令和５年度茨城大学大学院理工学研究科情報工学
専攻修士学位論文

CLIPを利用したマルチモーダル機械翻訳

所属　　情報工学専攻
著者　　YINXINYAN（22NM706X）
指導教員　新納浩幸教授

令和6年2月30日（火）

令和５年度茨城大学大学院理工学研究科情報工学専攻修士学位論文

CLIPを利用したマルチモーダル機械翻訳

著者

YINXINYAN (22NM706X)

指導教員

新納浩幸教授

論文要旨

近年、マルチモーダル機械翻訳モデルは、画像ラベル情報に関係なく、テキスト情報と画像情報の両方を使用している。さらに、モデルの訓練では、テキストエンコーダとビジュアルエンコーダは通常一緒に学習されるため、画像情報が実際に役割を果たしているのかどうかを判断することができない。

CLIPは、画像と言語の2つの領域の埋め込み表現を得ることができる事前学習モデルの一つです。CLIPでは、画像エンコーダーとテキストエンコーダーを一緒に訓練し、バッチ内のN個の実ペアの画像とテキストのembeddingのコサイン類似度を最大化する一方で、不正確なペアのembeddingのコサイン類似度を最小化することで、マルチモーダルなembedding空間を学習している。

本研究では、モデルの構築では、まずテキストモデルを訓練し、その後テキストエンコーダのパラメータを固定します。次にビジュアルエンコーダを導入し訓練を行います。この際、テキストエンコーダのパラメータは更新しない。 CLIP モデルは最初に "ViT-B-32" 事前トレーニング パラメーターで読み込まれ、画像エンコーダーは画像から特徴を抽出し、将来のネットワーク入力のために保存される。 Visual Genome データセットのオブジェクト クラス ラベルは、CLIP テキスト エンコーダーの入力テキストとして使用され、画像の特徴と 1600 のラベルがスコア付けされ、並べ替えられ、スコアが最も高い上位 5 つのラベルの特徴が保持される。画像の特徴抽出に CLIP モデルを利用し、画像内の最も関連性の高いラベルの特徴を特定して、機械翻訳の結果を改善する。

Master's Thesis in Scholastic 2024, Major in Computer and Information Sciences, Graduate School of Science and Engineering, Ibaraki University

Multimodal machine translation using CLIP

Author : YINXINYAN (22NM706X)

Adviser : Prof. Hiroyuki Shinnou

Abstract

In recent years, multimodal machine translation models have been utilizing both text and image information, irrespective of image label information. Moreover, during model training, text encoders and visual encoders are typically co-trained, making it challenging to determine the actual role of image information.

CLIP is one such pre-trained model that can obtain embedding representations for both image and language domains. Pre-trained models like CLIP are trained on large corpora and can perform various tasks in V&A fields with high accuracy in a zero-shot manner. CLIP utilizes an efficient Vision Transformer for image encoding, surpassing the efficiency of traditional CNNs. In CLIP, image encoders and text encoders are trained together. The training involves maximizing the cosine similarity of embeddings for N actual pairs of images and text within a batch while minimizing the cosine similarity of embeddings for inaccurate pairs, thus learning a multimodal embedding space.

In this study, the CLIP model is employed to extract label information from images, preserving relevant label information, and inputting it along with text into the text encoder. In the model construction, the text model is initially trained, and then the parameters of the text encoder are fixed. Subsequently, the visual encoder is introduced and trained, with the parameters of the text encoder remaining unchanged.

目次

第 1 章

序論

マルチモーダル機械翻訳は、複数の感覚やモードを統合的に活用する翻訳の手法です。従来の機械翻訳は主にテキストからテキストへの翻訳に焦点を当てていましたが、多模态翻訳はより広範な領域に進展し、テキスト、画像、音声などさまざまな形式の入力および出力の対象です。

この分野の典型的なタスクには、画像からテキストへの翻訳がある。これは画像を自然言語テキストに翻訳し、画像の内容を言葉で表現するものです。逆に、テキストから画像への翻訳では、テキストの説明を画像に変換し、テキストに対応する視覚的な表現を生成する。音声からテキストへの翻訳は、口頭の音声をテキストに変換し、テキストから音声への翻訳はテキストの音声合成を実現する。また、多モーダル融合は、画像やテキストなどの異なるモードの情報を統合し、複雑な情報の理解を向上させるために重要なタスクであり、翻訳の品質を向上する。

マルチモーダル機械翻訳における主要な研究分野は，例えば,

- 画像からテキストへの翻訳： 画像を対応する自然言語テキストに翻訳し、画像の内容を説明する。
- テキストから画像への翻訳： テキストの説明を画像に翻訳し、テキストの説明と一致する画像を生成する。
- 音声からテキストへの翻訳： 音声入力をテキストに翻訳し、音声中の情報をテキスト形式で表示する。
- テキストから音声への翻訳： テキストを音声に翻訳し、テキストの朗読や音声のヒントを実現する。

この分野の重要なタスクには、画像からのテキスト生成、テキストからの画像生成、音声からのテキスト変換などがある。これにより、単なる言

語だけでなく、視覚的な情報や音声情報も考慮に入れられ、より広範で自然なコミュニケーションが実現される可能性がある。

マルチモーダル機械翻訳の挑戦の一つは、異なるモダリティの情報をどのように統合するかです。例えば、画像とテキストの翻訳では、それぞれの情報が補完し合って意味の一貫性を持たせる必要がある。また、モデルの訓練において、異なる種類のデータを組み合わせて有益な表現を学習させることが求められる。

最近では、深層学習や強化学習の進展により、マルチモーダル機械翻訳の性能向上が著しくなりつつある[1]。例えば、CLIP（Contrastive Language-Image Pretraining）やDALL-Eなどのモデルが、異なるモダリティの情報を効果的に扱い、高度な翻訳を実現している[2][3][4][5]。

このような技術の進展は、多様なコンテンツやコンテキストに対応した翻訳の提供や、異なる言語や文化間での円滑なコミュニケーションを促進する可能性を秘めている。今後ますます発展が期待されるマルチモーダル機械翻訳は、言語処理技術の新たな展開を切り拓くです。

第 2 章

先行研究

2.1マルチモーダル機械翻訳の発展の歴史

2015年、深層畳み込みネットワーク層で構成されたモデルが画像解釈タスクを席巻している。時間的にも構成的である「深層」モデルが、視覚的なシーケンスまたはラベルのシーケンスを含むタスクに対して効果的かどうかを調査する。長期のRNNモデルは、変動長の入力を変動長の出力に直接マッピングし、複雑な時間動態をモデル化できるため魅力的です。モデルが、別々に定義およびまたは最適化された認識または生成のための最先端モデルよりも明確な利点を持っていることを示している。

2015年、Xu, Ran, et al.の論文では、ビデオとそれに対応するテキスト文を共同でモデリングする統一されたフレームワークを提案している。実験では、結果はSVM、CRF、CCAベースラインを上回り、Subject-Verb-Objectトリプレットと自然な文生成を予測する際に優れており、ビデオ検索および言語検索のタスクではCCAよりも優れている。

2016年、Yang, Xiaodong, Pavlo Molchanov, and Jan Kautz.の論文では、ビデオ分類のための深層ニューラルネットワークの複数のレイヤーとモダリティを組み合わせる新しいフレームワークを提案している。複数のレイヤーとモダリティを統一的な方法で融合するための強力なブースティングモデルを導入する。広範な実験では、公開されている2つのベンチマークデータセット、UCF101とHMDB51で最先端の結果を達成している。

2016年、Hendricks, Lisa Anne, et al. の研究では、画像と文章のデータセットに存在しない新しいオブジェクトの説明を生成するタスクに対処するために、Deep Compositional Captioner (DCC) を提案します。モデルの新しい概念を説明する能力を示すために、MSCOCOでその性能を経験

的に評価し、ImageNetの画像についてはペアの画像キャプションデータが存在しないオブジェクトについての定性的な結果である。

2016年、Singh, Aditya, et al.の論文では、伝統的なラベル付きデータセットを使用して、野生のVineでの非監督学習アクション分類の問題に焦点を当てている。このために、データ拡張ベースのシンプルなドメイン適応戦略を使用している。我々は、セマンティックワード2ベクトル空間を共通の部分空間として利用して、ラベル付きソースドメインとラベル無しターゲットドメインの両方からのビデオ特徴を埋め込む方法を採用する。

2017年、Deena, Salil, et al.の論文では、同梱の音声から派生した補助機能がNMTのために調査され、テキストから派生した機能と比較および組み合わせられる。これらの音響埋め込みは、翻訳の曖昧さを解消するのに役立ち、したがって出力を向上させる。結果は励みになり、音響情報はNMTに役立つことを示し、BLEUスコアで全体的な相対改善率が3.3%です。

2017年、Yang, Xitong, et al. の論文では、Correlational Recurrent Neural Network (CorrRNN) という新しい時間融合モデルを提案し、本質的に時間的なモダリティを結合するためのものです。提案されたCorrRNNモデルの異なるコンポーネントの貢献を実験によって検証し、複数のデータセットでの堅牢性、効果、最先端のパフォーマンスを実証する。

2017年、Liu, Lin, and Jianhou Ganの研究では、多モダルトピックモデリングの発展を概説し、国の文化資源システムにおける多モダルトピックモデリングの可能な応用について議論する。その応用には、クロスメディア検索、自動注釈、および推薦システムなどが含まれる。しかし、多言語と訓練データの不足といった要因から、既存の多モダルトピックモデルの改善を研究し探求する新たな需要が生まれている。

2018年、Liu, Zhun, et al.の論文では、効率を向上させるために低ランクテンソルを使用してマルチモーダルフュージョンを実行するLow-rank Multimodal Fusion方法を提案する。3つの異なるタスクでモデルを評価する：マルチモーダル感情分析、話者トレイト分析、感情認識。我々のモデルは、計算の複雑さを大幅に削減しながら、これらのすべてのタスクで競争力のある結果を達成する。

2018年、Yu, Xiaoming, et al.の研究では、画像から画像への変換の進歩により、単一のネットワークを介して多様な画像を生成するアプローチが増加している。1対多のマッピングの対象ドメインを示すために、潜在コードがジェネレーターネットワークに注入される。既存の正規化戦略は、特徴分布の不一致を引き起こすか、潜在コードの効果を排除する可能性がある。これらの問題を解決するために、多重マッピングモデルを設計するための一貫性内で多様性の基準を提案する。

2019年、Li, Xirong, et al. の論文は、データと基準方法の観点からのクロスリンガルな画像注釈と検索への貢献です。私たちは、MS-COCOを手作りの中国語の文章とタグで豊かにする新しいデータセットであるCOCO-CNを提案する。効果的な注釈取得のために、推奨支援型の共同注釈システムを開発し、画像の内容に関連すると思われるいくつかのタグと文章を注釈者に自動的に提供する。

2019年、Shen, Yiqing, and Yingbo Li. の論文では、音声、テキスト、ビジュアル指標などのマルチモーダルビデオ情報から抽出されたキーワード情報表現とビデオの時間情報を結合してビデオのスキミングの新しいアプローチを提案する。さらに、ブランドセーフなフィルタリングとセンチメント分析を導入し、ビデオスキムにユーザーフレンドリーなコンテンツのみを残す。

2019年、Pramanik, Subhojeet, Priyanka Agrawal, and Aman Hussain. の論文では、画像、テキスト、ビデオなどのさまざまなモダリティを含むタスクに使用できる拡張された統一アーキテクチャを紹介する。テンポラル入力シーケンスに対応する隠れ状態に加えて、入力の空間次元を学習できる時空間キャッシュメカニズムを提案される。

2020年、Wang, Dongyang, Junli Su, and Hongbin Yu.の研究では、マルチモーダルニューラルネットワークを提案する。各モードにはそれに対応する独立した構造を持つ複数層のサブニューラルネットワークがあり、異なるモードの特徴を同じモードの特徴に変換するために使用される。単語分割処理に関して、既存の単語分割方法ではテキストの意味の長期依存性と長いトレーニング予測時間をほとんど保証できない問題に鑑み、ハイ

ブリッドネットワーク英単語分割処理方法が提案される。

2021年、Kamath, Aishwarya, et al. の論文では、生のテキストクエリ（キャプションや質問など）に基づいて画像内のオブジェクトを検出するエンドツーエンドのモジュレーションされた検出器であるMDETRを提案する。事前トレーニングアプローチは、ラベル付けされたインスタンスが非常に少ないオブジェクトカテゴリの長い尾を処理する方法を提供する。

2021年、Bianchi, Federico, et al. の論文では、イタリア語用の初のCLIPモデル（CLIP-Italian）を紹介し、140万以上の画像テキストペアで訓練されました。結果は、CLIP-Italianが画像検索とゼロショット分類のタスクで多言語CLIPモデルを上回ることを示している。

2021年、Baldrati, Alberto, et al.の研究では、多モーダルなゼロショット表現学習の進歩に基づいて、最近のCLIPモデルから得られた特徴を使用して、条件付き画像検索を行う方法を探求する。CLIP特徴とシンプルなベースラインから始め、多モーダル特徴に基づいた慎重に設計されたCombinerネットワークが、非常に効果的であることを示す。

2022年、Bahng, Hyojin, et al.の論文には大規模なビジョンモデルを適応させるために視覚的なプロンプティングの効果を調査した。包括的な実験により、CLIPモデルに対して視覚的なプロンプティングが非常に効果的であり、分布の変化にも強力で、標準的な線形プローブと競合する性能を実現している。視覚的なプロンプティングの驚くべき効果は、ビジョンモデルの適応に新しい視点を提供している[6]。

2022年、Couairon, Guillaume, et al.の文章には画像表現では構造化された意味的関係は存在しないことが確認されました。結果によれば、バニラのCLIP埋め込みは画像をデルタベクトルで変換するのには適していないことが示され、COCOデータセットでの簡単なファインチューニングが著しい改善をもたらすことも示される[7]。

2022年、Wu, Ho-Hsiang, et alの研究にはWav2CLIPという、Contrastive Language-Image Pre-training (CLIP) から抽出した堅牢な音声表現学習手法を提案する。Wav2CLIPは、強力な音声表現学習手法で、音声分類、検索、生成などの多様なタスクで優れたパフォーマンスを示し

ました。Wav2CLIPは音声を画像とテキストと共有の埋め込み空間に投影し、ゼロショット分類やクロスモーダル検索などの多モーダルアプリケーションに適している[8]。

2023年、Myeongseob Ko, Ming Jin et al.の文章は、メンバーシップ推測攻撃（Membership Inference Attacks、MIAs）と、それらの大規模なマルチモーダル機械学習モデルへの適用について議論しており、特にCLIPなどのモデルに焦点を当てている。

2023年、Vaiani, Lorenzo, Luca Cagliero, and Paolo Garzaの文章はCLIPとテキストおよび画像の拡張技術を使用して、さまざまな領域と言語を横断したテキスト内の言語の曖昧さを解消する研究プロジェクトを説明している。主な目標は、単語の意味の曖昧さを解消する精度とシステムの性能を向上させることです。

2023年、Gao, Peng, et al. の論文では、プロンプト調整はテキスト入力に対して行われますが、著者はCLIP-Adapterを提案し、視覚または言語の分岐に対して特徴アダプターを使用して微調整を行います。具体的には、CLIP-Adapterは新しい特徴を学習するための追加のボトルネック層を採用し、元のプリトレーニング特徴との残差スタイルの特徴ブレンディングを実行する。

2.2中日データセットの研究

マルチモーダル翻訳は、テキスト、画像、音声など、複数の感覚モードにまたがる翻訳タスクを指す。このようなタスクにおいて、研究やモデルの性能評価において重要なのは、テキスト、画像、音声などのモードにわたるデータセットです。ただし、現在のところ、特定の中日多モーダル翻訳向けの専用データセットは比較的少なく、多くの多モーダルデータセットは主に英語を対象としている。

以下は、中日多モーダル翻訳の研究に使用できるいくつかの一般的な多モーダルデータセットです。これらのデータセットは主に英語向けですが、中日多モーダル翻訳の研究にも活用できる。

・MSCOCO（Microsoft Common Objects in Context）： これは画像の説

明生成に広く使　用されるデータセットで、多くの画像とそれに対応する説明文が含まれている。

・Multi30K： これは画像からテキストへの翻訳に使用されるデータセットで、画像、英語の説明、ドイツ語の説明の三モードデータが提供されている。

Flickr30K： Multi30Kと同様、これも画像からテキストへの翻訳に使用されるデータセットで、画像とそれに対応する英語の説明が含まれている。

・VoxCeleb： これは音声からテキストへの翻訳に使用されるデータセットで、多くの有名人の音声とそれに対応するテキストが含まれている。

・IAPR-TC12： これはテキストから画像への翻訳に使用されるデータセットで、複数の言語のテキストと関連する画像が含まれている。

中日多モーダル翻訳の研究を行うためには、特定のタスクや言語の要件に基づいて、データセットを作成または整理する必要がある。中文と日本語の既存のデータ、および関連する多モーダル情報を組み合わせて、研究目的に適したデータセットを構築する必要がある。

第 3 章

マルチモーダル機械翻訳

3.1機械翻訳

長い間、機械翻訳は文級翻訳を指してきた。主な理由は、文レベルの翻訳モデリングが問題を大幅に簡略化し、機械翻訳方法が実践され検証されやすいようにすることにある。しかし、人間が言語を使用する過程は、孤立して一つ一つの文の上で行われているわけではない。この問題は人間が言語を学習する過程に類することができる：子供が成長する過程で視覚、聴覚、触覚などの多種の信号を受け入れることができて、これらの信号の共通作用は彼らに客観的な世界に対する“認識”を発生させて、同時に彼らに“言語”を使って表現することを促す。この角度から言えば、言語能力は単一の要素によって形成されたものではなく、それは往々にして他の情報の相互作用を伴っており、例えば、人々が一言を翻訳する際に、見た画面、聞いたイントネーション、さらには前に言った文の中の情報を使うことができる。

3.2マルチモーダル機械翻訳

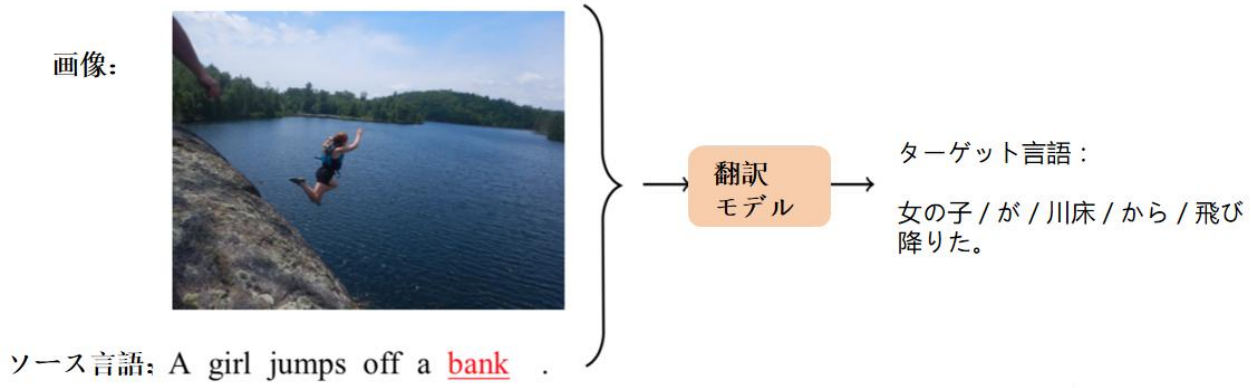

図 3.2.1

広義には、現在の文以外の情報はコンテキストと見なすことができる。例えば、次の図では、英語の文「A girl jumps off a bank.」を中国語に翻訳する必要がある。しかし、その中の「bank」には複数の意味があるため、英語の文自体の情報だけを使用すると、正しい訳文「河床」ではなく「銀行」に翻訳される可能性がある。しかし、この英語文に対応する画像は図3.2.1にも提供されており、画像に川床が直接表示されていることは明らかであり、この場合「bank」には曖昧さはない。通常、このような画像と文字を用いた機械翻訳のタスクをマルチモーダル機械翻訳（Multi-Modal Machine Translation）と呼ぶこともある。

モダリティ（Modality）とは、ある情報源のことです。例えば、視覚、聴覚、嗅覚、味覚はすべて異なるモダリティと見なすことができる。したがって、ビデオ、音声、文字などは、これらのモダリティを担持する媒体と見なすことができる。機械翻訳にマルチモーダルという概念を用いるのは、文字とは異なる情報を区別するためである。画像などの視覚モード情報のほか、機械翻訳にも聴覚モード情報を利用することができる。例えば、音声を直接翻訳したり、翻訳結果を音声で表現する。

異なる情報源が導入するコンテキストに加えて、機械翻訳も文字自体のコンテキストを利用することができる。例えば、1つの文章の中の1つの文を翻訳する場合、全体の文章の内容に基づいて翻訳することができる。明らかにこの章の文脈は機械翻訳に役立つ。本章の次の内容では、機械翻訳で異なるコンテキスト（マルチモーダルとページ情報）を使用する方法について議論する。

人間が受け取る情報の中で、視覚情報の比重は音声やテキスト情報に劣らず、さらに多いことが多い。視覚情報は通常画像として存在し、近年、画像を結合したマルチモーダル機械翻訳[9][10]が広く注目されている。多モード機械翻訳（下図（a））は、簡単に言えば、ソース言語と他のモード（例えば画像など）の情報を結合してターゲット言語を生成するプロセスである。このような画像を結合した機械翻訳は狭義の「翻訳」であり、本質的にはソース言語からターゲット言語、あるいはテキストからテキストへの翻訳である。実際に画像からテキスト（下図（b））への変換、す

なわち所与の画像が、画像コンテンツに関連する記述を生成し、広義の「翻訳」と呼ばれることもある。例えば、画像記述生成（Image Captioning）は典型的な画像からテキストへの翻訳である。もちろん、この広義

の翻訳形式には、画像からテキストへの変換だけでなく、画像から画像への変換（図©），テキストから画像への変換（下図（d））などもある。

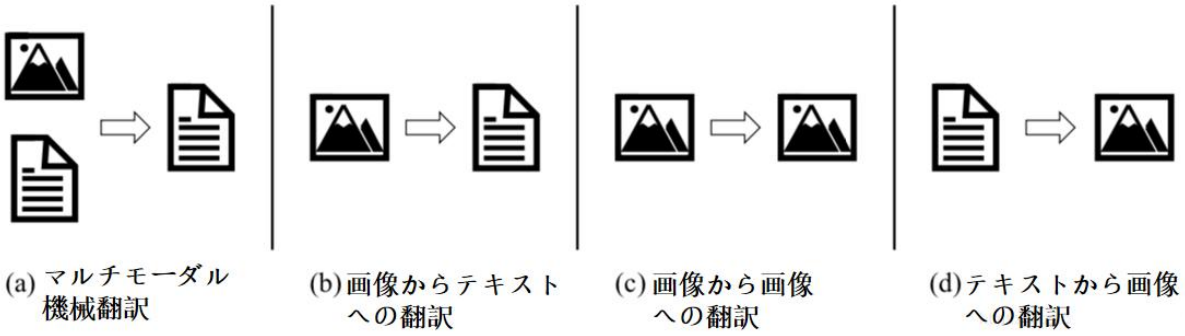

図 3.2.2

特徴融合に基づく方法は、通常、画像情報を入力文の一部としたり、エンコーダやデコーダの状態を初期化したりする.下図に示すように、図中のy<は現在時刻より前の単語列を示し、画像特徴の抽は通常、畳み込みニューラルネットワークに基づいている。畳み込みニューラルネットワークを介してグローバルな画像特徴を得て、次元変換を行った後、それをソース言語入力の一部または初期化状態としてモデルに導入する。

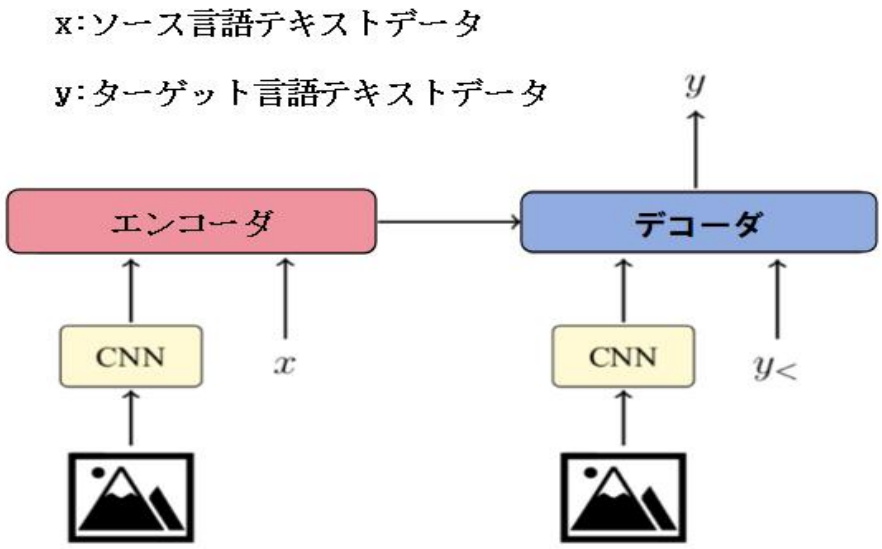

図 3.2.3

しかし、このような画像情報の導入方法には、以下の2つの欠点がある：

(1) 画像情報はすべて有用ではなく、グローバルな特徴としてノイズを導入するソース言語やターゲット言語とは関係のない情報が存在することが多い。

画像情報はソース言語の一部または初期化状態として間接的に翻訳文の生成に関与し、ニューラルネットワークの計算過程で画像情報に一定の損失が生じる。マルチモーダル機械翻訳は、従来の機械翻訳が主にテキストからテキストへの翻訳に焦点を当てていたのに対し、異なるモード（モード）の情報を統合してより包括的で豊かな翻訳を行う手法です。主にテキスト、画像、音声などの異なる情報源を扱う。

連合モデルに基づく方法は、通常、翻訳タスクを他の視覚タスクと結合し、連合訓練を行うことである。この方法はマルチタスク学習と見なすこともできますが、ここでは翻訳と視覚タスクだけに注目しています。1つの一般的な方法は、モデルの部分パラメータを共有して、異なるタスク間の類似部分を学習し、特定のモジュールを通じて各タスク固有の部分を学習することです。

下図に示すように、図中のy<は現在時刻より前の単語列を示し、マルチモーダル機械翻訳タスクを機械翻訳と画像生成の2つのサブタスクに分解することができる。その中で、機械翻訳をメインタスクとし、画像生成をサブタスクとする。ここでの画像生成とは、1つの画像記述から対応する画像を生成することであり、画像生成タスク原書について詳細に説明している。1つのエンコーダでソース言語データをモデリングし、2つのデコーダ（翻訳デコーダと画像デコーダ）で翻訳タスクと画像生成タスクをそれぞれ学習します。最上位レベルでは各タスクの独立した特徴が学習され、下位レベルの共有パラメータではより豊富なテキスト表現が学習される。

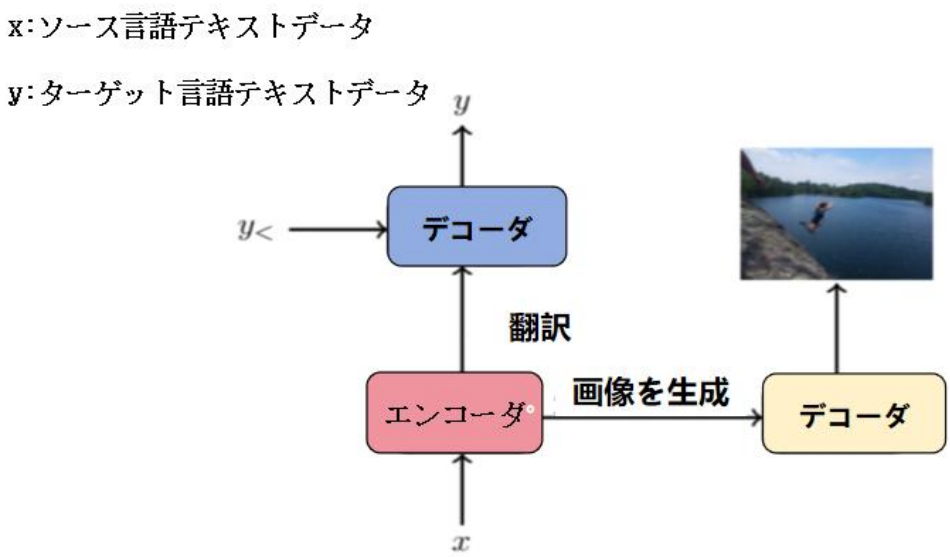

図 3.2.4

マルチモーダル機械翻訳は、従来の機械翻訳が主にテキストからテキストへの翻訳に焦点を当てていたのに対し、異なるモード（モード）の情報を統合してより包括的で豊かな翻訳を行う手法です。主にテキスト、画像、音声などの異なる情報源を扱う。

以下は、マルチモーダル機械翻訳の要点:

• 複数の入力モード： マルチモーダル機械翻訳は、通常、テキストだけでなく、画像や音声など、異なるモードからの情報を同時に処理します。これにより、より豊かで多様な情報を考慮できる。

• 多様な出力モード： 翻訳結果は、通常、テキストだけでなく、画像や音声にもなり得る。例えば、画像に対して説明文を生成するなどがある。

• モード間の情報統合： マルチモーダル機械翻訳は、各モードの情報を組み合わせ、相互に補完し合ってより正確で意味豊かな翻訳を実現しする。たとえば、画像の内容に基づいてテキストを生成するなどが考えられる。

• 深層学習とニューラルネットワーク： この手法は、通常、深層学習とニューラルネットワークを活用している。これにより、モードごとのパターンや相関関係を学習し、高度な翻訳を行うことが可能になる。

具体的な応用例として、英語の文章とその文章に関連する画像が与えら

れた場合、マルチモーダル機械翻訳はそれらの情報を同時に処理して、日本語の文章と対応する画像を生成することが期待される。これにより、言語と視覚情報の複合的なコンテキストを考慮した翻訳が可能になる。

第 4 章

CLIP: Contrastive Language-Image Pre-training

4.1導入と概要

多層畳み込みニューラルネットワーク（CNN）において、前半の畳み込み層は一般的で共有可能な視覚特徴（テクスチャ、エッジなど）を抽出する傾向がある。進むにつれて、タスクに特有の特徴（例: カテゴリ情報）を抽出するようになる。そのため、視覚タスクを分離することがあり、一般的にはバックボーンとヘッド（バックボーンは特徴抽出、ヘッドは目標検出に使用）と呼ばれる。

自然言語処理（NLP）の領域でも同様の考え方があり、まずモデルを大規模なコーパスで共通の意味表現能力を学習させ、その後に視覚タスクのようにタスクに関連した特徴を学習するというアプローチが取られる。これは転移学習と呼ばれ、微調整（Fine-tuning）とも呼ばれる[11][12]。

通常、視覚タスクではデータをフィットさせ、データとラベルができるだけよく対応するようにします。しかし、近年では別の手法もあり、データと対応するラベルの距離を縮めて対比学習を行う。

例えば、

·DPR（dense passage retrieval）では、インバッチネガティブサンプリングを使用し、単一の質問に対して標準答えを正例（距離をできるだけ近く）とし、他の質問の答えを負例（距離使用できません）としてデュアルエンコーダを最適化し、良好な結果を得ている。

·BERTのプレトレーニング課題は、人間の完全な埋め込み穴のようなものです。これにより、単語とそのコンテキスト単語との距離が十分に近くなるように模倣する。

·GPTは自己回帰的な方法を使用しており、モデルは現在の位置のトーク

ンを推論するために現在の位置以前のフィールドしか使用できない。

·UniLMモデルは、2つの異なる注意メカニズムを組み合わせてトレーニングされます。

NLP領域のいくつかのクラシックな手法は、研究にいくつかの洞察をもたらす可能性があります。異なるモードのデータを同じ形式に変換し、距離測定を使用して異なるモードのデータの対応関係を確立することができれば、複雑なタスクを達成するために複数のモードのデータを活用できるようになる。

Contrastive Language-Image Pretraining（CLIP）は、言語情報と画像情報を共同でトレーニングし、ダウンストリームのタスクでのゼロショット（学習時にそのクラスのトレーニングサンプルがない状態で、モデルが未知のクラスを識別できるようにする）の能力を実現している。具体的には、CLIPは画像の分類タスクを画像テキストマッチングのタスクに変換し、画像情報とセマンティック情報を同じマルチモードセマンティックスペースにマッピングし、対比学習の手法を使用してトレーニングする。

4.2モデル構造

CLIPのモデル構造は、VirTexと同様のCNNとテキストTransformerではなく、より効率的なVision Transformerを利用しています。正確な言葉を予測することは非常に難しく、一方で対象表現学習でも同程度の精度が得られることが最近の研究で明らかになっています。また、生成モデルはより多くのデータが必要とされることも理解されています。これらの理由から、テキスト中の正確な単語ではなく、単にテキスト全体がどの画像とペアになっているかを予測する、より簡単な代理タスクを解決するためのモデルが開発されました。

最終的なモデルでは、N個のペアのバッチが与えられると、CLIPはそのバッチ全体でN × N個の可能性のあるペアリングの中から、実際に発生したペアリングを予測するように訓練される。具体的には、CLIPでは画像エンコーダーとテキストエンコーダーを一緒に訓練し、バッチ内の

N個の実ペアの画像とテキストの埋め込みのコサイン類似度を最大化する一方で、不正確なペアの埋め込みのコサイン類似度を最小化することで、マルチモーダルな埋め込み空間を学習している（図4.2.1 参照）。

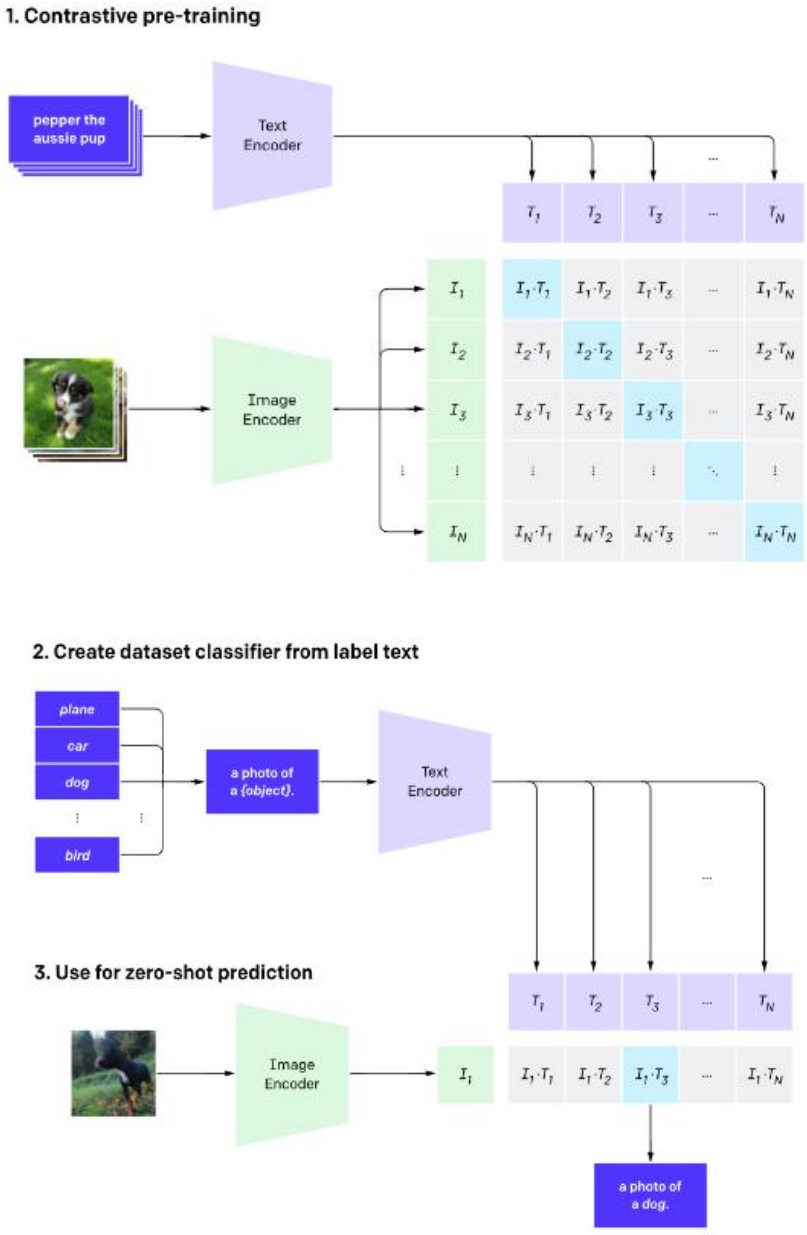

図 4.2.1：CLIP の事前学習時と推論時の全体像.

CLIP の事前学習の各ステップは，クラスごとに 1 つの例を含み，自然言語記述によって定義された総クラス数 32,768 のコンピュータビジョンデータセットに対して，ランダムに作成されたプロキシの性能を最適化していると見ることができる． zero-shot 評価のために，テキストエンコーダによって計算された zero-shot 分類器を一

度キャッシュし，その後のすべての予測のために再利用します．これにより，zero-shot 分類器の生成コストをデータセット内の全ての予測値に対して償却することが可能となる[13].

画像エンコーダ

モデル1　（比較ベースモデル）ResNet-50[15]の改良版を利用して。グローバル平均プールリング層を注意プールリング機構に置換する。注意プールリングは、画像に基づくグローバル平均プールリングの表現をクエリして条件を付加する「Transformerスタイル」の多重QKV注意の単一レイヤとして実装される。

モデル2　（より洗練されたモデル）は、基本的にVisionTransformerと同じモデルを使用します。違いは、Transformerの前にパッチと位置埋め込みの組み合わせに追加のレイヤ正規化を追加し、少し異なる初期化スキームを使用している点です。

テキストエンコーダ

基本的に Transformer を利用．ベースサイズとして，8 つのアテンションヘッドを持つ 63M パラメータの 12 層 512 ワイドモデルを使用．49,152 個の vocab サイズを持つテキストの小文字のバイトペアエンコーディング（BPE）表現で動作する．計算効率のため，最大シーケンス長は 76 に制限．テキストシーケンスは［SOS］と[EOS］トークンで括られ，[EOS］トークンにおけるトランスフォーマーの最上位層の活性化は，層を正規化した後，マルチモーダル埋め込み空間に線形投影されたテキストの特徴表現として扱われる．事前に学習した言語モデルで初期化したり，補助的な目的として言語モデルを追加したりする能力を維持するために，Masked self-attention が使用されている．

クロスロスエントロピー関数

まずlogicはコサイン類似度を、labelsはリアルラベルを定義する。ここで、logicはNXCサイズの配列であり、Nはいくつのサンプル（BatchSize）があるかを表し、Cはカテゴリの総数を表す。例えば、

$$logic = \begin{bmatrix} 0.5 & 0.4 & 0.2 & 0.7 \\ 0.2 & 0.5 & 0.3 & 0.8 \\ 0.1 & 0.5 & 0.4 & 0.1 \end{bmatrix},$$

3つのサンプル、4つのカテゴリがあることを示す。

一方、labelsは1次元配列で、サイズはNです。N個のサンプルに対応する実際のカテゴリを表す。label=[3,0,1]のように、1つに3つのサンプルがあることを示し、1つ目のサンプルに対応するリアルラベルは3、2つ目のリアルラベルは0、3つ目のリアルラベルは1である。

$$loss = -\sum_i^N labels[i] * \ln logic[i]$$

で、pytorch関数の内部では、labelはNXCサイズの配列に変換され、実際のカテゴリでは1、その他は0であると考えられます。例えば、label[3,0,1]を$\begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$、式を次のように簡略化できる。

$$loss = \frac{-\sum_i^N \ln logic[i][j]}{N}$$

各サンプルの実際のラベルの確率を平均化します。一方、 ln logicは、まずlogicをsoftmaxし、各サンプルが異なるカテゴリに対応する確率を得て、その後logを取るように近似することができる。

CosineAnnealingLR

ネットワークトレーニングのlossを最小にすると、重みの勾配を常に振動させ、トレーニングの損失値をグローバルな最下位にするのは難しいため、高い学習率を使用するのは適切ではない。だから学習率はやはり下がる必要があり、余弦関数を通じて学習率を下げることができる。余弦関数ではxが増加するにつれて余弦値が最初にゆっくりと低下し、それから加速低下し、再びゆっくりと低下する[14]。この下降パターンは学習率と組み合わせることができ、非常に効果的な計算方法で良い効果をもたらすことができる（図4.2.2参照）。

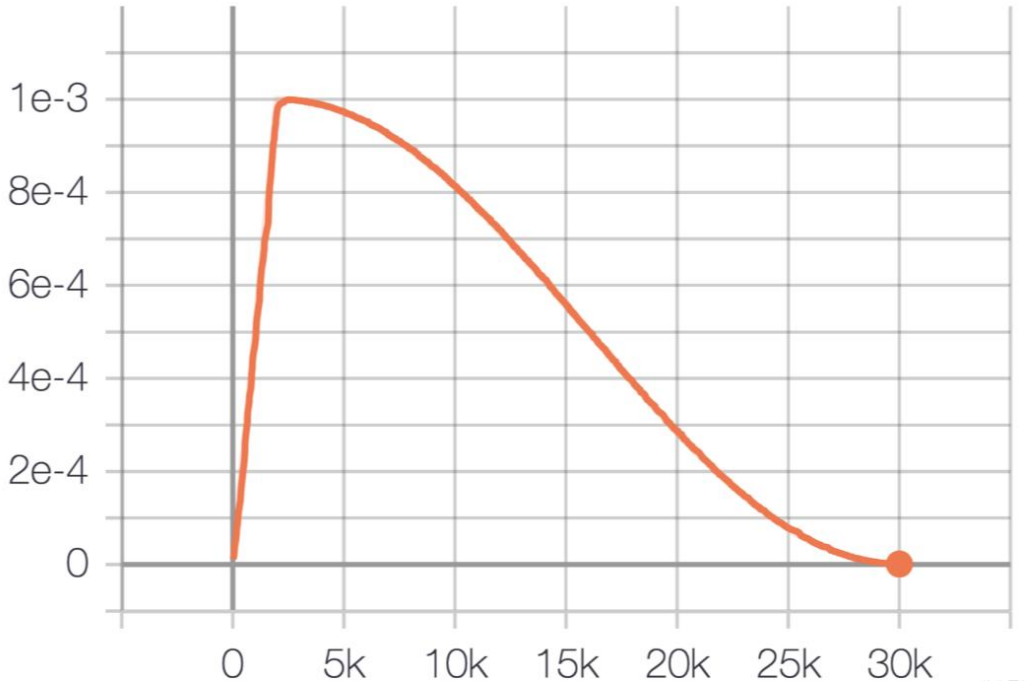

図 4.2.2:CosineAnnealingLR

事前トレーニング方法の選択：青い線のようなimage captionのような事前訓練タスクを使用すると、学習効率が低く、同じ精度を達成するには多くの画像データを使用する必要があります。図4.2.2に示すように黄色の線は基本的な訓練方法であり、画像の記述テキストを予測する単語であ

る。この2つの方法は、各画像に添付されている文字の正確な単語を予測しようとしていますが、種類が多いため難しいです。対照学習はよりよく特徴を学習することができ、テキスト全体と画像のペアを1つの目標とし、さらに効率を高めた。

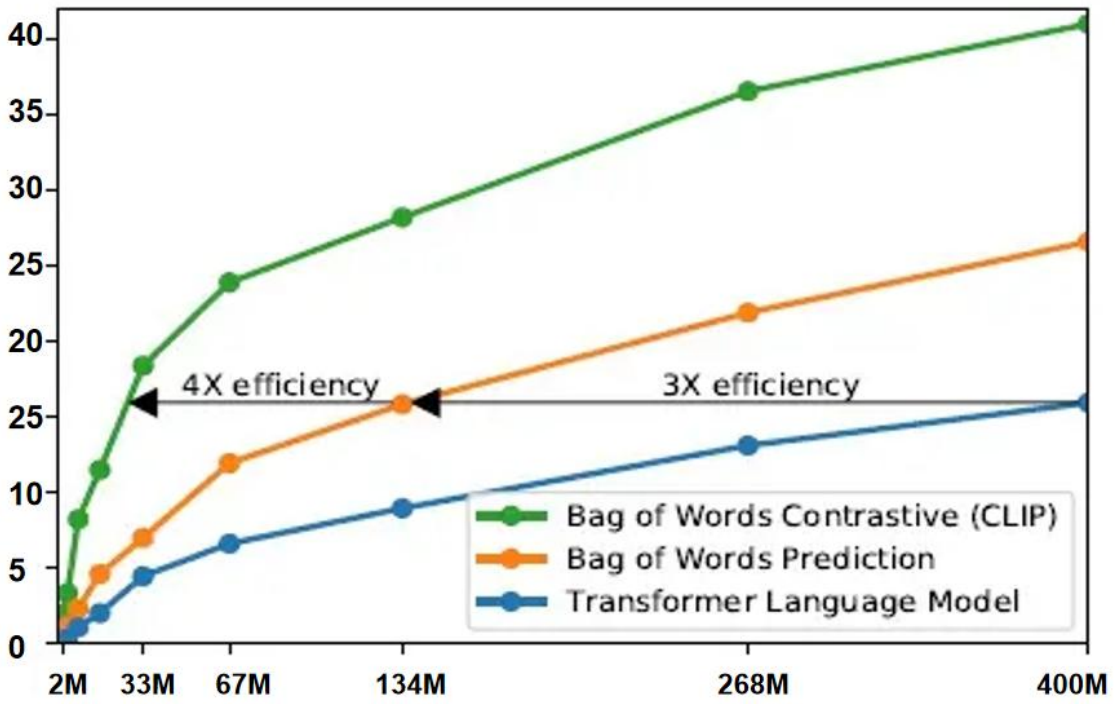

図 4.2.3

第 5 章

提案手法

本文の調査研究では、画像の特徴抽出に CLIP モデルを利用し、画像内の最も関連性の高いラベルの特徴を特定して、機械翻訳の結果を改善する。CLIP モデルは最初に ″ViT-B-32″ 事前トレーニング パラメーターで読み込まれ、画像エンコーダーは画像から特徴を抽出し、将来のネットワーク入力のために保存される。

Visual Genome データセットのオブジェクト クラス ラベルは、CLIP テキストエンコーダーの入力テキストとして使用され、画像の特徴と 1600 のラベルがスコア付けされ、並べ替えられ、スコアが最も高い上位 5 つのラベルの特徴が保持される。このプロセスは、画像の特徴ごとに個別に実行され、5 つの対応するラベルの特徴が生成される。これらは保存され、ネットワーク入力として利用される。

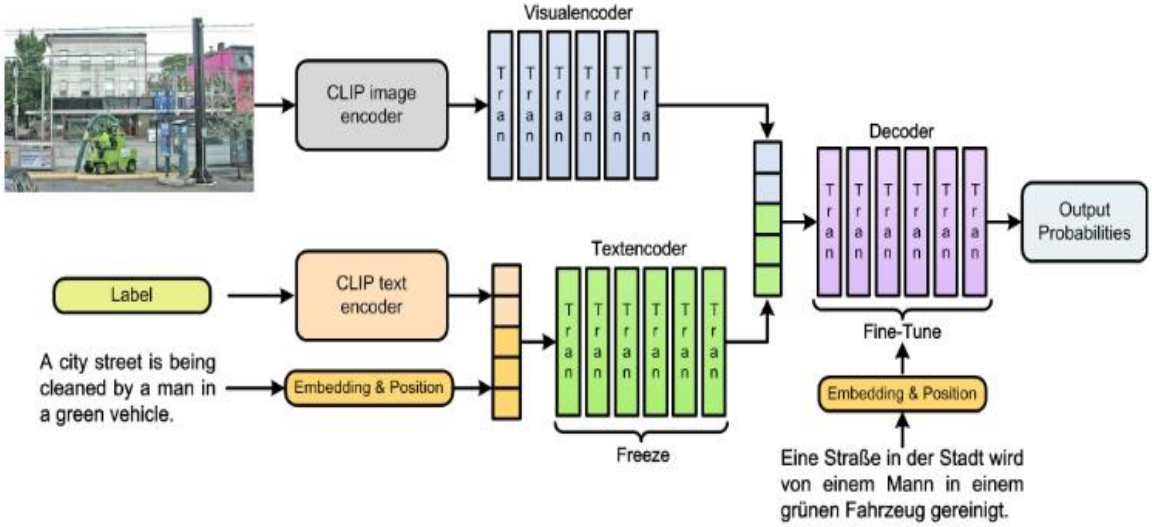

図 5.1

図5.1に示すよう、ラベル機能にとエンコードされたテキストは、「連結」操作を使用して連結され、ラベルとテキスト機能を結合するためにテキスト エンコーダーに渡される。結合された機能とビジュアルエンコー

ダーからの出力は、″Concat″ 操作を使用して連結されまする。 デコーダの変換モジュールが融合され、線形層と対数ソフトマックス関数を適用することで、最終的な出力確率が得られる。

モデルのトレーニング プロセスでは、最初にラベルとテキストの特徴のみを使用して基本的なテキスト翻訳モデルをトレーニングし、そのデルの重みを保存する。 次に、画像の特徴と、画像の特徴を処理するための追加の Transformer モジュールをモデルに追加する。 以前に保存した重みをモデルに読み込み、テキストとラベルの機能を処理する Transformer モジュールの重みを修正する。最後に、デコーダの重みを微調整する。

この研究では、BLEUとMeteorの自動評価メトリクスを使用して機械翻訳モデルの品質を評価する。モデルはPyTorch深層学習フレームワークを使用して実装され、すべての実験はNVIDIA Tesla P100 GPUと16GBのRAMを搭載したKaggleノートブックで実行される。バッチサイズは128、エポックは50であり、モデルの最適化にはAdamWオプティマイザーが使用される。学習率はCosineAnnealingLRを使用して調整され、ドロップアウトは0.5に設定される。予測にはビームサーチが採用され、ビームサイズは4に設定される。

第 6 章

実験

6.1実験設定

本研究では、CLIPモデルを用いて画像のラベル情報を抽出し、関連するラベル情報を保持し、テキストとともにテキストエンコーダに入力する。モデルの構築では、まずテキストモデルを訓練し、次にテキストエンコーダのパラメータを固定します。次に、ビジュアルエンコーダを導入してトレーニングを行います。このとき、テキストエンコーダのパラメータは更新されません。テキストエンコーダとビジュアルエンコーダは別々に学習するので、画像情報が実際に機能するかどうかを判断することができる。

実験データセット

本実験では、MSCOCOデータセットに基づいて新しい中日データセットを構築した。

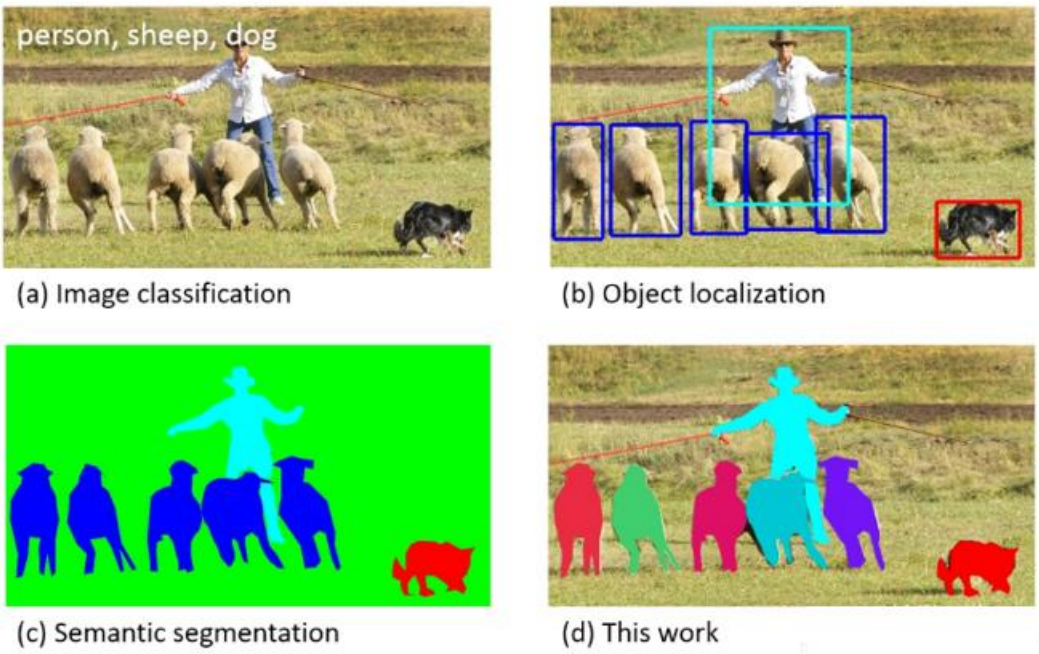

図6.1

MSCOCO 2014のデータセットには82783個の訓練サンプル、40504検証サンプル及び40775のテストサンプルがあり、他に27万の分割人像図と88.6万の分割物体図がある。このデータセットの主な特徴：
目標分割

- テキスト認識
- 各図に基づく多目標
- 30万図
- 2百万例
- 80種類のカテゴリ
- 平均1枚の図5目標。

MSCOCO 2014のデータを翻訳ソフトウェアを通じて中国語と日本語に変換し、それから人工的に比較検査を行い、中日翻訳の正確性、画像と文字の対応性を含み、最後に26477本のデータからなる中日データセットを得た。図6.2は、モデルBLEUと損失の暦による変化曲線を示している。

中日データセットには26,477件のデータが含まれている。そのうち24477件をトレーニングセットとし、1000件を検証セットとし、1000件をテストセットとした。

CLIP モデル

CLIPの事前学習済みモデルは，OpenAI が公開しているモデルを用いる．これは， Hugging Face 社の Transformers ライブラリから，モデル名’openai/clip-vit-base- patch32’で利用できるモデルである．このモデルは，画像エンコーダとしてViT-B/32 Transformer アーキテクチャ，テキストエンコーダとして12層の Transformerアーキテクチャ の Base モデル， 49,152 の語彙数と最大入力長が76に設定されている．

6.2実験結果

図6.2はモデルBLEUと損失に基づく暦の変化曲線を示す図である。事前に訓練されたテキストモデルをロードした後、私たちの機械翻訳モデルは高い初期BLEUスコアを獲得しました。10番目の暦の後、損失値に明らかな変化は観察されず、モデルが収束する可能性があることを示している。トレーニング中にBLEUスコアを38.3から46.90以上に上げることができます。これは、翻訳の質を高める上で当社のモデルが有効であることを示している。

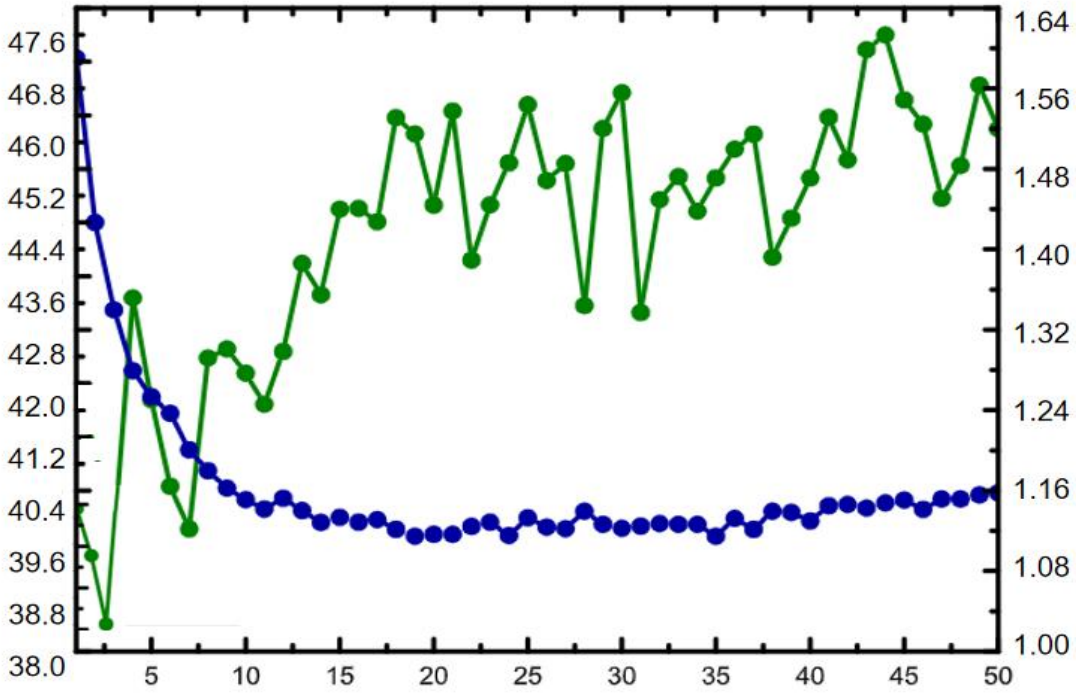

図　6.2

第7章

考察

7.1　実験対比

本実験では、英中文と日本語を含む中日データセット上の各種モデルの性能を比較した。その結果、私たちのモデルは2つの言語のBLEUスコアの比較でCLIPが他のモデルより優れており、画像の参加により翻訳結果がより正確になることが明らかになった。

トレーニングの過程で、BLEUスコアを38.3から46.9以上(表１)に上げることができ、これは私たちのモデルが翻訳品質を高める上で有効であることを示している。

Model	BLUE	Meteor
Transformer	38.3	55.6
Ours	46.91	58.74

表 1

本実験の過程で、多くのモデルが曖昧源語や複雑な文を正確に翻訳する際に困難に遭遇することが観察された対応する視覚的特徴。私たちのモデルの有効性を証明するために、私たちのモデルから生成された4つのケースの最適な翻訳とベースラインを表2に示す。モデルが予測対象に関連する動詞や形容詞を翻訳するのが難しい場合、視覚的特徴は特に有用である。例えば、第二の画像では、「ピザ」と「蔬菜」という言葉はオブジェクトとリンクしておらず、文のベースライン翻訳が逆になっています。しかし、私たちのモデルは画像の助けを得て正確な翻訳を生成することができる。

曖昧の名詞に関わる場合、領域視覚的特徴がより有用であることが証明された。例えば、2枚目の画像では、ベースラインはソース単語「白い衣」を「白衣」に翻訳し、私たちのモデルはそれを正確に「黒白衣」に翻訳します。これらのケースは、私たちのモデルが視覚情報を有効に利用して翻訳を修正し、さまざまな翻訳シーンの曖昧さを表現し、解消する上でより高い精度を実現できることを示している。

	ソース	青いフリスビーを持つ手があります。
	参照	有一只手拿着蓝色的飞盘。
	ベースライン	有一只手拿着一个蓝色的飞盘。
	私たち	有一只手拿着一个蓝色的飞盘。
	ソース	厚くて綺麗な焦げ目がついたピザです。
	参照	又厚又漂亮的棕色比萨饼。
	ベースライン	披萨很厚，呈漂亮的棕色。
	私たち	又厚又漂亮的棕色蔬菜比萨饼。
	ソース	みどりいろをした列車が線路をはしっています。
	参照	绿色列车在铁轨上行驶。
	ベースライン	一列绿色列车在铁轨上行驶。
	私たち	绿色的列车正在铁路上行驶。
	ソース	白い服を着た髪が黒く目が緑色の女性がいます。
	参照	有一个黑发绿眸的女子，身着一身白衣。
	ベースライン	有一个黑发绿眼穿着白衣的女人。
	私たち	有一个黑发绿眸的女子，穿着一身黑白衣。

表 2

第 8 章

結論

本論文では、CLIPモデルの支援を受けて機械翻訳の品質を向上させるための新しい手法を提案する。提案手法では、CLIPモデルを使用してラベルの特徴を抽出し、それを変換器を介してテキストの特徴と融合する。視覚モデルがテキストモデルに与える影響を調査するために、事前に訓練されたテキストの重みを読み込み、テキストエンコーダの重みを凍結し、視覚がテキストを向上させる効果を評価しました。さらに、CosineAnnealingLRアルゴリズムを導入して学習率を調整しました。中日翻訳タスクでの実験結果は、モデルの効果を強く示している。

将来的には、提案手法の他の翻訳タスクへの汎化可能性を調査し、視覚とテキストモデルを統合して翻訳品質をさらに向上させるための他の方法を探る予定です。

謝辞

本研究を進めるにあたって、多くのご指導を頂いた指導教員の新納浩幸教授に感謝致します。また、日常の議論を通して多くの知識、示唆を頂いた新納研究室の皆様に感謝します。

参考文献

[1] Myeongseob Ko, Ming Jin, Chenguang Wang, Ruoxi Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4871-4881

[2] Mirowski P, Banarse D, Malinowski M, et al. Clip-clop: Clip-guided collage and photomontage[J]. arXiv preprint arXiv:2205.03146, 2022.

[3] Yu K, Kim H, Kim J, et al. A Study on Webtoon Generation Using CLIP and Diffusion Models[J]. Electronics, 2023, 12(18): 3983.

[4] Vaiani L, Cagliero L, Garza P. PoliTo at SemEval-2023 Task 1: CLIP-based Visual-Word Sense Disambiguation Based on Back-Translation[C]//Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023). 2023: 1447-1453.

[5] Gao P, Geng S, Zhang R, et al. Clip-adapter: Better vision-language models with feature adapters[J]. International Journal of Computer Vision, 2023: 1-15.

[6] Bahng H, Jahanian A, Sankaranarayanan S, et al. Exploring visual prompts for adapting large-scale models[J]. arXiv preprint arXiv:2203.17274, 2022.

[7] Couairon G, Douze M, Cord M, et al. Embedding arithmetic of multimodal queries for image retrieval[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 4950-4958.

[8] Wu H H, Seetharaman P, Kumar K, et al. Wav2clip: Learning robust audio representations from clip[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 4563-4567.

[9] Baldrati A, Bertini M, Uricchio T, et al. Conditioned image retrieval for fashion using contrastive learning and CLIP-based features[M]//ACM Multimedia Asia. 2021: 1-5.

[10] Lei J, Li L, Zhou L, et al. Less is more: Clipbert for video-and-language learning via sparse sampling[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 7331-7341.

[11] Bianchi F, Attanasio G, Pisoni R, et al. Contrastive language-image pre-training for the italian language[J]. arXiv preprint arXiv:2108.08688, 2021.

[12] Kamath A, Singh M, LeCun Y, et al. Mdetr-modulated detection for end-to-end

multi-modal understanding[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 1780-1790.

[13] Wang D, Su J, Yu H. Feature extraction and analysis of natural language processing for deep learning English language[J]. IEEE Access, 2020, 8: 46335-46345.

[14] Pramanik S, Agrawal P, Hussain A. Omninet: A unified architecture for multi-modal multi-task learning[J]. arXiv preprint arXiv:1907.07804, 2019.