

AI 特許紹介(75)
AI 特許を学ぶ！究める！
～InstructBLIP 特許～

2025 年 4 月 10 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 11 月 9 日

公開日 2024 年 5 月 16 日

公開番号 US2024/0160858

発明の名称 視覚言語モデル指示チューニングのためのシステムおよび方法

858 特許は、マルチモーダルエンコーダにより、入力された画像の画像表現とテキスト指示をクロスアテンドし、その後言語モデルにより、テキスト指示に応答した視覚言語タスク出力を生成する InstructBLIP 技術に関する。

2.特許内容の説明

図 1 は、視覚言語指示モデルフレームワークを示す簡略図である。

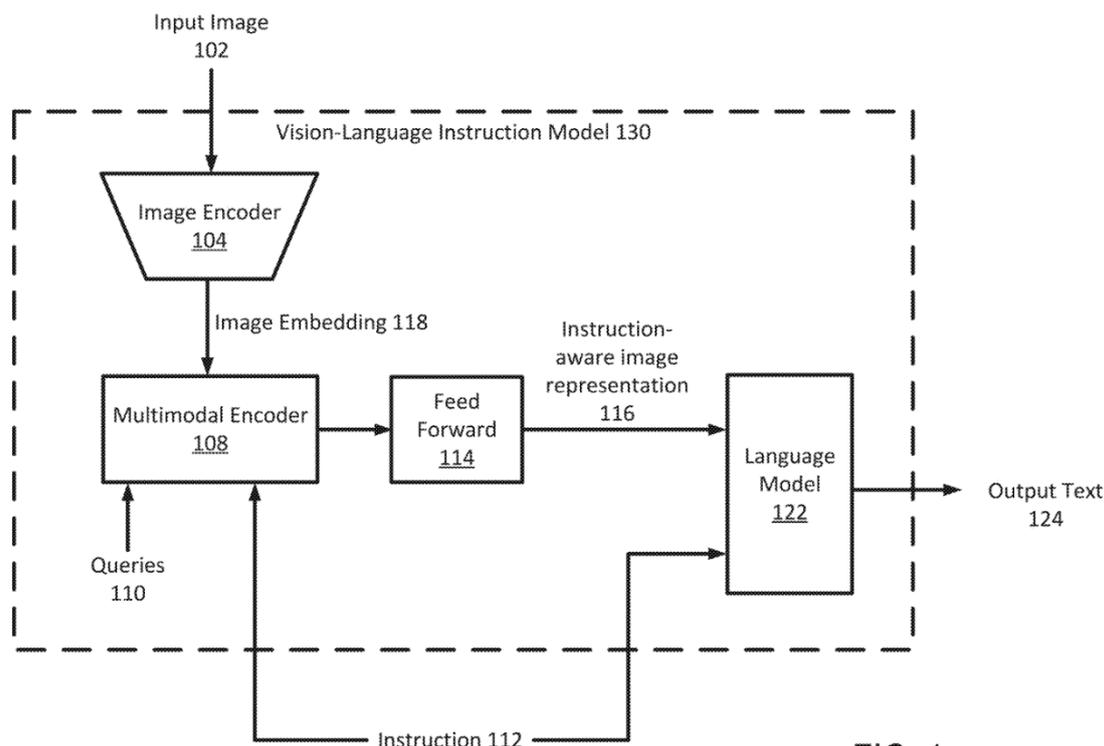


FIG. 1

視覚言語指示モデル 130 は、言語モデル 122 と、言語モデル 122 の指示認識画像表現 116 の生成を支援するマルチモーダルエンコーダ 108 とを含む。視覚言語指示モデル 130 は、入力画像 102 と指示 112 を受け取り、それらの入力に基づいて出力テキスト 124 を生成する。たとえば、入力画像 102 は、テーブル上のさまざまな野菜やその他の材料の画像である。指示 112 は、「この画像について詳しく教えてください。」である。これらの例示的な入力により、視覚言語指示モデル 130 は、次のような出力テキストを生成する。

「この画像は、テーブル上に並べられたニンジン、キュウリ、トマト、ナッツなどのさまざまな野菜のコレクションを示しています。ピーナッツ、カシューナッツ、カボチャの種など、さまざまな種類の材料が入った瓶がいくつかあります。これらの材料は、健康的な食事やスナックの一部である可能性があります。」

マルチモーダルエンコーダ 108 は、学習可能なクエリ 110 のセットを使用して、固定画像エンコーダ 104 から視覚的特徴を抽出する軽量のトランスフォーマーである。言い換えると、マルチモーダルエンコーダ 108 は、固定画像エンコーダ 104 と固定言語モデル 122 の間の情報ボトルネックとして機能し、入力画像 102 から最も有用な視覚的特徴を供給して、言語モデル 122 が目的のテキストを出力する。

入力画像 102 は、画像エンコーダ 104 によって、入力画像 102 のベクトル表現である画像埋め込み 119 にエンコードされる。画像エンコーダ 104 は、一般的な画像特徴を抽出する事前トレーニング済み画像エンコーダである。指示 112 は、テキストエンコーダによってテキスト特徴ベクトルにエンコードされる。画像特徴ベクトルおよびテキスト特徴ベクトルは、マルチモーダルエンコーダ 108 に入力される。

クエリ 110 は、トレーニングプロセスの一部として調整される可能性のある、ランダムに初期化されたベクトルである。マルチモーダルエンコーダ 108 は、指示 112 を使用して、入力画像 102 のうち指示 112 に最も関連する部分に注目することで、入力画像のベクトル表現（たとえば、命令認識画像表現）を生成する。フィードフォワードニューラルネットワークが対象のベクトル表現をさらに更新し、指示認識画像表現 116 を提供する。

指示認識画像表現 116 と指示 112 を組み合わせて、言語モデル 122 のプロンプトを生成する。言語モデル 122 は、プロンプトに基づいて、視覚言語タスク出力である出力テキスト 124 を生成する。

図 2 は、マルチモーダルエンコーダ 108 を示す簡略図である。

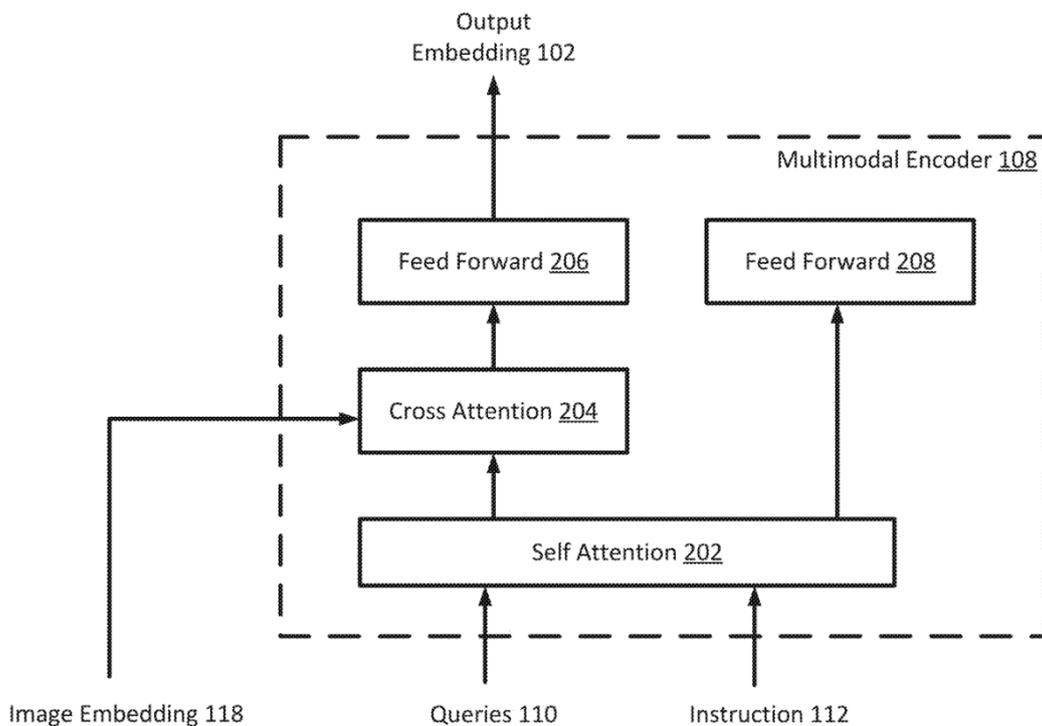


FIG. 2

マルチモーダルエンコーダ 108 は、同じセルフアテンション層 202 を共有する 2 つのトランスフォーマーサブモジュールで構成される。クロスアテンション 204 およびフィードフォワード 206 を備えた画像トランスフォーマーは、視覚的特徴抽出のためにクロスアテンション画像埋め込み 118 によって、フリーズ画像エンコーダ 104 と対話する。フィードフォワード 208 を含むテキストトランスフォーマーは、テキストエンコーダとテキストデコーダの両方として機能する。

画像エンコーダ 104 からの画像埋め込み 118 は、転送ブロックのスタックを含むクロスアテンション 204 に渡される。学習可能なクエリ埋め込み (クエリ) 110 の固定数がセルフアテンション 202 に入力される。クエリ 110 も調整可能であり、マルチモーダルエンコーダ 108 のパラメータとみなされ、トレーニング中にマルチモーダルエンコーダ 108 によって更新される。

クエリ 110 は、セルフアテンション層 202 を介して相互に対話し、セルフアテンション出力を生成する。クエリ 110 は、たとえばアテンションマスキングを介して、同じセルフアテンション層 202 を介して指示 112 とさらに対話する。次に、セルフアテンション出力は、クロスアテンション層 204 を介して、凍結画像の特徴、例えば凍結画像エンコーダ 104 からの画像表現と相互作用し、クロスアテンション出力を生成する。

クロスアテンション出力は、フィードフォワード層 206 に渡され、入力画像 102 の変換された画像表現として出力埋め込み 102 を生成する。たとえば、32 個のクエリが使用され、各クエリの次元は 768 である (マルチモーダルエンコーダ 108 の隠し次元と同じ)。出力埋め込み 102 のサイズ(32×768)は、凍結された画像特徴のサイズ (たとえば、一部の実施形態では 257×1024) よりもはるかに小さくなる。

一方、テキストトランスフォーマーは、入力指示 112 を受信してエンコードする。具体的には、指示 112 内のテキストトークンは、セルフアテンション層 202 を介して相互に作用し、セルフアテンション出力を生成する。次に、異なる視覚言語目標が採用され、クエリ 110 がテキスト指示 112 に最も関連のある画像表現から視覚情報を抽出するように強制される。フィードフォワード層 208 は、セルフアテンション出力からテキスト表現を生成する。トレーニング段階に応じて、指示 112 は、入力画像 102 に関連付けられた画像キャプションなどの別のテキスト入力になる。

3. クレーム

858 特許のクレーム 1 は以下の通りである。

1. 入力画像に関連するテキスト命令に対する視覚言語タスク出力を生成する方法であって、

データインターフェースを介して、入力画像と、入力画像に関連する指示を含むテキスト指示を受信し、

画像エンコーダを介して入力画像を第1の画像表現にエンコードし、

マルチモーダルエンコーダによって、第1の画像表現をテキスト指示にクロスアテンションすることに基づいて第2の画像表現を生成し、

ニューラルネットワークベースの言語モデルによって、第2の画像表現とテキスト指示とを組み合わせた入力に基づいて、テキスト指示に回答して視覚言語タスク出力を生成する。

4. 本特許に関連する論文

本特許に関する論文 “InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning”¹が、Salesforce の Wenliang Dai 氏らにより公表されている。

下記図は、画像及び指示を与えた場合に InstructBLIP により生成されるテキスト出力を示す。

¹ Wenliang Dai, et al. “InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning” arXiv:2205.11487v1 [cs.CV] 23 May 2022



What could have happened based on the current scene?

Based on the current scene in the image, it is possible that a hurricane or severe weather event caused significant damage to the buildings and infrastructure in the area. The person walking through the debris-covered area suggests that they are inspecting the damage caused by the disaster. The presence of palm trees indicates that the location may be in a tropical or sub-tropical region, which can be more susceptible to hurricanes and other natural disasters.



Introduce me this painting in detail.

The painting depicts a young girl with a pearl earring, also known as "The Girl with the Pearl Earring" or "Girl with a Pearl Earring". She is wearing a blue dress and has a pearl earring on her left ear. Her hair is pulled back into a bun, and she is looking directly at the viewer. This painting is a famous work of art by Dutch painter Johannes Vermeer, created in 1665. It is considered one of his masterpieces and is widely recognized for its intricate details and realistic portrayal of the young girl's facial expression.



Can you tell me about this image in detail?

The image depicts a collection of various vegetables including carrots, cucumbers, tomatoes, and nuts arranged on a stone surface. There are several jars filled with different types of ingredients, such as peanuts, cashews, sunflower seeds, and pumpkin seeds. These ingredients are likely to be part of a healthy meal or snack.

Can you give me an example of such meal?

An example of a healthy meal or snack that incorporates some of the ingredients in the image could be a salad made with mixed greens, carrots, cucumbers, tomatoes, and nuts. The salad can be dressed with a homemade vinaigrette made with olive oil, lemon juice, salt, and pepper. Additionally, the salad can be topped with pumpkin seeds or sunflower seeds for added crunch and nutritional value.

Sounds good! Show me steps of making this salad.

Sure! Here are the steps to make this healthy salad:
1. Wash and chop the mixed greens, carrots, cucumbers, and tomatoes into bite-sized pieces.
2. In a small bowl, whisk together olive oil, lemon juice, salt, and pepper to create a homemade vinaigrette dressing.
3. Place the chopped vegetables in a large bowl and toss them with the vinaigrette dressing.
4. Sprinkle pumpkin seeds or sunflower seeds on top of the salad for added crunch and nutritional value.
5. Serve the salad immediately or store it in the refrigerator until ready to enjoy.



Can you describe this image in detail?

The image depicts a person standing in front of an open doorway, looking out into space. They are wearing a jacket and appears to be contemplating the vastness of the universe beyond the doorway. In the background, there are multiple stars visible, creating a stunning celestial scene. The person's gaze is directed towards the center of the image, where a bright star can be seen shining brightly. There are also smaller stars scattered throughout the scene, adding to the overall cosmic atmosphere.

What if they jump over the doorway?

If the person were to jump over the doorway, they would be leaving the confines of their current location and venturing into the vastness of space. This could be a metaphorical representation of stepping out of one's comfort zone or embarking on a new adventure. However, it is important to note that such an action would be extremely dangerous and potentially life-threatening due to the harsh conditions of space.



What makes this dog special?

The dog in the image is wearing futuristic armor and has a glowing green eye, giving it a unique and futuristic appearance. This combination of elements makes the dog stand out from other dogs and adds an element of fantasy or sci-fi to the image. Additionally, the dog's posture and facial expression convey a sense of confidence and determination, further emphasizing its unique features.

論文には、他のモデルとの比較結果が下記の通り示されている。様々なデータセット

において InstructBLIP の性能が優れていることが理解できる。

	NoCaps	Flickr 30K	GQA	VSR	IconQA	TextVQA	Visdial	HM	VizWiz	SciQA image	MSVD QA	MSRVTT QA	iVQA
Flamingo-3B [4]	-	60.6	-	-	-	30.1	-	53.7	28.9	-	27.5	11.0	32.7
Flamingo-9B [4]	-	61.5	-	-	-	31.8	-	57.0	28.8	-	30.2	13.7	35.2
Flamingo-80B [4]	-	67.2	-	-	-	35.0	-	46.4	31.6	-	35.6	17.4	40.7
BLIP-2 (FlanT5 _{XL}) [20]	104.5	76.1	44.0	60.5	45.5	43.1	45.7	53.0	29.8	54.9	33.7	16.2	40.4
BLIP-2 (FlanT5 _{XXL}) [20]	98.4	73.7	44.6	68.2	45.4	44.1	46.9	52.0	29.4	64.5	34.4	17.4	45.8
BLIP-2 (Vicuna-7B)	107.5	74.9	38.6	50.0	39.7	40.1	44.9	50.6	25.3	53.8	18.3	9.2	27.5
BLIP-2 (Vicuna-13B)	103.9	71.6	41.0	50.9	40.6	42.5	45.1	53.7	19.6	61.0	20.3	10.3	23.5
InstructBLIP (FlanT5 _{XL})	119.9	84.5	48.4	64.8	50.0	46.6	46.6	56.6	32.7	70.4	43.4	25.0	53.1
InstructBLIP (FlanT5 _{XXL})	120.0	83.5	47.9	65.6	51.2	46.6	48.5	54.1	30.9	70.6	44.3	25.6	53.8
InstructBLIP (Vicuna-7B)	123.1	82.4	49.2	54.3	43.1	50.1	45.2	59.6	34.5	60.5	41.8	22.1	52.2
InstructBLIP (Vicuna-13B)	121.9	82.8	49.5	52.1	44.8	50.7	45.4	57.5	33.4	63.1	41.2	24.8	51.0

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「世界のソフトウェア特許(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。