AI 特許紹介(64)
AI 特許を学ぶ！究める！
～CoAtNets 特許～

2024 年 5 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
特許権者　Google
出願日　2022 年 5 月 27 日
登録日　2023 年 9 月 12 日
登録番号 US11755883
発明の名称　畳み込みとアテンションを備えた機械学習モデルのシステムと方法

883 特許は、畳み込み層とアテンション層とを相対アテンションを通じて統合し、これらを垂直に積み重ねることで、一般化、容量、及び、効率の向上効果をもたらす CoAtNets (「コート」ネットと発音)に関する。

2.特許内容の説明
畳み込みニューラルネットワーク（CNN）は、ニューラルネットワークで畳み込みフレームを使用する機械学習モデルのクラスである。トランスフォーマは、入力データの異なる部分に重みを付けるアテンションメカニズムを採用する機械学習モデルのクラスである。畳み込みとアテンションを組み合わせる既存のアプローチは、計算コストの

増加などの欠点に直面している。883 特許は、計算コストを削減し、精度を向上させてコンピュータビジョンを実行するためのコンピュータ実装方法を対象とする。

下記図は、畳み込みアテンションネットワーク（CoAtNet）モデル 200 を示すブロック図である。

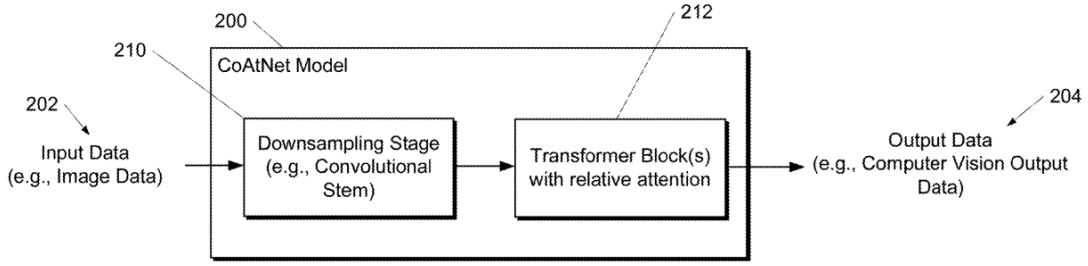


Figure 2

モデル 200 は、例えば画像データまたは他のタスク固有の入力データを記述する入力データのセット 202 を受信するように訓練され、コンピュータビジョンタスク（例えば、画像分類）などの特定の機械学習タスクに応答する出力データ 204 を提供する。モデル 200 は、ダウンサンプリングステージ 210 を含む。ダウンサンプリングステージ 210 は、入力データ 202 の空間解像度を低減する。

例えば、入力データ 202 がテンソルを含む場合、ダウンサンプリングステージ 210 は、ダウンサンプリングステージ 210 の出力が入力データ 202 のテンソルの次元または解像度よりも低い少なくとも 1 つの次元または解像度を有するように、空間解像度を低減する。ダウンサンプリングステージ 210 は、入力データに比べてチャネルの数を増加させることができる。ダウンサンプリングステージ 210 は、畳み込みステムを含む。畳み込みステムは、10 を超えるストライドなど、積極的なストライドを持つ。

モデル 200 は、 1 つまたは複数のアテンションブロック 212 を含む。アテンションブロック 212 は、ダウンサンプリングステージ 202 からダウンサンプリングされた入力データを受信し、出力データ 204 を生成する。アテンションブロック 212 は、相対的なアテンションメカニズムを実装する。アテンションブロック 212 は、トランスフォーマネットワークである。

アテンションブロック 212 は、相対アテンションメカニズムを含む。相対アテンションメカニズムには、静的畳み込みカーネルと適応的アテンション行列の合計を含む。この合計は、相対アテンションメカニズムによる SoftMax 正規化の前または後に適用される。相対アテンションメカニズム（たとえば、SoftMax 正規化の前に適用される）は、

数学的に次のように表すことができる。

$$y_i = \sum_{j \in \mathcal{G}} \frac{\exp\left(x_i^T x_j + w_{i-j}\right)}{\sum_{k \in \mathcal{G}} \exp\left(x_i^T x_k + w_{i-k}\right)} x_j$$

相対アテンションメカニズム（たとえば、SoftMax 正規化の後に適用される）は、数学的に次のように表すことができる。

$$y_i = \sum_{j \in \mathcal{G}} \left(w_{i-j} + \frac{\exp\left(x_i^T x_j\right)}{\sum_{k \in \mathcal{G}} \exp\left(x_i^T x_k\right)} \right) x_j$$

深さ方向の畳み込みカーネル w_{i-j} は、入力テンソル（i,j）の指定されたインデックスの静的な値の入力に依存しないパラメータであり(例えば、インデックス i-j 間の相対的なシフトであり、特定の値ではなく相対的なシフトへの依存は変換等価性と呼ばれ、限られたサイズのデータセットでの一般化を向上させることができる)、x_i と x_j は、それぞれ位置 i での入力と出力であり、g はグローバル受容野（たとえば、位置のセット全体）である。

大域的な受容野の使用（例えば、畳み込みネットワークで伝統的に使用されている限定された局所的な受容野とは対照的に）は、異なる空間位置間の複雑な関係的相互作用をキャプチャする能力の向上を提供することができ、これはより高いレベルの概念を処理するときに望ましい可能性がある 。分母項は、アテンション重み A_{i-j} とも呼ばれる。アテンション重みは、深さ方向の畳み込みカーネルと入力適応型入出力ペアの変換等価性によって共同で決定でき、さまざまな程度で両方の特性を提供し、モデルの一般化、容量、精度を向上させることができる。

3.クレーム

883 特許のクレーム 1 は以下の通りである。

1. 計算コストを削減し、精度を向上させてコンピュータビジョンを実行するコンピュータ実装の方法において、

１つまたは複数のコンピューティングデバイスを含むコンピューティングシステムによって、１つまたは複数の次元を有する入力テンソルを含む入力データを取得し、

コンピューティングシステムによって、機械学習された畳み込みアテンションネットワークに入力データを提供し、機械学習畳み込みアテンションネットワークは２つ以上のネットワークステージを含み、２つ以上のネットワークステージは１つ以上のアテン

ションステージと１つ以上の畳み込みステージを含み、

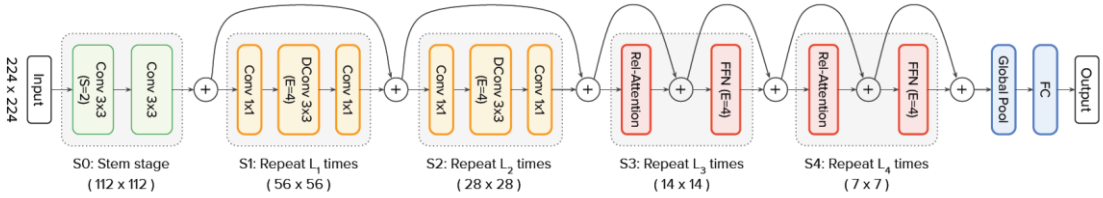
１つ以上のアテンションステージのうちの少なくとも１つは、ソフトマックス正規化を実行し、ソフトマックス正規化の実行の前または後のいずれかに、静的コンボリューションカーネルと適応アテンション行列との和を適用するように構成された相対アテンションメカニズムを備え、

入力データを機械学習畳み込みアテンションネットワークに提供することに応答して、コンピューティングシステムによって機械学習畳み込みアテンションネットワークから機械学習予測を受信する。

4. 本特許に関連する論文

本特許に関する論文“CoAtNet: Marrying Convolution and Attention for All Data Sizes”[1]が、Google の Zihang Dai 氏らにより公表されている。

下記図は、CoAtNet のネットワーク構成図である。



モデルは、S0、S1、S2、S3、および S4 ステージを含む。S0 ステージまたはステムステージは、２つの（例えば、3×3）畳み込み層（例えば、ストライド 2）を含む。さらに、畳み込み S1 ステージおよび S2 ステージはそれぞれ、1×1 畳み込み層、3×3 デコンボリューション層、および 1×1 畳み込み層を含む。

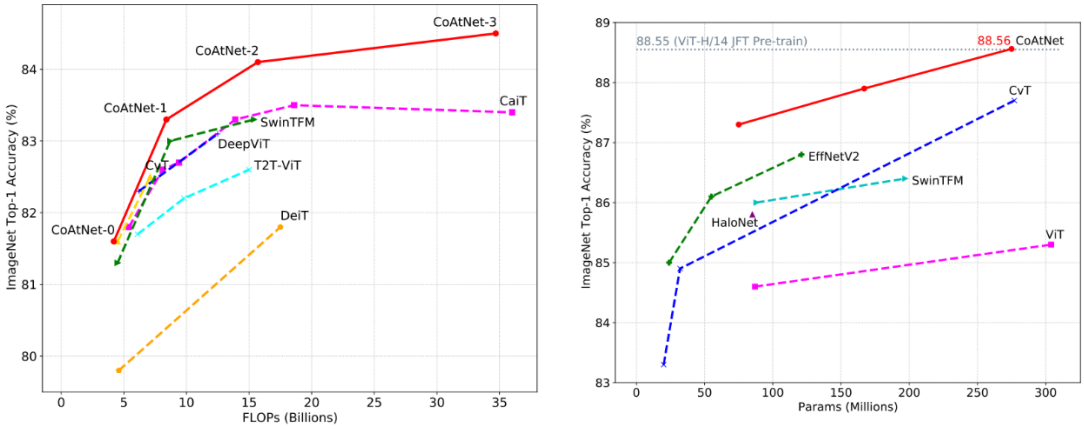
アテンション（S3 や S4 など）ステージはそれぞれ、相対アテンションメカニズムとフィードフォワードネットワークを含む。モデルは、モデル出力を生成するためのグローバルプーリング層と完全接続層をさらに含む。各段階は、設計された回数だけ繰り返すことができる。

CoAtNets は以下の２つの特徴を有する。

(1) 深さ方向の畳み込みとセルフアテンションは、単純な相対的なアテンションを通じて自然に統合できる。

[1] Zihang Dai, et al. “CoAtNet: Marrying Convolution and Attention for All Data Sizes” arXiv:2106.04803v2 [cs.CV] 15 Sep 2021

(2)畳み込み層とアテンション層を原則的な方法で垂直に積み重ねることは、一般化、容量、効率の向上に驚くほど効果的である。



左のグラフは、ImageNet-1K のみ 224x224 に設定した場合の精度対 FLOP(floating-point operations per second) のスケーリング曲線、右のグラフは、ImageNet-21K ⇒ ImageNet-1K 設定での精度対パラメータのスケーリング曲線である。

実験では、CoAtNet がさまざまなデータセットにわたるさまざまなリソース制約下で最先端のパフォーマンスを達成することが示されている。余分なデータがなければ、CoAtNet は ImageNet トップ 1 の 86.0% の精度を達成する。ImageNet-21K の 1,300 万画像で事前トレーニングした場合、CoAtNet は 88.56% のトップ 1 精度を達成し、使用するデータ量が 23 分の 1 でありながら、JFT-300M の 3 億画像で事前トレーニングされた ViT の巨大な精度に匹敵する。

特に、JFT-3B を使用して CoAtNet をさらにスケールアップすると、ImageNet で 90.88% のトップ 1 精度を達成し、新たな最先端の結果が確立された。

コードは、GitHub より入手することができる。
https://github.com/chinhsuanwu/coatnet-pytorch

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。