プロービングとアンサンブル学習による大規模視覚言語モデルのハルシネーション検出

宮里龍平 岡本一志 軽部幸起 柴田淳司 原田慧
電気通信大学 情報学専攻
{miyazato, harada}@uec.ac.jp

概要

大規模視覚言語モデル（VLM）は，多くのマルチモーダルタスクで高性能を示す一方，視覚的根拠に基づかない内容を生成するハルシネーションが課題である．本研究では，VLM の複数の内部表現を用いてハルシネーション検出器を学習し，それらをアンサンブルすることで視覚質問応答（VQA）におけるハルシネーションを検出する手法を提案する．実験の結果，提案手法は単一検出器や特徴量の単なる連結による検出手法を一貫して上回る性能を示し，内部表現ごとに検出器を学習してアンサンブルすることの有効性を確認した．さらに，注意機構に基づく検出器をアンサンブルすることで，未知ドメインに対する評価においても性能低下が緩和される傾向が確認された．

1 はじめに

大規模視覚言語モデル（Vision Language Model; VLM）は，視覚質問応答 (VQA)，画像キャプション生成などのタスクにおいて著しい性能向上を示している [1, 2, 3, 4]．一方，画像中に存在しない事実や視覚的根拠に基づかない内容をもっともらしく生成してしまう「ハルシネーション」は，信頼性や安全性の観点から深刻な問題 [5, 6, 7] であり，その検出や抑制は重要な課題となっている．

VLM のハルシネーション検出に関する既存研究は，主にモデルの出力 [8] と内部表現に基づく手法 [5, 9, 10, 11] に分類できる．出力ベースの手法では，生成テキストそのものや，生成確率や自己整合性などの指標を利用する．これらの手法は，モデルに依存しにくいが，リアルタイム検出が難しく，モデルが自信を持って誤っている場合は検出が難しいという問題がある．内部表現ベースの手法では，注意機構の出力や隠れ状態といったモデル内部を解析

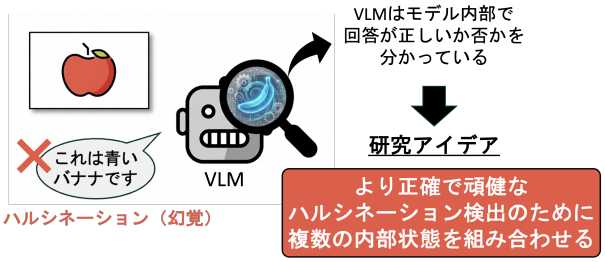

図 1 研究アイデア．VLM におけるハルシネーションをモデルの内部状態を複数利用して検出する．

（プロービング）し，ハルシネーションに特徴的な兆候を捉えることを目指している．これらの手法では，検出器の事前学習によりリアルタイム検出が可能であり，出力のみからは捉えにくい視覚と言語の対応関係や内部的な判断過程に関する情報も活用できる．しかし，既存研究の多くは単一の内部表現や単一検出器を使用しており，ハルシネーションの多様な性質を十分に捉えられていない可能性がある．

本研究では，複数の内部表現に基づく検出器をアンサンブルすることで，VLM のハルシネーション検出性能を向上させる手法を提案する（図 1）．具体的には，VQA に回答する際の VLM の注意機構の出力と隠れ状態を抽出して特徴量とし，生成中の回答がハルシネーションであるか否かを目的変数とするハルシネーション検出モデルを各層，各ヘッドごとに学習する．これらの検出器をアンサンブルすることで，個々の内部表現が捉える異なる情報を統合し，より頑健な検出を実現する．

VQA に関する CRAG-MM データセット [12] を用い，複数の VLM を対象として内部状態に基づくハルシネーション検出器を学習する．単一検出器，特徴量連結，およびアンサンブル（提案手法）を比較し，データ全体に加えて，学習ドメイン内（In Domain; ID）および未知ドメイン（Out-of-Domain; OOD）の両設定において検出性能を評価する．

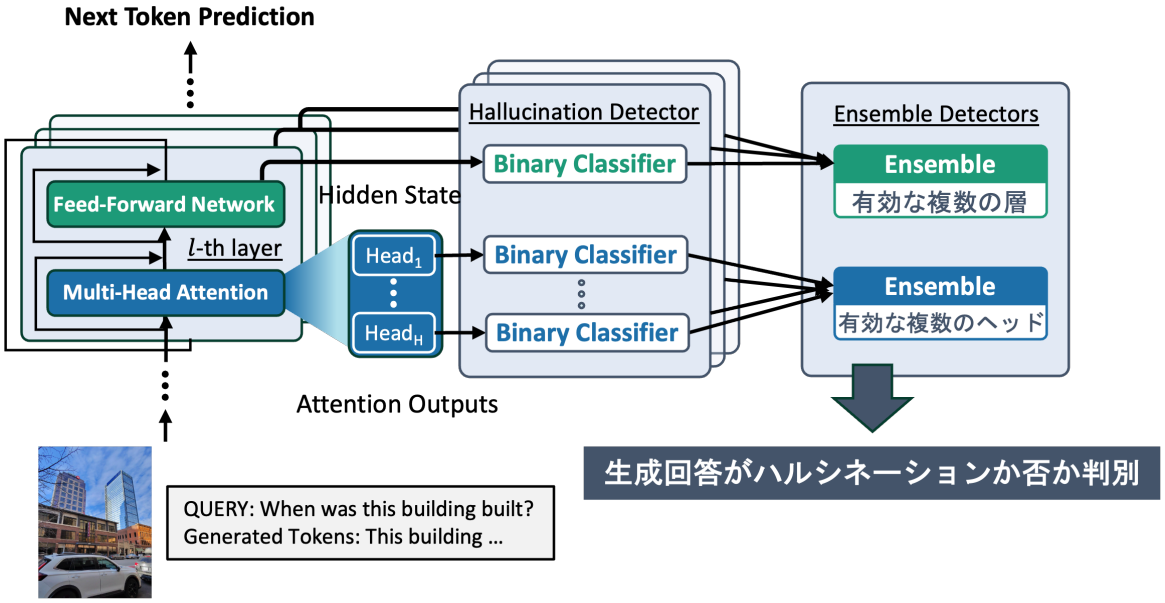

図 2　提案手法の概要図．画像に関する質問に対して回答を生成する際の VLM の内部状態を抽出して特徴量とし，生成回答がハルシネーションであるか否かをラベルとして，各層，各ヘッドごとにハルシネーション検出モデルを学習する．検証データによって，最も検出精度が高いアンサンブルの組み合わせを探索し，最終的な検出に使用する．

2　関連研究

本研究では，ハルシネーションの抑制ではなく，検出を対象とする．先行研究 [13] が指摘するように，ハルシネーションは言語モデルの学習および評価の枠組みに起因する不可避な現象である．また，正確な検出が可能になれば「分かりません」といった応答を可能にし，システム全体の信頼性を向上させることができる．そのため，本研究ではハルシネーションの検出自体に焦点を当てる．

言語モデルおよび VLM におけるハルシネーション検出の既存研究では，モデルの出力や内部状態に基づいた手法が提案されている．出力ベースの手法 [8] では，複数回生成した出力間の自己一貫性に基づいて生成回答がハルシネーションか否かを判定する．これらの手法は，モデルに依存しにくい一方で，推論時間が余分にかかり，リアルタイム検出に向かない．また，内部表現ベースの検出手法では，「言語モデルの内部状態には，現在生成している回答が正しいか否かに関する情報が含まれる [14]」という研究成果を根拠に，注意機構の出力や隠れ状態といった内部状態からハルシネーションの兆候を捉えようとしている．注意機構の出力や隠れ状態 [5, 9, 10]，画像埋め込み [11] を特徴量とした検出モデルの事前学習により，出力のみからは捉えにくい視覚と言語の対応関係や内部的な判断過程に関する情報を活用でき，リアルタイムな検出も可能である．しかし，既存研究の多くは単一の内部表現や単一検出器を使用しており，ハルシネーションの多様な性質を十分に捉えられていない可能性がある．本研究では，異なる複数の内部状態ごとに学習した検出モデルをアンサンブルすることで，より高精度で頑健なハルシネーション検出を目指す．

3　提案手法

3.1　内部表現の抽出

本提案手法では，VLM が VQA に回答する際の内部状態として，各マルチヘッド注意機構の出力および，隠れ状態を取り出す．画像と質問に対する回答トークン t を生成するごとに，それぞれの最終トークンのベクトル $\mathbf{e}_t$ を抽出し，最終的に T 個の全生成トークンにわたる平均 $\bar{\mathbf{e}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{e}_t$ を各内部状態の特徴量とする．隠れ状態 $\mathbf{e}^{HS}$ は，主成分分析で注意機構の出力と同じ次元まで圧縮している．

3.2　ハルシネーション検出モデルの学習

次に，各層，各ヘッドの注意機構の出力と各層の隠れ状態ごとにハルシネーション検出モデルを学習する．先に抽出した内部状態を特徴量，生成した回答がハルシネーションか否かをラベルとして，ロジ

スティック回帰による教師あり学習を行う．

$$y = \sigma(\mathbf{w}^\top \mathbf{e} + b)$$

ここで，$\sigma(\cdot)$ はシグモイド関数，$\mathbf{w}$ は重み，b はバイアス項である．学習データ数，モデルの軽量性と可読性を考慮し，ロジスティック回帰を採用した．

3.3 検出モデルのアンサンブル

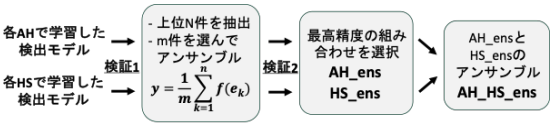

図 3　アンサンブルの概要図

図 3 にアンサンブルの概要図を示す．

まず，検証データを 2 つに分割する．検証データ 1 を用いて，注意機構の出力および隠れ状態のそれぞれについて，AUC が高い上位 N 個の検出モデルを選択する．次に，検証データ 2 を使って，選択したモデルのうち m 個による平均アンサンブルの性能を評価し，最適な組み合わせを決定する．さらに，隠れ状態と注意機構の出力のそれぞれで選んだ組み合わせを統合し，両者のアンサンブルも行う．

4 実験設定

本実験では，以下の 3 つの問いを検証する．

1. 複数の内部表現に基づいて学習した検出モデルをアンサンブルすることで，単一の検出モデルより検出精度は向上するか．
2. 検出モデルのアンサンブルは，複数の内部表現を単純に結合して学習する手法より検出精度が高いか．
3. 検出モデルのアンサンブルは，未知ドメイン（OOD）に対しても頑健か．

4.1 検出手法の定義

AH（Attention Head）は，各層・各ヘッドの注意機構の出力を用いた検出器を表し，HS（Hidden State）は，各層の隠れ状態に基づく検出器を表す．下付きの *one* は検証データで最も精度が高い単一の内部表現に基づく検出モデル，*ens* は複数の内部表現ごとに学習した検出モデルの平均アンサンブルを意味する．今回の実験では，使用モデルの層数を加味して，検証データ 1 を使って選択したモデル（$N_{AH} = 30, N_{HS} = 10,$）のうち，5 件（$m = 5$）のモデルをアンサンブルした．また，*concat* はアンサンブルに使用した組み合わせと同じ複数の内部表現を単純に連結した特徴量を用いて単一の検出モデルを学習する手法を表す．AH+HS_{ens} は，注意機構の出力と隠れ状態の両方に基づく検出モデルを統合したアンサンブル手法である．

4.2 使用モデル，データセット，評価指標

今回の実験では，VLM として，Llama-3.2-11B-Vision-Instruct，Qwen2.5-VL-7B-Instruct，および Pixtral-12B-2409 の 3 つを用いた．

また，データセットとして，CRAG-MM [12] を用い，ラベル分布を保つよう層化分割を行い，8:1:1 の比率で学習・検証・テストに分割した．CRAG-MM では，画像と画像に関する質問と答えが与えられており，回答が正解かどうかは，LLM-as-a-Judge で評価される．本実験では，LLM-as-a-Judge によって正解ではないと判定された生成回答をハルシネーションとし，評価を行った．

ハルシネーション検出精度は，生成解答がハルシネーションであるか否かを予測する 2 値分類タスクに対し，AUC を評価指標として用いた．

4.3 ID と OOD の評価方法

本研究では，CRAG-MM データセットで与えられているドメイン（自然，食べ物，物体/美術，車/景色）を使って，学習と評価に同一ドメインを用いる ID 設定と，異なるドメイン間で評価を行う OOD 設定における，検出モデルのアンサンブルの有効性の評価を行った．ID 設定では，学習/検証/テストを同一ドメインで行い，OOD 設定では，学習/検証/テストをそれぞれ異なるドメインで行った．

5 実験結果

5.1 アンサンブルの有効性

表 1 より，実験で使用した全ての VLM について，単一の検出モデルよりも，複数の検出モデルをアンサンブルしたモデルの方が検出精度が高く，隠れ状態と注意機構の出力で学習したモデルを組み合わせることでさらに検出精度が向上した．これは，各内部状態が捉えるハルシネーションの兆候が互いに異なり，それらを個別に学習した検出器を統合することで，相補的な情報を効果的に活用できたためであると考えられる．

また，同じ内部表現を使って，各々の内部状態を

表 1　VLM ごとのハルシネーション検出性能（AUC）

Model	AH_{one}	AH_{ens}	AH_{concat}	HS_{one}	HS_{ens}	HS_{concat}	$AH+HS_{ens}$
Llama-3.2-11B-Vision-Instruct	0.801	0.809	0.788	0.810	0.812	0.807	**0.825**
Qwen2.5-VL-7B-Instruct	0.806	0.842	0.796	0.846	0.854	0.847	**0.854**
Pixtral-12B-2409	0.718	0.784	0.715	0.778	0.798	0.793	**0.798**

表 2　ID / OOD 設定におけるハルシネーション検出精度（AUC）の検証 （ΔAUC = OOD − ID）

Model	Method	ID（AUC）	OOD（AUC）	ΔAUC
Llama-3.2-11B-Vision-Instruct	AH_{one}	0.637	0.570	-0.067
	AH_{ens}	0.638	0.622	-0.016
	HS_{one}	0.652	0.631	-0.021
	HS_{ens}	0.680	0.655	-0.025
	$AH+HS_{ens}$	**0.669**	**0.659**	-0.010
Qwen2.5-VL-7B-Instruct	AH_{one}	0.710	0.629	-0.081
	AH_{ens}	0.795	0.721	-0.074
	HS_{one}	0.685	0.698	+0.013
	HS_{ens}	0.748	0.757	+0.009
	$AH+HS_{ens}$	**0.804**	**0.743**	-0.061
Pixtral-12B-2409	AH_{one}	0.555	0.583	+0.028
	AH_{ens}	0.679	0.637	-0.042
	HS_{one}	0.633	0.629	-0.004
	HS_{ens}	0.682	0.650	-0.032
	$AH+HS_{ens}$	**0.695**	**0.659**	-0.036

特徴量として学習した検出モデルをアンサンブルした手法は，それぞれの特徴量を連結して単一の検出モデルを学習した手法よりも検出精度が高いという結果になった．この理由については今後の実験により明らかにしていきたい．

5.2　ID と OOD の結果

表 2 より，Llama-3.2-11B-Vision-Instruct と Qwen2.5-VL-7B-Instruct では，注意機構の出力単体に基づく検出は OOD への精度低下があるが，アンサンブルにより改善した．隠れ状態に基づく検出は，OOD への精度低下の傾向があまりなかった．Pixtral-12B-2409 では，同様の傾向が見られなかった．

これらの結果は，注意機構の出力から検出できるハルシネーションが特定のドメインに依存している可能性を示唆している．すなわち，単一ヘッドに基づく検出では，学習時に観測されたドメイン特有の挙動に過度に適合してしまい，未知ドメインへの汎化が困難になると考えられる．アンサンブルは，複数ヘッドが捉える異なる側面を統合することで，ドメイン依存性を緩和していると解釈できる．一方，隠れ状態に基づく検出モデルは，OOD 設定においても精度低下が比較的小さく，特定のドメインに依存しないハルシネーションの兆候を捉えている可能性が示唆される．

Appendix A の，各ドメインでアンサンブルに使用した内部状態の分布を見ると，注意機構はばらつきが大きいが，隠れ状態は特定の層に集中していることが分かる．この結果からも，注意機構と隠れ状態はハルシネーションの異なる性質を捉えており，注意機構の出力の方がよりドメインに特有の挙動を捉えているのではないかと考えられる．

6　おわりに

本研究では，VLM の内部表現に基づく複数の検出器をアンサンブルするハルシネーション検出手法を提案した．複数の VLM およびドメイン設定における実験の結果，提案手法は単一の検出器や，内部表現を単純に結合して学習する手法と比較して，一貫して高い検出性能を示すことを確認した．また，注意機構の出力に基づく検出手法は，OOD 設定における性能低下の傾向が見られたが，アンサンブルによって緩和された．一方で，本研究では，各内部状態が捉えるハルシネーションの具体的な性質は特定できていない．実験結果は，内部状態の種類や位置によって，検知可能な属性が異なる可能性も示唆しているため，今後は，各内部状態が捉えているハルシネーションの兆候をより詳細に分析する．

謝辞

本研究は，JST 経済安全保障重要技術育成プログラム JPMJKP24C3 の支援を受けたものです.

参考文献

[1] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-pro: A more robust multi-discipline multimodal understanding benchmark. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 15134–15186, 2025.

[2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In **Proceedings of the 37th International Conference on Neural Information Processing Systems**, 2023.

[3] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. **National Science Review**, Vol. 11, No. 12, p. nwae403, 11 2024.

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, 2024.

[5] Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In **Proceedings of the 13h International Conference on Learning Representations**, 2025.

[6] Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. Unified hallucination detection for multimodal large language models. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, 2024.

[7] Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In **Proceedings of the Computer Vision and Pattern Recognition Conference**, 2025.

[8] Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, 2023.

[9] Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, 2023.

[10] Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. Unsupervised real-time hallucination detection based on the internal states of large language models. In **Findings of the Association for Computational Linguistics: ACL 2024**, 2024.

[11] Anonymous. Detecting vision-language model hallucinations before generation. In **Submitted to ACL Rolling Review - July 2025**, 2025. under review.

[12] Jiaqi Wang, Xiao Yang, Kai Sun, Parth Suresh, Sanat Sharma, Adam Czyzewski, Derek Andersen, Surya Appini, Arkav Banerjee, Sajal Choudhary, et al. Crag-mm: Multi-modal multi-turn comprehensive rag benchmark. **arXiv preprint arXiv:2510.26160**, 2025.

[13] Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. Why language models hallucinate. **arXiv preprint arXiv:2509.04664**, 2025.

[14] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In **Proceedings of the 37th Conference on Neural Information Processing Systems**, 2023.

A　各ドメインごとのアンサンブル構成

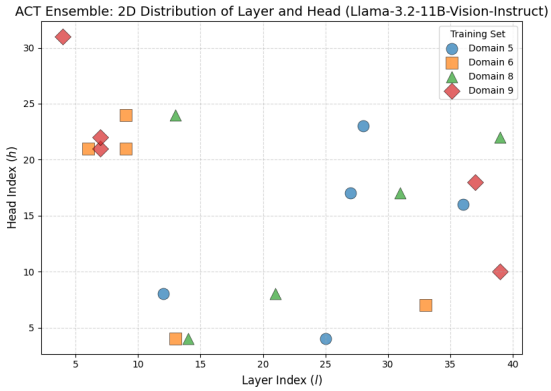

図 4　各ドメインごとにアンサンブルに使用した AH の分布（Llama-3.2-11B-Vision-Instruct）

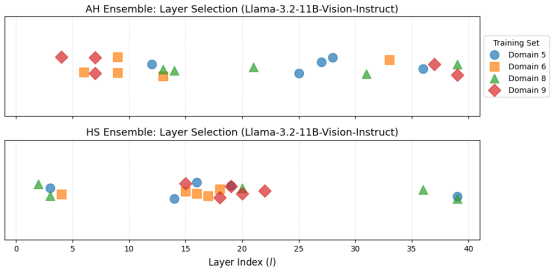

図 5　各ドメインごとにアンサンブルに使用した AH と HS の層の分布（Llama-3.2-11B-Vision-Instruct）

Llama-3.2-11B-Vision-Instruct において，各検出器をアンサンブルする際に使用した内部状態の分布が図 4 と図 5 である．図 4 の AH の分布を見ると，それぞれのドメインで同じヘッドの出力は使用しておらず，ドメインごとに異なる注意機構の出力を使ってハルシネーションを検出している．図 5 の AH と HS の層ごとの分布を見ると，隠れ状態に基づくハルシネーション検出では，より特定の層を使っていることが分かる．

これらの結果から，注意機構と隠れ状態は異なるハルシネーションの兆候を捉えており，注意機構の出力の方がよりドメインに特有の挙動を捉えているのではないかと考えられる．