トレースからユーザの意図を反映した画像キャプション生成

渡邊清子　　小林一郎
お茶の水女子大学
{watanabe.sayako,koba}@is.ocha.ac.jp

概要

近年，画像キャプション生成の研究は画像から得られる情報だけでなく，コントロールシグナルと呼ばれる追加情報を与えることにより，制御可能な画像キャプション生成へと発展している．本研究では，画像をなぞることをコントロールシグナルとみなし，なぞった軌跡（‘トレース’と呼ぶ）から推定されるユーザの意図に基づくキャプション生成手法を提案する．トレースの滞在時間からユーザの興味度合いを，トレースの順番から説明順序をユーザの意図として推定する．滞在時間を文長に比例させ，ユーザがトレースした情報を順番に漏らさずに生成文中に表現することを可能にする非自己回帰型のキャプション生成手法を提案する．

1　はじめに

コンピュータを操作するユーザインターフェイスは，キーボードを使った文字による CUI (Character-based User Interface) と，マウスを使った画像による GUI (Graphical User Interface) が主流であった．しかし，PC (Personal Computer) にタッチパッドが搭載されたことや，スマートフォンの出現により，「クリック」だけではない「スワイプ」「ピンチアウト」「フリック」など指先による画面操作方法が多様になってきた．また，近年メタバースが注目される中で，VR (Virtual Reality) や MR (Mixed Reality) などを使用する際には，ハンドコントローラーやハンドトラッキング技術により手の動きの軌跡を読み取り，画面を操作している．つまり，従来画面に合わせて決まった動かし方しかできなかったユーザインターフェイスは，ユーザの意図に合わせて多様な表現方法ができるように変容し続けている．更に今後は，「スワイプ」「ピンチアウト」「フリック」などルールに当てはめずともユーザの動きの軌跡から指示を推測するようなユーザインターフェイスに進化していくと考えている．そこで本研究では，ユーザが描く軌跡に焦点を当て，画像キャプション生成を題材に実験を行う．

2　関連研究

近年，画像キャプショニングの研究は，Faster R-CNN [1] や Semantic Segmentation [2] といった手法を用いて画像の内容を捉え，画像特徴量や言語モデルから画像の内容を深く捉える手法に基づくキャプション生成手法が提案されている [3, 4, 5, 6]．また，近年は画像特徴量や言語モデルだけでなく，コントロールシグナルと呼ばれる追加情報を与えて生成キャプションを制御する Controllable Image Captioning が盛んに研究されている．制御信号には様々な種類があり，バウンディングボックス [7]，シーングラフ [8]，文の長さ [9]，文の詳述さ [10] を用いた研究が挙げられる．しかし，コントロールシグナルはキャプションの内容やスタイルに言及したものが多く，ユーザの意図や興味に沿ったインタラクティブな題材についてはあまり研究されていない．このような背景から，本研究では，ユーザが画像をなぞる軌跡をコントロールシグナルとし，ユーザの説明意図を反映した新しい画像キャプション生成手法を提案する．説明をする際に指差しながら発話するということは，人間の自然な行動である．その為，トレースは画像キャプションのタスクと親和性が高い．また，トレースには画像内の位置を示す座標情報だけではなく，時間軸やトレースの形状などの情報も含まれるため，様々な応用が可能である．画像キャプション生成のコントロールシグナルにトレースを用いる先行研究として，Yan ら [11] は，トレースとバウンディングボックスの関係をスコアリングすることによって結び付けに成功し，高い精度を出した．これに対して本研究では，トレースの座標情報だけでなく，トレースの滞在時間に注目し，ユーザの興味に合わせて説明の詳述さをキャプションに反映する．

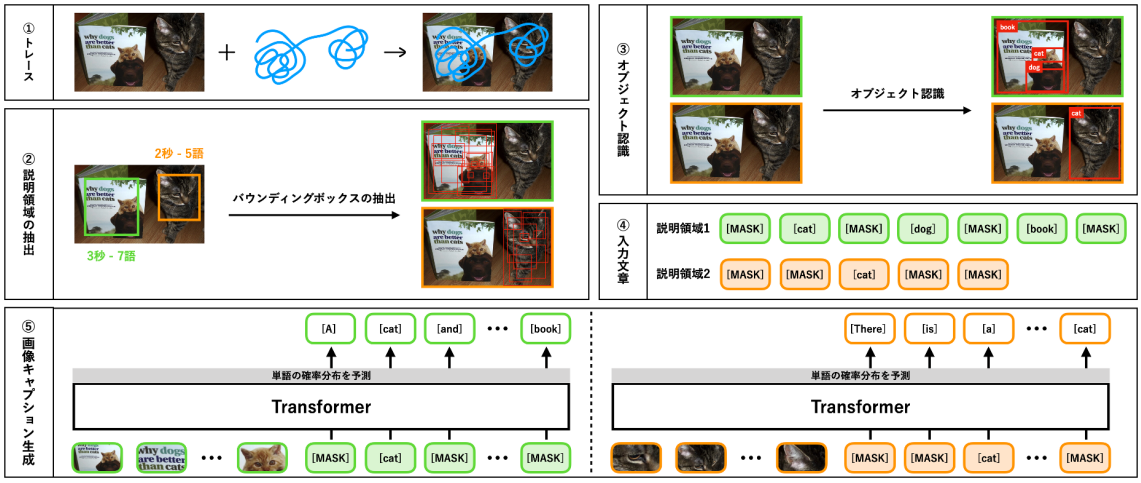


図 1　提案手法の概要

3　提案手法

図 1 に提案手法の概要を示す．以下，図 1 の各プロセスを説明する．

①画像のトレース　画像中の説明したい箇所をトレースする．その際，説明を詳細にしたい対象に対してはトレースを念入りに行う．

②説明領域の抽出　トレースの描画範囲から説明領域を抽出し，各領域のトレースの滞在時間から文長を推定する．また，各領域のバウンディングボックスを抽出する．

③オブジェクト認識　各領域内でオブジェクト認識を行う．

④入力文章　②で推定した文長と③で認識されたオブジェクトの単語から，画像キャプションを生成モデルに入力する文章を準備する．

⑤画像キャプション生成　②によって抽出した各バウンディングボックスの特徴量と④で準備した文章を入力とし，Deng ら [9] による文長制御可能な画像キャプション生成モデル LaBERT を用いて，それぞれの領域の画像キャプションを生成する．

3.1　使用データセット

Pont-Tuset ら [12] は，トレースデータセット Localized Narratives(LN) を構築した．LN は，マウスで画像をなぞりながら画像の内容を音声で説明するという実験によって収集されたデータセットである．LN は Open Images [13], Microsoft COCO [14], Flickr30k [15], ADE20k [16] の 4 つの画像データセットから構成されており，画像，トレース，画像キャプション，キャプションの音声が含まれている．

3.2　トレースによる説明領域の抽出

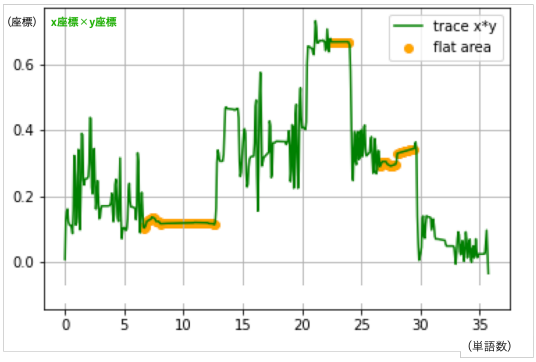


図 2　トレースの座標の変化量

ユーザが興味を示した範囲をその順番に沿ってキャプションを生成する為に，トレースから説明領域を抽出する．トレース時における人の行動特性を解明する為に，座標の変化量を特徴量として抽出した．横軸に単語数，縦軸にトレースの $(x \times y)$ 座標をとったグラフを図 2 に示す．グラフが平らになっている範囲が特徴的に表れていたので，発話データと参照してみたところ，ピリオドの部分に相当していた．同じような特徴が他のデータにも多く見受けられた．グラフが平らな範囲は，1 文の発話が終わり次の文章を発話するまでの間の時間かつ，次の説明物体に移動するまでの間の時間である．このことから，1 文ずつに分割することを説明領域の抽出方法の指針として，この部分を抽出する．$(x \times y)$ 座標の変化量が少なく，それが連続している部分を黄色い点でプロットした．この例の場合，説明領域は黄色い点を除き分割点とし，4 つに分かれる．画像説明時のトレースにおける人の行動特性は常時このようになるわけではなく，他の行動特性も観察された

が，本研究では上記の行動特性を基準とする．

3.3 文長とトレース滞在時間の関係

人は画像を説明する時，詳しく説明したい時にはトレースを長い時間滞在させ，簡単に説明したい時には短いトレースになると仮定する．キャプションの単語数とトレースの滞在時間の関係を明らかにするために，LN のデータを用いて定量的に検証した．横軸を文のキャプションの単語数，縦軸をトレースの滞在時間としたグラフを図 3 に示す．この結果より，キャプションの単語数が多くなるにつれて，トレースの滞在時間は対数的に伸びていることがわかる．今回は 7〜25 単語で生成するので，7〜25 単語に絞ってみてみると，比例的に伸びていることがわかる．

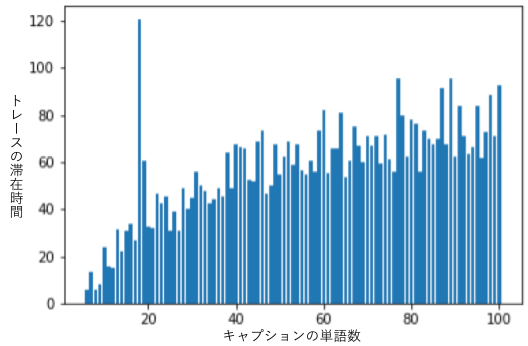


図 3 キャプションの単語数とトレースの滞在時間

トレースの滞在時間により説明の詳述さを決定する．外れ値を排除し，キャプションの単語数÷トレースの滞在秒数の中央値を計算すると，1 秒間あたり 2.35 単語発話していることがわかった．この結果から，トレースの滞在秒数を 2.35 倍したものを文の長さとして設定する．

3.4 非自己回帰型キャプション生成

ユーザの興味の度合いを反映した画像キャプションを生成する為に，トレースの滞在時間により説明の詳述さを決定することを考える．ここでは，説明の詳述さを生成文の長さとして捉える．一般に，逐次的に次の単語を予測する自己回帰的な文生成手法は，生成文の長さを制御できない．また，生成する文の長さが長くなると計算量は線形的に増加してしまうといった欠点がある．これに対し，Deng ら [9] は，長さ制御可能な画像キャプションの為の非自己回帰型デコーダを考案し，文長を制御する効率の良い文生成手法である LaBERT を提案している．

LaBERT では生成文に対して，各単語の確率分布を元にスコアが低いトークンを [MASK] に変更し，[MASK] 部分に対して再度単語を予測するという仕組みをとっている．本研究では，トレースを元に生成を行うため，ユーザの指した情報は生成キャプションでも出現させたい．LaBERT をそのまま使用すると，生成されたテキストが更新される過程で，残して欲しい単語に関しても [MASK] に変更されてしまう可能性があるため，常にその単語を含むキャプションを生成するようにデコーダを改良した．本研究では，LaBERT のデコーダを参考にし，トレースの滞在時間を文長に比例させ，ユーザの指した情報を順番に漏らさずに文中に挿入することを可能にする非自己回帰型のキャプション生成手法を提案する．概要を図 4 に示す．画像から物体認識されたラベルを取り出し，トレースがその物体のバウンディングボックス内を一定時間以上滞在した場合，挿入対象単語とみなす．そして，図 4 の Step1 において，[MASK] の代わりに挿入対象単語を挿入する．挿入位置は，発話中ずっと指していた場合は中央に，最後にトレースした場合は後方に配置するように，トレースがその物体を指していたタイミングに合わせて挿入する．

図 4 長さ L_{low}〜L_{high} の文生成

4 実験

画像とトレースを入力として，トレースからユーザーの意図を反映した画像キャプション生成を行った．

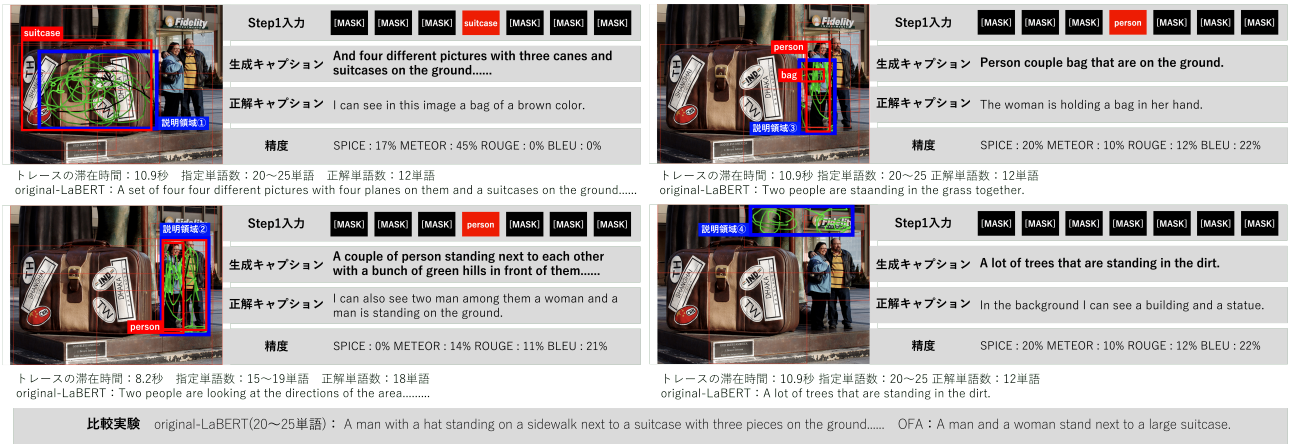


図 5 トレースを入力とした画像キャプション生成.

4.1 実験設定

表 1 に実験設定を示す．文生成の精度の評価指標には，BLEU@1 [17], ROUGE [18], METEOR [19], SPICE [20] を用いた．

表 1 実験設定

画像データセット	Microsoft COCO
画像特徴量	coco_detections.hdf5[1)]
単語数	7~9, 10~14, 15~19, 20~25
文のアップデート回数	10, 15, 20, 25
言語モデル	pre-trained $\text{BERT}_{\text{BASE}}$[2)]
バッチサイズ	256
イテレーション	100,000

*1 https://github.com/aimagelab/meshed-memory-transformer/coco_detections.hdf5
*2 https://huggingface.co/bert-base-uncased

4.2 実験結果

実験結果を 5 に示す．説明領域 3 では，トレースは女性のバックを強くトレースしており，正解キャプションも “bag”について言及している．物体認識の結果，Step1 での入力として “bag”が中央に追加され，生成キャプションに “bag”が含まれた．改良前の LaBERT で生成した場合，注目する物体を絞らずに周りの情報を多く取り込んでしまった結果，“bag”に関しての記述は無かった．LaBERT を改良したことにより，トレースの意図をよく捉えた結果になった．

また，トレースの滞在時間からキャプションの単語数を予測した結果，説明領域②, ③, ④では，単語数の指定範囲に正解単語数が収まっており，正しく予測出来た．説明領域①は，正解キャプション 12 単語に比べ，大幅に多く推定された．実際はトレースの滞在時間程，長い文による説明は求められていないことがわかる．しかし，トレースを見ると入念にトレースがされており詳しい説明を求めているような特性が見られる．生成されたキャプションは，“suitcase”を捉えるだけではなく，スーツケースに貼ってあるステッカーを “pictures”と捉え詳細な説明ができている．説明者の意図には沿えていないが，トレースの意図を汲み取った生成結果となっている．

比較実験として，統合マルチモーダル事前トレーニングモデル OFA [21] と改良前の LaBERT でキャプションを生成した．これらの結果は，画像の概要は説明できている．しかし，提案手法のように説明の順番や説明の詳述さなど，ユーザの意図に合せて文章を生成することはできていない．

付録に他の実験例を載せる．

5 おわりに

本研究では，画像に対してトレースを用いながら説明するデータセット LN と文長制御が可能な非自己回帰型テキスト生成を行う LaBERT のデコーダを改良し組み合わせ，トレースからユーザの説明意図を汲み取りインタラクティブに説明文を生成する画像キャプション生成手法を提案した．既存のキャプション生成手法と提案手法を比較すると，提案手法では説明の順序や詳述さを反映したキャプションを生成することに成功している．しかし，今回は画像を指す際に，渦巻きを描き移動し渦巻きを描きという特性を持つトレースに絞って実験を行ったが，実際のデータは様々な特性が見られた．

今後の課題として，トレースからユーザの特性を分析し，それに応じた処理がでいるようなトレース分析を考えている．

参考文献

[1] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. **CoRR**, Vol. abs/1506.01497, , 2015.

[2] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In **The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, June 2015.

[3] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. mplug: Effective and efficient vision-language learning by cross-modal skip-connections, 2022.

[4] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa, 2019.

[5] Ting-Yao Hsu, C. Lee Giles, and Ting-Hao 'Kenneth' Huang. Scicap: Generating captions for scientific figures, 2021.

[6] Junqi Jin, Kun Fu, Runpeng Cui, Fei Sha, and Changshui Zhang. Aligning where to see and what to tell: image caption with region-based attention and scene factorization, 2015.

[7] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. **CoRR**, Vol. abs/1811.10652, , 2018.

[8] Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs, 2020.

[9] Chaorui Deng, Ning Ding, Mingkui Tan, and Qi Wu. Length-controllable image captioning. **CoRR**, Vol. abs/2007.09580, , 2020.

[10] Zhangzi Zhu and Hong Qu. Improving image captioning with control signal of sentence quality, 2022.

[11] Kun Yan, Lei Ji, Huaishao Luo, Ming Zhou, Nan Duan, and Shuai Ma. Control image captioning spatially and temporally. In **ACL/IJCNLP (1)**, pp. 2014–2025, 2021.

[12] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In **ECCV**, 2020.

[13] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. **CoRR**, Vol. abs/1811.00982, , 2018.

[14] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. **CoRR**, Vol. abs/1405.0312, , 2014.

[15] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In **2015 IEEE International Conference on Computer Vision (ICCV)**, pp. 2641–2649, December 2015.

[16] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. **CoRR**, Vol. abs/1608.05442, , 2016.

[17] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

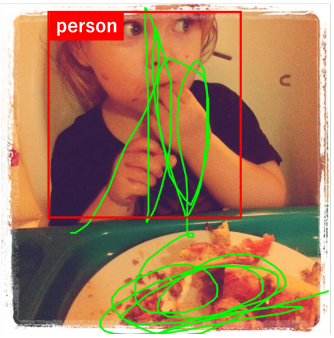
[18] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[19] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In **Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization**, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[20] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: semantic propositional image caption evaluation. **CoRR**, Vol. abs/1607.08822, , 2016.

[21] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework, 2022.

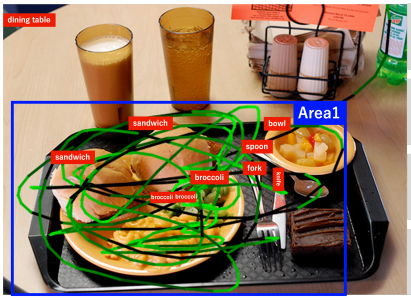
A　付録



Step1入力	[MASK] [MASK] [MASK] person [MASK] [MASK] [MASK]
生成キャプション	Person close up standing with a pizza in their hands.....
正解キャプション	In the image we can see there is a person who is sitting and in front on plate there is a food item.
精度	SPICE : 14% METEOR : 6% ROUGE : 11% BLEU : 8%

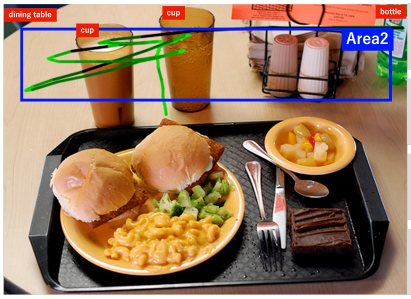
トレースの滞在時間：6.6秒　指定単語数：15〜19単語　正解単語数：23単語

比較実験	original-LaBERT： A little boy that is holding a plate of food..... OFA：A little kid eating pizza at a restaurant.



Step1入力	dining table sandwich broccoli sandwich broccoli spoon knife fork
生成キャプション	A plate of food sitting on a table next to a bowl of soup and a fork..
正解キャプション	In this image there are two burgers, pasta and salad on the plate.
精度	SPICE : 10% METEOR : 8% ROUGE : 7% BLEU : 20%

トレースの滞在時間：7.6秒 指定単語数：15〜19 正解単語数：13単語
original-LaBERT： A plate of food with a sandwich and a bowl of soup next to a glass of orange juice.....



Step1入力	dining table [MASK] [MASK] [MASK] [MASK] [MASK] cup cup
生成キャプション	Dining table with plates of food and cup and
正解キャプション	There are glasses and bottle.
精度	SPICE : 0% METEOR : 4% ROUGE : 15% BLEU : 11%

トレースの滞在時間：3.3秒 指定単語数：7〜9 正解単語数：5単語
original-LaBERT：A plate of food sitting on a table.

比較実験	OFA：A tray of macaroni and cheese, macaroni salad, and a sandwich with a drink.