AI 特許紹介(60)
AI 特許を学ぶ！究める！
～プロンプトアダプタ特許～

2024 年 1 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 Inc
出願日　2021 年 1 月 28 日
公開日　2023 年 3 月 16 日
公開番号 US20230083512
発明の名称　言語モデルから事実を抽出するシステムと方法

512 特許は、LLM（大規模言語モデル）の埋め込み層と、最初のアテンション層との間に設けられ、ユーザ間のプロンプトの相違に関わらず、事実情報を適切に抽出するプロンプトアダプタ（P アダプタ）技術に関する。

2.特許内容の説明
最近の研究（例: LAMA (Petroni et al., 2019)）では、大規模言語モデル（LLM）から抽出された事実情報の品質が、クエリに使用されるプロンプトに依存することが判明している。異なるユーザが異なる表現を使用して同じ情報を LLM に問い合わせるが、それに関係なく同じ正確な応答を受け取る必要があるため、この不一致が問題となる。

512 特許は、LLM の埋め込み層と第 1 のアテンション層との間に位置する軽量モデルである P アダプタを導入することで、この欠点に対処する。図 1 は、言語モデルから事実情報を抽出するための例示的なアーキテクチャを示す簡略図である。

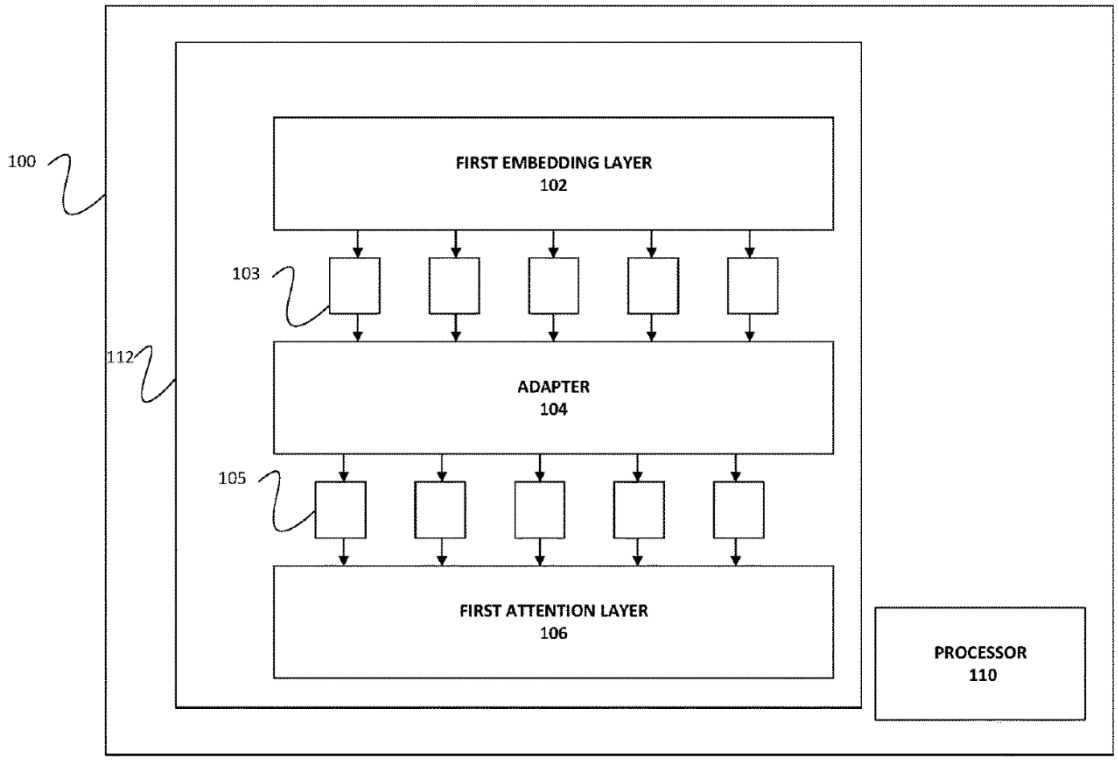


図 1 に示すように、システム 100 は、プロセッサ 110 およびメモリ 112 を含む。メモリ 112 は、事前訓練された言語モデルを格納する。事前トレーニングされた言語モデルは、第 1 の埋め込み層 102、第 1 のアテンション層 106、および第 1 の埋め込み層 102 と第 1 のアテンション層 106 との間に配置されたアダプタモデル 104 を備える。アダプタモデル 104 は、事前トレーニングされた言語モデルの第 1 の埋め込み層 102 からの第 1 の埋め込み 103 を入力として受け取り、連続プロンプトなどの第 2 の埋め込み 105 を出力する。

システム 100 は、通信インターフェースを介して、事実情報に対するクエリを受信する。システム 100 は、事前訓練された言語モデルを介して、自然言語プロンプトを第 1 の埋め込み 103 にエンコードする。例えば、システム 100 が通信インターフェース 108 を介して「アメリカの首都は」というクエリを受信すると仮定する。システム 100 は、事前訓練された言語モデルの埋め込み層 102 を介して、このクエリに対する第 1 の埋め込み 103 を決定する。

第１の埋め込み 103 は、n 次元空間における正規化後のクエリのベクトル表現である。アダプタモデル 104 は、埋め込み層 102 から第１の埋め込み 103 を受信する。アダプタモデル 104 は、第１の埋め込み 103 を第２の埋め込み 105 にエンコードする。第２の埋め込み 105 は連続表現である。連続表現は、第１の埋め込み 105 の代わりに、事前トレーニングされた言語モデルの第１のアテンション層 106 への入力として使用される。第２の埋め込み 105 は、第２の埋め込みが事前トレーニング済み言語モデルの第１のアテンション層に供給されるときに、第２の埋め込みが事実情報を一貫して、正確に、返す確率に基づく。

システム 100 は、事前訓練された言語モデルの第１のアテンション層 106 を介して、第２の埋め込み 105 をクエリに対する応答にデコードする。システム 100 は、クエリに対するデコードされた応答から事実情報を抽出する。

システム 100 は、通信インターフェースを介して、事実情報に対する同様の自然言語クエリの複数のセットを含むトレーニングデータセットを受信する。システム 100 は、類似の自然言語クエリの複数のセットから、類似の自然言語クエリのセットにおける各クエリの第１のエンティティおよび第２のエンティティを決定する。システム 100 は、類似の自然言語クエリのセット内の各クエリ内の少なくとも第１のエンティティまたは第２のエンティティをマスクする。

システム 100 は、類似の自然言語クエリのセット内のマスクされたクエリごとにマスクされた埋め込みを決定する。システム 100 は、事前トレーニングされた言語モデルの第１のアテンション層を介して、事前トレーニングされた言語モデルを介して、各マスクされた埋め込みに対する応答を決定する。システム 100 は、応答が自然言語クエリ内のマスクされたエンティティと一致するかどうかを判定する。

システム 100 は、マスクされた埋め込みが自然言語クエリ内のマスクされたエンティティと一致する応答を返す確率を決定する。システム 100 は、所定の言語モデルからの応答が、決定された確率に基づいてマスクされたエンティティと一致するクエリを選択するアダプタモデル 104 を更新する。

システム 100 は、式 : $e(x')$によって記述されるＰアダプタモデルなどのアダプタモデル 104 を訓練する。たとえば、P アダプタモデルは、自然言語プロンプトの埋め込み $e(x')$ を入力として受け取り、連続埋め込みの新しいシーケンス x'_{cont} を出力する。

トレーニング後、P アダプタモデルは、自然言語クエリ $e(x')$に対応する第 1 の埋め込みを入力として受け取り、連続埋め込み x'_{cont} を出力する。システム 100 は、第 1 のアテンション層 106 への入力として第 1 の埋め込み $e(x')$の代わりに置換される連続埋め込み x'_{cont} を決定する。

アダプタモデル 104 は、関数 $f_{P\text{-}Adapter}: e(x') \rightarrow x'_{cont}$ によって記述される。すなわち、P アダプタモデルは、第 1 の埋め込みを受信したときに連続埋め込みを返す。システム 100 は、確率 $P_{LM}(y|x'_{cont})$を最大化するようにアダプタモデル 104 を訓練する。

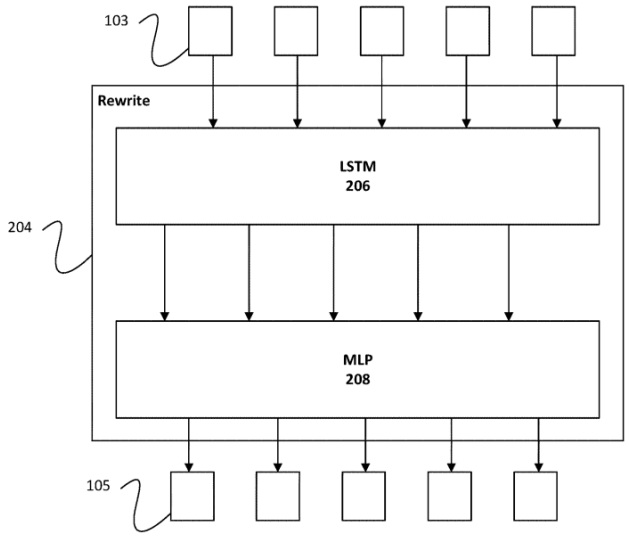
言い換えれば、アダプタモデル 104 は、連続埋め込み x'_{cont} が第 2 の埋め込み層 106 を介して事前トレーニングされた言語モデルへの入力として使用されるときに、事前トレーニングされた言語モデルが事実情報 y を返す確率を最大化するようにトレーニングされる。

システム 100 は、アダプタモデル 104 を介して、次の式によって記述される事前訓練された言語モデルからの予測を改善することができる。

$$\arg\max_{v \in \mathcal{V}} P_{LM}(v|f_{prompt}(e(x'))),$$

$\mathcal{V}$ は、事前トレーニングされた言語モデルの語彙である。

図 2 は、言語モデルから事実情報を抽出するための書き換え p アダプタモデルの簡略図である。



システム 100 は、長期短期記憶ネットワーク（LSTM）206 および一組の多層パーセプトロン（MLP）208 を含むアダプタモデル 204 のための書き換えＰアダプタ 204 を訓練する。書き換えＰアダプタモデル 204 は、LSTM206 と、LSTM206 の各出力に接続された MLP208 とを含み、LSTM206 の各出力に接続された MLP208 が、LSTM206 の出力からの対応する隠れ状態を処理する。

3.クレーム

512 特許のクレーム 1 は以下の通りである。

1. 言語モデルから事実情報を抽出する方法において、

通信インターフェースを介して事実情報のクエリを受信し、

事前にトレーニングされた言語モデルの埋め込み層を介して、自然言語プロンプトを第 1 の埋め込みにエンコードし、

アダプタモデルを介して、第 2 の埋め込みが事前トレーニング済み言語モデルの第 1 アテンション層に供給されたときに第 2 の埋め込みが事実情報を返す確率に基づいて、第 1 の埋め込みを、連続表現を含む第 2 の埋め込みにエンコードし、アダプタモデルは、事前トレーニングされた言語モデルの埋め込み層と、事前トレーニングされた言語モデルの第 1 のアテンション層との間に配置され、

事前トレーニングされた言語モデルの第 1 アテンション層を介して、第 2 の埋め込みを、クエリに対する応答にデコードし、

クエリに対するデコードされた応答から事実情報を抽出する。

4. 本特許に関連する論文

本特許に関する論文“P-ADAPTERS: ROBUSTLY EXTRACTING FACTUAL INFORMATION FROM LANGUAGE MODELS WITH DIVERSE PROMPTS”[1]が、Stanford 大学の Benjamin Newman 氏らにより公表されている。

LLM は事前トレーニング中に事実の知識を蓄積し、この知識を維持するためにパラメータをフリーズし、その後、手作りの自然言語プロンプトを使用してユーザによってクエリされる。たとえば、あるユーザがカナダの首都がどこであるかを知りたい場合、“The capital of Canada is [MASK]”というプロンプトを使用してフリーズモデルにクエリを実行するが、別のユーザは別の表現“Canada, which has the capital city [MASK]”

[1] Benjamin Newman et al. “P-ADAPTERS: ROBUSTLY EXTRACTING FACTUAL INFORMATION FROM LANGUAGE MODELS WITH DIVERSE PROMPTS” ICLR 2022

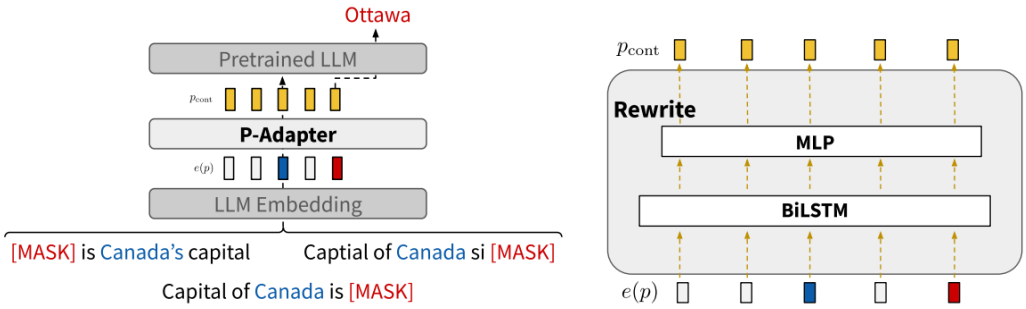
を使用する可能性がある。

LLM が効果的な知識ベースであるためには、ユーザが提供するさまざまなクエリに対して堅牢である必要がある。残念ながら、これまでの研究では、LLM は堅牢ではないことが示されている。意味的に同等のクエリは、一貫性のない予測につながる可能性がある。たとえば、上記のプロンプトの場合、最初のプロンプトは BERT Base から正しい答え「オタワ」を抽出するが、2 番目のプロンプトは間違った「ウィニペグ」を抽出する。

このことから、特定の関係に対する最適なプロンプトまたはプロンプトのセット、つまりモデルが事実情報を最もよく抽出できるプロンプトを見つけようとする多くの研究が行われてきた。本論文では、単一の最適なプロンプトを見つけることに重点を置くのではなく、自然言語プロンプトを連続表現に適応させ、LLM が事実情報を正確に予測できるようにすることで、LLM がこのばらつきを克服できるように支援することを目指している。

ユーザ重視の設定を動機としているため、適応方法では推論時に自然言語プロンプト（例: “The capital of Canada is [MASK]”）のみを必要とする。プロンプト内のエンティティペアの主語（Canada）と目的語（Ottawa）の間の関係（capital of）に対する追加のアノテーションは不要である。また、プロンプトとは別に主題の身元を知る必要もない。トレーニング時には、理想的な方法は（prompt, object）ペア（例: (“The capital of Canada is [MASK]”, “Ottawa”）のみに依存するため、新しいデータの収集に追加のアノテーションは必要ない。

本論文では、P アダプタモデルのクラスを導入する。P アダプタは、フリーズされた LLM の埋め込み層と最初のアテンション層の間に位置し、事実情報をより効果的に予測できるように LLM の埋め込みを変更する。これらはエンドツーエンドで最適化されており、トレーニング時に（プロンプト、オブジェクト）ペアのみを必要とし、変数トレーニングプロンプトを同じオブジェクトにマッピングすることを学習することで一貫性を促進する。P アダプタフレームワーク及び書換 P アダプタのネットワーク構成図を以下に示す。



P アダプタは、LLM の埋め込み層と最初のアテンション層の間に位置する。これらは LLM 埋め込みを入力として受け取り、LLM に供給される連続プロンプトを出力する。P アダプタのパラメータがトレーニングされている間、LLM はフリーズされる。アダプタは、プロンプト内のさまざまな語句間のばらつきや入力ミスを軽減するのに役立つ。

Subject は青、[MASK] 埋め込みは赤、P アダプタによって生成されたエンベディングを黄色で示す。その他の未変更の埋め込みは灰色である。 点線の矢印はモデルコンポーネントへの入力と出力を表し、実線の矢印は P アダプタの入力から出力へのコピーを表す。

Metric	P-Adapter	ID	OOD Prompts	OOD Objects	OOD KE
P@1	Baseline	0.157	0.157	0.069	0.092
	Rewrite	0.258 ± 0.00	0.247 ± 0.00	0.078 ± 0.00	0.167 ± 0.00
	Prefix	0.425 ± 0.01	0.415 ± 0.01	0.193 ± 0.00	0.326 ± 0.01
	P-Tuning	0.442 ± 0.00	0.422 ± 0.00	0.203 ± 0.00	0.325 ± 0.00
	MoE	0.488 ± 0.01	0.418 ± 0.00	0.237 ± 0.02	0.331 ± 0.00
	Oracle	0.496 ± 0.01	0.496 ± 0.01	0.238 ± 0.02	0.496 ± 0.01
Consistency	Baseline	0.126	0.133	0.097	0.068
	Rewrite	0.476 ± 0.01	0.448 ± 0.02	0.456 ± 0.01	0.223 ± 0.01
	Prefix	0.656 ± 0.02	0.613 ± 0.02	0.588 ± 0.03	0.452 ± 0.03
	P-Tuning	0.730 ± 0.00	0.646 ± 0.01	0.656 ± 0.01	0.476 ± 0.01
	MoE	0.947 ± 0.03	0.658 ± 0.03	0.916 ± 0.05	0.439 ± 0.01
	Oracle	1.000 ± 0.00	1.000 ± 0.00	1.000 ± 0.00	1.000 ± 0.00

上記表に示すように、P アダプタは、自然言語クエリのみを使用したベースラインと比較して、精度で 12 ～ 26% の絶対的な向上と一貫性の絶対的な向上が 36 ～ 50% 向上している。

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。