AI 特許紹介(79)
AI 特許を学ぶ！究める！
～BLIP-2 特許～

2025 年 8 月 8 日
河野特許事務所
所長弁理士　河野英仁

「AI 特許紹介」シリーズは、注目すべき AI 特許のポイントを紹介します。熾烈な競争となっている第４次産業革命下では AI 技術がキーとなり、この AI 技術・ソリューションを特許として適切に権利化しておくことが重要であることは言うまでもありません。

AI 技術は Google, Microsoft, Amazon を始めとした IT プラットフォーマ、研究機関及び大学から毎週のように新たな手法が提案されており、また AI 技術を活用した新たなソリューションも次々とリリースされています。

本稿では米国先進 IT 企業を中心に、これらの企業から出願された AI 特許に記載された AI テクノロジー・ソリューションのポイントをわかりやすく解説致します。

1.概要
特許出願人　Salesforce
出願日　2023 年 1 月 27 日
公開日　2024 年 5 月 16 日
公開番号 US20240161520
発明の名称　視覚言語事前学習フレームワークのためのシステムと方法

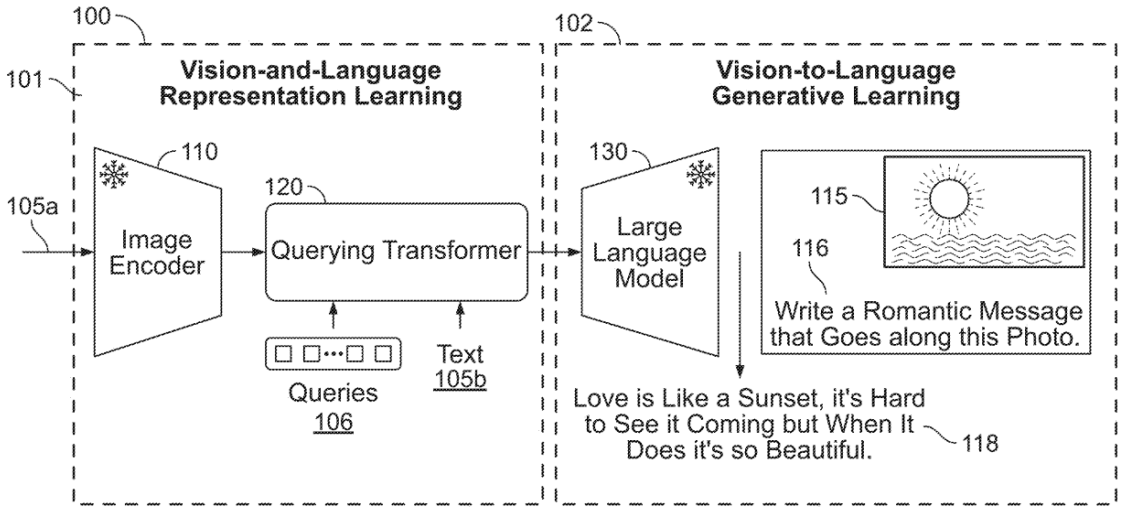
520 特許は画像及びクエリからテキストを生成するタスクを実行する画像エンコーダと言語モデルとの間にクエリトランスフォーマを設け、画像エンコーダ及び言語モデルのパラメータを固定したうえでクエリトランスフォーマを学習することで、学習効率を向上させることが可能な BLIP-2 技術に関する。

2.特許内容の説明
従来、視覚言語の事前学習は、大規模な画像とテキストのペアデータセットを用いてモデル全体をエンドツーエンドで学習する必要があった。しかし、パフォーマンス要求の高まりによりデータセットとモデルの両方の規模が拡大すると、従来のエンドツーエンドのフレームワークでは膨大な計算コストが発生し、視覚言語モデルのスケーラビリ

ティが低下する。

視覚言語モデルにおける効率性と多機能性の必要性に鑑み、画像エンコーダ、クエリトランスフォーマ、および事前学習済み言語モデルを含む、マルチモーダル視覚言語モデルの学習フレームワークを提供する。このフレームワークでは、軽量クエリトランスフォーマが唯一の学習可能なモジュールとなる。これにより、学習効率を大幅に向上させることができる。

図 1 は、2 段階視覚言語事前学習フレームワーク 100 のアーキテクチャ例を示すブロック図である。



画像エンコーダ 110、クエリトランスフォーマ 120、および(大規模)言語モデル(LLM) 130 から構成されるマルチモーダル視覚言語モデルは、視覚言語事前学習フレームワーク 100 によって学習される。

具体的には、画像エンコーダ 110 及び言語モデル 130 等のユニモーダルモデルは、学習中に凍結される。クエリトランスフォーマ 120 は、学習可能なクエリベクトル 106 のセットを用いて、凍結された画像エンコーダ 110 から視覚的特徴を抽出する軽量トランスフォーマである。言い換えれば、クエリトランスフォーマ 120 は、凍結された画像エンコーダ 110 と凍結された LLM130 の間の情報ボトルネックとして機能し、入力画像 105a から最も有用な視覚的特徴を LLM130 に供給して、目的のテキストを出力する。例えば、クエリトランスフォーマ 120 には 188M 個のパラメータが含まれる場合があるが、これは LLM 及び画像エンコーダと比較して、更新するパラメータ数が比較的少ない数である。

事前学習フレームワーク 100 は、2 つの段階 101 と 102 で構成される。最初の事前学習段階 101 では、視覚言語表現学習によって、クエリトランスフォーマがテキストに最も関連性の高い視覚表現を学習する。この段階では、画像エンコーダ 110 は固定され、クエリトランスフォーマ 120 とクエリ 106 のみが更新される。

第 2 の事前学習段階 102 では、更新されたクエリトランスフォーマ 120 の出力を、出力テキストを生成する LLM130 に接続することにより、視覚から言語への生成学習が実行される。クエリトランスフォーマ 120 は、その出力視覚表現が LLM130 によって解釈できるように再度学習される。第 2 段階では、画像エンコーダ 110 と LLM130 は固定され、クエリトランスフォーマ 120 とクエリ 106 のみが更新される。

2 つの学習段階 101-102 の後、フリーズ画像エンコーダ 110、学習済みクエリトランスフォーマ 120、およびフリーズ LLM130 からなるマルチモーダル視覚言語モデルは、ゼロショットファインチューニングを用いて、様々な視覚言語タスクを実行するために用いられる。例えば、入力画像 115 とガイドテキスト 116 が与えられた場合、マルチモーダル視覚言語モデル全体は、ガイドテキスト 116 に基づいて応答テキスト 118 を生成する。

3.クレーム

520 特許のクレーム 1 は以下の通りである。

1. 視覚言語タスク用のマルチモーダルフレームワークを事前学習する方法において、

通信インターフェースから画像と画像に付随するテキストを受信し、

画像エンコーダによって、画像を画像表現にエンコードし、

クエリトランスフォーマによって、画像表現とクエリのセットを変換された表現に変換し、

クエリトランスフォーマによって、少なくとも部分的に前記テキストに基づいてテキスト表現を生成し、

画像エンコーダを固定したまま、変換された表現とテキスト表現とに基づいて計算された 1 つ以上の視覚言語学習目標に従ってクエリトランスフォーマを学習し、

事前学習済み言語モデルによって、更新されたクエリトランスフォーマからの出力表現に基づいてデコードされた出力テキストを生成し、

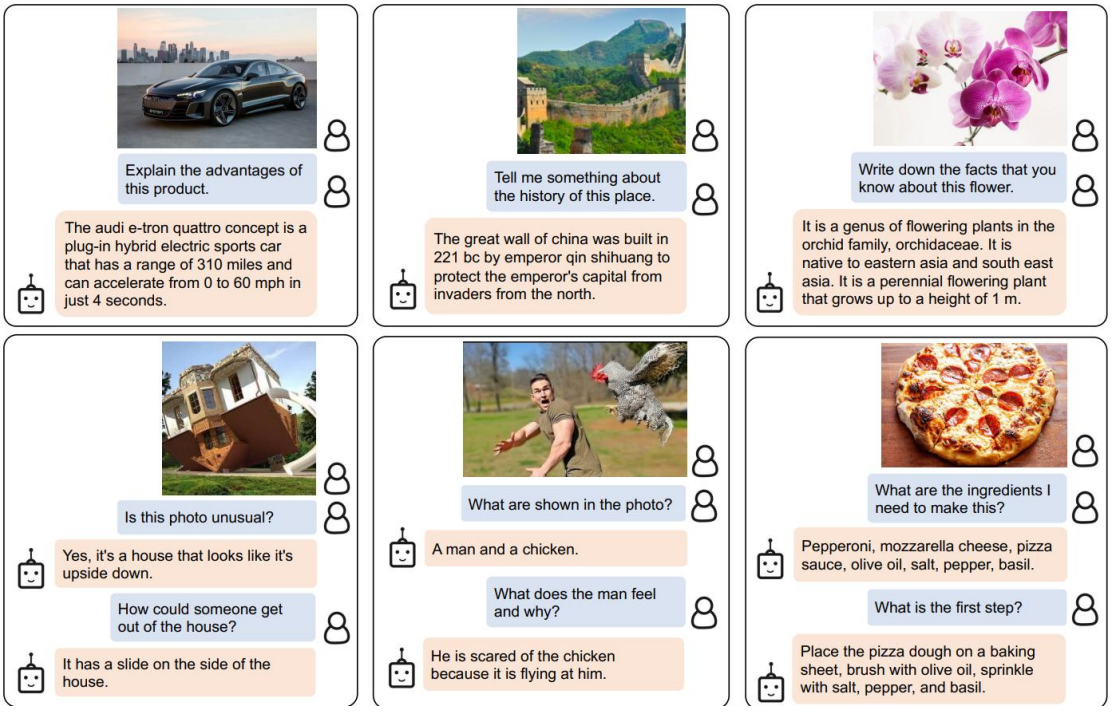
デコードされた出力テキストと画像に付随するテキストに基づいて損失を計算し、

画像エンコーダと事前学習済み言語モデルとを固定したまま、損失に基づいてクエリトランスフォーマを学習する。

4. 本特許に関連する論文

本特許に関する論文“Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models”[1]が、Salesforce の Junnan Li 氏らにより公表されている。

下記図に BLIP-2 モデルと ViT-g 及び FlanT5XXL を使用して、指示されたゼロショットの画像からテキストへの生成例を示す。



論文では、様々なゼロショット視覚言語タスクにおける BLIP-2 の結果の概要が以下の表のとおり示されている。

[1] Junnan Li, et al. “Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models” arXiv:2301.12597v3 [cs.CV] 15 Jun 2023

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

Models	#Trainable Params	Open-sourced?	Visual Question Answering VQAv2 (test-dev) VQA acc.	Image Captioning NoCaps (val) CIDEr	Image Captioning NoCaps (val) SPICE	Image-Text Retrieval Flickr (test) TR@1	Image-Text Retrieval Flickr (test) IR@1
BLIP (Li et al., 2022)	583M	✓	-	113.2	14.8	96.7	86.7
SimVLM (Wang et al., 2021b)	1.4B	✗	-	112.2	-	-	-
BEIT-3 (Wang et al., 2022b)	1.9B	✗	-	-	-	94.9	81.5
Flamingo (Alayrac et al., 2022)	10.2B	✗	56.3	-	-	-	-
BLIP-2	188M	✓	**65.0**	**121.6**	**15.8**	**97.6**	**89.7**

従来の最先端モデルと比較して、BLIP-2 は、視覚言語の事前学習において、学習可能なパラメータ数を最小限に抑えながら、最高のゼロショット性能を達成している。例えば、BLIP-2 はトレーニング可能なパラメータが 54 分の 1 であるゼロショット VQAv2 で Flamingo80B を 8.7%上回っている。

以上

著者紹介

河野英仁

河野特許事務所、所長弁理士。立命館大学情報システム学博士前期課程修了、米国フランクリンピアースローセンター知的財産権法修士修了、中国清華大学法学院知的財産夏季セミナー修了、MIT(マサチューセッツ工科大学)コンピュータ科学・AI 研究所 AI コース、生成 AI ビジネスコース修了。

AI 特許コンサルティング、医療 AI 特許コンサルティングの他、米国・中国特許の権利化・侵害訴訟を専門としている。著書に「世界のソフトウエア特許(共著)」、「FinTech 特許入門」、「AI/IoT 特許入門 3」、「ブロックチェーン 3.0(共著)」がある。