AI 特許紹介(90)
AI 特許を学ぶ！究める！
～Flamingo 特許～

2026 年 7 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
特許権者　GDM Holding
出願日　2023 年 4 月 28 日
登録日　2026 年 5 月 5 日
登録番号 US12619654
発明の名称　マルチモードクエリ入力を処理するための言語モデル

654 特許は、画像と言語の few shot サンプルを対象画像と共にトークン処理層及びアテンション層が交互に配置されたネットワークに入力することにより、対象画像に対する言語を出力する Flamingo 技術に関する。

2.特許内容の説明
下記図はクエリ処理システム 200 のブロック図である。

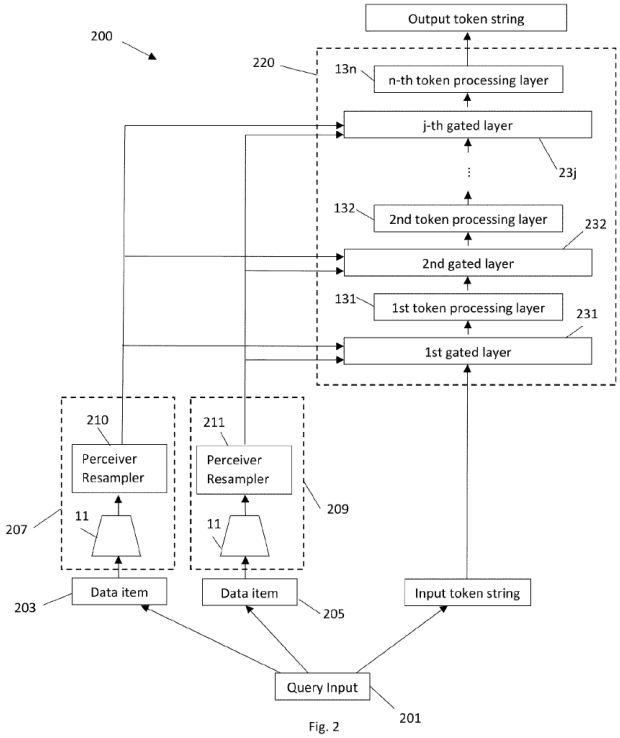


Fig. 2

クエリ処理システム 200 は、データ項目と入力トークン文字列を含むクエリ入力 201 を入力として受け取る。クエリ入力 201 は例えば 2 つのデータ項目 203、205 を含む。クエリ処理システム 200 は、クエリ入力 201 から 2 つのデータ項目 203、205 を抽出し、各データ項目を、モダリティネットワーク 207,209 を用いて処理する。

例えば、モダリティネットワーク 207 は、クエリ入力のデータ項目のうち、画像を受信し、モダリティネットワーク 209 は、クエリ入力のデータ項目のうち、音声データを受信する。各モダリティネットワーク 207、209 は、受信した各データ項目から、そのデータ項目に対応する圧縮表現を生成する。各モダリティネットワークは、事前学習済みのエンコーダネットワーク 11 と、それぞれのリサンプラー210、211 とを備える。各リサンプラー210、211 は、例えば、「Perceiver: General Perception with iterative attention」（A. Jaegle et al、2021）に記載されている「知覚器 Perceiver」を用いる。

リサンプラー210、211 は、対応するエンコーダネットワーク 11 によってデータ項目から生成された可変数の画像、ビデオ、またはオーディオの特徴を入力として受け取り、固定数の出力（例えば、視覚出力）を持つデータ項目の圧縮表現を生成する。

クエリ処理システム 200 はさらにデータ項目トークン処理モデル 220 を備える。データ項目処理モデル 220 は、トークン処理モデルのトークン処理層 131、132、...、13n

からなる層のスタックと、j 個のゲート付きクロスアテンション層 231、232、...、23j が交互に配置されている。リサンプラー210、211 からの出力は、ゲート付きクロスアテンション層への制御入力として使用される。入力トークン文字列は、データ項目トークン処理モデル 220 へのプロンプト入力として使用され、スタックの第 1 処理層（第 1 ゲート付きクロスアテンション層 231 など）に供給される。ゲート付きクロスアテンション層 231、232、…、23j はそれぞれ、モダリティネットワーク 207、209 から圧縮表現を受け取り、受け取った圧縮表現に基づいてゲーティング処理を実行する。

データは（図 2 では上向きに）データ項目トークン処理モデル 220 を通過し、出力トークン文字列が生成される。

下記図は、学習済みクエリ処理システム 200 の動作を示す説明図である。

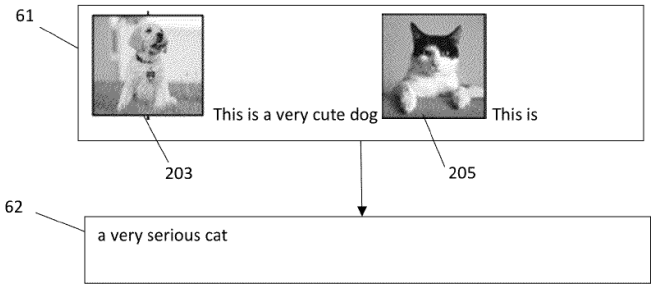


Fig. 6

クエリ入力 61 は、それぞれ猫と犬の静止画像である 2 つのデータ項目 203、205 を含む。データ項目 203、205 は、それぞれ対応するモダリティネットワーク 207、209 に入力され、データ項目 203、205 の対応する圧縮表現が生成される。クエリ入力 61 の単語は、クエリ処理システム 200 のクエリ入力用の入力トークン文字列を生成するために使用される。クエリ入力には、例えばクエリ入力 61 の各データ項目 203、205 に対応する、関連するデータ項目の存在を示す 1 つ以上の「マーカー」を含めることができる。例えば、入力トークン文字列は、次の ASCII 文字で構成される。
「<image> これはとてもかわいい犬です。<image> これは」。

クエリ処理システム 200 は、入力トークン文字列を層 231、131、232、132、...23j、13n に順に通過させることにより、入力トークン文字列を処理する。ここで、ゲート層 231、232、...、23j は、対応するデータ項目 203、205、すなわちクエリ入力 61 の 2 つの静止画像に基づいて、モダリティネットワーク 207、209 によって生成された視覚入力に基づいて制御される。このようにして、クエリ処理システム 200 は、図の出力トークン文字列 62 のような出力トークン文字列を生成する。出力トークン文字列は、「a very serious cat」と読める ASCII 文字の文字列である。これは、クエリ入力 61 の論

理的かつ文法的な続きである。最初の画像 203 とそれに関連付けられたテキスト「This is a very cute dog」は、「タスク例」を構成する。出力トークン文字列は、タスク例で例示されたタスクを第２のデータ項目 205（対象データ項目）に対して実行した結果である。

3.クレーム

654 特許のクレーム１は以下の通りである。

1. クエリ処理システムを学習するためのコンピュータ実装方法において、

前記クエリ処理システムは、入力トークン文字列と１つ以上のデータ項目とを含むクエリ入力に基づいて出力トークン文字列を生成するものであり、前記入力トークン文字列及び出力トークン文字列は、トークン語彙から選択されたトークンの文字列であり、前記データ項目は、前記トークン語彙から選択されたトークン以外のモダリティのものであり、

前記方法は、トークン処理層のスタックを含むトークン処理モデルを採用し、前記トークン処理層のスタックは、入力トークン文字列を受信し、対応する出力トークン文字列を生成するように構成され、また、各トレーニング例は少なくとも１つのデータ項目と少なくとも１つのトークン文字列を含むトレーニング例のデータベースを備え、

前記方法は、以下のステップを含み、

トークン処理モデルからのトークン処理層とゲート付きクロスアテンション層をインターリーブすることによりデータ項目トークン処理モデルを形成し、前記データ項目トークン処理モデルは、トークン文字列であるプロンプト入力を受信した際に、出力トークン文字列を生成するように構成され、前記トークン処理モデルは、トークン処理層のスタックを含み、前記トークン処理層のスタックは、入力トークン文字列を受信し、対応する出力トークン文字列を生成するように構成されており、トレーニング例のデータベースを含み、各トレーニング例は、少なくとも１つのデータ項目と少なくとも１つのトークン文字列を含み、

クエリ処理システムを形成し、該クエリ処理システムは、以下を含み、

（a）クエリ入力のデータ項目を受信し、各データ項目の１つ以上の圧縮表現を生成するように構成されたモダリティネットワーク；及び

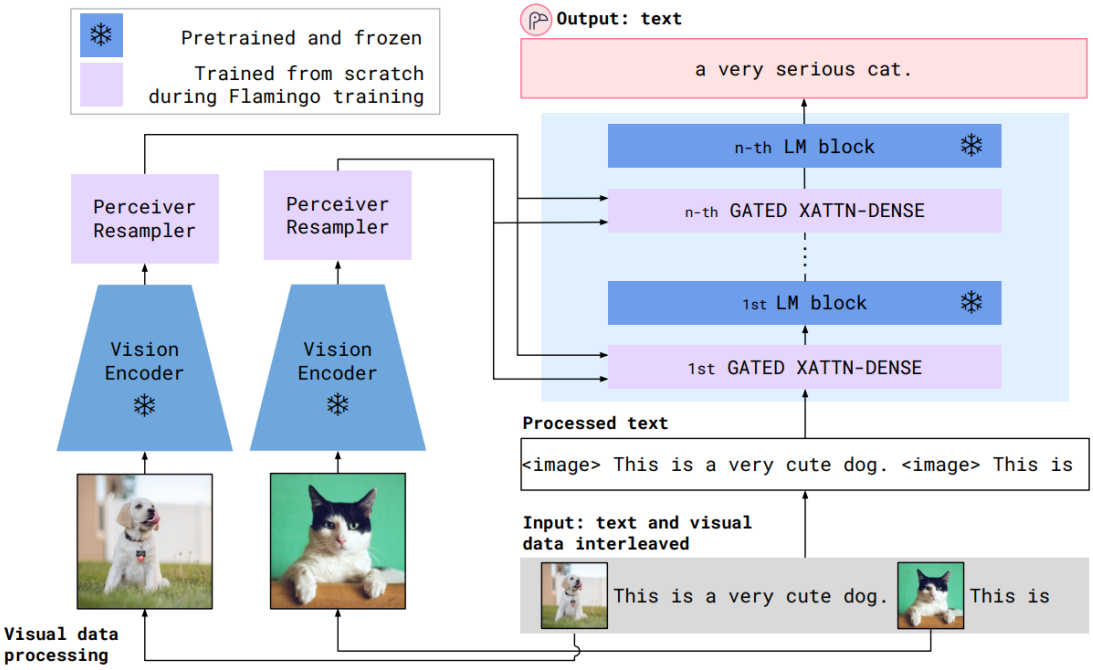
（b）クエリ入力の入力トークン文字列を含むプロンプト入力を受信するように構成されるデータ項目トークン処理モデルを備え、各ゲート付きクロスアテンション層は、圧縮表現の少なくとも１つを受信するように構成されており、

訓練データベースを用いて、モダリティネットワークと、複数のゲート付きクロスアテンション層を訓練する。

4. 本特許に関連する論文

本特許に関する論文“Flamingo: a Visual Language Model for Few-Shot Learning”[1]が、DeepMind の Jean-Baptiste Alayrac 氏らにより公表されている。

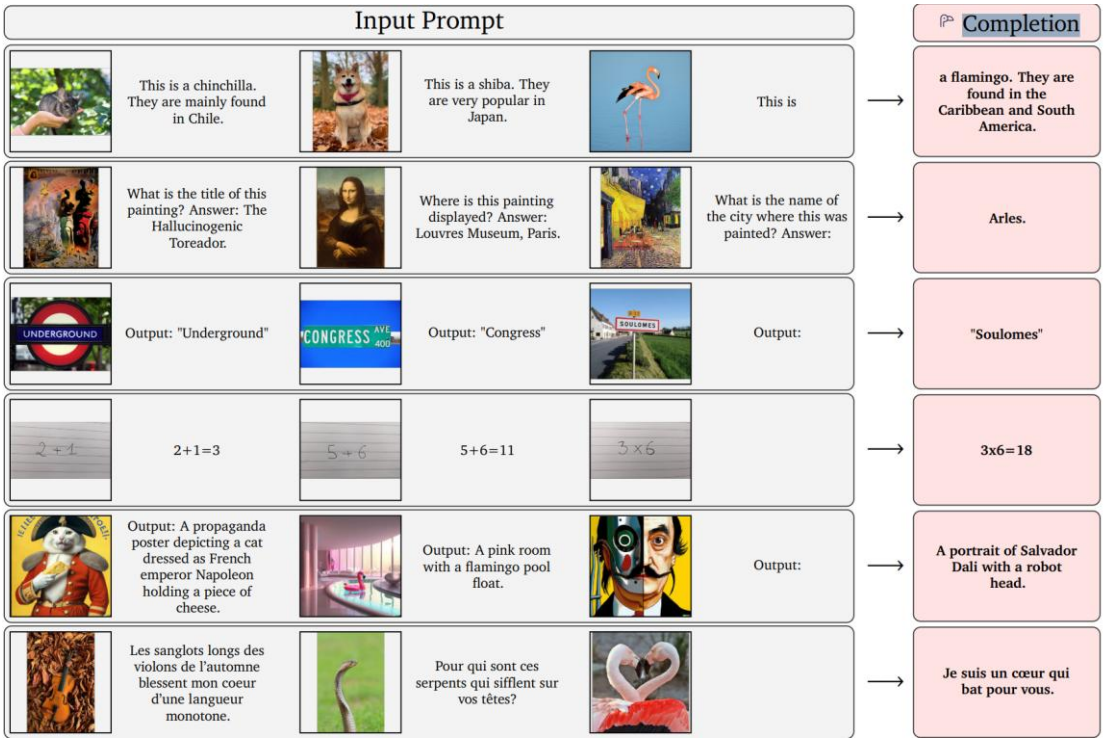
下記図は Flamingo のネットワーク構成図である。



Flamingo には犬の画像及び This is a very cute dog.のサンプルと、猫の画像及び This is のクエリとが入力される。２枚の画像はそれぞれビジョンエンコーダに入力され圧縮表現がリサンプラーに入力される。リサンプラーからの出力及びテキストはクロスアテンション層に入力される。またリサンプラーからの出力は LM ブロック層に対し交互に配置される後段のクロスアテンション層に入力される。なお、薄紫色はパラメータが隔週可能なブロックであることを示す。最終的に、ネットワークからは a very serious cat が出力される。

下記図は Flamingo の出力例を示す説明図である。

[1] Jean-Baptiste Alayrac, et al. “Flamingo: a Visual Language Model for Few-Shot Learning” arXiv:2204.14198v2 [cs.CV] 15 Nov 2022



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 4」、「ブロックチェーン 3.0(共著)」がある。