クラスラベルを仮定しない画像集合の多様性評価

岡本 夏旺[1,a)]　品川 政太朗[1,b)]　中村 哲[1,c)]

概要

画像生成モデルの生成画像の評価や，質の高い画像集合を用いたモデルの学習を行うために，画像集合の多様性を評価できる指標が重要である．しかし，既存の多様性評価指標では，特定のクラスラベルを用いて多様性の評価が行われているため，特定のクラスラベルに属さない画像集合に対しては正しい評価を行うことができない．そこで本研究では，任意の画像集合に対して，特定のクラスラベルを仮定せずに多様性を評価できる指標を提案し，提案する評価指標の有効性を既存の指標との比較により明らかにする．

1. はじめに

現在，階層的な概念において，概念の上下関係を表す方法として人間が定めた既知のグラフを用いる方法が一般的に考えられるが，ある概念に所属する画像集合を用いて，画像集合の多様性を測ることで画像の観点から概念の上下関係を推定する方法を考えることができる．そのため，画像集合における多様性を評価できる指標を検討することは重要である．画像集合の多様性を評価する方法として考えられる最も単純な方法の一つは画像認識モデルの予測されたクラスラベルの周辺分布のエントロピーを計算し多様性を評価する方法である．この方法は生成画像集合の評価である IS (Inception Score) [1] の一部でも定式化され用いられている．

画像認識モデルの予測されたクラスラベルの周辺分布のエントロピーを用いる方法の問題は，画像認識モデルのクラスラベルに属さない画像集合の多様性評価を行う場合に，正しく分類することができず，多様性を評価できないことである．これは，クラスラベルが既知のクラスラベルに限定されているからである．実際に，生成画像集合の多様性を考慮した評価指標である IS では，既知の 1000 クラスラベル以外に属するデータセットを用いて，多様性評価を行なった場合，正しく評価できないことが既存研究 [2] で報告されている．

1 奈良先端科学技術大学院大学
a) okamoto.natsuo.op8@is.naist.jp
b) sei.shinagawa@is.naist.jp
c) s-nakamura@is.naist.jp

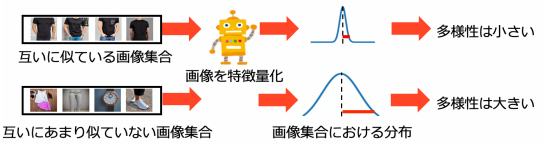

図 1: 分布の裾の長さを考慮し多様性を評価する図

生成画像集合において，クラスラベルを用いず，生成画像の質と多様性を切り分けて評価できる指標として Precision と Recall がある [6]．Precision は生成画像集合のうち実画像集合に属する割合を計算することで，生成画像の質を評価し，Recall は実画像集合のうち生成画像集合に属する割合を計算することで，生成画像の多様性を評価する指標である．Recall では生成画像における多様性評価に焦点を当てており，リファレンスとなる画像集合が必要である．

そこで本研究では，図 1 のように似ている画像で構成されている画像集合では，画像特徴量がそれぞれ類似したベクトルになるという考えの基，画像集合における分布の裾の長さに注目することで，画像認識モデルのように既知のクラスラベルに限定せず，リファレンスとなる画像集合とクラスラベルを用いず，画像集合の多様性を評価できる指標を提案し，実画像集合でエントロピーを計算し多様性を評価する手法との比較を行い，有効性を示す．また，提案指標が生成画像集合に対して，実画像集合とクラスラベルを用いず，多様性を評価することができるのか検証を行う．

2. 関連研究

クラスラベルを仮定せず，画像生成システムを生成画像の質を評価する指標では，FID (Fréchet Inception Distance) [3] が存在する．FID は，Inception v3 を用いて，実画像集合と生成画像集合における分布が多変量正規分布に従うと仮定し，両分布の距離を計算した評価指標である．FID は以下のように表せる．

$$\|\mu^r - \mu^g\|^2 + \mathrm{Tr}\left(\Sigma^r + \Sigma^g - 2\left(\Sigma^r\Sigma^g\right)^{1/2}\right)$$

$N(\mu^r, \Sigma^r)$, $N(\mu^g, \Sigma^g)$ は実画像集合と生成画像集合における平均ベクトル，分散共分散ベクトルを表す．KID (Kernel Inception Distance) [4] では，FID における分布が多変量正規分布に従うという仮定を緩和することで，FID の改善に努めた評価指標である．FID_∞, IS_∞ [5] は，サンプル

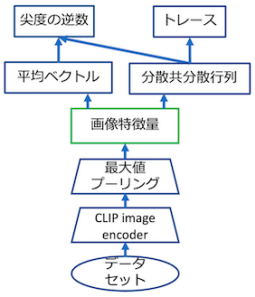

図 2: 提案手法の構造

数によるバイアスを削減するために，サンプル数が無限であった場合の推定値を計算する手法を提案した．これらの評価手法では，生成画像の質と多様性をそれぞれ明示的に計算しておらず，本研究では画像集合において多様性の観点のみに焦点を当てた評価指標に注目している．

3. 手法

図 2 は提案手法の構造を示しており，以下では提案手法について順に説明を行っていく．

3.1 画像特徴量の計算

画像を特徴量として表現するために，CLIP (Contrastive Language–Image Pre-training) [7] を用いる．CLIP を用いる理由は，CLIP が大規模な画像とテキストにより対照学習を行う手法であり，画像とテキストが組になった大規模データセットを用いて，テキストと画像間の表現が学習されており，さまざまなクラスに対応することができるからである．CLIP は画像の特徴量とテキストの特徴量を求めることができる image encoder と text encoder から成り立っており，image encoder では ResNet ベースのものと，Vision Transformer (ViT) ベースのものが存在する．本研究では，image encoder が ViT ベースで構成されている CLIP を用いる．

本研究では，画像特徴量の計算は以下の手順で行う．

(1) 画像を入力とし，ViT ベースの CLIP の image-encoder における隠れ層の最終層の 1 つ手前の層の潜在ベクトルを計算する
(2) 計算された隠れ層の潜在ベクトルに対し，[CLS] トークンに当たる値を削除し，カーネルサイズを 3×3，ストライド数 3 とし，最大値プーリングを行う
(3) 同上を画像の特徴量とする

最終層の 1 つ手前の層である隠れ層の値を用いる理由は，IS や FID の計算で用いられる Inception v3 でも最終層の 1 つ手前の層である隠れ層の潜在ベクトルが用いられており，また，CLIP をベースとした画像生成モデル Stable diffusion でも，最後の層で隠れ状態の値が急激に変化するため最終層の 1 つ手前の層を用いた方がよりプロンプトを反映することができると報告されているからである．[*1]

3.2 仮説と提案手法

本研究では，画像集合ごとに計算した画像特徴量が多変量正規分布に従っていると仮定し，多様性がある画像集合の多変量正規分布はより先端が尖った分布となり，多様性があまりない画像集合の多変量正規分布は裾の長い分布となると仮説を立て，既知のクラスラベルを用いない 2 つの多様性評価指標を提案する．この仮説は，似ている画像で構成されている画像集合では，画像特徴量がそれぞれ類似したベクトルとなり，あまり似ていない画像で構成されている画像集合では，画像特徴量がそれぞれ大きく異なるのではないかという考えに基づき立てた仮説である．以下ではある画像集合 c の画像 i における画像特徴量を $\mathbf{x}_i^c$ とし，画像集合 c の多変量正規分布の平均ベクトルを μ^c 分散共分散行列を $\mathbf{\Sigma}^c$ とする．

分散共分散行列のトレースを計算する指標

画像集合の分布の裾の長さを評価するため，画像集合 c における多様性評価指標として，分散共分散行列が $\mathbf{\Sigma}^c$ のトレースを計算を行う手法を提案する．この計算値は平均ベクトルがゼロベクトル，分散共分散行列がゼロ行列である多変量正規分布 $N(\mathbf{0}, \mathbf{0})$ との FID の計算値と同値となる．FID では，分布間の距離を計算しており，$N(\mathbf{0}, \mathbf{0})$ である分布と比較を行うため，画像集合 c における分布の形状のみを比較した値となり，画像集合 c がどれだけ裾の長い分布であるかを数値化した値と考えられる．そのため，トレースの計算値が大きいほど分布において裾が長く，多様性があると考えられる．

以下では，なぜ分散共分散行列 $\mathbf{\Sigma}^c$ のトレースの計算値が平均ベクトルがゼロベクトル，分散共分散行列がゼロ行列である多変量正規分布 $N(\mathbf{0}, \mathbf{0})$ との FID の計算値と同値になるのかについて説明をする．分散共分散行列 $\mathbf{\Sigma}^c$ の固有ベクトルを求め，固有ベクトルにより並行移動し，回転した分布の平均ベクトル $\mu^{c'}$ はゼロベクトル，分散共分散行列 $\mathbf{\Sigma}^{c'}$ は対角成分を固有値とする対角行列となり，$N(\mathbf{0}, \mathbf{\Sigma}^{c'})$ と表せる．$N(\mathbf{0}, \mathbf{0})$ である分布と，固有ベクトルにより並行移動し，回転した分布 $N(\mathbf{0}, \mathbf{\Sigma}^{c'})$ の FID の計算値は，$\mathbf{\Sigma}^{c'}$ のトレースをとった値となる．$\mathbf{\Sigma}^{c'}$ は対角成分を固有値とする対角行列であるため，トレースをとった値は，固有値の総和を取った値と同値である．$\mathbf{\Sigma}^c$ におけるトレースをとった値と固有値の総和は一致するので，$\mathbf{\Sigma}^c$ におけるトレースをとった値は平均ベクトルがゼロベクトル，分散共分散行列がゼロ行列である多変量正規分布 $N(\mathbf{0}, \mathbf{0})$ との FID を計算した値と同値であることがいえる．

[*1] https://blog.novelai.net/novelai-improvements-on-stable-diffusion-e10d38db82ac

(a) “clothing” (b) “jacket” (c) “leather jacket”

図 3: fashion-200k データセットのそれぞれのクラスに所属する画像例

尖度の逆数を計算する指標

尖度とは，分布が正規分布からどれだけ尖っているかを表す統計量であり，山の尖り度と裾の長さを示す指標であり，尖度の値が大きいほど，山の尖り度が高い正規分布となる．そのため多様性がある場合は，尖度は低くなり，多様性がない場合は，尖度は高くなると考えられる．画像集合 c における画像数を N としたとき，クラス c における多変量正規分布の尖度の計算式は以下のように表せる [8].

$$\frac{1}{N}\sum_{i=1}^{N}\left[\left(\mathbf{x}_i^c-\mu^c\right)'\left(\mathbf{\Sigma}^c\right)^{-1}\left(\mathbf{x}_i^c-\mu^c\right)\right]^2$$

本研究では，多様性が大きいほど，大きい値を出力する指標を構築し，昇順に並び替えた際に他指標との比較を行いやすくするため，尖度の逆数を取った値を提案手法とする．尖度の逆数の計算値が大きいほど分布において裾が長く，多様性があると考えられる．

4. 実験

本研究では，画像集合 c は，上位概念，中位概念，下位概念クラスのいづれかに属すると定義し，上位概念，中位概念，下位概念クラスの順に多様性があると仮定した．仮定を満たすために，fashion-200k データセットを分割して利用した．この仮定は fashion-200k データセットを用いると，たとえば，上位概念，中位概念，下位概念クラスをそれぞれ “clothing”, “jacket”, “leather jacket” クラスとした場合，“clohting” クラスの画像集合が一番多様性があり，“leather jacket” クラスの画像集合が一番多様性が少ないと仮定したものである．図 3 データセットのそれぞれのクラスに所属する画像例である．2 つの実験内容は以下のとおりである．

実験 1：提案手法は既存手法よりも多様性を正確に評価をすることができるのか

提案手法が既存手法よりも，実画像集合において多様性を測る有効な評価指標であるのかを検証する．

実験 2: 提案手法は生成画像集合においても多様性を測る有効な手法だといえるのか

提案手法が生成画像集合においても同様に多様性を測る有効な評価指標であるのかを検証する．

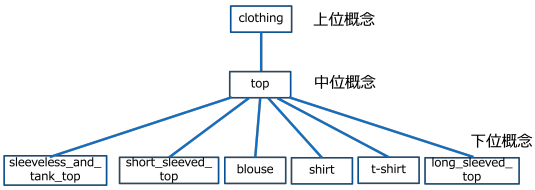

図 4: “top” を中位概念とした時の上位概念と下位概念の関係性

4.1 実験 1 における実験設定

多様性評価指標におけるベースライン

ベースラインを IS や FID で用いられる分類モデル Inception v3 を用いて，クラスラベルの予測を行い，予測されたクラスラベルの周辺分布のエントロピーを計算する手法と，Inception v3 と同様のラベルを用いて，image encoder が ViT をベースとしている CLIP のエントロピーを計算する手法とした．クラスラベルを y とした時，エントロピーの計算では，確率分布 $P(y)$ における，情報源の不確かさを数値化することができる．そのため，多様性がある場合はエントロピーの値は大きくなると考えられる．

データセット

fashion-200k データセット [9] を用いて実験を行った．fashion-200k データセットは，オンラインショッピングサイトから商品画像とその商品説明をクロールし作成されたデータセットである．fashion-200k データセットは 5 つのクラス (dress,top, jacket, pant, skirt) と，それぞれのクラスごとに更に細分化されたクラスから成り立っている．たとえば，“top” クラスにおいて更に細分化されたクラスには “blouse”, “shirt”, “short sleeved top”, “t-shirt” などが存在する．図 4 は “top” を中位概念とした時の上位概念と下位概念の関係性を表した図である．

本研究では，画像数が 3500 枚以上のクラスのみを対象にしており，fashion-200k からランダムに 3500 枚サンプリングした画像集合を上位概念クラスである “clothing” クラスとし，5 つのクラスにおいてランダムに 3500 枚サンプリングした画像集合をそれぞれ中位概念クラス，中位概念クラスを更に細分化されたクラスから 3500 枚サンプルした画像集合を下位概念クラスと定義した．

評価方法

本研究ではそれぞれ 5 つの中位概念クラスを中心に，上位概念である clothing クラス，中位概念クラス，中位概念の下位概念である下位概念クラスに対して，ベースラインである手法と提案手法の計算を行う．この計算をランダムシードの値を変更し，10 回行い，それぞれのクラスごとに平均値を算出する．そして，5 つの中位概念クラスを中心とし，平均値の大きさを基に大きい値順にランク付けをし，上位概念クラスと中位概念クラスにおける平均ランクを計算することによって評価を行う．平均ランクが高いほど，多様性のある実画像集合を多様性があると評価できて

手法	上位概念平均ランク	中位概念平均ランク
Inception v3 entropy	1.8	4.0
CLIP entropy	2.6	4.2
トレース	**1.0**	3.2
尖度の逆数	1.2	**2.4**

表 1: 実験 1 における既存手法と提案手法の比較

手法	上位概念ランク	中位概念ランク
Inception v3 entropy	1.6	3.4
CLIP entropy	3.4	3.8
トレース	1.6	4.6
尖度の逆数	3.4	4.0

表 2: 実験 2 における各指標のランク

いることを意味する．上位概念クラス，中位概念クラスの平均ランクはそれぞれ 1,2 になるのが理想的である．このようなランクを用いた評価を採用した理由は，各々指標において数値のスケールが大きく異なり，数値を比較した評価方法を行うことが困難だと考えたためである．

4.2 実験 2 における実験設定

多様性評価指標におけるベースライン

実験 1 と同様のベースラインを用いた．

生成画像の作成

本研究では，画像生成モデルである stable diffusion v1.5[*2]を用いて画像生成を行った． 実験 1 で用いた全てのクラスに対して，入力テキストを “a **クラス名** item, real” とし，それぞれのクラスごとに画像を 5000 枚生成した．たとえば，上位概念クラスである “clothing” クラスの場合，入力テキストは，“a clohting item, real” である．

評価方法

それぞれ 5 つの中位概念クラスを中心に，ベースラインである手法と提案手法の計算を行う．計算値をもとに大きい値順にランク付けをし，上位概念クラスと中位概念クラスにおける平均ランクを計算することによって，評価を行う．

5. 結果

表 1 は fashion-200k データセットにおいてそれぞれの評価指標による上位概念と中位概念クラスの平均ランクを表す．表 1 の結果から，上位概念クラスにおける平均ランクは分散共分散行列のトレースを計算する方法が一番良く，中位概念クラスにおける平均ランクは尖度の逆数を計算する方法が良いとわかる．2 つの提案手法は優劣をつけることは難しいが，ベースラインとして設定した Inception v3 と CLIP を用いてエントロピーを計算する方法よりも正しく多様性を評価できていることがわかる．

表 2 は，生成した画像集合において，それぞれの評価指標による上位概念と中位概念クラスの平均ランクを表す．理想としては，上位概念，中位概念クラスのランクはそれぞれ 1,2 であるが，表 2 の結果から，提案指標では，理想としているランクとはかけ離れており，生成画像の評価に使う際には問題があるように考えられる．

[*2] https://huggingface.co/runwayml/stable-diffusion-v1-5

図 5: “clothing” クラスにおける生成画像例

6. 考察

実画像集合においては，2 つの提案手法の方が多様性を正しく評価することができていると考えられ，ラベルを定義せず，多様性を評価することができるため，より汎用性がある評価指標であると考えられる．

生成画像集合において，多様性を正しく測ることができなかった原因として 2 つの可能性があると考えられる．1 つ目は，生成画像集合においては，多様性の大きさが上位概念，中位概念，下位概念順に従うという仮定が常に成り立つは限らない点である．図 5 は “clothing” クラスにおいて，生成された画像集合の例であるが，“dress” や “pant” の画像は少なく，“top” や “jacket” が多くかなり偏っていることがわかる．そのため，fashion-200k データセットで設計した “clothing” クラスよりも多様性が低くなってしまった可能性がある．2 つ目は背景の影響を受けているという可能性である．今回実画像集合で用いた fashion-200k データセットは背景色が白色で統一されており，生成された画像では，背景がかなりバラバラになっていることが図 5 から示唆されている．そのため，背景色に影響された可能性が考えられる．

7. まとめ

本研究では多様性を評価する既存の指標では，特定のクラスラベルに属さない画像集合に対しては正しい評価を行うことができないという問題を，クラスラベルとリファレンスとなる画像集合を用いない画像集合の多様性評価手法の提案により，解決しようとした．実験の結果から実画像集合においては，エントロピーを計算し多様性を評価する手法よりも提案手法の方が多様性を正しく評価することができ，画像集合に依存せず評価を行うことができる点で優位性があると考えられる．しかし生成画像においては，実験の結果から優位性が示せず，実験における仮定や，背景情報が原因になった可能性が考えられる．そのため，背景情報を考慮しないようにするために，物体検出技術を用いて提案手法に優位性があるのかを検証していく必要がある．

参考文献

[1] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Rad- ford, A., and Chen, X. “Improved techniques for training gans.” NIPS, pp. 2234–2242, 2016.

[2] Shane Barratt and Rishi Sharma. “A Note on the Inception Score.” arXiv:1801.01973, 2018.

[3] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. “GANs trained by a two time-scale update rule converge to a local Nash equilibrium.” NIPS, pp. 6626–6637, 2017.

[4] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. “Demystifying mmd gans.” ICLR, 2018.

[5] Min Jin Chong and David Forsyth. “Effectively unbiased fid and inception score and where to find them.” CVPR, pp. 6070–6079, 2020.

[6] Sajjadi, Mehdi S. M., Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. “Assessing Generative Models via Precision and Recall.” NIPS, pp. 5234–5243, 2018.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. “Learning transferable visual models from natural language supervision.” ICML, pp. 8748–8763, 2021.

[8] Mardia, K.V. “ Measures of Multivariate Skewness and Kurtosis with Applications.” Biometrika, 57(3), pp. 519-530, 1970.

[9] Han, Xintong and Wu, Zuxuan and Huang, Phoenix X. and Zhang, Xiao and Zhu, Menglong and Li, Yuan and Zhao, Yang and Davis, Larry S. “Automatic Spatially-aware Fashion Concept Discovery” ICCV, pp. 1472–1480, 2017.