複数の自己教師あり学習モデルを用いた下流タスクにおける知識蒸留に関する研究

TR22007 鈴木 渉起　　指導教授：藤吉 弘亘

1. はじめに

自己教師あり学習 (SSL) は，大量のラベルなしデータから下流タスクに転移しやすい特徴表現を獲得する事前学習法である．大規模なデータセットを用いた SSL は，教師あり学習を凌駕する性能を達成できるため注目されている．基盤モデル等の SSL モデルは，通常大規模であり，計算コストが高くエッジデバイスでの推論時間やメモリコストが問題となる．そのため，計算コストの低い小規模なモデルが必要とされる．しかし，小規模なモデルは表現能力に欠けているため性能が低い．

本研究では，小規模なモデルの性能向上を目的として，知識蒸留 (KD) [1] による SSL モデルからの知識の転移法を提案する．提案手法は，予備実験により確認した SSL の学習法による特徴表現の差異を基に，複数の SSL モデルを用いて，下流タスクにおける KD を行う．これにより，表現能力の高い小規模なモデルの獲得を目指す．

2. 自己教師あり学習 (SSL)

SSL の学習法には，Contrastive Learning (CL) と Masked Image Modeling (MIM) がある．CL は同じ画像間の特徴量は近づけ，異なる画像間の特徴量は遠ざけるように学習する．一方，MIM は画像の一部をマスクし，マスクした箇所の画素値または特徴量を予測するように学習する．これらの学習法は，獲得する特徴表現が大きく異なることが知られている．そこで，予備実験として CL と MIM の特徴表現の差異を確認する．

予備実験　SSL モデルの Attention Weight（Self-Attention における query と key の関係性）から，Attention distance を算出した結果を図 1 に示す．Attention distance は，層ごとに Attention Weight とピクセル間距離を乗算したものである．これにより，深い層に着目すると，赤色で示す CL は大域的な領域，青色で示す MIM は局所的な領域を認識していることがわかる．この結果から，SSL の学習法によって特徴表現に差異があることがわかる．そこで，それぞれの Attention Weight を小規模なモデルに伝達することで，小規模なモデルにおける表現能力の改善を図る．

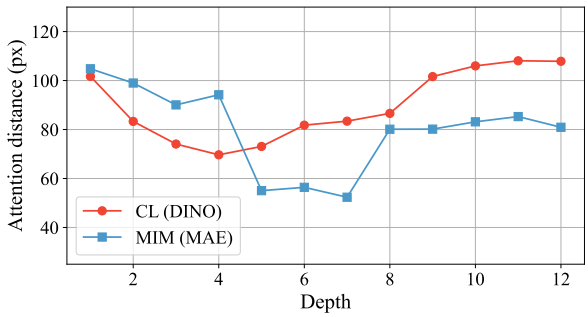

図 1: SSL モデルの Attention distance

3. 提案手法

予備実験を基に，複数の SSL モデルを用いて，下流タスクにおける KD を行うことで，より表現能力の高い小規模なモデルの獲得を目指す．KD は，学習済みモデル (Teacher) の知識を，未学習のモデル (Student) に伝達する手法である．図 2 に提案手法における KD を示す．提案手法では，Teacher として CL と MIM を用いて，2 つのモデルから KD をして Student に転移する．このとき，KD に用いる知識として，従来 KD で用いられるモデルの出力分布と，学習法による特徴表現の差異を確認した Attention Weight を考える．

3.1 Teacher の効率的な学習

学習済みモデルを下流タスクに適用する方法には，線形評価とファインチューニングがある．線形評価は，学習済みモデルにクラス分類のための全結合層 (Head) を結合し，Head 以外を固定して学習する．ファインチューニングは，学習済みモデルも含めて全てのパラメータを学習する．提案手法では，複数の大規模な学習済みモデルを Teacher として用いるため，メモリ消費量の少ない線形評価を用いる．しかし，MIM は線形評価で性能が低く，KD により Student に悪影響を与える．そこで，提案手法では Head に加えて，Transformer encoder の最終層も再学習する．これを Partial-1 と呼ぶ．これにより，MIM における出力分布を用いた KD の効果の改善を期待する．

3.2 出力分布を用いた KD

1 つ目の知識の伝達手段として，Teacher と Student に画像を入力し，得られた出力分布を用いて転移する．損失関数には，一般的な KD と同様に KL ダイバージェンスを用いる．CL と MIM の出力分布を用いた KD を次のように定式化する．

$$\mathcal{L}_{\mathrm{KD}} = \lambda_1 \mathrm{KL}(p^{\mathrm{CL}} \| p^S) + (1-\lambda_1)\,\mathrm{KL}(p^{\mathrm{MIM}} \| p^S), \quad (1)$$

ここで，$p^S, p^{\mathrm{CL}}, p^{\mathrm{MIM}}$ は各モデルの出力分布，λ_1 は CL と MIM の損失の比率を調整する係数である．λ_1 は 0.5 とする．

3.3 Attention Weight を用いた KD

予備実験より，Attention Weight は学習法によって特徴表現に差異があることが判明した（図 1）．そこで，それぞれの特徴表現を小規模なモデルに伝達することで，小規模なモデルにおける表現能力の改善を考える．しかし，モデル構造によって Attention Weight の形状が異なるため，Teacher から Student に Attention Weight を転移することはできない．そのため，2 つ目の知識の伝達手段として，重要な Attention Weight を選択して転移する．

Attention Weight の選択方法　Attention Weight の選択方法として，Attention Confidence (AC) [3] を用いる．ここで，AC は Multi-head attention の重要度を測る指標である．AC による評価値を次のように定式化する．

$$C_h = \frac{1}{|\mathbb{Q}|} \sum_{q \in \mathbb{Q}} \max_{k \in \mathbb{K}} A_h(q,k), \quad (2)$$

ここで，h は Head のインデックス，$\mathbb{Q}$ は query の集合，$\mathbb{K}$ は key の集合，A は Attention Weight である．Teacher の Attention Weight を AC を用いて評価し，最も高く評価された Attention Weight が下流タスクにおいて重要だと考えて Student に転移する．

損失関数　Attention Weight を用いた KD では，KL ダイバージェンスを用いて Student の Attention Weight が Teacher の Attention Weight に近づくように学習する．Attention Weight を用いた KD を次のように定式化する．

$$\mathcal{L}_{\mathrm{Attn}} = \lambda_2 \mathrm{KL}(A^{\mathrm{CL}} \| A_1^S) + (1-\lambda_2)\,\mathrm{KL}(A^{\mathrm{MIM}} \| A_2^S), \quad (3)$$

ここで，A_1^S は Student の 1 番目の Attention Weight，A_2^S は Student の 2 番目の Attention Weight，A^{CL} と A^{MIM} は AC を基準に選択された CL と MIM それぞれの Attention Weight である．λ_2 は CL と MIM の損失の比率を調整する係数であり，λ_2 は 0 とする．

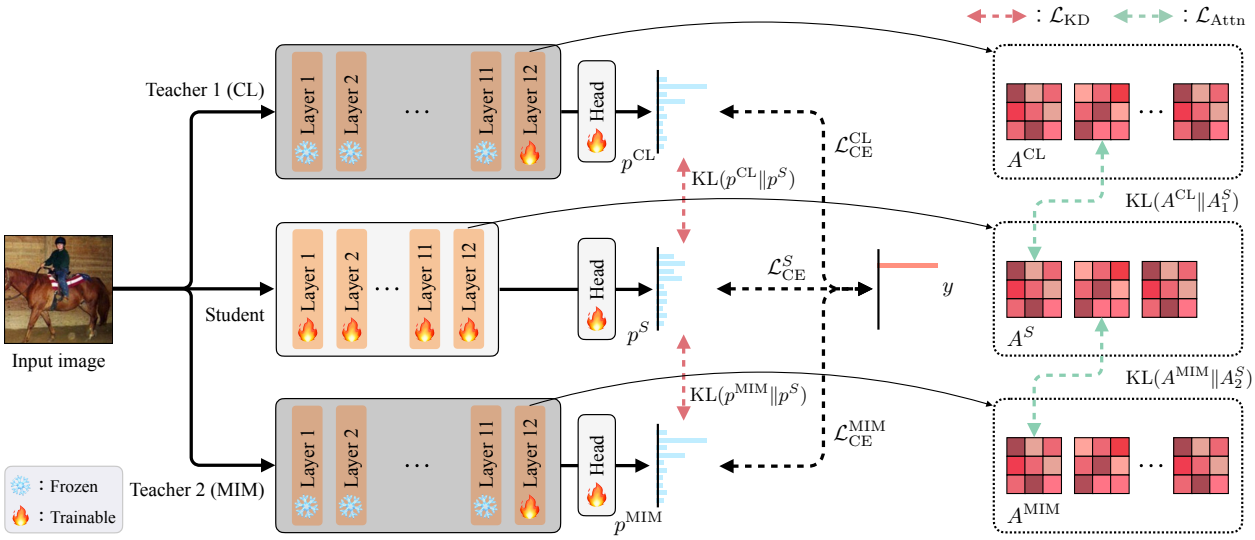

図 2: 提案手法の学習方法

3.4 Teacher の線形評価と KD の同時最適化

従来手法では，SSL モデルを線形評価やファインチューニングした後に KD を行うため，2 段階の学習になる．提案手法では，SSL モデルの線形評価と KD を同時に行うことで，1 段階の学習にする．全体の損失関数 $\mathcal{L}$ を次のように定式化する．

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}}^{S} + \mathcal{L}_{\mathrm{CE}}^{\mathrm{CL}} + \mathcal{L}_{\mathrm{CE}}^{\mathrm{MIM}} + \mathcal{L}_{\mathrm{KD}} + \mathcal{L}_{\mathrm{Attn}}, \qquad (4)$$

ここで，$\mathcal{L}_{\mathrm{CE}}^{S}, \mathcal{L}_{\mathrm{CE}}^{\mathrm{CL}}, \mathcal{L}_{\mathrm{CE}}^{\mathrm{MIM}}$ は正解ラベルと出力分布のクロスエントロピー損失である．

4. 評価実験

本節では，提案手法の有効性を検証するために様々な下流タスクにおける評価実験を行う．データセットとして，ImageNet-1k や CIFAR-10/100 などの多クラス分類用のデータセットを用いる．Student は ViT-Ti，Teacher は CL として DINO [4] と MIM として MAE [5] を用いる．

4.1 Teacher の学習方法の有効性

表 1 に SSL モデルの下流タスクへの適用方法を比較した結果を示す．その結果，提案手法に導入する Partial-1 は，最終層を再学習するのみでファインチューニングに匹敵する性能を達成している．

表 1: ImageNet-1k を用いた下流タスクへの適用 (%)

Method	Trainable Params (M)	CL	MIM
ファインチューニング	86.57	**81.92**	**81.70**
線形評価	0.77	76.31	54.02
Partial-1	7.86	80.26	79.30

4.2 従来手法との比較

ImageNet-1k を用いて評価した結果を表 2 に示す．Baseline は KD を行わずに Student が単体で学習したモデル，KD はファインチューニングした Teacher を用いて蒸留したモデル，Ours は複数の SSL モデルを用いて KD を行ったモデルである．Ours は Baseline から 2.2 pt，KD から 1.3 pt 向上した．

表 2: ImageNet-1k による評価結果 (%)

Method	Teacher	Top-1 Accuracy
Baseline (DeiT [2])	–	72.2
KD	DINO	73.1
	MAE	73.1
Ours	DINO, MAE	**74.4**

様々な下流タスクで評価した結果を図 3 に示す．Ours は，DTD と VOC2007 を除いたデータセットで KD よりも高い精度を達成した．

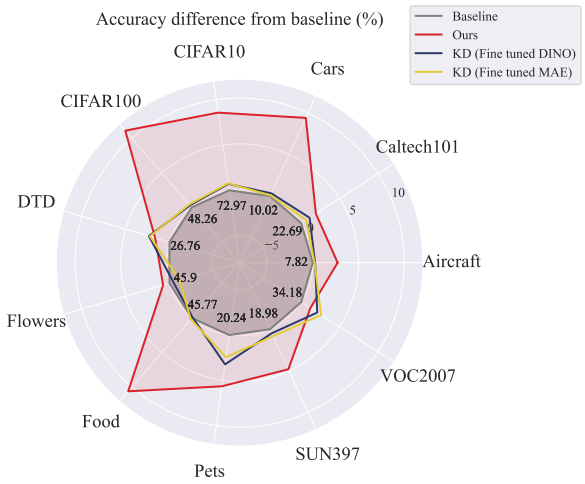

図 3: 様々な下流タスクにおける評価結果

5. おわりに

本研究では，小規模なモデルの精度向上を目的として，複数の自己教師あり学習モデルを用いた知識蒸留を提案した．評価実験により，様々な下流タスクにおいて提案手法が小規模なモデルの精度向上に有効であることを示した．今後は，さらなる精度向上を目指して，蒸留方法の多様化を行う．

参考文献

[1] G. Hinton, *et al.*, “Distilling the Knowledge in a Neural Network”, NIPS Workshop, 2015.

[2] H. Touvron, *et al.*, “Training data-efficient image transformers & distillation through attention”, ICML, 2021.

[3] E. Voita, *et al.*, “Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned”, ACL, 2019.

[4] M. Caron, *et al.*, “Emerging Properties in Self-Supervised Vision Transformer”, ICCV, 2021.

[5] K. He, *et al.*, “Masked Autoencoders Are Scalable Vision Learners”, CVPR, 2022.

研究業績

[1] 鈴木渉起 等, “Vision Transformer の相互学習による高精度化”, 画像の認識・理解シンポジウム, 2022.