AI 特許紹介(52)
AI 特許を学ぶ！究める！
～VGD-GPT 特許～

2023 年 5 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　Salesforce
出願日　2020 年 4 月 28 日
登録日　2022 年 11 月 1 日
登録番号 US11487999
発明の名称　ビデオベースの対話のための事前トレーニング済み言語モデルによる時空間推論

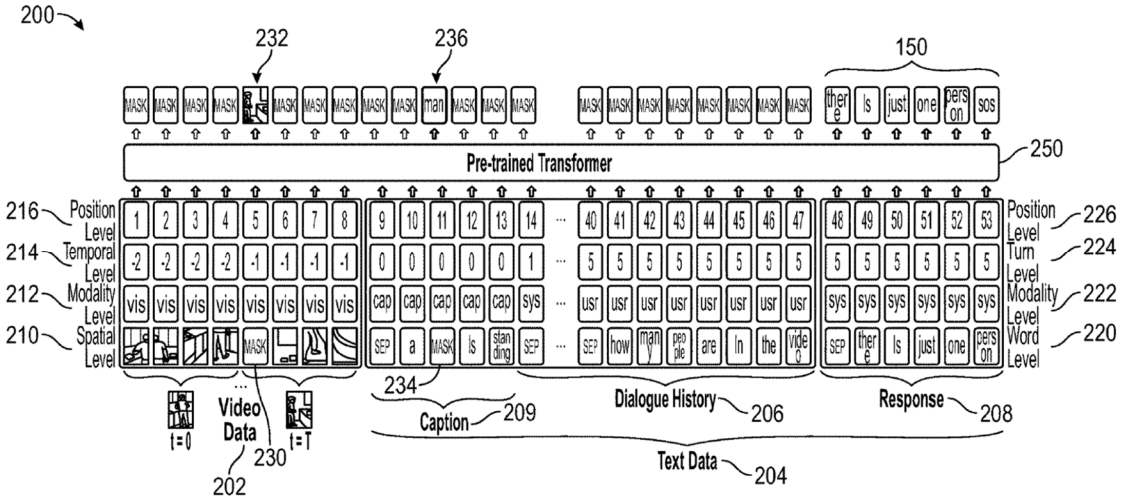
999 特許は、動画像及びテキストを符号化した上で事前トレーニング済み言語モデルに入力し､テキストを生成する VGD(Video-Grounded Dialogues)-GPT 技術に関する。

2.特許内容の説明
ビデオに基づく対話タスクは､対話エージェントがビデオに基づくニューラルネットワークを使用し、人間からの質問に複数回の対話で回答するビデオ質問応答（QA）の拡張である。ビデオの特徴表現には空間情報と時間情報の両方が含まれるため、ビデオに基づく対話はより困難となる。対話エージェントは、ビデオの「どこを見るか」(空間

的推論）と「いつ見るか」(時間的推論）という 2 つの主要な問題に対処する必要がある。ビデオの機能表現には空間情報と時間情報の両方が含まれるため、ビデオに基づく対話はより困難となる。

ビデオに基づいた対話の従来のアプローチは、入力ビデオの時間的分散に対する視覚と言語の推論に焦点を当てており、空間次元からの潜在的な信号を無視する。このような欠点は、非常に長く、多くのオブジェクトを含むビデオを推論する必要がある場合に、より顕著となる。

本発明は、GPT-2 のパワーを活用し、GPT-2 モデルを、空間次元と時間次元の両方にわたる異なるダイナミクスの複雑な特徴を含むビデオに基づく対話タスクに拡張する。下記図は、ビデオに基づく対話モデルのアーキテクチャまたはフレームワークの簡略図である。



視覚的表現とテキスト表現の組み合わせを含むビデオベースの対話タスクは、事前にトレーニングされた大規模なニューラルネットワーク言語モデル（Generative Pretrained Transformer 2 (GPT-2) モデルなど）に入力される前に、複数の符号化層を使用してファインチューニングされる。

ビデオに基づく対話モデル 200 は、ビデオデータ V202 およびテキストデータ T204 を入力 140 として受け取る。テキストデータ T204 は、モデル 200 と人間のユーザとの間の対話履歴 206 の一部を含む。対話履歴 206 は、人間の発話とモデル 200 の応答 208 との間の 1 つまたは複数のターンを含む。テキストデータ T204 はまた、ビデオデータ V202 に関連する、または対応するビデオキャプション C209 を含む。

モデル 200 フレームワークにより、言語モデルをファインチューニングして、ビデオの時空間レベルや対話コンテキストのトークン文レベルなど、さまざまなレベルの情報にわたって複数のモダリティにわたる依存関係をキャプチャできる。ビデオに基づく対話モデル 200 のフレームワークは、GPT モデルなどの事前に訓練されたトランスフォーマモデル 250 に基づくか、またはこれを用いて実装することができる。

事前トレーニング済みトランスフォーマモデル 250 は、GPT-2 アーキテクチャに基づいてトレーニングされる。GPT-2 モデルは、出力 150 である、現在の人間の発話に対するビデオに基づく対話応答を生成するように適合される。ビデオに基づく対話モデル 200 は、GPT モデル 250 の前段に、ビデオデータ V202 に対して様々なエンコーディングを実行するための層 210～216、およびテキストデータ T204 に対して様々なエンコーディングを実行するための層 220～226 を含む。

ビデオデータ V202 およびテキストデータ T204 は、モデル 200 の複数の符号化層 210～216 および 220～226 にわたって結合され、層 210～220 は、符号化された特徴に異なる属性を注入する。ビデオデータ V202 について、符号化層 210～216 は、空間レベル符号化層 210、モダリティレベル符号化層 212、時間レベル符号化層 214、および位置レベル符号化層 216 を含む。テキストデータ T204 の場合、符号化層 220～226 は、単語レベル符号化層 220、モダリティレベル符号化層 222、ターンレベル符号化層 224、および位置レベル符号化層 226 を含む。

モデル 200 の空間レベル符号化層 210 は、ビデオデータ V202 に対して空間レベル符号化を実行する。空間レベル符号化層 210 は、事前訓練された 2D CNN または 3D CNN ビデオモデルなどの事前訓練されたビデオモデルおよび RELU 活性化層を含む。各ビデオフレームまたはビデオセグメントは、事前にトレーニングされたビデオモデルを使用して抽出できる一連の空間領域として構造化できる。

入力ビデオ データ V 202 の場合、事前トレーニング済みの 2D CNN または 3D CNN ビデオモデルの出力は、Z_v^{re} と表すことができる。

$$Z_V^{re} \in \mathbb{R}^{F \times P \times d_{emb}}$$

d_{emb} は事前トレーニング済みのビデオモデルの特徴次元、F はサンプリングされたビデオフレームまたはビデオセグメントの結果の数、P は各ビデオフレームの空間領域の数である。

上記図を参照すると、F は T 個のサンプリングされたビデオフレームであり、P はビ

デオフレームごとの 4 つの空間領域である。2D CNN または 3D CNN ビデオモデルの出力 Z_V は、事前トレーニング済みのトランスフォーマモデル 250 の特徴次元 d に一致させるべく、出力 Z_V を調整線形アクティベーションユニット（ReLU）アクティベーションを使用した線形変換に渡すことにより、画像パッチのシーケンスとしてさらに再形成される。ReLU アクティベーションによる線形変換の出力は、ビデオデータ V202 の空間レベル特徴である。ビデオデータ V 202 の Z_V spatial への変換を以下に示す。

$$Z_V^{spatial}=\mathrm{ReLU}(Z_V^{pre}W_V)\in\mathbb{R}^{FP\times d} \qquad \text{Equation 1}$$

$W_V \in \mathbb{R}^{d_{emb}\times d}$

これは、入力ビデオの空間レベルの特徴として示される。

モダリティレベル符号化層 212 は、ビデオデータ V202 に対してモダリティ符号化を実行する。モダリティレベルの符合化は、ビデオデータ V202 である情報のタイプを通知する。 例えば、モダリティレベル符号化レイヤ 212 は、モダリティトークン vis を使用し、そのビデオデータ V202 を均一に表す。モダリティトークン vis は、情報タイプが視覚的であることを示す。

時間レベル符号化層 214 は、入力ビデオデータ V202 に対して時間符号化を実行する。時間符号化は、ビデオデータ V 202 内の入力特徴のフレームレベル（またはセグメントレベル）の位置に関連する。したがって、ビデオデータ V202 内の各フレームは、異なる時間符号化を有することができるが、各フレーム内のセグメントは、同じ時間符号化を有することができる。

位置レベル符号化層 216 は、ビデオデータ V202 に対して位置符号化を実行する。位置レベルの符合化には、フレームと各フレーム内のセグメントの空間レベルの順序付けが組み込まれている。したがって、各フレーム内およびフレーム間の各空間領域は、異なる位置レベルエンコーディングを持つことになる。入力ビデオデータ V202 の位置符号化は、BERT ベースの言語モデルに見られる文中のトークンの位置符号化と同等である。

モダリティレベル符号化層 212、時間レベル符号化層 214、および位置レベル符号化層 216 は、モデル 200 がビデオデータ V202 内の入力特徴のダイナミクスを学習できるようにするためのトレーニング可能なパラメータである。モダリティレベル符号化層 212、時間レベル符号化層 214、および位置レベル符号化層 216 は、事前訓練されたモデルと同じ特徴次元 d を有するようにモデル化される。符号化層 210～216 は、以下に示す符号化されたビデオ表現 Z_V である要素単位の合計によって結合される。

$Z_V = Z_V^{spatial} + Z_V^{mod} + Z_V^{temporal} + Z_V^{pos}$ Equation 2

さらに、モダリティレベル符号化層 212、時間レベル符号化層 214、および位置レベル符号化層 216 からの符号化の一部またはすべてが、符号化されたビデオ表現 Z_V に含まれる。ビデオに基づく対話モデル 200 はまた、1 つまたは複数のエンコーディング層を使用してテキストデータ T204 に対してトークンレベルのエンコーディングを実行することによって、エンコードされたテキスト表現 Z_T を生成する。符号化層は、単語レベル符号化層 220、モダリティレベル符号化層 222、ターンレベル符号化層 224、および位置レベル符号化層 226 である。

単語レベル符号化層 220 は、対話履歴 H206、応答 S208、およびキャプション C209 を入力として受け取る。単語レベル符号化層 220 は、一連の対話ターン $H=(H_1, H_2, \ldots, H_t)$として対話履歴 H206 を分解する。t は現在の対話ターンである。各対話ターンは、$H=((U_1, S_1), (U_2, S_2), \ldots, U_{t-1}))$を順次連結したユーザ発話Uとシステム応答 S208 との対として表される。S_t は、現在の人間の発話に応答してモデル 200 によって生成されるターゲット応答である。

目標応答 S_t は出力 150 である。次に、各発話は一連のトークン（単語）x として表され、対話履歴は $X_H=(x_1, x_2, \ldots x_{L\,H})$および $S_t(y_1, y_2, \ldots y_{L\,H})$として表され、$L_H$ と L_Y は、それぞれ対話履歴 H 206 とターゲット応答内のトークンの総数である。
ビデオキャプション C 209 は、別のテキスト入力である。ビデオキャプション C209 は通常、ビデオの言語的な要約を 1 つまたは 2 つの文で提供する。ビデオキャプション C 209 は、一連のトークン $X_C=(x_1, x_2, \ldots x_{L\,C})$として表すことができる。

テキストデータ 204T のすべてのテキスト入力シーケンスが組み合わされて、モデル 200 への入力として単一のシーケンス $X_T=(X_C, X_H, Y_{-1})$を形成する。Y_{-1} は、シフトされたターゲット応答である。単語レベル符号化層 220 の出力は、X_T の埋め込み特徴であり、入力テキストデータ T204 のトークンレベル符号化層である Z_T^{token} として表すことができる。

ビデオ機能と同様に、モダリティレベル符号化層 222、ターンレベル符号化層 224、および位置レベル符号化層 226 がビデオに基づく対話モデル 200 に追加され、X_T のさまざまな属性が注入される。

モダリティレベル符号化層 222 は、モダリティレベル符号化を実行する。モダリティ符号化は、X_T のセグメントを区別する。モダリティレベル符号化層 222 は、モダリテ

ィトークン「cap」、「sys」、および「usr」などのさまざまなモダリティトークンを使用して、テストデータ T 204 内の対応する位置にあるトークンがビデオキャプション C 209、システム応答 S208、またはユーザの発話 U の一部であるかどうかを指定する。ターンレベル符号化層 224 は、ターンレベル符号化を実行する。ターンレベル符号化は、トークンのターン数を対応する位置に符合化する。

例えば、図の例では、対話履歴 206 および応答 208 は 5 に設定され、これは、ユーザ発話Uおよびシステム応答 S208 の 5 つのペアがあり、現在のターンが 5 であることを示す。したがって、ターンレベル符号化層 224 は、発話Uおよびシステム応答 S208 に含まれるトークンを順番に 1 から 5 に設定する。

ビデオキャプション C 209 セグメントでは、ターンレベル符号化がゼロに設定される。位置レベル符号化層 226 は、トークン順序付けに関する属性を注入する位置レベル符号化を実行する。図に示すように、トークンの順序付けは、空間的順序付けからの次のトークン番号で継続する。ビデオ表現と同様に、エンコードされたテキストデータ T204 は、以下に示すエンコードされたテキスト表現 Z_T である要素ごとの合計によって結合される。

$$Z_T = Z_T^{token} + Z_T^{mod} + Z_T^{turn} + Z_V^{pos} \quad \text{Equation 3}$$

さらに、モダリティレベル符号化層 222、ターンレベル符号化層 224、および位置レベル符号化層 226 からの符号化の一部またはすべてが、符号化されたビデオ表現 Z_T に含まれる。ビデオに基づく対話モデル 200 は、エンコードされたビデオ表現 Z_V とエンコードされたテキスト表現 Z_T を単一の入力シーケンス Z_{VT} に連結する。入力シーケンス Z_{VT} の長さは、埋め込み次元 d を有する$(F\times P+L_C+L_H+L_Y)$である。単一の入力シーケンス Z_{VT} は、GPT-2 などの GPT モデル 250 をファインチューニングするために、事前訓練された GPT-2 への入力として使用される。

3.クレーム

999 特許のクレーム 1 は以下の通りである。

1. ビデオに基づく対話において、ビデオに基づく対話ニューラルネットワーク言語モデルによって応答を生成するための方法であって、

ビデオに基づく対話ニューラルネットワーク言語モデルで、ビデオ入力およびテキスト入力を受信し、ここで、テキスト入力は、ビデオに基づく対話ニューラルネットワーク言語モデルと人間のユーザとの間の対話履歴、および人間のユーザによる現在の発話を含み;

ビデオに基づいた対話ニューラルネットワーク言語モデルの複数のビデオ符号化層を使用して、エンコードされたビデオ入力を生成し、ここでエンコードされたビデオ入力は、空間レベルのエンコードと、モダリティレベルのエンコード、時間レベルのエンコード、または位置レベルのエンコードのうちの少なくとも 1 つとを含み、；

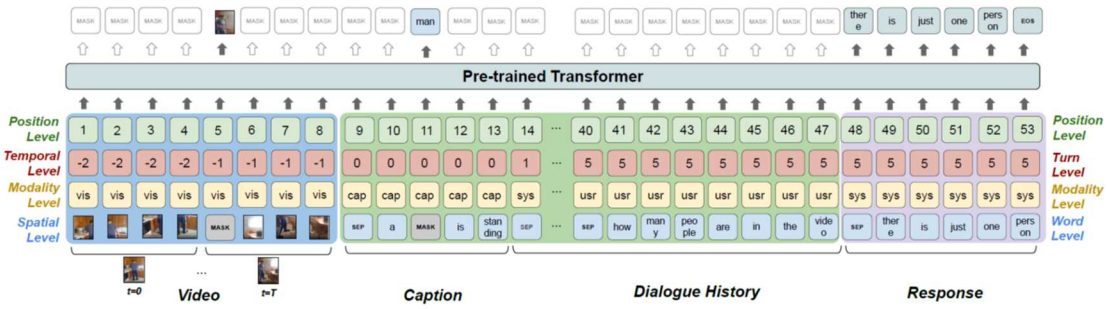
ビデオに基づく対話ニューラルネットワーク言語モデルの複数のテキストエンコーディング層を使用して、エンコードされたテキスト入力を生成し、エンコードされたテキスト入力は、トークンレベルのエンコードと、モダリティレベルのエンコード、ターンレベルのエンコード、または位置レベルのエンコードのうちの少なくとも 1 つとを含み、；

エンコードされたビデオ入力とエンコードされたテキスト入力を単一の入力シーケンスに連結し、；

単一の入力シーケンスから生成し、ビデオに基づいた対話ニューラルネットワーク言語モデルで生成済みの事前トレーニング済みトランスフォーマモデルを使用して、人間のユーザの現在の発話に対する応答を生成する。

4. 本特許に関連する論文

本特許に関する論文“Video-Grounded Dialogues with Pretrained Generation Language Models”[1]が、Salesforce の Hung Ley 氏らにより公表されている。



ビデオに基づく対話用 VGD-GPT2 アーキテクチャを上記図に示す。ビデオとテキストの入力は複数の符合化層で結合され、符合化された機能にさまざまな属性が挿入される。図の例では男性が室内を移動する動画像と How many people are in the video?の問いかけに対し、There is just one person.の応答が出力されている。

[1] Hung Ley, Steven C.H. Hoi “Video-Grounded Dialogues with Pretrained Generation Language Models” arXiv:2006.15319v1 [cs.CL] 27 Jun 2020

Model	Spatial	Temporal	MLM	MVM	MVT	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	CIDEr
Baseline		✓				0.626	0.485	0.383	0.309	0.215	0.487	0.746
AVSD Winner		✓				0.718	0.584	0.478	0.394	0.267	0.563	1.094
MTN		✓				0.731	0.597	0.490	0.406	0.271	0.564	1.127
VGD-GPT2 (S)	✓	✓	✓	✓	✓	0.750	**0.621**	0.516	0.433	**0.283**	0.581	**1.196**
VGD-GPT2 (S)	✓		✓	✓	✓	**0.753**	0.619	0.512	0.424	0.280	0.571	1.185
VGD-GPT2 (S)		✓	✓	✓	✓	0.750	0.616	0.511	0.427	0.280	0.579	1.188
VGD-GPT2 (S)	✓	✓	✓	✓		0.745	0.613	0.508	0.423	0.281	0.579	1.173
VGD-GPT2 (S)	✓	✓	✓		✓	0.749	0.613	0.505	0.419	0.274	0.571	1.153
VGD-GPT2 (S)	✓	✓		✓	✓	0.744	0.612	0.505	0.421	0.281	0.581	1.192
VGD-GPT2 (M)	✓	✓	✓	✓	✓	0.749	0.620	**0.520**	**0.436**	0.282	**0.582**	1.194

上記テーブルは、Visual Scene Dialog (AVSD)モデルおよび MTN モデルなどの従来のモデルに対する、ビデオに基づく対話モデルの改善を示す。改善は BLUE1, BLUE2, BLUE3, BLUE4, METEOR, ROUGE-L 及び CIDEr 等さまざまなデータセットで示されている。

テーブルでは、小規模または中規模の事前訓練済み GPT-2 モデルを使用するビデオに基づく対話モデル（VGD-GPT2）が、ベースライン、AVSD および MTN モデルよりも改善されていることを示している。また、GPT-2 のサイズを小から中に大きくすると、わずかに改善されることが理解できる。テーブルは、ビデオに基づいた対話モデルをファインチューニングすることでもパフォーマンスが向上することも示している。

たとえば、マルチタスクの目的で事前トレーニング済みのモデルをファインチューニングすると、応答生成のメインタスクにメリットがあることを示している。これらの補助的な目標は、事前トレーニング済みのモデルを現在のデータドメインであるビデオベースの対話に適応させるのに役立つ。特に、MLM（masked language modeling）および MVM(masked visual modeling)損失関数は、トークンおよび空間レベルでのローカル依存関係の学習を改善するために使用され、MVT(Matching Video-Text Pair)損失関数は、テキストと視覚モダリティ間のグローバル依存関係の学習を改善するために使用される。

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。