量子計算を用いた文字言語モデル

三輪拓真 [1,2]　河野誠也 [2,1]　吉野幸一郎 [2,1]
[1] 奈良先端科学技術大学院大学　[2] 理化学研究所ガーディアンロボットプロジェクト
miwa.takuma.mx0@is.naist.jp, {seiya.kawano, koichiro.yoshino}@riken.jp

概要

大規模言語モデルの登場により，言語モデルは近年多くの注目を浴びている．また言語モデルは機械翻訳や要約など，幅広いタスクに応用されている．一方で大規模言語モデルは未だ計算量に課題があるため，計算効率の改善は重要である．本研究では，限定されたタスクで量子計算の計算効率が古典コンピュータを上回る点に注目した．具体的には，量子計算を用いて文字言語モデルを構築し，その性能の分析を行った．結果として構築したすべてのモデルにおいて，同規模のマルコフモデルよりパープレキシティを減少させることが可能とわかった．

1　はじめに

近年大規模言語モデル (LLM) の貢献により，自然言語処理の関連領域研究は大きな発展を遂げている．モデルの学習データセットやパラメータ数を大規模化させることにより、モデルの予測性能が向上するスケーリング則が報告されている [1]．加えて言語モデルに与えるパラメータ数を一定数以上にして十分な訓練を行うことで，創発と呼ばれる言語モデリングタスク以上の能力を得られる現象も確認されている [2]．一方でハードウェアの制約や計算時間の問題により，現状モデルの大規模化はより困難になっている．そのためモデルのパラメータ数を増やすことなく，データセットやアルゴリズムの変更によって効率化することが重要である．本研究では，量子計算を用いた言語モデルの効率化を検討する．

量子計算では，量子ビット数に対して指数関数的に多くの値を同時に計算することが可能である．一方で現在の量子コンピュータは NISQ (Noisy intermediate-scale quantum) と呼ばれ，計算誤差や量子ビット数の制約が存在する [3]．このため、量子計算を機械学習で活用しようとする既存研究においては，多くはシミュレータを活用している [4]．ただし，シミュレータで計算誤差の問題を解消したとしても，シミュレーションには指数関数乗の空間計算量が必要とされ，利用可能な量子ビット数の制約が依然として存在する．

先に述べた通り，量子計算を用いた言語モデルの実装において、空間計算量は大きな課題である．そこで本研究では，量子ビットを用いた量子化により，現状のシミュレーター規模で実現可能な範囲の文字 n-gram モデルを構築し，その性能と限界を調査する．

2　量子計算

2.1　量子状態

量子計算では，量子の状態に関する情報を利用して演算を行う．各量子ビットは 0 であるという状態 $|0\rangle$ と，1 であるという状態 $|1\rangle$ の重ね合わせによって存在している．よって量子ビット単体の状態 $|\phi\rangle$ は以下のように表現される．

$$|\phi\rangle = \alpha|0\rangle + \beta|1\rangle \tag{1}$$

ただし α，β はそれぞれ複素数であり，これらの絶対値の 2 乗は状態の確率を表すため，制約条件 $|\alpha|^2 + |\beta|^2 = 1$ が成り立つ．回転ゲートと呼ばれる回路によって，量子の位相をずらすことで状態を変化させたり，量子ビット間に依存関係を持たせることが可能である．1 量子ビットの状態は 2 つの係数によって表せるため，2 次元ベクトルとし表記できる．

$$|\phi\rangle = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \tag{2}$$

この時量子回路は，(2) 式と行列の乗算によって表現可能である．ただし上記の制約により，行列はユニタリ行列である必要がある．量子回路によって生成された，各量子ビットの状態のみでは表現できない関係のことを，量子のもつれという．これは量子計算において，各量子ビットの状態を関連付けするために用いられる．

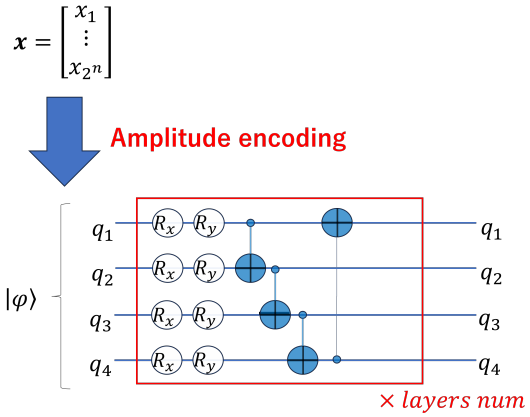

図 1　QNN の概要

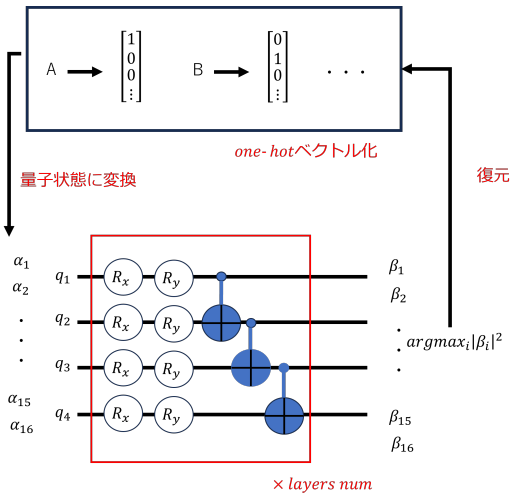

図 2　QNN を用いた言語モデル

2.2　量子機械学習

量子回路には，量子状態に与える変化をパラメータとして指定できるものがある．これをパラメータ付き量子回路と呼ぶ．パラメータ付き量子回路を組み込むことにより，より複雑な量子計算を行うことが可能になる．一方で，望ましい結果を得るには最適なパラメータを探索する必要があり，ランダムな探索では時間がかかりすぎてしまう．こうしたパラメータの探索には古典コンピュータのアルゴリズムが用いられており，この手法は量子古典ハイブリッドアルゴリズムと呼ばれる．その中でも特に量子ニューラルネットワーク (QNN) 等の，古典コンピュータの機械学習に近いアーキテクチャを実装し，そのパラメータを最適化することでタスクを学習する手法を量子機械学習と呼ぶ．

もっとも単純な QNN の一例を図 1 に示す．図中の n は量子ビット数であり，R_x，R_y はそれぞれ回転ゲートを表している．その他の回路は CNOT ゲートを表しており，量子のもつれを生成する．入力ベクトルは Amplitude encoding と呼ばれる手法により，量子状態の振幅へと変換する [5]．その後各量子ビットに対してパラメータを用いた演算を行い，出力は各ビットの値を確認することで得られる．ただし量子状態は，複数の状態の重なりとして存在するため，十分な回数の試行を行ったのち，最終出力はその状態同士の確率分布として得られる．

3　言語モデルへの適用

3.1　言語モデル

近年言語モデルは機械翻訳や質問応答、テキストの要約など，様々な自然言語処理のタスクに応用されている．これはトークンの系列（例えば，単語や文字，サブワードの系列）に対して，後続するトークンを予測を予測するタスクが基盤となっている．代表的な言語モデルである大規模言語モデルの多くは，self-attention を用いた Transformer モデルの構造を持ち、文章の一部を隠しながらその部分の予測を行う MLM（Masked Language Model）または CLM（Causal Language Model）によって事前学習を行っている [6]．一方で，文脈との関係に着目することができる Transformer モデルの機能は Multi-head Attention に依存し、計算量が大きいという難点がある．モデルの大規模化によってその課題はより顕著になっているため，精度を維持しつつ計算効率を向上させることが重要である．

3.2　QNN を用いた文字言語モデル

言語モデルは文の構成要素の並びに対してその同時確率を計算する．特に n-gram モデルでは，各文字は直前の $n-1$ 文字のコンテキストから決定されるように近似が行われる．そこで本研究では，直前の n 文字 $c_{t-n}, c_{t-n+1}, \ldots, c_{t-1}$ から次の単語 c_t の出現確率 $P(c_t|c_{t-n}, c_{t-n+1}, \ldots, c_{t-1})$ を予測する文字 n-gram 言語モデルを実装する．

QNN を適用した言語モデルの概要を図 2 に示す．モデルの入力には各文字の one-hot ベクトルを結合して使用する．その後最も出現頻度の高いビット列を求め，one-hot ベクトルへと復元することで順次文字予測を行う．例えば状態 $|10\ldots00\rangle$ の出力確率が最も高い場合，インデックス 0 の位置が 1 である文字 A へと復元され，$|01\ldots00\rangle$ の出力確率が最も高い場合には，インデックス 1 の位置が 1 である文

字 B へと復元される．

4 実験

4.1 モデルの実装

言語モデルの実装は PennyLane フレームワーク [7] を用いた．PennyLane は Python 用のフレームワークであり，量子シュミレーションにおける高速な計算を実現している．加えて TensorFlow や PyTorch とも併用可能であり，シュミレータ部分においては Amazon braket や IBM Quantum とも連携できる．

4.2 最適化手法

量子コンピュータにおける最適化には大きな課題がいくつか存在する．そこで本節では量子コンピュータの課題に触れつつ，それを補うための最適化手法について説明を行う．まず一つは，観測可能な値と実際の量子状態は異なるという点である．量子の状態は波の位相として表現されるが，観測できるのは 0 と 1 の出現確率のみである．そのため最適化を行う際には，出力から得られる観測値 $\hat{B}$ から量子状態 $\langle\hat{B}\rangle$ を推定する必要がある．パラメータ Θ を持つ量子回路と同値な行列を $U(\Theta)$ とすると，$\langle\hat{B}\rangle$ と $\hat{B}$ には次の関係が成り立つ．ただし，$\langle 0| = |0\rangle^\dagger$ である．また，$\dagger$ はエルミート共役を表す。

$$\langle\hat{B}\rangle(\Theta) = \langle 0|U(\Theta)^\dagger \hat{B} U(\Theta)|0\rangle \tag{3}$$

もう一つの課題は，計算過程の量子を観測すると情報量が減少するため，途中出力を利用できないことである．そのため実機による計算では，量子回路全体をブラックボックスとして，勾配降下法を利用したパラメータ 1 つ 1 つの最適化を順次行う必要がある．ただし通常の勾配降下法では局所的な解に収束する可能性があるため，量子機械学習では $\frac{\pi}{2}$ ずつパラメータをずらすパラメータシフト法が多用される．この手法により，各パラメータの勾配を以下のように得ることができる．

$$\nabla_{\theta_i}\langle\hat{B}\rangle(\theta) = \frac{1}{2}\left[\langle\hat{B}\rangle(\theta + \frac{\pi}{2}\hat{e}_i) - \langle\hat{B}\rangle(\theta - \frac{\pi}{2}\hat{e}_i)\right] \tag{4}$$

一方で本実験ではシュミレータを用いるため，量子情報を破壊することなく途中出力へのアクセスが可能である．PennyLane ではその利点を利用した誤差逆伝搬法が実装されている．そこで本実験では，誤差逆伝搬法を用いたモデルを最適化を行う．PennyLane では Qnode と呼ばれるオブジェクトが定義されており回路全体の演算をノード単位で区切ることができる．Qnode は古典コンピュータのニューラルネットワークにおける中間層に該当する．そしてノード間で勾配を伝搬することにより，全体のパラメータを更新していく．各ノードの勾配計算はデバイス毎に異なるが，シュミレータを用いる場合には効率的なアルゴリズムが存在する [8].

4.3 データセット

言語モデルの評価には，Marcus らが作成した，Wall Street Journal のテキストデータ [9] を使用する．訓練データとテストデータは別々に用意し，モデル規模の制限により，テキスト数はどちらも 10 文のみで計算を行う．

4.4 評価方法

訓練データを用いて学習を行う際には，クロスエントロピー誤差により損失を求める．また最終的な評価として，テストデータを用いて言語モデリングタスクにおける予測性能の指標であるパープレキシティ PPL を以下のように算出する．ただし $|j|$ は全文書の文字数の合計であり，N は文書数である．

$$H = -\frac{1}{|j|}\sum_N\sum_t \log P(c_t|c_{t-n}, c_{t-n+1}, \ldots, c_{t-1}) \tag{5}$$

$$PPL = 2^H \tag{6}$$

これらの指標を用い，$n = 1, 2, 4$ に対する n-gram 量子機械学習モデルの性能を比較する．またベースラインとして，単純に学習データの頻度から最尤推定によって構築した n-gram マルコフモデルによる PPL の算出も行い，参照する文字数を増やした際の PPL の挙動を確認する．

4.5 実験結果と考察

図 3 にマルフモデルと量子機械学習モデルの比較を示した。まずマルコフモデルについて、1-gram モデルは 0-gram より大きくテストセット PPL が改善されている．一方で 2-gram 以降では PPL は増加し続けている．それに対し量子ニューラルネットワークモデルの結果を確認すると，1-gram モデルではマルコフモデル以上に PPL が改善されている．2-gram モデルではそれほど変化がないが，4-gram モデルまで増やすとまた PPL が改善されることが確認できた．

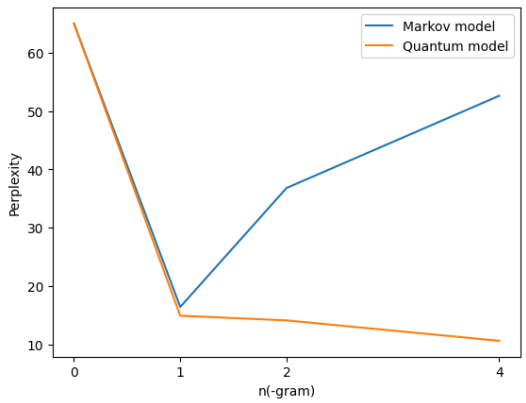

図 3 マルコフモデルと量子モデルのパープレキシティ比較

5 おわりに

本稿では言語モデルにおける量子計算の適用の有用性を検討するため，量子計算を用いた文字 n-gram モデルの実装を行った．実装においては，PennyLane フレームワークを利用した量子計算シュミレータを用いた．そのうえで量子ニューラルネットワークの実装を行い，PennyLane によって提供されている誤差逆伝搬法を使用して最適化を行った．結果として，量子機械学習モデルは $n = 1, 2, 4$ のすべてのモデルにおいて，ベースラインのマルコフモデルより良い精度を得ることができた．加えて参照する文字数を増やした際のパープレキシティの挙動から，量子機械学習モデルは最尤推定に基づくマルコフモデルよりも学習効率が良いことが示唆された．今回は最尤推定との比較を行ったが，今後はその代わりにニューラルネットワークを用いる言語モデルなど，より現代的な言語モデルに近い構造のモデルとの比較を行う．

今後の展望としては，文字ではなく単語予測モデルの実装が挙げられる．それに際して，量子計算とマッチしたエンベディング方法の探索を行う必要があると考えている．また先行研究では量子アテンション [10] も開発されているため，文脈を意識した言語モデルの研究も将来的な目標として残されている．そして最大の目標として，実機の量子コンピュータを用いたモデルの大規模化が挙げられる．量子コンピュータ最大の課題である計算誤差については，アルゴリズムによる誤差抑制へのアプローチもある程度行われている [11]．加えて量子機械学習モデルにおける勾配消失問題についても，観測値から量子状態を復元する shadow tomography[12] を利用した，実機での誤差逆伝搬法 [13] が提案されるなど，日々進展がみられている．最終的にはこれらの技術を活用し，大規模言語モデルと同等のサイズので同等の情報を用いるモデルを実装したうえで，性能の優劣を議論する必要があると考えている．

謝辞

本研究は、JST ムーンショット型研究開発事業 JPMJMS2236 の支援を受けたものです。

参考文献

[1] Kaplan Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. 2020.

[2] Wei Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, and Dani Yogatama. Emergent abilities of large language models. 2022.

[3] He-Liang Huang, Xiao-Yue Xu, Chu Guo, Guojing Tian, Shi-Jie Wei, Xiaoming Sun, Wan-Su Bao, and Gui-Lu Long. Near-term quantum computing techniques: Variational quantum algorithms, error mitigation, circuit compilation, benchmarking and classical simulation. **Science China Physics, Mechanics & Astronomy**, Vol. 66, No. 5, p. 250302, 2023.

[4] Asel Sagingalieva, Mohammad Kordzanganeh, Nurbolat Kenbayev, Daria Kosichkina, Tatiana Tomashuk, and Alexey Melnikov. Hybrid quantum neural network for drug response prediction. **Cancers**, Vol. 15, No. 10, p. 2705, 2023.

[5] Weikang Li, Zhide Lu, and Dong-Ling Deng. Quantum neural network classifiers: A tutorial. **Neurocomputing**, Vol. 470, pp. 457–461, 2022.

[6] Jiayi Fu, Lei Lin, Xiaoyang Gao, Pengli Liu, Zhengzong Chen, Zhirui Yang, Shengnan Zhang, Xue Zheng, Yan Li, Yuliang Liu, et al. Kwaiyiimath: Technical report. **arXiv preprint arXiv:2310.07488**, 2023.

[7] Ville Bergholm, Josh Izaac, Maria Schuld, Christian Gogolin, Shahnawaz Ahmed, Vishnu Ajith, M Sohaib Alam, Guillermo Alonso-Linaje, B AkashNarayanan, Ali Asadi, et al. Pennylane: Automatic differentiation of hybrid quantum-classical computations. **arXiv preprint arXiv:1811.04968**, 2018.

[8] Tyson Jones and Julien Gacon. Efficient calculation of gradients in classical simulations of variational quantum algorithms. **arXiv preprint arXiv:2009.02823**, 2020.

[9] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. **Computational Linguistics**, Vol. 19, No. 2, pp. 313–330, 1993.

[10] Guangxi Li, Xuanqiang Zhao, and Xin Wang. Quantum self-attention neural networks for text classification. **arXiv preprint arXiv:2205.05625**, 2022.

[11] Suppressing quantum errors by scaling a surface code logical qubit. **Nature**, Vol. 614, No. 7949, pp. 676–681, 2023.

[12] Scott Aaronson. Shadow tomography of quantum states. In **Proceedings of the 50th annual ACM SIGACT symposium on theory of computing**, pp. 325–338, 2018.

[13] Amira Abbas, Robbie King, Hsin-Yuan Huang, William J Huggins, Ramis Movassagh, Dar Gilboa, and Jarrod R McClean. On quantum backpropagation, information reuse, and cheating measurement collapse. **arXiv preprint arXiv:2305.13362**, 2023.