SCALABLE CONTINUAL LEARNING: ADAPTIVE MOES FOR EXPANDING TASK SETS

Adrian Candocia, Omer M. Inan, Raaghav Agarwal, Aamod Varma, Mark A. Davenport
Department of Electrical & Computer Engineering, Georgia Institute of Technology

ABSTRACT

Recently, the *Mixture-of-Experts* (MoE) model has been shown to be an effective strategy for *continual learning* because it can adapt to a range of tasks by employing an array of "experts" that each specialize on certain tasks. However, the MoE model lacks the ability to adapt to completely new tasks, particularly as the number of tasks grows to be large. In this work we develop a framework for expanding the number of experts as needed when new tasks arise. We also provide simulations demonstrating that our approach can effectively handle a growing number of tasks.

1 INTRODUCTION

Continual learning (CL) aims to enable models to learn sequentially from a stream of tasks while avoiding *catastrophic forgetting*, a phenomenon where previously acquired knowledge degrades as new tasks are introduced. Traditional approaches to CL have sought to mitigate this issue through various strategies. Regularization-based methods like Elastic Weight Consolidation (EWC) (Kirkpatrick et al. (2017)) constrain parameter updates based on prior task importance, while architectural approaches continually expand model capacity for each new task (Rusu et al. (2016)). Although these methods have shown promise, they often struggle with scalability and efficiency when faced with a large and growing number of tasks.

Recent advances in transformer-based architectures and the incorporation of Mixture-of-Experts (MoE) models have opened new possibilities for CL. MoE models leverage sparse activation of specialized experts, allowing them to adapt dynamically to new tasks while preserving previously learned knowledge. A recent theoretical study on MoE in CL (Li et al. (2024)) has suggested that selective routing and expert specialization can effectively mitigate catastrophic forgetting.

One of the most pressing challenges in CL is scaling to a large number of tasks, which is the most realistic scenario for real-world applications. Traditional methods often assume a fixed or a small number of tasks, whereas real-world deployment demands lifelong learning across an expanding set of tasks. Similarly, existing MoE models for continual learning consider a fixed number of experts, which ideally should be large to accommodate many tasks, but practically must often be small due to computational constraints. In this paper we describe a framework for dynamically adding experts to adapt to a growing number of tasks.

2 ADAPTIVE MIXTURES-OF-EXPERTS

We use the MoE architecture in the CL setting (Li et al. (2024)) while focusing on increasing the number of experts as more tasks are introduced to the task stream. At round t we observe task data $\mathcal{D}_t = (X_t, y_t)$ where the data is generated from a pool of N distinct tasks. Each task has s samples. The MoE model with M experts computes its output as

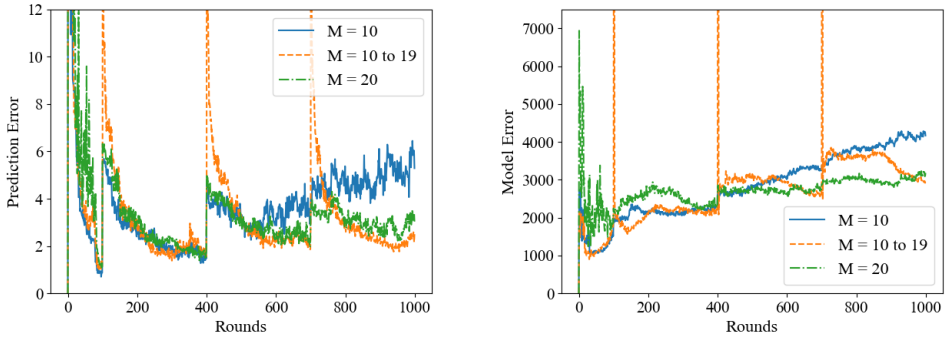
$$\widehat{\boldsymbol{y}}_{m_t}(\boldsymbol{X}_t) = \boldsymbol{X}_t^\intercal \boldsymbol{w}_t^{(m_t)} \quad \text{where} \quad m_t = \underset{m\in\{1,\dots,M\}}{\arg\max}\ \pi_m(\boldsymbol{X}_t, \boldsymbol{\Theta}_t)$$

and

$$\pi_m(\boldsymbol{X}_t, \boldsymbol{\Theta}_t) = \frac{\exp\left(\mathbf{1}_s^\intercal \boldsymbol{X}_t^\intercal \boldsymbol{\theta}_t^{(m)}\right)}{\sum_{m'=1}^{M} \exp\left(\mathbf{1}_s^\intercal \boldsymbol{X}_t^\intercal \boldsymbol{\theta}_t^{(m')}\right)}.$$

for $m = 1, \dots, M$, where $\mathbf{1}_s$ denotes the vector in $\mathbb{R}^s$ of all ones.

We adopt a similar approach to (Li et al. (2024)) in updating the network by, at each round, first updating the selected $\boldsymbol{w}_t^{(m_t)}$ and then updating the "gating weights" $\boldsymbol{\Theta}_t$. However, in our work we



(a) Prediction Error (b) Model Error

Figure 1: Empirical performance of adaptive MoE vs fixed MoE of different sizes. The fixed MoE with $M = 10$ struggles with a large number of tasks. The fixed MoE with $M = 20$ shows improved performance, but is surpassed by the adaptive MoE, which expands the number of experts only as new tasks are introduced.

select different loss functions. Specifically, we update $\boldsymbol{w}_t^{(m_t)}$ to minimize

$$\mathcal{L}_t^{\text{tr}}\left(\boldsymbol{w}_t^{(m_t)}, \mathcal{D}_t\right) = \frac{1}{s}\|\boldsymbol{X}_t^\intercal \boldsymbol{w}_t^{(m_t)} - \boldsymbol{y}_t\|_2^2. \tag{1}$$

The gating parameters $\boldsymbol{\Theta}_t$ are updated by minimizing

$$\mathcal{L}_t^{\text{gate}}\left(\boldsymbol{\Theta}_t, \mathcal{D}_t\right) = \alpha \mathcal{L}_t^{\text{load}}\left(\boldsymbol{\Theta}_t, \mathcal{D}_t\right) + \beta \mathcal{L}_t^{\text{align}}\left(\boldsymbol{\Theta}_t, \mathcal{D}_t\right).$$

with learning rates α and β. $\mathcal{L}_t^{\text{load}}\left(\boldsymbol{\Theta}_t, \mathcal{D}_t\right)$ is the load balancing loss common to MoE literature (Shazeer et al. (2017); Li et al. (2024)) that promotes exploration of experts. The second term is

$$\mathcal{L}_t^{\text{align}}\left(\boldsymbol{\Theta}_t, \mathcal{D}_t\right) = t \cdot \sum_{m=1}^{M} \pi_m(\boldsymbol{X}_t, \boldsymbol{\Theta}_t) \mathcal{L}_t^{\text{tr}}\left(\boldsymbol{w}_t^{(m)}, \mathcal{D}_t\right), \tag{2}$$

whose purpose is to incentivize the gate to place high priority on selecting an expert which minimizes (1). Note that $\mathcal{L}_t^{\text{align}}$ has a factor of t, and $\mathcal{L}_t^{\text{load}}$ has a factor of $1/t$. This accounts for the desire for the gate to explore experts early in training, and exploit given experts late in training.

To adapt and grow the MoE model, we first set $\boldsymbol{\Theta}_t$ to zero while adding a fixed number of new columns (also set to zero). Then, the corresponding experts are generated by sampling from a zero-mean multivariate Gaussian distribution with a variance equal to the average of all of the norms of existing experts. The original experts remain unchanged.

3 Experiments and Discussion

In the empirical evaluation of our adaptive MoE model we use the same task model as in (Li et al. (2024)). We begin with 16 tasks and add 16 additional novel tasks to the task stream at rounds $t = 100, 400$, and 700. We evaluate our performance in terms of prediction error, i.e., the expected error $\|\boldsymbol{y} - \widehat{\boldsymbol{y}}(\boldsymbol{X})\|_2^2$ where $\widehat{\boldsymbol{y}}(\boldsymbol{X})$ is the prediction for a randomly selected task and $\boldsymbol{y}$ is the ground truth. We also consider the model error which measures how accurately the learned $\boldsymbol{w}_t$ approximate the ground truth $\boldsymbol{w}$ that generate the observations across all tasks, i.e., $\mathbb{E}\|\boldsymbol{w}_t - \boldsymbol{w}\|_2^2$.

We consider two fixed MoEs with $M = 10$ and $M = 20$. The adaptive MoE begins with $M = 10$ but adapts by adding 3 experts whenever novel tasks are introduced, eventually growing to $M = 19$. While all methods perform similarly when the number of tasks is relatively small, as the number of tasks grows the MoE with $M = 20$ is superior to the $M = 10$ case. However, both fixed approaches are surpassed by the adaptive MoE which gradually increases in size as the number of tasks grows. This illustrates the benefit of an MoE model that can adaptively grow an environment with a increasing number of tasks.

Future work includes a study of the comparative benefits of model growth versus retraining when new tasks are introduced as well as automated strategies for network growth and consolidation.

REFERENCES

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Hongbo Li, Sen Lin, Lingjie Duan, Yingbin Liang, and Ness B. Shroff. Theory on mixture-of-experts in continual learning, 2024.

Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *CoRR*, abs/1606.04671, 2016.

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. In *International Conference on Learning Representations*, 2017.