Quantum Enhanced Knowledge Distillation

Simone Piperno[1], Leonardo Lavagna[1], Francesca De Falco[1], Andrea Ceschini[1], Antonello Rosato[1], David Windridge[2,3], and Massimo Panella[1]

[1]Department of Information Engineering, Electronics and Telecommunications,
University of Rome "La Sapienza", 00184 Rome, Italy,
Email: {simone.piperno, francesca.defalco,leonardo.lavagna, andrea.ceschini, antonello.rosato, massimo.panella}@uniroma1.it
[2]Department of Computer Science, Middlesex University, London, NW4 4BT, UK , Email: d.windridge@mdx.ac.uk
[3]Centre for Vision Speech and Signal Processing, University of Surrey, Surrey, GU2 7XH

1 Introduction

Knowledge distillation (KD) is a process of training a "student" machine learning system using the outputs of a pre-trained "teacher" model and it is a well-established practice in the optimization of classical Deep Neural Networks (DNNs) [1]. Usually, it is enacted via the substitution of the output softmax layer of the trained DNN teacher network with an equivalent layer of Boltzmann-temperature parameterized sigmoid functions, leveraging gradient information implicit in the softened logits for the training of the smaller student network. The smaller network is thus trained to replicate the output sigmoid layer of the larger network during training [2].

However, KD remains relatively unexplored within the quantum machine learning domain, with only a few pioneering studies [3], [4]. Some challenges in the quantum domain include the incongruence of the respective learning architectures and the transferability of gradient information in inter-domain approaches (e.g., classical-to-quantum distillation) or intra-domain transfer (e.g., quantum-to-quantum distillation). Additionally, the scarcity of quantum-to-quantum distillation research could be due to the current absence of sufficiently large and efficient quantum network architectures necessitating a distillation step a priori.

In this work, we focus on the classical-to-quantum paradigm and investigate the extent to which a hybrid quantum-classical architecture can effectively learn from the softmax outputs of a classical Multi-Layer Perceptron (MLP) in multi-class classification tasks. The multi-class scenario is chosen both for its representativeness of typical DNN usage and its inherently greater potential for meaningful gradient information transfer. In doing so, we demonstrate substantial empirical efficiency gains for classical-to-quantum KD in relation to an emblematic non-linearly separable 3-class problem. Our findings reveal that classical-to-quantum KD enhances the performances of standard hybrid quantum architectures and paves the way for the applicability of distillation techniques in the quantum realm.

2 Proposed Methodology

The implemented KD process involves a number of steps. For the specific case study illustrated in the following, the teacher MLP is trained by optimizing its parameters using standard techniques. Subsequently, the teacher outputs are utilized to distill the acquired knowledge into the smaller model, which can be either a hybrid quantum-classical student or a classical student. Distillation occurs through the use of a mixture of losses $\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{KL}}$, where $\mathcal{L}_{\text{CE}}$ represents the cross-entropy loss given by $\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{i,j} \log \hat{y}_{i,j}^{(s)}$, and $\mathcal{L}_{\text{KL}}$ denotes the KL-divergence loss expressed as $\mathcal{L}_{\text{KL}} = \sum_{i=1}^{N} \hat{y}_i^{(t)} \log \frac{\hat{y}_i^{(t)}}{\hat{y}_i^{(s)}}$. Here, $\hat{y}^{(t)}$ and $\hat{y}^{(s)}$ represent the outputs of the teacher and student models, respectively; N is the number of datapoints considered; C is the number of classes; $\hat{y}_{i,j}^{(t)}$ is the predicted probability for the ith datapoint and the jth class.

We have chosen a teacher architecture given by an MLP with $k^{(t)} = 1203$ parameters, consisting of an input layer with $n_i^{(t)} = 64$ neurons, two hidden layers with $h_1 = 16$ and $h_2 = 8$ neurons, respectively, and an output layer with $n_o^{(t)} = 3$ neurons. We use ReLU activation functions between the hidden layers, and a final softmax transformation to recover class probabilities, as is standard practice. The classical student architecture consists of an MLP with $k^{(s)} = 195$ parameters, a single layer having $n_i^{(s)} = 64$ input neurons and $n_o^{(s)} = n_o^{(t)}$ output neurons. Additionally, for the student architecture, the output layer is associated with a softmax transformation. We remark that the model described here is the smallest possible with respect to the number of parameters.

We defined multiple quantum students based on the same overall design principle, by leveraging a parameterized quantum circuit for each student to efficiently read the input features using amplitude embedding, enabling the encoding of $n_i^{(t)} = 64$ features into $\ln n_i^{(t)} = 6$ qubits. Following the principles of variational quantum algorithms, which allow for flexible choices of the ansatz, we employed a hardware-efficient ansatz (HEA) [5] and a universal circuit (Universal) [6] able to reach every unitary in $\mathfrak{su}(2^{\ln n_i^{(t)}})$. We also considered two other variants: a universal circuit with amplitude embedding followed by an Hadamard gate for each qubit (Universal + H); a universal circuit in which at each layer one of the qubits is measured (q_{n_d}), allowing for information compression from $2^{n_i^{(t)}}$ states to 2^{n_d} states, with $n_d \in \{2, ..., 5\}$ being the number of desired qubits in output.

In this setup, a measurement process is conducted on each qubit. Subsequently, the measurement results are fed into an MLP, with input neurons corresponding to the number of measured qubits and output neurons corresponding to the number of classes in the multi-class classification problem. This processing step ensures robust outputs in the form of class probabilities, achieved through the application of a softmax transformation. These hybrid structures, combined with amplitude encoding, exhibit promising characteristics such as smaller feature representations (from 64 features to 6), which allow for a notable reduction in parameter counts. i.e., from 195 of the classical MLP up to 74 when using the smallest of the proposed architectures q_2.

3 Experimental setup

To assess the effectiveness of classical-to-quantum KD, the introduced models were tested on an extended multidimensional XOR dataset with $N = 1000$ binary vectors of $n_i^{(t)} = 64$ bits and $n_o^{(t)} = C = 3$ classes, "zero", "one" or "two", depending on the value of the first two significant bits in each vector. This dataset was chosen in order to investigate the capabilities of the models in relation to the linear separability of the associated classification problem. We remark that, in contrast to 2D case, the generalized multidimensional XOR dataset is unbalanced: class zero has probability 1/2 of being sampled, whereas class one and two have both probability 1/4 of being sampled. For this reason, we chose to employ the F_1 score as a performance metric, given by $F_1 = \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}}$, where TP is the number of true positives, FN is the number of false negatives, and FP is the number of false positives.

The models were implemented in Python 3.10.12 with PyTorch 2.3.0+cu121, with the aid of Pennylane 0.36 for the implementation of the quantum models. The experiments were run on Google Colab's CPU. To ensure a good and fast convergence, the models were trained with the Adam optimization algorithm with learning rate set to 0.1, $n_e = 300$ epochs of training, early stopping set to 50 epochs, and 5 stacked layers of the chosen ansatz in the quantum student models.

The experiments were performed as follows. We first generated the generalized multidimensional XOR dataset and divided it into 80% train samples, 10% validation samples and 10% test samples. Then, we trained each of the architectures using $s = 12$ different initial seeds, in order to verify the models' consistency with respect to initialization of the parameters. Applying the trained models to the test set resulted in multiple F_1 test scores for each architecture; we then took the average of these scores to compute the mean F_1 test score and compare the architectures.

4 Results

We report our preliminary experimental results in Fig. 1. KD allows for a significant increase in the performance of the tested models, both in the classical-to-classical and classical-to-quantum approaches. The improvement is more significant for the classical-to-classical case, this may be due to the number of parameters involved (we utilize approximately 2× the number of parameters for the classical student compared to the hybrid counterparts). The main finding for consideration is that distilling knowledge in a hybrid architecture still results in an increased performance, even though these architectures are more susceptible to parameter initialization, as evidenced by the related variances.

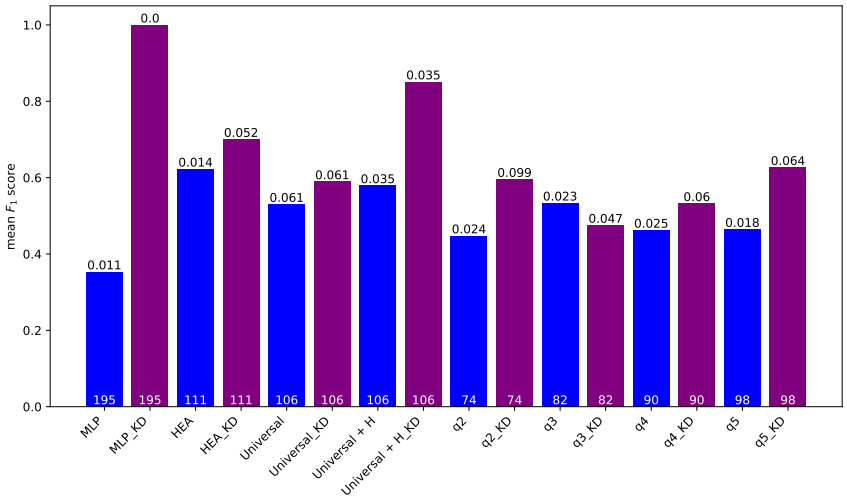

Figure 1: Experiments results. Individual bins report the mean F_1 score for each of the tested architectures: the classical MLP student with or without KD, and the quantum students with or without KD. On the top of the bin is reported the variance of the F_1 scores obtained while on the base of the bin is reported the number of trainable parameters for the architecture.

Current limitations of Noisy Intermediate-Scale Quantum (NISQ) devices in terms of hardware constraints and simulation capabilities prevent us from fully exploiting purely quantum architectures. Consequently, the most effective strategy at present time is to integrate variational circuits with classical architectures, where KD can play a crucial role. Our findings point precisely to the effectiveness of a teacher-student approach also in hybrid settings, expanding the applicability of distillation techniques to the quantum machine learning domain.

References

[1] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[2] Z. Allen-Zhu and Y. Li, "Towards understanding ensemble, knowledge distillation and self-distillation in deep learning," *arXiv preprint arXiv:2012.09816*, 2020.

[3] M. J. Hasan and M. Mahdy, "Bridging classical and quantum machine learning: Knowledge transfer from classical to quantum neural networks using knowledge distillation," *arXiv preprint arXiv:2311.13810*, 2023.

[4] M. Alam, S. Kundu, and S. Ghosh, "Knowledge distillation in quantum neural network using approximate synthesis," in *Proceedings of the 28th Asia and South Pacific Design Automation Conference*, 2023, pp. 639–644.

[5] L. Leone, S. F. E. Oliviero, L. Cincio, and M. Cerezo, *On the practical usefulness of the hardware efficient ansatz*, 2022.

[6] M. Larocca, P. Czarnik, K. Sharma, G. Muraleedharan, P. J. Coles, and M. Cerezo, "Diagnosing Barren Plateaus with Tools from Quantum Optimal Control," *Quantum*, vol. 6, p. 824, Sep. 2022.