AI 特許紹介(80)

AI 特許を学ぶ！究める！

～プロンプトチューニング特許～

2025 年 9 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　Google

出願日　2022 年 4 月 12 日

公開日　2023 年 10 月 12 日

公開番号 US20230325725

発明の名称　大規模で効率的なモデルのためのパラメータ効率的なプロンプトチューニング

725 特許は、調整可能なトークンを入力テキストの先頭に追加するソフトプロンプトを LLM に入力することにより、few-shot プロンプトよりも優れた性能を発揮し、かつ、モデルチューニングに対して品質の差を埋めることが可能なプロンプトチューニング技術に関する。

2.特許内容の説明

下記図はモデルチューニングとプロンプトチューニングとの対比を示す説明図である。

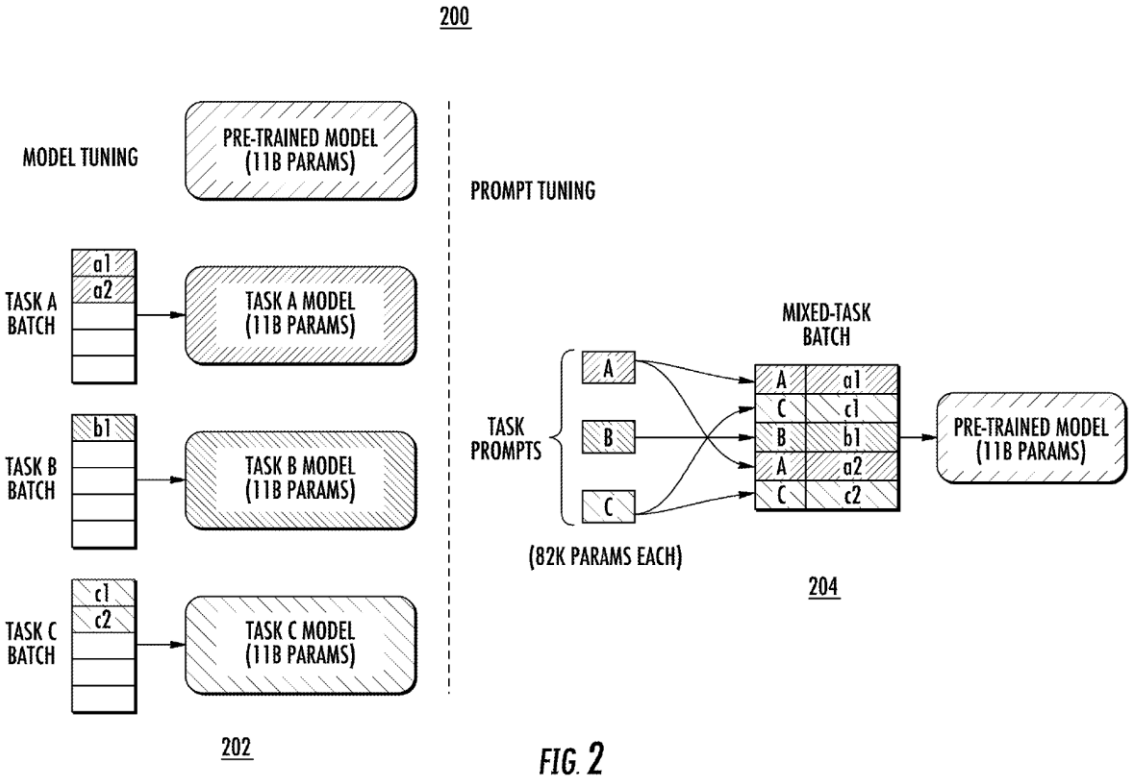


FIG. 2

図示されているように、モデルチューニング 202 は、機械学習モデルをタスクごとに再学習する。現在の言語モデルには数十億ものパラメータが含まれるため、学習プロセスおよび再学習プロセスは、時間と計算負荷の両方を消費する。このような処理は、計算リソースが限られたデバイスでは実行できない。加えて、異なるモデルの保存およびアクセスは、他の潜在的な問題を引き起こす。

プロンプトチューニング 204 は、事前学習済み機械学習モデルのパラメータを再学習させる代わりに、フリーズする。プロンプトチューニング 204 は、各特定のタスクについて少数のパラメータセット（例えば、タスクあたり 82,000 個）を学習し、これらのパラメータセットは入力データと共に事前学習済み機械学習モデルに入力され、その特定のタスクに対して事前学習済み機械学習モデルを準備する。したがって、プロンプトチューニング 204 では、学習と調整の対象となるパラメータが大幅に削減され、大規模な事前学習済み機械学習モデルを再学習することなく、様々なタスクに活用できるようになる。

モデルチューニング 202 は、各下流タスクについて、事前学習済みモデル全体のタスク固有のコピーを作成することに依存し、推論は別々のバッチで実行される。プロンプトチューニング 204 は、各タスクについて小さなタスク固有のプロンプトを保存することのみに依存し、元の事前学習済みモデルを使用して混合タスク推論を可能にする。

T5「XXL」モデルを使用すると、チューニングされたモデルの追加コピーごとに 110 億個のパラメータに依存する。対照的に、チューニングされたプロンプトは、プロンプト長が 20 トークン、埋め込み次元が 4,096 の場合、タスクあたり 81,920 個のパラメータのみに依存し、これは 5 桁以上の削減となる。

3.クレーム

725 特許のクレーム 1 は以下の通りである。

1. プロンプトチューニングのためのコンピューティングシステムにおいて、

１つ以上のプロセッサと、

前記１つ以上のプロセッサによって実行されると、前記コンピューティングシステムに演算を実行させる命令を集合的に記憶する１つ以上の非一時的なコンピュータ可読媒体とを備え、前記演算は以下を含み、

複数のトレーニング例と、それぞれのトレーニング例に対する複数のトレーニングラベルとを含むトレーニングデータセットを取得し、

事前学習済み機械学習モデルを用いて、トレーニング出力を生成するために、前記複数のトレーニング例のうちの１つ以上のトレーニング例およびプロンプトを処理し、事前学習済み機械学習モデルの複数の事前学習済みパラメータはプロンプトチューニング中に固定され、前記プロンプトは特定のタスクに関連付けられ、前記特定のタスクは前記１つ以上のトレーニング例に関連付けられ、

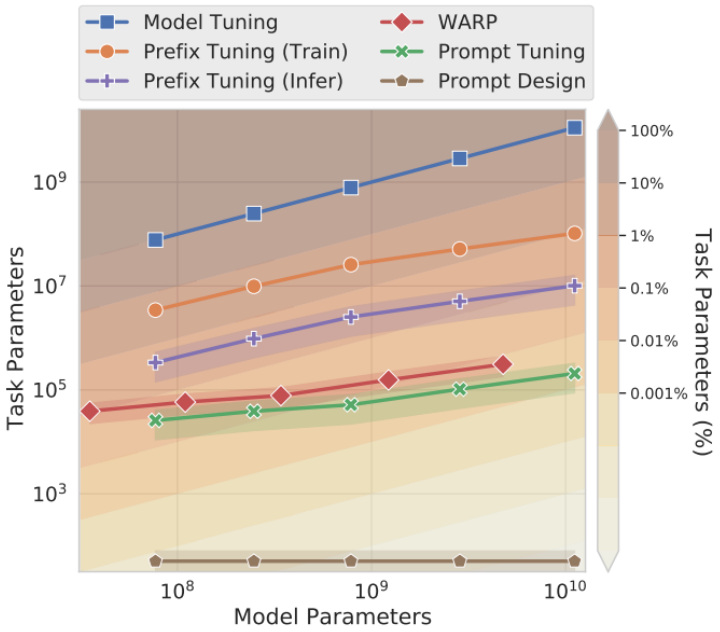
前記トレーニング出力と、前記１つ以上のトレーニング例に関連付けられた１つ以上のトレーニングラベルとの比較に少なくとも部分的に基づいて、プロンプト勾配を決定し、

前記プロンプト勾配に基づいて、前記プロンプトの１つ以上のプロンプトパラメータを調整する。

4. 本特許に関連する論文

本特許に関する論文“The Power of Scale for Parameter-Efficient Prompt Tuning”[1]が、Google の Brian Lester 氏らにより公表されている。論文には下記図に示すように、各手法とのパラメータの比較が示されている。

[1] Brian Lester, et al. “The Power of Scale for Parameter-Efficient Prompt Tuning” arXiv:2104.08691v2 [cs.CL] 2 Sep 2021



図には、複数の手法の結果が含まれており、各手法は、図に示されているタスクパラメータ（左縦軸）を、全体のモデルパラメータ（横軸）との関係で用いてトレーニングされる。さらに、使用されたモデルパラメータに対するタスクパラメータの相対的な割合を示すための二次スケール（右縦軸：タスクパラメータの割合）が示されている。図に示すように、プロンプトチューニングは、モデルサイズに関わらず、他の機械学習による代替手法よりも少ないパラメータを利用できる。

Model Tuning:すべてのパラメータはタスク固有である。

Prefix Tuning:活性化は各層のプレフィックスでチューニングされ、推論には 0.1～1%のタスク固有パラメータが必要であるが、学習にはより多くのパラメータが使用される。

WARP：入力層と出力層のみをチューニングすることで、タスクパラメータは 0.1%未満に削減される。

Prompt Tuning:プロンプト埋め込みのみをチューニングし、ほとんどのモデルサイズで 0.01%未満になる。

Prompt Design:プロンプト ID のシーケンス（500～2000 トークン）のみが必要である。

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、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 3」、「ブロックチェーン 3.0(共著)」がある。