社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

[ポスター講演] 特徴埋め込みを持つ Vision Transformer を用いたディープフェイク検知に関する研究

桑原　聡太郎†　　姜　　玄 浩††

† 東京工業高等専門学校電気電子工学専攻 †† 東京工業高等専門学校電子工学科　〒193-0997　東京都八王子市椚田町 1220–2
E-mail: †s18059@tokyo.kosen-ac.jp, ††kang@tokyo-ct.ac.jp

あらまし　近年の画像生成技術の向上により，芸術作品の生成や，リアルなスタントシーンの合成などができるようになった．一方で，Deep Fake を利用して顔認証を不正突破したり，他人の顔を使って嘘を拡散できるようになった．著者らの前研究では Deep Fake を今まで以上により高精度に検出するために，Vision Transformer(ViT) を応用し検証をした．論文中では，ViT をそのままの形で学習させた時の結果として 82% の検出精度を観測することもあったが，何度も計測していると数値があまり安定しないことがあった．そこで本論文では，まず前回の追実験を行いより状況を詳細に把握し，その後前回の反省を踏まえた改良点を検証する．

キーワード　Vision Transformer, Transformer, Deep Fake

Deep Fake Detection Using Vision Transformer with Feature Embedding

Sotaro KUWAHARA† and Hyunho KANG††

† National Institute of Technology ,Tokyo College,Department of Advanced Course of Electrical and Electronic Engineering †† Department of Electronic Engineering　1220–2 Kunugida, Hatchiozi-si, Tokyo, 105–0123 Japan
E-mail: †s18059@tokyo.kosen-ac.jp, ††kang@tokyo-ct.ac.jp

Abstract　Recent improvements in image generation technology have made it possible to generate works of art and composite realistic stunt scenes. On the other hand, Deep Fake can now be used to fraudulently break facial recognition and spread lies using other people's faces. In our previous study, we applied the Vision Transformer (ViT) to detect Deep Fake with higher accuracy than ever before. In the paper, the detection accuracy of 82% was observed when ViT was trained in its original form, but the values were not very stable after repeated measurements. In this study, we first follow up on the previous experiment to understand the situation in more detail, and then verify improvements based on the previous reflection.

Key words　Vision Transformer, Transformer, Deep Fake

1. ま え が き

年々発達する Deep Fake 技術は，我々に多大な恩恵と危険性をもたらした．よい活用例としては，映画を作る際に自然な形での吹き替えを作ったり，スタントマンを使用せずに危険なシーンの製作をしたりなどがある. 一方で悪用例としては，それを認証の不正突破や，悪意ある虚偽動画の作成に悪用する人間もいる．技術の活用方法は使用者の倫理に託されているので, 私が一方的に悪を叫ぶことはできないが, 問題はそのような虚偽情報を受け取ってしまう，もしくは自身の顔認証を不正に突破されるというリスクは身近であり，尚且つ起こる被害がただ事では済まないレベルであるということだ.

そのような問題を受け，昨年の著者らの研究 [1] ではそのリスクを少しでも防ぐために Deep Fake の検出に Vision Transformer（以下 ViT）[2] [3] を用いてより高精度な検出器を作成することができる可能性について論じた．当論文では 2019 年の Deep Fake Detection Challenge(以下 DFDC) [4] を対象として検証を行い，約 20% の精度の改善が見られた. しかし一方で，ほとんど学習が進まなかったり，精度が実験のたびに大きく変化することがあり，その研究結果の信頼性は高いとは言えない.

そこで本論文では，前回の研究の追実験を行い，まずは数値がどのくらい安定するのか（もしくは安定しないのか）を検証する．またそのうえで，昨年度の研究の考察で述べた ViT の埋め込み層に対する画像特徴の埋め込みに関して検証を行う．具体的には，VGG16 などをはじめとした CNN ベースのすでに学習されたモデルを用いて出力されたベクトルを Transformer

Encoder [5] の入力として学習させるという事である．

本論文の構成としては，第 2 章で背景や関連する研究について説明し，第 3 章で実験とその結果，第 4 章で展望と全体のまとめを述べる．

2. 原理・基礎的知見

2.1 転 移 学 習

転移学習（Transfer Learning）とは，機械学習分野における学習手法の一つで，ある領域で学習したモデルを別の領域に適用させる学習技術のことである．これは，元ドメインと目標ドメインの間に効率的に知識を再利用することで，目標ドメインにおける学習効率 (計算量，精度ともに) を上げることを目的として行われる．転移学習は主にデータセットが大きくなりがちな自然言語処理分野などでよく活用されるが，その他にも学習時間の短縮のため，もしくはデータセットの少なさを補うためにも用いられる。基本的に転移学習は大きなデータセットで先に学習したモデルを，小さなデータセットでの学習に転用する．その際に先に行われる学習のことを事前学習と言い，のちに行われる学習のことをファインチューニングと呼ぶ．転移学習には大まかに分けて 3 種類ある．以下の小節では転移学習の種類について述べる．

2.1.1 帰納転移学習

帰納転移学習（Inductive Transfer Learning）は元ドメインと目標ドメインの両者にラベル付きデータがある場合を想定したものである．帰納転移学習では，目標ドメインのデータが少ない時に，元ドメインのラベル付きデータを学習に用いることで予測精度の向上を図る．通常一般的に用いられる転移学習はこの帰納転移学習を指す．また，帰納転移学習を行う場合にも用法がいくつか存在し，元モデルを再学習するだけのもの，元モデルは再学習せずにさらに積層するもの，そして追加積層したうえで元モデルも追加層も両方学習する用法がある．

2.1.2 変換的転移学習

変換的転移学習は元ドメインにラベル付きデータがあるのに対して，目標ドメインにラベル無しのデータがある場合を想定している．本手法では，ラベルのないデータに対して元ドメインのデータから適切なラベルを予測する．

2.1.3 教師無し転移学習

教師無し転移学習（Unsupervised Transfer Learning）は，元ドメインと目標ドメインの両者にラベルがないデータが存在することを想定している．本手法では，目標ドメインのデータに対し，元ドメインのデータを用いてクラスタリングや次元削減などの教師無し学習を行う．

2.2 Transformer の画像処理応用

ViT は自然言語処理分野で発達した Transformer というモデルを画像分野に応用したものである．この技術のすごいところは，元となった Transformer のモデルに対してほとんど変更を加えず画像分野に応用したことにある．その方法としては，入力画像を任意のサイズにパッチ分割し，分割したパッチをベクトル化することで Transformer のエンコーダに入力している．自然言語処理分野では単語がベクトル，文章が行列となるようにとらている．そこで ViT ではパッチを単語，画像全体を文章としてとらえることで Transformer への適用を可能とした．ただし，それだけだと画像の位置情報が失われてしまうので，位置情報の埋め込みを行う．以下図 1 にパッチの処理の流れを示す．

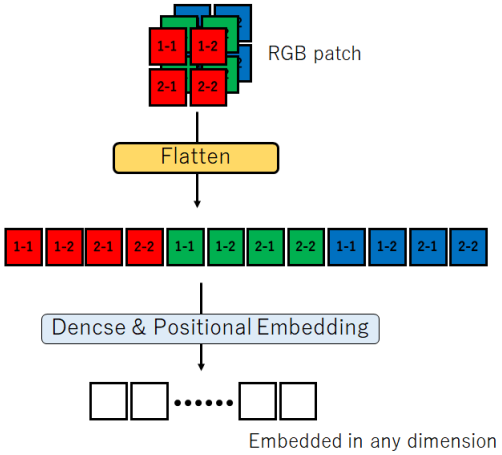

図 1 パッチのベクトル化

3. 実　　験

実験は Kaggle の DFDC で公開された喋っている人が映っている 400 本の動画データセットを用いて ViT を学習させ，それによる Fake 画像の検知精度を検証する．

3.1 前研究の追実験

著者らの前研究において実験結果が安定しないことがあったので，まずはその件についてさらに実験を行い検証する．また，実験のための事前処理済みのデータセットに関しては，昨年度使用したものを利用する．まず初めに，前回最も精度が良かった条件で 5 回検証を行った．実験結果を表 1 に示す．

上記の実行結果より，確かに数値が全く安定していないことが分かった．次に，前回は実験に大きく関与していたドロップアウト率を変化させて実験を行った．追加で 2 パターンの実験を各パターン 5 回ずつ実行した．結果は最大精度と最低精度の差が両方とも 64.7% となった．そのため，ドロップアウト率がどうであれど数値が安定しないことが分かる．最後に，ハイパーパラメータを組み替えて実験を行った．実験では毎回同条件下で 5 回行い，その際の最大値と最小値の差を記録した．結果として，学習率以外（mlp のヘッド数，transformer エンコードのレイヤ数含む）はすべて実験結果に変化さえ起こさなかった．

表 1 最高精度条件での実行

Table 1 Execution under highest precision conditions

実行精度 [%]	最終損失	学習したエポック数 [回]
69.6	2.72	23
17.7	2.72	23
46.5	2.72	23
46.0	2.72	23
82.35	2.72	23

表 2　学習率を変えた際の精度

Table 2　Accuracy when changing learning rate

学習率	実行精度 [%]	最終損失	学習したエポック数 [回]
0.001	69.6	2.72	23
	17.65	2.72	23
	46.51	2.72	23
	45.96	2.72	23
	82.35	2.72	23
0.002	81.5	2.78	21
	81.5	2.78	21
	81.5	2.78	21
	81.5	2.78	21
0.003	82.4	2.78	21
	82.4	2.78	21
	82.4	2.78	21
	82.4	2.78	21
	82.4	2.78	21
0.005	82.4	12.5	21
	82.4	12.5	21
	82.4	12.5	21
	82.4	12.5	21
	82.4	12.5	21
0.007	82.4	12.5	21
	82.4	12.5	21
	17.6	12.5	21
	17.6	12.5	21
	17.6	12.5	21
0.01	17.6	12.5	21
	17.6	12.5	21
	82.4	12.5	21
	17.6	12.5	21
	82.4	12.5	21
0.02	82.4	12.5	21
	17.6	12.5	21
	82.4	12.5	21
	82.4	12.5	21
	17.6	12.5	21

表 3　学習率を変えた際の平均値と標準偏差

Table 3　Average value and standard deviation when changing learning rate

学習率	平均精度 [%]	標準偏差	最大-最小誤差 [%]
0.001	52.4	24.9	64.7
0.002	81.5	0.00	0.00
0.003	82.4	0.00	0.00
0.005	82.4	0.00	0.00
0.007	43.5	35.4	64.7
0.01	43.5	35.4	64.7
0.02	56.46	35.4	64.7

しかし，学習率を変えたときはデータのばらつき方が変わった．以下の表 2 及び表 3 に，学習率ごとの実験結果を示す．

上記の実験結果より，学習率が 0.002 から 0.005 までの範囲では数値がかなり安定しており，そのほかの範囲（特に 0.007 以上）においては実験結果が著しく変化しばらつくことが分かった．

3.2　データセットの修正

昨年度の研究では，データセットのクラスの偏りがあることは学習率に大きな影響を表していると考察した．実際，使用しているデータセットの約 8 割の動画は FAKE ラベルが指定されており、リアルの動画は 80 本ほどしかない．そこで，データセットの事前処理を一部改変して実験を行う．具体的には，動画からフレームを切り出す際の枚数を FAKE 動画では少なく，REAL 動画では多くなるように調整した．以下に事前処理を一部改良した際の実験結果を示す．尚，実験の際の条件は，著者らの前研究の最高精度条件で学習率のみ 0.001 として実行している．実験結果は，平均値が 43.5[%]，標準偏差が 9.33 だった．平均値が下がったもののばらつきが減った．

表 4　学習率を変えた際の平均値と標準偏差

Table 4　Average value and standard deviation when changing learning rate

精度 [%]	損失	実行エポック数
39.4	9.50	21
39.4	9.50	21
60.6	9.50	21
39.4	9.50	21
39.4	9.50	21

3.3　特徴埋め込みの適用

Deep Fake 動画をはじめとした生成された動画や画像は，人間がそれを目的のものと認識させようとして生成されるため，そのため，大まかな形状よりも局所的な特徴にこそ生成された動画特有の特徴があると考えた．

そこで，ViT に VGG などの学習済みモデルによる特徴埋め込みを用いた転移学習を行い実験を行う．画像の特徴抽出には，ResNet50，DenceNet, Inception, MobileNet, VGG16 の 5 つの学習済みモデルを用いた．また，これらはいずれも ImageNet によって学習されたモデルである．本来は画像のすべてのパッチから特徴抽出を行う予定であったが，著者の技術不足と時間的な制約によりできなかった．そのため，今回は画像をそのまま特徴抽出に通し，出力される 1000 次元のベクトルを（10, 100）型に分解して ViT の入力として用いた．以下の図 2 に全体の処理概要を示す．

また本実験では，先ほど用いたデータセットのラベル比を修正したデータセットに対しても同様に実験を行った．今回は，各条件下で 10 回実験を行い，それらの実験結果をまとめた．具体的には，毎回の実験による最終的な精度，損失，学習にかかったエポック数について測定した．測定結果のうち，各条件下における平均精度，標準偏差，最高精度を以下の表 5，6 にまとめた．精度は安定したが，いまだに学習が不安定だった．以下に DenceNet を特徴埋め込みに用いた際の学習曲線を示す．他の実験結果も参照すると，学習そのものは安定しておらず，またあまり学習が進んでいないという結果になった．

表 5　特徴埋め込みありの ViT の実行結果

Table 5　Execution results of ViT with feature embeddinge

使用した特徴抽出	平均精度 [%]	精度標準偏差	最高精度 [%]
DenceNet	82.6	0	82.6
Inception	82.6	0	82.6
MobileNet	82.6	0	82.6
ResNet	82.6	0	82.6
VGG	82.6	0	82.6

表 6　特徴埋め込みありの ViT の実行結果（バランス補正あり）

Table 6　Execution results of ViT with feature embeddinge (with balance corrected data set)

使用した特徴抽出	平均精度 [%]	精度標準偏差	最高精度 [%]
DenceNet	59.0	0	59.0
Inception	59.0	0	59.0
MobileNet	59.0	0	59.0
ResNet	59.0	0	59.0
VGG	59.0	0	59.0

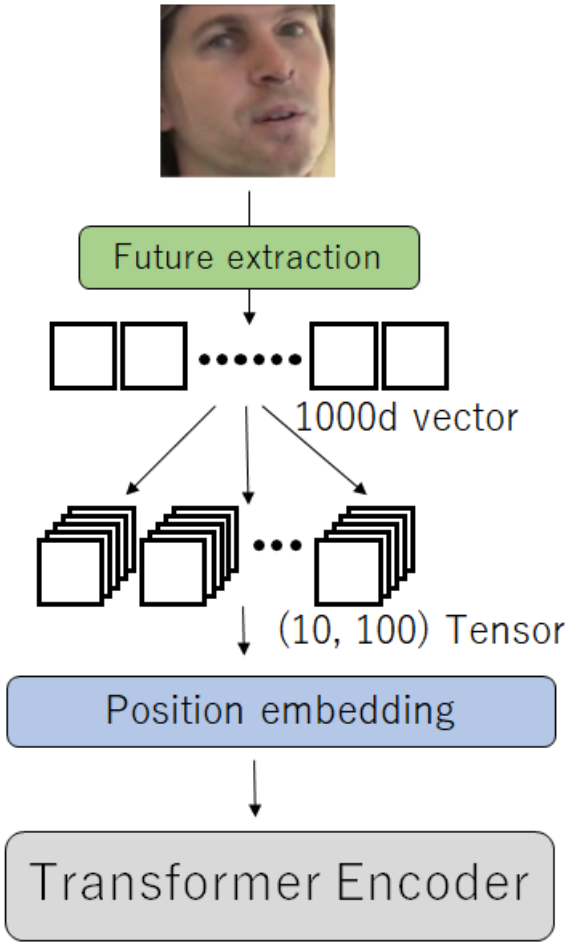

図 2　特徴埋め込み付き ViT

Fig. 2　ViT with feature embedding

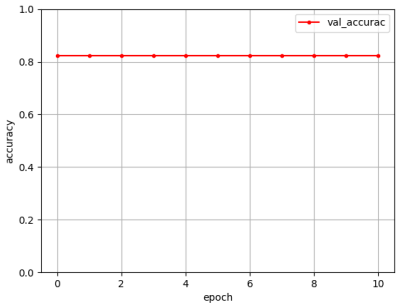

図 3　ラベル比をそろえたデータセットでの実験精度

Fig. 3　Experimental accuracy on datasets with unarranged label ratios

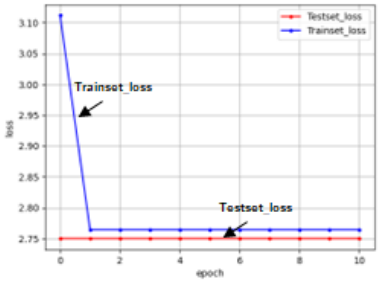

図 4　ラベル比をそろえてないデータセットでの損失推移

Fig. 4　Loss transition in a dataset with the unarranged label ratio

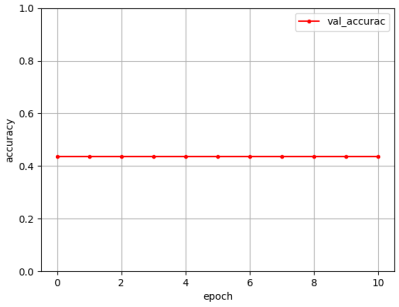

図 5　ラベル比をそろえたデータセットでの実験精度

Fig. 5　Experimental accuracy on datasets with uniform label ratios

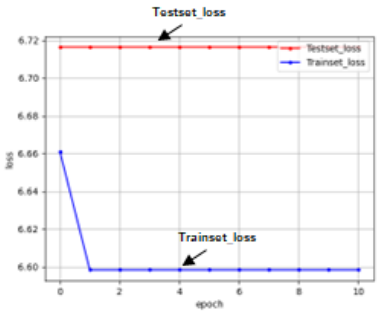

図 6　ラベル比をそろえたデータセットでの損失推移

Fig. 6　Loss transition in a dataset with the same label ratio

4.　考察および展望

4.1　結 果 考 察

まず初めに著者らの前研究の追実験について述べる．まず，

前回の最高精度条件下での実験で数値がばらついていたことと，そのほかの条件下でほぼ同様の最高精度を測定できていたことより，前回の実験結果はたまたまそのような計測結果が出たという可能性が非常に高い．しかし一方で，学習率が 0.002 から 0.005 までの範囲では数値が安定していたので，最高精度 82.4% という数値に関してはある程度信頼がおけると考えられる．

次に，データセットのラベル比率に関して考察する．前回研究ではデータセットの比率により学習が安定しないという考察をした．そして，今回の実験では標準偏差が少々下がったことから，ばらつきの抑制にはそれなりに効果があると考えられる．しかし，抑えられたといっても最大精度と最低精度の差が 30% 近くあるので，原因や修正できる点がまだ存在すると考えられる．また，3 章 2 節，3 節でラベルの比率を修正したデータセットによる結果が，比率調整を行っていない実験の結果に比べて平均精度が下回っていることが分かる．これは，REAL データ（少ないほうのラベル）を増やすためにより多くのフレーム数を取得したことが原因であると考えられる．実際，著者らの前研究の論文では平均精度を見たときに，最も数値が良かったのは取得フレーム数が 1 のデータセットで学習したモデルであったことが分かる．よって，同じ動画から何度もフレームを取得すると，過学習を起こし実験結果に学習精度に悪影響を与えてしまうと考えられる．また，この考察に関しての検証のために，FAKE 動画の数を減らして実験を行うなどの検証をする必要がある．

次に，特徴埋め込みをしたモデルに関して考察する．第 3 章 3 節の実験結果では，すべての条件下で標準偏差が 0 であった．これにより，画像の特徴の埋め込みは Deep Fake の画像（もしくは動画の 1 フレーム）の検知に関して，学習精度のばらつきを抑えるのに有効である可能性が高い．また，実験結果では最高精度はどれも変わらなかったことから，これらが精度の改善に対しては有効である可能性は低いと考えられる．またこれらの事実をもとに，ViT 単体では局地的な形状特徴に関する理解のための能力が不安定であると考察できる．そもそも Deep Fake は人間にそれっぽく認識させることを目的として作られているので，広い視野での形状的な特徴を理解することに特化している ViT はそもそも Deep Fake の検出に向いていないのかもしれない．また，特徴抽出器を変えても精度が一貫して変わらなかったことから，少なくとも画像の特徴抽出と ViT を組み合わせた Deep Fake 検知のためのモデルでは，MobileNet と ViT の組み合わせが最も効率が良いということが分かる，

4.2 実験の反省点

今回の研究のための実験には，以下の点に関して検証を行うことができなかった．

（1） パッチごとに特徴抽出を行う

（2） 色空間の特徴の利用

（3） 時間的な特徴の利用

（4） Transformer エンコーダのレイヤ数，埋め込みの次元，mlp のヘッド数などのハイパーパラメータの検証に時間がかかりすぎる範囲に関して

（5） 特徴埋め込みを ViT に含めた際の，ハイパーパラメータの調整

（6） 他のデータセットを使用しての検証

パッチごとの特徴抽出に関しては，Tensorflow に機能を埋め込む方法が分からず，事前処理の一部として行おうとしたが，実行時間の予測が約 2 週間ほどになってしまうので行わなかった．しかし，そもそも Transformer の強みである位置情報とパーツの位置の関係性の学習という強味をつぶしてしまっている．これについて検証をすることができれば，今回の検証結果と比較することで ViT の広域的な特徴抽出が Deep Fake 動画の抽出に関して有効であるかどうかに関して確認することができる．また，パッチごとの特徴抽出ができれば精度の改善ができる可能性もあるだろう．色空間や時間的特徴は前回論文中でも触れたように，先行研究で生成画像の検知精度に大きく関与することが示されているので，検証するべきだった．また，特徴埋め込みを ViT と組み合わせた際のハイパーパラメータの調整に関しては，モデルを作り直した影響で最適なパラメータが変わっている可能性があるので，調整によって精度を改善できる可能性が高い．

4.3 あ と が き

本研究では，前回の研究内容に関しての検証並びに新しい提案手法についての検証を行た．結果としては，今までできなかった安定した数値の測定ができるようになった．それによって，前回提示した最高測定精度がかなり安定して提示できるようになった，一方で，いまだに学習の過程が安定しない，過学習が起きる，あまり学習が進まないという問題が起きており，解決策は以前模索中である．本研究は，ViT を用いて Deep Fake 動画を検知するためのスタートラインに立ったに過ぎない．そのため，今後精度改善と学習力の向上のためまだまだ研鑽を続ける必要がある．

謝辞 謝辞

本研究をするにあたって，容量が悪く行動の遅い私を根気強く導いてくださり，また論文当提出物の校正をしてくださった国立東京工業高等専門学校電子工学科姜玄浩准教授に深く感謝申し上げる．また，親身になり的確なアドバイスをくださった同研究室の小林佐助先輩並びに，協力と助言いただいた同研究室メンバーの方々に感謝の意を表する．

文　　献

[1] 姜玄浩桑原聡太郎，“Vision transformer を用いた deep fake 検出に関する研究，” 信概福 vol.122，no.412，pp.61–65，2023.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” arXiv preprint arXiv:2010.11929, pp.••–••, 2020.

[3] 徳永匡臣 箕浦大晃 Q Yue 品川政太朗片岡裕雄，Vision Transforemr 入門，技術評論社，2022.

[4] Kaggle, “Deepfake detection challenge ‖ kaggle,” 2023. https://www.kaggle.co m/c/deepfake-detection-challenge

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems, vol.30, pp.••–••, 2017.