画像物体間の構造情報を用いた深層学習によるキャプション生成

野口 敬輔　　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
{noguchi@ai., ninomiya@, tamura@}cs.ehime-u.ac.jp

1 はじめに

近年，自然言語処理分野の様々な領域においてニューラルネットワークに基づく手法が盛んに研究されており，従来手法よりも高い性能を実現することから注目されている．ニューラルネットワークに基づく手法は，自然言語処理だけではなく画像認識や音声認識など人工知能の諸分野で非常に高い性能を実現しているため，入力画像に対して説明文を生成するキャプション生成など，マルチモーダルなタスクにおいても高い性能を実現することが期待されている．

ニューラルネットワークに基づくキャプション生成の多くは，エンコーダ・デコーダモデルに基づいており，特に，画像から獲得された特徴ベクトルとデコーダの内部状態の類似度によるアテンション付きニューラルキャプション生成 [8] が知られている．アテンション付きニューラルキャプション生成を拡張する手法として，画像から物体の領域抽出処理を行い，それによって獲得された特徴ベクトルをアテンションとして活用する手法 [1] が提案されている．一方，キャプション生成と同様にエンコーダ･デコーダモデルを用いるニューラル機械翻訳では，入力文の句構造または係り受け構造に対する内部表現を生成し，それらをアテンションとして利用する手法 [2, 4] が提案されており，構文構造を表す内部表現に対するアテンションが有効であることが示されている．しかし，キャプション生成において画像内にある物体の関係性をアテンションに利用する手法に関する研究はまだ取り組まれていない．

本研究は，画像内にある物体の関係性を表すニューラルネットワークを部分構造として持つ新しいアテンション付きニューラルキャプション生成モデルを提案する．提案するモデルでは，まず，入力となる画像を畳込みニューラルネットワークに与えることで，抽象化された画像表現を得る．抽象化された画像表現は，画像としては解像度の低い粗い表現となるが，その各画素に対し特徴ベクトルが与えられている．本研究は，抽象化された画像の各画素間の全ての組み合わせに対し，その特徴ベクトルのペアを入力とする新しい畳込み層を提案する．キャプション生成のデコーディングにおいて，新しい畳込み層の出力に対しアテンションを適用することで，抽象化された画素間の関係性を表す文の生成を実現する．

キャプション生成の実験を行い，従来手法と比較し 0.98 ポイントの BLEU スコアの上昇を確認できた．また，入力画像に描写されている物体間に関する情報やそれらの修飾表現に関する情報を獲得できていることが確認できた．

2 エンコーダ・デコーダモデル

2.1 ニューラルキャプション生成

ニューラルキャプション生成 [7] は，入力の画像から，その説明文をエンコーダ・デコーダモデルにより生成する．エンコーダとデコーダにはそれぞれ畳込みニューラルネットワーク (CNN) と再帰型ニューラルネットワーク (RNN) が使用されている．RNN には Gated Recurrent Unit (GRU) や Long Short Term Memory (LSTM) が使用されている．本研究では LSTM を使用する．CNN エンコーダでは入力画像 I からその中間表現ベクトル $h_{enc} \in \mathcal{R}^{m\times 1}$ を生成する：

$$h_{enc} = CNN(I) \tag{1}$$

LSTM デコーダは，入力画像の中間ベクトルを初期状態として説明文 $t = (t_1, \cdots, t_l)$ を逐次的に生成する：

$$h_0 = W h_{enc}. \tag{2}$$
$$(t_i, h_{t_i}) = LSTM_{dec}(t_{i-1}, h_{i-1}). \tag{3}$$

ただし，h_{t_i} は LSTM デコーダの内部状態，$W \in \mathcal{R}^{n\times m}$ は重み行列である．

2.2 アテンション付きニューラルキャプション生成

アテンション付きニューラルキャプション生成は，説明文を生成する際に注目する画像領域をアテンションを用いて獲得する [8]．これまでアテンションとして様々な機構が提案されているが，本研究では Luong ら [3] のアテンション機構を利用する．Luong ら [3] のアテンション機構では，h_{t_j} の文脈ベクトル c_j を式 (4) で算出する：

$$c_j = \sum_{i=1}^{T} \alpha_j(s_i) h_{s_i}, \tag{4}$$

$$\alpha_j(s_i) = \frac{\exp(h_{t_j} \cdot h_{s_i})}{\sum_{s_k} \exp(h_{t_j} \cdot h_{s_k})}. \tag{5}$$

ここで，α は出力単語と入力（画像の領域）との関係度を表す．また，$s = (s_1, s_2, ..., s_T)$ は入力系列を表している．この文脈ベクトルを用いて以下の通り単語を生成する：

$$\bar{h_{t_j}} = tanh(W_c[h_{t_j}; c_j]), \tag{6}$$

$$p(t_j | t_{<j}, s) = softmax(W_o \bar{h_{t_j}}). \tag{7}$$

ただし，$W_c \in \mathcal{R}^{n \times 2n}$ は重み行列，$W_o \in \mathcal{R}^{V \times n}$，$V$ は語彙サイズである．

ニューラルキャプション生成では，アテンション機構で用いる内部状態 h_{s_i} は CNN の畳込み層の中間状態から導出する．畳込み層のフィルタ数を K，畳込みフィルタにより抽出された画像要素のサイズを $k \times k$，画像の内部状態の総数を $L(= k \times k)$ としたときの画像の内部状態を $h_{img} = \{h_1, h_2, ..., h_L\}, h_i \in \mathcal{R}^K$ で表す．図 1 は，画像における内部状態を示している．図のように複数の畳込み結果を積層することで，畳込み層の中間状態は総数 L の内部状態の集合となる．つまり，図 1 で着色している要素が一つの画像の内部状態となる．この画像の内部状態 h_{img} が上式の h_{s_i} に相当する．

また，このときのアテンション付きニューラルキャプション生成におけるデコーダ側 LSTM の初期状態 h_0 及び初期内部セル $cell_0$ は，式 (8),(9) で示すことが出来る：

$$h_0 = f_{MLPs,h}(\frac{1}{L} \sum_{i=1}^{L} h_{img_i}), \tag{8}$$

$$cell_0 = f_{MLPs,c}(\frac{1}{L} \sum_{i=1}^{L} h_{img_i}). \tag{9}$$

ここで，$f_{MLPs,h}, f_{MLPs,c}$ はそれぞれ LSTM の内部状態と内部セルを導出する関数を表している．

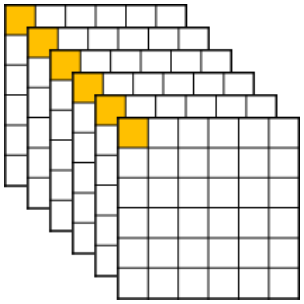

図 1　畳込み層における内部状態

3　提案手法

近年，エンコーダ・デコーダモデルを用いるニューラル機械翻訳では，入力文の構文構造をアテンション機構で利用することで翻訳精度を改善している [2, 4]．例えば，Eriguchi ら [2] は原言語の文の句構造解析結果から得られた 2 分木に従い TreeLSTM[6] を用いて親ノードとなる句に関する内部状態を生成し，アテンション機構に利用している．本研究では，TreeLSTM に基づいたユニット（以降，ペアユニット（PU）と呼ぶ）により画像内の複数の物体の組み合わせに対する内部状態を生成し，アテンション機構に活用することで，物体間の構造を考慮する新たなアテンション付きニューラルキャプション生成モデルを提案する．図 2 に提案モデルを示す．PU は，画像中の 2 つの物体の内部状態 (h_{img_i}, h_{img_j}) から，その組み合わせに対する内部状態 h_{pair} を生成する：

$$h_{pair} = PU(W_h h_{img_i}, W_h h_{img_j}). \tag{10}$$

ここで，$W_h \in \mathcal{R}^{n \times K}$ は重み行列である．提案手法では，式 (11) の組み合わせに対する内部状態を PU ユニットにより生成し，生成した内部状態を追加した $\bar{h_{img}}$ に対してアテンションを算出する：

$$\bigcup_{i,\, j(1 \le i < j \le L)} (h_{img_i}, h_{img_j}) \tag{11}$$

$$\bar{h_{img}} = [h_{img}; h_{pair}]. \tag{12}$$

なお，$[;]$ はベクトルの結合を表す．

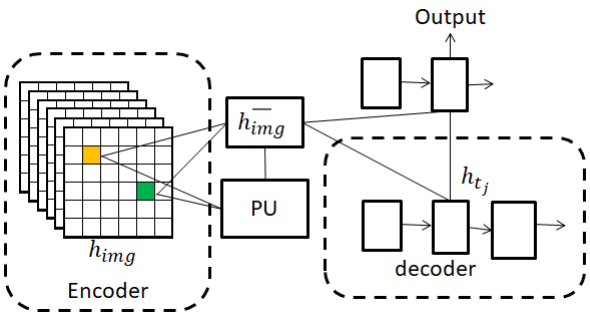

図 2　提案モデル

PU の具体的な構成は下記の通りである：

$$i = \sigma(\sum_{l=1}^{2} U_l^{(i)} h_l + b^{(i)}), \tag{13}$$

$$f_k = \sigma(\sum_{l=1}^{2} U_{kl}^{(f)} h_l + b^{(f)}), \tag{14}$$

$$o = \sigma(\sum_{l=1}^{2} U_l^{(o)} h_l + b^{(o)}), \tag{15}$$

$$\tilde{c} = tanh(\sum_{l=1}^{2} U_l^{(\tilde{c})} h_l + b^{(\tilde{c})}), \tag{16}$$

$$c = i \odot \tilde{c} + \sum_{l=1}^{2} f_l \odot c_l, \tag{17}$$

$$h = o \odot tanh(c). \tag{18}$$

ここで，$i, f_l, o, \tilde{c}, c \in \mathcal{R}^{n\times 1}$ はそれぞれ，入力ゲート，内部状態 l 用の忘却ゲート (l = 1,2)，出力ゲート，更新用メモリセル，メモリセルを表しており，$U_l^{(i)}, U_{kl}^{(f)}, U_l^{(o)}, U_l^{(\tilde{c})} \in \mathcal{R}^{n\times n}$ は，それぞれ対応するゲート及びメモリセルの重み行列である．$b^{(i)}, b^{(f)}, b^{(o)}, b^{(\tilde{c})} \in \mathcal{R}^{n\times 1}$ はそれぞれのバイアスである．$\odot$ は，ベクトルの要素積である．

4　実験

4.1　実験設定

実験データには Microsoft COCO データセットから 82,783 件の画像データを使用した．説明文データは各画像に対して 5 件以上あるが，学習時には各画像につき 1 文ずつ使用した．画像サイズは 224×224 でトリミングした．説明文のデータ整形は小文字化，ピリオド，改行の削除を行った．辞書サイズは 10,000 単語とし，出現頻度の高いものから選択した．開発用データ及びテストデータはそれぞれ Xu らの研究 [8] に従い 5,000 件ずつ用意した．評価時には，各画像に対して正解文を 5 文ずつ用意し，6 文以上あるものに対しては無作為に 5 文を選択した．学習はデコーダとアテンション機構のみ行い，エンコーダとなる CNN 層は ILSVRC-2014 の Classification+localization 分野で 1 位となった事前学習済みの vgg19 モデル [5] を用いた．デコーダ層の内部状態，PU の内部状態の次元数はそれぞれ 512 とし，エポック数は 20 とした．最適化手法は Adam を使用，学習率は 0.001 を初期値とした．1 エポック終了毎に開発用データを使用して汎化状態の確認を行い，前エポックより単語予測一致率が低下した場合のみ学習率を 0.5 倍した．Dropout 確率は 0.5 とした．バッチサイズは 40 とした．精度評価については BLEU スコアを採用した．アテンション機構への入力は Xu らの手法に従い，1 時刻前のデコーダ側内部状態を使用した．

ベースライン手法にはアテンション付きニューラルキャプション生成を採用し，アテンションを適用する畳込み中間層には 5 番目の畳込み層群の 4 番目の層 (Conv5_4 layer) を用いた．提案手法は，Conv5_4 layer に加え，Conv5_4 layer に最大値プーリングを施した中間層に対し PU を適応し，得られた内部状態に対しアテンションを適用した．また，本実験では高速化のために PU に与える内部状態の全組み合わせのうち，無作為に選ばれた半分の内部状態だけを生成しアテンションの対象とした．

4.2　実験結果

表 1 に実験結果を示す．ベースライン手法がアテンション付きニューラルキャプション生成であり，提案手法はそれに更に PU によるアテンション機構を追加したモデルである．表 1 より，提案手法では従来手法よりも 0.98 ポイント BLEU スコアが上昇したことが確認できた．

また，図 3 に入力画像と正解説明文，従来手法による生成文及び提案手法による生成文を示す．従来のアテンション付きニューラルキャプション生成 (ベースライン手法) によって生成されている文 “a cat sitting on the table” となっており，画像内の動物に関する説明のみ生成されており，猫の周辺にある物体についてには言及されていない．それに対し，提案手法による生成文は “a black cat sitting on a table with a red plate” となっており，生成文自体は誤っているものの猫の周辺にある物体に関する説明文が生成されていることが確認できる．また，猫に関する説明も “a cat” から “a black cat” のようにより詳しい表現が生成されている．

正解文 1:a black cat sitting on desk with two teddy bears
正解文 2:a black cat is standing on the desk with a printer and small stuffed bears
ベースライン手法による翻訳文：a cat sitting on a table
提案手法による翻訳文：a black cat sitting on a table with a red plate

図 3　入力画像と各モデルによる生成文の比較

表 1　テストデータに対する精度比較

手法	BLEU
ベースライン手法	11.12
提案手法	12.10

4.3　考察

CNN は本来，画像のどの位置に何が描写されているのかの学習を行っており，それら物体の関係情報までは完全に考慮されていない．そのため，従来手法では “a black cat” という正解文に対して “black” と “cat” の関係性までは獲得できず，生成文は “a cat” となったと考えられる．これは，生成文で “with” 以下が欠落していることからも示すことができ，従来手法では物体同士の位置関係の学習が出来ていないことがわかる．それに対して，提案手法では画像から学習できた物体の特徴量を PU によって結合させることで物体同士の状態に関する情報が獲得されていることが “a black cat” や “with” 以下の猫の周辺にある物体に対しての説明を含めた文が生成できたことから確認できる．

5　終わりに

本研究では，自然言語処理分野におけるエンコーダ・デコーダモデルに注目し，画像に対して疑似的に構造情報を獲得し，文生成を行う新たなニューラルキャプション生成モデルを提案した．実験では従来のアテンション付きニューラルキャプション生成モデルと比較し，生成文の改善が確認でき有用性を示すことが出来た．提案手法は，従来の画像解析を必要とせず画像から構造情報の獲得が実現できる．今後はより大量のデータセットに対する実験について取り組みたい．

謝辞

本研究は JSPS 科研費 25280084 の助成を受けたものである．ここに謝意を表する．

参考文献

[1] 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 VQA. *CoRR*, Vol. abs/1707.07998, , 2017.

[2] Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. Tree-to-sequence attentional neural machine translation. *CoRR*, Vol. abs/1603.06075, , 2016.

[3] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. *CoRR*, Vol. abs/1508.04025, , 2015.

[4] Rico Sennrich and Barry Haddow. Linguistic input features improve neural machine translation. *CoRR*, Vol. abs/1606.02892, , 2016.

[5] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, Vol. abs/1409.1556, , 2014.

[6] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. *CoRR*, Vol. abs/1503.00075, , 2015.

[7] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *CoRR*, Vol. abs/1411.4555, , 2014.

[8] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, Vol. abs/1502.03044, , 2015.