トレースにより画像の注視点を与えるインタラクティブなテキスト生成

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に対する前処理から得られた情報を元に画像キャプションを生成するだけでなく，画像処理情報に対して，コントロールシグナルと呼ばれる視点に相当する追加情報を与えることで，画像に対してユーザの興味に基づくキャプションを生成する研究へと発展している．本研究では，人は一般的に画像の内容を説明する際，大抵説明したい対象に注意を促す為に指でそのものを差しながら説明することに着目し，画像への指差しをコントロールシグナルと捉える．また，このような行為で生じる指差しの軌跡のことをトレースと呼び，トレースに込められた意味を反映することで，より説明者の意図に沿ったインタラクティブな画像キャプション生成手法を提案する．

1 はじめに

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，その結果から画像内の物体間の関係を捉えるシーングラフ [3, 4] を構築し，そのグラフに基づきキャプションを生成するものなど，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 [5, 6, 7]．一方で，生成されるキャプションは多くの場合，用いられる学習データに依存しており，画像内容を説明する者の意図が反映される結果ではない場合が多い．このことを踏まえて，近年では，キャプション生成を制御する為のコントロールシグナルと呼ばれる追加情報を与えて，説明者の意図に近いキャプションを生成する研究なども取り上げられてきている [8, 9, 10]．しかし，与えられるコントロールシグナルはキャプション内容に言及したものが多く，説明者の感覚や興味に沿うインタラクティブな画像キャプション生成の報告はあまりない．このことから，本研究では，音声で画像を説明する際に説明の描画領域を指したトレースデータをもつ Localized Narratives(LN) [11] を用いて，トレースにより画像の注視点を与えた画像キャプショニング手法を提案する．

2 提案手法

2.1 概要

図 1 に提案手法の概要を示す．

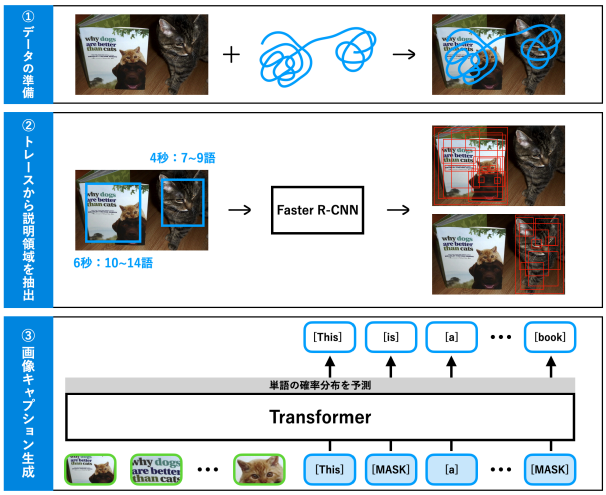

図 1 提案手法の概要

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．Fast R-CNN を用いて，各領域のバウンディングボックス（B.Box）を抽出する．

③画像キャプション生成　②によって抽出した B.Box の特徴量と生成する文の長さを示す語数を入力とし，Deng ら [12] による文長を制御可能な画像

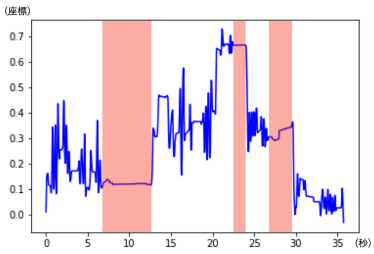

図 2　トレースの座標変化量

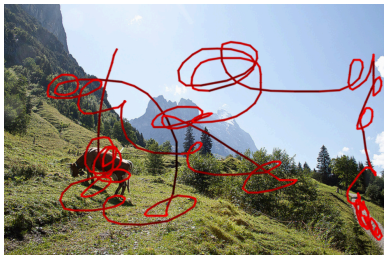

図 3　速さに注目したトレースの可視化

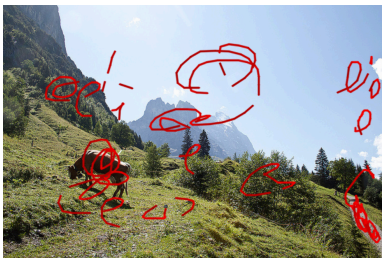

図 4　速いトレースを削除した結果

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．

2.2　Localized Narratives

Pont-Tuset ら [11] は，視覚と言語を繋ぐマルチモーダルな画像キャプションを目標として，音声で画像を説明する際に，説明の描写領域を指したトレースデータセット Localized Narratives（LN）を構築した．LN は，人が画像をペンでトレースしながら音声でその内容を説明するという実験を通じて収集したデータセットである．データ数も多く，Open Images [13]・Microsoft COCO [14]・Flickr30k [15]・ADE20k [16] の 4 つのデータセットから成っている．LN には，画像・トレース・画像キャプション・音声の 4 つが含まれている．表 1 にデータの内容を示す．

timed_caption の情報から，時間当たりのの平均発話単語数を計算したところ，1 秒間に 1.94 単語発話している事がわかった．この結果は，画像キャプションの長さコントロールの際に参考にする．

表 1　Localized Narratives データ内訳

dataset_id	データセットの種類
image_id	画像 ID
annotator_id	アノテーションの ID
timed_caption	単語・開始秒数・終了秒数
traces	x 座標・y 座標・秒数
voice_recording	音声データの URL

2.3　トレースによる説明領域の抽出

画像の中でも，説明者が注目した部分からキャプションを生成する為に，トレースの描画範囲から説明領域を抽出する．画像説明時における人の特性を解明する為に，以下の 2 つの特徴量を抽出した．

座標の変化量　横軸に時間をとり，縦軸にトレースに関する変化量（x 座標 ×y 座標）をとったグラフを図 2 に示す．赤色でハイライトしている部分は，説明者が説明を止めている文と文の間の時間であり，画像の説明に一区切りついた箇所に相当する．赤い部分のグラフは平らになっており，トレースの変化量はほぼ変化がない事がわかった．同じような現象が他のデータにも多く見受けられた．この結果は，説明者は 1 文説明し終わった際，少し止まってから次の文の説明に移るという特性があることを示唆する．画像説明時のトレースにおける人の行動特性は必ずしも常に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．

トレースの移動速度　トレースによる画像説明時における人の 2 つ目の行動特性として，特別に動きが速いトレースは，説明対象となるオブジェクト間の移動の為のトレースであり，その部分を説明をしている訳ではないという点に着目した．トレースの動作の速度を，速いほど黒に，遅いほど赤に対応させて可視化したものを図 3 に示す．図 4 は，図 3 において，オブジェクト間の移動とみなされた部分を削除して可視化した．実際に，芝生から木への移動，木から山への移動，空から雲への移動の箇所が削除され，説明対象となった各オブジェクトのみを指しているトレースが抽出されたことがわかる．

2.4　LaBERT [12]

説明者の注目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手法として，文長制御が可能な非自己回帰型のキャプション生成を行う LaBERT [12] のデコーダを用いる．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 [12] 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

	1	2	3	4
説明領域				
生成キャプション	a set of four four different pictures with four planes on them and a suitcases on the ground.....	two people are looking at the direction of the sky.....	two people are standing in the grass together.	a room filled with lots of furniture..
正解キャプション	I can see in this image a bag of a brown color.	I can also see two man among them a woman and a man is standing on the ground.	The woman is holding a bag in her hand.	In the background I can see a building and a statue.
トレースの滞在時間	10.9	8.2	4.0	4.2
推定単語数	21.15	15.91	7.76	8.15
文長の指定	20〜25	15〜19	7〜9	7〜9
BLEUスコア	1.267	1.199	2.552	2.573

図 5　トレースを入力とした説明者の意図に沿った画像キャプション生成結果

を実現した．

LaBERT のデコーダにおける処理の概要を図 6 に示す．また、そのアルゴリズムを Algorithm1 に示す．

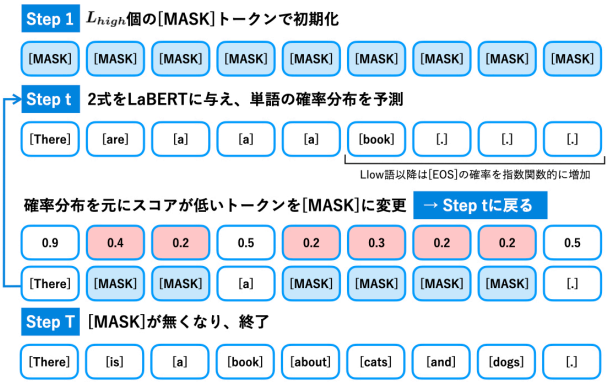

図 6　LaBERT を用いた長さ L_{low}〜L_{high} の文生成概要

Algorithm 1 Length-Controllable Caption Generation

Require: キャプション S を L_{high} 個の [MASK]token で初期化．T: 文のアップデート回数 e_i: 文長埋込ベクトル, $e_{w,s}$: 単語埋込ベクトル, e_p: 位置埋込ベクトル, f_e: 範囲属性, f_c: 分類確率, f_l: 局所属性, LN:Layer Normalization [17], e_{img}: 画像領域とトークンを区別する学習可能な埋込ベクトル,W_e, W_p:, 対応する特徴を d-D 空間に射影する 2 つの学習可能な射影行列 i is 1〜L_{high}.

while $t \leq T$ **do**
  if s_i is $[MASK]$ **then**
    $x_{s_i} = e_i + e_{w,s_i} + e_{p,i}$
    $x_{r_i} = W_e^T f_{e,i} + W_p^T [LN(f_{c,i}), LN(f_{l,i})] + e_{img}$
    $p_i \leftarrow LaBERT(x_{s_i}, x_{r_i})$
  end if
  $c_i \leftarrow max_s p_i(s_i = s)$
  if $c_i \leq \min(\frac{T-t}{T} L_{high}, c)$ **then**
    $s_i \leftarrow [MASK]$
  end if
  $t = t + 1$
end while

3　実験

画像とトレースを入力として，トレースに沿った画像キャプションを生成する．

3.1　実験設定

画像データは Microsoft COCO を使用し， Cornia ら [18] が使用している Faster-RCNN を各画像に施し取得した B.Box の特徴量をまとめたデータセット coco _detections.hdf5[1] から，2048 次元の画像特徴量，B.Box の位置を示す 4 次元の座標，オブジェクトを示す 1601 次元のカテゴリーを用いた．生成文の単語数は，7〜9words, 10〜14words, 15〜19words, 20〜25words の 4 種類を設定し，文のアップデート回数は 4 種類に対し，10 回, 15 回, 20 回, 20 回に設定した．また，言語モデルは事前学習済み $\text{BERT}_{\text{BASE}}$[2] を用いた．この実験設定の下，バッチサイズ 256，イテレーション 100,000 回でモデルの学習を行なった．

3.2　実験結果

トレースを入力として，説明者の意図に沿うように画像の説明したい箇所にトレースを与えた情報に基づきキャプション生成を行った．結果の例を図 5 に示す．図 5 中の各項目について以下に説明する．

説明領域　トレースの座標変化量と速さを元に説明対象となる領域（図中，青い四角形の領域）を抽出したものとなる．各画像には，抽出されたトレースと B.Box（赤い四角形）が写っており，説明対象領域を $\frac{1}{4}$ 以上含む B.Box から画像特徴量を取得している．また，黒いトレースはオブジェクト間の移動

1）　https://github.com/aimagelab/meshed-memory-transformer

2）　https://huggingface.co/bert-base-uncased

とみなし，削除した．

正解キャプション　実際に説明者が発話したキャプション．

トレースの滞在時間　説明領域内にトレースが滞在した秒数．

推定単語数　生成文を構成する単語数．トレースの滞在秒数に，LN における 1 秒間あたりの平均発話単語数 1.94 を掛けて推定した．

文長の指定　推定単語数から選ばれる文の長さ．現在，3.1 節で示した 4 つの範囲を設定している．

BLEU スコア　文生成の精度の評価指標として，BLEU スコア [19] を示している．

各説明領域での結果を以下に示す．

図 *5-1*：suitcase というオブジェクトは捉えていたものの，推定単語数が正解キャプションの単語数に比べて多くなっていた．

図 *5-2*：オブジェクトの認識やキャプションの長さなどトレースの意図を捉えられ，期待した結果が得られた．

図 *5-3*：正解キャプションでは「女性のカバン」について言及しているが，周りの情報を取り込んで two people の説明になってしまっている．

図 *5-4*：背景のみの描写なので，認識できるオブジェクトが少ないこともあり，想定していたキャプションが生成されなかった．

3.3　考察

実験結果と正解キャプションおよび付録 A に示した様々な文長の下での生成結果と比較しながら考察する．

図 *7-1*：トレースの滞在時間から推測した 20〜25 単語ではなく，7〜9 単語の結果の方が正解キャプションに近いと考えられる．また，この場合 BLEU スコアも 1.72 と少し上がっていた．正解キャプションも 12 単語である為，トレースの滞在時間ほど長い文による説明は求められていない．これは，図中の「カバン」というオブジェクトが画像全体の比率の大部分を占めていることから，ひとつのオブジェクトを指し示すのにトレースの滞在時間が長くなってしまった為と考えられる．このことは，文長を 1 秒あたりの平均発話語数のみで決めるのではなく，説明対象となる領域に含まれるオブジェクトや領域の面積も考慮すべきであると考える．

図 *7-2*：全ての文長の場合で 2 人の人間を捉えられていたが，どこに立っているかなど，背景の情報は不揃いで正確に捉えられていないことがわかった．

図 *7-3*：「女性のカバン」を言及出来ていない点について，このように誰かの所有物や食べ物の具など，特定の部分に注目して説明する場合は Dense captioning [20] などのような局所的な説明を可能にする画像キャプション生成方法などを参考に改良する必要があると考える．

図 *7-4*：10〜14words, 15〜19words の生成途中で，正解キャプションに含まれる building が出現したにも関わらず，最終的な生成キャプションとして残らなかった様子が見られた．これは，画像特徴量よりも言語モデルが優先されてしまった結果生じた事例だと考える．一概に t の値が大きくなる程，良い単語に置き換わる訳ではないことがわかった．

また，今回，キャプション生成の評価指標として採用した BLEU のスコアはとても低くなっている．この原因の一つとして，正解文とする LN のキャプションは，画像全体を一括して説明するものになっており，今回のように部分的な領域での説明と整合性がとれない部分もあることが考えられる．

4　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 Localized Narratives 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 を組み合わせ，トレースから説明者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．説明対象となる領域の選択やキャプションの長さは，LN の統計量から求めた値を採用したが，説明意図を表現するキャプション生成に必要となる画像特徴量の適切な抽出や，個々の説明領域内の物体の在り様などをより踏まえて、キャプション生成する必要があることがわかった．

今後の課題として，これらの問題に取り組みつつ，説明領域全体を俯瞰する観点からのキャプション生成も可能にしたい．

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] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[4] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph R-CNN for scene graph generation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part I*, Vol. 11205 of *Lecture Notes in Computer Science*, pp. 690–706. Springer, 2018.

[5] Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. Improving image captioning with better use of caption. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7454–7464, Online, July 2020. Association for Computational Linguistics.

[6] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10677–10686, 2019.

[7] Wenqiao Zhang, Haochen Shi, Siliang Tang, Jun Xiao, Qiang Yu, and Yueting Zhuang. Consensus graph representation learning for better grounded image captioning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 3394–3402. AAAI Press, 2021.

[8] 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.

[9] Long Chen, Zhihong Jiang, Jun Xiao, and Wei Liu. Human-like controllable image captioning with verb-specific semantic roles. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 16846–16856. Computer Vision Foundation / IEEE, 2021.

[10] Kun Yan, Lei Ji, Huaishao Luo, Ming Zhou, Nan Duan, and Shuai Ma. Control image captioning spatially and temporally. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2014–2025, Online, August 2021. Association for Computational Linguistics.

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

[12] Chaorui Deng, Ning Ding, Mingkui Tan, and Qi Wu. Length-controllable image captioning. *CoRR*, Vol. abs/2007.09580, , 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] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016.

[18] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[19] 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.

[20] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

A　付録

		1	2	3	4
説明領域					
生成キャプション		**a set of four four different pictures with four planes on them and a suitcases on the ground.....**	**two people are looking at the direction of the sky.....**	**two people are standing in the grass together.**	**a room filled with lots of furniture..**
正解キャプション		I can see in this image a bag of a brown color.	I can also see two man among them a woman and a man is standing on the ground.	The woman is holding a bag in her hand.	In the background I can see a building and a statue.
トレースの滞在時間		10.9	8.2	4.0	4.2
推定単語数		21.15	15.91	7.76	8.15
文長の指定		20～25	15～19	7～9	7～9
BLEUスコア		1.267	1.199	2.552	2.573
7~9 words	t=2	a suitcase of with with of of of.	a couple of walking a of a a.	two people are are a of of..	a lot of with with of of it.
	t=3	a suitcase of with with of of it.	a couple of people standing a a together.	two people are are a of of together.	a lot of with with of of it.
	t=4	a suitcase is is a of the..	a couple of people walking in a together.	two people are are a of of together.	a large filled with many and and it.
	t=5	a suitcase is sitting in the the..	a couple of people standing in a together.	two people are standing in a of together.	a room filled with lots of and..
	t=6	a suitcase is sitting on the green..	a couple of people standing some some together.	two people are standing on the field together.	a room filled with lots of and..
	t=7	a suitcase is sitting on the green..	a couple of people standing by some field.	two people are standing on the empty together.	a room filled with lots of different..
	t=8	a suitcase is sitting on the wooden..	a couple of people standing by some grass.	two people are standing in the field together.	a room filled with lots of green..
	t=9	a suitcase is sitting on the ground..	a couple of people standing by some trees.	two people are standing in the grass together.	a room filled with lots of furniture..
	t=10	a suitcase is sitting on the ground outside.	a couple of people standing by some trees.	two people are standing in the grass together.	a room filled with lots of furniture..
10~14 words	t=2	a is of of of of of of of of....	a couple of are a of of of of of....	a people of are a of of of of of....	a are of with with with with and a a....
	t=3	a is of of with of of of of the....	a people of are a of of of of of....	two people of are a of of of of of....	a are of with with with with and a a....
	t=4	a suitcase of with with of of of of....	a people of are a of of of of the....	two people of are a of of of of the....	a picture of with with of of and of a....
	t=5	a suitcase of with with of of of of it....	a couple of a a of of of of of....	two people are are a of of of of of....	a picture of with with of of and the it....
	t=6	a bunch of four that are are in a a....	a couple of people that are a of a the....	two people are in a of of of of of....	a bunch of five that are are in a a....
	t=7	a bunch of four that are on on a area....	a couple of people that are a of a the....	two people are looking the the of of the area....	a bunch of five that are in a a area....
	t=8	a bunch of different that are in in a other....	a couple of people that are a of some some....	two people are looking the the of of the area....	a bunch of green that are in in a building....
	t=9	a bunch of animals that are on in the other....	a couple of people that are standing in some some....	two people are looking at the directions of the area....	a lot of signs that are on in a building....
	t=10	a bunch of animals that are standing in the dirt....	a couple of people that are standing in some grass....	two people are looking at the direction of the area....	a lot of signs that are standing in the dirt....
	t=11	a bunch of animals that are standing in the dirt....	a couple of people that are standing in some grass....	two people are looking at the direction on the picture....	a lot of trees that are standing in the dirt....
	t=12	a bunch of animals that are sitting in the dirt....	a couple of people that are standing in some grass....	two people are looking at the directions on the wall....	a lot of trees that are out in the dirt....
	t=13	a bunch of animals that are standing in the dirt....	a couple of people that are standing in some grass....	two people are looking at the directions on the wall....	a lot of trees that are standing in the dirt....
	t=14	a bunch of animals that are standing in the dirt....	a couple of people that are standing in some dirt....	two people are looking at the signs on the wall....	a lot of trees that are standing in the dirt....
	t=15	a bunch of animals that are standing in the dirt....	a couple of people that are standing in some dirt....	two people are looking at the signs on the wall....	a lot of trees that are standing in the dirt....
15~19 words	t=2	a is of of of of of of a a.....	a couple of a a of of of of of.....	a couple of a a of of of of of.....	a are of with with of with a a a.....
	t=3	a are of of of of of of a a.....	a couple of a a a of of of of.....	a couple of a a a of of of of.....	a are of with with of with a a a.....
	t=4	a are of of of of of of a a.....	a couple of in a of of of of of.....	a couple of a a a of of of of.....	a are of with with of with a a a.....
	t=5	a are of of of of of of a a.....	a couple of in a of of of of of.....	a couple of a a a of of of of.....	a are of with with of with a a a.....
	t=6	a are of of of of of of a the.....	a couple of are a of of of of a.....	a couple of are a of of of of of.....	a are of with with of with in a a.....
	t=7	a are of of of of of of a the.....	a couple of are a of of of of a.....	a couple of are a of of of of of.....	a are of with with of with in a a.....
	t=8	a are of of of of of of a the.....	two couple of are a of of of of a.....	a couple of are a of of of of of.....	a are of with with of with in a a.....
	t=9	a are of of of of of of a the.....	two couple of are a of of in a a.....	two couple of are a of of of of the.....	a are of with with of of in a the.....
	t=10	a suitcase of with with of of of a it.....	two couple of are a are of of a area.....	two couple of are a are of of the area.....	a picture of with with of of and a it.....
	t=11	a suitcase of of with of of of the it.....	two people are are a a of a a area.....	two people are are a of of a a area.....	a picture of with with of of and the it.....
	t=12	a bunch of six that are on in a area.....	two people are looking the the of of a area.....	two people are in a of of of of area.....	a bunch of five that are in a a building.....
	t=13	a bunch of green that are sitting in a other.....	two people are looking at the of of the area.....	two people are standing together a a lots of them.....	a bunch of green that are sitting in a building.....
	t=14	a bunch of animals that are sitting in the other.....	two people are looking at the directions of the area.....	two people are standing together on a lot of area.....	a bunch of signs that are sitting in a building.....
	t=15	a bunch of animals that are standing in the dirt.....	two people are looking at the direction in the area.....	two people are standing in a a full of land.....	a lot of trees that are standing in the dirt.....
	t=16	a bunch of animals that are standing in the dirt.....	two people are looking at the direction of the building.....	two people are standing in a a lot of land.....	a lot of trees that are standing in the dirt.....
	t=17	a bunch of animals that are standing in the dirt.....	two people are looking at the direction of the wall.....	two people are standing in to a lot of land.....	a lot of trees that are standing in the dirt.....
	t=18	a bunch of animals that are standing in the dirt.....	two people are looking at the direction of the wall.....	two people are standing next to a lot of trees.....	a lot of trees that are standing in the dirt.....
	t=19	a bunch of animals that are standing in the dirt.....	two people are looking at the direction of the sky.....	two people are standing next to a lot of trees.....	a lot of trees that are standing in the dirt.....
	t=20	a bunch of animals that are standing in the dirt.....	two people are looking at the direction of the sky.....	two people are standing next to a lot of trees.....	a lot of trees that are standing in the dirt.....
20~25 words	t=2	a is of of a of of a a a a and and and and a the the.....	a man of a a of one one one one one one one one one one the the.....	a man of a a of of one one one one one and one one one one the the.....	a view of with with with with and and and and and and and and and and and the the.....
	t=3	a is of of a of of a a a a and and and and a a the the.....	a man of a a of one one one one one one one one one one the the.....	a man of a a of one one one one one one one one one one one the the.....	a view of with with with with and and and and and and and and and and a the the.....
	t=4	a is of of a of of with a a a and and and and a a the the.....	a man of a a of one one one one one one one one one one the the.....	a man of a a of one one one one one one one one one one one the the.....	a view of with with with with and and and and and and and and and and a on the the.....
	t=5	a is of of a of of with a a a and and and and a a the the.....	a man of a a of one one one one one one one one one one the the.....	a man of a a of one one one one one one one one one in the the.....	a view of with with with with and and and and and and and and and and a on the the.....
	t=6	a brown of with four of of with and and and and and and a and on the it.....	a man of a a of of one one one one and and one one one in the them.....	a man of a a of of a one one one and and and one one in the them.....	a photo of with with with with and and and and and and and and and and on the it.....
	t=7	a view of four with with four with with with a and and and a a on the it.....	a pair of fours in a one one one one one one one one one on the them.....	a man is in a of a a of of of of of of of of the them.....	a view of four with with with with with with with and and and one on on the them.....
	t=8	a set of four different of four a one with and and and and and and on the them.....	a couple of three are in in of with with with with with with with in front of them.....	a man is on a of a a of of of with with of thes on the distance.....	a view of a four with with with with with with and and and and and and on the it.....
	t=9	a set of four four of four four with with and and and and and and on the them.....	a group of four standing in a a with with with with with a on in front of them.....	a man is on a of a a of of of with with with one on the distance.....	a view of a room with with with with with with and and and and and on the it.....
	t=10	a set of four four different with with four with and and and and the one on the background.....	a couple of three standing in a of with with with with with with in front of them.....	a man stands on a window with with a of of with with with with one on the distance.....	a view of a building with with four and and and and and and and and and the it.....
	t=11	a set of four four different with with with with with and and and ands on the ground.....	a couple of people standing on a of of with with with with ofs in front of them.....	a man leans on a window sill a a of with with with with seen on the distance.....	a view of a building with with four different and and and and and and and and the it.....
	t=12	a set of four four different with with with with with and and and is still on the ground.....	a couple of people standing in a each with with with of of ofs in front of them.....	a man sits on a window sill with a of with with with withs on the distance.....	a view of a building with with four different and one one and as one on the it.....
	t=13	a set of four four different with with with with thes and four suitcases on the ground.....	a couple of people standing next to each other with a of of ofs in front of them.....	a man sitting on a window sill with a of of with with with are on the distance.....	a view of a building with with four different and the one and a suitcases on the background.....
	t=14	a set of four four pictures with with three on the them and four suitcases on the ground.....	a couple of people standing next to each other with a lot of ofs in front of them.....	a man sitting on a window sill looking front of of area with fours on the distance.....	a view of a building with with four faces on the left and three browns on the left.....
	t=15	a set of four four differents with fours on them and four suitcases on the ground.....	a couple of people standing next to each other with a of of ofs in front of them.....	a man sitting on a window sill looking at a a window with fours on the distance.....	a view of a building with with four faces on the left and three browns on the left.....
	t=16	a set of four four different four with fours on them and a suitcases on the ground.....	a couple of people standing next to each other with a of of withs in front of them.....	a man sitting on a window sill looking at a distance of with fours on the distance.....	a view of a building with with four faces on the left and three different one on the left.....
	t=17	a set of four four different four with fours on them and a suitcases on the ground.....	a couple of people standing next to each other with a row of green running in front of them.....	a man sitting on a window sill looking at a window window with four sky on the distance.....	a view of a building with with four planes on the right and three different one on the left.....
	t=18	a set of four four different pictures with four planes on them and a suitcases on the ground.....	a couple of people standing next to each other with a bunch of green running in front of them.....	a man sitting on a window sill looking at a round area with four planes on the distance.....	a view of a building with with four planes on the right and three different one on the left.....
	t=19	a set of four four different pictures with four planes on them and a suitcases on the ground.....	a couple of people standing next to each other with a bunch of green running in front of them.....	a man sitting on a window sill looking at a grassy area with four planes on the distance.....	a view of a building with with four planes on the right and three different directions on the left.....
	t=20	a set of four four different pictures with four planes on them and a suitcases on the ground.....	a couple of people standing next to each other with a bunch of green hills in front of them.....	a man sitting on a window sill looking at a grassy area with four planes on the distance.....	a view of a building with with four directions on the right and three different directions on the left.....

図 7　様々な文長での画像キャプション生成結果