Vision Language Model を用いた走行画像認識性能の検証

表野理人[1]　針屋慶吾[1]　福田有輝也[1]　米陀佳祐[1]　菅沼直樹[1]
[1]金沢大学
masato1226@stu.kanazawa-u.ac.jp, k.yoneda@staff.kanazawa-u.ac.jp,
{hariya1103, fukuda-yukiya, suganuma}@se.kanazawa-u.ac.jp

概要

自動運転技術の関連研究に用いられる走行データを効率良く応用する場合，交通状況を理解するための詳細な認識が重要である．そこで，走行データの認識に，走行中の自車から撮影されたカメラ画像を深層学習ベースモデルによる分類で得られる情報を用いることが有効な手段として挙げられる．本研究では，画像情報のみを扱う Convolutional neural network(CNN)ベースモデルと画像とテキストの両方の情報を扱える Vision Language Model(VLM)の走行画像認識性能を評価・比較する．実験の結果，VLM が詳細なラベルである走行エリア，走行場所，路面状況，渋滞状況，掠れた白線の有無，工事規制されている車線の有無のラベルにおいて，CNN ベースモデルの精度を上回ったことが確認された．

1 はじめに

近年，自動運転に関する技術は過疎地域の新たな交通手段の提供として広く研究されている．自動運転車にはカメラや LiDAR といった複数のセンサが搭載されており，その情報をもとに自車や周辺環境を把握し，自動運転が可能になる．そのような自動運転車を市街地などで走行させたデータは，走行データとして自動運転分野に広く利用されている．この走行データを有効に活用するには，交通状況の詳細な認識が重要である．そこで，走行データに含まれる全方位カメラ画像の分類情報を深層学習ベースモデルで得ることが，走行データの認識に有効な手段として挙げられる．深層学習ベースの画像分類とは，ニューラルネットワークを用いて抽出された画像の特徴量から，画像が属するクラスを分類するタスクである．画像分類には一般的に CNN ベースモデルなどが用いられるが，近年では，画像エンコーダとテキストデコーダから構成される Transformer ベースの VLM が登場しており，画像の特徴量のみ扱う CNN ベースモデルと画像と文章の特徴量を扱う VLM では，どちらが画像分類の性能が高いのかには検証の余地がある．図 1 の(a)は CNN ベースの画像認識モデルであり，畳み込み層で構成されるエンコーダで画像の特徴量を抽出し，線形層と活性化関数で構成されるデコーダでクラスごとの確率を出力する．(b)は CNN ベースの物体検出モデルであり，畳み込み層で構成されるエンコーダで画像の特徴量を抽出し，畳み込み層とアップサンプリング，活性化関数で構成されるデコーダで特徴量から，物体領域の座標(bounding box：BBOX)とそのクラスごとの確率を出力する．(c)は Transformer ベースの VLM であり，Transformer と Multilayer perceptron(MLP)で構成される画像エンコーダで抽出された画像の特徴量を別の MLP でテキストの特徴量に変換し，テキストと組み合わせて，Transformer と MLP，線形層，活性化関数で構成されるテキストデコーダ(Large Language Model : LLM)に入力してテキストトークンの確率を出力する．

本研究では，事前学習されたこれらのモデルに対して走行画像のデータセットでファインチューニングを行い，画像認識性能を評価・比較し，優れているモデルの検証を行う．

2 関連研究

本研究では，CNN ベースモデルの画像認識・物体検出モデルに ConvNext V2[1]，YOLOv11[2]，VLM の画像認識モデルに LLaVA[3]を用いる．

ConvNext V2： CNN ベースの画像認識モデルである ConvNext V2 は，Vision Transformer の構造を取り入れており， ImageNet-21k[4]での事前学習後に，ImageNet-1k[5]でファインチューニングされたモデルである．

YOLOv11： CNN ベースの物体検出モデルである YOLOv11 は，小さな物体や遮られた物体を高精度に検出でき，COCO[6]で学習されたモデルである．

LLaVA：VLM である LLaVA は，画像エンコーダに CLIP[7]の画像エンコーダ部分を，LLM には Mistral

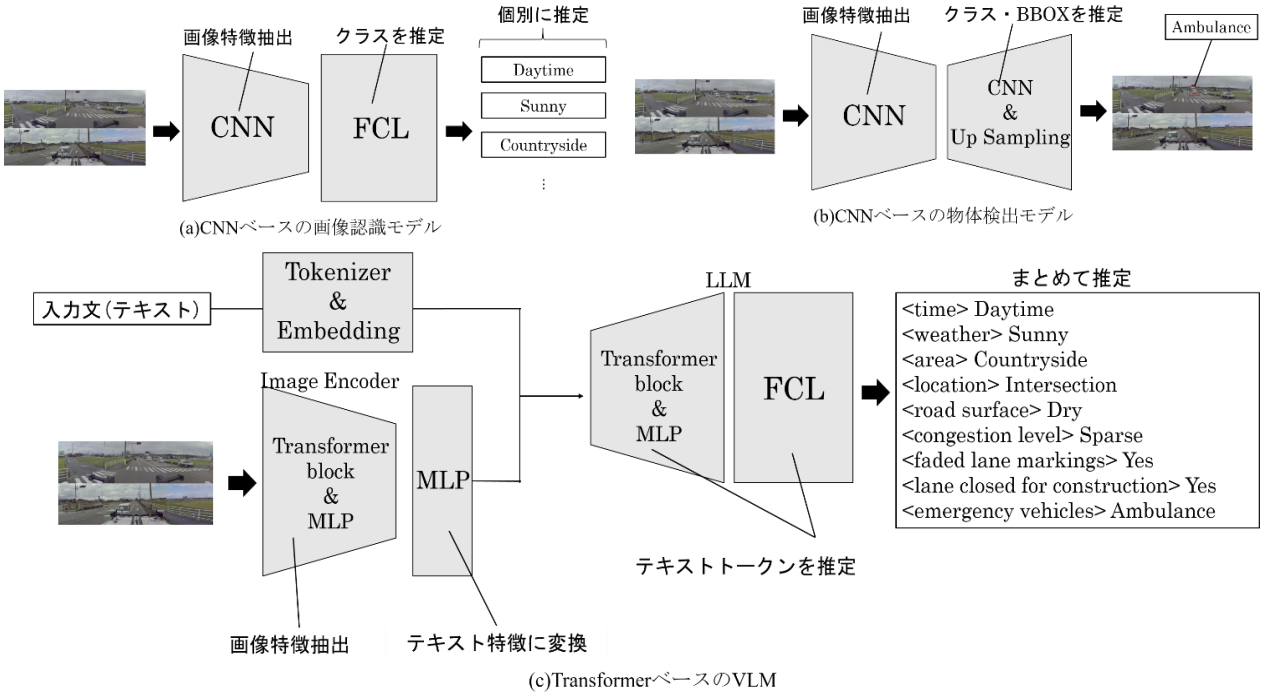

図 1　CNN ベースモデルと VLM の構造

表 1　ラベルとクラスの一覧

label	class
time	Daytime, Evening, Night, Unidentifiable
weather	Sunny, Rainy, Snowy, Unidentifiable
area	Urban, Residential, Countryside, Mountain, Tunnel, Highway, Ramp
location	Straight, Curvy, Intersection, Parking
road surface	Dry, Wet, Snow covered
congestion level	Dense, Normal, Sparse
faded lane markings	Yes, No
construction	Yes, No
emergency vehicles	Ambulance, Fire truck, Police car, None

7B[8]を採用しているモデルである．LLaVA では，画像の特徴量をテキストの特徴量に変換する MLP の学習を，他の重みは固定したまま，事前学習している．その後，画像エンコーダの重み以外の重みを，ファインチューニングしている．

また，VLM による走行画像認識を行った研究[9, 10, 11]では，時間や天候，走行エリアなど単純な交通状況の認識は行われているが，路面状況や渋滞状況，掠れた白線の有無などの詳細な交通状況の認識は行われていない．そのため，このような詳細な交通状況の認識に関しても，VLM による画像認識が有効であるかを調査する必要がある．

3 VLM による画像分類

VLM には，既存の CNN ベースの画像認識・物体検出モデルモデルでは扱えないテキストの特徴量を扱える利点がある．また，CNN ベースモデルでは，ラベルの特性ごとのに特化させたモデルを個別に用意しなければ十分な精度を出すことが難しいが，VLM では画像の複雑なコンテクストの予測をテキストとして一度に出力でき，効率良く走行画像の詳細な分類が可能である．

本研究で VLM に行わせるタスクは，表 1 に示すような 9 項目のラベルに関して，走行画像がどのクラスに属しているかを予測させる画像分類である．図 2 に，簡易的な入・出力文の例を示す．赤文字で記した部分は prompt 設計を工夫した箇所である．データを取り扱いやすくするために，「フォーマットに従って回答しなさい」と指示文を与え，図 2 で示されている出力文の形式で回答させる．また，マルチラベルの回答をさせるために，「area, emergency vehicles ラベルの質問は，複数回答して

もよい」と加えた．さらに，area ラベルでは Highway・Mountain などのマルチラベルである場合があり，画像だけでそのようなラベルを予測することは困難であると考えられる．そのため，自車の速度情報が有効であると考え，「質問にはあなたの車が v km/h で走行していることを考慮しなさい」と加える．v [km/h]は，全方位カメラ画像が撮影されたタイミングでの自車の速度である．

入力文

You are an excellent driver behind the wheel.

Analyze six connected images of a given driving scene captured by a your car's surround view camera.

Answer the following questions by selecting from the options given, following the format given. However, for the <area> and <emergency vehicle> questions, you may use AND to make multiple selections. Also, for the questions, take into account that your vehicles is traveling at v km/h.

<time> When is the time of day?
options: Daytime, Evening, Night, Unidentifiable

<weather> How is the weather?
options: Sunny, Rainy, Snowy, Unidentifiable

<area> What is the area in which your car is traveling? (You may select multiple options in this question.)
options: Urban area, Residential area, Countryside, Mountain roadway, Tunnel, Highway, Highway ramp

<emergency vehicles> Which of the emergency vehicles listed in the options are driving near the roadway your car is traveling on? If none o f the emergency vehicles listed in the options are present, respond with None. (You may select multiple options in this question.)
options: Ambulance, Fire truck, Police car, None

出力文

<time> Daytime <Weather> Sunny <area> Urban area <location> Intersection <road surface> Dry <congestion level> Sparse <faded lane markings> No <lane closed for construction> No <emergency vehicles> None

図 2　簡易的な入・出力文の例

4 実験

CNN ベースモデルは走行画像，VLM は走行画像とテキストのペアのデータセットでファインチューニングを行い，評価と比較を行う．

4.1 ファインチューニング概要

ConvNext V2 および YOLO v11 では，出力層である線形層の出力次元数を各クラス数に変更し，全パラメータのフルファインチューニングを行う．一方で，LLaVA では，モデルが大規模なため，Low Rank Adaption(LoRA)[12]で画像エンコーダの注意機構，LLM の注意機構・FFN 層のファインチューニングを行う．LoRA は効率的なファインチューニング手法であり，元のモデルに少数の重みを追加し，その重みのみを学習させるものである．

4.2 データセット

表 1 に走行画像データセットのラベルとそのクラスの一覧を示す．データセットの画像枚数については，train が 3493 枚，validation が 679 枚，test が 684 枚である．

4.2.1 走行画像の作成

走行画像は図 3 に示すように，自車に取り付けられた全方位カメラから撮影された 8 方向の画像のうち，前・後方 3 枚の計 6 枚を連結させる．また，繋ぎ目を滑らかにするために，左右斜め前・後方の画像を 1/4 カットしてから連結させている．

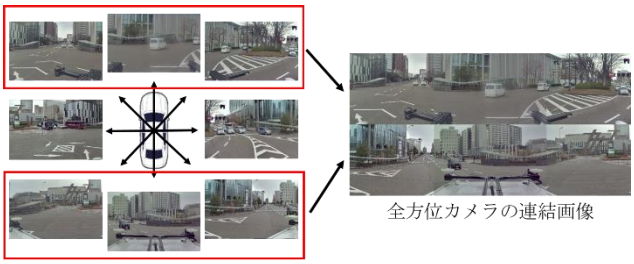

図 3　走行画像の作成方法

4.2.2 画像データの拡張

モデルの汎化性能向上，過学習の防止，データ不足の補間のためにデータ拡張を行う．図 2 中の time ～construction ラベルの学習については，ConvNext V2 をラベルごとにフルファインチューニングするが，faded lane markings ラベル以外のラベルでは，画像をランダムに左右反転，-15° ～15° 回転させる．faded lane markings ラベルでは，白線を強調するために，先述した処理に加え，ランダムにグレースケール変換，明度・輝度・コントラスト変更させる．また，emergency vehicles ラベルの学習については，YOLOv11 をフルファインチューニングするが，回転を行うことで緊急車両が見えなくなる場合を考慮して左右反転のみを行う．さらに，LLaVA の LoRA ではすべてのラベルに関して一度に学習を行うが，データ数の少ない construction，emergency vehicles ラベルのみ左右反転を行う

4.2.3 テキストデータの拡張

テキストデータに関しても 4.2.2 節と同様の目的でデータ拡張を行う．拡張方法としては，図 2 のラベル質問と各ラベルの options の順序をランダムに入れ替え，出力文は入れ替え後のラベル質問の順序に合わせる．

表 2 モデルごとの各ラベルに対する Macro F1 値： road→road surface，congestion→congestion level，markings→faded lane markings，construction→lane closed for construction，emg veh→emergency vehicles を表す．

	Macro F1(%)								
モデル	time	weather	area	location	road	congestion	markings	construction	emg veh
ConvNext V2(FT)	88.8	**81.7**	73.0	64.2	69.8	72.5	60.2	77.5	–
YOLOv11(FT)	–	–	–	–	–	–	–	–	**80.0**
LLaVA	57.4	56.5	44.3	32.1	70.0	52.9	46.3	54.5	2.72
LLaVA(LoRA)	87.9	75.5	74.1	83.8	**89.9**	**81.9**	**80.0**	92.7	73.9
LLaVA(LoRA) + velocity	**91.7**	76.0	**85.1**	**84.9**	87.5	79.7	69.4	**93.8**	77.6

4.4 評価方法

本研究では，4.2 節で述べたテストデータを用いて，CNN ベースモデルである ConvNext V2 および YOLOv11 と VLM である LLaVA の走行画像分類性能の評価と比較を行う．評価指標として，画像分類の評価に広く用いられる適合率(Precision)，再現率(Recall)，Macro F1 値を利用する．

4.5 実験結果

表 2 に，モデルごとの各ラベルに対する Macro F1 値の結果を示す．また，表 2 中の LLaVA は元の重みのモデル，LLaVA(LoRA)は図 1 に示す入力文に速度に関する文章を含まない場合での LoRA モデル，LLaVA(LoRA)＋velocity は入力文に速度に関する文章を含む場合での LoRA モデルである．表 2 から，CNN ベースモデルと比較して，LLaVA の LoRA モデルが全体的に高精度であることが分かった．しかし，weather ラベルの精度は，LLaVA の LoRA モデルが ConvNext V2 を下回った．また，LLaVA(LoRA)に比べて，LLaVA(LoRA)＋velocity では，area ラベルの精度の向上が確認された．

5 考察

表 3 に，ConvNext V2 および LLaVA(LoRA)＋velocity の weather ラベル内の Rainy クラスにおける Precision と Recall を示す．また，表 4 に，LLaVA(LoRA)および LLaVA(LoRA)＋velocity の area ラベル内の Highway･Ramp クラスにおける Precision と Recall を示す．表 3 から，LLaVA(LoRA)＋velocity の Rainy クラスの Recall が ConvNext V2 を大幅に下回っており，未検出が多いことが分かる．そこで，図 4 の左図の Rainy クラスの走行画像に対する入・出力文に Chain-of-Thought Prompting[13]である「How did you determine the answers for the questions? Let's think step by step．」を加えた入力で，再度推論を行うと，「空からの日差しがあり，視界が明瞭であることから晴れだと判断した．」と結果が得られた．このことから，LLaVA では画像のより詳細な部分まで考慮しすぎてしまい，CNN ベースモデルの ConvNext V2 に比べて，全体的な特徴から予測が行えなかったと考えられる．また，表 4 の LLaVA(LoRA)＋velocity の結果から，図 4 の右図の Tunnel・Highway クラスなどの画像の特徴だけでは Highway クラスだと予測することが難しい走行画像も正しく予測ができていたことから，入力文に速度に関する文章を加えることの有効性が確かめられた．

表 3 Rainy の Precision と Recall

	Precision(%)	Recall(%)
	weather	
モデル	Rainy	
ConvNext V2	47.0	**84.9**
LLaVA(LoRA) + vel	**95.8**	24.7

表 4 Highway・Ramp の Precision と Recall

	Precision(%)		Recall(%)	
	area			
モデル	Highway	Ramp	Highway	Ramp
LLaVA(LoRA)	**75.0**	50.0	47.1	8.16
LLaVA(LoRA) + vel	66.7	**86.8**	**74.5**	**67.3**

図 4 Rainy(左)および Tunnel・Highway(右)の画像

6 おわりに

本研究では，自動運転車の走行画像の認識性能の評価･比較を CNN ベースモデルと VLM を用いて行った．結果として，画像だけでなくテキストの特徴量も扱える VLM の走行画像認識性能が CNN ベースモデルに比べて高いことが判明した．また，VLM には，より詳細な交通状況の認識を高精度に行えることも確かめられた．今後の検討として，天候，掠れた白線の有無のラベルなどの精度改善には，prompt 設計のさらなる工夫や損失関数の変更などを行うことなどが挙げられる．

参考文献

[1] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, Saining Xie. ConvNext V2: Co-designing and Scaling ConvNets with Masked Autoencoders. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 16133-16142, 2023

[2] Rahima Khanam, Muhammad Hussain. YOLOv11: An Overview of the Key Architectural Enhancements. **arXiv preprint arXiv:2410.17725**, 2024.

[3] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 26286-26296, 2024.

[4] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, Lihi Zelnik-Manor. ImageNet-21K Pretraining for the Masses. **arXiv preprint arXiv:2104.10972**, 2021.

[5] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. **International Jounal of Computer Vsion(IJCV)**, Vol. 115, pp. 211-252, 2015.

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence ´ Zitnick. Microsoft coco: Common objects in context. **Computer Vision-ECCV 2014**, pp. 740-755, 2014.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. **arXiv preprint arXiv:2103.00020**, 2021.

[8] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, William El Sayed. Mistral 7B, **arXiv preprint arXiv:2310.06825**, 2023.

[9] Francisco Romero, Caleb Winston, Johann Hauswald, Matei Zaharia, Christos Kozyrakisv. Zelda: Video Alalytics using Vision-Language Models. **arXiv preprint arXiv:2305.03785**, 2023

[10] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, Hang Zhao. DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models. **arXiv preprint arXiv:2402.12289**, 2024

[11] Zian Guo, Zakhar Yagudin, Artem Lykov, Mikhail Konenkov, Dzmitry Tsetserukou. VLM-Auto: VLM-based Autonomous Driving Assistant eith Human-like Behavior and Understanding for Complex Road Scenes. **arXiv preprint arXiv:2405.05885**, 2024

[12] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen. LoRA: Low-Rank Adaption of Large Language Models. **arXiv preprint arXiv:2106.09685**, 2021.

[13] Jason Wei，Xuezhi Wang，Dale Schuurmans，Maarten Bosma，Brian Inchter，Fei Xia，Ed Chi，Quoc Le，Denny Zhou．Chain-of-Thought Prompting Elicits Reasoning in Large Language Models．**arXiv preprint arXiv:2201.11903**, 2022.