# 潜在拡散モデルによる変換画像を用いるマルチモーダルニューラル機械翻訳

湯浅 亮也 [1]　田村 晃裕 [1]　梶原 智之 [2]　二宮 崇 [2]　加藤 恒夫 [1]
[1] 同志社大学　[2] 愛媛大学
[1]{ctwh0190@mail4,aktamura@mail,tsukato@mail}.doshisha.ac.jp
[2]{kajiwara,ninomiya}@cs.ehime-u.ac.jp

## 概要

本研究では，潜在拡散モデルによる画像変換で生成した擬似画像を用いる新しいマルチモーダルニューラル機械翻訳（MNMT）を提案する．MNMTは，原言語文と関連画像に基づき翻訳を行うが，関連画像は必ずしも原言語文の内容のみを表しているわけではなく，関連画像内には原言語文とは無関係な箇所も存在する場合が多い．そのため，関連画像は翻訳の補助情報として最適とは限らない．そこで提案手法では，原言語文に基づき関連画像を潜在拡散モデルで画像変換することで，原言語文の内容に沿った疑似画像を生成し，生成した疑似画像を用いて翻訳を行う．Multi30k データセットを用いた英独翻訳実験を行い，提案手法の有効性を確認した．

## 1　はじめに

近年，機械翻訳の分野では，原言語文に加えて関連画像を翻訳に使用する MNMT [1] が注目されている．関連画像により，翻訳時の曖昧性を解消したり，テキスト情報だけでは捉えることが困難な情報を補完したりすることで，翻訳性能の改善が期待されている．MNMT では，通常，画像キャプション生成のデータセットを多言語に拡張した原言語文と目的言語文，関連画像の 3 つ組データからなるデータセットを使用する．しかし，画像には様々な内容が含まれており，原言語文は関連画像が表す内容の一部になっている場合が多く，原言語文とは無関係な箇所が関連画像に存在する場合も多い．図 1 に MNMT の標準的なデータセットである Multi30k データセット [2] の実例を示す．図 1 のように，Multi30k では一つの画像に内容の異なる複数の原言語文が関連づけられている．そして，例えば，原言語文 2 には関連画像に写っている家に関する内容が含まれていない．そのため，関連画像は翻訳の補助情報として最適とは限らない．

そこで本研究では，潜在拡散モデルによる画像変換で生成した擬似画像を用いる新たな MNMT を提案する．具体的には，関連画像と原言語文を用いて潜在拡散モデルにより画像変換を行い，関連画像から原言語文とは無関係な内容を排除し，関連画像を原言語文に即した内容に変換する．そして，変換された擬似画像を関連画像として用いて MNMT モデルで翻訳を行う．このように原言語文の内容をより反映した関連画像を翻訳の補助情報をして用いることで翻訳の性能改善を試みる．

提案手法の有効性を，Multi30k [2] と Ambiguous COCO [3] を用いた英独翻訳タスクで検証した．その結果，提案手法により，Multi30k の教師データをそのまま用いる従来の MNMT と比較して，Multi30k Test2016 と Test2017 でそれぞれ 0.14，Ambiguous COCO で 0.39，BLEU スコアが改善することを確認した．また，CLIPScore [4] を用いて関連画像と原言語文の類似度を算出した結果，提案手法で用いた関連画像の方が，変換前の関連画像よりも原言語文に即した画像であることを確認した．

## 2　従来 MNMT モデル

近年の MNMT では，Transformer [5] ベースの MNMT モデルが主流となっている．その翻訳性能を改善するため，これまで，視覚的注意機構の導入 [6] や共有エンコーダを用いてテキストと画像の特徴表現を同時に学習する手法 [7] など様々な試みがなされている．その一つに，関連画像のパッチと原言語文の単語の関連を捉える注意機構である Selective Attention を用いる Transformer MNMT [8] が提案されている．本研究ではベースの MNMT モデルとして，この Selective Attention MNMT モデルを

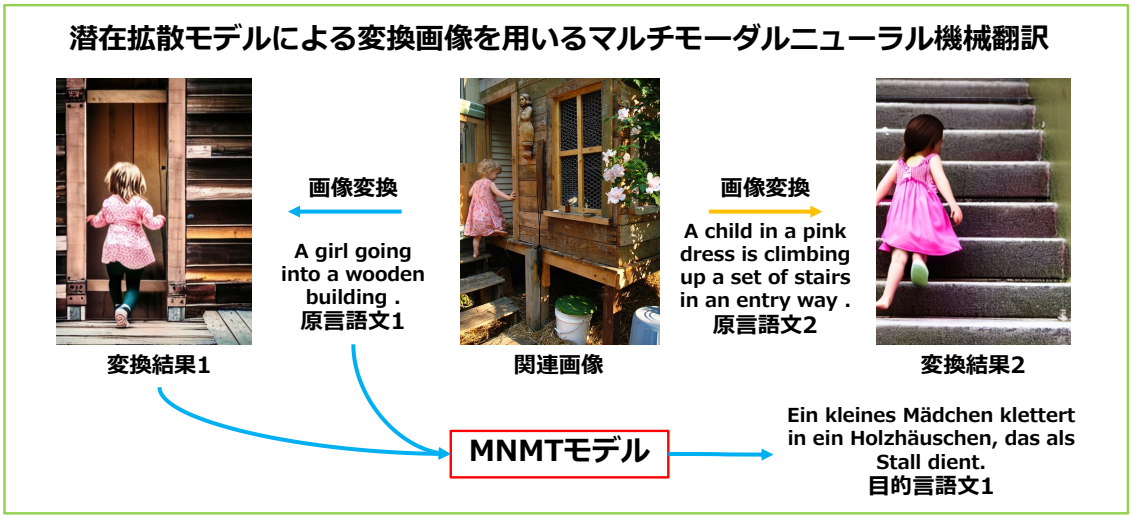


**図 1**　提案手法の概要図

使用する．以降では，Selective Attention MNMT モデルを概説する．

Selective Attention MNMT モデルでは，まず，原言語文 $X^{\text{text}}$ と関連画像 $X^{\text{img}}$ をそれぞれ，式 (1) と式 (2) で特徴表現 $H^{\text{text}}$ と $H^{\text{img}}$ に変換する．

$$H^{\text{text}} = \text{TextEncoder}(X^{\text{text}}) \tag{1}$$

$$H^{\text{img}} = W\,\text{ImageEncoder}(X^{\text{img}}) \tag{2}$$

ここで，$W$，TextEncoder，ImageEncoder は，それぞれ，パラメータ行列，Transformer Encoder，Vision Transformer [9] である．

そして，式 (3) の通り，Selective Attention で画像のパッチと原言語の単語の関連を注意機構で捉える．

$$H^{\text{img}}_{attn} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

ここで，$Q$，$K$，$V$ は，それぞれ，$H^{\text{text}}$，$H^{\text{img}}$，$H^{\text{img}}$ であり，$d_k$ は $H^{\text{text}}$ の次元数である．

その後，Gated Fusion 機構 [10] により，関連画像の影響度を制御して原言語文と関連画像を表す特徴ベクトル $H^{\text{out}}$ を生成する．具体的には，原言語文と画像の特徴ベクトルから式 (4) を用いて関連画像の影響度をコントロールする $\lambda$ を算出し，式 (5) のように $\lambda$ で重み付けした特徴表現 $H^{\text{out}}$ を生成する．

$$\lambda = \text{Sigmoid}(UH^{\text{text}} + VH^{\text{img}}_{attn}) \tag{4}$$

$$H^{\text{out}} = (1-\lambda)\cdot H^{\text{text}} + \lambda \cdot H^{\text{img}}_{attn} \tag{5}$$

ここで，$U$ と $V$ は学習可能なパラメータ行列である．そして，$H^{\text{out}}$ を Transformer Decoder に入力し，翻訳文を生成する．

## 3　提案手法：潜在拡散モデルによる変換画像を用いる MNMT

本節では，原言語文に基づいて関連画像を変換した擬似画像を使って翻訳を行う MNMT モデルを提案する．図 1 に提案手法の概要を示す．

MNMT のデータセットは，原言語文と目的言語文，関連画像の 3 つ組データで構成される．従来使われているデータセットでは，各原言語文は関連画像が表す内容の一部になっている場合が多く，原言語文とは無関係な内容が関連画像に存在する場合も多い．例えば，図 1 の画像は，ピンク色のドレスを着た少女が階段を上って木造の家に入っている場面を表しているが，原言語文 1 には階段を上っているという内容は含まれていない．また，原言語文 2 には家に関する内容は含まれていない．そのため，関連画像は翻訳の手がかりとして最適とは限らない．

そこで提案手法は，まず，潜在拡散モデルにより，関連画像から原言語文とは無関係な内容を排除して原言語文に即した疑似画像を生成する．そして，生成した疑似画像と原言語文を用いて従来の MNMT モデルにより翻訳を行う．これにより，関連画像とテキストの関連が翻訳時に捉えやすくなり，翻訳性能が向上することが期待できる．3.1 節では，関連画像を変換する潜在拡散モデルを説明する．ベースとなる従来 MNMT モデルは，実験では，2 節で説明した Selective Attention MNMT モデルを用いた．

### 3.1　画像変換：潜在拡散モデル

提案手法の画像変換で使用する潜在拡散モデル [11] を説明する．潜在拡散モデルは，VAE [12] の潜

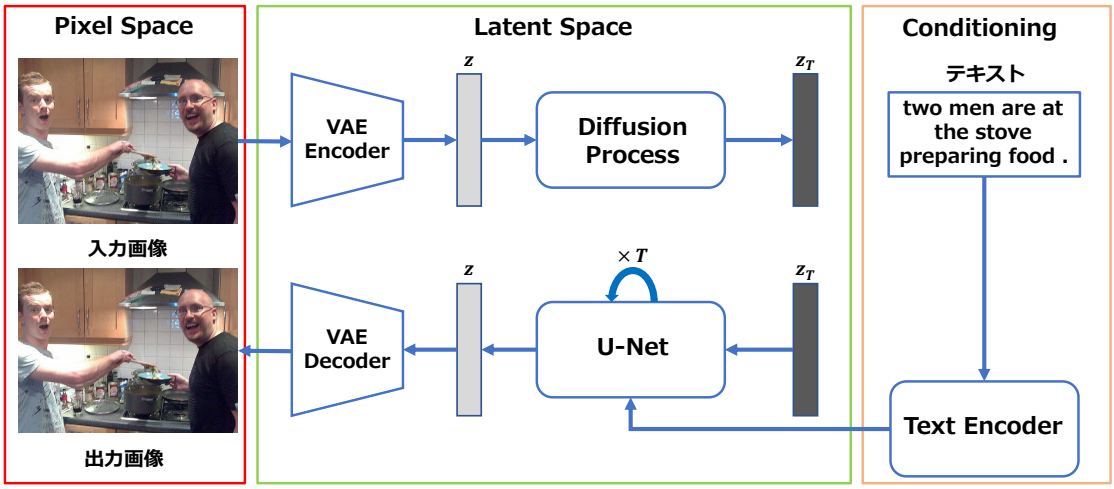


**図 2**　潜在拡散モデルの学習の概要図

在空間に対して拡散モデル [13] を適用させたモデルで，主に VAE，U-Net [14]，Text Encoder で構成される（図 2 参照）．潜在拡散モデルでは，VAE Encoder を用いて入力画像をピクセル空間から低次元の潜在空間に射影し，その潜在表現を得る．そして，拡散プロセスで潜在表現に対してガウシアンノイズを連続的に付与する．その後，逆拡散プロセスで，ノイズが含まれた潜在表現からノイズを除去する．その際，U-Net に対して，Text Encoder で変換したテキストの特徴表現で条件付けを行う．この条件付けは，Cross Attention で実現する．このプロセスを複数回実施することで，ノイズが含まれた潜在表現から徐々にノイズを除去する．最後に，VAE Decoder により，ノイズが除去された潜在表現を潜在空間からピクセル空間に射影して出力画像を得る．

損失関数は式 (6) の通りである．

$$L_{\mathrm{LDM}} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2] \quad (6)$$

ここで，$\mathcal{E}$，$\epsilon_\theta$，$\tau_\theta$，は，それぞれ，VAE Encoder，U-Net，Text Encoder を表し，$x$，$y$，$\epsilon$，$t$，$z_t$，は，入力画像，テキスト，ガウシアンノイズ，時刻，時刻 $t$ の潜在表現である．

提案手法では，この潜在拡散モデルにおいて，関連画像と原言語文を，それぞれ，VAE Encoder と Text Encoder の入力にすることで，関連画像を原言語文に即した画像に変換する．

# 4　実験

## 4.1　実験設定

本実験では，Multi30k と Ambiguous COCO を用いた英独翻訳タスクで提案手法の有効性を検証した．学習・検証データとして，Multi30k の学習データ（29,000 組）と検証データ（1,014 組）を使用し，テストデータには，Multi30k の Test2016（1,000 組）と Test2017（1,000 組），Ambiguous COCO（461 組）を使用した．

本実験では，3 節で説明した提案手法の翻訳性能を，関連画像を用いない NMT モデル，関連画像としてデータセットの画像をそのまま使用した MNMT モデル，原言語文のみから生成した画像を関連画像として使用した MNMT モデルの翻訳性能と比較した．以降では，各モデルをそれぞれ，MNMT（変換画像），NMT，MNMT（原画像），MNMT（生成画像）と記す．

NMT モデルは，Transformer-Tiny [1] を用いた．このモデルは，通常の Transformer モデルよりも，レイヤー数や隠れ層のサイズ，注意機構のヘッド数などを削減した小規模なデータセットに適した NMT モデルである [2]．ハイパーパラメータは，先行研究 [15] の実験設定に倣い，エンコーダとデコーダのレイヤー数を 4，隠れ層のサイズを 128，順伝播層の入力サイズを 256，注意機構のヘッド数を 4，Dropout を 0.3，label smoothing weight を 0.1 とした．最適化手法は Adam [16] を使用し，$\beta_1 = 0.9, \beta_2 = 0.98$ とした．また，学習率は，最初の 2,000 ステップで $1e^{-7}$ から $5e^{-3}$ まで warmup を線形に行い，その後，学習回数に応じて学習率を減少させた．語彙辞書は原言語と目的言語で共有し，Byte Pair Encoding [17] によりマージ操作を 10,000 回として作成した．

MNMT モデルは，2 節で説明した Selective Attention

1）https://github.com/LividWo/Revisit-MMT
2）文献 [15] では，Multi30k データセットにおいて Transformer Base/Small よりも翻訳性能が高いことが報告されている．

表 1　実験結果：英独翻訳（BLEU[%]）

| モデル | Test 2016 | Test 2017 | Ambiguous COCO |
|---|---|---|---|
| NMT | 40.50 | 31.31 | 27.81 |
| MNMT（原画像） | 41.06 | 32.06 | 27.91 |
| MNMT（生成画像） | 40.81 | 31.81 | **28.54** |
| MNMT（変換画像） | **41.20** | **32.20** | 28.30 |

MNMT [3] を用いた．画像の特徴抽出には，Vision Transformer の vit_base_patch16_384 [4] を使用した．MNMT（生成画像）や MNMT（変換画像）で行う関連画像の生成や画像変換には，潜在拡散モデルをベースとする Stable Diffusion [5] を採用し，モデルは stable-diffusion-v1-5 [6] を使用した．実装には，それぞれ，diffusers [7] の StableDiffusionPipeline と StableDiffusionImg2ImgPipeline を用いた．画像生成では，デフォルトのパラメータを使用し，guidance_scale を 7.5，num_inference_steps を 50 として原言語文から画像を生成した．また，画像変換でもデフォルトのパラメータを使用し，strength を 0.8，guidance_scale を 7.5，初期値を関連画像として原言語文を用いて画像変換した．学習時のハイパーパラメータと最適化手法，語彙辞書作成方法は，NMT モデルの実験設定と同じである．

全モデルにおいて，学習終了直前の 10 エポックのモデルに対して checkpoint averaging を行い，ビーム幅 5 のビーム探索により翻訳文を生成した．評価指標は，BLEU [18] を使用した．また，異なる 5 つの random seed で学習・検証し，最も検証データに対する BLEU が高いモデルを評価した．

### 4.2　実験結果

表 1 に実験結果を示す．表 1 より，画像情報を使用しない NMT モデルより，画像情報を使用する 3 種類の MNMT モデルの方が全てのデータセットにおいて BLEU が高いことが分かる．これより，本実験で用いたデータセットにおいては，画像情報が翻訳性能の改善に寄与することが確認できた．

また，MNMT モデル間の比較では，Test2016 と Test2017 においては提案手法である MNMT（変換画像）が最も高い翻訳性能を達成した．Ambiguous COCO においては，MNMT（生成画像）の方が MNMT（変換画像）よりも高い翻訳性能となったが，全体的には，MNMT（変換画像）の方が良い結果となり，提案手法の有効性を確認した．

3）https://github.com/libeineu/fairseq_mmt
4）https://github.com/rwightman/pytorch-image-models
5）https://github.com/CompVis/stable-diffusion
6）https://huggingface.co/runwayml/stable-diffusion-v1-5
7）https://github.com/huggingface/diffusers

表 2　CLIPScore：原言語文と関連画像の類似度

| モデル | Test 2016 | Test 2017 | Ambiguous COCO |
|---|---|---|---|
| MNMT（原画像） | 79.59 | 78.32 | 78.17 |
| MNMT（変換画像） | 79.74 | 79.35 | 80.08 |

## 5　考察

本節では，提案手法で用いた疑似画像（潜在拡散モデルにより原言語文に基づき関連画像を変換した画像）を分析する．変換した画像の例は付録 A に示す．MNMT で使用した画像が原言語文をどれだけ反映していたかを調査するため，使用した画像と原言語文の類似度を，式（7）の通り算出される CLIPScore [4] を用いて評価する．

$$\mathrm{CLIPScore}(\boldsymbol{c}, \boldsymbol{v}) = w \cdot \max(\cos(\boldsymbol{c}, \boldsymbol{v}), 0) \qquad (7)$$

ここで，$\boldsymbol{c}$ と $\boldsymbol{v}$ は，それぞれ，CLIP [19] の Text Encoder と Image Encoder の特徴ベクトルである．また，$w$ は出力をリスケーリングするために使用され，ここでは先行研究 [4] に倣い 2.5 とする．

表 2 に評価結果を示す．表 2 より，全てのデータセットにおいて，提案手法で画像変換した擬似画像の方が，元々の関連画像よりも原言語文との類似度が高いことが分かる．特に，曖昧性を含む Ambiguous COCO では類似度が改善し，CLIPScore が 1.91 向上した．これらの結果から，提案手法では原言語をより反映した関連画像を翻訳の手がかりとして利用できることを確認した．

## 6　おわりに

本研究では，潜在拡散モデルにより関連画像を原言語文に即した画像に変換し，変換した画像を用いる新たな MNMT を提案した．Multi30k を用いた英独翻訳タスクによる実験を行い，提案手法の方が従来手法より高い翻訳性能を実現できることを確認し，提案手法の有効性を確認した．また，MNMT で使用した画像と原言語文の類似度を CLIPScore で評価し，提案手法で用いた画像の方が元の画像よりも原言語文に類似した画像であることを確認した．今後は英独以外の言語対に対しても有効性を検証したい．

## 謝辞

本研究の一部は JSPS 科研費 JP22K12177, JP21K12031 の助成を受けたものである．また，本研究成果の一部は，国立研究開発法人情報通信研究機構の委託研究 (No. 225) により得られたものである．ここに謝意を表する．

## 参考文献

[1] Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. A shared task on multimodal machine translation and crosslingual image description. In **Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers**, pp. 543–553, 2016.

[2] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30K: Multilingual English-German image descriptions. In **Proceedings of the 5th Workshop on Vision and Language**, pp. 70–74, 2016.

[3] Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. Findings of the second shared task on multimodal machine translation and multilingual image description. In **Proceedings of the Second Conference on Machine Translation**, pp. 215–233, 2017.

[4] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 7514–7528, 2021.

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

[6] 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.

[7] Desmond Elliott and Ákos Kádár. Imagination improves multimodal translation. In **Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 130–141, 2017.

[8] Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. On vision features in multimodal machine translation. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 6327–6337, 2022.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In **International Conference on Learning Representations**, 2021.

[10] Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. Neural machine translation with universal visual representation. In **International Conference on Learning Representations**, 2020.

[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 10684–10695, 2022.

[12] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In **2nd International Conference on Learning Representations**, 2014.

[13] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In **Proceedings of the 32nd International Conference on Machine Learning**, Vol. 37 of **Proceedings of Machine Learning Research**, pp. 2256–2265, 2015.

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In **Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015**, pp. 234–241, 2015.

[15] Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 6153–6166, 2021.

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In **International Conference on Learning Representations (Poster)**, 2015.

[17] 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 (Volume 1: Long Papers)**, pp. 1715–1725, 2016.

[18] 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.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In **Proceedings of the 38th International Conference on Machine Learning**, Vol. 139 of **Proceedings of Machine Learning Research**, pp. 8748–8763, 2021.

# A　付録

**原言語文**
**a man grilling meat on an outdoor grilling pit .**

**原言語文**
**a man wearing black and white stripes is trying to stop a horse .**

**原言語文**
**a young girl in a red dress is wearing a black cowboy hat .**

**原言語文**
**one man holds another man's head down and prepares to punch him in the face .**

**図 3**　原言語文を用いて Multi30k の関連画像を画像変換した場合の成功例（左）と失敗例（右）