AI 特許紹介(61)

AI 特許を学ぶ！究める！

～ViViT(Video Vision Transformer)特許～

2024 年 2 月 7 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　Google

出願日　2021 年 7 月 8 日

公開日　2023 年 1 月 19 日

公開番号 US20230017072

発明の名称　ビデオの理解を改善するためのシステムと方法

072 特許は、ビデオデータから、ビデオデータ内の時空間のビデオトークンを抽出し、複数のビデオトークンを、ビデオトランスフォーマエンコーダモデルを含むビデオ理解モデルで処理する ViViT(Video Vision Transformer)技術に関する。

2.特許内容の説明

下記図は、ビデオ理解モデル 200 のブロック図を示す。

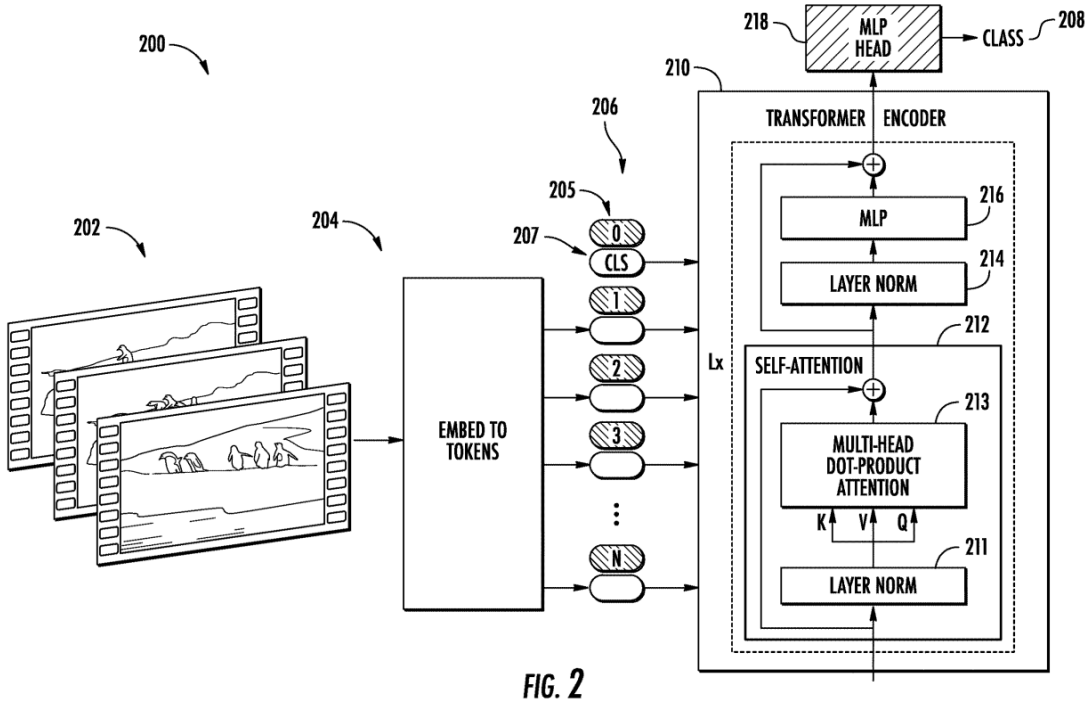

FIG. 2

ビデオ理解モデル 200 は、入力データ 202（例えば、ビデオデータ）を受信し、入力データ 202 の受信に応答して、出力データ 208（例えば、分類出力）を生成する。S204 で、モデル 200 は、ビデオデータ 202 から複数のビデオトークン 206 を抽出する。

ビデオトークン 206 は、ビデオデータ 202 の時空間情報の表現（例えば、埋め込み表現）である。ビデオ理解モデル 200 は、多数のビデオトークンを処理する。トークン 206 は、位置埋め込み 205 を含むことができる。さらに、トークン 206 は、分類トークン 207 を含む。

ビデオ理解モデル 200 は、ビデオトランスフォーマエンコーダモデル 210 を含む。トランスフォーマエンコーダモデル 210 は、アテンションメカニズム 212（例えば、セルフアテンションメカニズム）、正規化層 214、および多層パーセプトロン層 216 を含む。セルフアテンション機構 212 は、例えば、マルチヘッドドット積アテンション機構 213 に供給する正規化層 211 を含む。

ビデオ理解モデルは、トークン 206 のシーケンス全体を直接処理する。例えば、一連のトークン 206（例えば、位置埋め込み 205、分類(CLS)トークン 207 などを含む）は、トランスエンコーダモデル 210 に直接入力される。ビデオ理解モデル 200 は、入力シーケンス内の時空間トークンのすべてのペア間のペアごとの相互作用をモデル化することができる。トランスエンコーダ 210 からの出力は、出力を分類してビデオ分類出力 208 を提供する分類モデル（例えば、多層パーセプトロンヘッド）218 に提供する。

下記図 3A は、ビデオデータをトークン化するための均一フレームサンプリング手法のデータフロー図を示す。

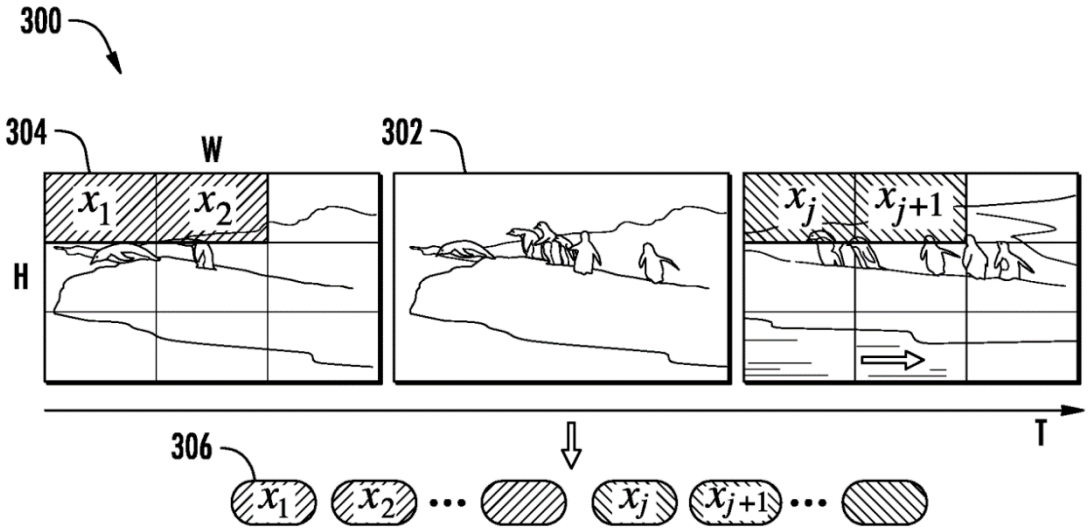

FIG. 3A

ビデオデータ 300 は、複数のビデオフレーム 302 を含む。各フレームは、複数の「パッチ」304 に分割される。各パッチ 304 は、それぞれのトークン 306 に投影またはラスタライズすることができる。トークン 306 は、シーケンス内のフレーム 302 およびパッチ 304 によって順序付けすることができる。

図３Ｂは、ビデオデータをトークン化するためのチューブレット埋め込みアプローチのデータフロー図を示す。

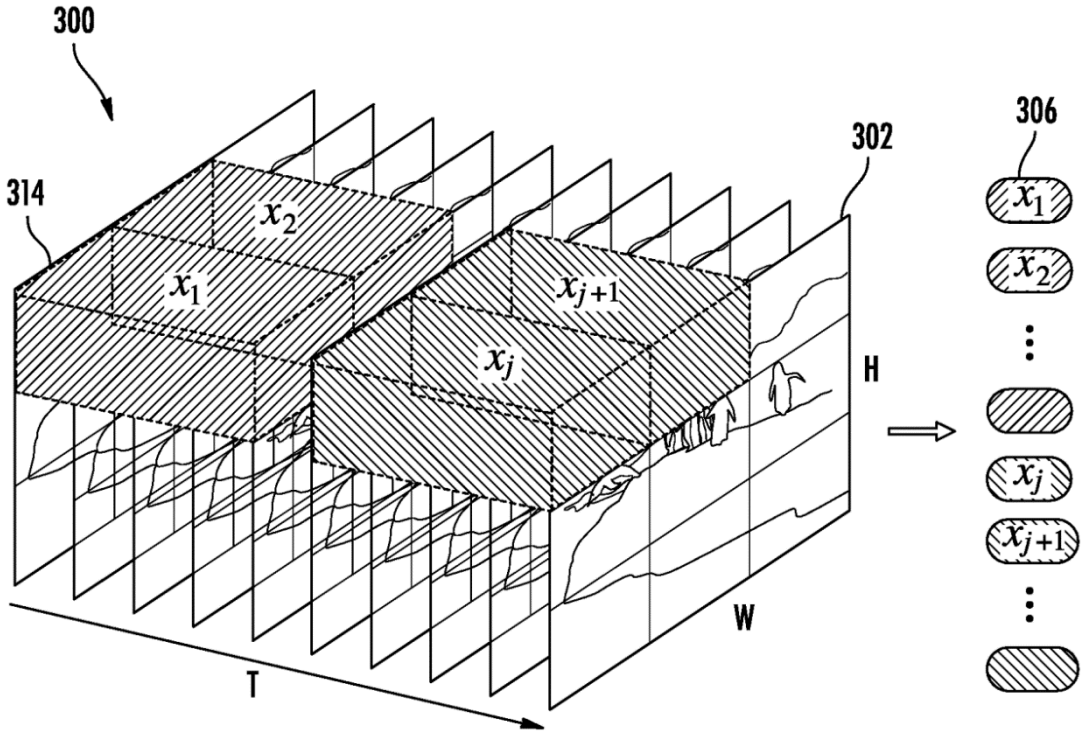

FIG. 3B

図３Ａと同様に、ビデオデータ 300 は複数のビデオフレーム 302 を含む。ビデオデータは、チューブレット 314 に分解することができる。各チューブレットは、１つ以上のビデオフレーム 302 にまたがる。例えば、チューブレット 314 は、複数のフレームにわたる共通の空間領域をカバーする。各チューブレット 314 は、対応するトークン 306 に射影することができる。

例えば、いくつかの実装形態では、ビデオトークン 306 は、長さ（例えば、l）および幅（例えば、w）を有し、多数のビデオフレーム 302（例えば、t）にまたがるチューブレット 314 から形成され、テンソル表現 (d 次元ベクトルなど)に投影される。。

次に、この一連の時空間トークン 306 は、ビデオ理解モデル 200 を通過する。図４Ａは、因数分解エンコーダ 400 のブロック図、図４Ｂは、ビデオ理解モデル 450 に組み込まれた因数分解エンコーダのブロック図を示す。

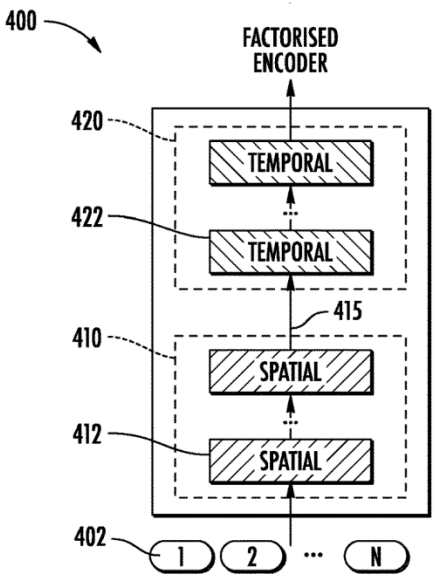

FIG. 4A

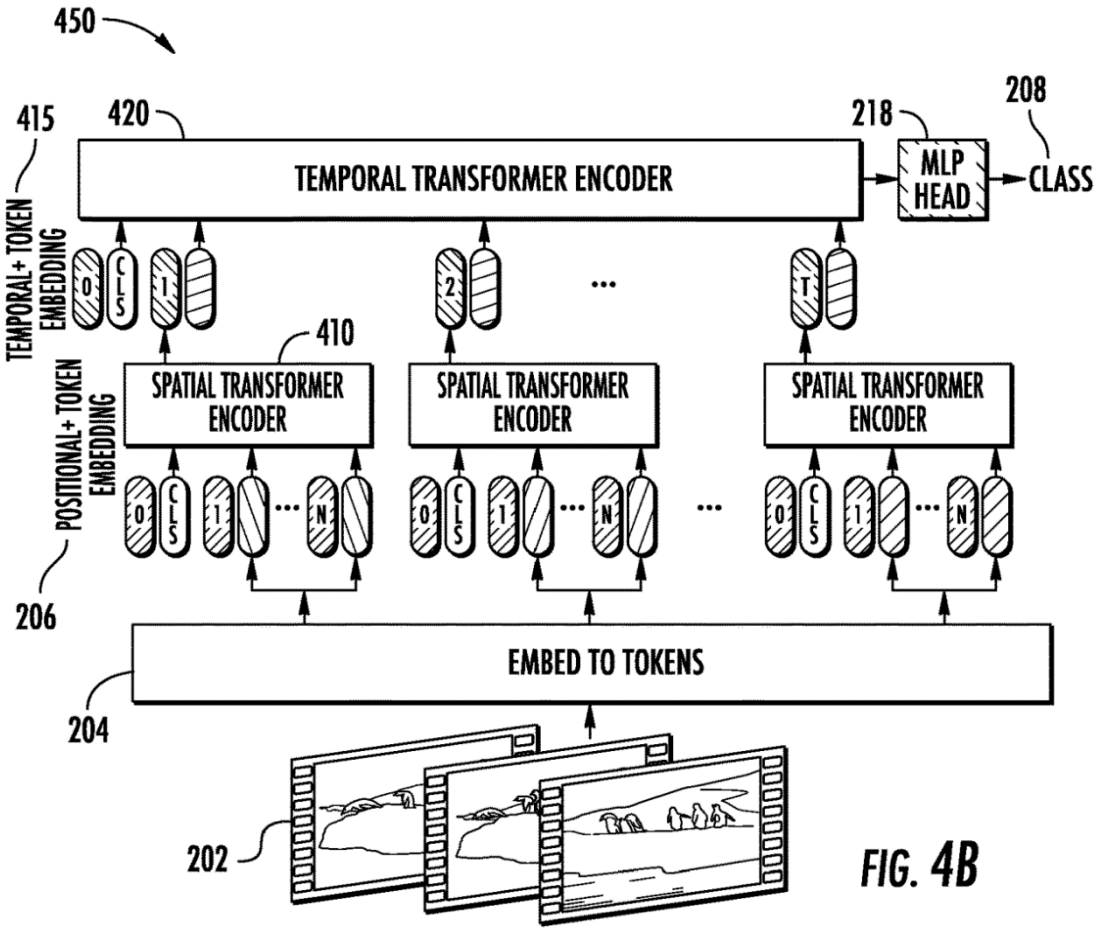

FIG. 4B

因数分解エンコーダ 400 は、例えば、空間変換エンコーダ 410 および時間変換エンコーダ 420 を含む、２つの別個のサブエンコーダ（例えば、変換器）を直列に含む。空間変換エンコーダ 410 は、１つまたは複数の空間変換エンコーダ層 412 を含み、時間

変換エンコーダ 420 は、１つまたは複数の時間変換エンコーダ層 422 を含む。

空間変換エンコーダ 410 は、複数のビデオトークン 402 を受信し、複数のビデオトークン 402 の受信に応答して、複数の時間表現 415 を生成する。時間変換エンコーダ 42 は、複数の時間表現 415 を受信し、複数の時間表現の受信に応答して、ビデオデータの時空間表現を生成し、時空間表現は分類されて、分類出力を生成する。

例えば、空間エンコーダ 410 は、同じ時間インデックスから抽出されたトークン 402 間の相互作用をモデル化する。各時間インデックス（例えば、フレーム）のトークン表現 415 は、空間エンコーダ 410 から取得される。

その後、時間エンコーダ 420 は、表現 415 間の相互作用をモデル化する。例えば、時間表現 415 は、時間エンコーダ 420（例えば、L_t 層 422 を含む）を通じて転送され、異なる時間インデックスからのトークン 402 間の相互作用をモデル化する。次いで、時間エンコーダ 420 からの出力は、（例えば、多層パーセプトロンモデルなどの分類モデルによって）分類することができる。

3.クレーム

072 特許のクレーム 1 は以下の通りである。

1. 精度を向上させてビデオデータを分類するためのコンピュータ実装方法において、

１つまたは複数のコンピューティングデバイスを含むコンピューティングシステムによって、複数のビデオフレームを含むビデオデータを取得し、

コンピューティングシステムによってビデオデータから、ビデオデータ内の時空間情報の表現を含む複数のビデオトークンを抽出し、

コンピューティングシステムによって、複数のビデオトークンを、ビデオトランスフォーマエンコーダモデルを含むビデオ理解モデルへの入力として提供し、

コンピューティングシステムによって、ビデオ理解モデルからの分類出力を受信する。

4. 本特許に関連する論文

本特許に関する論文“ViViT: A Video Vision Transformer”[1]が、Google Research の Anurag Arnab 氏らにより公表されている。

[1] Anurag Arnab et al. “ViViT: A Video Vision Transformer” arXiv:2103.15691v2 [cs.CV] 1 Nov 2021

論文では、他のモデルと比較した結果が下記テーブルに示されている。

(c) Moments in Time

	Top 1	Top 5
TSN [72]	25.3	50.1
TRN [86]	28.3	53.4
I3D [8]	29.5	56.1
blVNet [19]	31.4	59.3
AssembleNet-101 [54]	34.3	62.7
ViViT-L/16x2 FE	**38.5**	**64.1**

表 c に示すように、本論文の手法は、動画の種類が多様であり、ラベルノイズの多い本データセットにおいても、TSN（Temporal segment networks）等の最先端技術を大幅に上回っている。本論文では徹底したアブレーション研究を実施し、Kinetics 400 および 600、Epic Kitchens、Something-Something v2、Moments in Time を含む複数のビデオ分類ベンチマークで最先端の結果を達成し、ディープ 3D 畳み込みネットワークに基づく従来の手法を上回るパフォーマンスを実現している。

ViViT のコードは Github に公開されている。
https://github.com/google-research/scenic

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。