物体間の関係性を考慮したTransformerに基づくキャプション生成

中本 裕大　　田村 晃裕　　二宮 崇
愛媛大学 大学院理工学研究科 電子情報工学専攻
{nakamoto@ai., tamura@, ninomiya@}cs.ehime-u.ac.jp

1 はじめに

近年, 自然言語処理分野の様々なタスクにおいてニューラルネットワーク (以降，NN) に基づいた研究が盛んに行われている．NN に基づく手法は，画像分類や物体検出 [1] などの自然言語処理以外の分野においても広く利用され，各領域において高い精度を実現している．

入力画像に対する説明文を自動生成する画像キャプション生成においても，NN に基づいた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モデルは，エンコーダ･デコーダモデル [2] が主流であり，特に，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注意機構に基づくモデル [3] が高い精度を実現することで知られる．従来の注意機構モデルでは，CNN から抽出した特徴マップを利用しており，物体の大きさなどに関わらず，等しい大きさと形状のグリッドに対応する特徴を用いるため, 画像上の物体や顕著な領域への重みづけが考慮されていない．そこで，Faster R-CNN[1] などの物体検出手法を特徴抽出に用いることで，オブジェクトとして注目すべき領域群を特徴量とし，視覚的注意機構に利用する手法 [4] が提案されている．しかし，これらの視覚的注意機構では，画像内の物体や物体の位置については考慮されているが，物体間の関係性の情報に関しては考慮されていないという問題点がある．

近年，キャプション生成モデルと同様にエンコーダ･デコーダモデルを扱う機械翻訳タスクにおいて，自己注意機構 (Self-Attention) を用いた Transformer モデル [5] が RNN や CNN に基づくモデルと比較し，高い精度を実現したことで注目を浴びている．自己注意機構は Transformer モデルのエンコーダとデコーダにそれぞれ用いられ，入力文中の単語間の関係性を考慮した潜在表現の獲得を可能とする．画像キャプション生成においても，エンコーダに CNN，デコーダに Transformer デコーダを使用したモデル [6] が提案され，いくつかの SOTA の手法と比較して同程度の精度を達成したことが報告されている．

本研究は，画像内の各物体の関係性を考慮する視覚的注意機構モデルを提案する．具体的には，提案モデルでは，入力画像全体を CNN で畳み込むことで獲得した特徴マップに加え，[4] の手法で用いられた bottom-up attention model により獲得した領域群による特徴マップを同時に使用する．これらの特徴マップを自己注意機構を用いて潜在表現へエンコードすることで，物体間の関係性を考慮した画像表現を獲得する．キャプション生成時に，獲得した画像表現に対して視覚的注意機構を適用することで，顕著なオブジェクト間の関係性を考慮した文の生成を実現する．

MSCOCO image captioning 2015 challenge データセットで提案モデルの評価を行い，キャプション評価の各指標において，大幅な精度向上の確認ができた．

2 関連研究

本節では，2.1 節で Transformer モデルについて説明する．2.2 節では，ベースラインである Transformer モデルを利用したキャプション生成モデルを説明する．最後に 2.3 節で，物体検出による画像の領域抽出手法に基づいた，特徴量抽出の手法 (bottom-up attention model) について説明する．

2.1 Transformer モデル [5]

Transformer モデルは，エンコーダとデコーダで構成されており，エンコーダを用いて原言語 $X = (x_1, ..., x_n)$ を潜在表現 $Z = (z_1, ..., z_n)$ へと変換する．デコーダは Z を受け取り，目的言語 $Y = (y_1, ..., y_m)$ を生成するネットワークを学習する．エンコーダとデ

コーダはそれぞれエンコーダレイヤとデコーダレイヤを N 個スタックすることで構成される．エンコーダレイヤは，自己注意機構と単語位置ごとの全結合層 (Feed-Forward Network) の 2 つのサブレイヤから構成される．一方，デコーダは自己注意機構と単語位置ごとの全結合層，ソース・ターゲット注意機構の 3 つのサブレイヤから構成される．各サブレイヤ間には，残差接続と層正規化が適用される．

自己注意機構とソース・ターゲット注意機構は，複数ヘッドの注意により実現され，注意機構の計算には縮小付き内積注意が用いられる．縮小付き内積注意の計算は式 (1) により算出される．

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}}) \cdot V \quad (1)$$

Q，K，V はそれぞれ $query$，key，$value$ に対応する．$query$ と key の内積計算により各要素の類似度を算出し softmax により確率値にすることで，要素間の関係の強さを算出する．算出したスコアと $value$ 要素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．自己注意機構では，エンコーダ，デコーダそれぞれにおいて自身の隠れ層の出力を $query$，key，$value$ として用いるため同一文内の単語間関係の強さを計算することができる．一方，デコーダのソース・ターゲット注意機構では，$query$ はデコーダの隠れ層の出力，key と $value$ はエンコーダの最終出力を用いるため，原言語の各単語表現と関係が強いと判断された目的言語の単語表現による重み付き加重和を計算することができる．

単語位置ごとの全結合層の計算は次式 (2) により行われる．

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \quad (2)$$

ここで，$FFN(x)$ への入力次元は d_{model}，中間層の次元は d_{ff} となる．また，$W_1 \in R^{d_{model} \times d_{ff}}$，$W_2 \in R^{d_{ff} \times d_{model}}$，$b_1 \in R^{d_{ff} \times 1}$，$b_2 \in R^{d_{model} \times 1}$ である．

Transformer は，RNN や CNN のように時系列データの順序を考慮していないため，系列情報を付与する必要がある．そこで，Transformer モデルでは，入力文の単語埋め込み行列に対し位置エンコーディングの行列 PE を各要素に加算することで系列情報を付与する．位置エンコーディングの各成分は異なる周波数の sin 関数と cos 関数を用いて次式 (3),(4) により算出される．

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}}) \quad (3)$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{model}}) \quad (4)$$

ここで，pos は単語の位置，i は各成分の次元を指す．

2.2 Captioning Transformer[6]

Captioning Transformer は，エンコーダに CNN，デコーダに Trasnsformer のデコーダを用いるキャプション生成モデルである．CNN を用いて入力画像 I を特徴マップへとエンコードする．高品質な空間的情報を獲得するため，畳み込み層群の最終レイヤーの後に適応プーリングレイヤー，全結合層，ReLU 関数を適用し，$V = \{v_1, ..., v_{k \times k}\}, v_i \in R^{d_{model}}$ の特徴マップを得る．ここで，$k \times k$ は畳み込み後の画像領域の数であり，v_i は画像の各領域表現に対応する．デコーダは，2.1 節で示したデコーダと同一のものを使用する．特徴マップ V をデコーダのソース・ターゲット注意機構への入力とすることで，キャプション生成時に画像領域に着目するモデルとなる．

2.3 Faster R-CNN に基づく特徴抽出 [4]

画像内のオブジェクトや顕著な領域を特徴マップとして抽出する手法に Anderson ら [4] が提案した bottom-up attention model(図 1) がある．この手法は，画像内の物体の位置やクラスを予測する物体検出タスクに用いられる Faster R-CNN を拡張したモデルである．Faster R-CNN による物体検出は 2 つのステップに区別される．1 つ目は，領域提案ネットワーク (RPN) である．CNN により畳み込まれた画像に対して RPN をスライドさせ，特徴マップの各要素位置に対して，複数の形状パターンを用いて物体であるか否かの判断を行う．ここで，物体であれば畳み込み前の画像上の位置を予測する．予測された領域の重なり具合を示す IOU 値を用いて，閾値処理を行い同一の物体を示す予測領域を 1 つに絞る．2 つ目は RoI pooling により各予測領域に対する特徴量を抽出する．各物体を示す領域とそれに付与されたクラス識別確率が信頼値以上の場合，各領域の特徴量を CNN へと入力し特徴 v_i を得る．

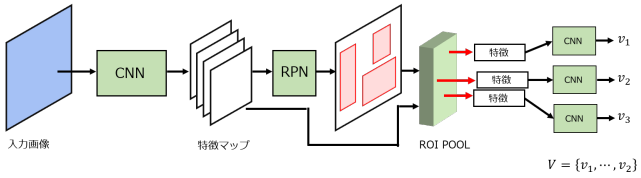

図 1: Faster R-CNN に基づく特徴量抽出の例

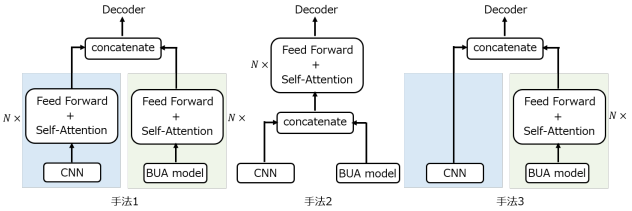

図 2: 提案手法の全体図

3 提案手法

近年，エンコーダ·デコーダモデルを用いる機械翻訳タスクにおいて，自己注意機構を用いた Transformer モデルが高い翻訳精度を実現している．また，キャプション生成タスクにおいては，物体検出アルゴリズムを利用した bottom-up attention (BUA) model により獲得した特徴マップを視覚的注意機構に利用することで，画像内のオブジェクトと顕著な領域に対しての注意を算出し，高い精度を達成している．本研究では，画像全体を畳み込む従来の CNN と画像内のオブジェクトや顕著な領域に着目した bottom-up attention model の特徴を同時に使用し，Transformer の自己注意機構に取り込むことで各物体間の関係性をより考慮した潜在表現を獲得し，獲得した表現に着目してキャプションを生成するモデルを提案する．エンコーダの構造として以下の 3 つのパターン (図 2) を提案する．

1. Transformer のエンコーダに bottom-up attention model から抽出した特徴を入力．もう一つエンコーダを用意し，CNN の特徴をそのエンコーダへと入力する．

2. Transformer のエンコーダに bottom-up attention model の特徴と CNN の特徴を concat したものを入力する．

3. Transformer のエンコーダに bottom-up attention model の特徴を入力．エンコーダの出力と CNN の特徴を concat したものをデコーダへの入力とする．

ここで，デコーダは 2.1 節のものと同一であり，エンコーダ出力を全結合層に通し d_{model} 次元へと変換することに留意する．また，提案手法のエンコーダには位置エンコーディングは使用していない．

4 実験

4.1 実験設定

本節では，各手法を MSCOCO 2014 captions データセット [7] で評価し，提案手法の有効性を検証する．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つけられている．MSCOCO オフライン評価でモデル評価を行うために，Karpathy らの手法 [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，残りの画像を含めた 113,287 件の画像を訓練用画像に用いた．また，テキストの前処理として，空白スペースによる単語分割，文字の小文字化，辞書サイズは 10,000 とし，出現頻度の高いものから選択した．各画像に対するキャプションは 6 文以上あるものに対しては 5 文に制限した．画像の前処理として, 最短辺が 256 になるようにリサイズ処理を施した上で，224×224 サイズになるようにクロップ処理を施した．bottom-up attention model は, Anderson らの手法 [4] で事前学習されたモデルを利用した．モデルに使用する CNN は，Faster R-CNN に ResNet101[9] が用いられているため，提案手法で使用する CNN とベースラインのエンコーダにも ResNet101 を用いた．訓練時に CNN の fine-tuning は行わず，推論時におけるキャプション生成は貪欲法により行った．CNN から取得する特徴マップは，畳み込み層群の最終レイヤーの出力を使用した．したがって，抽出される特徴マップのサイズは $7 \times 7 \times 2048$ となる．Transformer のパラメータ設定は，Vaswani ら [5] に倣い，エンコーダ·デコーダレイヤをそれぞれ 6 個スタックし， ヘッド数は 8，d_{model}，d_{ff} はそれぞれ 512， 2048 とした．また，モデルの最適化手法は Adam を用い，Transformer の学習率算出に関わる $warmup\_steps$ は 20,000 とした．

表 1: 各手法の精度比較

手法	BLEU-1	BLEU-4	METEOR	CIDEr
ベースライン	70.3	27.9	24.0	91.4
提案手法 1	74.3	32.5	27.4	111
提案手法 2	74.9	33.0	27.1	109.5
提案手法 3	74.9	33.1	27.5	110.1
Up-Down[4]	77.2	36.2	27.0	113.5

4.2 評価手法

モデルの評価には，BLEU-n と METEOR，CIDEr を用いた．ここで，BLEU-n は，BP を考慮しない各 n-gram 以下のスコアの幾何平均であることに留意する [3]．CIDEr スコアは，各文中の n-gram に対して TF-IDF による重みづけを行いスコアを算出する．学習は 30epoch 行い，開発用データに対して最も性能の良いモデルに対し，テストデータにより評価を行った．

4.3 実験結果

ベースライン手法と提案手法の精度評価の結果を表 1 に示す．提案手法はベースライン手法と比較して, BLEU-1 スコアが最大 4.6 ポイント向上したことが確認できる．BLEU-4，METEOR，CIDEr においてもそれぞれ，5.2，3.5，19.6 ポイントの向上が確認できた．特に，CIDEr スコアにおいては 20 ポイント近いスコアの向上から大幅なモデルの性能向上が伺える．一方，同じく bottom-up attention model を用いた LSTM ベースの手法 (表中では Up-Down と表記)[4] と比較すると，BLEU-1 では 2.3 ポイント，BLEU-4 では 3.1 ポイント，CIDEr では 2.4 ポイント下回っていたが、METEOR では最大 0.5 ポイントのスコアの向上が確認できた．

5 おわりに

本研究では，自己注意機構をもつ Transformer モデルに着目し，画像内のオブジェクトや顕著な領域の関係性を考慮したモデルを提案した．bottom-up attention model より抽出した特徴マップを同時に使用することで，オブジェクトの位置関係を埋め込んだ情報と顕著なオブジェクト間での関係性を考慮した表現の獲得を目的とした．実験から，Transformer のデコーダを用いたベースライン手法と比較してキャプション生成の各指標において，スコアの向上が確認できた．今後は，2 つの特徴マップから顕著なオブジェクト領域の注目度を調整する仕組み等を取り入れることを目指したい．

謝辞

本研究成果は，国立研究開発法人情報通信研究機構の委託研究により得られたものである．ここに謝意を表する．

参考文献

[1] R. Girshick S. Ren, K. He and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pp. 91–99, 2015.

[2] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *The IEEE Conference on CVPR*, June 2015.

[3] 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, 2015.

[4] 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 *The IEEE Conference on CVPR*, June 2018.

[5] 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 NIPS*, pp. 5998–6008, 2017.

[6] Xinxin Zhu, Lixiang Li, Jing Liu, Haipeng Peng, and Xinxin Niu. Captioning transformer with stacked attention modules. *Applied Sciences*, Vol. 8, No. 5, p. 739, 2018.

[7] M. Belongie S.J. Hays J. Perona P. Ramanan D. Dollár P. Zitnick C Lawrence Lin, T. Maire. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

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

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on CVPR*, pp. 770–778, 2016.