors that can be computed efficiently using Sinkhorn's algorithm (Cuturi, 2013). This formulation allows to control the smoothness of the transport plan through the entropy ϵ, making it particularly useful in cases where distributions often need to be compared or aligned efficiently.

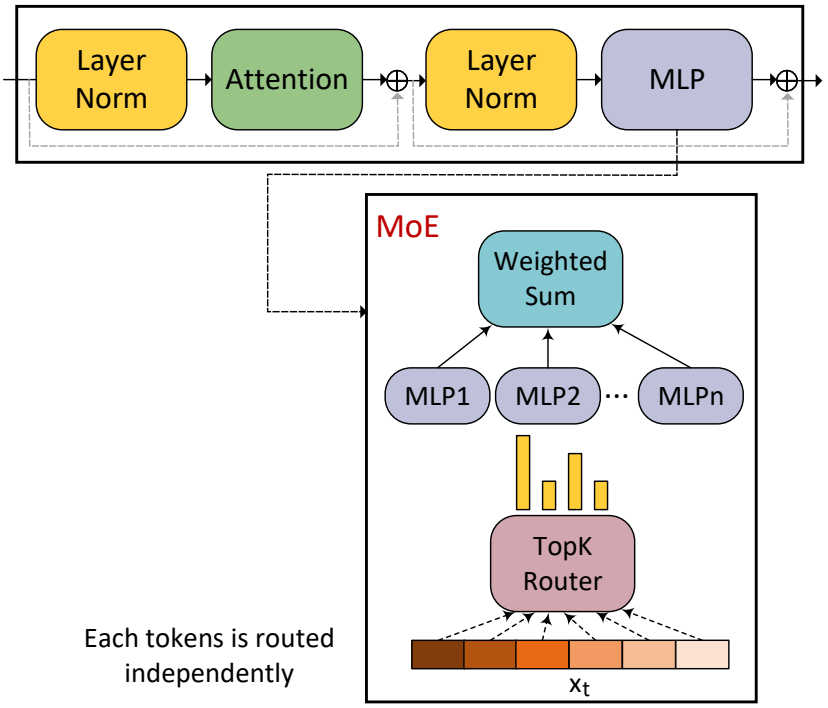

Figure 1: The VMoE model takes the basic design of a Vision Transformer (ViT) and replaces some of its standard dense feedforward layers, MLPs, with sparsely-gated mixture-of-experts (MoE) layers. Here, each token is evaluated independently, and only the most relevant experts (MLP) for that particular token (x) are activated. The TopK router is the decision-maker. It compares it with the experts (MLP) and then creates a sparse gating matrix to find which experts are the best match for that token. In our model (SVMoE), we used regularized optimal transport as the TopK router.

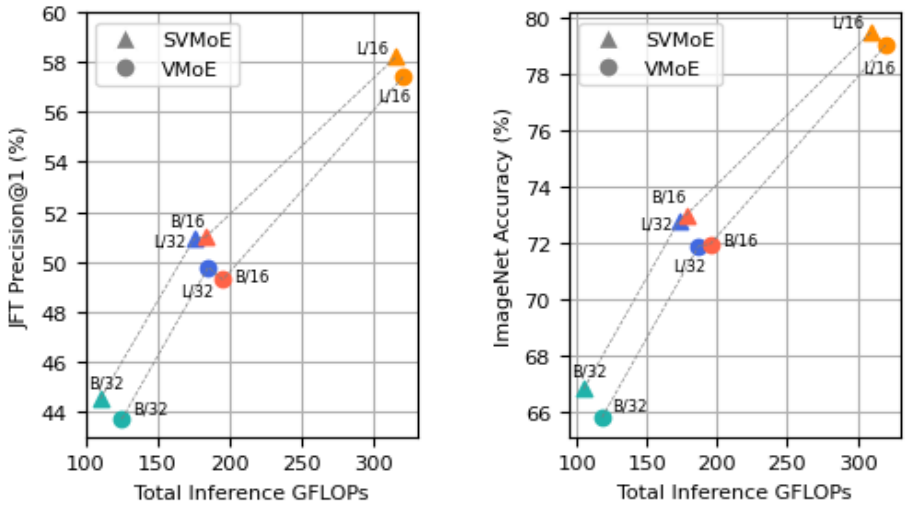

Figure 2: Performance vs. inference FLOPs for VMoEs and SVMoEs. Each color represents a different VMoE variant (B/32, B/16, L/32, L/16). Our model slightly outperforms the VMoEs with fewer FLops. VMoEs with the original routing are represented by ●, while ▶ shows SVMoEs.

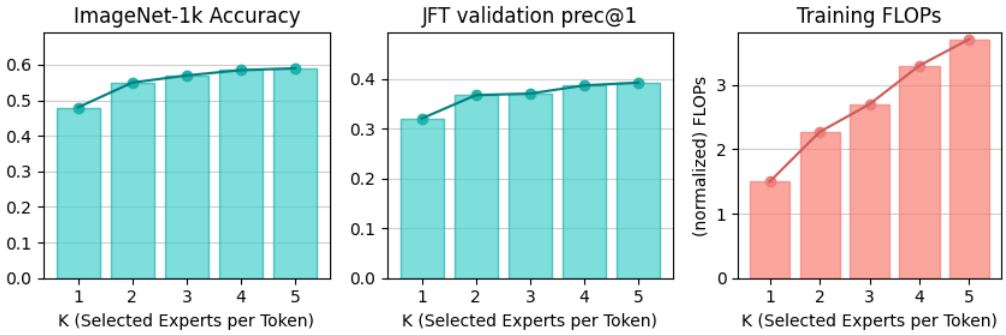

Figure 3: Training FLOPs versus k (experts per token) using VMoE-S/32 (by replacing the TopK router with the regularized OT. Our model's cost depends on the number of selections per token. We used k=2 in our experiments, as choosing $k > 2$ doesn't significantly improve performance but notably increases computation cost.

B COMPARING THE SPEED OF ROUTERS

The architecture of our model is detailed in Figure 1. Additionally, Figures 2 and 3 illustrate the comparative impact of using the optimal transport router versus the TOPk router in VMoE models, with a specific focus on the ImageNet and JFT datasets. OT router uses a cost-efficient strategy to assign tokens to experts, aiming to minimize the transport cost in terms of computational resources. TOPk routing mechanism selects the top k experts for each token based on predefined criteria such as similarity or capacity. Our findings indicate that using OT leads to a consistent and slight improvement in the efficiency of all VMoE architectures during both the training and inference process, which is characterized by reduced FLOPs. As demonstrated in Figure 3, the computational cost of our model is influenced by the number of experts selected per token. In our experiments, setting $k = 2$ provided an optimal balance between performance and computational cost. It's noteworthy that selecting $k > 2$ does not yield a substantial enhancement in performance, yet it leads to a considerable increase in computational demand.