画像キャプション生成におけるシーングラフ特徴量の効果

田屋 侑希 † 欅 惇志 ‡ 塚原 裕史 ‡ 小林 一郎 †
† お茶の水女子大学 ‡ 株式会社デンソーアイティーラボラトリ

†{g1620525,koba}@is.ocha.ac.jp, ‡{akeyaki,htsukahara}@d-itlab.co.jp

1 はじめに

近年，深層学習を用いた画像のキャプション生成に関する研究が盛んに行われている [5]．初期の研究では画像特徴量を用いたキャプション生成が中心として取り組まれていたが，画像中に含まれる人や物などの物体とその属性，及び，物体間の関係に注目したキャプション生成を目指し，シーングラフを用いたキャプション生成の研究も進められている [8][6][7]．シーングラフを用いた研究は多数存在するものの，それを用いることで得られる効果について，特に物体 · 属性 · 関係ごとに着目した検証は十分に議論されておらず，主に BLEU などの定量的な評価尺度の改善についての議論が中心である．そこで本研究では，画像特徴量のみから生成されたキャプションと画像特徴量とシーングラフの両方を用いて生成されたキャプションの比較を行うことで，シーングラフを用いることで得られる効果について分析を行う．

2 関連研究

2.1 画像キャプション生成

画像を入力として，その画像の説明文（キャプション）を生成する言語処理タスクは画像キャプション生成と呼ばれる．主流なアプローチは，まず入力画像に対して CNN を適用することで画像特徴量を抽出し，次に得られた画像特徴量を LSTM に入力することで説明文を生成する [5]．

このような技術によって人間に理解可能な画像キャプション生成が実現しつつある．しかし，生成されるキャプションの性質の観点において，必ずしも人間の直感と一致する物体について言及されているとは限らず，更には，画像内における複数の物体間の関係やその物体の補助的な説明 (属性記述) 能力も不十分である．そこで，画像中の物体とそれらの関係を認識して構築されるシーングラフを用いることで，より記述力の高いキャプション生成を目指す研究が取り組まれている [8][6][7]．

2.2 シーングラフを用いたキャプション生成

シーングラフの構築に先立ち，まずは画像内に含まれる物体の認識が行われる．その際，Faster-RCNN[4] など画像処理技術が多用される．認識された各物体は矩形 (画像中の座標) とそのラベル (man, phone, tree など) で表現されることが一般的である．物体は独立して認識されるため，矩形の間の重畳は認められている．このような重畳，もしくは近接した物体間の中には何らかの関係 (near, next to など) を持つものがあり，物体認識技術を拡張して取得可能である．また，物体の持つ属性 (young, tall など) においても同様に認識される．

物体や関係，属性が認識されれば，シーングラフが構築される．シーングラフとは，画像中の物体 · 関係 · 属性をノード，物体-関係間と物体-属性間の関係を有向エッジとして表現した有向グラフのことである．シーングラフの例を図 1 に示す．2 つの物体とそれら

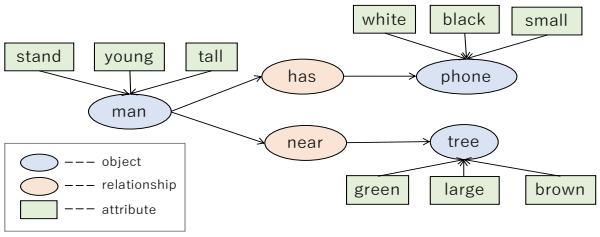


図 1: シーングラフの例

を結ぶ関係から構成される三つ組は <*subject, predicate, object*>として表現され，subject から predicate, predicate から object にエッジが張られている．ただし，subject は文において主語となる物体，object は目的語になる物体，predicate は subject と object の間に存在する関係を表す動詞または前置詞とする．なお，本研究において predicate は relationship と記述し，subject と object を区別せずに “物体” のことを object と表記する．また，ある物体が任意の個数 N 個の属性を持ち (<*object, attribute 1, ···, attribute N* >)，属性から物体にエッジが張られている．これらの構成要素からなる画像内容を捉えたグラフを 1 つのシーングラフとして定義する．図 1 の例においては，<*man, has, phone*>と<*man, near, tree*>の 3 つ組，<*man, stand, young, tall*>，<*phone, white, black, small*>，<*tree, green, large, brown*>の 4 つ組が存在する．

続いて，Marcheggiani ら [3] などの手法が用いられてシーングラフからグラフ特徴量が抽出される．最終的なキャプション生成において，画像内で認識された各物体の画像特徴量とグラフ特徴量が LSTM に入力される手法 [6] や，外部知識で拡張されたグラフ特徴量を LSTM に入力することでキャプション生成を行う研究 [8] などが存在する．

本研究では，まずは単純にシーングラフを用いた場合の効果を検証するため，画像特徴量とシーングラフから抽出されたグラフ特徴量をナイーブに連結することで画像のキャプションへの影響を検証する．

3 キャプション生成手法

本研究では，画像特徴量とグラフ特徴量を用いて画像キャプション生成を目指す．図 2 に提案手法の概要を示す．入力画像を CNN に適用して画像特徴量を抽出する．その一方で，画像から構築されるシーングラフからグラフ特徴量を抽出する．これら特徴量を連結し LSTM に入力することで説明文を生成する．この一連の流れは， Vinyals ら [5] の手法を参考にした．また，画像に対するシーングラフは Yang ら [8] の研究において生成されたものを利用した．Yang らの研究では，画像とそれに対応するシーングラフが正解文として与えられているデータセット Visual Genome[1] を利用して，Microsoft COCO のデータに対してシーングラフを生成している．シーングラフを生成するにあたり，object は 305 語，predicate は 64 語，attribute は 103 語に絞り学習している．また，Yang らの研究で用いられている spatial Graph Convolutional Networks(GCNs)[2][3] を参考にして，シーングラフをベクトルに変換した．その手順を図 3 に示す．

Relationship Embedding:

2 つの object とその間の relationship の三つ組 $<o_i, r_{ij}, o_j>$, 得られる relationship のベクトルを $\boldsymbol{x}_{r_{ij}}$ と定義する．また，object, relationship, attribute, それぞれ単語埋め込みベクトルで表現する際に，pre-trianed モデルの word2vec[1]を使用する．変換されたベクトルをそれぞれ $\boldsymbol{e}_{o_i}$，$\boldsymbol{e}_{r_{ij}}$，$\boldsymbol{e}_{o_j}$ とし，それぞれ 300 次元のベクトルで表される．

$$\boldsymbol{x}_{r_{ij}} = (\boldsymbol{e}_{o_i}, \boldsymbol{e}_{r_{ij}}, \boldsymbol{e}_{o_j}) \tag{1}$$

この時，() 内のベクトルをそれぞれ縦に連結させる．そのため，$\boldsymbol{x}_{r_{ij}}$ は 900 次元となる．

Attribute Embedding:

オブジェクト o_i に対する attribute を a_{il} $(l = 1 \sim N_{a_i})$ と定義し，得られる attribute のベクトルを $\boldsymbol{x}_{a_i}$ と定義する．$\boldsymbol{x}_{a_i}$ は o_i に対する属性を表したベクトルになる．また， Relationship Embedding と同様に変換されたベクトルを $\boldsymbol{e}_{o_i}$, $\boldsymbol{e}_{a_{il}}$ $(l = 1 \sim N_{a_i})$ とし，それぞれ 300 次元のベクトルで表す．

$$\boldsymbol{x}_{a_i} = \frac{1}{N_{a_i}} \sum_{l=1}^{N_{a_i}} (\boldsymbol{e}_{o_i}, \boldsymbol{e}_{a_{i,l}}) \tag{2}$$

N_{a_i} は o_i に対して付与されている属性の個数とする．$\boldsymbol{x}_{a_i}$ は 600 次元となる．

Object Embedding:

本研究においてシーングラフは有向グラフであるため，オブジェクト o_i は <*subject, predicate, object*>における object の場合と subject の場合がある．それぞれの場合において区別し embedding を行い，得られる object のベクトルを $\boldsymbol{x}_{o_i}$ と定義する．$\boldsymbol{x}_{o_i}$ は o_i に隣接した relationship と object を含んだベクトルになる．

$$\boldsymbol{x}_{o_i} = \frac{1}{N_{r_i}}[\sum_{o_j \in sbj_{o_i}} (\boldsymbol{e}_{o_i}, \boldsymbol{e}_{o_j}, \boldsymbol{e}_{r_{ij}})] + [\sum_{o_k \in obj_{o_i}} (\boldsymbol{e}_{o_k}, \boldsymbol{e}_{o_i}, \boldsymbol{e}_{r_{ki}})] \tag{3}$$

N_{r_i} は o_i に隣接している relationship の個数とする．また sbj_{o_i} は<*subject, predicate, object*>の 3 つ組において，o_i の subject となる “物体” のことであり，同様に，obj_{o_i} は o_i の object となる “物体” を指す．$\boldsymbol{x}_{r_i}$ は 900 次元となる．

4 実験

シーングラフから得られるグラフ特徴量を追加することで得られる影響を調査するため，画像特徴量のみを用いた場合と，画像特徴量とグラフ特徴量を用いた場合のキャプション生成結果を比較して，どのように変化するかを分析する．定性的な評価としてシーングラフを追加することによってキャプションの性能が向上した例を挙げ，定量的な評価として BLEU を計測した．

4.1 使用データ

本研究ではデータセットとして Microsoft COCO を使用する．静止画像とその説明文のペアのデータセットであり，1 画像あたり 5, 6 文の英語の説明文が付与されている．訓練データは 82,783 画像， テストデータは 40,504 画像用意されている．

4.2 実験設定

画像キャプション生成におけるコードは深層学習フレームワークである chainer[2] を用いて実験を行った．ハイパーパラメータの設定を表 1 に示す．

4.3 実験結果

入力画像に対して，画像特徴量のみからキャプションを生成した結果と，画像特徴量とシーングラフにお

[1] https://code.google.com/archive/p/word2vec/

[2] https://github.com/chainer/chainer

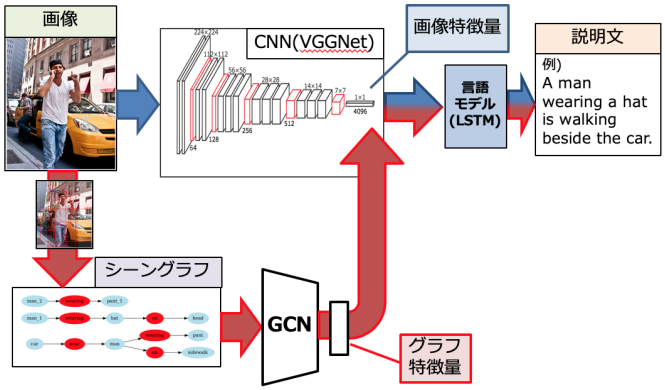


図 2: 本研究の概要図

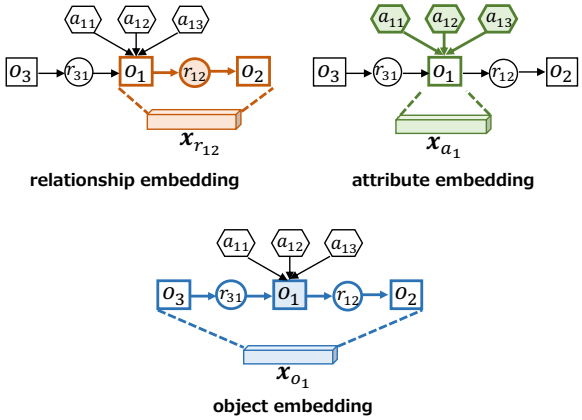
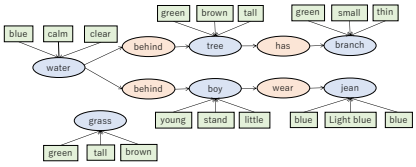


図 3: 本研究での Graph Embedding

表 1: 実験設定

画像→画像特徴量→説明文モデル	
データセット	Microsoft COCO
学習量	82,783 画像 ×100 epoch
アルゴリズム	Adam
学習率	0.001
eps	1e-8
勾配閾値	1
L2 正則化項	0.005
語彙	頻出語 3,469 語
誤差関数	交差エントロピー

表 2: object に着目した結果



画像特徴量のみから生成：
A man standing in the sand with an umbrella.

画像特徴量＋グラフ特徴量から生成：
A man is playing **frisbee** in the **grass**.

正解文
Chubby little kid is making a face near a lake
a child in a field with a **frisbee** near a tree
a child is standing outside in the **grass**
A young boy is throwing a **frisbee** on the **grass** by a lake and by a tree.
A boy is playing with a **frisbee** under the tree next to the water.

けるグラフ特徴量から出力したキャプションの例を挙げる．出力結果から，画像 100 枚を無作為にサンプリングして，キャプション生成結果を確認した．その中で，最も尤度の高いキャプションに着目した．以下に，object, attribute, relationship それぞれに着目して，画像特徴量とグラフ特徴量からのキャプション生成の品質が顕著に改善した結果を報告する．その際，差分となる語を**太字**で表記する．なお，画像特徴量のみから生成したキャプションと画像特徴量とグラフ特徴量から生成したキャプションでほぼ同一の文が出力された画像も多数観察された (同等:改善 = 8:2)．シーングラフは一部抜粋したものを示す．

4.3.1 OBJECT に着目した結果

object に着目し，シーングラフを利用した場合に結果が改善した例を表 2 に示す．

正解文中に出現する “grass” は，画像特徴量のみを用いたキャプション中には出現しない．その一方で，シーングラフにおいて “grass” は object として認識され，グラフ特徴量を用いたキャプション中にも “grass” は出現している．この結果から，画像特徴量のみを用いた場合よりも正しい object をキャプションに取り入れていることが確認できる．また，同様に正解文に出現している “frisbee” は，シーングラフとしては認識されていないものの，グラフ特徴量を用いたキャプションには正しく出現している．これは，グラフ特徴量として “grass” が認識されることで，画像特徴量が “frisbee” を認識する際の補助を行っていると推測される．

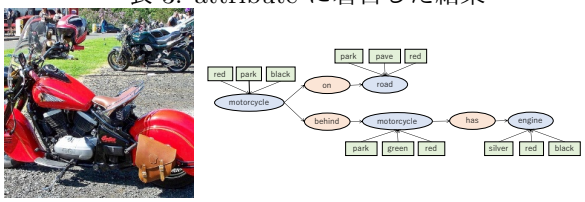
4.3.2 ATTRIBUTE に着目した結果

attribute に着目した結果を表 3 に示す．シーングラフ内において，“motorcycle” に対する attribute として “parked” が捉えられており，その結果，グラフ特徴量を用いたキャプション中において “parked” が出力されている． 画像特徴量のみから生成した文では実際には存在しない “man” が出力されている点からも，グラフ特徴量を用いることでより適切な文を生成することができている．

4.3.3 RELATIONSHIP に着目した結果

最後に relationship に着目した結果を表 4 に示す．画像特徴量によるキャプション生成では “kitchen”, “oven stove”, “refrigerator” といった object が “with” や “and” で列挙されている．それに対して，グ

表 3: attribute に着目した結果

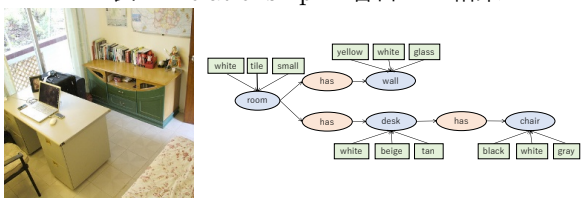


画像特徴量のみから生成：
A man riding on the back of an old motorcycle.

画像特徴量＋グラフ特徴量から生成：
A motorcycle is **parked** on the street.

正解文
a red motorcycle **parked** on some gravel next to grass
A red motorcycle **parked** in a parking lot space.
A red and black motorcycle with a brown satchel **parked** in a lot.
A red motorcycle **parked** close to other motorcycles.
A custom red motorcycle left unattended in a **parking** lot.

表 4: relationship に着目した結果



画像特徴量のみから生成：
A kitchen with an oven stove and refrigerator.

画像特徴量＋グラフ特徴量から生成：
A living room **filled with** furniture and chairs.

正解文
A very orderly office with a bed in it was very well lit cause of open windows.
a laptop sitting on a desk in a small office
A home office **has** a desk and a laptop.
A room with a desk and a dresser.
A bedroom scene complete with a desk and laptop

ラフ特徴量を用いた結果では，object 間の relationship である <*room*, *has*, *desk*>や <*desk*, *has*, *chair*>などの 3 つ組から “living room” と “furniture and chairs” の間に成立する関係として “filled with” がキャプションとして生成できていると推測される．“filled with” は “with” よりも詳細な記述であるため，グラフ特徴量を用いることでキャプションの表現力がより高まっていると考えられる．

4.3.4 定量的結果

BLEU による評価結果について報告する．BLEU@1(平均) は unigram において各正解文とキャプションどれだけ一致しているかを計測する尺度であり，以下の式によって算出される． ただし，I はテスト用画像のうちの無作為にサンプリングした 500 枚，REF は正解文 (5, 6 文)，cap は本研究で生成した文とする．

$$\frac{1}{|I|}\sum_{i\in I}\frac{1}{|REF|}\sum_{r\in REF}(BLEU(cap,r)) \quad (4)$$

また，BLEU@1(最大) は以下の式で定義する．

$$\frac{1}{|I|}\sum_{i\in I}max(BLEU(cap,REF)) \quad (5)$$

表 5: BLEU による評価

	BLEU@1(平均)	BLEU@1(最大)
画像特徴量	29.0	41.9
画像特徴量＋グラフ特徴量	**29.4**	**42.4**

画像特徴量から生成されたキャプションに対する BLEU と画像特徴量とグラフ特徴量から生成されたキャプションの BLEU を表 5 に示す．表の結果より，画像特徴量とグラフ特徴量から生成したキャプションの性能がわずかに高い結果となった．

4.4 考察

主観による定性評価と BLEU による定量評価ともに，グラフ特徴量を用いた場合に，画像特徴量のみを用いた場合と同等もしくはより良い結果が確認できた．また，2 節の object に着目した結果の “frisbee” の出力のように，シーングラフで認識された物体が別の物体認識の補助を行うことができる例も確認された．

5 おわりに

本研究では，画像キャプション生成において，画像特徴量のみを用いた場合と画像特徴量とグラフ特徴量を用いた場合のキャプションを比較することで，シーングラフを用いることの効果を検証した．実験の結果，シーングラフとして認識される物体，属性，関係それぞれにおいて有用な特徴が獲得できる例が確認された．本稿によってシーングラフを用いることの有用性が部分的に検証されたため，今後はより高度なシーングラフの利用を目指す．指定したオブジェクトに着目したキャプション生成はその一つである．

参考文献

[1] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and F.-F. Li. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123:32–73, 2017.

[2] Y. Li, R. Zemel, M. Brockschmidt, and D. Tarlow. Gated Graph Sequence Neural Networks. In *Proc. of ICLR*, 2017.

[3] D. Marcheggiani and I. Titov. Encoding Sentences with Graph Convolutional Networks for Semantic Role Labeling. In *EMNLP*, 2017.

[4] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*, 2015.

[5] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and Tell: A Neural Image Caption Generator. In *Proc. of CVPR*, 2015.

[6] D. Wang, D. Beck, and T. Cohn. On the Role of Scene Graphs in Image Captioning. In *EMNLP*, 2019.

[7] D. Wang, D. Beck, and T. Cohn. Unpaired Image Captioning via Scene Graph Alignments. In *ICCV*, 2019.

[8] X. Yang, K. Tang, H. Zhang, and J. Cai. Auto-Encoding Scene Graphs for Image Captioning. In *Proc. of CVPR*, 2019.