AI 特許紹介(59)
AI 特許を学ぶ！究める！
～GSahrd 特許～

2023 年 12 月 8 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　2021 年 6 月 30 日
公開日　2023 年 7 月 13 日
公開番号 US20230222318
発明の名称　条件付き計算を備えたアテンション ニューラル ネットワーク

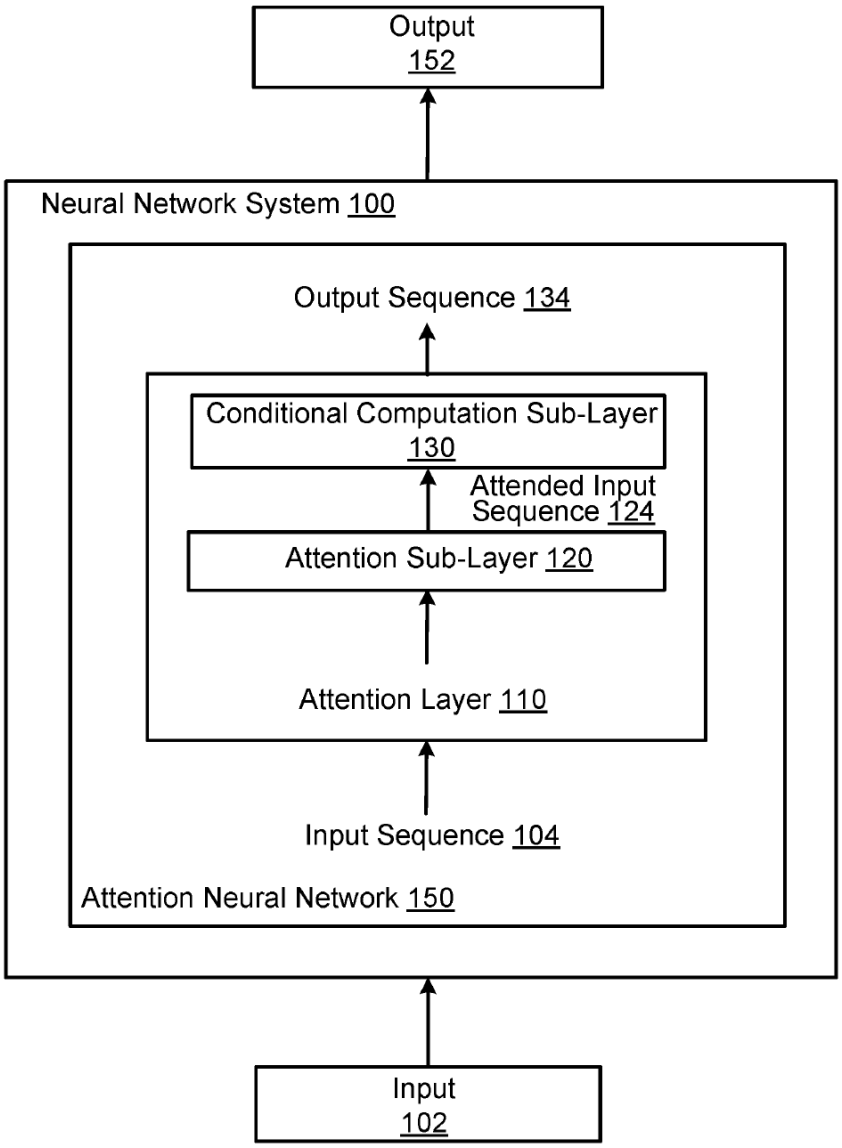
318 特許は、トランスフォーマーモデルにおけるフィードフォワード層を条件付き計算サブ層に置き換え、自動シャーディングを使用することにより、トランスフォーマーモデルを大幅にスケールアップする技術に関する。

2.特許内容の説明
ニューラルネットワークのスケーリングは、膨大な量のトレーニングデータとコンピューティングを使用する機械学習アプリケーションでモデルの品質を向上させるために重要である。このスケーリングの傾向は、モデルの品質を向上させるための確実なアプローチであることが確認されているが、その過程には、計算コスト、プログラミング

の容易さ、並列デバイスでの効率的な実装などの課題がある。

318 特許は、トランスフォーマーモデルにおけるフィードフォワード層を条件付き計算サブ層に置き換え、自動シャーディングを使用することにより、トランスフォーマーモデルを大幅にスケールアップする。

図 1 は、例示的なニューラルネットワークシステム 100 を示す。



ニューラルネットワークシステム 100 は、入力 102 を受信し、入力 102 に対して機械学習タスクを実行して出力 152 を生成する。ニューラルネットワークシステム 100 は、複数のアテンション層 110 を含むアテンションニューラルネットワーク 150 を含む。各アテンション層 110 は、入力シーケンス 104 に対して動作し、対応する出力シーケンス 134 を生成する。

入力シーケンス 104 から出力シーケンス 134 を生成するために、各アテンション層 110 はアテンションサブ層 120 およびフィードフォワードサブ層を含む。アテンションサブ層 120 は、層 110 の入力シーケンス 104 を受信し、その層の入力シーケンスにアテンション機構を適用して、アテンション入力シーケンス 124 を生成する。一般に、アテンション機構を適用するために、サブ層 120 は 1 つまたは複数のアテンションヘッドを使用する。各アテンションヘッドは、クエリのセット、キーのセット、バリューのセットを生成し、クエリ、キー、バリューを使用してクエリ-キー-バリュー(QKV)アテンションの様々なバリエーションを適用して出力を生成する。

フィードフォワードサブ層は、アテンション入力シーケンス 124 に作用して、層 110 の出力シーケンス 134 を生成する。より具体的には、アテンションニューラルネットワーク内の層 110 の一部またはすべてについて、フィードフォワードサブ層は、対象入力シーケンス 124 内の異なる対象層入力を処理するために異なるコンポーネントを使用する条件付き計算サブ層 130 である。

図 2 は、従来のフィードフォワードサブ層 210 を含むアテンションニューラルネットワーク層(左側)と、条件付き計算サブ層 250 を含むアテンションニューラルネットワーク層（右側）の一例を示す。

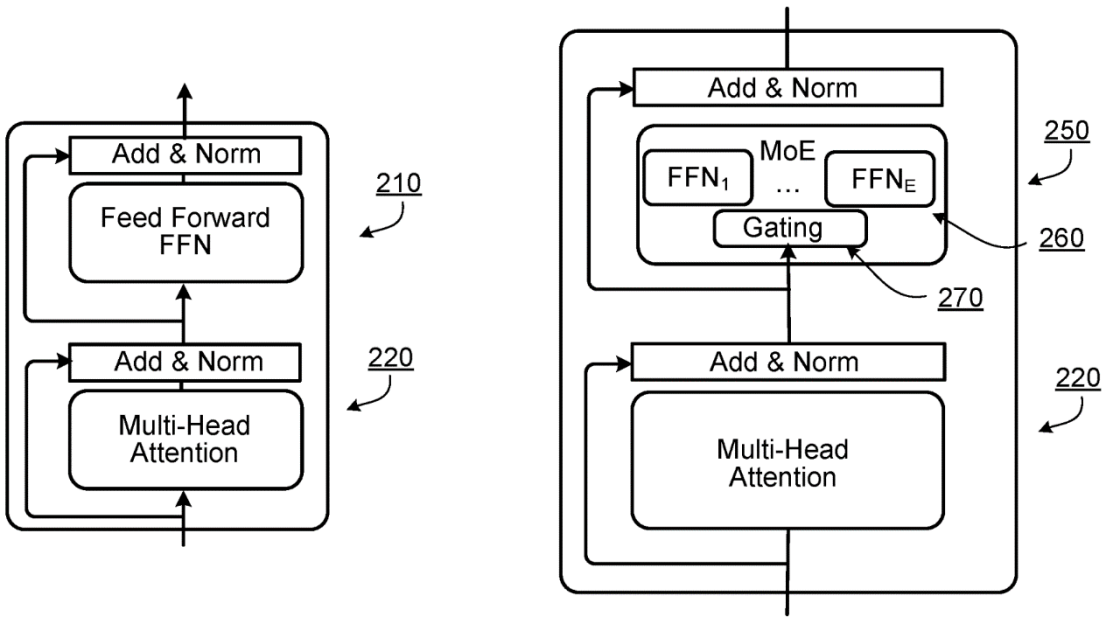


FIG. 2

両方の層は、層の入力シーケンスにアテンション機構を適用し、その後、アテンション入力シーケンスを生成する「加算＆ノルム“add & norm”」操作を適用するアテンシ

ョンサブ層 220 を含む。「加算およびノルム」操作には、層正規化操作が後に続く残留接続が含まれる。フィードフォワードサブ層 210 と条件付き計算サブ層 250 は両方とも、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各アテンション層入力（トークン）に対するそれぞれの出力を含むアテンションニューラルネットワーク層 250 の出力シーケンスを生成する。

フィードフォワードサブ層 210 は、対象入力シーケンス内の各位置を個別に、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、入力位置ごとに、フィードフォワードサブ層 210 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。

より具体的には、フィードフォワードサブ層 210 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フィードフォワードニューラルネットワーク（FFN）を含む。特に、入力位置ごとに、フィードフォワードサブ層 210 は、入力位置でアテンション層入力を受信し、FFN を使用してアテンション層入力を処理して、入力位置に対する初期出力を生成するように構成される。次に、フィードフォワードサブ層 210 は、「加算＆ノルム」演算を初期出力に適用して、アテンション層の出力シーケンスを生成する。

条件付き計算サブ層 250 も位置に関する問題で動作するが、同じ FFN を使用して各対象層入力を処理する代わりに、複数のエキスパート FFN260（「エキスパート」とも呼ばれる）を維持する。各エキスパートは一般に同じアーキテクチャを持っているが、アテンションニューラルネットワークのトレーニングの結果として異なるパラメータ値を持つ。例えば、各エキスパートは、例えば ReLU または GeLU 活性化関数を備えた全結合層の多層、例えば２層または３層のニューラルネットワークである。

任意の所与の入力位置について、条件付き計算サブ層 250 は、その位置にあるそれぞれのトークンにゲート関数 270 を適用して、複数のエキスパート 260 のそれぞれについて各ゲートスコアを生成する。一般に、ゲート関数 270 は、学習されたパラメータを有し、学習されたパラメータに従ってトークンをそれぞれのゲートスコアにマッピングする関数である。ゲート関数 270 は、エキスパートの学習ベクトルと位置のアテンション層入力との間の内積を計算することによって、各エキスパートのそれぞれの初期スコアを生成し、その後、初期スコアにソフトマックス関数を適用することによってゲートスコアを計算する。

次に、各トークンについて、条件付き計算サブ層 250 は、複数のエキスパート FFN か

ら、少なくとも各ゲートスコアに基づいて適切なサブセットを選択する。一般に、条件付き計算サブ層 250 は、適切なサブセットに含まれるEのエキスパートのうち最大 k 個を選択する。k はエキスパート E の総数に比べて小さい正の整数である。

特定の例として、k は 2 または 10 未満の別の小さな整数に等しく、E は少なくとも 100 に等しくなる。場合によっては、E は少なくとも 500 に等しくなり、これらの場合によっては、E は少なくとも 2000 に等しくなる。したがって、特定のアテンション層入力に対して、サブ層は、たとえば最大 2%、場合によっては最大 0.1% のエキスパートを選択する。言い換えれば、多数のエキスパート E を維持することで、アテンションニューラルネットワークに大幅な追加容量が与えられ(ニューラルネットワークのパラメータの総数が増加する)、与えられた入力にはごく一部のみが使用されるため、ニューラルネットワークを使用した入力の処理が計算効率を維持できる。特に、アテンションニューラルネットワークのトレーニング後、サブ層 250 は、上位 k のエキスパート、すなわち、最も高いゲートスコアを有する k のエキスパートを、適切なサブセット内のエキスパートとして選択する。

所与のトレーニングバッチ内のすべてのネットワーク入力を処理するには、条件付き計算サブ層 250 は合計N個のトークンを処理する必要がある。つまり、N は、指定されたトレーニングバッチ内のすべてのネットワーク入力にわたってアテンションサブ層によって生成されたすべてのアテンション入力シーケンスに含まれるトークンの総数である。アテンションニューラルネットワークによって処理または生成されるネットワーク入力、ネットワーク出力、またはその両方のサイズにより、N は数百万になる。

特定のトレーニングバッチでのニューラルネットワークのトレーニング中に、サブ層は、各グループに S=N/G トークンが含まれるように、特定のバッチ「内の」合計 N 個のトークンを G 個のグループに分割できる。次に、サブ層は各グループに最大数（容量）を割り当て、グループ内の各トークンのエキスパートを選択する。これにより、最大数のトークンがグループ内のすべてのトークンの中から特定のエキスパートにディスパッチされるが、各トークンに対して最大 k のエキスパートが選択される。たとえば、k が 2 に等しい場合、各グループには 2N/(GE) に等しい容量を割り当てることができる。

より一般的には、容量は O(N/GE) に等しくなる。これにより、バッチの N 個のトークンすべてのうち、トレーニングバッチのトークンのうち O(N/GE) を超えるエキスパートが割り当てられないことが保証される。さらに、システムは各グループのトークンの処理を独立して並行して実行できる。各グループ内で制約が確実に満たされるよう

にするために、システムは各グループ内でエキスパートの選択を順番に実行できる。

次に、サブ層 250 は、適切なサブセット内のエキスパート FFN のそれぞれを使用して、所与の位置でそれぞれのアテンション層入力を処理する。すなわち、適切なサブセット内に存在しないエキスパート FFN を使用せず、適切なサブセット内の各エキスパート FFN についてそれぞれのエキスパート出力を生成する。

次に、サブ層 250 は、それぞれのエキスパート出力を結合して、結合されたエキスパート出力を生成する。サブ層は、選択された各エキスパートフィードフォワードニューラルネットワークのそれぞれの正規化されたゲートスコアを生成し、エキスパート出力を生成した選択されたエキスパートフィードフォワードニューラルネットワークの正規化されたゲートスコアによって各エキスパート出力に重み付けされた、それぞれのエキスパート出力の加重和を計算する。

サブ層は、k エキスパートのゲートスコアの合計が 1 になるように、最も高いゲートスコアを持つ k エキスパートのゲートスコアを正規化することにより、選択された各エキスパートのそれぞれの正規化ゲートスコアを計算する。サブ層 250 は、結合されたエキスパート出力からの位置でそれぞれの層出力を生成する。たとえば、フィードフォワード層は、位置で結合されたエキスパート出力をそれぞれの層出力として使用することも、入力位置の結合されたエキスパート出力に「加算およびノルム」演算を適用して、それぞれの層出力を生成することもできる。

3.クレーム

318 特許のクレーム 1 は以下の通りである。

1.ネットワーク出力を生成するために、ネットワーク入力に対して機械学習タスクを実行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複数の層を含むアテンションニューラルネットワークを備え、各層はアテンションサブ層とフィードフォワードサブ層で構成され、該アテンションサブ層は、以下のように構成されており：

1 つ以上の位置のそれぞれで各層入力を含む層の入力シーケンスを受信し、

層の入力シーケンスにアテンションメカニズムを適用することによって、少なくとも部分的にアテンション入力シーケンスを生成し、該アテンション入力シーケンスは、1 つ以上の位置のそれぞれにおける各アテンション層入力を含み、

前記フィードフォワード層は次のように構成されており、:

アテンション入力シーケンスを受信し、

アテンション入力シーケンスから層の出力シーケンスを生成し、該出力シーケンスは、１つ以上の位置のそれぞれにおける各層出力を含み、

複数の層のうちの少なくとも１つについて、フィードフォワードサブ層は、(i)複数のエキスパートフィードフォワードニューラルネットワークを備え、(ii)処理を実行することによってその層の出力シーケンスを生成するように構成され、前記処理は、層の入力シーケンス内の各位置について、以下を含み：

その位置で各アテンション層入力を受信し、

複数のエキスパートフィードフォワードニューラルネットワークのそれぞれについて各ゲートスコアを生成するために、その位置で各アテンション層入力にゲート関数を適用し、

複数のエキスパートフィードフォワードニューラルネットワークから、少なくともそれぞれのゲートスコアに基づいて適切なサブセットを選択し、

各エキスパートフィードフォワードニューラルネットワークに対するそれぞれのエキスパート出力を生成するために、適切なサブセット内のエキスパートフィードフォワードニューラルネットワークのそれぞれを使用して、その位置での各アテンション層入力を処理し、

結合されたエキスパート出力を生成するために、それぞれのエキスパート出力を結合し、

結合されたエキスパート出力からその位置でそれぞれの層出力を生成する。

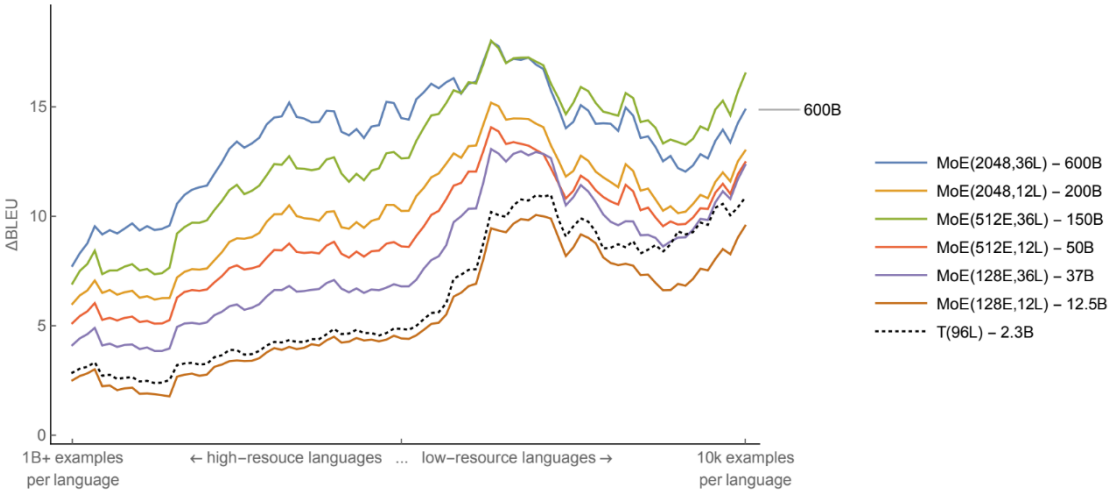
4. 本特許に関連する論文

本特許に関する論文“GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding” [1]が、Google の Dmitry Lepikhin 氏らにより公表されている。

GShard により、自動シャーディングを使用して、まばらにゲートされたエキスパートの混合による Transformer モデルを 6,000 億パラメータを超えてスケールアップしている。このような巨大なモデルを 2048 TPU v3 アクセラレータで 4 日間効率的にトレーニングし、100 言語から英語への翻訳において従来技術と比較してはるかに優れた品質を達成できることを実証している。下記グラフは GShard ベースラインと単

[1] Dmitry Lepikhin et al. “GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding” arXiv:2006.16668v1 [cs.CL] 30 Jun 2020

言語ベースラインでトレーニングされた多言語 MoE（Mixture-of-Experts）Transformer モデルの翻訳品質の比較を示す。



Id	Model	BLEU avg.	ΔBLEU avg.	Weights
(1)	MoE(2048E, 36L)	**44.3**	13.5	600B
(2)	MoE(2048E, 12L)	41.3	10.5	200B
(3)	MoE(512E, 36L)	43.7	12.9	150B
(4)	MoE(512E, 12L)	40.0	9.2	50B
(5)	MoE(128E, 36L)	39.0	8.2	37B
(6)	MoE(128E, 12L)	36.7	5.9	12.5B
*	T(96L)	36.9	6.1	2.3B
*	Baselines	30.8	-	100×0.4B

X 軸は、高リソースから低リソースまでの言語を表す。縦軸に ΔBLEU を示す。これは、特定の言語用にトレーニングおよび調整された単一言語の Transformer モデルと比較した、単一の多言語モデルの品質向上を表す。GShard でトレーニングされた MoE Transformer モデルは、実線の傾向線でレポートされている。

破線の傾向線は、同じデータセット上で GPipe を使用してトレーニングされた単一の 96 レイヤー多言語 Transformer モデル T(96L) を表す。

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。