AI 特許紹介(67)
AI 特許を学ぶ！究める！
～LoRA 特許～

2024 年 8 月 9 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　Microsoft
出願日　2021 年 5 月 19 日
公開日　2022 年 12 月 1 日
公開番号 US20220383126
発明の名称　ニューラルネットワークモデルの低ランク適応

126 特許は、事前トレーニング済みのモデルの重みを固定し、トレーニング可能なランク分解マトリックスを Transformer アーキテクチャの各層に挿入して、下流のタスクのトレーニング可能なパラメータの数を大幅に削減する Low-Rank Adaptation (LoRA) 技術に関する。

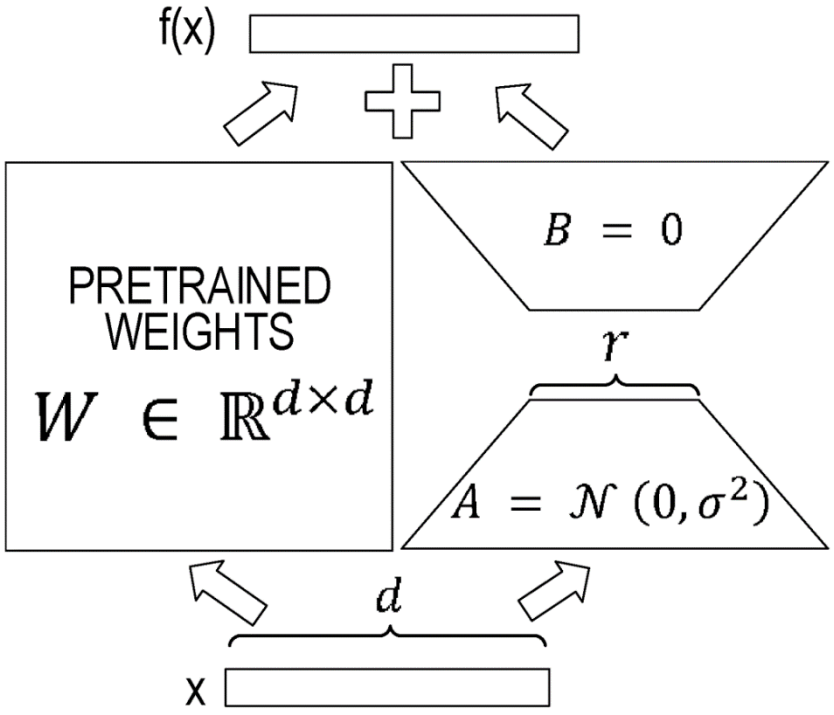
2.特許内容の説明
大規模なモデルでは、モデルサイズが大きくなるにつれてタスクのパフォーマンスが向上し続けることが実証されている。ただし、特定のタスクまたはドメインの一般モデルを完全にファインチューニングするには、ファインチューニングされたモデルに元の

一般モデルと同じ数のパラメータを保存する必要がある。事前トレーニング済みのモデルが大きくなると、さまざまなタスク固有のモデルを保存し、実稼働環境でそれらのモデルを切り替えることが難しくなる。

タスク間で高価な処理リソースを共有し、タスク固有のモデルを切り替えるには、毎回非常に大きなチェックポイントを VRAM にロードする必要がある。このような切り替えは、時間がかかり、リソースを大量に消費する。つまり、従来のファインチューニングでは、膨大な事前トレーニング済みモデルを扱う場合、拡張できない。

そこで、事前トレーニング済みのモデルの重みを固定し、トレーニング可能なランク分解マトリックスを Transformer アーキテクチャの各層に挿入することにより、下流のタスクのトレーニング可能なパラメータの数を大幅に削減する Low-Rank Adaptation (LoRA) を導入する。

図 1 は、入力ベクトル x を関数 f(x)で処理することを示すブロック図である。



言語モデルは、Transformer ベースの深層学習言語モデルである。事前トレーニングされた重みは、ネットワーク全体が一般的なドメインデータでトレーニングされた結果生じる d×d の次元を持つ行列の形式である。入力ベクトル x は、単語またはその他の言語コンポーネントを表すトークンであり、次元も d である。

入力ベクトルは、ランク分解行列のペアである行列 A と行列 B によっても処理される。行列 A は長さ d の入力ベクトル x を受け取り、それを長さ r のベクトルに変換する。マトリックス B は長さ r のベクトルを受け取り、それを長さ d のベクトルに変換し、事前トレーニングされた重みマトリックスの結果と組み合わせて、ニューラルネットワークの次の層への入力である f(x)を提供する。

マトリックス A と B は、一般的なモデルを特定のタスクまたはドメインに適応させるため、適応マトリックスと呼ばれる。LoRA では、事前トレーニング済みの重みの元の行列を変更せずに、ランク分解行列 A と B を挿入して最適化することで、ニューラルネットワーク内の複数の密な層のそれぞれを間接的にトレーニングできる。実際には、フルランクが高くても非常に低いランクで十分であり、LoRA はスペース効率と計算効率の両方に優れている。

LoRA にはいくつかの重要な利点がある。1 つの事前トレーニング済みモデルを共有し、さまざまなタスクに合わせて多数の小さな適応を構築するために使用できる。これにより、トレーニング中に勾配を計算したり、膨大な元のモデルの最適化状態を維持したりする必要がないため、トレーニングがより効率的になる。共有された元のモデルは、VRAM（揮発性ランダムアクセスメモリ）またはその他の選択されたメモリに保存され、積み重ねられたマトリックス A と B で構成される大幅に小さい LoRA モデルを効率的に切り替えることで、プロセッサの使用率が大幅に向上する。

完全なファインチューニングとは異なり、適応マトリックスを使用すると、適応されたモデルがバイパスされて元のモデルに戻るため、一般的なドメインに対する元のモデルの機能が損なわれることはない。適応マトリックスを使用すると、展開中に更新マトリックスを元の重みと組み合わせることができるため、推論の遅延は発生しない。

大規模な事前トレーニング済みモデルを特定のタスクに適応させる作業は、適応マトリックスのごく少数のパラメータを最適化しながら実行できる。従来のファインチューニングと比較して、これによりトレーニングのハードウェア障壁が下がり、推論の遅延を追加することなく、サービスコストが大幅に削減される。

入力トークンの長さ、つまり重みマトリックスの幅は d=10,000 である。重み行列のトレーニング可能なパラメータの数は以下のとおりである。

|W|=d^2=100,000,000.

ランク r は d よりもはるかに小さくなる。

適応マトリックスの数は、|A|+|B|=d*r+r*d=2*10,000*8=160,000 個のトレーニング可能なパラメータである。現在普及している言語モデルでは、r の一般的な値は 1 より大きく 100 未満の範囲である。将来的には、より大きなニューラルネットワークモデルで、より大きな r が使用される可能性がある。ランク r は、実際には経験的に決定される。

適応マトリックスの使用によって解決される技術的な問題を説明するために、一般的なドメインモデルを適応させる一般的な問題について説明する。事前トレーニング済みの大規模言語モデルを、要約、機械読解（MRC：machine reading comprehension）、自然言語から SQL（NL2SQL）などの条件付きテキスト生成タスクに適応することを検討すると、ここで、トレーニングインスタンスはコンテキストとターゲットのペアである：$\{(x_i,y_i)\}$ i=1, . . . , N; x_i と y_i はどちらもトークンのシーケンスである。たとえば、x_i は自然言語クエリであり、 y_i は自然言語を構造化シーケンス言語クエリ（NL2SQL と呼ばれる）に変換するタスク内の SQL である。

古典的な適応フレームワークでは、モデルは事前トレーニングされたパラメータ Φ_0 で初期化され、条件付き言語モデリングの目的を最大化することで Φ'にファインチューニングされる。

$$\Phi' = \text{argmax}_{\Phi} \sum_{i=1}^{N} \sum_{t=1}^{|y_i|} \log p_{\Phi}(y_{i,t} \mid x_i, y_i, < t) \quad (1)$$

N は例の数であり、式(1)は入力と既知の出力が与えられた場合に正しいトークンを生成するように動作する。

従来のファインチューニングアプローチでは、パラメータ空間全体を更新するため、計算とメモリの面で非効率的である。そこで、効率的な重み保存モデル適応アプローチを提案する。元の事前トレーニング済みモデルパラメータ Φ_0 は保持され、完全なモデルのファインチューニングと比較してパフォーマンスを低下させることなく、タスク指定の小さなサイズのパラメータセット Θ, $|\Theta| >> \Phi_0$ を追加で学習する。

$$\ominus' = \text{argmax}_{\ominus} \sum_{i=1}^{N} \sum_{t=1}^{|y_i|} \log p_{(\ominus,\Phi_G)}(y_{i,t} \mid x_i, y_i, < t) \quad (2)$$

典型的なニューラルネットワークには、行列乗算を実行する多数の密な層が含まれている。これらの層の重み行列は、通常、フルランクを持つことが許可される。ただし、事前トレーニング済みモデルのその後の更新はランク不足になる傾向があり、低ランク

の再パラメータ化にもかかわらず、効率的に学習することができる。

3.クレーム

126 特許のクレーム 1 は以下の通りである。

1. コンピュータ実装方法において、

複数のニューラルネットワーク層のそれぞれについて、ニューラルネットワークベースのモデルの基本モデル重み行列を取得し、

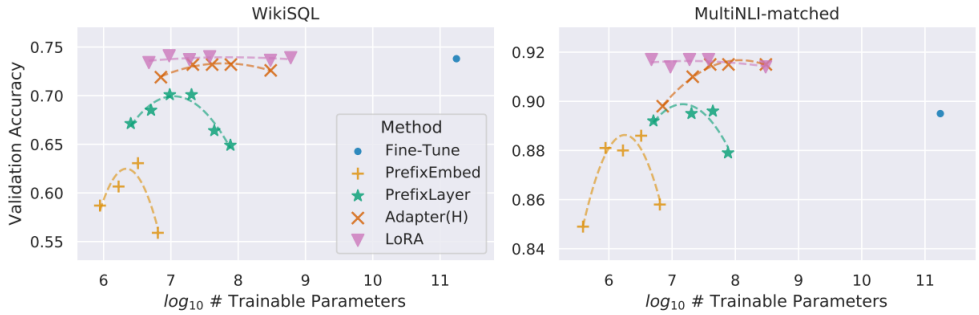
第 1 のドメインモデルを形成するために、ベースモデルの重み行列に、トレーニング可能なパラメータとして扱われる対応する第 1 の低ランク因数分解行列を追加し、

ベースモデルの重み行列を変更せずに、最初のドメイン固有のトレーニングデータを使用して第 1 のドメインモデルをトレーニングする。

4. 本特許に関連する論文

本特許に関する論文 “LORA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS”[1]が、Microsoft の Edward Hu 氏らにより公表されている。

論文には LoRA と他の方法との比較が示されている。下記グラフは、GPT-3 175B 検証精度と、WikiSQL および MNLI(Multi Natural Language Inference) マッチングにおける複数の適応方法のトレーニング可能なパラメータ数の比較を示すものである。



グラフに示すように LoRA は、他の方法と比較してより優れたスケーラビリティとタスクパフォーマンスを示している。

LoRA に関する各種コードは下記サイトに開示されている。

https://github.com/microsoft/LoRA

[1] Edward Hu, et al. “LORA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS” arXiv:2106.09685v2 [cs.CL] 16 Oct 2021

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。