注意機構制約を用いたマルチモーダルNMT

西原 哲郎[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

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

1 はじめに

近年，自然言語処理の多くのタスクにおいてニューラルネットワークに基づく手法が主流になっている．機械翻訳においてもニューラルネットワークに基づいた機械翻訳 (NMT) が主流であり，様々な手法が提案されている．初期から広く使われてきたNMTモデルがRNNベースの注意機構付きNMT[1]である．このモデルは，原言語文内の単語と目的言語文内の単語の関連性を捉える注意機構を用いることで，従来のRNNベースのNMTより高い精度を実現した．また，近年，従来の言語間の注意機構に加えて自己注意機構を導入したTransformerモデル [2] がRNNやCNNを用いた手法と比べて高い精度を実現し，注目されている．自己注意機構では，エンコーダ及びデコーダでそれぞれ，原言語文及び目的言語文内の単語間の関連性を考慮することが可能となっている．機械翻訳の性能を改善する手法については様々な研究がされているが，その内の一つに上述の言語間注意機構に制約を与える研究がある [3]．彼らの研究ではあらかじめアライメントツールを用いて原言語文内の単語と目的言語文内の単語間の対応関係を取得し，その対応関係を教師データとして与えて注意機構を学習させることで翻訳性能が向上している．

機械翻訳の手法の一つとして，原言語文に加えて画像を入力することで翻訳性能の改善を目指すマルチモーダルNMT（MNMT）がある [4]．入力画像が翻訳時の曖昧性解消や省略補完の手がかりなどとして役立つと考えられている．MNMTモデルとして，Helclら [5] はCNNにより抽出した画像の特徴量を翻訳に活用するためにTransformerモデルのデコーダ内に，文中の単語と画像の領域との対応を捉える視覚的注意機構を導入したモデルを提案している．また，Delbrouckら [6] は，RNNベースのNMTモデルのエンコーダ内に視覚的注意機構を導入したモデルを提案している．

本研究は，MNMTの性能改善のために，言語間注意機構と視覚的注意機構に制約を与える手法を提案する．従来のMNMTの視覚的注意機構は，注意すべき領域が教師データとして与えられるわけではなく，MNMTの学習を通じて自動的に学習されている．本研究は，原言語文中の単語と画像内のオブジェクトとの対応関係を示した教師データを用意することで，Transformerモデルのエンコーダ内で視覚的注意機構を直接学習させることを行う．

実験では，国立研究開発法人情報通信研究機構（NICT）の委託研究で作成されているMulti30k英日データセットを使用した英日翻訳の評価実験を行い，視覚的注意機構及び言語間の注意機構に制約を加えて学習することにより，MNMTの翻訳性能が改善することを示す．

2 関連研究

2.1 Transformerモデルに基づくマルチモーダルNMTモデル

MNMTモデルとして色々なネットワーク構造のモデルが提案されているが，近年は，Transformerモデル [2] に基づくMNMTモデルが盛んに研究されている．例えば，宅島ら [7] はCNNとTransformerエンコーダからなる画像エンコーダをTransformer NMTに導入したMNMTモデルを提案している．彼らのモデルでは，画像エンコーダで，まず，入力画像に対してCNNを適用し，入力画像の特徴量を獲得する．その後，獲得した特徴量をTransformerモデルのエンコーダに入力し，自己注意機構によって画像の領域間の関

係を考慮したエンコードを行う．画像のエンコードと並行して Transformer エンコーダにより原言語文をエンコードし，画像と原言語文のエンコード結果を結合した中間表現から Transformer モデルのデコーダにより目的言語文を生成する．

以降では，Transformer に基づく MNMT のベースとなっている Transformer NMT モデル [2] を説明する．Transformer NMT モデルは，原言語文を中間表現に変換するエンコーダと，変換された中間表現を受け取って目的言語文を生成するデコーダから構成されている．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複数スタックした構成となっている．エンコーダレイヤは自己注意機構と位置毎の全結合層の 2 つのサブレイヤで構成されている．デコーダレイヤは上記の 2 つのサブレイヤの間に，言語間注意機構を加えた 3 つのサブレイヤから構成されている．なお，各サブレイヤ間では残差接続と層の正規化が行われる．

自己注意機構及び言語間注意機構 (Att) は，以下の式で表される．

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

ここで，Q,K,V はエンコーダ及びデコーダでの隠れ状態を表し，d_k は Q,K,V の次元数を表す．式 (1) において，Q,K,V が全て前のサブレイヤの出力から与えられる場合は自己注意機構となる．また，デコーダにおいて Q のみが前のサブレイヤの出力から与えられ，K と V がエンコーダの出力から与えられる場合は言語間注意機構となる．

Transformer モデルにおける注意機構は，複数ヘッドの注意機構とすることで，様々な部分空間から情報を取り入れることができるようになり，性能が向上することが知られている．複数ヘッドの注意機構 (MHA) は以下の式で表される．

$$MHA(Q,K,V) = [head_1;\ldots;head_h]W^O \quad (2)$$

$$head_i = Att(QW_i^Q, KW_i^K, VW_i^V) \quad (3)$$

ここで，$W_i^Q \in \mathbb{R}^{d_{model}\times d_k}, W_i^K \in \mathbb{R}^{d_{model}\times d_k}, W_i^V \in \mathbb{R}^{d_{model}\times d_k}, W_i^O \in \mathbb{R}^{hd_k\times d_{model}}$ はパラメータ行列である．なお，d_{model} は埋め込み次元数を表しており，$d_k = d_{model}/h$ である．

Transformer モデルは，RNN や CNN ベースのモデルとは異なり，単語の語順が考慮されていない．そこで，単語の語順を考慮するために以下の式で表される位置エンコーディングが用いられている．

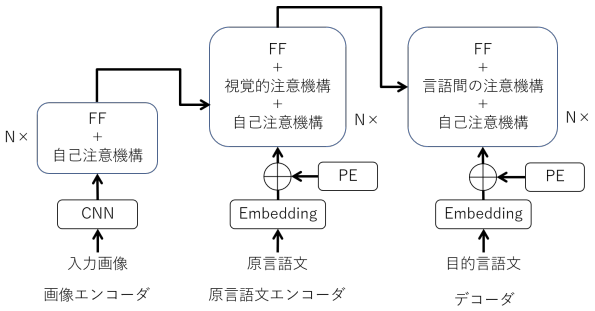

図 1: 提案手法の MNMT モデル

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

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}}) \quad (5)$$

ここで，pos は単語の位置，i は次元を表す．この位置エンコーディングを単語の埋め込み表現に加算することで，単語の語順の情報を付与する．

2.2 言語間注意機構に対する制約

Liu ら [3] の研究では，NMT において言語間の注意機構に対して教師を与えて学習させる手法を提案している．アライメントツールを用いることで言語間の対応関係を取得し，注意機構によって計算される注意行列との誤差を計算することで学習を行う．誤差の計算は以下に示す交差エントロピーによって行う．

$$L_a(A) = -\sum_m \sum_n G_{m,n} \times log(A_{m,n}) \quad (6)$$

ここで，A は注意機構によって計算された注意行列を，G は教師となる行列を表す．この $L_a(A)$ を通常の翻訳誤差 L_t に加えた誤差関数に基づきモデルを学習することで，言語間注意機構に対する制約を用いた NMT を実現している．全体の誤差関数 L は以下のようになる．

$$L = L_t + \lambda L_a(A) \quad (7)$$

ここで，λ はハイパーパラメータである．

3 提案手法

本研究における，マルチモーダル Transformer NMT モデルのモデル図を図 1 に示す．図中の PE は位置エンコーディングを表す．本モデルは，原言語文エンコー

ダ内において，画像エンコーダの出力と自己注意機構の出力との視覚的注意機構により，原言語の単語と画像のオブジェクト間の関連を捉える．このモデルの原言語文エンコーダ内の視覚的注意機構及びデコーダ内の言語間注意機構に対して制約を加える．視覚的注意機構に対する制約では原言語文内の単語と対応する画像の領域との対応関係を示した教師を用いて注意機構のヘッドを学習させる．また，言語間注意機構に対する制約では，原言語文内の単語と対応する目的言語文内の単語との対応関係を示した教師を用いて注意機構のヘッドを学習させることで，翻訳性能の改善を目指す．

誤差関数 L については，マルチモーダル Transformer モデルの誤差関数を L_T，視覚的注意機構における注意行列と制約行列との誤差関数を L_{img_src}，言語間注意機構における注意行列と制約行列との誤差関数を L_{src_tgt} とすると以下のようになる．

$$L = L_T + \lambda_1 L_{img\_src} + \lambda_2 L_{src\_tgt} \tag{8}$$

ここで，λ_1 および λ_2 はハイパーパラメータである．

以降では，3.1 節で視覚的注意機構に対する制約行列の作成方法について説明し，3.2 節で言語間注意機構に対する制約行列の作成方法について説明する．

3.1 視覚的注意機構に対する制約

本研究では，Flickr30k entities データセット [8] を用いて視覚的注意機構に対する制約を作成した．Flickr30k entities データセットは，一つの画像に対して 5 つのキャプション文がつけられており，各キャプション文中の単語が画像内のオブジェクトと関係がある場合，その単語が画像内のどの領域と関係があるかがアノテーションされたデータセットとなっている．今回はこのデータセットから原言語文内の単語と画像内のオブジェクトとの対応関係を抽出し，制約を作成する．

まず，Flickr30k entities データセットに付与されている単語とオブジェクト間の対応関係を CNN で畳み込んだ際の領域にスケールさせる．例えば，画像エンコーダに用いる CNN によって画像を 3×3 に畳み込んだ場合，画像内の各物体と 9 つの領域との対応関係を求める．複数の領域に対応する場合は，各領域に等しく対応が張られるように値を平均化する（値を「1/対応付いた領域数」とする）．その後，2 次元の領域を 1 次元に線形化し，先頭に特殊トークンを付与する．画像内のオブジェクトと対応がない単語については，先頭に設置した特殊トークンに対して対応付けを行う．この工程を原言語文のすべての単語に対して行うため，最終的な制約行列のサイズは「原言語文の単語数 ×（畳み込んだ領域数+1）」となる．

3.2 言語間注意機構に対する制約

本研究で使用する，NICT の委託研究で作成されている Multi30k 英日データセットには，原言語文内の単語と目的言語文内の単語との対応関係が人手で付与されている．そこで，言語間注意機構に対する制約は，アライメントツールではなくデータセットに付与されたアノテーションから抽出して作成する．なお，目的言語文内のある 1 単語が複数の原言語文内の単語と対応関係にある場合，等しく対応が張られるように値を平均化する（値を「1/対応づいた単語数」とする）．また，対応関係がない目的言語文内の単語については，原言語文の先頭に特殊トークンを置き，そのトークンに対して対応関係が張られるようにする．

4 実験

4.1 実験設定

本研究では，NICT の委託研究で作成されている Multi30k 英日データセットを用いて英日翻訳実験を行う．このデータセットは，Multi30k データセット [9] の英文を人手で日本語に翻訳し，英文と日本文の単語の対応関係を人手で付与した英日の対訳コーパスである．Multi30k データセットのテキストの前処理に倣い，英文には文字の小文字化，句読点の正規化，Moses のトーカナイザ[1]を施している．日本文については Kytea[2]を用いて単語分割を行った．また，訓練データには日英共に 100 単語以下の対訳文のみを用いた．訓練データは 59,516 文，開発データは 2,017 文，テストデータは 2,000 文である．画像に対する前処理として，画像サイズを 256×256 になるようにリサイズした後，224×224 となるように中央部にクロップ処理を施した．画像エンコーダにおいて使用する CNN は ResNet50 [10] を用いた．なお，ResNet50 から取得する画像特徴量は最終の畳み込み層の出力を用いた．したがって抽出さ

[1] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

[2] http://www.phontron.com/kytea/index-ja.html

表 1: 実験結果

制約を加えた箇所	BLEU(%)
制約なし	44.39
視覚的注意機構	44.26
言語間注意機構	44.82
視覚的注意機構+言語間注意機構	45.03

れる画像特徴量のサイズは $7 \times 7 \times 2048$ である．また，学習時に CNN の fine-tuning は行わない．画像エンコーダ，原言語文エンコーダおよびデコーダレイヤはそれぞれ 6 層スタックし，ヘッド数は 8，埋め込み次元数は 512 とした．モデルの最適化手法は Adam を使用した．モデルの学習時にはミニバッチサイズを 128 とし，40 エポックの学習を行った．推論時には目的言語文の生成を貪欲法により行った．

翻訳性能は BLEU で評価した．開発データに対する BLEU 値が最も高かったエポックのモデルを選択し，テストデータに対する性能を評価した．実験では，制約を与えていないマルチモーダル Transformer モデル，視覚的注意機構にのみ制約を与えたモデル，言語間注意機構にのみ制約を与えたモデル，視覚的注意機構及び言語間注意機構の両方に制約を加えたモデルを比較した．なお，式 (8) におけるハイパーパラメータは $\lambda_1 = 0.03$，$\lambda_2 = 0.05$ とした．また，視覚的注意機構にのみ制約を与えたモデル及び言語間注意機構にのみ制約を与えたモデルにおけるハイパーパラメータは $\lambda = 0.05$ とした．

4.2 実験結果

表 1 に実験結果を示す．視覚的注意機構にのみ制約を与えた場合は，制約を与えなかったモデルと比較して BLEU 値が 0.13 ポイント下がった．しかし，言語間注意機構にのみ制約を与えたモデル及び視覚的注意機構と言語間注意機構の両方に制約を加えたモデルでは，それぞれ BLEU 値が 0.43 ポイント，0.64 ポイントの向上を確認した．このことより，提案手法のように注意機構に制約を与えることで MNMT の翻訳性能を改善できることが実験的に確認できた．

5 おわりに

本研究では，言語間注意機構及び視覚的注意機構に対する制約を用いた MNMT モデルを提案した．Multi30k 英日翻訳実験により，視覚的注意機構と言語間注意機構の制約を同時に与えることで翻訳性能を改善できることを確認した．今後は画像と目的言語間の注意機構に対する制約を導入したい．

6 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．また，本研究の一部は JSPS 科研費 18K18110 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] T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. In *Proc. of EMNLP*, pp. 1412–1421, 2015.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Proc. of NIPS*, pp. 5998–6008. 2017.

[3] L. Liu, M. Utiyama, A. Finch, and E. Sumita. Neural machine translation with supervised attention. In *Proc. of COLING*, pp. 3093–3102, 2016.

[4] L. Barrault, F. Bougares, L. Specia, C. Lala, D. Elliott, and S. Frank. Findings of the third shared task on multimodal machine translation. In *Proc. of WMT*, pp. 304–323, 2018.

[5] J. Helcl, J. Libovický, and D. Variš. CUNI system for the WMT18 multimodal translation task. In *Proc. of WMT*, pp. 616–623, 2018.

[6] J. B. Delbrouck and S. Dupont. Modulating and attending the source image during encoding improves multimodal translation. 2017.

[7] 宅島寛貴, 田村晃裕, 二宮崇, 中山英樹. 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. 言語処理学会第 25 回年次大会, pp. 743–746, 2019.

[8] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, Vol. 123, No. 1, pp. 74–93, 2017.

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

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. of CVPR*, pp. 770–778, 2016.