AI 特許紹介(65)

AI 特許を学ぶ！究める！
～スイッチトランスフォーマ特許～

2024 年 6 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　2023 年 7 月 7 日
公開日　2023 年 11 月 2 日
公開番号　US20230351188
発明の名称　スイッチ層を備えたニューラルネットワーク

188 特許は、入力されるサンプルごとに異なるパラメータを選択する MoE（Mixture of Experts）ルーティングアルゴリズムを簡素化し、通信コストと計算コストを削減するスイッチトランスフォーマ技術に関する。

2.特許内容の説明

深層学習では、モデルは通常、すべての入力に対して同じパラメータを再利用する。Mixture of Experts (MoE) モデルはこれを無視し、代わりに、入力されるサンプルごとに異なるパラメータを選択する。その結果、膨大な数のパラメータを備えた、まばらにアクティブ化されたモデルが作成されるが、計算コストは一定である。

しかし、MoE はいくつかの注目に値する成功を収めているにもかかわらず、複雑さ、通信コスト、トレーニングの不安定さによって広範な導入が妨げられている。188 特許はスイッチトランスフォーマの導入によってこれらの課題を解決するものである。

図 3A は、フィードフォワードサブ層の代わりにスイッチ層 330（スイッチング FFN 層）を含むアテンションニューラルネットワーク層の例を示す。

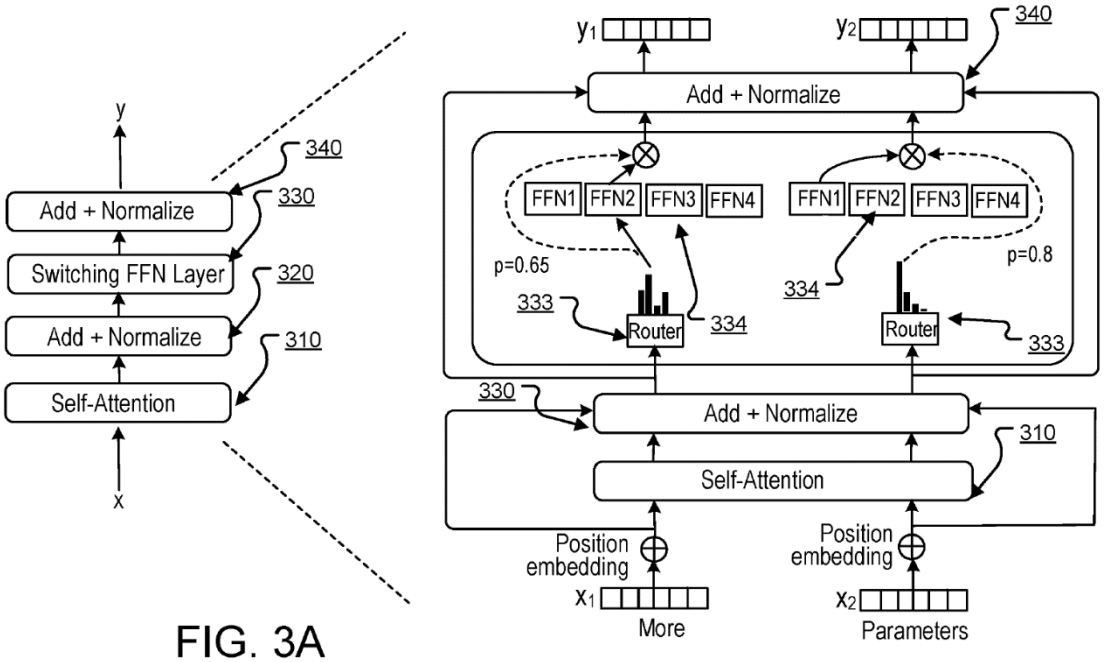


FIG. 3A

アテンションニューラルネットワーク層は、入力シーケンスにアテンション機構を適用するアテンションサブ層 310 と、その後、アテンション入力シーケンスを生成するための「add & norm」オペレーション 320 を含む。「add & norm」オペレーション 330 は、レイヤ正規化オペレーションが後に続く残差接続を含む。

アテンション層は、アテンション入力シーケンスを処理して、アテンション入力シーケンス内のアテンション層入力ごとのそれぞれの出力を含むアテンション層の出力シーケンスを生成する。スイッチ層 330 は、対象入力シーケンス内の各位置に対して個別に、すなわち位置ごとに動作するように構成されている。特に、各入力位置について、スイッチ層 330 は、入力位置でアテンション層入力を受信し、入力位置でアテンション層入力に一組の変換を適用して、入力位置に対する出力を生成する。

1 つのスイッチ層入力の計算は、他のスイッチ層入力の計算から独立しているため、一般に、スイッチ層 330 は、所与の入力シーケンス内、またはニューラルネットワーク

への所与の入力バッチ内で各スイッチ層入力の処理を並行して実行する。

次に、アテンション層は、初期出力に“add&norm”操作 340 を適用して、アテンション層の出力シーケンスを生成する。より具体的には、図３Ａは、アテンション入力シーケンスの２つの位置にある２つのトークンに対するスイッチ層 330 の動作を示す(1 つのトークン x1 は単語「more」に対応し、もう 1 つのトークン x2 は単語「parameters」に対応する)。

上記例のアテンション層はニューラルネットワーク内の最初のアテンション層であるため、アテンション層は、アテンションサブ層 310 および add&norm 操作 330 を使用してトークンを処理する前に、まず各トークンに位置埋め込みを適用する。

スイッチ層 330 は、ルーティング機能 333 と４つのエキスパートニューラルネットワーク 334 を含む。各エキスパートニューラルネットワーク 334 は、フィードフォワードニューラルネットワーク（FFN)、例えば多層、例えば 2 層または 3 層の、例えば ReLU または GeLU 活性化関数を備えた全結合層のニューラルネットワークである。

第１のトークンに関して、スイッチ層 330 はルーティング機能 333 を第１のトークンに適用してエキスパートのスコアのセットを生成し、最高スコアは第２のエキスパート（「FFN2」）のスコア 0.65 である。これに基づいて、スイッチ層 330 は、FFN2 のみを使用して第 1 のトークンを処理して、FFN2 のエキスパート出力を生成し、その後、FFN2 のスコア（0.65）と、FFN2 によって生成されたエキスパート出力との積を計算することによって、第１のトークンの最終出力を生成する。

第２のトークンの場合、スイッチ層 330 はルーティング機能 333 を第２のトークンに適用してエキスパートのスコアのセットを生成し、最高スコアは第１のエキスパート（「FFN1」）のスコア 0.8 である。これに基づいて、スイッチ層 330 は、FFN1 のみを使用して第２のトークンを処理して、FFN1 のエキスパート出力を生成し、次いで、FFN1 のスコア（0.8）と、FFN1 によって生成されたエキスパート出力との積を計算することによって、第１のトークンの最終出力を生成する。

3.クレーム

188 特許のクレーム１は以下の通りである。

1. ネットワーク入力に対して機械学習タスクを実行してネットワーク出力を生成するシステムにおいて、該システムは、１つまたは複数のコンピュータと、１つまたは複数

のコンピュータによって実行されると、1つまたは複数のコンピュータに以下を実行させる命令を格納する1つまたは複数の記憶装置とを備え、

機械学習タスクを実行するように構成されたニューラルネットワークであって、1つまたは複数のスイッチ層を含むニューラルネットワークを備え、各スイッチ層は以下を含む：

(i)ルーティングパラメータを持つそれぞれの学習されたルーティング関数と、

(ii) それぞれの複数のエキスパートニューラルネットワークであって、それぞれがエキスパートパラメータのそれぞれのセットを有し、スイッチ層に対するスイッチ層入力を受信し、スイッチ層に対するスイッチ層入力を、エキスパートニューラルネットワークのエキスパートパラメータのそれぞれのセットに従って、スイッチ層のそれぞれの初期スイッチ層出力を生成すべく、処理するように構成されており、各スイッチ層は以下のように構成されており、

スイッチ層のスイッチ層入力を受信し、

スイッチ層内の複数のエキスパートニューラルネットワークのそれぞれについてのそれぞれのルーティングスコアを含むスコア分布を生成するために、それぞれの学習されたルーティング関数のルーティングパラメータの現在値に従って、スイッチ層のそれぞれの学習されたルーティング関数をスイッチ層入力に適用し、

複数のエキスパートニューラルネットワークから、最も高いルーティングスコアを有するエキスパートニューラルネットワークのみを選択し、

層入力に対する初期スイッチ層出力を生成するために、選択されたエキスパートニューラルネットワークのみを使用し、選択されたエキスパートニューラルネットワークのエキスパートパラメータの現在値に従ってスイッチ層入力を処理し、

選択されたエキスパートニューラルネットワークのルーティングスコアと、選択されたエキスパートニューラルネットワークによって生成された初期スイッチ層出力との積を計算することを含み、スイッチ層の最終スイッチ層出力を生成する。

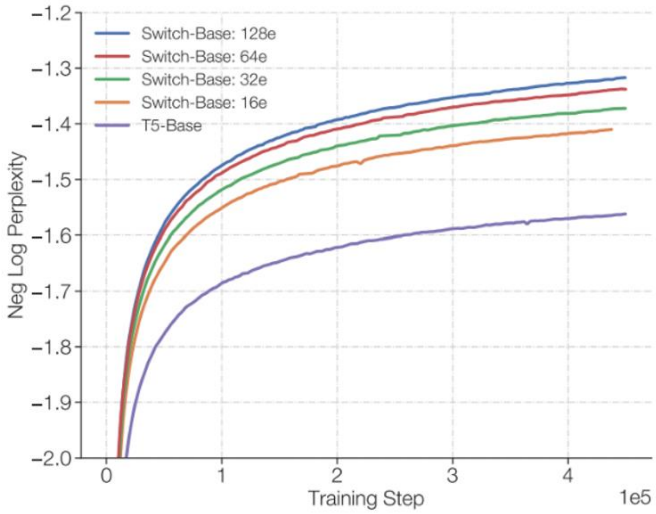
4. 本特許に関連する論文

本特許に関する論文“Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity”[1]が、Google の William Fedus 氏らにより公表されている。

論文ではスイッチトランスフォーマと MoE との比較がなされている。

[1] William Fedus, et al. “Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity” arXiv:2101.03961v3 [cs.LG] 16 Jun 2022

Model	Capacity Factor	Quality after 100k steps (↑) (Neg. Log Perp.)	Time to Quality Threshold (↓) (hours)	Speed (↑) (examples/sec)
T5-Base	—	-1.731	Not achieved†	1600
T5-Large	—	-1.550	131.1	470
MoE-Base	2.0	-1.547	68.7	840
Switch-Base	2.0	-1.554	72.8	860
MoE-Base	1.25	-1.559	80.7	790
Switch-Base	1.25	-1.553	65.0	910
MoE-Base	1.0	-1.572	80.1	860
Switch-Base	1.0	-1.561	**62.8**	1000
Switch-Base+	1.0	**-1.534**	67.6	780



上記テーブル及びグラフは、MoE トランスフォーマおよび T5 高密度ベースラインに対するスイッチトランスフォーマのメリットをステップごとおよび時間ごとに測定して比較したものである。品質は、負の対数複雑度（Negative log perplexity）と、任意に選択された品質しきい値 Neg. Log Perp.=-1.50 に到達するまでの時間によって測定される。 T5-Base と比較して大幅に品質が向上していることが理解できる。

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。