機械知覚&ロボティクスグループ
中部大学 工学部　ロボット理工学科

MPRG
MACHINE PERCEPTION AND ROBOTICS GROUP

2023年度　藤吉研究室　卒業論文発表　アブストラクト

Long-tailed Classification，Contrastive Learning，Representation Learning

混合比率を用いた対照学習による Long-tailed 物体認識における表現学習の向上

柴田 祐汰

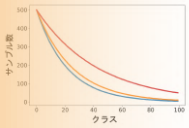
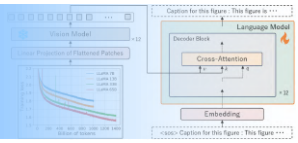


Image Captioning，Vision-Language，Transformer

Attention Weight の操作による注目領域に着目したグラフ図のキャプション生成

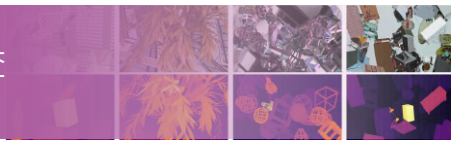
増田 大河



Stereo Matching，CG Datasets Analysis

深層学習ベースのステレオマッチングに有効な学習データに関する調査

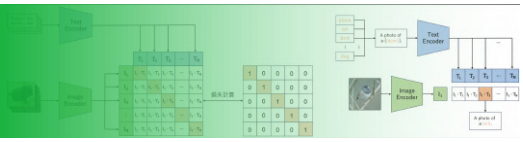
松原 太地



Contrastive Learning，Vision and Language，Data Augmentation

CLIP 学習における画像データ拡張の有効性の調査

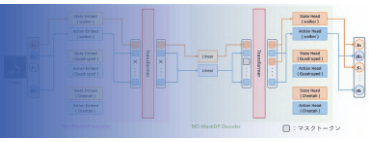
西尾 優希



Reinforcement Learning，Self-Supervised Learning

深層強化学習における MaskDP を用いたマルチドメイン事前学習

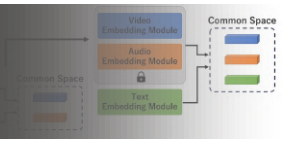
鈴木 佳三



Multimodal Learning，Self-Supervised Learning，Contrastive Learning

マルチモーダル自己教師あり学習におけるモーダルの組み合わせによる学習効果

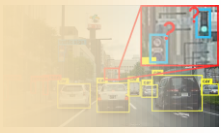
小林 亮太



Semi-Supervised Learning，Object Detection

教師なしデータを活用した半教師あり学習による遠方の物体検出

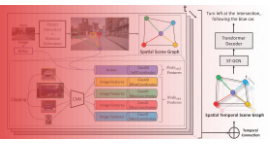
田中 隆聖



Deep Learning，Graph Neural Networks，Human-like Guidance

物体の詳細情報を考慮した時空間シーングラフによる案内文生成

鈴木 颯斗



Deep Learning，Position Estimation，Robotics

Transformer モデルを用いた 2D 手書き指示による把持位置推定

野田 修平

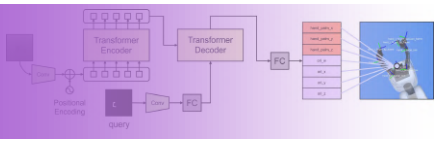
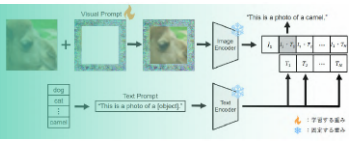


Image Classification，Visual Prompt

Visual Prompt の付与方法による物体認識精度への影響

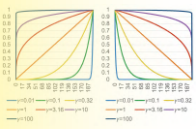
松﨑 恵也



Collaborative Learning，Knowledge Distillation

知識転移の時間軸方向への重み付け戦略の最適化

原田 優輝



URL: http://mprg.jp

混合比率を用いた対照学習による Long-tailed 物体認識における表現学習の向上

ER20042 柴田祐汰　　指導教授：藤吉弘亘

1. はじめに

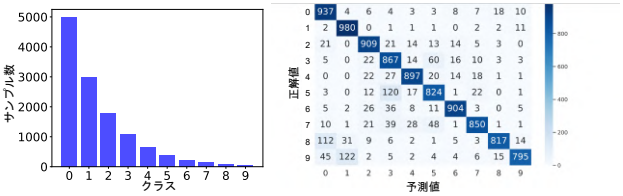
Long-tailed 物体認識は，クラス毎のサンプル数に偏りがあるデータを対象とする．代表的な手法である GLMC[1] は，教師あり学習と表現学習のマルチタスク学習によって，Long-tailed 物体認識の精度を向上した．表現学習では，mixup と CutMix を施した画像同士の特徴量の類似度を最大化する．これにより，混合したクラスの特徴表現の一貫性を向上する一方で，異なるクラス間の特徴分布が近くなる．特に，混合する頻度が少ない少数派クラスに対して表現学習が十分に発揮されず，多数派クラスの認識精度と大きな差が生じる．そこで，本研究では，対照学習を用いた特徴表現を向上する学習法を提案する．

2. 従来手法 (GLMC)

GLMC [1] は，入力画像に mixup と CutMix というデータ拡張を施し，それらに対して教師あり学習と表現学習を行う手法である．教師あり学習では，通常の分類損失とサンプル数を考慮した損失の割合をエポック毎に調整し，多数派クラスに対する過度な学習を緩和する．表現学習では，混合した画像同士の特徴量のコサイン類似度を最大化し，クラス毎の一貫した特徴表現を獲得する．GLMC は，この 2 つのタスクを同時に学習することで，Long-tailed 物体認識における高精度化を実現した．

3. GLMC におけるモデルの予測傾向の調査

GLMC が高精度化を達成した一方で，多数派クラスと少数派クラスの精度に大きな差が存在する問題がある．GLMC で CIFAR-10-LT を評価した際の評価結果を図 1 に示す．図 1(a) は，学習データの分布である．図 1(b) より，少数派クラスを多数派クラスと多く誤認識する傾向があることが分かる．これは，mixup と CutMix が，混合したクラスの中間的な特徴を学習して識別能力を向上させる一方で，混合したクラスの特徴分布が近くなるからである．また，各クラスの混合する頻度は，クラスのサンプル数に依存する．そのため，少数派クラスと多数派クラス間での識別境界が学習しきれず，特徴分布同士が近づくことで誤分類を誘発したと考える．以上より，混合した画像の特徴量の類似度を最大化する表現学習だけでは不十分であると考える．



(a) 学習に使用したデータ分布　(b) 評価データに対する混同行列

図 1: 傾向調査の結果

4. 提案手法

本研究では，少数派クラスの精度向上を目的とし，特徴表現を向上する手法を提案する．提案手法では，3 節で述べた問題を解決するために，混合前後の画像の特徴量に対する対照学習を GLMC に導入する．この時，混合前後の画像の特徴量に対する類似度を混合比率に基づいて最大化し，それ以外の特徴量に対する類似度は最小化する．これにより，混合前後の画像の特徴関係を考慮した学習と，混合前の画像を基準としたクラス毎の特徴分布を離す学習を可能にする．提案手法の流れを図 2 に示す．具体的には，各イテレーション毎に 3 つのステップを行う．

Step1. 混合画像の生成

mixup，CutMix を混合比率 λ に基づいて異なるミニバッチ間で行い，混合画像を生成する．

Step2. 特徴量を抽出

モデルに混合前後の画像を入力し，特徴量を抽出する．ここで，混合前の画像の特徴量を $z^{(1)}$ と $z^{(2)}$，mixup と CutMix した画像の特徴量を $\tilde{z}^{(m)}$，$\tilde{z}^{(c)}$ とする．

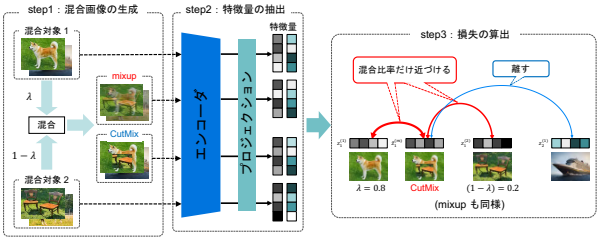


図 2: 提案手法の概念図

Step3. 損失の算出

まず，各特徴量のコサイン類似度を算出後，式 (1) に示す simCLR の NT-Xent Loss に基づいて混合前後の損失を算出する．ここで，sim(z, $\tilde{z}$) は z と $\tilde{z}$ のコサイン類似度，τ は温度パラメータ，N はバッチサイズを表す．

$$L_{(z,\tilde{z})} = \frac{1}{2N}\sum_{i=1}^{2N}(-\log\frac{\exp(\mathrm{sim}(\mathrm{z_i},\tilde{\mathrm{z}}_i)/\tau)}{\sum_{k=1}^{2N}\exp(\mathrm{sim}(\mathrm{z_i},\tilde{\mathrm{z}}_k/\tau)}) \quad (1)$$

式 (1) と混合比率 λ を使用して，mixup した画像及び CutMix した画像に対する損失 L_{mixup}，L_{cutmix} を算出する．損失 L_{mixup} と L_{cutmix} を式 (2) に示す．

$$\begin{aligned} L_{mixup} &= \lambda L_{(z^{(1)},\tilde{z}^{(m)})} + (1-\lambda)L_{(z^{(2)},\tilde{z}^{(m)})} \\ L_{cutmix} &= \lambda L_{(z^{(1)},\tilde{z}^{(c)})} + (1-\lambda)L_{(z^{(2)},\tilde{z}^{(c)})} \end{aligned} \quad (2)$$

最後に，これらの平均を求めて，提案手法の総損失を算出する．提案手法の総損失 L_{CLR} を式 (3) に示す．

$$L_{CLR} = (L_{mixup} + L_{cutmix})/2 \quad (3)$$

5. 評価実験

Long-tailed 物体認識における提案手法の有効性を評価する．

5.1. 実験概要

本実験で使用するデータセットは，不均衡率 ρ を [100,50] とした CIFAR-100-LT である．エンコーダには，ResNet-32 を使用する．ミニバッチサイズは 128，エポック数は 200 とする．比較手法は，GLMC，提案手法とする．評価指標は，学習サンプル数に基づいて 3 つのクラスに分割した Many-shot，Med-shot，Few-shot グループの Accuracy とする．また，各評価指標は seed 値を 5 回変更して学習を行った結果の平均値とする．

5.2. 実験結果

比較結果を表 1 に示す．表 1 より，どちらの不均衡率でも Few-shot，Med-shot の精度が向上したことが分かる．この結果より，提案手法は少数派クラスの精度改善が可能であると考えられる．一方，Many-shot の精度低下は，提案手法による同じクラスに属する画像の特徴量も離す作用が，多数派クラスに大きく影響するためと考える．

表 1: CIFAR-100-LT における各精度 [%]

ρ	手法	Many-shot	Med-shot	Few-shot	平均
100	GLMC	**74.55**	57.36	37.96	57.34
	提案手法	71.41	**58.01**	**41.04**	**57.44**
50	GLMC	**75.14**	57.58	42.55	61.92
	提案手法	74.92	**58.19**	**43.14**	**62.19**

6. おわりに

本研究では，混合比率を用いた対照学習による表現学習を向上する手法を提案し，少数派クラスの認識精度が向上することを確認した．今後の展望として，Mnay-shot の精度低下を抑制するために，特徴空間で離す対象を学習データのラベル情報を用いて選択するように改善する．

参考文献

[1] F. Du, *et al.*, “Global and Local Mixture Consistency Cumulative Learning for Long-tailed Visual Recognitions”, CVPR,2023.

Attention Weight の操作による注目領域に着目したグラフ図のキャプション生成

ER20080 増田大河　　指導教授：藤吉弘亘

1. はじめに

研究開発における生産性向上のため，学術論文等の文献から研究内容を理解して説明する AI の実現が期待されている．学術論文のグラフ図は多くの文字情報を含んでおり，研究内容を理解するための重要な要素である．画像を言語化する Vison-Language の研究が盛んに取り組まれているが，多くは一般物体画像を対象としてる．そのため，グラフ図を言語で説明することは困難である．そこで本研究では，大量のグラフ図を用いて，グラフ図の理解に必要な知識を獲得した基盤モデルをファインチューニングして，グラフ図のキャプションを生成するモデルを提案する．また，グラフ図の注目領域を明示することでより詳細なキャプションの生成を目指す．

2. MatCha

MatCha[1] は，グラフ図に対する多くのタスクに適応できる基盤モデルとして提案されている．MatCha は，グラフ図を理解する上で必要な (i) グラフ図からその描画プログラム，テーブルデータへの逆変換，(ii) 数学的問題推論の 2 つのタスクによって事前学習している．MatCha のモデル構造を図 1 に示す．事前学習によりグラフ図の理解や数学的推論に特化しているが，グラフ図に対する質問応答を行う ChartQA などの下流タスクに適応するにはファインチューニング (FT) が必要となる．

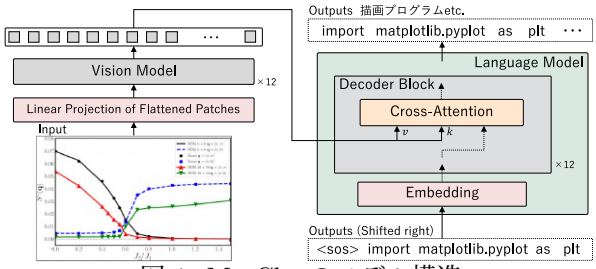


図 1: MatCha のモデル構造

3. 提案手法

グラフ図のキャプション生成を行うモデルを構築し，グラフ図の注目領域を強調することで詳細なキャプションを生成する方法を提案する．

モデルの構築　MatCha を FT することでグラフ図のキャプションを生成するモデルを構築する．MatCha によるグラフ図の表現を活用してキャプションを生成するために，画像エンコーダ部分は凍結して言語生成モデル部分のみを FT する．

注目領域の強調方法　人間が指定した任意の領域に対する詳細なキャプションを生成するために，指定した注目領域に対応するパッチの Attention Weight の値に対して γ を加算する．これにより注目領域に着目した特徴抽出を行う．注目領域を強調した Attention を式 (1) に示す．

$$\mathrm{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \left(\mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right) + \mathbf{M}\times\gamma\right)\mathbf{V} \tag{1}$$

ここで，$\mathbf{M}$ は注目領域に対応するパッチ部分が 1，それ以外のパッチ部分が 0 になっているマスクである．注目領域の強調方法を図 2 に示す．これにより，画像全体としての特徴を捉えつつ，注目領域にも着目したキャプションが生成できる．

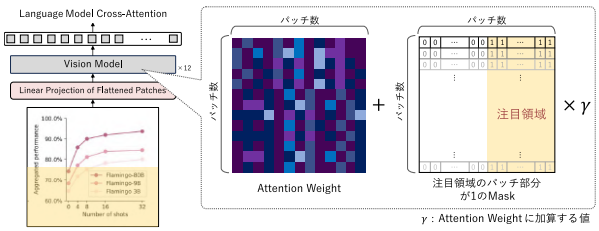


図 2: 注目領域の強調方法

4. 評価実験

提案手法の有効性を示すために，従来手法とのキャプション生成性能の比較を行う．また，指定領域の強調による生成文の変化について確認する．従来手法には M4C-Captioner を用いる．提案手法と従来手法はグラフ図とキャプションなどから構成される SciCap+データセットを用いて FT した．

4.1. キャプション生成性能の比較

キャプション生成の評価指標による比較結果を表 1 に示す．これより，ROUGE-L で 5.7pt，CIDEr で 26.1pt の精度の向上を確認した．BLEU-4 の評価が低下しているが，CIDEr の評価が向上していることから，重要な単語を抜き出すことが出来ていることがわかる．

表 1: キャプション生成の評価指標による比較

モデル	BLEU-4	ROUGE-L	CIDEr
M4C-Captioner †	**1.5**	15.4	4.6
Ours	1.3	**21.1**	**30.7**

†:SciCap+論文値

4.2. 注目領域の強調による生成文の変化

生成したキャプションに対応するグラフ図の領域を定量的に評価する．生成文中に評価領域に存在する単語が含まれる割合を評価する．x，y 軸周辺にそれぞれ注目することを期待して，強調する領域を画像の下半分と左半分とした．加算する値 γ を 12.5×10^{-5} とした時の結果を表 2 に示す．ここで，γ の値はあらかじめグリッドサーチによって探索した値である．表 2 より，強調なしと比較して，画像の下半分を強調した場合は下半分の数値が増加している．一方で，画像の左半分を強調した場合は左半分の数値が減少し，下半分の数値が増加した．

表 2: 生成文中に評価領域の単語が含まれる割合 (%)

強調する領域	評価領域		
	全体	左半分	下半分
強調なし	21.5	**22.6**	21.6
左半分	22.0	21.8	**22.8**
下半分	**22.1**	22.2	**22.8**

注目領域の強調による生成文の変化を図 3 に示す．これより，修正なしの場合と比べて，注目領域として強調した部分に含まれる単語を生成文に多く含んでいることが確認できる．これらの結果より，注目領域に対応する画像エンコーダの Attention Weight に値を加算することで，注目領域に合わせたキャプションの生成が可能である．

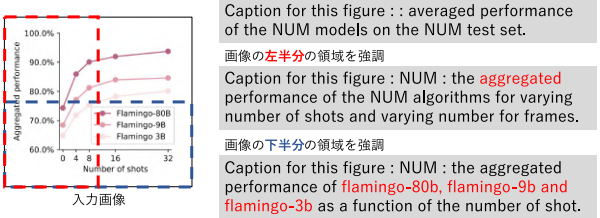


図 3: 注目領域の強調による生成文の変化

5. おわりに

本研究では，グラフ図を詳細に説明するモデルの構築と，その注目領域の強調方法を提案した．評価実験より，構築したモデルは既存手法と比較してグラフ図のキャプション生成精度が高いことを確認した．また，注目領域の強調によって，より詳しい文章生成が可能になることを確認した．今後はモデルに関連文も入力することで，キャプションの生成精度の向上を図る．

参考文献

[1] Fangyu Liu, *et al.*, “MatCha: Enhancing Visual Language Pretraining with Math Reasoning and Chart Derenderingn”, In ACL, 2023.

深層学習ベースのステレオマッチングに有効な学習データに関する調査

ER20084 松原太地　　指導教授：藤吉弘亘

1. はじめに

深層学習ベースのステレオマッチング手法に，Cascaded REcurrent Stereo matching network (CREStereo) [1] がある．CREStereo は独自に作成した大規模な CG データセット (CREStereo データセット) を用いて学習することで，従来手法と比較して高い精度とロバスト性を確保している．CREStereo データセットは，CG 環境に様々なオブジェクトをランダムに配置して，多様性の高いデータで構成されている．本研究では，CREStereo データセットを作成したシーンに焦点を当て，どのようなシーンが精度向上に寄与するかを調査する．

2. CREStereo データセット

CREStereo はモデルとデータセットの両者に対して精度向上の観点からアプローチを提案している．モデルは低解像度から高解像の特徴マップを推論時に階層的に利用する．低解像度画像の特徴で全体構造を理解し，高解像度画像の特徴で細部の特徴の抽出を実現している．データセットは 200,000 枚の学習データから構成されている．図 1 に CREStereo データセットの例を示す．従来のデータセットと比較して，CREStereo データセットは生物，木，穴があるものなどの多種多様な物体，照明条件，反射などの実環境の要素をランダムに含むシーンを生成している．

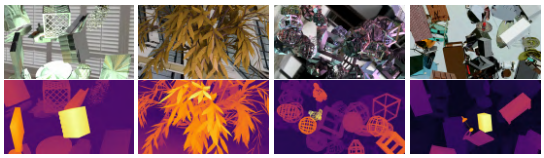


図 1: CREStereo データセットの例

3. データの傾向調査

従来研究では，データセットの改善による精度・ロバスト性の向上を主張するにも関わらず，データセットの工夫点について十分な議論が行われていない．そこで，CREStereo データセットを要素ごとに分割し，各要素がモデルの学習に与える影響を調査する．CREStereo データセットは，Hole，Reflective，Shapenet，Tree の 4 つの要素から構成されている．本研究では，データの傾向の調査の手法として，すべての要素で学習，要素を 1 つずつ除外し学習した結果を比較して構成要素の学習に与える影響を調査する．評価データには，ステレオマッチングのベンチマークで広く利用される屋内と屋外のシーンを多く含む実環境データである Middlebury Stereo データセットを利用する．

4. 評価実験

本実験では，データセットが学習に与える影響を検証をするために，データセットを分割し，分割後のデータで学習・推論を行う．定量的な評価指標には Average Error と Bad を使用する．Average Error は予測した視差と正解画像間における差の平均値を評価する．Bad はピクセルごとに予測視差と正解画像の値の差を求め，差の絶対値が N 以上あるピクセルの割合を評価する．

4.1. 実験条件

学習用データセットには CREStereo データセット，評価用データセットには Middlebury Stereo データセットを用いる．データ分割による傾向調査のために学習データの 4 つの構成要素をそれぞれ除外して学習を行う．最適化手法は Adam，バッチサイズは 4，イテレーション数は 300,000 回とする．評価指標の Bad の N は 2 とする．

4.2. 実験結果

表 1 に学習データを要素ごとに分割した際の定量的評価結果を示す．表 1 より，Reflective の要素を含むデータを除外した場合と全データを学習した場合を比較すると，精度が約 3.5pt 向上していることが確認できる．Reflective は多様なシーンを表現するためランダムに反射条件を決定し，図 2 のような実環境では再現性のない過剰な反射を含む画像がある．このような再現性のない画像を除外することが精度向上に有効であることが確認できた．加えて再現性のない画像は特徴マップ上で Reflective の重心付近に多く分布しており，そこを除外することで精度が向上することを確認できた．また，Shapenet と Hole を除外した際に精度が約 8.0pt 低下した．これより，2 つの要素が実環境データに適しており，CREStereo データセットの精度向上に大きく影響していると考えられる．以上より，実環境データにおいて Hole と Shapenet が学習に有効であることを確認した一方で，Reflective の要素は学習に有効でないことが分かった．

表 1: 定量的評価

学習データ	Average Error ↓	Bad 2.0 ↓
全データ	30.373	58.883
Reflective 除外	**26.791**	**57.013**
Hole 除外	38.355	64.895
Shapenet 除外	38.211	64.123
Tree 除外	30.942	52.373

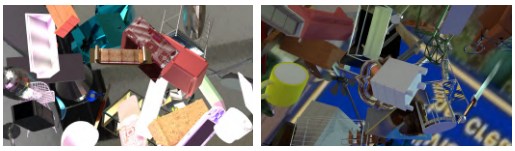
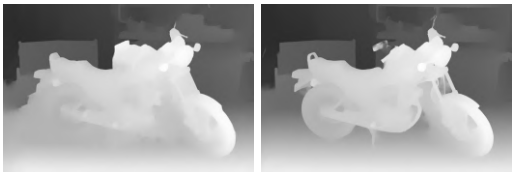


図 2: Reflective データの例

4.3. 定性的評価

図 3 に定性的評価結果を示す．図 3(b) は正解画像，図 3(c) と図 3(d) はそれぞれ CREStereo データセットを全て学習した結果，Reflective を除外したデータで学習した結果である．図 3(c) および図 3(d) より，Reflective を除外したデータで推論を行った際に物体の輪郭や背景領域のぼやけの低減が確認できた．これは，Reflective を除外することで，背景領域を含め再現性のない反射条件の画像を除外し，物体の輪郭や背景領域に対してより有効なデータで学習することができたからである．

(a) 元画像　　(b) 正解画像



(c) 全データ　　(d) Reflective 除外

図 3: 定性的評価

5. おわりに

本研究では，CREStereo データセットの傾向調査を行い，深層学習ベースのステレオマッチングの精度向上につながるシーンを調査した．実験結果から，学習データから Reflective を除くことで精度が向上することを確認し，各データセットを定性的に確認した．今後はより詳細なデータの分析を行い，汎化能力と精度を両立するデータの選定を行う予定である．

参考文献

[1] Jiankun Li, *et al.*, “Practical Stereo Matching via Cascaded Recurrent Network with Adaptive Correlation”, In CVPR, 2022

CLIP 学習における画像データ拡張の有効性の調査

ER20062 西尾優希　　指導教授：藤吉弘亘

1. はじめに

CLIP は，画像とテキストを組み合わせた対照学習手法であり，大量の画像と画像に関連するテキストを用いて学習を行う．学習方法として，ミニバッチ内の画像とそれに対応するテキストを正例，それ以外を負例として対照学習を行う．学習済みの CLIP は，事前学習で使用していないデータに対する zero-shot の画像分類が可能である．一方で，CLIP の学習にはどのようなデータ拡張が精度向上に寄与するか分かっていない．そこで本研究では，CLIP のモデルを学習する時のデータ拡張が zero-shot 分類に有効であるかを調査する．

2. CLIP

CLIP は，画像とテキストのペアを用いたマルチモーダルな対照学習法の 1 つである．CLIP における損失関数を式 (1) に示す．

$$E = -\log \frac{\exp(\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{N} \exp(\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)} \tag{1}$$

N はバッチサイズ，$\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_j)$ は正例ペアの特徴ベクトルの類似度，$\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_k)$ は正例と負例ペアの特徴ベクトルの類似度を表す．CLIP の学習方法を図 1 に示す．CLIP では，画像と画像に対応するテキストを使用して，ミニバッチ内の画像に対応するテキストを正例，それ以外のテキストを負例として対照学習を行い，画像とテキストに共通する特徴表現を学習する．学習済みモデルを画像分類タスクに適用する場合，Image Encoder には画像を，Text Encoder には，「a photo of a [CLS]」のような文章をクラス数分入力し，画像と文章の類似度が最も高いものを分類したクラスとする．CLIP には，事前学習時にどのようなデータ拡張をすることで精度が向上するか明確ではないという問題がある．そこで，本研究では異なるデータ拡張を画像に対して適用し，効果的なデータ拡張方法について調査する．

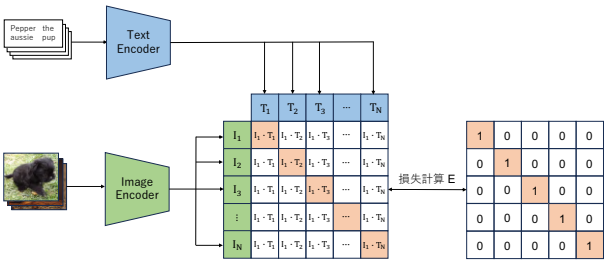


図 1：CLIP の学習方法

3. 実験概要

本研究では，データ拡張に RandomResizedCrop (RRC) とコントラスト変換を使用して CLIP の事前学習を行う．本実験で用いる RRC は画像の 90%から 100%の領域をランダムで切り抜き，コントラスト変換は画像のコントラスト値を-30%から 30%の間でランダムに変化させる．データ拡張を適用する前の画像を図 2(a)，RRC を元画像に適用した例を図 2(b)，コントラスト変換を元画像に適用した例を図 2(c) に示す．図 2 では，実験で用いるより強いデータ拡張を適用し，元画像との変化を明確にする．

(a) 元の画像　　(b) RRC　　(c) コントラスト変換

図 2：データ拡張後の画像

モデルには ResNet-50，学習用データセットには，Conceptual 12M (CC12M)[2]，評価用データセットには ImageNet-1K (IN-1K)，CIFAR-10 (CI-10)，CIFAR-100 (CI-100)

図 3：CC12M データセット

を使用する．CC12M データセットの画像とテキストの例を図 3 に示す．CC12M データセットでは，Google Cloud Natural Language API を用いてテキストに含まれる人物に関する情報を ⟨PERSON⟩ に置換する．評価タスクは zero-shot 分類である．学習時に加えるデータ拡張の種類とエポック数に応じてデータ拡張の種類を変化させた場合の精度変化を調査する．また，学習時のバッチサイズは 256，エポック数は 32，学習率は 0.0001 である．

3.1 評価結果

データ拡張の種類による zero-shot 画像分類の精度変化を表 1 に示す．表 1 より，RRC やコントラスト変換を適用した学習は，データ拡張なしと比べて精度が向上しており，CLIP においてデータ拡張が有効だとわかる．また，データ拡張の種類により精度が異なり，RRC が効果的であることがわかる．

表 1：データ拡張の種類による精度比較 [%]

データ拡張		評価用データセット		
RRC	コントラスト変換	IN-1K	C-10	C-100
		27.3	29.0	12.1
✓		**31.1**	**48.5**	**15.5**
	✓	29.3	40.6	12.0

次に，RRC とコントラスト変換の組み合わせによる zero-shot 画像分類の精度変化を表 2 に示す．表 2 より，RRC とコントラスト変換の両方を適用した学習は，RRC のみの適用と比べて精度が低い．一方で，コントラスト変換を 16epoch 目から導入すると，0epoch 目から導入した場合と比べて CIFAR-10 以外の精度が低下する．これは，複数のデータ拡張を組み合わせることで困難な学習となり，精度が低下したと考えられる．そのため，データ拡張の強度や適用のタイミングを学習状況に合わせて動的に変更する等の処理が必要である．

表 2：エポック数に応じてデータ拡張を変更した場合の精度

データ拡張		評価用データセット		
RRC	コントラスト変換	IN-1K	CI-10	CI-100
✓		**31.1**	**48.5**	15.5
✓	0epoch から導入	30.5	38.1	**17.4**
✓	16epoch から導入	29.9	44.6	12.9
✓	24epoch から導入	30.4	12.9	12.1

4. おわりに

本稿では，CLIP におけるデータ拡張による精度変化について調査した．データ拡張として RRC を用いることで，データ拡張を適用しない場合と比較して精度が向上し，CLIP においてデータ拡張が有効であると分かった．しかし，データ拡張を加えるタイミングにより精度が著しく変わるため，基準を設けてデータ拡張の強度を変えていく必要があると考えられる．今後は，他の種類のデータ拡張の検証や，パラメータ探索により決定したデータ拡張の強度での事前学習，学習時の誤差に応じてデータ拡張の強度や確率を変更するカリキュラム学習に取り組み，精度向上を目指す．

参考文献

[1] Alec Radford, *et al.* "Learning Transferable Visual Models From Natural Language Supervision ", ICML, 2021.

[2] Soravit Changpinyo, *et al.* "Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts", CVPR, 2021.

深層強化学習における MaskDP を用いたマルチドメイン事前学習

ER20045 鈴木佳三　　指導教授：藤吉弘亘

1. はじめに

自然言語処理やコンピュータビジョン分野では，多種多様なタスクに適応可能な基盤モデルが注目されている．基盤モデルは，膨大なデータを用いた大規模な事前学習後，特定のタスクにファインチューニングすることで様々な下流タスクに適応できる．同様に，深層強化学習においても様々なタスクを解くことができる基盤エージェントモデルの実現が期待されている．その 1 つのアプローチとして，深層強化学習における自己教師あり学習法である Masked Decision Prediction (MaskDP) [1] が提案されている．MaskDP では事前学習と下流タスクが同一のドメインである必要がある．そのため，適応可能な下流タスクが限定される．そこで本研究では，異なる複数ドメインの経験を用いた Multi-Domain MaskDP を提案する．様々なドメインの経験を用いた学習により各ドメインに共通する特徴を獲得し，下流タスクにおける学習効率の向上及び基盤エージェントモデルの実現を図る．

2. Masked Decision Prediction (MaskDP)

MaskDP は，深層強化学習における自己教師あり学習法である．MaskDP のネットワーク構造を図 1 に示す．MaskDP は，状態と行動のシーケンスデータに対して Masked Autoencorder (MAE) [3] のようにマスク箇所の再構成を行う事前学習手法である．Encoder では一部の領域をマスクしたシーケンスデータから特徴抽出を行う．Decoder では Encoder で抽出した特徴とマスクトークンから元の状態及び行動を再構成する．

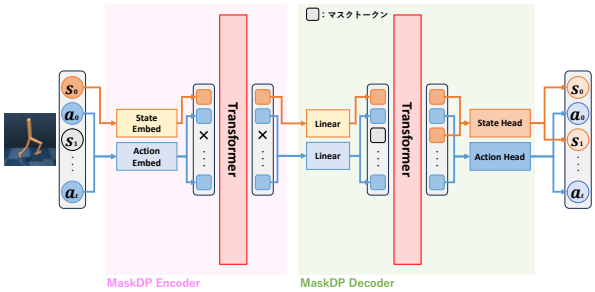


図 1：MaskDP の概略

3. 提案手法

異なるドメインのシーケンスデータから共通した特徴を獲得できる，自己教師あり事前学習手法 Multi-Domain MaskDP (MD-MaskDP) を提案する．MD-MaskDP のネットワーク構造を図 2 に示す．MaskDP による事前学習に異なるドメインのシーケンスデータを用いた場合，ドメイン間における入出力次元数の違いが問題となる．そこで，Encoder 部の埋め込み層をドメインごとに独立して構築する．入力データに適した埋め込み層を用いることで，Transformer Block へ入力する特徴の次元数を統一する．また，Decoder 部も同様に，状態及び行動を再構成する Head をドメインごとに構築し，ドメインに適した再構成を行う．これにより，ドメイン間の入力及び出力次元数の違いによる問題を解決する．

4. 評価実験

事前学習なしである scratch，MaskDP，MD-MaskDP におけるファインチューニング時の報酬推移を比較する．

4.1 実験条件

強化学習環境には DeepMind Control Suite [2] の Walker, Quadruped, Cheetah を使用する．事前学習の対象ドメインとして，MaskDP では Walker，MD-MaskDP では Walker, Quadruped, Cheetah を用いる．学習条件はバッチサイズを 384，イテレーション数を 400,000，学習率を 1e-4 とする．事前学習データセットは，内部報酬に基づきエージェントが収集した対象ドメインの経験データを使用

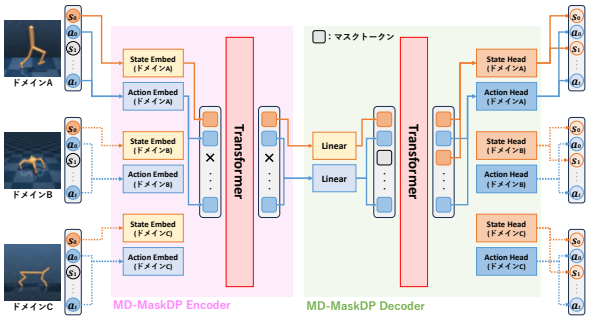


図 2：Multi-Domain MaskDP の概略

する．本データはタスクに依存しないドメイン固有の経験である．下流タスクとして，Walker の stand, walk, run タスクを用いたオフライン強化学習によるファインチューニングを行う．学習条件はバッチサイズを 384，イテレーション数を stand タスクでは 100,000，walk タスクでは 200,000，run タスクでは 1,000,000，学習率を 1e-4 とする．ファインチューニングデータセットは，外部報酬に基づきエージェントが収集した Walker における対象タスクでの経験データを使用する．本データはタスクに依存した経験データである．

4.2 報酬推移による比較

ファインチューニング時の報酬推移を図 3 に示す．図 3 より，stand 及び walk タスクにおいて，MaskDP と MD-MaskDP は同等の報酬を獲得しており，学習初期で scratch より高い報酬を獲得している．このことから，提案手法によりドメイン間で共通した特徴を捉えることができ，Walker のみを対象とした事前学習と同等の効果を得られたと考えられる．一方，run タスクではいずれのエージェントも学習初期から中期にかけて同等の報酬を獲得している．これは，MD-MaskDP 及び MaskDP による事前学習が不足しており，stand 及び walk タスクより難易度の高い run タスクでの事前学習の効果は低いといえる．また，学習終盤では MD-MaskDP は MaskDP より高い報酬を獲得したが，scratch より報酬が低下している．この現象の原因は現時点では明らかになっておらず，今後の追加実験で原因調査を行う予定である．

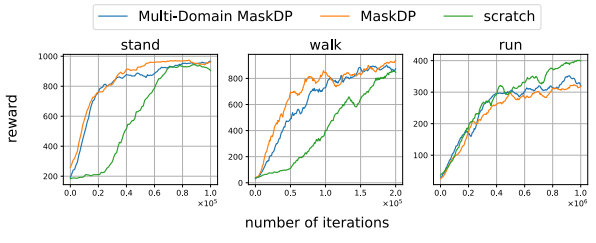


図 3：Walker における対象タスクでの報酬推移

5. おわりに

本研究では，基盤エージェントモデルの実現を目的とし，複数ドメインの経験を用いた自己教師あり学習法である MD-MaskDP を提案した．評価実験では，提案手法による下流タスクでの学習効率向上を確認できた．今後は，run タスクでの問題について調査を行う．また，ドメイン数の増加及び他ドメインでの評価実験に取り組む．

参考文献

[1] F. Liu, *et al.*, “Masked Autoencoding for Scalable and Generalizable Decision Making”, NeurIPS, 2022.

[2] Y. Tassa, *et al.*, “dm_control: Software and tasks for continuous control”, Software Impacts, 2020.

[3] H. Kaiming, *et al.*, “Masked Autoencoders Are Scalable Vision Learners”, CVPR, 2022.

マルチモーダル自己教師あり学習におけるモーダルの組み合わせによる学習効果

ER20038 小林亮太　　指導教授：藤吉弘亘

1. はじめに

深層学習では，人間の認知プロセスのように複数の感覚 (モーダル) を組み合わせたマルチモーダル学習が注目されている．マルチモーダル学習は，同時刻に収集した異なるモーダルデータを用いて学習する．しかしながら，同時刻における各モーダルデータは完全一致していないことも多い．例えば，テキストにはオーディオやビデオに含まれる雑音のようなノイズとなる情報が存在せず，モーダルごとに性質が異なる．そのため，モーダルの持つ特性を考慮したモデルの設計を行うことで，雑音を軽減した学習が期待できる．また，マルチモーダル学習では使用するデータの多さから，ラベルなしデータのみを用いる自己教師あり学習が用いられることが多い．そこで，本研究ではマルチモーダル自己教師あり学習において，モーダルの組み合わせ方による学習への影響調査を行う．

2. Multimodal Clustering Network (MCN)

マルチモーダル自己教師あり学習の代表的な手法として Multimodal Clustering Network (MCN) [1] がある．MCN のアーキテクチャを図 1 に示す．MCN では，各モーダルのデータを特徴抽出器に入力して特徴を抽出し，各モーダルの特徴を線形射影をして特徴空間に埋め込む．

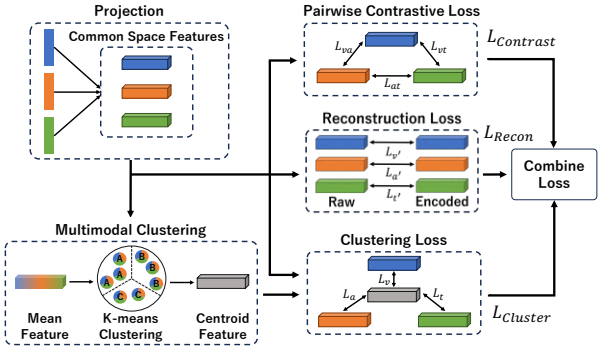


図 1：MCN のアーキテクチャ

損失計算には，式 (1) に示すような $L_{Contrast}$ と $L_{Cluster}$ と L_{Recon} の 3 つの損失関数の総和 $L_{Combine}$ を用いる．

$$L_{Combine} = L_{Contrast} + L_{Cluster} + L_{Recon} \quad (1)$$

$L_{Contrast}$ は，各モーダルのペア（テキスト，オーディオ），（オーディオ，ビデオ），（ビデオ，テキスト）において，動画内で同じ時刻に存在するもの同士を近づけ，そうでないもの同士を遠ざける損失である．$L_{Cluster}$ は K-means 法でクラスタリングを行って各クラスタの重心を算出した上で各モーダルと重心を意味的に近づける損失を表す．L_{Recon} は各モーダルにおいてオートエンコーダによる再構築の前後のデータを近づける損失を表す．これらの損失を同時に使用することによって，異なる時系列における行動や動作の類似性を確保することができる．MCN では，全てのモーダルで単一の空間を利用していることにより検索タスクなどをモーダル間の隔たりなく実行できるという利点が存在する．しかし，共通の空間を利用することは，全てのモーダルが同じ表現力や詳細度もつことを暗黙のうちに仮定しているため，モーダルごとの性質が十分に考慮されていない．

3. 調査実験

ビデオ，オーディオ，テキストのモーダルの組み合わせによるマルチモーダル自己教師あり学習への影響を調べることを目的とする．

3.1 実験概要

本実験では，3 つのモーダルを二段階に分けて学習する．一段階目では，2 モーダルのみで学習をし，二段階目では学習済みの 2 つのモーダルを凍結させ，未使用のモーダルを追加して学習する．以上の手順での学習を 3 つのパターンで比較する．

3.2 実験条件

学習用データセットには HowTo100M[2]，評価用データセットには YouCook2[3] と MSR-VTT[4] を用いて，テキストからのビデオの検索タスクでゼロショット評価を行う．HowTo100M は，Youtube 上のビデオを利用する大規模なデータセットである．一部が削除や非公開されているため，利用可能なビデオの数は減少している．今回の実験では，現時点で残存する約 90 万本のビデオを用いる．ビデオは全て 454 × 256 ピクセルの解像度と 30FPS のフレームレートに統一する．ビデオの特徴の抽出には ResNet152，オーディオの特徴の抽出には DaveNet，テキストの特徴の抽出には Word2vec で学習済みのモデルを用いる．学習条件は，学習率が 0.0001，バッチサイズが 128，エポック数が 30，最適化手法は Adam とした．

4. 実験結果

実験の結果を表 1 に示す．ビデオを V，オーディオを A，テキストを T を表す．最初の 2 つのアルファベットは一段階目の学習，最後のアルファベットは二段階目の学習に用いるモーダルを示す．R は再現率であり，@の後の値は，総正解数に対する各クエリの上位 k 個の予測のうちの正解数の割合を表す．表 1 より，YouCook2 では AT_V が最も高い精度である．YouCook2 は料理の調理手順を扱うデータセットであるため，手順を表現する上でテキストが重要な情報を持つ可能性が高く，一段階目でテキストを用いることが効果的だったと考える．また，オーディオも料理手順の順序性を表現できるため、テキストとオーディオの組み合わせが最適だったと考える．一方で，MSR-VTT では AV_T が最も高い精度である．MSR-VTT はビデオとテキストの説明文からなるデータセットであるため，ビデオそのものの内容の理解が重要であると考えられることから，一段階目でビデオを用いることが効果的だったと考える．このことから，タスクによって適切なモーダルの組み合わせは異なると言える．

表 1：精度比較

	YouCook2			MSR-VTT		
	R@1	R@5	R@10	R@1	R@5	R@10
AV_T	1.61	5.97	9.43	**0.18**	**0.92**	**1.63**
AT_V	**5.16**	**11.3**	**15.3**	0.14	0.61	1.35
VT_A	1.10	5.13	7.13	0.14	0.53	1.14

5. おわりに

本研究では，マルチモーダル自己教師あり学習において，モーダルの組み合わせ方による学習効果への影響についての調査を行った．データセットにより適切なモーダルの組み合わせが異なることが分かった．今後は，引き続きモーダルの組み合わせ方による学習効果への影響について調査を行う，最終的にはマルチモーダル自己教師あり学習におけるモーダルの性質に応じた対照学習の設計による事前学習の性能改善を行う．

参考文献

[1] B. Chen, *et al.*, “Multimodal Clustering Networks for Self-supervised Learning from Unlabeled Videos”, ICCV, 2021.

[2] A. Miech, *et al.*, “HowTo100M:Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips”, ICCV, 2019.

[3] L. Zhou, *et al.*, “Towards automatic learning of procedures from web instructional videos”, AAAI, 2018.

[4] J. Xu, *et al.*, “MSR-VTT: A large video description dataset for bridging video and language”, CVPR, 2016.

教師なしデータを活用した半教師あり学習による遠方の物体検出

ER20057 田中隆聖　　指導教授：藤吉弘亘

1. はじめに

車載カメラ画像からの物体検出では，遠方物体の検出精度が低下するという問題がある．これは，学習データに含まれる遠方物体のラベル付きデータが少ないためである．そのため，大量の遠方物体が含まれるラベル付きデータが必要である．しかしながら，遠方物体のデータ収集にはコストがかかり，正解ラベルの付与も負担が大きい．このコストを低減するための手法として半教師あり学習がある．半教師あり学習は，ラベル付きデータとラベルなしデータの両方を利用してモデルを学習する．しかし，半教師あり学習手法によって遠方物体の学習データは増加するが，近方物体のデータ数も増加する．その結果，近方物体を優先的に学習するため遠方物体の学習に良いデータの影響が少なくなる．そこで本研究では，物体の距離に応じた重みを損失関数に重み付けする Far Object Weight 関数 (FOW) の提案する．そして，半教師あり学習に FOW を導入し，ラベルなしデータを用いて遠方物体の検出精度の向上を図る．

2. Unbiased Teacher v2

Unbiased Teacher v2 [1] は，代表的な半教師あり学習による物体検出手法である．図 1 に Unbiased Teacher v2 の構造を示す．Unbiased Teacher v2 は，Burn-In Stage と Teacher-Student Mutual Learning Stage の 2 つのステージで構成される．Burn-In Stage では擬似ラベルを生成するための Teacher を学習する．Teacher-Student Mutual Learning Stage では Teacher が生成した擬似ラベルを用いて Student の学習を行う．このとき，Listen 2 Student により，Student の予測よりも Teacher の予測が正しい場合にのみ重みを更新する．Student の重みを更新後，指数移動平均を用いて Student の重みをもとに Teacher の重みを更新する．

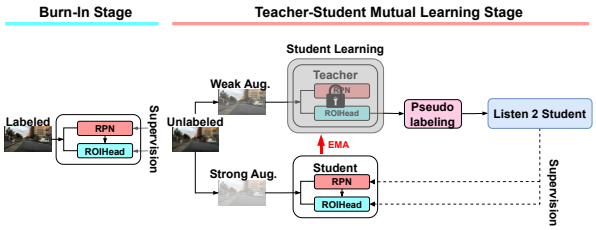


図 1：Unbiased Teacher v2 の構造

3. 提案手法：Far Object Weight 関数

遠方物体の検出精度を向上させるため，物体の距離に応じた重みを損失関数に重み付けする Far Object Weight 関数 (FOW) を提案する．FOW では，クラスごとに基準面積を求め，遠方と近方を判別する．基準面積よりも Bounding Box(BBox) の面積が大きいほど近方，小さいほど遠方と仮定して重みを計算する．半教師あり学習における重みの計算には，教師あり学習の段階では正解ラベルを用い，教師なし学習の段階では擬似ラベルを用いて重みを計算する．クラス (cls) の $FOW(cls)$ を式 (1) に示す．

$$FOW(cls) = \begin{cases} \frac{n}{N(cls)} \log(\frac{A_r(cls)}{A}) + 1.0 & \text{if } A \leq A_r(cls) \\ 1.0 & \text{otherwise} \end{cases} \quad (1)$$

ここで A は BBox の面積，A_r はクラスごとの基準面積，N はクラスごとのサンプル数，n はクラス内の最小サンプル数である．基準面積はラベル付きデータ集合に含まれる BBox の面積をクラスごとで昇順に並び替え，下位数%以下の面積を平均した値で決定する．

4. 評価実験

提案手法の有効性を評価するために，提案手法導入前と導入後の精度比較を行う．また，本実験では半教師あり学習手法として Unbiased Teacher v2，データセットには BDD100K を用いて実験を行う．BDD100K は米国で撮影された約 10 万枚の車載画像で構成されたデータセットである．

4.1. 実験条件

ラベル付きデータ数を 1%，5%，10%の場合で実験を行う．Burn-In Stage のイテレーション数を 30,000 回，Teacher-Student Mutual Learning Stage のイテレーション数を 75,000 回に設定する．提案手法を教師あり学習の損失関数と教師なし学習の損失関数に導入して実験を行う．また，式 (1) の基準面積を下位 1%に設定する．評価指標には mAP を用いる．

4.2. 実験結果

表 1 にラベル付きデータの数が異なる場合の検出精度を示す．表 1 から提案手法を導入することでラベル付きデータ 1% では AP，AP_S，AP_M の精度が向上した．5%では AP，AP_S，AP_M の精度が向上し，10%では全体的に精度が向上した．AP_S は小さな物体に対する精度であるため，遠方の小さな物体に対する精度が向上していることが確認できる．また，表 2 にクラス別の検出精度を示す．表 2 から 1%の場合は bus，truck，bicycle 以外のクラスの精度が向上，5%では rider 以外の精度が向上，10%では rider 以外の精度が向上した．

表 1：ラベル付きデータ数ごとの検出精度

ラベル付きデータ数	1%				5%				10%			
FOW	AP	AP_S	AP_M	AP_L	AP	AP_S	AP_M	AP_L	AP	AP_S	AP_M	AP_L
	21.69	9.09	26.27	**40.19**	24.86	10.33	28.99	**44.74**	27.55	12.15	31.65	49.07
✓	**21.73**	**9.25**	**26.30**	39.77	**24.96**	**10.65**	**29.27**	44.58	**27.72**	**12.59**	**32.13**	**49.29**

表 2：クラス別の検出精度

ラベル付きデータ数	FOW	pedestrian	car	rider	bus	truck	bicycle	motorcycle	traffic light	traffic sign
1%		24.60	43.25	14.56	**30.50**	**27.58**	**15.48**	11.23	19.77	30.57
	✓	**24.81**	**43.48**	**15.31**	29.59	27.06	15.24	**11.39**	**20.15**	**31.04**
5%		26.76	45.23	**18.32**	36.65	35.21	17.66	14.34	20.61	32.66
	✓	**26.96**	**45.85**	17.86	**36.79**	**35.59**	**18.13**	**15.77**	**20.92**	**32.91**
10%		**29.85**	47.57	**20.76**	41.11	38.68	19.87	18.53	22.93	35.35
	✓	**29.99**	**47.71**	20.56	**41.49**	**39.06**	**20.49**	**19.76**	**23.13**	**35.47**

ラベル付きデータ 10%での定性的評価を図 2 に示す．図 2 から，遠方にある car と traffic sign を従来手法では検出していないが，提案手法では検出していることが確認できる．

図 2：定性的評価

5. おわりに

本研究では，遠方の物体の検出精度の向上を図るために Far Object Weight 関数を提案した．提案手法を導入することでラベル付きデータ数 1%，5%，10%において遠方の小さな物体に対する検出精度の向上が確認できた．しかしながら，近方物体の検出精度が低下した．今後は，近方物体の検出精度低下を抑制する方法を検討する予定である．

参考文献

[1] Yen-Cheng Liu, Chih-Yao Ma, Zsolt Kira, "Unbiased Teacher v2: Semi-supervised Object Detection for Anchor-free and Anchor-based Detectors", CVPR, 2022.

物体の詳細情報を考慮した時空間シーングラフによる案内文生成

ER20046 鈴木颯斗　　指導教授：藤吉弘亘

1. はじめに

自動車に広く利用されるナビゲーションシステムは，GPS と地図データを基盤として周辺情報をルールベースに処理して案内を行う．そのため，複雑な周辺状況では運転者の誤解を招く可能性がある．一方で，人間が案内を行う場合，視界に含まれる動的な物体などの情報を用いて指示することが可能である．こうした人間のような案内を行うことでドライバーの直感的な理解を目指したアプローチが Human-like Guidance(HlG) である．本研究では，HlG の実現に向け，シーン画像から認識した物体をノードとしたシーングラフとして表現して文章生成モデルにより視界情報に基づいた案内文の生成を行う．

2. Transformer を用いた文章生成

Transformer は，Attention 機構を導入した言語モデルである．その柔軟性と拡張性は高く，言語とは異なる種類のデータに適応することができる．CPTR[1] は，画像入力に対応した Transformer ベースモデルであり，与えられた画像について説明する文章の生成が可能である．一方で，画像に多数の物体が含まれる場合，曖昧な文章が生成される可能性が高い．HlG のような特定の物体に注目した文章を生成するには画像の入力では不十分である．

3. 提案手法

本研究では，空間情報及び時系列情報を時空間シーングラフとして表現し，HlG に適した案内文の生成手法を提案する．本手法の概略図を図 1 に示す．

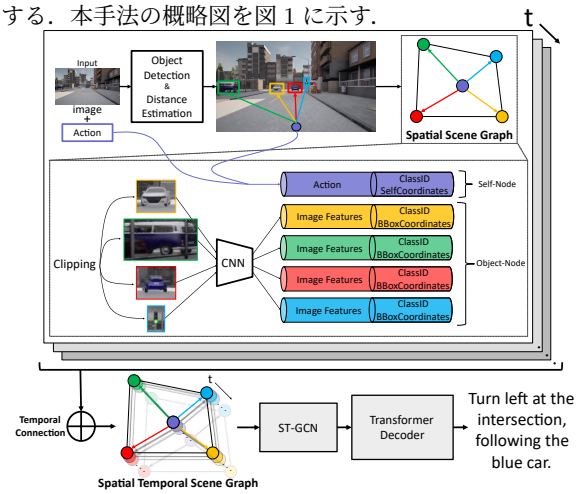


図 1: 提案手法

3.1. 時空間シーングラフの生成

始めに，YOLOv8 を用いて画像に含まれる物体（車両や信号機）の検出を行う．検出した物体をオブジェクトノードとしてグラフ化する．各ノードには，YOLOv8 が出力した ClassID と Bounding Box(BBox) 座標，抽出した BBox 領域内の特徴を用いる．また，自車両との関係性を考慮するために自車両を示すセルフノードを追加する．セルフノードには，ClassID，自車両座標とともに自車両の動作情報“right”，“left”，“straight”）を用いる．オブジェクトノードとセルフノードを接続するエッジは，BBox サイズと実際の物体サイズを考慮して算出した物体間距離を重みとして与える．このシーングラフを各時刻で求めて時空間シーングラフとする．時系列接続を行うことで時間情報を考慮したグラフとなる．

3.2. 文章生成モデル

時空間シーングラフを ST-GCN[2] に入力し，物体の位置関係や時間的な変化を捉えた特徴量を抽出する．そして，ST-GCN で抽出した特徴量を Transformer-Decoder に入力し，案内文を生成する．学習時，Transformer-Decoder は入力シーンに対応した案内文と抽出したグラフ特徴量から単語列を逐次的に推測してデータセットに含まれる案内文の統計的な特徴を獲得する．推論時，グラフ特徴量と文章の開始を示すトークンが入力され，獲得した特徴を基に単語列を逐次的に推測することで案内文を生成する．

4. データセット

本研究の有効性を評価するために，CARLA Simulator を用いて HlG 用のデータセットを作成した．全 160 シーン，計 10,219 フレームで構成されており，交差点付近の走行シーンと案内文，交差点における動作情報が含まれる．案内文は，先頭に “Turn left” 等の動作情報，その後ろに適切な物体を中心とする指示を含む形式とする．

5. 評価実験

提案手法の有効性を検証する評価実験を行う．

5.1. ベースラインと比較条件の設定

シーン画像のみを用いた手法をベースラインとする．また，両方のアプローチにおいて時系列情報の有無による比較を行う．各モデルの詳細を表 1 に示す．

表 1: 提案手法とベースラインのモデル構造

	時系列	Encoder	Decoder
ベースライン		ResNet-18	Transformer
	✓	ResNet3D-18	Transformer
提案手法		GCN	Transformer
	✓	ST-GCN	Transformer

5.2. 実験設定

エポック数を 100，学習率を 0.0001 とし，時系列情報を含む場合は入力を 5 時刻分とする．また，評価指標として BLEU，METEOR，ROUGE を用いる．

5.3. 実験結果

定量的評価を行った結果を表 2 に示す．表 2 より，時系列情報を含む提案手法が最も高精度であることが確認できる．

表 2: 評価結果

	時系列	Bleu-4	METEOR	ROUGE
ベースライン		0.202	0.512	0.533
	✓	0.180	0.514	0.566
提案手法		0.209	0.575	0.598
	✓	**0.224**	**0.603**	**0.624**

案内文生成結果の例を図 2 に示す．図 2 より，提案手法において時系列情報を含めることで注目対象の進行方向を考慮した詳細な案内方法による文章が生成された．これは，時空間シーングラフにより注目対象とその動作の関係を捉えることが可能になったためと考えられる．

図 2: 各モデルの案内文の生成例

6. おわりに

本研究では，シーン画像から時空間シーングラフを生成する方法を提案し，文章生成モデルによる案内文生成と精度の評価について検証を行った．結果より，注目対象を中心とした適切な案内文の生成を実現した．今後は，案内方法の表現を上げるために正解案内文の増強を検討する．また，HlG に適した評価指標の検討を行う．

参考文献

[1] W. Liu, *et al.*, “CPTR: Full Transformer Network for Image Captioning”, CVPR, 2021.

[2] S. Yan, *et al.*, “Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition”, AAAI, 2018.

Transformer モデルを用いた 2D 手書き指示による把持位置推定

ER20066 野田修平　　指導教授：藤吉弘亘

1. はじめに

生活支援ロボットの実現には，誰でも簡単に操作可能なインターフェースが重要である．岩永らは，人の手書きのスケッチ画像からロボットの把持位置を CNN で推定する手法を提案している [1]．シーンの深度画像とスケッチ画像を同時に入力して畳み込み処理を行うため，スケッチの位置のみを変更した場合に対応できない場合がある．そこで，本研究では Transformer[2] を用いた Encoder-decoder モデルによる把持位置推定を提案する．Encoder では，深度画像からシーンの中の各物体を解析する．Decoder では，Encoder のシーン解析を結果とスケッチ画像との関係性を Cross-Attention により求めることで，物体とスケッチの位置関係を正確に算出する．

2. スケッチ画像を用いた把持位置推定

岩永らは，簡単かつ直観的な指示が可能なスケッチ画像を用いた把持位置推定を提案している [1]．岩永らのモデル構造は ResNet18 をベースにしており，入力はシーンの深度画像と物体に対して書かれたスケッチ画像をチャンネル方向に重ね合わせた 2 チャンネル画像である．出力は，グリッパの手首，左指先，右指先の 3 点の 3 次元座標である．本手法は，深度画像とスケッチ画像のペアをあらかじめ用意する必要がある．しかし，物体を把持する位置は 1ヶ所とは限らない．そのため，学習していない深度画像とスケッチ画像のペアを入力した場合，位置関係を正確に理解できない問題がある．

3. 提案手法

本研究では，Transformer をベースとする把持位置推定モデルを提案する．提案手法のモデル構造を図 1 に示す．まずシーンの深度画像を 3 層の畳み込み層で処理する．そして，得られた特徴マップに Positional Encoding を加えたものを Encoder に入力する．Encoder では物体や床の位置関係を特徴量として抽出する．スケッチ画像は 3 層の畳み込み層で処理した後，全結合層でベクトルに変換して Decoder に入力する．出力は，グリッパの手首の位置の 3 次元座標と姿勢のクォータニオンである．Decoder では，Encoder で得た特徴量とスケッチの入力データとの Cross-Attention を求めることで，正確かつ汎化性能の高い把持位置推定を実現する．

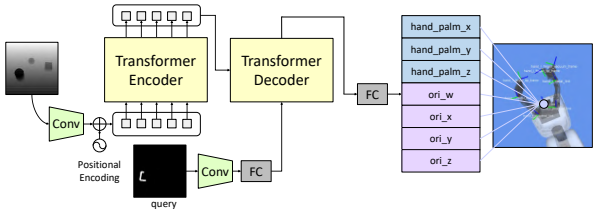


図 1: スケッチ画像を入力とする Transformer モデル

4. 評価実験

従来手法 [1] と提案手法の比較により，有効性を検証する．

4.1. 実験条件

学習の最適化手法には Adam，エポック数を 200，バッチサイズを 128，学習率を 0.0001 として学習を行う．損失関数は位置の 3 次元ユークリッド距離の誤差と姿勢の平均二乗誤差の和を用いる．

4.2. データセットの作成

本実験では Unity 環境で作成したデータを用いる．把持対象物体の深度画像は，3 種類の物体をランダムな位置に出現させ，オブジェクトの位置に応じてカメラの角度を変えて撮影した画像である．スケッチ画像はグリッパを模したコの字型オブジェクトを把持対象物体の周辺に配置し，スケッチ線を中心に円を配置した画像とする．1 枚の深度画像に対して異なる 10 通りのスケッチを用意する．データ数は それぞれ訓練データ 104,000 枚，検証データ 13,000 枚，テストデータ 13,000 枚である．

4.3. 実験結果

表 1 に従来手法と提案手法の位置と姿勢の推定誤差の平均値と最小値，最大値を示す．表 1 から平均値は同程度であるが，提案手法の最大値が大きく低下していることが確認できる．このことから従来手法では対応できなかった場合も対応できるようになったと考えられる．この例を図 2，図 3 に示す．

表 1: 誤差の比較

	平均値	最大値	最小値
従来手法	0.060	**2.168**	0.007
提案手法	0.061	**0.619**	0.004

図 2 は従来手法による推論結果の例，図 3 は提案手法による推論結果である．左 2 枚と右 2 枚はそれぞれ同一のシーンに異なるスケッチを与えた時の結果である．スケッチの正解姿勢をオレンジ色，出力結果を青色のオブジェクトで表す．左 2 枚のシーンでは，従来手法は把持できない誤った位置と姿勢を予測しているが，提案手法は把持可能な位置を予測できた．以上により，提案手法は様々なスケッチ入力に対応可能であるといえる．

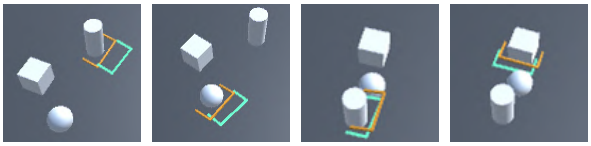


図 2: 従来手法による推論結果

図 3: 提案手法による推論結果

Attention の可視化結果の例を図 4 に示す．Encoder では，シーン全体を注視している．一方で，Decoder ではスケッチした指示周辺の物体を注視していることが確認できる．

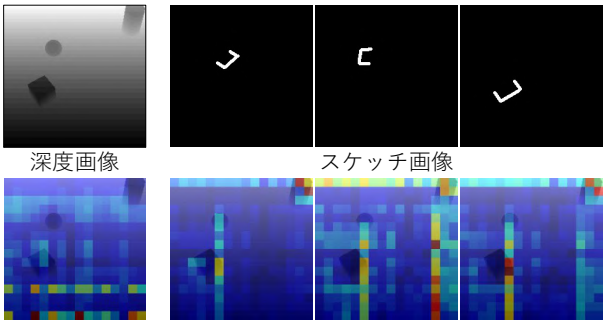


図 4: Attention の可視化結果

5. おわりに

本研究では，物体とスケッチの位置関係を正確に求めるため，Encoder-decoder 型の Transformer モデルを用いて把持位置推定を行い，汎化性能向上の効果を確認した．今後は推定した把持位置を用いたロボット動作の生成を行う予定である．

参考文献

[1] 岩永優香 等，“2D 手書き指示でロボットに人の意図を伝えるインタフェースの開発と評価～深層学習を用いた把持位置姿勢指示手法の検討～”，日本ロボット学会学術講演会，2022.

[2] A.Vaswani, *et al.*,“ *Attention is all you need* ”, *NearIPS, 2017.*

Visual Prompt の付与方法による物体認識精度への影響

ER20082 松崎恵也　　指導教授：藤吉弘亘

1. はじめに

大規模な事前学習モデルでは，下流タスクに適応する方法としてモデル全体を再学習することが難しい．そこで，入力に学習可能なパラメータであるプロンプトを追加し，プロンプトを学習するプロンプトチューニングがある．Visual Prompt は，画像にパディング状のプロンプトを付加して学習する手法である．これにより，プロンプトのみを学習するだけで，モデル全体を再学習する手法に準じる性能を達成できる．一方で，Visual Prompt が学習にどのような影響を与えているか明確ではない．そこで本研究では，Visual Prompt を付加する位置を変更したときの影響を調査する．具体的には，分類精度の変化とアテンションマップによる評価を行う．

2. Visual Prompt

Visual Prompt [1] は，CLIP [2] を下流タスクに適応させるために，入力画像に対して学習可能なパラメータを持つ共通のプロンプトを挿入する手法である．図 1 に Visual Prompt の概要図を示す．Visual Prompt を付加した画像を重みが固定された事前学習済みモデルに入力し，正解ラベルとの尤度を最大化するように Visual Prompt を学習する．

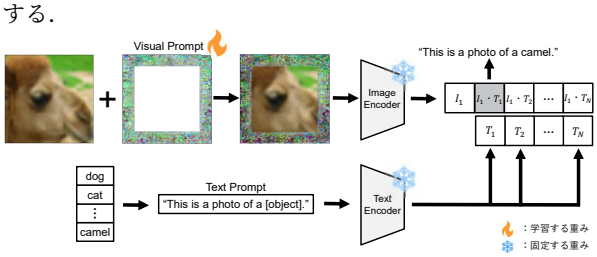


図 1：Visual Prompt の概要

3. Visual Prompt の改善

Visual Prompt には問題点がある．例として図 2 に Grad-CAM を用いて Visual Prompt を付与する前後のアテンションマップを可視化した結果を示す．図 2(a) のように Visual Prompt を付与する前は識別対象に注視していたのに対し，図 2(b) のように Visual Prompt を付与した後は識別対象に注視していないことが確認できる．

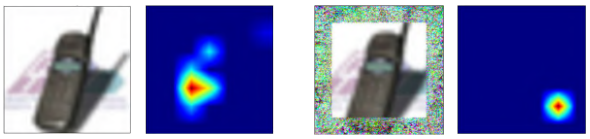


図 2：アテンションマップの可視化の結果

Visual Prompt を画像に重ねることで，図 3(b) のように識別対象の物体の一部が Visual Prompt により隠れてしまうことがある．これにより，図 2 のように認識に有効な特徴量が抽出されなくなると考えられる．そこで，図 3(c) のように Visual Prompt を画像の外側に付与することで，Visual Prompt が対象物体に重ならないようにして，プロンプトチューニングを行う．

図 3：Visual Prompt の改善

4. 実験

本研究では，Visual Prompt を画像の内側または外側に付与した場合の分類精度を比較する．加えて，それぞれの場合のアテンションの可視化結果を比較する．

4.1. 実験条件

データセットに CIFAR100 を用いて，CLIP で事前学習した ResNet50 と ViT-B/32 を評価する．学習するエポック数は 1,000，バッチサイズは 256，最適化は SGD を使用する．

4.2. 実験結果

表 1 より，Visual Prompt を画像の外側に付与すると，モデルに ResNet50 を用いた場合は精度が 5.90pt，ViT-B/32 を用いた場合は精度が 4.93pt 向上した．このことから，Visual Prompt は画像の外側に付与すると良いことを確認した．

表 1：各 Visual Prompt における分類精度

Image Encoder	Visual Prompt の位置		
	プロンプトなし	従来手法	提案手法
ResNet50	40.47	47.28	**53.18**
ViT-B/32	63.11	75.09	**80.02**

Visual Prompt を付与した各位置におけるアテンションマップの可視化結果を図 4 に示す．図 4(a) のように ResNet50 の場合，画像の外側に Visual Prompt を付与することで物体を注視するようになった．一方で，図 4(b) のように物体を注視しないことも確認された．ViT-B/32 の場合，画像の外側に Visual Prompt を付与することで，図 4(c) のように画像領域内の物体に注視した．これにより，Visual Prompt の位置が精度向上に影響することを確認した．

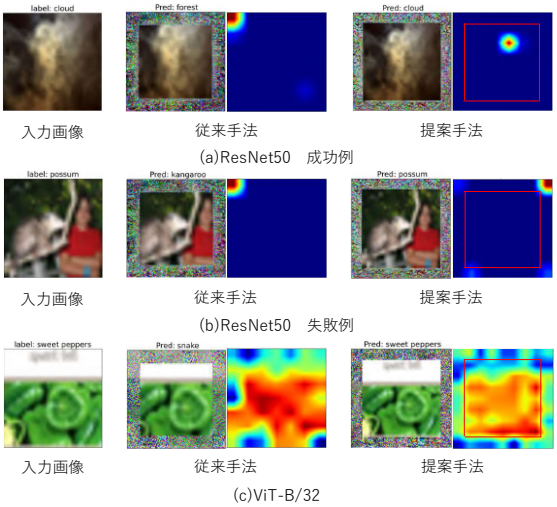


図 4：プロンプトの位置の変更に伴うアテンションマップ

5. おわりに

本研究では，Visual Prompt の位置が認識精度及び解釈性に与える影響を分析した．実験により，Visual Prompt を画像の外側に付与することで精度の改善が確認された．一方，アテンションマップは十分に改善されていない画像も確認できた．今後は，さらなるアテンションマップの改善を目的とした Visual Prompt を考案する予定である．

参考文献

[1] H. Bahng, *et al.* “Exploring Visual Prompts for Adapting Large-Scale Models”, arXiv, 2022.

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

知識転移の時間軸方向への重み付け戦略の最適化

ER20069 原田優輝　　指導教授：藤吉弘亘

1. はじめに

知識転移グラフ [1] は，知識蒸留や相互学習を含む多様な学習手段を Gate 関数を用いたグラフとして表現し，複数モデルを用いた共同学習である．Gate 関数はハイパーパラメータ探索により，最適な組合せとなるグラフを探索する．4 種類の Gate 関数により多様な知識転移を表現しているが，学習過程に伴う変化が少ないことから，時間軸方向に対する知識転移は限定的である．そこで本研究では，Gate 関数の時間軸方向に対する表現力向上を目的として，Positive-Gamma Gate，Negative-Gamma Gate を提案する．これらの Gate 関数は 0.01 から 100 までの連続値で表現される γ 値を持つことで時間軸方向への重み付け戦略を表現することが可能となる．これにより，従来の 4 種類であった Gate 関数が 2 種類に減少し，ハイパーパラメータ変数が削減され，ベイズ最適化による探索コストの削減が期待できる．

2. 知識転移グラフ

知識転移グラフは，モデルを表現するノードとモデル間の知識転移を表現するエッジで構成され，各エッジに適応する Gate 関数を「Through Gate，Cutoff Gate，Correct Gate，Linear Gate」の 4 種類から選択することでモデル間の知識転移を制御する．Gate 関数の組み合わせをハイパーパラメータとして自動探索することで，人手で設計した手法より効果的な知識転移グラフを発見することができる．これにより 1 つのモデルで学習する場合よりも高い精度を達成している．

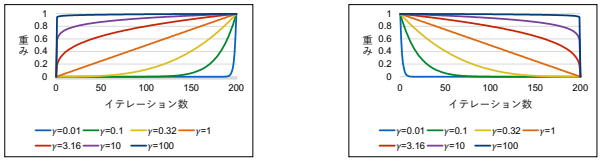
3. 提案手法

Gate 関数の時間軸方向への重み付け戦略により，共同学習の精度向上を目的とし，γ 値を導入した「Positive-Gamma Gate (PG Gate) と Negative-Gamma Gate (NG Gate)」を提案する．PG Gate を式 (1)，NG Gate を式 (2) に示し，

$$G_{s,t}^{\mathrm{PG}}(a) = \left(\frac{k}{k_{\mathrm{end}}}\right)^{\frac{1}{\gamma}}(a) \tag{1}$$

$$G_{s,t}^{\mathrm{NG}}(a) = \left(\frac{k_{\mathrm{end}}-k}{k_{\mathrm{end}}}\right)^{\frac{1}{\gamma}}(a) \tag{2}$$

2 つの Gate 関数のグラフを図 1 に示す．式 (1) は，学習初期から学習後半にかけて損失に乗算する重みを増加させる．一方で式 (2) は学習初期から学習後半にかけて重みを減少させる．提案する Gate 関数は γ 値により，線形な関数では表現不可能な多様な重み付け戦略を表現する．これらの Gate 関数を導入した知識転移グラフを多変量ベイズ最適化で探索することで時間軸方向へ最適な知識転移グラフを獲得できることが期待できる．



(a) PG Gate　　(b) NG Gate

図 1: PG Gate と NG Gate のグラフ

4. 評価実験

PG Gate と NG Gate を知識転移グラフに導入し，その後ベイズ最適化を行いの精度を評価する．ベイズ最適化による探索回数は 500 回とする．データセットには CIFAR-100，評価対象のノードに ResNet32，補助ノードには WRN-28 を使用する．

4.1 精度比較

学習結果を表 1 に示す．従来の知識転移グラフ（DCL）の精度 72.88% に対して，提案手法では 73.34% と 0.46*pt* 精度が向上した．

表 1: 教師あり学習における精度比較

手法	Vanilla	DCL	Ours
Acc[%]	70.71	72.88	**73.34**
Gate の種類	0	4	2
評価ノード	ResNet32	ResNet32	ResNet32
補助ノード	-	WRN-28	WRN-28
探索回数	1	1500	500

図 2 に，ベイズ最適化で獲得された知識転移グラフを示す．獲得された知識転移グラフの Gate 関数から，時間軸方向への重み付けが戦略が最適化されていることが分かる．最適化されたグラフから学習前半は教師ラベルから学習し，学習後半にかけて徐々に相互学習に切り替えていく学習方法が適していたことが分かる．

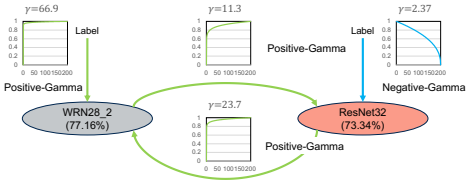


図 2: 最適化済みの知識転移グラフ

4.2 自己教師あり学習への適用

自己教師あり学習である SimCLR，SimSiam，MoCo，SwAV，BYOL，BarlowTwins，DINO の 7 種の学習手法をノード，提案した Gate 関数をエッジとし，これらの組合せをベイズ最適化で探索する．データセットは CIFAR-10，評価対象モデルは ResNet20，補助ノードは ResNet32 を使用する．知識転移には DoGo[2] で提案されたサンプル間の類似度と DisCO[3] で提案された同サンプル間の特徴量の差異を使用する．学習結果を表 2 に示す．従来の SimCLR の KNN 精度 64.16% と比較して，提案手法は 66.80% と 2.64*pt* 精度が向上した．

表 2: 自己教師あり学習における精度比較

手法	SSL1	SSL2	KNN Acc[%]
Vanilla	SimCLR	-	64.16
DML	SimCLR	SimCLR	64.24
Ours	SimCLR	BarlowTwins	**66.80**

5. おわりに

本研究では，時間軸方向への重み付け戦略が多様な深層共同学習を実現するために，PG Gate と NG Gate を提案した．提案手法では，学習経過に伴う損失に乗算する重みの変化により，従来手法に比べて，0.46*pt* 精度が向上した．また自己教師あり学習においても効果を確認した．今後は，ViT をベースとした，基盤モデルへの活用に取り組む予定である．

参考文献

[1] S. Minami, *et al.*, “Knowledge Transfer Graph for Deep Collaborative Learning”, In ACCV, 2020.

[2] P. Bhat, *et al.*, “Distill on the Go: Online knowledge distillation in self-supervised learning ”, In CVPR, 2021.

[3] Yuting Gao, *et al.*, “DisCo: Remedy Self-supervised Learning on Lightweight Models with Distilled Contrastive Learning ”, In ECCV, 2022.