質問文から連想した画像特徴量を用いた質問応答モデル

石橋 陽一 §　森 泰 †　木村 輔 ‡　宮森 恒 ‡
§奈良先端科学技術大学院大学 先端科学技術研究科
†京都産業大学 コンピュータ理工学部
‡京都産業大学 先端情報学研究科
§ishibashi.yoichi.ir3@is.naist.jp, †‡{g1545466, i1658047, miya}@cc.kyoto-su.ac.jp

1　はじめに

近年, 質問文から応答を生成する研究が盛んに行われている. しかしこれらの研究はテキスト情報のみを元にした応答を生成するため, 正答するために視覚情報が有用に働くと考えられる質問に必ずしも効果的な応答ができるとは限らない. 例えば, “how many legs does a spider have ?”という問に対して, “three , i think .”と応答したことが報告されている [1]. 近年は画像を用いた質問応答の研究も行われている. VQA[2] は画像を用いた質問応答タスクであり, ある画像についての質問に回答することを目指している. しかし一般的に質問応答システムにおいて画像が常に与えられるケースは限定されており, VQA に有用なモデルを, 通常のテキスト質問文に応答するシステムにそのまま活用することは困難である. したがって, 視覚情報は質問応答に有用であると考えられるが, テキスト質問のみに応答するシステムでは VQA での画像に相当する情報を何らかの形で活用することが正答率改善につながるのではないかと考えた.

そこで本研究では, 画像の代用として, 入力テキストから生成（連想）した視覚情報を用いることで, テキスト入力テキスト出力の構造を保持したまま視覚情報を活用する質問応答を目指した.

我々は, 連想を用いた質問応答のためのデータセット (iNatVQA 図 1) を新たに作成し, 実験を行った. 実験では, 連想を行う機構として単純な多層パーセプトロンを用いて視覚情報を生成し, 応答予測に用いた. 現時点での結果として, 提案手法はベースラインの精度を下回っている. 原因として, 多層パーセプトロンが視覚情報の生成に成功していないことが考えられる.

図 1: iNatVQA データセットの一例

2　データセット

2.1　目的

画像を用いた質問応答のタスクとして VQA[2] があるが, 正解するためには画像が必要不可欠である. 例えば, 図 2 において, VQA の質問は画像中の特定の物体 (白い風船) についての質問であり, 正しく回答するためには画像が必要不可欠である. したがって, 文の情報 (What color are the balloons ?) から 特定の物体 (白い風船) の視覚情報を生成することは困難であるため, 連想の効果を検証するには適したデータセットではない. そこで, 従来手法と提案手法の効果を定量的・定性的に比較するため, 生物の画像と学名のデータセットである iNaturalist[3] のデータを利用して新たな VQA データセット (iNatVQA) を作成した. このデータセットでは, 画像は与えられているが, 生物名に紐付いた一般的な視覚情報という位置づけでの画像である. したがって生物名と画像が自然に関連

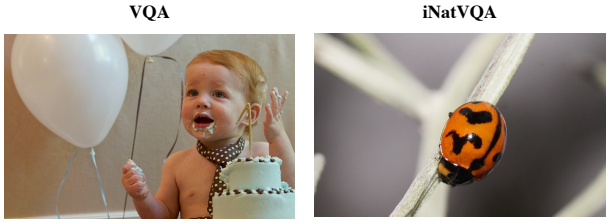

図 2: VQA と iNatVQA の比較. VQA は特定の画像についての質問であり, 回答するためには特定の画像情報 (白い風船)が必要不可欠である. 一方, iNatVQA では質問文に含まれる生物名 (*Coccinella-Transversalis*) の一般的な画像情報 (生物画像)が付与されているため, 質問文から画像情報を連想することが可能である.

しており, 質問文から画像特徴量を連想する際に利用することが可能である (図 2). また, iNatVQA の質問は SQuAD[4] などの画像を利用しないデータセットとは異なり, 画像情報が正答に有用であると期待される質問から構成されている. したがって iNatVQA を用いることで 3 つの手法 (画像を用いない質問応答手法 (seq2seq など), 画像を用いる手法 (VQA), 画像の代わりに連想を用いる手法 (提案手法)) の定量的・定性的比較が可能になる.

2.2 作成方法

表 1 の質問タイプごとにテンプレートを用意し, 生物名と iNaturalist のデータから抽出した回答をペアとし質問と回答を作成した. また我々は iNaturalist のデータに加え, 生物の視覚的特徴を陽に表す「色情報」を質問に加えるため人手でアノテーションを行った. データ数は, 学習データとして 2.7M, テスト用と開発用データが各 0.1M である. また重要な制約として, 生物に関して 学習用, テスト用, 開発用データに同じタイプの質問は存在しないように設計されている. 例えば, *Araneus Diadematus* という生物に関して, 鱗があるか問う質問 (Does a *Araneus Diadematus* have scales ?) は, 学習データにのみ存在する. また, 色の質問 (What color is a *Araneus Diadematus* ?) はテストデータにのみ存在する. また, 毛があるか問う質問 (Does a *Araneus Diadematus* have hairs ?) は開発データにのみ存在する. つまり, 学習データで学習したモデルは, *Araneus Diadematus* に鱗があるか

表 1: 各質問タイプごとのテンプレート. [name] は学名を表す.

Question type
What color is a [name] ?
What super-category does a [name] belong ?
How many legs does a [name] have ?
Where does a [name] live ?
Does a [name] have a beak ?
Does a [name] have wings ?
Does a [name] have feathers ?
Does a [name] have scales ?
Does a [name] have fins ?
Does a [name] have hairs ?
Does a [name] have legs ?
Does a [name] have eyes ?
Does a [name] have leaves ?
Does a [name] have branches ?

どうか学習しているが, 色や枝があるかどうかは学習していない. したがって, 画像も連想も用いない手法 (seq2seq) では, 色や枝の有無についての質問に正解できない. しかし連想を用いる手法 (提案手法) では, 生物名からその生物の視覚情報を連想できるため, 色や枝の有無の質問に正解することが可能であると考えられる. つまり iNatVQA データセットは, 視覚情報を用いない場合, 他の網羅的な生物学的知識を活用するなどしない限り, テスト・開発データには正解できない設計になっている. 以上のように, iNatVQA データセットは, 提案手法の効果を定量的・定性的に評価するためのデータセットである.

3 連想を用いた質問応答

3.1 概要

図 3 にモデルの概要を示す. 本研究の最終目的は, 質問文から生成された画像特徴量を用いることで, テキスト情報だけでは得られない, より適切な応答の生成を可能にすることである. そこで, 我々は以下のアプローチを取ることでこの問題の解決を目指した.

- はじめに, 入力文を Encoder によって文ベクトルに符号化する

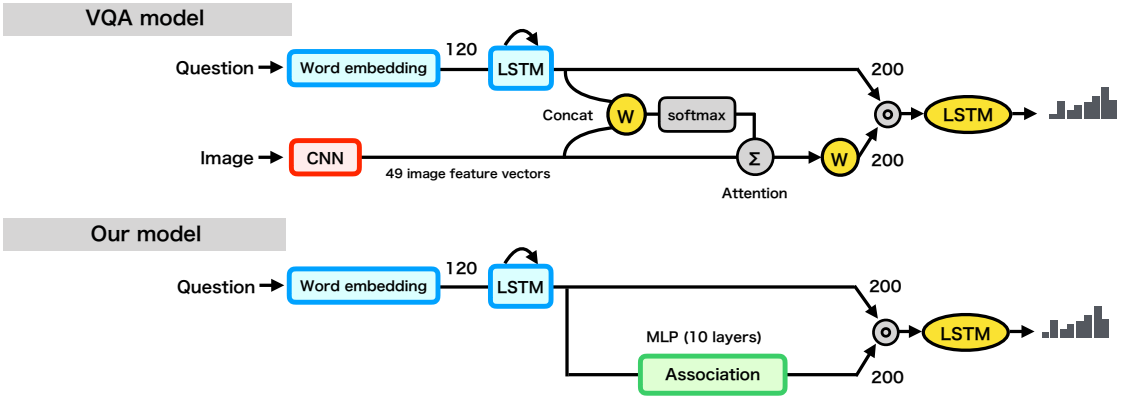

図 3: 一般的な VQA モデル (上) と提案手法 (下) のネットワーク構造. VQA モデルは入力に画像を用いるが, 提案手法では画像入力のない質問応答に対応するため, 画像の代わりに文ベクトルから画像特徴量を生成する連想符号化器 (Association) を備えている.

- 次に, 文ベクトルから文と関連した画像特徴量を生成する
- 最後に, 文ベクトルと画像特徴量を元に応答を生成する

一般的な VQA モデルは文を文ベクトルに符号化する Encoder と, 文ベクトルをクエリにし画像の特徴マップに Attention を行い画像特徴量を得る機構, そして文ベクトルと画像特徴量を用いて応答を予測する多層パーセプトロンからなる. しかし本研究では画像の与えられない質問応答システムに対応するため, 質問文しか入力されない. そこで提案手法では文ベクトルから画像特徴量を生成する連想符号化器を用いることで, 文入力のみであるが視覚情報も考慮し応答を予測可能にする. この連想符号化器は画像を用いる学習済みの VQA モデルから抽出した文ベクトルと画像特徴量を用いて独立に学習を行う. 最終的に学習された連想符号化器を用いて提案手法を構築し応答を学習する. したがって学習は 3 ステップに別れ, ステップ 3 で提案手法の学習を行うために, ステップ 1 とステップ 2 において 2 つのモデルの学習をする必要がある.

3.2 ステップ 1 : 文ベクトルと画像特徴量の学習と抽出

ここで利用するモデルは一般的な VQA モデルである (図 3 上). このステップで利用する学習データは, 質問文, 質問文に対応する動画, および, 質問文に対する応答である. 質問文 X_{txt} と, 質問文に対応する画像 X_{vis} を入力とし, 応答 Y を出力する一般的な VQA モデルの学習を行う. この学習済みモデルを用いて文ベクトル C^{txt} と画像特徴量 C^{vis} を抽出する.

3.3 ステップ 2 : 連想符号化器の学習

ステップ 2 では文ベクトル C^{txt} を連想符号化器に入力し, 画像特徴量 C^{vis} を予測させる学習を行う (図 3 緑). 以前行った連想を用いた対話応答文生成の研究 [5] では連想符号化器として多層パーセプトロンを用いて, その効果が確かめられたため, 今回も同様に, 連想符号化器として多層パーセプトロンを用いた. 連想符号化器は入力が文のベクトル C^{txt}, 教師が画像特徴量のベクトル C^{vis} の回帰問題を解くモデルである.

$$\hat{C}^{vis} = MLP(C^{txt}) \tag{1}$$

損失関数には平均二乗誤差を用いた.

3.4 ステップ 3 : 連想を用いた応答生成

ステップ 3 では, 提案手法が画像特徴量のかわりに連想を用いて応答を学習する (図 3 下). 提案手法への入力は質問文 X_{txt}, 出力は応答 Y である. 提案手法ではステップ 1 の事前学習モデルの CNN を除去し, 代わりに連想符号化器を用いる. なお, 質問文の Encoder と連想符号化器の重みは固定し, 学習しない.

4 実験

この章ではベースラインと提案手法の比較結果について述べる．ベースラインには seq2seq[6] と，ステップ 1 の画像を用いる VQA モデルを用いる．表 2 はテストデータでの精度を示す．画像を用いる VQA モデルの精度が最も高く，画像を用いる手法が最良であることがわかる．次に画像を用いない手法（seq2seq・提案手法）では，提案手法よりも視覚情報を一切用いない seq2seq の精度が高い．

表 2: 各手法の精度比較

Model	Accuracy (%)
seq2seq	55.74
VQA	59.54
提案手法	53.44

ここで提案手法の精度が seq2seq よりも低かった原因は，連想符号化器が画像特徴を完全に予測できていない点であると考えられる．仮に 連想符号化器が画像特徴を完全に予則できていたとすると，VQA と同程度の精度が出るはずである．今回，連想符号化器として単純な多層パーセプトロンを用いたが，今後は GAN を用いて生物名から生物の画像そのものを連想し，応答に用いる手法を計画している．

5 関連研究

文生成アルゴリズムでは，画像と文を同時に与えることで有益な文生成をする試みがある [7] [8]．本研究では VQA で成功している手法をベースに，連想を用いてテキストのみの入力に対応させることを想定している．したがってこれらの研究で用いられた手法は現段階では検討していない．

訓練時にのみ画像を用いるというアプローチでは，[8] が本研究と類似した手法を採用している．Elliott らのネットワーク構造は，文の Encoder と文の Decoder，画像の予測器からなり，文を入力し文の Decoder で翻訳文を予測させる．同時に，Encoder の隠れ層を平均して，画像の予測器に入力し，画像特徴を予測させる．[8] と提案手法とで精度にどの程度差があるか，今後検証していく予定である．

6 結論

一般的に質問応答システムにおいて画像が常に与えられるケースは限定されており，VQA に有用なモデルをそのまま活用することは困難である．したがって，視覚情報は質問応答システムに有効であるが実際のシステムで画像を用いるケースは少ない，という相違が生じている．そこで本研究では，画像の代用として，入力テキストから生成（連想）した視覚情報を用いることで，テキスト入力テキスト出力の構造を保持したまま視覚情報を活用する質問応答を目指した．実験では，連想を行う機構として単純な多層パーセプトロンを用いて視覚情報を生成し，応答予測に用いた．現段階の結果として，提案手法はベースラインの精度を下回っている．原因として，多層パーセプトロンが視覚情報の生成に成功していないことが考えられる．

謝辞

本研究の一部は科研費 18K11557 の助成を受けたものです．ここに記して感謝の意を表します．

参考文献

[1] Oriol Vinyals and Quoc V. Le. A neural conversational model. *CoRR*, Vol. abs/1506.05869, , 2015.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

[3] Grant Van Horn, Oisin Mac Aodha, Yang Song, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist challenge 2017 dataset. *arXiv preprint arXiv:1707.06642*, 2017.

[4] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[5] 石橋陽一, 宮森恒. 連想対話モデル: 発話文から連想した視覚情報を用いた応答文生成. *DEIM Forum*, 2018.

[6] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 3104–3112, 2014.

[7] Hideki Nakayama and Noriki Nishida. Zero-resource machine translation by multimodal encoder-decoder network with multimedia pivot. *Machine Translation*, Vol. 31, No. 1-2, pp. 49–64, 2017.

[8] Desmond Elliott and Ákos Kádár. Imagination improves multimodal translation. *CoRR*, Vol. abs/1705.04350, , 2017.