画像キャプショニングのための制約語の抽出法

芳賀あかり [1]　平尾努 [2]　帖佐克己 [2]
本多右京 [1] 出口祥之 [1] 渡辺太郎 [1]
[1] 奈良先端科学技術大学院大学　[2]NTT コミュニケーション科学基礎研究所
{haga.akari.ha0,honda.ukyo.hn6,deguchi.hiroyuki.db0,taro}@is.naist.jp
{tsutomu.hirao.kp,katsuki.chousa.bg}@hco.ntt.co.jp

概要

従来の画像キャプショニングには画像とは無関係の語を含むキャプションをしばしば生成するという問題がある．これを解決するため，画像に関連する語をあらかじめ与えた上でキャプションを生成する手法が提案されているが，その自動決定法については議論がされていない．本研究では，物体検出器が出力するラベル (物体名) をその信頼度スコアと顕著性スコアを組み合わせてランキングすることでキャプションに含めるべき語を決定する手法を提案する．提案法で得た単語と人手生成の正解キャプション中の単語を比較した結果，自動抽出した単語のうち半数程度はキャプションに含まれていた．さらに人手評価を行ったところ，キャプションには含まれない単語であっても，その多くはキャプション生成が可能な程度に画像に関連した語であることがわかった．

1　はじめに

画像キャプショニングとは，画像を説明する一文を生成するタスクであり，Vision and Language 分野における主要なタスクの一つである．近年では，Vision transformer による系列変換モデルで実現されることが多くなってきたが，Transformer に基づく系列変換モデルには，流暢な文を生成する一方，入力とは関連のない文をしばしば生成するという問題がある [1]．画像キャプショニングも例外では無く，画像には写っていないものや画像とは関連ない語を含むキャプションを生成する Object hallucination という問題が生じることが知られている [2]．この問題を抑制するための一つの方法として語彙に制約を与えてキャプションを生成する手法が提案されている [3]．しかし，この手法ではキャプションに含めるべき語 (以下，制約語) をどのように決定するか議論していない．もちろん，人手で与えることも可能ではあるが大量の画像を処理する場合にはコストの観点から現実的でない．

そこで，本研究では制約語を画像から自動的に抽出する方法を提案する．人間は画像の構図，視覚的顕著性などに基づきキャプションに含める語を決定していると考えられることから，画像に対して物体検出器が出力するラベルをその信頼度スコアと顕著性スコアを組み合わせることでランキングし，その上位 n 件を制約語とする手法を提案する．

MSCOCO キャプションの検証セットのキャプションに提案法で抽出した制約語が含まれる割合，制約語の平均正解率は 49.9%であった．ただし，人手評価を行ったところ，正解率の低い画像でもキャプションに含めてよい語は 50%を超え，正解率が高い画像ではそれが約 80%に達した．

2　準備

物体検出とは画像に写る物体を矩形 (以下，バウンディングボックス) で特定し，その名称をラベルとして出力する．ラベルはあらかじめ定められているので，その決定は多値分類問題に帰結する．

物体検出器により決定できるラベル数は大きく異なる．たとえば，Detic [4] は CLIP [5] を使用することにより，学習データにはないラベルを検出可能としており，21,000 ラベルまでの異なる粒度で物体検出を行うことができるという特徴がある．Detic は COCO，Objects365，OpenImages，Lvis の 4 つのカスタムラベルセットを提供しており，容易に異なるラベルセットで物体検出を行うことができる．各ラベルセットはそれぞれ 80，365，500，1203 カテゴリである．使用するラベルセットにより検出する物体の粒度が異なり，カテゴリ数が少ないほどカテゴリの抽象度が高くなり，カテゴリ数が多いほど具体的になる．例えばカテゴリ数 80 の COCO ラベルセッ

トで行った物体検出では「bird」ラベルを与える物体を，カテゴリ数 365 の objects365 ラベルセットでは「goose」や「duck」と，より具体的なラベルを与える．

一方，画像には人間の視覚注意により注視を集めやすい領域とそうでない領域がある．これを画像の各ピクセルに対してスコアを与えることで表現したものを顕著性マップと呼び，画像中の顕著な場所，すなわち目立つ領域の特定に活かすことができる [6]．Hou らが提案した Spectral Residual アプローチ [7] は，画像データを周波数に変換して分析を行うことで顕著性マップを作成する．この手法は特徴量やカテゴリなどの物体に関する事前知識を利用せず，高速かつ頑健に顕著性を検出することができる．また，Montabone らは解像度を下げない，きめ細かい顕著性マップを作成する方法を提案しており，このマップは明確な境界線を抽出するという特徴がある [8]．本研究では Spectral Residual アプローチを用いて顕著性マップを作成する．

人間が画像に対しキャプションを生成する際，多くの場合は画像に写っている物体に言及するであろうし，複数の物体が写っているのであればどこに焦点をあてるべきかの取捨選択も行うであろう．たとえば，図 1 には様々な物体が写っているが，人間がキャプションを生成する際には「cat」，「glove」，「ball」といった物体に言及するだろう．実際，この画像に対するキャプションには「cat」，「baseball glove」，「ball」という語が利用されている．一方，フォーカスされていない右後方の机や左後方のサイドテーブルには言及していない．つまり，人間は画像中の顕著性の高い物体について言及すると考えるのが自然である．そこで，本稿では物体検出と画像の顕著性マップを組み合わせることで人間がキャプションで言及すると考えられる顕著なラベルを特定し，制約語として用いる．

3 提案手法

本稿では，物体検出器が検知した物体のラベルのスコアをその信頼度スコアと物体のバウンディングボックス内部の顕著性スコアの重み付き線形和で計算し，ランキングすることでより確度の高い制約語を得る．

制約語候補の抽出　画像に対して物体検出器を適用し，それが出力するすべてのバウンディングボックス (b) とラベル (ℓ) の対を得る．この時，バウンディングボックスに対しては割り当てられたラベルに対する信頼度スコア Conf.(b,ℓ) が与えられる．そして，信頼度の低いラベルは削除するため Conf.$(b,\ell) \geq 0.5$ となるラベルのみを制約語候補として利用する．

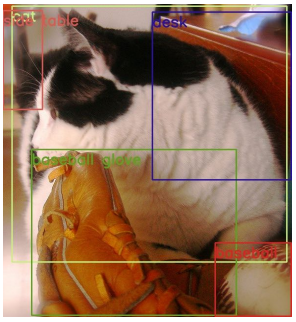

図 1　物体検出例[1)]

制約語候補の顕著性スコアの計算　顕著性スコアは画像中のすべてのピクセルに対して与えられるので，制約語候補の顕著性スコアを計算するにあたっては，対応するバウンディングボックス内部の各ピクセルの顕著性スコアを利用する．本稿では，バウンディングボックス内部における顕著性スコアの最大値を制約語候補の顕著性スコア Sal.(b,ℓ) とし，次の式で計算する．

$$\mathrm{Sal.}(b,\ell) = \max_{r_k \in b} \mathrm{Sc}(r_k) \tag{1}$$

ここで r_k は物体検出器が検出したバウンディングボックス内の $k \times k$ ピクセルのボックスをあらわし，$\mathrm{Sc}(r_k)$ はその顕著性スコアである．バウンディングボックス内の i 行 j 列目の顕著性スコアの値を $x_{i,j}$ とするとき，$\mathrm{Sc}(r_k)$ は次のように計算する．

$$\mathrm{Sc}(r_k) = \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} x_{i,j} \tag{2}$$

顕著性スコアの算出までの流れを図 2 に示す．図 2 の例では制約語候補が bird であり，太枠の長方形が制約語候補 bird に対応するバウンディングボックスを示す．バウンディングボックス内の顕著性スコアの最大値を 1 ピクセルを基本単位として与えるとどのようなバウンディングボックスに対しても非常に高い値をとる可能性がある．これを緩和するため，バウンディングボックス内にそれよりも小さな正方形のボックスを設け内部のピクセルに対する顕著性スコアの和 $\mathrm{Sc}(r_k)$ を計算し，バウンディングボックス内のすべての可能な正方形ボックスの顕著性スコア $\mathrm{Sc}(r_k)$ の最大値を制約語候補の顕著性スコアと

1)　MSCOCO val2014 COCO_val2014_000000518974.jpg

する．図 2 の例では，バウンディングボックス内に $k \times k (k = 3)$ ピクセルの正方形ボックスを設け，各ピクセルの顕著性スコア，$0.7, 0.5, ...0.7$ を合計し，正方形ボックスの顕著性スコアとする．次に，正方形ボックスをバウンディングボックス内部で 1 ピクセルスライドさせ，同じように合計を計算する．これを繰り返して全ての $k \times k$ ボックスのスコアを計算し，この最大値を制約語候補 bird の顕著性スコアとする．

制約語候補のランキング　制約語候補のスコア F を物体検出器の信頼度スコアと顕著性スコアの重み付き線形和として以下の式で定義する．

$$F(b, \ell) = \alpha \mathrm{Sal.}(b, \ell) + (1 - \alpha)\mathrm{Conf.}(b, \ell) \qquad (3)$$

ここで，α $(0 \leq \alpha \leq 1)$ は 2 つの項のバランスをとるためのパラメータである．画像中の各制約語候補に対し上記スコアを計算しその上位 n 件を制約語とみなす．

4　実験

制約語の正解データはないため，提案法で得た制約語が人手で作成された正解のキャプションにどれだけ含まれるかで評価を行った．

4.1　実験設定

データセットは MSCOCO の val2014 を採用し，自動評価には 40500 画像を使用した．物体検出には imagenet データセット [9] を用いて学習されたモデルである Detic[4] を用いた．顕著性マップは Spectral Residual アプローチを実装した OpenCV の SpectralResidual 関数を用いて取得した．パラメータは $k = 20, \alpha = 0.2$ を採用している．

4.2　自動評価

抽出した制約語候補のうち，人手作成のキャプションに含まれる語の割合，正解率を用いて自動評価を行った．なお，Detic のラベルセットは Object365，Open Images，Lvis の 3 種を用いた．ラベル数はそれぞれ，365，500，1203 となる．ラベル数が少ないほど抽象的で多いほど具体的な名称ラベルを与える．

キャプションに制約語が含まれているか否かは語の完全一致で行う．しかし，完全一致で評価した場合，同義語，類義語が無視されてしまうため，類語も正解とする処理も取り入れた．ここでの類語とは異なるラベルセットで物体検出を行った結果，同じバウンディングボックスを指しているラベルを指し，これらも正解として扱い評価を行う．例えば，あるバウンディングボックスの物体を bird と表すラベルセットと，goose と表すラベルセットがあるとする．この場合，bird と goose を類語とする．なお，バウンディングボックスの一致については物体検出で使用される一般的な評価指標である IoU (Intersection over Union) が 0.8 以上である場合に一致とみなした．二つの任意の図形 A，B の IoU は以下のように計算される [10]．

$$IoU = \frac{|A \cap B|}{|A \cup B|} \qquad (4)$$

表 1　カテゴリ数による正解率の変動

ラベルセット	カテゴリ数	正解率	正解率（類語処理あり）
Objects365	365	0.446	0.499
Open Images	500	0.440	0.478
Lvis	1203	0.408	0.490

顕著性スコアの効果　制約語のスコアを計算する際のパラメータ α を動かして，顕著性スコアの効果を調べた．その結果を図 3 に示す．図の縦軸は正解率，横軸はパラメータ α の値であり，$\alpha = 0$ が物体検出の信頼度スコアのみ，$\alpha = 1$ は顕著性スコアのみを利用することを表す．全てのカテゴリ数において $\alpha = 0.2$ 付近で正解率が向上していることから，物体検出の信頼度スコア，顕著性スコアのみではなく，双方を考慮することの有効性がわかる．以下は $\alpha = 0.2$ を採用した結果である．

カテゴリ数による正解率の変動　各ラベルセットでの評価結果を表 1 に示す．表より，完全一致の正解率でみた場合，Lvis を用いた場合が最も低く，Object365 を用いた場合が最も高い．この結果は，MSCOCO データセットのキャプションが比較的抽象的な語を用いて記述されている，つまり，Livs のような具体的なラベルを利用すると語の完全一致では不利になるからであると考える．実際，類語処理を導入すると正解率はどのラベルセットでも 5 割近くになっており，顕著な差はみられなくなることがこれを支持している．

4.3　人手評価

自動評価は類語も含めて評価をしているが，すべての類語を網羅できているとは限らず，不当に低いスコアになっている可能性がある．そこで，自動評価の結果が悪いもの・良いもののうちそれぞれラン

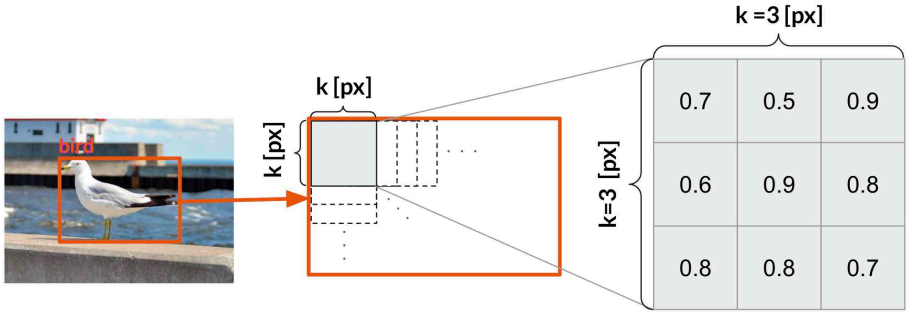

図 2　制約語候補の顕著性スコアの算出[2)]

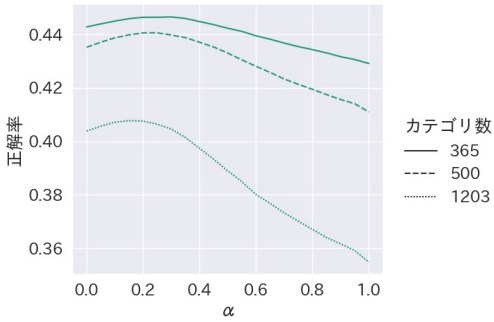

図 3　パラメータ α による正解率の変動

図 4　人手評価に使用した画像の例[3)]

表 2　人手評価結果（正解率 ≤ 0.2）

α	キャプションに含まれて良いラベル	キャプション必須のラベル
0	0.538	0.382
0.2	0.522	0.376
0.5	0.509	0.353
1.0	0.458	0.322

表 3　人手評価結果（正解率 ≥ 0.8）

α	キャプションに含まれて良いラベル	キャプション必須のラベル
0	0.779	0.628
0.2	0.799	0.641
0.5	0.781	0.617
1.0	0.739	0.560

ダムに 25 件サンプリングし，人手評価を行った．自動評価の結果が悪いものとは正解率が 0.2 以下，良いものとは正解率が 0.8 以上のものをさす．評価指標として，キャプションに含まれて良いラベルの割合，またキャプションに必須のラベルの割合を主観で評価した．

例えば図 4 のような画像の場合，surf, surfboard, wet suit, person などの語をキャプションに含まれる可能性があるラベルとみなし，surf や person のように中心的な語を必須のラベルとみなした．自動評価の結果が悪いものの結果を表 2，良いものの結果を表 3 に示す．表 2 より，自動評価の結果が悪いのものでも，$\alpha = 0.2$ ではキャプションに含まれてもいいだろうというラベルは 5 割以上あることがわかる．$\alpha = 0$ のとき，差は小さいもののもっとも良い結果となった理由は，物体を多く検知するような場合，顕著性スコアの高い領域と人間がキャプションを生成する際に着目する領域にずれが生じるからではないかと考える．さらに，表 3 より，自動評価の結果が良いものでは，キャプションに含まれてもいいだろうというラベルは 8 割近くあり，キャプション必須のラベルは 6 割以上含まれている．ここで，$\alpha = 0$ は顕著性スコアのみ，$\alpha = 1$ は物体検出の信頼度スコアのみとなっており，自動評価の結果が良いものでは $\alpha = 0.2$ で最も高い割合となったことから，両者を用いることの有効性が手動評価でも明らかとなった．

5　おわりに

本論文では，物体検出と画像の顕著性を示す顕著性マップを用いて制約語を自動的に抽出・ランキングする手法を提案した．さらにランキング上位の制約語がキャプションにどの程度含まれているか評価を行った．その結果，自動評価と人手評価双方により物体検出スコアと顕著性スコア双方を考慮することの有効性を示した．本手法で自動抽出した制約語を実際に画像キャプショニングへ組み込むことが今後の課題である．

2）　MSCOCO val2014 COCO_val2014_000000324670.jpg
3）　MSCOCO val2014 COCO_val2014_000000374010.jpg

参考文献

[1] Guangxiang Zhao, Junyang Lin, Zhiyuan Zhang, Xuancheng Ren, Qi Su, and Xu Sun. Explicit sparse transformer: Concentrated attention through explicit selection. **CoRR**, Vol. abs/1912.11637, , 2019.

[2] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In **EMNLP**, 2018.

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. In **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**, pp. 936–945, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[4] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In **ECCV**, 2022.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In **International Conference on Machine Learning**, pp. 8748–8763. PMLR, 2021.

[6] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. **IEEE Transactions on pattern analysis and machine intelligence**, Vol. 20, No. 11, pp. 1254–1259, 1998.

[7] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In **2007 IEEE Conference on Computer Vision and Pattern Recognition**, pp. 1–8, 2007.

[8] Sebastian Montabone and Alvaro Soto. Human detection using a mobile platform and novel features derived from a visual saliency mechanism. **Image and Vision Computing**, Vol. 28, No. 3, pp. 391–402, 2010.

[9] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In **Conference on Fairness, Accountability, and Transparency**, 2020.

[10] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, pp. 658–666, 2019.