Vol.2020-NL-246 No.6
Vol.2020-SLP-134 No.6
2020/12/2

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機械翻訳

岩本 裕司[1,a)]　田村 晃裕[2,b)]　二宮 崇[1,c)]

概要：近年，機械翻訳の際に，原言語文に加えて関連画像情報を利用することで，翻訳精度の向上を図る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．しかし，このような MNMT モデルの学習には，原言語文，目的言語文，関連画像で構成される 3 つ組データが必要となり，データ数不足が問題となっている．そこで，本研究では 3 つ組データを必要とせず，対訳テキストデータと画像キャプションデータを用いて MNMT モデルを学習する新たな手法を提案する．本手法の大まかな流れとしては，まず画像キャプションデータの各文を対訳テキストデータから訓練されたニューラル機械翻訳 （Neural Machine translation; NMT）モデルで翻訳し，擬似 3 つ組データを作成する．そして，作成した擬似 3 つ組データを用いて，対訳文ペアから画像を生成する text-to-image モデルと MNMT モデルを初期化する．その後，テキストから画像へ変換する text-to-image モデルと MNMT モデルを，逆翻訳形式のフレームワークで交互に繰り返し学習する．本研究の実験では，対訳テキストデータとして Multi30k データセットを，画像キャプションデータとして MSCOCO データセットを使用した．結果として英独翻訳タスクの評価では，提案した MNMT モデルは画像入力なしの NMT モデルよりも優れており（+1.18 BLEU スコア），また，提案した反復逆翻訳学習方式は初期の MNMT モデルの性能を向上させる（+6.93 BLEU スコア）ことが示された．

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．

一部の研究では，このような 3 つ組データを必要としない教師なし MNMT モデルを提案している [1], [2], [3]．これらの研究では，2 つの独立した画像キャプションデータ（原言語キャプションデータと目的言語キャプションデータ）から MNMT モデルの学習を行う．画像情報を原言語空間と目的言語空間の間のピボットとして利用しているが，原言語空間と目的言語空間の間のアラインメント情報は教師として与えられるのではなく自動的に学習されるため，原言語空間と目的言語空間の整合性は保証されない．そこで，本研究では，既存の対訳テキストデータは，既存の画像キャプションデータよりも言語や領域の多様性が高く大規模であることを考慮し，MNMT の教師なし学習における対訳テキストデータの利用に焦点を当てる．

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

[1] 愛媛大学
[2] 同志社大学
[a)] iwamoto@ai.cs.ehime-u.ac.jp
[b)] aktamura@mail.doshisha.ac.jp
[c)] ninomiya@cs.ehime-u.ac.jp

Vol.2020-NL-246 No.6
Vol.2020-SLP-134 No.6
2020/12/2

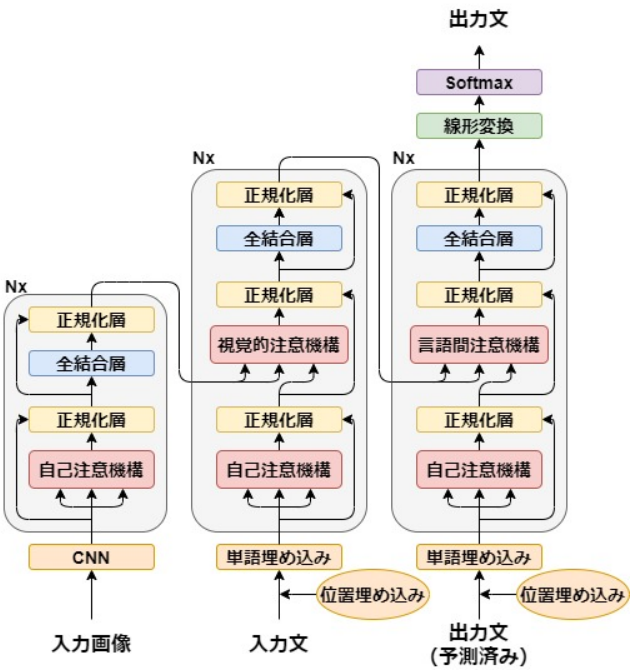


図 1: MNMT モデルの構造

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 [4] の英独対訳テキストデータと MSCOCO データセット [5] 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（+1.18BLEU 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（+6.93BLEU 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．

2. 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 [6] 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 [7] 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語との関係の強さを計算する視覚的注意機構 [8] 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 [9] 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。

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**

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，原言語側の画像キャプションデータを学習した NMT によって翻訳することで，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 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

[*1] 実験では，アルゴリズム 1 の 3 行目における N の値は 20 に設定した．

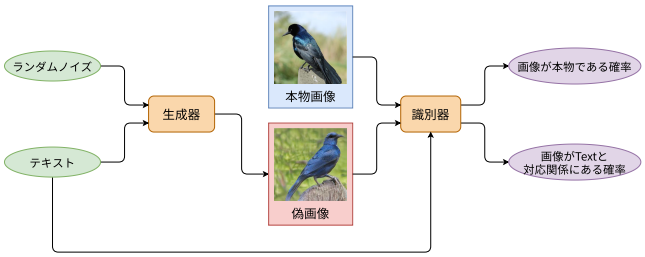


図 2: Text-to-Image モデルの基本構造

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 モデルの基本的な構造図を**図 2** に示す．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 G と，画像が本物であるかを識別する識別器 D の 2 つのモデルで構成されている．

生成器 G は以下のように，入力として文 T_{true} とランダムノイズ z を受け取り，偽画像 I_{fake} を生成する．

$$I_{fake} = G(T_{true}, z)$$

一方，識別器 D は以下のように，画像 I を受け取り画像 I が本物である確率 P_R を出力する．

$$P_R(I) = D(I)$$

また，識別器 D は以下のように，画像 I と文 T を受け取り，画像 I と文 T が対応関係にある確率 P_S も出力する．

$$P_S(I, T) = D(I, T)$$

生成器 G は識別器 D に対し，自身が生成した偽画像 I_{fake} が本物であり，文 T_{true} と対応関係にあると識別させるように学習が行われる．すなわち，以下の式 (1)(2) の確率がそれぞれ最大となるように学習が行われる．

$$P_R(I_{fake}) = D(I_{fake}) \tag{1}$$

$$P_S(I_{fake}, T_{true}) = D(I_{fake}, T_{true}) \tag{2}$$

よって，生成器 G の学習誤差 L_G は，以下の式 (3) に示すような 2 値交差エントロピーにより算出される．

$$L_G = -\frac{1}{2}\log P_R(I_{fake}) - \frac{1}{2}\log P_S(I_{fake}, T_{true}) \tag{3}$$

一方で識別器 D は，本物画像 I_{real} を本物，生成器が生成した偽画像 I_{fake} を偽物と識別するように学習が行われる．すなわち，以下の式 (4) の確率が最大，式 (5) の確率が最小となるように学習が行われる．

$$P_R(I_{real}) = D(I_{real}) \tag{4}$$

$$P_R(I_{fake}) = D(I_{fake}) \tag{5}$$

また，識別器 D では画像 I に対して，文 T_{true}, T_{false} がそれぞれ対応関係にあるかどうかを正しく識別させるように学習が行われる．すなわち，以下の式 (6) の確率が最大，式 (7) の確率が最小となるように学習が行われる．

$$P_S(I, T_{true}) = D(I, T_{true}) \tag{6}$$

$$P_S(I, T_{false}) = D(I, T_{false}) \tag{7}$$

よって，識別器 D の学習誤差 L_D は，以下の式 (8) に示すような 2 値交差エントロピーにより算出される．

$$\begin{aligned} L_D = &-\frac{1}{2}\log P_R(I_{real}) - \frac{1}{2}\log\left(1 - P_R(I_{fake})\right) \\ &-\frac{1}{2}\log P_S(I, T_{true}) - \frac{1}{2}\log\left(1 - P_S(I, T_{false})\right) \end{aligned} \tag{8}$$

このように，T2I モデルでは 2 つのモデルを互いに競い合わせることで，より文の意味に沿った本物に近い画像が生成されるように学習が行われる．

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 [10] 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．

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

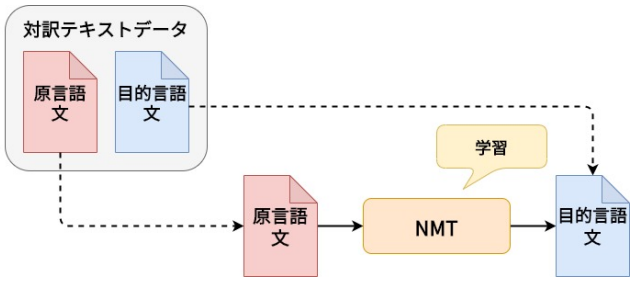
そして，2 つの特徴量を連結したもの（$[e_{src}; e_{tgt}]$ や $[\bar{e}_{src}; \bar{e}_{tgt}]$）をテキスト特徴量として用いる．また，以下のように注意機構を用いて，画像とテキストとの関連性を反映した画像特徴量 h' を用いる．

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

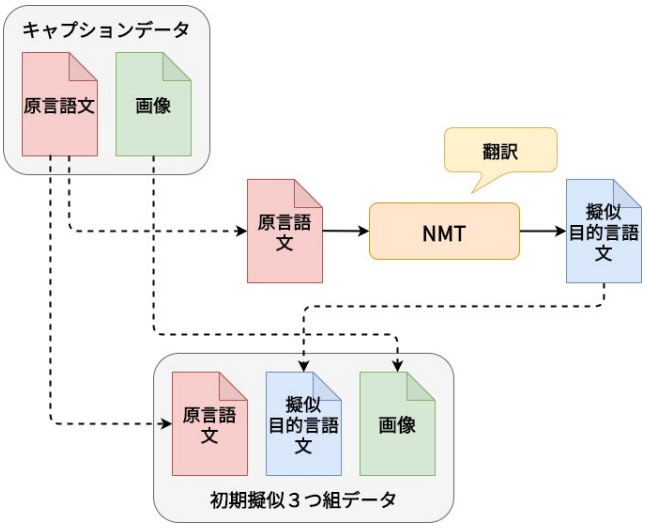
ここで，h と $Attn$ は，それぞれテキストエンコーダの隠れ状態と注意機構である．

3.2 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．その流れを**図 3** に示



(a) Transformer NMT モデルの学習



(b) 擬似目的言語文の生成

図 3: 初期擬似 3 つ組データ作成の流れ

す．まず，対訳テキストデータから Transformer NMT モデルを学習する（**図 3a**）．そして，学習させた NMT モデルを用いて，画像のキャプションデータを目的言語の文に翻訳する（**図 3b**）．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．

3.3 MNMT の再学習

BiAttnGAN モデルを用いて，MNMT モデルの再学習を行う際のモデル図を**図 4** に示す．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擬似目的言語文 y が目的言語文 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。

3.4 T2I の再学習

MNMT モデルを用いて，BiAttnGAN モデルの再学習

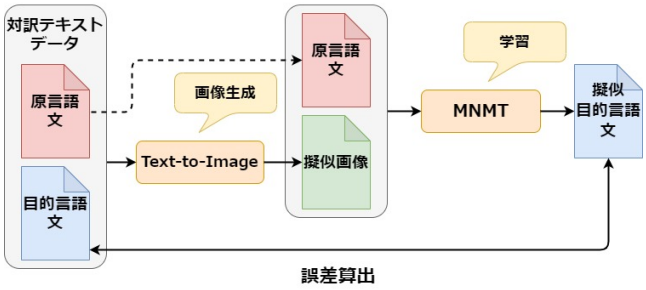


図 4: MNMT モデルの再学習

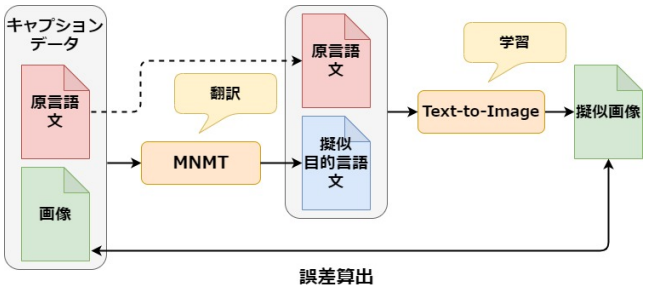


図 5: BiAttnGAN モデルの再学習

を行う際のモデル図を**図 5** に示す．まず，MNMT モデルを用いて，画像キャプションデータから擬似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 [4] の英独対訳文 29,000 文と，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 [5] 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．

初期擬似 3 つ組データを生成する際に用いる Transformer

表 1: 実験結果

モデル	BLEU (%)
NMT	37.98
$MNMT_{init}$	32.23
$MNMT_{prop}$	39.16
$MNMT_{gold}$	38.29

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 [6] に倣い，レイヤー数を 6 層，注意機構のヘッド数を 8 個，隠れ次元を 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 [10] 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 [11] を使用した．BiAttnGAN モデルは，ミニバッチサイズ 20 とし，エポック数は初期化時は 100，再学習時は 20 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 20 で学習を行った．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，以下の 4 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（$MNMT_{gold}$）

提案モデルと比較するベースラインの 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．また，$MNMT_{gold}$ 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 [12] 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の裏付けとなっている．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．このことについては，5 節で考察する．

図 6: 本物画像

図 7: BiAttnGAN モデルが生成した偽画像

5. 考察

本節では，$MNMT_{prop}$ が $MNMT_{gold}$ よりも性能が高くなった理由について考察する．**図 6** に Multi30k データセットに含まれる本物画像を，**図 7** に BiAttnGAN モデルが生成した偽画像の 1 例を示す．図 7 は，図 6 の画像に対する原言語文 “group of people walking on the heavy snow.” および目的言語文 “eine gruppe geht durch den tiefschnee.” を入力として，本手法で学習された BiAttnGAN モデルが生成した偽画像である．

真の 3 つ組データを用いて学習する場合，毎エポックごとに図 6 のよう同一の画像を用いて学習が行われることになる．そのため，モデルは多様な分布であるはずの画像空間の，ごく一部分しか学習することができない可能性がある．一方で，擬似 3 つ組データを用いて学習する場合，BiAttnGAN モデルはランダムノイズによって生成する画像の背景やオブジェクトの配置を変えるため，エポックごとに図 7 のような多種多様な画像が用いられる．そのため，画像生成を利用した擬似 3 つ組データを用いて学習したモデルの性能が，真の 3 つ組データを用いて学習したモデルに比べて向上したと考えられる．

6. 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 モデルを交互に学習し，もう一方のモデルを利用して生成された疑似 3 つ組データに基づいて学習を行うことで，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．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] Nakayama, H. and Nishida, N.: Zero-resource machine translation by multimodal encoder-decoder network with multimedia pivot, *Machine Translation*, Vol. 31, No. 1-2, pp. 49–64 (2017).

[2] Chen, Y., Liu, Y. and Li, V. O.: Zero-Resource Neural Machine Translation with Multi-Agent Communication Game, *Proc. of the Thirty-Second AAAI Conference on Artificial Intelligence(AAAI-18), the 30th innovative Applications of Artificial Intelligence(IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18)* (McIlraith, S. A. and Weinberger, K. Q., eds.), pp. 5086–5093 (2018).

[3] Huang, P.-Y., Hu, J., Chang, X. and Hauptmann, A.: Unsupervised Multimodal Neural Machine Translation with Pseudo Visual Pivoting, *CoRR*, Vol. abs/1207.0016 (2020).

[4] Elliott, D., Frank, S., Sima'an, K. and Specia, L.: Multi30K: Multilingual English-German Image Descriptions, *Proc. of the 5th Workshop on Vision and Language (VL16)*, pp. 70–74 (2016).

[5] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L.: Microsoft coco: Common objects in context, *European conference on computer vision*, Springer, pp. 740–755 (2014).

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. and Polosukhin, I.: Attention is All you Need, *Advances in Neural Information Processing Systems 30*, pp. 5998–6008 (online), available from ⟨http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf⟩ (2017).

[7] Grönroos, S.-A., Huet, B., Kurimo, M., Laaksonen, J., Merialdo, B., Pham, P., Sjöberg, M., Sulubacak, U., Tiedemann, J., Troncy, R. and Vázquez, R.: The MeMAD Submission to the WMT18 Multimodal Translation Task, *Proc. of the Third Conference on Machine Translation: Shared Task Papers(WMT18)*, pp. 603–611 (online), DOI: 10.18653/v1/W18-6439 (2018).

[8] 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).

[9] He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, *Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pp. 770–778 (2016).

[10] Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X. and He, X.: AttnGAN: Fine-Grained Text to Image Generation With Attentional Generative Adversarial Networks, *Proc. of 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, pp. 1316–1324 (2018).

[11] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, *CoRR*, Vol. abs/1412.6980 (2014).

[12] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: Bleu: a Method for Automatic Evaluation of Machine Translation, *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318 (online), DOI: 10.3115/1073083.1073135 (2002).