Mixture of experts for predicting properties of quantum data

Ernesto Campos[1,*], Andrey Kardashin[1], Konstantin Antipin[1,2]

*ernesto.campos@skoltech.ru

[1]Skolkovo Institute of Science and Technology, Moscow, Russia

[2]Lomonosov Moscow State University, Moscow, Russia

Abstract

Variational quantum algorithms (VQAs) have become the standard for the era of noisy intermediate-scale quantum computers. These algorithms consist of a parameterized quantum circuit that is iteratively optimized by a classical co-processor to minimize a given cost function. VQAs share similarities with artificial neural networks, and as such some techniques have found applicability in both quantum and classical neural networks.

One such strategy is *mixture of experts* (MoE), a form of ensemble learning which has seen success in multiple machine learning tasks [1], and notably has been recently employed in some of the best performing *large language models* (LLMs) [2]. It consists of a set of *expert* networks specialized in different regions of the input space, and a *gating* network that determines the weight of each expert.

For a specific region of the input space, an expert may require a smaller network to achieve the same performance of a larger one trained on the whole input space. Additionally smaller networks result in faster training and inference for individual experts.

Here we describe the use of a MoE for predicting properties of quantum data and test this approach by predicting the *negativity* of quantum states [3]. Our numerical results show that in this setting the use of MoE can perform better than some predictors using a single larger quantum circuit.

Mixture of Experts

Given a training set $\mathcal{T} = \{\rho_j, \alpha_j\}_{j=1}^{T}$, where ρ have a corresponding label $\alpha \in \mathbb{R}$, the task is to use regression to predict α from a state ρ.

The MoE approach to this problem is the following:

1. The training set is divided into classes $C_l = \{\rho_{jl}, \alpha_{jl}\}_j$, such that $\mathcal{T} = \bigcup_{l=1}^{k} C_l$, and the data from each class is used to train its respective expert E_l to predict α from ρ.
2. The gating function G is trained on $\mathcal{T}_{\text{class}} = \{\rho_j, l_j\}_{j=1}^{T}$, where ρ_j belongs to the class with index l_j. $G(\rho)$ returns the weights $\{w_l\}_{l=1}^{k}$ corresponding to the probabilities of ρ belonging to each class.
3. The final prediction is calculated by weighting the experts output as $\tilde{\alpha} = \sum_{l=1}^{k} w_l E_l(\rho)$.

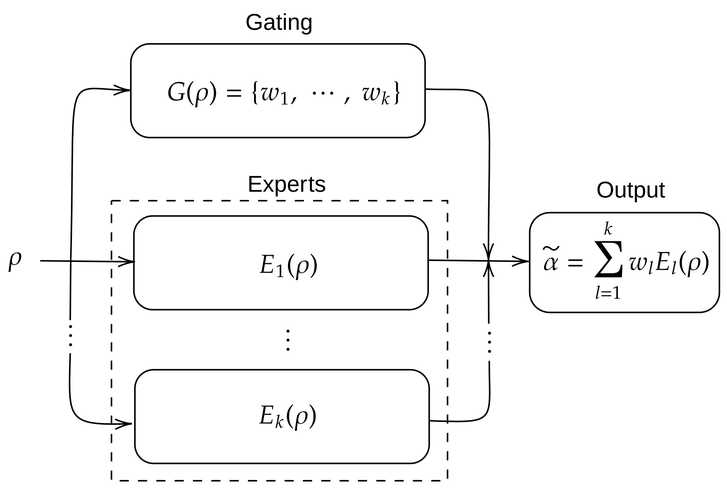

Fig. 1: Schematic of a mixture of experts for quantum data.

Prediction

Experts

Given a state ρ, each expert E_l gives a prediction of α plus a bias $b(\alpha)$, the later being small when $\{\rho, \alpha\} \in C_l$,

$$E_l(\rho) \equiv \mathrm{Tr}(H_l \rho) = \alpha + b(\alpha). \tag{1}$$

We parameterize H_l by $\boldsymbol{x}_l, \boldsymbol{\theta}_l \subset \mathbb{R}$ as

$$H_l(\boldsymbol{x}_l, \boldsymbol{\theta}_l) = \sum_j x_{jl} \Pi_{jl}(\boldsymbol{\theta_l}), \tag{2}$$

where $\boldsymbol{x}_l = \{x_{jl}\}_j$ are the eigenvalues, and $\Pi_{jl}(\boldsymbol{\theta}_l) = U_l(\boldsymbol{\theta}_l) |j\rangle\langle j| U_l^{\dagger}(\boldsymbol{\theta}_l)$ are projectors onto the jth computational basis state transformed by a variational quantum circuit $U_l(\boldsymbol{\theta}_l)$.

Gating

The gating function G calculates the probabilities $\{w_l\}_{l=1}^{k}$ as:

$$G(\rho) = \left\{ w_l \;\middle|\; w_l = \frac{p_l[g(\rho)]}{\sum_{l=1}^{k} p_l[g(\rho)]} \right\}_{l=1}^{k} \tag{3}$$

where $g(\rho) = \mathrm{Tr}(H_{\mathrm{g}}\rho) = l + b(l)$ gives a prediction of the class l where ρ belongs, and H_{g} is parameterized as $H_{\mathrm{g}}(\boldsymbol{x}_{\mathrm{g}}, \boldsymbol{\theta}_{\mathrm{g}})$ similar to (2). Each p_l is a function that fits the probability distributions h_l followed by $\{g(\rho_{jl})\}_{j=1}^{|C_l|}$ for states in class C_l, as in the example illustrated in Figure 2.

Optimization

To find optimal parameters $\{\boldsymbol{x}_l^*, \boldsymbol{\theta}_l^*\}_{l=1}^{k}$ for experts $\{E_l\}_{l=1}^{k}$, and $\boldsymbol{x}_{\mathrm{g}}^*, \boldsymbol{\theta}_{\mathrm{g}}^*$ for g, we solve the following minimization problem over the datasets $\{C_l\}_{l=1}^{k}$ and $\mathcal{T}_{\text{class}}$ respectively:

$$\boldsymbol{x}^*, \boldsymbol{\theta}^* = \arg\min_{\boldsymbol{x},\boldsymbol{\theta}} \Big(w_{\mathrm{ls}} F_{\mathrm{ls}}(\boldsymbol{x}, \boldsymbol{\theta}) + w_{\mathrm{var}} F_{\mathrm{var}}(\boldsymbol{x}, \boldsymbol{\theta}) \Big), \tag{4}$$

where

$$F_{\mathrm{ls}}(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{j=1}^{|D|} \Big(\gamma_j - f\big(\rho_j, \boldsymbol{x}, \boldsymbol{\theta}\big) \Big)^2, \quad F_{\mathrm{var}}(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{j=1}^{|D|} \Delta_{\rho_j}^2 H(\boldsymbol{x}, \boldsymbol{\theta}),$$

with $w_{\mathrm{ls}}, w_{\mathrm{var}} > 0$ being weights, $\Delta_\rho^2 H \equiv \langle H^2 \rangle_\rho - \langle H \rangle_\rho^2$, and $D = \{\rho_j, \gamma_j\}_j$, f and H being the appropriate dataset, function and observable respectively.

Application: predicting entanglement

As a measure of entanglement, we chose the negativity $N(\rho_{AB}) = \left\| \rho_{AB}^{T_B} \right\|_1 - 1$. We train on a data set $\mathcal{T} = \left\{\rho_j, N_j\right\}_{j=1}^{1000}$ with random mixed two-qubit states ρ_j and their negativities N_j. We divide $\mathcal{T}$ into 5 classes of evenly spaced ranges of N on which we train each expert. We allow each expert E_l and g to process $c = 2$ copies of the labeled states, and use two layers of the hardware efficient ansatz (HEA). Figure 2 illustrates the distributions $\{h_l\}_{l=1}^{k}$ obtained by evaluating g on the states of each class, and function fits $\{p_l\}_{l=1}^{k}$ used to calculate G. Figure 3 illustrates the predictions $\tilde{N}_{\mathrm{MoE}}$ and $\tilde{N}_{\mathrm{s}}$ obtained by using MoE and a single predictor respectively.

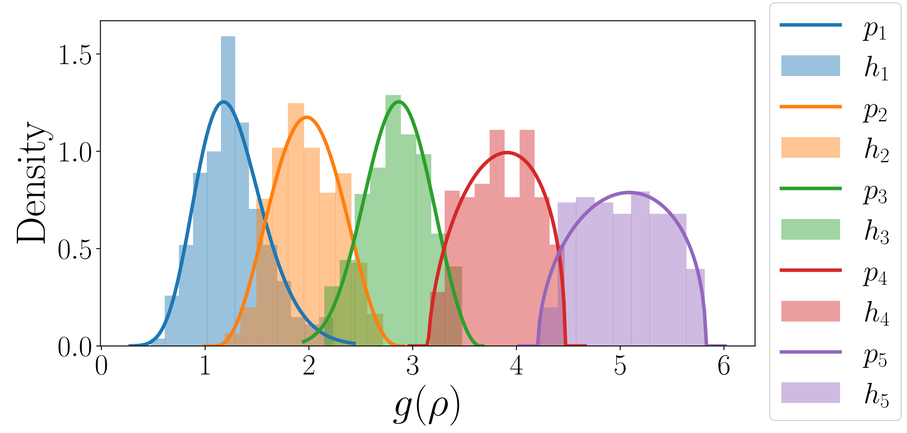

Fig. 2: Probability distributions from the training set and fits used in the gating function.

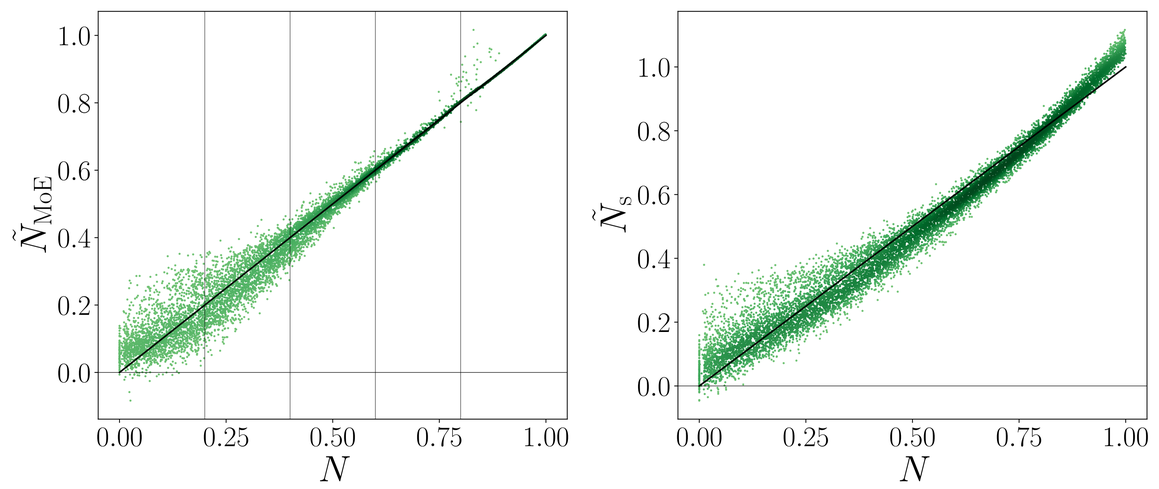

Fig. 3: Predicted versus exact negativity of 10^4 random mixed states. Left: Predictions using a mixture of 5 experts resulting in a mean absolute error of 0.02. Each expert and gating use $c = 2$ copies of the input state and 2 layers of HEA. Right: Predictions of a single predictor resulting in a mean absolute error of 0.04. The predictor uses $c = 3$ copies of the input state and 3 layers of HEA.

References

[1] S. Masoudnia and R. Ebrahimpour. Mixture of experts: a literature survey. Artificial Intelligence Review, 42:275293, 2014.

[2] A. Jiang, et al. Mixtral of experts. arXiv preprint arXiv:2401.04088, 2024.

[3] A. Kardashin, et al. Predicting properties of quantum systems by regression on a quantum computer. arXiv preprint arXiv:2407.08847, 2024.