Representation learning for Qunatum States

Xie, Shubing

Representation learning for Quantum states-using Classical shadow transformer

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE
in
PHYSICS

Author :	Shubing Xie
Student ID :	3554309
Supervisor :	Evert van Nieuwenburg
2^{nd} corrector :	Jordi Tura

Leiden, The Netherlands, January 30, 2024

Representation learning for Quantum states-using Classical shadow transformer

Shubing Xie

Institute-Lorentz, Leiden University
P.O. Box 9500, 2300 RA Leiden, The Netherlands

January 30, 2024

Abstract

Quantum computing stands at the forefront of modern science, where understanding quantum states is crucial for progress. This thesis introduces the Classical Shadow Transformer (CST), an AI model crafted to reconstruct quantum states from classical data. Trained on GHZ, W, and Zero states, each with unique entanglement levels, the CST employs shadow tomography and a transformer architecture with a variational bottleneck to interpret measurement outcomes. The CST shines in decoding less entangled states and is challenged by the intricacies of maximally entangled states, particularly the W state. This contrast in learning efficiency reveals key differences in entanglement types. The investigation into the CST's latent space provides insights into the interpretability of quantum states, showcasing how AI can unravel quantum complexity. These insights pave the way for future quantum computing advancements, positioning AI as a tool for demystifying quantum phenomena.

Contents

Chapter 1

Introduction

1.1 Context and Motivation

Quantum computing, heralding a new era in technological advancement, offers unparalleled computational capabilities. Central to this field are further understanding of quantum states, which embody the capabilities and dynamics of quantum systems. Despite their importance, the high-dimensional and probabilistic nature of these states poses significant challenges in their interpretation and analysis[1] [2][3].

Traditionally, quantum state tomography [4]has been the mainstay in analyzing these states. However, as quantum systems grow in size, traditional methods face scalability issues[5], presenting a major barrier to harnessing the full potential of quantum computing. In response, the field of machine learning, particularly representation learning, emerges as a promising solution. It aims to develop interpretable representations of complex data, offering new avenues for quantum state analysis.

A notable development in this context is the Classical Shadow Transformer (CST), inspired by Yi Zhuang's work [6]. The CST innovatively combines quantum physics and machine learning techniques to interpret quantum information through classical data. This approach addresses scalability issues and provides fresh insights into quantum state interpretability.

A key aspect of quantum computing that poses a particular challenge is the complex entanglement in quantum states, especially notable in GHZ and W states. Numerous debates and analyses focus on the extent of entanglement in these two states[7] [8] [9], despite being termed 'maximally entangled states' in a 3-qubit case. [10] Understanding these entangled states' properties and their implications for quantum computing is crucial. Additionally, the CST's ability to analyze and represent these states in its latent space further underscores the importance of machine learning in quantum state analysis.

1.2 Research Objectives and Structure

This research aims to evaluate the effectiveness of the Classical Shadow Transformer in understanding and representing quantum states. The study focuses on applying CST to various quantum states, such as GHZ and W states, to explore their entanglement properties and learnability. A significant component of the research is analyzing the CST's latent space representations to understand how these representations capture the complexities of different quantum states.

The thesis is organized as follows:

- Introduction to key concepts in quantum information science, including the fundamentals of quantum states and their entanglement characteristics.
- Detailed exploration of the Classical Shadow Transformer, inspired by YiZhuang's work, and its application in quantum state analysis.
- Experimental analysis of CST's performance in representing various quantum states, with a focus on GHZ and W states.
- Interpretation of latent space representations within the CST and their implications for understanding quantum states.
- Concluding remarks discussing the implications of the findings and directions for future research in quantum computing and machine learning.

This thesis, situated at the intersection of quantum information science and machine learning, makes a substantial contribution to our comprehension of quantum states. By integrating machine learning techniques in the analysis of quantum states, this work not only deepens our understanding in this field but also paves the way for novel developments in quantum computing and information processing.

Chapter 2

Preliminaries

2.1 Basic Concepts

Quantum information science is a field of study based on the principles of quantum mechanics. It encompasses quantum computing, quantum communication, and quantum cryptography, among others. At the heart of quantum information science are the concepts of quantum states, quantum gates, Pauli operations, measurement bases, and various metrics such as fidelity and entropy.

2.1.1 Quantum States

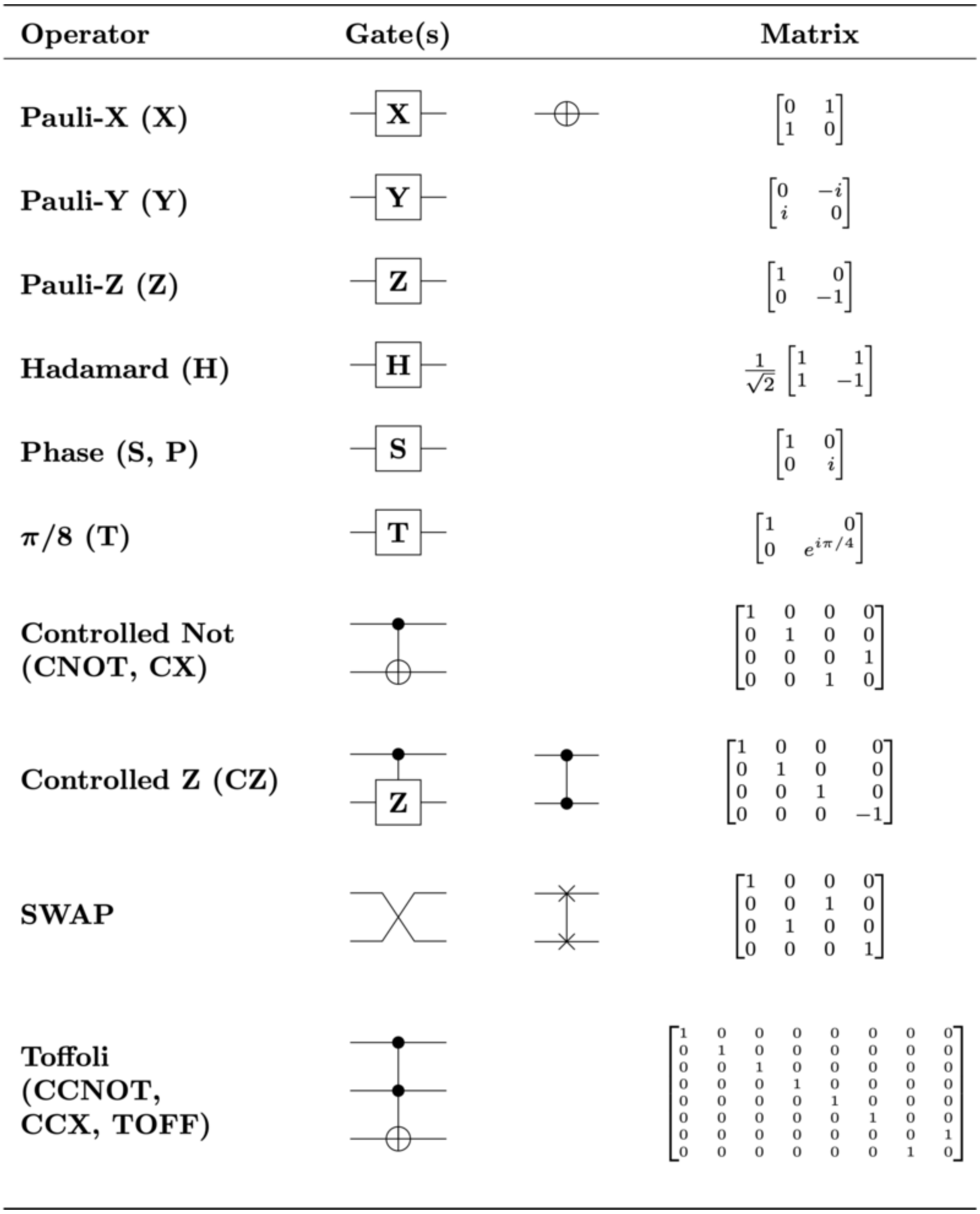
A quantum state represents the state of a quantum system and is typically denoted by the ket notation $|\psi\rangle$. The state $|\psi\rangle$ is a vector in a Hilbert space, and for a qubit, the simplest quantum system, it can be represented as a linear combination of the basis states $|0\rangle$ and $|1\rangle$:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle, \tag{2.1}$$

where α and β are complex numbers that satisfy the normalization condition $|\alpha|^2 + |\beta|^2 = 1$.

2.1.2 Quantum Gates

Quantum gates are operations that change the state of qubits. They are the quantum analogue of classical logic gates and are represented by unitary matrices. A unitary operation U that acts on a state $|\psi\rangle$ transforms it to a new state $U|\psi\rangle$.Some common quantum gates and their corresponding matrix representation can be seen in Table2.1



Operator	Gate(s)	Matrix
Pauli-X (X)	**X**	$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$
Pauli-Y (Y)	**Y**	$\begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}$
Pauli-Z (Z)	**Z**	$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$
Hadamard (H)	**H**	$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$
Phase (S, P)	**S**	$\begin{bmatrix} 1 & 0 \\ 0 & i \end{bmatrix}$
π/8 (T)	**T**	$\begin{bmatrix} 1 & 0 \\ 0 & e^{i\pi/4} \end{bmatrix}$
Controlled Not (CNOT, CX)		$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$
Controlled Z (CZ)	**Z**	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$
SWAP		$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$
Toffoli (CCNOT, CCX, TOFF)		$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$

Figure 2.1: *Common Quantum Logic Gates (Source: Wikipedia)*

2.1.3 Pauli Operators and Measurement Bases

Pauli operators are a set of matrices that form a basis for the space of 2×2 Hermitian matrices. They are defined as:

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{2.2}$$

Measurement in quantum mechanics is the process of determining the state of a qubit. The measurement basis typically refers to the eigenbasis of the Pauli operators.

2.1.4 Fidelity

Fidelity is a measure of similarity between two quantum states, which can be either pure or mixed states represented by density matrices.

For two mixed states described by density matrices ρ and σ, the fidelity $F(\rho, \sigma)$ is defined as:

$$F(\rho, \sigma) = \left(\mathrm{tr} \sqrt{\sqrt{\rho}\sigma\sqrt{\rho}} \right)^2. \tag{2.3}$$

If at least one of the states (say ρ) is pure, the fidelity simplifies to:

$$F(\rho, \sigma) = \mathrm{tr}(\sigma\rho). \tag{2.4}$$

And if both states are pure, the fidelity further simplifies to the square of the modulus of the inner product of the two state vectors. Fidelity is symmetric and bounded between 0 and 1. It is unitarily invariant, meaning it remains unchanged under unitary transformations of the quantum states.

2.1.5 Entropy

Entropy, in the context of quantum information, is a measure of uncertainty or disorder within a quantum system. The von Neumann entropy of a quantum state ρ is defined as:

$$S(\rho) = -\mathrm{Tr}(\rho \log_2 \rho). \tag{2.5}$$

2.1.6 Quantum Entanglement

Introduction

Quantum entanglement is a remarkable phenomenon where the quantum states of two or more particles become intertwined in such a way that the state of each particle cannot be described independently of the others, regardless of the distance separating them. This counterintuitive aspect challenges classical intuitions about separability and locality. Entanglement was first critically discussed in the context of the Einstein-Podolsky-Rosen (EPR) paradox in 1935 [11], questioning the completeness of quantum mechanics. John Bell, in 1964, introduced inequalities [12] that experimentally distinguished between quantum entanglement and classical correlations, further illuminating the peculiar nature of quantum entanglement.

2.1.7 Characterizing Quantum Entanglement

Quantum entanglement is characterized by specific criteria that distinguish entangled states from separable ones. Two primary criteria are the Peres-Horodecki (PPT) criterion and the inseparability of product states.

Peres-Horodecki (PPT) Criterion

The Peres-Horodecki, or Positive Partial Transpose (PPT), criterion provides a necessary condition for a bipartite state to be separable. Introduced by Asher Peres and refined by MichaÅ and Ryszard Horodecki, the criterion asserts that if the partial transpose of a density matrix of a separable state has non-negative eigenvalues, then the state is considered separable. Mathematically, for a bipartite system described by a density matrix ρ_{AB}, the PPT criterion is expressed as:

$$\text{if } \rho_{AB}^{T_B} \text{ has negative eigenvalues, then } \rho_{AB} \text{ is entangled} \tag{2.6}$$

where $\rho_{AB}^{T_B}$ denotes the partial transpose of ρ_{AB} with respect to subsystem B. The partial transpose is defined by the operation:

$$\rho_{AB}^{T_B} = (I \otimes T)(\rho_{AB}) = \sum_{ijkl} p_{kl}^{ij} |i\rangle\langle j| \otimes (|k\rangle\langle l|)^T \tag{2.7}$$

This can be visualized more clearly when ρ_{AB} is represented as a block matrix, and the partial transpose with respect to subsystem B is taken across these blocks. If the resulting matrix $\rho_{AB}^{T_B}$ has any negative eigenvalues, it indicates that ρ_{AB} is entangled.

Inseparability of Product States

Another fundamental aspect of characterizing entangled states is their non-decomposability into direct product states. A bipartite quantum state

$|\psi\rangle$ is entangled if it cannot be represented as a product of individual states of its subsystems, mathematically expressed as:

$$|\psi\rangle \neq |\phi_A\rangle \otimes |\phi_B\rangle \tag{2.8}$$

for any $|\phi_A\rangle$ in the state space of subsystem A and $|\phi_B\rangle$ in the state space of subsystem B. This inseparability criterion is fundamental to the definition of entanglement, implying that the properties of the whole system cannot be described merely by its parts [13].

Concurrence

Concurrence is a quantitative measure of entanglement for a pair of qubits. For a mixed state ρ, concurrence $C(\rho)$ is defined as:

$$C(\rho) = \max\{0, \lambda_1 - \lambda_2 - \lambda_3 - \lambda_4\} \tag{2.9}$$

where λ_i (with $i = 1, 2, 3, 4$) are the eigenvalues, in decreasing order, of the matrix $\sqrt{\sqrt{\rho}\tilde{\rho}\sqrt{\rho}}$, and $\tilde{\rho} = (\sigma_y \otimes \sigma_y)\rho^*(\sigma_y \otimes \sigma_y)$ [14].

Negativity

Negativity, another entanglement measure, is applicable to mixed states of any dimension. It quantifies the degree to which the PPT criterion is violated. Negativity is defined as the sum of the negative eigenvalues of the partial transpose ρ^{T_B} of the density matrix ρ:

$$\mathcal{N}(\rho) = \frac{||\rho^{T_B}||_1 - 1}{2} \tag{2.10}$$

where $||\cdot||_1$ denotes the trace norm [15].

Schmidt Number

The Schmidt number is a basis-independent measure of entanglement for bipartite pure states. It is defined by the number of non-zero terms in the Schmidt decomposition:

$$|\psi\rangle = \sum_i \sqrt{\lambda_i} |u_i\rangle |v_i\rangle \tag{2.11}$$

where λ_i are the Schmidt coefficients [16]. The Schmidt number gives the minimum number of product states needed to express the entangled state, with higher numbers indicating stronger entanglement.

2.2 Shadow Tomography

In quantum physics, there are multiple "languages" to describe a quantum system, such as wave functions and density matrices. These descriptions are uniquely quantum, not directly observable or intuitively understandable in classical terms, and their complexity grows exponentially with the size of the system. The concept of "classical Shadow" emerges from the idea of describing quantum systems using classical information. It leverages the act of observation, which, under the Copenhagen interpretation, collapses the wave function to a specific value. Classical Shadowã[4] refers to using the outcomes of certain measurements and their probability distributions to characterize a quantum state, thereby forming a 'classical language' for describing quantum phenomena.

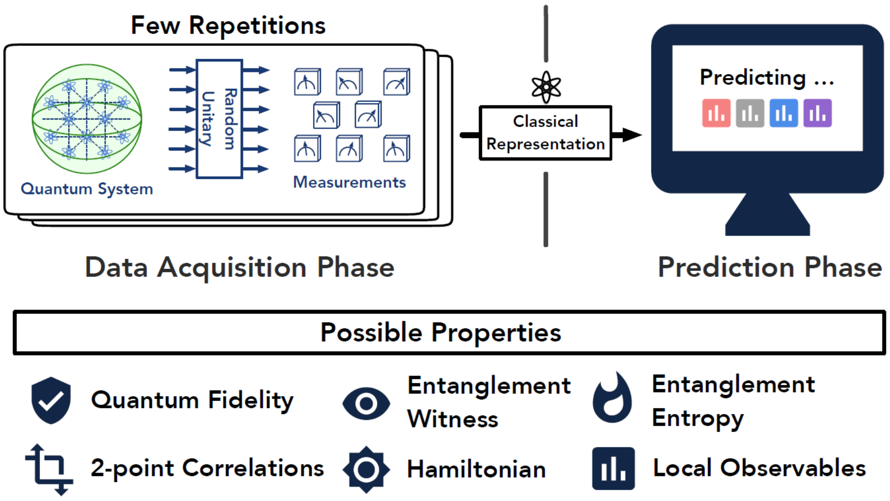


Figure 2.2: *A figure from[4] that represents the procedure of shadow tomography which means predicting the properties of a quantum system from randomized measurements. First, we apply a set of Unitary transformations and random measurements on copies of a n-qubit system to obtain the classical shadow data set. Then we get the prediction of the system using median-of-means protocol on these datasets*

The concept of "classical shadow" in quantum physics offers a novel approach to tackling the complexity of large-scale quantum systems. It bypasses the limitations of traditional prediction methods like quantum state tomography, which suffers from the exponential growth of parameters with system size. Shadow tomography enables efficient predictions of various quantum properties, requiring significantly fewer samples and resources. This technique, focusing on properties that are linear functions of the density matrix, involves applying unitary transformations and measurements to form a classical representation of the quantum state. The method is optimally efficient, adhering to quantum information theory's lower bounds, and is versatile enough to predict a wide range of quantum properties effectively.

A classical shadow is generated through a repetitive process involving unitary transformations $U\rho U^\dagger$ on a quantum state ρ, followed by measuring all qubits in the computational basis. The frequency of this process determines the shadow's size. The transformation U is randomly chosen from a set of unitaries, each set offering unique strengths and limitations. This method is designed to be implementable via efficient quantum circuits. Notably, random n-qubit Clifford circuits and products of single-qubit Clifford circuits are key examples, providing complementary advantages in practical applications. Figure 2.2 demonstrates the process and key properties we can predict from shadow tomography.

2.2.1 Procedure of Shadow Tomography

For example, if we focus on n-qubit quantum systems within a fixed state in $d = 2^n$ dimensions. To decipher the state, we apply a random unitary U from a predetermined set, measure in a computational basis(Pauli Basis or Clifford gate can be a complete set which has been proved), and store the outcome. Repeating this yields a classical snapshot of the quantum state, which, through post-processing, can predict various properties of the system. The 'classical shadow' consists of these snapshots, and their number N determines the prediction's accuracy. This process, involving median-of-means, is efficient and circumvents the need for full quantum descriptions.

We consider an n-qubit quantum system with state ρ in a 2^n-dimensional space. The extraction of information from ρ is performed via a series of measurements with random unitaries U from a predefined set, creating a classical snapshot of the quantum state. The classical description of each outcome is stored, and the mapping to a classical snapshot is treated as a quantum channel:

$$\mathbb{E}\left[U^\dagger|b\rangle\langle b|U\right] = M(\rho) \Rightarrow \rho = \mathbb{E}\left[M^{-1}\left(U^\dagger|b\rangle\langle b|U\right)\right]. \tag{2.12}$$

The quantum channel M arises from averaging over unitary transformations and measurement outcomes. Despite M^{-1} not being physically realizable, it is used classically to obtain snapshots:

$$\hat{\rho} = M^{-1}(U^{\dagger}|\hat{b}\rangle\langle\hat{b}|U), \tag{2.13}$$

yielding a classical snapshot from a single measurement. These snapshots collectively form the classical shadow:

$$S(\rho; N) = \{\hat{\rho}_1, \ldots, \hat{\rho}_N\}, \tag{2.14}$$

allowing for efficient property prediction. This framework is compatible with various measurement strategies, such as Clifford and Pauli-based measurements. For random n-qubit Clifford circuits, we need $n^2/log(n)$ entangling gates to sample from n-qubit Clifford unitaries[4], and the corresponding quantum channel is $M_n^{-1}(X) = (2^n + 1)X - I$. For random Pauli basis, which is easier to implement in various platforms, and the corresponding quantum channel is $M_P^{-1} = \bigotimes_{i=1}^{n} M_1^{-1}$.

[Shadow Tomography Theorem] Given an unknown D-dimensional quantum mixed state ρ, and a set of 2-outcome measurements $\{E_i\}_{i=1}^{M}$, shadow tomography enables us to estimate the probabilities $\mathrm{Tr}(E_i\rho)$ within an error margin ε, succeeding with probability $1 - \delta$. This can be achieved using[*]

$$k = \widetilde{O}\left(\frac{\log^{1/\delta}}{\varepsilon^4} \cdot \log^4 M \cdot \log D\right)$$

copies of ρ, with $\widetilde{O}$ encompassing polylogarithmic factors in M, D, and $1/\varepsilon$.

This theorem showcases the efficiency of shadow tomography, revealing that a logarithmic number of measurements relative to the system size is sufficient. It determines the computational complexity of predicting a quantum system's properties and establishes the feasibility of such predictions using a surprisingly small number of quantum state copies. This makes shadow tomography a potent tool in quantum computing, providing a pragmatic approach to the study of quantum systems.

2.2.2 Applications for Shadow Tomography

Quantum Fidelity Estimation

Classical shadows allow for efficient fidelity estimation between an experimental n-qubit state and a target state, improving upon traditional methods by requiring significantly fewer samples. This approach is scalable and can estimate multiple fidelities simultaneously.

[*] the provement can be seen in [5]

Entanglement Verification

Fidelity measurements can also act as entanglement witnesses[17]. Classical shadows enable simultaneous verification of multiple entanglement witnesses, providing an efficient method for confirming entanglement in bipartite states.

Predicting Expectation Values

Classical shadows are instrumental for calculating expectation values of local observables in quantum systems, particularly in near-term applications. They offer a highly efficient method for evaluating many-body Hamiltonians, as an alternative to the repetitive direct measurement approach. This has significant implications for fields such as quantum chemistry and lattice gauge theory[18, 19].

Classical shadows prove advantageous for local observables but encounter challenges when applied to global observables due to scaling issues. Consider a non-local observable in a spin chain, characterized by the Pauli expectation value:

$$\langle P_{i_1} \otimes \cdots \otimes P_{i_n} \rangle_\rho = \mathrm{tr}(O_1 \rho), \tag{2.15}$$

where the observables' Hilbert-Schmidt norm is given by:

$$\mathrm{tr}(O_1^2) = 2^n, \tag{2.16}$$

and the locality parameter k is equal to n. For such non-local observables, a classical shadow may require an exponentially large number of samples for accurate prediction. In contrast, direct measurements can achieve similar accuracy with only:

$$\mathcal{O}\left(\frac{1}{\varepsilon^2}\right) \tag{2.17}$$

copies of the state ρ, where ε denotes the desired precision.

2.3 Generative Models

To approximate the probability distributions associated with randomized measurements and their outcomes, we employ generative models. This is informed by the work "Explainable Representation Learning of Small Quantum States" [1], which leverages the encoder of β-VAE as a method for information scrambling to visualize the latent space of quantum entanglement. Complementarily, "Observing SchrÃ¶dingerâs Cat with Artificial Intelligence: Emergent Classicality from Information Bottleneck" [6] utilizes a transformer architecture to learn from classical data, subsequently reconstructing the quantum state from numerous random measurements. In our work, we specifically incorporate the β-VAE and Transformer as our chosen models.

2.3.1 VAE and β-VAE

A Variational Autoencoder (VAE) is a powerful generative model that has gained considerable attention in the field of deep learning and artificial intelligence. VAEs are particularly noted for their ability to generate new data instances that resemble a given dataset. They find applications in a wide range of tasks, including image generation, and anomaly detection, and as a tool for understanding and visualizing complex data distributions.

The foundational concept of a VAE was introduced in two seminal papers: "Auto-Encoding Variational Bayes" by Diederik P. Kingma and Max Welling[20], and "Stochastic Backpropagation and Approximate Inference in Deep Generative Models" by Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra[21]. Both papers were published in 2014 and laid the groundwork for the development and understanding of VAEs.

VAEs are a type of autoencoder, a neural network architecture used for unsupervised learning. However, unlike traditional autoencoders that aim to learn a fixed representation of the input data, VAEs introduce a probabilistic twist. They encode input data into a distribution in a latent space, rather than a single point. This process involves learning the parameters of the distribution - typically the mean and variance. The VAE then samples from this distribution to generate new data points.

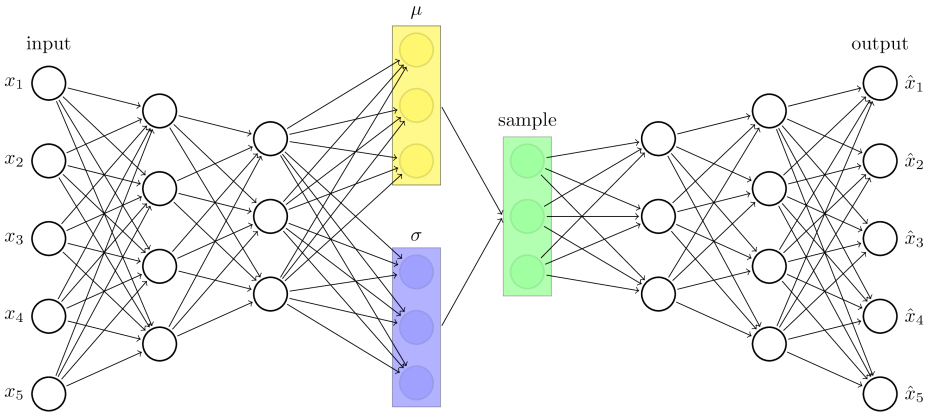


Figure 2.3: *A diagrammatic representation of a Variational Autoencoder (VAE). The left section illustrates the Encoder where the input image x is compressed into a lower-dimensional latent space Z characterized by the learned mean $\mu_{z|x}$ and standard deviation $\Sigma_{z|x}$. In the central section, the latent space Z serves as a bottleneck, encapsulating the essential information required for reconstruction. The right section depicts the Decoder, which reconstructs the original input into the output image $\hat{x}$ using the sampled latent variables. The entire process aims to minimize a composite loss function comprised of Reconstruction Loss and the Kullback-Leibler (KL) Divergence, ensuring a balance between accurate reconstruction and a well-formed latent distribution.*

The architecture of a VAE consists of two main components:

- **Encoder:** The encoder network transforms the input data into a latent space representation, approximating the posterior distribution of latent variables conditional on the input.
- **Decoder:** The decoder network reconstructs the input data from the sampled latent variables, aiming to capture the conditional probability of the data given the latent representation.

The training of a VAE involves optimizing the variational lower bound, or evidence lower bound (ELBO), which balances two aspects: the quality of the reconstruction (how well the output matches the input) and the regularity of the learned latent space (typically enforced through a Kullback−Leibler divergence term).

VAEs stand out for their ability to generate new samples that are both diverse and similar to the original data, making them particularly useful in fields like image generation where they can produce new, plausible images that do not appear in the training set.

The β-Variational Autoencoder (β-VAE), an extension of the Variational Autoencoder (VAE) proposed by Kingma and Welling [20], introduces a regularization hyperparameter, β, into the VAE framework [22]. This hyper parameter plays a pivotal role in learning disentangled representations in the latent space. Disentanglement in this context refers to the separation of the distinct, interpretable factors of variation in the data.

The architecture of a β-VAE is similar to that of a standard VAE, comprising an encoder and a decoder. The encoder in a β-VAE, parameterized by ϕ, maps the input data x to a latent space representation characterized by two parameters: the mean μ and the standard deviation σ, which are used to define a Gaussian distribution over the latent variables z:

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \mu_\phi(\mathbf{x}), \sigma_\phi(\mathbf{x})^2 I), \tag{2.18}$$

where $\mu_\phi(\mathbf{x})$ and $\sigma_\phi(\mathbf{x})$ are outputs from the encoder network, and I is the identity matrix. The specific expression for the Gaussian distribution of the latent space in a β-VAE is given by:

$$q_\phi(\mathbf{z}|\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k |\Sigma_\phi(\mathbf{x})|}} \exp\left(-\frac{1}{2}(\mathbf{z} - \mu_\phi(\mathbf{x}))^\top \Sigma_\phi(\mathbf{x})^{-1}(\mathbf{z} - \mu_\phi(\mathbf{x}))\right), \tag{2.19}$$

where $\mu_\phi(\mathbf{x})$ is the mean vector, $\Sigma_\phi(\mathbf{x}) = \sigma_\phi(\mathbf{x})^2 I$ is the covariance matrix, and k is the dimensionality of the latent space vector $\mathbf{z}$.

The decoder, parameterized by θ, then attempts to reconstruct the input data from the latent representation:

$$p_\theta(\mathbf{x}|\mathbf{z}) = \text{Decoder}_\theta(\mathbf{z}). \tag{2.20}$$

The loss function of the β-VAE, $\mathcal{L}(\mathbf{x}; \phi, \theta)$, comprises two terms: the reconstruction loss $\mathcal{L}_R$ and a regularization term $\mathcal{L}_{KL}$, the latter being weighted by the hyperparameter β:

$$\mathcal{L}(\mathbf{x}; \phi, \theta) = \mathcal{L}_R(\mathbf{x}; \phi, \theta) + \beta \cdot \mathcal{L}_{KL}(\mathbf{z}; \phi). \tag{2.21}$$

Here, $\mathcal{L}_R(\mathbf{x}; \phi, \theta)$ is typically a mean squared error or binary cross-entropy between the input $\mathbf{x}$ and its reconstruction. The term $\mathcal{L}_{KL}(\mathbf{z}; \phi)$ is the Kullback-Leibler divergence, given by:

$$\mathcal{L}_{KL}(\mathbf{z}; \phi) = D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})), \tag{2.22}$$

(2.23)

which measures the difference between the encoder's distribution $q_\phi(\mathbf{z}|\mathbf{x})$ and the prior distribution $p(\mathbf{z})$, typically assumed to be a standard Gaussian distribution $\mathcal{N}(0,1)$. When we assume both $q_\phi(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{z})$ are multivariate Gaussian distributions, the Kullback-Leibler divergence can be expressed in a closed form. The encoder's distribution is given by a multivariate Gaussian $\mathcal{N}(\mu_\phi(\mathbf{x}), \sigma_\phi(\mathbf{x})^2 I)$, and the prior distribution is often chosen as a standard multivariate Gaussian $\mathcal{N}(0, I)$.

Given these distributions, the KL divergence is calculated using the formula:

$$D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})) = \frac{1}{2}\left(\mathrm{tr}(\sigma_\phi(\mathbf{x})^2) + \mu_\phi(\mathbf{x})^\top \mu_\phi(\mathbf{x}) - k - \log\left|\sigma_\phi(\mathbf{x})^2 I\right|\right), \tag{2.24}$$

where tr denotes the trace of a matrix, $\top$ indicates the transpose, k is the dimensionality of the latent variable space, and $|\cdot|$ denotes the determinant of the matrix. This expression quantifies the difference between the encoder's learned distribution of the latent variables $\mathbf{z}$ given $\mathbf{x}$, and the prior distribution of $\mathbf{z}$. This divergence acts as a regularizer, encouraging the encoder to learn distributions of latent variables that are close to the prior, which promotes the learning of independent and interpretable latent representations.

2.3.2 Transformer

The Transformer architecture, introduced by Vaswani et al. [23], forms the structural foundation of our model. Renowned for its self-attention mechanism, the Transformer allows for differential weighting of input data elements, enhancing the model's interpretative power. In contrast to conventional sequence-to-sequence models that process data sequentially, our implementation of the Transformer processes inputs in parallel, leading to substantial gains in computational efficiency.

As illustrated in Figure 2.4, our model's encoder captures the input sequence into a series of continuous representations. These are then utilized by the decoder to synthesize an output sequence. Both the encoder and decoder consist of a series of N identical layers, each designed to maintain the integrity of sequential data.

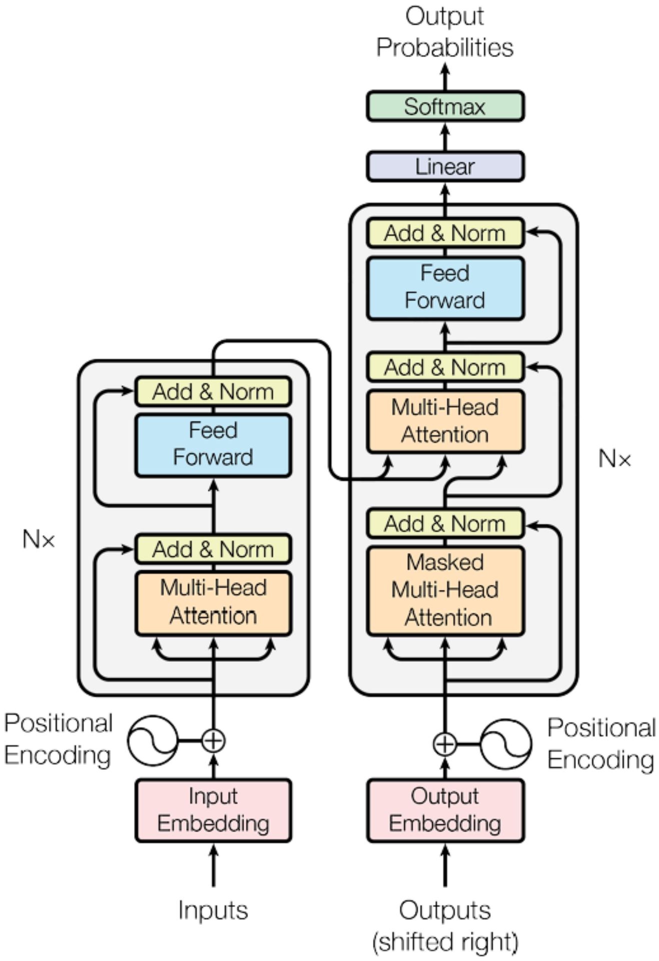


Figure 2.4: *The Transformer architecture as implemented in our model. The encoder processes the input sequence in parallel, while the decoder synthesizes the output sequence, both employing self-attention and feedforward layers.*

In our design, each encoder layer, instantiated by the `EncoderLayer` class, integrates a multi-head self-attention mechanism with a position-wise fully connected feed-forward network. This configuration enables the simultaneous consideration of the entire input sequence, thereby preserving the contextual relationships inherent within the data. The `DecoderLayer` mirrors this approach while additionally incorporating cross-attention with the encoder's outputs, thus harmonizing the information flow between input and output sequences.

To ensure stability during training, we incorporate residual connections followed by layer normalization within each layer, a strategy realized through the `torch.nn.LayerNorm` module. This approach not only aids in training deeper networks but also enhances the model's ability to generalize from training data.

The optimization of our model is centred around a composite loss function, characteristic of the Transformer-based VAE. This function amalgamates a reconstruction loss component with a Kullback−Leibler divergence term, fostering a latent space that adheres to a predefined distribution while also promoting accurate reconstruction of the input sequence. The `Transformer` class encapsulates this optimization process, with the `Randomizer` module effectuating the reparameterization step, thereby imposing an information bottleneck vital for the extraction of salient features.

Through the integration of the Transformer architecture within the VAE framework, our model emerges as a potent tool for generative modelling, adept at learning and replicating complex data patterns. To have a deeper understanding of the model we use for the whole work, the hyper-parameters which are possible to be tuned are listed in the following table: 2.1

Hyperparameter	Description
`vocab_size`	Number of tokens in the vocabulary
`outtk_size`	Number of tokens valid for output
`n_tokens_max`	Maximum number of tokens
`n_layers`	Number of layers in the transformer
`embed_dim`	Dimensionality of the embeddings
`num_heads`	Number of attention heads
`dropout`	Probability of dropout after applying the MLP
`symmetric`	Boolean to respect permutation symmetry

Table 2.1: *Transformers' Hyper-parameters in implementation of Classical Shadow Transformer*

Chapter 3

Implementing the Classical shadow Transformer

To adhere to the principles of Shadow Tomography and integrate them with Machine Learning methodologies, we use the generative model termed the "Classical Shadow Transformer." This model is specifically designed to evaluate the efficacy of reconstructing quantum states from classical data, thereby establishing a benchmark for the capacity of artificial intelligence to understand and interpret quantum phenomena. Our training regimen incorporates three distinct yet archetypal quantum states: the GHZ state, the W state, and a separable state (All-Zero state), to probe the intricate relationship between entangled and non-entangled states, with a particular focus on the variances between maximally entangled states (W-like and GHZ-like).

Our research is driven by several pivotal inquiries:

- How proficiently can an artificial intelligence agent assimilate knowledge from a quantum environment and effectively translate this ’quantum language’ into a comprehensible format within the constraints of information processing limitations?

- What delineates the threshold between the quantum and classical realms, particularly in the context of system size and information processing capabilities?

- From the vantage point of language modelling, what are the discernible distinctions between entangled and non-entangled states, and furthermore, what are the specific differences between various forms of maximal entanglement (W-like versus GHZ-like)?

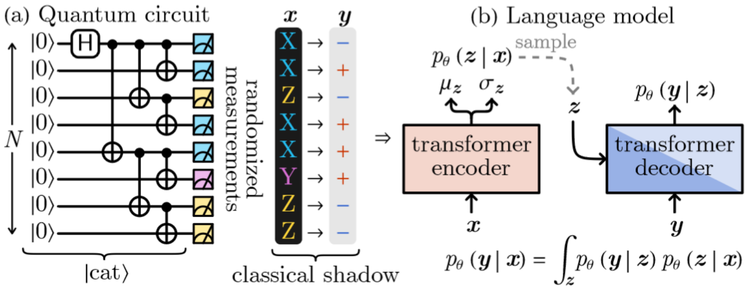


Figure 3.1: *A schema for the classical shadow. The quantum Circuits were built up to prepare the quantum state(we select GHZ-state, W-state and a non-entangled state(ALL zero-state). Randomized Pauli Measurements was applied to the pre-prepared stated to generate the shadow data-set (x, y). Then the generative model, classical shadow transformer(transformer-based β -VAE architecture) is trained to learn the distribution of $p(y|x)$. As a result, we follow the specific quantum channel M for Random Pauli Measurement to reconstruct the original state and extract related quantum information from classical data. [6]*

These questions aim to deepen our understanding of the interplay between quantum mechanics and artificial intelligence, shedding light on the potential and limitations of AI in interpreting and utilizing quantum information.

3.1 State Preparation and Measurement Simulation

The working mechanism of the Classical Shadow Transformer (CST) is elaborated in Figure3.1. The initial phase of this procedure involves the construction of a quantum circuit to prepare three specific types of quantum states: the GHZ state, the W state, and the Zero state. For the implementation of our experiment, we employ the **qiskit** package, a renowned quantum computing software development framework. The detailed circuit structures for each state are outlined as follows:

3.1.1 Zero-state and GHZ state

The Zero state, or a separable state, is the simplest among the three. This circuit typically involves initializing all qubits in their ground state $|0\rangle$ and does not require entangling operations.

$$|Zero\rangle = |0\rangle^{\otimes N}$$

The Greenberger-Horne-Zeilinger (GHZ) state represents a prominent example of a multi-partite entangled state in quantum computing and quantum information theory. It is particularly notable for its applications in tests of quantum nonlocality and quantum teleportation. A GHZ state of N qubits is defined as:

$$|\mathrm{GHZ}\rangle = \frac{1}{\sqrt{2}}\left(|0\rangle^{\otimes N} + |1\rangle^{\otimes N}\right), \tag{3.1}$$

where $|0\rangle^{\otimes N}$ and $|1\rangle^{\otimes N}$ denote the tensor product of N qubits all in the state $|0\rangle$ or $|1\rangle$, respectively. This state exhibits maximal entanglement among the qubits, making it a subject of extensive study in the field of quantum mechanics [24].

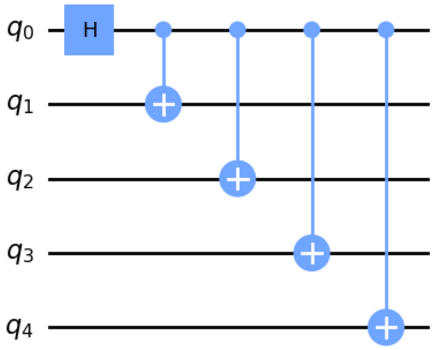


Figure 3.2: *An example diagram for a 5-qubit GHZ-state Circuit*

Creating a GHZ (Greenberger-Horne-Zeilinger) state in a quantum circuit is a fundamental operation in quantum computing, crucial for demonstrating quantum entanglement and quantum teleportation. The procedure begins with initializing qubits in the ground state, $|0\rangle$. A Hadamard gate is then applied to the first qubit to generate a superposition of $|0\rangle$ and $|1\rangle$. This is followed by applying Controlled-NOT (CNOT) gates between the first qubit and the subsequent ones. For instance, in a five-qubit system, a CNOT gate is applied from the first to the other 4 qubit, as the Figure 3.2shows. These operations entangle the qubits, resulting in the

GHZ state, $\frac{1}{\sqrt{2}}(|000\rangle + |111\rangle)$, which is a superposition of all qubits being either in the 0 or 1 state. This state exemplifies the essence of quantum entanglement.

3.1.2 W-state

The W state is another important class of entangled states in quantum computing, distinct from the GHZ state. It is particularly notable for its robustness against qubit loss, retaining entanglement even when a part of the system is discarded. A W state for a system of N qubits is defined as:

$$|\mathrm{W}\rangle = \frac{1}{\sqrt{N}}(|100\ldots0\rangle + |010\ldots0\rangle + \ldots + |000\ldots1\rangle), \tag{3.2}$$

where each term in the superposition involves exactly one qubit in the state $|1\rangle$ and the rest in the state $|0\rangle$. This state is a symmetric superposition of all possible states with exactly one qubit in the $|1\rangle$ state and the rest in the $|0\rangle$ state, making it fundamentally different from the GHZ state. The W state has been studied extensively for its applications in quantum information processes [25].

The circuit designed to generate a n-qubit W-state is more complex than the GHZ state circuit, involving a combination of Hadamard gates, CNOT gates, and other quantum gates to achieve the required superposition and entanglement. The paper "Deterministic construction of arbitrary W states with quadratically increasing number of two-qubit gates" [26] presents a method for creating W states in a quantum system. The construction of W states is achieved through quantum circuits that use a specific number of two-qubit gates, with the number of gates increasing quadratically with the number of qubits in the system.

To implement the algorithm of constructing a quantum circuit that can generate a n-qubit W-state, referenced by the paper by Firat Diker in 2016 [26]. We First need to define a F Gate.

The F Gate is a kind of control Operation which involves a control qubit q_k and target qubit q_{k+1}, here k represents the specific wire the qubit is located and we can denote such an operation q_k apply to q_{k+1} as $F_{k,k+1}$. The equivalent combined Operation of F can be regarded as the following:

- First a rotation round Y-axis $R_y(-\theta_k)$ applied on q_{k+1}
- Then a controlled Z-gate cZ in any direction between the two qubits q_k and q_{k+1}
- Finally a rotation round Y-axis $R_y(\theta_k)$ applied on q_{k+1}

The matrix representations of a $R_y(\theta)$ is :

$$R_y(\theta) = \begin{pmatrix} \cos(\frac{\theta}{2}) & -\sin(\frac{\theta}{2}) \\ \sin(\frac{\theta}{2}) & \cos(\frac{\theta}{2}) \end{pmatrix} \tag{3.3}$$

And CZ (control-Z) gate is given as:

$$CZ = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

The value of θ_k depends on n and k following the relationship:

$$\theta_k = \arccos\left(\sqrt{\frac{1}{n-k+1}}\right)$$

where n represents the number of qubits for the system size and k represents the order of the wire since the lowest significant bit (LSB). To have a better understanding, the F gate has the same function as the circuit in Figure3.3

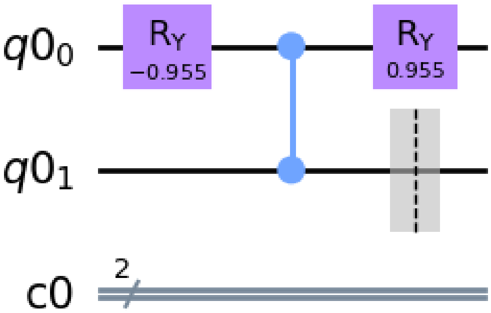


Figure 3.3: *equvivalent diagram for a F_{01} Gate for 2-qubit W-state*

We have already defined a $F_{k,k+1}$ gate for an n-qubit circuit between the wires k and $k+1$. The procedure to obtain the circuit for an n-qubit W-state is as follows:

1. Initially, the qubits are set in the state $|\varphi_0\rangle = |10\ldots0\rangle$.

2. This is followed by the application of $n-1$ successive F gates:

$$|\varphi_1\rangle = F_{n-1,\,n} \,...\, F_{k,\,k+1} \,...\, F_{2,\,3}\, F_{1,\,2}\, |\varphi_0\rangle \;=\; \sqrt{\frac{1}{n}}\,(\,|10...0\rangle + |11...0\rangle + ... + |11...1\rangle\,) \tag{3.4}$$

3. Then, $n-1$ CNOT gates are applied, leading to the final circuit:

$$|W_n\rangle = \text{CNOT}_{n,n-1}\text{CNOT}_{n-1,n-2}\ldots\text{CNOT}_{k,k-1}\ldots\text{CNOT}_{2,1}|\varphi_1\rangle$$

Takes a circuit for a 4-qubit W-state for instance, follow the rules we introduce, just as Figure 3.4 shows. It is composed of 3 F gates and 5 CNOT gates. The entire circuit corresponds to:

$$|W_4\rangle \;=\; cNOT_{4,3}\; cNOT_{3,2}\; cNOT_{2,1}\; F_{3,4}\; F_{2,3}\; F_{1,2}\; |\varphi_0\rangle$$

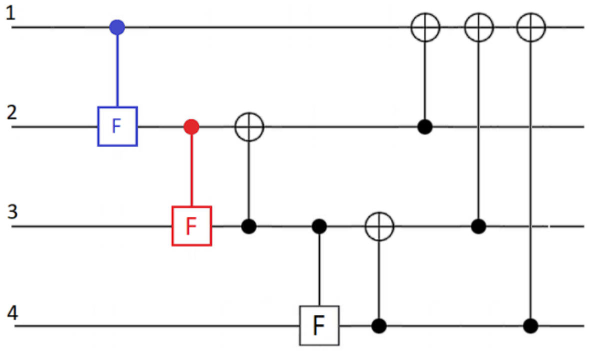


Figure 3.4: *The circuit used for creating a four-qubit W state.*

Figure 3.5 shows a qiskit version of visualization of the same circuit only in the Rotation Y(R_y)gate, Control Z(CZ) gate and Hadamard (H) gate

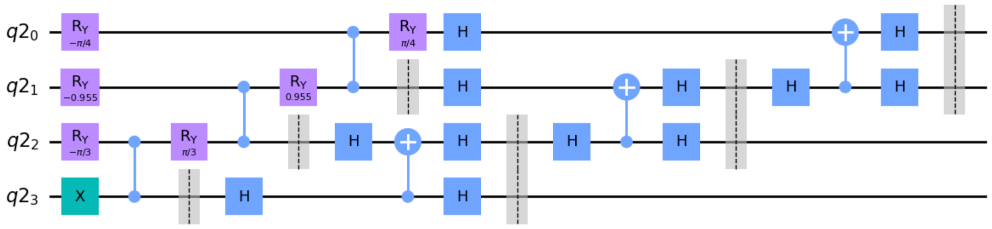


Figure 3.5: *qiskit visualization for a four-qubit W state circuit*

3.1.3 Measurement

In a quantum circuit, measurements in the Pauli bases can be performed as follows:

- **Pauli X Basis:** Apply a Hadamard gate H before measurement:
$$\text{Measure in X} = H$$
- **Pauli Y Basis:** Apply a sequence of $S^\dagger$ (S-dagger) and H gates:
$$\text{Measure in Y} = S^\dagger H$$
- **Pauli Z Basis:** Measure directly in the standard computational basis (Z basis):
$$\text{Measure in Z}$$

These operations transform the state to the Z basis, where a standard measurement can be performed. Figure3.6 shows the diagram of a quantum circuit for Pauli XYZ measurement

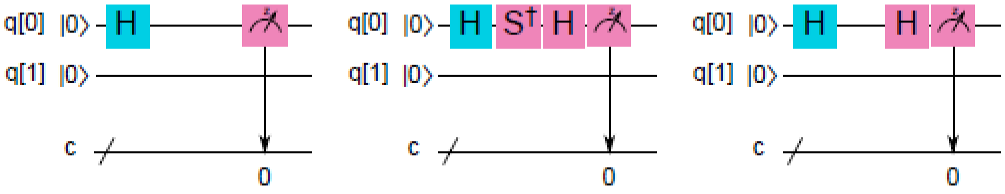


Figure 3.6: *Quantum circuit used to transform Pauli Basis to Z basis.The order from left to right is corresponding to Pauli X,Y,Z. The plot is cited from "Quantum Algorithm Implementations for Beginners"[27]*

3.2 Classical Shadow Data Structure

After we can generate the quantum circuit for corresponding state and their measurements, we now have the ability to get the data-set (x, y), where x represents the Pauli list of observables $x \in \{X, Y, Z\}^{\times \mathbb{N}}$ and y represents the measurement outcomes of corresponding measurement for each qubit $y \in \{\pm 1\}^{\times \mathbb{N}}$.

First, we need to derive the probability distribution of outcomes y when observable x is given, denoted as $P(y|X)$. According to the reference [28], the quantum-mechanical state of any two-level system can be expressed as a 2×2 density matrix:

$$\hat{\rho} = \frac{1}{2}\left(1 + x\hat{\sigma}_x + y\hat{\sigma}_y + z\hat{\sigma}_z\right) = \frac{1}{2}\left(1 + \mathbf{r}\cdot\hat{\boldsymbol{\sigma}}\right) \tag{3.5}$$

where $\mathbf{r} = (x, y, z) \in \mathbb{R}^3$ and $\hat{\sigma} = (\hat{\sigma}_x, \hat{\sigma}_y, \hat{\sigma}_z)$. The Bloch vector only represents a physical state if $\mathbf{r}^2 < 1$ and it represents a pure state when $\mathbf{r}^2 = 1$. Consider the relationship of Kronecker product for Pauli operators:

$$\mathrm{tr}(\sigma_\alpha) = 2\delta_{0\alpha}$$

$$\mathrm{tr}(\sigma_\alpha \sigma_\beta) = 2\delta_{\alpha\beta}$$

The identity matrix will have a trace of 2 while the identity has a trace of 0, the product of any two different Pauli matrices is also 0, and the product of any Pauli matrix with itself is 2. Thus, we can derive that:

$$\begin{aligned}
\langle \sigma_x \rangle &= \mathrm{tr}\left(\frac{1}{2}(I + r_x\sigma_x + r_y\sigma_y + r_z\sigma_z)\sigma_x\right) \\
&= \frac{1}{2}\mathrm{tr}(I\sigma_x + r_x\sigma_x\sigma_x + r_y\sigma_y\sigma_x + r_z\sigma_z\sigma_x) \\
&= \frac{1}{2}\left(\mathrm{tr}(I\sigma_x) + r_x\mathrm{tr}(\sigma_x\sigma_x) + r_y\mathrm{tr}(\sigma_y\sigma_x) + r_z\mathrm{tr}(\sigma_z\sigma_x)\right) \\
&= \frac{1}{2}(r_x \cdot 2) = r_x
\end{aligned}$$

According to Born's rules, the probability that the result of the measurement lies in a measurable set M for a certain projection-valued measure Q is given by $\langle\psi|Q(M)|\psi\rangle$. Thus, we can derive that the probability distribution $P(y|x)$ for the outcomes of a certain Pauli measurement $x \in \{X, Y, Z\}$ is given by:

$$p(y|x) = \left\langle \psi \middle| \bigotimes_i \left(\frac{1 + y_i\hat{x}_i}{2}\right) \middle| \psi \right\rangle \tag{3.6}$$

We've designed a classical simulator capable of generating pairs of sequences (x, y) according to the probability distribution $P(y|x)$ when prompted. This essentially replicates the cyclical procedure of generating the quantum state, allowing it to undergo decoherence, and then recording the classical data it imparts to the surroundings. Instances of (x, y) pairs are accessible in the Supplementary Information. These pairs, termed the classical shadows of the quantum state, represent stochastic projections of the quantum state on a random measurement basis, similar to how a three-dimensional object casts a shadow onto a two-dimensional plane. The method of classical shadow tomography provides a structured approach for inferring the quantum state from its classical shadows through classical post-processing. Given the random Pauli measurement framework described earlier[29][30], the state reconstruction is denoted by:

$$\rho_\psi = \mathbb{E}_{(x,y)\sim p_{\text{dat}}} \bigotimes_i \left(\frac{1 + 3y_i x_i}{2} \right). \tag{3.7}$$

where $p_{\text{dat}}(x, y)$ represent the certain probability a (x,y) pair may occur. this equation illustrates that with ample classical information regarding multiple instances of a quantum state, one can theoretically reconstruct the complete quantum entity with precision.

3.3 Train The Generative Model

The first 2 sections above introduce how a classical shadow tomography works. First, we need to initialize the system to prepare the required state and randomly sample sufficient Paulilist as our measurements to get the classical data pair (x,y). After that, with the equation3.6, and equation 3.7. We can reconstruct the original quantum state by purely classical data (x,y) and underlying distribution $p_{\text{dat}}(x, y)$. To integrate shadow tomography with artificial intelligence, our goal is to train a generative model that approximates the distribution $p_{\text{dat}}(x, y)$. This approach aims to extract relevant information from the quantum state, a process we refer to as the 'Classical Shadow Transformer'."

In formulating the probabilistic model $p_{\text{mdl}}(x, y) = p_\theta(y|x)p(x)$, we concentrate on the conditional distribution $p(y|x)$ parameterized by θ. Given that $p(x) = 3^{-N}$ is a simple uniform distribution, it necessitates no further modelling. Envisioning the observable sequence x as a question, and the measurement outcome sequence y as its answer from a quantum experiment, the modelling of $p(y|x)$ can be likened to a conversation completion task in natural language processing. This analogy indicates that a generative language model would be an intuitive approach. Once adequately trained, the language model is capable of replacing the quantum experiment to provide responses about the quantum state $|\psi\rangle$, effectively "speaking" the language of quantum mechanics. The learning process emulates how an intelligent agent acquires knowledge by observing its surroundings. The transformer architecture, introduced in Chapter 2, is notably suited for modelling $p(y|x)$, with a slight modification in its latent space to include a variational information bottleneck, derived from the β-VAE model, which enables tuning of the model's information processing capacity for our subsequent research.

Our Classical Shadow Transformer consists of two probabilistic models: an encoder $p_\theta(\mathbf{z}|\mathbf{x})$ that deduces latent variables $\mathbf{z}$ from the input sequence $\mathbf{x}$, and a decoder $p_\theta(\mathbf{y}|\mathbf{z})$ that constructs the output sequence $\mathbf{y}$ predicated on $\mathbf{z}$. Hence, we define

$$p_\theta(\mathbf{y}|\mathbf{x}) = \sum_{\mathbf{z}} p_\theta(\mathbf{y}|\mathbf{z}) p_\theta(\mathbf{z}|\mathbf{x}). \tag{3.8}$$

A thorough explication of the architecture is available in Chapter 2.3. The objective is to converge $p(\mathbf{y}|\mathbf{x})$ in Eq3.8 through the optimization of the model parameters θ. Training the model involves minimizing the β-VAE loss function

$$\mathcal{L} = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{\text{dat}}}[\mathcal{L}(\mathbf{x},\mathbf{y})] \tag{3.9}$$

on the training dataset composed of classical shadows from the quantum state, where the loss function for each shadow $(\mathbf{x}, \mathbf{y})$ is

$$\mathcal{L}(\mathbf{x},\mathbf{y}) = -\mathbb{E}_{\mathbf{z}\sim p_\theta(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{y}|\mathbf{z})] + \beta D_{KL}[p_\theta(\mathbf{z}|\mathbf{x}) \| p_N(\mathbf{z})]. \tag{3.10}$$

The first term corresponds to the negative log-likelihood, and the second term is a Kullback-Leibler (KL) divergence regularization, with $p_N(\mathbf{z})$ symbolizing the normal distribution. The hyperparameter β allows for the adjustment of the variational bottleneck within the transformer. A larger β coerces $p_\theta(\mathbf{z}|\mathbf{x})$ to resemble $p_N(\mathbf{z})$ irrespective of $\mathbf{x}$, thereby constraining the model's aptitude to encode details about $\mathbf{x}$ within the latent variables $\mathbf{z}$. Thus, an increment in β enforces a more robust information bottleneck, concomitantly reducing the model's information processing capability.

In this chapter, we introduce how to build up our machine learning model "classical shadow transformer" and why we select certain tricks and architecture. In the following chapter, we will test how well it works and compare the model performance for different quantum states(GHZ, W, Separable) to analyse different extents of entanglement from the perspective of states' Learnability

Chapter 4

Representation Learning for Quantum Entanglement

In the preceding chapter, we laid the groundwork for implementing the Classical Shadow Transformer by delving into essential concepts of quantum information. This foundation included a detailed examination of the generative architecture employed in our approach. We paid special attention to the process of reconstructing a density matrix via Pauli measurements, a critical aspect of shadow tomography. Furthermore, we explored the interaction between the Classical Shadow dataset and the transformer, highlighting how these elements integrate to achieve our objectives in quantum information processing. In this chapter, we delve into the experimental results obtained for three distinct quantum states: GHZ, W, and Zero. These states are chosen for their varying extents and types of entanglement, providing a comprehensive basis for analysis. Our primary objective is to evaluate the effectiveness of our reconstruction method. This assessment considers various constraints, including the limitations imposed by the system size, the information bottleneck, and the structural design of our model.

We will conduct a comparative analysis of the results to uncover specific learnability attributes associated with each of these quantum states. This comparison aims to identify any unique or shared properties that emerge from the learning process applied to these states.

Towards the end of the chapter, we engage in a thorough discussion about the observed differences among these states. Our goal is to offer plausible explanations for these variations, thereby contributing to a deeper understanding of their underlying characteristics.

This exploration is pivotal as it attempts to unveil the secrets of the quantum realm through classical data and measurements. By interpreting these findings through the lens of what we term a "quantum large language model," we endeavour to decipher the complex language of quantum mechanics using classical methodologies. This approach not only bridges the gap between classical and quantum information theory but also paves the way for future advancements in quantum computing and information processing.

4.1 Model performance

We prepare three distinct quantum states: GHZ, W, and the Zero state. For each of these states, a dedicated Classical Shadow Transformer is trained. The training dataset consists of 400 groups of classical shadow data(x, y), each tailored for this purpose.

Each dataset comprises two main components:

1. **PauliList**: This serves as the measurement basis$x \in \{X, Y, Z\}^{\times \mathbb{N}}$, containing 200 different Pauli operators, essential for acquiring a diverse range of measurement outcomes.

2. **Measurement Outcomes**: For each Pauli operator, the measurement outcomes$y \in \{\pm 1\}^{\times \mathbb{N}}$ for a specific qubit are recorded.

The training process is designed to be comprehensive and adaptive, taking into account:

- The system size n, denoted by the number of qubits$(0 < n < 6)$.
- Different levels of information processing strength, parameterized by β in the encoder of our model$(-6 < \log \beta < 6)$.
- The dimensionality of the latent space within the Classical Shadow Transformer, adjusted as a key architectural parameter ($dim \in [128, 64, 32, 16]$).

For each combination of system size, β level, and latent space dimension, the model is trained separately. This approach allows for a precise assessment of the impact of each variable on the model's performance and the optimization of the model architecture for each specific scenario. We record the loss history for each model, and we firstly take an over look at the learning curve for 3 states as Figure 4.1 and Figure 4.2

Observing the loss history from the Classical Shadow Transformer, we note that upon convergence, the final loss values are ordered as $L_W >$

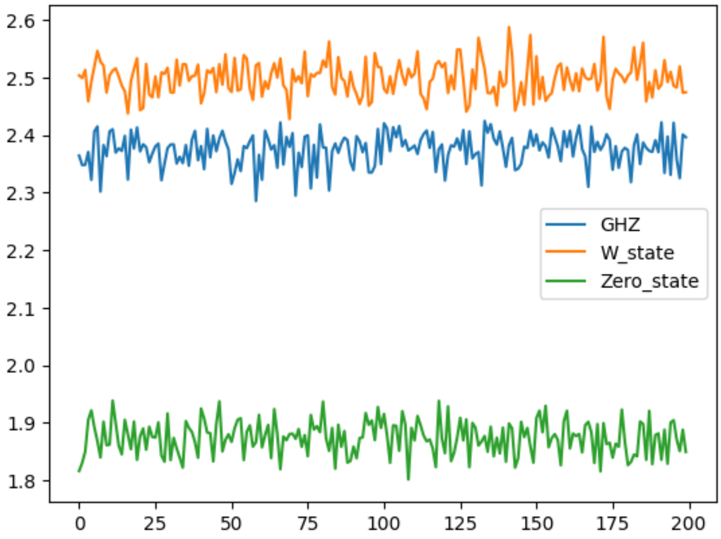


Figure 4.1: *The loss history for 3 states in the last 200 training epoch, the model parameters: n-qubits = 4,* $\log \beta = -6$*, dim for latent space = 64*

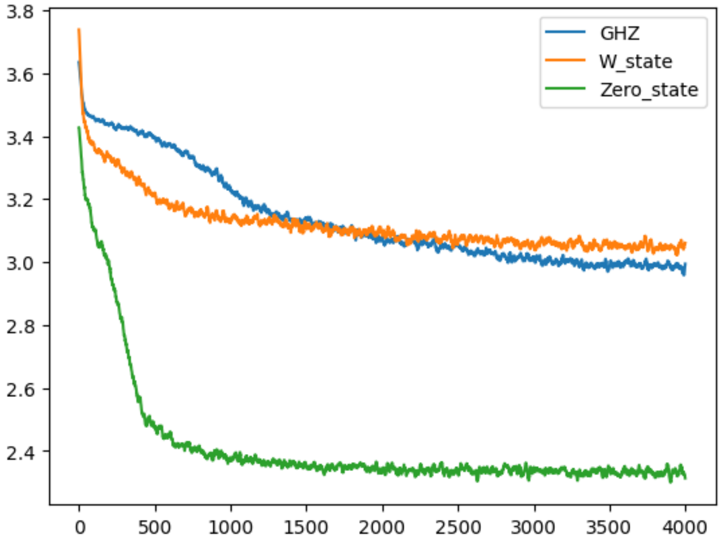


Figure 4.2: *The loss history for 3 states in all 4000 training epochs, the model parameters: n-qubits = 4,*$\log \beta = -6$ *dim for latent space = 64, we apply the Moving Average Filter to the curve and* $window\_size = 15$

$L_{ghz} > L_{Zero}$. This suggests that the transformer achieves better learning performance with the non-entangled Zero state than with the entangled GHZ and W states. Furthermore, it appears that the model finds the W state more challenging to learn and reconstruct compared to the GHZ state.

Besides, we can compare the learning rates of the three quantum states: GHZ, W, and Zero state. In the early stages of training, the loss for all states drops sharply, indicating that the model is rapidly learning and gaining an initial understanding of each quantum state. Specifically, in Figure 4.2, the GHZ and Zero states exhibit a quicker decline in loss, showing a higher learning rate, while the W state's loss decreases more slowly, suggesting that the model may require more time to learn this state. In the later stages of training, the losses for all three states stabilize, but the loss for the W state remains relatively higher, indicating that despite the slowdown in learning rate, the model's final performance on learning the W state is not as effective as for the other two states.

To have a better visualization of the model's performance. We can also draw the phase diagram for the three states in terms of Quantum fidelity (Figure4.3) and Von Neumann entropy(Figure4.4), they are defined as follows:

Quantum Fidelity (F) between two quantum states represented by density matrices ρ and σ is defined as:

$$F(\rho,\sigma) = \left(\mathrm{Tr}\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}\right)^2 = \langle\psi|\rho|\psi\rangle$$

ρ here stands for the density matrix reconstructed by our CST(classical shadow transformer) and ψ is the state vector for the original state. The F ranges from 0 to 1, with 1 indicating identical states and 0 indicating orthogonal states.

Von Neumann Entropy (S) for a quantum state with density matrix ρ is defined as:

$$S(\rho) = -\mathrm{Tr}(\rho \log_2 \rho)$$

This entropy measures the disorder or uncertainty of the quantum state, with higher values indicating more disorder and potentially greater entanglement.

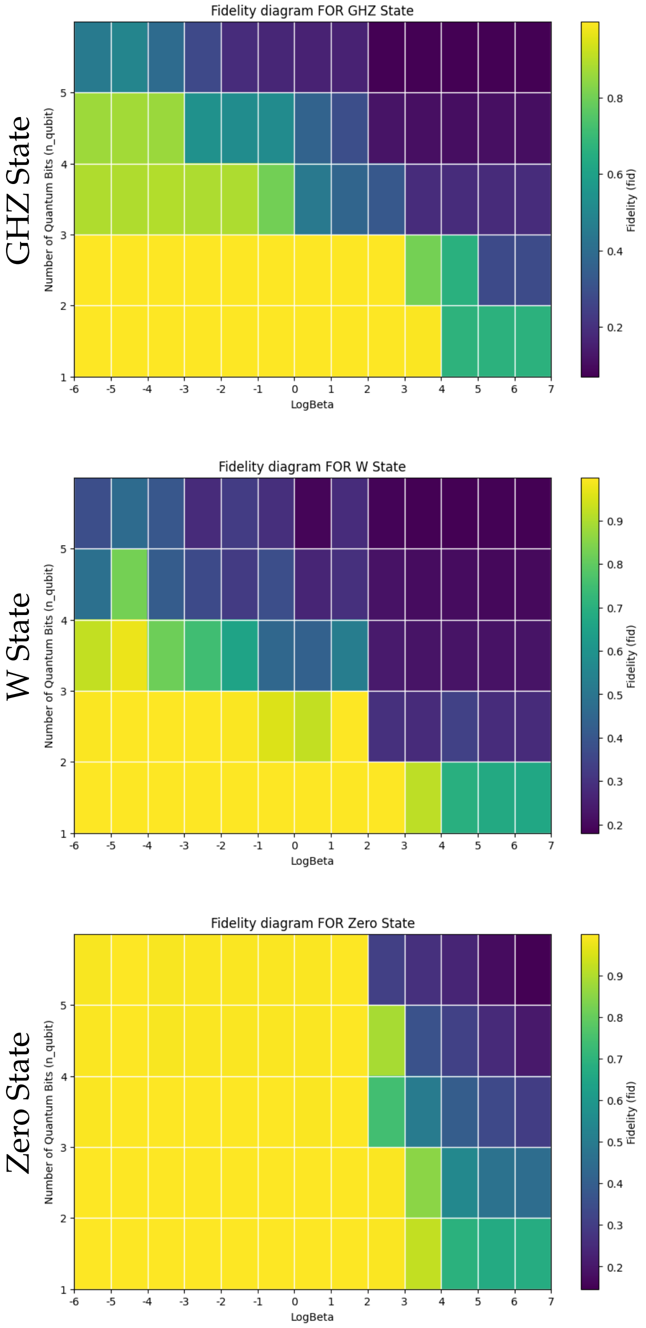


Figure 4.3: *Dependence of Quantum Fidelity on Information Bottleneck and System Size for GHZ, W, and Zero States. The x-axis quantifies the information bottleneck strength with* $\log_2 \beta$ *ranging from -6 to 6, where a higher value indicates a more pronounced limitation on information processing. The y-axis corresponds to the number of qubits,* n*, with the CST trained on systems such that* $0 < n < 6$*. Throughout, the generative model's dimension is held constant at* $dim = 64$*.*

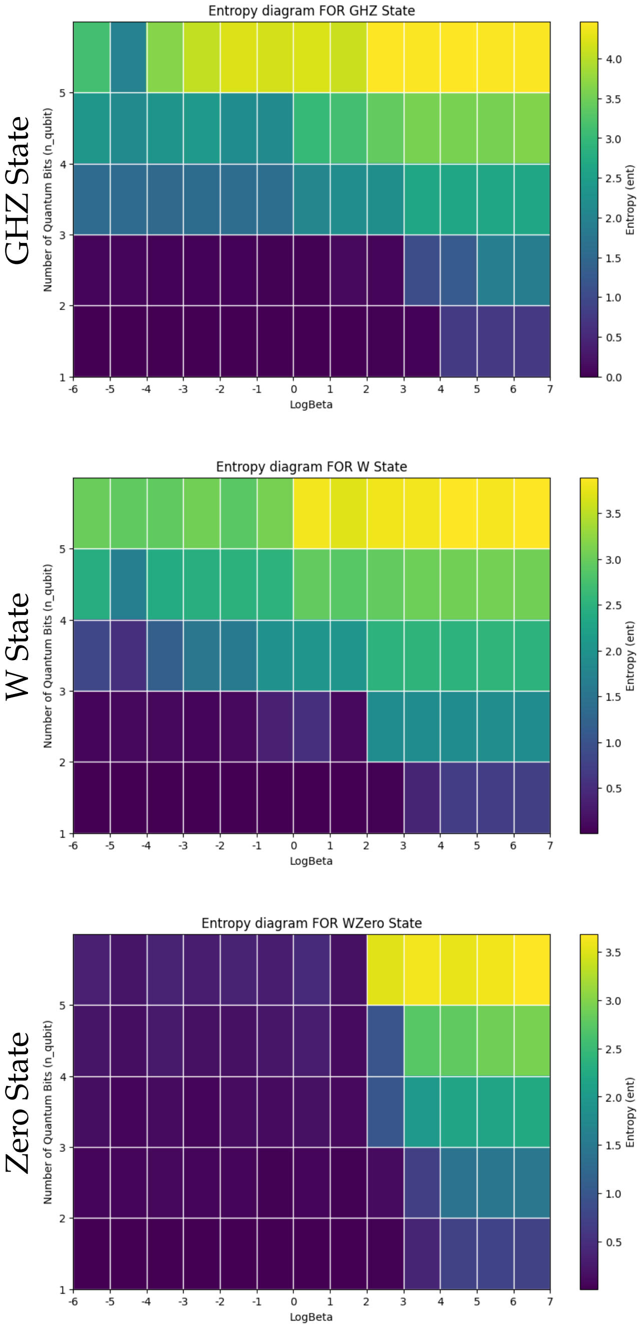


Figure 4.4: *Dependence of Von Neumann Entropy on Information Bottleneck and System Size for different states. The x-axis quantifies the information bottleneck strength with* $\log_2 \beta$ *ranging from -6 to 6, where a higher value indicates a more pronounced limitation on information processing. The y-axis corresponds to the number of qubits,* n*, with the CST trained on systems such that* $0 < n < 6$*. Throughout, the generative model's dimensionality is held constant at* $dim = 64$*.*

Firstly, from a global perspective, an overarching trend can be observed across the GHZ, W, and Zero states: as the system size increases and the information processing capability decreases, there is a general decline in fidelity and an increase in entropy. This pattern can be interpreted as a gradual loss of our decoding capabilities of the quantum world. This observation might provide an insight that, although the principles of quantum mechanics are typically observable at the microscopic level and elusive at the macroscopic scale, the limitations are not solely due to the scale of observation. Our capacity to decode information also plays a crucial role. Both factors contribute to the limitations of our understanding of the language of the quantum world. It may be posited that if we possess sufficiently advanced quantum information processing and decoding capabilities, we could gain a better understanding of quantum phenomena on a larger scale, potentially blurring the boundary between quantum and classical realms.

Furthermore, a horizontal comparison of the measurement results for the three states reveals that the Zero state exhibits the best reconstruction outcomes. For most conditions where $\log_2 \beta < 2$, regardless of system size, the Zero state maintains a fidelity close to 1 and a Von Neumann entropy approaching 0. This indicates that our Classical Shadow Transformer (CST) has nearly perfectly learned the characteristics of the Zero state, fully decoding its quantum information. However, for the GHZ and W states, we observe a heightened sensitivity to both the system size and the information bottleneck, making them more susceptible to these factors' limitations. These findings are consistent with those observed in the loss history, which may be attributable to the more complex information structures of entangled states compared to non-entangled states. Indeed, entangled states exhibit more complex and unpredictable probability distributions in measurement outcomes. Consequently, the degree of entanglement appears to correlate positively with the difficulty of performing state tomography via CST in terms of learnability. Beyond the distinction between entangled and non-entangled states, we are more concerned with the differences between various types of entangled states, such as W and GHZ states. In the following section, we will attempt to discuss the differences between these two so-called maximally entangled states and endeavour to provide explanations. Furthermore, we will explore the data structure within the CST's latent space to discern what characteristics can be revealed between different types of quantum states and what conclusions can be drawn from these observations.

4.2 Representation learning and latent dimension analysis

If we look more thoroughly at the phase diagram of Figure4.3 and 4.4. We can observe subtle yet discernible distinctions in the phase diagrams between the W and GHZ states. Specifically, as the system size increases (n > 2), the W state consistently exhibits lower fidelity and higher entropy across various values of β compared to the GHZ state, indicating a greater discrepancy between the reconstructed state and the true state. This suggests, to a certain degree, that the W state presents a higher learning difficulty and greater informational complexity than the GHZ state. However, here we ignore one important hyper-parameter, the dimension of the latent space of the transformer. To have a further understanding, we also take a look of the model's performance in terms of Fidelity and Entropy under the effect of different dimensions of latent space as Figure 4.5 and Figure4.6 show

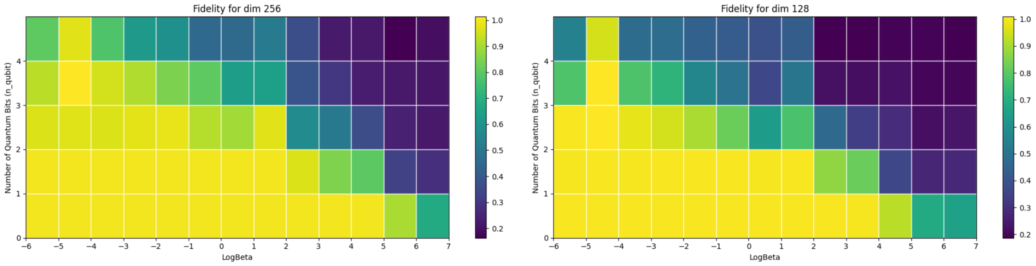


Figure 4.5: *Fidelity heatmap for reconstructed W-state from CST in higher latent dimension(the left one is 256 and the right one is 128)*

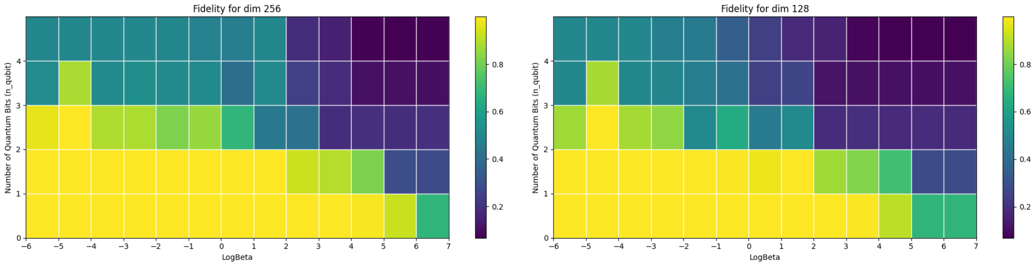


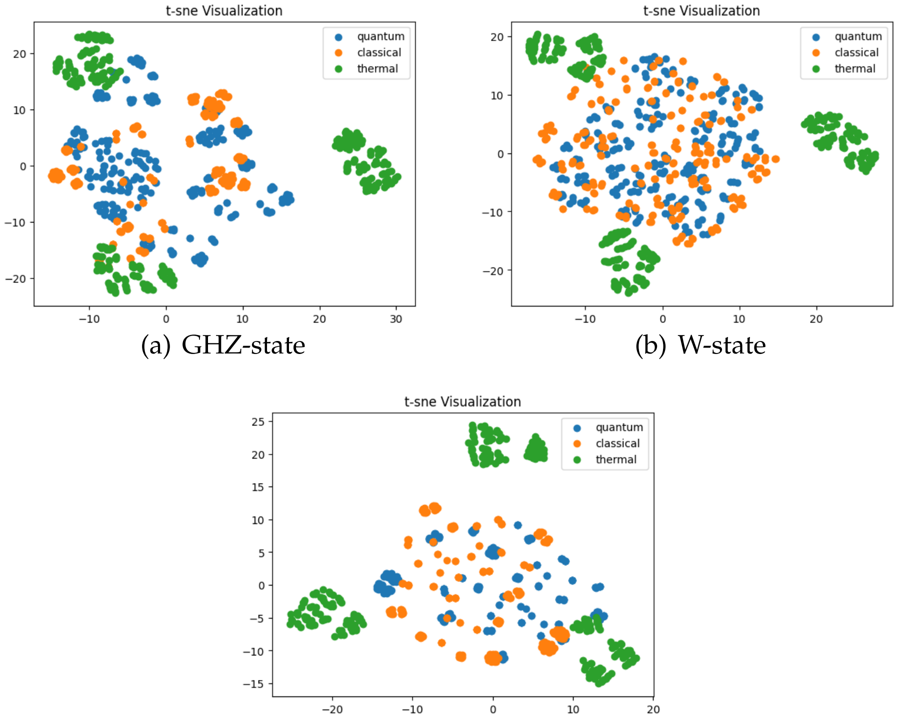
Figure 4.6: *Fidelity heatmap for reconstructed ghz-state from CST in higher latent dimension(the left one is 256 and the right one is 128)*

We discovered that as the dimension of the latent space increases, the model's performance in reconstructing the quantum density matrix of W-states surpasses that of GHZ-states. This implies that machine learning models with more complex structures and a greater number of parameters are seemingly more adept at learning W-states than GHZ-states. This finding is in stark contrast to conclusions drawn from models with smaller latent spaces and fewer parameters. It suggests that a model's ability to handle the complexity of quantum states varies with its structural and parametric complexity, highlighting a nuanced aspect of quantum state machine learning.

Given our experiments and analysis, the latent space plays a crucial role in our Classical Shadow Transformer (CST). In related research, attempts have been made to use the shape of latent space data as an entanglement classifier[31], to learn global quantum properties[2], or to assign interpretability to the shape of the latent space[1]. Consequently, we also focus on the latent space of our CST. It is essential to clarify that our CST learns from the Classical Shadow dataset (x, y), which represents the joint probability distribution of the observation basis and the results for a specific quantum state. Thus, the parameters of our different CSTs should encode the probability characteristics of the quantum state under Pauli measurements.

Specifically, we first divide the model into three regions: *'quantum'*, *'classical'*, and *'thermal'*, corresponding to different information processing capabilities, with $\log_2 \beta$ = [-5, -1, 6]. Since the model learns the probability distribution of the observed results under a given Pauli measurement basis, for the 5-qubit case, we generate $\{x, y, z\}^{\times 5}$, which amounts to $3^5 = 243$ combinations of all Pauli measurement bases. We observe the behaviour of the model's latent space. To project it onto a two-dimensional plane, we employ the T-SNE(t-distributed Stochastic Neighbor Embedding) method for dimensionality reduction.

To have a comparison both on the inherent information processing ability of the model(different β value) and the states the model is trained on, we draw the plot as Figure4.7, which show 3 different regions (*'quantum'*, *'classical'*, and *'thermal'*) in each subplot for 3 different states and also draw the plot as Figure 4.8, which show all 3 states(GHz,w zero) in each subplot for different information bottleneck.



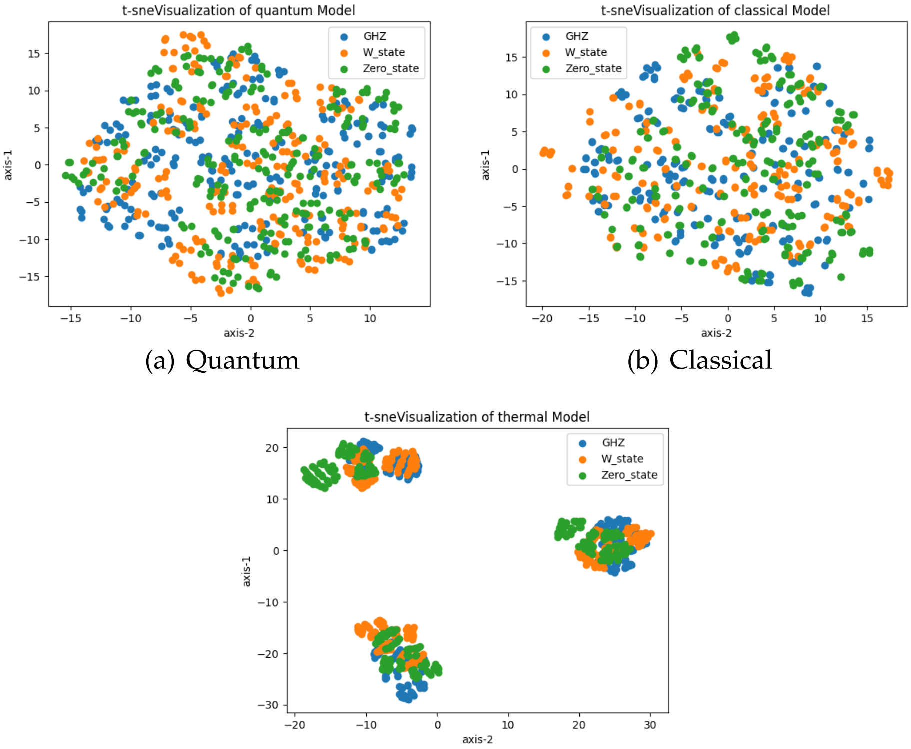
(a) GHZ-state

(b) W-state

(c) Zero-state

Figure 4.7: *T-sne plot of the latent space of all possible observables for CST trained on different states, each dot represents a Pauli sequence and model with different information bottleneck are labelled in a different colour of Green(thermal,$\log_2 \beta = 6$), Orange(classical,$\log_2 \beta$=-1) and Blue(quantum,$\log_2 \beta$=-5), all model is trained on a 5-qubit system and dimension of latent space is 64*

According to Figure 4.7, We observe that for the t-SNE latent images of the three states, when in the thermal region, where the model's information processing capability is at its weakest, the clusters tend to be distributed towards the periphery of the plane. Conversely, as the decoding ability of the CST (Classical Shadow Transformer) is enhanced, situated in the quantum and classical regions, the clusters corresponding to these observed sequences become more centralized in the plane



(a) Quantum

(b) Classical

(c) thermal

Figure 4.8: *T-sne plot of the latent space of all possible observables of three states for different information bottleneck regions, each dot represents a Pauli sequence and model are labelled in different color of blue(GHZ-state). orange (W-state). and green (Zero-state), all model is trained on a 5-qubit system and the dimension of latent space is 64*

If we focus on the behaviour for different states but in the same information bottleneck as Figure 4.8 shows. We can find that when the model has the highest limitation on decoding information as the thermal region, all three states tend to cluster more closely together, indicating a diminished capacity for state discrimination. This is in stark contrast to the quantum and classical regions, where the clusters corresponding to each state are more spread out and distinct. In the quantum region, this spread is indicative of a superior ability to differentiate between states, likely due to the higher information processing capabilities of the model. Moving to the classical region, while the differentiation between states is still apparent, it is less pronounced than in the quantum region, suggesting a moderate level of information processing capability. The thermal region's tight clustering underscores the challenges the model faces when information processing is highly constrained, leading to a potential overlap in the representation of different quantum states within the latent space.

Moreover, consistent with the model's loss history and the phase diagrams depicted for the reconstruction of the density matrix, the difference between entangled states (W, GHZ) and separable states (Zero) is significant, yet the distinction between W and GHZ states remains elusive. Finally, providing a quantitative and deeper interpretation of the t-sne images is challenging; the preceding analysis offers only a preliminary and intuitive understanding. The interpretability of the latent space's morphology is also a direction for our future research.

4.3 Discussion and analysis

In light of our discussions and the outcomes observed from the Classical Shadow Transformer (CST), we initially explored the fundamental concept of shadow tomography and implemented it using a transformer-incorporated Variational Autoencoder (VAE) model with an information bottleneck parameter, β. It was observed that learning entangled quantum states poses a greater challenge than their non-entangled counterparts, which aligns with intuition. This is attributed to the complex correlations between measurements across qubits in entangled states, compared to non-entangled states.

Moreover, discerning the degree of entanglement between GHZ and W states does not lend itself to a straightforward conclusion. Typically, when the number of copies of the original state for measurement is limited-a scenario analogous to restricted sample sizes and training durations-the W state appears more difficult to learn than the GHZ state. However, expanding the dimensionality of the latent space within the CST, thereby increasing the model's parameters, results in an improved reconstruction performance for the W-state CST. In contrast, the performance of the GHZ-state CST does not exhibit a significant enhancement with an increase in parameters. We want to give some possible explanations and analysis on such phenomena and first, we just take a look at the structure of each state's density matrix as Figure 4.9 and Figure 4.10

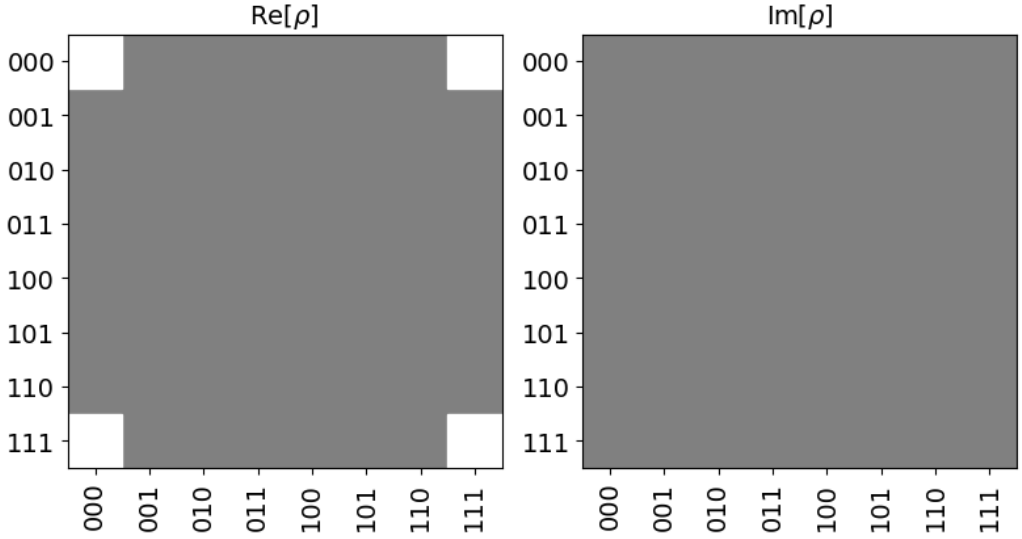


Figure 4.9: *Density Matrix Structure for a 3-qubit ghz-state*

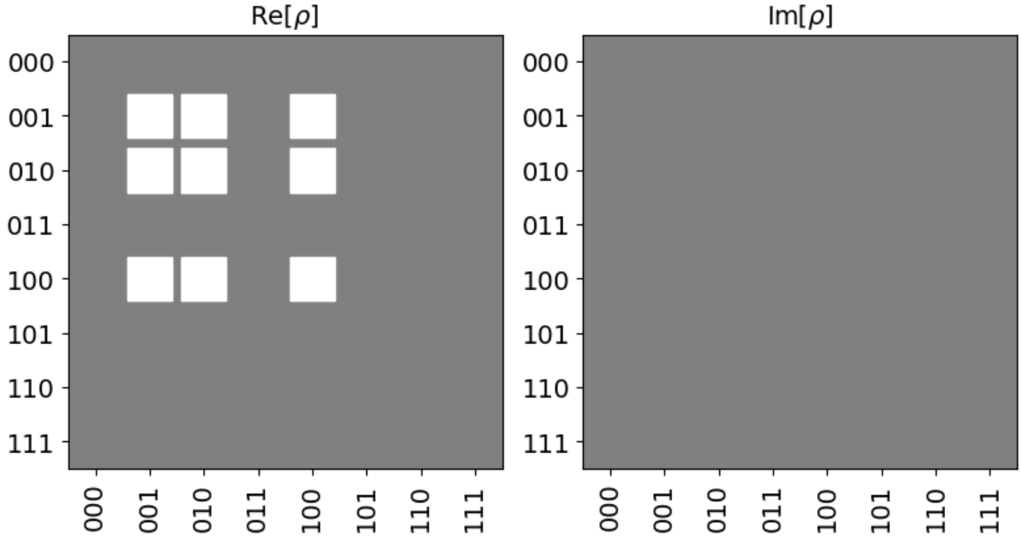


Figure 4.10: *Density Matrix Structure for a 3-qubit w-state*

Upon initial examination of the density matrices, it is evident that the W-state possesses a more intricate structure with less symmetry compared to the GHZ-state. This complexity arises irrespective of dimensionality, as the GHZ-state consistently displays only four non-zero elements located at the matrix's corners. Such a sparse and symmetric distribution inherent to the GHZ-state contributes to its relatively simpler characterization.

However, for the phemona, the w-state will be more easily learned with the increase of the number of parameters for the model and a longer training time, while ghz-state behaves less sensitive to the increasing of complexity of the model. It's quite intriguing but hard to give an overall conclusion on it. According to recent research, the W-state may not have received as much attention as other entangled states, but its unique characteristics are noteworthy. For instance, its entanglement is exceptionally resilient- more so than GHZ states, as it typically necessitates a greater number of measurements to disentangle the system.[8] [7]. Another property for the w-like entanglement we can conclude is that the W-state is known for its robustness of entanglement; even if one of the three qubits is lost, the remaining two-qubit system is still entangled. This is a stark contrast to the GHZ state, which becomes completely separable upon the loss of one qubit[3]. They represent two distinctly different kinds of entanglement and cannot be transformed into each other[10], even probabilistically, via local quantum operations. So, we can draw a conclusion, from the persistence perspective, the W-state is more entangled than the GHZ-state [9].

Above all, the characteristics of the W-state in terms of its persistence and robustness reflect its complex internal structure. This persistence is evident as the W-state maintains entanglement even after the loss of a qubit, demonstrating a stronger resistance to environmental disturbances. Therefore, when using the Classical Shadow Transformer (CST) to decode information from the W-state, this process can be viewed as a response to external information leakage or environmental interference.

In this context, compared to the GHZ-state, the W-state indeed requires a higher number of observations and a more complex model structure for effective decoding and learning. While it's not straightforward to define which state, W or GHZ, is more entangled, we can infer from the analysis of the properties of the W-state that it is more complex in terms of quantum information decoding.

This understanding helps in delving deeper into the applications and potential of different quantum states in quantum information processing. Each state has its unique properties and advantages, which are crucial considerations in the design of quantum algorithms and communication protocols.

Chapter 5

Conclusion

Bridging Quantum and Classical Worlds through Machine Learning We successfully demonstrate the potential of machine learning models, particularly the Classical Shadow Transformer, in interpreting and decoding quantum information. The study reveals that the boundary between the quantum and classical worlds is not absolute. This distinction hinges not only on the scale of the quantum system but also on our ability to process information and the complexity of our computational agent. Consequently, if equipped with a sufficiently powerful AI agent with advanced computational capabilities, it may become possible to perfectly decode quantum information, regardless of the system size.

Entanglement Complexity and Learnability Our experiments with different quantum states indicate that the complexity and type of entanglement significantly influence the model's learning and reconstruction abilities. Notably, the W state, characterized by its robust entanglement, presents more challenges in learning and reconstruction compared to the GHZ and Zero states. This finding suggests a direct correlation between the degree of entanglement in a quantum state and its learnability in quantum state tomography.

The Role of Latent Space in Quantum State Analysis The research underscores the crucial role of the latent space in the transformer model, providing valuable insights into the distinct characteristics of various quantum states. The exploration of latent space morphology through T-SNE plots further emphasizes the potential of machine learning in quantum information science. This exploration not only enhances our understanding of quantum states but also points to the significance of latent features in advancing the applications of quantum computing.

Chapter 6

Future Direction

Exploring Alternative Measurement Bases Having successfully trained our Classical Shadow Transformer (CST) using the Pauli measurement basis, we now consider exploring other groups of measurement bases and corresponding quantum channels. An immediate approach is to replace the Pauli basis with Clifford gates to examine their efficacy in shadow tomography.

Deeper Analysis of Quantum Entanglement Our attempt to characterize W-like versus GHZ-like entanglement from a learning ability perspective is intriguing. However, providing a universal statement about which form of entanglement is more complex remains challenging. The W-state, noted for its persistence, demands more observations for accurate information extraction. A more precise and quantitative analysis is needed to make definitive statements about these types of entanglement.

Insights into Latent Structure The visualization of the latent structure for different states and varying levels of information processing reveals complexity. However, the representations are not immediately interpretable. To gain a deeper understanding of the underlying physics, we may consider clustering analysis or alternate methods of dimensionality reduction on the latent structures.

Inherent Limitations of Shadow Tomography A common issue in the generative modelling of quantum states is the potential creation of non-physical states, characterized by non-positive semi-definiteness and negative eigenvalues. This inherent limitation in shadow tomography arises from the inability to perfectly learn the real distribution of observables and measurements. Current solutions, such as forcing negative eigenvalues to zero or employing Generative Adversarial Networks (GANs) to filter out non-physical states, can disrupt the density matrix structure and potentially lead to information loss about the quantum state.

Scaling Up the Quantum Language Model The primary goal of this research is to develop an artificial intelligence that can comprehend ′quantum language′. However, the scope of our current model is somewhat limited and does not demonstrate a clear superiority over traditional methods. A potential extension could involve applying the model to larger quantum systems without performing full-state tomography. By extracting only partial information from the quantum system, we aim to enhance the efficiency of our learning process.

Acknowledgements

This work has been supported by many individuals to whom I wish to express my deepest gratitude.

First and foremost, I would like to extend my sincere thanks to Evert van Nieuwenburg, my first supervisor. Your patience, insightful conversations, and timely encouragement have been invaluable. Although the outcomes of this work might be preliminary or elementary, the knowledge and experience I have gained from this journey far exceed my expectations.

I am also grateful to Jordi Tura, my second supervisor, for his interesting ideas and suggestions that greatly contributed to this project.

Special thanks go to Felix Frohnert, whose significant contributions to the details of my project cannot be overstated. His patient answers to even my simplest questions have been of great benefit to me.

I would also like to thank the condensed AI group, of course with Meng Yao and Xiaoyu from AQA, for their company and the discussions that have been immensely helpful. Of course, I must also express my gratitude to YI Zhuang You from UCSD, the author of a core reference paper, whose discussions with me resolved many of my doubts.

Finally, I would like to thank my friends and family, and certainly not least, Ocean, for their emotional support, without which I could not have completed this endeavour.

Bibliography

[1] F. Frohnert and E. van Nieuwenburg, *Explainable Representation Learning of Small Quantum States*, 2023.

[2] Y.-D. Wu, Y. Zhu, Y. Wang, and G. Chiribella, *Learning Global Quantum Properties from Local Measurements with Neural Networks*, arXiv e-prints , arXiv (2023).

[3] P. MigdaÅ, J. Rodriguez-Laguna, and M. Lewenstein, *Entanglement classes of permutation-symmetric qudit states: Symmetric operations suffice*, Physical Review A **88** (2013).

[4] H.-Y. Huang, R. Kueng, and J. Preskill, *Predicting many properties of a quantum system from very few measurements*, Nature Physics **16**, 1050â1057 (2020).

[5] S. Aaronson, *Shadow Tomography of Quantum States*, 2018.

[6] Z. Zhang and Y.-Z. You, *Observing Schrödinger's Cat with Artificial Intelligence: Emergent Classicality from Information Bottleneck*, 2023.

[7] H. J. Briegel and R. Raussendorf, *Persistent Entanglement in Arrays of Interacting Particles*, Physical Review Letters **86**, 910â913 (2001).

[8] E. D'Hondt and P. Panangaden, *The computational power of the W and GHZ states*, arXiv preprint quant-ph/0412177 (2004).

[9] M. Enríquez, I. Wintrowicz, and K. Życzkowski, *Maximally entangled multipartite states: a brief survey*, in *Journal of Physics: Conference Series*, volume 698, page 012003, IOP Publishing, 2016.

[10] W. DÃŒr, G. Vidal, and J. I. Cirac, *Three qubits can be entangled in two inequivalent ways*, Physical Review A **62** (2000).

[11] A. Einstein, B. Podolsky, and N. Rosen, *Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?*, Phys. Rev. **47**, 777 (1935).

[12] J. S. Bell, *On the Einstein Podolsky Rosen paradox*, Physics Physique Fizika **1**, 195 (1964).

[13] R. F. Werner, *Quantum states with Einstein-Podolsky-Rosen correlations admitting a hidden-variable model*, Phys. Rev. A **40**, 4277 (1989).

[14] W. K. Wootters, *Entanglement of Formation of an Arbitrary State of Two Qubits*, Phys. Rev. Lett. **80**, 2245 (1998).

[15] G. Vidal and R. F. Werner, *Computable measure of entanglement*, Phys. Rev. A **65**, 032314 (2002).

[16] E. I. G. Schmidt, *Zur Theorie der linearen und nichtlinearen Integralgleichungen*, Mathematische Annalen **63**, 433 (1907).

[17] O. Gühne and G. Tóth, *Entanglement detection*, Physics Reports **474**, 1â75 (2009).

[18] C. Kokail, C. Maier, R. van Bijnen, T. Brydges, M. K. Joshi, P. Jurcevic, C. A. Muschik, P. Silvi, R. Blatt, C. F. Roos, and P. Zoller, *Self-verifying variational quantum simulation of lattice models*, Nature **569**, 355â360 (2019).

[19] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, *Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets*, Nature **549**, 242â246 (2017).

[20] D. P. Kingma and M. Welling, *Auto-Encoding Variational Bayes*, 2022.

[21] D. J. Rezende, S. Mohamed, and D. Wierstra, *Stochastic Backpropagation and Approximate Inference in Deep Generative Models*, 2014.

[22] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner, *beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework*, in *International Conference on Learning Representations*, 2016.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, *Attention Is All You Need*, 2023.

[24] D. M. Greenberger, M. A. Horne, and A. Zeilinger, *Going beyond Bellâs theorem*, in *Bellâs theorem, quantum theory and conceptions of the universe*, pages 69–72, Springer, 1989.

[25] W. DÃŒr, G. Vidal, and J. I. Cirac, *Three qubits can be entangled in two inequivalent ways*, Physical Review A **62** (2000).

[26] F. Diker, *Deterministic construction of arbitrary W states with quadratically increasing number of two-qubit gates*, 2022.

[27] A. J. et al., *Quantum AlgorithmÂ Implementations for Beginners*, ACM Transactions on Quantum Computing **3**, 1â92 (2022).

[28] R. Schmied, *Quantum state tomography of a single qubit: comparison of methods*, Journal of Modern Optics **63**, 1744â1758 (2016).

[29] M. Guţă, J. Kahn, R. Kueng, and J. A. Tropp, *Fast state tomography with optimal error bounds*, Journal of Physics A: Mathematical and Theoretical **53**, 204001 (2020).

[30] M. Ohliger, V. Nesme, and J. Eisert, *Efficient and feasible state tomography of quantum many-body systems*, New Journal of Physics **15**, 015024 (2013).

[31] N. Sa and I. Roditi, *β-variational autoencoder as an entanglement classifier*, Physics Letters A **417**, 127697 (2021).