5X-03

基盤モデルの潜在表現を用いた生成モデルによる知識蒸留

佐々木 達也 [1], 坂井 俊介 [1], 長谷川 達人 [1]

[1] 福井大学

1 はじめに

基盤モデル [1] とは，多様な下流タスクに極めて少ない訓練事例で適応できる大規模な深層学習モデルの総称である．基盤モデルの代表的な例として，Contrastive Language Image Pre-training (CLIP) [2] がある．CLIP は大規模なキャプション付き画像データセット上で，対応する画像とキャプションの埋め込み空間上のコサイン類似度を最大化するようにモデルを訓練する．CLIP は多くの下流タスクにおいて訓練データが与えられない場合でも高い分類性能を示している [2]．CLIP を下流タスクに適用する際の問題は，推論に著しい計算コストを要する点である．この問題は計算資源が限られていたり，リアルタイム性が求められる環境における基盤モデルの活用を妨げている．本研究では CLIP の知識を特定のタスクに特化した上で，より小規模なモデルへと転移する知識蒸留に取り組む．

知識蒸留とは，既に有用な知識を持った教師モデルの持つ知識を別の生徒モデルに効果的に伝達することを目指す学習パラダイムである [3]．本研究では，教師モデルの訓練データセットが大規模であるため，より小規模な下流タスクのデータセットからの知識蒸留を目指す．

Knowledge Distillation Using Generative Models Based on Latent Representations of Foundation Models

[1] Tatsuya Sasaki, Shunsuke Sakai, Tatsuhito Hasegawa (University of Fukui).

しかしながら，下流タスクのデータが十分に得られない場合，知識蒸留による性能改善が小さくなる問題が生じる．本研究では，この問題を画像生成モデル [4, 5] による訓練画像の拡充によって緩和する．既存の基盤モデルからの知識蒸留と比較して，このアプローチは以下のような利点を持つ．

1. 知識蒸留に大規模なデータセットを必要としない
2. 下流タスクに近いドメインのデータセットで蒸留できる
3. 生成時に用いた条件情報を再利用できる

2 提案手法

本研究の概要を図 1 に示す．提案手法では，より下流タスクに特化した知識蒸留を実現するため，生成モデルを用いて訓練画像を拡充する．異なる生成手法と蒸留手法の効果を検証するため，2 種類の生成アプローチ（cls2img, img2img）と 3 種類の知識蒸留（Feature Distillation, Logit Distillation, Condition Prediction）を比較する．

cls2img (A) では，クラス名によるプロンプトに基づき訓練画像を生成し，img2img (B) では，実画像に基づき訓練画像を生成する．知識蒸留のアプローチとしては，(i) Feature Distillation（FD）, (ii) Logit Distillation（LD）の 2 種類を検証する．(i) では入力画像の教師モデルの潜在表現を予測する．(ii) では入力画像に対して計算された出力分布をマッチングする．また，(B) による生成画像を用いた知識蒸留では，各生成画像に対応する実画像が存在するため，対応する実画像の潜在表現を予測する (iii) Conditon Prediction（CP）を

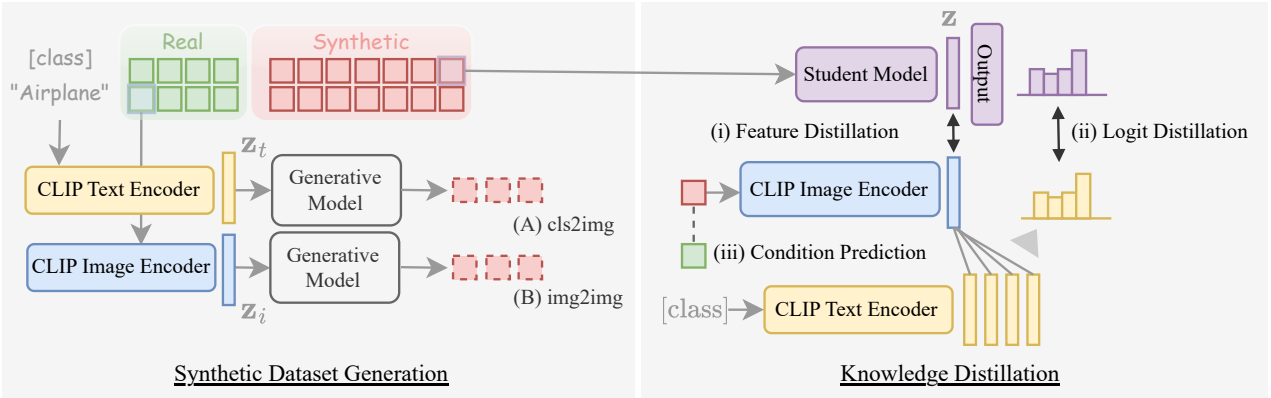

図 1: 本研究で取り組むアプローチの概要

表 1: ResNet50 の代表的な画像分類データセットにおけるテスト精度

	Real			Synthetic [cls2img]		Synthetic [img2img]		
Dataset	**SUP**	**FD**	**LD**	**FD**	**LD**	**FD**	**LD**	**CP**
STL10	69.5	75.9 [+6.4]	69.2 [-0.3]	82.5 [+13.0]	76.9 [+7.4]	**82.9** [+13.4]	76.3 [+6.8]	82.1 [+12.6]
SVHN	95.0	98.0 [+3.0]	98.0 [+3.0]	**98.1** [+3.1]	98.0 [+3.0]	98.0 [+3.0]	98.0 [+3.0]	97.7 [+2.7]
ImageNette	83.3	87.2 [+3.9]	82.7 [-0.6]	90.8 [+7.5]	87.5 [+4.2]	**90.9** [+7.6]	86.6 [+3.3]	89.9 [+6.6]
EuroSAT	94.2	97.5 [+3.3]	96.6 [+2.4]	**98.5** [+4.3]	97.5 [+3.3]	98.3 [+4.1]	97.7 [+3.5]	98.2 [+4.0]
Flowers102	26.2	39.8 [+13.6]	36.9 [+10.7]	**54.3** [+28.1]	48.1 [+21.9]	53.9 [+27.7]	46.6 [+20.4]	50.8 [+24.6]
Food101	70.9	80.4 [+9.5]	72.4 [+1.5]	84.7 [+13.8]	79.6 [+8.7]	**85.1** [+14.2]	78.4 [+7.5]	82.9 [+12.0]
Average	73.2	79.8	76.0	84.8	81.3	**84.9**	80.6	83.6

検証する．本実験では，生成モデルとして Stable Diffusion v2.1 [4] を用いて，実画像の訓練データセットのおよそ 4 倍の生成画像を各データセットについて作成した．生成画像は知識蒸留においてのみ用いて，知識蒸留により事前訓練したモデルを実画像のみを用いて教師あり学習する．

3 実験

我々は提案手法の有効性を検証するため，汎用的な画像分類タスクのデータセットである STL10, SVHN, Imagenette, EuroSAT, Oxford102_Flowers，Food101 を用いた．生徒モデルとして ResNet50 [6] を，教師モデルとして LAION-5B において訓練した OpenCLIP ViT-H [7] を用いた．

ベースラインとして，実画像のみを用いて初期状態から生徒モデルを教師あり学習する場合（SUP）を考える．教師あり学習の事前訓練として知識蒸留を行うことにより，ベースラインを超える性能を達成することを目指す．

表 1 に ResNet50 のテストデータセットにおける分類精度を示す．CLIP からの知識蒸留により，ベースライン（SUP）と比較すると大幅に精度が向上している．また，実画像のみによる知識蒸留と比較して合成画像による訓練データを拡充することでさらに性能を改善できる．FD と LD を比較すると，FD の方が性能が高い傾向がある．

LD については，出力分布を得るためには，画像に加えて CLIP のテキストエンコーダに入力するクラス名を含んだテキストを必要とする．そのクラス名を含むテキストをどのように用意するか (a photo of ”classname”, a painting of ”clasnaame” など) によって出力分布が変化するため，目的のタスクで効果的な蒸留をするためにはテキストエンコーダーに与えるテキストを調整しなければならないという煩雑さが生じる．この点においても，テキストを必要とせず，潜在表現のみで学習をすることができる FD や CP のほうが優れている．

また，CP は LD よりも性能が高いが，FD よりは性能が低い．一方で，CP は各生成画像についての教師モデルの特徴の再計算を必要とせず，画像生成時に用いた潜在表現を教師として扱うことができるため，知識蒸留に要する計算コストが少ない．FD については，各生成画像につき新たに潜在表現を計算するため，計算コストは増加するものの，知識蒸留における教師信号がより多くの情報を含んでおり，CP と比較して性能が高くなる傾向がある．

FD と LD それぞれについて，(A) と (B) の生成方法を比較すると同程度の性能であった．これは生成モデルが検証に用いたデータセットの知識を十分に有しており，プロンプト及び潜在表現による条件付けのいずれにおいても人間が意図している訓練画像を適切に生成していることに基因するものとみられる．

4 おわりに

本研究では，生成モデルを用いて基盤モデルである CLIP からの知識蒸留を行うアプローチを検証した．生成モデルによる生成画像を用いた CLIP からの知識蒸留は多様な画像分類ベンチマークにおいて大幅な性能改善に寄与することを示した．今後は，教師モデルの特徴量に対する後処理や生成モデルの変更などを検証していく．

References

[1] Rishi Bommasani et al. “On the Opportunities and Risks of Foundation Models”. In: *ArXiv* (2021).

[2] Alec Radford et al. “Learning Transferable Visual Models From Natural Language Supervision”. In: *ICML*. 2021.

[3] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. *Distilling the Knowledge in a Neural Network*. 2015.

[4] Robin Rombach et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2021.

[5] Aditya Ramesh et al. *Zero-Shot Text-to-Image Generation*. 2021.

[6] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *CVPR*. 2016.

[7] Mehdi Cherti et al. “Reproducible scaling laws for contrastive language-image learning”. In: *CVPR* (2022).