NESYA
LABORATORY

A generalized framework for Quantum Knowledge Distillation

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

Department of Information Engineering, Electronics and Telecommunications, University of Rome “La Sapienza”, Rome, Italy,
Department of Computer Science, Middlesex University, London, NW4 4BT, UK

Knowledge Distillation

Classical Knowledge Distillation (KD)

- Model compression technique used to transfer knowledge from **‘teacher’** to **‘student’** architectures;
- Obtain a smaller model with reduced computational complexity that maintains similar performances.

How does it work?

- The idea is to teach the student to reproduce the probability class distribution of the teacher by adding a term in the loss:

$$\text{Loss}_{\text{KD}} = \mathcal{L}_{\text{CE}}(\underline{y}_{true}, P_s) + \mathcal{L}_{\text{KL}}(P_t, P_s)$$

Why should we make it quantum?

- **NISQ Technology** doesn’t allow for deep circuits: an application where is required a small architecture could be suitable to make quantum machine learning relevant;

- **Amplitude encoding:** Allows to encode many features in a small number of qubits thus making the architecture small;

- **Hybrid architectures tend to have far fewer parameters** than classical neural networks and **achieve comparable results** for some specific tasks.

Quantum KD - state of the art

We started working on this topic when there was nothing in literature, now…

1. Bridging Classical and Quantum Machine Learning (M. Hasan and M.R.C. Mahdy)
2. Hybrid Quantum Classical Machine Learning with Knowledge Distillation (M. Li, L. Fan, A. Cummings, X. Zhang, M. Pan, and Z. Han)

Proposal: show multiple classification KD settings and prove how and why they are working. These distillation settings are:

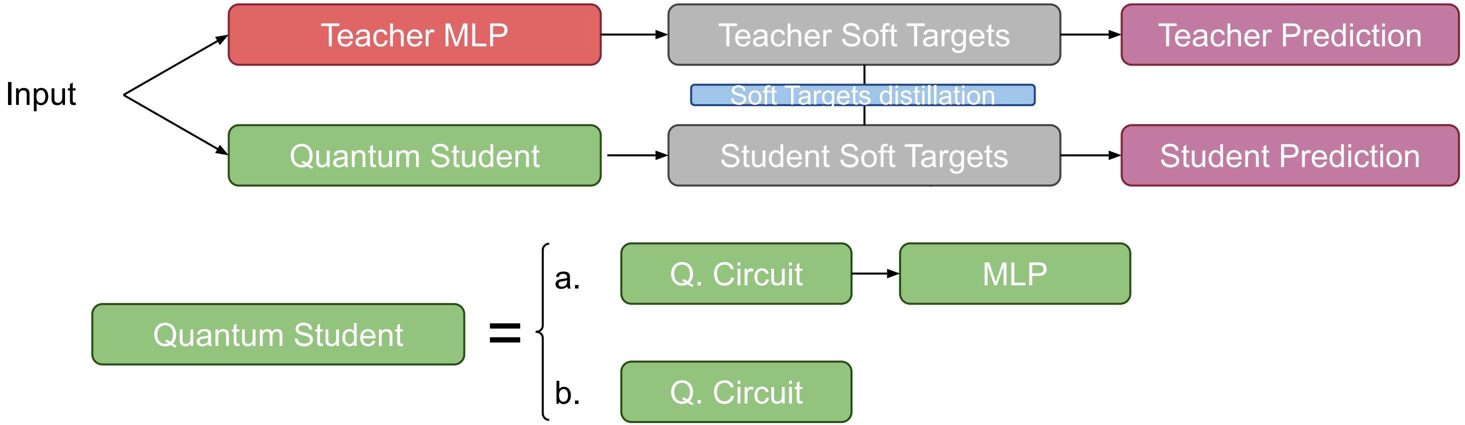
- **Classical to Quantum KD (C2Q KD):**
  a. Distillation from a classical teacher to a hybrid student model
  b. Distillation from a classical teacher to a fully quantum student model
- **Quantum to Classical KD (Q2C KD):**
  a. Distillation from a hybrid architecture to a classical one

Classical to Quantum KD - C2Q KD

Classical to Quantum KD (CQKD): we introduce two different frameworks:

a. Distillation from a classical teacher to a hybrid student model
   - Reduce #parameters compared to classical student

b. Distillation from a classical teacher to a fully quantum student model
   - Removing the MLP allows to further reduce the #parameters



C2Q KD - How is the distillation performed?

Given a classification problem in which a single datapoint $\underline{x}_i$ represents y_i, one of the n_c possible classes, if we want to perform KD we will need two discrete probability distributions:

$$\boxed{\underline{\hat{y}}_i^{\,t} = \mathrm{Softmax}(t(\underline{x}_i))} \quad \text{and} \quad \boxed{\underline{\hat{y}}_i^{\,s} = \mathrm{Softmax}(s(\underline{x}_i))}$$

pretrained classical teacher prediction

student model prediction

We are then able to define the loss function:

$$\mathrm{Loss}_{\mathrm{KD}} = \mathcal{L}_{\mathrm{CE}}(\underline{y}_i, \underline{\hat{y}}_i^{\,s}) + \mathcal{L}_{\mathrm{KL}}(\underline{\hat{y}}_i^{\,t}, \underline{\hat{y}}_i^{\,s})$$

Distillation from a classical teacher to a hybrid student model: MLP makes straightforward applying the method described above;

Distillation from a classical teacher to a fully quantum student model: Usually the states' probability distribution is such that $2^{n_q} >> n_c$ thus making it hard to compute the KL-Divergence.

We can sum some of the state probabilities together to perform distillation which:

- yields a probability distribution;
- forces the states to represent a specific class, once trained with KL-Divergence.

C2Q KD - How is the distillation performed?

Dataset: Generated, non linearly separable dataset with 3 classes, 64 features per sample

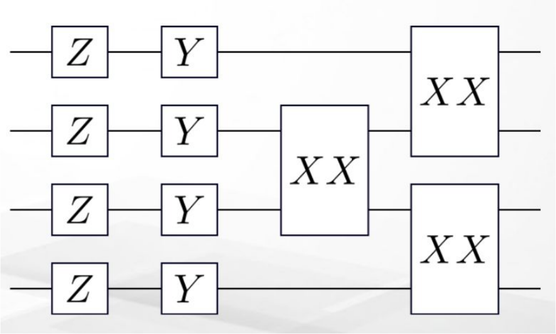
Teacher architecture: MLP

Student Architectures:

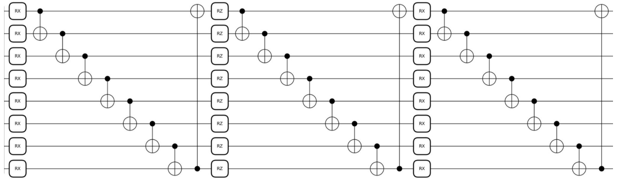
Classical: MLP(input=num_feature,output=num_classes)

a. **Hybrid student model**
Amplitude encoding with variable ansatz and final MLP for prediction

b. **Fully quantum student model**
Amplitude encoding with Universal ansatz

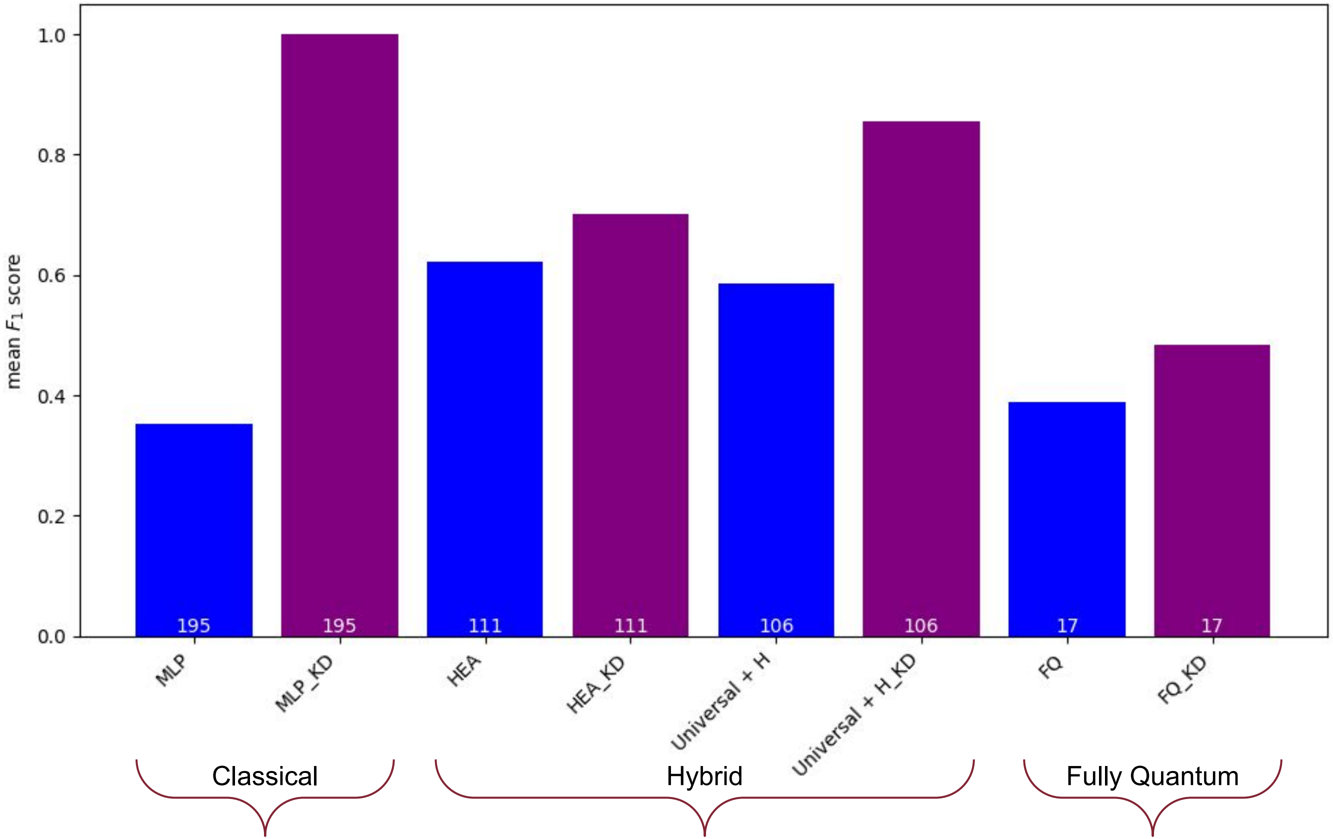


Universal ansatz



Hardware Efficient ansatz

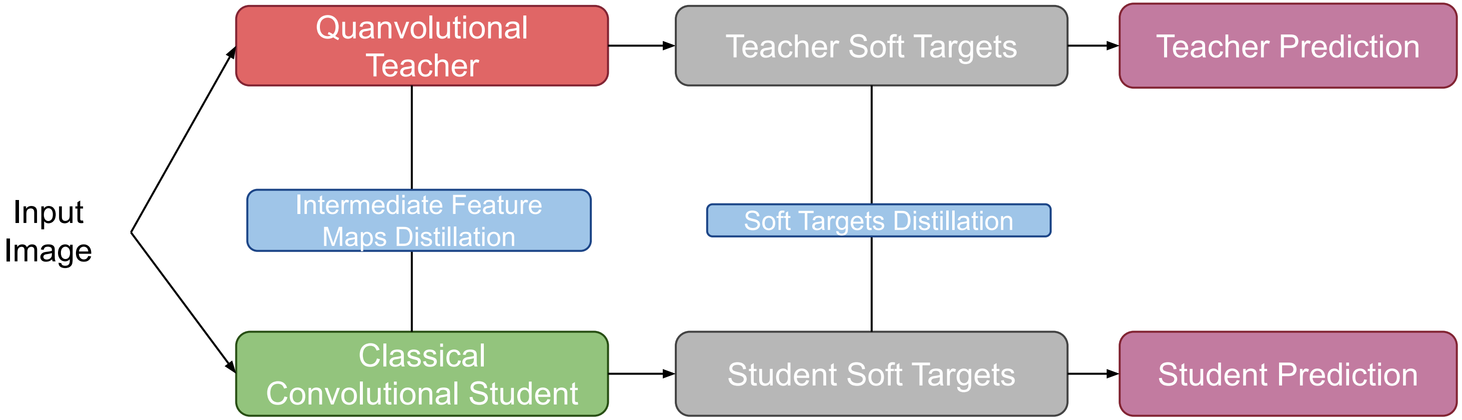
C2Q KD - Results



Quantum to Classical KD - Q2C KD

In the Q2C framework the idea is to extend distillation to intermediate layers:

- Employing a image-like dataset it's possible to distill feature maps obtained by using quanvolutional layers [3].



[3] M. Henderson and S. Shakya, "*Quanvolutional Neural Networks: Powering Image Recognition with Quantum Circuits,*" Semantic Scholar.

Q2C KD - How is the distillation performed?

Given a classification problem over an image dataset, in which a single datapoint $\underline{x}_i$ represents y_i, one of the possible classes, if we want to perform KD we will need :

- two discrete probability distributions:

$$\boxed{\underline{\hat{y}}_i{}^t = \mathrm{Softmax}(t(\underline{x}_i))} \quad \text{and} \quad \boxed{\underline{\hat{y}}_i{}^s = \mathrm{Softmax}(s(\underline{x}_i))}$$

pretrained classical teacher prediction

student model prediction

- two feature maps, which are the output of the quanvolutional layer of the teacher, and the output of the classical convolutional layer of the student, we will have respectively:

$$\boxed{\mathbf{f}^t = \tilde{t}(\underline{x}_i)} \quad \text{and} \quad \boxed{\mathbf{f}^s = \tilde{s}(\underline{x}_i)}$$

pretrained hybrid teacher feature map

student model feature map

We are then able to define the loss function as:

$$\mathrm{Loss}_{\mathrm{KD}} = \mathcal{L}_{\mathrm{CE}}(\underline{y}_i, \underline{\hat{y}}_i{}^s) + \mathcal{L}_{\mathrm{KL}}^{T_B}(\underline{\hat{y}}_i{}^t, \underline{\hat{y}}_i{}^s) + \mathcal{L}_{\mathrm{MSE}}(\mathbf{f}^t, \mathbf{f}^s)$$

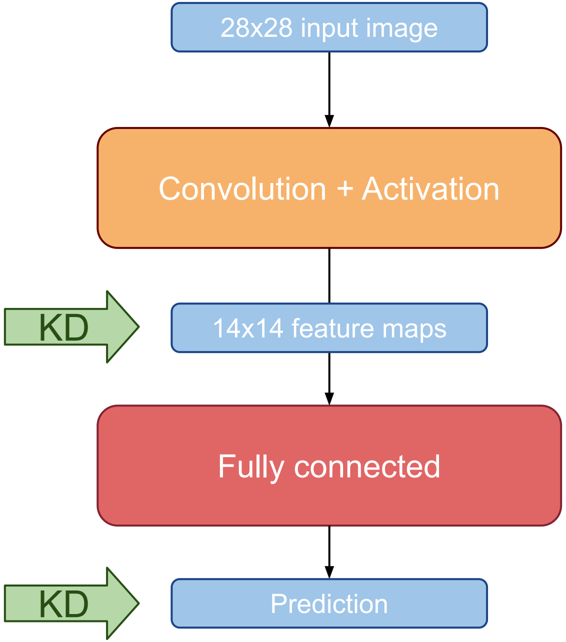
Q2C KD - How is the distillation performed?

Dataset: FASHION-MNIST, randomly sampled K datapoints

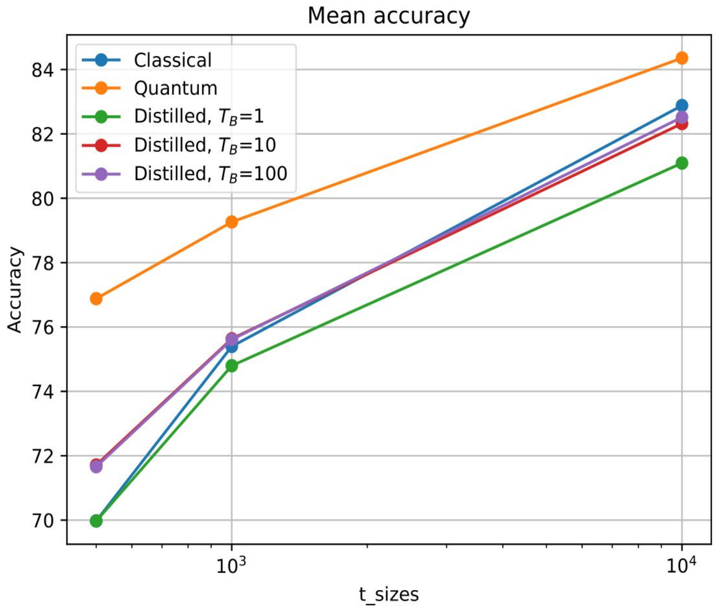
Here, to study the effectiveness of distilling feature maps we used the same architecture for teacher and student, in which we change only the type of convolution applied (quanvolutional for teacher, classical convolution for student).

Architecture details:

1. **Convolution:** 3 filters, 2x2 kernel size, stride =2, padding=0;
2. **Fully Connected:** MLP(3x14x14, 10);
3. **# Parameters:** The quantum architecture has 5923 parameters, while the classical one has 5905.



Q2C KD - Results



- Improved performances w.r.t. classical architecture when we have a small number of samples, could be useful in some specific applications

Summary

- Quantum KD looks like a promising application for QML;
- More tests have to be performed;
- This could be an interesting topic to further investigate and our code will be made available for the community to more easily approach it.

Thanks for your attention!

Simone Piperno

simone.piperno@uniroma1.it