Vision Transformer によるだまし絵の認識および生成

1 はじめに

近年，深層学習の登場により, 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は困難である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に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した人工知能による画像認識について検討する．また, 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．

以上本研究では, 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．最終的には識別の発展的応用である多義図形の自動生成を目標とし, そのために GAN の技術を用いた実験によって多義図形生成における GAN の有効性を示す．

2 要素技術

2.1 Vision Transformer

Vision Transformer (ViT) [1] は画像処理で一般的な CNN を利用せずに, 自然言語処理で有効であった Transformer をそのままの形で画像に適用する手法である．二次元の画像データを分割し, 各データを単語のように扱うことで Transformer に適用することを可能としている．事前学習データセットとモデルを大きくすることで性能が向上することが報告されている．ViT-B_16 は，“ImageNet21k ”と呼ばれる大規模データセットで事前学習された ViT モデルである．

2.2 LSGAN

Generative Adversarial Networks (GAN) [2] は Generator (生成器) と Discriminator (識別器) という 2 つのネットワークを対立的に学習させる生成モデルである．実データとの類似性と実用上の新規性が高いデータを生成可能という点で注目を集めるモデルである．しかし, GAN の損失関数は学習中に消失勾配問題を引き起こす可能性があることが知られている．

Least Squares Generative Adversarial Networks (LSGAN) [3] は 消失勾配問題を克服するために損失関数として最小二乗損失関数を適用した GAN である．LSGAN は通常の GAN と比べて高品質の画像を生成することができ, 学習過程において安定して動作することが知られているため, 本研究では LSGAN を用いて画像生成をした．

3 従来研究

だまし絵の一種に多義図形がある．多義図形とは, 人の視覚系によって 2 通り以上に解釈される図形である．本研究では画像中に存在する各オブジェクトのラベルが一意に定まるものを一義図形, だまし絵のようにラベルとして複数の解釈が可能なものを多義図形と定義する．認知科学の分野において, 多義図形の解釈に影響を与える要因は注視点や選択的注意とされており図形の解釈過程は山村ら [4] によって報告されている．多義図形の解釈について, 計算機を用いた手法は堀江ら [5][6] によって報告されている．しかしながら従来研究では CNN による実験しかなされていない．また, 多義図形の自動生成に関する研究もなされていない．

3.1 深層学習の構成

本研究では, 既存の深層学習手法を用いて, 多義図形を理解するために必要な問題の枠組みについて提案し, 具体的な実験方法に基づいて実験した．

本研究では, 計算機が多義図形を理解するとはどのようなことであるかを示すために, ViT を用いた多義図形を理解する手法および実験の枠組みについて提案する．データセットに対して Data Augmentation を使用し, ViT-B_16 の転移学習を適用し, Optuna [7] を用いて学習率とドロップアウト率のパラメータを調整した．

3.2 実験の構成

以下に今回の実験について示す.

実験 1 ViT を用いて, 風景と人の顔の多義図形, 風景画, 肖像画の 3 クラス識別をした. ViT において, 識別するうえでの判断根拠を視覚化した.

実験 2 本研究では GAN を使うことで多義図形を生成することができるか否かを確認するため, LSGAN による多義図形の生成をした. また, 実験 1 における ViT を用いて多義図形を識別することで多義図形生成の評価をした.

3.3 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. 「多義図形」クラスの訓練データ 257 枚に対して 10 倍に Data Augmentation し, 2570 枚にした. また, 一義図形のデータセットとして, WikiArt [8] 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. 最後に各クラスの画像をグレースケール化することにより訓練データを 2 倍にし, 5140 枚にして使用した.

4 数値実験

4.1 実験 1

4.1.1 実験条件

表 1 に ViT の実験条件を示す.

表 1: 実験 1 ViT 実験条件

クラス	3 クラス (多義図形, 風景画, 肖像画)
エポック	25
バッチサイズ	8
訓練枚数	5140 枚/クラス
評価枚数	36 枚/クラス
テスト枚数	72 枚/クラス
データサイズ	384 × 384 × 3(RGB)
活性化関数	GELU
最適化関数	Adam
損失関数	交差エントロピー
ドロップアウト率	0.025113
学習率	3.1565e-05

4.1.2 実験結果

ViT による風景と人の顔の多義図形, 風景画, 肖像画の 3 クラス識別の識別率はベースライン 33.33% に対して, 95.83% となった. 表 2 に縦軸を真値, 横軸を ViT による予測値とした混同行列を示す. 図 1 に ViT の Attention の可視化による判断根拠結果を示す.

表 2: 実験 1 ViT による混同行列

真値	多義図形	65	5	2
	風景画	0	72	0
	肖像画	0	2	70
		多義図形	風景画	肖像画
		ViT による予測値		

4.2 実験 2

4.2.1 実験条件

実験 1 で用いた訓練データの中でグレースケール画像を含まない多義図形画像及び評価データ, テストデータの多義図形画像をまとめて LSGAN における訓練画像として合計 2678 枚を用いた. 本実験では 5000 epoch 学習させ, 50 枚の画像生成をした.

4.2.2 実験結果

図 2 に LSGAN によって生成された画像例を示す. LSGAN を用いて 50 枚を生成し, 生成画像を入力データとした ViT による多義図形識別率は % となった.

5 結果と考察

実験 1 より ViT の 3 クラス識別は従来の手法より一部識別率が高くなった. 識別率が向上したことは実験 2 において生成された画像を従来よりも識別可能となったことが言えるため生成に関しても重要な役割を担っている. 図 1 より ViT が識別する際に判断根拠としている部

分は, 主観ではあるが人間が多義図形であると判断している部分と類似している. また図 1 の D, E の画像は複数の顔や人物像の要素が含まれているため, 主観ではあるが人間にとっても多義図形か識別しづらい画像となっているにも関わらず, 判断根拠及び識別結果が正しいことと識別率が高いことから ViT は多義図形識別において優位性があるといえる.

実験 2 より LSGAN を用いることで多義図形が生成可能であることを示した. ViT によって多義図形が生成されているか否かを識別することで LSGAN におけて多義図形生成が成功していることを示せた. 生成画像に対する判断根拠については生成画像が 64 × 64　であるため見づらく, うまく視覚化をすることができなかった. 本実験では GAN による多義図形生成が可能か否かを示す実験だったが可能であることが示されたため, 今後はより高画質な画像生成をする GAN を実装することで判断根拠の視覚化をすることができると考えられる.

6　まとめと今後の課題

本研究ではだまし絵の一種である多義図形を計算機に理解させる手法を提案し, 数値実験及び視覚化を通してその有効性を示した. 実験 1 では計算機に多義図形と多義図形を構成する 2 要素の一義図形とを ViT を用いて 95.83% の識別率で識別させることに成功し, 判断根拠の視覚化によって計算機の識別に関する妥当性, 及び主観ではあるが人間と計算機の判断根拠の比較をすることで類似していることが示された. 実験 2 では GAN を使うことで多義図形を生成できることを示し, 生成画像に関して多義図形か否かを ViT を用いて %の識別率で識別させることに成功した.

今後の課題としては, 他の DA による識別率の向上, より多義図形に適した GAN の実装, 他のだまし絵や錯視画像への応用などが挙げられる.

参考文献

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, Vol. 27, , 2014.

[3] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2794–2802, 2017.

[4] 山村毅, 大洞将嗣, 大西昇, 杉江昇. 多義図形の解釈過程のシミュレーション. 計測自動制御学会論文集, Vol. 31, No. 8, pp. 1242–1244, 1995.

[5] 堀江紗世, 森直樹. 人工知能による多義図形認識手法の提案及び解析. 人工知能学会全国大会論文集, Vol. JSAI2020, pp. 3D1OS22a05–3D1OS22a05, 2020.

[6] 堀江紗世. 深層学習によるだまし絵認識手法の提案および解析. 後期研究発表会発表資料, 2020.

[7] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2623–2631, 2019.

[8] WikiArt. `https://github.com/cs-chan/ArtGAN/tree/master/WikiArt%20Dataset`.

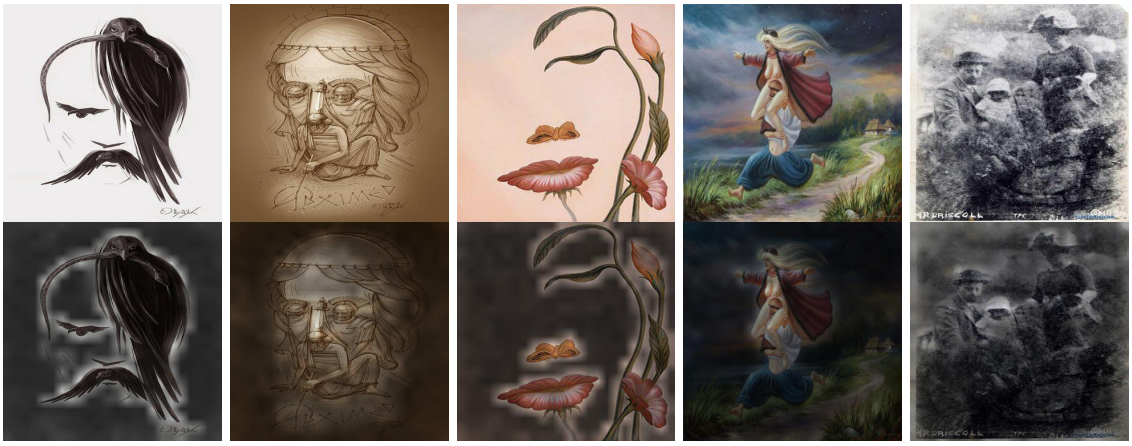

図 1: ViT による判断根拠の視覚化の例, 左から順に A, B, C, D, E とする

図 2: LSGAN による生成画像例