マルチモーダル機械翻訳の技術動向調査

A Survey of Technological Trends in Multi-Modal Machine Translation

愛媛大学　大学院理工学研究科　教授
二宮　崇

2001 年東京大学大学院理学系研究科情報科学専攻博士課程修了。博士 (理学)。2017 年より愛媛大学大学院理工学研究科教授。自然言語処理の研究に従事。

ninomiya.takashi.mk@ehime-u.ac.jp　TEL 089-927-9954

1 はじめに

深層学習技術の進展に伴い、LSTM 等の RNN ベースの機械翻訳 (Sutskever et al. 2014; Bahdanau et al. 2015; Luong et al. 2015) や Transformer ベースの機械翻訳 (Vaswani et al. 2017) が発展し、従来の統計機械翻訳を超える高い翻訳精度と流暢性を有する機械翻訳が実現されるようになった。現在主に用いられている機械翻訳では、翻訳元言語 (原言語) と翻訳先言語 (目的言語) の翻訳例 (パラレルコーパス) から機械翻訳モデルの学習を行っているが、これらの学習にはテキストデータだけが用いられており、画像などのその他の情報源は扱われてこなかった。しかし、画像情報を用いることができれば、例えば、「Crane」（クレーン、鶴）や「Pepper」（胡椒、ピーマン、パプリカ、チリペッパーなど）、「Bank」（銀行、土手）のように曖昧性のある単語であっても、画像情報から正しく翻訳することができる。ある翻訳家に伺うと、例えば、観光のためのパンフレットを翻訳する場合、翻訳対象のテキストだけではなく、パンフレットの対象に関する写真を集めたり、さらに場合によっては実物そのものを入手したり、直接赴き目にすることで、対象に対する理解を深め、より良い翻訳を作成することができるのだという。このようにテキスト以外の情報を活用することでより良い翻訳を生成する機械翻訳が実現される可能性が十分考えられる。

2016 年頃から、画像等のマルチモーダル情報を用いたニューラル機械翻訳の研究が行われるようになり、そのような機械翻訳は「マルチモーダル機械翻訳」と呼ばれている。特に Multi30K (Elliott et al. 2016) と呼ばれる、画像とその対訳文対（英独）を約 31,000 件集めたマルチモーダルパラレルコーパスを対象に研究が進められ、大きく発展したが、その道のりは平坦なものではなかった。図 1 は Multi30K のデータ例を表す。英独翻訳のマルチモーダル機械翻訳の場合、これらの画像と英語文（原言語文）を入力として与え、ドイツ語文（目的言語文）を生成することが目的となる。最初の 2016 年は、画像情報を用いないテキスト情報だけを用いた機械翻訳の方が性能が高く、画像情報の有効性が疑問視された。

英語	a red-haired man with dreadlocks is sitting playing and acoustic guitar .
ドイツ語	ein rothaariger mann mit dreadlocks sitzt und spielt auf einer akustischen gitarre .

図 1　マルチモーダルパラレルコーパス（Multi30K）の例

2017 年以降は、画像情報を有効に活用することで、通常の機械翻訳よりも高い性能を実現できるようになっ

てきた。しかしながら、画像情報を活用したマルチモーダル機械翻訳には、いくつかの課題や疑問が指摘されている。まず、(i) 画像情報を処理するため、通常のテキストのみを扱う機械翻訳モデルと比べてモデルの規模が非常に大きくなりやすいという問題がある。それにもかかわらず、精度の向上はわずかであり、コストパフォーマンスの面で効率が非常に悪いという問題がある。次に、(ii) マルチモーダルパラレルコーパスとして広く使われている Multi30K（約３万件）は、通常の機械翻訳で用いられる大規模な対訳パラレルコーパス（数千万件規模）と比べて極めて小規模であり、機械翻訳のための学習用データとして不十分という指摘がある。つまり、機械翻訳の学習のためにテキストデータが絶対的に不足しており、そのため過去の研究では画像情報の有効性が確認できたのではないかということである。さらに、(iii) 数千万件規模のテキストデータから学習された機械翻訳モデルや、事前学習済みモデルに対し、Multi30K のテキスト部分だけを用いてファインチューニング（追加学習）を行うことで、十分に高性能な機械翻訳を実現できることが考えられる。つまり、原言語文には翻訳のために必要な情報が十分に載っており、大規模な対訳パラレルコーパスを用いて学習すれば、わざわざ画層情報まで用いなくても良いだろう、ということである。このように、マルチモーダル機械翻訳に対し、「本当に画像情報は役に立つのか？立っているのか？」ということが問われるようになってきた。

Wu らによる 2021 年の論文「Good for Misconceived Reasons: An Empirical Revisiting on the Need for Visual Context in Multimodal Machine Translation」(Wu et al. 2021) では、画像情報の必要性に対する辛辣な批判が展開されており、ゲーティング機構を有するマルチモーダル機械翻訳において画像情報の貢献が極めて小さいことを示し、さらに従来の画像情報による精度向上は正則化 (regularization) の効果と同程度であることを示した。特に後者については、画像情報の代わりに、ノイズ情報を入力しても同程度の精度向上が得られることが確認され、画像情報の有効性が疑問視されることになった。また、同じく 2021 年の Caglayan らによる論文「Cross-lingual Visual Pre-training for Multimodal Machine Translation」(Caglayan et al. 2021) では、マルチモーダル事前学習モデルを用いたマルチモーダル機械翻訳を提案しており、従来の Multi30K だけから学習したモデルに比べ非常に高い精度向上を実現し、事前学習モデルの有効性を示した。しかし、実験結果より画像情報による若干の性能向上が確認できるものの、テキストだけを扱う機械翻訳モデルとマルチモーダル機械翻訳の間に大きな性能差は無く、大規模マルチモーダルパラレルコーパスが存在する場合において、画像情報の有効性は疑わしいものとなった。

2021 年はマルチモーダル機械翻訳に対する否定的な研究が続いたが、しかしその後も多くの研究者によってマルチモーダル機械翻訳の研究は続けられ、さらに高い精度を実現するようになった。正則化効果に対する疑念については、実験において白色画像やノイズ画像、ランダムに選択してきた画像と比較することで払拭されてきており、マルチモーダル機械翻訳における画像の有効性が明らかになってきている。現在は、大規模言語モデル（Large Language Model (LLM)）の進展により、LLM を用いた機械翻訳の研究が盛んに行われている。画像情報を取り入れた Vison LM (VLM) を活用することで大規模言語モデル等の事前学習を用いたマルチモーダル機械翻訳が今後期待される。

マルチモーダル機械翻訳については、Shen と Shao らによるサーベイ論文「A Survey on Multi-modal Machine Translation: Tasks, Methods and Challenges」(Shen et al. 2024) にまとめられており、詳細についてはこちらを参照されたい。本稿では、このサーベイ論文に従って、マルチモーダル機械翻訳の分類と解説を行う。２節では、シーン画像マルチモーダル機械翻訳のモデルについて解説し、３節ではその学習手法について解説する。最後に４節で本稿のまとめを述べる。

２ シーン画像マルチモーダル機械翻訳モデル

マルチモーダル機械翻訳というと、本来、静止画像だけではなく、音、動画等の様々なモダリティを利用した機械翻訳を意味すると思われるが、Multi30K を中心に発展してきた経緯があって、マルチモーダル機械翻訳というと、何らかの状況を表す静止画像が一枚あり、それに対する説明文を翻訳するタスクとして認識されている。ここでは区別をするためにそういった Multi30K を用いたマルチモーダル機械翻訳を「シーン画像マルチモーダル機械翻訳」と呼ぶことにする。

2.1　ダブルアテンション機構

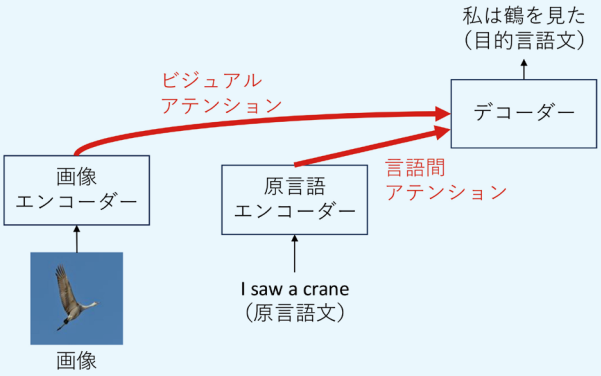

図２　ダブルアテンション機構

シーン画像マルチモーダル機械翻訳において最初に提案されたモデルがダブルアテンション機構によるマルチモーダル機械翻訳モデルである。このタイプのモデルは Caglayan らによって最初に提案された (Caglayan et al. 2016)。図２はダブルアテンション機構の概要図を示す。ダブルアテンション機構では、原言語文も画像も同じように扱い、デコーダーが翻訳文を生成する際に原言語文および画像の両方に対するアテンションをそれぞれ計算し、それらから得られるコンテキストベクトルを合成することで画像情報を組み入れた機械翻訳を実現する。

Caglayan らによるモデルでは、原言語文を双方向 GRU で解析し、画像を ResNet-50 で解析し、デコーダーはそれぞれの内部状態に対しアテンションの計算を行い、コンテキストベクトルを合成する。多くのマルチモーダル機械翻訳はこのようなダブルアテンション機構をベースにしてそれらを改良することによって高性能なマルチモーダル機械翻訳を実現している (Calixto et al. 2017; Libovický and Helcl 2017; Delbrouck and Dupont 2017b; Su et al. 2021; Zhao et al. 2022a; Zhao et al. 2022b)。

2.2　補助情報としての画像

ダブルアテンション機構では、原言語情報も画像情報も同じように扱い、それぞれに対してアテンションを計算するというモデルであった。その他の方法として、画像情報をテキストへの補助情報として用いる方法が考案されている。つまり、画像情報よりもテキスト情報がより重要であり、原言語の翻訳の際に補助的に画像情報も用いるという発想である。

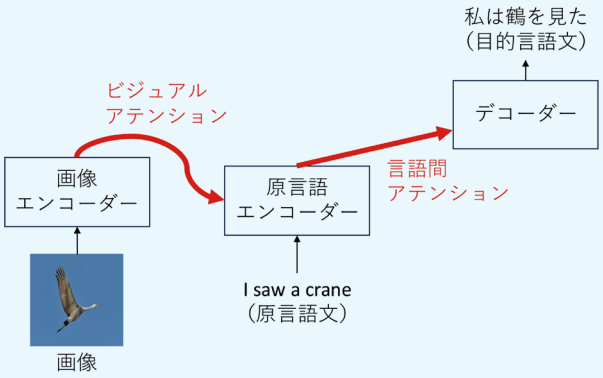

図３　エンコーダー部におけるビジュアルアテンション

Delbrouck と Dupont (2017a) は、エンコーダーにおけるビジュアルアテンションを提案している。図３はその概要図を示す。原言語に対するエンコーダーの計算を行う際に、画像エンコーダーに対しアテンションの計算を行い（ビジュアルアテンション）、原言語エンコーダーの内部状態の計算を行う。その後は通常の機械翻訳モデルとみなして、デコーダーは原言語エンコーダーに対し、言語間アテンションの計算を行う。Nishihara ら (2020) は教師付きアテンション制約を与えるマルチモーダル機械翻訳手法を提案しているが、彼らのモデルはこの Delbrouck と Dupont の手法をベースにしている。

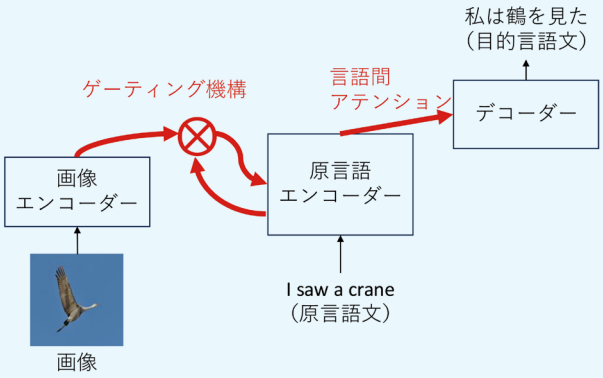

図４　ゲーティング機構によるマルチモーダル機械翻訳

Zhang ら (2020) や Wu ら (2021) は、マルチモーダル機械翻訳のエンコーダー部にゲーティング機構を用いたマルチモーダル機械翻訳を提案した。図４はゲーティング機構を用いたマルチモーダル機械翻訳の概要図を表す。ゲーティング機構は、２つの情報源があったときに、片方を主流、もう片方を支流とみなし、支流からの流入を制御しながら主流に支流を流し込む計算を行う融合手法である。ここでは、原言語情報を主流、画像情報を支流とみなし、画像情報をどれだけ流入させるかゲーティング機構により制御する。

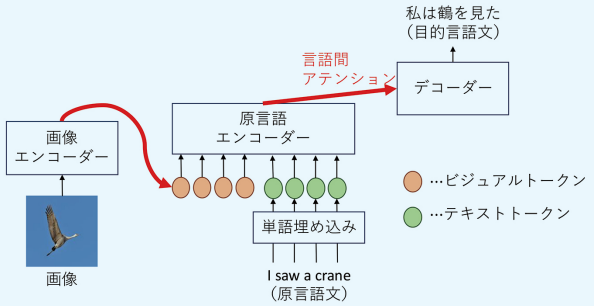

図５　ビジュアルトークンによるマルチモーダル機械翻訳

Huang ら (2016) は画像特徴量を原言語文と一緒にエンコーダーに与える手法を提案している（図 5）。物体領域に対する特徴量と画像全体の特徴量を原言語文の先頭に並べ、通常のエンコーダーデコーダーモデルを適用する。モデルが複雑にならないため、実装が比較的容易であり、自己アテンションを用いれば、アテンション機構を個別に用意する手法と同等の性能が期待できる。Vision Transformer や LLM との相性も良いため、VLM でもよく用いられている方式である。今後 LLM ベースのマルチモーダル機械翻訳において広く活用されることが期待される。

これらの手法は画像情報を積極的に利用するとは限らないので、通常のテキストベースの機械翻訳から性能を落としにくいという利点があり、ダブルアテンション機構と同様多く用いられている。

2.3　テキストからの画像生成と画像検索

2.1 節及び 2.2 節で説明したマルチモーダル機械翻訳手法はいずれも学習時及び推論時（翻訳時）に原言語文に関連する画像が手に入ることを想定している。しかし、翻訳する時に関連する画像が手に入るという想定は、現実問題として必ずしも一般的ではない。そこで、テキストから画像を生成する / 検索することで画像を得るマルチモーダル機械翻訳手法が提案されている。

Long ら (2021) は GAN を用いて、原言語文から画像特徴量を生成し、それらの画像特徴量を用いてマルチモーダル機械翻訳を行う手法を提案している。Yuasa ら (2023) は、潜在拡散モデルを用いて、オリジナル画像と原言語文から、より原言語文の内容に沿った画像を生成することでマルチモーダル機械翻訳の精度を上げる手法を提案している。ただし、Yuasa らの手法はオリジナル画像を必要とするため、原言語文だけで翻訳できるようになる手法にはなっていない。Tang ら (2022) は画像検索エンジンを用いて、原言語文に関連する画像を取得し、それらを用いてマルチモーダル機械翻訳の学習および翻訳を行う手法を提案した。

3　学習手法

次にマルチモーダル機械翻訳において用いられている学習手法について紹介する。Multi30K のように画像と原言語文、目的言語文の 3 つ組が揃ったコーパスが大量に存在すれば問題は無いが、そのようなコーパスは希少であり、大量に存在する対訳パラレルコーパスや、画像 - テキストを対とするキャプショニングデータを用いたマルチタスク学習や、教師なし学習、対訳動画データを利用したマルチモーダル機械翻訳が検討されている。

3.1. マルチタスク学習

複数のタスクに対し、単一のモデルで学習することをマルチタスク学習と呼ぶ。関連する複数のタスクを同時に学習することにより、単一のタスクだけ学習するよりも、より多くの学習が可能となる。

Elliott と Kádár による 2017 年の論文「Imagination Improves Multimodal Translation」(Elliott and Kádár 2017) において、対訳パラレルコーパスと単言語のキャプショニングデータ（画像とその説明文を対とするデータ）を用いたマルチタスク学習が提案されている。Elliott と Kádár による手法は、対訳パラレルコーパスを用いた通常の機械翻訳タスクと、画像特徴量をキャプションから生成するタスクの 2 つのタスクを想定しており、機械翻訳タスクにおける原言語エンコーダー部分と、画像特徴量生成タスクにおけるキャプションエンコーダー部分を共有することで、マルチタスク学習が行われる。この手法の利点としては、大量に存在する対訳パラレルコーパスおよびキャプショニングデータを利用して学習できるということ、および翻訳時には画像データを必要としないということがある。キャプショニングデータを用いているが、キャプションから画像を生成する必要はなく、何らかの画像エンコーダーを用いて得られる画像特徴量をキャプションから生成する学習だけ行えれば良い。

Nishihara ら (2020) は、マルチモーダル対訳コー

パスを用いたマルチモーダル機械翻訳の学習と、画像領域 - 単語間のアライメント情報を用いたアテンションの教師付き学習のマルチタスク学習を提案している。

Zuo ら (2023) は VQA を補助タスクとして、マルチモーダル機械翻訳とのマルチタスク学習を行う手法を提案している。VQA は Visual Question Answering の略であり、ある状況を示す画像に対し、その画像に関連する質問に対して回答するタスクのことである。彼らの手法では、LLM が活用されており、Multi30K の画像と説明文から、LLM を用いて自動的に擬似 VQA タスクデータを生成し、その VQA タスクとマルチモーダル機械翻訳のマルチタスク学習を行う。近年、LLM ベースの手法が急速に発展しており、自動的に生成された QA データや VQA データにより、LLM の学習を進める手法が盛んに研究されている。視覚言語モデル (VLM) において有効な手法として注目されている。

3.2　対照学習

対照学習は似ているものをベクトル空間において近づけ、異なるものを遠ざけることによって意味的な表現を学習する手法である。様々な自然言語処理タスクにおいて対照学習の研究が進んでおり、画像と言語の間の意味的関連性を対照学習により学習することで、マルチモーダル機械翻訳の性能が向上することが期待されている。

Ji ら (2022) は相互情報量に基づくテキスト - 画像間の対照学習を用いたマルチモーダル機械翻訳を提案した。原言語文と画像の相互情報量の下限を最大化するために InfoNCE 損失の最小化を行い、画像と対応する原言語文（正例）を近づけ、同時に画像と対応しない原言語文（負例）を遠ざける学習を行う。目的言語文と画像の間は、オリジナル画像を用いた翻訳確率と破損した画像を用いた翻訳確率が乖離するように学習を行う。全体の学習は、マルチモーダル機械翻訳に対する負対数尤度の最小化、原言語文に関する負相互情報量の最小化、目的言語文に関する負相互情報量の最小化により学習が行われる。

3.3　事前学習

BERT (Devlin et al. 2019) 等の事前学習モデル（言語モデル）が提案され、事前学習モデルを用いることで様々な自然言語処理タスクの性能が大きく向上することが知られるようになった。大量のデータから学習された事前学習モデルを用いることで、比較的少ない教師データからでも十分な学習（ファインチューニング）が行えるようになったことも大きい。マルチモーダル機械翻訳において主に使われてきた Multi30K は対訳コーパスとしてみたときに、そのデータサイズが非常に小さいという問題があったが、事前学習モデルを用いることによって、この問題が解消されることが期待されている。

Caglayan ら (2021) は BERT に基づき、原言語文 - 目的言語文 - 画像特徴量列に対するマスク言語モデルの事前学習を行った。彼らは、CC (Conceptual Captions) と呼ばれる、インターネットから集めた約 330 万件の画像とその画像に付随する alt-text（英語）を対とするキャプショニングデータを用い、この alt-text 英語文をドイツ語に自動翻訳することで、事前学習モデルのためのデータを作成した。この事前学習モデルに対し、ファインチューニングを行うことでマルチモーダル機械翻訳の学習が行われる。彼らの手法により、従来のマルチモーダル機械翻訳の性能を大きく上回る高い翻訳精度を得ることができたが、一方で、テキストだけを用いた事前学習モデルと画像も用いたマルチモーダル事前学習の間に大きな性能差が無く、画像情報の有効性が疑問視されることになった。

4　まとめ

本稿は、Shen と Shao らによるサーベイ論文「A Survey on Multi-modal Machine Translation: Tasks, Methods and Challenges」(Shen et al. 2024) に基づき、マルチモーダル機械翻訳の技術動向についてまとめた。マルチモーダル機械翻訳のモデルはダブルアテンション機構、エンコーダーでのビジュアルアテンション、ゲーティング機構による融合が主なモデルとして用いられている。マルチモーダル機械翻訳における大きな問題点として、主に使われている Multi30K が非常に小規模であるということが挙げられる。この問題を解決するために画像生成や画像検索を用いたマルチモーダル機械翻訳や、事前学習モデルを用いたマルチモーダル機械翻訳が提案されてきた。

マルチモーダル機械翻訳の研究に対して、「本当に画像情報は役に立っているのか？」ということが問われ、様々

な検証が行われてきている。Wu らによる 2021 年の論文「Good for Misconceived Reasons: An Empirical Revisiting on the Need for Visual Context in Multimodal Machine Translation」(Wu et al. 2021) において、画像情報は正則化の効果の域を出ないことが示された一方、原言語文に情報が少ない場合には画像情報が有用であることも示されており、性や数の無い言語や省略の多い言語においては画像情報が有用であると考えられている。また、文書ベースの機械翻訳においては有用ではないかもしれないが、画像内テキストの翻訳や、同時通訳、ビデオ翻訳など、特殊な機械翻訳タスクにおいては画像情報が有用であると考えられている。

マルチモーダル機械翻訳の研究は継続して行われており、トップカンファレンスにおける論文数は少なくなるどころか増加傾向にある (Shen et al. 2024)。シーン画像マルチモーダル機械翻訳だけではなく、商品説明の機械翻訳、画像内テキストの翻訳、ビデオ翻訳、漫画翻訳、マルチモーダル同時通訳、マルチモーダルチャット翻訳など様々なマルチモーダル機械翻訳の応用先が研究されており、マルチモーダル機械翻訳の技術が今後広がっていくことが期待される。

参考文献

D. Bahdanau, K. Cho and Y. Bengio. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. Proceedings of ICLR 2015.

O. Caglayan, M. Kuyu, M. S. Amac, P. Madhyastha, E. Erdem, A. Erdem and L. Specia. (2021). Cross-lingual Visual Pre-training for Multimodal Machine Translation. Proceedings of EACL 2021, pp. 1317–1324.

O. Caglayan, L. Barrault and F. Bougares. (2016). Multimodal Attention for Neural Machine Translation. arXiv:1609.03976 [cs.CL].

I. Calixto, Q. Liu and N. Campbell. (2017). Doubly-Attentive Decoder for Multi-modal Neural Machine Translation. Proceedings of ACL 2017, Volume 1: Long Papers, pp. 1913–1924.

J.-B. Delbrouck and S. Dupont. (2017a). Modulating and attending the source image during encoding improves Multimodal translation. arXiv:1712.03449 [cs.CL].

J.-B. Delbrouck and S. Dupont. (2017b). Multimodal Compact Bilinear Pooling for Multimodal Neural Machine Translation. arXiv:1703.08084 [cs.CL].

J. Devlin, M.-W. Chang, K. Lee, K. Toutanova. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT 2019, pp. 4171-4186.

D. Elliott, S. Frank, K. Sima' an and L. Specia. (2016). Multi30k: Multilingual English-German Image Descriptions. Proceedings of the 5th Workshop on Vision and Language, pp. 70-74.

D. Elliott and Ákos Kádár. (2017). Imagination Improves Multimodal Translation. Proceedings of IJCNLP 2017, Volume 1: Long Papers, pp. 130-141.

P.-Y. Huang, F. Liu, S.-R. Shiang, J. Oh and C.Dyer. (2016). Attention-based Multimodal Neural Machine Translation. Proceedings of WMT 2016, Volume 2, Shared Task Papers, pp. 639-645.

B. Ji, T. Zhang, Y. Zou, B. Hu and S. Shen. (2022). Increasing Visual Awareness in Multimodal Neural Machine Translation from an Information Theoretic Perspective. Proceedings of EMNLP 2022, pp. 6755-6764.

J. Libovický, J. Helcl. (2017). Attention Strategies for Multi-Source Sequence-to-Sequence Learning. Proceedings of ACL 2017, Volume 2: Short Papers, pp. 196–202.

Q. Long, M. Wang and L. Li. (2021). Generative Imagination Elevates Machine Translation. Proceedings of NAACL, pp. 5738-5748.

T. Luong, H. Pham and C. D. Manning. (2015). Effective Approaches to Attention-based Neural Machine Translation. Proceedings of EMNLP 2015, pp. 1412–1421.

T. Nishihara, A. Tamura, T. Ninomiya, Y. Omote

and H. Nakayama. (2020). Supervised Visual Attention for Multimodal Neural Machine Translation. Proceedings of COLING 2020, pp. 4304–4314.

H. Shen, L. Shao, W. Li, Z. Lan, Z. Liu and J. Su. (2024). A Survey on Multi-modal Machine Translation: Tasks, Methods and Challenges. arXiv:2405.12669 [cs.CL].

J. Su, J. Chen, H. Jiang, C. Zhou, H. Lin, Y. Ge, Q. Wu and Y. Lai. (2021). Multi-modal neural machine translation with deep semantic interactions. Information Sciences, Volume 554, pp. 47–60.

I. Sutskever, O. Vinyals and Q. V. Le. (2014). Sequence to Sequence Learning with Neural Networks. Proceedings of NeurIPS.

Z. Tang, X. Zhang, Z. Long and X. Fu. (2022). Multimodal Neural Machine Translation with Search Engine Based Image Retrieval. Proceedings of the 9th Workshop on Asian Translation, pp. 89-98.

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

Z. Wu, L. Kong, W. Bi, X. Li and B. Kao. (2021). Good for Misconceived Reasons: An Empirical Revisiting on the Need for Visual Context in Multimodal Machine Translation. Proceedings of ACL/IJCNLP 2021, Volume 1: Long Papers, pp. 6153–6166.

R. Yuasa, A. Tamura, T. Kajiwara, T. Ninomiya and T. Kato. (2023). Multimodal Neural Machine Translation Using Synthetic Images Transformed by Latent Diffusion Model. Proceedings of ACL Student Research Workshop 2023, pp. 76–82.

Z. Zhang, K. Chen, R. Wang, M. Utiyama, E. Sumita, Z. Li and H. Zhao. (2020). Neural Machine Translation with Universal Visual Representation. Proceedings of ICLR 2020.

Y. Zhao, M. Komachi, T. Kajiwara and C. Chu. (2022a). Region-attentive multimodal neural machine translation. Neurocomputing, 476, pp. 1–13.

Y. Zhao, M. Komachi, T. Kajiwara and C. Chu. (2022b). Word-Region Alignment-Guided Multimodal Neural Machine Translation. IEEE ACM Transactions on Audio, Speech, and Language Processing, Volume 30, pp. 244–259.

Y. Zuo, B. Li, C. Lv, T. Zheng, T. Xiao and J. Zhu. (2023). Incorporating Probing Signals into Multimodal Machine Translation via Visual Question-Answering Pairs. Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 14689-14701.