AI 特許紹介(63)

AI 特許を学ぶ！究める！

～ロバストビジョントランスフォーマ特許～

2024 年 4 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　NVIDIA

出願日　2023 年 3 月 9 日

公開日　2023 年 9 月 14 日

公開番号 US20230290135

発明の名称　ロバストビジョントランスフォーマ

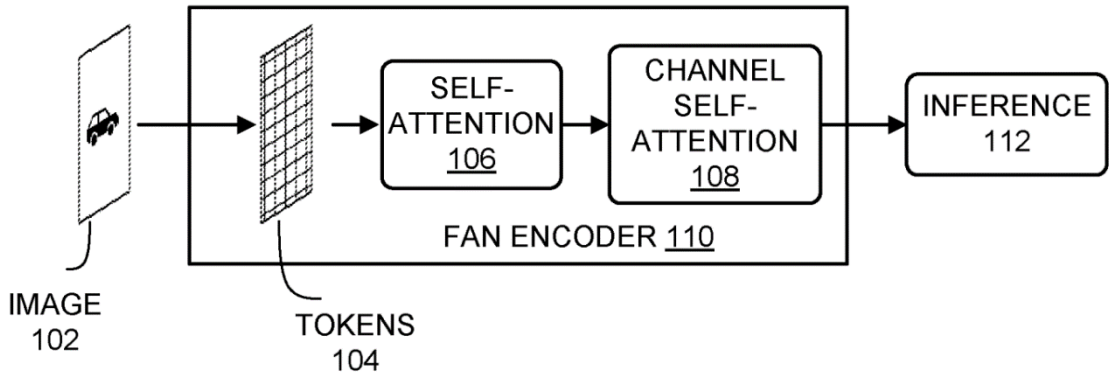
135 特許は、ビジョントランスフォーマにおけるトークンセルフアテンションの後段にチャネルアテンションを設け、ネットワーク全体を完全アテンションにすることによりノイズ及びデータ破損に対する堅牢性を高めたロバストビジョントランスフォーマ技術に関する。

2.特許内容の説明

最近の研究では、ビジョントランスフォーマ（ViT）がさまざまな破損に対して強力な堅牢性を示すことが示されている。この特性は部分的にはセルフアテンションメカニズムに起因すると考えられているが、体系的な理解はまだ不足している。135 特許は、

さらにアテンションチャネル処理設計を組み込むことで堅牢性機能を強化する完全アテンションネットワーク（FAN; fully attention network）を提案している。

図１は、フルアテンションネットワーク(FAN)ビジョントランスフォーマ（ViT）システムを示す。



ViT は、集約された特徴表現に基づいて入力画像に関する推論を生成する。FAN は、入力画像 102、一連のトークン 104、セルフアテンションブロック 106、チャネルセルフアテンションブロック 108、FAN エンコーダ 110、および推論 112 を含む。

ViT は入力画像 102 を n 個のパッチに均一に分割し、各パッチをトークン埋め込みテンソルにエンコードする。例えば $x_i \in \mathbb{R}^d$, i=1, . . . , n.これらすべてのトークン 104 は、セルフアテンションブロック 106 およびチャネルセルフアテンションブロック 108 を含むトランスフォーマブロックのスタックに入力される。トランスフォーマブロックは、トークン混合にセルフアテンションを利用し、チャネルごとの特徴変換に多層パーセプトロン（MLP）を利用する。

セルフアテンションブロック 106 を使用する ViT は、セルフアテンションを利用してグローバル情報を集約する。たとえば、テンソルを埋め込む入力トークンが $X=[x_1, \ldots, x_n] \in \mathbb{R}^{d \times n}$ の場合、セルフアテンションは、パラメータ W_K, W_Q, W_V を使用して線形変換を適用し、それぞれキー $K=W_KX \in R^{d \times n}$, クエリ $Q=W_QX \in \mathbb{R}^{d \times n}$ 及びバリュー$V=W_VX \in \mathbb{R}^{d \times n}$ に埋め込む。

次に、セルフアテンションブロック 106 は、以下のようにアテンションマトリックスを計算し、トークン特徴を集約する。

$$Z^T = SA(X) = \text{Softmax}\left(\frac{Q^T K}{\sqrt{d}}\right) V^T W_L,$$

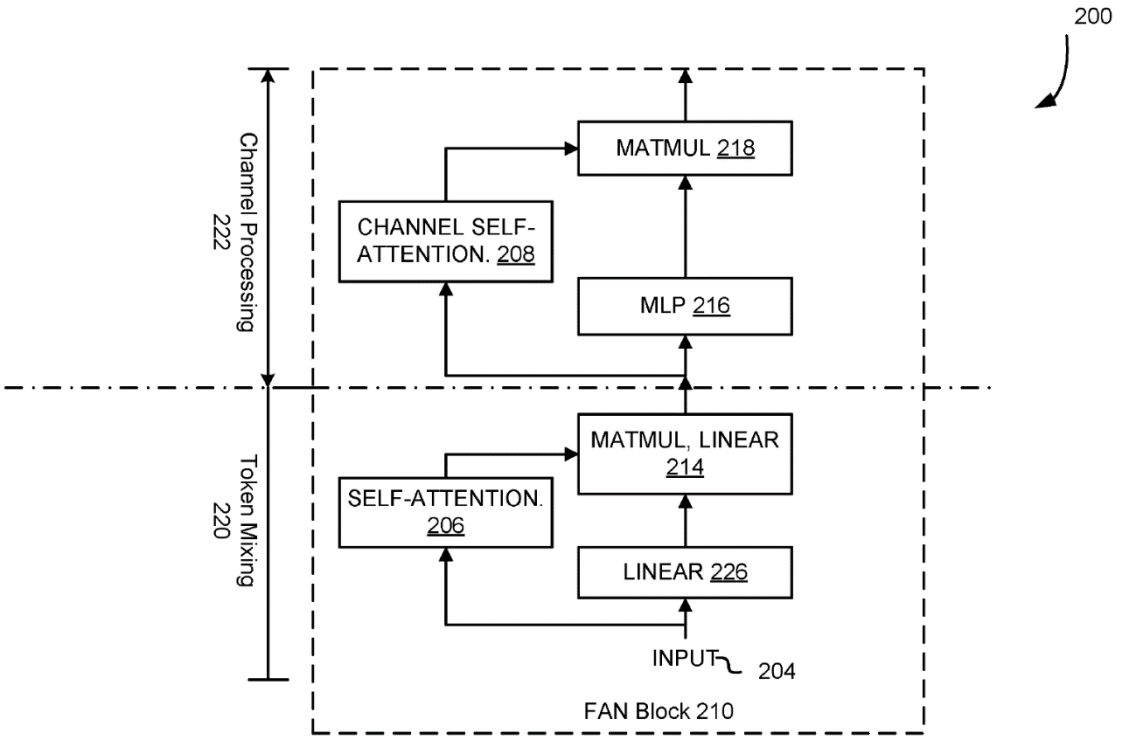
$W_L \in \mathbb{R}^{d \times d}$ は線形変換であり、$Z=[z_1, \ldots, z_n]$ は集約されたトークン特徴であり、$\sqrt{d}$ はスケーリング係数である。セルフアテンションモジュールの出力は正規化されて MLP に供給され、次のブロックへの入力を生成する。

次にチャネル処理を使用し、ViT は、MLP ブロックを採用して、入力トークンを特徴 Z に変換する。

$\mathcal{Z}=\text{MLP}(\mathcal{Z})$, (2)

FAN エンコーダ 110 などの FAN ブロックは、特徴変換を実行するために、チャネルアテンションブロック 108 などのチャネルワイズセルフアテンション（CSA）モジュールを含む。推論 112 は、FAN エンコーダ 110 の出力であってもよい。推論 112 には、入力画像の変換、オブジェクトの画像セグメンテーション、オブジェクト検出、またはオブジェクト識別を示すデータが含まれる。

図 2 は、FAN ブロックの環境例を示す。



FAN ブロック 210 では、入力トークン 204 はセルフアテンション層 206 によって集

約され、その後に線形投影 214 が続く。その後、チャネルセルフアテンションブロック 208 および多層パーセプトロン（MLP）ブロック 216 などの追加のセルフアテンション層がセルフアテンションブロック 206 に付加される。これにより、完全アテンションのチャネル処理が可能になり、従来の ViT の線形投影層 226 が除去される。

FAN ブロック 210 のセルフアテンション層の後には多層パーセプトロン（MLP）ブロックが続き、これはチャネル情報交換を可能にするが、セルフアテンション動作ではない。このセルフアテンションの影響をさらに強化し、表現学習におけるその堅牢性を活用するために、ViT に基づくモデルを完全アテンションにする。この新しいファミリーのモデルは、完全アテンションネットワーク（FAN）として説明される。

FAN ブロック 210 は、チャネル処理のためのチャネルセルフアテンションモジュール 208 を含む。FAN ブロック 210 は、MLP ブロック 216 を別のセルフアテンションモジュールに移動し、完全なアテンションチャネル処理を可能にし、アテンション行列の後の追加の射影線形層を除去する。チャネルセルフアテンションモジュール 208 は、トークン次元ではなくチャネル次元に沿って行列乗算（matmul）演算 218 を計算する。

完全アテンションネットワークは、セルフアテンションを利用してモデルの堅牢性を向上させる。これにより、セルフアテンションをより広範囲に活用してモデルの堅牢性をさらに向上させることができる。

図 2 に示すように、FAN は、トークン混合 220 にセルフアテンションを適用する。チャネル処理 222 に関して、FAN は、チャネルごとのセルフアテンション（CSA）を使用して、MLP ブロックをセルフアテンションブロックに移動することによって特徴を変換するアテンション設計を採用する。

FAN ブロックは、次のように定式化される特徴変換を実行するために、次のチャネルごとのセルフアテンション（CSA）を含む。

$$CSA(Z) = \mathrm{Softmax}\left(\frac{(W'_Q Z)(W'_K Z)^T}{\sqrt{n}}\right) MLP(Z). \tag{3}$$

ここで $W'_Q \in \mathbb{R}^{d\times d}$ 及び $W'_K \in \mathbb{R}^{d\times d}$ は線形変換パラメータである。

CSA は、トークン次元ではなくチャネル次元に沿ってアテンション行列を計算することができ、これにより特徴変換に特徴共分散（線形変換後 W'_Q, W'_K）が活用される。より大きな相関値を有する相関特徴チャネルは集約されるが、低い相関値を有する外れ

値特徴は分離される。

これは、モデルが無関係または重要でない情報を除外するのに役立つ。チャネルごとのセルフアテンションを使用することで、FAN モデルは無関係な特徴をフィルタリングし、フォアグラウンドトークンとバックグラウンドトークンのより正確なトークンクラスタリングを形成できる。

3.クレーム

135 特許のクレーム 1 は以下の通りである。

1. システムにおいて、

少なくとも 1 つのプロセッサと、

少なくとも 1 つのプロセッサによる実行に応答して、システムに少なくとも下記の処理を実行させる命令を含む少なくとも 1 つのメモリとを備え、

少なくとも 1 つのセルフアテンション部分および少なくとも 1 つのチャネルセルフアテンション部分を備える 1 つ以上のエンコーダで入力の 1 つ以上の入力トークンを受け取り、該入力トークンは入力のそれぞれの部分に対応し、

少なくとも 1 つのセルフアテンション部分において、1 つ以上の入力トークンを集約し、

1 つ以上の入力トークンの集約に少なくとも部分的に基づいて、入力トークンの 1 つ以上の特徴表現を生成し、

少なくとも 1 つのチャネルセルフアテンション部分において、入力トークンの 1 つ以上の特徴表現を集約し、チャネルセルフアテンション部分は、入力のチャネル次元に沿ってアテンション行列を計算し、1 つまたは複数の特徴表現は、閾値を超える特徴チャネル間の相関値に少なくとも部分的に基づいて集約され、

集約された 1 つ以上の特徴表現に少なくとも部分的に基づいて、入力に関する推論を生成し、

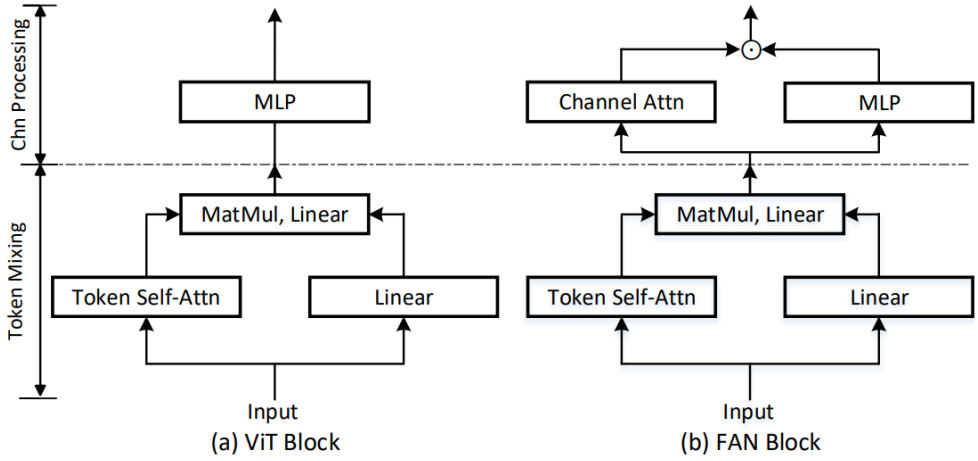
少なくとも部分的に推論に基づいて出力を生成する。

4. 本特許に関連する論文

本特許に関する論文“Understanding The Robustness in Vision Transformers”[1]が、NVIDIA の Daquan Zhou 氏らにより公表されている。

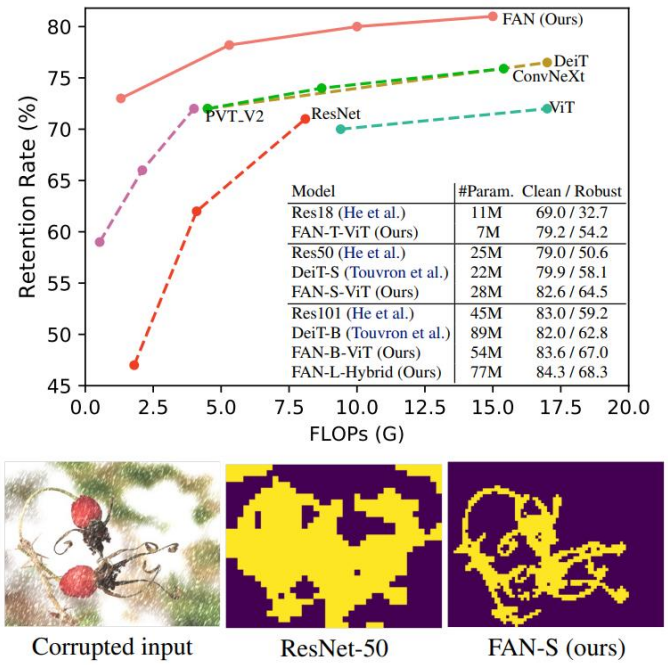
[1] Daquan Zhou et al. “Understanding The Robustness in Vision Transformers” arXiv:2204.12451v4 [cs.CV] 8 Nov 2022

下記図は従来の ViT ブロックと提案 FAN ブロックの比較である。



(a) の ViT block では、入力トークンはまずセルフアテンションによって集約され、続いて線形投影が行われ、特徴変換のために MLP がセルフアテンションブロックに追加される。 (b)の FAN block では、トークンセルフアテンションとチャネルアテンションの両方が適用されるため、ネットワーク全体が完全にアテンションになる。チャネルアテンション後の線形投影層は除去される。

下記図は ImageNet-C の主な結果（上の図）とクラスタリングの視覚化（下の行)とを示す。



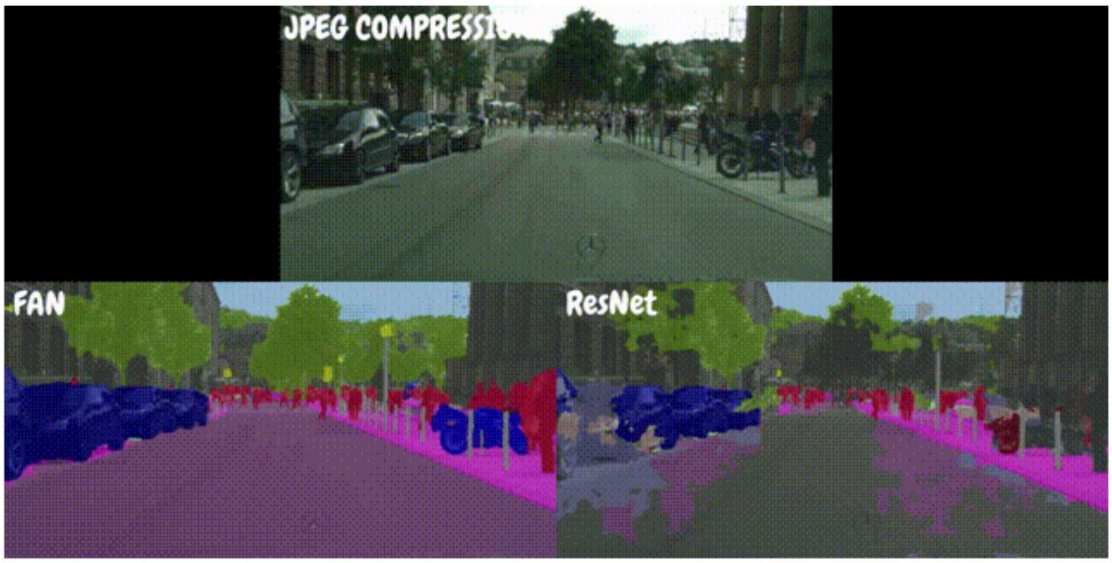
Model	#Param.	Clean / Robust
Res18 (He et al.)	11M	69.0 / 32.7
FAN-T-ViT (Ours)	7M	79.2 / 54.2
Res50 (He et al.)	25M	79.0 / 50.6
DeiT-S (Touvron et al.)	22M	79.9 / 58.1
FAN-S-ViT (Ours)	28M	82.6 / 64.5
Res101 (He et al.)	45M	83.0 / 59.2
DeiT-B (Touvron et al.)	89M	82.0 / 62.8
FAN-B-ViT (Ours)	54M	83.6 / 67.0
FAN-L-Hybrid (Ours)	77M	84.3 / 68.3

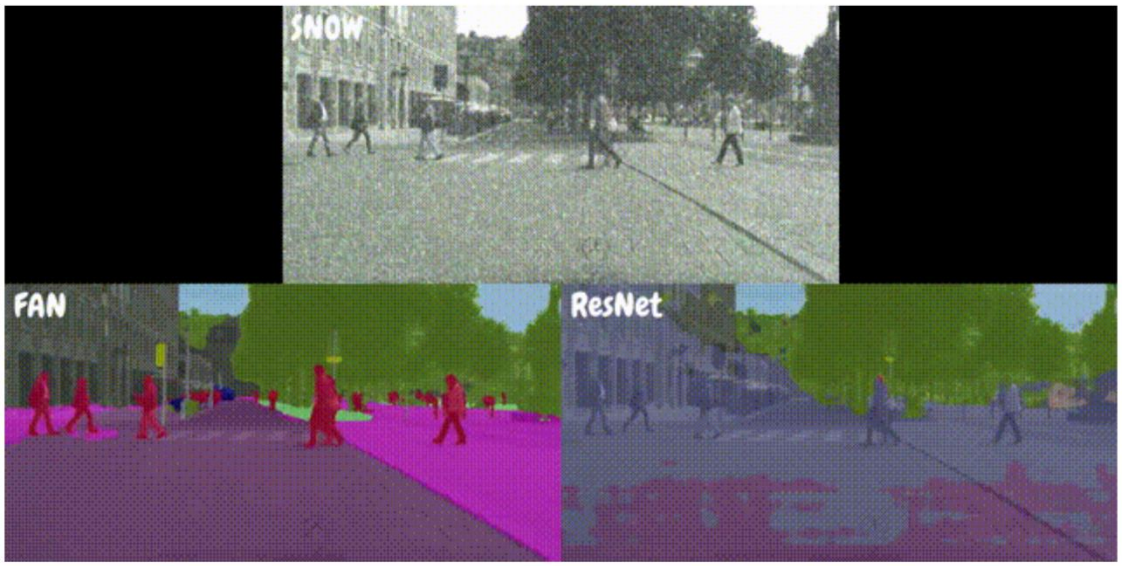
Corrupted input　ResNet-50　FAN-S (ours)

上記グラフの縦軸の保持率(Retention)は、堅牢精度/クリーン精度(robust accuracy / clean accuracy)として定義される。FAN の保持率が他のモデルよりも上回っていることが理解できる。

下の図は左から右にかけて、 破損（雪）によって汚染された入力イメージと ResNet 及び FAN による可視化されたクラスタを示している。視覚化は、最後から 2 番目のレイヤーの出力特徴（トークン)に対して実行された。本論文の FAN の方が破損に対する堅牢性が高いことが視認できる。なお、すべてのモデルは ImageNet-1K で事前トレーニングされている。

NVIDIA は自動運転車両の画像認識に本ロバスト ViT を使用した。下記図は FAN と ResNet との比較である。





FAN は、ResNet と比較して路面上のノイズ、及び、画像全体に付加されたノイズに対して堅牢であることが示されている。ロバスト ViT のコードは以下から入手できる。
https://github.com/NVlabs/FAN

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。