画像生成による疑似教師データを用いたマルチモーダルニューラル機械翻訳

岩本 裕司 [1]　田村 晃裕 [2]　二宮 崇 [1]
[1] 愛媛大学　[2] 同志社大学
[1] {iwamoto@ai., ninomiya@}cs.ehime-u.ac.jp　[2] aktamura@mail.doshisha.ac.jp

1 はじめに

近年，ニューラル機械翻訳（Neural Machine Translation; NMT）の性能を向上させる手段の 1 つとして，マルチモーダルニューラル機械翻訳（Multimodal Neural Machine Translation; MNMT）が注目されている．MNMT は翻訳元の文（原言語文）だけでなく関連画像も用いることで，これらの画像を手がかりに状況に即したより自然な翻訳文（目的言語文）を生成することを目的としている．MNMT モデルの学習には通常，対訳テキストデータに加えて関連画像が必要となるが，そのような原言語文，目的言語文，関連画像で構成される 3 つ組の対訳データは通常の対訳データに比べ非常に小規模なものしか存在していない．また，通常の対訳データに比べ，MNMT 学習のための 3 つ組データが存在する言語ペアや領域は非常に限られている．

本研究では，従来の MNMT 用学習データ（3 つ組データ）に比べて比較的入手が容易な対訳テキストデータと，原言語側の画像キャプションデータから MNMT の学習を行う方法を提案する．提案手法では，まず対訳テキストデータから NMT モデルを学習し，画像キャプションデータの原言語文を NMT モデルで翻訳することで初期疑似 3 つ組データを生成する．次に，MNMT モデルと，対訳文ペアから画像を生成する text-to-image（T2I）モデルの 2 つのモデルを，初期擬似 3 つ組データから学習し，両モデルを初期化する．最後に，T2I モデルと MNMT モデルを逆翻訳形式のフレームワークを用いて交互に再学習する．このフレームワークでは，MNMT モデルは対訳テキストデータと T2I モデルによって生成された画像による疑似 3 つ組データで学習し，T2I モデルは画像キャプションデータと MNMT モデルによって生成された目的言語文による疑似 3 つ組データで学習する．

実験では，学習データとして Multi30k データセット [1] の英独対訳テキストデータと MSCOCO データセット [2] の画像キャプションデータを用いた．そして，テストデータとして Multi30k テストデータセットを用いて，英独翻訳タスクで提案手法の評価を行った．その結果，提案の MNMT モデルは入力画像を用いない NMT モデルよりも優れた翻訳性能を持つことを確認し（+1.38 BLEU スコア），提案する逆翻訳形式の学習方法は MNMT の翻訳性能を向上させる（+2.8 BLEU スコア）ことが示された．また，実験を通じて，提案の学習方法により訓練された MNMT は，真の 3 つ組データ（Multi30K 訓練データセットの 3 つ組データ）から訓練された MNMT モデルよりも優れていることが示された．さらに，WMT14 データセットおよび GoodNews データセット [3] を用いた事前学習を行った結果，翻訳精度がさらに改善される（+1.07 BLEU スコア）ことが示された．

2 関連技術

2.1 Transformer ベースの MNMT

近年，ニューラルネットワークをベースとしたモデルがマルチモーダル機械翻訳のためによく用いられており，特に Transfomer NMT モデル [4] をマルチモーダル機械翻訳に拡張した Transformer ベースの MNMT モデル [5] が非常に高い性能を実現している．本研究でも Transformer ベースの MNMT モデルを使用する．

本研究で使用する Transformer ベースの MNMT モデルの構造を図 1 に示す．このモデルは，Transformer NMT モデルに入力画像用のエンコーダが追加され，画像エンコーダ，テキストエンコーダ，テキストデコーダで構成されている．また，テキストエンコーダには，画像特徴と原言語文の各単

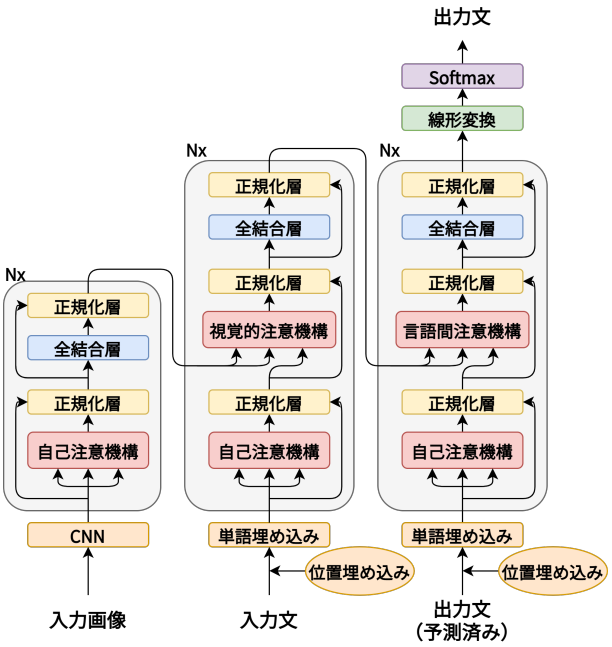

図 1: MNMT モデルの構造

語との関係の強さを計算する視覚的注意機構 [6] が組み込まれている．画像エンコーダは，まず入力画像から CNN を用いて画像特徴量を抽出し，その後，線形変換を施すことで画像を画像特徴ベクトルにエンコードする．なお，本研究では CNN として ResNet50 [7] を用いた．テキストエンコーダとテキストデコーダは，テキストエンコーダ内の各レイヤーが，画像と原言語文の各単語との間のマルチヘッドアテンション（視覚的注意機構）を有することを除いて，Transformer NMT と同じである．

2.2 Text-to-Image モデル

Text-to-Image（T2I）モデルは，入力として文およびランダムノイズを受け取り，受けとった文の意味に沿った本物に近い画像を生成するモデルである．ランダムノイズは，背景やオブジェクトの位置，向きなどの，文には現れない情報を決定するために入力される．T2I モデルは敵対的生成ネットワークを用いて学習され，主に画像を生成する生成器と，画像が本物であるかを識別する識別器の 2 つのモデルで構成されている．

従来の T2I モデルでは，1 つの文から 1 つの画像を生成する．しかし，本提案手法では対訳テキスト（原言語文と目的言語文）から 1 つの画像を生成する T2I モデルを用いる．具体的には，最先端の T2I モデルの 1 つである AttnGAN モデル [8] をバイリンガルな設定（対訳文ペアを入力にするモデル）に拡張する．本研究では，AttnGAN のテキストエンコーダーと注意機構を改良し，AttnGAN をバイリンガルな設定に拡張する．以降では，改良した AttnGAN を BiAttnGAN と呼ぶ．BiAttnGAN では，原言語文と目的言語文のそれぞれに対してエンコーダと注意機構を導入し，これら 2 つのエンコーダと注意機構の出力をそれぞれ連結したものを生成器と識別器で用いる．具体的には，原言語／目的言語文エンコーダ $Enc_{src/tgt}$ は，以下の式のように原言語／目的言語文 $x_{src/tgt}$ を単語特徴量 $e_{src/tgt}$ と文特徴量 $\bar{e}_{src/tgt}$ に符号化する．

$$e_{src}, \bar{e}_{src} = Enc_{src}(x_{src})$$
$$e_{tgt}, \bar{e}_{tgt} = Enc_{tgt}(x_{tgt})$$

そして，2 つの特徴量を連結したもの（$[e_{src}; e_{tgt}]$ や $[\bar{e}_{src}; \bar{e}_{tgt}]$）をテキスト特徴量として用いる．また，以下のように注意機構を用いて，画像とテキストとの関連性を反映した画像特徴量 h' を用いる．

$$h' = [Attn_{src}(h, e_{src}, e_{src}),\ Attn_{tgt}(h, e_{tgt}, e_{tgt})]$$

ここで，h と $Attn$ は，それぞれテキストエンコーダの隠れ状態と注意機構である．

3 MNMT のための逆翻訳学習

本節では，対訳テキストデータ $B = (B_{src}, B_{tgt})$ と，原言語側の画像キャプションデータ $C = (C_{img}, C_{src})$ から MNMT モデルを学習する手法を提案する．以降は，接尾辞の src, tgt, img は，それぞれ原言語文，目的言語文，画像を表す．提案手法の流れをアルゴリズム 1 に示す．本手法では，まず対訳テキストデータから NMT モデルを学習し，学習した NMT によって原言語側の画像キャプションデータのキャプション文を翻訳することで，初期疑似 3 つ組データを生成する（1 行目）．次に，生成した初期擬似 3 つ組データを用いて MNMT モデルと T2I モデルの初期化を行う（2 行目）．最後に，MNMT モデルと T2I モデルを交互に反復逆翻訳フレームワークを用いて際学習する（3 から 5 行目）[1]．

3.1 モデルの初期化

逆翻訳形式による学習の準備として，Transformer ベースの MNMT モデルと BiAttnGAN モデルの初期化を行う．これらの初期化に用いる 3 つ組データは，Transformer NMT を用いて擬似的に作成する．

1） 実験では，アルゴリズム 1 の 3 行目における N の値は 15 に設定した．

アルゴリズム 1：学習アルゴリズム

入力： $B = (B_{src}, B_{tgt})$, $C = (C_{src}, C_{img})$

1. **初期擬似 3 つ組データの生成：** まず，NMT モデル $P_{src \to tgt}$ を B から学習する．その後，3 つ組データ $(C_{src}, C_{tgt'}, C_{img})$ を生成する．ただし $C_{tgt'} = P_{src \to tgt}(C_{src})$ である．

2. **モデルの初期化：** MNMT モデル $P^{(0)}_{(src,img) \to tgt}$ と T2I モデル $P^{(0)}_{(src,tgt) \to img}$ を初期擬似 3 つ組データ $(C_{src}, C_{tgt'}, C_{img})$ を用いて学習する．

3. **for k=1 to N do**

4. **MNMT の再学習：** MNMT モデル $P^{(k)}_{(src,img) \to tgt}$ を擬似 3 つ組データ $(B_{src}, B_{img'}, B_{tgt})$ を用いて再学習する．
   ただし $B_{img'} = P^{(k-1)}_{(src,tgt) \to img}(B_{src}, B_{tgt})$ である．

5. **T2I の再学習：** T2I モデル $P^{(k)}_{(src,tgt) \to img}$ を擬似 3 つ組データ $(C_{src}, C_{img}, C_{tgt'})$ を用いて再学習する．
   ただし $C_{tgt'} = P^{(k-1)}_{(src,img) \to tgt}(C_{src}, C_{img})$ である．

6. **end**

まず，対訳テキストデータから Transformer NMT モデルを学習する．そして，学習させた NMT モデルを用いて，原言語側の画像キャプションデータのキャプション文を目的言語の文に翻訳する．このようにして作成した初期擬似 3 つ組データを用いて，MNMT モデルおよび BiAttnGAN モデルを学習することで，両モデルの初期化を行う．

3.2 MNMT の再学習

MNMT モデルの再学習では，BiAttnGAN モデルを用いて再学習を行う．まず，BiAttnGAN モデルを用いて，対訳テキストデータの各対訳文ペアから画像を生成する．次に，対訳テキストデータと生成した画像で構成される疑似 3 つ組データから MNMT モデルを学習する．学習では，原言語文と生成画像から予測された目的言語文 y が疑似 3 つ組データの目的言語文 t と同じになるように，以下のクロスエントロピー損失 L_M を最小化する．

$$L_M = -\sum_{i=0}^{l-1} t_i \times \log P(y_i)$$

ここで，l は目的言語文の長さである．

3.3 T2I の再学習

BiAttnGAN モデルの再学習では，MNMT モデルを用いて再学習を行う．まず，MNMT モデルを用いて，画像キャプションデータから目的言語文を生成する．次に，画像キャプションデータと生成した疑似目的言語文で構成される擬似 3 つ組データから BiAttnGAN モデルを学習する．

学習では，原言語文と擬似目的言語文から生成された画像 y_{img} と，本物画像 t_{img} が同じになるように，次のクロスエントロピー損失 L_G を最小化する．

$$L_G = -\frac{1}{2}\log P_R(y_{img}) - \frac{1}{2}\log P_S(y_{img}, x_{src})$$

ここで，P_R は生成された画像が本物かどうかを表す確率であり，P_S は生成された画像と文が一致するかどうかを表す確率である．

4 実験

4.1 実験設定

本提案手法を英独翻訳による実験で評価した．実験では，Multi30k データセット [1] の英独対訳文 29,000 文対と，画像ごとに 5 つのキャプションが付与されている MS COCO 2014 データセット [2] の 82,783 枚の画像とそのキャプションを，2 種類の学習データセット（対訳テキストデータと画像キャプションデータ）として使用した．Multi30k データセットの開発データ (1,014 組) とテストデータ (1,000 組) をそれぞれ開発データとテストデータとして使用した．Multi30k データセットの各データは原言語文，目的言語文，画像の 3 つ組で構成されるが，提案手法の学習では目的言語文は使用していないことに注意されたい．

また，提案手法を事前学習に用いる場合の有効性を調べるため，Multi30k データセットとは異なるドメインの大規模データセットを用いて MNMT モデルを事前学習した場合の性能も評価した．事前学習では，対訳テキストデータとして WMT14 英独データセットを，画像キャプションデータセットとして GoodNews データセットを用いた．

初期擬似 3 つ組データを生成する際に用いる Transformer NMT モデルのハイパーパラメータと，Transformer ベースの MNMT モデルのハイパーパラメータは，Vaswani ら [4] に倣い，レイヤー数を 6 層，注意機構のヘッド数を 8 個，隠れ層の次元数

表 1: 実験結果

モデル	BLEU (%)
NMT	38.18
$MNMT_{init}$	36.76
$MNMT_{prop}$	39.56
$MNMT_{gold}$	38.54
$MNMT_{pre\_un}$	40.63
$MNMT_{pre\_semi}$	42.18

を 512 に設定した．また，BiAttnGAN のハイパーパラメータに関しては，オリジナルの AttnGAN [8] に倣い，生成器の次元数を 48，識別器の次元数を 96 とした．最適化手法としては Adam [9] を使用した．BiAttnGAN モデルは，ミニバッチサイズ 32 とし，エポック数は初期化時は 100，再学習時は 15 で実験を行った．Transformer NMT モデルは，ミニバッチ数を 128，エポック数を 40 として学習を行った．そして，Trasformer MNMT モデルは，ミニバッチ数を 128，エポック数は初期化時は 25，再学習時は 15 で学習を行った．また，単語辞書の作成には Byte Pair Encoding [10] を使用し，トークンサイズを英独合わせて 7,000 とした．なお，これらの翻訳モデルを用いた目的言語文の推論時には貪欲法を用いた．

4.2 実験結果

実験では，次の 6 つのモデルを評価した．(1) 提案 MNMT モデル（$MNMT_{prop}$），(2) 画像入力なし NMT モデル（NMT），(3) 初期化時の MNMT モデル（$MNMT_{init}$），(4) 真の 3 つ組データを用いた MNMT モデル（$MNMT_{gold}$），(5) 大規模データセットによる事前学習を行い，擬似データを用いて教師無し再学習を行った MNMT モデル（$MNMT_{pre\_un}$），(6) 大規模データセットによる事前学習を行い，真の 3 つ組データを用いて教師あり再学習を行った MNMT モデル（$MNMT_{pre\_semi}$）である．$MNMT_{prop}$，NMT，$MNMT_{init}$，$MNMT_{gold}$ では事前学習を行っていない．また，NMT は，Multi30k データセットの学習データの画像を使わずに対訳文から学習したモデルである．$MNMT_{gold}$ は，Multi30k データセットの学習データに含まれる 29,000 組の 3 つ組データから学習した MNMT モデルであり，$MNMT_{init}$ は，初期疑似 3 つ組データで学習した MNMT モデル（アルゴリズム 1 の $P^{(0)}_{(src,img)\to tgt}$）である．

各モデルの翻訳性能は，開発データの BLEU スコアが最も高いエポックモデルを選択し，テストデータの case-insensitive BLEU4 [11] で測定した．実験結果を表 1 に示す．

表 1 から分かる通り，$MNMT_{prop}$ は $MNMT_{init}$ よりも高い性能を示している．このことは，擬似 3 つ組データを用いて MNMT モデルと T2I モデルを交互に学習することで，MNMT の性能が向上することを示している．すなわち，提案手法である逆翻訳形式のフレームワークが有効であることを示している．また，$MNMT_{prop}$ は NMT よりも性能が優れており，画像情報の有効性が示されている．さらに，真の 3 つ組データを用いた $MNMT_{gold}$ よりも，擬似 3 つ組データを用いた $MNMT_{prop}$ の方が性能が高くなっている．これは擬似画像生成による多様な画像によって学習が行われたためであると考えられる．さらに，$MNMT_{pre\_un}$ および $MNMT_{pre\_semi}$ は事前学習を行わないモデルによりも性能が向上しており，提案手法による大規模データセットを用いた事前学習が有効であることを示している．

5 まとめ

本研究では，マルチモーダル機械翻訳における低リソース問題を解決するために，対訳テキストデータと画像キャプションデータから MNMT モデルを学習するための新しい逆翻訳形式のフレームワークを提案した．提案手法では，T2I モデルと MNMT モデルを用いて，一方のモデルにより生成された疑似 3 つ組データに基づいて他方のモデルを学習することで，T2I モデルと MNMT モデルを交互に学習する．英独翻訳タスクでの評価の結果，提案した逆翻訳形式のフレームワークは MNMT の性能を向上させ，提案手法によって学習された MNMT モデルは，真の 3 つ組データから学習した MNMT モデルよりも性能が優れていることが示された．また，提案手法による大規模データセットを用いた事前学習が有効であることも示された．今後は異なる規模やドメイン，言語対のデータセットを用いた実験を行い，提案手法の有効性を確認したい．

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．また，本研究の一部は JSPS 科研費 20K19864 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] 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 (VL16)*, pp. 70–74, 2016.

[2] 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.

[3] Ali Furkan Biten, Lluis Gomez, Marcal Rusinol, and Dimosthenis Karatzas. Good news, everyone! context driven entity-aware captioning for news images. In *Proc. of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, June 2019.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pp. 5998–6008, 2017.

[5] Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, and Raúl Vázquez. The MeMAD submission to the WMT18 multimodal translation task. In *Proc. of the Third Conference on Machine Translation: Shared Task Papers(WMT18)*, pp. 603–611, October 2018.

[6] 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.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pp. 770–778, 2016.

[8] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proc. of 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, pp. 1316–1324, 2018.

[9] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, Vol. abs/1412.6980, , 2014.

[10] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, jul 2002.