人工知能による多義図形認識手法の提案及び解析

1 はじめに

近年，画像識別を中心に人工知能研究が積極的になされている．従来の画像認識の研究では，描かれるオブジェクトが一義的な画像が対象とされており，複数の解釈が可能な多義図形に対して人工知能による画像認識を適用した例は殆ど報告されていない．本研究では，多義図形に描かれた二つのオブジェクトの認識と，多義図形に対する人工知能の認識における注視点の解析手法について提案する．これは，これまでの人工知能研究において対象とされてきた唯一の解があるタスクではないため難易度が高い．本研究では，多義図形を対象とした複数の実験を通じて，提案手法の有効性を確認する．

2 要素技術

2.1 多義図形

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意図形，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 [1][2]. 特に，本研究でも中心的に扱う鴨と兎両方に見える多義図形に関しては，川端らの研究 [3] より，図 1 の点 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．

本研究では，人間の多義図形認識が注視点の差によるとすれば，人工知能の多義図形の認識の注視点はどうなるかについて検討する．

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，Convolutional Neural Network (CNN) が注目されており，本研究では ImageNet で学習済みの VGG-16 を使用する．

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ることでが分類予測を行う時に重要な箇所であると判断する．勾配に関しては一般的に最終 Convolution 層の予測クラスの loss 値に対する勾配が用いられる．

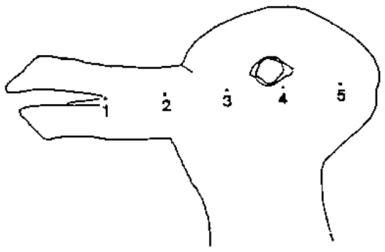

図 1: 兎鴨画像の注視点

3 提案手法

本研究では，人工知能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，CNN を用いた多義図形を理解する手法および実験方法について提案する．本研究では書籍や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．収集した画像について Data Augmentation を施し，それらを用いた各タスクに対して VGG-16 の 最終 Convolution 層以降の Transfer Learning を適用した．

3.1 実験 1

本実験の目的は，風景と人間の顔に関する多義図形と風景画および肖像画の人工知能による識別と，その判断根拠の解析である．データセットとして，WikiArt 中の “landscape”, “cityscape”, “portrait” の画像を用いた．風景と顔の多義図形については，本研究のために作成したデータセットを用いた．表 1 に実験条件を，表 2 に Optuna による最適化結果を示す．

結果として，多義図形はすべて多義図形として識別することに成功した．また図 2 に Grad-CAM 結果を示す．Grad-CAM の結果より，風景画については全体に満遍なく注視点があり，肖像画については人間の顔の領域に注視点があることがわかった．多義図形についても肖像画と同様に，人間の顔の領域に注視点が集まっていることが確認された．

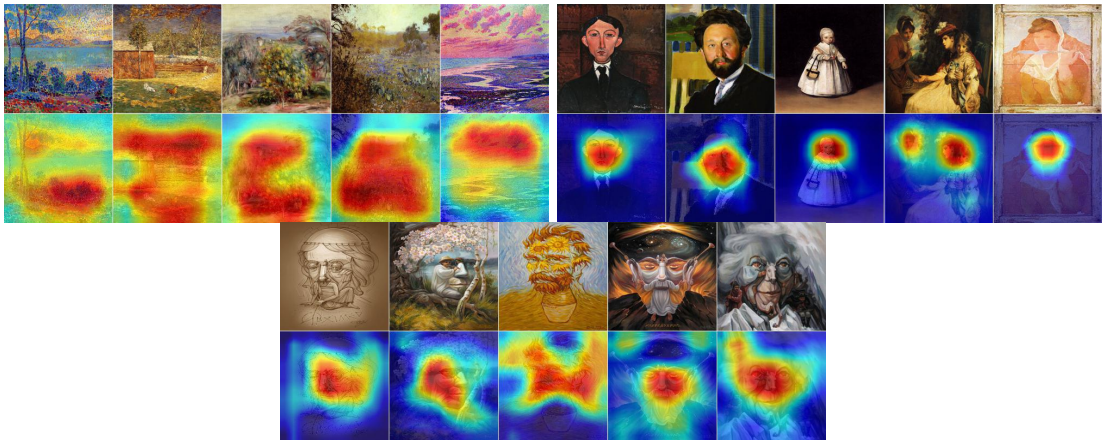

図 2: 実験 1 Grad-CAM の結果の例

表 1: 実験 1 の実験条件

クラス	3 クラス
Epoch	300
バッチサイズ	20
Train 枚数	380 枚/クラス
Valid 枚数	20 枚/クラス
Test 枚数	20 枚/クラス
データサイズ	200 × 200 × 3(RGB)
活性化関数	Softmax
最適化関数	Adam
損失関数	Categorical cross entropy

表 2: 実験 1 の Optuna による最適化結果

ドロップアウト率	0.64198
学習率	1.3410e-05
中間層のユニット数	400

表 3: 実験 2, 実験 3 の実験条件

クラス	2 クラス
Epoch	300
バッチサイズ	32
Train 枚数	800 枚/クラス
Valid 枚数	100 枚/クラス
Test 枚数	100 枚/クラス
データサイズ	200 × 200 × 3(RGB)
活性化関数	Softmax
最適化関数	Adam
損失関数	Categorical cross entropy

表 4: 実験 2 Optuna による最適化結果 (馬・蛙)

中間層のユニット数	100
ドロップアウト率	0.042160
学習率	1.8731e-05

3.2 実験 2

本実験の目的は，馬と蛙のイラスト画像を人工知能に学習させた場合の多義図形内の馬と蛙の認識と，その判断根拠の解析である．データセットとして，Pinterest の “horse illustlation”, “frog illustlation” の検索結果をスクレイピングし，馬・蛙とラベル付けしたものを用いた．表 3 に実験条件を，表 4 に Optuna による最適化結果を示す．

図 3 に蛙優位画像の Grad-CAM 結果，図 4 に馬優位画像の Grad-CAM 結果の例を示す．Grad-CAM の結果より，蛙優位画像では蛙の足部分と水面との境目を注視し，馬優位画像では馬の耳やたてがみ部分を注視していた．90 度回転によって馬・蛙に変化して見える絵については，94% 以上の精度でそのオブジェクトの変化を人工知能に認識させることに成功した．

3.3 実験 3

本実験の目的は，兎と鴨のイラスト画像を人工知能に学習させた場合の，多義図形内の兎と鴨の認識と，その判断根拠の解析である．データセットとして，Pinterest の “rabbit illustlation”, “duck illustlation” の検索結果をスクレイピングし，兎・鴨とラベル付けしたものを用いた．実験条件については表 3 に示した実験 2 におけるものと同一である．また表 5 に Optuna による最適化結果を示す．

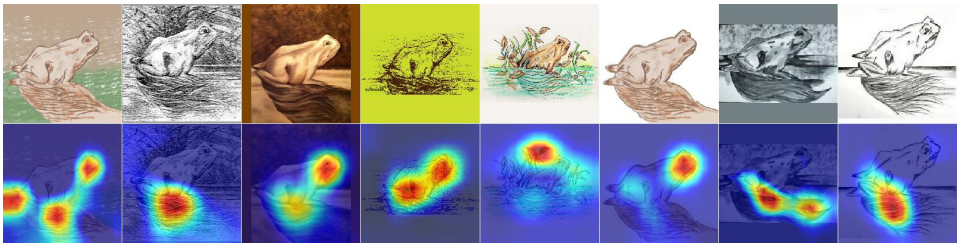

図 3: 実験 2 蛙優位画像の Grad-CAM の結果

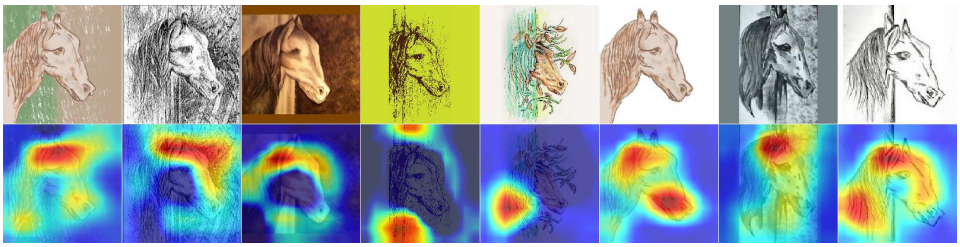

図 4: 実験 2 馬優位画像の Grad-CAM の結果

表 5: 実験 3 Optuna による最適化結果 (兎・鴨)

中間層のユニット数	500
ドロップアウト率	0.30511
学習率	5.0936e-05

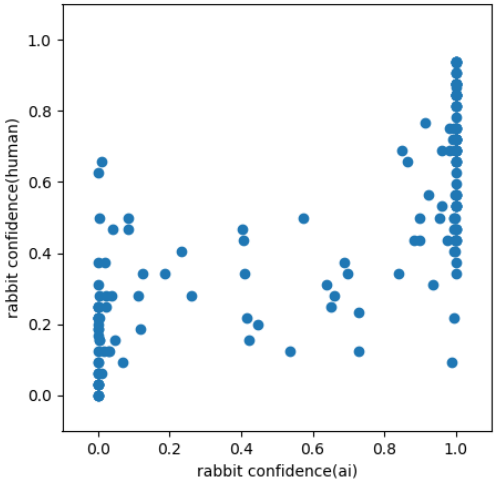

図 5: 実験 3 人間と CNN の認識の散布図

また，16 人に 7 枚の兎と鴨が描かれた多義図形について 15 度ずつ 360 度回転させた画像を見せ,「兎，鴨，どちらでもない」のいずれかを選択してもらうアンケートを取り,「0: 鴨，0.5: どちらでもない，1: 兎」と正規化して平均値を取った．図 5 に縦軸を人間へのアンケート結果，横軸を CNN による認識結果とした散布図を示す．アンケート結果と人工知能による認識結果の相関係数は 0.772 であったため，多義図形に対する人間による認識結果と人工知能による認識には強い相関があるといえる．

また，Grad-CAM の結果より，兎優位画像の注視点は耳の付け根あたりに集中していること，鴨優位画像の注視点は目〜胴に注視点が集中していることがわかった．図 6 に兎優位画像の Grad-CAM の結果，図 7 に鴨優位画像の Grad-CAM の結果の例を示す．Grad-CAM の結果と先行研究 [3] の比較より，人間の注視点と人工知能の注視点は異なるという結果となった．

3.4 実験 4

本実験の目的は，人間の注視状態に近い状況を作り出すことで CNN と人間の多義図形の注視点を比較することである．

以下人間と CNN の違いについて考察する．一つ目として，人間は注視点の周囲は徐々にぼやけて見える一方，人工知能は画面全体について見え方が均一であることが挙げられる．よって，Grad-CAM 領域に合わせてマスクをかけることで人間に近い状態を設定できるのではないかと考えた．二つ目として，人間は同じ画像を見ても注視点推移により違うオブジェクトが描かれているように見える一方，人工知能は同じ画像の捉え方は 1 通りしかないことが挙げられる．よって，マスクのかけ方で人工知能による test 結果が変われば人間らしいと考えることができる．

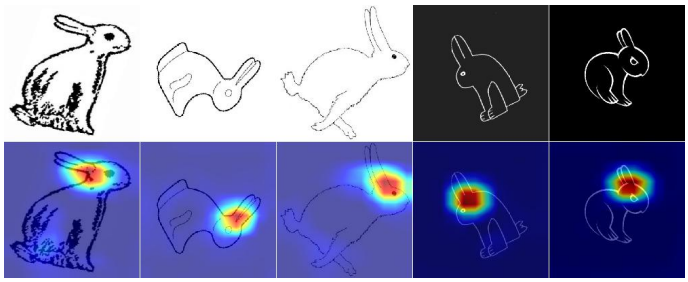

図 6: 実験 3 兎優位画像の Grad-CAM の結果

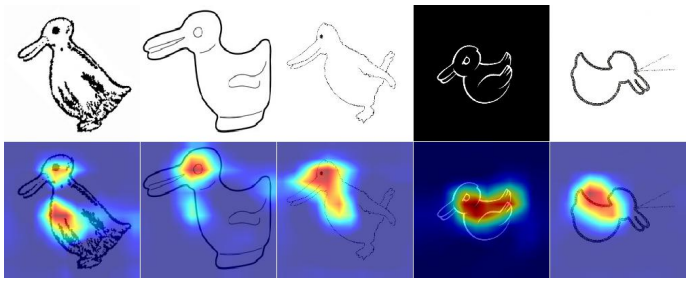

図 7: 実験 3 鴨優位画像の Grad-CAM の結果

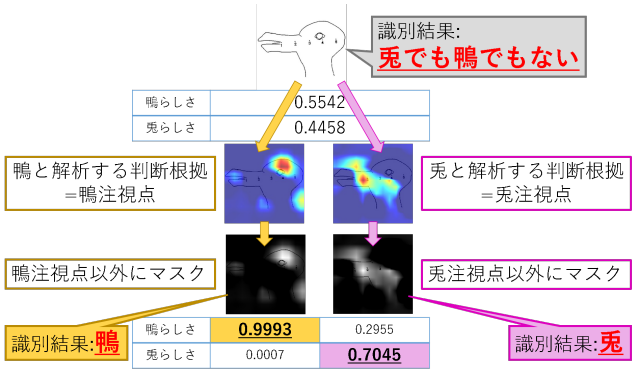

図 8: マスク解析実験

以上の点を踏まえて，実験 4 としてマスク解析について実験をした．まずは，実験 3 で作成した鴨と兎の 2 クラス識別モデルを用意し，test 結果の兎らしさ，鴨らしさが 0.4-0.6 である，つまり兎らしくも鴨らしくもある画像を抽出した．その枚数は 2016 枚中 24 枚だった．次に，抽出した画像の鴨の判断根拠となる Grad-CAM の CAM 値および兎の判断根拠の Grad-CAM の CAM 値と元画像の各ピクセルの値をそれぞれ掛け合わせる．この操作によって，鴨の CNN の注視領域以外にマスクがかかった状態の画像，兎の CNN の注視領域以外にマスクがかかった状態の画像を実験 4 では作成する．これらの画像の test 結果と加工前の画像の test 結果を比較した．

結果として，マスクによる鴨から兎への識別の変化があったのは 16 枚であった．約 67 % の画像で識別の変化が見られたことから，Grad-CAM による CNN の判断根拠領域の可視化は，人間の注視状態と近いことが考察される．

4　まとめと今後の課題

本研究では多義図形を人工知能に理解させるための手法を提案し，複数の数値実験を通じてその有効性を示した．実験 1 では，人工知能に多義図形と一義的な画像を識別させることに成功した．また，Grad-CAM を用いた解析により，人工知能による肖像画と多義図形の注視点における類似性を確認した．実験 2 では，90 度回転により馬・蛙に変化して見える多義図形について，94% 以上の精度で人工知能に認識させることに成功した．実験 3 では，回転によって兎・鴨に変化して見える多義図形についてアンケート調査を実施した結果，人間による認識と人工知能による認識の相関係数が 0.772 と非常に強い相関を示した．また実験 3, 4 では，Grad-CAM による解析の結果，兎らしさや鴨らしさを認識する際に人工知能と人間が注視する点は異なることがわかった．

今後の課題と研究予定としては, 画像の見せ方や周囲の環境に配慮したアンケート実験，アイトラッキング技術の活用, 他のだまし絵・錯視画像への応用を考えている．また、多義図形か一義図形かを識別するモデルを用いて、一義図形の中から多義図形らしい箇所を抽出する多義図形の生成タスクにも取り組みたい。

参考文献

[1] Kiyoshi Noaki Morikazu Kawabata N, Nobuo Yamagami. Visual fixation points and depth perception. 1977.

[2] Kawabara N. Attention and depth perception. 1986.

[3] 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.

[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.