深層学習によるだまし絵認識手法の提案および解析

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も報告されており，今後も計算機による画像識別能力は向上していくと考えられている．一方, 人間の感性や認知に関わる分野で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明確な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が現在の人工知能の枠組みでは難しいためである．そこで，本研究では，人間と人工知能の比較を通じて視覚認知の多義性を理解することを目的としている．ここで，多義的とは一つの対象に対して二つ以上の見方を与える性質のこととし，一義的とは一つの対象に対して明らかな一つの見方を与える性質のこととする．

今回の発表では，多義図形と一義図形の画像分類のための深層学習モデルを学習・テストするデータセットの画像の情報量を段階的に落とすことにより，人間と人工知能の多義性の認識の相違や傾向を調べた．

2　従来研究

2.1　多義図形

だまし絵の一種に，多義図形と呼ばれるものがある．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．認知科学の分野では，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いる．図 1 に示す兎と鴨の両方に見える多義図形に関して，点 1 を注視しているときは 98 % 「鴨」と認識される一方, 点 5 を注視している時は 94% 「兎」として解釈されることが報告されている [1]．このような多義図形の解釈について，計算機を用いたアプローチは堀江らによって報告されている [2].

2.2　Convolutional Neural Network

画像認識分野において視覚野における受容野の性質に着想を得た深層学習手法として，Convolutional Neural Network (CNN) が注目されている．CNN の持つ層として，主にフィルタによって画像の局所的な特徴抽出をする Convolution 層, 特徴を統合する Pooling 層, 特徴量に基づいた分類をする全結合層がある．本研究では，2019 年に発表された EfficientNet [3] を用いて実験した．EfficientNet は，モデルの「深さ」と「広さ」と「解像度」の 3 つをバランスよく調整することにより，少ないパラメータと比較的簡素なモデル構造で高い精度を達成したことで知られる．

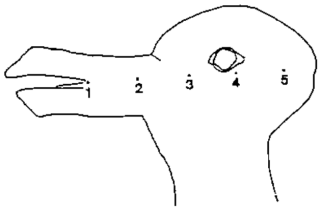

図 1: 兎鴨画像の注視点

2.3　Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) [4] とは，CNN が分類のために注視している範囲をカラーマップで表示する CNN の判断根拠の可視化技術である．Grad-CAM は, 予測クラスに対する勾配の大きさを寄与の大きさと考え, 分類予測を行う時に重要な箇所であると判断する．寄与の計算の際には, 一般的に最終畳み込み層の予測クラスの損失値に対する勾配が用いられる．本研究では, 勾配を可視化することで, 多義図形の判断根拠の解析をしている．本研究では Grad-CAM の図としてヒートマップを用いて勾配の大きさを可視化し, 勾配の大きい部分を赤色, 小さい部分を青色とした．本研究では, 赤色が最も CNN の最終畳み込み層の勾配が大きい, つまり判断に大きな影響を及ぼしているとし, また, 青色が最も CNN の最終畳み込み層の勾配が小さい, つまり判断にあまり影響を及ぼしていない, となるように可視化した．

3　提案手法

本実験の目的は，人間が認識できる多義性と人工知能が認識できる多義性の類似点および相違点を求めることによって，人間と人工知能それぞれの多義性の認識につ

いて理解を深めることである．よって本研究の提案手法は次の三つである．

- 多義図形データセット作成
- 人間から見た一義性・多義性を保った画像変換
- 人工知能による多義図形・一義図形の識別およびその評価

3.1 多義図形データセット作成

本研究ではインターネットから多義図形画像を収集し，データセットを作成した．

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の多義図形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．筆者の判断根拠は以下 4 点である．

- 回転しなくても多義性がある
- 「顔」が明確にわかる
- 「顔の中に含まれている別のオブジェクト」が明確にわかる
- 「顔」を構成する顔以外のオブジェクトが不自然な配置では無い

「多義図形」クラスの訓練データのみ，137 枚を 10 倍に Data Augmentation して使用した．また，一義図形のデータセットとして，WikiArt [5] 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．

3.2 画像変換

表 1 に今回採用した画像の変換方法と，その変換を施した際の一画像内の最大色数，色の候補数をまとめた．また，図 2 に画像変換の一例を示す．

3.3 多義図形・一義図形の識別及びその評価

本実験では，人の顔と風景の多義図形，肖像画，風景画の識別には ImageNet で学習済みの EfficientNetB3 [3] の最終 20 層の転移学習を，その評価には正解率及び多義図形を識別できたか否かの F1 値を用いた．また，その判断根拠の可視化には Grad-CAM を用いた．

表 1: 画像変換方法

	一画像内の最大色数	色の候補数
オリジナル	256^3 色	256^3 色
グレースケール化	256 色	256 色
減色 (カラー 3 色)	3 色	256^3 色
減色 (グレースケール 3 色)	3 色	256 色
減色 (カラー 2 色)	2 色	256^3 色
2 値化	2 色	2 色
エッジ検出	2 色	2 色

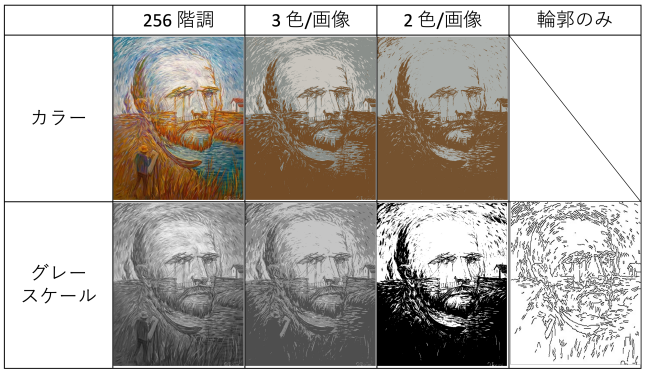

図 2: 画像変換の一例

4 数値実験

4.1 実験条件

表 1 に示した画像変換を適用したデータを学習モデルとして，それぞれ実験した．表 2 に実験条件を示す．

4.1.1 実験結果

図 3 に正解率を，図 4 に多義図形クラスの F1 値を示す．図 3，図 4 より，識別モデルに関してはノーマルデータで学習したモデル (以下ノーマルモデル)，カラー 3 色モデル，カラー 2 色モデル，グレースケールモデル，グレースケール 3 色モデルと正解率，F1 スコアが下がっていくことから，オブジェクトの形状情報より，色彩の情報が多義図形識別に及ぼす影響が大きいことがわかる．

4.1.2 考察

図 5 に風景画の Grad-CAM 結果例，図 6 に肖像画の Grad-CAM 結果例，図 7 に多義図形の Grad-CAM 結果例を示す．これらより，風景画は判断根拠が画面の端に集中している傾向があること，肖像画は判断根拠が顔部分に集中している傾向があること，多義図形は学習し

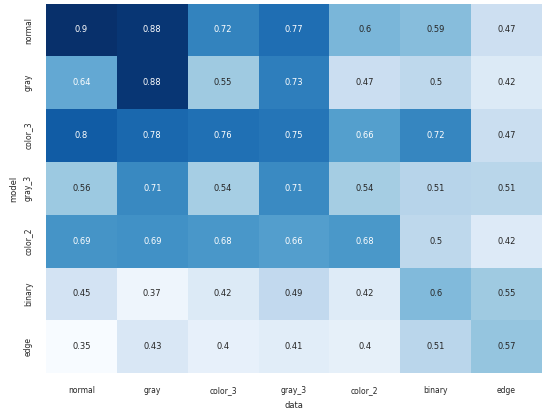

図 3: 正解率

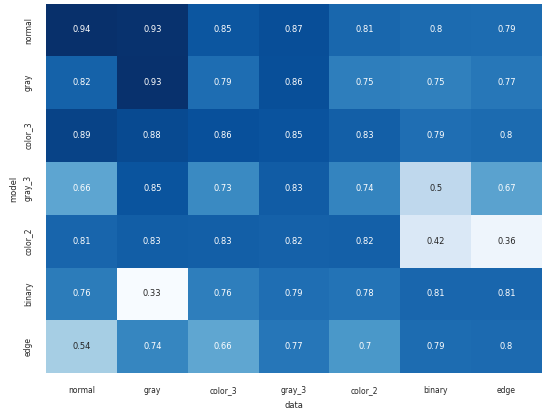

図 4: F1 値

表 2: 実験条件

使用モデル	EfficientNetB3
クラス	3 クラス (多義図形, 風景画, 肖像画)
エポック	Early Stopping
バッチサイズ	128
訓練枚数	1370 枚/クラス
評価枚数	135 枚/クラス
テスト枚数	135 枚/クラス
データサイズ	300 × 300 × 3
活性化関数	Softmax
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	0.2
学習率	0.00001

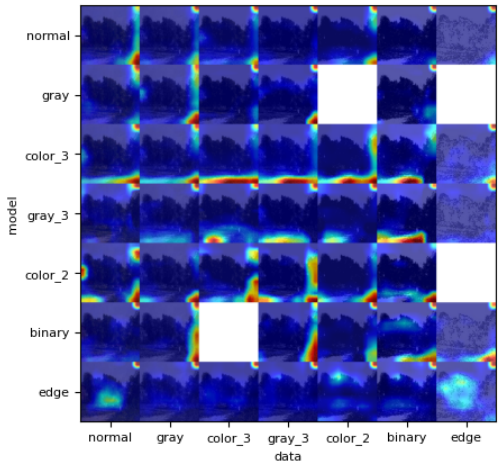

図 5: 風景画の Grad-CAM 結果例

たデータやテストデータの種類によって大きく判断根拠が変化することがわかった.

また，図 8 に学習モデル 7 種類，テストデータ 7 種類の合計 49 試行中，7 回以下しか多義図形として正しく識別されなかった画像を示す．この図より，バイナリデータやエッジデータのみによって識別が可能な多義図形も存在することがわかる．これより，バイナリデータやエッジデータのように極端に情報量を減らしたものを学習データやテストデータに加えたり，今回の各識別モデルを並列に結合することで，ノーマルデータでは識別できない多義図形を識別することが可能になると考えられる．

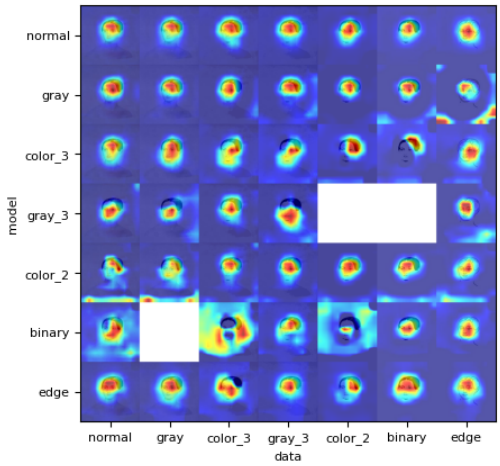

図 6: 肖像画の Grad-CAM 結果例

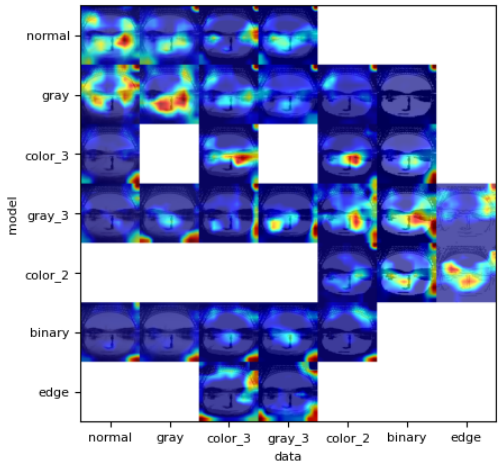

図 7: 多義図形の Grad-CAM 結果例

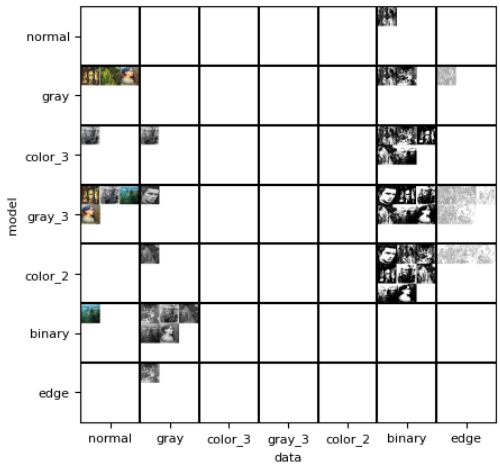

図 8: 49 試行中 7 回以下しか多義図形として識別されなかった画像群

5　まとめと今後の課題

本実験では，データセットの情報量を段階的に減らすことにより，色彩情報およびオブジェクトの形状が多義図形に与える影響を考察した．また，判断根拠の可視化により，一義図形は比較的勾配の分布のブレが少ないことに対し，多義図形は大きいことがわかった．また，極端に情報量を減らす画像変換を加えることで，多義図形の識別率が向上する可能性を示唆した．今後の課題としては，画像の見せ方や周囲の環境に配慮したアンケート実験，パラメータの調整，Vision Transformer による識別率向上などが挙げられる．

参考文献

[1] 岸本充史, 川端信男. 局所的・大域的情報選択モデルによる多義図形の非あいまい化. テレビジョン学会誌, Vol. 50, No. 5, pp. 594–598, 1996.

[2] 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.

[3] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114. PMLR, 09–15 Jun 2019.

[4] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

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