6章 視覚・言語のデータセット

吉川友也†, 重藤優太郎†, 竹内彰一†

キーワード：画像・動画の説明文生成，キャプション生成，ビジュアル質問応答，データセット

1．まえがき

これまでの章では，画像・動画と言語を融合した研究課題とそれらを解決するための手法が主に紹介されてきた．これらの手法の出力結果は，その手法を学習するために用いたデータセットによって異なるため，用途に合わせてデータセットを変える必要がある．また，用途に合うデータセットが存在しなければ，データセットの構築も検討するべきだろう．したがって，さまざまなデータセットを知っておくことは，所望の出力を得るためには重要である．

本稿では，画像・動画と言語を融合した研究課題の一例として，画像・動画の説明文生成と画像に対する質問応答を取り上げ，これらの研究において最近構築されたデータセットを比較するとともに，どのように構築され，どのようなアノテーションが付与されているのかを概説する．ページ数制約のため，各データセットの詳細については元論文を参照されたい．

2．画像の説明文データセット

画像から説明文（またはキャプション）を生成するモデルは，教師あり学習により訓練される．さまざまなシーンや物体を認識し，それを充分な語彙で説明するためには，画像と説明文のペアが大量に必要となる．

最近の主要なデータセットの特徴を**表1**にまとめた．画像の説明文生成のベンチマークとして標準的に用いられているのはMS-COCOやFlickr8k/30kである．これらは画像に対して単文の説明文が付けられている．画像中のさまざまな箇所に説明文を付与したデータセットとしてはVisual Genomeが有名である．また，Visual Genomeの画像に対して1パラグラフの説明文を付与したデータセットもある[16]．

最近の画像の説明文データセットの多くは，①説明文を付与する画像を収集，②クラウドソーシングで説明文を付与，③付与された説明文の選別や修正，の各ステップによって構築されている．

表1　画像の説明文データセットの一覧

データセット	説明対象	単位	画像数	平均説明数
Pascal Sentence[22]	画像全体	文	1,000	5
MS-COCO[3]	画像全体	文	123,287	5
Flickr8k[8]	画像全体	文	8,092	5
Flickr30k[32]	画像全体	文	31,783	5
Visual Genome[18]	部分領域	文	108,077	50
Krause et al.[16]	画像全体	段落	19,551	1

それぞれのステップにおいて，データセット毎に異なる特徴が見られる．画像に関しては，Flickrから収集した画像がよく利用される．ただし，画像の著作権に関してはさまざまな物が混在しているため，使用に関しては注意が必要となる．Flickr以外の画像を用いたケースとして，Abstract Scenes Dataset[36]では，クリップアートを合成して作成した画像を用いている．

クラウドソーシングによる説明文付与に関しては，英語のデータセットが多いということもあり，Amazon Mechanical Turk（AMT）がよく利用されている．日本語説明文を付与したケースでは，YJ Captions[20]はYahoo!クラウドソーシング，STAIR Captions[31]はクラウドワークスとCROWDが用いられている．説明文付与の際，ワーカーにはガイドラインが示される．例えば，MSCOCOでは，説明文は，①8単語以上，②重要なシーンの説明だけ，③過去や未来について書かない，④人の固有名詞は書かない，等が指示されている．また，日本語特有のガイドラインとしては，STAIR Captionsは「だ・である」調で説明文を書くように指示している．

上記ステップで付与された説明文には，スペル誤りや文法的誤り，ガイドライン違反等が見られる．このような説明文を除去または修正するために，新たにクラウドソーシングを行い，スペル誤りや文法的誤りがある場合は修正をしたり，ガイドラインに沿っているかどうかの判定を行い，沿ってい

†千葉工業大学 人工知能・ソフトウェア技術研究センター
"Datasets for Vision and Language Research" by Yuya Yoshikawa, Yutaro Shigeto and Akikazu Takeuchi (STAIR Lab, Chiba Institute of Technology, Chiba)

映像情報メディア学会誌 Vol. 72, No. 5, pp. 672～675 (2018)

ないものを除去したりする作業が行われている．この作業は最終的なデータセットの品質に関わるものであるが，データセット毎に対応が異なるため注意が必要である．

3．動画の説明文データセット

動画の説明文生成は，基本的には画像の説明文生成の入力が画像から動画に変わったものとなる．したがって，データセットの作成方法も画像の説明文と同様のステップで行われる．

表2には最近の主要な動画の説明文データセットとそれらの統計量をまとめた．動画の説明文生成において標準的と言えるデータセットはまだないが，MSR-VTT, ActivityNet Captions, LSMDCは国際会議で行われたコンペティションで使用された実績がある．

説明文を付与する対象の動画は，データセット毎にさまざまなリソースから収集されている．YouCook，MSR-VTT，ActivityNet Captionsは，YouTubeから収集された動画が利用されている．特に，YouCookは料理動画だけを集め，MSR-VTTはシーンが偏らないようにさまざまなクエリで動画を検索して収集している．LSMDCは，DVDやBlu-rayの映画から動画を収集している．このデータセットでは，映画の脚本や音声解説の書き起こし文も利用可能である．上記のように既存の動画を集めるのではなく，新たに撮影した動画を利用するケースもある．TACoS multi-levelの動画は，ワーカーに料理の手順を示してその指示通りに料理をしてもらった動画である．また，Charadesはキーワードからの台本作成をAMTで行い，その台本に基づいてAMTで動画撮影を依頼し著作権フリーの動画を収集している．

動画に対する説明文付与は，画像の説明文と同様にクラウドソーシングで行われる．MSR-VTTは動画を数十秒の短い動画にして，それらに対して単文の説明文付与を行っている．このような処理を行わず，動画全体に説明文を付与するケースでは，TACoS Multi-levelは，動画1本に対して詳細に記述した複数の説明文，短く記述した複数の説明文，動画全体を表す1つの説明文の3種類が付与されている．ActivityNet Captionsでは，動画1本を説明する1パラグラフ分の説明文をワーカーに付与してもらい，その後，そのパラグラフの各文が動画中のどの時間区間に対応するかを付与してもらうことで，時間区間の重複を含む説明文が付けられている．

表2　動画の説明文データセットの一覧
TACoS M-LはTACoS multi-levelの略．AN CaptionsはActivityNet Captionsの略．

データセット	動画数	クリップ数	合計説明文	平均クリップ長
MSVD[2]	−	1,970	70,028	10秒
YouCook[5]	88	−	2,668	−
TACoSML[25]	127	14,105	52,593	360秒
Charades[26]	10,000	10,000	16,129	30秒
MSR-VTT[30]	7,180	10,000	200,000	20秒
LSMDC[24]	202	118,081	118,081	4.8秒
ANCaptions[17]	19,994	100,000	100,000	180秒

4．画像を対象とした質問応答データセット

キャプション生成と同様に，画像・動画を対象とした質問応答に関する研究も盛り上がりを見せている．それに伴い，新たなデータセットが次々に構築されている．本節では，主に画像を対象とした質問応答データセットの紹介を行う．

表3に，代表的な画像を対象とした質問応答データセットを示す．各データセットは，画像に対して質問とその応答が付与されている．質問応答を付与する画像には，実画像を用いる場合と人工画像を用いる場合の2通りの方法がある．実画像を用いたデータセットに注目した場合，DAQUARを除いたすべてのデータセットにおいて実画像としてMS-COCOを用いている．

FM-IQA[6]とVisual7W[35]には，他のデータセットにはな

表3　質問応答データセットの比較
アノテーション列における，Aは質問応答を自動で作成，Mは人手で作成したデータセットであることを示している．タスク列における，MCは応答を候補から選択する（Multiple Choice）タスクであり，OEは応答を生成する（Open-Ended）タスクである．

データセット	画像数	質問応答数	アノテーション	画像	タスク
DAQUAR[19]	1,449	12,468	A·M	NYU Depth Dataset V2	OE
FM-IQA[6]	316,193	158,392	M	MS-COCO	OE
COCO-QA[23]	123,287	117,684	A	MS-COCO	OE
Visual7W[35]	47,300	327,939	M	MS-COCO	MC
Visual Genome[18]	108,077	1,773,258	M	MS-COCO,YFCC100M	MC
VQA（abstract）[1]	50,000	150,000	M	人工画像（クリップアート）	OE
VQA（balanced-abstract）[7]	31,325	33,383	M	人工画像（クリップアート）	OE
VQA（balanced-real）[34]	204,721	1,105,904	M	MS-COCO	OE
CLEVR[10]	100,000	864,968	A	人工画像（3Dレンダリング）	OE
TDIUC[11]	167,437	1,654,167	A·M	MS-COCO	OE

い特色がある．FM-IQAは，質問応答が英語だけではなく中国語でも付与されている*1．Visual7Wは，「Which pillow is farther from the window?」のような物体を特定する質問に対しては，言語による応答に加えて，画像中の対応する物体にバウンディングボックスも付与してある．

VQA[1]は，頻繁にベンチマークとして利用されている．VQAは，real imagesデータセットとabstract scenesデータセットの2種類のデータセットから構成されている．Real imagesデータセットは，MS-COCOに含まれる写真を基にしたデータセットであり，abstract scenesデータセットは，クリップアートを使って人工的に生成した画像で構成されている．これらのデータセットは，AMTを使い人手で質問と応答が付与されている*2．作業者には「人間には簡単に答えられるが，ロボットには答えられない質問」を付与するように指示されており，「What is ...」，「Is there ...」，「How many ...」，「Does the ...」などのフレーズで始まる質問がデータセットに存在する．

現在公開されている最新のVQAは第2版（balanced VQAデータセット）[34)7)]となっている．第2版は，第1版に存在したデータセットに存在するバイアスを除去するようにデータセットの改善を行っている．このバイアスとは，データセットに存在する応答の偏りのことを意味している．例えば，第1版において「What sport is ...?」から始まる質問に対する正しい応答の41%が「tennis」である．また「How many ...?」に対する正しい応答の39%が「2」となっている．その結果として，質問応答システムが実際に画像を理解せずとも，偏った回答を記憶（例えば「What spor ...?」に対して常に「tennis」と回答）することにより高い性能に達して（いるように見えて）しまう．このようなシステムは本質的に役に立たず，実応用上でも好ましくない．VQA第2版では，原因となるバイアスを除去しており，質問に正しく応答するためには画像の理解が必要である，と主張している．同様にバイアスが抑えられたデータセットとしてCLEVR[10)]とTDIUC[11)]が構築されている．

表には記していないが，特色がある質問応答データセットを簡単に紹介する．図中に含まれるダイアグラムの理解を目的としたAI2D[13)]，教科書に記載された文章と図の理解を目的としたTQA[14)]，棒グラフや折れ線グラフなどのグラフを質問の対象としたFigureQA[12)]，常識や外部知識に注目したKB-VQA[29)]，FBQA[28)]，質問と応答を対話として付与したVisDial[4)]，質問応答を文の穴埋め問題として取り扱っているVisual Madlibs[33)]がある．

画像に対する質問応答に注目して説明を行ったが，動画に対する質問応答も盛んに研究されている．データセットとしてMovieQA[27)]，MarioQA[21)]，TGIF-QA[9)]，PororoQA[15)]が構築されている．動画に対する質問応答では，画像では扱えない事象に注目し，質問が付与されている．例えば，TGIF-QAデータセットには「How many times does the man wrap string?」，「How many times does the cat touch the dog?」や「What does the bear on right do after sitting?」などの画像からでは応答できないような質問が含まれている．

5．むすび

本稿では，画像・動画の説明文生成と画像を対象とした質問応答における最近のデータセットについて紹介した．最近のデータセットは，クラウドソーシングを駆使してアノテーションすることにより，非常に大規模になっている．このような大規模なデータセットが存在するからこそ，ニューラルネットワークのような大量のパラメータを持つ複雑なモデルを学習可能にしていると言えるだろう．参考文献に目を通していただくとわかるが，毎年新たなデータセットが公開されている．深層学習の発展，クラウドソーシングプラットフォームの整備などもあり，今後も新たなデータセットが構築され続けると思われる．実際に今年（2018年）のCVPRにおいても，新しいデータセットが発表されるようである*3．そのため，ベンチマークの定番であるMS-COCOやVQAデータセットが今後も定番であり続けるとは限らず，新たに構築されるデータセットの情報を継続して調査する必要がある．

謝辞　この成果は，国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の委託業務の結果得られたものです．

（2018年5月31日受付）

〔文　献〕

1）A. Agrawal, et al.: "VQA: Visual Question Answering", In ICCV（2015）
2）D.L. Chen, et al.: "Collecting Highly Parallel Data for Para phrase Evaluation", In ACL（2011）
3）X. Chen, et al.: "Microsoft COCO Captions: Data Collection and Evaluation Server", In ECCV（2015）
4）A. Das, et al.: "Visual Dialog", In CVPR（2017）
5）P. Das, et al.: "A Thousand Frames in Just a Few Words: Lingual Description of Videos through Latent Topics and Sparse Object Stitching", In CVPR（2013）
6）H. Gao, et al.: "Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering", In NIPS（2015）
7）Y. Goyal, et al.: "Making the V in VQA matter: Elevating the role of image understanding in visual question answering", In CVPR（2017）
8）M. Hodosh, et al.: "Framing Image Description as a Ranking Task Data, Models and Evaluation Metrics", JAIR（2013）
9）Y. Jang, et al.: "TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering", In CVPR（2017）
10）J. Johnson, et al.: "CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning", In CVPR（2017）

*1　正確には，はじめに中国語で質問応答を付与し，その後，中国語を英語に翻訳している．
*2　Abstract scenesデータセットに対しては，各画像に対して5個のキャプションも付与されている．
*3　執筆段階において，CVPR 2018の会議録が公開されていないため，今回は説明する対象から除外した．

11) K. Kae, et al.: "An Analysis of Visual Question Answering Algorithms", In ICCV（2017）
12) S.E. Kahou, et al.: "FigureQA: An Annotated Figure Dataset for Visual Reasoning", In Workshop on ICLR（2018）
13) A. Kembhavi, et al.: "A Diagram is Worth a Dozen Images", In ECCV（2016）
14) A. Kembhavi, et al.: "Are You Smarter Than a Sixth Grader? Textbook Question Answering for Multimodal Machine Comprehension", In CVPR（2017）
15) K.M. Kim, et al.: "DeepStory: Video Story QA by Deep Embedded Memory Networks", In IJCAI（2017）
16) J. Krause, et al.: "A Hierarchical Approach for Generating Descriptive Image Paragraphs", In CVPR（2017）
17) R. Krishna, et al.: "Dense-Captioning Events in Videos", In ICCV（2017）
18) R. Krishna, et al.: "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations", IJCV（2017）
19) M. Malinowski, et al.: "A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input", In NIPS（2014）
20) T. Miyazaki, et al.: "Cross-Lingual Image Caption Generation", In ACL（2016）
21) J. Mun, et al.: "MarioQA: Answering Questions by Watching Gameplay Videos", In ICCV（2017）
22) C. Rashtchian, et al.: "Collecting Image Annotations Using Amazon's Mechanical Turk", In NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk（2010）
23) M. Ren, et al.: "Exploring Models and Data for Image Question Answering", In NIPS（2015）
24) A. Rohrbach, et al.: "Movie Description", IJCV（2017）
25) M. Rohrbach, et al.: "Recognizing Fine-Grained and Composite Activities Using Hand-Centric Features and Script Data", IJCV（2016）
26) G.A. Sigurdsson, et al.: "Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding", In ECCV（2016）
27) M. Tapaswi, et al.: "MovieQA: Understanding Stories in Movies through Question-Answering", In CVPR（2016）
28) P. Wang, et al.: "FVQA: Fact-based Visual Question Answering", TPAMI（2017）
29) P. Wang, et al.: "Explicit Knowledge-Based Reasoning for Visual Question Answering", In IJCAI（2017）
30) J. Xu, et al.: "MSR-VTT: A Large Video Description Dataset for Bridging Video and Language", In CVPR（2016）
31) Y. Yoshikawa, et al.: "STAIR Captions: Constructing a Large-Scale Japanese Image Caption Dataset", In ACL（2017）
32) P. Young, et al.: "From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions", TACL（2014）
33) L. Yu, et al.: "Visual Madlibs: Fill in the blank Image Generation and Question Answering", In ICCV（2015）
34) P. Zhang, et al.: "Yin and Yang: Balancing and Answering Binary Visual Questions", In CVPR（2016）
35) Y. Zhu, et al.: "Visual7W: Grounded Question Answering in Images", In CVPR '16（2016）
36) C.L. Zitnick, et al.: "Bringing Semantics Into Focus Using Visual Abstraction", In CVPR（2013）

吉川　友也（よしかわ　ゆうや）　2011年，静岡県立大学経営情報学部卒業．2013年，奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．2015年，同研究科博士後期課程修了．2015年より，千葉工業大学人工知能・ソフトウェア技術研究センター主任研究員．博士（工学）．

重藤優太郎（しげとうゆうたろう）　2017年，奈良先端科学技術大学院大学博士課程修了．同年より，千葉工業大学人工知能・ソフトウェア技術研究センター主任研究員．博士（工学）．

竹内　彰一（たけうち　あきかず）　1979年，東京大学工学系研究科計数工学専門課程修了．同年，三菱電機（株）中央研究所に入社．1991年，（株）ソニーCSLに入社．2015年より，千葉工業大学人工知能・ソフトウェア技術研究センター主席研究員．工学博士．