CLIP におけるモーダル間の知識蒸留による構造化枝刈り

ER21061 西川実希　　指導教授：藤吉弘亘

1. はじめに

Contrastive Language-Image Pre-Training（CLIP）[1] は，画像とテキスト間の共通する特徴表現を学習する Vision-Language Model（VLM）である．このモデルは画像エンコーダとテキストエンコーダが独立した構造であるため，モデル全体のパラメータ数が膨大である．そのため，精度を維持しつつモデルを軽量化することが求められる．本研究では CLIP が持つ画像とテキスト間の共通する特徴表現を維持する構造化枝刈り手法を提案する．この手法によりモデルの軽量化と精度維持の両立を実現する．

2. Unified and Progressive Pruning（UPop）

CLIP を対象とした枝刈り手法として UPop [2] がある．UPop は枝刈り箇所の探索を行い，枝刈り後に微調整を行うことでモデル精度を回復させる手法である．UPop の枝刈りではまず，モデルに学習可能なマスク $\boldsymbol{\zeta}$ を作成する．CLIP の損失関数 L_{CLIP} とハイパーパラメータ w から構成する式 (1) の損失関数を用いて，モデル精度への寄与が少ない枝刈り箇所の探索を行う．

$$L_{\text{UPop}} = L_{\text{CLIP}} + w \sum_{\boldsymbol{\zeta}_i \in \boldsymbol{\zeta}} \| \boldsymbol{\zeta}_i \|_1 \tag{1}$$

UPop は，モデル精度への寄与が少ないパラメータに対してマスクの値を 1 から 0 に向けて段階的に減少させるとともに，更新するマスクの範囲，すなわち枝刈り箇所を増加させることで，最終的に設定した枝刈り率に到達させる．しかし，UPop は枝刈り箇所が増えるにつれてモデル精度が著しく低下してしまうという課題がある．この要因として枝刈り箇所の探索によってモデル精度への寄与が大きい構造が削除されることで，枝刈り前のモデルが捉えていた有益な情報や複雑なパターンが失われることが考えられる．

3. 提案手法

大規模なモデルの知識を転移しながら別のモデルを学習する技術として知識蒸留がある．本研究では，UPop において枝刈り箇所を増やす際に生じる精度低下を抑制するために，CLIP モデルに合わせた知識蒸留により枝刈り前の知識を活用する手法を提案する．具体的には，図 1 に示すように，枝刈り前モデルの特徴表現を維持するために，以下の 3 つのモーダル間蒸留損失を加えて UPop による枝刈りを行う．

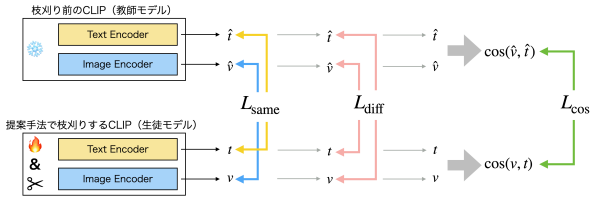

図 1：CLIP のモーダル間の特徴表現を考慮した知識蒸留

- 同一モーダル間の蒸留損失：L_{same}
  枝刈り前モデルと枝刈り中のモデルが抽出した特徴ベクトルを一致させることで，枝刈り前モデルの特徴表現を維持することを目的とする．枝刈り前モデルが抽出した特徴ベクトル $\hat{\boldsymbol{v}}$，$\hat{\boldsymbol{t}}$ と，枝刈り中のモデルが抽出した特徴ベクトル $\boldsymbol{v}$，$\boldsymbol{t}$ 間の MSE 誤差を計算する．同一モーダル間の蒸留損失を式 (2) に示す．

$$L_{\text{same}} = \frac{1}{2}(\text{MSE}(\hat{\boldsymbol{v}}, \boldsymbol{v}) + \text{MSE}(\hat{\boldsymbol{t}}, \boldsymbol{t})) \tag{2}$$

- 異なるモーダル間の蒸留損失：L_{diff}
  異なるモーダル間で知識蒸留を行うことにより，CLIP の目的に合った蒸留損失を計算することを目的とする．教師と生徒モデルが抽出した特徴ベクトルを用いてコサイン類似度 $\cos(\hat{\boldsymbol{v}}, \boldsymbol{t})$，$\cos(\hat{\boldsymbol{t}}, \boldsymbol{v})$ を計算し，クロスエントロピー誤差関数により蒸留損失を計算する．異なるモーダル間の蒸留損失を式 (3) に示す．

$$L_{\text{diff}} = \frac{1}{2}(\text{CE}(\cos(\hat{\boldsymbol{v}}, \boldsymbol{t})) + \text{CE}(\cos(\hat{\boldsymbol{t}}, \boldsymbol{v}))) \tag{3}$$

- コサイン類似度の蒸留損失：$L_{\cos}$
  枝刈り前モデルが捉えた特徴表現を直接維持することを目的とする．例えば，画像とテキストのペアは必ずしも 1 対 1 の関係ではなく，複数のテキストが正例である場合がある．そのような表現は枝刈り前モデルがそれまでの学習過程で獲得していると考えられるため，枝刈り前モデルが捉えた関係性を維持するように知識蒸留を行う．枝刈り前のコサイン類似度 $\cos(\hat{\boldsymbol{v}}, \hat{\boldsymbol{t}})$ と枝刈り後の $\cos(\boldsymbol{t}, \boldsymbol{v})$ を計算し，MSE 誤差により蒸留損失を計算する．コサイン類似度の知識蒸留を式 (4) に示す．

$$L_{\cos} = \text{MSE}(\cos(\hat{\boldsymbol{v}}, \hat{\boldsymbol{t}}), \cos(\boldsymbol{v}, \boldsymbol{t})) \tag{4}$$

3 つの蒸留損失と UPop による枝刈り箇所探索の損失を集約した損失関数を式 (5) のように定義する．

$$L_{\text{total}} = L_{\text{UPop}} + L_{\text{same}} + L_{\text{diff}} + L_{\cos} \tag{5}$$

4. 評価実験

本研究の有効性を検証するために UPop と提案手法の比較を行う．

4.1. 実験条件

枝刈り対象モデルには，画像エンコーダが Vision Transformer，テキストエンコーダが Transformer で構成される CLIP を使用する．データセットには COCO データセットを用いる．枝刈り率は 75%，バッチサイズは 6，マスクの探索と微調整はどちらも 6 エポックとする．タスクは画像検索タスク（t2i 検索）とテキスト検索タスク（i2t 検索）で，Top-1 の正解率を用いてモデル精度を評価する．また特徴表現を維持できているか，2 つのデータにおける相関関係の一致度を測るピアソン相関係数 ρ を用いて確認する．

4.2. 実験結果

枝刈りモデルとそれの微調整後のモデル精度を表 1 に示す．表 1 より，提案手法を用いることで，枝刈り後の精度低下を抑制できた．また，微調整後の精度も提案手法が優れている．ピアソン相関係数 ρ を比較すると，提案手法が高いため，枝刈り前モデルの特徴表現を維持できる枝刈り箇所の探索が出来ているといえる．

表 1：COCO データセットによる検索タスクの精度

	微調整	t2i 検索	i2t 検索	ρ
UPop		15.68	15.43	
	✓	62.48	46.54	0.53
提案手法		22.04	32.30	
	✓	**65.28**	**53.94**	**0.93**

5. おわりに

本研究では，UPop による枝刈りと知識蒸留を組み合わせた学習を行うことでモデルの精度を向上させることができた．今後は，より最適な知識蒸留を行いモデル精度の向上を目指す．

参考文献

[1] A. Radford, *et al.*,“Learning Transferable Visual Models From Natural Language Supervision”, ICML,2021

[2] D. shi, *et al.*,“UPop: Unified and Progressive Pruning for Compressing Vision-Language Transformers”, ICML,2023