Vol.2021-NL-248 No.44
Vol.2021-CVIM-226 No.44
2021/5/21

VQAでの質問文の接頭辞とバイアスの関係についての検討

河野 凌我[1,a)]　数藤 恭子[1,b)]

概要：Visual Question Answering(VQA) は，画像とその画像に関する質問の提示に対して，質問に対する正しい答えを導き出すタスクで，近年ある程度の性能が確認されたモデルや学習データセットも公開されている．しかし，VideoQA(VQA の動画版) のバイアス分析の研究では，特定の接頭辞の質問は，言語 (質問文) におけるバイアスが大きくなる傾向があるため，その他の接頭辞の質問カテゴリよりも高い正解率になることが明らかにされている．そこで，VQA においてテスト時の質問文に対応する画像情報を欠落させた場合をそうでない場合と比較すると，特定の接頭辞の質問に対する正答率の低下は，その他の接頭辞の質問カテゴリに対する正答率の低下より少ないと予測した．これについて実験を行い，質問文の接頭辞とバイアスの関係についての考察を試みた結果，実験結果は必ずしも予測と一致しなかった．

キーワード：VQA，バイアス，接頭辞

Examination of identifying bias in learning VQA model

Ryoga Kono[1,a)]　Kyoko Sudo[1,b)]

Abstract: Visual Question Answering (VQA) is one of the tasks mainly dealt with in deep learning, which is to derive the correct answer to an image and the question about the image when presented, and some performance has been confirmed in recent years. However, in the study of videoQA (video version of VQA) bias analysis, questions with specific prefixes tend to be highly biased in language (question text). Therefore, comparing the case where the image information corresponding to the question text at the time of the test is omitted in VQA with the case where it is not, we predicted that the decrease in the correct answer rate for questions with a specific prefix would be less than the decrease in the correct answer rate for questions with other prefixes. We verified this by experiments and tried to consider the relationship between the prefix of the question sentence and the bias.

Keywords: VQA, bias, prefix

1. はじめに

Visual Question Answering（VQA) は，ある画像とその画像に関する質問の提示に対して，質問に対する回答を出力する．2016 年から国際コンペティションが開催されており [4] ，ある程度の性能が確認されたモデル [5] や学習データセット [6] も公開されている．VQA モデルは，画像の解析・文の解析・文の生成を含む複雑なネットワークであるため，誤答の要因や，正答時に画像や文章をそれぞれ理解していたかどうかを確認することは困難である．また，質問文の単語と画像に対する答えの統計的な傾向（バイアス）の学習が疑われている [1]．一方，VQA の動画版である，動画と質問の提示に対する答えを導く VideoQA において，特定の接頭辞の質問は，言語 (質問文) におけるバイアスが大きくなる傾向があるため，その他の接頭辞の質問カテゴリよりも高い正解率になることが報告されている [2].

そこで本研究では，VQA においても同様に接頭辞に依存した質問文のバイアスがあるのか実験により検証する．具体的には，特定の接頭辞の質問に対する正答率の低下は，その他の接頭辞の質問カテゴリに対する正答率の低下より

[1] 東邦大学 〒 274–8510 千葉県船橋市三山 2–2–1
Toho University Miyama 2–2–1, Funabashi-shi, Chiba 274–8521, Japan
[a)] 6521004k@st.toho-u.jp
[b)] kyoko.sudo@sci.toho-u.ac.jp

少ないと予測し，テスト時の質問文に対応する画像情報を欠落させた場合をそうでない場合と比較する実験を行った．

2. 関連研究

バイアス分析の研究には，VQA では Manjunatha らの方法 [1] が，VideoQA(VQA の動画版) には Yang らの方法 [2] がある．Manjunatha らは，モデルにおける注意マップの関心領域に対応する視覚的な単語，質問，回答が格納されたデータベースで単純なルールマイニングアルゴリズムを実行することにより，VQA モデルは質問と画像の要素を回答に直感的に関連付けているように見えることを明らかにした．また，画像を使用せず言語のみで推論を行った場合でも，43 %の確率で正解することが可能であり，88 %の確率でもっともらしい答えを出力することが可能であるということを示した．

Yang らは，Liu らの学習済み RoBERTa[8] を用いて VideoQA データセットの QA（質問・正解）バイアスの分析を行った．データセットに重要なバイアスが存在し，そのバイアスはアノテーターの分割と質問のタイプから生じる可能性があることを示した．また，質問の接頭辞ごとの精度を導くことで，質問が抽象的な場合には，事実に基づく直接的な質問の場合よりもバイアスがあることから，バイアスは質問の種類に依存する可能性を示した．具体的には，”why”や”how”などの質問は推論的で抽象的な質問であり，回答がより複雑であるため，言語モデルが利用できるバイアスが大きくなる傾向がある一方、“what”や“who”，“where” の質問は事実に基づいた直接的なものであり，バイアスが少なくなる傾向を示した．一方，VQA モデルのバイアスを減らすことを目的とした研究として,Cadene らが提案した RUBi [9] がある．このモデルでは，VQA モデルに 2 つの入力モダリティを使用するように暗黙的に強制することにより，画像を見なくても正しく分類できる例の重要性を低下させた．

3. 提案手法

3.1 提案手法の概要

Yang らの VideoQA の研究 [2] で明らかにされたのと同様の接頭辞に依存した質問文へのバイアスは，VQA でも起きている可能性が考えられる．その場合，特定の接頭辞の質問に対する正答率の低下は，その他の接頭辞の質問カテゴリに対する正答率の低下より少ないと予測される．これを検証する 1 つの方法として，本研究では，同じ学習済み VQA モデルに対して，テスト時に，通常通り入力にテストデータとして画像と質問文を使用した場合と，画像を使わず質問文のみを使用した場合で各々に接頭辞ごとの正解率を出す．その 2 つの接頭辞ごとの正解率を比較することで，考えられた予測が事実であるか否かを示す．VQA や VideoQA では，画像・動画と質問文から答えを生成するように学習されているため，テスト時に画像を与えず質問文のみを与えた場合，基本的には正解率が低下すると考えられる．一方，Yang らの VideoQA のバイアス分析の研究 [2] では，特定の接頭辞の質問は他の接頭辞の質問カテゴリよりも高い正解率になっており，言語（質問文）におけるバイアスが大きくなる傾向があることが明らかにされている．その事実から，特定の接頭辞の質問文については，その他の場合と異なり，画像を与えず質問文のみを与えても正解率が低下しないと予測した．

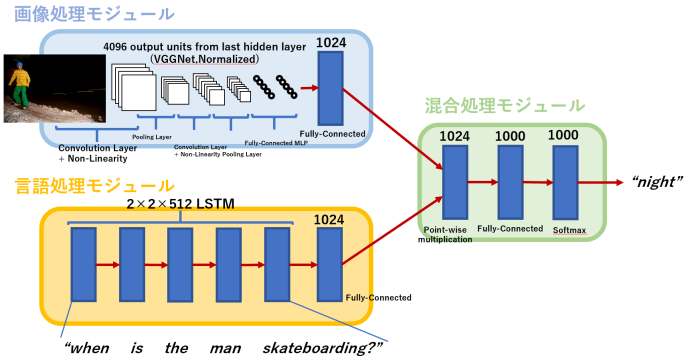

図 1 実験に用いた VQA モデル [4] の構成の概略．

3.2 モデルとデータセット

VQA モデル

本研究では，VGGNet と LSTM で構成される基本的な VQA モデル [4] を用いる (図 1)．画像処理モジュールでは，VGGNet の最後の非表示層 [7] を使用して画像をエンコードし，画像の特徴を L2 正規化する．言語処理モジュールでは，2 層の LSTM を使用して質問をエンコードする．その後，画像処理モジュールと言語処理モジュールの両方で，画像と質問の両方の特徴ベクトルはその要素数を合わせるために全結合層を通過する．最後に，混合処理モジュールにて，質問と画像の両方の特徴は要素ごとの乗算によって融合され，全結合層とソフトマックス層を通過し，回答全体の分布が取得される．

回答パターンの仕様

VQA モデルでは，回答の種類を 1000 パターンに限っている．この 1000 パターンとは，”yes”や”hat”などの単語や”playing game”などの句である．そのため実験で使用した VQA1.0 データセットの中の，画像・質問文・回答ワンセットのデータの回答がその 1000 パターンの中になければ，その回答データを前処理段階で”yes”に置き換える仕様になっている．

データセット

実験では，画像と質問文と回答がセットになっている VQA1.0 データセット [6] を使用した．このデータセットは，Training 用画像が 82,783 枚，Validation 用画像が 40,504 枚，Testing 用画像が 81,434 枚，Training 用質問と

回答が 248,349 個，Validation 用質問と回答が 121,512 個，Testing 用質問と回答が 244,302 個である．

4. 検証実験

4.1 実験方法

VQA モデルへの入力データを次のように定義する．

A. Image-Text データ：画像と質問文のセットが揃っている入力データ．

B. Text-Only データ：画像無しで質問文のみの入力データ．

また，質問の接頭辞の種類は，以下の 11 種類に分類した．

1．”yes/no”:接頭辞が動詞か助動詞である，yes か no で答えられる質問．

2．”what”:接頭辞が what である質問．

3．”why”:接頭辞が why である質問．

4．”where”:接頭辞が where である質問．

5．”which”:接頭辞が which である質問．

6．”who”:接頭辞が who である質問．

7．”when”:接頭辞が when である質問．

8．”if”:接頭辞が if である質問．

9．”how”:接頭辞が”how”+動詞か助動詞である質問（”どのように…？”といった質問）．

10．”how+adj”:接頭辞が”how”+形容詞である質問（how many などの数値で答えられる質問）．

11．”OTHER”:上記の 10 つの分類に当てはまらなかった質問．

実験には，[3] の VQA モデルを使用する．

実験 1

実験では，まず，A.Image-Text データの train 用データを使用し学習を行う．次に，学習済み VQA モデルに対して A.Image-Text データの test 用データでテストを行い，接頭辞ごとに正解率を出す．評価には同じデータセットの回答のデータを使用した．次に，学習済み VQA モデルに対して B.Text-Only データの test 用データでテストを行い，接頭辞ごとに正解率を出す．評価には同じデータセットの回答のデータを使用した．最後に，その 2 つの接頭辞ごとの正解率を比較を行った．

なお，実験の都合上，A.Image-Text データを使用してのテストの際，この VQA モデルの画像処理部分（図 1 の上部）の入力には，その画像処理部分終わりの全結合層手前までの，VGGnet の特徴抽出部分をあらかじめ計算したものを使用した．B.Text-Only データを使用してのテストの際は，その画像処理部分終わりの全結合層への入力をすべて 1 として行った．

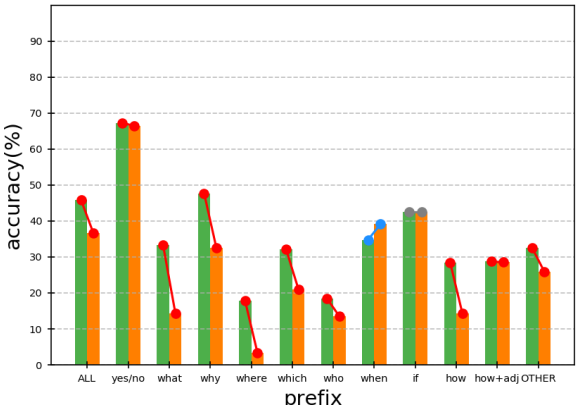

図 2 テスト時に画像有り (緑色)vs 画像無し (橙色)，接頭辞ごとの正解率の比較．

表 1 テスト時に画像有り vs 画像無し，接頭辞ごとの正解率の比較．

接頭辞ごとの正解率	画像有り	画像無し
”ALL”	**45.83**	36.64
”yes/no”	**67.29**	66.39
”what”	**33.33**	14.32
”why”	**47.69**	32.50
”where”	**17.79**	3.43
”which”	**32.12**	21.05
”who”	**18.40**	13.48
”when”	34.78	**39.13**
”if”	42.55	42.55
”how”	**28.45**	14.23
”how+adj”	**28.73**	28.54
”OTHER”	**32.52**	25.82

実験 2

次に，以下の理由からデータセットの一部を除外して同様の実験を行った．セクション 3.1 で述べた通り，[3] の Git の VQA モデルは回答の種類を 1000 パターンに限っている．この 1000 パターンとは，”yes”や”hat”などの単語や”playing game”などの句である．そのため実験で使用した VQA1.0 データセットの中の，画像・質問文・回答ワンセットのデータの回答がその 1000 パターンの中になければ，その回答データを前処理段階で”yes”に置き換える仕様になっている．

この仕様がここまでの実験に与えた影響を調査するため，回答データが”yes”に置き換えられる画像・質問文・回答ワンセットのデータを除外して，その仕様による影響を無くし，接頭辞ごとの正解率を出した．

4.2 実験結果

実験結果 1

実験 1 の結果を図 2 と表 1 に示す．表の”ALL”のカテゴリは，接頭辞ごとに 11 種類に分類した質問カテゴリをすべて含めた正解率である．

この実験結果では，画像有りの正解率の場合，”yes/no”のカテゴリが最も高かった．それ以外では，”why”や”if”の

Vol.2021-NL-248 No.44
Vol.2021-CVIM-226 No.44
2021/5/21

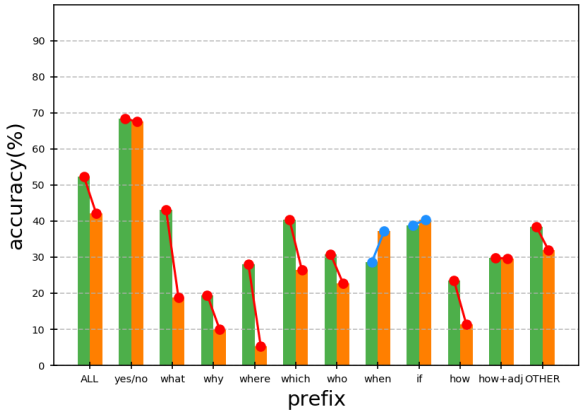

図 3 テスト時に画像有り (緑色)vs 画像無し (橙色)，接頭辞ごとの正解率の比較（データを一部除外）.

表 2 テスト時に画像有り vs 画像無し，接頭辞ごとの正解率の比較（データを一部除外）.

接頭辞ごとの正解率	画像有り	画像無し
”ALL”	**52.34**	42.14
”yes/no”	**68.37**	67.52
”what”	**43.20**	18.83
”why”	**19.49**	10.04
”where”	**28.05**	5.39
”which”	**40.29**	26.52
”who”	**30.82**	22.80
”when”	28.57	**37.14**
”if”	38.89	**40.28**
”how”	**23.48**	11.34
”how+adj”	**29.79**	29.62
”OTHER”	**38.46**	31.87

カテゴリが高くなった．画像有りと画像無しの正解率を比較すると，画像有りの正解率の中で最も高かった”yes/no”のカテゴリや”if”,”how+adj”のカテゴリはほとんど低下しなかった一方，”when”のカテゴリは僅かに上昇し，そのほかのカテゴリは 10 ～ 20% 前後の低下となった．

実験結果２

実験２の結果を図 3 と表 2 に示す．

この実験の結果では，画像有りの正解率の場合，実験結果１と同様に，”yes/no”のカテゴリが最も高かった．また，”what”,”which”,”if”のカテゴリが高くなった．一方，実験結果１とは異なり，”why”のカテゴリは他のカテゴリと比べても特に低い正解率となった．画像有りと画像無しの正解率を比較すると，実験結果１とほとんど同様に，画像有りの正解率の中で最も高かった”yes/no”のカテゴリや”how+adj”のカテゴリはあまり低下しなかった一方，”when”のカテゴリは僅かに上昇し，そのほかのカテゴリは 10 ～ 20% 前後の低下となった．

5. まとめ

5.1 考察と今後の課題

本研究では，以下の２つの実験を行った．

実験１）テスト時に A.Image-Text データを使用した場合と，B.Text-Only データを使用した場合での接頭辞ごとの正解率の差異を比較する．

実験２）セクション 3.1 で述べた仕様が実験１に与える影響を調査するため，回答データが”yes”に置き換えられる画像・質問文・回答ワンセットのデータを除外して，その仕様による影響を無くし，実験１を行う．

Yang らの VideoQA のバイアス分析の研究 [2] では，”why”や”how”などの接頭辞を接頭辞とする，推論的で抽象的な質問は，回答が複雑になることが予測され，そのため言語（質問文）におけるバイアスが大きくなる傾向があることが明らかにされている．その事実から，テスト時に A.Image-Text データを使用した場合と比較して，B.Text-Only データを使用した場合は全体的に正解率は低下するが，”why”や”how”などの接頭辞の質問については正解率が低下しない，と予測した．しかし，実験結果からは，”why”や”how”などの接頭辞の質問についても，ほかの質問と同様に正解率が低下することがわかった．これは，VideoQA が１つの質問に対して回答の選択肢が 5 つ用意されているのに対して，本研究で使用した VQA モデルはすべての質問が 1000 パターンの語句から選択する必要があり，正解のしやすさが異なっているためだと考えられる．つまり，推論的・抽象的な質問であっても回答に選択肢がある場合は回答が容易になるため，選択肢がない場合ほど言語へのバイアスはかからないと考えられる．

また，同じく Yang らの VideoQA のバイアス分析の研究 [2] では，”why”や”how”などの接頭辞の質問は他の接頭辞の質問カテゴリよりも高い正解率になっていることが確認されている．実験１では，”why”のカテゴリは他のカテゴリよりも正解率は高くなっていたが，”how”のカテゴリは他のカテゴリとで正解率に大きな差はなかった．しかし，実験２では，”why”のカテゴリは他のカテゴリよりも正解率は低くなるという結果になった．実験２を行った理由は，画像・質問文・回答ワンセットのデータの回答がモデルが出力できる限られた 1000 パターンの中になければ，その回答データを前処理段階で”yes”に置き換える仕様が実験に与える影響を無くすためである．これらのことから，接頭辞が”why”の質問に対する正答はその限られた 1000 パターンに含まれないものが多くあるということ，そして，実験に使用した VQA モデルは，そのような正答が”yes”に置き換わったものに対して多く正解していることが明らかである．

一方，”when”や”if”のカテゴリの正解率は，実験１と実験２の両方とも，テスト時に A.Image-Text データを使用した場合から B.Text-Only データを使用した場合で，同等か僅かな上昇になっていた．直感的には”when”のような質問に正しく回答するには視覚情報を必要とすると考えられるが，実験ではその直感とは逆の結果になった．今後はこのような結果になった原因を調査することが求められる．

5.2 結論

本研究では，”why”や”how”などの接頭辞の質問文については，他の接頭辞の質問カテゴリと異なり，画像を与えず質問文のみを与えても正解率が低下しないと予測した．実験でテスト時の質問文に対応する画像の有無による接頭辞ごとの正解率の差を比較することで，その予測を検証した．実験の結果，”why”や”how”などの接頭辞の質問についても，ほかの質問と同様に正解率が低下することが明らかとなった．これは，VideoQA と VQA で回答の形式に違いがあり，推論的・抽象的な質問であっても回答に選択肢がある場合は回答が容易になるため，選択肢がない場合ほど言語へのバイアスはかからないためだと考えられる．そのため，VideoQA と VQA の回答形式を一致させるなどの方法をとることで，違った結果が出る可能性がある．一方，本研究ではあまり着目していなかった”when”や”if”のカテゴリの正解率は，同等か僅かな上昇になっていた．直感的には”when”のような質問に正しく回答するには視覚情報を必要とすると考えられるが，実験ではその直感とは逆の結果になった．今後はその接頭辞の質問や対応する画像の内容を精査するなどのことを行い，このような結果になった原因を調査することが求められる．

参考文献

[1] Varun Manjunatha1,Nirat Saini, Larry S. Davis,Adobe Research ,University of Maryland, College Park:Explicit Bias Discovery in Visual Question Answering Models
[2] Jianing Yang,Yuying Zhu,Yongxin Wang,Ruitao Yi,Amir Zadeh, Louis-Philippe Morency:What Gives the Answer Away?Question Answering Bias Analysis on Video QA Datasets
[3] https://github.com/anantzoid/VQA-Keras-Visual-Question-Answering
[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Batra, D. Parikh, ”VQA: Visual Question Answering,” ICCV2015.
[5] https://github.com/facebookresearch/grid-feats-vqa
[6] https://visualqa.org/download.html
[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014.
[8] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis,Luke Zettlemoyer, and Veselin Stoyanov. 2019.Roberta: A robustly optimized bert pretraining approach. ArXiv, abs/1907.11692.
[9] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. 2019. Rubi: Reducing unimodal biases for visual question answering. In Advances in Neural Information Processing Systems, pages 839–850.