Reporting Bias and LLMs' Commonsense Knowledge

2026. 01. 29.

양 건

Scaling Laws Do Not Solve Everything

- 대규모 언어 모델(LLMs) 에 대한 현재 인식
  - Scaling Laws를 따름 → 사전학습 규모가 커질수록 성능 향상
    ↳전제 조건: 모델이 필요로 하는 지식이 사전학습 **텍스트에 충분히** 존재해야 함

하지만. . . 텍스트로 기록되지 않는 지식이 있다면?

The Nature of Human Communication

- 인간의 의사소통은 **효율성**을 추구
  - Paul Grice's Maxim of Quantity
    "상대에게 필요한 만큼의 정보만 제공하라"

  - 청자 / 독자 가 이미 알거나 상식으로 추론 가능한 정보는 생략
    → 많은 정보가 Implicit 하게 남겨짐

예시 대화)
A: "Is there a gas station nearby?"
B: "There's one around the corner."
→ B의 대답에서 생략된 정보: The gas station should be open and operational

From Reality to Language: Information Bottlenecks

- Reporting Bias (보도 편향)
  - 커뮤니케이션은 본질적으로 선택적/압축적
    - 소통 시 효율성을 추구하는 인간의 특성으로 인해 발생
  - 따라서, 텍스트는 현실을 그대로 기록하지 않음
    - 글에 등장하는 빈도 ≠ 현실 세계 발생 빈도

  - 뉴스 기사
    - 눈길을 끄는 것 위주로 보도 (사소하고 평범한 이벤트 보도 X)
      예) 건물 화재 발생 (멀쩡한 건물은 언급 X)

  - 일상 대화
    - 서로에게 당연한 정보 언급 X
      예) “주말에 연구실 건물 정전이래요" (정전이 아닌 날은 굳이 불 들어온다고 언급 X)

- LLMs
  - Web Text 를 crawling 해서 세상을 배움
    - Reporting Bias 문제로 인해 사람과 동일한 추론 / 판단 을 하는데 한계 발생

Understanding the World Through Google N-grams

- "Reporting Bias and Knowledge Acquisition" (Gordon & Van Durme, 2013)
  - presented textual frequencies based on the Google Web 1T N-gram data
  - Teraword (1조 개 단어) 중 "사람" 다음 나오는 단어 N-gram 카운트 한 결과
  - 텍스트로만 세상을 배우면
    - 사람이 살인을 저지른 빈도 > 숨을 쉰 빈도
    - 사람이 지각한 빈도 > 제 때 도착한 빈도
  - 결과적으로,
    mundane events → text 내 under-represented
    sensational events → text 내 over-represented

Word	*Teraword*
Spoke	11,577,917
Laughed	3,904,519
Murdered	2,843,529
Inhaled	984,613
Breathed	725,034
Was late	368,922
Exhaled	168,985
Was on time	23,997

The “Black Sheep Problem”

- Reporting Bias 현상은 ML 쪽에서 The “black sheep problem” 으로도 잘 알려져 있음
  - 현실 세계 빈도? 흰 양 >>>>> 검은 양
    - 양은 흰색 털을 유도하는 유전자가 우성 → 자연적으로 흰 양 비율 높음
    - 흰 털이 염색이 용이해 상품성이 높음 → 사육 시에도 주로 흰 양만 사육
  - 영어 텍스트 내 빈도? 검은 양 >>>>> 흰 양 (25 : 1)
    - 이유: “black sheep” 이라는 idiomatic expression 이 있음
      - '어떤 집단 (특히 가족) 내에서 환영받지 못하는 천덕꾸러기’ 라는 뜻
    - 유래: 흰 양 사이에서 검은 양이 태어날 확률은 약 0.1%
      → 양 사육 농가에서 검은 양이 태어나면 골칫덩어리로 여기던 관행에서 비롯됨
  - 따라서 영어 텍스트 빈도만으로 “양은 보통 무슨 색?”을 추정하면 현실과 반대로 왜곡될 수 있음
  - 이러한 현상은 언어 별로 다르게 발생
    - 한국어 corpus에서는 ‘흰 양' : ‘검은 양' 이 1.5 : 1 로 흰 양이 우세
    - 한국어 사례: 세잎 클로버 vs 네잎 클로버
      - 현실 빈도: 10,000 : 1 (0.01% 확률로 네잎 클로버가 돌연변이로 발생)
      - 텍스트 빈도: 1 : 2 (네잎 클로버가 행운의 상징이라 더 자주 언급됨)

미니 테스트: 빈칸 채우기

- 사용 모델: ChatGPT 5.2 Thinking

Please fill in the [blank] from the below sentences, limiting answer choices only to colors.

1. The [blank] sheep wandered away from the flock and into the tall grass.
2. We spotted a [blank] sheep standing alone on the hillside at sunrise.
3. The farmer fed the [blank] sheep by hand every morning.
4. She knitted a scarf while watching the [blank] sheep graze quietly.
5. The dog barked once, and the [blank] sheep trotted back toward the barn.

1. The **black** sheep wandered away from the flock and into the tall grass.
2. We spotted a **black** sheep standing alone on the hillside at sunrise.
3. The farmer fed the **black** sheep by hand every morning.
4. She knitted a scarf while watching the **black** sheep graze quietly.
5. The dog barked once, and the **black** sheep trotted back toward the barn.

미니 테스트: 빈칸 채우기

- 사용 모델: Empathetic KULLM (Default)

아래 문장의 [빈칸]에 올 단어로 가장 알맞은 걸 골라줘.
나는 길을 가다가 클로버를 발견했다. 당연하게도, 그 클로버는 [빈칸]잎 클로버였다.

나는 길을 가다가 클로버를 발견했다. 당연하게도, 그 클로버는 **네** 잎 클로버였다.

아래 문장의 [빈칸]에 올 단어로 가장 알맞은 걸 골라줘.
나는 길을 가다가 클로버를 발견했다. 그 클로버는 평범한 [빈칸]잎 클로버였다.

나는 길을 가다가 클로버를 발견했다. 그 클로버는 평범한 **네** 잎 클로버였다.

아래 문장의 [빈칸]에 올 단어로 가장 알맞은 걸 골라줘.
네 잎 클로버는 행운을 가져다 주지만 [ ] 잎 클로버는 행복을 가져다 준대.

네 잎 클로버는 행운을 가져다 주지만 **네** 잎 클로버는 행복을 가져다 준대.

How Reporting Bias affects LLMs' Commonsense Knowledge

- Common Sense?
  - 특별한 전문 지식이나 교육 없이도 **사회 구성원이 일반적으로 공유** 하는, 합리적이고 실용적인 판단 능력
  - 한 사회/집단의 대다수에 의해 "common" 하다고 여겨지기에 (Shared Understanding),
    말/글로 explicit 하게 언급되는 빈도가 낮음

- Reporting Bias 가 LLM 의 commonsense reasoning 에 영향을 주는 사례들
  - Typicality / Rarity
    - Event Outcomes (인과 관계)
      - 한 이벤트의 결과로 일어날 다음 이벤트를 예측할 때, 흔하지 않은 결과도 발생할 확률 높다고 판단
        예) 수도꼭지를 틀자 → 물줄기가 일정하게 흘러나왔다 (typical)
        물줄기가 세면대 밖으로 솟구쳤다 (rare)
    - Entity Properties (e.g., colors, sizes, and shapes)
      - "yellow" banana

  - Procedural Knowledge
    - Event Sequence Description 생성에 취약 (과정이 누락되거나 진행 순서에 오류)
      - 양치하는 과정
      - 신발끈 묶는 과정

VISUALLY-AUGMENTED LANGUAGE MODELING

Weizhi Wang[†], **Li Dong**[‡], **Hao Cheng**[‡], **Haoyu Song**[‡],
Xiaodong Liu[‡], **Xifeng Yan**[†], **Jianfeng Gao**[‡], **Furu Wei**[‡]
[†]University of California, Santa Barbara [‡]Microsoft Research
weizhiwang@ucsb.edu, {lidong1, haocheng}@microsoft.com

VaLM (ICLR 2023)

Improving the Efficiency of Visually Augmented Language Models

Paula Ontalvilla **Aitor Ormazabal** **Gorka Azkune**
HiTZ Center - Ixa, University of the Basque Country (UPV/EHU)
{paula.ontalvilla, aitor.ormazabal, gorka.azcune}@ehu.eus

BLIND-VaLM (COLING 2025)

VaLM: Motivation

- Text-only pre-training 의 한계
  - 텍스트는 “당연한 시각적 정보(색/모양/크기 등)”를 자주 생략
  - Visual commonsense가 필요한 문제에서 LM이 약해질 수 있음

- 시각 정보는 해당 지식을 직접 포함
  - 이를 활용하면 visual-knowledge-intensive commonsense 성능 향상 기대

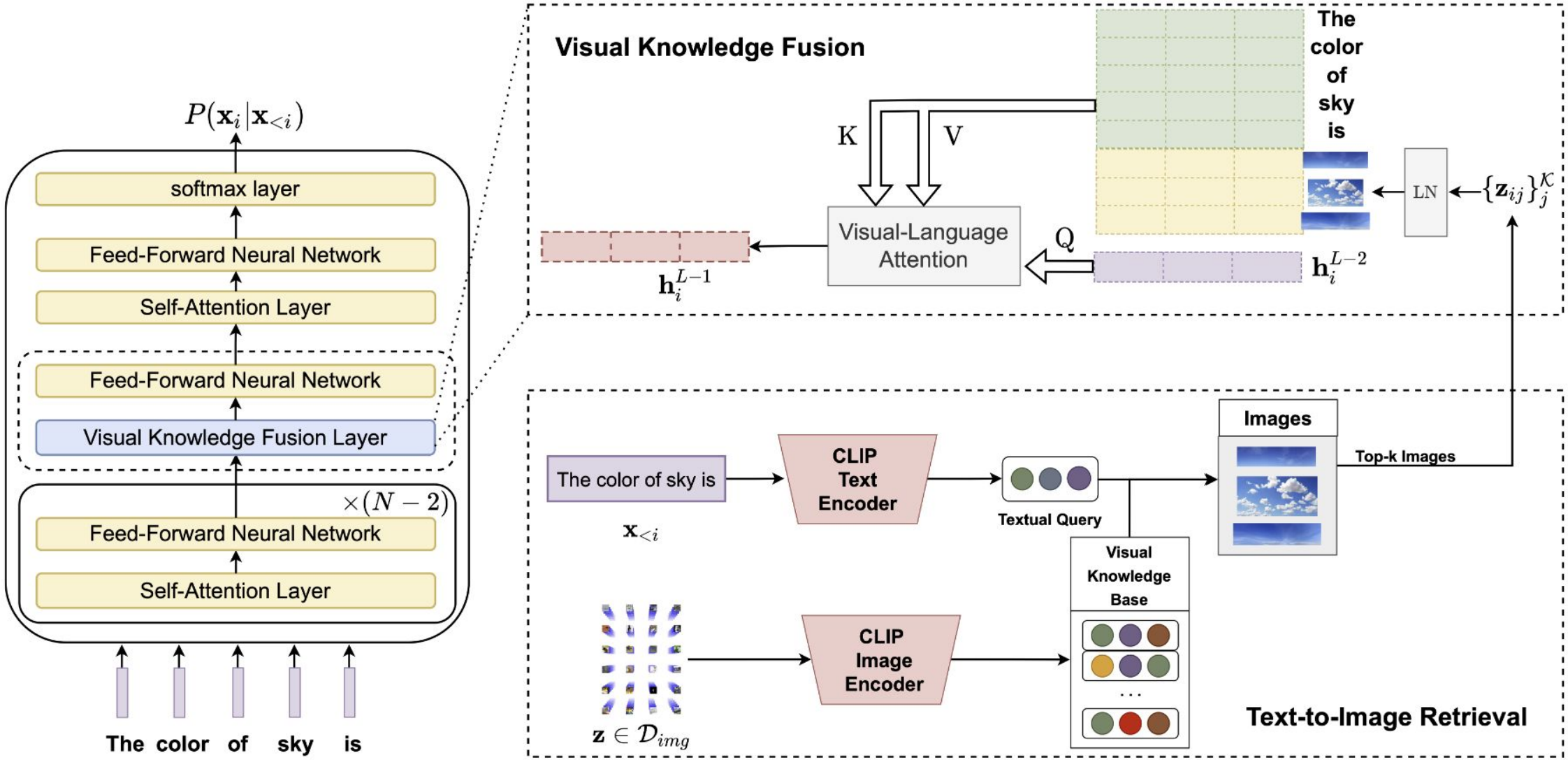
- 기존 VLM의 한계
  - VLM pre-training 은 주로 (이미지, 캡션) pair 기반의 explicit alignment 에 의존
  - 따라서 텍스트만 주어지는 추론 환경에서는 cross-modality fusion 이 제한적
    → Text-only 환경에서도 필요할 때 시각 지식을 끌어와 결합하는 메커니즘이 필요

Methods

- 핵심 아이디어: “이미지를 입력으로 받지 않아도, 필요하면 가져와서 붙이자"
  → 텍스트 입력만 주어져도, 토큰별로 관련 이미지를 검색해 시각적 지식 증강

- Step 0: Backbone LM
  - 텍스트를 표준 Decoder-Only Transformer LM에 입력
  - 중간 레이어에서 각 토큰의 Hidden Representation 을 생성 (이후 단계의 '쿼리' 역할)
- Step 1: Token-level Image Retrieval
  - Step 0 의 Hidden Representation 을 쿼리로 사용
  - CLIP 임베딩 공간을 활용해, 대규모 이미지 DB에서 문맥과 가장 관련성 높은 Top-K 이미지 검색
    - CLIP: 대규모 이미지-텍스트 쌍으로 학습된 contrastive VLM
      - text↔image가 같은 임베딩 공간에 정렬되어 있어 retrieval이 가능
- Step 2: Visual Knowledge Fusion Layer
  - LM 블록 내부에 추가된 특수 계층
  - Joint Attention 메커니즘 구현
    - Query: 텍스트 토큰의 Hidden Representation
    - Key & Value: [텍스트 토큰; 검색된 이미지 패치 특징]
      → 결과: 다음 토큰을 예측할 때 텍스트 맥락뿐 아니라 시각적 단서도 함께 참조
- Step 3: Output
  - 이후는 일반 LM과 동일: next-token prediction (standard LM loss)

VaLM Overview



Visual Knowledge Fusion

hidden state $\mathbf{H}_i^{L-1}$ is computed as:

$$\mathbf{Q} = \mathbf{H}^{L-2}W^Q + b^Q, \mathbf{K} = \mathbf{H}^{L-2}W^K + b^K, \mathbf{V} = \mathbf{H}^{L-2}W^V + b^V, \tag{2}$$

$$\dot{\mathbf{k}}_{ik} = \mathrm{LN}_{img}(\mathbf{z}_{ik})W^K + b^K_{img}, \dot{\mathbf{v}}_{ik} = \mathrm{LN}_{img}(\mathbf{z}_{ik})W^V + b^V_{img}, \tag{3}$$

$$e_i = \frac{\mathbf{Q}_i\mathbf{K}^T}{\sqrt{d}}, a_i = \frac{\exp(e_i)}{\sum_{j=1}^{\mathcal{L}}\exp(e_{ij}) + \sum_{k=1}^{\mathcal{K}}\exp(e_{ik})}, \tag{4}$$

$$e_{ik} = \frac{\mathbf{Q}_i\dot{\mathbf{k}}_{ik}^T}{\sqrt{d}}, a_{ik} = \frac{\exp(e_{ik})}{\sum_{j=1}^{\mathcal{L}}\exp(e_{ij}) + \sum_{k=1}^{\mathcal{K}}\exp(e_{ik})}, \tag{5}$$

$$\mathbf{H}_i^{L-1} = a_i\mathbf{V} + \sum_k a_{ik}\dot{\mathbf{v}}_{ik}, \tag{6}$$

Results

- Visual knowledge-intensive commonsense (Object property reasoning)
  - 색/모양/크기 같은 시각적 속성 추론 벤치마크 4종
    MemoryColor / ColorTerms / ObjectShape / RelativeSize

Model	$\mathcal{K}$	Color (ACC↑)		Shape (ACC↑)	Size (ACC↑)
		MEMORYCOLOR	COLORTERMS	OBJECTSHAPE	RELATIVESIZE
GPT-2*	N/A	44.14%	39.10%	51.09%	47.22%
BERT	N/A	24.34%	26.33%	31.86%	34.78%
CaptionBERT	N/A	24.84%	28.40%	38.14%	66.05%
CLIP	N/A	26.25%	23.08%	13.66%	47.99%
OSCAR	N/A	20.32%	16.86%	33.14%	50.14%
VisualBERT	N/A	26.68%	38.02%	11.14%	67.23%
VALM	4	53.99%	**52.66%**	**62.77%**	**85.03%**
VALM	8	**58.64%**	50.19%	59.41%	62.35%

Results

- Commonsense QA (물리 상호작용)
  - PIQA (Physical Interaction QA)

Model	GPT-2*	BERT	CaptionBERT	VALM ($\mathcal{K}$=4)	VALM ($\mathcal{K}$=8)
PIQA (ACC↑)	62.53%	54.73%	53.97%	64.31%	**64.64%**

BLIND-VaLM: Motivation

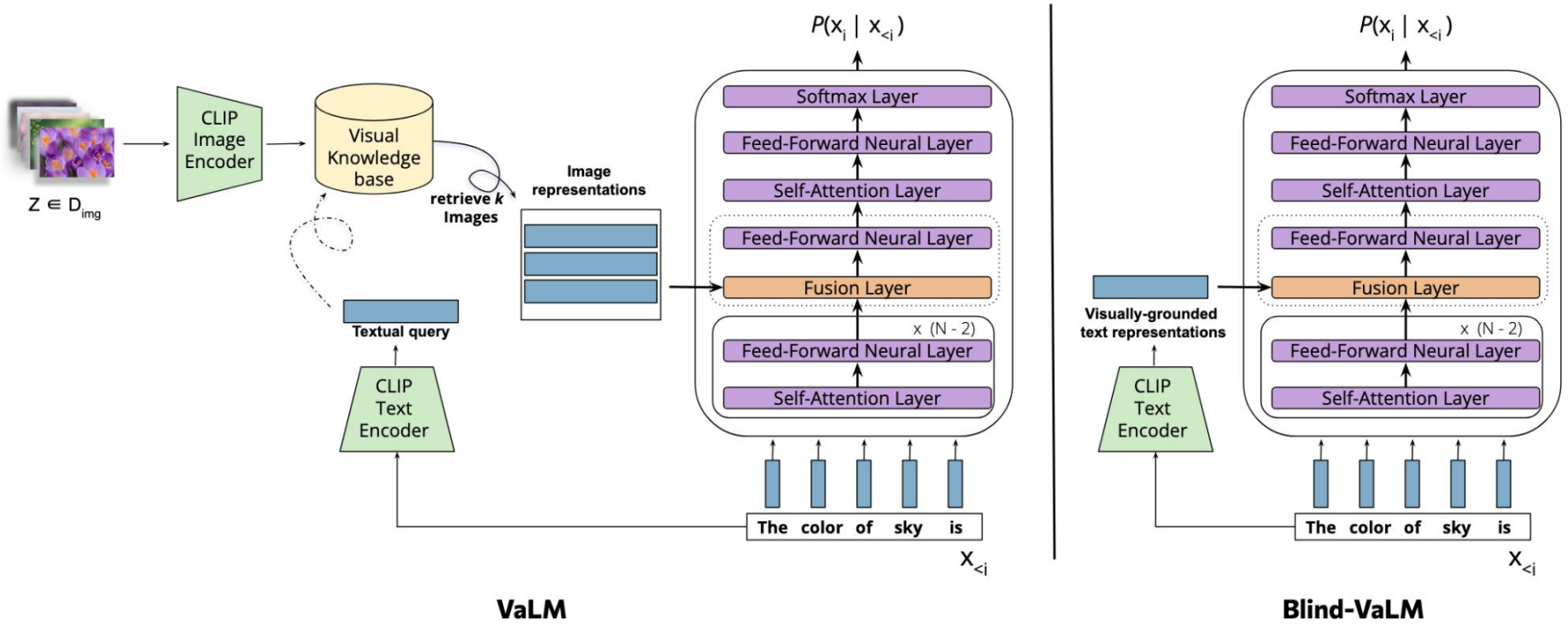
- Image Retrieval 이 가져오는 Overhead
  - VaLM: LM 에 visual knowledge를 주입하기 위해 Image Retrieval 사용
  - Image Retrieval + Image Encoding 로 인해 훈련/추론 모두 매우 비싼 비용(시간·자원) 요구
    → 실용성 낮음

- RQ: LM 에 visual knowledge 를 주입하려면 반드시 이미지가 필요할까?

Methods

- 핵심 아이디어: “이미 CLIP 이 이미지와 텍스트를 정렬시켜 놨으니 이 텍스트를 시각 정보 대용으로 사용해보자"
  → 실제 이미지를 보지 않고 (Blind), 오직 CLIP 텍스트 인코더가 이미 학습한 시각–언어 연관 지식만 활용

- BLIND-VaLM 의 architecture: VaLM 의 병목 (이미지 검색 + 이미지 인코딩) 지점을 CLIP-text 로 대체



Results

- 동일 조건으로 VaLM 과 BLIND-VaLM 생성 후 학습 효율성 및 성능 비교
  - Base: GPT2-SMALL
  - GPU: A100 8장
  - Text Corpus: CC-100 English 중 10.5B tokens (약 19%)

- Efficiency (2.2x faster)
  - VaLM: 1.2K GPU-hours
  - BLIND-VaLM: 530 GPU-hours

- 벤치마크 결과
  - 이미지 없이도 VaLM과 동급 (또는 소폭↑) / 효율성은 극적 개선

Model	Color (ACC ↑)			Shape (ACC ↑)	Size (ACC ↑)		AVG
	MemoryC	ColorTerms	VCT	ShapeITC	RelativeS	TNWT	
VALM	47.09	41.88	20.46	**40.45**	**26.03**	23.94	33.30
BLIND-VALM	**47.20**	**46.37**	**22.60**	40.07	25.43	**25.18**	**34.48**

Results

- Compute-matched Scaling
  - 주장: VaLM 이 사용한 만큼의 자원을 쓰면 우리 성능도 더 좋아질 수 있다
    - BLIND-VaLM+ (pre-training compute scaling, 1.17K GPU-hours)
    - BLIND-VaLM-M (model size scaling, GPT2-Medium, 595 GPU-hours)

Model	Color (ACC ↑)			Shape (ACC ↑)	Size (ACC ↑)		AVG
	MemoryC	CTerms	VCT	ShapeITC	RelativeS	TNWT	
VALM	47.09	41.88	20.46	40.45	**26.03**	23.94	33.30
BLIND-VALM	47.20	46.37	22.60	40.07	25.43	25.18	34.47
BLIND-VALM+	45.97	48.71	20.51	**43.64**	25.33	**25.40**	34.93
BLIND-VALM-M	**47.60**	**48.93**	**24.09**	43.33	24.36	24.93	**35.54**

Limitations

- CLIP 의존성 (VaLM / BLIND-VaLM 공통)
  - 두 방법 모두 CLIP 이 학습한 text-image 정렬 표현에 크게 의존
  - 따라서 “시각적 이해”의 상한은 결국 기반 VLM/CLIP의 품질에 의해 결정됨
    → 근본 개선은 CLIP(또는 대체 VLM) 의 발전에 달림

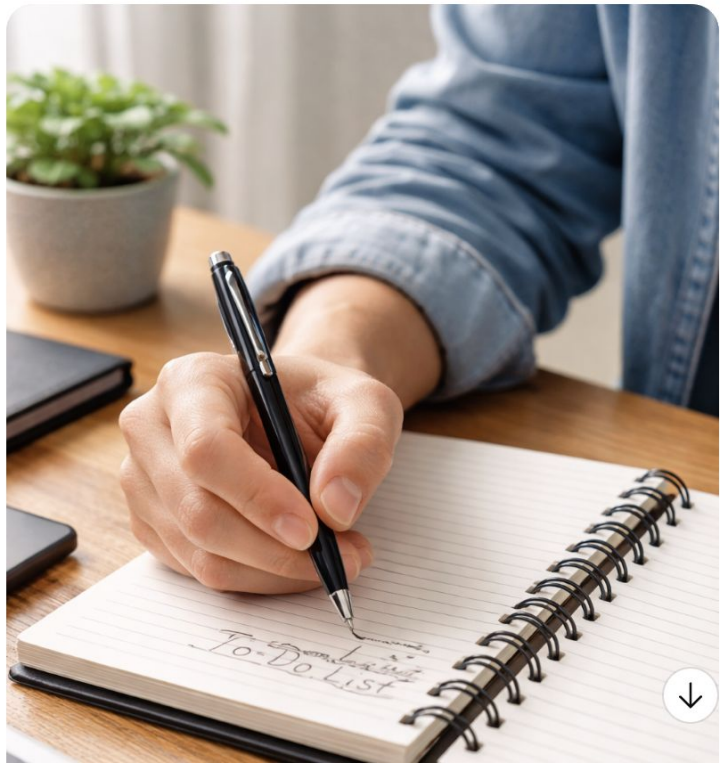
- VLM 도 Reporting Bias 로 부터 자유롭지 못함
  - 텍스트의 reporting bias를 완화하려고 이미지를 쓰지만, 이미지, 캡션 데이터 모두 bias 존재

Future Directions: Reporting Bias in Images

- 현실에서 다수인 것이 이미지 데이터에서도 대부분을 차지
  → Minority Representation 힘듦

create an image of someone writing with their left hand.

Image created • 할 일 목록 작성 중



Future Directions: Reporting Bias in Image Captions

- 캡션이 이미지의 핵심 속성(색/형태/상태/배경 등)을 자주 생략
  - 동일한 캡션이 서로 다른 이미지에 붙는 사례
  - 모델이 object-attribute 구분을 못 하거나 잘못 정렬하게 됨

?

slices ◄·····► fish

“salmon *(slices)* on the table”

“salmon *(fish)* on the table”

Thank you