CNNとTransformerエンコーダを用いたマルチモーダルニューラル機械翻訳

宅島 寛貴[1]　田村 晃裕[2]　二宮 崇[2]　中山 英樹[3]

[1]愛媛大学 工学部情報工学科

[2]愛媛大学 大学院理工学研究科 電子情報工学専攻

[3]東京大学 大学院情報理工学系研究科

{takushima@ai., tamura@, ninomiya@}cs.ehime-u.ac.jp,
nakayama@nlab.ci.i.u-tokyo.ac.jp

1 はじめに

近年，自然言語処理の分野ではニューラルネットワークを利用した研究が盛んに行われている．機械翻訳においてもニューラルネットワークを用いた機械翻訳（NMT）に関する様々な手法が提案されている．特に，自己注意機構を用いるTransformerモデル [1] は，RNNやCNN (Convolution Neural Network) ベースのNMTの性能を上回り，注目されている．Transformerモデルは，RNNベースやCNNベースのモデル同様，原言語文から中間表現を生成するエンコーダと中間表現から目的言語文を予測するデコーダから構成される．エンコーダやデコーダは，自己注意機構という，同一文中の単語間の関連を捉える構造を有しており，エンコーダ内の自己注意機構で入力された文中の単語間，デコーダ内の自己注意機構で出力文中の単語間の関連を考慮して翻訳することができる．

これまで機械翻訳の性能を改善するため様々な研究が行われているが，その中の一つに，マルチモーダル学習によりNMTの性能改善を目指す研究がある [2]．マルチモーダル学習とは，複数のモダリティから学習する手法全般のことを指し，人間が言葉の意味を視覚や聴覚，触覚など複数のモダリティから理解するのと同様な統合的理解を計算機上で実現させることを目的としている．マルチモーダル学習は様々なタスクで利用されており，機械翻訳においては，対訳コーパスに加えて，画像を利用することで翻訳性能が改善されている．

画像を用いるマルチモーダルNMTでは，入力として，原言語文に加えて，原言語文に関連する画像を与える．この入力画像は，原言語文中の多義語を翻訳する際の曖昧性解消の手がかりになるなど，翻訳の性能改善に寄与すると考えられている．例えば，“bank”という英単語の日本語訳は”銀行”と“土手”という2つの意味が考えられるが，金融関連の画像が入力されれば“銀行”に，川岸の画像が入力されれば“土手”に正しく翻訳されることが期待できる．

マルチモーダルNMTモデルとして，Barraultら [3] は，入力画像からCNNを用いて抽出した画像特徴量を，目的言語文のデコード時に，NMTのデコーダとエンコーダ間の注意機構内で考慮する手法を提案している．また，入力画像から抽出した画像特徴量を，NMTエンコーダで原言語文をエンコードする際に活用する手法 [4, 5] 等も提案されている．しかし，これらの従来手法では画像内の特徴を捉えることができても，画像の領域間の関係を捉えることができない．

そこで本研究では，画像の領域間の関係を考慮するマルチモーダルNMTモデルを提案する．具体的には，提案モデルでは，画像と原言語文をそれぞれエンコードし，それらの中間表現を合成したものをデコーダの入力として目的言語文を生成する．その際，画像のエンコーダでは，CNNにより画像を領域毎に特徴ベクトル化した後，その画像の特徴ベクトルをTransformerエンコーダに入力することで，Transformerエンコーダの自己注意機構により，画像の領域間の関係を考慮したエンコードを可能とする．

Multi30kデータセットを使用した英独翻訳の評価実験を行い，CNNによる画像特徴量を，Transformerエンコーダに通さずにそのまま画像の中間表現として用いるベースラインと比較して，提案モデルはBLEUスコアで0.47ポイント向上したことを確認した．

2 Transformer モデル

Transformer モデル [1] は原言語文として $X = (x_1, x_2, ..., x_M)$，目的言語文として $Y = (y_1, y_2, ..., y_N)$ が与えられた時，X を Y に変換する目的関数 $p(Y|X)$ を学習するニューラルネットワークである．Transformer エンコーダは原言語文 X から中間表現 h_t を生成し，Transformer デコーダは中間表現 h_t から目的言語文 Y を推測する:

$$h_t = TransformerEncoder(X) \quad (1)$$

$$Y = TransformerDecoder(h_t) \quad (2)$$

Transformer エンコーダ，Transformer デコーダは，それぞれ，レイヤを複数層スタックしたモデルである．エンコーダの各レイヤは，下位から順に，自己注意機構，単語位置毎の全結合層 (FeedForward，以下「FF」と記す) の 2 つのサブレイヤで構成される．一方，デコーダの各レイヤは，エンコーダの 2 つのサブレイヤの間に，原言語と目的言語間の注意機構 (以下,「言語間の注意機構」と呼ぶ) を加えた 3 つのサブレイヤで構成される．エンコーダ，デコーダの各サブレイヤ間には，レイヤの正規化と残差接続が行われる．

自己注意機構及び言語間の注意機構の各注意機構 $Attention(\cdot)$ の計算は次式で表される:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_{model}}})V \quad (3)$$

ここで，Q, K, V はエンコーダ／デコーダの内部表現を表す．また，d_{model} は内部表現の次元のサイズである．Q と K の内積は Q と K の各要素間の類似度であり，softmax 関数によって確率値にすることで Q の K に対する注意の重みを算出できる．注意機構は，この注意の重みを用い，V との重み付き荷重和を取る操作であり，その結果，Q と K の間の単語の関連の強さを考慮した表現を獲得できる．自己注意機構では，Q, K, V として同一の入力源（エンコーダはエンコーダ内の内部表現，デコーダはデコーダ内の内部表現）を用いることで，同一文内の単語間の関連の強さを計算することができる．言語間の注意機構では，Q にはデコーダの内部表現，K, V はエンコーダからの最終出力を用いることで，原言語文の各単語と目的言語の単語との関連の強さを重み付き加重和により表現できる．

また，注意機構は単一ではなく複数ヘッドの注意機構を取ることで異なる空間での処理が可能となり，よりよい性能を発揮することが知られている．h 個の複数ヘッドの注意機構 $MultiHead(\cdot)$ は次式で表される:

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^o \quad (4)$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (5)$$

ここで，$W_i^Q \in \mathcal{R}^{d_{model} \times d_k}, W_i^K \in \mathcal{R}^{d_{model} \times d_k}, W_i^V \in \mathcal{R}^{d_{model} \times d_k}$ は d_{model} 次元の Q, K, V を d_k $(= d_{model}/h)$ 次元に線形変換するための重み行列，$Concat$ は行列の結合を示す関数である．このように，複数ヘッドの注意機構は，h 個のヘッドそれぞれで注意機構を計算し，それらを連結して線形変換する機構である．

また，Transformer では単語の位置情報を考慮するために Positional Encoding(以下,「PE」と記す) が用いられる．PE は正弦波関数と余弦波関数によって計算された行列の各成分を入力文の埋め込み成分にそれぞれ加算する:

$$PE(pos, 2i) = sin(pos/10000^{2i/d_{model}}) \quad (6)$$

$$PE(pos, 2i+1) = cos(pos/10000^{2i/d_{model}}) \quad (7)$$

ここで pos は単語の位置，i は成分の次元を示す．

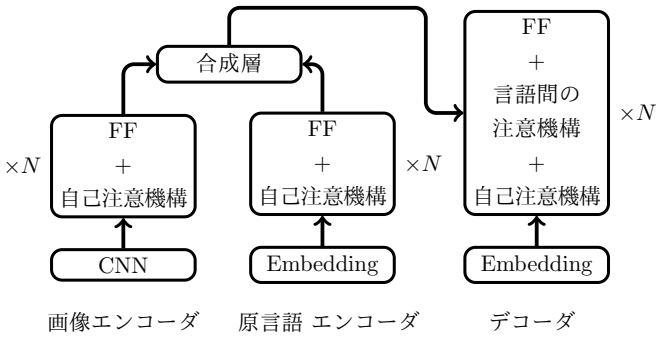

図 1: 提案手法の全体図

3 提案手法

本研究では CNN と Transformer エンコーダを用いたマルチモーダル機械翻訳を提案する．画像の領域間の関連の強さを Transformer エンコーダを用いて考慮することで翻訳精度の改善を目指す．提案手法の全体図を図 1 に示す．提案手法では，入力画像のエンコーダ (以下,「画像エンコーダ」と記す) と原言語文のエンコーダ (以下,「原言語エンコーダ」と記す) の 2 つのエンコーダを用いて入力をエンコードする．そして，

それぞれのエンコーダが出力する中間表現を連結したものを，通常の Transformer のデコーダの入力し，目的言語文を生成する．

以降では，まず，3.1 節で画像エンコーダについて説明し，3.2 節で画像エンコーダと原言語エンコーダ間の中間表現の合成について詳細に説明する．デコーダについては従来の Transformer モデルと同様なので説明は割愛する．

3.1 画像エンコーダ

画像エンコーダでは，まず，入力画像から領域毎の画像特徴量を得るために CNN を用いる．CNN とは畳み込み層とプーリング層を多層に繰り返すニューラルネットワークである．実験では，CNN として VGG16[6] を用いた．VGG16 とは畳み込み層 13 層と全結合層 3 層の計 16 層から構成されるニューラルネットワークである．その後，CNN が出力した画像特徴量を Transformer エンコーダに入力する:

$$feature = CNN(image) \tag{8}$$

$$h_v = TransformerEncoder(W \cdot feature) \tag{9}$$

ここで，$image$ は入力画像，$feature$ は CNN が抽出した画像特徴量，h_v は画像エンコーダが出力する中間表現，$W \in \mathcal{R}^{d_{feature} \times d_{model}}$ は $d_{feature}$ 次元から d_{model} 次元に線形変換するための重み行列を表す．また，画像エンコーダでは PE による位置情報の付与は行わない．

3.2 中間表現の合成

式 1 と式 9 より，原言語エンコーダ，画像エンコーダで原言語文と入力画像の中間表現を生成した後，合成層でそれぞれの中間表現を連結し，その結果得られる中間表現 h を提案手法のデコーダの入力とする:

$$h = Concat(h_v, h_t) \tag{10}$$

4 実験

4.1 実験設定

本実験では，Multi30k[7] をデータセットとして用いる. 言語対は独英とした．訓練データとして 29,000 文，検証データとして 1,014 文，そしてテストデータとして 1,000 文を用意した．テキストデータに対しては BPE (Byte Pair Encoding)[8] を用い，出現頻度の低い単語をサブワード単位に分解してエンコーダとデコーダの辞書の共有を行った．語彙サイズは 6,150 となった．提案モデルの画像エンコーダにおける，画像の各領域の画像特徴量は，VGG16[6] の 13 個目の畳み込み層の出力を用いた．Transformer のパラメータは Vaswani ら [1] に倣って，エンコーダとデコーダのレイヤをそれぞれ 6 層スタックし，ヘッド数は 8 つ，埋め込み次元は 512 次元とした．また，optimizer は Adam を使用した．モデル学習時はミニバッチサイズを 80 に設定し，50 エポック繰り返した．また，学習時には CNN の fine-tuning は行わず，推論時は貪欲法で目的言語文を出力した．

提案手法の評価は BLEU を用いた．エポックごとの検証データの BLEU スコアが最も良いモデルを選択し，テストデータに対する性能を評価した．画像を用いていない Transformer（以下,「画像なし $Transformer$」と記す)，CNN による画像特徴量を Transformer エンコーダへ入力せずに原言語エンコーダと中間表現を合成したモデル（以下,「$CNN + Trans_{Dec}$」と記す）と提案モデル（以下,「$CNN + Trans_{Enc} + Trans_{Dec}$」と記す）を比較した．

表 1: 実験結果

手法	BLEU
画像なし $Transformer$	34.30
$CNN + Trans_{Dec}$	34.79
$CNN + Trans_{Enc} + Trans_{Dec}$	35.26

4.2 結果

表 1 より，提案モデルは画像なし $Transformer$ と比較して BLEU スコアで 0.96 ポイントの向上，$CNN + Trans_{Dec}$ と比較して BLEU スコアで 0.47 ポイントの向上が見られた．

5 画像情報の有効性に関する考察

表 2 に各モデルで翻訳した実例を示す．提案モデル，$CNN + Trans_{Dec}$ に入力した画像は図 2 である．表 2 より，画像なし $Transformer$ の翻訳では果物 (obst)

表 2: 各手法の翻訳例

入力文	two young boys putting fruit on the bike .
参照訳	zwei jungen packen obst auf das fahrrad . （2 人の男の子が自転車に果物を詰め込んでいます.）
画像なし $Transformer$	zwei kleine jungen legen auf das bike . （2 人の男の子が自転車に乗っています.）
$CNN + Trans_{Dec}$	zwei kleine jungen bringen obst auf dem fahrrad an . （2 人の男の子が自転車に果物を入れています。）
$CNN + Trans_{Enc} + Trans_{Dec}$	zwei jungen stellen obst auf das fahrrad . （2 人の男の子が自転車に果物を入れます.）

図 2: 原画像

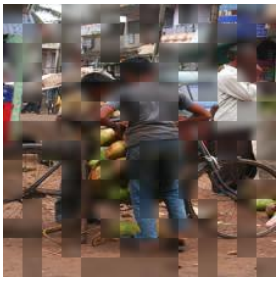

図 3: 果物への注意の可視化

の情報が欠落し，約抜けしていることが分かる．それに対して，画像を活用する $CNN + Trans_{Dec}$ や提案手法では果物 (obst) の情報が抜け落ちることなく翻訳できている．画像情報が実際に有効であるのかを調べるためにデコーダの最終レイヤの果物という意味を表している単語 (obst) と画像の間の注意を図 3 に可視化した．鮮明に見えている部分が注意の重みが大きい部分である．図 3 より，実際に果物のある領域に注意が向けられていることが分かり，画像情報の有効性が確認できる．

6 おわりに

本研究は，CNN と Transformer エンコーダを用いたマルチモーダル機械翻訳モデルを提案した．そして，画像の領域間の関係を取ることによる翻訳精度の向上が確認できた．今後は，物体検出技術を組み込むことでより精度の向上を目指していきたい．

7 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] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in NIPS*, pp. 5998–6008. Curran Associates, Inc., 2017.

[2] L. Barrault, F. Bougares, L. Specia, C. Lala, D. Elliott, and S. Frank. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pp. 304–323, 2018.

[3] I. Calixto, Q. Liu, and N. Campbell. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1913–1924. ACL, 2017.

[4] P.-Y. Huang, F. Liu, S.-R. Shiang, J. Oh, and C. Dyer. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pp. 639–645. ACL, 2016.

[5] I. Calixto and Q. Liu. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 992–1003. ACL, 2017.

[6] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[7] D. Elliott, S. Frank, K. Sima'an, and L. Specia. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pp. 70–74. ACL, 2016.

[8] R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.