Information Bottleneck of Quantum Neural Networks

Juexiao Wang
Department of Computer Science
Purdue University
wang5360@purdue.edu

Myeongsu Kim
Department of Computer Science
Purdue University
kim2167@purdue.edu

Sabre Kais
Department of Electrical and Computer Engineering
North Carolina State University
skais@ncsu.edu

Abstract

Occam's Razor, as a principle of model selection, favors simpler models that achieve the same empirical error rate to more complex models. Structural risk minimization finds such optimal model with a structural search by minimizing a high probability bound on the true error rate, enjoying the trade-off between minimizing the empirical error rate and the generalization bound. Tishby and Zaslavsky [2015] suggested to quantify such trade-off via the theoretical framework of the Information Bottleneck theory. It claims that Deep Neural Networks learn a compressed bottleneck representation that preserves only the relevant information about the target label Y from the source data X. The optimal compression-accuracy trade-off is characterized by the information-bottleneck function, providing a tight achievable lower bound on the representation length.

An open quantum system is equivalent to a noisy neural network that can perform machine learning tasks with non-unitary evolution, leveraging the potential power of entanglement and superposition. Therefore understanding the learnability of a quantum system is crucial in the noisy intermediate-scale quantum (NISQ) era. In this paper, we are generalizing the Information Bottleneck theory to quantum Hilbert space to characterize the learnability of a quantum system with quantum mutual information as the complexity measure of a quantum concept class and the lower bound of the prediction accuracy of a quantum hypothesis. We compared the training dynamics of quantum neural networks with the classical analog in the information plane and, through numerical simulations, demonstrated that the learned quantum representations are more expressive than classical representations.

1 Introduction

The Information-Bottleneck (IB) method, introduced in Tishby et al. [2000], considers extracting relevant information about the target signal Y through a correlated observable X with a compressed representation T. The objective is to find such compressed representation with minimal average description length, characterized by the mutual information $I(X;T)$, subject to the constraint on the expected distortion as the relevant information $I(T;Y)$ about the target signal Y. The optimal trade-off between compressing the representation and preserving meaningful information is characterized by the information bottleneck function that separates the achievable and non-achievable region of representations.

38th Second Workshop on Machine Learning with New Compute Paradigms at NeurIPS 2024(MLNCP 2024).

Information-Bottleneck method suggests a learning algorithm known as structural risk minimization, minimizing a high probability bound on the true error rate of the learned hypothesis with respect to a distribution, consisting of the empirical error rate and generalization bound.

Mutual information provides a robust measure of the model complexity and the prediction accuracy of a hypothesis. $I(X;T)$ characterizes the maximal achievable rate of the distinguishable typical sequences with asymptotic zero probability of error which provides an ϵ-cover for the representation space. Relevant information $I(T;Y)$ provides a lower bound of the prediction accuracy Gilad-Bachrach et al. [2003], Cover and Thomas [2006]. Therefore minimizing the Lagrangian is equivalent to minimizing the the high probability bound.

Shwartz-Ziv and Tishby [2017] empirically demonstrated that neural network trained with stochastic gradient descent (SGD) implicitly achieves the Information Bottleneck bound. The optimization dynamics of SGD is a Wiener process with two distinct phases in resonance to the phase transition in thermal annealing dynamics. Empirical risk minimization (ERM) phase minimizes the empirical risks with an increase in $I(T;Y)$. Representation compression phase has decreasing rate $I(X;T)$ where the input feature is compressed into a finer grained clustered representation of X such that it can generalize better.

Quantum Neural Networks (QNN) harness the power of entanglement and superposition, to perform machine learning task. A quantum model is a kernel function that maps data into a quantum Hilbert space with dimension exponentially growing with the number of qubits. Quantum channels characterized by unitary evolution transforms the computational basis followed by a projective measurement in such basis defines a linear decision boundary in the quantum feature space. Training a quantum model is equivalent to finding the optimal Positive-Operator Valued Measure (POVM) measurement that minimizes a data-dependent cost function Schuld [2021].

In this paper, we compare the learning dynamics of quantum and classical neural network in the information plane and conclude from the numerical simulation that the learned quantum representations are more expressive than the classical representations.

2 Information Bottleneck Theory for Deep Learning

2.1 Mutual Information

Definition 2.1. (Classical Mutual information) Shannon's mutual information between two random variables X and Y with joint probability distribution P_{XY}, is defined as the relative entropy between the joint distribution P_{XY} and the product of its marginal distributions $P_X \otimes P_Y$.

$$\begin{aligned} I(X;Y) &= D(P_{XY}||P_X \otimes P_Y) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}$$

The quantum generalization of the mutual information is defined with quantum relative entropy.

Definition 2.2. (Quantum Mutual Information) Quantum mutual information is defined as the quantum relative entropy between the joint state ρ_{XY} and the product state of $\rho_X \otimes \rho_Y$ of individual subsystems.

$$\begin{aligned} I(X;Y)_\rho &= S(\rho_{XY}||\rho_X \otimes \rho_Y) \\ &= H(X)_\rho + H(Y)_\rho - H(X,Y)_\rho \end{aligned}$$

2.2 Information Bottleneck of classical DNNs

Information Bottleneck method introduced in Tishby et al. [2000] finds the optimal code for the source signal that preserves the relevant information about a target signal with minimal codeword length. Let X denote the signal space with a probability distribution P_X and let T denote its quantized codebook. For every signal element $x \in X$, it is mapped to a codeword $T \in T$ via a stochastic map $P_{T|X}$, which induces a soft partition of the measurable space of X where each block is associated with the codebook elements $T \in T$, with probability $P_{T|X}(T|x)$ Tishby et al. [2000].

The optimal representation T is an approximate minimal sufficient statistics of X with respect to Y that trades off between compressing the representation of X and preserving the meaningful information about Y. Tishby et al. [2000] characterizes this trade-off through the information bottleneck function $R^I(I_Y)$, defined as the minimal achievable rate under a given constraint on the expected distortion measured by $I(T;Y)$ Gilad-Bachrach et al. [2003].

$$\min_{p(T|x):I(T;Y)\geq I_Y} I(X;T) \tag{1}$$

Finding the information-bottleneck function is a variational problem that can be solved by minimizing the Lagrangian functional subject to the Markov chain constraint $Y \to X \to \tilde{X}$ with joint distribution $P_{X,Y}P_{T|X}$

$$L[\,p(T|x)] \;= I(T;X) - \beta I(T;Y) \tag{2}$$

The positive Lagrange multiplier β operates as a tradeoff parameter between the complexity (rate) of the representation $I(X;T)$ and the amount of preserved relevant information $I(T;Y)$. By varying the parameter β one can explore the trade-off between the preserved meaningful information and compression at various resolutions Tishby et al. [2000].

The output of DNN layers form a Markov chain of successive internal representations of the input features. Tishby and Zaslavsky [2015] suggested that the optimal DNN induces the simplest representations that preserves the relevant information for predicting the target label. The optimal representations of the input feature achieves the information bottleneck bound. The mutual information between the input and output of a DNN layer can be estimated with training samples $S = \{(x_i, y_i)\}_{i=1}^n$ using an empirical plug-in estimator Goldfeld and Polyanskiy [2020].

2.3 Information Bottleneck of quantum DNNs

The quantum generalization to the information bottleneck theory is to substitute the Shannon entropy in the definition of classical relative entropy with von Neumann entropy in the definition of quantum relative entropy Catli and Wiebe [2022].

Definition 2.3. (Quantum Relative Entropy) Let ρ_A and ρ_B be density operators acting on C^{2^n} for integer n, the quantum relative entropy is defined to be

$$S(\rho_A||\rho_B) = Tr(\rho_A log(\rho_A)) - Tr(\rho_A log(\rho_B)) \tag{3}$$

Definition 2.4. (Quantum Information Bottleneck Objective Function) The quantum information bottleneck objective function for a distribution ρ_{TXY} and a quantum channel ϕ is denoted $L[\phi]$ which is defined to be

$$L[\phi] = I(T;X) - \beta I(T;Y) = S(\rho_{XT}||\rho_X \otimes \rho_T) - \beta S(\rho_{XY}||\rho_X \otimes \rho_Y) \tag{4}$$

The training samples $S = \{(x_i, y_i)\}_{i=1}^n$ is mapped to a joint mixed state $\rho_{XY} = \sum_{x,y} \hat{P}_{XY}(x,y)\,|x\rangle\langle x| \otimes |y\rangle\langle y|$ where $\hat{P}_{XY}(x,y)$ is the empirical joint distribution.

$$\rho_{XY} = \sum_{x,y} \hat{P}_{XY}(x,y)\,|x\rangle\langle x| \otimes |y\rangle\langle y| \tag{5}$$

$$\rho_X = \mathrm{Tr}_Y\,\rho_{XY} = \sum_x \hat{P}_X(x)\,|x\rangle\langle x| \tag{6}$$

$$\rho_{\hat{X}} = \mathrm{Tr}_Y\,\rho_{\hat{X}Y} = \sum_x \hat{P}_{\hat{X}}(T)\,|\hat{x}\rangle\langle \hat{x}| \tag{7}$$

$$\rho_Y = \mathrm{Tr}_X\,\rho_{XY} = \sum_y \hat{P}_Y(y)\,|y\rangle\langle y| \tag{8}$$

3 Mutual Information Estimation

Mutual information is a nonlinear functional of a joint probability distribution P_{XY} over the product measurable space $\mathcal{X} \times \mathcal{Y}$ and it is invariant under invertible transformations. In a deterministic deep

neural network, mutual information between the input X and the output of one of its internal layer T_l is provably a constant if X is a discrete random variable or infinite if X is a continuous random variable Goldfeld et al. [2019]. Therefore it is meaningless to track the flow of information in the training of a deterministic deep neural network. Goldfeld et al. [2019] develops a framework for tracking the flow of information in DNNs by injecting isotropic gaussian noise over the domain of the learned representation T_l such that mutual information is sensitive to the complexity of the class of functions parameterized by DNNs. As an analog to the classical noisy DNNs, we inject quantum noise to QNNs, in particular we employ the amplitude damping channel as a layer of QNNs. And we estimate the classical mutual information $I(X;T_l)$ during the training of both quantum and classical neural networks such that we can compare the model complexity and prediction accuracy of two different models under the same information measure.

Estimating the mutual information in a classical noisy DNNs is reduced to estimating the differential entropy of the representation distribution under Gaussian convolution. We apply the non-parametric Kernel Density estimator and k-NN based estimator in Saxe et al. [2018] for mutual information estimation in the training of classical neural networks.

To estimate the classical information of a quantum system, we employed the method of orthogonal states. Estimating the mutual information in a quantum noisy DNNs, we simplifies the estimation of joint distribution using the Pennylane package Bergholm et al. [2018], which calculates the density matrix of the joint state ρ_{XT} of the composite system on the input and the output system of the amplitude damping channel. And by estimating the mutual information of a mixture of orthogonal quantum states, the calculation of quantum mutual information using Pennylane Bergholm et al. [2018] is equivalent to estimating the classical mutual information of the joint distribution over an ensemble of product states.

4 Numerical simulation

We compared the learning dynamics of classical neural networks and quantum neural networks on a constructed dataset in the information plane.

4.1 Dataset

The dataset is constructed to be a set of standard normal basis vectors $S = \{(x_i, y_i)\}_{i=1}^N, x_i^T x_j = 0, \forall i \neq j$ and we use the amplitude encoding to map the classical training example x_i to a quantum state $|x_i\rangle = \sum_{d=1}^N x_i^d |d\rangle$. Since the classical mutual information provides a robust measure of the epsilon-cover of the class of representations, we would like to estimate the classical mutual information across the training of both classical and quantum neural networks to compare the generalization capability of the learned representations.

The construction of the set of orthogonal basis vectors simplifies the estimation of classical mutual information of quantum representations in the sense that the von Neumann entropy of a mixture of orthogonal quantum states with density matrix $\rho = \sum_i p_i |x_i\rangle \langle x_i|$ is reduced to the Shannon entropy of the probability distribution over the ensemble of states $S(\rho) = -\text{Tr}(\rho \log \rho) = -\sum_i p_i \log p_i$. The convenience of calculating the density matrix provided by the Pennylane package Bergholm et al. [2018] under a specified quantum noise (Markov transition kernel) saves us the effort of estimating the joint distribution using finite samples.

Moreover, the training samples are linearly separable with a full rank feature matrix X thus there exists a unique linear separator that is consistent with the training samples $w = X^{-1}y$ which provides a good sanity check on the training accuracy achieved by the two models.

4.2 Neural Network Architecture

To have a fair comparison between the expressiveness of quantum neural networks and classical neural networks, we employ a classical neural network architecture with identity activation function and have the same number of parameters as the quantum neural networks. Our dataset is a set of 32 and 64 standard normal basis vectors and classical neural network has architecture 32-32-32-32-32-32 with cross entropy loss function. Quantum neural network is composed of a set of

6 non-unitary operators composed of rotation gates on each qubits and pairwise CNOT gates and amplitude damping channels.

4.3 Classical vs Quantum information plane

We compared the classical training trajectories with quantum training trajectories. We trained with 80 percent of the dataset and the mutual information is estimated on the full dataset as the test accuracy. The training trajectories is shown in Fig.1.

4.3.1 Sample complexity determined by $I(X;T)$

For large typical input patterns Shwartz-Ziv and Tishby [2017], $2^{I(X;T)}$ measures the number of typical patterns in the representation space, thus determining the complexity of the class which in turn determines the sample complexity of agnostic PAC learning the class using the empirical risk minimization (ERM) learning algorithm $\mathcal{M}_{ERM}^{AG} = O(\frac{2}{\epsilon^2}(2^{I(X;T)} + \ln\left(\frac{2}{\delta}\right)))$.

At the end of the training, the mutual information between quantum representation and the input is shown to be greater than the mutual information between classical representation and the input $I(X;T_Q) \approx 4.6 > I(X;T_C) \approx 1$. The result can be interpreted in that classical representations are clustered into two clusters $I(X;T_C) = H(T_C) - H(T_C|X) \approx H(T_C) = 1$ while the quantum representations have more patterns, which implies that the sample complexity to achieve the same generalization guarantee for quantum neural networks is higher than that for the classical neural networks.

We conjecture that the higher degree of the expressiveness of quantum representations is a result of the richness of quantum Hilbert space, which is a complex vector space with more elements than the real vector space and we can potentially prove it using the chain rule of mutual information under the assumption that the input distribution P_X and the noise model $P_{T|X}$ are the same and the real component T_R and imaginary component T_I of the representation are distributed independently and identically by decomposing the mutual information into $I(X;T_Q) = I(X;T_R) + I(X;T_I) \geq I(X;T_R) = I(X;T_C)$ Xie et al. [2014].

4.3.2 Test accuracy determined by $I(T;Y)$

For large typical input patterns Shwartz-Ziv and Tishby [2017], $I(T;Y)$ provides a lower bound on the testing accuracy. At the end of the training, the mutual information between quantum representation and the target label is less than that between classical representation and the target label $I(T_Q;Y) \approx 0.9 > I(T_C;Y) \approx 0.7$, which implies that quantum representations generalizes better than the classical representations.

4.3.3 Loss landscape

We also observe that there exists a flatten area in the quantum information plane, as shown in Fig. (1), where there is high degeneracy in $I(Y;T)$ which might be an indication of barren plateau in the quantum loss landscape Mcclean et al. [2018].

5 Conclusion

In order to study the fundamental question in learning theory, the quantification of the basic tradeoff between the complexity of a model and its predictive accuracy, Tishby and coworkers introduce the concept of "Information Bottleneck". Thus, Shannon's mutual information allows us to quantify the tradeoff and measure both the model's complexity and its prediction accuracy. Their work focuses on classical machine learning by studying the nature of this complexity-accuracy tradeoff and discuss some of its theoretical properties. Furthermore, they present relations to classical information theoretic problems, such as rate-distortion theory, cost-capacity tradeoff and source coding with side informationTishby and Zaslavsky [2015].

In this paper, we compared the training dynamics quantum neural network and classical neural network in the information plane. We demonstrated through numerical simulations that quantum neural

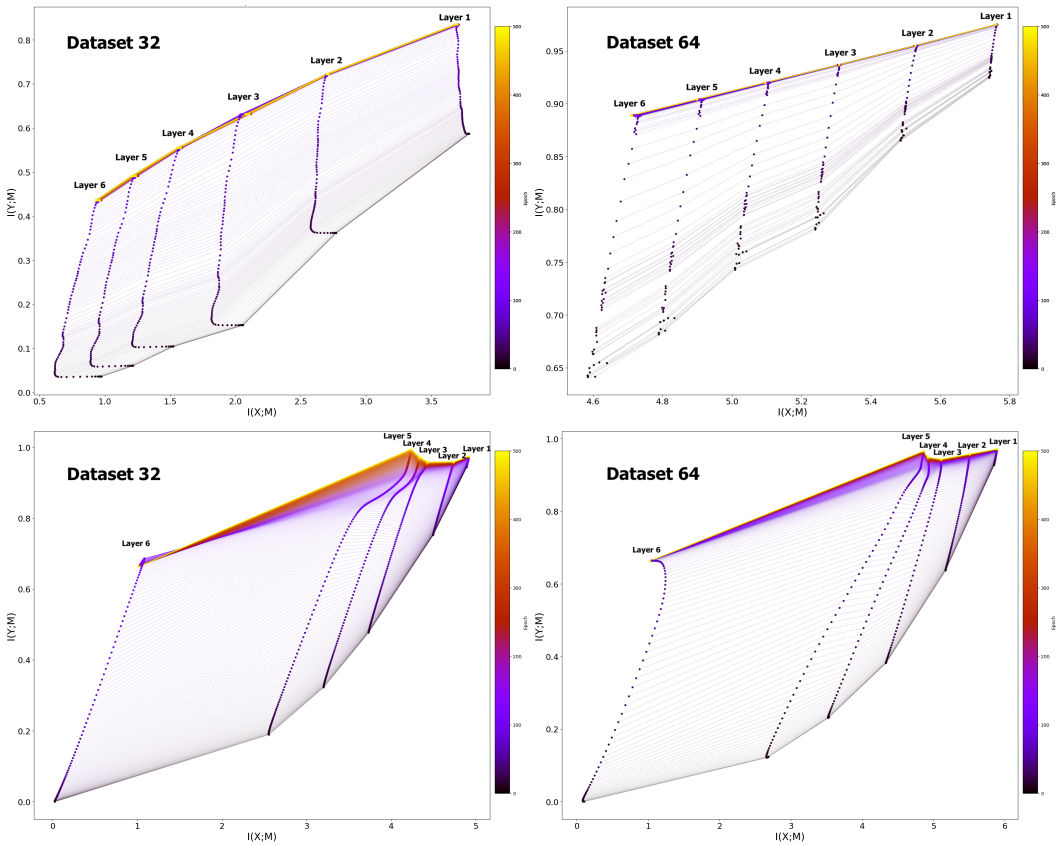

Figure 1: Information Plane of the training dynamics of a Quantum Neural Network (upper) and Classical Neural Network (lower). The x-axis (I(X;M)) shows the mutual information between the input X and the compressed representation M, while the y-axis (I(Y;M)) represents the mutual information between the target Y and the representation M, illustrating how much relevant information is retained for prediction.

network is a more expressive concept class which can fit dataset with more complicated structures thus decreasing the approximation error, while taking more samples to achieve the same generalization guarantee thus increasing the estimation error. For future work, further simulation results might be needed to provide more convincing conclusions and exact bounds. The implementation and related resources for this work are available at `https://github.com/kim2167/Information_Bottleneck_QNN`.

6 Acknowledgement

We would like to thank Dr. Manas Sajjan for fruitful discussions. J.W. thanks the members from Kais group for helpful discussions and the precious suggestions and feedback received from the discussions with several professors throughout the project.

References

Ville Bergholm, Josh A. Izaac, Maria Schuld, Christian Gogolin, Ankit Khandelwal, and Nathan Killoran. Pennylane: Automatic differentiation of hybrid quantum-classical computations. *ArXiv*, abs/1811.04968, 2018. URL `https://api.semanticscholar.org/CorpusID:53289556`.

Ahmet Burak Catli and Nathan Wiebe. Training quantum neural networks using the quantum information bottleneck method, 2022.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.

Ran Gilad-Bachrach, Amir Navot, and Naftali Tishby. An information theoretic tradeoff between complexity and accuracy. *Lecture Notes in Computer Science*, 07 2003. doi: 10.1007/978-3-540-45167-9_43.

Ziv Goldfeld and Yury Polyanskiy. The information bottleneck problem and its applications in machine learning. *CoRR*, abs/2004.14941, 2020. URL `https://arxiv.org/abs/2004.14941`.

Ziv Goldfeld, Ewout Van Den Berg, Kristjan Greenewald, Igor Melnyk, Nam Nguyen, Brian Kingsbury, and Yury Polyanskiy. Estimating information flow in deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2299–2308. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/goldfeld19a.html`.

Jarrod Mcclean, Sergio Boixo, Vadim Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9, 11 2018. doi: 10.1038/s41467-018-07090-4.

Andrew Michael Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan Daniel Tracey, and David Daniel Cox. On the information bottleneck theory of deep learning. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=ry_WPG-A-`.

Maria Schuld. Supervised quantum machine learning models are kernel methods, 2021.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *CoRR*, abs/1703.00810, 2017. URL `http://arxiv.org/abs/1703.00810`.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle, 2015.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method, 2000.

Qiuliang Xie, Zhaocheng Wang, and Zhixing Yang. Polar decomposition of mutual information over complex-valued channels. *IEEE Transactions on Information Theory*, 60(6):3163–3171, 2014. doi: 10.1109/TIT.2014.2313081.