令和 2 年 12 月 22 日 (火)　　伊藤　薫

Vision Transformer によるだまし絵認識の提案および評価

1　はじめに

近年，深層学習によって画像認識の分野は急速な発展を遂げている．単純な物体認識では既に人を凌駕する成果が報告されている一方で，人間の感性や認知に関わる分野においては深層学習を用いても十分な学習ができないという課題が報告されている．これは，人間の感性や認知といった唯一の正解が存在しない抽象的な概念を定量的に評価することは現在の人工知能の枠組みでは難しいためである．このため，従来の深層学習研究では，画像分類や物体検出など，答えが一意に定まる問題が主として扱われ，だまし絵認識のような唯一の正解を定義しにくい問題については十分な研究がなされてこなかった．そこで今回は複数の意味解釈をもつだまし絵のような画像を対象とする．だまし絵ではないが，意図せずにそのような性質を持つ可能性が考えられるため，以降本研究では複数の意味解釈が可能な図形を多義図形と定義する．

そこで本研究では，人の視覚認知の多様性を計算機に理解させることを目的とし，計算機による多義図形の識別に必要な実験の枠組みおよび計算機による識別手法を提案する．また，実際のだまし絵を用いた数値実験によって提案手法の有効性を示す．

2　要素技術

2.1　Vision Transformer

Vision Transformer (ViT) とは一般的な Transformer をそのままの形で画像に適用し，かつ高い精度を達成することを目的とする手法である．実際に Convolutional Neural Network (CNN) を完全に排除した状態で高い精度を達成する成果が報告されている．ViT の特徴として二次元の画像を用いるために画像パッチを単語のように変換してから入力する．

3　従来研究

4　提案手法

4.1　深層学習の構成

4.2　数値実験の構成

4.3　データセット

本研究ではインターネットから多義図形として解釈できるだまし絵画像を収集し，データセットを作成した．

実験では, 風景と人の顔をモチーフにした多義図形が多く，また多義図形と似た肖像画や風景画が多く存在することに着目した．そこで, 風景と人の顔の多義図形については，著者が風景と人の顔であると判断した画像を集めて作成したデータセットを 「多義図形」クラスとした．また, 一義図形のデータセットとして, WikiArt [**?**] 中の “landscape”, “cityscape” ラベルの画像を「風景画」クラス, “portrait” ラベルの画像を「肖像画」クラスとして用いた．

5　数値実験

6　結果と考察

7　おわりに