量子計算を用いたダイレクトモデル

三輪拓真 [1,3]　小田 悠介 [2,3]　河野誠也 [3,1]　吉野幸一郎 [4,3,1]
[1] 奈良先端科学技術大学院大学　[2] 国立情報学研究所 大規模言語モデル研究開発センター
[3] 理化学研究所ロボットガーディアンプロジェクト　[4] 東京科学大学 情報理工学院

概要

複数のモジュールでの処理を順次通過するようなタスクでの一般的な実装として，カスケードモデルと End-to-End モデルが存在する．カスケードモデルは汎用性が高く，学習データも豊富に存在する一方で，中間出力において情報が一部欠損してしまう．End-to-End モデルは入力から最終出力まで一貫して計算を行うため中間出力における情報欠損が少ないが，既存の訓練データが少なく，学習コストが高い．本研究では両者の課題にアプローチするため，量子計算の性質に着目した．量子計算を用いた量子機械学習モデルをそれぞれ独立に訓練し，推論時のみ回路を結合させることで，量子状態の観測を行わずに中間出力の受け渡しを行う．古典コンピュータに比べて豊富な表現力を持つ量子ビットを中間出力に用いることで，モデル間の情報欠損を抑制する．本研究では対話状態追跡タスクを用いて提案法の検証を行い，今後の課題について考察を示した．

1　はじめに

近年機械学習モデルは，実社会適用を意識した複雑なタスクでの成果を期待されている．そうしたタスクは一般に複数のモジュールを用いて実装される．複数のモジュールでの処理を順次通過するようなタスクでの一般的な実装として，カスケードモデルと End-to-End モデルが挙げられる．

カスケードモデルは複数のモジュールを結合させてタスクに適用する．この手法は各モデルを分離して学習可能なため汎用性が高い．一方で各モデル間の結合はモデルの入出力であるテキストなどの離散シンボル列を用いるため，機械学習モデルが内部に持つパラメータの情報は一部欠損し [1]，またモジュールごとの誤差が伝搬していく [2] という問題がある．前モデルの出力を複数入力する n-best[3] や，出力を実数ベクトルのまま次モデルに入力する tight integration[4] 等の対策が存在するが，依然として情報欠損の課題は解決されていない．

End-to-End モデルはディープラーニング研究の発展により，大きな成果を得た手法である．ニューラルネットワークを用いてモジュールを結合・通貫して学習させることで，入力から最終出力までを一度に計算できる．この手法は中間出力による情報の欠損が無く，前モデルからの誤り伝搬も無いため，カスケードモデルより高い精度が期待される．実際に音声翻訳の領域では，韻律情報を保ったまま翻訳を行うことで，カスケードモデルより精度を向上させたケースも存在する [5]．しかしながら標準的な評価データセットでは，カスケードモデル優位になることが多い [2]．これは End-to-End モデルの学習には入出力に対応したペアデータが必要で，これまで個々のモジュール用に整備されてきたデータをうまく利活用できないためである [6]．このように End-to-End モデルは高い性能を示す可能性がある一方で，その取り回しに課題がある．

本研究ではカスケードモデルの情報欠損と End-to-End モデルの学習コストという課題にアプローチするため，量子計算の性質に着目する．量子機械学習モデルはカスケードモデルのように各モデルを独立に学習しつつ，中間出力を量子状態で受け渡し可能である．そのため End-to-End モデルが抱えるモジュールの取り回しの問題を解決しながら，モジュール間のエラー伝搬や情報欠損の問題を解消できる可能性がある．具体的には前段モデルの最終層及び後段モデルの先頭層に量子機械学習モデルを取り入れることで，中間出力における情報欠損の抑制を行う．本論文では，前段モデルを音声認識，後段モデルを対話状態追跡とするようなタスク設定において，量子機械学習を用いたダイレクトモデルを提案し，その効果を検証した．

2　対話状態追跡

対話状態追跡（Dialogue State Tracking; DST）は特定のタスク解決を行うタスク指向対話システム

（Task-oriented Dialogue; TOD）を構築する際のタスクである．TOD はレストランやホテルの予約などの固定されたタスクでユーザを対話的に支援するものである [7]．DST は TOD において音声で与えられるユーザ発話の系列から, 現在のユーザの状態をフレームとして推定・追跡するタスクである．

DST は slot と呼ばれる情報のカテゴリと，value と呼ばれる各カテゴリの値によって情報を管理する．また value は分類なものと非分類なものに分けられる．例えば曜日は分類なものだが，日付は非分類なものである．こうした情報の組み合わたフレームを認識することが DST のゴールである．

実装方式については，近年音声 DST はカスケードモデルによるものが主流である．これは音声認識モデル (Automatic Speech Recognition; ASR) とテキスト DST モデルは別々に扱うことができ，両者は効果的に統合できると仮定されているためである [8]．一方で本論文で課題としているように，カスケードモデルには情報の欠損が発生する．加えて近年ディープラーニングの発展により，ASR モデルとテキスト DST モデルはいずれも大きく改善されたが，音声 DST モデルは未だ大きな改善を見せていない [9]．そこで本研究では音声 DST を対象としたダイレクトモデルの検証を行い，提案法の有用性について考察を行う．

3　量子計算

3.1　量子の性質

量子計算は量子力学の性質を利用し，古典コンピュータとは根本的に異なる手法で情報処理を行う技術である．量子計算で用いるビットは量子ビットと呼ばれ，古典ビットとは大きく異なる性質を持つ [10]．本節では量子機械学習において特に重要となる重ね合わせの性質に触れながら，量子計算の基本について述べる．

3.1.1　重ね合わせの性質と量子ビット演算

量子ビットは値 0 と 1 が確率的に共存しており，観測することで初めて値が確定する．量子の状態はブラケット記号 $|*\rangle$ を用いて表記される．よって 1 量子ビットの量子状態を $|\phi\rangle$ とすると，$|\phi\rangle$ は複素数係数 α，β を用いて，次のように表記できる．

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

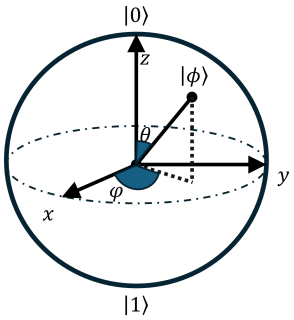

図 1　ブロッホ球

確率分布は α，β それぞれの絶対値の 2 乗に等しい．よって α，β の間には，制約条件 $|\alpha|^2 + |\beta|^2 = 1$ が成り立つ．この制約条件と (1) 式から次式が導ける．

$$|\phi\rangle = e^{i\gamma}\left(\cos\frac{\theta}{2}|0\rangle + e^{i\varphi}\sin\frac{\theta}{2}|1\rangle\right). \tag{2}$$

ただし，$e^{i\gamma}$ は観測可能な効果をもたらさないため，本研究では取り除いて考える．その場合 (2) 式により，$|\phi\rangle$ は θ 及び φ によって定義される，3 次元単位球上の点とみなすことができる．それを示した図 1 をブロッホ球と呼ぶ．このように，量子状態 $|\phi\rangle$ は $|0\rangle$ と $|1\rangle$ の連続した状態の中に存在している [11]．本研究では量子ビットの持つ豊富な表現力を利用し，カスケードモデルにおけるモデル間の情報欠損の抑制が可能かの検証を行う．

3.2　量子機械学習

量子計算の理論研究は多様な発展を期待される一方で，現在のハードウェアの制限から，実際の適用は非常に限られたものになっている．ハードウェア上の大きな課題として，量子ビット数の制限や，計算誤差を多く含んでしまう点が挙げられる．このような性質を持つ量子コンピュータを NISQ (Noisy Intermediate-Scale Quantum) と呼び，量子コンピュータの過渡期とされている [12]．NISQ 時代における量子計算の適用先として，近年では量子機械学習の研究が行われている．量子機械学習は量子への作用を変数を用いて指定可能な変分回路 $U(\theta)$ を用いて実装される．目的タスクにおける最適パラメータ θ_{optim} を求め，初期状態 $|00\ldots0\rangle$ に対して適用することで，最適解 $|\phi_{optim}\rangle = U(\theta_{optim})|00\ldots0\rangle$ を得る．一方で，そのような最適パラメータを経験則から得るのは非常に困難である．そこで観測結果と理想的な出力の差分を求める損失関数 $\mathcal{L}(\theta)$ を定義し，それを最小化することで近似解を求める．量子計算

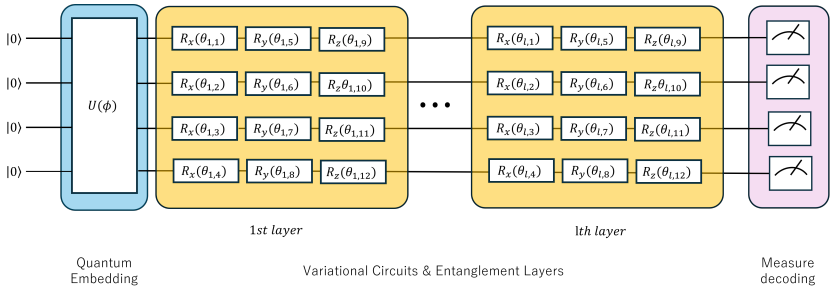

図 2　中間層に利用する量子回路

と古典コンピュータの両方を用いることから，この手法はハイブリッド法と呼ばれる [13]．量子機械学習は要求する量子ビット数が比較的少なく，計算誤差に対してもパラメータ調整によってある程度対策できるため，NISQ の適用先として期待されている [14]．以降では量子機械学習の構成要素となる量子エンベディングとパラメータ付き回路について説明する．

3.2.1　量子エンベディング

量子計算を行うには古典ビットから量子ビットへ入力データを変換する必要があり，この処理を量子エンベディングと呼ぶ．古典データを正しく量子状態に変換することは，学習モデルの精度にとって非常に重要である [15]．本研究では入力ベクトルを量子状態の振幅へと変換する Amplitude embedding[16] を採用する．Amplitude embedding の処理を以下に示す．

$$|\phi\rangle = \sum_{i=1}^{N} x_i |i\rangle \tag{3}$$

Amplitude embedding は結果の観測に $O(2^n)$ の時間が必要であり，ノイズが入りやすいという欠点がある [17]．一方で Amplitude embedding は N 次元ベクトルの入力 x を，$n = \log_2(N)$ 量子ビットで扱うことができ，量子エンベディングの中で最も量子ビット効率の良い手法である．本研究では量子ビット効率の観点からこの手法を採用した．

3.2.2　パラメータ付き回路

与えられたパラメータに応じて量子の状態を変化させる回路をパラメータ付き回路と呼ぶ．量子機械学習ではパラメータ付き回路を用いてモデルの実装を行う．代表的なパラメータ付き回路として R_x ゲート [11] が挙げられる．R_x ゲートはブロッホ球上の x 軸を軸に量子の位相をパラメータ θ だけ回転させる．同様の回路 R_y，R_z ゲートを用いて回路の実装を行う．本研究では量子ニューラルネットワークモデル (QNN) を参考に図 2 のような回路を用いた．QNN はニューラルネットワーク (NN) をもとに提案されたアーキテクチャであり，量子状態を調節するパラメータ付き回路と，量子ビット間に関連性を持たせる CNOT ゲートによって構成される．CNOT ゲートは 2 量子ビットを入力とし，決められた一方の量子ビットが 1 であればもう一方を反転させる．CNOT ゲートは量子もつれを発生させ量子回路の情報量を増加させる利点がある一方で，量子の状態を大きく変化させてしまう．本研究では古典モデルを事前学習モデルとして用いており，ファインチューニングには繊細なパラメータ調整が必要となるため，CNOT ゲートは用いない．この回路を各タスクにおいて独立に訓練し，推論時にはパラメータを連結することで，モデル間で量子状態の観測を行わずに中間出力の受け渡しを行う．

4　量子計算を用いた音声 DST

本研究で提案するダイレクトモデルの概要を説明する．図 3 に提案法の概要を示した．ダイレクトモデルはカスケードモデル同様各モデルを独立に訓練する．本研究では音声認識モデルと対話状態追跡モデルをそれぞれ独立に訓練する．そのため一般的なデータセットのみで訓練可能であり，かつ各モデルは並列に訓練できる．その後学習を行った前段モデルの最終層と，後段モデルの第一層に図 2 量の回路を持つ量子機械学習モデルを追加し，それぞれファインチューニングを行う．

推論の際には各モデルの量子回路を連結し，回路に学習済みのパラメータを適用する．モデル間では量子状態を観測せず，重ね合わせの状態を保ったまま出力を次モデルへと受け渡す．これにより中間出力の観測による情報の欠損を抑制し，最終モデルの出力のみ観測して目標タスクに適用する．

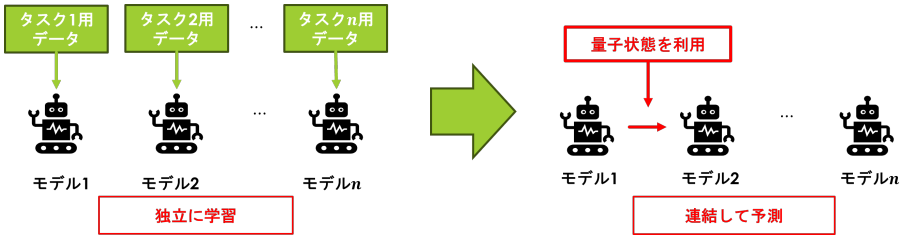

図 3　ダイレクトモデルの概要

5　実験

5.1　実験詳細

古典モデルのみからなるカスケードモデルと量子・古典ハイブリッドモデルをそれぞれ実装し，精度の比較を行った．比較検証には DSTC2 を用いた．DSTC2[18] は人間と機械の対話データを用いた対話状態追跡タスクであり，音声データも含まれる．

まず古典 ASR モデルには Conformer[19] を使用した．Conformer は CNN と Transformer を組み合わせたアーキテクチャであり，ASR タスクにおいて優れた性能を発揮する．本研究では量子回路と結合するため，リニア層を用いて次元数を調整した．学習率 5×10^{-6}，エポック数 400 で学習を行った．その後 10 層の量子機械学習モデルを用いてハイブリッドモデルのファインチューニングを行った．この際学習率 1.0×10^{-3}，エポック数 20 とした．

次に DST モデルについて，古典モデルには CNN モデルを採用した．学習率 1.0×10^{-3}，エポック数 20 で学習を行った．その後 5 層の量子機械学習モデルを用いてファインチューニングを行った．また事前学習を行った古典モデルの精度を維持するため，量子モデルの初期値は全て 0 とした．

両モデルの学習には DSTC2 に含まれるデータ個の学習には DSTC2 に含まれる音声データ 8579 件を用いた．ASR におけるモデル単体での音声認識の精度を示す WER は 37.7 であり，DST モデルのテキスト DST 単体での精度 (Accuracy) は 88.4 であった．性能評価には DSTC2 に含まれるテストデータのうち 8966 件を用いた．

5.2　結果と考察

実験結果を表 1 に示した．ただし，DST only は音声認識が正しく行われた場合の DST モデルの精度である．まず DST モデルのみの精度と古典カスケードモデルの精度を比較すると，精度は著しく低下した．これは ASR モデルの認識誤りや中間出力における情報欠損が原因と考えられる．この問題に対して，提案を行った量子・古典ハイブリッドモデルは更に 4.9 ポイント精度が低下した．ただし，量子回路の連結による出力の受け渡しには成功していることが見て取れる．今後はスコアの差として考えられるパラメータ数の制約や，よりエラーの伝搬に強いアーキテクチャを考察しつつ古典カスケードモデルを上回る精度を目指す．

表 1　DSTC2 タスクにおける精度比較

Task	Model	Accuracy
テキスト DST	DST only	88.4
音声 DST	古典カスケード	46.9
	量子・古典ハイブリッド	41.8

6　結論

本研究では量子状態を維持したまま中間出力の受け渡しを行うダイレクトモデルを提案し，音声を入力とする対話状態追跡タスクにおいて検証を行った．DSTC2 データセットを用いて音声認識モデル，対話状態追跡モデルの訓練を行い，その後量子機械学習モデルを用いて入出力の結合を行うダイレクトモデルの構築を行った．結果として量子状態を維持することによる精度の改善は確認できなかったものの，量子状態の受け渡しには成功した．精度が低下した原因の一つとしては，音声認識とテキスト DST の前後に挟んだ量子機械学習モデル自体の誤差が挙げられる．そこで今後の方針としては，対話状態追跡の先行研究でも行われているように，訓練データに対してノイズを追加することで，ノイズに耐性を持たせる手法の追加実装が考えられる．また量子計算の観点からも，量子計算には計算誤差が多く発生するという背景から，ノイズへのアプローチが多く存在する．そうした手法が本研究のアーキテクチャへ有用かの検証も行っていく．

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] Evgeny Matusov, Stephan Kanthak, and Hermann Ney. On the integration of speech recognition and statistical machine translation. In **Interspeech**, pp. 3177–3180, 2005.

[2] Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. Cascade versus direct speech translation: Do the differences still make a difference? In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 2873–2887, Online, August 2021. Association for Computational Linguistics.

[3] Jihao Yin, Xiaozhong Fan, Kaixuan Zhang, and Jiangde Yu. Chinese organization name recognition using chunk analysis. In **The 20th Pacific Asia Conference on Language, Information and Computation: Proceedings of the Conference**, Vol. 20, pp. 347–353. Waseda University, 2006.

[4] Parnia Bahar, Tobias Bieschke, Ralf Schlüter, and Hermann Ney. Tight integrated end-to-end training for cascaded speech translation. In **2021 IEEE Spoken Language Technology Workshop (SLT)**, pp. 950–957. IEEE, 2021.

[5] Giulio Zhou, Tsz Kin Lam, Alexandra Birch, and Barry Haddow. Prosody in cascade and direct speech-to-text translation: a case study on Korean wh-phrases. In Yvette Graham and Matthew Purver, editors, **Findings of the Association for Computational Linguistics: EACL 2024**, pp. 674–683, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[6] Thierry Etchegoyhen, Haritz Arzelus, Harritxu Gete, Aitor Alvarez, Iván G. Torre, Juan Manuel Martín-Doñas, Ander González-Docasal, and Edson Benites Fernandez. Cascade or direct speech translation? a case study. **Applied Sciences**, Vol. 12, No. 3, 2022.

[7] Jianfeng Gao, Michel Galley, and Lihong Li. Neural approaches to conversational AI. In Yoav Artzi and Jacob Eisenstein, editors, **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts**, pp. 2–7, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[8] Yue Feng, Yang Wang, and Hang Li. A sequence-to-sequence approach to dialogue state tracking. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 1714–1725, Online, August 2021. Association for Computational Linguistics.

[9] Lucas Druart, Léo Jacqmin, Benoît Favre, Lina Maria Rojas-Barahona, and Valentin Vielzeuf. Are cascade dialogue state tracking models speaking out of turn in spoken dialogues? Submitted to IEEE ICASSP 2024© 2023 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works., November 2023.

[10] Josh Schneider and Ian Smalley. What is quantum computing?

[11] Michael A. Nielsen and Isaac L. Chuang. **Quantum Computation and Quantum Information: 10th Anniversary Edition**. Cambridge University Press, 2010.

[12] John Preskill. Quantum computing in the nisq era and beyond. **Quantum**, Vol. 2, p. 79, 2018.

[13] Maria Schuld and Francesco Petruccione. **Machine learning with quantum computers**, Vol. 676. Springer, 2021.

[14] Yunfei Wang and Junyu Liu. A comprehensive review of quantum machine learning: from nisq to fault tolerance. **Reports on Progress in Physics**, 2024.

[15] Seth Lloyd, Maria Schuld, Aroosa Ijaz, Josh Izaac, and Nathan Killoran. Quantum embeddings for machine learning. 1 2020.

[16] Mansoor A. Khan, Muhammad N. Aman, and Biplab Sikdar. Beyond bits: A review of quantum embedding techniques for efficient information processing. **IEEE Access**, Vol. 12, pp. 46118–46137, 2024.

[17] Emmanuel Ovalle-Magallanes, Dora E Alvarado-Carrillo, Juan Gabriel Avina-Cervantes, Ivan Cruz-Aceves, and Jose Ruiz-Pinales. Quantum angle encoding with learnable rotation applied to quantum–classical convolutional neural networks. **Applied Soft Computing**, Vol. 141, p. 110307, 2023.

[18] Matthew Henderson, Blaise Thomson, and Jason D Williams. The second dialog state tracking challenge. In **Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)**, pp. 263–272, 2014.

[19] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. **arXiv preprint arXiv:2005.08100**, 2020.