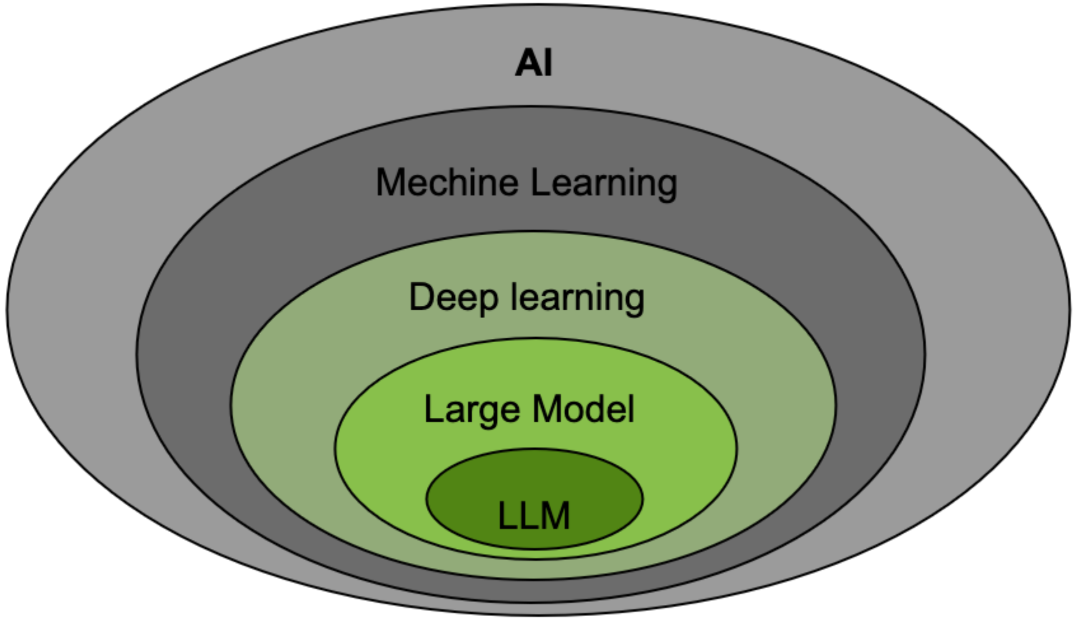
华大生命科学研究院 BGI·Research

Quantum Meets Large Model

马 腾

2024.5

Concepts



卷积神经网络（CNN）图像数据

递归神经网络（RNN）序列数据

变换器网络（Transformer）注意力机制，序列数据

$$\min_{\theta} D\{H[\ W(\theta),\ x\ ],\ y\ ]\}$$

optimization Model ansatz input output

Training cost

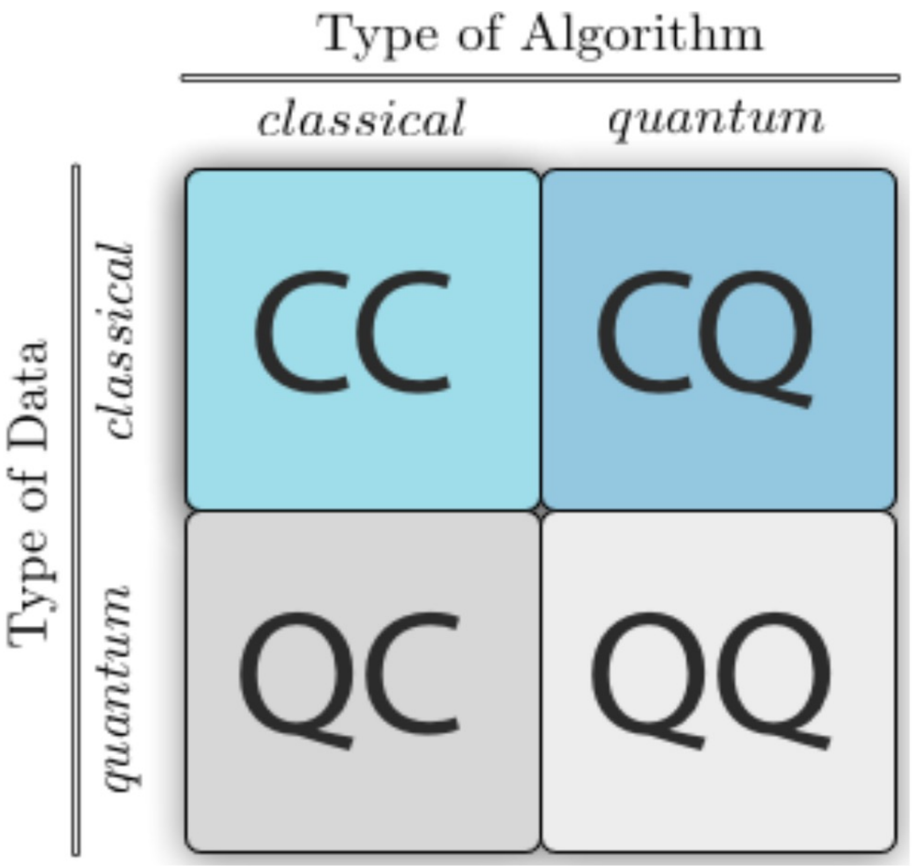
华大生命科学研究院 BGI·Research

- 大模型
- 生成对抗网络
- 多模态模型
- 预训练
- GPT (generative pretrained transformer)
- 扩散模型

	数据集	参数
GPT-1	5GB	1.17亿
GPT-2	40GB	15亿
GPT-3	45TB	1750亿

Enormous computational power and time for training !!!

Quantum * Machine Learning



Four different ways to combine quantum computing with machine learning

- Quantum algorithms for training neural networks
- Quantum optimization algorithms
- Quantum machine learning
- ….

[1] A. Zeguendry, Z. Jarir, and M. Quafafou, Entropy **25**, 287 (2023).

Attention Is All You Need [1]

Transformer [Chat-GPT]

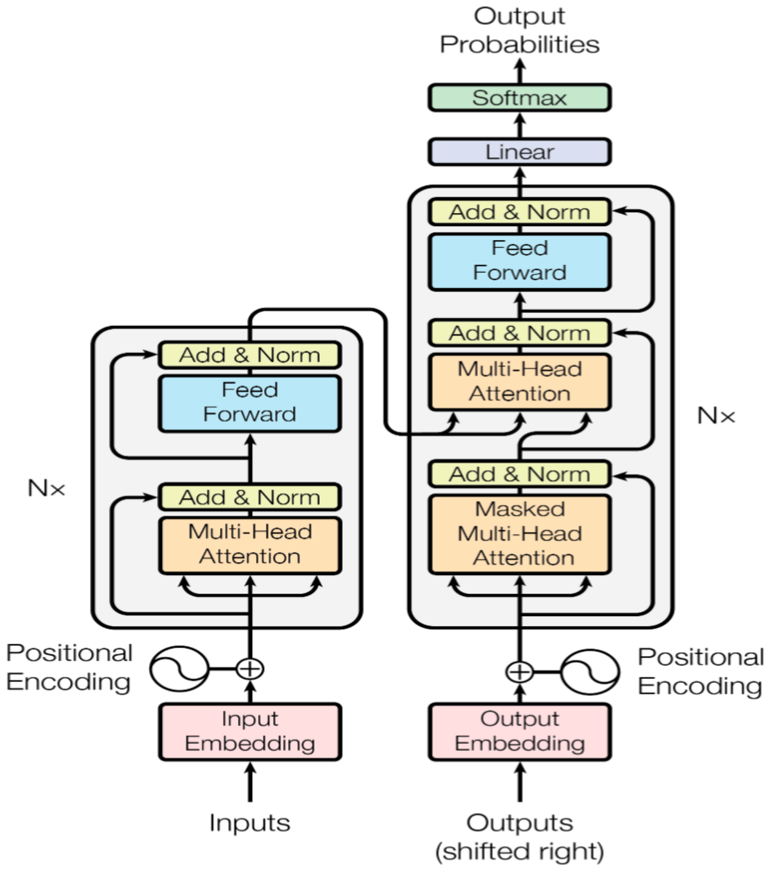
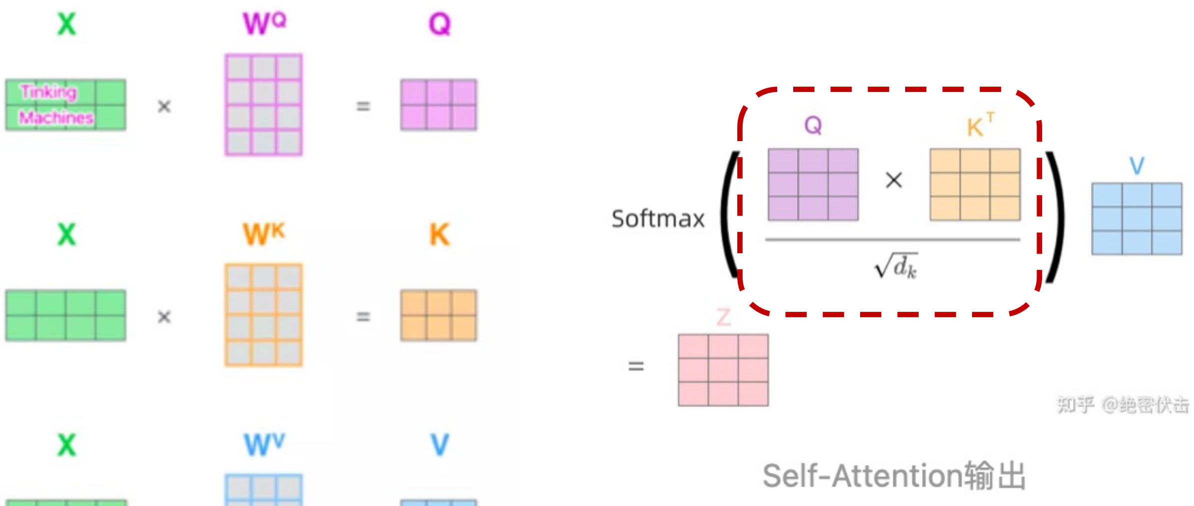


Figure 1: The Transformer - model architecture.

Attention Matrix



$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Vol. 30 (Curran Associates, Inc., 2017).

Quantum * attention

华大生命科学研究院 BGI·Research

Fast Quantum Algorithm for Attention Computation [1]

- **Grover Search** algorithm to compute a sparse **attention computation matrix**
- A polynomial **quantum speed-up** over the classical method.

Quantum linear algebra is all you need for Transformer architectures [2]

- how to prepare a **block encoding** of the self-attention matrix
- combine quantum subroutines

[1] Y. Gao, Z. So`ng, X. Yang, and R. Zhang, arXiv:2307.08045
[2] N. Guo, Z. Yu, A. Agrawal, and P. Rebentrost, arXiv:2402.16714.

Quantum * neural networks

Quantum algorithms for feedforward neural networks [1]

- quantum **inner-product estimation** to speed up the training of feedforward neural networks and convolutional neural networks

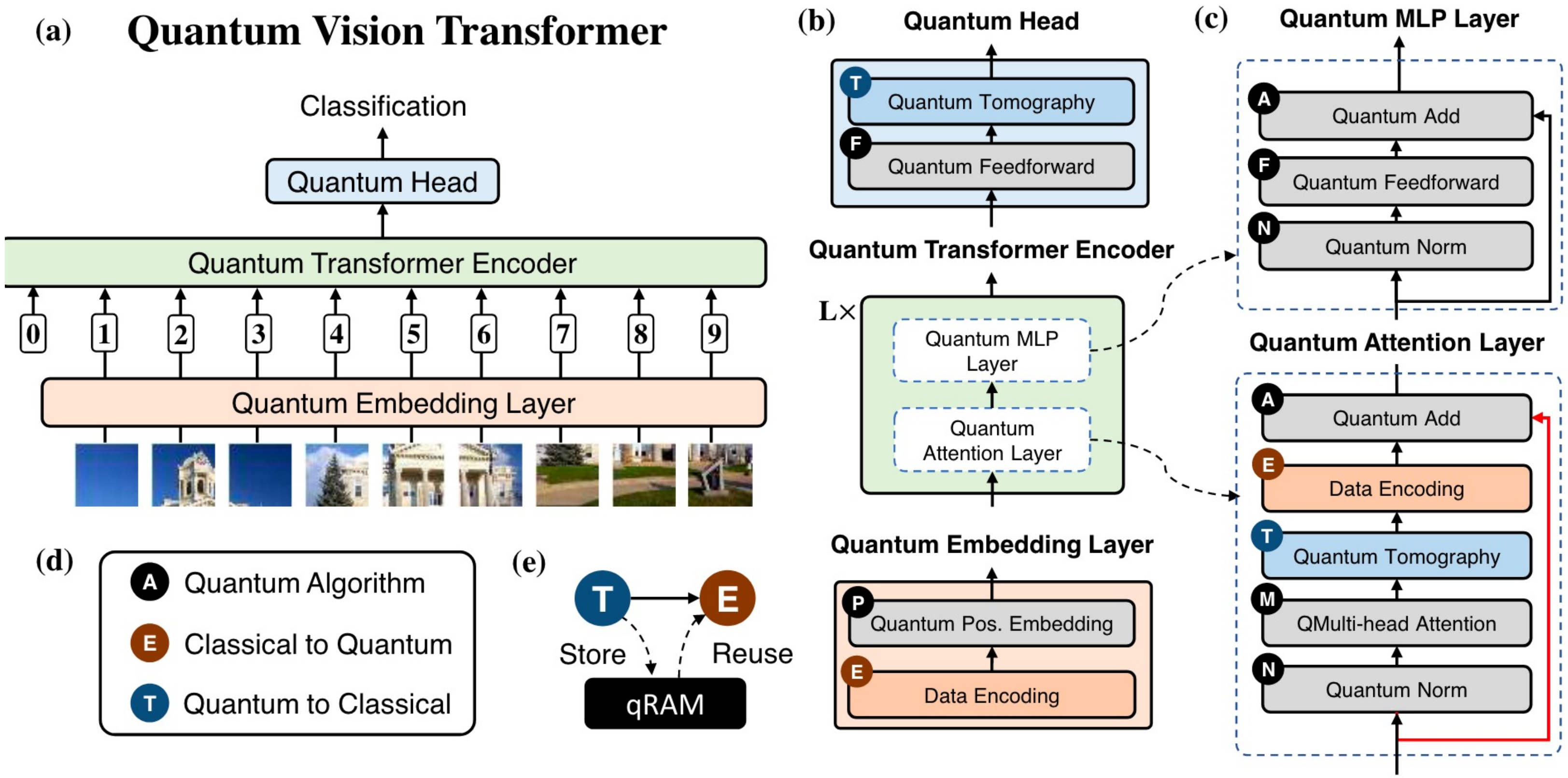
Towards provably efficient quantum algorithms for large-scale machine-learning models [2]

- fault-tolerant quantum algorithms could potentially contribute to most state-of-the-art, large-scale machine-learning problems.

[1] ACM Transactions on Quantum Computing, 1(1):1–24, 2020.
[2] J. Liu, M. Liu, J.-P. Liu, Z. Ye, Y. Wang, Y. Alexeev, J. Eisert, and L. Jiang, Nat. Commun. **15**, 434 (2024).

End-to-End Quantum Vision Transformer: Towards Practical Quantum Speedup in Large-Scale Models[1]

- Information lose in quantum tomography --> quantum residual connection technique



[1] C. Xue, Z.-Y. Chen, X.-N. Zhuang, Y.-J. Wang, et.al , arXiv:2402.18940.

The Quantum Leap: How Quantum Computing Will Shape the Future of Large Language Models [1]

Faster Training： Quantum computing's inherent ability to handle complex algorithms efficiently makes it a compelling candidate to expedite the training process. Quantum computers can explore multiple possibilities simultaneously, significantly reducing the time it takes to train large language models.

Enhanced Natural Language Processing： Quantum neural networks can potentially uncover deeper linguistic patterns and nuances, leading to more accurate language models. They have the potential to excel in tasks like sentiment analysis, language translation, and context-aware natural language understanding.

Improved Security and Privacy: Quantum cryptography, particularly quantum key distribution (QKD) protocols, offers an unparalleled level of security. By integrating quantum cryptography into language models, organizations can ensure end-to-end encryption and protect sensitive data.

[1] https://www.appypie.com/blog/role-of-quantum-computing-in-llms

参考文献

华大生命科学研究院
BGI·Research

[1] J. Liu, M. Liu, J.-P. Liu, Z. Ye, Y. Wang, Y. Alexeev, J. Eisert, and L. Jiang, Nat. Commun. **15**, 434 (2024).
[2] D. Kartsaklis, I. Fan, R. Yeung, A. Pearson, R. Lorenz, A. Toumi, G. de Felice, K. Meichanetzidis, S. Clark, and B. Coecke, arXiv:2110.04236.
[3] H. Pan, N. Mudur, W. Taranto, M. Tikhanovskaya, S. Venugopalan, Y. Bahri, M. P. Brenner, and E.-A. Kim, arXiv:2403.03154.
[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Vol. 30 (Curran Associates, Inc., 2017).
[5] Y. Gao, Z. Song, X. Yang, and R. Zhang, arXiv:2307.08045.
[6] C. Xue, Z.-Y. Chen, X.-N. Zhuang, Y.-J. Wang, T.-P. Sun, J.-C. Wang, H.-Y. Liu, Y.-C. Wu, Z.-L. Wang, and G.-P. Guo, arXiv:2402.18940.
[7] J. Bausch, arXiv:2006.14619.
[8] M. Parigi, S. Martina, and F. Caruso, arXiv:2308.12013.
[9] D. Widdows, W. Aboumrad, D. Kim, S. Ray, and J. Mei, arXiv:2403.19758.
[10] F. Chen, Q. Zhao, L. Feng, C. Chen, Y. Lin, and J. Lin, arXiv:2403.02871.
[11] N. Guo, Z. Yu, A. Agrawal, and P. Rebentrost, arXiv:2402.16714.
[12] C. Chen and Q. Zhao, arXiv:2401.07039.
[13] K. Huang, Z.-A. Wang, C. Song, K. Xu, H. Li, Z. Wang, Q. Guo, Z. Song, Z.-B. Liu, D. Zheng, D.-L. Deng, H. Wang, J.-G. Tian, and H. Fan, Npj Quantum Inf. **7**, 165 (2021).
[14] C. Zoufal, A. Lucchi, and S. Woerner, Npj Quantum Inf. **5**, 103 (2019).
[15] A. M. Smaldone, G. W. Kyro, and V. S. Batista, Quantum Machine Intelligence **5**, 41 (2023).
[16] I. Kerenidis, J. Landman, and A. Prakash, arXiv:1911.01117.
[17] Y. Chen, arXiv:2110.15667.
[18] H.-L. Huang, Y. Du, M. Gong, Y. Zhao, Y. Wu, C. Wang, S. Li, F. Liang, J. Lin, Y. Xu, R. Yang, T. Liu, M.-H. Hsieh, H. Deng, H. Rong, C.-Z. Peng, C.-Y. Lu, Y.-A. Chen, D. Tao, X. Zhu, and J.-W. Pan, Phys. Rev. Appl. **16**, 024051 (2021).
[19] Y. Liao and C. Ferrie, arXiv:2403.09418.
[20] B. Zhang, P. Xu, X. Chen, and Q. Zhuang, Phys. Rev. Lett. **132**, 100602 (2024).
[21] M. Kölle, G. Stenzel, J. Stein, S. Zielinski, B. Ommer, and C. Linnhoff-Popien, arXiv:2401.07049.
[22] J. Ho, A. Jain, and P. Abbeel, (n.d.).
[23] A. Cacioppo, L. Colantonio, S. Bordoni, and S. Giagu, arXiv:2401.07049v1 (n.d.).

THANKS

OMICS FOR ALL

基 因 科 技 造 福 人 类