機械知覚&ロボティクスグループ
中部大学 工学部 情報工学科

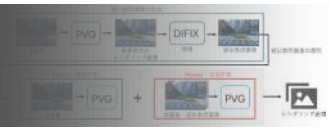
MPRG
MACHINE PERCEPTION AND ROBOTICS GROUP

2025年度　山下研究室　卒業論文発表　アブストラクト

Periodic Vibration Gaussian, New trajectory

疑似教師画像を用いた追加学習による PVG 新規視点画像の品質調査

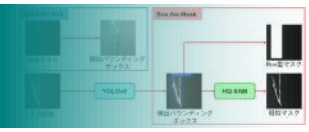
市川 真夏大



Multitask learning, Object detection, Semantic segmentation

マルチタスク学習による落雷検出に向けた疑似ラベル生成フレームワークの構築

大河内 那樹



Automatic Dataset Generation

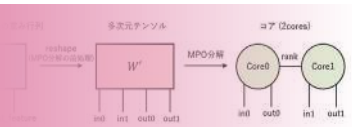
自動運転向け道路インフラ改善画像データセットの自動生成パイプラインの提案

大西 郁美

Pruning, ViT, Matrix Product Operator

枝刈り率を用いた適応的層選択と MPO 分解によるモデル軽量化

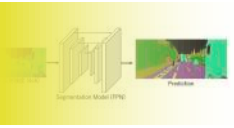
大原 琉生



High-Density Point Cloud, Semanteic Segmentation, Flare Detection

高密度 LiDAR 点群と反射強度を用いたセマンティックセグメンテーションによるフレア検出

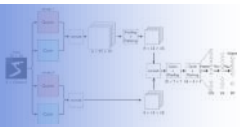
奥谷 俊宏



Quantum Machine Learning, Image Classification

量子・古典特徴の階層的融合によるマルチスケール画像分類

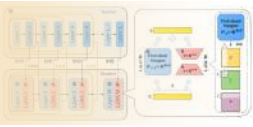
加藤 靖大



Distillation, LoRA

SVD を用いた LoRA と段階的知識蒸留による効率的なモデル圧縮

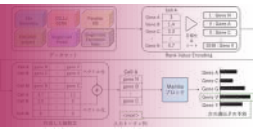
熊澤 綱佑



scRNA-seq, Selective State Space Model

Mamba を用いたマウスの単一細胞解析のための基盤モデル構築

小林 岳隼



Multimodal Large-Language Model (MLLM), Autonomous Driving

運転判断の因果関係を考慮した自動運転のための言語説明データセットの構築

佐藤 克昭

MRI, Self supervised learning

自己教師あり学習による脳 MRI 解析基盤モデルの構築と評価

羽澄 直弥

Multimodal Large Language Model, Autonomous Driving, Scenario

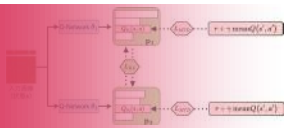
MLLM と BEV 映像を用いた走行映像からの OpenSCENARIO 自動生成

服部 美月

Reinforcement Learning, Deep Mutual Learning

Mutual DDQN: 深層強化学習エージェントの相互学習

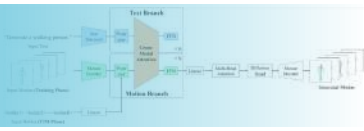
前田 登生



Text-to-Motion, 2D Keypoints

時間的整合性を考慮したテキストからの 2 次元モーション生成

川本 寛和



URL: http://mprg.jp

疑似教師画像を用いた追加学習による PVG 新規視点画像の品質調査

EP22015 市川真夏大　　指導教授：山下隆義

1. はじめに

自動運転システムの閉ループ型評価を目的として，実世界シーンを再現し，任意視点の映像を生成する研究が行われている．中でも 3D Gaussian Splatting（3DGS）は，シーンを多数の 3 次元ガウス分布により表現し，写実性の高い映像を生成できる．3DGS の代表的な手法として Periodic Vibration Gaussian（PVG）[1] が提案されている．PVG は学習時に観測されていない新規視点の生成時，アーティファクトが生じやすいという課題がある．そこで本研究では，新規視点の生成時に，高品質な疑似生成画像を学習にフィードバックすることで，生成した新規視点の品質向上を目指す．

2. 関連研究

3DGS を時間方向に拡張した PVG とレンダリング画像を高品質化させる DIFIX について述べる．

2.1. Periodic Vibration Gaussian (PVG)

3DGS は 3 次元空間に，色や不透明度の 3 次元ガウス分布を分布を配置し，特定の視点からの画像を生成する．PVG[1] は，各 3 次元ガウス分布に時間表現を導入することで，動的シーンの映像生成を可能にしている．これにより，動的シーンにおいて時間的に滑らかな再構成が実現される．一方，新規視点においては視覚的品質が低下しやすく，アーティファクトが発生しやすい課題がある．

2.2. DIFIX

DIFIX[2] は，3DGS などの 3D 再構成手法によって生成された画像に含まれるアーティファクトを低減することを目的とした手法である．従来の多段階拡散モデルではなく，単一ステップの拡散モデルを用いることで，高速かつ実用的なアーティファクト除去を実現している．

3. 提案手法

本研究では，PVG による新規視点の生成画像の品質を向上させるために，DIFIX を用いて修復した生成画像を疑似教師画像として PVG の学習にフィードバックする手法を提案する．図 1 に提案手法の流れを示す．まず，実画像のみで 30,000 イテレーション学習した PVG を用いて新規視点の画像を生成する．次に，物体の一部が破綻しているかを目視で確認し，破綻していない生成画像のみを疑似教師画像とする．そして，元の視点の画像とともに PVG を学習する．これにより構造的破綻を学習することを防ぐ．

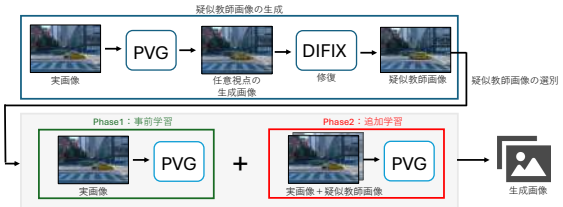


図 1: 提案手法の学習の流れ

4. 評価実験

追加する疑似教師画像の生成条件を変えて，PVG を学習した時の画像を比較し，提案手法の有効性を検証する．

4.1. 実験概要

図 2 に評価実験の学習の流れを示す．DIFIX による疑似教師画像の生成条件は以下の 3 つである．

1) 実画像のみ
2) 右へ 20cm 移動
3) 右へ 5〜20cm 移動

条件 1 では，実画像のみを用いて 30,000 イテレーション学習した．条件 2 では，実画像で 15,000 イテレーション学習後，DIFIX で生成した新規視点画像を選別・追加し，15,000 イテレーションで学習した．条件 3 では，実画像で 15,000 イテレーション学習後，DIFIX で生成した新規視点画像を選別・追加し，15,000 イテレーション学習した．各条件で学習した PVG を用いて右に 20cm ずらした新規視点画像の視覚的品質を比較する．評価には 4 シーンを用いる．

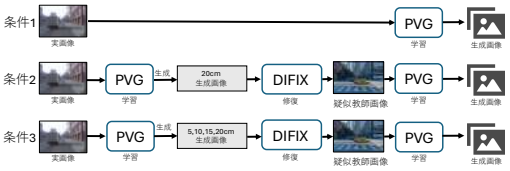


図 2: 各条件におけるレンダリング画像の生成

4.2. 実験結果

表 1 に 4 シーンの定量的評価を示す．表 1 より，条件 3 は条件 1 と比較して PSNR が約 0.7pt 向上し，SSIM も約 0.06pt 向上していることが分かる．また，条件 3 は条件 2 と比較して PSNR が約 2.8pt 向上し，SSIM が約 0.09pt 向上していることが分かる．

表 1: 4 シーンにおける平均値および分散

	条件 1（実画像のみ）	条件 2（20cm のみ）	条件 3（5〜20cm）
NIQE↓（Mean）	**3.284**	3.525	3.585
NIQE↓（Var）	0.108	0.120	0.116
PSNR↑（Mean）	18.594	16.454	**19.265**
PSNR↑（Var）	3.207	2.809	11.129
SSIM↑（Mean）	0.497	0.471	**0.556**
SSIM↑（Var）	0.00377	0.00539	0.00831
LPIPS↓（Mean）	**0.308**	0.459	0.365
LPIPS↓（Var）	0.00034	0.00013	0.00814

図 3 に 1 シーン目の定性的評価を示す．図 3 より，条件 2 と比較して条件 3 は，アーティファクトが含まれていないことが分かる．このことから，複数の軌跡の疑似教師画像を用いることで，新規視点画像における画素値の再現性および構造的一貫性が改善されることが考えられる．しかし，条件 3 は条件 1 と比較して，白い車にアーティファクトを含んでいる．これより，提案手法による改善は不十分であることが分かった．

条件 1：実画像のみ

条件 2：疑似教師あり (20cm のみ)

条件 3：疑似教師あり (5〜20cm)

図 3: 定性的評価

5. おわりに

本研究では，DIFIX で修復した新規軌跡画像を疑似教師画像として PVG の学習にフィードバックする手法を提案した．評価実験より，条件 2 と比較して条件 3 の方が定性的・定量的評価で良いスコアが確認できた．今後は，より高品質な新規視点の疑似教師画像を生成することで，新規視点のレンダリング精度の向上に取り組む．

参考文献

[1] Y. Huang *et al.*, “Periodic Vibration Gaussian for Dynamic Scene Reconstruction”,IJCV,2026.

[2] J. Z. Wu *et al.*, “DIFIX3D+: Improving 3D Reconstructions with Single-Step Diffusion Models”,CVPR,2025.

マルチタスク学習による落雷検出に向けた疑似ラベル生成フレームワークの構築

EP22033　大河内 那樹　　指導教授：山下 隆義

1. はじめに

非剛体で形状が多様な物体クラスの検出に対して，セマンティックセグメンテーションは有効な手段の一つである．しかしながら，正解ラベルを画素単位で付与するためアノテーションコストが高い．そこで本研究では，物体検出を併用して，未ラベルデータに対してアノテーションを自動的に行うフレームワークを構築する．そして，疑似的にアノテーションしたデータを用いて学習することで，不完全なデータ状況下における落雷検出モデルの高精度化を目指す．

2. 部分的教師ありマルチタスク学習

マルチタスクにおいて，全タスクの正解ラベルが揃わないデータでは，損失計算ができず学習が困難となる．この課題に対し，欠損ラベルを補完する部分的教師あり学習が提案されており，代表的な手法に BoMBo[1] がある．BoMBo の構造を図 1 に示す．BoMBo は，Backbone と Neck の Encoder と各タスクの Head で構成される．学習時，バウンディングボックスのみのデータには Box-for-Mask を行う．ここでは，GrabCut による疑似マスクを教師信号とするほか，Box 型マスクとアテンションマップの MSE を最小化して Segmentation head を学習する．一方，マスクのみのデータには Mask-for-Box を適用する．ここでは，マスクから外接矩形を算出して疑似ラベルとし，Detection head を学習する．

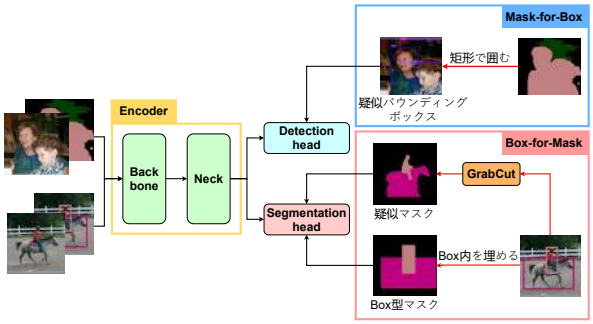


図 1: BoMBo の構造

3. 提案手法

本研究では，未ラベル画像とマスクのみの画像に対して両タスクの疑似ラベルを自動生成するフレームワークを構築する．そして，本フレームワークにより生成された疑似ラベルを用いて BoMBo による落雷検出の高精度化を図る．

3.1. 疑似ラベル生成フレームワーク

提案する疑似ラベル生成フレームワークを図 2 に示す．まず，未ラベル画像に対し，落雷画像でファインチューニングを行った YOLOv8 を用いて検出バウンディングボックスを生成する．次に，Segment Anything in High Quality(HQ-SAM)[2] を用いて疑似マスクを生成する．この際，事前学習済み重みを固定し，LoRA により追加された少数のパラメータのみを学習させることで，計算コストを抑えつつモデルを落雷画像に適応させる．

3.2. 部分的教師あり学習の適用

部分的教師あり学習の手法である BoMBo をもとに，検出バウンディングボックスから Box 型マスクを生成する．マスクのみのデータに対しては Mask-for-Box により疑似バウンディングボックスを補完し，これらを用いたマルチタスク学習を実施する．

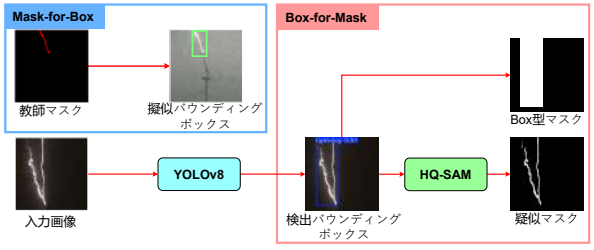


図 2: 疑似ラベル生成フレームワーク

4. 評価実験

画像処理や既存の基盤モデルを用いた疑似マスク作成と比較し，構築したアノテーションフレームワークによって生成されたラベルの有効性を検証する．

4.1. 実験概要

本実験では，学習用データとして教師マスクありデータ 56 枚，検出バウンディングボックスおよび疑似マスクありデータ 9,660 枚を用いる．評価用データは両タスク共通の 40 枚をデータセットとして用いる．モデルは RetinaNet，バッチサイズは 32，学習回数は 80epoch，評価指標は物体検出は AP，セグメンテーションは IoU とする．

4.2. 実験結果

定量的評価を表 1 に示す．表 1 より，LoRA を適用した HQ-SAM が最も高精度であり，通常の HQ-SAM と比較して，雷道の IoU が 8.75pt 向上した．

表 1: 定量的評価 [%]

LoRA	Method	AP	IoU	
			lightning	background
-	二値化	60.08	34.07	98.31
-	GrabCut	58.43	42.86	97.96
-	SAM	62.18	61.98	99.02
-	HQ-SAM	60.15	63.31	99.10
✓	HQ-SAM	**65.99**	**72.06**	**99.41**

HQ-SAM および LoRA を適用した HQ-SAM により生成したマスク画像を学習に使用した際のセグメンテーションの定性的評価を図 3 に示す．図 3 から，LoRA を適用することで雷道の詳細を捉えることができた．図 4 に物体検出結果を示す．図 4 より雷道の検出が可能であることを確認した．

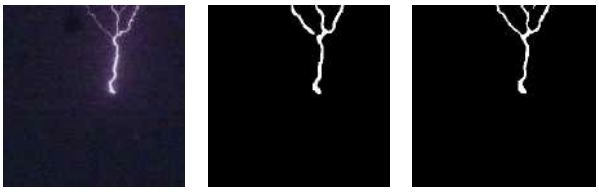


図 3: セグメンテーションの定性的評価

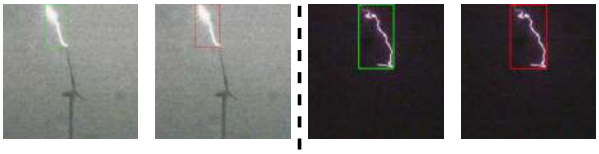


図 4: 物体検出の定性的評価

5. おわりに

本研究では，未ラベルデータを有効活用するため，自動アノテーションフレームワークを構築し，疑似ラベルを用いた部分的教師あり学習による落雷検出の高精度化を実現した．実験結果より，LoRA を適用した HQ-SAM による疑似マスクの導入が，落雷検出およびセグメンテーション精度の向上に有効であることを確認した．今後は，より高精度な疑似ラベル生成に向けた手法の検討を行う．

参考文献

[1] H.Lê *et al.*, “Box for Mask and Mask for Box: weak losses for multi-task partially supervised learning”, BMVC, 2024.

[2] L.Ke *et al.*, “Segment Anything in High Quality”, NeurIPS, 2023.

自動運転向け道路インフラ改善画像データセットの自動生成パイプラインの提案

EP22036 大西郁美　　指導教授：山下隆義

1. はじめに

道路インフラは，人間の運転を前提としており，自動運転システムには適していない環境もある．例えば，経年劣化により視認性が低下した車線や，街路樹等により遮蔽された標識などが挙げられる．こうした環境に対するインフラ改善案を提示する手法として OD-RASE [1] が提案されている．OD-RASE は，過去の交通事故要因をもとに交通事故を引き起こす要因となる道路構造を検出し，インフラ改善案とその改善画像を生成する．これにより，専門知識を持たない場合でも改善すべき道路構造とその改善案を直感的に把握可能となる．OD-RASE の生成画像は直感的に改善内容を理解できるが，物理的整合性を欠く場合がある．これは，生成モデルがインフラ改善に関するドメイン知識を学習していないからである．また，そのための学習データセットが存在しないことに起因する．そこで本研究では，インフラ改善が可能な画像生成モデルの学習のためにインフラ改善前後のペア画像データセットを自動的に構築するパイプラインを提案する．

2. OD-RASE

OD-RASE は，自動運転システムの安全性向上を目的として，過去の交通事故要因をもとに道路インフラ改善案を生成する手法である．しかし，図 1 のように改善内容の画像化において，物理的整合性の確保に課題が残されている．これは，使用する生成モデルが，インフラ改善に関するドメイン知識を学習していないこと，ならびにその学習に必要なデータセットが整備されていないことに起因する．そのため，インフラ改善前後のペア画像データセットを新たに構築する必要がある．

(a) 改善前　　(b) 改善後

図 1: OD-RASE によるインフラ改善後画像

3. データセット自動生成パイプライン

本研究では，インフラ改善前後のペア画像データセットを自動構築するためのパイプラインを提案する．パイプラインの構成を図 2 に示す．入力となる改善テキストは OD-RASE を用いて自動生成する．パイプラインは，サブタスク分割，プロンプト付与，画像生成の 3 つのステップから構成される．以下では，3 つのステップについて詳細に述べる．

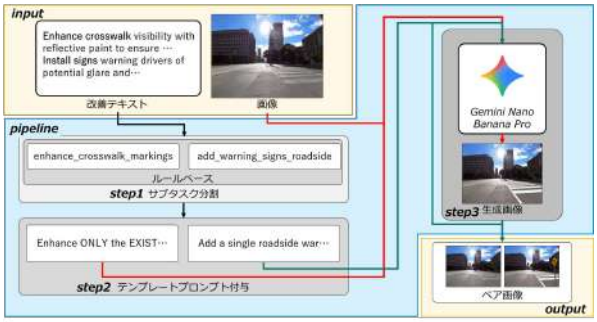


図 2: データセット構築パイプライン

Step1 サブタスク分割　OD-RASE が生成した改善テキストを，ルールベースでサブタスクへ変換する．この変換は，改善テキストが自由記述形式であり，そのままでは改善内容の分布把握や体系的なデータセット構築が困難であるためである．そこで，Mapillary Vistas の Validation に含まれる OD-RASE の出力結果（計 1253 枚）を分析し，本研究で対象とする改善パターンを網羅できるよう，ルールを設計した．改善テキストに含まれる動詞や改善対象を手がかりに，その改善内容を表すサブタスクを付与する．

Step2 プロンプト付与　各サブタスクに対して，改善内容と維持すべき背景要素を明記したテンプレートプロンプトを付与する．事前に設計した固定のテンプレートを一律に適用することで，プロンプトの表現の揺らぎによる生成品質のばらつきを排除し，一貫した画像生成指示を与える．

Step3 画像生成　生成モデルには，予備実験で高い性能を示した Gemini Nano Banana Pro を使用する．複数サブタスクに対しては，前段の生成画像を次段の入力として逐次的に処理を行い，全改善案を反映したインフラ改善後の画像を得る．

4. データセットの品質評価

提案手法により構築したデータセットが，インフラ改善タスクにおいて妥当な品質を有するか検証する．

4.1. 評価条件

データセットは Mapillary を使用する．OD-RASE が生成するインフラ改善案をプロンプトとして画像生成を行うものをベースラインとする．評価は目視により，改善内容の描画と物理的整合性の維持の 2 点を成功したか否かで成功率を算出する．

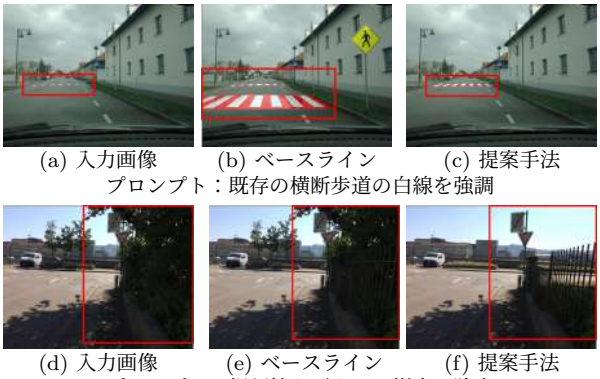
4.2. 評価結果

評価の結果を表 1 に示す．表 1 より，ベースラインは全 250 枚のうち 168 枚が成功，全体の成功率は 67.2%である．一方，提案手法は全 250 枚のうち 208 枚が成功，全体の成功率は 83.2%である．これより，提案手法の有効性が定量的に示された．

表 1: 成功率による比較結果

手法	ベースライン	提案手法
成功率 (%)	67.2	**83.2**

生成結果例を図 3 に示す．図 3 より，横断歩道標示の強調や樹木の除去といった異なる改善内容に対し，物理的整合性を維持した画像が生成されていることが分かる．これより，提案手法による指示の明確化とテンプレートプロンプトが生成品質の向上に寄与したといえる．



(a) 入力画像　(b) ベースライン　(c) 提案手法

プロンプト：既存の横断歩道の白線を強調

(d) 入力画像　(e) ベースライン　(f) 提案手法

プロンプト：視認性を悪くする樹木の除去

図 3: 構築したデータセットのサンプル

5. おわりに

本研究では，インフラ改善前後のペア画像データセットを自動構築するパイプラインを提案した．提案手法は，OD-RASE が出力したインフラ改善文に基づくサブタスク分割と，テンプレートプロンプトを用いた画像生成により，改善内容を画像へ反映できた．今後は，作成したデータセットを用いた生成モデルの学習を検討する．

参考文献

[1] K. Shimomura, *et al.*, “OD-RASE: Ontology-Driven Risk Assessment and Safety Enhancement for Autonomous Driving”, ICCV, 2025.

枝刈り率を用いた適応的層選択と MPO 分解によるモデル軽量化

EP22038 **大原琉生**　　**指導教授：山下隆義**

1. はじめに

Transformer をベースとした深層学習モデルが注目されている．画像認識分野においても，画像エンコーダに Transformer を用いたモデルである Vision Transformer (ViT) が高精度を示している．一方で，ViT は重みパラメータ数が多いため，計算資源の限られたエッジデバイスへの展開が困難である．この課題に対し，モデルの重みパラメータを削減する手法として行列積演算子 (MPO) 分解による低ランク近似がある [2]．MPO 分解はモデルのパラメータ削減と推論の高速化が可能であるが，全ての層に一律で適用するとモデル性能が著しく低下する課題がある．そこで本研究では，層ごとにパラメータの冗長性が異なる点に着目し，各層の重要度に応じて動的に低ランク近似する手法を提案する．これにより，精度維持と高速化を両立した軽量化手法の実現を目指す．

2. 従来手法

本章では，提案手法の基盤となる非構造枝刈り手法である Adaptive Feature Retaining (AFR) と，低ランク近似手法である Matrix Product Operator (MPO) 分解について述べる．

2.1. Adaptive Feature Retaining

Adaptive Feature Retaining (AFR)[1] は，事前学習モデルの知識維持と下流タスク適用を両立した非構造枝刈り手法である．AFR では知識維持とタスク適応の両観点から重みの重要度を評価する指標を導入している．具体的には，i 番目の重みパラメータ w_i に対する評価値 $S_{\mathrm{AFR}}(w_i)$ を式 (1) のように定義する．

$$S_{\mathrm{AFR}}(w_i) = \mathcal{Z}\left(\left|\frac{\partial \sum_{l=1}^{L} F_{\mathrm{SVD}}^{l}}{\partial w_i} w_i\right|\right) + \mathcal{Z}\left(\left|\frac{\partial \mathcal{L}}{\partial w_i} w_i\right|\right) \quad (1)$$

ここで，$\mathcal{Z}(\cdot)$ は標準化を表す．第 1 項は知識維持の観点に基づいた指標であり，l 層目の出力特徴量に対する特異値の平均 F_{SVD}^{l} を用いて，w_i が特徴空間の情報量や分布にどの程度寄与しているかを評価する．第 2 項はタスク適応の観点に基づく項であり，損失関数 $\mathcal{L}(\cdot)$ に対する勾配を用いて，下流タスクの精度に対する重みの影響を評価している．これにより，事前学習で獲得した知識を保持しながら，下流タスクに対する適応性も考慮した効果的な枝刈りが実現される．

2.2. 行列積演算子によるモデル圧縮

行列積演算子 (MPO) 分解によるモデル圧縮 [2] は，重み行列を低ランクテンソルネットワークに近似することで深層学習モデルを効率的に圧縮する手法である．MPO 分解では入力次元 N，出力次元 M を持つ重み行列 $\boldsymbol{W} \in \mathbb{R}^{M \times N}$ に対し，それぞれの次元を $N = \prod_{k=1}^{n} I_k$, $M = \prod_{k=1}^{n} J_k$ となる n 個の因子の積に分解し，高階テンソル $\boldsymbol{W}_{j_1 \ldots j_n, i_1 \ldots i_n}$ に変形する．このとき，$\boldsymbol{W}_{j_1 \ldots j_n, i_1 \ldots i_n}$ は式 (2) のように，n 個のコアテンソル $\boldsymbol{w}^{(k)}$ の縮約として近似できる．

$$\boldsymbol{W}_{j_1 \ldots j_n, i_1 \ldots i_n} \approx \mathrm{Tr}\left(\boldsymbol{w}^{(1)}[j_1, i_1] \cdots \boldsymbol{w}^{(n)}[j_n, i_n]\right) \quad (2)$$

ここで，各コアテンソル $\boldsymbol{w}^{(k)}$ は隣接するコアテンソルと接続するための次元であるボンドインデックス（ボンドランク）を持ち，その大きさを調整することで表現力とパラメータ数のトレードオフを調整できる．

3. 提案手法

本研究では，非構造枝刈りにおける枝刈り率を指標として，MPO 分解による低ランク近似を適用する層を動的に選択する手法を提案する．具体的には，事前学習済みモデルに対する非構造枝刈り手法として有効な AFR を用いて各層の枝刈り率を算出する．そして，枝刈り率が高い上位 K 個の層の集合 L_{topk} に対して MPO 分解による低ランク近似を適用する．層 l に含まれる重み行列を $\boldsymbol{W}$，提案手法適用後の重みを $\hat{\boldsymbol{W}}$ とすると，動的な層選択は式 (3) のように定式化される．

$$\hat{\boldsymbol{W}} = \begin{cases} \Phi_{\mathrm{MPO}}(\boldsymbol{W}) & (l \in L_{\mathrm{topK}}) \\ \boldsymbol{W} & (\text{otherwise}) \end{cases} \quad (3)$$

ここで，$\Phi_{\mathrm{MPO}}(\cdot)$ は MPO 分解による低ランク近似を表す．これにより，重要層は保持し冗長層のみ圧縮することで，精度維持と高速化を実現する．また，本手法では非構造枝刈りを不要な重みを除去する前処理として利用する．枝刈りによってノイズが低減された行列に対して MPO 分解を行うことで，元の行列を近似する場合と比較して，重要な情報を損なわずに低ランク近似が可能になることを期待する．

4. 評価実験

ImageNet-1k で事前学習された ViT-B/16 を baseline とし，提案手法による画像分類精度の変化およびモデル圧縮と高速化の効果を評価する．

4.1. 実験概要

MPO 分解を適用する対象は MLP とし，適用する層数は 12 ブロック中 6 ブロックとする．MPO 分解はコア数 2，ランク 4 とする．また，提案手法の有効性を検証するため，以下 2 つの選択方法において比較を行う．

固定選択 偶数番目のブロックに MPO 分解を適用する．なお，前処理として AFR による枝刈り率を 70%とする．

動的選択 提案手法に基づき，ブロックごとの枝刈り率が高い上位 6 ブロックを選択して MPO 分解を適用する．AFR における枝刈り率を，10%および 70%とする．

評価データセットは CIFAR-10 と CIFAR-100，Stanford Cars を用い，提案手法を適用後に 150 エポックのファインチューニングを行う．

4.2. 実験結果

ViT-B/16 に対して提案手法を適用した際の画像分類精度と推論の高速化率を表 1 に示す．表 1 より，固定選択に比べ枝刈り率 70%の動的選択が高い精度を示した．これは，枝刈り率を指標とすることで，圧縮による影響が少ない層を適切に選択できていることを示している．また，枝刈り率 10%と 70%の比較では，全てのデータセットで 70%の方が高い精度を示した．これは，非構造枝刈りにより低ランクでの近似のしやすさが向上していることが考えられ，低ランク表現化における枝刈りの有効性を示している．さらに，推論速度についてはベースラインと比較して最大 1.06 倍の向上が確認されたものの，大幅な改善には至らなかった．

表 1: 圧縮後モデルの分類精度 [%]

Dataset	baseline	Fixed	Ours (10%)	Ours (70%)
CIFAR-10	98.39	94.34	94.06	**94.51**
CIFAR-100	89.53	76.78	77.03	**80.84**
Stanford Cars	68.37	17.91	13.05	**24.39**
Speed-up	1.00x	1.05x	1.06x	1.06x

5. おわりに

本稿では，非構造枝刈りにおける枝刈り率を指標として，MPO 分解を適用する層を動的に選択する手法を提案した．評価実験の結果，固定的な層選択と比較して，提案手法による動的選択が精度の向上に寄与することを確認した．特に，非構造枝刈り率が高い設定においてその効果が顕著であった．今後は，非構造枝刈りで生じた 0 値が低ランク近似に与える影響の調査と，より高度な層の選択指標の検討を行う．

参考文献

[1] 新田常顧, et al. “事前学習済みモデルの知識維持と下流タスク適応を両立した Single-shot Foresight Pruning”, 画像の認識・理解シンポジウム, 2025.

[2] Ze-Feng Gao, et al. “Compressing deep neural networks by matrix product operators”, Physical Review Research, 2020.

高密度 LiDAR 点群と反射強度を用いたセマンティックセグメンテーションによるフレア検出

EP22044 奥谷俊宏　　指導教授：山下隆義

1. はじめに

SPAD 型 LiDAR は，高密度な点群を取得できるため，従来のレーザスキャン型 LiDAR の弱点である垂直方向の解像度の限界を克服している．しかしながら，原理的な問題で，実際の物体の形状と異なる形状が取得されるフレアが発生する．これは，本来存在しない障害物の誤検知を誘発することになり，自動運転では事故の危険につながる．本研究では，セマンティックセグメンテーションによるフレア検出を実現する．

2. フレアの発生現象と課題

フレアは，高密度点群の取得過程で，発光素子と受光素子のミスマッチに起因し，実際の物体の形状と異なる形状が発生する現象である．フレアが発生した時の Ambient 画像と点群のマスク画像を図 1 に示す．図 1(a) では赤枠で囲む領域のようにフレアが発生していないのに対し，図 1(b) ではフレアが発生している．

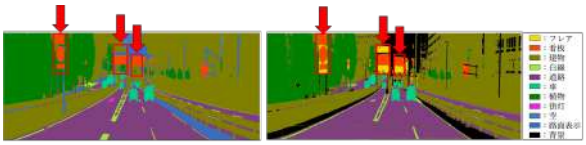


図 1: Ambient 画像とフレアが発生した点群の比較

セマンティックセグメンテーションの学習では，画像中の大きな範囲を占めるクラスを優先して学習する傾向がある．そのため，フレアのように占める割合が低いクラスの精度は低くなる．そこで，フレアの検出に反射強度が有効であるかを調査する．反射強度の可視化結果を図 2 に示す．図 2(a) より，看板領域では反射強度が高い数値で観測される一方で，フレア領域では反射強度が低い傾向にあることが確認できる．この結果から，フレアと実物体との間には反射強度に明確な差が存在し，フレア検出において反射強度情報が重要であると考えられる．

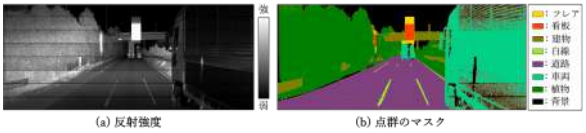


図 2: 反射強度の可視化

3. 提案手法

提案手法の全体構造を図 3 に示す．高密度点群（1 サンプルあたり約 60 万点）は，従来の点群と比較して，1 サンプルあたりに含まれる点数が約 20 倍と非常に多く，直接セマンティックセグメンテーションを適用することは計算コストの観点から困難である．そこで高密度点群を，解像度 512 × 1200 ピクセルの画像形式に変換する．具体的には，各点の三次元座標を画像の RGB チャンネルに対応させ，疑似的なカラー画像を生成する．これにより，2 次元畳み込み処理による CNN モデルでの学習が可能となる．さらに，前述の事前調査に基づいて，点群から生成した疑似画像に各点における反射強度をチャンネル方向に追加した 4 チャンネル画像を入力として用いる．

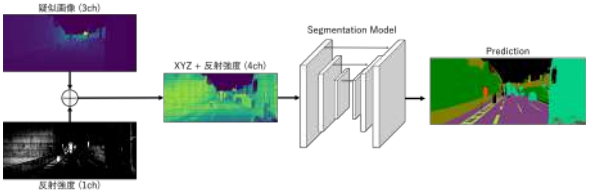


図 3: 提案手法の概要

4. 評価実験

提案手法におけるフレア検出の有効性を検証する．ここで，疑似画像のみ，反射強度のみを入力する手法をベースラインとする．データセットには，フレアが含まれる独自の高密度点群データセットを使用する．データセットは高速道路や市街地の走行シーンから取得した，全 41 シーン，6,400 フレームで構成される．これを学習用に 4,100 フレーム，検証用に 1,900 フレーム，評価用に 400 フレームに分割する．

4.1. 実験条件

セグメンテーションを行うベースモデルとして，Feature Pyramid Network(FPN)[1] を用いる．学習条件として，エポック数を 50，学習率を 0.001，損失関数は Cross Entropy と，不均衡なクラス分布に対応した Focal Loss とで比較を行う．評価指標として mIoU，フレアクラスの IoU を用いる．

4.2. 定量的評価

定量的評価結果を表 1 に示す．表 1 より，疑似画像と反射強度の場合の mIoU は Focal Loss で 0.5965 と疑似画像のみの場合より 7.52pt，反射強度のみの場合より 1.61pt 向上した．また，フレアクラスの IoU は Focal Loss で 0.2210 と疑似画像のみの場合より 12.52pt，反射強度のみの場合より 1.61pt 向上した．これより，疑似画像と反射強度がフレアの検出に有効であることがわかった．

表 1: 定量的評価

疑似画像	反射強度	損失関数	mIoU↑	Flare-IoU↑
✓		Cross Entropy	0.4788	0.1109
✓		Focal Loss	0.5213	0.1089
	✓	Cross Entropy	0.4865	0.2132
	✓	Focal Loss	0.5645	0.2049
✓	✓	Cross Entropy	0.5000	**0.2341**
✓	✓	Focal Loss	**0.5965**	0.2210

4.3. 定性的評価

提案手法および疑似画像のみを入力する場合の定性的評価を図 4 に示す．図 4(c) より，ベースラインではフレアと看板の領域の境界が曖昧になっている．一方で図 4(d) より，反射強度の入力によってフレアと看板の境界識別が正しくなったことが確認できる．

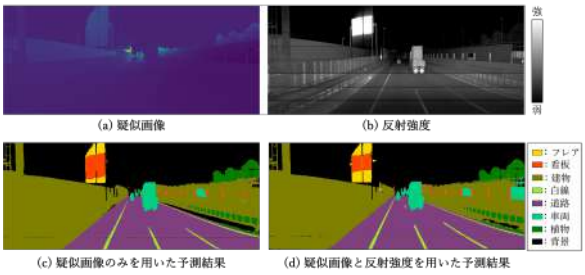


図 4: モデルの定性的評価

5. おわりに

本研究では，高密度点群の取得時に発生するフレアの検出を実現するため，疑似画像と反射強度を活用したセマンティックセグメンテーションを提案した．実験結果より，反射強度の入力により点群のフレア検出を実現した．今後はよりフレア検出に適したモデルの選定や，時系列の考慮による効果についての検証を行う予定である．

参考文献

[1] A.kirillov, *et al.*, “Panoptic Feature Pyramid Networks”, CVPR, 2019.

量子・古典特徴の階層的融合によるマルチスケール画像分類

EP22052　加藤 靖大　　指導教授：山下 隆義

1. はじめに

機械学習モデルの大規模化に伴い，計算資源や学習時間が増大し，計算コストの増加が問題となっている．解決策として，量子コンピュータを利用した量子機械学習が注目されている．量子機械学習の代表的な手法である HQNN-Quanv[1] は，量子畳み込み層を構成する量子回路のパラメータを学習で最適化し，タスクに適した特徴抽出が可能である．しかし，単一の量子畳み込み層では，使用可能な量子ビット数や量子回路の深さに制約があり，局所的・大域的な特徴を同時に抽出することが困難である．そこで本研究では，局所的・大域的な特徴を同時に抽出するために，量子・古典双方の特徴を活用した MS-HQNN を提案する．

2. HQNN-Quanv

HQNN-Quanv は，量子回路にパラメータ化量子回路 (Parameterized Quantum Circuits; PQC) を導入した手法である．HQNN-Quanv の構造を図 1 に示す．本手法は，量子ゲートのパラメータを学習により最適化することで，タスクに適した特徴の抽出が可能である．まず，入力データを量子状態にエンコードし，量子畳み込み層で特徴抽出を行う．その後，量子ビットの状態を測定し，全結合層で分類を行う．HQNN-Quanv は，使用可能な量子ビット数や量子回路の深さに制約があるため，局所的・大域的な特徴を同時に抽出することが困難である．

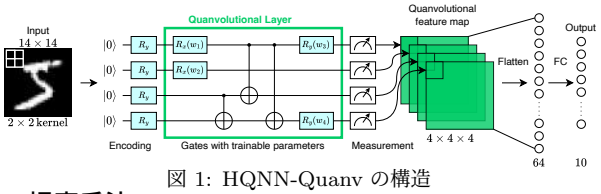


図 1: HQNN-Quanv の構造

3. 提案手法

本研究では，HQNN-Quanv における課題を解決する Multi Scale-HQNN (MS-HQNN) を提案する．MS-HQNN は，タスクに適応した特徴抽出を行うため，PQC を導入した量子畳み込み層と古典畳み込み層を並列に組み合わせた手法である．MS-HQNN の構造を図 2 に示す．量子畳み込み層の高次元な特徴と古典畳み込み層の線形な特徴を結合することにより，表現力を向上させる．さらに，マルチストライド構造を導入することで，異なるスケールで抽出した特徴を結合する．これにより，局所的な特徴と大域的な特徴を同時に抽出できる．学習においては，損失関数から得られる勾配を古典畳み込み層と量子畳み込み層へ逆伝播させ，両層のパラメータを最適化する．

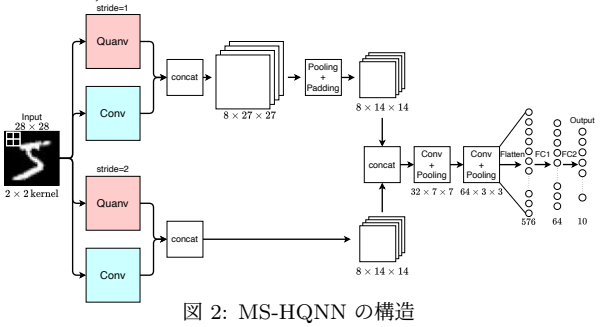


図 2: MS-HQNN の構造

4. 評価実験

提案手法の有効性を検証するため，複数のデータセットでの既存手法との分類精度比較および推論時に各層・カーネルの出力を全て 0 に置換するマスク実験により，各層・カーネルの寄与度分析を行う．

4.1. 実験概要

本実験では，提案手法，CNN，HQNN-Quanv，提案手法の畳み込み層を量子層または古典層に置換し，マルチストライド構造を維持した場合 (Quantum Only, Classical Only) の分類精度を比較する．ここで，分類精度の比較のみでは，量子層および古典層が相補的な役割を果たしているかを十分に確認できない．そこで，各構成要素の貢献度を明らかにするため，推論時に各層および各カーネルの出力を 0 に置換するマスク実験を行う．エポック数は 50，バッチサイズは 100，最適化手法は Adam，損失関数は Cross Entropy Loss を用いる．データセットについては，分類精度の比較実験では MNIST，Fashion-MNIST および CIFAR-10，マスク実験では CIFAR-10 を用いる．

4.2. 実験結果

実験結果を表 1 に示す．MNIST では，Classical Only および提案手法が 99.02%の最高精度を達成した．CNN の精度は 98.85%であり，CNN より 0.17 ポイント向上している．Fashion-MNIST では，Classical Only が 91.22%の最高精度を達成し，提案手法は 90.53%であった．MNIST より複雑な分類タスクである CIFAR-10 で，提案手法は 67.77%の分類精度を達成し，他手法を上回る精度となった．以上より，提案手法の有効性を確認した．

表 1: 各手法における分類精度

Model	Test acc [%]		
	MNIST	Fashion MNIST	CIFAR-10
CNN	98.85	88.72	65.10
HQNN-Quanv	86.49	81.07	32.53
Quantum Only	98.93	88.92	65.67
Classical Only	**99.02**	**91.22**	66.23
MS-HQNN	**99.02**	90.53	**67.77**

次に，層単位のマスク実験結果を図 3 に示す．q1，c1 はストライド 1，q2，c2 はストライド 2 における量子および古典層を表す．量子層をマスクした場合，古典層よりも精度低下が大きく，量子層の寄与が大きいことが確認された．また，ストライド 1 の層をそれぞれマスクした際に精度低下が顕著であり，詳細な特徴抽出の重要性が示唆された．

最後に，カーネル単位のマスク実験結果を図 4 に示す．括弧内の数字 0~3 はカーネル番号を表す．ストライド 1 では，単一カーネルのマスクによる精度低下の幅が大きく，代替困難な特徴を抽出していると考えられる．一方，ストライド 2 では精度低下が小さく，特徴が複数カーネルに分散して寄与していると考えられる．

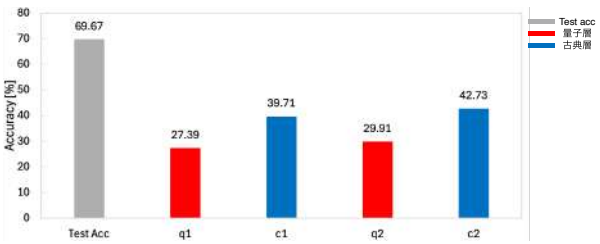


図 3: 畳み込み層単位のマスク実験結果

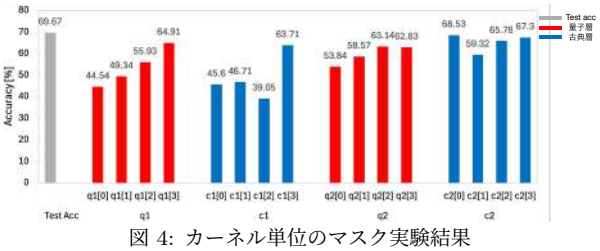


図 4: カーネル単位のマスク実験結果

5. おわりに

本稿では，量子・古典特徴の階層的融合によるマルチスケール画像分類として，MS-HQNN を提案した．比較実験の結果，提案手法は，複雑な特徴を持つ CIFAR-10 において全ての比較手法を上回る精度を達成した．今後は，量子層と古典層の配置・組み合わせの変更による分類精度向上や特徴抽出差異の分析を行う．

参考文献

[1] Senokosov, et al. Quantum machine learning for image classification. *Machine Learning: Science and Technology*, 5(1):015040, March 2024.

SVD を用いた LoRA と段階的知識蒸留による効率的なモデル圧縮

EP22061 熊澤綱佑　　指導教授：山下隆義

1. はじめに

Vision Transformer（ViT）は画像認識分野を代表する深層学習モデルの 1 つであり，高い認識性能を示す．その一方で，ViT はパラメータ数が多く，学習・推論時の計算量やメモリ消費が大きいという課題がある．この課題に対し，LoRA による学習パラメータ削減と知識蒸留によるモデル圧縮を組み合わせた手法として WeCoLoRA が提案されている．WeCoLoRA は，生徒モデル構築時に中間層を間引くため，中間層が持つ重要な特徴表現が失われる可能性がある．そこで本研究では，間引かれた層の知識を LoRA の初期化時に効果的に活用する．さらに，段階的な知識蒸留により，教師モデルの間引かれた層の特徴表現を継承した生徒モデルを構築する手法を提案する．

2. WeCoLoRA

WeCoLoRA[1] は，図 1 に示すように，事前学習済み ViT を教師モデルとし，一部の層を間引いて生徒モデルを構築する知識蒸留手法である．WeCoLoRA では，間引き率 r に基づいて教師モデルの層を等間隔に選択し，その重みをコピーして生徒モデルを構築する．構築後，生徒モデルの Transformer block 全体に LoRA を適用し，教師モデルとの知識蒸留を行う．LoRA の低ランク行列はランダムに初期化され，学習を通じて更新される．損失関数を式 (1) に示す．

$$\mathcal{L}\left(E_i^T, E_i^S\right) = \left|E_i^T - E_i^S\right| \tag{1}$$

ここで，E_i^T は教師の特徴ベクトル，E_i^S は生徒の特徴ベクトルを表す．

WeCoLoRA は，間引いた層の重みを蒸留や初期化に利用しないため，教師モデルが中間層で獲得した特徴表現を生徒モデルに十分に継承できないという課題がある．

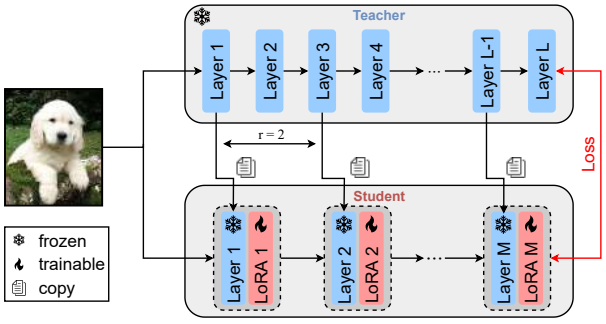


図 1：WeCoLoRA の概要

3. 提案手法

提案手法の概要を図 2 に示す．本研究では，間引いた層の知識を活用するために，その重み行列に SVD を適用して LoRA のパラメータの初期値に活用する．また，LoRA を浅い層から深い層へ段階的に適用することで，教師モデルの中間出力を再現する段階的蒸留手法を提案する．

3.1. SVD を用いた LoRA の初期化

生徒モデル構築時に，間引く層の重み行列に SVD を適用し，主要成分を LoRA の初期値として利用する．重み行列 W は式 (2) のように分解できる．

$$W = U\Sigma V^\top \tag{2}$$

ここで，U は出力空間，V は入力空間における基底ベクトルを表す．特異値の大きい成分が主要情報を担うことから，Σ の上位 k 成分を用いて低ランク近似を行う．低ランク行列 A, B を式 (3) に示す．

$$A = \sqrt{\Sigma_k}\, V_k^\top, \qquad B = U_k \sqrt{\Sigma_k} \tag{3}$$

3.2. 知識蒸留を用いた段階的 LoRA チューニング

間引いて構築した生徒モデルに対し，教師モデルの中間出力を用いた段階的蒸留を行う．段階的蒸留では，生徒モデルの浅い層から順に蒸留対象層を追加し，対応する教師モデルの中間層出力を用いて学習することで，各層が段階的に教師モデルの表現に近づくよう促す．

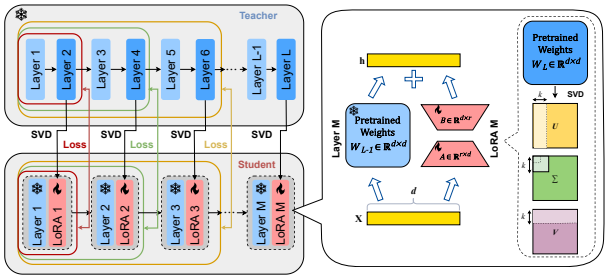


図 2：提案手法

4. 評価実験

事前学習済みの教師モデルから間隔 2 で間引いて生徒モデルを構築する．構築した生徒モデルに対して，提案手法および WeCoLoRA をそれぞれ適用し，精度を定量的に比較する．

4.1. 実験概要

本実験では，事前学習済み ViT-B を教師モデルとする．蒸留には ImageNet-1k の 1% のデータを用いる．WeCoLoRA は 15 エポック蒸留を行うのに対し，提案手法では 5 層分の段階的蒸留後，残りのエポック数で全体の蒸留を行う．総蒸留エポック数は WeCoLoRA と同じである．下流タスクの評価には ImageNet-1k または CIFAR-100 を用いて，50 エポック学習する．

4.2. 精度比較

各手法における精度を表 1 に示す．表 1 より，ImageNet-1k データセットにおける Top-1 accuracy は，WeCoLoRA が 64.50% であるのに対し，提案手法は最大 68.87% を達成し，4.37pt の精度向上が確認された．また，CIFAR-100 データセットでは，WeCoLoRA の 62.46% に対して，提案手法は最大 68.13% を達成し，5.67pt の精度向上が得られた．さらに，学習時間は，WeCoLoRA と比較して約 15.8% 短縮した．これにより，精度の向上と効率的なモデル圧縮の両立という観点から，本手法の有効性を確認した．

表 1：各手法の精度比較

手法	LoRA の初期値		学習数		学習時間	Top-1 accuracy	
	ランダム	SVD	段階的蒸留	蒸留		ImageNet-1k	CIFAR-100
教師モデル	-	-	-	-	-	81.37	80.57
WeCoLoRA	✓		-	15	0:20:33	64.50	62.46
提案手法	✓		1×5	10	0:17:17	66.27	67.39
		✓				**68.87**	**68.13**

5. おわりに

本研究では，間引く層の知識を活用し，生徒モデルを教師モデルの中間出力により近づける蒸留手法を提案した．今後は，様々なデータセットやモデルサイズを変更して，より汎化性能の高い生徒モデルの構築を目指す．

参考文献

[1] D. Grigore, *et al.*, “Weight Copy and Low-Rank Adaptation for Few-Shot Distillation of Vision Transformers,” arXiv preprint arXiv:2404.09326, 2024.

Mamba を用いたマウスの単一細胞解析のための基盤モデル構築

EP22064 小林 岳隼　　指導教授：山下 隆義

1. はじめに

次世代シーケンサを用いた Single-cell RNA sequencing (scRNA-seq) 解析の進展により，単一細胞内の遺伝子発現量の取得が可能となった．これに伴い，深層学習を用いた遺伝子解析技術も発展し，マウスの単一細胞データで事前学習を行った基盤モデルとして Mouse-Geneformer[1] が提案されている．本モデルは，遺伝子間の複雑な関係性を学習しており，細胞型の分類や in silico 摂動実験において高い性能を示す．しかし，そのアーキテクチャの基礎である Transformer は，計算量が入力長の 2 乗に比例して増大する．これにより，膨大なメモリ使用量がボトルネックとなり，現実的な計算リソースでは扱える遺伝子数に実質的な制約が生じる．そこで本研究では，入力長に対して線形な計算量で動作し，長い入力長でも効率的な学習が可能な Mamba[2] モデルを採用した，Mouse-GeneMamba を提案する．

2. Mouse-Geneformer

Mouse-Geneformer はマウスの単一細胞データの遺伝子解析を目的とした基盤モデルである．学習には大規模なマウスの単一細胞データセットである Mouse-Genecorpus-20M を用いる．Mouse-Genecorpus-20M では，各細胞内の遺伝子発現量の上位 2,048 個の遺伝子を抽出し，遺伝子トークン列に変換することで細胞文とする．作成した細胞文に対して，Masked Language Modeling (MLM) で学習を行うことで，正常なマウスの遺伝子間の関係を学習できる．さらに，このモデルを特定の臓器の単一細胞データで細胞型分類タスクにファインチューニングすることで，従来手法より正確な細胞型分類ができることを示した．

3. 提案手法：Mouse-GeneMamba

本研究では，Mouse-Geneformer の Transformer Encoder を Mamba ブロックに置換した Mouse-GeneMamba を提案する．本手法の全体概要を図 1 に示す．入力データの構築において，順位情報を持つ遺伝子トークンと正規化した遺伝子発現量をそれぞれベクトル化して統合することで，各遺伝子の順位と大きさを両方含む細胞文を作成する．次に，この細胞文を Mamba ブロックへ入力して，長大な遺伝子配列の大域的な文脈学習を行う．学習タスクには Next Token Prediction (NTP) を採用し，過去の遺伝子配列から次の遺伝子を予測することで，遺伝子ネットワークの因果関係を獲得する．また，本モデルの学習には， Mouse-Genecorpus-20M を拡張し，正規化された遺伝子発現量の数値を保持した大規模データセットを用いる．

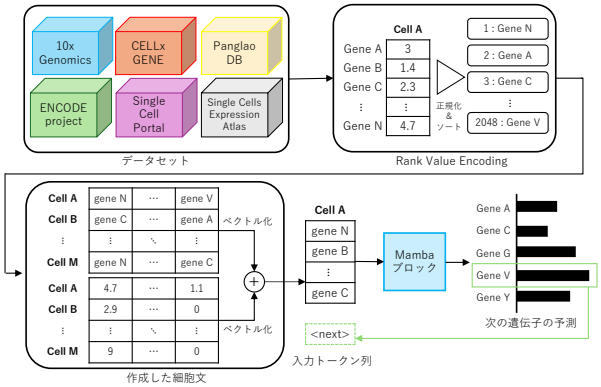


図 1：Mouse-GeneMamba の学習方法

4. 評価実験

提案手法の有用性を検証するために，複数の評価実験を行う．いずれの実験においても，事前学習モデルを細胞型分類タスクのデータセットでファインチューニングし，分類精度により評価する．事前学習タスクとして NTP を用いたモデルと MLM を用いたモデルを用いて，タスクの違いがモデル性能に与える影響を比較する．また，遺伝子発現量をモデル入力に統合することの有効性を検証する．具体的には，発現量の有無による精度比較に加え，Mouse-Geneformer との比較を行う．

4.1 評価実験結果

事前学習タスクとして NTP と MLM で学習したモデルの細胞型分類タスクの結果を表 1 に示す．表 1 より，事前学習タスクとして NTP を採用したモデルは多くの臓器で最も高い精度を達成した．

表 1：入力長 2,048 における事前学習タスクによる性能比較

事前学習タスク	脳	四肢の筋肉	腎臓	胸腺	舌	乳腺	心臓	脾臓	大腸	平均
NTP	97.6	99.6	95.3	96.8	94.2	98.8	98.1	98.7	94.2	97.0
MLM	97.9	99.5	94.6	97.2	94.7	98.9	97.5	98.5	93.1	96.9

提案手法の発現量の有無と Mouse-Geneformer の細胞型分類タスクの結果を表 2 に示す．各臓器の分類において，最も高い精度を赤色，低い精度を青色で示す．表 2 より，遺伝子発現量の有無の観点では，発現量を考慮しない設定においてより高い分類精度を示した．この結果から，遺伝子発現量を学習に用いる場合，本研究で採用した方法とは異なる利用方法を検討する必要がある．一方で，Mouse-Geneformer との比較においては，提案手法の方が平均精度において最も高い精度を達成し，有用性を確認した．

各入力長における精度変化に着目すると，Mouse-Geneformer は入力長を 2,048 から 8,192 に拡張した際，平均精度が 0.53 ポイント低下したのに対し，提案手法は 0.20 ポイントの低下に留まった．以上の結果から，本手法に用いている Mamba モデルは長い入力長に対しても情報の損失を抑えつつ特徴を抽出できることを示した．

表 2：提案手法と Mouse-Geneformer の細胞型分類精度

モデル	Mouse-GeneMamba						Mouse-Geneformer		
発現量	なし			あり			なし		
入力長	2,048	4,096	8,192	2,048	4,096	8,192	2,048	4,096	8,192
脳	97.6	98.0	97.8	96.7	96.4	95.7	96.7	96.2	95.3
四肢の筋肉	99.6	99.7	99.6	99.5	99.4	99.3	99.6	99.5	99.7
腎臓	95.3	94.4	95.1	93.8	93.2	92.5	94.8	94.7	94.1
胸腺	96.8	97.3	96.9	94.9	96.1	96.2	96.8	96.8	97.3
舌	94.2	94.4	93.7	92.8	92.2	91.1	94.6	94.4	93.9
乳腺	98.8	98.8	98.7	98.5	98.1	98.4	98.9	99.1	98.9
心臓	98.1	97.3	97.2	96.5	97.0	96.9	96.8	96.4	96.9
脾臓	98.7	98.5	98.6	98.3	97.9	97.8	98.6	98.6	98.5
大腸	94.2	94.3	93.9	93.3	93.4	93.4	92.3	92.3	89.8
平均	97.0	97.0	96.8	96.0	96.0	95.7	96.6	96.4	96.0

事前学習におけるメモリ使用量を表 3 に示す．表 3 より，提案手法は Mouse-Geneformer に比べてメモリ効率が向上しており，Mamba モデルを用いる有効性を確認した．

表 3：事前学習におけるメモリ使用量

モデル	Mouse-GeneMamba			Mouse-Geneformer		
入力長	2,048	4,096	8,192	2,048	4,096	8,192
メモリ使用量 (↓)	10.4GB	24.4GB	36.6GB	16.8GB	32.5GB	out of memory

5. おわりに

本研究では，Mouse-Geneformer の高いメモリ使用量や入力長の制限という問題を解決するためのモデルである Mouse-GeneMamba を提案した．また，発現量の値を考慮した新たなデータセットを構築し，そのデータで学習および細胞型の分類実験を行うことで，発現量を考慮する学習の有効性を検証した．

今後は，別の発現量の入力方法での学習や Mamba の内部構造の変更，データセットの大規模化を実施することで，モデルの分類精度向上を目指す．加えて，多様な下流タスクによる検証を行うことで，基盤モデルとしての汎用性と有用性を実証していく．

参考文献

[1] Keita Ito, *et al.*, “Mouse-Geneformer: A deep learning model for mouse single-cell transcriptome and its cross-species utility”, PLOS, 2025.

[2] Albert Gu, *et al.*, “Mamba: Linear-Time Sequence Modeling with Selective State Spaces”, COLM, 2024.

運転判断の因果関係を考慮した自動運転のための言語説明データセットの構築

EP22071 佐藤克昭　　指導教授：山下隆義

1. はじめに

視覚情報と言語知識を統合した Multimodal Large Language Model (MLLM) を用いた自動運転手法は，運転判断の根拠を言語として表現可能な手法として注目されている．MLLM の学習には，映像に対する判断とともに，その判断の根拠に対応する言語キャプションを大量に用いた学習が必要となる．既存のデータセットの言語キャプションには，道路構造や周辺物体の状態などの運転判断に関与する外界情報が多く含まれる．しかし，運転判断に至る過程を構成要素として明示的に表現することが経路予測精度にどのような影響を与えるのかについては，明らかになっていない．本研究では，運転判断に至る推論過程を自然言語として明示的に付与したデータセットを構築し，各説明要素の有無が経路予測精度に与える影響を比較する．

2. MLLM を用いた End-to-End 自動運転

MLLM を用いた自動運転の代表的なアプローチとして EMMA [1] が提案されている．EMMA は，複数のカメラで撮影した全周囲の画像と自車両の走行履歴およびナビゲーション指示を自然言語で入力する．視覚情報と言語情報を MLLM で共通の特徴表現へ変換することで，経路計画などの自動運転タスクを，タスク固有のプロンプトに基づき統一的に処理可能である．しかし，このような説明可能な自動運転を実現するためには，運転判断の根拠を明示的に含んだデータセットが必要となる．既存の自動運転データセットは，知覚・運動タスクを中心に構成されており，運転判断に至る因果関係が明示的に記述されていない課題がある．

3. 提案手法

本研究は，運転判断に至る推論過程を 3 つの説明タスクとして定義する．マルチセンサ情報を収録した nuScenes[2] に運転判断に関する自然言語記述を付与することで，推論過程を段階的に表現した自動運転データセットを構築する．データセット構築のフレームワークの概要を図 1 に示す．

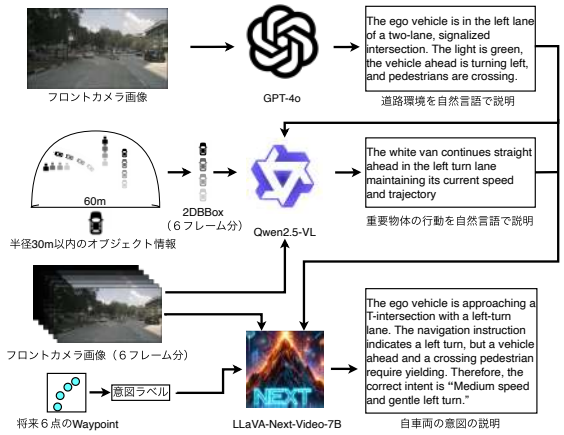


図 1: データセット構築のフレームワーク

道路構造の記述

フロントカメラ画像から，交差点，車線区分，信号，自車両位置などを言語キャプションで生成する．生成には空間理解能力に優れる GPT-4o を用いる．

重要物体の検出および将来行動説明

nuScenes に収録されている画像内の各オブジェクトの BBox，クラスラベルや属性情報を用いる．将来 3 秒間の自車両経路から半径 30 m 以内に存在する物体を重要物体と定義し，道路構造の説明を補助情報として与え，これらの重要物体の状態および将来行動を言語キャプションで生成する．生成には Qwen2.5-VL-7B-Instruct を用いる．

自車両の運転意図の説明

速度および軌道形状をもとに 16 種類の運転意図にルールベースで分類する．さらに，道路構造，重要物体の将来行動およびナビゲーション指示を根拠として，運転意図に対する判断理由を生成する．生成には LLaVA-Next-Video-7B を用いる．これらの説明要素を MLLM に推論させることで，経路予測精度の向上を図る．

4. 評価実験

提案手法の有効性を検証するため，評価実験を行う．LLaVA-Next-Video-7B をベースモデルとし，学習率 1e-5，エポック数 5 で学習を行う．比較手法として，追加学習を行わずに推論を行わせるプロンプトチューニングをベースラインとして用いる．評価指標は経路予測における平均 L2 誤差を用いる．また，各説明要素の有無を変化させ，それらが経路予測精度に与える影響を比較する．

4.1. 定量的評価

表 1 に各説明要素の有無による L2 誤差を示す．結果から，プロンプトチューニングのみでは，「意図説明＋経路予測」と比較して経路予測精度が低い傾向が見られた．また，「経路予測のみ」と比較すると，説明要素の付与により経路予測精度が向上し，特に「意図説明＋経路予測」で顕著な改善が確認された．これは，意図説明が意図ラベルに加えて道路構造および重要物体の将来行動を統合した表現であり，経路生成に有効に作用したためと考えられる．一方，「重要物体＋経路予測」のみの場合は，半径 30 m 以内の物体を重要物体と定義しているため，経路予測に直接関与しない物体が含まれ，精度が低下したと考えられる．

表 1: 説明要素の有無による平均 L2 誤差の比較

提案手法				平均 L2 誤差 [m] ↓
道路構造	重要物体	意図説明	経路予測	
			✓	2.45
✓			✓	2.29
	✓		✓	2.49
✓	✓		✓	2.37
		✓	✓	**2.26**
プロンプトチューニング				4.96

4.2. 定性的評価

図 2 に比較結果を示す．「経路予測のみ」では，横断歩行者が存在する状況においても直進する経路が生成されている．一方，「意図説明＋経路予測」では，横断歩行者を認識した上で「停止」という意図を選択しており，状況認識と運転行動の因果関係が明確に表現されている．

図 2: 説明要素が寄与した一例

5. おわりに

本研究では，自動運転タスクに特化したマルチタスクデータセットを構築し，有効性を検証した．その結果，説明要素の活用により経路予測精度が向上することが判明した．今後は，ベースモデルの変更や異なるドメインのデータセットでの検証を行うとともに，交通ルールを収録した Retrieval Augmented Generation を用いて未学習の地域の交通ルールに対応可能な手法を目指す．

参考文献

[1] J.-J. Hwang, et al., “EMMA: End-to-End Multimodal Model for Autonomous Driving”, arXiv preprint arXiv:2410.23262, 2024.

[2] H. Caesar, et al.,“nuScenes: A Multimodal Dataset for Autonomous Driving”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

自己教師あり学習による脳 MRI 解析基盤モデルの構築と評価

EP22114 羽澄 直弥　　指導教授：山下 隆義

1. はじめに

脳の皮質厚や皮質下領域の体積は，加齢や精神・神経疾患と関連していることが知られている．Magnetic Resonance Imaging（MRI）を用いた脳画像解析の従来の研究では，脳年齢予測や疾患分類等のタスクに特化した個別モデルで実現されている．そのため，複数の下流タスクに適用可能な基盤モデルの実現が期待されている．画像や自然言語処理の分野では，自己教師あり学習による基盤モデルの構築が有効とされている．そこで，本研究では，脳 MRI を対象とした基盤モデルの構築を目的とする．また，下流タスクにおいて基盤モデルがどの程度有用な表現を獲得できているかを検証する．

2. 従来研究

Siegel らは，MRI からの脳年齢予測タスクにおいて，CNN と Transformer の有用性を評価している [1]．結果として，いずれのモデルにおいても高精度な脳年齢予測が可能であることが示唆された．しかし，従来手法は教師あり学習を前提としており，大規模なラベル付きデータの確保や他タスクへの展開という課題がある．これに対し，自己教師あり学習を用いることで，ラベルなしデータからさまざまなタスクに共通する汎用的な特徴を抽出でき，基盤モデルの作成が可能と考える．これにより，脳年齢予測や疾患分類といった複数のタスクに特化したモデルを構築するためのデータ収集やアノテーションコストを低減できる．

3. 提案手法

本研究では，教師あり学習による脳 MRI 解析手法で必要となる大規模なラベル付きデータを用いず，自己教師あり学習により基盤モデルを構築する．

3.1 データセットと前処理

基盤モデルの構築に向けて，8 つのデータベースおよび公開サイトから 54,521 件のデータを収集した．データの内訳は，事前学習用データ 47,699 件，年齢予測タスク用データ 8,824 件，3 クラス分類の疾患データが 1,421 件，2 クラス分類の疾患データが 571 件である．なお，分類タスクの健常は年齢予測タスクと共有されている．各データベースでは撮影条件や保存形式が異なり，学習に悪影響を及ぼす可能性が高いため，3 つの前処理を実施した．図 1 に本研究で用いた前処理を示す．1 つ目に RAS 座標系の統一，2 つ目に 1mm ボクセルへのリサンプリング，3 つ目に Z スコアによる正規化を行う．これらの処理により，複数のデータバンク間に存在する差異を低減した統一的な MRI データを得ることができる．

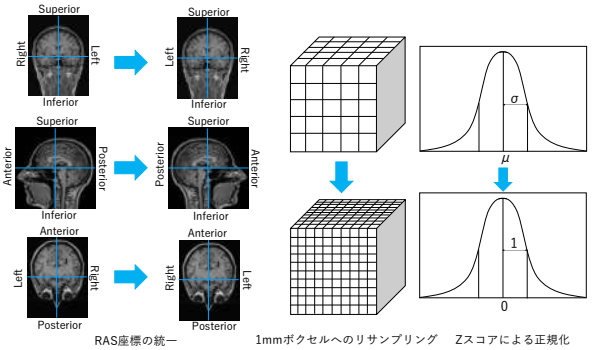


図 1：本研究における MRI データの前処理

3.2 基盤モデルの構築

事前学習では，自己教師あり学習に基づく手法として 3D Masked Autoencoder（3D MAE）を採用する．入力となる 3 次元 MRI を 3D パッチに分割し，ランダムに選択したパッチをマスクする．マスクされていないパッチのみを 3D Vision Transformer (3D ViT) の Encoder に入力し，得られた潜在表現から Decoder によりマスクされたパッチの復元を行う．このとき，Decoder で出力した復元パッチと元画像中の対応するパッチとの差を平均二乗誤差として計算し，その誤差を最小化するように学習を行う．この事前学習によって得られた Encoder を基盤モデルとして用いる．

4. 実験概要

事前学習は 47,699 件のデータを使用し，マスク率 75% で学習する．基盤モデルの汎用性を検証するために下流タスクとして脳年齢予測タスクと疾患分類タスクを実施し，教師あり学習で学習を行った ViT-Large と比較検証する．

4.1 脳年齢予測タスク

MRI から被験者の実年齢を予測する脳年齢予測タスクを行う．脳年齢予測タスクには，8,824 件のデータを使用している．脳年齢予測タスクにおける教師あり学習と自己教師あり学習の精度比較を表 1 に示す．表 1 より，教師あり学習モデルに対し，自己教師あり学習で事前学習を行ったモデルでは MAE が 1.59pt 低下し，決定係数 R^2 は 0.16pt 向上した．

表 1：脳年齢予測タスク

学習方法	MAE (↓)	R^2 (↑)
教師あり	6.11 ± 0.37	0.48 ± 0.06
自己教師あり	**4.52 ± 0.12**	**0.64 ± 0.02**

4.2 アルツハイマー分類タスク

疾患分類として，健常者と認知機能障害，アルツハイマーの 3 クラス分類を行う．3 クラス分類では，2,037 件の MRI データを 8:2 に分割した交差検証を用いている．表 2 に，アルツハイマーの 3 クラス分類タスクの結果を示す．表 2 より，自己教師あり学習で事前学習を行ったモデルの方が，教師あり学習で学習を行ったモデルと比較して精度が向上した．

表 2：3 クラス分類の精度比較

学習方法	Accuracy (%)	Macro F1 (%)
教師あり	44.31 ± 2.24	33.01 ± 1.30
自己教師あり	**65.43 ± 1.58**	**57.36 ± 1.72**

4.3 自閉スペクトラム症分類タスク

疾患分類として，健常者と自閉スペクトラム症の 2 クラス分類を行う．2 クラス分類では，1,099 件の MRI データを 8:2 に分割した交差検証を用いている．表 3 に，自閉スペクトラム症の 2 クラス分類タスクの結果を示す．表 3 より，自己教師あり学習で事前学習を行ったモデルの方が，教師あり学習で学習を行ったモデルと比較して精度が向上した．

表 3：2 クラス分類の精度比較

学習方法	Accuracy (%)	Macro F1 (%)
教師あり	55.16 ± 4.09	54.74 ± 4.14
自己教師あり	**58.90 ± 2.04**	**58.71 ± 2.99**

これらの結果から，大規模な事前学習を行うことでそのタスクに特化した学習と比較して，より良い精度を得られることが確認できた．

5. おわりに

本研究では，MRI を用いた脳画像解析のための基盤モデル構築に向け，3D Vision Transformer による大規模データを用いた自己教師あり学習の有効性を検証した．MRI データを用いて自己教師あり学習を行うことで，従来研究である教師あり学習を用いた手法と比較して精度の向上を確認した．このことから，特定のタスクに特化させるための教師あり学習よりも，自己教師あり学習による汎用的な特徴獲得の有効性が示された．今後は，拡散モデルによるデータ拡張や，他の事前学習手法の検討を行う．

参考文献

[1] N. T. Siegel, *et al.*“Do Transformers and CNNs Learn Different Concepts of Brain Age?”, Wiley, 2025.

MLLM と BEV 映像を用いた走行映像からの OpenSCENARIO 自動生成

EP22115 服部美月　　指導教授：山下隆義

1. はじめに

自動運転システムの安全性評価には，多様なシナリオでの検証が必須である．事故等の危険な状況を実環境で再現することは安全性の面で難しいため，シミュレータの活用が検討されている．しかしながら，人手による多様なシナリオ作成には多大なコストを要する．そこで，本研究では Multimodal Large Language Model (MLLM) を用いたシナリオの自動生成法を提案する．提案手法により，事故につながる危険なシナリオを効率的に作成でき，自動運転システムの安全性評価に貢献することを目指す．

2. OpenSCENARIO XML

OpenSCENARIO XML は，自動化システムと測定システムの国際標準化団体によって策定された，自動運転検証用の標準シナリオ記述フォーマットである．本規格は，車両や歩行者などの交通参加者を定義する Entities や，シナリオ推移を記述する Storyboard などのタグにより階層的に構成される．OpenSCENARIO が動的な挙動を記述するのに対し，OpenDRIVE は静的な道路環境を提供する．シナリオ生成には道路構造を定義する OpenDRIVE との連携が不可欠であり，両者の統合により整合性のとれた交通状況が再現可能となる．

3. 提案手法

本研究では，MLLM として Gemini 3 Pro [1] を用い，実環境の走行動画から OpenSCENARIO XML を自動生成する手法を提案する．本手法の概要を図 1 に示す．入力情報には，実環境のフロントカメラ映像に加え，逆透視投影変換を用いて生成した俯瞰（Bird's Eye View: BEV）映像を使用する．BEV 映像を補助的に用いることで，フロントカメラ映像だけでは困難な自他車両間の距離や挙動を捉えられると考える．

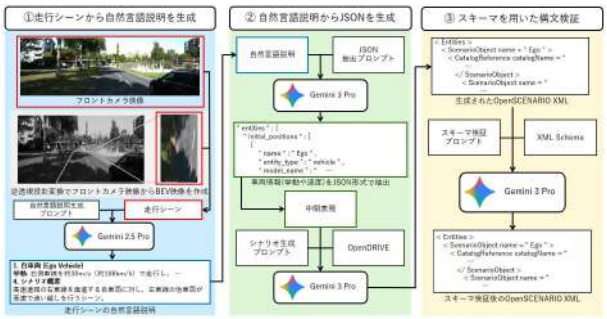


図 1: 提案手法の概要

3.1. 走行シーンから中間表現の生成

フロントカメラ映像と BEV 映像を MLLM に入力し，走行シーンの自然言語説明を生成する．次に，再度 MLLM を用いて自然言語説明から車両情報 (車両の挙動や速度) のパラメータを抽出し，JSON 形式に出力する．深い階層構造を持つ XML を MLLM に直接生成すると，タグの不整合や構文エラーが多発する．そこで，JSON 形式にすることで，中間表現としてデータの構造化を行う．MLLM の学習データ特性との親和性を活かし，生成エラーを抑制しながら複雑なシナリオ情報を正確に構造化することを可能にしている．

3.2. 中間表現から OpenSCENARIO XML の生成

生成した JSON 形式の中間表現を MLLM を用いて OpenSCENARIO XML 形式に変換する．MLLM に対し前段の中間表現に加え，OpenSCENARIO XML の記述例と道路情報が記述された OpenDRIVE ファイルを入力する．次に，生成された OpenSCENARIO XML に対し，データ型などが定義された XML Schema を用いた構文チェックを行う．この構文チェックを行うことで，シミュレータでの実行可能率を向上させる．

4. 定量評価

本実験では，OpenSCENARIO XML の生成の成功率とともに，走行軌跡の再現性を定量的に評価する．

4.1. 実験条件

本実験では，提案手法の有効性を検証するため，入力情報の違いによるシナリオ再現精度の比較を行った．比較手法として，図 2(a) のような前方映像のみを入力とする場合と，前方映像に加え図 2(b) のような BEV 映像を補助情報として入力する場合の 2 パターンで比較実験を行う．評価には KITTI Odometry Dataset [2] の直進，右左折を含む計 15 シーン（10 秒間）を用いる．評価指標には，生成した OpenSCENARIO XML をシミュレータ（Esmini）で実行したときの走行軌跡と実走行軌跡 (Ground Truth) との平均位置誤差 (ADE) と最終地点誤差 (FDE) を用いて比較する．

(a) フロントカメラ　(b) BEV

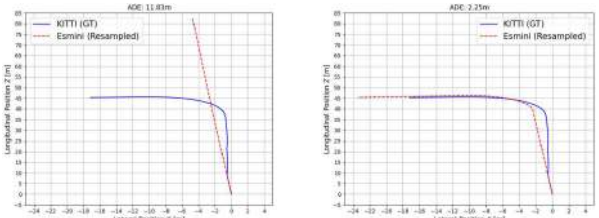
図 2: 入力シーンの例

4.2. 実験結果

表 1 にシーンごとの評価結果を，図 3 に生成したシナリオの軌跡と正解軌跡を示す．表 1 より，動画情報のみを用いた場合は，右左折シーンにおける精度が悪く，動作の再現性が低い結果となった．しかし補助情報（BEV）を用いた場合は，右左折時の ADE が改善した．また，全体平均においても ADE が半減し，FDE は 6 分の 1 程度となり，誤差の低減が確認できた．さらに，図 3 の通り，補助情報を追加した場合の軌跡は，正解軌跡により近いことが分かる．以上の結果から，OpenSCENARIO XML の生成は，右左折時の軌跡再現精度の向上に効果的である．

表 1: 生成シナリオの軌跡評価と生成成功率

シーン	Video Only			Video + BEV		
	ADE(m)	FDE(m)	成功率 (%)	ADE(m)	FDE(m)	成功率 (%)
右折	12.4	43.1	60.0	6.8	8.7	60.0
直進	2.6	6.3	80.0	2.6	6.2	100.0
左折	12.1	46.7	20.0	4.8	1.2	80.0
平均	9.0	32.0	66.7	4.7	5.4	**86.6**



(a) Video Only　(b) Video + BEV

図 3: 軌跡の比較

5. おわりに

本研究では，実環境映像から MLLM を用いて OpenSCENARIO XML を自動生成する手法を提案した．評価実験において，フロントカメラ映像に加えて BEV 映像の補助情報を入力することで，直進のみならず右左折シーンにおいても高精度なシナリオ生成が可能であることを確認した．今後はより長時間のシナリオへの対応や，他車両を含めた動的環境の再現精度の向上が課題である．

参考文献

[1] Gemini Team, Google, "Gemini 3 Pro Model Card.", Google DeepMind, 2025.

[2] A. Geiger, *et al.*, "Are we ready for autonomous driving? The KITTI vision benchmark suite", CVPR, 2012.

Mutual DDQN：深層強化学習エージェントの相互学習

EP22126 前田登生　　指導教授：山下隆義

1. はじめに

深層強化学習 (DRL) は，環境との相互作用を通じて累積報酬を最大化する方策を深層ニューラルネットワークで学習する枠組みである．代表的な手法である DQN は，状態 s において行動 a を選択したときの累積報酬の期待値を表す行動価値（Q 値）をニューラルネットワークで近似する．しかし，実際より Q 値を大きく見積もり，学習が不安定化する問題がある．この問題は，Q 値に含まれる誤差（推定誤差）とその誤差が増幅されやすい学習構造に起因する．Double DQN(DDQN) [1] は，この学習構造を改善する代表的な手法であるが，推定誤差を直接抑制する手法ではない．そこで本研究では，相互学習 (DML) を DDQN に導入し，学習の安定化を図る Mutual DDQN を提案する．

2. DDQN

従来の DQN は，選択した行動の Q 値が目標値 y に近づくように学習する．しかし，目標値 y は，次状態 s' における最大の Q 値に依存して求められる構造のため，推定誤差により大きく見積もられた Q 値が目標値 y に反映されやすい．DDQN では，目標値 y を求める際に，行動の選択と Q 値の算出に異なるネットワークを用いる．具体的には，学習によって逐次更新されるオンラインネットワーク Q_θ で次状態 s' において Q 値が最大の行動 a' を選択し，更新を遅らせるターゲットネットワーク Q_{θ^-} で行動 a' の Q 値を算出し，報酬 r と合わせて目標値 y を求める．目標値 y を式 (1) に示す．

$$y = r + \gamma Q_{\theta^-}(s', \mathrm{argmax}_{a'} Q_\theta(s', a')) \tag{1}$$

行動の選択に影響を与えた推定誤差が，Q 値の算出には反映されないため，推定誤差により大きく見積もられた Q 値を目標値 y に用いる傾向を抑制できる．

3. 提案手法：Mutual DDQN

本研究では，異なる初期パラメータを持つ 2 つの独立したネットワーク（$Q_{\theta_1}, Q_{\theta_2}$）が互いの出力を参照しながら学習する DML を DDQN に導入し，推定誤差を直接抑制する Mutual DDQN を提案する．提案手法の概要図を図 1 に示す．ネットワーク Q_{θ_i} の学習には，式 (2) および式 (3) で定義する 2 つの損失関数 $L^{(i)}_{MTD}$ と $L^{(i)}_{KL}$ の和を用いる．ここで $i \in \{1, 2\}$，$j \neq i$ とする．

$$L^{(i)}_{MTD} = \left(r + \gamma \,\mathrm{mean} Q(s', a') - Q_{\theta_i}(s, a)\right)^2 \tag{2}$$

$$L^{(i)}_{KL} = D_{KL}(p_j \| p_i) \tag{3}$$

$L^{(i)}_{MTD}$ は，平均 Q 値を用いた目標値と現在の Q 値の二乗誤差である．ここで Mean TD(MTD) は，DDQN の目標値 y に用いる次状態の Q 値を，2 つのターゲットネットワークで算出した平均 Q 値 $\mathrm{mean}Q(s', a')$ に置き換える手法である．これにより，目標値に含まれる推定誤差を緩和し，学習の安定化を図る．

$L^{(i)}_{KL}$ では，状態 s における 2 つのネットワークの出力に softmax 関数を適用して得られる行動分布 p_1 および p_2 の間で KL ダイバージェンスを最小化する．softmax 関数により行動間の相対関係を分布として表現し，すべての行動で Q 値の相対的な大きさと順位関係を学習する．これにより，ネットワーク間の行動分布を近づけ，行動選択の一致を促進する．

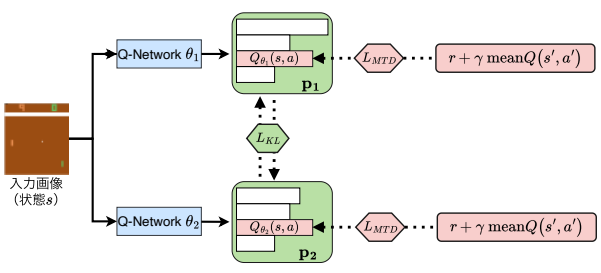


図 1: Mutual DDQN

4. 評価実験

4.1. 実験概要

行動数が異なる複数の Atari 2600 環境のタスク（AirRaid, Enduro, Seaquest）を用い，提案手法の評価を行った．比較手法は，標準的な DDQN および提案手法とする．提案手法に導入した損失関数の分析として MTD のみ，KL のみを使用した比較を行う．学習ステップ数は合計 500 万ステップとする．提案手法の評価には，並列に学習された 2 つのネットワークのうち高い累積報酬を達成した単一ネットワークを用いる．報酬は学習，評価ともに-1 から 1 にクリップする．学習過程は DDQN と提案手法の累積報酬の推移を比較し，学習後の評価指標として，複数エピソードの平均累積報酬，および Q 値と実際の累積報酬の差である Q 値の推定誤差を用いる．

4.2. 実験結果

図 2 に示した累積報酬の推移より，提案手法は学習初期の立ち上がりが早く，全タスクで DDQN を上回る累積報酬の推移を示した．

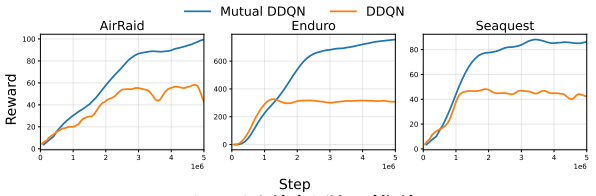


図 2: 累積報酬の推移

表 1 に示すように，全タスクにおいて提案手法は DDQN を上回る平均累積報酬を獲得した．また，MTD のみ，KL のみのいずれも DDQN から性能が向上しており，各損失の導入が有効であることを確認した．

表 1: 平均累積報酬

手法	MTD	KL	AirRaid	Enduro	Seaquest
DDQN			39	323	50
提案手法	✓		88	483	95
		✓	74	834	81
	✓	✓	**119**	**895**	**101**

表 2 に示した Q 値の推定誤差より，全ての条件で推定誤差は正の値を示し，実際の累積報酬よりも推定する Q 値が高くなる傾向を確認した． 提案手法は全タスクで推定誤差が最小であり，Q 値を過剰に見積もる傾向を改善している． MTD のみ，KL のみでも，DDQN と比較して推定誤差は小さいが，KL と比べて MTD がより推定誤差の減少に寄与している．

表 2: Q 値の推定誤差

手法	MTD	KL	AirRaid	Enduro	Seaquest
DDQN			2.1	8.2	2.9
提案手法	✓		0.6	2.5	0.9
		✓	1.0	6.7	2.9
	✓	✓	**0.4**	**2.4**	**0.8**

5. おわりに

本研究では，Q 値推定を改善するために DDQN に DML を導入した Mutual DDQN を提案した．評価実験では，提案手法は従来手法を上回る累積報酬と推定誤差の改善も見られ，提案手法の有効性が示された．今後の展望として，単一ネットワークではなく複数ネットワークの平均 Q 値に基づく行動選択を DML で改善することや Actor-Critic 等の DRL 手法への応用が考えられる．

参考文献

[1] H. van Hasselt et al., *Deep Reinforcement Learning with Double Q-learning*, AAAI, 2016.

時間的整合性を考慮したテキストからの 2 次元モーション生成

EP22058 川本寛和　　指導教授：山下隆義

1. はじめに

テキストからの 2 次元モーション生成の先行研究として，2CM-GPT[1] が提案されている．2CM-GPT は，3 次元モーション生成モデルと異なり，収集が容易な 2 次元モーションのデータセットを学習に利用できる．そのため，生成に失敗した 2 次元モーションを収集して，データセットを動的に拡張することにより，効果的なファインチューニングが実現できる．一方で，2CM-GPT はいくつかの課題もある．1 つ目は，各フレームのモーションを独立で生成するため，時間的な整合性が低い．2 つ目は，テキストとモーションを対応付けることなく混在して学習するため，テキストとモーションの整合性が損なわれる．これらの課題を解決するために，本研究ではテキストとモーションをアテンション機構で関連付けるとともに，時間的な整合性を考慮する手法を提案する．

2. 2 次元モーション生成

2CM-GPT は，2 次元のモーションを生成する代表的な手法である．2CM-GPT のモデル構造を図 1 に示す．2CM-GPT は，Motion Tokenizer で人間のモーションを離散的なトークンに変換する．そして，テキストも同様に Text Tokenizer でトークンに変換する．これらを連結させた Mixed Tokens を Language Encoder に入力して潜在ベクトルを獲得する．潜在ベクトルをもとに Language Decoder が出力した Output Tokens を Motion Tokenizer に入力して 2 次元のモーションを生成する．

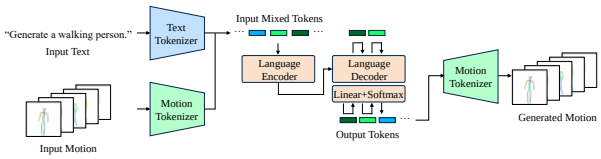


図 1：2CM-GPT のモデル構造

3. 提案手法

本研究では，2CM-GPT の学習アプローチが抱える「時間的な整合性」と「テキストとモーションの整合性」の不一致を解決する手法を提案する．提案手法のモデル構造を図 2 に示す．なお，提案手法がテキストと 2 次元モーションの関係性を学習する過程を Training Phase，テキストから 2 次元モーションを生成する過程を T2M Phase とする．

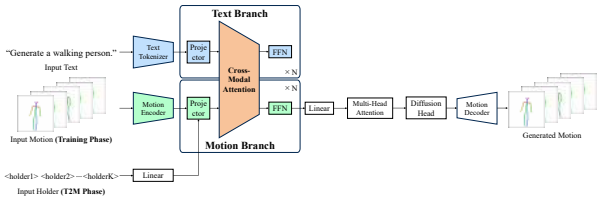


図 2：提案手法のモデル構造

3.1. 学習

Training Phase では，VAE で学習された Motion Encoder を用いて人間の連続的な動作を離散化しない形で潜在空間に埋め込む．この埋め込み特徴を Motion Branch に入力する．テキストは，Text Tokenizer で埋め込み特徴とし，Text Branch に入力する．各ブランチにおいて，埋め込み特徴を Projector に入力し，Query, Key, Value ベクトルを抽出する．そして，Motion Branch と Text Branch の各ベクトルを同じ Cross-Modal Attention に入力して，Motion Branch の Value ベクトルにテキスト特徴を反映させる．Motion Branch の出力を Multi-Head Attention に入力して潜在ベクトルを獲得する．潜在ベクトルを Diffusion Head に入力し，後述の T2M Phase のための逆拡散過程を学習する．損失関数には，生成モーションと実モーションの差を評価する特徴再構成損失，対応関係にあるモーションとテキストがクロスモーダルな特徴空間上で近接するよう制約する分類損失，逆拡散過程における潜在表現の再構成誤差を評価する拡散損失を用いる．

3.2. モーション生成

T2M Phase では，Motion Branch と Text Branch の Cross-Modal Attention によって，Motion Branch に入力した Holder にテキスト特徴を反映させる．Motion Branch の出力を Multi-Head Attention に入力して，潜在ベクトルを獲得する．潜在ベクトルを Diffusion Head に入力して，逆拡散過程によるノイズ除去を行う．その後，VAE で学習された Motion Decoder を用いて，潜在ベクトルから 2 次元のモーションを生成する．

4. 評価実験

2CM-GPT との比較実験により，提案手法の有効性を示す．

4.1. 定量的評価

表 1 より，提案手法は FID が低いことから，2CM-GPT と比べて 2 次元モーション生成精度の向上を確認した．一方，2CM-GPT の Diversity は提案手法よりも高い．その要因として，モーション生成精度が十分でないために，類似の指示文に対しても多様なモーションを生成することが考えられる．

表 1：テキストからの 2 次元モーション生成精度の比較

Method	FID ↓		Diversity ↑	
	real	gen	real	gen
2CM-GPT	-1.37×10^{-9}	32.36	16.96	19.28
提案手法	-5.24×10^{-9}	12.15	16.69	12.92

そこで，多様性の要因を検証するために，表 1 の Diversity と，後述の定性的評価で用いる図 3 の指示文のみを与えて生成されたモーションの Diversity を比較する．表 2 より，2CM-GPT は同一の指示文のみを与えた場合でも Diversity が高い傾向を示したことから，上記の可能性を裏付ける傾向が確認された．

表 2：指示文ごとの生成モーションの Diversity の比較

Method	Multi Instruction	Single Instruction
2CM-GPT	19.28	18.85
提案手法	12.92	6.07

4.2. 定性的評価

2CM-GPT と提案手法に同じ指示文を与えて生成させた 2 次元モーションを図 3 に示す．図 3 の手の動きに注目すると，提案手法の生成モーションは 2CM-GPT と比較して，指示文の動作内容を正確に反映していると判断できる．

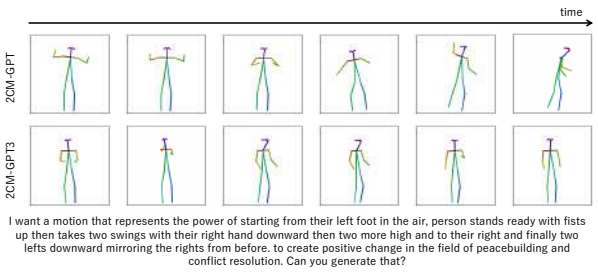


図 3：テキストからの 2 次元モーション生成結果の可視化

5. おわりに

本研究では，2CM-GPT と提案手法の評価実験を行い，提案手法の有効性を示した．今後は，提案手法で生成したモーションを用いてポーズ誘導による人物動画生成を実施し，実用性を検証する．

参考文献

[1] R. Inoue, *et al.*, "2D Motion Generation Using Joint Spatial Information with 2CM-GPT", VISIGRAPP, vol.2, pp.582-590, 2025.