7T-02

マルチモーダルモデルによる生態特徴観察と画像生成モデルを用いたデータ拡張の検証

芹澤栞苑† 岡山充希‡ 中野雄太‡ 長谷川達人‡
福井大学† 福井大学大学院‡

1. はじめに

ニューラルネットワーク（NN）のスケーリング則によると，NN の性能はモデルサイズとデータセットのデータ数に対してべき乗則的に向上する [1]．したがって，大規模なデータセットで大規模なモデルを訓練することが性能向上に直結する．一方， Fine-Grained Image Recognition（FGIR）は，画像認識の中でも，柴犬や秋田犬の識別のように同じカテゴリ内の細かい違いを見分けるタスクである．FGIR においてもデータセットを増強することが必要となるが，FGIR では詳細なラベルまで付与された画像データセットを大量に集めるのが困難という課題がある．

本研究では，FGIR 用のデータセットを構築するのが困難という課題に対し，画像生成モデルでデータを生成することで解決を図る．関連研究として，Stöckl ら[2]は画像生成モデル Stable Diffusion（SD）を用い画像データセット生成を評価した結果，SD は自然言語のプロンプトに対して，多くの場合で正確な画像を生成する事が出来たとしている．一方，生成画像の中にはプロンプトの意図と異なる画像も存在する事が分かっており，FGIR に対しても適用可能かは明らかになってない．

本研究では，画像生成モデルの FGIR への応用可能性を模索する．また，生成画像を既存のデータセットに追加して画像分類モデルの学習を行い，その推定精度や影響を調べることを目的とする，特に，対象クラスの外観から読み取れる形態学的特徴（以降，生体特徴と呼ぶ）や，画像の背景情報の工夫による，以下の 3 手法を試みた結果を報告する．

- 「対象クラス名と多様な背景状況」を，入力プロンプトとして画像を生成する．
- 画像とテキストのマルチモーダルモデルによって得られた「対象クラスの生態特徴」を入力プロンプトとして画像を生成する．
- 背景置換に特化した画像処理パイプラインシステムを使用して，データセットの学習データを描き直すことで画像を生成する．

2. 検証実験

実験では CUB-200-2011 という鳥類分類の FGIR データセットを用いる．学習用 5994 枚，テスト用 5794 枚で構成され，計 200 の鳥のカテゴリを有する．FGIR 画像分類モデルに High-temperaturE Refinement and Background Suppression (HERBS) [3]を使用する．HERBS は CUB-200-2011 の State-of-The-Art であるため採用した．画像生成モデルには SDXL1.0 を使用する．Stability AI により公開された SDXL1.0 は，従来の SD からモデルアーキテクチャーが改良され，パラメータ数と学習データが増加した画像生成モデルである．本実験では時間短縮のために，CUB-200-2011 から種名に「Sparrow」を含む 21 種類のクラスに限定して検証する．

2.1 鳥の種名を基づいた画像生成

本節では，(A)～(D)の異なる入力プロンプトに対する画像生成の影響を考察する．対象の学習データの各クラスに対して，既存の画像数に応じた割合で生成画像を追加し，モデルの学習を行う．

画像生成時の入力プロンプトを以下の通り定義する．(A)はベースラインとして設定し，データセットに追加画像を加えないケースとする．(B)は鳥の種名のみを用いる．(C)は種名を基に GPT-4 で現実的な背景状況を考案し，種名とともに用いる．(D)は(C)に加えて現実にはあり得ない背景状況を GPT-4 で考案して用いる．更に，(B)～(D)において「Bird」というキーワードをプロンプトに加えることで，確実に鳥が生成されるようにする．これは，事前検証で特定の鳥の種名に対して，鳥以外の生物や無機物，人間などが生成される事例があったからである．

実験結果を表 1 に示す．表 1 より，NN のスケーリング則に反して，精度が下がる傾向がある事がわかる．生成画像を目視で確認したところ，入力した種名に対して，生態特徴が定まらない例が確認された．SDXL1.0 は，種名に基づいた正確な知識が欠如している可能性を示唆している．そのため，種名による画像生成は，FGIR の詳細な種ごとの特徴を生成画像に反映させることができず，データ拡張に利用できないことがわかった．

2.2 鳥の画像の観察を基づく画像生成

次に，GPT-4V による鳥の生体特徴の観察に基づく画像生成手法について検討する．これは

表 1. 実験結果（Accuracy:%）

Add Image per Class (%)	Experiment				
	(A)	(B)	(C)	(D)	(E)
0	**92.85**	-	-	-	-
25	-	**92.85**	92.36	**92.85**	92.36
50	-	92.20	92.68	92.36	91.87
75	-	91.89	91.55	91.71	91.71
100	-	91.22	91.55	91.06	91.55

Data Augmentation Using Image Generation Models Based on Ecological Feature Observation with Multimodal Models
†Shion Serizawa, University of Fukui
‡Mitsuki Okayama, Yuta Nakano, Tatsuhito Hasegawa , Graduate School of Engineering, University of Fukui

SDXL1.0 が鳥の種名を知らない可能性を考慮し，鳥の特徴を言語化してプロンプトにする手法である．具体的な方法は次のとおりである．

1. 各クラスの画像を 1 枚ランダムに選択する．
2. GPT-4V を使用してこれらの鳥種の生態特徴を観察し，それぞれの顔，翼，くちばし，体の色，形，特徴などの情報をテキスト化する．（以後，観察テキストと呼ぶ．）
3. GPT-4 を用いて観察テキストを SDXL1.0 のプロンプト形式に要約し，画像を生成する．

以上を手法(E)とし，表 1 に結果を載せている．表 1 より，手法(E)も有効に働かない事がわかる．生成画像を確認すると，GPT-4V による観察で得られた特徴が反映されていない鳥が存在した．例えば「Harris's Sparrow」は，頭から胸にかけて，中心に黒い帯のような柄が特徴なのだが，生成画像では，その特徴が無いという例があった．この原因は，SDXL 内の CLIP のトークン数の制限や，CLIP の性能に起因していると考えている．このため，詳細な特徴や微妙な違いが画像生成で適切に表現されない可能性がある．

2.3 One-Shot 学習の場合の検証

前節の手法(E)における手順 1 で用いた，各クラス 1 枚の画像を学習データに割り当て One-shot 学習を行う．また，各クラスに対し，(B)～(E)で生成された画像を 5 枚，10 枚追加し学習を行い比較する．

実験結果を表 2 に示す．表 2 より，One-Shot 学習では，FGIR でも画像生成によるデータ増強が有効に働いたことがわかる．更に，(D)でやや精度向上が見られることから，背景の多様性が性能向上に寄与する可能性が示された．(B)，(D)の比較より，モデルが画像に対する前景と背景を正確に区別し，鳥に注目する能力の向上に寄与した可能性がある．一方で，(E)の GPT-4V による鳥の生体特徴の観察は手法(B)～(D)より精度が低い．原因としては，2.2 節で指摘した，CLIP による表現の制約と，観察による SDXL1.0 プロンプトが入力トークンを消費することにより，背景に関するプロンプトを含められず，背景の多様性を実現できなかったと考えられる．

表 2. One-Shot 実験における結果

Add Image Num per Class	Experiment				
	(A)	(B)	(C)	(D)	(E)
0	7.81	-	-	-	-
5	-	76.59	69.11	76.42	45.53
10	-	78.86	78.54	**81.95**	45.04

表 3. 背景置換時の実験結果

Add Image per Class (%)	Experiment		
	(A)	(B)	(C)
0	92.85	-	-
25	-	92.52	**93.17**
50	-	92.68	92.36
75	-	92.68	92.20
100	-	92.20	**93.17**

2.4 背景置換システムを用いた画像生成

正規の訓練データに対して，背景を描き換える手法について検討する．これは鳥の生態特徴を正確に描写しながら，背景に多様性を生むことができない点に対処するためである．背景描き換えには，Shopify による背景置換に特化した画像処理パイプラインシステム，Shopify Image Background Replacement[4]を用いる．このシステムは，対象のオブジェクトを認識・切り抜きを行い，奥行き関係を踏まえた画像生成を行うことで背景置換を実現するものである．実験に用いる入力プロンプトは手法(B),(C)の 2 種について行う．

実験結果を表 3 に示す．表より，わずかに推定精度が向上する条件があることと，生成画像の数を増やしても精度低下の傾向が見られなかったことがわかる．以上より，画像生成モデルの FGIR 活用は，FGIR の詳細な特徴を表現する点に現状は問題があるが，背景の多様化には活用可能であることが明らかとなった．

3. まとめ

本研究では，ニューラルネットワークのスケーリング則を考慮し，画像生成モデルにより FGIR 用のデータセットのデータ数を増やす観点で検証を行った．入力プロンプトの工夫によるデータ増強に関して，SDXL1.0 による画像生成では，FGIR のための鳥画像を正確に生成することができないことが分かった．しかし，One-Shot においては精度向上に寄与することが示された．また，学習データの画像に対する背景多様化として，画像生成モデルを使うことは有用であることが示唆された．しかし，本研究の課題として安定的に精度向上を行う方法の構築が残った．

謝辞

本研究は，JST ACT-X(JPMJAX20AJ) の支援を受けたものであり，ここに感謝の意を表する．

参考文献

[1] Bahri, Y, et al. "Explaining neural scaling laws." arXiv preprint arXiv:2102.06701 (2021).
[2] Stöckl, A. "Evaluating a synthetic image dataset generated with stable diffusion." International Congress on Information and Communication Technology. Singapore: Springer Nature Singapore (2023).
[3] Chou, Po-Yung, et al. "Fine-grained Visual Classification with High-temperature Refinement and Background Suppression." arXiv preprint arXiv:2303.06442 (2023).
[4] Gözükara, F. "SDXL Background Replacement for Product Images ". GitHub. https://github.com/FurkanGozukara/background-replacement （最終アクセス日：2024 年 1 月 10 日）