物体構成要素間の階層的な注意を用いた画像キャプション生成

平川 幸佑, 田村 晃裕, 加藤 恒夫
同志社大学 大学院理工学研究科
{ctwf0106@mail4, aktamura@mail, tsukato@mail}.doshisha.ac.jp

概要

画像キャプション生成タスクでは，Faster R-CNN のような物体検出器により抽出した物体の特徴ベクトルを用いるモデルが高い精度を達成している．従来モデルでは Transformer Encoder の自己注意機構において物体間の関連を捉えられる．しかし，物体間の関連は学習の中で自動的に学習されるため，位置が近い物体同士が纏まりやすいという傾向は陽にモデルに取り入れられていない．そこで本研究では，Transformer Encoder において隣接する物体間の関連を階層的に捉える自己注意機構「構成要素注意機構」を提案する．

1 はじめに

近年，ニューラルネットワーク (NN) に基づいた画像キャプション生成の研究が盛んに行われている．NN に基づくキャプション生成は Encoder-Decoder モデル [1] が主流であり，CNN を用いて畳み込んだ画像の各領域への注意を学習する視覚的な注意機構に基づくモデルが高い精度を実現している [2]．また，Faster R-CNN[3] などの物体検出器により抽出された画像上の物体や顕著な領域を用いることで，物体の位置やサイズを考慮したキャプション生成モデルが提案されている [4]．この従来モデルでは Encoder の自己注意機構において物体間の関連を捉えられる．しかし，物体間の関連は学習の中で自動的に学習されるため，位置が近い物体同士が纏まりやすいという傾向は陽にモデルに取り入れられていない．

そこで本研究では，Transformer に基づく画像キャプション生成モデルの Encoder において，隣接する 2 つの物体構成要素間の纏まりを捉えながら物体間の関連を階層的に捉える自己注意機構「構成要素注意機構」を提案する．構成要素注意機構により隣接する物体構成要素同士を階層的に纏めていくことで，位置が近い物体同士を関連づけられるため，キャプション生成の精度向上が期待できる．

MSCOCO 2014 captions データセット [5] を用いた画像キャプション生成タスクによる評価実験の結果，Transformer をベースとした古典的な画像キャプション生成モデルに提案の構成要素注意機構を導入することで，有意な性能改善を確認できた．一方で，Object-Relation-Transformer[4] に構成要素注意機構を導入した場合，評価指標のスコアはわずかに向上したが，有意差は確認できなかった．

2 関連研究

2.1 節で Transformer について説明する．2.2 節では，本研究でベースモデルとして利用する，物体検出により抽出した物体間の関連を Encoder で捉える Object-Relation-Transformer について説明する．

2.1 Transformer

Transformer[6] は，入力系列を潜在表現へと変換し，変換した潜在表現から出力系列を生成する Encoder-Decoder モデルである．Encoder と Decoder はそれぞれ Encoder レイヤと Decoder レイヤを複数個スタックすることで構成される．Encoder レイヤは自己注意層と全結合層の 2 つのサブレイヤで構成され，Decoder レイヤは自己注意機構と全結合層，ソース・ターゲット注意機構の 3 つのサブレイヤで構成されている．各サブレイヤ間には，残差接続と層正規化が適用される．自己注意機構は H 個の同一のヘッド (マルチヘッド) で構成され，各ヘッドの計算には式 (1) の縮小付き内積注意が用いられる．

$$\mathrm{head}_i(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

ここで，Q，K，V はそれぞれ $query$，key，$value$ に対応しており，$d_k = d_{model}/h$ である．d_{model} は元々の入力の埋め込み次元である．$query$ と key の内積により算出した各要素の類似度を softmax により確率化することで，要素間の関係の強さを表すスコア

を算出する．このスコアと *value* との内積を算出することで，*query* の各要素と関係の強い *value* の各要素の重み付き加重和による特徴が抽出できる．式 (1) は各ヘッドで独立に計算され，式 (2) のように H 個のヘッドの出力を 1 つのベクトルに連結し，重み行列 W^O と乗算される．

$$\mathrm{MultiHead}(Q,K,V) = \mathrm{Concat}(\mathrm{head}_1,\ldots,\mathrm{head}_H)W^O \tag{2}$$

また，全結合層は式 (3) により行われる．

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{3}$$

ここで，W_1，b_1 と W_2，b_2 は全結合層の重みとバイアスである．

Transformer は再帰や畳み込みを用いていないため，入力系列や出力系列の各要素の位置情報を捉えることができない．そこで Encoder と Decoder では，埋め込み層の後で sin 関数と cos 関数を用いた位置エンコーディングを行い，各要素の位置情報を付加する．ただし，物体検出ベースの画像キャプション生成 Transformer の場合，入力となる特徴ベクトルの並びは画像における位置に基づくものではない．そのため，エンコーダでは位置エンコーディングを用いた埋め込み処理を行っていない．

2.2 Object-Relation-Transformer

Object-Relation-Transformer[4] は 2.1 節で紹介した Transformer Encoder の注意機構で，物体間の相対的な位置やサイズの比率を組み込んだ幾何学的注意を計算する．まず，画像中の 2 つの物体 m と n の中心座標 $(x_{m/n}, y_{m/n})$，幅 $(w_{m/n})$，高さ $(h_{m/n})$ から変位ベクトル $\lambda(m,n)$ を式 (4) により算出する．

$$\lambda(m,n) = \left(\log\left(\frac{d_x}{w_m}\right), \log\left(\frac{d_y}{h_m}\right), \log\left(\frac{w_n}{w_m}\right), \log\left(\frac{h_n}{h_m}\right)\right) \tag{4}$$

ここで，$d_x = |x_m - x_n|$，$d_y = |y_m - y_n|$ である．そして，幾何学的注意重みを式 (5) により算出する．

$$\omega_G^{mn} = \mathrm{ReLU}(\mathbf{Emb}(\lambda)W_G) \tag{5}$$

ここで，**Emb** は，Transformer の sin 関数と cos 関数による位置エンコーディングを行う関数である．位置エンコーディングの結果は重みベクトル W_G と乗算することでスカラー値に変換し，非線形関数 ReLU を適用する．その後，幾何学的注意重み ω_G^{mn} は，式 (6) によりオリジナルの注意機構に組み込む．

$$\omega^{mn} = \frac{\omega_G^{mn}\exp(\omega_A^{mn})}{\sum_{l=1}^{N_I}\omega_G^{ml}\exp(\omega_A^{ml})} \tag{6}$$

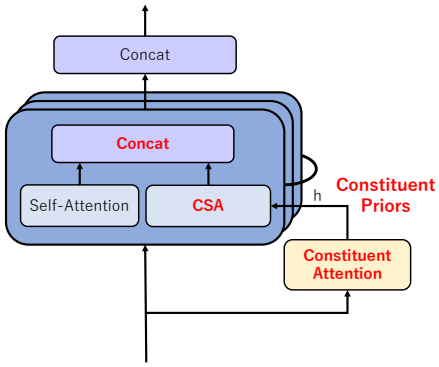

図 1　構成要素注意機構を導入した自己注意機構

ここで，ω_A^{mn} は *query* と *key* の縮小付き内積により算出された行列の要素である．また，N_I は画像中の物体の個数である．最終的なヘッドの出力は式 (7) になる．

$$\mathrm{head}_i(Q,K,V) = \Omega V \tag{7}$$

ここで，Ω は ω^{mn} を要素に持つ $N_I \times N_I$ 行列である．

3 提案手法

従来の物体検出ベースの Transformer モデルは，Encoder の自己注意機構で物体間の関連を捉える．しかし，その関連は学習の中で自動的に学習されるため，近い位置の物体同士が纏まりやすい傾向は陽にモデルに取り入れられていない．そこで本研究では，Transformer Encoder で隣接する物体間の関連を階層的に捉える構成要素注意機構を提案する．

構成要素注意機構を組み込んだ Transformer Encoder の自己注意機構の概略図を図 1 に示す．図 1 において，赤い部分が提案の構成要素注意機構に関する部分である．以降，3.1 節で構成要素注意機構について説明し，3.2 節で構成要素注意機構を画像キャプション生成 Transformer モデルへ組み込む方法を説明する．

3.1 構成要素注意機構

構成要素注意機構は，Transformer の自己注意機構において文の句構造を捉えるために提案された Constituent Attention[7] を画像キャプション生成モデルに応用したものである．構成要素注意機構では，隣接する物体間の関連を表す隣接注意スコアに基づき，特定範囲の物体が同じ構成要素に属する確率を表す構成要素事前確率を階層的に求める．そして，その構成要素事前確率で自己注意機構を条件付けることで，自己注意機構において隣接する要素間をより関連付けることが可能になる．

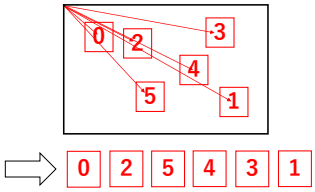

図 2　物体集合の系列化の例

提案手法では，まず，物体検出器により検出された物体の集合を物体系列に変換する．この際，画像中で隣接する物体は物体系列においても隣接するように系列化する．具体的には，図 2 のように，各物体の中心座標に基づき，原点から距離が近い順に物体を並べる．この物体系列を基に構成要素事前確率を計算することで，隣接する物体同士を同じ構成要素として関連づけることが可能になる．

構成要素注意機構は物体系列に対して，まず，物体 o_i が右隣りの物体 o_{i+1} と左隣りの物体 o_{i-1} に結びつく確率を式 (8) のように算出する．ただし，$s_{i,i+1}$ と $s_{i,i-1}$ は，それぞれ，o_i が o_{i+1} と $o_{i,i-1}$ に結びつく度合いを表すスコアであり，式 (9) の通り，縮小付き内積注意により算出される．

$$p_{i,i+1}, p_{i,i-1} = \mathrm{softmax}(s_{i,i+1}, s_{i,i-1}) \tag{8}$$

$$s_{i,i+1} = \frac{q_i \cdot k_{i+1}}{d},\ s_{i,i-1} = \frac{q_i \cdot k_{i-1}}{d} \tag{9}$$

ここで，q_i は o_i の $query$ であり，k_{i-1} と k_{i+1} は，それぞれ，o_{i-1} と o_{i+1} の key である．また，$d = d_{model}/2$ である．

その後，o_i と o_{i+1} が結びつく度合いを表す隣接注意スコアを，式 (10) の通り，$p_{i,i+1}$ と $p_{i+1,i}$ の 2 つの確率の幾何平均により求める．

$$\hat{a}_{i,i+1} = \sqrt{p_{i,i+1} \times p_{i+1,i}} \tag{10}$$

そして，この隣接注意スコアを，式 (11) のように階層的に取り込んだ構成要素注意スコアを求める．構成要素注意スコア $a_{i,i+1}$ は物体 o_i と隣接物体 o_{i+1} が同じ構成要素に属する度合いを表している．

$$a^l_{i,i+1} = a^{l-1}_{i,i+1} + \left(1 - a^{l-1}_{i,i+1}\right) \hat{a}^l_{i,i+1} \tag{11}$$

ここで，a^l は l 層目の構成要素注意スコア，$\hat{a}^l$ は l 層目の隣接注意スコアである．初期状態では各物体は異なる構成要素とみなし，a^0=0 とする．

構成要素注意スコア $a_{i,i+1}$ を算出後，算出した $a_{i,i+1}$ をもとに，式 (12) のように構成要素事前確率 $C_{i,j}$ を算出する．$C_{i,j}$ は物体 o_i から o_j までが同じ構成要素である確率を表す．

$$C_{i,j} = e^{\sum_{k=i}^{j-1} \log\left(a_{k,k+1}\right)} \tag{12}$$

この構成要素事前確率を用いて，式 (13) のように自己注意を条件づけることで，隣接する関連の高い物体をより注目させる．

$$E = C \odot \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{13}$$

ここで，C は $C_{i,j}$ を要素とする行列である．

3.2　構成要素注意機構の導入

構成要素注意機構（式 (13)）のみを用いると，隣接しないが関連度の高い物体が注目されない問題が生じ得る．例えば，図 2 において，物体 “0” と物体 “1” が関連ある物体であっても，構成要素注意機構では関連強いと判定できずに正しいキャプションを生成できない可能性がある．そこで提案モデルでは，図 1 のように，Transformer Encoder の各ヘッドで通常の自己注意機構と構成要素注意機構を結合し，2 つの自己注意を考慮する．各ヘッドでの演算と最終的なマルチヘッドの出力はそれぞれ式 (14)，式 (15) の通りである．式 (14) の head_i は式 (1) や式 (7) で計算されるものである．

$$\hat{\mathrm{head}}_i(Q,K,V) = \mathrm{Concat}(\mathrm{head}_i(Q,K,V), EV) \tag{14}$$

$$\mathrm{MultiHead}(Q,K,V) = \mathrm{Concat}(\hat{\mathrm{head}}_1, \ldots, \hat{\mathrm{head}}_H)W^O \tag{15}$$

4　実験

4.1　実験設定

本研究では，MSCOCO 2014 captions データセット [5] を用いて，提案手法の有効性を検証した．MSCOCO データセットは，82,783 件の訓練用画像と 40,504 件の開発用画像，40,775 件のテスト画像からなる．各画像には最低 5 つのキャプションが付けられている．本実験では，Karpathy ら [8] に倣い，開発用画像のうち 5,000 件をモデル開発用，5, 000 件をテスト用，残りを訓練用画像に足し，113,287 件を学習に用いた．また，入力に用いる特徴マップは，Herdade ら [4] に倣い，Anderson ら [9] が公開している事前学習済みの bottom-up attention model により抽出した MS COCO 画像の領域特徴群[1)] を採用した．Transformer のパラメータ設定は，Vaswani ら [6] に倣い，Encoder，Decoder レイヤは 6 個スタックし，ヘッド数は 8，d_{model} は 512 とした．また，モデルの最適化手法は Adam，ウォームアップステップを

1)　https://github.com/peteanderson80/bottom-up-attention

表 1　各手法による評価比較

評価手法		B@1	B@4	M	C
T	平均値	75.7	34.6	27.7	112.8
	最大値	75.8	35.0	27.8	113.2
T+CSA	平均値	76.1	35.3	27.9	114.4
	最大値	**76.2**	**35.4***	**27.9**	**114.9***
ORT	平均値	76.3	35.1	27.9	114.4
	最大値	76.6	35.5	28.0	115.4
ORT+CSA	平均値	76.3	35.4	28.0	114.8
	最大値	**76.7**	**35.8**	**28.1**	**116.0**

20,000，バッチサイズを 15 とし，クロスエントロピー誤差を損失関数として 30 エポック学習を行った．キャプション文はビームサイズ 2 のビームサーチで生成した．

提案の構成要素注意機構（CSA）の効果を確認するため，通常の画像キャプション生成 Transformer モデル（T），Object-Relation-Transformer（ORT），それぞれのモデルに構成要素注意機構を導入したモデル（T+CSA，ORT+CSA）の性能を比較した．各モデルで 3 回ずつ学習した後，テストデータに対する性能の平均値と最大値で比較した．評価指標は，BLEU-n[10] と METEOR[11]，CIDEr-D[12] を用いた．

4.2　実験結果

それぞれのモデルの性能を表 1 に示す．なお，「*」はベースラインモデル（T ／ ORT）との差が，対応ありの両側 t 検定で有意（有意水準 5%）であったことを表す．Transformer と構成要素注意機構を導入した Transformer を比較すると，Bleu4 スコアにおいては，平均値で 0.7 ポイント，最大値で 0.4 ポイントの向上が確認できた．CIDEr-D スコアにおいても，平均値で 1.6 ポイント，最大値で 1.7 ポイントの向上し，有意な改善を確認できた．

Object-Relation-Transformer と構成要素注意機構を導入した Object-Relation-Transformer を比較すると，Bleu4 スコアにおいては，平均値，最大値共に 0.3 ポイントの向上が確認できた．CIDEr-D スコアにおいても，平均で 0.4 ポイント，最大値で 0.6 ポイントの向上が確認できた．しかし，これらの結果に対して対応ありの両側 t 検定を行った結果，有意水準 5% でいずれの指標においても有意差を確認できなかった．

4.3　考察

提案の構成要素注意機構で算出した自己注意の例を図 3，4 に示す．各画像において，枠内の画像が

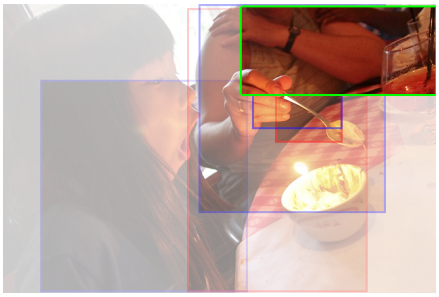

図 3　構成要素注意機構で算出した注意の例 (1 層目)

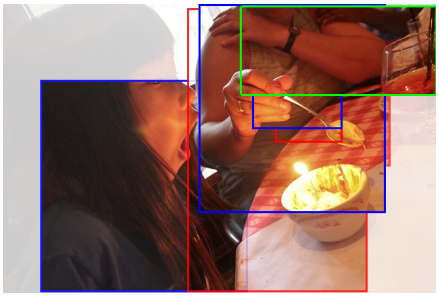

図 4　構成要素注意機構で算出した注意の例 (6 層目)

鮮明なほど，その画像に対する注意が高くなることを示す．ここで，一番右上の緑枠は注目する物体，赤枠は緑枠より原点から遠い物体，青枠は緑枠より原点に近い物体であり，緑枠に隣接する両隣 3 個の物体を表している．

これらの図より，右上の緑枠の腕に関して，1 層目では，周辺のスプーンや持ち手に注意が向き，最上層では，左の女性も 1 つの要素として纏めていることが確認できる．このことから，構成要素注意機構は，周辺の物体を纏めながら階層的に注意を計算できたことが分かる．

5　おわりに

本研究では，Faster-R-CNN のような物体検出器により獲得した物体や顕著な領域の特徴ベクトルを用いた画像キャプション生成 Transformer モデルに対して，画像上で隣接する物体を物体構成要素の纏まりとして階層的にとらえる自己注意機構「構成要素注意機構」を提案した．MSCOCO データを用いた実験結果より，Transformer に構成要素注意機構を導入することで，統計的に有意にキャプション生成性能を改善できることが確認できた．一方，Object-Relation-Transformer に構成要素注意機構を導入した場合も，いくつかの評価スコアが向上したが，検定の結果，有意ではなかった．今後は，構成要素注意機構で行う物体群の系列化の方法を変更するなどして，提案モデルの性能を改善させたい．

参考文献

[1] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.

[2] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048–2057. PMLR, 2015.

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, Vol. 28, pp. 91–99, 2015.

[4] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. In *Advances in neural information processing systems*, 2019.

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[7] Yaushian Wang, Hung-Yi Lee, and Yun-Nung Chen. Tree transformer: Integrating tree structures into self-attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1061–1070, November 2019.

[8] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

[9] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6077–6086, 2018.

[10] 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, 2002.

[11] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pp. 376–380, 2014.

[12] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.