# テキストと関連画像の視覚的要素を活用した質問応答

森 泰§　　石橋　陽一†宮森 恒‡
§京都産業大学 コンピュータ理工学部
†奈良先端科学技術大学院大学 先端科学技術研究科
‡京都産業大学大学院 先端情報学研究科

§g1545466@cc.kyoto-su.ac.jp, †‡ishibashi.yoichi.ir3@is.naist.jp, miya@cc.kyoto-su.ac.jp

## 1　はじめに

本稿では，質問応答システムにおいて，テキストだけでなく，それに関連した画像を取得し，その視覚的要素を活用して自動回答する手法を提案する．従来，質問応答システムでは，質問がテキストで与えられ，テキストで記述された知識源を参照し，テキストの回答が出力されることが多い．一方，我々は，質問によっては，色や形，数など，一旦，視覚的要素を連想することで，より正確な回答に繋がると期待されるタイプの質問も存在すると考えた．そこで，本稿では，質問文から情報検索で得られる関連画像の特徴表現 (視覚的要素) を，質問文と融合的に活用することで，回答を生成する手法を提案する．実験では，テキストのみを用いた回答手法と，視覚的要素を用いた提案手法とで，正解率がどの程度改善されるか検証する．

## 2　データセット

画像を用いた質問応答のタスクとしては VQA[2] が知られているが, VQA では, 特定の画像とそれに関連する質問が入力として与えられることを前提としている．そのため, 本稿で想定するような, 質問文の情報から一般的, 典型的な視覚情報を連想して活用するためのデータセットとしては適切ではない．そこで, 一般的, 典型的な視覚情報を連想し活用した場合の効果を検証するため, 動植物の画像と学名のデータセットである iNaturalist[5] 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．iNatVQA では, 質問タイプごとにテンプレートを用意し, 生物名と iNaturalist のデータから作成された質問と回答から構成される．データ数は, 学習データとして 2.7M, 開発用とテスト用データが各 0.1M である [3].

## 3　視覚的要素を用いた質問応答

提案手法の目的は, 質問文から連想された画像の特徴表現を用いることで, テキスト情報だけでは得られない, より適切な回答が得られることを示すことである．そこで, 我々は質問文から情報検索で得られる関連画像の特徴表現 (視覚的要素) を，質問文と融合的に活用することで，より適切な回答を生成することを目指した．一般的な VQA モデルでは, 文を文ベクトルに符号化する Encoder と, 文ベクトルをクエリとし, 画像の特徴マップに Attention を適用し画像の特徴表現を得る機構, そして文ベクトルと画像の特徴表現を用いて回答を予測する全結合ニューラルネットからなる．しかし, 本稿では, 特定の画像が与えられず, 質問文のみが与えられる質問応答を想定している．そこで, 提案手法では, 一般的な VQA モデルをベースとし, 質問文をクエリとして画像検索することで得られる画像を用いて, 連想した視覚的要素を学習する.

提案手法の概要を図 1 に示す．提案手法では, 質問文 $x^{txt}$ を入力とし, $x^{txt}$ を用いた画像検索で得られる $x^{img}$ を用いて視覚的要素 $X_{vis}$ を生成する．$x^{txt}$ をエンコードした $X_{txt}$ と $X_{vis}$ を融合した表現 $C_t$ を用いて, 応答 $Y$ を出力する.

## 4　実験

質問文から連想された画像の特徴表現を融合的に利用することがどの程度質問応答に効果があるかを調べる．ベースラインには seq2seq[7] と, VQA モデルを用いる．VQA では, iNaturalist で利用可能な生物名と紐づいた画像を, 連想で得られる典型的な正解画像として与えた．表 1 にテストデータでの正答率を示す.

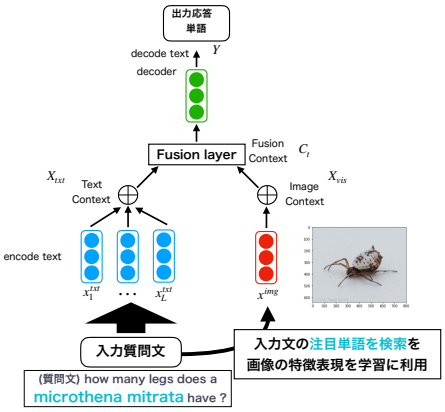


図 1: 提案するモデルの概要図

表 1: 各手法の精度比較

| Model | Accuracy (%) |
|---|---|
| seq2seq | 58.78 |
| VQA | 57.54 |
| 提案手法 | 60.10 |

結果より, 提案手法の精度が最も高く, 画像検索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）では, 画像を用いる VQA モデルよりも精度が高い. ここで提案手法の精度が seq2seq よりも高かった理由としては, 画像検索により, 視覚的要素を有効に使用できたためだと考えられる.

## 5 関連研究

文生成アルゴリズムにおいて, 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 [9] [10]. 本研究では VQA で成功している手法をベースとした上で, 画像検索に基づく視覚的要素を用いてのテキストのみの入力に対応させることを想定している. ゆえ, それらの研究で用いられた手法は用いていない.

## 6 まとめ

本研究では, 画像の代用として, 入力テキストから画像検索を行い, その結果得られる視覚的要素（画像特徴量）を学習に用いることで, 入出力がテキストとなる質問応答の構造を目指した. 実験では, 連想を行う機構として単純な全結合ニューラルネットを用いて視覚情報を生成し, 応答予測に用いた. 結果として, 提案手法はベースラインの精度を上回った. 理由として, 画像検索により, 視覚的要素を有効に使用できたためだと考えられた.

## 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] 石橋 陽一, 森 泰, 宮森 恒. 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. In *NLP2019*, .

[4] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2, 2017.

[5] 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.

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

[7] 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.

[8] Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P. Spithourakis, and Lucy Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. *CoRR*, Vol. abs/1701.08251, , 2017.

[9] 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.

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