6C-06

特徴重要度の評価を用いたViTによる表情認識手法の提案

呉強　　浜田宏一
帝京大学大学院理工学研究科

1. はじめに

表情認識は，コンピュータビジョンの重要な研究方向の1つとして，既に広く日常生活に応用されている．例えば，運転手の顔表情によって居眠り運転を防ぐ．

しかし，深層学習を用いた顔表情認識には，まだ多くの問題が残っている．特に大規模なデータセットには，モデルの学習に影響を与えるラベルノイズが存在する．ラベルノイズは，曖昧な表情に付与された誤ったラベルである．

本研究では，このラベルノイズに対する頑強性を高める目的で，表情画像の特徴重要度を評価できる Vision Transformer（ViT）[1]表情認識モデルを提案する．そして，モデルの有効性を検証するため，顔表情データセット（RAF-DB, FERPlus）を使用して評価を行った．

2. ViT モデル

ViT モデルは，自然言語処理に用いられるTransformerのモデルをベースに構築された画像分類モデルであり，従来の CNN をベースとした手法と比較して，画像の分類と認識の精度を大幅に向上させた．

近年の研究により，ViT モデルは多くの他の研究分野にも応用されている．本研究では，ViT モデルを改善し，表情認識タスクの認識精度を向上させることを目的としている．

3. 提案手法

深層学習の手法で表情を認識する場合は，データセットを使用することが必要である．データ量の少ないデータセットを使用すると，タスクに効果的な特徴をモデルが十分に学習できず，過学習の原因にもなる．一方，大規模なデータセットを用いれば，モデルは有効な特徴をより適切に学習することができる．

しかし，大規模なデータセットでは，誤ったラベルが付与されたデータも存在する．例えば，顔の表情データセット内の曖昧な表情画像などが挙げられる．そこで，Wang ら[2]は畳み込みニューラルネットワークに特徴重要度を追加することで，この問題を解決する Self-Cure Network（SCN）モデルを提案した．我々は，この SCN モデルで提案された特徴重要度が ViT モデルにも適用できるのではないかと考え，ViT モデルに特徴重要度の評価機能を追加する検討を行った．

図1に，提案手法の詳細を示す．各表情画像の重要度を計算するステージと，ViT モデルを用いて表情認識を行う2つのステージから構成されており，これらは学習中に同時に実行される．

最初のステージでは，各表情画像の重要度を計算する．まず，特徴抽出器を利用して，各表情画像の特徴量を抽出する．ここで利用される特徴抽出器は CNN である．得られた特徴を特徴重要度の計算モジュール Attention Importance Weighting Module（AIWM）に入力する．AIWM は，シグモイド関数を用いて，各特徴の重要度重みを計算することができる．計算式を式（1）に定義する[2].

$$\lambda_i = \sigma\left(W_\lambda^{\mathrm{T}} x_i\right), \qquad (1)$$

λ_iは特徴重要度，W_λ^{T}は学習できるパラメータ，σはシグモイド関数，x_iは入力された表情画像の特徴を表している．曖昧な表情画像は式（1）によって得られた特徴重要度が低くなる．特徴重要度の低い曖昧な表情画像をより正確に見つけるために，重要度の調整モジュール Self-Adjustment Module（SAM）を設計し，重要度によって表情画像を分ける．

次のステージでは，最初のステージで得られた特徴重要度を使用し，ViT モデルで表情認識を行う．特徴重要度を最大限に活用するために，提案手法には SCN の Relabeling Module（RM）を導入する．RM は，モデルが一定回数で学習した後，特徴重要度の低い曖昧な表情画像のラベルを修正することができる．

表1 データセットの詳細

Dataset	Image size	Number	Classes
RAF-DB	100×100	15339	7
FERPlus	48×48	28127	7

Proposal of Facial Expression Recognition Method Using ViT with Feature Importance Evaluation

†Qiang Wu, Kouichi Hamada, Graduate School of Science and Engineering, Teikyo University

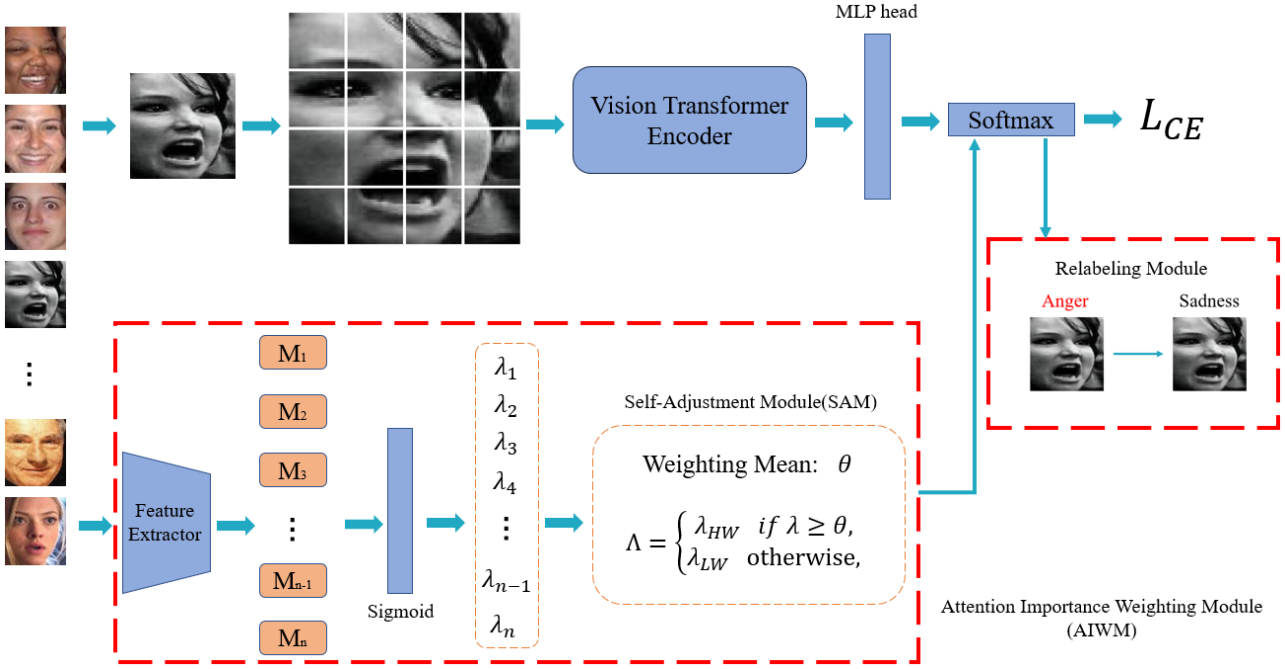

図 1 提案手法の概要

4. データセット

本研究では，2つのデータセットRAF-DB[3], FERPlus[4]に対して実験を行った．データセットの詳細は表 1 に示す．

5. 実験結果と考察

表 2 RAF-DB の表情認識結果

手法	精度(%)
ResNet-18(Baseline)	84.20
ViT(Baseline)	85.95
SCN[2]	87.03
提案手法	**87.16**

表 2 に，RAF-DB に対して行った表情認識結果を示す．特徴重要度を使用すると，畳み込みニューラルネットワークとViTモデルの性能がある程度向上していることが分る．さらに，特徴重要度をViTモデルに組み合わせることで，ViTモデルの性能が 1.21%向上するだけでなく，SCNモデルを上回っていることが分かる．

表 3 に，FERPlus を用いた実験結果を示す．

表 3 FERPlus の表情認識結果

手法	精度(%)
ResNet-18(Baseline)	86.80
ViT(Baseline)	87.35
SCN[2]	88.01
提案手法	**89.42**

表 3 によれば，提案手法の精度はベースラインの精度よりも高く，2.07%向上した．また，RAF-DB と比較すると，FERPlus に対する提案手法のパフォーマンスは SCN よりも優れていることが分かる．FERPlus がより複雑で，より多くの曖昧な表情画像が含まれているためであると考えられる．提案手法はより複雑なデータセットに対して畳み込みニューラルネットワークよりも優れたパフォーマンスを発揮することが分かった．最後に，表 2 と表 3 の結果を見ると，画像の特徴重要度を ViT モデルに適用可能であり，ViT モデルが曖昧な表情画像の特徴を学習するのを防げることが分かる．

6. おわりに

本論文では，2 つのデータセット（FERPlus, RAF-DB）対して、表情認識を行い、画像の特徴重要度をViTモデルに適用した．今後は，モデルの最適化と，顔表情認識アプリの開発などを検討していく．

参考文献

[1] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." In ICLR2021.
[2] Wang Kai, et al. "Suppressing uncertainties for large-scale facial expression recognition." In CVPR2020, pp. 6897–6906.
[3] Li Shan, et al. "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild." In CVPR2017, pp. 2852-2861.
[4] Barsoum Emad, et al. "Training deep networks for facial expression recognition with crowd-sourced label distribution." In Proc. 18th ACM Int. Conf. Multimodal Interact, pp. 279–283. 2016.