Patch-level Routing in Mixture-of-Experts is Provably Sample-efficient for Convolutional Neural Networks

Mohammed Nowaz Rabbani Chowdhury [1], Shuai Zhang [1], Meng Wang [1], Sijia Liu[2,3], Pin-Yu Chen[3]

[1]Rensselaer Polytechnic Institute

[2]Michigan State University

[3]IBM Research

International Conference on Machine Learning (ICML 2023)
July, 2023

Background

- Scaling conventional deep models
  - Linear increase of training cost with model parameters
- Mixture-of-Experts (MoE)
  - Only sublinear increase of training cost[1]

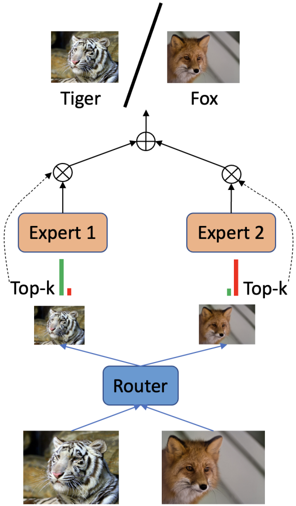
[1] Noam Shazeer et al. (2017). "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer". In: *International Conference on Learning Representations*.

Background

- Routing in MoE
  - Sample-level Routing[a][b]

[a] Prajit Ramachandran and Quoc V Le (2019). "Diversity and depth in per-example routing models". In: *International Conference on Learning Representations*.

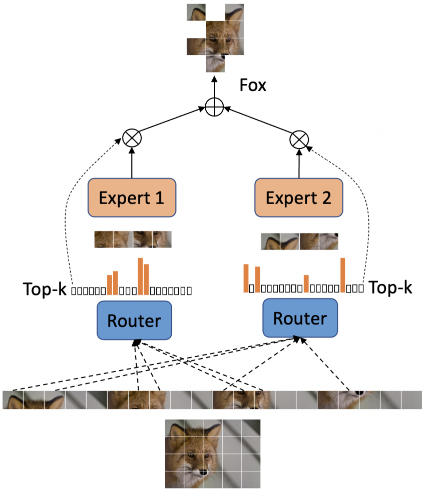
[b] Brandon Yang et al. (2019). "Condconv: Conditionally parameterized convolutions for efficient inference". In: *Advances in Neural Information Processing Systems 32*.



Sample-level MoE

Background

- Routing in MoE
  - Patch-level Routing
    - ★ Patch-wise Routing[a][b]
    - ★ Expert-choice Routing[c]



Patch-level MoE (Expert-choice) (pMoE)

---

[a] William Fedus, Barret Zoph, and Noam Shazeer (2022). "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity". In: *The Journal of Machine Learning Research* 23.1, pp. 5232–5270.

[b] Carlos Riquelme et al. (2021). "Scaling vision with sparse mixture of experts". In: *Advances in Neural Information Processing Systems* 34, pp. 8583–8595.

[c] Yanqi Zhou et al. (2022). "Mixture-of-experts with expert choice routing". In: *Advances in Neural Information Processing Systems* 35, pp. 7103–7114.

Motivation

- Patch-level MoE (pMoE)
  - Significant empirical success, but no theoretical guarantee
- Compared to conventional models:
  - Why does pMoE provide similar generalization with low compute?
  - How much computational resource does pMoE save?

Contributions

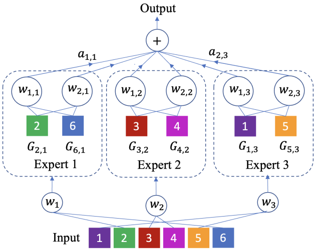
- First convergence and generalization analysis of pMoE for CNN
  - Polynomial reduction of **time**, **sample**, and **model complexity**
- Characterization of the desired property of the pMoE router
- Experimental demonstration of sample efficiency of pMoE in deep CNN models

Setup for Theoretical Analysis

- **Binary supervised classification**
- Given: N i.i.d. training samples $\{(x_i, y_i)\}_{i=1}^N$ generated by a unknown distribution $\mathcal{D}$
- Goal: To learn a NN model that can map x to y ($y \in \{+1, -1\}$) for any $(x, y) \sim \mathcal{D}$

- **The analyzed pMoE model:** Two-layer mixture of CNNs

$$f_M(\theta, x) = \sum_{s=1}^{k} \sum_{r=1}^{m/k} \frac{a_{r,s}}{l} \sum_{j \in J_s(w_s, x)} \mathrm{ReLU}(\langle w_{r,s}, x^{(j)} \rangle) G_{j,s}(x)$$

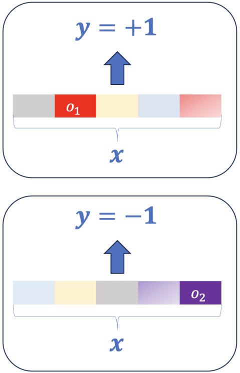


The analyzed model

  - Each input $x \in \mathbb{R}^{nd}$: divided into n disjoint patches, $x^{(j)}$ denotes j-th patch
  - k **experts** and k corresponding **routers**, each selecting l out of n patches ($l < n$)

Setup for Theoretical Analysis

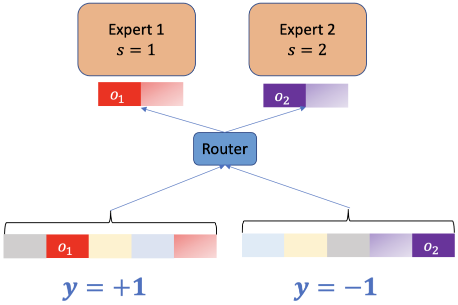
- **Two modes of training:**
  - **Separate training** of the routers and experts
  - **Joint training** of the routers and experts
- **Loss Function:** Binary cross-entropy
- **Training Algorithm:** SGD
- **Data model:** Among the n patches of a sample (x, y)
  - **one** class-discriminative pattern
    - denoted as o_1, if $y = +1$
    - denoted as o_2, if $y = -1$
  - **(n-1)** class-irrelevant patches



Data model

Theoretical Results: Router Property

- Sends similar class-discriminative patches to the same expert
  - o_1 to Expert 1
  - o_2 to Expert 2
- Drop class-irrelevant patches
  - Efficient learning in experts
- Sample complexity: $\Omega(n^2)$ (Separate training)



The proved router property

Theoretical Results: Complexity

To achieve ϵ generalization error	CNN	pMoE		Savings in pMoE	
		Separate training	Joint training	Separate training	Joint training
Sample Complexity	$\Omega(n^8/\epsilon^{16})$	$\Omega(l^8/\epsilon^{16})$	$\Omega(k^4 l^6/\epsilon^{16})$	$\Theta(n^8/l^8)$	$\Theta(n^8/k^4 l^6)$
Iteration Complexity	$O(n^4/\epsilon^8)$	$O(l^4/\epsilon^8)$	$O(k^2 l^2/\epsilon^8)$	$\Theta(n^4/l^4)$	$\Theta(n^4/k^2 l^2)$
Model Complexity	$\Omega(n^{10}/\epsilon^{16})$	$\Omega(l^{10}/\epsilon^{16})$	$\Omega(k^3 n^2 l^6/\epsilon^{16})$	$\Theta(n^{10}/l^{10})$	$\Theta(n^{10}/k^3 n^2 l^6)$
Computational Complexity	$O(Bmn^5 d/\epsilon^8)$	$O(Bml^5 d/\epsilon^8)$	$O(Bmk^2 l^3 d/\epsilon^8)$	$\Theta(n^5/l^5)$	$\Theta(n^5/k^2 l^3)$

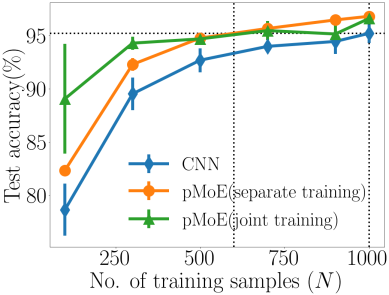
Theoretical Results: Complexity

To achieve ϵ generalization error	CNN	pMoE		Savings in pMoE	
		Separate training	Joint training	Separate training	Joint training
Sample Complexity	$\Omega(n^8/\epsilon^{16})$	$\Omega(l^8/\epsilon^{16})$	$\Omega(k^4 l^6/\epsilon^{16})$	$\Theta(n^8/l^8)$	$\Theta(n^8/k^4 l^6)$
Iteration Complexity	$O(n^4/\epsilon^8)$	$O(l^4/\epsilon^8)$	$O(k^2 l^2/\epsilon^8)$	$\Theta(n^4/l^4)$	$\Theta(n^4/k^2 l^2)$
Model Complexity	$\Omega(n^{10}/\epsilon^{16})$	$\Omega(l^{10}/\epsilon^{16})$	$\Omega(k^3 n^2 l^6/\epsilon^{16})$	$\Theta(n^{10}/l^{10})$	$\Theta(n^{10}/k^3 n^2 l^6)$
Computational Complexity	$O(Bmn^5 d/\epsilon^8)$	$O(Bml^5 d/\epsilon^8)$	$O(Bmk^2 l^3 d/\epsilon^8)$	$\Theta(n^5/l^5)$	$\Theta(n^5/k^2 l^3)$

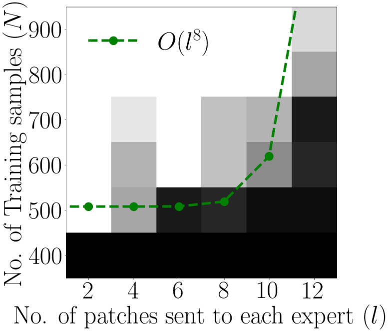
Experimental Results: pMoE of Two-layer CNN

- MNIST characters are used as patterns **(I)**
- pMoE saves almost *half* of the training samples used for CNN **(II)**
- *poly*(l) sample complexity verified **(III)**

(I)



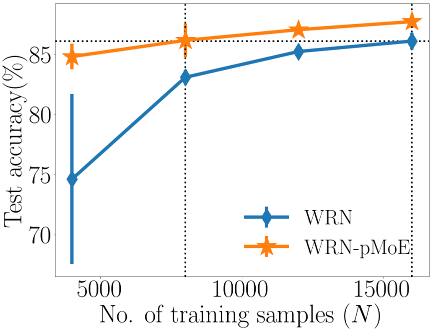
(II)



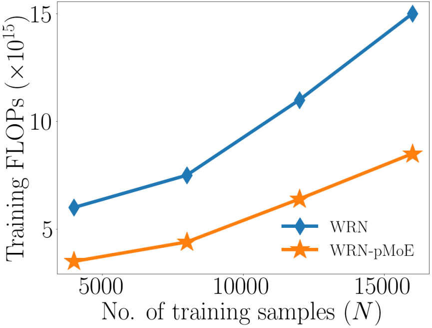
(III)

Experimental Results: pMoE of Wide Residual Networks (WRN)

- 10 layers, Widening factor of 10
- Dataset: CelebA; Multiclass classification
- WRN-pMoE saves
  - 60% of the training samples **(I)**
  - 50% of the training FLOPs **(II)**



(I)



(II)

References

- Fedus, William, Barret Zoph, and Noam Shazeer (2022). "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity". In: *The Journal of Machine Learning Research* 23.1, pp. 5232–5270.
- Ramachandran, Prajit and Quoc V Le (2019). "Diversity and depth in per-example routing models". In: *International Conference on Learning Representations*.
- Riquelme, Carlos et al. (2021). "Scaling vision with sparse mixture of experts". In: *Advances in Neural Information Processing Systems* 34, pp. 8583–8595.
- Shazeer, Noam et al. (2017). "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer". In: *International Conference on Learning Representations*.
- Yang, Brandon et al. (2019). "Condconv: Conditionally parameterized convolutions for efficient inference". In: *Advances in Neural Information Processing Systems* 32.
- Zhou, Yanqi et al. (2022). "Mixture-of-experts with expert choice routing". In: *Advances in Neural Information Processing Systems* 35, pp. 7103–7114.