[2026 winter seminar]

Object hallucinations in VLM

Pusan National University
Computer Vision & Signal Processing Lab, Dept. of Electronic Engineering

Presented By
김소현

CVSP Lab
Computer Vision & Signal Processing

Index

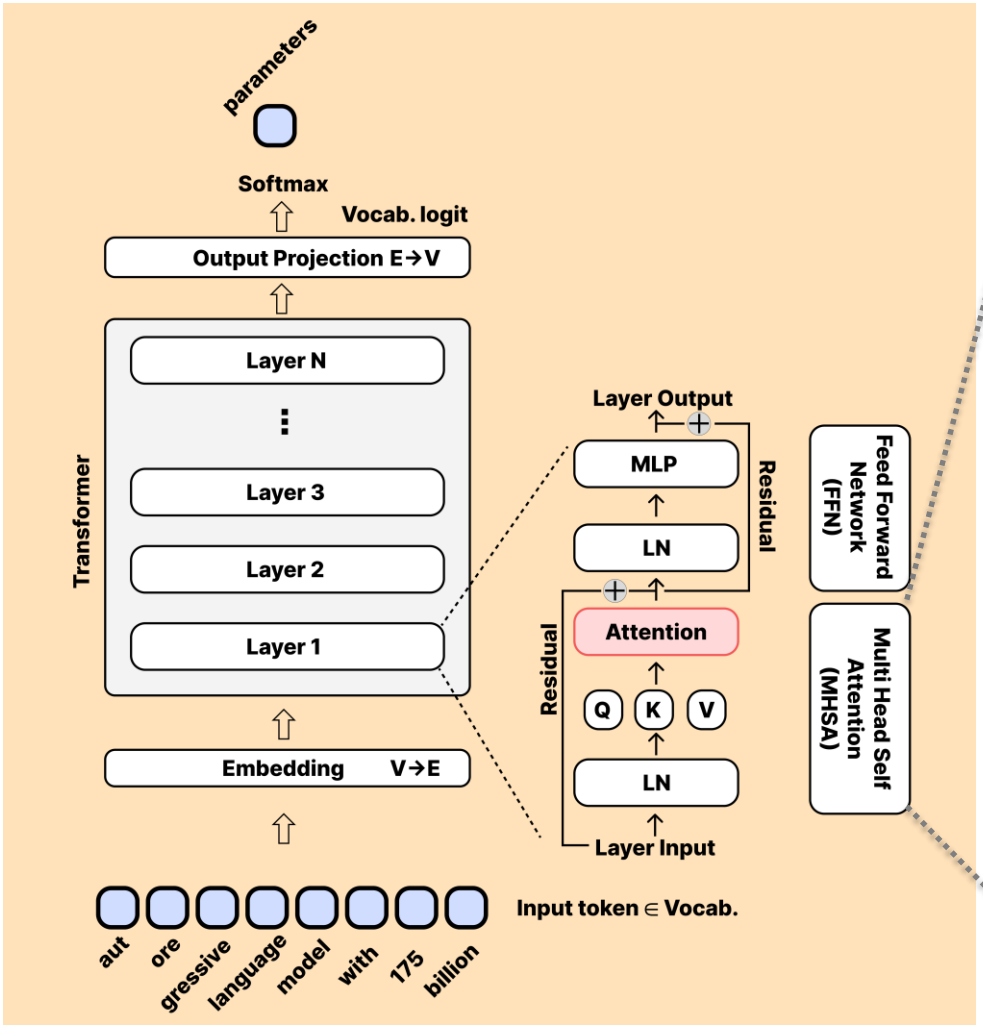
1) Jiang, Zhangqi, et al. "Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.
2) Seo, Hoigi, et al. "On Epistemic Uncertainty of Visual Tokens for Object Hallucinations in Large Vision-Language Models." arXiv preprint arXiv:2510.09008 (2025).

- Background
- 논문 소개
  - Devils in Middle Layers of Large Vision-Language Models: Interpreting, Detecting and Mitigating Object Hallucinations via Attention Lens [CVPR 2025] 1)
  - On Epistemic Uncertainty of Visual Tokens for Object Hallucinations in Large Vision-Language Models [NeurIPS 2025] 2)
- Conclusion

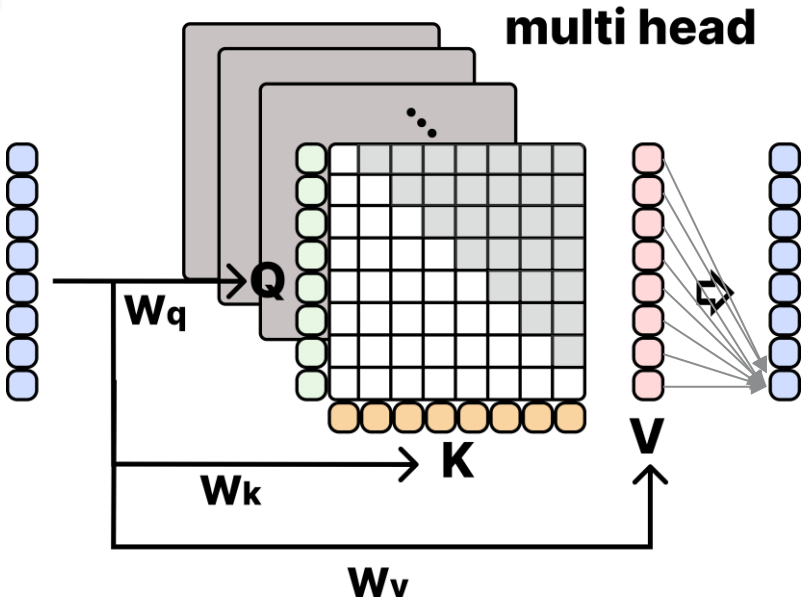
CVSP Lab
Computer Vision & Signal Processing

Background

- LLM 및 Attention



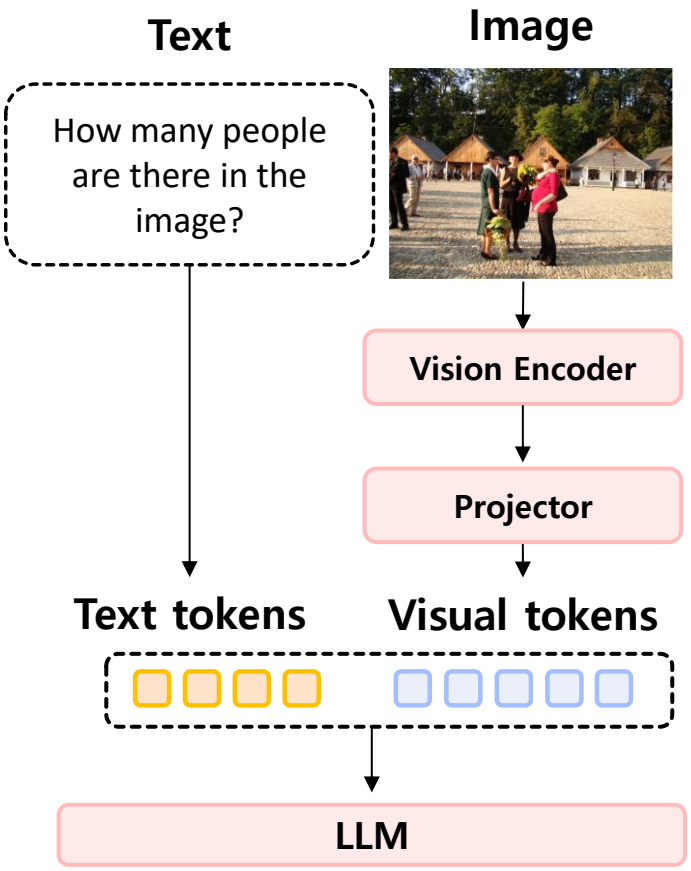
$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



CVSP Lab
Computer Vision & Signal Processing

Background

- VLM 구조



Background

- Object hallucination
  - VLM이 이미지에 실제로 있지 않은 객체를 있다고 말하거나 존재하는 객체를 잘못 설명하거나 왜곡하는 현상
  - 분석된 핵심 이유
    - 언어 사전 확률 과도
      - 학습 중 자주 등장한 사전 지식에 의존
    - 시각 언어 정렬 불완전
      - 비전 토큰이 충분한 객체 증거를 제공하지 못하는 경우, 언어 쪽이 주도권을 가져 추측으로 채움
    - 비전 토큰 품질 문제
      - 중요 토큰이 약화되거나 배경, 중복 토큰이 과도해지는 경우



A few cars can be seen parked or driving around the square, and a bicycle is also present in the scene. Additionally, there are two other people in the background, possibly observing the man lying on the ground or going about their own activities.

1) Jiang, Zhangqi, et al. "Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.

Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens[1)]

CVPR 2025

Introduction

- Motivation
  - 시각 처리 과정의 중요성 고려
    - 기존 object hallucination을 완화하는 방법들은 환각 생성의 원인을 언어 편향에서 찾으며 시각 처리 과정의 중요성을 간과
    - LVLM이 추론과정에서 이미지 토큰으로부터 시각 정보를 처리하는 방식과 이것이 객체 환각 생성에 어떤 영향을 미치는지 분석

CVSP Lab
Computer Vision & Signal Processing

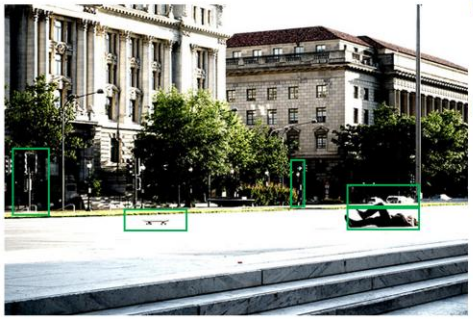
Introduction

- Findings
  - LVLM의 추론 과정에서 시각 정보 처리에 핵심적인 단계가 중간 레이어임을 파악
  - 중간 레이어의 시각 정보 처리 과정을 두가지 핵심 단계로 구분
    - 시각 정보 확장 단계: 시각 정보가 객체 토큰으로 전달
    - 의미 정제 단계: 인코딩 된 시각 정보를 텍스트를 통해 해석
      - 실제로 객체 토큰을 형성하는데 기여하는 단계는 어디인지 파악
  - 환각 토큰과 실제 객체 토큰의 중간 단계에서의 어텐션 패턴 비교 분석하여 환각 탐지
    - 시각 정보 확장 단계의 어텐션 패턴에서 어떤 차이가 환각을 일으키는가

Analysis

- Experimental Setup for Case study
  - Dataset
    - COCO 2014 validation set에서 무작위로 선택한 2000개의 이미지에 대해 수행
  - Models
    - LLaVA-1.5-7B(main)/Shikra/MiniGPT-4
  - 토큰 분류
    - 생성된 토큰들을 정답 annotation을 기준으로 실제 이미지에 존재하는 '실제 객체 토큰'과 존재하지 않는 '환각 객체 토큰'으로 구분
    - 총 1842개의 환각 토큰과 4397개의 실제 토큰 확보

COCO class
person
bicycle
car
motorcycle
spoon
…
chair



GT: traffic light / car / person / skateboard

User

Please help me describe the image in detail.

Model

The image features a man lying on the ground in a city square, possibly taking a break or resting. He is positioned in the center of the scene, with a skateboard nearby, suggesting that he might have been skateboarding before taking a break. The city square is surrounded by tall buildings, giving the impression of an urban environment. There are multiple traffic lights visible in the scene, indicating that the area is well-regulated for vehicle and pedestrian traffic. A few cars can be seen parked or driving around the square, and a bicycle is also present in the scene. Additionally, there are two other people in the background, possibly observing the man lying on the ground or going about their own activities.

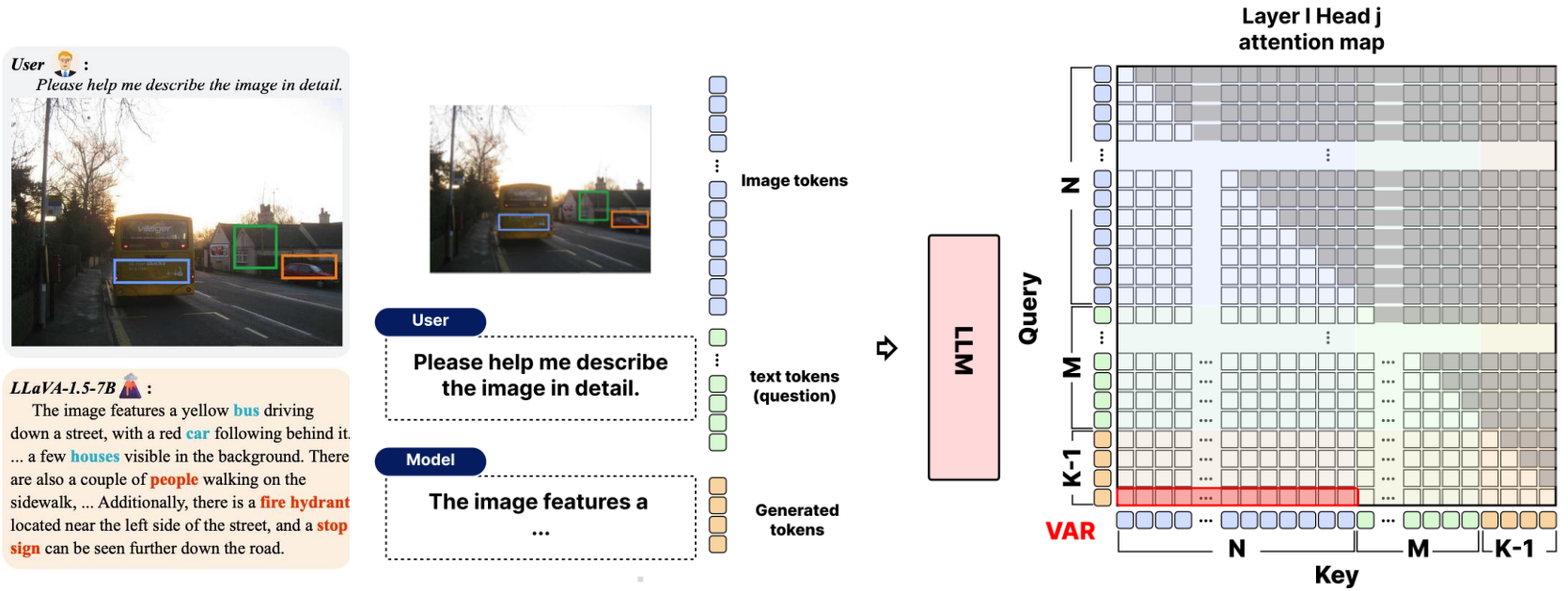
CVSP Lab
Computer Vision & Signal Processing

Analysis

- Preliminary
  - Vision Attention Ratio(VAR)
    - K번째 토큰 y_k에 대해 레이어 l 헤드 h에서의 시각 어텐션 비율 VAR

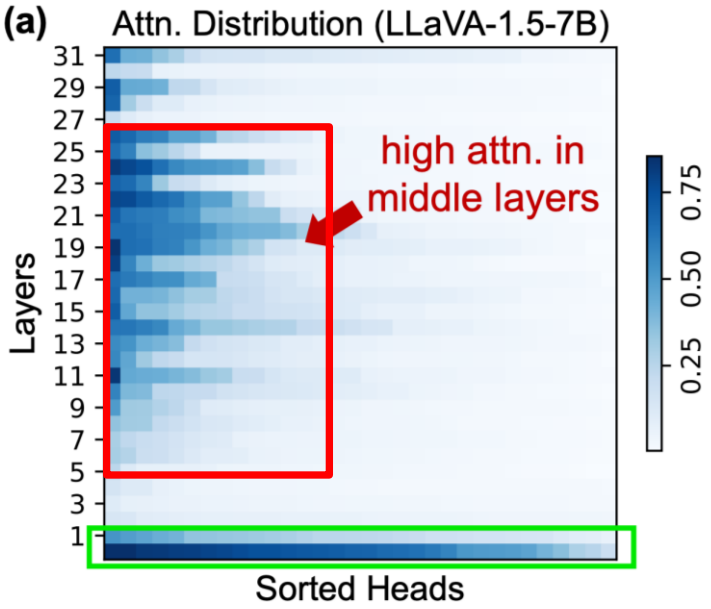
$$VAR^{(l,h)}(y_k) \triangleq \sum_{i=1}^{n} A_k^{(l,h)}(a_k, i)\ (a_k = N + M + (k-1))$$

새로 생성된 토큰 y_k가 이미지 토큰들에 할당한 어텐션 가중치로 토큰 생성과정에서 시각 정보와 얼마나 강하게 상호작용하는지를 나타내는 지표



Analysis

- Experimental Setup for Case study
  - Finding1
    - 실제 토큰이 생성될 때 이미지 토큰들을 참조하는 정도인 VAR을 분석
      - 모든 레이어와 헤드에 대해 계산하고 모든 실제 토큰들에 대한 값을 평균
      - 모델에 관계없이 일관되게 VAR이 높은 값을 가지는 레이어는 중간 레이어(5~26레이어)
    - 시각 정보와의 상호작용이 주로 중간 레이어에서 일어남

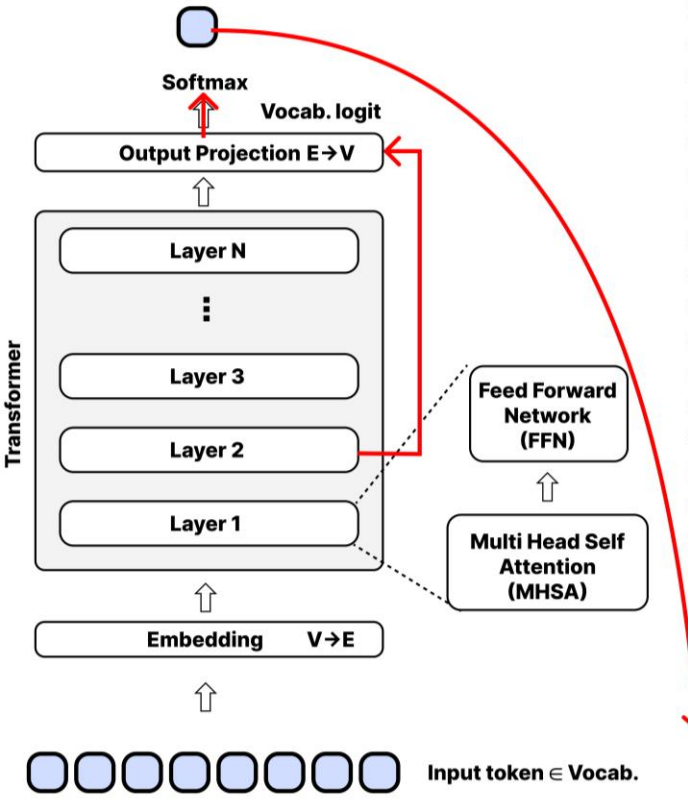


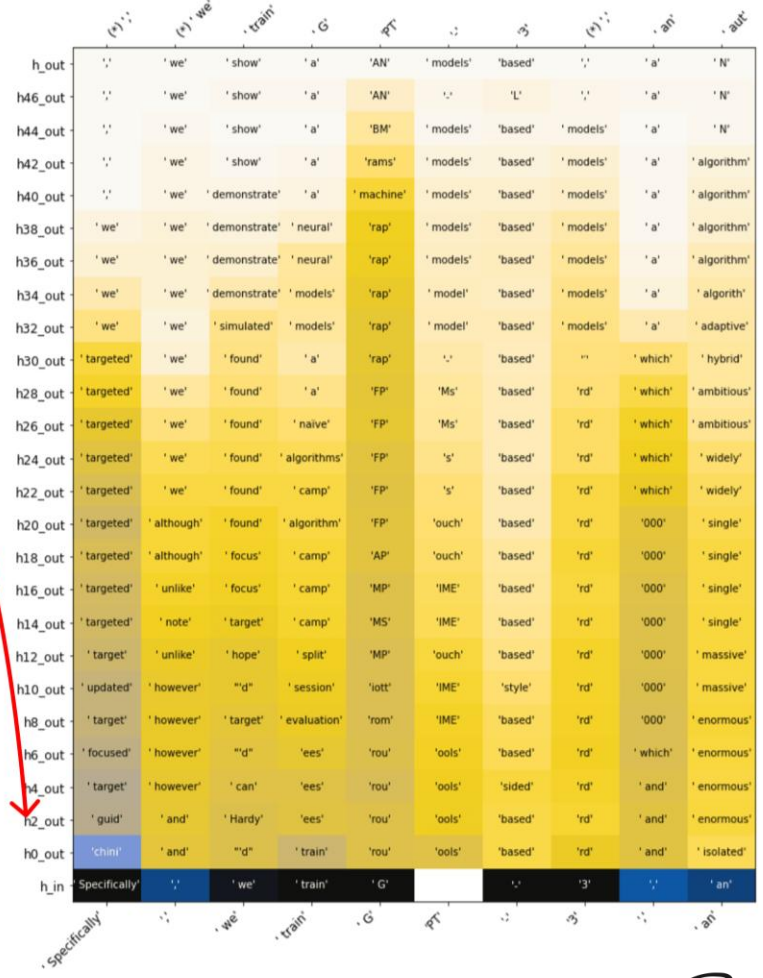
<Distribution of VAR for real object tokens across heads and layers in LLaVA-1.5-7B>

CVSP Lab
Computer Vision & Signal Processing

Analysis

- Preliminary
  - Logit Lens

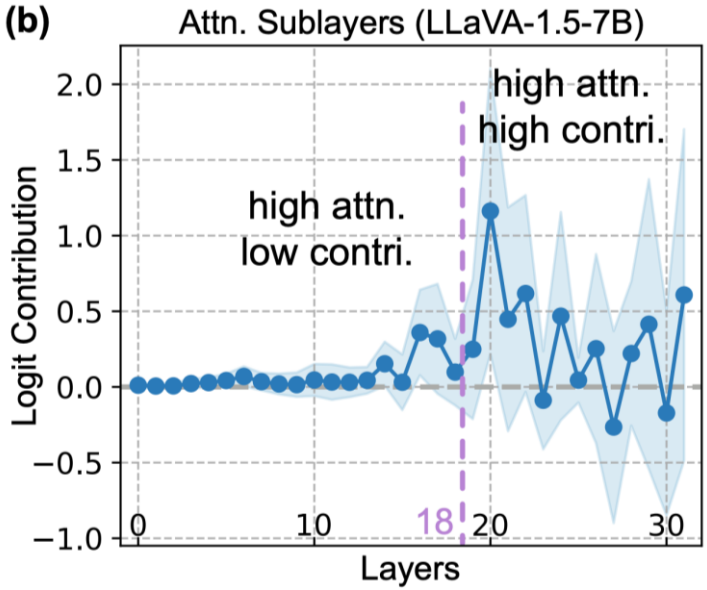




CVSP Lab
Computer Vision & Signal Processing

Analysis

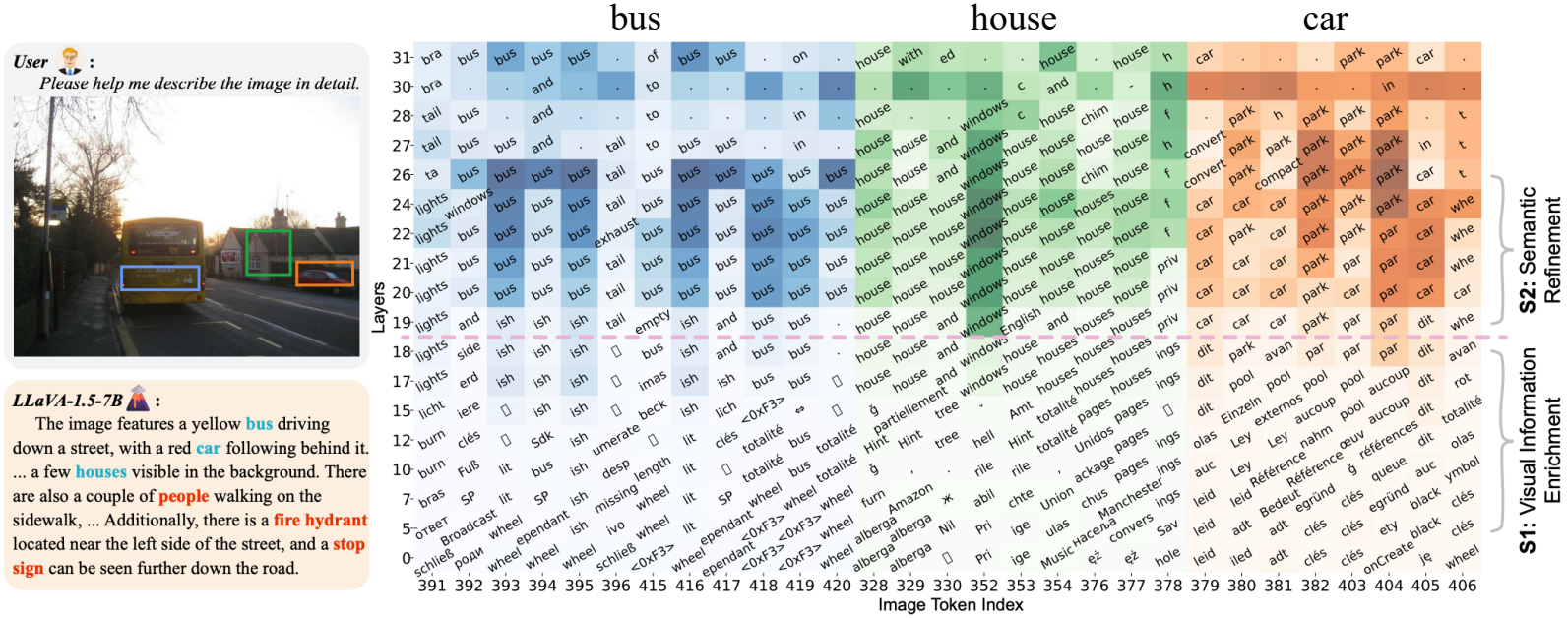
- Experimental Setup for Case study
  - Finding1
    - 중간 레이어를 실제 객체 토큰 예측에 기여하는 정도에 따라 두 가지 핵심 단계로 구분
      - 실제 토큰 예측에 대한 1번째 MHSA 레이어의 기여도: $P_O(V|a_r^l)$
      - 5~18번째 레이어는 기여도가 낮고 19~26 레이어는 기여도가 높게 나타남



<The logit contribution of attention sublayers to real token prediction>

Analysis

- Experimental Setup for Case study
  - Finding1
    - 중간 레이어를 실제 객체 토큰 예측에 기여하는 정도에 따라 두가지 핵심 단계로 구분
      - Logit lens를 사용하여 각 레이어 l에서 이미지 토큰 v_i^l에 대해
        $\arg max_{1 \le j \le |V|}\{p_j(V|v_i^l)\}$ 에 해당하는 텍스트 추출

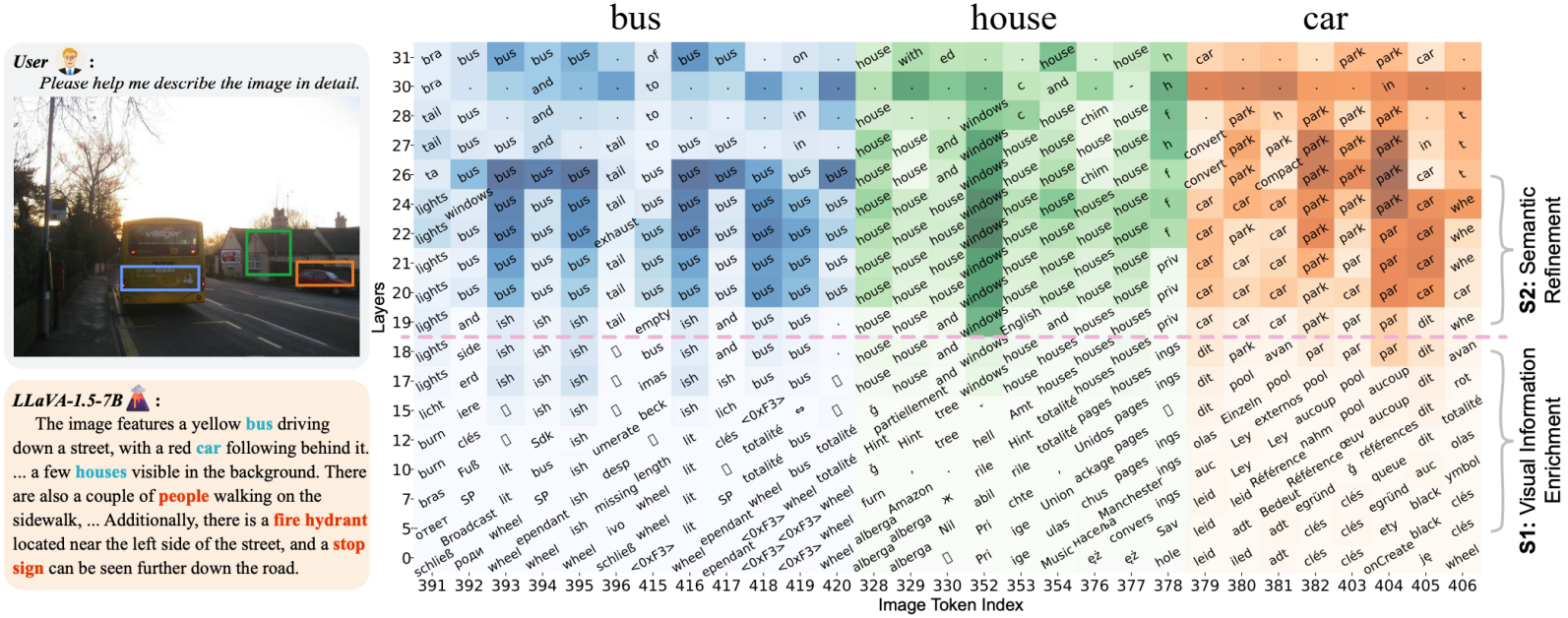


< Case study of image hidden state interpretation in LLaVA-1.5-7B via logit lens >

CVSP Lab
Computer Vision & Signal Processing

Analysis

- Experimental Setup for Case study
  - Finding1
    - 중간 레이어를 실제 객체 토큰 예측에 기여하는 정도에 따라 두가지 핵심 단계로 구분
      - 5~18번째 레이어: 추출된 텍스트 토큰이 이미지 패치와의 연관성이 낮음
        - ✓**시각 정보 확장 단계**: 어텐션을 통해 시각 정보를 누적하는 단계이므로 VAR이 높게 나타남
      - 19~26번째 레이어: 추출된 텍스트 토큰이 이미지 패치와 의미적으로 일관됨
        - ✓**의미 정제 단계**: 이미지 토큰에 인코딩된 시각 정보를 해석하고 활용할 수 있는 단계

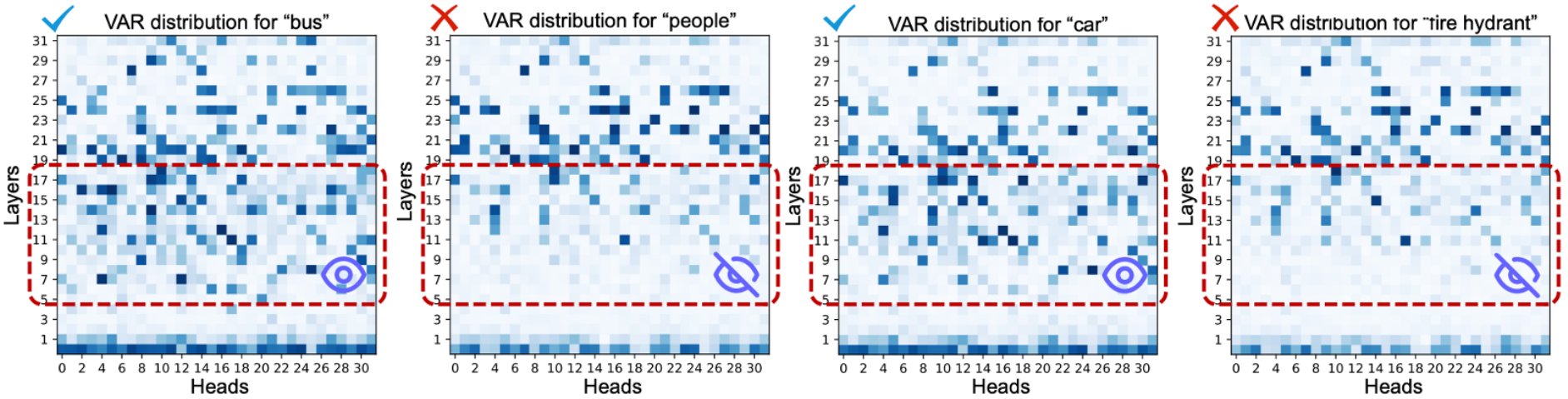


< Case study of image hidden state interpretation in LLaVA-1.5-7B via logit lens >

Analysis

- Experimental Setup for Case study
  - Finding2
    - 중간 레이어에서 활성화되지 않는 시각 어텐션은 환각을 시사
      - 실제 객체 토큰에 대한 VAR과 환각 객체 토큰에 대한 VAR 비교
      - 환각 객체 토큰의 경우 시각 정보 확장 단계인 5~18레이어에서 실제 객체 토큰에 비해 비활성화된 어텐션 패턴 보임

Real object | Hallucinated object | Real object | Hallucinated object



< Case study of the visual attention ratio distribution over heads and layers in LLaVA-1.5-7B. >

CVSP Lab
Computer Vision & Signal Processing

Analysis

- Experimental Setup for Case study
  - Finding3
    - 시각 정보 확장 단계에서 다중 객체와 상호작용하는 어텐션 헤드는 환각을 유발함
      - 실제 객체 토큰의 경우 헤드별 어텐션 분포가 해당 객체의 공간적 범위와 어느정도 일관되게 정렬
      - 환각 객체 토큰의 경우 이 단계에서 헤드들이 이미지 내 일관되지 않은 객체들과 상호작용함
    - 이러한 불안정한 헤드 동작이 여러 객체로부터 혼합된 정보를 객체 토큰에 인코딩하게 만들며 환각으로 이어질 가능성 높다고 분석



< Visualization of attention maps over image for 'bus' (real) and 'people' (hallucinated) object tokens in heads (ℓ, h) >

CVSP Lab
Computer Vision & Signal Processing

Method

- Heads guided attention intervention
  - 추론 시점에서 중간 레이어에 위치한 각 헤드에 대해 이미지 토큰에 할당되는 어텐션 가중치 $\{A_k^{(l,h)}(a_k, i)\}_{i=1}^{n}$ 를 조정함
    - 추론 시점 k에서 softmax연산 이전의 어텐션 점수 행렬 추출

$$S_k^{(l,h)} = \frac{Q_{l,h}K_{l,h}^{\intercal}}{\sqrt{d_k}} (Q_{l,h}, K_{l,h} \in \mathbb{R}^{a_k \times d_k})$$

    - 시각 정보 확장 단계의 레이어에서 원래의 어텐션 점수에 양의 값을 더하여 시각 정보 간 상호작용 강화(finding1+2)
    - 이때, 서로 다른 헤드들이 일관되게 상호작용하는 영역만 큰 증폭을 받도록 헤드별 절댓값 스코어의 평균을 더해줌(finding3)

$$S_k^{(l,h)}(a_k, i) = S_k^{(l,h)}(a_k, i) + \alpha \frac{1}{H} \sum_{h=1}^{H} |S_k^{(l,h)}(a_k, i)|$$

      - α는 개입 강도를 조절하는 균형 계수

CVSP Lab
Computer Vision & Signal Processing

Experiments

- Quantitative Results
  - CHAIR-환각 평가 벤치마크의 성능과 F1 score
    - 모델: LLaVA-1.5-7B,13B/ Shikra-7B/ MiniGPT-4-7B
    - 기존 환각 완화 기법들과 비교 결과 제시

Method	LLaVA-1.5-7B			LLaVA-1.5-13B			Shikra-7B			MiniGPT-4-7B			Avg.	
	$C_S \downarrow$	$C_I \downarrow$	F1↑	$C_S \downarrow$	$C_I \downarrow$	F1↑	$C_S \downarrow$	$C_I \downarrow$	F1↑	$C_S \downarrow$	$C_I \downarrow$	F1↑	$C_S \downarrow$	$C_I \downarrow$
Decoding Strategy														
Greedy	53.0	15.6	76.7	49.8	14.6	78.2	57.6	15.7	75.3	31.8	12.0	71.1	48.1	14.5
Beam	55.6	15.4	77.5	50.4	13.8	79.0	59.0	16.3	74.6	29.2	9.9	71.2	48.6	13.9
OPERA	45.6	13.1	79.1	42.6	13.2	77.8	41.4	13.7	73.5	25.4	9.6	71.2	38.8	12.4
Contrastive Decoding														
VCD†	58.6	18.2	72.8	53.6	15.3	75.8	56.4	15.5	75.2	41.4	14.1	68.2	52.5	15.8
PAI†	**24.2**	7.1	75.2	33.0	9.2	77.8	38.6	10.1	76.2	23.2	8.2	71.4	29.8	8.7
Ours†	25.0	**6.7**	76.1	**25.8**	**8.8**	77.3	**23.8**	**9.4**	72.7	**21.4**	**8.0**	70.8	**24.0**	**8.2**
Δ%	+3.3%	-5.6%		-21.8%	-4.3%		-38.3%	-6.9%		-7.8%	-2.4%		-19.5%	-5.7%

< CHAIR hallucination evaluation and F1 results on three LVLMs with max new token set to 512 >

CVSP Lab
Computer Vision & Signal Processing

2) Seo, Hoigi, et al. "On Epistemic Uncertainty of Visual Tokens for Object Hallucinations in Large Vision-Language Models." arXiv preprint arXiv:2510.09008 (2025).

On Epistemic Uncertainty of Visual Tokens for Object Hallucinations in Large Vision-Language Models[2)]

NeurIPS 2025

Introduction

- Motivation
  - 환각 완화를 위한 기존 연구들
    - LLM의 디코딩 전략을 수정하는 방식/ 어텐션 메커니즘을 조절하는 방식 등
      - LVLM에서 가장 계산 비용이 큰 구성 요소인 LLM을 여러 번 추론해야 함
  - 시각적 인지를 담당하는 비전 인코더를 대상으로 하는 객체 환각 완화 기법은 충분히 연구되지 않음
    - 비전 인코더로부터 생성되는 시각 토큰의 불확실성을 분석
      - 불확실성은 기존에 Monte Carlo(MC) dropout과 같이 수천 번의 순전파를 요구하는 계산 집약적 방법에 의존했음
      - 이론적 분석을 통해 시각 토큰의 불확실성을 추정할 수 있는 효율적 대안 제시

=> 비전 인코더만을 수정하여 객체 환각을 완화하는 효과적 전략 제시

CVSP Lab
Computer Vision & Signal Processing

Preliminary

- Adversarial attack on LVLMs
  - 이미지에 인간이 인지하기 어려운 미세한 섭동을 추가하여 모델로 하여금 잘못된 예측을 하도록 유도하는 기법
    - 초기 연구들은 분류나 객체 검출과 같은 과제에 이를 활용하였으나 최근에는 모델의 강건성 향상을 위해 LVLM을 대상으로 한 의도적 공격까지 확장
  - 미세하게 섭동된 이미지를 생성하기 위한 방법으로 PGD가 가장 널리 활용됨
    - Projected Gradient Descent (PGD)의 최적화 과정

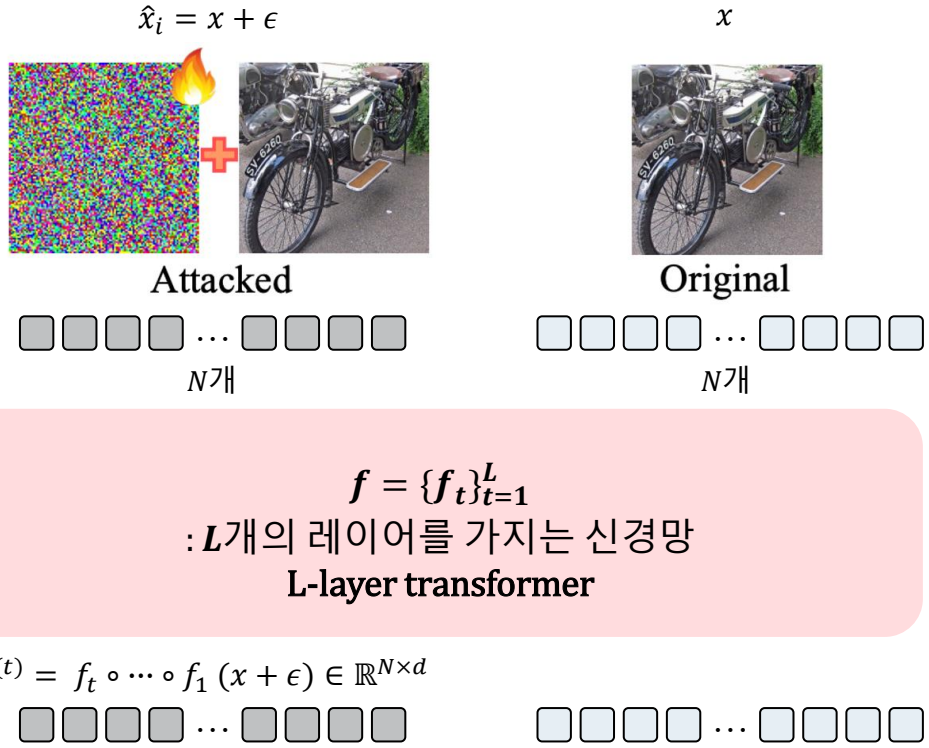
$$\hat{x}_{i+1} = \Pi\left(\hat{x}_i + \alpha \cdot sign\left(\nabla_{\hat{x}_i}\mathcal{L}\left(F(\hat{x}_i), F(x)\right)\right)\right)$$

      - α: learning rate/ F: 대상 신경망 / x: 원본 이미지/ $\hat{x}_i$: i번째 반복 섭동된 이미지 /
      - Π: $\|\hat{x}_{i+1} - x\|_\infty \leq k$를 만족하는 제약집합으로의 투영 연산
    - 섭동된 이미지와 원래 이미지에 대한 표현 차이가 미세하도록 섭동 이미지를 최적화하는 과정

CVSP Lab
Computer Vision & Signal Processing

Method

- 효율적인 시각 토큰 불확실성 분석을 위한 이론 제시
  - Lemma1
    - $f \in C^2$이고 ϵ가 충분히 작으면 $Z^{(t)}$는 $z^{(t)}$를 중심으로 하는 가우시안 분포로 국소적으로 근사될 수 있음



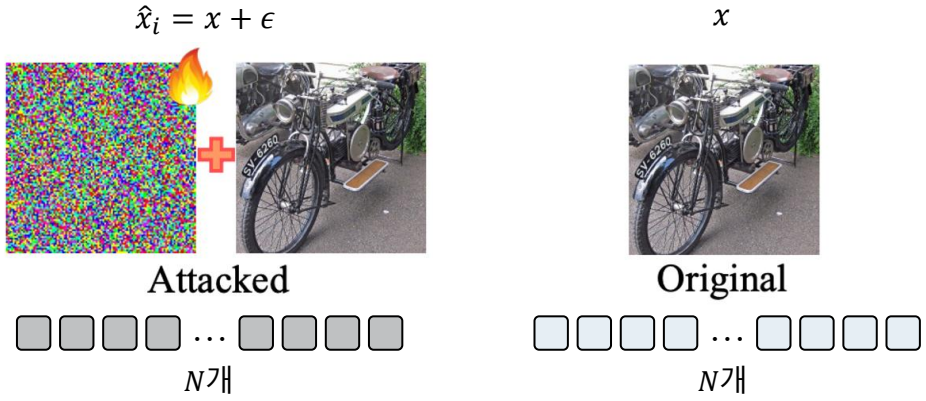
CVSP Lab
Computer Vision & Signal Processing

Method

- 효율적인 시각 토큰 불확실성 분석을 위한 이론 제시
  - Lemma2

$-Z_i^{(t)}$ 가 작은 ϵ에 대해 부드럽게 변화한다면 $Z_i^{(t)}$의 미분 엔트로피 상한은

$\mathbb{E}_\epsilon\left[\left\|Z_i^{(t)} - z_i^{(t)}\right\|_2^2\right]$ 가 증가함에 따라 증가



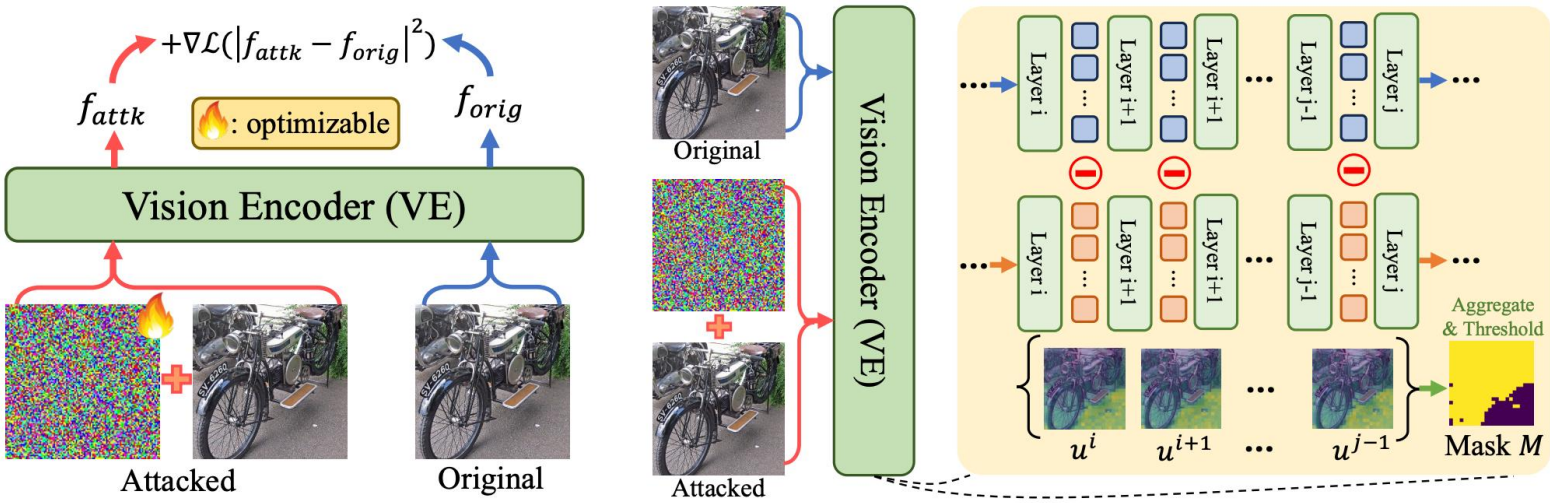
CVSP Lab
Computer Vision & Signal Processing

Method

- 효율적인 시각 토큰 불확실성 분석을 위한 이론 제시
  - Lemma1+Lemma2
    - 충분히 작은 섭동 ϵ가 가해진 이미지를 만들 수 있다면 t레이어에서의 토큰의 불확실성은 레이어의 두 이미지로부터 형성된 은닉상태의 편차 norm으로 근사 가능
      - MC dropout가 같이 많은 순전파 연산을 거치지 않고도 불확실한 시각 토큰을 판별할 수 있음
      - 원본 이미지에 대한 레이어의 은닉상태를 기준으로 이미지에 적대적 섭동을 가했을 때 은닉상태의 변화가 큰 시각 토큰을 불확실한 토큰으로 판별

Method

- Proposed method
  - Step1: Adversarial attack process
    - PGD 기반 적대적 섭동 하에서 표현 편차가 크게 나타나는 시각 토큰을 불확실한 토큰으로 식별
  - Step2: Uncertainty mask generation process
    - 비전 인코더 중간 블록에 위치한 어텐션 레이어에서 불확실한 시각 토큰을 마스킹하여 영향력 억제



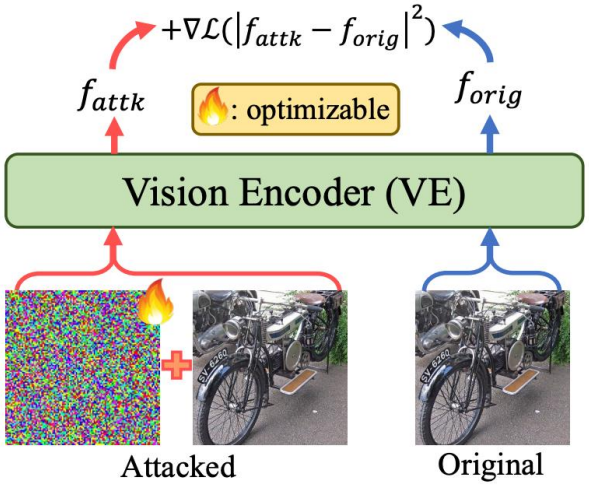
(a) Adversarial attack process

(b) Uncertainty mask generation process

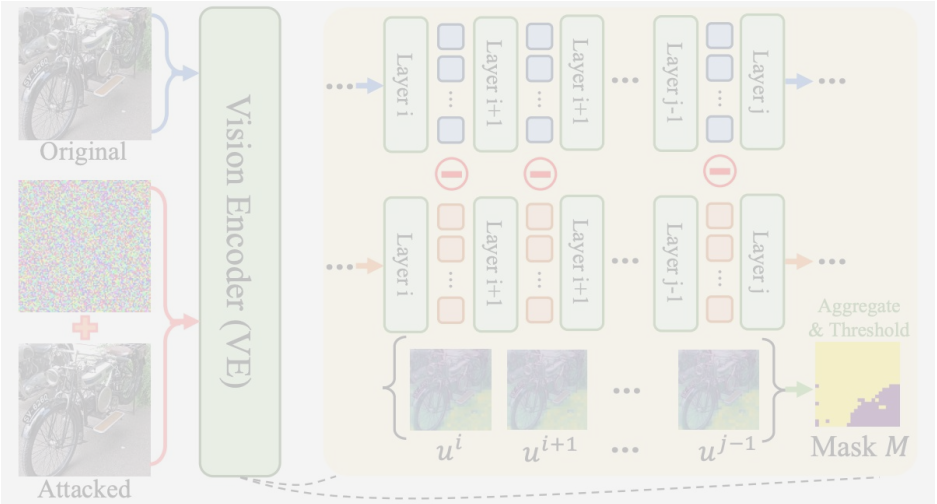
Method

- Proposed method
  - Step1: Adversarial attack process

    - $f_{orig.}$과 $f_{attk.}$사이의 평균제곱오차를 적대적 목적 함수로 정의하고 PGD 최적화 식 $\hat{x}_{i+1} = \Pi\left(\hat{x}_i + \alpha \cdot sign\left(\nabla_{\hat{x}_i}\mathcal{L}\left(F(\hat{x}_i), F(x)\right)\right)\right)$를 활용하여 I번 동안 ϵ를 최적화하여 최종 공격 이미지를 얻음



(a) Adversarial attack process



(b) Uncertainty mask generation process

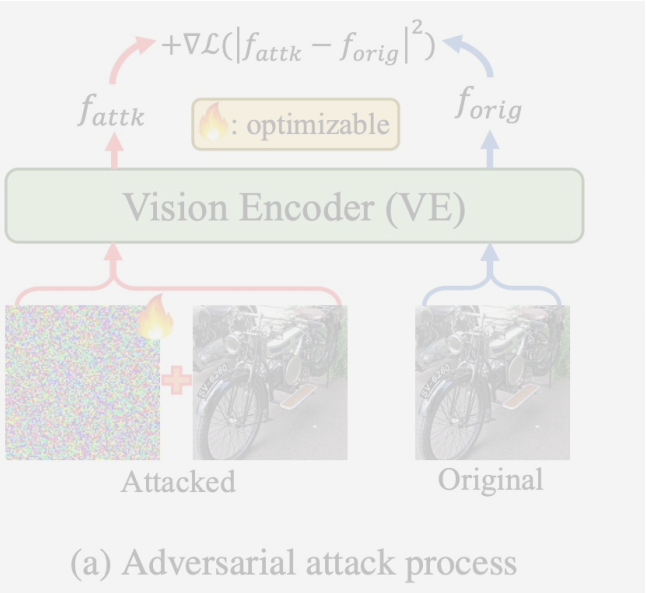
Method

- Proposed method
  - Step2: Uncertainty mask generation process
    - 연속적인 레이어 집합 $S = \{i, \dots, j-1\}$에 대해 원본 이미지와 섭동된 이미지 각각에 대해 비전 인코더 F_V 의 각 레이어 은닉 상태 추출

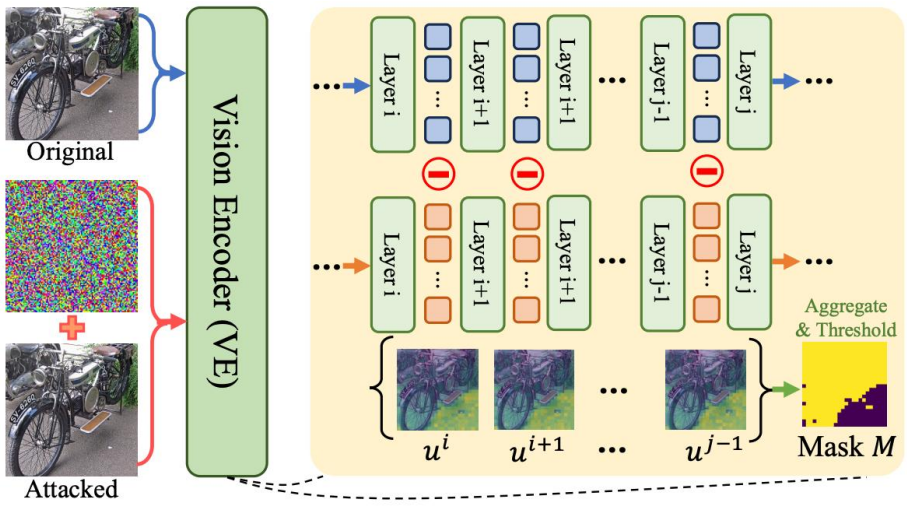
$$F_{orig} = \left\{ f^i_{orig.}, \dots, f^{j-1}_{orig.} \right\}, \quad F_{attk} = \left\{ f^i_{attk.}, \dots, f^{j-1}_{attk.} \right\}$$

    - 각 레이어 $l \in S$에 대해 대응되는 은닉상태 간 편차 norm을 정의

$$u^l = \left\| f^l_{attk.} - f^l_{orig.} \right\|_2$$



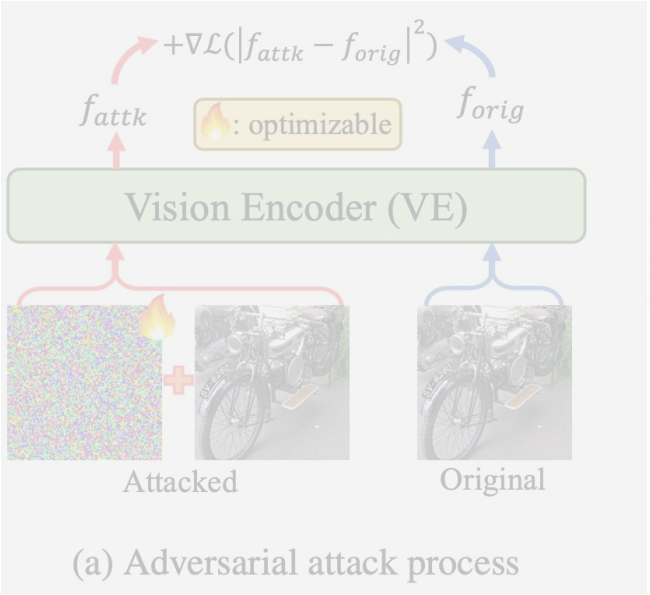
(a) Adversarial attack process

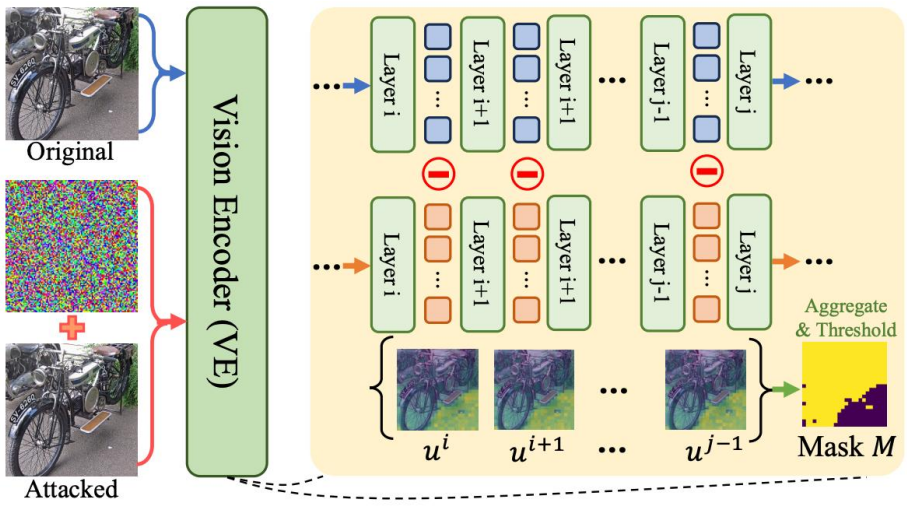


(b) Uncertainty mask generation process

Method

- Proposed method
  - Step2: Uncertainty mask generation process
    - 레이어별 불확실성 맵 집합을 얻음
      - $\mathcal{U} = \{u_i, \ldots, u_{j-1} | \forall l \in S\}$
    - 각 u_l에 대해 min-max 정규화하고 레이어 수에 대해 평균하여 불확실성 맵 U 얻음
      - $\mathrm{U} = \frac{1}{j-i} \sum_{l=i}^{j-1} \frac{u_l - u_l^{min}}{u_l^{max} - u_l^{min}}$

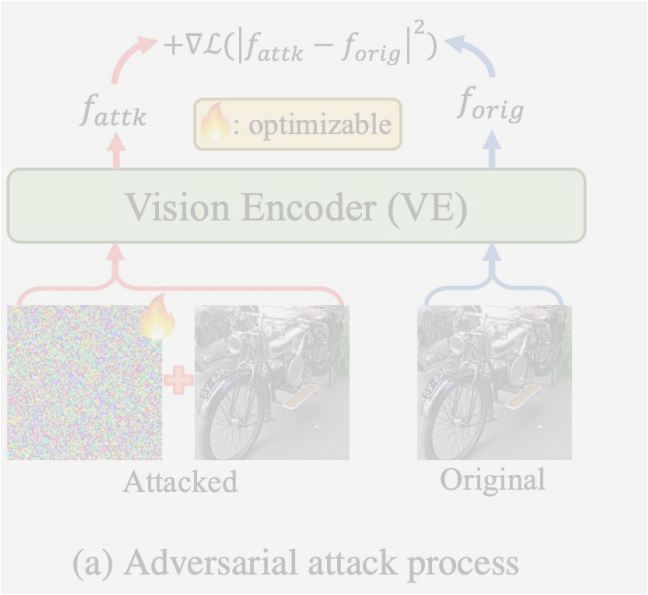


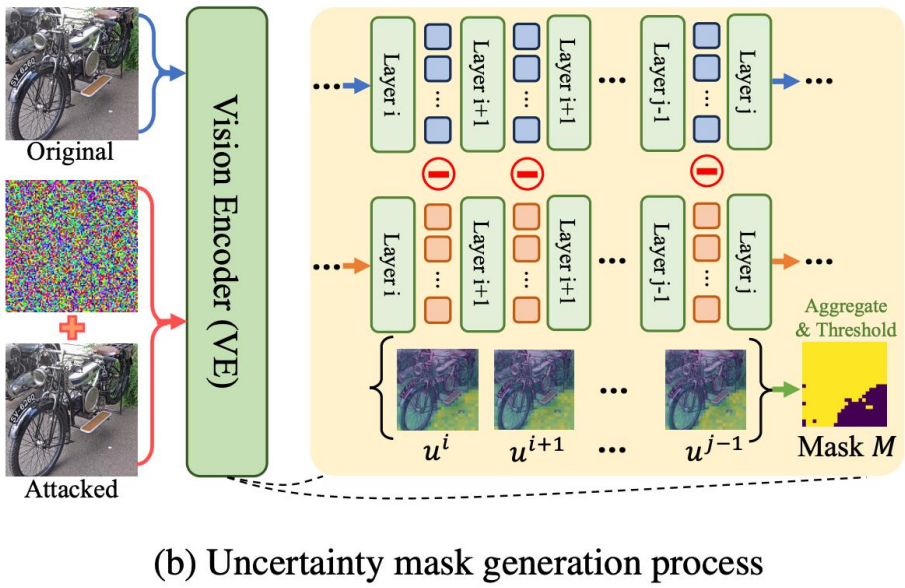


(b) Uncertainty mask generation process

Method

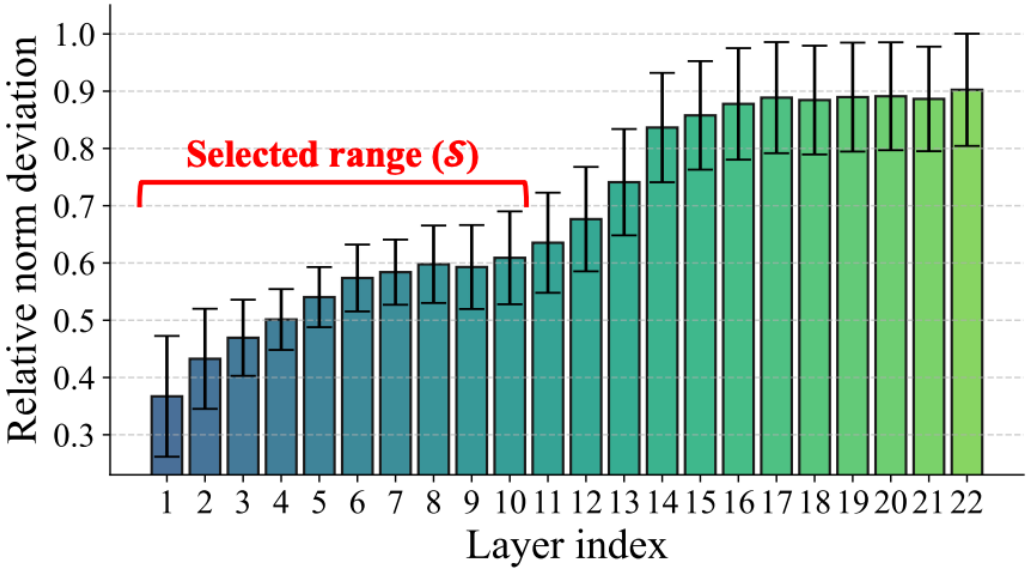
- Proposed method
  - Step2: Uncertainty mask generation process
    - 최종 불확실성 마스크 M 정의
      - U의 평균 μ_U와 표준편차 σ_U 를 사용해 정규화한 뒤 임계값 파라미터 σ_{th}를 활용하여 이진화
      - $M = 1 - \frac{1}{2}[sign\left(\frac{U-\mu_U}{\sigma_U} - \sigma_{th}\right) + 1]$





Method

- Proposed method
  - Mask를 생성할 때 활용할 레이어 설정
    - 불확실성 마스크 M 정의 시 초기 레이어들로 특징 추출
      - 제안 방법은 Lemma1~2에서 적대적 공격으로 유도된 시각 특징의 편차 norm이 작다는 가정에 의존하므로 가해진 섭동이 제한된 범위 내에 있어야 함
      - 시각 특징 편차 norm은 후반 레이어로 갈수록 증가



CVSP Lab
Computer Vision & Signal Processing

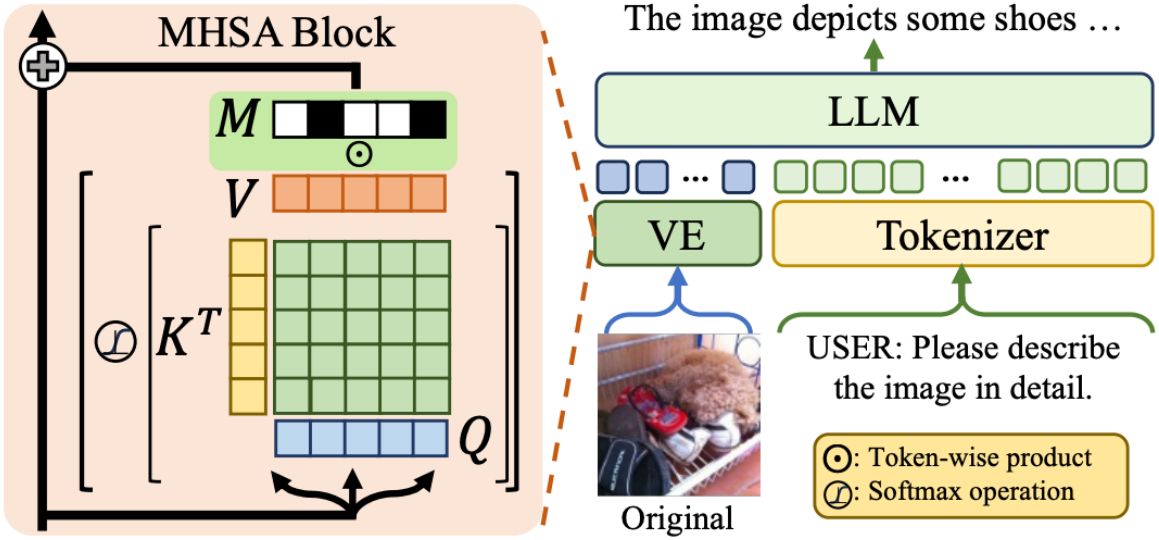
Method

- Proposed method
  - 불확실한 시각 토큰 Masking
    - 불확실성 마스크 M을 중간 레이어에 적용하여 어텐션 계산 수정

$$Attention(Q,K,V,M) = (Softmax\left(\frac{QK^{T}}{\sqrt{d'}}\right)V)\odot M$$

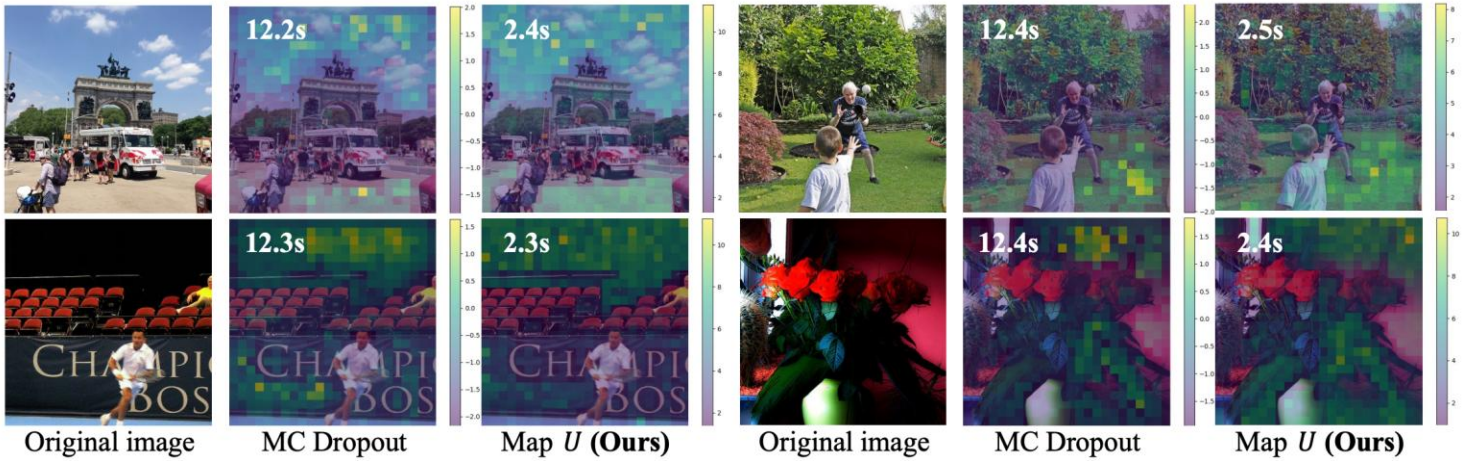
✓⊙: 토큰 단위 곱 연산

마스킹은 redsidual connection 구조 내에서 적용되며, 어텐션 가중치 및 토큰 간 상호작용은 유지하여 안정적으로 불확실한 토큰이 어텐션 출력에 미치는 영향 억제

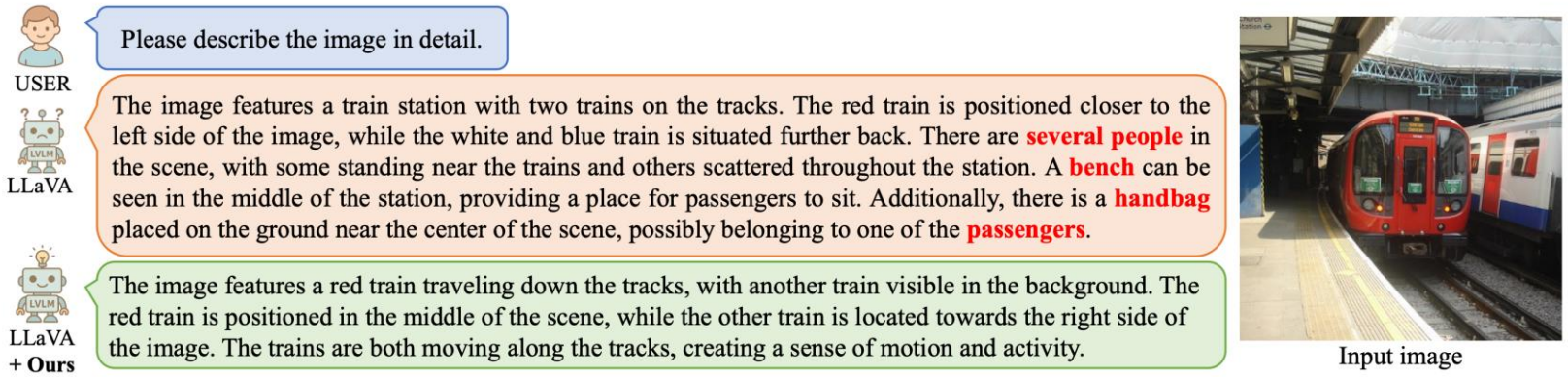


Experiments

- Qualitative Results



< Visual comparison of estimated uncertainty from MC dropout and proposed method>



< Qualitative results of proposed method on LLaVA-1.5-7B >

CVSP Lab
Computer Vision & Signal Processing

Experiments

- Quantitative Results
  - CHAIR 및 POPE 벤치마크의 성능
    - 모델: LLaVA-1.5-7B/ Shikra-7B/ MiniGPT-4
    - 호환성 평가를 위해 기존 기법(LLM)들과 결합한 실험 결과 제시

	Method	Greedy			OPERA			VCD			PAI			Devils		
		Orig.	+Ours	Δ%	Orig.	+Ours	Δ%	Orig.	+Ours	Δ%	Orig.	+Ours	Δ%	Orig.	+Ours	Δ%
LLaVA-1.5-7B	$C_s \downarrow$	47.4	29.2	↓38.4%	47.8	29.4	↓38.5%	53.8	35.2	↓34.6%	33.2	26.0	↓21.7%	27.0	**23.0**	↓14.8%
	$C_i \downarrow$	12.2	9.3	↓23.8%	12.8	9.5	↓25.8%	15.2	10.7	↓29.6%	8.5	7.9	↓7.1%	6.6	**5.6**	↓15.2%
	F1 ↑	77.9	78.2	↑0.4%	77.7	**78.4**	↑0.9%	75.2	75.2	↑0.0%	78.3	77.2	↓1.4%	78.3	78.0	↓0.4%
	Rand. ↑	89.3	89.3	↑0.0%	89.2	88.6	↓0.7%	84.6	86.2	↑1.9%	89.4	89.2	↓0.2%	89.6	**90.0**	↑0.4%
	Pop. ↑	85.8	85.8	↑0.0%	85.8	85.2	↓0.7%	82.4	82.9	↑0.6%	86.0	86.4	↑0.5%	86.4	**87.2**	↑0.9%
	Adv. ↑	79.3	80.0	↑0.9%	**80.3**	79.6	↓0.9%	77.0	78.1	↑1.4%	79.5	79.9	↑0.5%	78.6	79.6	↑1.3%
Shikra-7B	$C_s \downarrow$	58.0	43.2	↓25.5%	34.8	28.8	↓17.2%	56.2	47.2	↓16.0%	32.4	22.2	↓31.5%	24.4	**20.6**	↓15.6%
	$C_i \downarrow$	15.6	11.7	↓25.0%	11.1	9.6	↓13.5%	16.1	12.8	↓20.5%	7.8	**6.1**	↓21.8%	7.6	6.8	↓10.5%
	F1 ↑	74.7	**76.9**	↑2.9%	74.2	74.2	↑0.0%	74.4	75.2	↑1.1%	76.7	75.1	↓2.1%	73.3	72.2	↓1.5%
	Rand. ↑	83.2	85.1	↑2.3%	84.8	**85.4**	↑0.7%	82.1	82.7	↑0.7%	83.9	84.0	↑0.1%	83.8	82.5	↓1.6%
	Pop. ↑	82.3	82.6	↑0.4%	82.8	82.1	↓0.8%	79.7	80.7	↑1.3%	**83.1**	80.7	↓2.9%	79.9	78.2	↓2.1%
	Adv. ↑	78.2	78.8	↑0.8%	79.2	**79.7**	↑0.6%	77.3	77.1	↓0.3%	78.8	77.4	↓1.8%	77.7	76.7	↓1.3%
MiniGPT-4	$C_s \downarrow$	28.6	27.4	↓4.2%	23.8	22.6	↓5.0%	32.0	30.6	↓4.4%	19.6	**17.8**	↓9.2%	21.6	20.8	↓3.7%
	$C_i \downarrow$	8.5	8.3	↓2.4%	8.8	8.5	↓3.4%	9.7	9.1	↓6.2%	6.2	**6.0**	↓3.2%	7.5	7.0	↓6.7%
	F1 ↑	71.5	71.3	↓0.3%	69.8	70.0	↑0.3%	70.2	71.3	↑1.7%	**71.7**	**71.7**	↑0.0%	70.1	70.4	↑0.4%
	Rand. ↑	**82.8**	82.5	↓0.4%	74.2	74.4	↑0.3%	59.2	59.3	↑0.2%	82.1	82.0	↓0.1%	77.4	77.8	↑0.5%
	Pop. ↑	75.1	74.6	↓0.7%	71.3	71.8	↑0.7%	54.9	55.0	↑0.2%	**75.8**	75.2	↓0.8%	68.4	68.6	↑0.3%
	Adv. ↑	71.8	71.2	↓0.8%	69.7	69.4	↓0.4%	53.8	54.2	↑1.1%	**72.1**	71.6	↓0.7%	65.2	65.3	↑0.2%

CVSP Lab
Computer Vision & Signal Processing

Experiments

- Ablation studies
  - 마스크를 추출하는 레이어와 적용하는 레이어에 관한 실험
    - 마스크 추출: 초기 레이어에서 적대적 섭동에 대한 특징 변화가 더 적어 불확실성을 추정하기에 더 적합
    - 마스크 적용 : 중간 레이어(13-17)에 마스킹을 적용할 때 환각이 가장 크게 감소하며 F1스코어도 최고를 기록함

Mask Source Layer	$C_s\downarrow$	$C_i\downarrow$	F1 $\uparrow$
Greedy	47.4	12.2	77.9
Layers 1–10	**29.2**	**9.3**	**78.2**
Layers 11–20	44.2	12.7	77.4
Layers 21–22	41.8	12.1	77.9

< Impact of vision encoder layers on generating the uncertainty mask. >

Masking Layer Range	$C_s\downarrow$	$C_i\downarrow$	F1 $\uparrow$
Greedy	47.4	12.2	77.9
Layers 1–8	45.0	12.6	77.9
Layers 8–12	55.8	15.5	75.7
Layers 13–17	**29.2**	**9.3**	**78.2**
Layers 18–22	45.8	13.0	77.7

< Effect of applying the uncertainty mask to different layers in the vision encoder. >

CVSP Lab
Computer Vision & Signal Processing

Conclusion

- Future work
  - 경량화기법에 적용할 수 있는 환각 완화 기법 연구
    - 불확실한 이미지 토큰들의 비전 인코더 어텐션 패턴 분석
      - 헤드별 특징
      - 레이어별 특징
    - 프로젝터 이후 어떻게 변하는지 분석

Appendix

- 미분 엔트로피
  - 연속확률변수에 대한 엔트로피
    - $h(Z) = -\int p(z) log p(z) \mathrm{dz}$
      - 값이 클수록 불확실성이 큼
    - 논문에서의 확률 변수 $Z_i^{(t)}$:적대적 섭동에 의해 변하는 i번째 시각 토큰표현
      - 미분 엔트로피의 직접 계산이 거의 불가능
      - 가우시안 분포의 엔트로피 성질 활용
        - ✓같은 분산을 가지는 분포 중 미분 엔트로피가 최대인 것은 가우시안 분포
        - ✓가우시안을 따르는 확률 변수의 미분 엔트로피 구할 수 있음
        - ✓이를 상한으로 하여 시각 토큰 표현의 불확실성 표현 가능

CVSP Lab
Computer Vision & Signal Processing