NJESR/September-2023/Volume-5/Issue-9 E-ISSN-2582-5836
DOI-10.53571/NJESR.2023.5.9.76-83

From Black Box To Glass Box: Quantum-Assisted Explainable Generative Models For High-Stakes Applications

Sabu.K.J[*1]Antony.P.V[*2]

Lecturer[*1&*2]

Department Of Computer Engineering

Government Polytechnic College

Kothamangalam

Kerala

(Received-25August2023/Revised-13September2023/ Accepted-24September 2023/ Published-30September2023)

Abstract

Governance, transparency, and explainability are becoming essentialas artificial intelligence becomes more and more integrated into high-stakes decision-making, especially in industries like healthcare, banking, and human resources.Institutions are moving away from opaque, black-box approaches towards more interpretable, accountable systems due to regulatory pressures and ethical commitments.In artificial intelligence (AI), the phrase "From Black Box to Glass Box" describes the transition from opaque, incomprehensible models to transparent, intelligible "glass box" models. The term, "Quantum-Assisted Explainable Generative Models for High-Stakes Applications," refers to a state-of-the-art field of study that focusses on how quantum computing can assist in developing AI systems that produce justifications for their choices, making them appropriate for crucial applications in industries like healthcare and finance where accountability and trust are crucial.For high-stakes applications like healthcare and finance, quantum-assisted explainable generative models combine explainability (XAI) and generative AI (Gen AI) to maximise their potential.These models provide benefits including enhanced interpretability and faster, more sophisticated data production by utilising quantum events.The use of Retrieval-Augmented Generation architectures to transform AI from a "black box" to a "glass box" an understandable, interrogable system that satisfies institutional governance standards is examined in this article.We go over use cases, design patterns, implementation difficulties, and institutional alignment frameworks.

Keywords: Deep Learning, Machine Learning, Quantum Neural Networks, Quantum Mechanics, Artificial Intelligence(AI), Quantum Computing(QC).

Introduction

Utilising quantum mechanics to effectively produce complicated quantum states or data samples[3], quantum generative models [1] have demonstrated considerable promise in domains including materials research and drug discovery [2]. However, quantum characteristics like superposition and entanglement, which lead to high entropy in the state space and make output interpretation more difficult, provide special interpretability issues for these models.

Interpreting the link between input factors (such as Hamiltonian parameters) and output quantum states is a major challenge [4, 5]. Unlike classical models, where layers and parameters can be directly analyzed, quantum models operate within highly abstract quantum state spaces. The "black box" nature of quantum models limits our understanding of their internal operations, making it challenging to control or verify specific outcomes, especially in high-stakes applications. This lack of interpretability has significant practical implications. In critical fields like drug discovery and materials science, understanding how quantum states are generated is vital for ensuring trustworthy results that meet stringent requirements[6]. Without the ability to interpret the relationships between input parameters and generated quantum states, controlling these models to achieve desired outcomes becomes extremely difficult, undermining their reliability and broader adoption. While classical generative models also face interpretability challenges, various techniques, such as visualization tools, simplified architectures, and model inversion, have been developed to enhance understanding. However, these methods are often insufficient or inapplicable to quantum models due to their probabilistic nature and complex non-linearities, creating a critical gap in our ability to understand and control quantum generative models[7]. To address these challenges, we propose applying model inversion techniques to quantum generative models. Model inversion maps generated quantum states back to their latent variables, revealing the underlying relationships between inputs and outputs. This approach allows users to trace how specific quantum states are produced, offering a mechanism to control and fine-tune outputs by adjusting latent variables. Moreover, it enables the identification of specific directions in the latent space that correlate with distinct features in the generated samples, allowing for targeted interventions in the generative process. Explainable AI (XAI) refers to methods and techniques in the application of AI such that the results of the solution can be understood by human experts. It contrasts with the concept of the "black box" in AI, where the internal workings of the model are not easily accessible or interpretable[5-7]. These

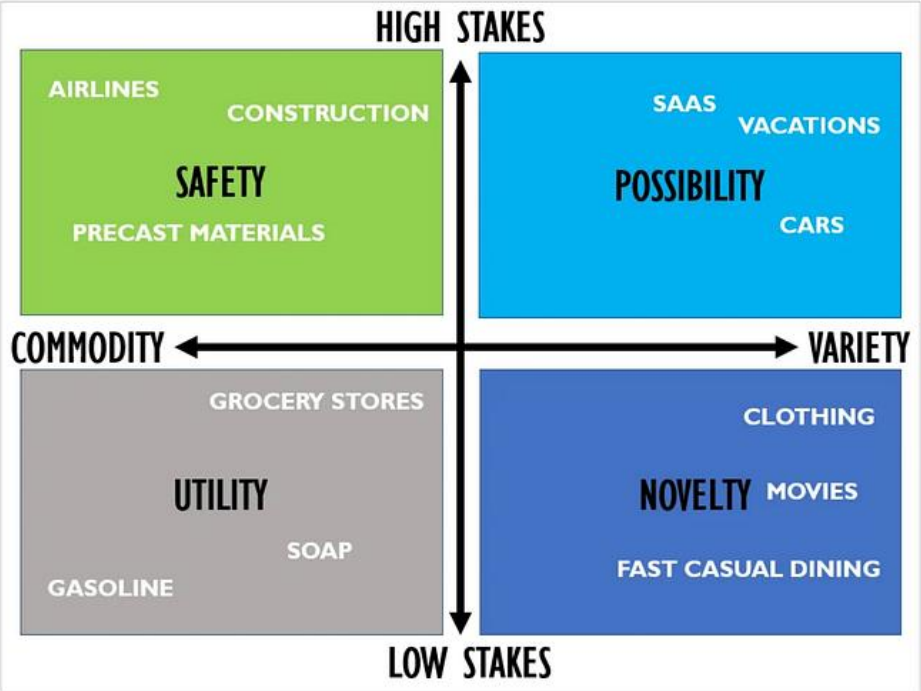
models make predictions based on input data, but the decision-making process and reasoning behind the predictions are not transparent to the user. This lack of transparency makes it strenuous for users to understand the model's behavior, detect potential biases or errors, or hold the model accountable for its decisions. In XAI, the term "black box" is often used to contrast with "white box" or "transparent" models, where the internal workings and reasoning behind the predictions are easily accessible and interpretable[7]. Overall, it helps users deeply understand and trust the decisions made by these systems. In general, highly successful prediction models, such as deep neural networks (DNNs), have some inherited drawbacks in terms of transparency that need to be addressed to justify the use of these models in many scenarios. This paper aims to explore the emerging relationship between AI and quantum computing, analyzing how these technologies can complement and accelerate each other. We delve into the current state of research, identify key areas of synergy, and discuss real-world applications where the fusion of AI and quantum computing could unlock transformative outcomes[8].

High Stakes- Low Stakes Decisions

Understanding your machine learning model is important, but how much you need to understand depends on what it's being used for. In some cases, knowing just a little might be enough, but when the model is used for important or sensitive tasks, you need to know a lot more. High-stakes decisions are those where the model's choices can have big effects on people's lives or money. It's really important to make sure the model works as it should and doesn't cause problems[8]. For example, in healthcare, when a model is used to diagnose diseases or suggest treatments, it's important to know how it comes up with its answers to make sure patients are safe and get good care. In finance, when models are used to decide who gets a loan or spot fraud, understanding how they work helps ensure decisions are fair and follow the law. Bias and fairness are also important. If a model is used in areas where treating people equally is key, knowing how it makes decisions can help spot and fix unfairness.[6-8]

In public policy or legal matters, people might need to know how decisions are made. Without this, it's hard to improve the model or fix any problems it might have. Trust is really important when people interact with computers. If they don't trust a system, they're less likely to use it. Explaining how the system makes decisions can help build trust, especially for people who aren't experts and might be worried about "black box" systems that don't show their thinking. Some

industries have rules that require systems to be transparent. For example, in the European Union, the General Data Protection Regulation (GDPR) gives people the right to ask for explanations about decisions made by machines that affect them.[9]



For low-stakes applications like movie recommendations, an inaccurate recommendation is not a severe issue. The main priority in such cases might be speed in delivering results rather than ensuring complete transparency. During the initial stages of model development (prototyping), the objective might be to test several models rapidly. Here, the focus might not be on understanding each model deeply but on iterating quickly. When developing machine learning models for high-stakes decision-making, achieving high accuracy is essential but not enough. Moreover, the data used for training and testing phases might not fully capture the diversity and complexity of real-world data the model will face.[8] Implementing continuous monitoring mechanisms is crucial to track the model's performance over time and make necessary adjustments to account for changes in data distribution. In these contexts, model decisions often have significant and far-reaching consequences, making it imperative to consider additional factors. These additional factors (criteria) are difficult to measure precisely due to their subjective and context-dependent nature.

For example, transparency requires that the model's workings are open and understandable to stakeholders. However, quantifying transparency is challenging as it may mean different things to different people. For instance, a data scientist may require detailed algorithmic information, while an end-user may need a simple, high-level explanation of how decisions are made.[9]

Interpretable Machine Learning

The first thought that comes to mind when discussing black-box models is usually a basic interpretation of them. When machine learning models are used in a product, interpretable systems are often a key factor. In machine learning, interpretability is an essential aspect. However, it remains unclear how to measure it. Due to this ambiguity, academics often confuse the terms "interpretability" and "explainability." Only when machine learning models are explainable can they be audited and debugged. Even in a reliable field, such as movie reviews, it is challenging to determine if a review is positive or negative since the movie rating and the sentiment do not always match [2, 3]. When a product is deployed, things can go wrong. Interpreting an incorrect prediction helps in identifying the root cause. It offers guidance on how to fix the system. A notable example of ambiguity is the task of classifying wolves versus Siberian huskies from [3], where a deep neural network incorrectly labels some canines as wolves. The experiment predicts "Wolf" if there is snow and "Husky" otherwise, regardless of color, position, or pose. The experiment starts as follows: First, a wolf without a snowy background is presented (classified as a husky), and then a husky with a snowy background is presented (classified as a wolf) [4]. Another example of an incorrect prediction by ML that could be resolved through interpretability is the case of a deep learning model designed to predict which patients would benefit from antidepressants. When the model was evaluated on a new set of patients, it made inaccurate predictions. Specifically, it predicted that some patients who benefited from the medication would not, and vice versa. This could have serious consequences for patients as prescribing the wrong medication could lead to ineffective treatment and potentially dangerous side effects[9]. The researchers used the SHapley Additive exPlanations (SHAP) technique to create an interpretable version of the deep learning model for predicting treatment outcomes in depression. SHAP is a method that can be applied to any machine learning model to provide explanations for specific predictions. Using SHAP, the researchers were able to identify the most influential factors affecting the model's predictions for each patient. These factors included demographic variables such as age and gender, along with

genetic markers associated with treatment response[8]. By providing these explanations to clinicians, the researchers aimed to improve the accuracy and reliability of the model's predictions and to detect potential errors or biases in the underlying data. The interpretable version of the model achieved approximately 70% accuracy in identifying patients who were more likely to benefit from escitalopram. Modern methods are being developed daily to make AI more understandable. Trying to keep up with all the published research would be absurd and impossible[5].

Quantum Generative Models

Quantum generative models have become important tools for creating and simulating quantum states in different quantum systems. These models take advantage of quantum computing's ability to manage complex and high-dimensional data, offering various methods for generating quantum states. One type of model is the Quantum Circuit Born Machine (QCBM), which uses quantum circuits to learn and represent the probabilities of different quantum states. QCBMs are good at learning from quantum data and can be used for preparing quantum states and performing quantum machine learning[9]. Another type is Quantum Generative Adversarial Networks (QGANs), which work like classical GANs but use quantum processes. In QGANs, a quantum generator creates states that are checked by a discriminator, which can be either classical or quantum, helping to produce high-quality quantum states. These models have been successful in tasks like creating quantum states, analyzing them through tomography, and simulating quantum systems. However, there are still big challenges. It's difficult to understand and control the quantum states that these models produce because quantum systems are complex and the outcomes are based on probability. To better study these models, researchers look at specific quantum systems, such as the transverse-field Ising model and generalized cluster Hamiltonian. These systems are well-known for showing quantum phase transitions and can serve as useful benchmarks for testing and understanding quantum generative models.[8-9]

Understanding Black Box vs. Glass Box Models[9]

Black Box Models:

These are complex AI models, similar to many deep learning networks, which produce accurate results but lack transparency in their internal decision-making processes. It is difficult to understand why they make certain predictions.

Glass Box Models:

Also known as "white box" or transparent models, these are systems where the algorithms, training data, and model logic are visible and understandable. Examples include decision trees and linear regression.

Quantum-Assisted Explainable Generative Models

Quantum Computing's Role:

This section examines how quantum mechanics and quantum algorithms can improve generative AI models.[10]

Generative Models:

These AI systems produce new data, such as images or text, and can often be quite complex.

Quantum Assistance:

Quantum computing may offer new methods for processing information and identifying complex patterns, potentially leading to more powerful and manageable generative models.

The "Explainable" Aspect:

By combining quantum approaches with explainable AI principles, researchers are striving to make these quantum-enhanced generative models more transparent.

High-Stakes Applications:

The goal is to develop AI systems for important fields (e.g., healthcare, finance) that not only generate advanced outputs but also provide clear and reliable explanations for how those outputs are generated.[10]

Conclusions

In conclusion, our study offers a solution to interpretability, which is one of the most important problems in the field of quantum generative models. By creating and using model inversion methods, we offer a framework that improves our comprehension and management of quantum state creation. Explainable AI (xAI) is a fast developing topic that offers a variety of approaches and strategies to improve the transparency and understandability of sophisticated machine learning algorithms. This blog has explored a wide range of application cases of model understanding and differentiated between high-stakes and low-stakes judgements, delving deeply into the complex realm of xAI. We've seen that striking a balance between interpretability and accuracy is a philosophical as well as a technological challenge that calls for a sophisticated approach to model selection and justification. Deep learning, neural networks, and machine learning (ML) algorithms can all be better understood and explained by humans with the aid of explainable AI. Many people

believe that machine learning models are unintelligible black boxes. Deep learning neural networks are among the most difficult for humans to comprehend.

References

[1]. Dalal, A. (2022). Design and application of small-scale quantum information processors.

[2]. Kop, M., Aboy, M., De Jong, E., Gasser, U., Minssen, T., Cohen, I.G., ... & Laflamme, R. (2023). Towards responsible quantum technology. Harvard Berkman Klein Center for Internet & Society Research Publication Series, 1.

[3]. Vuffray, M., Coffrin, C., Kharkov, Y.A., & Lokhov, A.Y. (2022). Programmable quantum annealers as noisy Gibbs samplers. PRX Quantum, 3(2), 020317.

[4]. Borle, A. (2022). Quantum Optimization for Linear Algebra Problems (Doctoral dissertation, University of Maryland, Baltimore County).

[5]. Marshall, J., Mossi, G., & Rieffel, E.G. (2022). Perils of embedding for quantum sampling. Physical Review A, 105(2), 022615.

[6]. Vandenbroucque, A., Chiacchio, E.I. R., & Munro, E. (2022). The Houdayer algorithm: overview, extensions, and applications. arXiv preprint arXiv:2211.11556.

[7]. Yarkoni, S. (2022). Applications of quantum annealing in combinatorial optimization (Doctoral dissertation, Leiden University).

[8]. Ng, E., Onodera, T., Kako, S., McMahon, P.L., Mabuchi, H., & Yamamoto, Y. (2022). Efficient sampling of ground and low-energy Ising spin configurations with a coherent Ising machine. Physical Review Research, 4(1), 013009.

[9]. Kehoe, A., Wittek, P., Xue, Y., & Pozas-Kerstjens, A. (2021). Defence against adversarial attacks using classical and quantum-enhanced Boltzmann machines. Machine Learning: Science and Technology, 2(4), 045006.

[10]. McClean, J. R., Harrigan, M.P., Mohseni, M., Rubin, N.C., Jiang, Z., Boixo, S., ... & Neven, H. (2021). Low-depth mechanisms for quantum optimization. PRX Quantum, 2(3), 030312.