自動生成した画像による
教師なしマルチモーダルニューラル機械翻訳

岩本 裕司

愛媛大学 大学院理工学研究科
電子情報工学専攻

iwamoto@ai.cs.ehime-u.ac.jp

概要

近年，機械翻訳の際に，原言語文（翻訳元言語の文）に加えて翻訳に関連する画像を利用することで，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（MNMT）が注目されている．しかし，MNMT モデルの学習には，原言語文，目的言語文（翻訳先言語の文），関連画像で構成される 3 つ組データが必要となり，学習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，3 つ組データより用意しやすい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を用いて MNMT モデルを学習する新たな手法を提案する．提案手法では，まず対訳テキストデータからニューラル機械翻訳モデル（NMT）を学習し，学習した NMT モデルで画像キャプションデータの各文を翻訳することで，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，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．具体的には，MNMT モデルは，T2I モデルによって生成された画像と対訳テキストデータによる疑似 3 つ組データで学習し，T2I モデルは，MNMT モデルによって生成された目的言語文と画像キャプションデータによる疑似 3 つ組データで学習する．提案手法の有効性を，対訳テキストデータとして Multi30k データセット，画像キャプションデータとして MSCOCO データセットを用いた英独翻訳タスクで検証した．その結果，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（+1.38 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（+2.8 BLEU スコア）ことを確認した．さらに，大規模なデータセットを用いた事前学習により，さらなる翻訳精度の向上が可能であることを実験的に示した（+1.07 BLEU スコア）．

目次

第 1 章　はじめに

近年，機械翻訳の分野において，ニューラルネットワークを用いた機械翻訳（Neural Machine translation; NMT）が注目を集めている．NMT は従来のルールベース機械翻訳や統計的機械翻訳に比べて高い翻訳精度を実現しており，様々な手法が研究・提案されている．中でも，自己注意機構という文内の単語間の関連を捉える機構を備える Transformer モデル [1] は，これまでの再帰型ニューラルネットワーク（Recurrent Neural Network; RNN）を用いた NMT モデル [2] の性能を上回り， 機械翻訳のデファクトスタンダードとなっている．

このような NMT の性能を向上させる手段の一つとして，マルチモーダル学習がある．マルチモーダル学習とは，単一のモダリティではなく複数のモダリティを統合的に処理することで，性能の向上を図る手法である．NMT においては，翻訳元の文（原言語文）だけでなく関連画像も用いて翻訳先の文（目的言語文）を予測するマルチモーダルニューラル機械翻訳（Multimodal Neural Machine translation; MNMT）が注目されており，翻訳性能の向上が期待されている [3]．しかし，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 つ組データは通常の対訳データに比べて非常に小規模なものしか存在していない．例えば，英語とドイツ語（英独）や英語と日本語（英日）の言語対の場合，関連画像を含まない対訳データであれば，数千万レベルの対訳文対のデータが存在して容易に利用できる．一方で，3 つ組データの場合，数十万レベルのデータしか存在しない．また，通常の対訳データに比べて，MNMT 学習のための 3 つ組データが存在する言語対や領域は非常に限られている．

この学習データ不足の問題を解決するため，近年，このような 3 つ組データを必要としない教師なし MNMT モデルが提案されている [4, 5, 6, 7]．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（原言語側のキャプションデータと目的言語側のキャプションデータ）から MNMT モデルの学習を行う．これらの従来手法では，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するが，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に着目し，MNMT の教師なし学習において対訳テキストデータを利用する．対訳テキストデータを用いることで，従来の教師なし MNMT より，適用範囲が広がるとともに，原言語空間と目的言語空間のアラインメント情報を教師として与えることができる．

本研究では，従来の MNMT 用学習データ（3 つ組データ）や従来の教師なし MNMT 用学習データ（2 種の画像キャプションデータ）に比べて入手が容易な，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．原言語側の画像キャプションデータと対訳テキストデータに基づく教師なし MNMT モデルはこれまで提案されておらず，本研究が初めての試みであることを特筆しておく．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 T2I モデルによって生成された画像と対訳テキストデータによる疑似 3 つ組データで学習し，T2I モデルは MNMT モデルによって生成された目的言語文と画像キャプションデータによる疑似 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 [8] の英独対訳テキストデータと MSCOCO データセット [9] 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（+1.38BLEU 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（+2.8BLEU スコア）ことを確認した．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を示した．さらに，大規模データセットである WMT14 データセットおよび GoodNews データセット [10] を用いて事前学習を行い，教師なしおよび半教師あり実験を行った．その結果，事前学習によりさらに翻訳性能が向上することを実験的に示した（+1.07BLEU スコア）．

第 2 章　関連研究

本章では関連研究として，2.1 節で現在の NMT のデファクトスタンダードである Transformer モデルについて述べ，2.2 節で Transformer モデルに基づいた MNMT モデルについて述べる．そして，2.3 節では T2I モデルについて述べる．

2.1　Transformer モデル

Transformer モデル [1] は，入力として受け取った原言語文から目的言語文を予測，生成する NMT モデルである．Transformer モデルの基本的な構造図を図 2.1 に示す．Trasformer モデルは原言語文を受け取り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り目的言語文を予測するデコーダから成るエンコーダ・デコーダモデルで実現されている．

エンコーダは主に単語埋め込みと N 層スタックされたエンコーダレイヤで構成されている．エンコーダレイヤは自己注意機構および全結合層の 2 つのサブレイヤで構成されており，各サブレイヤの後には残差接続および正規化が適用される．残差接続は深い層で構成されるニューラルネットワークの学習を促進するための手法であり，勾配消失問題の影響を軽減できることが知られている．また，正規化層では各レイヤの入力分布を一定にすることで，学習の安定性および速度を向上できることが知られている．

一方，デコーダは主に単語埋め込みと N 層スタックされたデコーダレイヤ，そして線形変換層と Softmax 層で構成されている．デコーダレイヤは自己注意機構および言語間注意機構，全結合層の 3 つサブレイヤで構成されており，エンコーダレイヤと同じく各サブレイヤの後には残差接続および正規化が適用される．

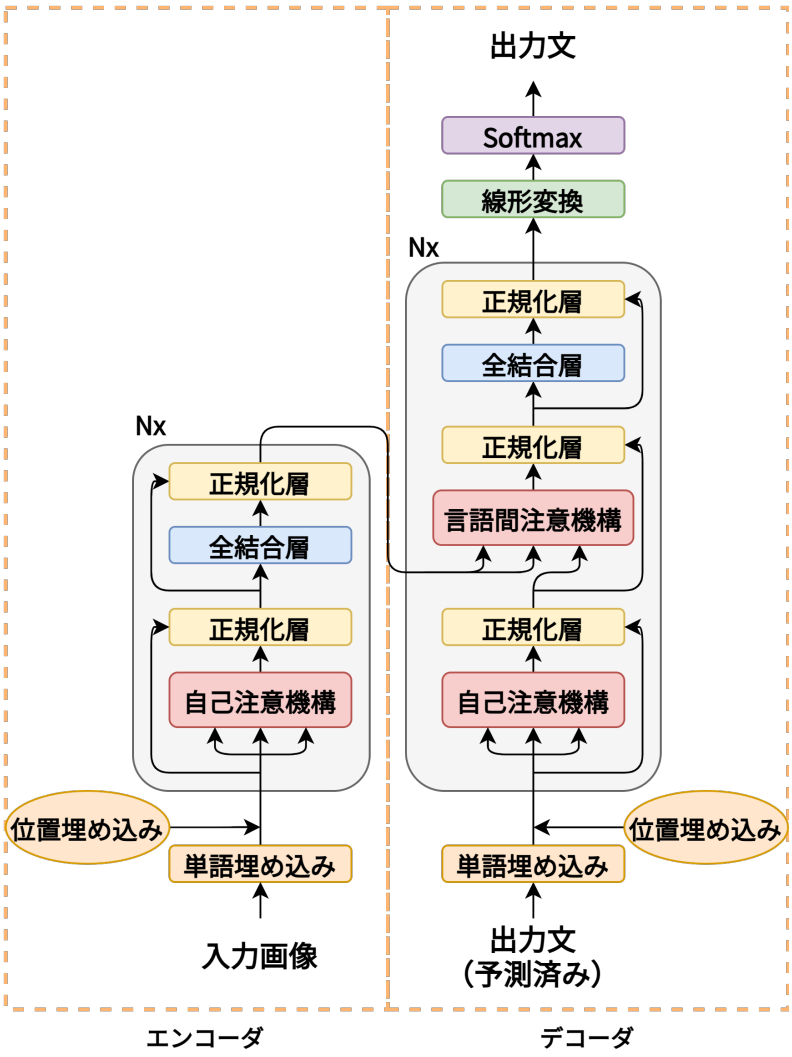


図 2.1: Transformer モデルの基本構造

エンコーダおよびデコーダでは，最初に単語埋め込み層によって各単語を分散表現に変換する．Transformer モデルは RNN のような再帰構造を備えておらず，このままでは各単語の位置情報を考慮できない．そこで，Transformer モデルでは，位置埋め込み層によって単語の位置情報を分散表現に加算する．このようにして獲得した単語の位置情報を含んだ分散表現ベクトルが，エンコーダレイヤやデコーダレイヤの入力となる．

エンコーダレイヤおよびデコーダレイヤには自己注意機構が，さらにデコーダレ

イヤには言語間注意機構が設けられており，各単語間の関連性の強さを考慮できる構造になっている．自己注意機構および言語間注意機構は，以下の式 (2.1) で表される．

$$Attn(Q,K,V) = softmax\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{2.1}$$

ここで，Q,K,V はエンコーダもしくはデコーダの隠れ状態を表し，d_k は Q,K,V の次元数を表す．エンコーダの自己注意機構では，Q,K,V の全てにエンコーダの隠れ状態が与えられ，デコーダの自己注意機構では，Q,K,V の全てにデコーダの隠れ状態が与えられる．一方で，デコーダの言語間注意機構では，Q にデコーダの隠れ状態が，K,V にエンコーダの隠れ状態が与えられる．結果として，自己注意機構では同一文内の各単語間の関連性が考慮され，言語間注意機構では原言語文内の各単語と目的言語文内の各単語間の関連性が考慮される．

Transformer モデルにおけるこれらの注意機構は，マルチ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．h 個のヘッドからなるマルチヘッドの注意機構は以下の式 (2.2) で表される．

$$\begin{aligned} MultiHead(Q,K,V) &= Concat(head_1,\ldots,head_h)W^O \\ head_i &= Attn(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{2.2}$$

ここで，$Concat$ はベクトルを結合する関数である．また，$W_i^Q \in \mathbb{R}^{d_{model}\times d_k}, W_i^K \in \mathbb{R}^{d_{model}\times d_k}, W_i^V \in \mathbb{R}^{d_{model}\times d_k}, W_i^O \in \mathbb{R}^{hd_k\times d_{model}}$ 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．

デコーダの最後には，線形変換層および Softmax 層が設けられており，線形変換層でデコーダの隠れ状態の次元数を目的言語の単語の種類数に変換する．そして，Softmax 層は各次元の値を確率値に変換することで，目的言語の全単語に対する出力確率を得る．翻訳する際は，この出力確率に基づき，入力された原言語文に対して確率が最大となる目的言語の単語系列を予測・生成する．

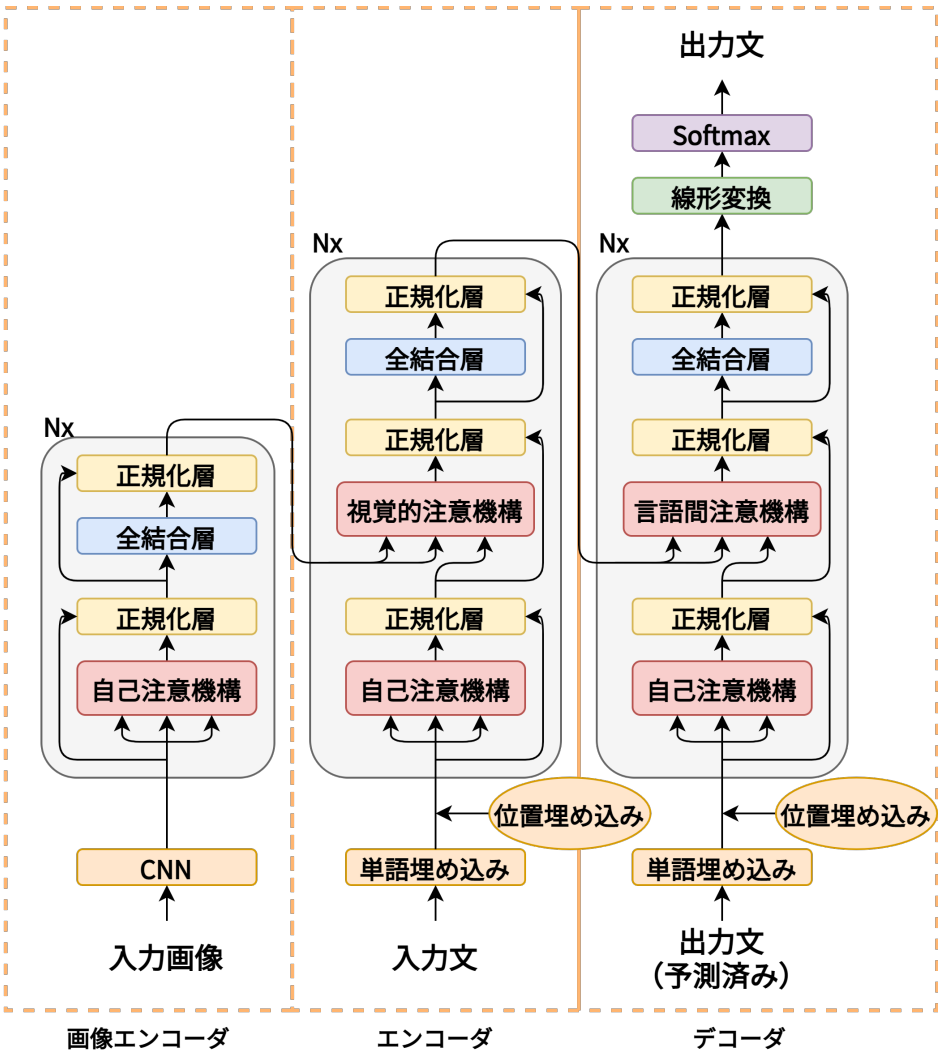


図 2.2: MNMT モデルの構造

2.2 Transformer ベースの MNMT モデル

マルチモーダル機械翻訳の主流はニューラルネットワークを用いた手法であり，特に Transfomer NMT モデル [1] 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 [11, 12] が非常に高い性能を実現している．本研究では Nishihara らの Transformer ベースの MNMT モデル [12] を使用する．本節では，その使用する MNMT モデルを概説する．

本研究で使用する Transformer ベースの MNMT モデルの構造を図 2.2 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．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 [13] を用いている．このモデルでは，自己注意機構，言語間注意機構，視覚的注意機構の 3 つの注意機構がある．視覚的注意機構も，式 (2.1) で表され，Q がテキストエンコーダの自己注意機構の出力であり，K, V は画像エンコーダの出力である．この視覚的注意機構により，単語と画像領域の関連性を考慮できる．

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．

2.3 Text-to-Image モデル

Text-to-Image（T2I）モデル [14, 15, 16, 17] 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 モデルの基本的な構造図を図 2.3 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 は以下の式 (2.3) 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．

$$I_{fake} = G(T_{true}, z) \tag{2.3}$$

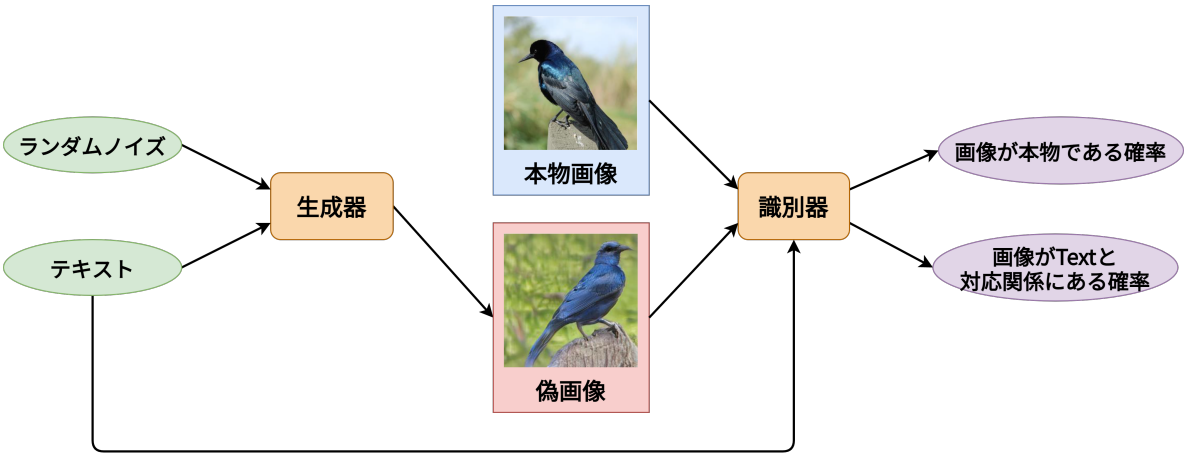


図 2.3: Text-to-Image モデルの基本構造

一方，識別器 D は以下の式 (2.4) のように，画像 I を受け取り画像 I が本物である確率 P_{real} を出力する．

$$P_{real}(I) = D(I) \tag{2.4}$$

また，識別器 D は以下の式 (2.5) 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_{match} も出力する．

$$P_{match}(I, T) = D(I, T) \tag{2.5}$$

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 (2.6), (2.7) の確率がそれぞれ最大となるように学習が行われる．

$$P_{real}(I_{fake}) = D(I_{fake}) \tag{2.6}$$

$$P_{match}(I_{fake}, T_{true}) = D(I_{fake}, T_{true}) \tag{2.7}$$

よって，生成器 G の学習誤差 L_G は，以下の式 (2.8) に示す 2 値交差エントロピー

により算出される．

$$
\begin{aligned}
L_G = &-\frac{1}{2}\log P_{real}(I_{fake}) \\
&-\frac{1}{2}\log P_{match}(I_{fake}, T_{true})
\end{aligned}
\tag{2.8}
$$

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 (2.9) の確率が最大，式 (2.10) の確率が最小となるように学習が行われる．

$$
P_{real}(I_{real}) = D(I_{real}) \tag{2.9}
$$

$$
P_{real}(I_{fake}) = D(I_{fake}) \tag{2.10}
$$

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 (2.11) の確率が最大，式 (2.12) の確率が最小となるように学習が行われる．

$$
P_{match}(I, T_{true}) = D(I, T_{true}) \tag{2.11}
$$

$$
P_{match}(I, T_{false}) = D(I, T_{false}) \tag{2.12}
$$

よって，識別器 D の学習誤差 L_D は，以下の式 (2.13) に示す 2 値交差エントロピーにより算出される．

$$
\begin{aligned}
L_D = &-\frac{1}{2}\log P_{real}(I_{real}) \\
&-\frac{1}{2}\log\left(1 - P_{real}(I_{fake})\right) \\
&-\frac{1}{2}\log P_{match}(I, T_{true}) \\
&-\frac{1}{2}\log\left(1 - P_{match}(I, T_{false})\right)
\end{aligned}
\tag{2.13}
$$

このように，T2I モデルでは生成器と識別器の 2 つのモデルを互いに競い合わせることで，より入力文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．

2.3.1 AttnGAN モデル

近年提案された最先端の T2I モデルの 1 つとして AttnGAN モデル [17] がある．AttnGAN モデルは従来の T2I モデルに対して注意機構を導入し，入力文と画像特徴量の関連性を考慮することで，性能の向上を実現したモデルである．以下，2.3.1.1 節で AttnGAN における生成器 G の構造を述べ，2.3.1.2 節で AttnGAN における識別器 D の構造を述べる．

2.3.1.1 生成器の構造

AttnGAN モデルにおける生成器の基本的な構造図を図 2.4 に示す．生成器は解像度ごとにステージが 3 つに分けられており，それぞれ 64×64，128×128，256×256 の画像が生成される．生成器は主に Upsample 層と残差接続層で構成されている．Upsample 層では画像特徴量を 4 倍（縦に 2 倍，横に 2 倍）に拡張した後，畳み込みおよび正規化が行われる．残差接続層では入力された画像特徴量と，畳み込みおよび正規化が 2 回行われた画像特徴量との残差接続が行われる．

ステージ 1 はテキストエンコーダと 4 つの Upsample 層，そして畳み込み層で構成されている．ステージ 1 では入力としてテキストを受け取り，入力テキストがテキストエンコーダに渡される．テキストエンコーダでは LSTM [18] を用いて，テキスト特徴量を抽出する．抽出されたテキスト特徴量はランダムノイズと結合され，4×4 の画像特徴量が生成される．その後，4×4 の画像特徴量は 4 つの Upsample 層を経て，64×64 の画像特徴量に拡張される．そして，最後の畳み込み層により，画像特徴量の次元がカラー画像の次元である 3 に変換される．

ステージ 2 およびステージ 3 は，注意機構と 3 つの残差接続層，Upsample 層，そして畳み込み層で構成されている．ステージ 2 では入力として，テキスト特徴量および 64×64 の画像特徴量を受け取り，それらを注意機構に入力する．注意機構では Transformer モデルの場合と同じく，式 (2.1) によってテキスト特徴量と画像特徴

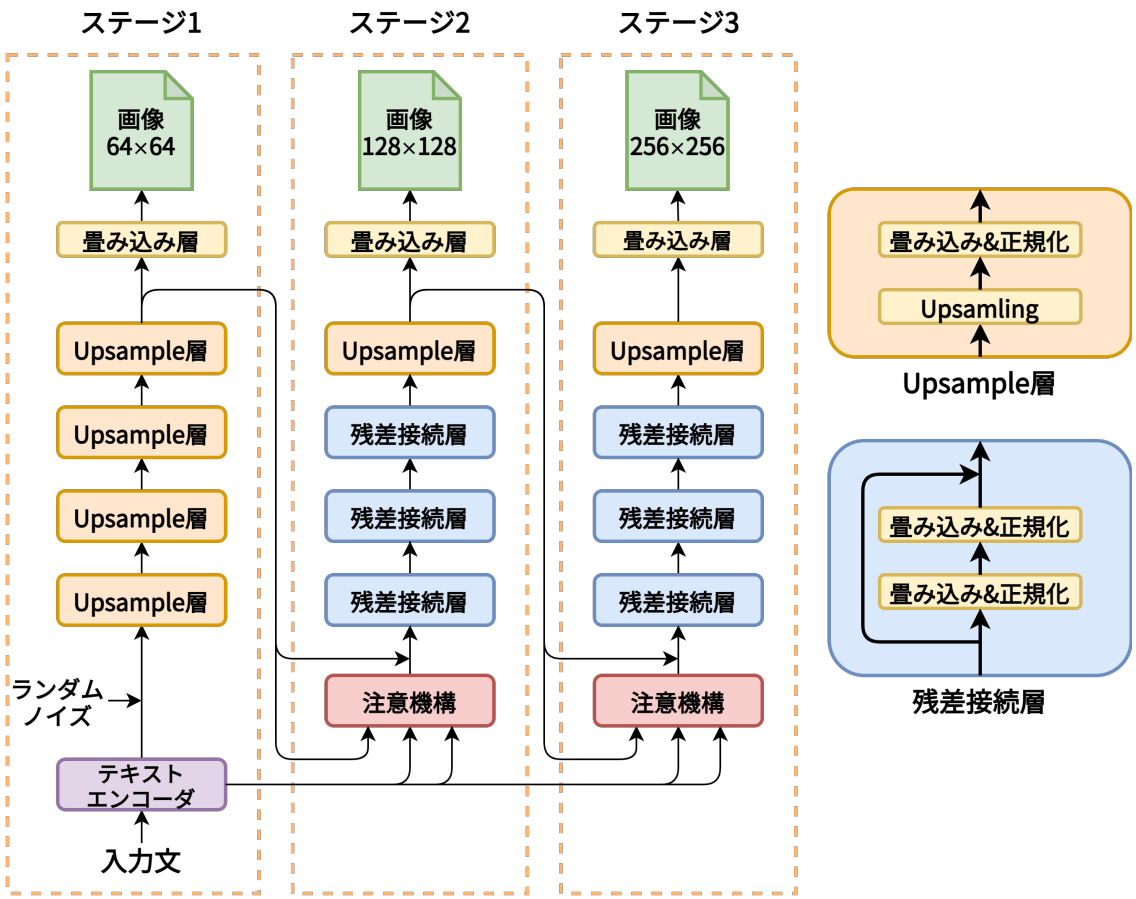


図 2.4: AttnGAN の生成器

量の関連性の高さが考慮される．ただし，Q は画像特徴量，K と V はテキスト特徴量であり，$d_k = 1$ である．その後，64×64 の画像特徴量を結合し，3 つの残差接続層と Upsample 層を経て，128×128 の画像特徴量に拡張する．最後の畳み込み層はステージ 1 と同様に，画像特徴量の次元をカラー画像の次元に変換するために用いられる．

ステージ 3 ではステージ 2 と同様にして，128×128 の画像特徴量を 256×256 の画像特徴量に拡張する．

以上のようにして，生成器では 64×64，128×128，256×256 の画像を生成する．

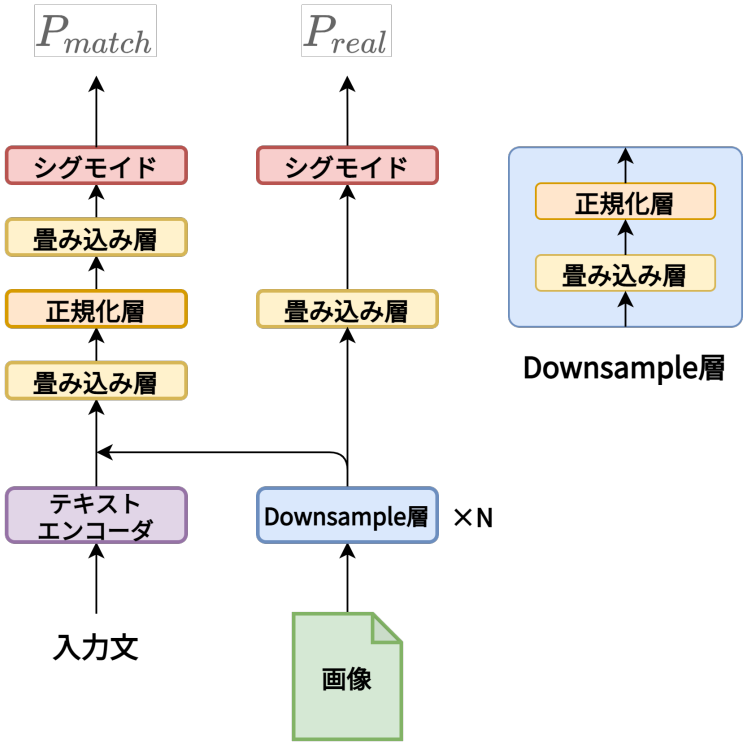


図 2.5: AttnGAN の識別器

2.3.1.2 識別器の構造

AttnGAN モデルにおける識別器の基本的な構造図を図 2.5 に示す．識別器は 64×64，128×128，256×256 の画像それぞれに対し，画像が本物である確率 P_{real} と画像とテキストが対応関係にある確率 P_{match} を出力する．

識別器は主に Downsample 層とテキストエンコーダ，畳み込み層，正規化層で構成されている．テキストエンコーダは生成器と同じものが用いられる．Downsample 層では，畳み込み層によって画像特徴量を 1/4（縦に 1/2，横に 1/2）に圧縮した後，正規化が行われる．Downsample 層の数 N は入力画像のサイズによって決まっており，64×64 の場合は 4 層，128×128 の場合は 5 層，256×256 の場合は 6 層となっている．

入力された画像は Downsample 層で 4×4 の画像特徴量に変換される．その後，畳

み込み層およびシグモイド関数によって，画像が本物である確率 P_{real} を算出する．また，4×4 の画像特徴量はテキストエンコーダから出力されたテキスト特徴量と結合される．その後，畳み込み層と正規化層，シグモイド関数によって画像がテキストと対応関係にある確率 P_{match} を算出する．

第 3 章　提案手法

3.1　BiAttnGAN モデル

従来の T2I モデルは 1 つの文から 1 つの画像を生成する．一方，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，2.3.1 節で述べた最先端の T2I モデルの 1 つである AttnGAN モデル [17] 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 (3.1) 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 $\bar{e}_{src/tgt}$ へと符号化する．

$$\begin{aligned} e_{src}, \bar{e}_{src} &= Enc_{src}(x_{src}) \\ e_{tgt}, \bar{e}_{tgt} &= Enc_{tgt}(x_{tgt}) \end{aligned} \tag{3.1}$$

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，注意機構では，以下の式 (3.2) のように画像とテキストとの関連性を反映した画像特徴量 h' を算出する．

$$h' = [Attn_{src}(h, e_{src}, e_{src});\ Attn_{tgt}(h, e_{tgt}, e_{tgt})] \tag{3.2}$$

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 $M_{src \to tgt}$ 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 $C_{tgt'} = M_{src \to tgt}(C_{src})$ である．
2. **モデルの初期化：** MNMT モデル $M^{(0)}_{(src,img) \to tgt}$ と T2I モデル $M^{(0)}_{(src,tgt) \to img}$ 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 $M^{(k)}_{(src,img) \to tgt}$ 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 $B_{img'} = M^{(k-1)}_{(src,tgt) \to img}(B_{src}, B_{tgt})$ である．
5. **T2I の再学習：** T2I モデル $M^{(k)}_{(src,tgt) \to img}$ 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 $C_{tgt'} = M^{(k-1)}_{(src,img) \to tgt}(C_{src}, C_{img})$ である．
6. **end**

ここで，h はテキストエンコーダの隠れ状態であり，$Attn$ は式 (2.1) に示した注意機構である．ただし，Q は画像特徴量，K と V はテキスト特徴量であり，$d_k = 1$ である．

3.2 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．なお，アルゴリズム 1 における M は，NMT,

MNMT，T2I モデルのいずれかを表す．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再学習する（3 から 5 行目）[1)]．

3.2.1 初期擬似 3 つ組データの作成

MNMT および T2I モデルの初期化の際に用いる初期擬似 3 つ組データは，Transformer NMT モデルを用いて擬似的に作成する．その流れを図 3.1 に示す．まず，対訳テキストデータから Transformer NMT モデルを学習する（図 3.1(a)）．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 3.1(b)）．この翻訳された目的言語文と画像キャプションデータを合わせて初期擬似 3 つデータとする．

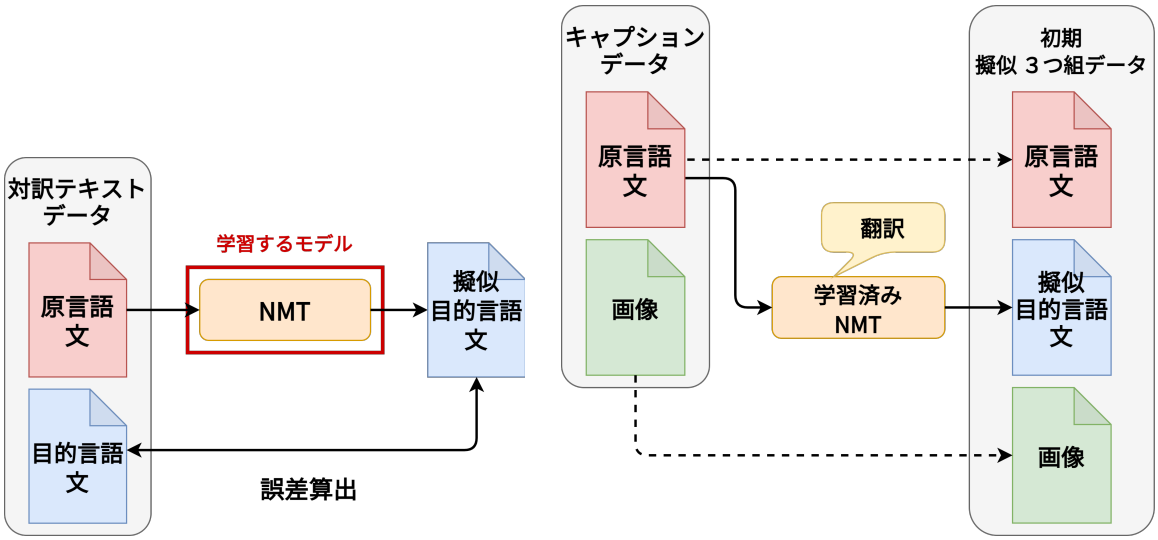
3.2.2 モデルの初期化

学習の安定化と高速化を図るため，Transformer ベースの MNMT モデルと BiAttnGAN モデルの初期化を行う．具体的には，3.2.1 節で作成した初期擬似 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．

3.2.3 MNMT の再学習

BiAttnGAN モデルを用いて，MNMT モデルの再学習を行う過程の概要図を図 3.2 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

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．



(a) Transformer NMT モデルの学習　　(b) 擬似目的言語文の生成

図 3.1: 初期擬似 3 つ組データ作成の流れ

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 が本来の目的言語文 t と同じになるように，以下の式 (3.3) のクロスエントロピー損失 L_M を最小化する．

$$L_M = -\sum_{i=0}^{l-1} t_i \times \log P(y_i) \tag{3.3}$$

ここで，t_i, y_i は t, y 中の i 番目の単語であり，l は目的言語文の長さである．

3.2.4 T2I の再学習

MNMT モデルを用いて，BiAttnGAN モデルの再学習を行う過程の概要図を図 3.3 に示す．まず，MNMT 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（原言語文）を目的言語文に翻訳する．そして，原言語側の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．

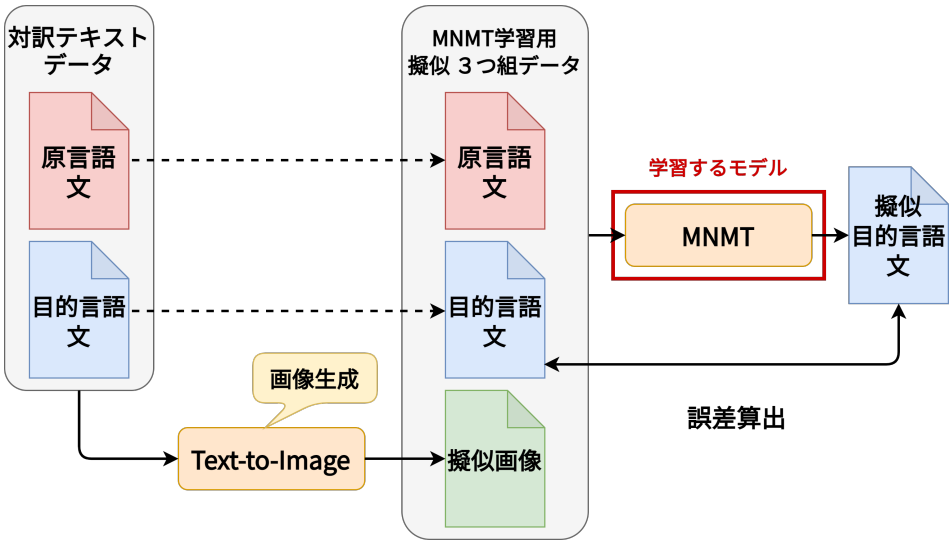


図 3.2: MNMT モデルの再学習

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，以下の式 (3.4) のクロスエントロピー損失 L_G を最小化する．

$$L_G = -\frac{1}{2}\log P_{real}(y_{img}) - \frac{1}{2}\log P_{match}(y_{img}, x_{src}) \tag{3.4}$$

ここで，P_{real} は生成された画像が本物かどうかを表す確率であり，P_{match} 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．また，識別器 D が正しく判別できるようにするために，すなわち P_{real} と P_{match} の精度を上げるために，以下の式 (3.5) のクロスエントロピー損失 L_D が最小になるように判別器を訓練する．

$$\begin{aligned} L_D = &-\frac{1}{2}\log P_{real}(t_{img}) - \frac{1}{2}\log\left(1 - P_{real}(y_{img})\right) \\ &-\frac{1}{2}\log P_{match}(t_{img}, x_{src}) - \frac{1}{2}\log\left(1 - P_{match}(y_{img}, x_{src})\right) \end{aligned} \tag{3.5}$$

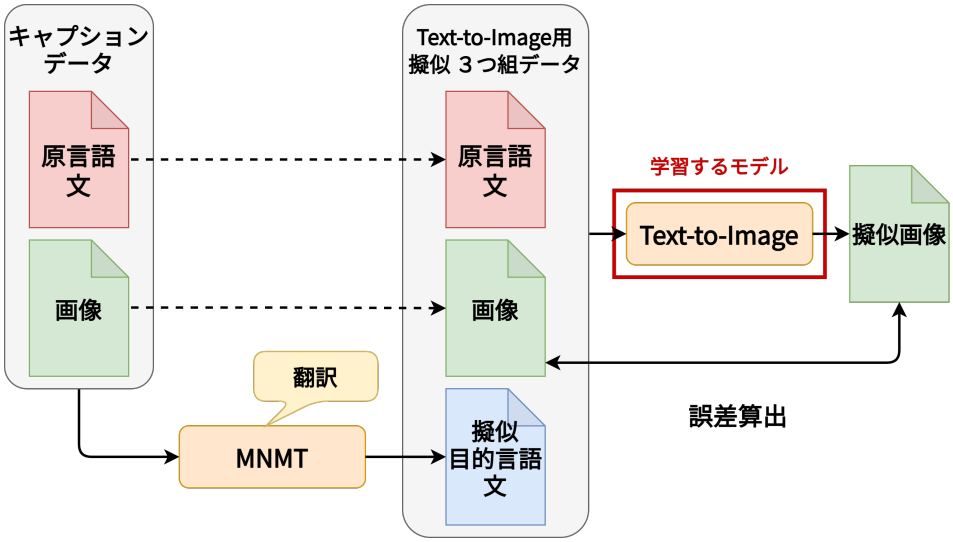


図 3.3: BiAttnGAN モデルの再学習

第 4 章　実験

4.1　実験設定

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 [8] 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 [9] 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原言語側の画像キャプションデータ）として使用した．すなわち，アルゴリズム 1 の B として Multi30k データセットを，C として MS COCO データセットを使用した．なお，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．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．

MNMT モデルで使用する画像に対する前処理として，各画像をサイズが 256×256 にリサイズした後，224×224 になるように中央でクリップ処理を施した．また，T2I モデルで使用する本物画像に対しては，サイズが 304×304 になるようにリサイズした後，256×256 になるようにランダムクリップ処理を施した．そして，T2I モデルの出力である偽画像のサイズは 256×256 に設定した．

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 [1] 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

表 4.1: 実験結果

モデル	BLEU (%)
NMT	38.18
$MNMT_{init}$	36.76
$MNMT_{prop}$	39.56
$MNMT_{gold}$	38.54

の AttnGAN [17] 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 [19] 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，Transformer 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．ドロップアウトの確率を 0.3 に設定し，Byte Pair Encoding（BPE） [20] を適応した． 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，比較した．

1. 提案 MNMT モデル（$MNMT_{prop}$）
2. 画像入力なし NMT モデル（NMT）
3. 提案手法における初期化時の 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 $M^{(0)}_{(src,img)\rightarrow tgt}$）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 [21] で測定した．実験結果を表 4.1 に示す．

表 4.1 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 章で考察する．

第 5 章　考察

5.1　生成された偽画像の例

図 5.1 に本物画像と BiAttnGAN モデルが生成した偽画像との比較を示す．図 5.1 (a) は Multi30k データセットの本物画像であり，図 5.1 (b) は BiAttnGAN モデルが生成した偽画像である．この偽画像は，英文「group of people walking on the heavy snow. 」とその独文「eine gruppe geht durch den tiefschnee.」の文対から生成された画像である．これらの図より，生成された偽画像には多様性があり，すべての図は入力文のペアに関連していることが分かる．BiAttnGAN モデルの性能については次の節で述べる．

$MNMT_{gold}$ の学習では，各エポックで同じ画像（例えば 図 5.1 (a)）の本物画像を使用している．これにより $MNMT_{gold}$ は画像空間上の多様な分布のごく一部しか学習できていないと考えられる．一方で，$MNMT_{prop}$ の学習では，図 5.1 (b) に示すように，BiAttnGAN はエポック毎に背景や物体の配置を異なるランダムノイズで変化させるため，エポック毎に様々な画像を使用することになる．したがって，我々の疑似 3 つ組データを用いた逆翻訳フレームワークによる MNMT モデル $MNMT_{prop}$ は，学習時に多様な画像から学習されるため，結果として真の 3 つ組データを用いたモデル $MNMT_{gold}$ よりも高い翻訳精度を実現したと考えられる．

5.2　BiAttnGAN モデルの性能

本節では BiAttnGAN の性能の調査を行う．従来の AttnGAN モデル [17]（$AttnGAN$）と，提案手法の初期化時（アルゴリズム 1 の step 2）の BiAttnGAN モデ

(a) 本物画像

(b) BiAttnGAN が生成した偽画像

図 5.1: 本物画像と BiAttnGAN モデルが生成した偽画像との比較

ル（$BiAttnGAN_{init}$）と，提案の BiAttnGAN モデル（$BiAttnGAN_{prop}$）とを比較した．

各モデルの性能は，GAN の評価方法である inception スコア [22] を用いて，文献 [14] に従い MS COCO の検証セットで評価した．具体的には以下の式 (5.1) でスコアを算出した．

$$I = exp\left(\mathbb{E}_{\boldsymbol{x}} D_{KL}(p(y|\boldsymbol{x}) \,||\, p(Y))\right), \tag{5.1}$$

ここで，D_{KL} は Kullback-Leibler ダイバージェンス，$\boldsymbol{x}$ は生成された画像，y は Inception モデルで予測されたラベルである．Inception モデルには MS COCO データセットによって事前に学習された Inception モデルを使用した．直感的にはスコアが高いほど，より多様で意味のある画像が生成されていることを示している．

AttnGAN では英語文のみから画像を生成した．$BiAttnGAN_{init}$ と $BiAttnGAN_{prop}$

表 5.1: BiAttnGAN の性能

Model	Inception score
$AttnGAN$	25.89 ± .47
$BiAttnGAN_{init}$	26.41 ± .40
$BiAttnGAN_{prop}$	26.47 ± .37

の評価では，英語文を MNMT モデルで独文に翻訳し，その翻訳文ペアを入力とし画像を生成した．用いる MNMT モデルとしては，$BiAttnGAN_{init}$ では初期化時の MNMT モデル $MNMT_{init}$，$BiAttnGAN_{prop}$ では提案の MNMT モデル $MNMT_{prop}$ を用いた．

表 5.1 に実験結果を示す．表 5.1 より，我々の 2 つの BiAttnGAN モデル（$BiAttnGAN_{init}$ と $BiAttnGAN_{prop}$）が $AttnGAN$ モデルよりも高い性能を達成していることが分かる．これは追加の入力，すなわち目的言語文を使うことの有効性を示している．また，表 5.1 より，$BiAttnGAN_{prop}$ が $BiAttnGAN_{init}$ よりも優れていることが分かる．

5.3 翻訳例

図 5.2 に実際の翻訳例を示す．図 5.2 を見ると，NMT と $MNMT_{gold}$ が原言語文の情報を反映できていないことが分かる．具体的には，図 5.2 (a) の「parade」と図 5.2 (b) の「red guitar」の情報が失われている．一方，提案手法 $MNMT_{prop}$ ではこれらの情報が反映されている．この実例からも，我々の提案する逆翻訳フレームワークが有効であり，翻訳品質の向上に貢献していることが分かる．

5.4 大規模データセットによる事前学習

提案する逆翻訳フレームワークは，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 モデルを学習するもので，3 つ組データを必要としない．

Source :	a dance group performs in a **parade** in china .
Reference :	eine tanztruppe bei einer aufführung während einer parade in china .
NMT :	eine gruppe tanzt in china .
$MNMT_{gold}$:	eine gruppe von tanzern in china .
$MNMT_{prop}$:	eine gruppe tanzt in einer **parade** in china .

(a) 翻訳例 (1)

Source :	guitar player performs at a nightclub **red guitar** .
Reference :	gitarristin spielt in einem nachtklub auf einer roten gitarre .
NMT :	ein gitarrespieler spielt in einem nachtclub .
$MNMT_{gold}$:	ein gitarrespieler tritt in einem nachtclub auf einer nachtclub auf .
$MNMT_{prop}$:	ein gitarrespieler tritt in einem nachtclub auf einer **roten gitarre** .

(b) 翻訳例 (2)

図 5.2: 実際の翻訳例

そのため，従来の 3 つ組データを用いた手法に比べて，より多様で大規模なデータセットを利用することが可能である．この利点の有用性を検証するために，Multi30k データセットとは異なる領域の大規模データセットも用いる実験を行った．対訳テキストデータとして WMT14 英独データ，画像キャプションデータとして GoodNews データを用いた．以下，この大規模データセットを用いて学習した MNMT モデルを $MNMT_{pre}$ とし，このモデルを事前学習モデルとする．

事前学習時のパラメータは，以下の点を除いて 4.1 節と同じである．事前学習

表 5.2: 事前学習モデルによる実験結果 (BLEU (%))

Model	Validation	Test
$MNMT_{gold}$	40.36	38.54
$MNMT_{prop}$	41.03	39.56
$MNMT_{pre}$	27.21	27.48
$MNMT_{pre\_un}$	42.59	40.63
$MNMT_{pre\_semi}$	42.98	42.18

時，NMT モデルでは，ドロップアウト確率を 0.1 とし，ミニバッチサイズを約 25,000 トークンとした．MNMT モデルでは，ドロップアウト確率を 0.3，ミニバッチサイズを 256 とした．BiAttnGAN モデルでは，ミニバッチサイズを 64 とした．NMT モデルは 10 万ステップ，初期の MNMT モデルは 20 エポック，初期の T2I モデルは 60 エポック，再訓練は 15 エポックで学習を行った．

事前学習モデル $MNMT_{pre}$ の性能は，表 5.2 からわかる通り，学習データの領域がテストデータ (Multi30k) の領域と異なるため，非常に低いことに注意する必要がある．そこで Multi30k データを用いた教師なしと教師ありの方法で，事前学習モデルをファインチューニングした．ファインチューニング時には各モデルは 15 エポックで学習した．教師なしのファインチューニングについては 5.4.1 節，教師ありのファインチューニングについては 5.4.2 節で述べる．

5.4.1 教師なし MNMT モデル

本節では，$MNMT_{pre}$ を提案手法による逆翻訳形式フレームワークを用いて，教師なしでファインチューニングする．すなわち，3 つ組データは使わず，Multi30k の対訳テキストデータと MSCOCO の画像キャプションデータを用いてファインチューニングした．教師なし MNMT モデル $MNMT_{pre\_uns}$ の性能を表 5.2 に示す．この表から，$MNMT_{pre\_uns}$ が $MNMT_{prop}$ を上回る（+1.07 BLEU ポイント）ことが

分かる．この結果は，本提案フレームワークによる事前学習により，3 つ組データを利用しない教師なしの環境で，翻訳品質のさらなる向上が達成できることを示している．

5.4.2 半教師あり MNMT モデル

本節では，教師なし学習によって得られた事前学習モデル $MNMT_{pre}$ を，テストデータの領域の 3 つ組データを用いて教師ありの方法でファインチューニングする．具体的には，Multi30k データセットの真の 3 つ組データ（Multi30k データセットの対訳文とその画像）を用いてファインチューニングした．したがって，このモデルは半教師ありモデルである．この半教師あり MNMT モデル $MNMT_{pre\_semi}$ の性能を表 5.2 に示す．表 5.2 より, $MNMT_{pre\_semi}$ が $MNMT_{gold}$ よりも優れていることがわかる（+3.64 BLEU ポイント）．この結果は，提案するフレームワークによる事前学習が，3 つ組データが利用できる教師ありの環境でも有効であることを示している．

第 6 章　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原言語側の画像キャプションデータから MNMT モデルを学習する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実験を通じて，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を示した．また，大規模な領域外データセットを用いた事前学習により，教師なし及び半教師ありの環境でも翻訳品質をさらに向上させることができることを確認した．今後は異なる規模や領域，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．

参考文献

[1] 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 *Proceedings of Advances in Neural Information Processing Systems 30*, pp. 5998–6008, 2017.

[2] T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, 2015.

[3] 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 2018 Third Conference on Machine Translation*, pp. 304–323, 2018.

[4] Hideki Nakayama and Noriki Nishida. Zero-resource machine translation by multimodal encoder-decoder network with multimedia pivot. *Machine Translation*, Vol. 31, No. 1-2, pp. 49–64, 2017.

[5] Yun Chen, Yang Liu, and Victor O.K. Li. Zero-resource neural machine translation with multi-agent communication game. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, the 30th innovative Applications of Artificial Intelligence, and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence*, pp. 5086–5093, 2018.

[6] Yuanhang Su, Kai Fan, Nguyen Bach, C.-C. Jay Kuo, and Fei Huang. Unsupervised multi-modal neural machine translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[7] Po-Yao Huang, Junjie Hu, Xiaojun Chang, and Alexander Hauptmann. Unsupervised multimodal neural machine translation with pseudo visual pivoting. *CoRR*, Vol. abs/1207.0016, , 2020.

[8] 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, 2016.

[9] 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 *Proceedings of the 13th European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

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

[11] 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 *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pp. 603–611, 2018.

[12] Tetsuro Nishihara, Akihiro Tamura, Takashi Ninomiya, Yutaro Omote, and Hideki Nakayama. Supervised visual attention for multimodal neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4304–4314, 2020.

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

[14] Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 1060–1069, 2016.

[15] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang,

and Dimitris N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[16] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *arXiv: 1710.10916*, 2017.

[17] 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 *Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1316–1324, 2018.

[18] Sepp Hochreiter, JürgenSchmidhuber. Long short-term memory. *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

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

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

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

[22] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In *Proceedings of Advances in Neural Information Processing Systems 29*, pp. 2234–2242, 2016.