一般論文

教師付き視覚的注意を用いたマルチモーダルニューラル機械翻訳

西原　哲郎[†]・田村　晃裕[††]・二宮　　崇[†]・表　悠太郎[†]・中山　英樹[†††]

本稿では，マルチモーダルニューラル機械翻訳 (MNMT) のための教師付き視覚的注意機構を提案する．提案手法は，人手で付与された画像内の領域と単語との対応関係を視覚的注意の教師データとして与え，これらの対応関係を制約にして直接視覚的注意機構の学習を行う．教師なしで学習される従来の視覚的注意機構に比べてより正確に単語と画像領域との関係性を捉えることが期待される．実験では Multi30k データセットを用いた英独・独英翻訳，Flickr30k Entities JP データセットを用いた英日・日英翻訳を行い，提案する教師付き視覚的注意機構によって Transformer ベースの MNMT モデルの性能が改善することが確認できた．また，教師付きの言語間注意機構と組み合わせることにより，さらに性能が改善され，最大で BLEU スコアが 1.61 ポイント，METEOR スコアが 1.7 ポイント向上することが確認できた．

キーワード：マルチモーダル機械翻訳，視覚的注意機構

Supervised Visual Attention for Multimodal Neural Machine Translation

Tetsuro Nishihara[†], Akihiro Tamura[††], Takashi Ninomiya[†], Yutaro Omote[†] and Hideki Nakayama[†††]

This paper proposed a supervised visual attention mechanism for multimodal neural machine translation (MNMT), trained with constraints based on manual alignments between words in a sentence and their corresponding regions of an image. The proposed visual attention mechanism captures the relationship between a word and an image region more precisely than a conventional visual attention mechanism trained through MNMT in an unsupervised manner. Our experiments on English-German and German-English translation tasks using the Multi30k dataset and on English-Japanese and Japanese-English translation tasks using the Flickr30k Entities JP dataset show that a Transformer-based MNMT model can be improved by incorporating our proposed supervised visual attention mechanism and that further improvements can be achieved by combining it with a supervised cross-lingual attention mechanism (up to +1.61 BLEU, +1.7 METEOR).

Key Words: *Multimodal Machine Translation, Visual Attention Mechanism*

[†] 愛媛大学大学院理工学研究科, Graduate School of Science and Engineering, Ehime University
[††] 同志社大学, Doshisha University
[†††] 東京大学, The University of Tokyo

1 はじめに

現在，ニューラルネットワークを用いた機械翻訳（ニューラル機械翻訳）が機械翻訳の主流となっている．注意機構を用いた再帰型ニューラルネットワーク (Recurrent Neural Network; RNN) に基づくニューラル機械翻訳モデルは初期のころから広く使用されてきたモデルであり，原言語文内の単語と目的言語文内の単語間の関係を捉える言語間注意機構を用いることで，従来の RNN ベースのニューラル機械翻訳よりも高い精度を実現した (Bahdanau et al. 2015; Luong et al. 2015)．また，従来の言語間注意機構に加えて，同じ文中の単語間の関係を捉える自己注意機構を導入した Transformer モデル (Vaswani et al. 2017) が提案され，RNN や畳み込みニューラルネットワーク (Convolutional Neural Network; CNN) を用いた手法と比べて高い精度を実現することから，近年注目されている．

ニューラル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なされているが，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 (Liu et al. 2016; Mi et al. 2016; Garg et al. 2019)．これらの研究では，アライメントツールを用いて原言語文と目的言語文間の単語の対応関係を予め取得し，その対応関係を教師データとして与えて言語間注意機構を学習させることで翻訳性能の向上を実現している．

機械翻訳手法の一つとして，原言語文とそれに対応する内容の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ニューラル機械翻訳 (Barrault et al. 2018) が提案されている．翻訳時に与えられる画像は，翻訳の曖昧性解消や省略補完の手がかりとして役立つと考えられ，画像を参照することでより質の高い翻訳が実現されることが期待されている．マルチモーダルニューラル機械翻訳のモデルとして，Helcl ら (Helcl et al. 2018) は，CNN によって抽出した画像の特徴量を翻訳に活用するために，文中の単語と画像の領域との対応関係を捉える視覚的注意機構を Transformer モデルのデコーダ内に導入したモデルを提案している．また，Delbrouck ら (Delbrouck and Dupont 2017) は，RNN ベースのマルチモーダルニューラル機械翻訳モデルのエンコーダに視覚的注意機構を導入したモデルを提案している．しかし，これらの視覚的注意機構は，マルチモーダルニューラル機械翻訳の訓練時に教師なしで自動的に学習が行われている．そのため，本来捉えるべき対応関係を常に捉えられているとは限らない．

本稿では，マルチモーダルニューラル機械翻訳の性能改善のために，人手により与えられた文中の単語と画像領域との対応関係に基づいて教師付き学習を行う制約付き視覚的注意機構を提案する．具体的には，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が付与されたデータを教師データとして用いることで，Transformer モデルエンコーダ内の視覚的注意機構を直接学習させることを行う．

Multi30k データセット (Elliott et al. 2016) を用いた英独翻訳および独英翻訳と Flickr30k Entities JP データセット (Nakayama et al. 2020) を用いた英日翻訳および日英翻訳の評価実験

を行い，提案する教師付き視覚的注意機構によって Transformer ベースのマルチモーダル機械翻訳モデルの翻訳性能が改善することが確認できた．また，教師付きの言語間注意機構と組み合わせることにより，さらに翻訳性能が改善されることを確認した．

本稿の構成は以下の通りである．2 節で本研究の背景について述べ，3 節で提案手法について説明する．4 節では実験について述べ，5 節で実験結果の考察を行う．6 節で関連研究について述べ，最後に 7 節でまとめと今後の課題について述べる．

2 Transformer ベースのニューラル機械翻訳

本節では，本研究で提案するマルチモーダル機械翻訳モデルの基礎となる Transformer ベースのニューラル機械翻訳を説明する．まず最初に Transformer モデルの概要を述べる．次に，Transformer モデルにおける教師付き言語間注意機構について説明する．

2.1 Transformer モデルの概要

Transformer モデルは，原言語文を受け取って中間表現に変換するエンコーダと，その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．各エンコーダレイヤは自己注意機構と位置毎の全結合レイヤの 2 つのサブレイヤを持っている．また，各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．これらのサブレイヤ間では，残差接続 (He et al. 2016) とレイヤの正規化 (Ba et al. 2016) が用いられる．

自己注意機構と言語間注意機構 (Att) は以下の式で表される．

$$\mathrm{Att}(Q, K, V) = AV \tag{1}$$

$$A = \mathrm{softmax}(\frac{QK^\top}{\sqrt{d_k}}) \tag{2}$$

ここで，A は注意行列と呼ばれる．また，Q, K, V はエンコーダ及びデコーダにおける隠れ状態を表し，d_k は Q, K, V の次元数を表す．自己注意機構では，上式の Q, K, V として直前のサブレイヤの出力を用いる．また，言語間注意機構では，Q としてデコーダ内の直前のサブレイヤの出力，K と V としてエンコーダの出力を用いる．自己注意機構では同一文中の単語間の関係を計算することができる．また，言語間注意機構では原言語文内の単語と目的言語文内の単語間の関係を計算することができる．

Transformer の特徴として，隠れ状態を部分空間に分割し，各部分空間において様々な情報を表現することを可能にするマルチヘッド注意機構がある．h 個のヘッドからなるマルチヘッド

注意機構 (MHA) は以下のように表される．

$$\mathrm{MHA}(Q,K,V) = [head_1;\ldots;head_h]W^O \tag{3}$$

$$head_i = \mathrm{Att}(QW_i^Q, KW_i^K, VW_i^V) \tag{4}$$

ここで，[ ] はベクトルを結合することを表している．$W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d_{model}\times d_k}$ はそれぞれヘッド毎に定義されるパラメータ行列であり，d_{model} 次元ベクトルを線形変換により d_k 次元に縮退させる．この d_k 次元ベクトルが各ヘッドに渡される．$W^O \in \mathbb{R}^{hd_k\times d_{model}}$ はパラメータ行列であり，各ヘッドの出力を結合したベクトルに対し線形変換を行う．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．

単語位置毎の全結合レイヤ (FFN) における計算は，以下の式で表される．

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{5}$$

ここで，$W_1 \in \mathbb{R}^{d_{model}\times d_{ff}}$，$W_2 \in \mathbb{R}^{d_{ff}\times d_{model}}$ はパラメータ行列，$b_1 \in \mathbb{R}^{d_{ff}}$，$b_2 \in \mathbb{R}^{d_{model}}$ はバイアス項である．また，全結合レイヤへの入力と出力の次元数は d_{model}，中間レイヤの次元数は d_{ff} である．

Transformer では語順の情報を組み込むために以下で表される位置エンコーディングが導入されている．

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

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

ここで，pos は単語の位置，i は各成分の次元を表す．この位置エンコーディングを単語の埋め込み表現に加えることで，単語の語順の情報を付与することができる．

2.2 教師付き言語間注意機構

Garg ら (Garg et al. 2019) は，Transformer モデルの言語間注意機構に原言語と目的言語間の単語の対応関係を教師データとして与えて学習を行う手法を提案している．アライメントツールを用いて言語間の対応関係を取得し，マルチヘッド言語間注意機構のある 1 つのヘッドとの間で計算される誤差を最小化することによって注意機構の学習を行う．誤差は以下の交差エントロピーによって計算される．

$$\mathcal{L}_a(A) = -\frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N} G_{m,n} \times \log(A_{m,n}) \tag{8}$$

ここで，M は目的言語文の文長，N は原言語文の文長，A は式 (4) によって計算される言語間注意機構の注意行列，G は教師データとなる単語の対応関係を表した行列である．なお，n 番

目の原言語文の単語と m 番目の目的言語文の単語が対応関係にある場合は，$G_{m,n}$ は 1 となり，それ以外は 0 となる．この機械翻訳モデルの目的関数 $\mathcal{L}$ としては，上述の $\mathcal{L}_a(A)$ を翻訳の誤差 $\mathcal{L}_t$ に加えた以下の損失関数を用いる．

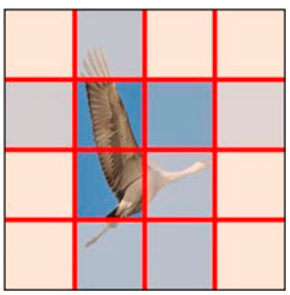
$$\mathcal{L} = \mathcal{L}_t + \lambda \mathcal{L}_a(A), \tag{9}$$

ここで，λ はハイパーパラメータである．

2.3　Transformer ベースのマルチモーダルニューラル機械翻訳に関する研究

マルチモーダルニューラル機械翻訳モデルとして様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 モデルベースのモデルが盛んに研究されている．例えば，Takushima ら (Takushima et al. 2019) は CNN と Transformer エンコーダからなる画像エンコーダを Transformer モデルに導入したマルチモーダルニューラル機械翻訳モデルを提案している．彼らのモデルでは，まず，画像エンコーダで入力画像に対して 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．また，Helcl ら (Helcl et al. 2018) は，Transformer デコーダの内部で言語間注意機構の出力と CNN の出力を用いた視覚的注意機構によって画像情報を活用するモデルを提案している．図 1 に視覚的注意機構の例を示す．図において，原画像がより鮮明に見えている部分に強く注意が向けられていることを表している．入力された画像は，CNN によって高次元の目の粗い格子状の特徴量へと変換される．特徴量中の各領域は，元の画像の領域に対応しており，高次元の特徴量を持っている．例えば，図 1(b) は 4×4 の領域に変換されていて，各領域が 2,048 次元の特徴量

(a) 原画像



(b) 単語 crane に対する視覚的注意機構の例

図 1　視覚的注意機構の例

を保有している．視覚的注意機構では，各単語はこれらの画像特徴量の領域に対して注意が向けられる．このため，視覚的注意は，領域のヒートマップとして可視化することが出来る．

3　教師付き視覚的注意機構を用いたマルチモーダルニューラル機械翻訳

本節では，教師付き視覚的注意機構を用いたマルチモーダルニューラル機械翻訳を提案する．まず，ベースラインとして用いる Transformer ベースのマルチモーダルニューラル機械翻訳モデルを説明する．その後，マルチモーダルニューラル機械翻訳の翻訳性能を向上させるための教師付き視覚的注意機構を提案する．

3.1　Transformer ベースのマルチモーダルニューラル機械翻訳モデル

図 2 に本研究で用いる Transformer ベースのマルチモーダルニューラル機械翻訳モデルの概要図を示す．本モデルは，原言語文エンコーダとデコーダに加えて，画像エンコーダを持つ．画像エンコーダでは，まず，入力した画像に対して CNN を適用し，画像の特徴量を得る．次

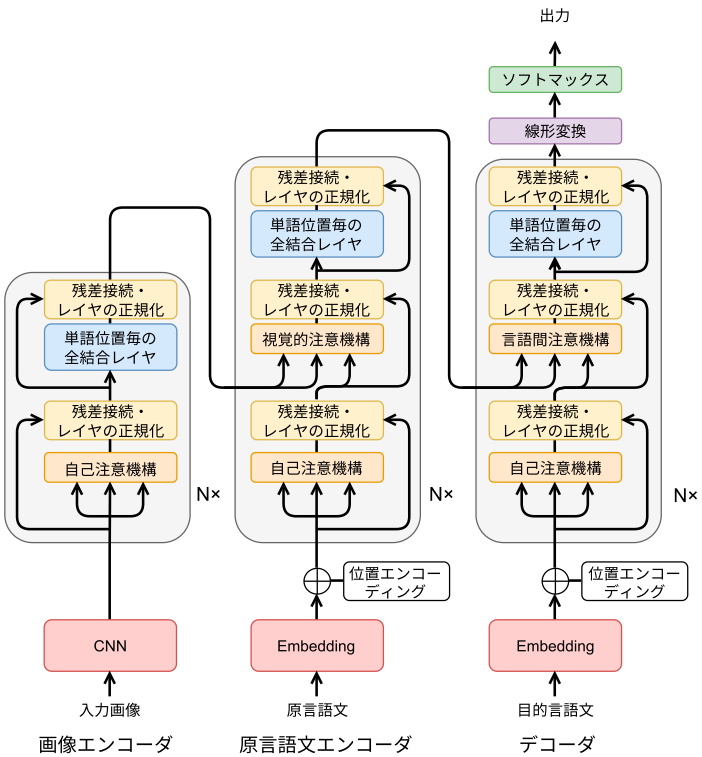


図 2　マルチモーダルニューラル機械翻訳モデル

に，CNN の出力に対して自己注意機構を適用する．この自己注意機構によって画像の領域間の関係性が計算される．最後に，自己注意機構の出力に対して位置毎の全結合レイヤを適用したものが画像エンコーダ全体の出力となる．

本モデルでは，従来のモデルとは異なり，原言語文エンコーダの内部において原言語文に対する自己注意機構の出力と画像エンコーダの出力を用いて，注意機構の一種である視覚的注意機構 (Libovický et al. 2018) により，原言語文の単語と画像の領域との関係性を計算する．視覚的注意機構では，式 (1) において，Q として原言語文エンコーダの自己注意機構の出力，K と V として画像エンコーダの出力を用いる．

3.2 教師付き視覚的注意機構

提案する教師付き視覚的注意機構は，原言語文の単語と画像内のオブジェクトとの対応関係を人手で付けたものを教師データとして与えて視覚的注意機構を学習する．具体的には，画像エンコーダの出力と原言語文エンコーダ内の自己注意機構の出力との間の視覚的注意機構が教師データに近づくような制約を与える．視覚的注意機構に対する制約は，教師となる対応関係を示した行列と，式 (2) で計算される注意行列との間の誤差が最小となるように適用される．誤差は以下の交差エントロピーによって計算される．

$$\mathcal{L}_{img\_src}(A) = -\frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N} G_{m,n} \times \log(A_{m,n}) \tag{10}$$

ここで，M は原言語文の文長を，N は CNN によって畳み込まれた画像の領域数を表す．また，A は注意行列を，G は教師データとなる原言語文内の単語と画像内のオブジェクトの対応関係を示した行列を表す．原言語文内の m 番目の単語が画像の n 番目の領域に対応しているとき，$G_{m,n}$ は 1 となり，それ以外の時は 0 となる．

本研究では，Flickr30k Entities データセット (Plummer et al. 2017) を用いて視覚的注意機構に対する教師データを作成した．このデータセットは Flickr30k データセット (Young et al. 2014) から作られたデータセットである．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，図 3 のようにその単語が画像内のどの領域と関係があるかが示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，教師データを作成する．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 4×4 に畳み込んだ場合，画像内の各オブジェクトと 16 個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．すなわち，値を「1 / 対応付いた領域数」とする（図 4(a)）．その後，

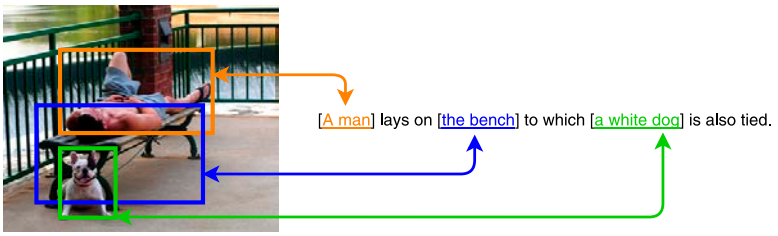
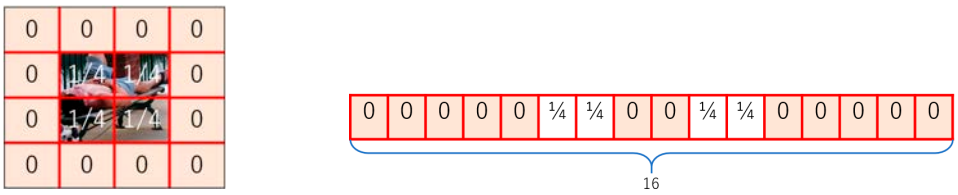


図 3　Flickr30k Entities の例



(a) 単語 man と画像領域との対応関係の例　　(b) 線形化

図 4　教師付き視覚的注意機構に対する教師データの作成例

2 次元の領域を 1 次元に線形化する（図 4(b)）．この手続きを原言語文のすべての単語に対して行う．画像内のオブジェクトと対応がない単語については，Liu ら (Liu et al. 2016) や Mi ら (Mi et al. 2016) に倣い，特殊トークンを用いて処理を行った．Liu らや Mi らは教師付き言語間注意機構を学習する際に特殊トークンを導入している．本手法では，図 4(b) に示した行列の先頭に特殊トークンを付与し，画像内のどのオブジェクトとも対応関係にない単語は特殊トークンに対応付ける．

教師付き視覚的注意機構を用いたマルチモーダルニューラル機械翻訳モデルの目的関数 $\mathcal{L}$ は以下のように表される．

$$\mathcal{L} = \mathcal{L}_T + \lambda_1 \mathcal{L}_{img\_src} \tag{11}$$

ここで，$\mathcal{L}_T$ はマルチモーダルニューラル機械翻訳モデルの損失関数，$\mathcal{L}_{img\_src}$ は視覚的注意機構における注意行列と教師データとなる行列との損失関数を表している．また，λ_1 は翻訳誤差 $\mathcal{L}_T$ と教師付き視覚的注意機構に関する誤差 $\mathcal{L}_{img\_src}$ の重みを制御するためのハイパーパラメータである．

3.3　教師付き視覚的注意機構と教師付き言語間注意機構の組み合わせ

本研究では，教師付き視覚的注意機構に加えて，2.2 節で説明した教師付き言語間注意機構を我々のマルチモーダルニューラル機械翻訳モデルに導入し，翻訳性能の改善を図る．教師付

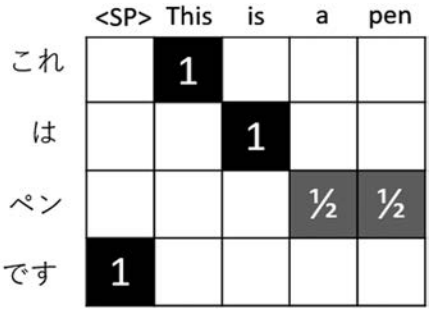


図 5　言語間注意機構に対する教師データの例

き言語間注意機構を学習するためには，原言語文内の単語と目的言語文内の単語との対応関係を取得する必要がある．本研究では，Liu ら (Liu et al. 2016) や Garg ら (Garg et al. 2019) のように，アライメントツールを用いて単語間の対応関係を取得した．アライメントツールとしては，GIZA++ (Och and Ney 2003) をマルチスレッドで動作させることを可能とした MGIZA (Gao and Vogel 2008) を用いた．

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．すなわち，値を「1 / 対応付いた単語数」とする．また，3.2 節のように，どの単語とも対応関係がない単語については，特殊トークンを用いて処理を行った．具体的には，図 5 に示すように原言語文の先頭に特殊トークンを設置し，原言語文内のどの単語とも対応関係を持たない目的言語文の単語はこの特殊トークンに対応が張られるようにする．

教師付き視覚的注意機構と教師付き言語間注意機構の両方を用いるマルチモーダルニューラル機械翻訳モデルの目的関数 $\mathcal{L}$ は次式のように表される．

$$\mathcal{L} = \mathcal{L}_T + \lambda_1 \mathcal{L}_{img\_src} + \lambda_2 \mathcal{L}_{src\_tgt}, \tag{12}$$

ここで，$\mathcal{L}_{src\_tgt}$ は言語間注意機構における注意行列と教師データとの損失関数を表している．また，λ_1 は，翻訳誤差 $\mathcal{L}_T$ と視覚的注意機構に関する誤差 $\mathcal{L}_{img\_src}$ との重みを制御するハイパーパラメータであり，λ_2 は，翻訳誤差 $\mathcal{L}_T$ と言語間注意機構に関する誤差 $\mathcal{L}_{src\_tgt}$ との重みを制御するハイパーパラメータである．

4　実験

4.1　実験設定

本研究では，英独翻訳，独英翻訳，英日翻訳，日英翻訳の 4 つの実験を行った．英独翻訳および独英翻訳では，Multi30k データセット (Elliott et al. 2016) を用いた．このデータセットは，

画像とその説明文が対になったもので，訓練データは 29,000 文対，開発データは 1,014 文対である．また，テストデータとしては test2016（1,000 文対）を用いた．

英日翻訳および日英翻訳では，Flickr30k Entities JP データセット (Nakayama et al. 2020) を用いた．Multi30k データセットのテキストの前処理に倣い，英文には小文字化，句読点の正規化，Moses のトークナイザ (Koehn et al. 2007) を施している．日本語文については KyTea (Neubig et al. 2011) 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 文対，開発データは 2,027 文対，テストデータは 2,000 文対である．

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 (He et al. 2016) 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レイヤの出力を用いた．したがって，抽出さ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 レイヤから成る．マルチヘッド注意機構におけるヘッド数は 8，埋め込み次元数は 512 とした．また，単語位置毎の全結合レイヤの中間レイヤの次元数は 2,048 とした．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．最適化手法には Adam (Kingma and Ba 2014) を用いた．英独および独英実験では BPE (Sennrich et al. 2016) を適用した．英側の単語と独側の単語を合わせて BPE を学習し，マージのオペレーション数は 6,000 とした．推論時には目的言語文の生成を貪欲法により行った．実験では機械学習ライブラリである PyTorch のバージョン 1.1.0 を用いて実装したモデルを使用した．

翻訳性能は BLEU (Papineni et al. 2002) と METEOR (Denkowski and Lavie 2014) を用いて評価した．テスト時には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．また，ブートストラップ・リサンプリング法 (Koehn 2004) により統計的有意差の検定を行った．BLEU の検定には paired_bootstrap_v13a[1] を，METEOR の検定には jhclark/multeval[2] を fork した ozancaglayan/multeval[3] を用いた．実験では，画像無しの Transformer モデル (NMT)，画像無しの Transformer モデルに教師付き言語間注意機構を適用したモデル (NMT+SCA)，教師付き注意機構を用いないマルチモーダル Transformer モデル (MNMT)，教師付き視覚的注意機構のみを用いるモデル (MNMT+SVA)，教師付き言語間注意機構のみ用いるモデル (MNMT+SCA)，教師付き視覚的注意機構と教師付き言語間注意機構の両方を用いるモデル (MNMT+SVA+SCA) を比較した．また，教師付き視覚的注意機構及び教師付き言語間注意機構では，6 レイヤスタックされた原言語文エンコーダレ

[1] http://www.cs.cmu.edu/~ark/MT/paired_bootstrap_v13a.tar.gz
[2] https://github.com/jhclark/multeval
[3] https://github.com/ozancaglayan/multeval

イヤおよびデコーダレイヤの内，最終の 6 レイヤ目の注意機構の一つのヘッドに対して制約を与えて教師付き学習を行った．なお，3.3 節で説明した目的関数 $\mathcal{L}$ におけるハイパーパラメータは Garg らに倣い，$\lambda_1 = 0.05, \lambda_2 = 0.05$ とした．また，教師付き視覚的注意機構のみを用いるモデルのハイパーパラメータは $\lambda_1 = 0.05$ とした．本実験で利用したデータセットには，単語と画像内のオブジェクトとの対応は英語のキャプション文にのみ与えられている．英独翻訳と英日翻訳については，人手によって付けられた，英語のキャプション文内の単語と画像内のオブジェクトとの対応関係を直接利用して，視覚的注意機構に対する教師データを作成した．独英翻訳と日英翻訳については，初めに人手によって付けられている英語のキャプション文内の単語と画像内のオブジェクトとの対応関係を，MGIZA によって得た言語間の対応関係を用いて独語および日本語の文と画像内のオブジェクトとの対応関係に変換し，視覚的注意機構に対する教師データを作成した．なお，今回英日および日英実験で用いた Flickr30k Entities JP 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間に人手で対応関係がつけられている．そこで，NMT+SCA，MNMT+SVA，MNMT+SCA，MNMT+SVA+SCA の場合について，この人手での対応関係を用いた実験も行った．

4.2 実験結果

実験結果を表 1 に示す．表を見ると，MNMT+SVA はすべての実験において MNMT のスコアを上回っており，英独および英日・日英実験については検定によりその有意差が確認できる．また，MNMT+SCA については，英独および独英実験においては MNMT のスコアを上回っており，

	英 → 独		独 → 英		英 → 日		日 → 英	
	B	M	B	M	B	M	B	M
NMT	38.76	57.59	42.58	39.19	43.69	59.27	44.21	40.03
+SCA	40.34	58.91	43.38	39.54	44.36	59.89	44.80	40.38
+SCA（人手）	—	—	—	—	44.44	59.88	44.88	40.45
MNMT	38.89	57.35	42.29	39.13	44.09	59.59	44.42	40.03
+SVA	39.91*	58.11	42.52	38.86	44.51*	60.03*	44.76*	40.40*
+SCA	39.98*	58.20*	43.75*	39.63*	44.09	59.63	44.18	39.73
+SVA+SCA	**40.50***	**59.05***†‡	**43.76***†	**39.71***†	**44.79***‡	**60.23***‡	**45.36***†‡	**40.65***†‡
+SVA（人手）	—	—	—	—	—	—	44.76	40.40
+SCA（人手）	—	—	—	—	45.11*‡	60.36*‡	45.39*‡	40.55*‡
+SVA+SCA（人手）	—	—	—	—	44.85*‡	60.22*‡	44.90‡	40.41*‡

表 1 実験結果．B と M はそれぞれ BLEU と METEOR を表す．$*$ マークは MNMT と比較して有意水準 5% ($p \leq 0.05$) で検定を行い，有意と判定されたことを表す．また，† マークは MNMT+SVA と比較して有意水準 5%で検定を行い，有意と判定されたことを表す．さらに，‡ マークは MNMT+SCA と比較して有意水準 5%で検定を行い，有意と判定されたことを表す．

有意差もあることが確認できる．MNMT+SVA+SCA では，英独実験については MNMT+SVA や MNMT+SCA と比較してそれぞれ同程度スコアが向上しており，教師付き視覚的注意機構と教師付き言語間注意機構を組み合わせて利用することの有効性が確認できた．また，英日および日英実験については，MNMT+SCA では MNMT と比較して同程度もしくは下がる結果であるが，教師付き視覚的注意機構と合わせた MNMT+SVA+SCA では MNMT+SVA を上回っており，教師付き言語間注意機構と教師付き視覚的注意機構を組み合わせることの有効性が確認できた．

次に，人手による対応関係を用いた実験結果を見ると，NMT+SCA（人手）および MNMT+SCA（人手）では MGIZA によって対応関係を取得した NMT+SCA および MNMT+SCA と比較して性能が向上していることが確認できた．しかし，MNMT+SVA+SCA と MNMT+SVA+SCA（人手）を比較すると，MNMT+SVA+SCA（人手）のスコアは同程度かもしくは少し下がる結果となった．この結果より，人手で付けられた対応関係を用いて言語間注意機構を教師付き学習する際には，教師付き視覚的注意機構を同時に用いることの相乗効果は見られなかった．

5　考察

5.1　視覚的注意の例

図 6 と図 7 はそれぞれ英日翻訳におけるテストデータ中の単語 man と単語 lamp に対する視覚的注意を表している．これらの図において，より暗くなっている部分はより強く注意が向けられていることを表している．図を見ると，視覚的注意機構に制約を与えない通常の視覚的注意機構では画像全体に等しく注意が向けられている（図 6(b)，図 7(b)）．これに対し，制約を与えて学習させた教師付き視覚的注意機構では，それぞれ対応する領域に注意が向けられており，視覚的注意が改善する事例があることが確認できた（図 6(c)，図 7(c)）．これらの結果より，教師付き視覚的注意機構が，各単語の注意を関連する領域に向けさせた可能性があると考えられる．

(a) 正解データ　　(b) 制約なし　　(c) 制約あり

図 6　単語 man に対する視覚的注意

(a) 正解データ　(b) 制約なし　(c) 制約あり

図 7　単語 lamp に対する視覚的注意

5.2　翻訳例

図 8 は日英翻訳と独英翻訳のテストデータに対する翻訳結果の例を表している．図を見ると，教師付き注意機構を用いないマルチモーダル機械翻訳モデル (MNMT) によって翻訳された文は，原言語文のいくつかの情報が抜け落ちていることが分かる．例えば，図 8(a) では，画像内の男性が持っている「携帯電話」および男性の状態を表す「話しながら」という情報が抜け落ちている．また，図 8(b) では，「a yellow shovel」という情報が，図 8(c) では画像内の男性の特徴である「dreadlocks」という情報が抜け落ちている．これに対し，教師付き視覚的注意機構を用いたモデル (MNMT+SVA) ではこれらの抜け落ちていた情報を正しく翻訳できる事例があることが確認できた．これは，教師付き視覚的注意機構によって原言語文の各単語と画像内の関連する領域が対応付けられた可能性があると考えられる．

6　関連研究

ニューラル機械翻訳は，原言語の単語と目的言語の単語間での自動もしくは人手による対応関係に基づいて言語間注意機構を訓練することによって，その性能が改善されている．Liu ら (Liu et al. 2016) や Mi ら (Mi et al. 2016) は RNN ベースのニューラル機械翻訳モデルに，Garg ら (Garg et al. 2019) は Transformer ベースのニューラル機械翻訳モデルにおいて制約付き言語間注意機構を提案している．

ニューラル機械翻訳の性能改善に画像が有効であることが示されている (Elliott 2018; Caglayan et al. 2019)．マルチモーダルニューラル機械翻訳の研究も盛んに行われ，様々なモデルが提案されている．初期のころは，RNN ベースのニューラル機械翻訳 (Bahdanau et al. 2015) を拡張させた RNN ベースのマルチモーダルニューラル機械翻訳 (Calixto et al. 2017; Caglayan et al. 2017; Delbrouck and Dupont 2017) が主流であった．近年は，Transformer ベースのマルチモーダルニューラル機械翻訳の研究が盛んに行われている (Helcl et al. 2018; Libovický et al. 2018;

Source: an asian man walking down an alley talking on his cellphone with brightly colored fabric behind him .

MNMT: 明るい色の布を背にして路地を歩いているアジア人男性。

MNMT+SVA: 明るい色の布を背にして携帯電話で話しながら路地を歩くアジア人男性。

Reference: 明るい色の布を背に携帯電話で話しながら路地を歩くアジア系の男性。

(a) 英日翻訳の例

Source: 赤いシャツを着た男の子が、黄色いシャベルで砂を掘っている。

MNMT: a boy in a red shirt is shoveling sand .

MNMT+SVA: a boy in a red shirt is digging in the sand with a yellow shovel .

Reference: a boy wearing a red shirt digs into the sand with a yellow shovel .

(b) 日英翻訳の例

Source: ein rothaariger mann mit dreadlocks sitzt und spielt auf einer akustischen gitarre .

MNMT: a man with red-hair sits on an acoustic guitar .

MNMT+SVA: a red-haired man with dreadlocks is sitting and playing an acoustic guitar .

Reference: a red-haired man with dreadlocks is sitting playing and acoustic guitar .

(c) 独英翻訳の例

図 8　翻訳例

Grönroos et al. 2018; Ive et al. 2019; Zhang et al. 2020)．ほとんどのマルチモーダルニューラル機械翻訳モデルでは，視覚的注意機構によって画像の特徴を組み込んでいる．原言語文の単語と画像領域との関係を捉えるために視覚的注意機構を利用している研究 (Delbrouck and Dupont 2017; Zhang et al. 2020) や，目的言語文の単語と画像領域を捉えるために視覚的注意機構を利用している研究 (Calixto et al. 2017; Helcl et al. 2018; Libovický et al. 2018; Ive et al. 2019; Takushima et al. 2019) がある．なお，これらの研究で利用されている視覚的注意機構は訓練時に自動的に学習が行われており，視覚的注意機構に制約を加えた手法ではない．

7 結論

本稿では，教師付き視覚的注意機構を用いるマルチモーダルニューラル機械翻訳モデルを提案した．提案手法では，画像領域とその説明文中の単語との間に人手で付けられている対応関係を用いて視覚的注意機構の教師データを作成し，その教師データによってマルチモーダルニューラル機械翻訳モデルのエンコーダ内の視覚的注意機構に制約を与えて学習する．実験では，Multi30k データセットを用いた英独翻訳および独英翻訳と Flickr30k Entities JP データセットを用いた英日翻訳および日英翻訳を行い，提案手法によって Transformer ベースのマルチモーダルニューラル機械翻訳モデルの性能が改善できることを確認した．今後は，本実験で用いた Transformer ベースのマルチモーダルニューラル機械翻訳モデル以外のモデルに対しても，提案手法である制約を与えた視覚的注意機構が有効であるかどうかを検証していきたい．

謝 辞

本論文は国際会議 The 28th International Conference on Computational Linguistics (COLING2020) に採択された論文 (Nishihara et al. 2020) に基づいて日本語で書き直し，説明や評価を追加したものである．

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 JP20K19864 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

Ba, J., Kiros, J. R., and Hinton, G. E. (2016). “Layer Normalization.” *ArXiv*, **abs/1607.06450**.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). “Neural Machine Translation by Jointly Learning to Align and Translate.” In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*.

Barrault, L., Bougares, F., Specia, L., Lala, C., Elliott, D., and Frank, S. (2018). “Findings of the Third Shared Task on Multimodal Machine Translation.” In *Proceedings of the 3rd Conference on Machine Translation: Shared Task Papers*, pp. 304–323.

Caglayan, O., Aransa, W., Bardet, A., García-Martínez, M., Bougares, F., Barrault, L., Masana, M., Herranz, L., and van de Weijer, J. (2017). “LIUM-CVC Submissions for WMT17 Multimodal Translation Task.” In *Proceedings of the 2nd Conference on Machine Translation*, pp. 432–439, Copenhagen, Denmark. Association for Computational Linguistics.

Caglayan, O., Madhyastha, P., Specia, L., and Barrault, L. (2019). “Probing the Need for Visual Context in Multimodal Machine Translation.” In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4159–4170, Minneapolis, Minnesota. Association for Computational Linguistics.

Calixto, I., Liu, Q., and Campbell, N. (2017). “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, Vancouver, Canada. Association for Computational Linguistics.

Delbrouck, J.-B. and Dupont, S. (2017). “Modulating and Attending the Source Image During Encoding Improves Multimodal Translation.” *CoRR*, **abs/1712.03449**.

Denkowski, M. and Lavie, A. (2014). “Meteor Universal: Language Specific Translation Evaluation for Any Target Language.” In *Proceedings of the 9th Workshop on Statistical Machine Translation*, pp. 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Elliott, D. (2018). “Adversarial Evaluation of Multimodal Machine Translation.” In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2974–2978, Brussels, Belgium. Association for Computational Linguistics.

Elliott, D., Frank, S., Sima’an, K., and Specia, L. (2016). “Multi30K: Multilingual English-German Image Descriptions.” In *Proceedings of the 5th Workshop on Vision and Language*, pp. 70–74.

Gao, Q. and Vogel, S. (2008). “Parallel Implementations of Word Alignment Tool.” In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pp. 49–57, Columbus, Ohio. Association for Computational Linguistics.

Garg, S., Peitz, S., Nallasamy, U., and Paulik, M. (2019). “Jointly Learning to Align and Translate with Transformer Models.” In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4452–4461, Hong Kong, China. Association for Computational Linguistics.

Grönroos, S.-A., Huet, B., Kurimo, M., Laaksonen, J., Merialdo, B., Pham, P., Sjöberg, M., Sulubacak, U., Tiedemann, J., Troncy, R., and Vázquez, R. (2018). “The MeMAD Submission to the WMT18 Multimodal Translation Task.” In *Proceedings of the 3rd Conference on Machine Translation: Shared Task Papers*, pp. 603–611, Belgium, Brussels. Association for Computational Linguistics.

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

Helcl, J., Libovický, J., and Variš, D. (2018). "CUNI System for the WMT18 Multimodal Translation Task." In *Proceedings of the 3rd Conference on Machine Translation: Shared Task Papers*, pp. 616–623.

Ive, J., Madhyastha, P., and Specia, L. (2019). "Distilling Translations with Visual Awareness." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6525–6538, Florence, Italy. Association for Computational Linguistics.

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

Koehn, P. (2004). "Statistical Significance Tests for Machine Translation Evaluation." In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 388–395, Barcelona, Spain. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). "Moses: Open Source Toolkit for Statistical Machine Translation." In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pp. 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Libovický, J., Helcl, J., and Mareček, D. (2018). "Input Combination Strategies for Multi-Source Transformer Decoder." In *Proceedings of the 3rd Conference on Machine Translation: Research Papers*, pp. 253–260, Brussels, Belgium. Association for Computational Linguistics.

Liu, L., Utiyama, M., Finch, A., and Sumita, E. (2016). "Neural Machine Translation with Supervised Attention." In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 3093–3102.

Luong, T., Pham, H., and Manning, C. D. (2015). "Effective Approaches to Attention-based Neural Machine Translation." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421.

Mi, H., Wang, Z., and Ittycheriah, A. (2016). "Supervised Attentions for Neural Machine Translation." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2283–2288, Austin, Texas. Association for Computational Linguistics.

Nakayama, H., Tamura, A., and Ninomiya, T. (2020). "A Visually-Grounded Parallel Corpus with Phrase-to-Region Linking." In *Proceedings of The 12th Language Resources and Evaluation*

Conference, pp. 4204–4210, Marseille, France. European Language Resources Association.

Neubig, G., Nakata, Y., and Mori, S. (2011). "Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 529–533, Portland, Oregon, USA. Association for Computational Linguistics.

Nishihara, T., Tamura, A., Ninomiya, T., Omote, Y., and Nakayama, H. (2020). "Supervised Visual Attention for Multimodal Neural Machine Translation." In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4304–4314, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Och, F. J. and Ney, H. (2003). "A Systematic Comparison of Various Statistical Alignment Models." *Computational Linguistics*, **29** (1), pp. 19–51.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). "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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2017). "Flickr30K Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models." *International Journal of Computer Vision*, **123** (1), pp. 74–93.

Sennrich, R., Haddow, B., and Birch, A. (2016). "Neural Machine Translation of Rare Words with Subword Units." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Takushima, H., Tamura, A., Ninomiya, T., and Nakayama, H. (2019). "Multimodal Neural Machine Translation Using CNN and Transformer Encoder." In *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2019)*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). "Attention Is All You Need." In *Advances in Neural Information Processing Systems 30*, pp. 5998–6008.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). "From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference Over Event Descriptions." *Transactions of the Association for Computational Linguistics*, **2**, pp. 67–78.

Zhang, Z., Chen, K., Wang, R., Utiyama, M., Sumita, E., Li, Z., and Zhao, H. (2020). "Neural Machine Translation with Universal Visual Representation." In *8th International Conference*

on Learning Representations, ICLR 2020, Apr 26–May 1.

略歴

西原　哲郎：2020 年愛媛大学工学部情報工学科卒業．2020 年より同大学院理工学研究科修士課程に在学．

田村　晃裕：2005 年東京工業大学工学部情報工学科卒業．2007 年同大学院総合理工学研究科修士課程修了．2013 年同大学院総合理工学研究科博士課程修了．日本電気株式会社，国立研究開発法人情報通信研究機構にて研究員として務めた後，2017 年より愛媛大学大学院理工学研究科助教，2020 年より同志社大学理工学部准教授となり，現在に至る．博士（工学）．言語処理学会，情報処理学会，人工知能学会，ACL 各会員．

二宮　　崇：1996 年東京大学理学部情報科学科卒業．1998 年同大学大学院理学系研究科修士課程修了．2001 年同大学大学院理学系研究科博士課程修了．同年より科学技術振興事業団研究員．2006 年より東京大学情報基盤センター講師．2010 年より愛媛大学大学院理工学研究科准教授，2017 年同教授．博士（理学）．言語処理学会，アジア太平洋機械翻訳協会，情報処理学会，人工知能学会，電子情報通信学会，日本データベース学会，ACL 各会員．

表　悠太郎：2019 年愛媛大学工学部情報工学科卒業．2019 年より同大学院理工学研究科修士課程に在学．

中山　英樹：2006 年東京大学工学部機械情報工学科卒業．2011 年同大学大学院情報理工学系研究科知能機械情報学専攻博士課程修了．博士（情報理工学）．2012 年より東京大学大学院情報理工学系研究科創造情報学専攻講師．2018 年より同准教授，現在に至る．画像認識，自然言語処理，機械学習の研究に従事．情報処理学会，電子情報通信学会，IEEE, ACM 各会員．

（2020 年 11 月 1 日　受付）
（2021 年 2 月 8 日　再受付）
（2021 年 3 月 10 日　採録）