AI 特許紹介(47)
AI 特許を学ぶ！究める！
～Conformer 特許～

2022 年 12 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　Google
出願日　2020 年 12 月 31 日
公開日　2022 年 6 月 30 日
公開番号 US20220207321
発明の名称　畳み込み拡張トランスフォーマーモデル

321 特許は、畳み込みニューラルネットワークモデルとトランスフォーマーモデルとを組み合わせたコンフォーマー技術に関する。

2.特許内容の説明
最近では、セルフアテンションベースのモデル（トランスフォーマーモデルなど）と、畳み込みニューラルネットワークベースのモデルが、自動音声認識（ASR）で有望な結果を示しており、リカレントニューラルネットワーク（RNN）よりも優れている。

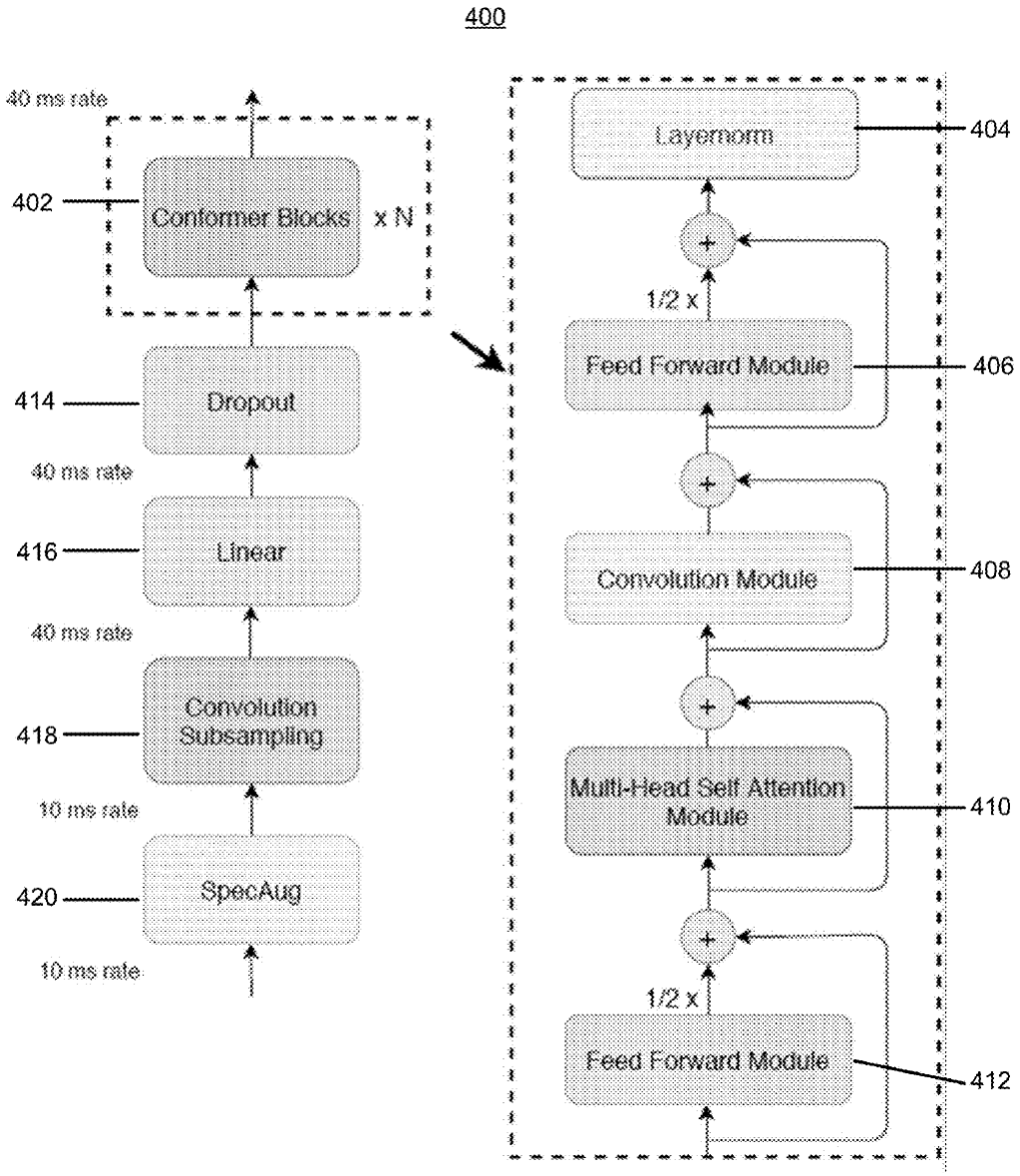
トランスフォーマーモデルは、コンテンツベースのグローバルインタラクションのキ

ャプチャに優れているが、CNN はローカル特徴を効果的に引き出す。しかしながら、セルフアテンションまたは畳み込みを含むモデルには、それぞれ独自の制限がある。トランスフォーマーは、長期にわたるグローバルコンテキストのモデル化に優れているが、きめ細かいローカル特徴パターンを抽出する能力は低くなる。

一方、畳み込みニューラルネットワークはローカル情報を活用し、視覚における事実上の計算ブロックとして使用される。それらは、ローカルウィンドウ上で位置ベースの共有カーネルを学習する。これにより、平行移動の等価性が維持され、エッジや形状などの特徴をキャプチャできる。

ローカル接続を使用する際の制限の 1 つは、グローバル情報を取得するためにさらに多くのレイヤーまたはパラメーターが必要になることである。この問題に対処するために、特定の既存の手法は、より長いコンテキストをキャプチャするために、各残差ブロックに Squeeze-and-Excitation モジュールを採用している。ただし、シーケンス全体にグローバル平均化を適用するだけであり、動的なグローバルコンテキストのキャプチャにはまだ制限がある。

321 特許は上記問題を解決すべく畳み込みニューラルネットワークモデルとトランスフォーマーモデルとを組み合わせたものである。下記図は、コンフォーマーモデル 400 のブロック図である。



コンフォーマーモデル 400 は、人間の音声の入力データのセットを受信し、入力データの受信の結果として、音声をテキストに解釈する出力データを提供するようにトレーニングされる。

その他、コンフォーマーモデル 400 は、タンパク質を記述する入力データのセットを受け取り、入力データの受け取りの結果として、タンパク質合成データを含む出力データを提供するようにトレーニングすることもできる。

コンフォーマーモデル 400 は、データを処理するように動作可能な第 1 のフィードフォワードブロック 412、セルフアテンションブロック 410、畳み込みブロック 408、および第 2 のフィードフォワードブロック 406 を含む。

特に、上記図は、コンフォーマーエンコーダモデルアーキテクチャの１つの実装を示している。この実装では、コンフォーマーブロック 402 は、SpecAugment ブロック 420、畳み込みサブサンプリングブロック 418、線形変換ブロック 416、およびドロップアウトブロック 414 によって最初に処理された入力データを取得する。

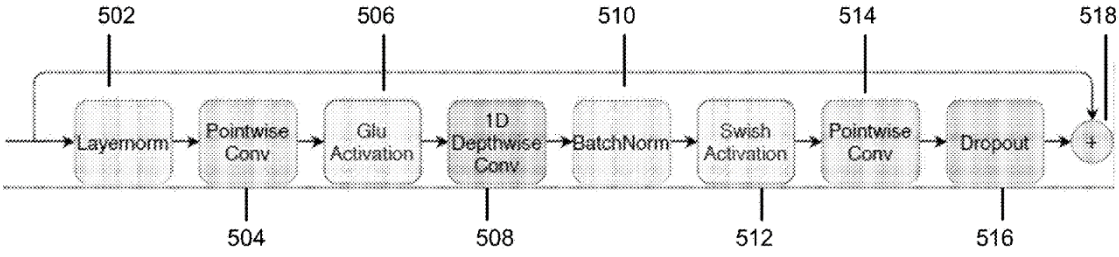
SpecAugment ブロック 420 は、プレコンフォーマーブロック処理ステップを入力データに提供する。SpecAugment ブロック 420 は、コンフォーマーブロックが音声認識のためにデータを処理するのを助ける（音声データを画像データに変換してから水増しする）。

畳み込みサブサンプリングブロック 418 は、コンフォーマーブロック 402 によって処理される前に、入力データをサブサンプリングする。次の線形変換ブロック 416 およびドロップアウトブロック 414 は、コンフォーマーブロック 402 処理のための入力データをさらに準備する。

コンフォーマーブロック 402 は、マルチヘッドセルフアテンションブロック 410 と畳み込みブロック 408 を挟むハーフステップ残差結合を有する２つのマカロン状フィードフォワードブロック 406 および 412 を含む。マカロン構造の後には、ポストレイヤーノルム 404 が続く。

ハーフステップフィードフォワードブロックは、計算効率を高めることができ、セルフアテンションブロック 410 と畳み込みブロック 408 のペアリングは、キャプチャされるローカル相関とグローバル相関の両方を提供することができる。最後に、レイヤノルムブロック 404 は、サンドイッチ構造の出力を正規化する。

下記図は、畳み込みブロック 500 のブロック図を示す。

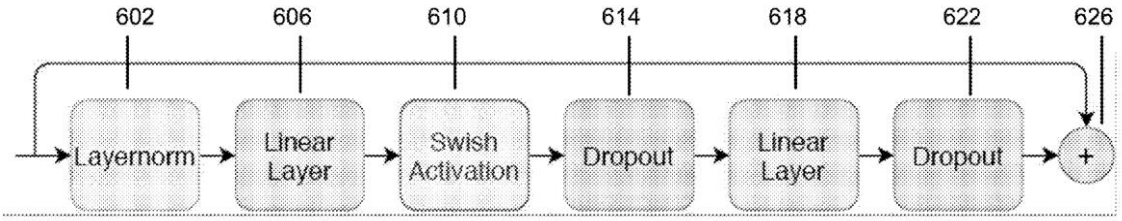


畳み込みブロック 500 は、畳み込みブロック 408 として示されているコンフォーマーモデル 400 に実装することができる。畳み込みブロック 500 は、相対オフセットに基づく局所相関を学習し、キャプチャする。畳み込みブロック 500 は、チャネル入力の

半分をゲーティングする GLU 活性化層 506 を有するチャネル数を射影する 2 の拡張係数を有する点ごとの畳み込み 504 にフィードするレイヤノルムブロック 502 を含む。そして、その後に 1 次元深さ方向畳み込み 508 が続く。

1D 深さ方向畳み込み 508 の後に、バッチノルム 510 が続き、その後、スウィッシュ活性化層 512 が続く。次に、スウィッシュ活性化層 512 は、点ごとの畳み込み 514 に供給され、その後にドロップアウトブロック 516 が続く。その後、セルフアテンションブロックの出力を取り込み、フィードフォワードブロックによって処理できる畳み込み出力データ 518 を生成する。

下記図は、フィードフォワードブロック 600 のブロック図である。



フィードフォワードブロック 600 は、第 1 のフィードフォワードブロック 412 または第 2 のフィードフォワードブロック 406 のいずれかとして、コンフォーマーモデル 400 に実装することができる。

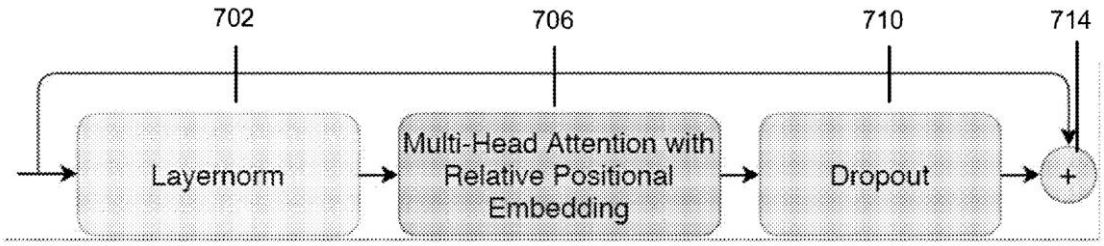
特に、上記図は、6 つのサブブロックを有するフィードフォワードブロックの 1 つの実装を示している。この実装における 6 つのサブロックは、レイヤノルムブロック 602、第 1 の線形変換ブロック 606、スウィッシュ活性化ブロック 610、ドロップアウトブロック 614、第 2 の線形変換ブロック 618、およびドロップアウトブロック 622 を含む。

第 1 の線形変換ブロック 606 は拡張係数 4 を使用することができ、第 2 の線形変換ブロック 618 はデータをモデル次元に射影する。システムは、両方のフィードフォワードブロックで、スウィッシュ活性化ブロック 610 およびプレノルム残差ユニットを使用することができる。

第 1 のフィードフォワードブロックの場合、出力データ 626 は、セルフアテンションブロックによって処理できる第 1 のフィードフォワード出力データ 626 である。第 2 のフィードフォワードブロックについて、ブロックは、入力データを取得し、入力データを処理し、第 2 のフィードフォワード出力データを生成することができ、これはコン

フォーマーブロックの最終出力データである。

下記図は、セルフアテンションブロック 700 のブロック図である。



セルフアテンションブロック 700 は、セルフアテンションブロック 410 としてコンフォーマーモデル 400 に実装することができる。上記図は、3 つのサブブロックを有するマルチヘッドセルフアテンションブロックの 1 つの実装を示す。この実施形態では、3 つのサブブロックは、レイヤノルムブロック 702、マルチヘッドアテンションブロック 706、およびドロップアウトブロック 710 を含む。

システムは、基準前の残差ユニットに相対的な位置を埋め込むマルチヘッドセルフアテンションを使用する。さらに、フィードフォワードブロックの出力を取り込み、畳み込みブロックによって処理できるアテンション出力データ 714 を生成する。

3.クレーム

321 特許のクレーム 1 は以下の通りである。

1. ローカルとグローバルの両方の依存関係を説明するデータを効率的に処理するためのコンピュータ実装方法において、

1 つまたは複数のコンフォーマーブロックを含む機械学習されたコンフォーマーモデルを記述するデータにアクセスし、各コンフォーマーブロックはブロック入力を処理してブロック出力を生成するように構成されており、各コンフォーマーブロックは以下を含み、

ブロック入力を処理して第 1 のフィードフォワード出力を生成するように構成された第 1 のフィードフォワードブロックと、

第 1 のフィードフォワード出力を処理してアテンション出力を生成するセルフアテンションを実行するように構成されたセルフアテンションブロックと、

畳み込みフィルタを使用して畳み込みを実行し、セルフアテンションブロックのアテンション出力を処理して畳み込み出力を生成するように構成された畳み込みブロックと、

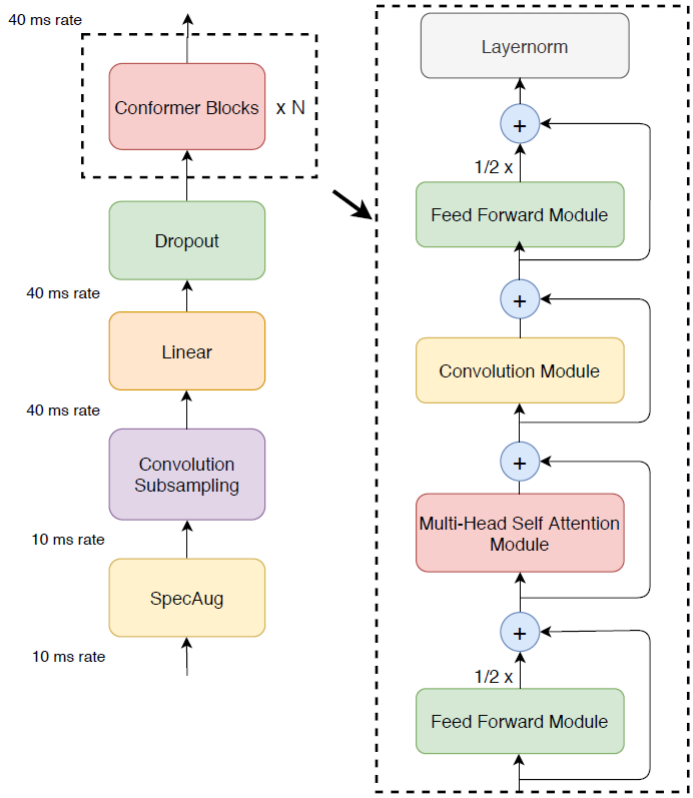
畳み込みブロックの畳み込み出力を処理して第 2 のフィードフォワード出力を生

成するように構成された第２のフィードフォワードブロックとを備え、
入力データを取得し、
入力データを機械学習したコンフォーマーモデルで処理して、出力データを生成する。

4. 本特許に関連する論文

本特許に関する論文“Conformer: Convolution-augmented Transformer for Speech Recognition”[1]が、Google の Anmol Gulati 氏らにより公表されている。

論文の図１はコンフォーマーエンコーダーモデルのアーキテクチャを示す。



上述した通り、コンフォーマーは、マルチヘッドのセルフアテンションモジュールと畳み込みモジュールを挟むハーフステップ残差接続を備えた 2 つのマカロンのようなフィードフォワード層で構成され、これにポストレイヤーノルムが続く。

[1] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, Ruoming Pang “Conformer: Convolution-augmented Transformer for Speech Recognition” arXiv:2005.08100v1 [eess.AS] 16 May 2020

テーブル 2 は、LibriSpeech test-clean/testother でのモデルの単語誤り率 (WER) の結果をいくつかの最先端のモデルと比較したものである。

Method	#Params (M)	WER Without LM		WER With LM	
		testclean	testother	testclean	testother
Hybrid					
Transformer [33]	-	-	-	2.26	4.85
CTC					
QuartzNet [9]	19	3.90	11.28	2.69	7.25
LAS					
Transformer [34]	270	2.89	6.98	2.33	5.17
Transformer [19]	-	2.2	5.6	2.6	5.7
LSTM	360	2.6	6.0	2.2	5.2
Transducer					
Transformer [7]	139	2.4	5.6	2.0	4.6
ContextNet(S) [10]	10.8	2.9	7.0	2.3	5.5
ContextNet(M) [10]	31.4	2.4	5.4	**2.0**	4.5
ContextNet(L) [10]	112.7	**2.1**	4.6	**1.9**	4.1
Conformer (Ours)					
Conformer(S)	10.3	**2.7**	**6.3**	**2.1**	**5.0**
Conformer(M)	30.7	**2.3**	**5.0**	**2.0**	**4.3**
Conformer(L)	118.8	**2.1**	**4.3**	**1.9**	**3.9**

コンフォーマーとの比較対象は、ContextNet, Transformer transducer 及び QuartzNet であり、評価結果はすべて小数点以下 1 桁に切り上げている。コンフォーマーモデルは、さまざまなモデルパラメーターサイズの制約に対して一貫して改善を示す。10.3M のパラメーターで、モデルは、既存技術 ContextNet(S).と比較した場合、testother で 0.7% 優れたパフォーマンスを発揮する。30.7M のモデルパラメーターで、コンフォーマーモデルは、139M のパラメーターを持つ Transformer Transducer の以前に公開された最先端の結果を大幅に上回ることができる。

言語モデルがない場合、中型モデルのパフォーマンスは、test/testother で 2.3/5.0 という競争力のある結果を既に達成しており、最もよく知られている Transformer、LSTM ベースのモデル、または同様のサイズの畳み込みモデルよりも優れている。言語モデルを追加すると、モデルは既存のすべてのモデルの中で最も低い単語エラー率を達成できる。このように、単一のニューラルネットワークで Transformer と畳み込みを組み合わせることの有効性が示されている。

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。