シーングラフ特徴量を反映した画像キャプション生成への取り組み

田屋侑希　(指導教員：小林一郎)

1　はじめに

近年，深層学習を用いた画像の説明文（キャプション）生成に関する研究が盛んに行われている [1]．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 [2][3]．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　シーングラフ

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] などの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, 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 (on, 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で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．

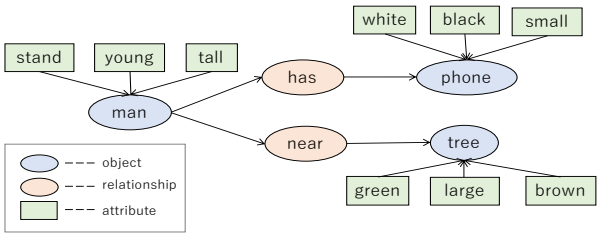


図 1: シーングラフの例

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生成される画像のキャプションへの影響を検証する．

3　シーングラフを用いたキャプション生成

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．まず，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．これらの特徴量を連結し 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 [2] 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与えられているデータセット Visual Genome[6] 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．また，Yang らの研究で用いられている spatial Graph Convolutional Networks(GCNs)[7][8] を参考にして，シーングラフをグラフ特徴量に変換した．

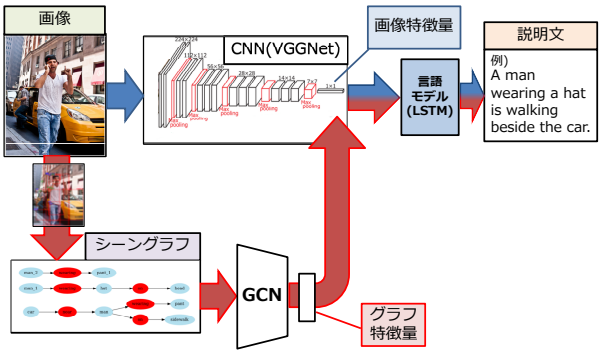


図 2: 本研究の概要図

4　実験

シーングラフから抽出された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る．

4.1　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のツールである chainer[1] を用いて実験を行った．データセットは，画像とそのキャプションのペアのデータセットである Microsoft COCO を使用し，1 画像あたり 5, 6 文（英文）キャプションが付与されている．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．ハイパーパラメータの設定を表 1 に示す．

4.2　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．以下に，object, attribute 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．グラフ特徴量も用いたことで改善された箇所を**太字**で表記する．

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

表 1: 実験設定

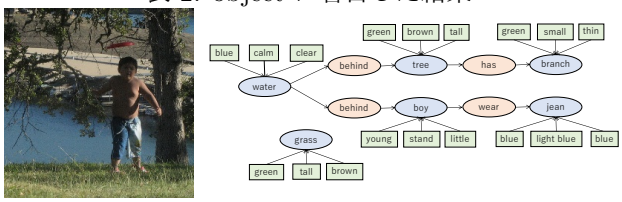
画像→画像特徴量→説明文モデル	
データセット	Microsoft COCO
学習量	82,783 画像 ×100 epoch
勾配法	Adam
語彙	頻出語 3,469 語
損失関数	交差エントロピー

また，シーングラフは一部抜粋したものを示す．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：悪化 = 5:4:1)．

4.2.1 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．

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**.

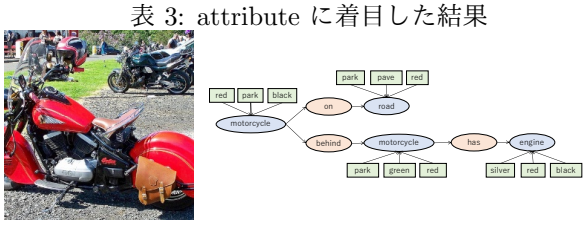
正解文
Chubby little kid is making a face near a lake
a child in a field with a **frisbee** near a tree
a child is standing outside in the **grass**
A young boy is throwing a **frisbee** on the **grass** by a lake and by a tree.
A boy is playing with a **frisbee** under the tree next to the water.

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，シーングラフに含まれていないものの，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” の要素が足されることで，画像特徴量から “frisbee” をキャプションに出現させる際の補助となっていることが推測される．

4.2.2 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 “park” 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．

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.

4.3 考察

主観による定性評価では，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，4.2.1 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効果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．

参考文献

[1] Raimonda Staniute and Dmitrij Sesok. A systematic Literature Review on Image Captioning. 2019.

[2] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-Encoding Scene Graphs for Image Captioning. In *Proc. of CVPR*, 2019.

[3] Dalin Wang, Daniel Beck, and Trevor Cohn. On the Role of Scene Graphs in Image Captioning. In *EMNLP*, 2019.

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

[5] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and Tell: A Neural Image Caption Generator. In *Proc. of CVPR*, 2015.

[6] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, Vol. 123, pp. 32–73, 2017.

[7] Yujia Li, Richard Zemel, Marc Brockschmidt, and Daniel Tarlow. Gated Graph Sequence Neural Networks. In *Proc. of ICLR*, 2017.

[8] Diego Marcheggiani and Ivan Titov. Encoding Sentences with Graph Convolutional Networks for Semantic Role Labeling. In *EMNLP*, 2017.