Analysis of Loss of Plasticity in Quantum Continual Learning

Haruto Tanaka
Department of Computing Science
University of Alberta
Edmonton, Canada
haruto@ualberta.ca

***Abstract*—Continual Learning is a type of machine learning (ML) problems that challenge intelligent systems to learn incrementally from the stream of non-i.i.d. data. Within the continual learning settings, the conventional (classical) ML systems struggle with two major challenges, *catastrophic forgetting* and *loss of plasticity*. Both issues were extensively investigated with classical ML settings and remain active research fields. In contrast, these phenomena are rarely investigated with quantum ML (QML) systems. Especially, to the best of our knowledge, no explicit work on the loss of plasticity with QML has been done. In this report, we attempt to fill out this gap by providing an empirical analysis of whether the loss of plasticity occurs in QML settings. Additionally, we reveal the potential similarities and differences with the one that occurs with classical ANNs.**

***Index Terms*—Continual Learning, Machine Learning, Quantum Computing, Loss of Plasticity**

I. Introduction

The basic concept of *continual learning* is a relaxation of conventional ML settings, where the data distribution no longer needs to be stationary nor independent and identically distributed (i.i.d.). Relaxing this, perhaps the most fundamental assumption of traditional ML settings, raises the demand for the algorithms which allow ML models to incrementally learn from non-stationary data streams in a lifelong manner. Such algorithms would free conventional ML systems from the agony of periodically retraining after the new data is sufficiently accumulated.

While realizing such algorithms could greatly expand the practicality of modern ML systems, the direct application of conventional ML algorithms in continual learning settings often causes some critical issues. Two representative issues are the *catastrophic forgetting* and the *loss of plasticity*. Former refers to the phenomenon where the ML model fails to retain the previously learned knowledge after learning from the newly incoming data streams. The latter refers to the phenomenon where the ML model loses its ability to fit the new incoming data. Both issues are extensively investigated and remain active research fields in classical ML. However, these issues are relatively unexplored with QML settings. Especially for the loss of plasticity, to the best of our knowledge, none of the past literature specifically discussed whether it emerges in the context of QML. Therefore, the ultimate purpose of our work is to partially fill this gap by providing a small empirical analysis of the loss of plasticity in quantum continual learning settings. In particular, we address the following two questions through our experiments:

1) ***Does the loss of plasticity occur with the VQC?***
2) ***Are there any similarities/differences from the one with classical ANNs?***

This report consists of the following sections. In section II, we briefly summarize the previous investigations on the loss of plasticity in conventional ML systems and the catastrophic forgetting in quantum continual learning. Section III covers the necessary foundations to understand the elements used in Section IV. Section IV also covers all the necessary components for the experiments. The analysis of the outcomes of the experiments is conducted and summarized in Section V. Lastly, we wrap up the report with the conclusion and potential future works in Section VI.

II. Related Works

The loss of plasticity in ML emerged as a major academic interest relatively recently and remains one of the most active research fields in continual learning, with classical ANNs. The majority of research focuses on either establishing the potential root causes of the loss of plasticity and/or developing a method to mitigate it. The root causes of the loss of plasticity have been extensively investigated and multiple possibilities are proposed, such as neuron dormancy [Sokar et al., 2023], increase in the parameter norms [Nikishin et al., 2022], decrease/vanish of the update [Abbas et al., 2023], and so on. However, these explanations are often not globally applicable and only partially explain the phenomenon [Lewandowski et al., 2024]. Hence, the root cause or the underlying mechanism is still an open question.

Another track of the research focuses on the development of algorithms to mitigate the loss of plasticity. The regularization-based method, such as the L^2-regularization [Goodfellow et al., 2016], is one of the most popular methods. Kumar et al., 2023 introduced the regenerative regularization, which

The code for this project is available at: https://github.com/WINUprj/lop-qcl .

Also, we would like to give a special mention to the work by Lan, 2021, which we have been heavily influenced in terms of the contents/structure of the report and the way source code is organized.

regularizes towards the initial parameters. The intuition behind this is to retain the ability to learn by remembering the randomly initialized state as much as possible, assuming that the network plasticity is maximal when the parameters are randomly initialized. Similar ideas can be observed in the work by Lewandowski et al., 2024, which utilizes the Wasserstein regularization. Another idea of retaining the property of randomly initialized parameters is to perturb the parameters. Ash and Adams, 2020 came up with the technique called "shrink and perturb", which constantly decreases the norm and adds random noise to the parameters. As an extension, Elsayed and Mahmood, 2024 developed the Utility-based Perturbed Gradient Descent (UPGD) algorithm, which controls the amount of perturbation with the accumulated utility measure of each parameter. Overall, the loss of plasticity is explored and investigated extensively with the classical ML systems.

While the loss of plasticity remains unexplored in the context of QML, catastrophic forgetting with QML systems has been investigated by some literature. Jiang et al., 2022 revealed that the catastrophic forgetting occurs with the variational quantum classifier. The work also provided the approach to mitigate the forgetting by adopting the elastic weight consolidation (EWC) method, the regularization technique where the regularization term is proportional to the distance between the current and previous solution weights [Kirkpatrick et al., 2016]. Additionally, Situ et al., 2023 has shown that catastrophic forgetting can be mitigated by incorporating the gradient episodic memory (GEM) Lopez-Paz and Ranzato, 2017.

III. Foundations

In this section, we briefly formalize the problem settings for continual learning and cover some mathematical backgrounds on the QML techniques, such as VQC.

A. Continual Learning

[1]Generally, the continual learning problems consist of the (potentially unending) stream of data. Each data is given as the tuple of random vectors from the abstract domains $\mathcal{X} \times \mathcal{Y}$

$$S = (X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \dots$$

At the arbitrary timestep t, random vectors X_t and Y_t are fed into the model as the input and corresponding targets, respectively. The random vector Y_t is sampled according to the hidden, unknown distribution $p(Y|X_t, t)$. The goal of the continual learning algorithms is to produce the parameterized function f_θ which approximates this hidden probability distribution p through the loss minimization. This formulation roughly aligns with that of the conventional supervised learning problems. The only and most significant difference from the conventional supervised learning settings is that the target distribution p is explicitly conditioned on timestep t. This implies the possibility of the target distribution changing depending on the timestep and therefore the loss landscape [Lyle et al., 2023]. Thus, considering the chronological effects on the data distribution and loss landscape is one of the necessary conditions for the ideal continual learning algorithms to satisfy.

As a side note, it is worth mentioning that this formulation is precisely for the *continual supervised learning* settings and indeed does not consider the other possible types of continual learning settings. For instance, reinforcement learning [Abbas et al., 2023], and many others have their own unique formulation which does not necessarily agree with the above. In this sense, our formulation fails to provide a unified platform for all the possible continual learning problems. However, at least in this report, our formulation is sufficient, as our interest only lies in continual supervised learning and nothing else. Due to this reason, we use the terms continual supervised learning and continual learning interchangeably.

B. Quantum Encoding

The quantum algorithms are designed to interact with the qubits and take full advantage of their properties, such as the high expressivity through the entanglements. However, this is also a source of dilemma because most of the available data are stored in classical bits which cannot be readily used in quantum algorithms. One therefore demands the methods to bridge between the classical data and quantum systems, namely the quantum encoding methods.

Out of the diverse set of quantum encoding methods, we limit our focus to the *amplitude encoding*:

Definition 1: [LaRose and Coyle, 2020, Huang et al., 2021, Schuld et al., 2020] *Let the N-dimensional real-valued normalized feature vector be* $\mathbf{x} \in \mathbb{R}^N$ *and the number of qubits be* $n = \lceil \log_2(N) \rceil$. *Then, the amplitude encoding map* $\phi : \mathbb{R}^N \to \mathbb{C}^n$ *maps classical input feature vector* $\mathbf{x}$ *to the input state* $|x\rangle$ *as:*

$$\mathbf{x} \to \sum_{k=1}^{N} x_k |k\rangle .$$

The clear advantage of amplitude encoding is the exponential reduction of the number of required qubits to express the original classical bit sequences. This exponential size reduction property is favorable for our experiments, considering the size of the classical input data that we use. For further details on the dataset, see Section IV.

C. Variational Quantum Classifier (VQC)

Here, we first describe the variational ansatz, then proceed to explain the VQC. *Variational Ansatz* is the stacked layers of the unitary operators where each operator is equipped with various quantum gates. Those quantum gates are associated with the adjustable parameters which could be freely adjusted to the desired values. In the context of QML, these variational ansatz are used in a way the classical ANN is employed in ordinary ML settings. In other words, the variational ansatz

[1]The formulation here is heavily inspired by Javed and White, 2019, and Lopez-Paz and Ranzato, 2017.

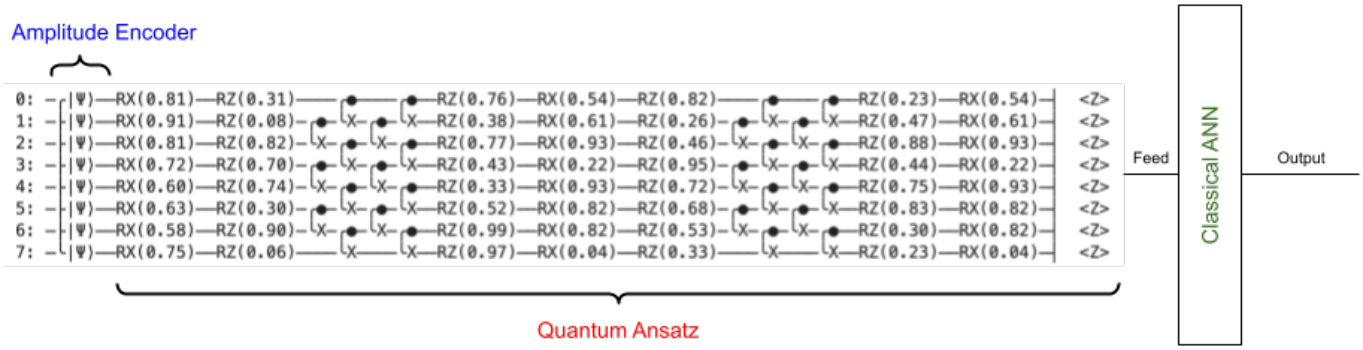

Fig. 1. Example VQC architecture with 8 qubits and 2 layers of variational ansatz. The first layer of Ψ corresponds to the amplitude encoding. Preceding layers up until the expectation value Z is the quantum ansatz. The classical ANN takes in the expectation value Z from variational ansatz and produces the final output.

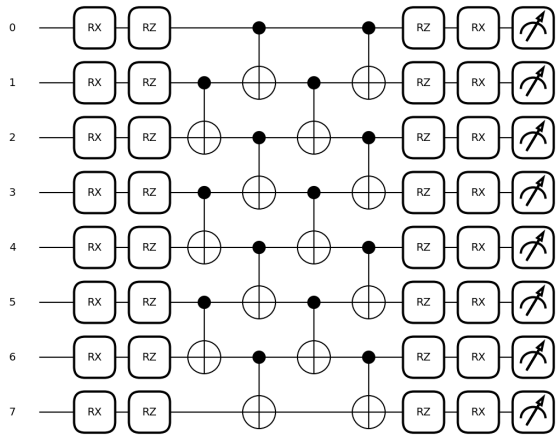

Fig. 2. Single layer variational ansatz over 8 qubits.

plays the role of function approximator, and the associated adjustable parameters are optimized to minimize the loss between the prediction and the targets. We do not go over the details of the optimization algorithms for the variational ansatz (such as parameter-shift by Benedetti et al., 2019) but rather focus on its architecture.

In this report, we utilize the variational quantum ansatz provided by Jiang et al., 2022. The main reason for choosing this is that the architecture is designed in a hardware-efficient way and it is used in the context of quantum continual learning. Like in the other variational quantum layers, it consists of multiple rotations and controlled-NOT (CNOT) operators. The precise definition of each layer is given as follows:

Definition 2: [Situ et al., 2023] *A single layer of the variational ansatz is defined as:*

$$U_j = \bigotimes_{i=1}^{n} R^i_{\mathbf{z}}(\alpha_{j,i,3}) \prod_{k=1}^{2} \left(\bigotimes_{i=1}^{\lfloor \frac{n}{2} \rfloor} \mathrm{CX}^{2i-1}_{2i} \bigotimes_{i=1}^{\lfloor \frac{n-1}{2} \rfloor} \mathrm{CX}^{2i}_{2i+1} \right) \bigotimes_{i=1}^{n} R^i_{\mathbf{z}}(\alpha_{j,i,2}) \bigotimes_{i=1}^{n} R^i_{\mathbf{x}}(\alpha_{j,i,1})$$

where n is the number of qubits, R^i denotes the rotation operators on the i-th qubit, CX^k_l are the CNOT operators acting over the k-th and l-th qubits, and α is the matrix of adjustable parameters for the rotation operators. The stack of these layers followed by the $R_{\mathbf{x}}$ gates produces the variational ansatz. Formally, the ansatz can be defined as:

Definition 3: [Situ et al., 2023] *Given the layer in Definition 2, overall variational ansatz is defined as a map from encoded input state $|x\rangle \in \mathbb{C}^n$ to the output state $|x'\rangle \in \mathbb{C}^n$:*

$$|x'\rangle = \bigotimes_{i=1}^{n} R^i_{\mathbf{x}}(\beta_i) \prod_{i=1}^{N_l} U_i |x\rangle .$$

where β_i are the adjustable parameters of the associated rotation operators and N_l is the number of layers. Finally, the outputs of the variational ansatz are the expectation values of each qubit: $z_i = \langle x'| \sigma^i_z |x'\rangle$, where σ^i_z is the measured observable on qubit i (implementation-wise, the observable here is the PauliZ). Figure 2 illustrates an example instance of a complete single layer quantum ansatz over 8 qubits.

Given the complete definition of the variational ansatz, we now define the VQC. VQC here is defined as the quantum-classical hybrid classifier [Benedetti et al., 2019, Alam et al., 2021]. In other words, it is the stack of amplitude encoder (Definition 1), variational ansatz (Definition 3), and postprocessing classical ANN. An example sketch is depicted in Figure 1.

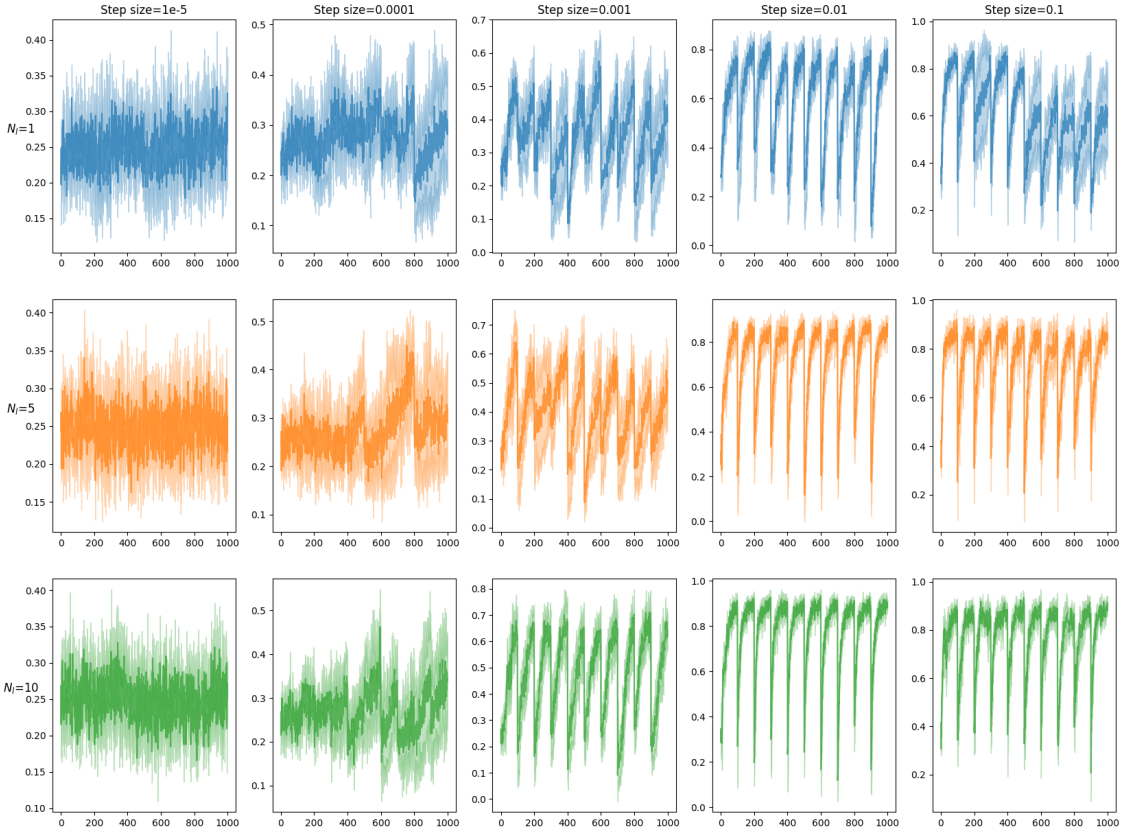

Fig. 3. Average online accuracy plots for VQC with different hyperparameter settings. Each column and row represents the different step-sizes and number of layers.

IV. Experiment Settings

In this section, we describe the details of each component of the experiments and the procedures. As a testbed, we utilize the label-permuted letter EMNIST, which is the modified version of the popular continual learning problem; label-permuted MNIST [Elsayed and Mahmood, 2024, Lyle et al., 2023]. At each timestep, the mini-batch of image-label pairs from the letter EMNIST dataset is provided. Each image is the grayscale handwritten alphabet (both lower and upper case), and the corresponding label is represented by the index between 1 through 26. All the images are resized into the size of 16×16 so the number of total pixels equals the power of 2 [Jiang et al., 2022]. After the resizing, images are normalized and flattened into a vector. Due to the lack of computational resources, we reduced the number of classes to 3 classes instead of the full 26 classes. Particularly, here, we chose the labels 23, 24, 25, and 26, namely the classes 'W', 'X', 'Y', and 'Z' since the lower and upper case of these characters do not significantly vary. These labels do not remain constant but permute periodically with the interval of M timesteps. Permute here means that the correspondence between the image classes and labels gets shuffled. For instance, after M steps of learning, the label 24 could indicate 'Z' instead of 'X', and so on. We also halved the number of samples per label, due to the computational constraints. So instead of including 19200 data points, we include 9600 data points.

The VQC architecture directly follows from the one described in Section III-C. The number of qubits is 8 (by $\log_2(16 \times 16) = 8$), and we stack N_l layers of variational ansatz. We do not specify the value of N_l here as it is the target of grid-search. The postprocessing classical ANN is a fully connected NN with a single hidden layer of size 64, activated with ReLU. We also test the performance of the classical ANN model as a reference. It consists of the 2 hidden fully connected layers of size 300 and 150 [Elsayed and Mahmood, 2024]. The outputs of each hidden layer are also activated by ReLU.

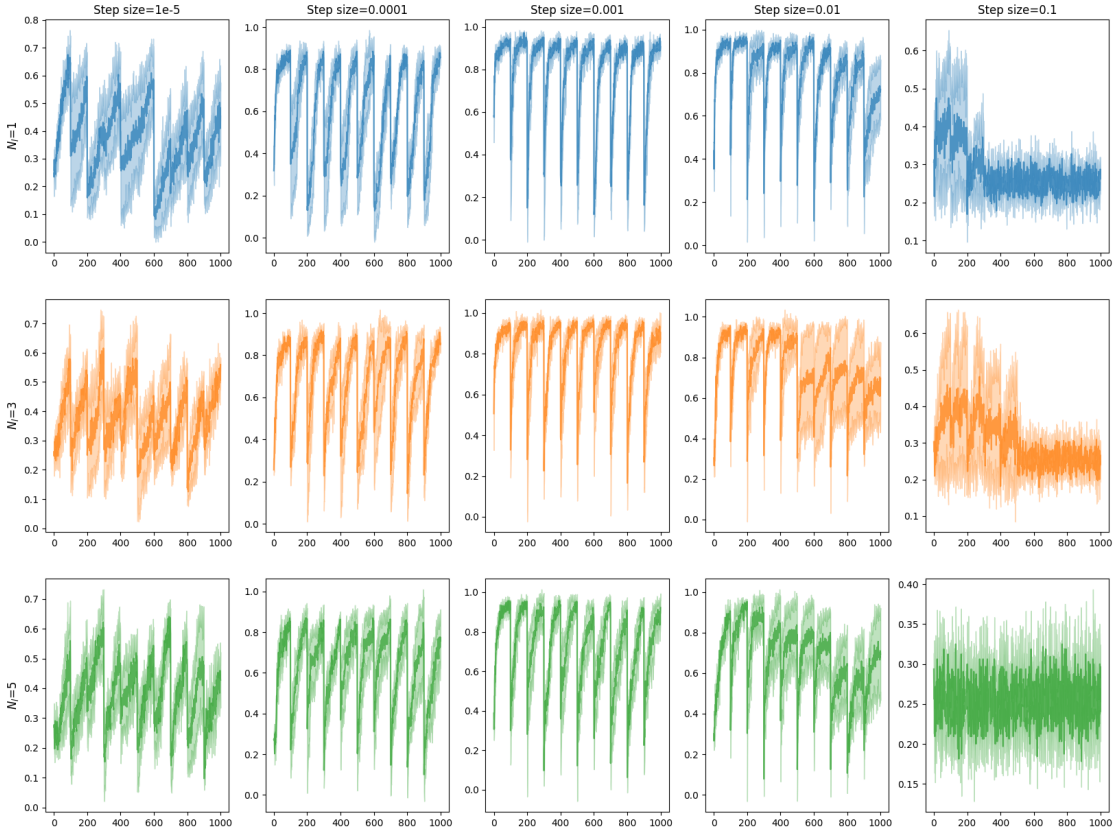

Fig. 4. Average online accuracy plots for classical ANN with different hyperparameter settings. Each column and row represents the different step-sizes and number of layers, respectively.

The experiment design is as follows. Each experiment consists of 5 independent runs[2]. We feed the mini-batch of label-permuted EMNIST data to VQC or classical ANN for 1000 timesteps on each run. The mini-batch size here is 64, and the label index is permuted per 100 timesteps. This implies that the model will witness the 10 different tasks on each run. The models are optimized with Adam optimizer [Kingma and Ba, 2015]. Since it is known that the step-sizes could be a crucial factor for the model plasticity [Berariu et al., 2021, Dohare et al., 2023], we conduct the grid search for the step-sizes over the values of $\{0.1, 0.01, 0.001, 0.0001, \text{1e-5}\}$, for all the models. The online training losses and online training accuracies are measured throughout the run. No measurements on the validation or test sets are obtained since those held-out datasets are not accessible in continual learning settings. Also, for both VQC and classical ANN, we conduct the additional grid-search on the number of layers. For the VQC, the search will be conducted over the range of $\{1, 5, 10\}$. For the classical ANN, we vary the number of layers with a size of 300 over the range of $\{1, 3, 5\}$. Finally, all the experiments are implemented and simulated with Pennylane and PyTorch [Bergholm et al., Paszke et al., 2019].

[2]"Independent" here implies that each run is processed with different random seeds.

V. Results & Analysis

In this section, we provide the results from the experiments described in the previous section and provide some insights.

A. *Loss of Plasticity with VQC*

Figure 3 depicts the online accuracy with the VQC for all the possible hyperparameter combinations. From the plots, it is evident that the VQC also suffers from the loss of plasticity with some hyperparameter settings. The most apparent one is when the layer number is 1 and step-size is 0.1 (i.e. a plot at the top-right corner). While the online accuracy reaches near 90% for the first two tasks, the maximum accuracy decreases as the timestep proceeds. Eventually, the accuracy only reaches up to 60%, which is a dramatic loss of plasticity considering

the initial reachable accuracy. On the other hand, the loss of plasticity is not apparent with other valid hyperparameter settings[3]. All the plots on 4th and 5th columns of Figure 3 except for the top-right corner consistently reach the same level of average accuracy on every task. The 5th column is particularly interesting as it indicates the mitigation of the loss of plasticity by increasing the number of ansatz layers. This phenomenon is extensively discussed in the next section, as this does not occur with the classical ANNs. Overall, we still observe the loss of plasticity occurs under a quantum continual learning setting.

B. Loss of Plasticity in VQC vs. Classical ANN

We now investigate the similarity and difference in the loss of plasticity between VQC and classical ANN. According to Figure 3 and 4, the dominant influence of the step-sizes on the loss of plasticity is evident for both VQC and classical ANN. While there are slight differences, almost all the accuracy curves in both figures are column-wise similar, indicating that the step-sizes are a more dominant factor in deciding the overall performance of the models (including the plasticity) under the continual learning settings. This common trait stresses the necessity of step-size adaptation algorithms (e.g. Jacobsen et al., 2019) regardless of the model architectures.

One of the most clear differences is the influence of the number of layers on the plasticity loss. From the right-most (5th) column of Figure 3, it is clear that the loss of plasticity has been relaxed and eventually mitigated as the number of layers in quantum ansatz increases. On the other hand, this pattern did not apply to the classical ANNs. This is evident from the 4th column of Figure 4, where all the plots show the severe degradation of the ability to learn regardless of the number of hidden layers. However, notice that simply varying the number of parameters by stacking additional hidden layers does not necessarily lead to performance improvements of the classical ANNs. To accurately attribute the correlation between the classical network size and loss of plasticity, one needs to conduct the hyperparameter search over the width, depth, resolution, and other various components of the network [Tan and Le, 2019]. Nevertheless, at least from our results, the number of layers is potentially an important decision factor in mitigating the loss of plasticity with VQC.

VI. Conclusion

In this report, we discovered the loss of plasticity in the QML settings and empirically analyzed their characteristics in comparison with the classical settings. The experimental results suggest that the quantum-classical hybrid VQC also loses plasticity under specific conditions, especially when the capacity of the quantum ansatz is small and the unright selection of step-size. We also presented that the step-size is the dominant decision factor on the plasticity loss for both classical ANN and VQC, and indicated the possibility that the number of layers in the variational ansatz is another important decision factor for the VQC. While the experimental results are plausible, the outcomes presented here are produced from relatively small-scaled experiments and thus it is impossible to completely exclude the possibility of having different results/insights by running the large-scaled experiments. Therefore, large-scale experiments (using full data from EMNIST, larger VQC/ANN architectures, etc) remain as future work.

[3] Valid hyperparameters here refers to those allowing the models to achieve at least 80% of accuracy. For example, in Figure 3, all the combinations of step-sizes $\{0.1, 0.01\}$ and number of layers $\{1, 2, 5\}$ are considered to be the valid hyperparameter settings (namely, the 4th and 5th column of the plot). We ignore other results since those fail to sufficiently fit the data.

References

Z. Abbas, R. Zhao, J. Modayil, A. White, and M. C. Machado. Loss of plasticity in continual deep reinforcement learning. In *Conference on Lifelong Learning Agents*, 2023.

M. Alam, S. Kundu, R. O. Topaloglu, and S. Ghosh. Quantum-classical hybrid machine learning for image classification (iccad special session paper). In *IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, 2021.

J. T. Ash and R. P. Adams. On warm-starting neural network training. In *International Conference on Neural Information Processing Systems*, 2020.

M. Benedetti, E. Lloyd, S. Sack, and M. Fiorentini. Parameterized quantum circuits as machine learning models. *Quantum Science and Technology*, 2019.

T. Berariu, W. Czarnecki, S. De, J. Bornschein, S. L. Smith, R. Pascanu, and C. Clopath. A study on the plasticity of neural networks. *CoRR*, 2021.

V. Bergholm, J. Izaac, M. Schuld, C. Gogolin, S. Ahmed, V. Ajith, M. Sohaib Alam, G. Alonso-Linaje, B. Akash-Narayanan, A. Asadi, J. M. Arrazola, U. Azad, S. Banning, C. Blank, T. R. Bromley, B. A. Cordier, J. Ceroni, A. Delgado, O. Di Matteo, A. Dusko, T. Garg, D. Guala, A. Hayes, R. Hill, A. Ijaz, T. Isacsson, D. Ittah, S. Jahangiri, P. Jain, E. Jiang, A. Khandelwal, K. Kottmann, R. A. Lang, C. Lee, T. Loke, A. Lowe, K. McKiernan, J. J. Meyer, J. A. Montañez-Barrera, R. Moyard, Z. Niu, L. J. O'Riordan, S. Oud, A. Panigrahi, C.-Y. Park, D. Polatajko, N. Quesada, C. Roberts, N. Sá, I. Schoch, B. Shi, S. Shu, S. Sim, A. Singh, I. Strandberg, J. Soni, A. Száva, S. Thabet, R. A. Vargas-Hernández, T. Vincent, N. Vitucci, M. Weber, D. Wierichs, R. Wiersema, M. Willmann, V. Wong, S. Zhang, and N. Killoran. PennyLane: Automatic Differentiation of Hybrid Quantum-classical Computations. *arXiv e-prints*.

S. Dohare, Q. Lan, and A. R. Mahmood. Overcoming policy collapse in deep reinforcement learning. In *Sixteenth European Workshop on Reinforcement Learning*, 2023.

M. Elsayed and A. R. Mahmood. Addressing loss of plasticity and catastrophic forgetting in continual learning. In *International Conference on Learning Representations*, 2024.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

H.-Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, and J. Mcclean. Power of data in quantum machine learning. *Nature Communications*, 2021.

A. Jacobsen, M. Schlegel, C. Linke, T. Degris, A. White, and M. White. Meta-descent for online, continual prediction. *AAAI Conference on Artificial Intelligence*, 2019.

K. Javed and M. White. Meta-learning representations for continual learning. In *Neural Information Processing Systems*, 2019.

W. Jiang, Z. Lu, and D.-L. Deng. Quantum continual learning overcoming catastrophic forgetting. *Chinese Physics Letters*, 2022.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell. Overcoming catastrophic forgetting in neural networks. *National Academy of Sciences*, 2016.

S. Kumar, H. Marklund, and B. V. Roy. Maintaining plasticity in continual learning via regenerative regularization, 2023.

Q. Lan. Variational quantum soft actor-critic, 2021.

R. LaRose and B. Coyle. Robust data encodings for quantum classifiers. 2020.

A. Lewandowski, H. Tanaka, D. Schuurmans, and M. C. Machado. Directions of curvature as an explanation for loss of plasticity, 2024.

D. Lopez-Paz and M. A. Ranzato. Gradient episodic memory for continual learning. In *Neural Information Processing Systems*, 2017.

C. Lyle, Z. Zheng, E. Nikishin, B. Avila Pires, R. Pascanu, and W. Dabney. Understanding plasticity in neural networks. In *International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2023.

E. Nikishin, M. Schwarzer, P. D'Oro, P.-L. Bacon, and A. Courville. The primacy bias in deep reinforcement learning. In *International Conference on Machine Learning*, 2022.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. *PyTorch: An Imperative Style, High-performance Deep Learning Library*. 2019.

M. Schuld, A. Bocharov, K. M. Svore, and N. Wiebe. Circuit-centric quantum classifiers. *Phys. Rev. A*, 2020.

H. Situ, T. Lu, M. Pan, and L. Li. Quantum continual learning of quantum data realizing knowledge backward transfer. *Physica A: Statistical Mechanics and its Applications*, 2023.

G. Sokar, R. Agarwal, P. S. Castro, and U. Evci. The dormant neuron phenomenon in deep reinforcement learning. In *International Conference on Machine Learning*, 2023.

M. Tan and Q. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, 2019.