# 複数モデルの勾配を統合したデータセット蒸留による下流タスクの高精度化

ER22005 市川翔　　指導教授：藤吉弘亘

## 1. はじめに

Contrastive Language-Image Pre-Training (CLIP) や Distillation with no labels (DINOv1) に代表される事前学習済みモデルは，下流タスクでの評価において高い性能を示しており，多様な画像認識タスクの基盤として広く利用されている．事前学習済みモデルを下流タスクへ転移学習をする際，計算コストを抑制するために限られた学習データによる効率的な学習が重要である．その 1 つのアプローチとして学習データを少数の合成データに凝縮するデータセット蒸留が注目されている．従来のデータセット蒸留手法は単一モデルに基づく手法であり，学習目的の違いに起因する特徴の多様性を同時に反映できない可能性がある．そこで本研究では，Linear Gradient Matching [1] を拡張し，特性の異なる複数の事前学習済みモデルを用いたデータセット蒸留手法を提案する．本手法により，未知のモデルに対しても分類に有効な特徴を保持する合成画像の作成を目指す．

## 2. Linear Gradient Matching

データセット蒸留の代表的な手法として，Linear Gradient Matching (LGM) [1] が提案されている．LGM は，事前学習済みモデルを固定した状態で，実画像と合成画像に対する線形分離器の勾配が一致するように合成画像を最適化する．これにより，分類に有効な特徴を合成画像に凝縮できる．しかし，LGM は単一モデルの勾配に基づいて最適化を行うため，得られる合成画像は蒸留に使用したモデルの特徴表現に依存する．その結果，異なるアーキテクチャを持つモデルに対しては，転移性能が十分に発揮されないという課題がある．

## 3. 提案手法

本研究では図 1 に示すように，LGM を拡張し，複数の事前学習済みモデルを同時に用いるデータセット蒸留手法を提案する．学習目的の異なるモデルの勾配を同時に用いることで，特定のモデルに依存しない合成画像の作成を目的とする．本研究では，自己教師あり学習に基づく DINO 系モデルと，画像と言語対照学習に基づく CLIP 系モデルを組み合わせる．提案手法における勾配損失を式 (1) に示す．本手法では，式 (1) を最小化するように，合成画像の作成のパラメータを最適化する．具体的には，事前学習済みモデルおよび線形分離器のパラメータは学習せずに，誤差逆伝播法により合成画像に対する線形分離器の損失勾配を計算し，実画像から得られる勾配との類似度が高くなるように，合成画像を更新する．ここで，$g^{\mathrm{DINO}}$ および $g^{\mathrm{CLIP}}$ は，それぞれ DINO 系モデルおよび CLIP 系モデルで得られる線形分離器の勾配を表す．

$$\mathcal{L}_{\mathrm{grad}} = (1 - \cos(g^{\mathrm{DINO}})) + (1 - \cos(g^{\mathrm{CLIP}})) \qquad (1)$$

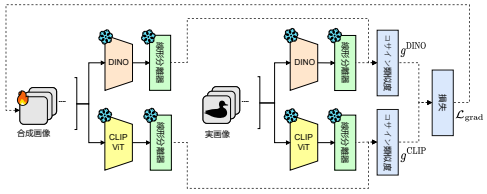


図 1: 提案手法の概要

## 4. 評価実験

本研究では，作成した合成画像が蒸留に使用していないモデルに対しても分類に有効な特徴表現を保持しているかを検証するため，事前学習済みモデルのバックボーンを凍結し，最終層の線形分離器のみを学習する Linear Probing を用いて分類精度評価を行う．評価モデルには，蒸留に使用していない CNN アーキテクチャである ResNet50 を用いた．評価には 20 クラスの画像を用い，各クラス 5 枚の画像から線形分離器を学習した．実験は 5 回実施し，その平均値と標準偏差を評価値とした．また，比較として全学習データを用いた場合の精度を上限値として併記する．画像の作成には，DINOv1，DINOv2，CLIP，Sigmoid Loss for Language Image Pre-training（SigLIP）の単体モデルおよびそれらの組み合わせを用いた．

### 4.1. 定量的評価

分類精度の評価結果を表 1 に示す．単一モデルの場合と比較して，複数モデルを統合した場合では，多くの組み合わせにおいて分類精度の向上が確認できる．具体的に，DINOv2 と SigLIP の組み合わせを DINOv2 単体と比較した場合を除き，精度の向上が確認できる．さらに，複数モデルを統合した場合，DINOv2 と SigLIP を組み合わせた場合を除いて，実画像を用いた場合を上回る精度となった．また，単一モデルでは分類精度が低かった SigLIP においても，他のモデルと統合することで分類精度の向上がみられた．この結果から，複数モデルの勾配を統合することで，分類に有効な特徴をより効果的に合成画像へ反映することができることがわかった．

表 1: Top1 Accuracy [%]

| 画像の種類 | 画像作成に使用したモデル | | | | 評価モデル |
|---|---|---|---|---|---|
| | DINO 系 | | CLIP 系 | | ResNet50 |
| | v1 | v2 | CLIP | SigLIP | |
| 上限値 | – | – | – | – | 98.46 ± 0.05 |
| 実画像 | – | – | – | – | 95.86 ± 0.05 |
| 合成画像 (単一モデル) | ✓ | – | – | – | 96.24 ± 0.19 |
| | – | ✓ | – | – | 95.68 ± 0.15 |
| | – | – | ✓ | – | 92.28 ± 0.12 |
| | – | – | – | ✓ | 72.38 ± 0.34 |
| 合成画像 (複数モデル) | ✓ | – | ✓ | – | 96.68 ± 0.07 |
| | ✓ | – | – | ✓ | 97.06 ± 0.15 |
| | – | ✓ | ✓ | – | 96.68 ± 0.24 |
| | – | ✓ | – | ✓ | 95.68 ± 0.07 |

### 4.2. 定性的評価

図 2 にフラミンゴの合成画像例を示す．単一モデルによる合成画像では，モデルごとに着目する視覚的要素が異なる．図 2 (a) の DINOv2 では，フラミンゴのシルエットや形状が強調されている．一方，図 2 (b) の CLIP では，背景の水面などのテクスチャ表現が強調されている．これに対し，図 2 (c) の DINOv2 と CLIP の組み合わせでは，形状とテクスチャ情報が含まれているようにも見受けられるが，単一モデルと比較して視覚的な差異は明確ではない．この結果は，分類に有効な特徴が，必ずしも視覚的に識別可能な形で現れないことを示唆している．

(a) DINOv2　(b) CLIP　(c) v2+CLIP

図 2: 作成した合成画像

## 5. おわりに

本研究では，LGM を拡張し，複数の事前学習済みモデルの勾配を同時に最適化する LGM を提案した．定量的評価により，提案手法で作成した合成画像が，単一モデル蒸留と比較して高い分類精度を示すことを確認した．この結果は，提案手法が複数の事前学習済みモデルの特徴を統合し，分類に有効な特徴表現を抽出可能であることを示唆している．今後の展望として，モデル統合時の各バックボーンの重み最適化や中間層の特徴活用による合成画像の質向上，および対象クラス数やデータセットの拡大が挙げられる．さらに，Linear Probing 以外の手法への適用における課題と可能性についても，検証を行う予定である．

## 参考文献

[1] G. Cazenavette, *et al.*, “Dataset Distillation for Pre-Trained Self-Supervised Vision Models”, NeurIPS, 2025.