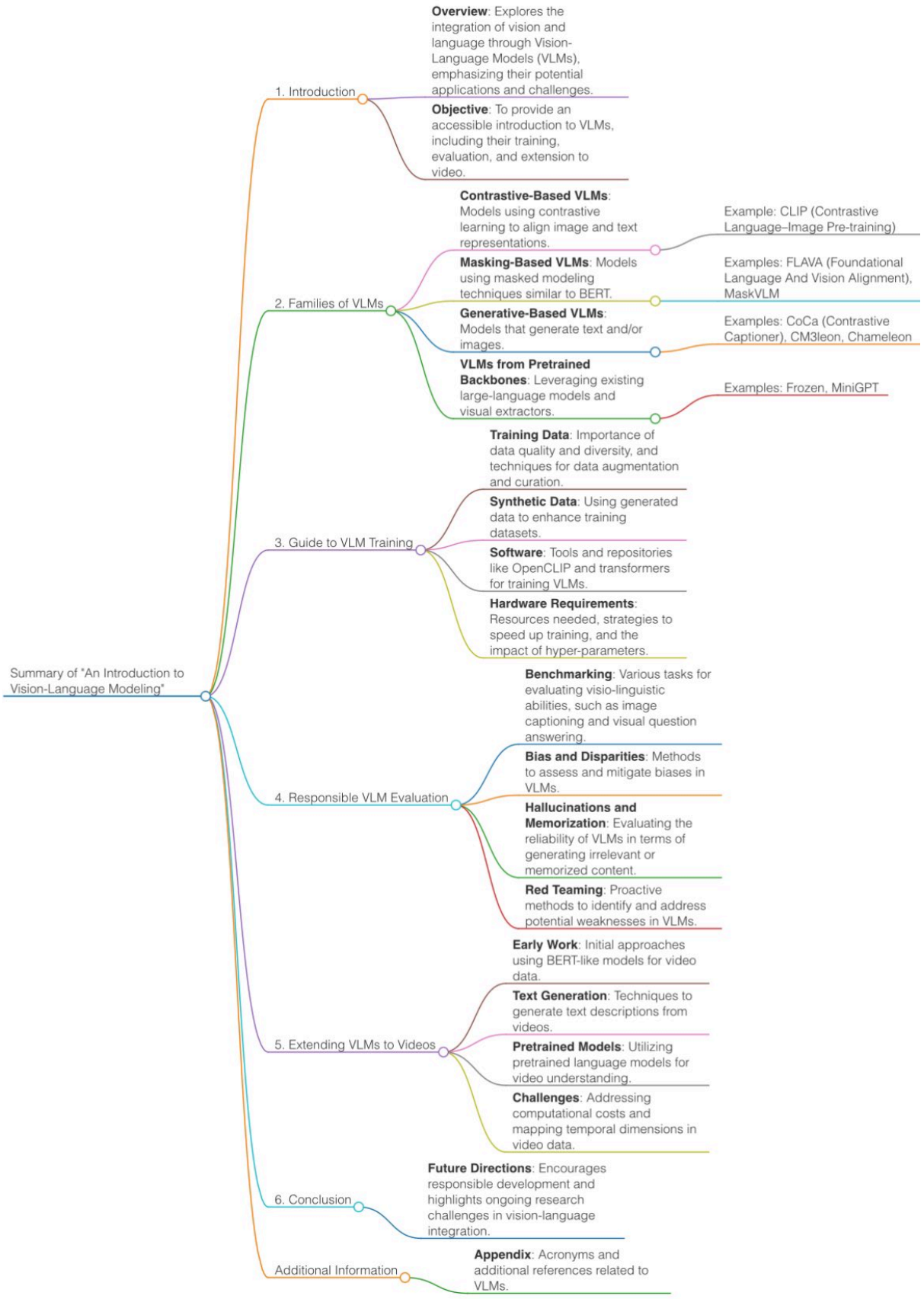
Summary of VLMs



출처: https://x.com/MindBranches/status/1795501373179498941

1. 서론 (Introduction)

최근 대규모 언어 모델(LLMs)의 인기와 함께 이를 시각적 도메인으로 확장하려는 여러 시도가 이루어지고 있습니다. 비전-언어 모델(VLMs)의 응용은 단순한 텍스트 설명을 통해 이미지를 생성하는 생성형 모델에서부터 낯선 환경에서 길을 안내하는 시각적 보조 도구까지, 기술과 우리의 관계에 큰 영향을 미칠 것입니다. 하지만 이러한 모델의 신뢰성을 높이기 위해 해결해야 할 과제가 많습니다.

언어는 이산적인 특성을 갖고 있지만, 비전은 개념이 항상 쉽게 이산화될 수 없는 훨씬 높은 차원의 공간에서 진화합니다. 비전을 언어로 매핑하는 메커니즘을 더 잘 이해하기 위해, 우리는 이 문서에서 VLMs에 대한 입문서를 제공합니다. 이 문서는 이 분야에 진입하려는 사람들에게 도움이 되기를 바랍니다.

우리는 먼저 VLM이 무엇인지, 어떻게 작동하는지, 그리고 어떻게 훈련할 수 있는지를 소개합니다. 그런 다음, 다양한 연구 목표에 따라 VLM을 평가하는 접근 방식을 제시하고 논의합니다. 이 작업은 주로 이미지를 언어로 매핑하는 데 중점을 두지만, VLM을 비디오로 확장하는 방법도 논의합니다.

최근 몇 년 동안 언어 모델링 분야에서 인상적인 발전이 있었습니다. Llama나 ChatGPT와 같은 많은 대규모 언어 모델(LLMs)은 이제 매우 다양한 작업을 해결할 수 있으며, 그 사용이 점점 더 보편화되고 있습니다. 텍스트 입력에 주로 국한되었던 이러한 모델은 이제 시각적 입력으로 확장되고 있습니다. 비전과 언어를 연결하는 것은 현재 AI 기반 기술 혁명의 핵심이 될 여러 응용 프로그램을 가능하게 할 것입니다.

이미 여러 연구가 대규모 언어 모델을 비전으로 확장했지만, 언어를 비전으로 연결하는 것은 아직 완전히 해결되지 않았습니다. 예를 들어, 대부분의 모델은 공간적 관계를 이해하거나 복잡한 엔지니어링 오버헤드 없이 계산하는 데 어려움을 겪습니다. 많은 비전-언어 모델(VLM)은 속성과 순서를 이해하는 데도 부족합니다. 또한 입력 프롬프트의 일부를 무시하고, 원하는 결과를 생성하기 위해 상당한 프롬프트 엔지니어링 노력이 필요합니다. 일부는 환각(hallucination)을 일으켜 요구되지 않거나 관련 없는 콘텐츠를 생성하기도 합니다. 결과적으로, 신뢰할 수 있는 모델을 개발하는 것은 여전히 매우 활발한 연구 분야입니다.

이 작업에서는 비전-언어 모델(VLMs)에 대한 소개를 제공합니다. 우리는 VLM이 무엇인지, 어떻게 훈련되고, 다양한 연구 목표에 따라 이를 효과적으로 평가할 수 있는 방법을 설명합니다. 이 작업은 VLM 연구 분야의 설문 조사나 완전한 가이드로 간주되어서는 안 됩니다. 따라서 우리는 VLM 연구 분야의 모든 작업을 인용하는 것을 목표로 하지 않습니다. 이 작업은 이 분야에서 효과적인 연구를 위해 명확하고 이해하기 쉬운 소개를 제공하며, 특히 다른 분야의 연구자들이 이 분야로 진입하는 데 유용합니다.

우리는 VLM 훈련 패러다임을 먼저 소개합니다. 대조적 방법이 어떻게 이 분야를 변화시켰는지 논의합니다. 그런 다음 마스킹 전략이나 생성적 요소를 활용하는 방법을 제시합니다. 마지막으로, 사전 학습된 백본(예: LLM)을 사용하는 VLM을 제시합니다. VLM을 다양한 범주로 나누는 것은 쉽지 않습니다. 대부분의 모델이 중첩된 구성 요소를 가지고 있기 때문입니다. 하지만 이러한 범주화가 새로운 연구자들이 이 분야를 탐색하고 VLM의 내부 메커니즘을 밝히는 데 도움이 되기를 바랍니다.

다음으로, VLM을 훈련하기 위한 전형적인 레시피를 제시합니다. 예를 들어, 특정 연구 목표에 따라 어떤 데이터셋이 적합한지, 데이터 큐레이션 전략은 무엇인지, 텍스트 인코더를 훈련해야 하는지 아니면 사전 학습된 LLM을 활용할 수 있는지, 비전 이해를 위해 대조 손실이 충분한지 또는 생성적 요소가 핵심인지 등을 다룹니다. 또한 모델 성능을 향상시키는 데 사용되는 일반적인 기술뿐만 아니라 정렬과 보다 나은 맞춤화 기술도 제시합니다.

이어서 VLM의 니즈를 더 잘 이해하기 위해 모델을 훈련하는 레시피를 제공하는 것이 중요한 단계지만, 이러한 모델을 견고하고 신뢰할 수 있도록 평가하는 것도 동일하게 중요합니다. VLM을 평가하는 데 사용되는 많은 벤치마크가 최근 도입되었습니다. 그러나 이러한 벤치마크 중 일부는 연구자가 인지해야 할 중요한 한계가 있습니다. VLM에 대한 이해를 심화하기 위해 앞으로의 과제를 밝히고자 이러한 벤치마크의 강점과 약점을 논의합니다.

다음 세대의 VLM은 비디오를 언어로 매핑하여 비디오를 이해할 수 있게 될 것입니다. 그러나 비디오와 이미지는 여러 면에서 다릅니다. 계산 비용이 훨씬 높을 뿐만 아니라 텍스트를 통해 시간적 차원을 매핑하는 방법에 대한 다른 고려사항도 존재합니다.

VLM 연구의 진입 장벽을 낮춤으로써, 우리는 비전 이해의 경계를 확장하면서 더 책임감 있는 VLM 개발을 위한 기초를 제공하기를 바랍니다.

2. VLM의 유형 (The Families of VLMs)

딥러닝이 컴퓨터 비전 및 자연어 처리 분야에서 눈부신 발전을 이루면서, 두 도메인을 연결하려는 여러 시도가 이루어졌습니다. 이 논문은 트랜스포머 기반 최신 기술들에 중점을 둡니다【Vaswani et al., 2017】. 이러한 기술들을 네 가지 주요 훈련 패러다임으로 분류합니다(그림 1 참조).

1. 대조 학습**(Contrastive Learning)**:
    - 긍정(positive)과 부정(negative) 쌍의 데이터를 활용하는 일반적인 전략입니다.
    - 모델은 긍정 쌍에 대해 유사한 표현을 예측하도록 훈련되며, 부정 쌍에 대해서는 상이한 표현을 예측합니다.
2. 마스킹 학습**(Masking Learning)**:
    - 텍스트의 일부 단어나 이미지의 패치를 마스킹하고, 이를 복원하도록 모델을 학습시킵니다.
    - 캡션의 단어를 마스킹하거나 이미지를 마스킹하여 두 모달리티 간 관계를 학습합니다.
3. 생성 기반 학습**(Generative Learning)**:
    - 텍스트와 이미지를 생성할 수 있도록 학습합니다.
    - 이러한 모델들은 종종 훈련 비용이 많이 듭니다.
4. 사전 학습된 백본 활용**(Pretrained Backbone)**:
    - 공개된 LLM(예: Llama【Touvron et al., 2023】)과 같은 사전 학습된 모델을 활용하여, 이미지 인코더와 LLM 간 매핑을 학습합니다.
    - 텍스트 및 이미지 인코더를 처음부터 훈련하지 않아도 되는 장점이 있습니다.

이러한 패러다임은 상호 배타적이지 않으며, 많은 접근법이 대조 학습, 마스킹, 생성 기준을 혼합하여 사용합니다. 각 패러다임별로 하나 또는 두 개의 모델을 소개하며, 이들 모델의 설계에 대한 고급 정보를 제공합니다.

2.1 트랜스포머 기반 초기 VLM 연구 (Early Work on VLMs Based on Transformers)

트랜스포머 아키텍처【Vaswani et al., 2017】는 언어 모델링에서 비약적인 성능 향상을 이루었습니다. BERT【Devlin et al., 2019】는 당시 모든 언어 모델링 접근법을 능가했으며, 이를 시각적 데이터 처리로 확장한 연구도 이루어졌습니다.

- **Visual-BERT**【Li et al., 2019】와 **ViLBERT**【Lu et al., 2019】:
    - 텍스트와 이미지 토큰을 결합하여 학습.
    - 주요 학습 목표:
        1. 마스킹 모델링**(Masked Modeling)**: 입력의 누락된 부분을 예측.
        2. 문장**-**이미지 예측**(Sentence-Image Prediction)**: 캡션이 이미지 콘텐츠를 실제로 설명하는지 예측.

이러한 모델들은 주의(attention) 메커니즘을 통해 단어와 시각적 단서를 연결하는 능력 덕분에 강력한 성능을 발휘합니다.

2.2 대조 기반 VLMs (Contrastive-Based VLMs)

대조 기반 학습은 종종 에너지 기반 모델**(Energy-Based Models, EBM)**【LeCun et al., 2006】의 관점에서 설명됩니다. 모델 E_\theta는 관찰된 변수에는 낮은 에너지를, 관찰되지 않은 변수에는 높은 에너지를 할당하도록 훈련됩니다.

수식**:**

- 목표: 대상 분포 $P_D(x)$를 추정하는 전통적인 최대 우도(maximum likelihood) 목적:

$$\arg\min_{\theta} \mathbb{E}_{x\sim P_D(x)}[-\log P_\theta(x)]$$

- 그래디언트:

$$\frac{\partial \mathbb{E}_{x\sim P_D(x)}[-\log P_\theta(x)]}{\partial \theta} = \mathbb{E}_{x^+\sim P_D(x)}\frac{\partial E_\theta(x^+)}{\partial \theta} - \mathbb{E}_{x^-\sim P_\theta(x)}\frac{\partial E_\theta(x^-)}{\partial \theta}$$

CLIP:

- CLIP은 InfoNCE 손실을 활용하여 이미지와 텍스트를 공통 표현 공간에 매핑.
- 4억 개의 이미지-캡션 쌍으로 학습되었으며, 제로샷 학습에서 강력한 성능을 보임.
- 손실 함수:

$$L_{\text{InfoNCE}} = -\sum_{(i,j)\in P} \log \frac{\exp(\text{CoSim}(z_i, z_j)/\tau)}{\sum_{k=1}^{N}\exp(\text{CoSim}(z_i, z_k)/\tau)} \quad (2)$$

InfoNCE 손실을 사용하는 일반적인 대조 방법은 Contrastive Language-Image Pre-training (CLIP) [Radford et al., 2021]입니다. 긍정적인 예시 쌍은 하나의 이미지와 그에 해당하는 ground truth 캡션으로 정의되며, 부정적인 예시는 동일한 이미지이지만 미니 배치에 포함된 다른 이미지를 설명하는 다른 모든 캡션과 함께 사용됩니다. CLIP의 새로운 점은 비전과 언어를 공유 표현 공간에 통합하는 모델을 훈련시킨다는 것입니다. CLIP은 무작위로 초기화된 비전 및 텍스트 인코더를 훈련하여 대조 손실을 사용하여 이미지와 캡션의 표현을 유사한 임베딩 벡터에 매핑합니다. 웹에서 수집한 400만 개의 캡션-이미지 쌍에 대해 훈련된 원본 CLIP 모델은 놀라운 제로샷 분류 전이 기능을 보여주었습니다. 특히, ResNet-101 CLIP은 지도 학습된 ResNet [He et al., 2015] 모델의 성능과 일치했으며(76.2% 제로샷 분류 정확도 달성), 여러 견고성 벤치마크에서 이를 능가했습니다.
SigLIP [Zhai et al., 2023b]은 CLIP과 유사하지만, InfoNCE 기반의 CLIP 다중 클래스 목표 대신 이진 교차 엔트로피(binary cross-entropy) 기반의 원래 NCE 손실을 사용한다는 점에서 차이가 있습니다. 이러한 변경을 통해 CLIP보다 작은 배치 크기에서 더 나은 제로샷 성능을 얻을 수 있습니다.
잠재 언어 이미지 사전 훈련**(Latent language image pretraining**, Llip) [Lavoie et al., 2024]은 하나의 이미지에 대해 여러 가지 다른 방식으로 캡션을 달 수 있다는 사실을 고려합니다. Llip은 교차 주의 모듈을 통해 대상 캡션에 대한 이미지 인코딩을 조건화하는 방법을 제안합니다. 캡션 다양성을 고려하면 표현의 표현력이 높아지고 일반적으로 다운스트림 제로샷 전이 분류 및 검색 성능이 향상됩니다.

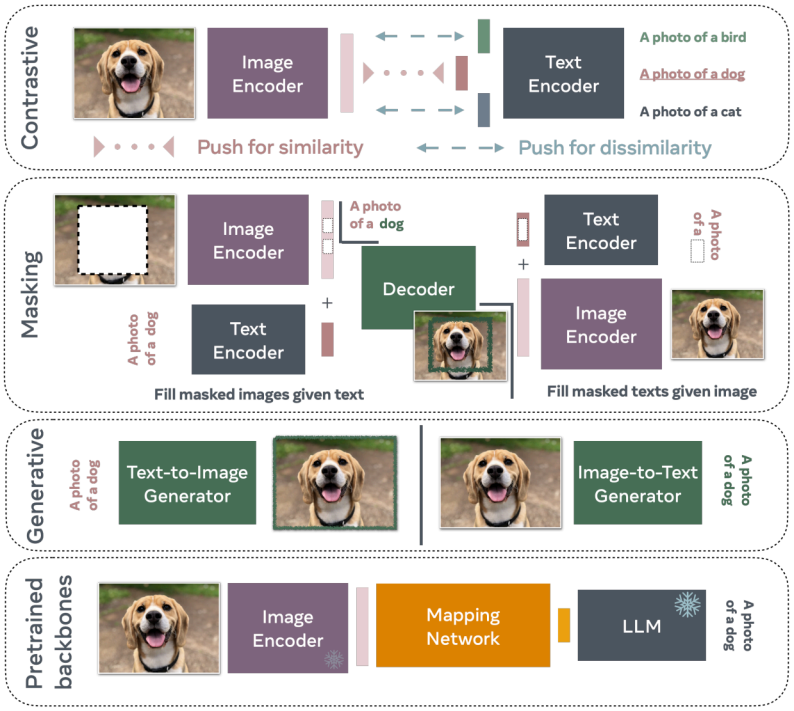


그림 **1: VLMs**의 패러다임

설명: 대조 학습, 마스킹 학습, 생성 기반 학습, 사전 학습된 백본 기반 학습의 구조를 시각적으로 설명합니다.
대조 학습은 긍정적 및 부정적 예 쌍을 사용하는 일반적인 전략입니다. VLM은 긍정 쌍에 대해 유사한 표현을 예측하고 부정 쌍에 대해 다른 표현을 예측하도록 훈련됩니다. 마스킹은 마스크되지 않은 텍스트 캡션이 주어졌을 때 누락된 패치를 재구성하여 VLM을 훈련하는 데 활용할 수 있는 또 다른 전략입니다. 마찬가지로 캡션의 단어를 마스크하면 마스크되지 않은 이미지가 주어졌을 때 해당 단어를 재구성하도록 VLM을 훈련할 수 있습니다. 이러한 접근 방식의 대부분은 중간 표현 또는 부분 재구성을 활용하지만, 생성 VLM은 전체 이미지 또는 매우 긴 캡션을 생성할 수 있도록 훈련됩니다. 이러한 모델의 특성상 훈련 비용이 가장 많이 드는 경우가 많습니다. 사전 훈련된 백본 기반 VLM은 종종 Llama와 같은 오픈 소스 LLM을 활용하여 이미지 인코더(사전 훈련될 수도 있음)와 LLM 간의 매핑을 학습합니다. 이러한 패러다임이 상호 배타적이지 않다는 점을 강조하는 것이 중요합니다. 많은 접근 방식은 대조, 마스킹 및 생성 기준을 혼합하여 사용합니다.

2.3 마스킹 기반 VLMs (VLMs with Masking Objectives)

마스킹은 딥러닝 연구에서 자주 사용되는 기법으로, 특정 입력의 일부를 제거(마스킹)한 후 이를 복원하도록 학습합니다. 이는 특정 공간적 구조를 가진 노이즈를 복원하는 디노이징 오토인코더(Denoising Autoencoder)【Vincent et al., 2008】로 볼 수 있습니다. 이 방법은 다음과 같은 두 가지 측면에서 발전해 왔습니다:

1. 마스킹 언어 모델링**(Masked Language Modeling, MLM)**:
    - BERT【Devlin et al., 2019】은 문장의 누락된 토큰을 예측하는 학습 방식을 도입.
2. 마스킹 이미지 모델링**(Masked Image Modeling, MIM)**:
    - MAE【He et al., 2022】는 이미지 패치를 마스킹한 후 이를 복원하도록 학습.

FLAVA:

- **FLAVA**는 텍스트와 이미지를 결합한 마스킹 기법을 활용하는 모델입니다.
- 아키텍처:
    - 이미지 인코더: Vision Transformer(ViT)을 기반으로 이미지 패치를 처리.
    - 텍스트 인코더: 텍스트를 벡터로 임베딩.

    - 멀티모달 인코더: 이미지와 텍스트 인코더의 출력(hidden states)을 결합하여 시각-언어 정보를 통합.
- 학습 전략:
    - 마스킹 기반 손실(masking-based losses)과 대조적 손실(contrastive loss)을 결합.

MaskVLM:

- **MaskVLM**은 FLAVA의 한계를 극복하고자 설계된 모델로, 기존 사전 학습된 인코더 의존도를 줄이기 위해 픽셀 수준에서 직접 마스킹을 적용합니다.
- 텍스트와 이미지가 서로 정보를 교환하며 복원 작업을 수행.

아래는 **2.3.3** 정보 이론적 관점에서 본 **VLM** 목표 **(Information Theoretic View on VLM Objectives)** 및 **2.4.3 생성 기반 텍스트-이미지 모델을 활용한 다운스트림 비전-언어 작업 (Using Generative Text-to-Image Models for Downstream Vision-Language Tasks)**의 설명과 수식 번역 및 해석입니다.

2.3.3 정보 이론적 관점에서 본 **VLM** 목표

핵심 개념:

- VLM의 학습 목표를 정보 이론의 관점에서 해석합니다.
- 모델은 **정보율(rate)**과 왜곡**(distortion)** 간의 균형을 최적화하여 데이터를 효율적으로 표현하고 예측 성능을 극대화합니다.

수식**:**

1. 목표 최적화:
    - 데이터 X와 해당 표현 Z 간 관계를 다음과 같이 정의합니다:
2. $\arg\min_{p(z|x)} I(f(X);Z) + \beta \cdot H(X|Z)$
    - $I(f(X);Z)$: 입력 데이터 XX에서 표현 ZZ로 전달되는 정보의 양.
    - $H(X|Z)$: ZZ를 통해 복원할 때 발생하는 왜곡.
    - β: 정보율과 왜곡 간의 균형을 조정하는 상수.
3. 구체적인 손실 함수:
    - 위 수식을 근사하여 다음 손실 함수로 변환:
4. $L = -\sum_{x \in D} \mathbb{E}_{p(f)p(Z|f(x))} [\log q(z) + \beta \cdot \log q(x|z)]$
    - $\log q(z)$: 엔트로피 병목(entropy bottleneck)으로 불필요한 정보를 제거.
    - $\log q(x|z)$: 복원 손실로 정보 왜곡을 줄이고 예측 성능을 향상.

해석:

- **대조 손실(Contrastive Loss)**와 **오토인코딩 손실(Auto-Encoding Loss)**은 왜곡(distortion)을 구현하는 방식으로 해석됩니다.
- 대조 손실은 데이터 복원 없이 정보를 압축하는 반면, 마스킹 및 오토인코딩 손실은 복원을 통해 예측 정보를 극대화합니다.

2.4 생성 기반 **VLMs (Generative-Based VLMs)**

생성 기반 학습은 텍스트나 이미지를 생성할 수 있도록 모델을 학습시키며, 다음과 같은 주요 사례가 있습니다:

1. **CoCa**:
    - 텍스트 생성과 대조 손실을 결합하여 학습.
    - 주어진 이미지에서 캡션을 생성하고, 추가적인 멀티모달 이해 작업을 수행할 수 있음.
2. **CM3Leon**:
    - 텍스트-이미지 및 이미지-텍스트 생성이 가능한 모델.
    - 토큰화된 텍스트와 이미지를 단일 트랜스포머 아키텍처로 처리.
3. **Stable Diffusion** 및 **Parti**:
    - 텍스트를 조건으로 이미지를 생성하는 데 주력하지만, 이미지-텍스트 이해 작업에도 응용 가능.

아래는 **2.4.3 Using Generative Text-to-Image Models for Downstream Vision-Language Tasks** 전체 내용을 수식과 함께 번역하고 설명한 내용입니다.

2.4.3 Using Generative Text-to-Image Models for Downstream Vision-Language Tasks

핵심 개념:

- 생성 기반 텍스트-이미지 모델은 원래 텍스트 설명으로 이미지를 생성하는 데 사용되지만, 이를 다운스트림 비전-언어 작업(예: 분류, 캡션 생성)에 활용할 수 있습니다.
- 이러한 접근 방식은 **Bayes** 정리를 통해 모델의 조건부 확률을 활용하여 분류 작업을 수행합니다.

조건부 우도 계산**(Likelihood Estimation)**

생성 모델은 이미지 x와 텍스트 조건 c에 대해 조건부 우도 p_\theta(x | c)를 추정하며, 이를 통해 분류 작업을 수행할 수 있습니다.

1. 조건부 확률 계산:
    - n개의 텍스트 클래스 $\{c_i\}_{i=1}^n$가 주어졌을 때, 클래스 c_i에 대한 확률은 다음과 같이 계산됩니다:
2. (5) $$p_\theta(c_i|x) = \frac{p(c_i) \cdot p_\theta(x|c_i)}{\sum_{j=1}^n p(c_j) \cdot p_\theta(x|c_j)}$$
    - $p_\theta(x|c_i)$: 모델이 조건 cic_i에서 이미지 xx를 생성할 확률.
    - $p(c_i)$: 클래스 c_i의 사전 확률.

Likelihood Estimation with Autoregressive Models

오토회귀 모델은 다른 모달리티(예: 텍스트, 음성)에서 토큰화된 데이터에 기반하여 작동하며, 이미지를 연속적인 토큰 $(t_1, \ldots, t_K)$으로 변환합니다.

수식**:**

1. 오토회귀 모델의 우도**(likelihood)**:
    - 이미지 x의 우도를 계산:
2. $$\log p_\theta(x|c_i) = \sum_{j=1}^K \log p_\theta(t_j|t_{<j}, c_i)$$
    - t_j: jj번째 토큰.

    - $t_{<j}$: jj번째 이전의 모든 토큰.
    - c_i: 조건으로 주어진 텍스트(예: 클래스 레이블).
3. 이미지 토큰화:
    - 이미지는 연속적인 디지털 토큰으로 변환되며, 이를 위해 벡터 양자화-변분 오토인코더(VQ-VAE) 프레임워크를 사용.
    - VQ-VAE는 이미지 압축 표현을 생성하고 이를 이산적 토큰으로 매핑.

VQ-VAE 기반 손실 함수**:**

1. 픽셀 공간 재구성 손실:
   $L_{\text{recon}} = \|x - \hat{x}\|^2$
    - xx: 입력 이미지, $\hat{x}$: 재구성된 이미지.
2. 코드북(embedding table) 손실:
    - 코드북 임베딩과 인코더 출력 간의 거리를 최소화.

장점:

- 오토회귀 방식은 이산적 토큰에 대한 직접적인 모델링이 가능하며, 텍스트와 이미지 간의 관계를 세밀하게 캡처합니다.

Likelihood Estimation with Diffusion Models

확산 모델은 노이즈를 점진적으로 제거하면서 이미지를 생성하며, 이를 통해 조건부 우도를 간접적으로 추정합니다.

수식**:**

1. 우도 하한**(Lower Bound)**:
    - 확산 모델에서 조건부 이미지 우도를 다음과 같이 정의:
2. $\log p_\theta(x|c_i) \propto -\mathbb{E}_{t,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, c_i)\|^2\right]$
    - x_t: 노이즈가 추가된 이미지.
    - ϵ: 실제 노이즈.
    - $\epsilon_\theta(x_t, c_i)$: 모델이 예측한 노이즈.
3. **Monte Carlo** 샘플링:
    - 우도 계산은 여러 번의 샘플링을 통해 Monte Carlo 방식으로 추정.
    - 계산 효율성을 위해 클래스별로 샘플을 동적으로 할당하거나 노이즈 매칭 기법을 사용.

장점:

- 확산 모델은 고급 이미지 표현을 학습하며, 이미지 생성 작업뿐 아니라 분류 작업에서도 경쟁력을 가집니다.
- 노이즈 예측 오차가 적을수록 우도가 높아지며, 이는 이미지-텍스트 일치도를 향상시킵니다.

제한점:

- 계산 비용이 매우 높으며, 클래스 수가 많을수록 계산 복잡도가 증가.
- 최적화를 위해 추가적인 알고리즘 개선이 필요.

Advantages of Generative Classifiers

생성 기반 분류기의 주요 장점은 다음과 같습니다:

1. **Out-of-Distribution** 성능:
    - 분포 밖 데이터에 대한 더 나은 강건성(effective robustness)을 제공합니다.
    - 동일한 분포 내 정확도에서 더 우수한 일반화 성능을 보입니다.
2. 조합적 추론 성능:
    - Winoground와 같은 조합적 추론(compositional reasoning) 작업에서 차별적인 성능을 보여줍니다.
    - 생성 모델은 더 높은 형태적 편향(shape bias)을 가지며, 인간의 판단과 더 잘 일치.
3. 테스트 단계에서의 적응:
    - 라벨이 없는 시험 데이터를 사용해 생성 분류기를 조정하여 성능을 개선 가능.
    - 분류, 세분화(segmentation), 깊이 예측(depth prediction)과 같은 다양한 작업에서 유연하게 사용 가능.
4. 협력 학습:
    - 생성 분류기와 판별 분류기를 결합하여 더욱 강력한 모델을 구축 가능.

2.5 사전 학습된 백본 활용 VLMs (VLMs from Pretrained Backbones)

사전 학습된 백본(Backbone)을 활용한 VLM 접근 방식은 기존의 강력한 언어 또는 비전 모델을 활용하여 효율적으로 비전-언어 모델을 구축하려는 아이디어에서 출발했습니다. 이 접근법은 사전 학습된 모델의 능력을 최대한 활용하면서, 텍스트와 이미지 간의 매핑을 추가로 학습하는 방식으로 이루어집니다. 대표적인 예는 **Frozen**, **MiniGPT**, 그리고 다양한 하이브리드 모델들입니다.

2.5.1 Frozen

Frozen 모델은 사전 학습된 대규모 언어 모델(LLM)을 활용하여 고정된(frozen) 상태로 유지하면서 이미지와 언어 간 매핑을 수행하는 방식으로 동작합니다.

1. 핵심 아이디어:
    - 언어 모델은 고정 상태로 유지되며, 별도의 학습이 필요 없습니다.
    - 이미지를 처리하기 위한 인코더와 언어 모델 사이의 연결 매핑 네트워크를 훈련합니다.
2. 설계:
    - 이미지 인코더: 이미지 데이터를 처리하여 임베딩으로 변환.
    - 매핑 네트워크: 이미지 임베딩과 언어 모델의 입력 공간을 정렬.
    - 언어 모델: 기존의 사전 학습된 언어 모델(예: GPT 계열)로, 텍스트 생성을 담당.
3. 장점:

    - 사전 학습된 언어 모델의 언어적 지식을 유지.
    - 계산 효율성 증가: 언어 모델을 다시 학습하지 않아도 됨.
    - 텍스트 기반 다운스트림 작업에서 우수한 성능.
4. 제한점:
    - 매핑 네트워크의 설계에 따라 성능이 크게 좌우됨.
    - 이미지를 언어 표현으로 변환하는 데 추가적인 정밀 조정이 필요.

2.5.2 MiniGPT

MiniGPT는 기존의 대규모 언어 모델(LLM)과 사전 학습된 이미지 인코더 사이를 간단한 매핑 네트워크로 연결하여 효율적으로 텍스트-이미지 작업을 수행하는 모델입니다.

1. 핵심 아이디어:
    - 사전 학습된 CLIP 이미지 인코더와 LLM을 간단한 매핑 네트워크로 연결하여 효율적인 학습을 수행.
    - 대규모 데이터에 대한 추가 학습 없이 뛰어난 성능을 달성.
2. 구조:
    - **CLIP** 이미지 인코더: 이미지 데이터를 벡터로 변환.
    - 매핑 네트워크: 간단한 선형 계층으로 구성되어 CLIP 임베딩을 LLM 입력 공간으로 변환.
    - 언어 모델: LLM은 텍스트 생성 및 추론 작업을 수행.
3. 학습 전략:
    - 매핑 네트워크만 훈련하므로 계산 비용이 낮음.
    - LLM 및 CLIP은 고정 상태로 유지.
4. 장점:
    - 매우 적은 계산 비용으로 강력한 비전-언어 능력을 발휘.
    - 다중 모달 입력(텍스트 + 이미지) 처리 가능.
    - 실시간 추론 작업에 적합.
5. 응용 사례:
    - 이미지 캡션 생성, 시각적 질문 응답(VQA), 텍스트 조건부 이미지 생성 등.

2.5.3 Other Popular Models Using Pretrained Backbones

Qwen. MiniGPT-4와 유사하게, Qwen-VL 및 Qwen-VL-Chat [Bai et al., 2023b] 모델은 LLM, 시각적 인코더, 그리고 시각적 표현을 LLM의 입력 공간에 맞추는 메커니즘을 사용합니다. Qwen에서 LLM은 Qwen-7B [Bai et al., 2023a]에서 초기화되고, 시각적 인코더는 ViT-bigG를 기반으로 하며, 1-layer 교차 주의 모듈을 사용하여 시각적 표현을 고정된 길이(256)의 시퀀스로 압축하여 LLM에 입력합니다.

BLIP2. Li et al. [2023e]는 이미지를 입력으로 받아 텍스트 출력을 생성하는 비전-언어 모델인 BLIP-2를 소개합니다. BLIP-2는 사전 훈련된 고정 모델을 활용하여 훈련 시간을 크게 단축합니다. CLIP와 같은 비전 인코더는 이미지 임베딩을 생성하고, 이는 OPT와 같은 LLM의 입력 공간에 매핑됩니다. Q-Former라고 불리는 비교적 작은(～100-200M 매개변수) 구성 요소가 이 매핑을 위해 훈련됩니다. Q-Former는 무작위로 초기화된 고정된 수의 "쿼리" 벡터를 입력으로 받는 Transformer입니다. 순방향 패스에서 쿼리는 Q-Former에서 교차 주의를 통해 이미지 임베딩과 상호 작용한 다음, 쿼리를 LLM의 입력 공간에 투영하는 선형 계층을 거칩니다. 문헌에는 사전 훈련된 LLM을 기반으로 하는 더 많은 모델이 있습니다. 각 LLM은 결국 VLM 버전으로 확장되므로, 이러한 주제에 대한 특정 조사의 범위는 매우 넓어질 것입니다. 이 소개에서는 모두 표현 간의 매핑 학습이라는 동일한 원칙에 의존하므로 몇 가지 선택된 모델을 제시하는 것을 목표로 합니다.

---

3. A Guide to VLM Training

이 섹션에서는 비전-언어 모델(VLM)을 훈련하기 위한 전반적인 가이드라인을 제시합니다. 여기에는 적절한 데이터셋 선택, 데이터 품질 관리, 훈련 전략, 최적화 기술 등이 포함됩니다. 연구 목표에 따라 어떤 데이터를 선택하고 어떤 훈련 전략을 사용할지에 대한 통찰을 제공합니다.

3.1 Training Data

훈련 데이터는 VLM(비전-언어 모델)의 성능에 결정적인 역할을 합니다. 이 섹션에서는 데이터 품질을 보장하고 모델의 학습을 최적화하기 위해 필요한 주요 방법론을 다룹니다.

3.1.1 Data Diversity and Balance

- 다양성의 중요성:
  - 데이터 세트는 다양한 개념, 주제, 언어, 시각적 특성을 포함해야 합니다.
  - 다양한 데이터를 학습함으로써 모델은 제로샷 학습(zero-shot learning) 능력을 향상시킬 수 있습니다.
- 균형 유지:
  - 특정 클래스나 주제가 과도하게 포함되지 않도록 데이터 균형을 유지하는 것이 중요합니다.
  - 예: ImageNet 기반 데이터 세트는 개체 분류에 강점을 보이며, COCO 데이터 세트는 복잡한 장면 설명에 적합합니다.

3.1.2 Data Curation

- 저품질 데이터 제거:
  - 텍스트-이미지 쌍의 정렬 품질이 낮은 데이터는 제거해야 합니다.
  - 잘못된 텍스트 설명이나 이미지 간 불일치는 모델 학습을 방해합니다.
- 정렬 품질 평가:
  - 데이터의 정렬 품질을 평가하기 위해 **CLIPScore**와 같은 정량적 메트릭을 사용합니다.

- **CLIPScore**는 텍스트와 이미지의 유사성을 점수로 나타내는 도구입니다.

3.1.3 Synthetic Data

- 합성 데이터 생성:
  - 학습에 필요한 텍스트-이미지 쌍을 직접 생성함으로써 데이터 다양성을 증가시킬 수 있습니다.
  - 예: BLIP(Bootstrapping Language-Image Pretraining) 모델은 텍스트 생성 모델을 사용해 이미지 설명을 자동 생성합니다.
- 장점:
  - 합성 데이터는 특정 도메인이나 언어에서 부족한 데이터를 보완할 수 있습니다.
  - 실시간 생성된 데이터는 새로운 상황에 빠르게 적응하는 데 유용합니다.

3.1.4 Assessing Multimodal Data Quality

- 멀티모달 데이터 품질 평가:
  - 텍스트-이미지 쌍의 품질을 평가하는 것은 매우 중요합니다.
  - 데이터 품질이 낮을 경우, 모델의 학습 과정에서 편향이나 성능 저하를 초래할 수 있습니다.
- 평가 메트릭:
  - **CLIPScore**: 텍스트와 이미지의 정렬 품질을 측정.
  - **Human Evaluation**: 사람의 판단을 통해 텍스트와 이미지의 의미적 일치를 평가.
- 자동 평가 방법:
  - **Language Understanding-based Scoring**: 언어 모델을 활용하여 텍스트 설명의 일관성과 문법적 정확성을 평가.
  - **Visual Content-based Scoring**: 이미지 콘텐츠와 텍스트 설명의 일치를 평가하는 알고리즘적 접근법.

3.1.5 Harnessing Human Expertise: The Power of Data Annotation

- 데이터 라벨링의 중요성:
  - 사람의 전문성을 활용해 데이터 라벨링 품질을 보장할 수 있습니다.
  - 텍스트-이미지 쌍의 품질 향상뿐만 아니라, 모델 학습의 신뢰성도 높아집니다.
- 라벨링 방법:
  - **Direct Annotation (**직접 라벨링**)**:
    - 사람 annotator가 직접 이미지와 텍스트를 일치시키거나 설명을 추가합니다.
  - **Validation and Correction (**검증 및 수정**)**:
    - 자동 생성된 텍스트-이미지 쌍을 사람 annotator가 검토하여 품질을 보장합니다.
  - **Iterative Feedback Loops (**반복 피드백 루프**)**:
    - 모델의 예측 결과를 사람이 검토하고 수정함으로써, 데이터 품질과 모델 성능을 점진적으로 개선합니다.
- 장점:
  - 특정 도메인에서의 데이터 품질 보장을 위해 필수적.
  - 멀티모달 학습에서 언어적/시각적 정밀성을 향상.

요약 및 표

표**: 3.1 Training Data**의 주요 전략

전략	설명	장점
데이터 다양성과 균형	다양한 주제와 언어를 포함하여 데이터 세트를 구성.	모델의 제로샷 학습 능력 향상.
데이터 큐레이션	정렬 품질이 낮은 데이터를 제거.	학습 데이터의 신뢰성 및 정확성 증가.
합성 데이터 생성	텍스트-이미지 쌍을 자동으로 생성하여 부족한 데이터를 보완.	특정 도메인의 데이터 부족 문제 해결.
멀티모달 데이터 품질 평가	CLIPScore와 사람 검토를 통해 데이터 품질을 평가.	데이터 신뢰도 및 정렬 품질 개선.
사람의 전문성 활용	사람이 직접 라벨링하거나, 자동 생성 데이터를 검토 및 수정.	도메인 특화 데이터 세트 구축 및 모델 성능 최적화.

3.2 Software

이 섹션에서는 VLM(비전-언어 모델) 훈련과 평가를 위해 사용할 수 있는 주요 소프트웨어와 관련 기술을 다룹니다. 또한, 훈련에 필요한 자원과 최적화 방법에 대한 논의도 포함됩니다.

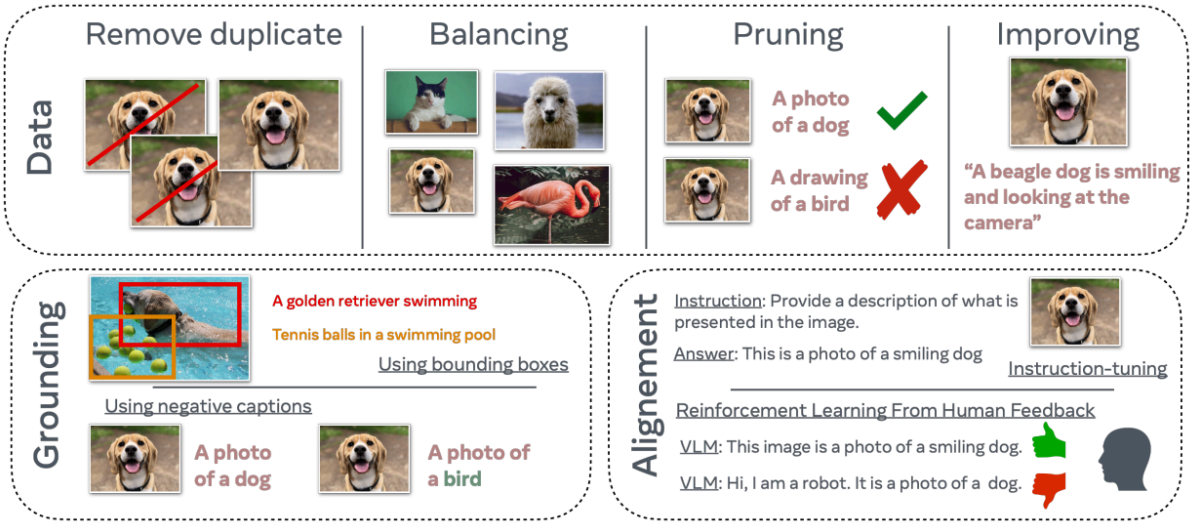


Figure 2: VLM(비전**-**언어 모델**)** 훈련 시 고려해야 할 중요한 요소는 모델의 성능과 학습 효율성을 극대화하기 위해 중요한 데이터 처리 및 정렬 전략을 강조합니다. 아래는 그림의 내용을 기반으로 한 주요 설명입니다:

데이터 관련 고려사항

1. 데이터 다양성과 균형**(Diversity and Balance)**:
    - 훈련 데이터는 다양한 개념, 주제, 그리고 언어를 포함해야 하며, 이를 통해 모델이 폭넓은 세계 모델(world model)을 학습할 수 있습니다.
    - 데이터의 균형을 유지해 특정 클래스 또는 주제에 치우치지 않도록 해야 합니다.
2. 중복 제거**(Removing Duplicates)**:
    - 대규모 데이터 세트에서는 중복된 항목이 자주 발생합니다.
    - 중복 데이터를 제거하면 계산 시간을 절약하고, 모델이 불필요한 내용을 암기(memorizaton)하는 문제를 줄일 수 있습니다.
3. 데이터 가지치기**(Data Pruning)**:
    - 데이터 가지치기는 이미지와 캡션이 실제로 관련이 있는지 확인하는 작업입니다.
    - 캡션이 이미지 콘텐츠와 일치하지 않을 경우, 이를 제거해야 합니다.
4. 캡션 품질 개선**(Caption Quality Improvement)**:
    - 훈련 데이터의 캡션 품질을 향상시키는 것은 모델 성능을 높이는 데 필수적입니다.
    - 더 정확하고 풍부한 의미를 가진 캡션은 모델의 학습 효과를 크게 증가시킵니다.

VLM의 그라운딩**(Grounding VLMs)**

- **그라운딩(Grounding)**은 모델이 특정 단어를 특정 개념과 올바르게 연관짓도록 하는 과정입니다.
- 대표적인 방법:
    1. **Bounding Boxes**:
        - 이미지 내 개체를 나타내는 경계 상자를 활용해 단어와 시각적 개체 간의 관계를 학습.
    2. **Negative Captions**:
        - 이미지와 관련 없는 캡션을 제공해 잘못된 연관성을 학습하지 않도록 유도.

정렬**(Alignment)**

- 정렬의 중요성:
    - 모델이 인간의 기대에 부합하는 답변을 생성하도록 만드는 과정입니다.
    - 정렬은 모델이 생성하는 응답이 사용자 관점에서 이해 가능하고 의미 있게 보이도록 보장합니다.

요약

Figure 2는 다음과 같은 메시지를 전달합니다:

- 데이터 품질과 구조는 모델 성능에 핵심적이다. 중복 제거, 캡션 품질 개선, 데이터 가지치기를 통해 효율성을 높이고 암기를 방지한다.
- 그라운딩과 정렬은 필수 단계로, 모델이 인간과 상호작용 시 올바른 개념을 이해하고 적절한 응답을 제공하도록 돕는다.

3.2.1 Using Existing Public Software Repositories

- 주요 소프트웨어:

    - OpenCLIP: VLM 구현을 지원하는 오픈 소스 프로젝트로, CLIP 기반 모델 훈련과 비교를 위한 플랫폼 제공.
    - Hugging Face Transformers: 다양한 미리 학습된 VLM과 언어 모델을 제공하며, 손쉽게 사용 가능.
- 활용 예시:
    - 사용자는 사전 학습된 모델을 활용하거나, 주어진 다운스트림 작업에서 서로 다른 VLM 성능을 비교 가능.

3.2.2 How Many GPUs Do I Need?

- 훈련에 필요한 자원:
    - CLIP와 같은 대규모 모델 훈련에는 500개 이상의 GPU가 필요했으며, 이는 공공 클라우드에서 수십만 달러의 비용이 소요됨.
- 효율적인 접근법:
    - 고품질 데이터 세트와 마스킹 전략을 결합하면, 64 GPU만으로 수백만 이미지로 구성된 CLIP와 같은 모델을 훈련 가능.
    - 사전 학습된 이미지 또는 텍스트 인코더를 사용하는 경우, 훈련 비용이 더욱 절감됨.

3.2.3 Speeding Up Training

- 최적화 기법:
    - PyTorch의 **torch.compile** 기능과 **xformers** 라이브러리를 통해 훈련 속도를 개선.
- 데이터 로딩 병목 문제:
    - 대규모 미니배치 로드 과정이 훈련 속도를 저하시킬 수 있음.
    - **FFCV(Fast Forward Computer Vision)**: 데이터를 최적화된 형식으로 저장하여 로딩 속도를 대폭 향상.
    - 압축되지 않은 파일 형식을 사용하는 것이 훈련 속도에는 유리하지만, 스토리지 비용 증가 가능.

3.2.4 Importance of Other Hyper-parameters

- 하이퍼파라미터 중요성:
    - 이미지 해상도와 비전 인코더 용량, 시각적 사전 학습 데이터가 모델 성능에 가장 큰 영향을 미침.
    - 데이터 형식(텍스트-이미지 쌍, 텍스트 전용 데이터 등)의 적절한 조합이 제로샷 분류와 시각적 질문 응답(VQA)에서 최적의 성능을 제공.

3.3 Which Model to Use?

이 섹션에서는 다양한 VLM(비전-언어 모델) 훈련 방법 중에서 특정 상황에 적합한 모델을 선택하는 방법을 다룹니다. 모델 선택은 훈련 목표, 데이터 규모, 계산 자원에 따라 달라집니다. 주요 접근 방식으로는 대조 학습**(Contrastive Learning)**, 마스킹**(Masking)**, 생성 모델**(Generative Models)**, 사전 학습된 백본 기반 모델이 있습니다.

3.3.1 When to Use Contrastive Models Like CLIP

- 특징:
    - CLIP(Contrastive Language-Image Pretraining)은 텍스트와 이미지 표현을 대조 학습을 통해 정렬합니다.
    - 단순한 학습 파라다임으로, 텍스트와 이미지 쌍을 임베딩 공간에서 일치하도록 훈련.
    - 텍스트 설명으로부터 관련 이미지를 검색하거나 반대로 이미지를 통해 텍스트를 검색하는 작업에 적합.
- 장점:
    - 대규모 데이터 세트에서 학습 시 강력한 표현 학습 가능.
    - 제로샷 학습과 검색 작업에서 우수한 성능.
    - 데이터 큐레이션 파이프라인(예: MetaCLIP)을 통해 텍스트-이미지 정렬 데이터 생성 가능.
- 제한점:
    - CLIP은 생성 모델이 아니므로 특정 이미지에 대한 텍스트 캡션 생성 불가.
    - 대규모 데이터 세트와 큰 미니배치가 필요하여 계산 자원이 많이 소모됨.

3.3.2 When to Use Masking

- 특징:
    - 마스킹은 텍스트 또는 이미지 일부를 가리고 이를 복원하도록 학습합니다.
    - 대조 모델과 달리, 입력 데이터 공간으로 복원하기 위한 디코더가 필요.
- 장점:
    - 배치 종속성이 없으므로 작은 미니배치를 사용할 수 있음.
    - 부정적인 샘플이 필요 없기 때문에 대조 학습보다 하이퍼파라미터 튜닝이 단순화됨.
- 제한점:
    - 추가 디코더가 필요하여 계산 비용이 증가할 수 있음.

3.3.3 When to Use a Generative Model

- 특징:
    - 생성 모델(예: 확산 모델 또는 오토회귀 모델)은 텍스트 프롬프트로 포토리얼리스틱 이미지를 생성 가능.
    - 텍스트-이미지 간의 암묵적 결합 분포를 학습.
- 장점:
    - 모델이 학습한 표현을 입력 데이터 공간으로 디코딩할 수 있어 직관적인 이해 가능.
    - CLIP과 달리, 생성 모델은 입력 데이터 공간에서 직접 결과를 출력 가능.
    - 텍스트와 이미지 간의 암묵적 결합 분포를 학습하여 보다 강력한 표현 생성.
- 제한점:
    - 대조 학습 모델보다 계산 비용이 더 높음.
    - 학습 및 추론 시간이 상대적으로 오래 걸릴 수 있음.

3.3.4 When to Use LLM on Pretrained Backbone

- 특징:
    - 사전 학습된 LLM(예: GPT, LLaMA)과 시각적 인코더를 결합하여 학습.
    - LLM과 시각적 인코더 간의 매핑만 학습하므로, 전체 모델을 처음부터 학습하는 것보다 효율적.
- 장점:
    - Flamingo, MiniGPT 등과 같은 모델은 사전 학습된 LLM을 활용해 계산 비용을 절감.
    - 기존의 강력한 텍스트 모델과 시각적 데이터를 결합하여 멀티모달 성능 향상.
- 제한점:
    - LLM 및 시각적 인코더의 사전 학습 품질에 크게 의존.

3.4 Improving Grounding

Grounding은 VLM(비전-언어 모델) 및 생성 모델 연구에서 중요한 도전 과제입니다. 이 과정은 모델이 텍스트 프롬프트를 제대로 이해하지 못하거나, 프롬프트에 없는 내용을 상상(환각)하여 생성하는 문제를 해결하는 데 초점을 맞춥니다. 대표적인 문제는 다음과 같습니다:

- 텍스트로 표현된 관계(예: 객체가 왼쪽에 있는지 오른쪽에 있는지)의 이해.
- 부정어, 카운팅, 속성(색상, 텍스처 등)의 인지.

이 섹션에서는 이러한 문제를 해결하기 위해 일반적으로 사용되는 방법을 소개합니다.

3.4.1 Using Bounding Boxes Annotations

- 활용 방법:
    - X-VLM [Zeng et al., 2022]은 바운딩 박스 주석을 활용하여 객체를 정확히 찾고 시각적 개념을 텍스트와 정렬합니다.
    - 바운딩 박스 회귀(regression)와 IoU(Intersection over Union) 손실을 통합하여 시각적 단서를 텍스트 설명과 일치시키는 능력을 개선합니다.
- 훈련 데이터:
    - COCO, Visual Genome, SBU, Conceptual Captions 데이터 세트를 포함해 약 1600만 개의 이미지로 구성된 데이터셋에서 훈련.
    - 이러한 대규모 데이터셋은 이미지-텍스트 검색, 시각적 추론, 시각적 정렬, 이미지 캡션 생성 등의 작업에서 X-VLM의 성능을 향상.
- 대체 접근법:
    - Kosmos-2 [Peng et al., 2024]는 공개 모델을 사용하여 자체 이미지-텍스트 데이터셋을 생성합니다.
    - 예: spaCy를 사용해 텍스트 캡션에서 명사를 추출하고 GLIP 모델을 활용해 해당 명사에 연결된 바운딩 박스를 예측.
    - 제한점: 바운딩 박스 탐지를 위한 기본 모델이 드문 명사나 사례에서 실패하면 다운스트림 모델도 유사한 실수를 할 가능성이 있음.

3.4.2 Negative Captioning

- 역할:

    - 대조 학습 목표에서 부정 샘플은 모델의 붕괴(collapse)를 방지하고 일반화 성능 및 판별적 특징 학습을 향상하는 데 널리 사용됨.
    - 부정 샘플(비슷하거나 관련 없는 샘플)을 활용해 모델이 데이터의 미묘한 차이를 이해하도록 강제.
- 적용 사례:
    - ARO 벤치마크 [Yuksekgonul et al., 2023]는 모델이 이미지와 캡션을 올바르게 연결할 수 있는 능력을 평가.
    - 부정 샘플을 사용하여 모델이 잘못된 또는 비논리적인 쌍을 이해하도록 테스트.
- 효과:
    - 부정 샘플에 노출된 모델은 더 정교한 차별화 능력을 개발해 보다 정확하고 상황을 잘 이해하는 모델 생성 가능.

3.5 Improving Alignment

이 섹션은 VLM(비전-언어 모델)의 **멀티모달 정렬(alignment)**을 개선하기 위한 방법론을 다룹니다. 정렬 개선은 모델 출력이 인간의 기대와 더 잘 맞도록 하며, 특히 멀티모달 대화 및 과제 수행에서 중요한 역할을 합니다.

3.5.1 A LLaVA Story

- **LLaVA (Large Language and Vision Assistant)**:
    - 언어 도메인에서의 명령어 튜닝**(Instruction Tuning)** 성공에서 영감을 받아 멀티모달 대화 능력을 개선하기 위해 LLaVA가 개발되었습니다.
    - 훈련 데이터:
        - 약 15만 개의 합성 시각적 명령어 데이터를 생성하여 훈련.
        - LLaVA는 **Vicuna** 언어 모델과 **CLIP ViT-L/14** 비전 인코더를 결합.
        - 출력은 동일한 차원 공간으로 통합하여 정렬.
    - 결과:
        - LLaVA는 질적 대화와 과학 질문 응답(SciQA) 벤치마크에서 개선된 성능을 보였습니다.
- **LLaVA 1.5**:
    - LLaVA의 개선된 버전으로, 교차 모달 완전 연결 **MLP** 레이어와 학술적 VQA 데이터를 활용하여 훈련.
    - 60만 개의 이미지-텍스트 쌍으로 훈련하며, 8개의 A100 GPU에서 약 하루만에 훈련 가능.
    - 학술 VQA 및 명령어 수행 벤치마크에서 우수한 성능을 발휘.
- **LLaVA-RLHF**:
    - 시각적 명령어 튜닝 데이터가 부족한 문제를 해결하기 위해 강화 학습을 활용한 인간 피드백**(RLHF)** 알고리즘을 사용.
    - **Factually Augmented RLHF**를 통해 보상 모델을 보강하고, 이미지 캡션 및 정답을 활용해 보상 해킹(reward hacking)을 방지.
    - 결과:
        - GPT-4 성능의 94%에 해당하는 결과를 기록.
        - 특히 환각(hallucination) 감소에 중점을 둔 MMHAL-BENCH에서 60% 향상된 성능을 보임.
- **LLaVA-NeXT (v1.6)**
    - LLaVA-NeXT는 **LLaVA-v1.5**에서 다음과 같은 여러 방면에서 개선된 최신 버전입니다【51:0†source】.
    - 주요 개선 사항

    - 이미지 해상도 증가:
        - 전체 이미지와 작은 이미지 패치의 시각적 특징을 결합하여, 각각을 비전 인코더에 독립적으로 입력.
        - 이를 통해 모델의 시각적 표현력을 향상시킴.
    - 시각적 명령어 튜닝 데이터 개선:
        - 기존 데이터셋에 다음과 같은 새로운 예제를 추가:
            - 시각적 추론(Visual Reasoning)
            - OCR(광학 문자 인식)
            - 세계 지식(World Knowledge)
            - 논리적 추론(Logical Reasoning)
    - 더 큰 **LLM** 백본 사용:
        - 가장 큰 모델 변형은 **Nous-Hermes-2-Yi-34B**라는 34B 파라미터의 대규모 언어 모델(LLM)을 백본으로 사용.

3.5.2 Multimodal In-Context Learning

- 정의:
    - 멀티모달 인-컨텍스트 학습은 다양한 예제와 명령어를 활용하여 모델이 특정 작업을 수행하도록 가이드합니다.
    - 모델은 명시적으로 훈련되지 않은 새로운 작업에서도 잘 일반화될 수 있음.
- 활용:
    - **OpenFlamingo**:
        - 멀티모달 인-컨텍스트 학습을 활용하여 텍스트와 이미지의 관계를 이해.
        - 예제 기반 학습을 통해 다양한 다운스트림 작업 수행 가능.
    - **InstructBLIP**:
        - 사용자 명령을 더 잘 수행하기 위해 명령어 튜닝을 통합.
- 장점:
    - 모델이 미리 훈련된 데이터에서 벗어나도 유연한 학습 능력을 제공.
    - 다양한 사용 사례에서 멀티모달 기능을 확장.

3.6 Improving Text-Rich Image Understanding

텍스트 중심 이미지 이해의 중요성

텍스트를 이해하는 능력은 시각적 인식의 중요한 측면입니다. **멀티모달 대형 언어 모델(Multimodal Large Language Models, MLLMs)**의 성공은 이러한 능력을 활용하여 제로샷 작업을 다양한 실제 시나리오에 적용하는 가능성을 열었습니다. 예를 들어, **Liu et al. (2023e)**는 MLLMs이 OCR(광학 문자 인식) 관련 도메인에서 명시적으로 훈련하지 않았음에도 불구하고, "제로샷" 성능에서 우수함을 보였다고 보고합니다. 그러나 많은 모델이 이미지 내 텍스트를 이해하는 데 있어 복잡한 관계나 데이터 유형 간 상호작용에서 어려움을 겪는 경우가 있습니다. 이는 주로 **Conceptual Captions** 및 **COCO**와 같은 자연 이미지 중심의 훈련 데이터로 인해 발생합니다.

3.6.1 Instruction Tuning with Fine-Grained Text-Rich Data

LLaVAR (Zhang et al., 2023c)는 텍스트가 풍부한 이미지(예: 영화 포스터, 책 표지)를 사용하여 시각적 명령어 튜닝 파이프라인을 개선합니다.

- 데이터 수집:
    - 공개 OCR 도구를 사용해 LAION 데이터셋에서 42만 2천 개의 텍스트 중심 이미지를 처리.
    - 텍스트와 이미지 캡션을 추출한 후, **GPT-4**를 활용하여 각 이미지에 대해 16,000개의 대화 데이터를 생성.
    - 대화 데이터는 텍스트 중심 이미지에 대한 질문-응답 쌍을 포함.
- 결과:
    - LLaVAR 모델은 이전의 멀티모달 명령어 데이터와 결합하여 학습.
    - 텍스트 기반 VQA 데이터셋에서 최대 20% 정확도 향상.
    - 자연 이미지에서 약간의 성능 향상도 달성.

3.6.2 Dealing with Fine-Grained Text in High-Resolution Images

Monkey (Li et al., 2023h)는 고해상도 이미지에서의 텍스트 세부 사항을 다루기 위한 새로운 접근 방식을 제공합니다.

- 문제점:
    - 대부분의 멀티모달 LLM은 입력 이미지 크기가 224x224로 제한되어 텍스트 중심 작업(예: 장면 텍스트 중심 VQA, 문서 지향 VQA, 주요 정보 추출 등)에서 세부 사항을 추출하는 데 어려움을 겪음.
- 해결 방법:
    - 슬라이딩 윈도우**(sliding window)** 방식으로 입력 이미지를 패치로 나누어 처리.
    - 각 패치는 정적 비전 인코더에서 독립적으로 처리되며, **LoRA** 조정 및 훈련 가능한 비전 리샘플러로 강화.
    - 최대 1344x896 픽셀의 고해상도를 처리 가능.
- 결과:
    - 멀티레벨 설명 생성 방법을 활용하여 장면-객체 연관성을 풍부하게 만듦.
    - 텍스트 중심 캡션 생성과 장면의 복잡한 세부 사항 캡처를 위한 포괄적인 접근법을 제공.

Decoupled Scene Text Recognition Module and MM-LLM

Lumos: Decoupled Scene Text Recognition Module

- 개요: Lumos는 멀티모달 어시스턴트로, 텍스트 이해를 강화하기 위해 설계되었습니다. Lumos는 디바이스 상에서 실행 가능한 디커플드**(Scene Text Recognition, STR)** 모듈과 클라우드에서 작동하는 멀티모달 LLM 모듈을 결합합니다【59:0†source】.

- 구조 및 기능:
    - **STR** 모듈 구성 요소:
        - **ROI(Region of Interest) Detection**:
            - 시각적 관심 영역을 감지하고, STR 입력으로 해당 영역을 자릅니다.
        - **Text Detection**:
            - 자른 이미지를 입력으로 받아 단어를 감지하고, 각 단어의 바운딩 박스 좌표를 반환합니다.
        - **Text Recognition**:

        - ROI 검출된 이미지와 텍스트 감지 결과를 입력으로 받아 인식된 단어를 반환합니다.
    - **Reading-order Reconstruction**:
        - 인식된 단어를 레이아웃에 따라 읽기 순서로 정렬합니다.
- 클라우드 통합:
    - **LLM** 모듈: 클라우드에서 실행되며, STR 모듈에서 추출된 텍스트와 바운딩 박스 좌표를 활용합니다.
    - 이점:
        - 고해상도 이미지를 클라우드로 전송하지 않아도 되므로 전송 지연과 전력 소비를 줄일 수 있습니다.
        - Lumos의 STR 모듈은 최대 3k×4k 해상도의 이미지를 처리할 수 있어, 복잡한 텍스트 이해 작업에서 성능을 크게 향상합니다.

성능 및 응용

Lumos는 Monkey와 유사하게 복잡한 텍스트 중심 작업에서 향상된 성능을 제공합니다. 이 모듈은 OCR, 문서 중심 VQA(Document-Oriented VQA), 장면 텍스트 기반 VQA(Scene Text-Centric VQA)와 같은 작업에서 활용됩니다.

3.7 Parameter-Efficient Fine-Tuning

훈련된 VLM(비전-언어 모델)을 다운스트림 작업에 적응시키는 과정에서, 모델의 전체 매개변수를 세부 조정(fine-tuning)하는 것은 계산 비용과 자원 사용 측면에서 비효율적일 수 있습니다. 이를 해결하기 위해 **Parameter-Efficient Fine-Tuning (PEFT)** 방법론이 개발되었습니다. PEFT는 대규모 모델의 모든 매개변수를 조정하지 않고, 필요한 하위 집합만을 조정하여 계산 비용을 절감하고 효율성을 높입니다. 이 섹션에서는 주요 PEFT 접근 방식을 설명합니다.

주요 **PEFT** 접근 방식

1. **Low Rank Adapters (LoRA)** 기반 접근법
    - **LoRA** [Hu et al., 2022]:
        - 언어 및 비전-언어 모델에 널리 사용되는 매개변수 효율적인 조정 기법입니다.
        - LoRA는 기존 모델 매개변수를 동결한 상태에서 학습 가능한 저랭크(low-rank) 행렬만 추가로 학습합니다.
    - 확장된 변형:
        - **QLoRA** [Dettmers et al., 2023]:
            1. LoRA를 4비트 양자화된 백본과 통합하여, 동결된 모델에서도 그래디언트 역전파(back-propagation)가 가능.
        - **VeRA** [Kopiczko et al., 2024]:
            1. 모든 계층에서 공유되는 단일 저랭크 행렬 쌍을 사용하여 학습 가능한 매개변수를 줄이면서 성능 유지.
        - **DoRA** [Liu et al., 2024]:
            1. 사전 학습된 가중치를 크기와 방향으로 분해해 LoRA를 VLM 벤치마크에 일반화.
2. **Prompt** 기반 접근법
    - **Context Optimization (CoOp)** [Zhou et al., 2022]:
        - CLIP 같은 대규모 VLM을 다운스트림 작업에 맞게 조정하며, 학습 가능한 벡터를 통해 프롬프트를 최적화.
        - 두 가지 구현 방식:

        1. 통합 컨텍스트**(unified context)**: 모든 클래스에서 동일한 컨텍스트 사용.
        2. 클래스별 컨텍스트**(class-specific context)**: 각 클래스에 고유한 컨텍스트 사용.
    - **Visual Prompt Tuning (VPT)** [Jia et al., 2022]:
        - 비전 트랜스포머에서 입력 공간에 학습 가능한 파라미터를 도입하며, 모델 백본은 동결 상태로 유지.
3. **Adapter** 기반 접근법
    - **CLIP-Adapter**:
        - 시각적 또는 언어적 지점에서 피처 어댑터(feature adapter)를 활용해 조정.
        - 잔차 스타일의 피처 혼합을 통해 새로운 피처를 학습.
    - **VL-Adapter**:
        - 이미지-텍스트 및 비디오-텍스트 작업에서 멀티태스크 프레임워크 내에서 어댑터 기반 접근법을 평가.
        - 태스크 간 가중치 공유(weight-sharing)를 통해 효율성을 극대화.
4. **Mapping** 기반 접근법
    - 사전 학습된 모델에 학습 가능한 모듈을 주입하거나, LoRA를 사용하여 매개변수를 조정.
    - 간단한 매핑 학습만으로 다중 모달 작업 수행 가능.

---

4. Approaches for Responsible VLM Evaluation

이 섹션에서는 비전-언어 모델(VLMs)의 평가를 위한 다양한 접근 방식을 설명합니다. VLMs의 주요 능력은 텍스트와 이미지를 정렬하고 매핑하는 것입니다. 따라서 모델이 실제로 시각적 단서를 올바르게 이해하고 이를 텍스트와 매핑하는지를 평가하는 것이 중요합니다 .

그림 **3: VLM** 평가를 위한 다양한 방법**.** 시각적 질문 응답(VQA)은 가장 일반적인 방법 중 하나이지만, 모델과 정답을 정확한 문자열 매칭을 통해 비교하므로 모델 성능이 과소평가될 수 있습니다. 추론은 VLM에 캡션 목록을 제공하고 이 목록에서 가장 가능성 있는 캡션을 선택하도록 하는 것으로 구성됩니다. 이 범주에서 널리 사용되는 두 가지 벤치마크는 Winoground[Diwan et al., 2022] 및 ARO[Yuksekgonul et al., 2023]입니다. 최근에는 밀집된 인간 주석을 사용하여 모델이 캡션을 이미지의 올바른 부분에 얼마나 잘 매핑할 수 있는지 평가할 수 있습니다[Urbanek et al., 2023]. 마지막으로, PUG[Bordes et al., 2023]와 같은 합성 데이터를 사용하여 다양한 구성으로 이미지를 생성하여 특정 변형에 대한 VLM의 견고성을 평가할 수 있습니다.

4.1 Benchmarking Visio-Linguistic Abilities

VLM을 평가하는 첫 번째 방법은 **Visio-Linguistic Benchmarks**를 활용하는 것입니다. 이 벤치마크는 모델이 특정 단어나 구문을 시각적 단서와 연관 지을 수 있는지를 평가하도록 설계되었습니다.

평가 과제

- 단순한 시각적 단서(예: "이미지에 개가 보입니까?")에서부터 복잡한 장면(예: "이미지에 몇 마리의 개가 있으며, 무엇을 보고 있습니까?")까지 다양한 수준의 질문을 포함합니다.
- 기본적인 캡션에서 공간적 이해와 추론이 필요한 더 복잡한 캡션으로 발전하는 문제를 통해, 모델의 강점과 약점을 평가할 수 있습니다.

4.1.1 Image Captioning

- 소개:
    - COCO 캡셔닝 데이터셋(Chen et al., 2015)은 주어진 VLM이 생성한 캡션의 품질을 평가합니다.
    - 외부 평가 서버를 활용하여 모델의 캡션이 다음 기준을 충족하는지 확인:
        - 정확성**(Accuracy)**: 캡션이 이미지 내용을 정확히 설명하는가?
        - 유창성**(Fluency)**: 언어적으로 자연스러운가?
        - 정보의 적절성**(Relevance)**: 주요 시각적 정보를 반영하는가?
- 지표:
    - BLEU, METEOR, CIDEr, SPICE와 같은 다중 언어 평가 지표를 사용하여 생성된 캡션의 품질을 정량화.

4.1.2 Text-to-Image Consistency

- 평가 목적:
    - 텍스트 프롬프트가 제공될 때 생성된 이미지가 프롬프트 내용을 정확히 반영하는지 평가.
    - CLIPScore를 사용하여 생성된 이미지와 텍스트 프롬프트 간의 일치성을 점수화.
- 활용 사례:
    - 생성형 텍스트-이미지 모델(예: DALL-E, MidJourney)의 품질 평가.

4.1.3 Visual Question Answering (VQA)

- 개요:
    - VQA는 이미지에 대한 질문에 답변하는 모델의 능력을 평가.
    - Antol et al. (2015)에 의해 처음 소개된 이후, 여러 확장 벤치마크가 제안됨.
- 평가 방식:
    - Open-Ended VQA: 모델이 텍스트로 답변.

    - Multiple-Choice VQA: 주어진 선택지 중에서 정답 선택.
- 지표:
    - 정확도(Accuracy), 응답의 논리적 일관성.

4.1.4 Text-Centric Visual Question Answering

- 특징:
    - 텍스트가 포함된 이미지(예: 표지판, 문서)를 활용하여 모델의 OCR 능력과 텍스트 이해 능력을 평가.
- 활용 데이터:
    - TextVQA, DocVQA와 같은 데이터셋에서 모델 성능을 벤치마크.
- 지표:
    - 모델이 텍스트를 읽고 의미적으로 이해할 수 있는지 평가.

4.1.5 Zero-shot Image Classification

Zero-shot classification은 모델이 명시적으로 학습되지 않은 분류 작업에서 평가받는 방식을 의미합니다. 이는 모델이 특정 작업에 대한 학습 데이터 없이도 일반화된 성능을 보일 수 있는지를 확인합니다. Few-shot 학습과는 달리, Zero-shot은 다운스트림 작업의 미세 조정을 위해 적은 양의 학습 데이터를 필요로 하지 않습니다【71:0†source】.

CLIP을 활용한 **Zero-shot Classification**

- **Radford et al. (2021)**:
    - CLIP은 특정 작업에 맞는 프롬프트 구조를 변경함으로써 Zero-shot 분류 성능을 크게 개선할 수 있음을 보여주었습니다.
    - 예를 들어, "a photo of a {class}" 또는 "a demonstration of a {class}"와 같은 프롬프트를 활용하여 ImageNet 벤치마크에서 경쟁력 있는 성능을 달성.
    - 이는 VLM이 표준 분류 훈련과 경쟁할 수 있음을 처음으로 보여준 사례입니다.

평가 데이터셋

- CLIP은 ImageNet 외에도 다양한 추가 분류 데이터셋에서 평가됩니다:
    - CIFAR10/100
    - Caltech 101
    - Food101
    - CUB (Caltech-UCSD Birds)
    - Stanford Cars
    - EuroSAT
    - Flowers102

    - OxfordPets
    - FGVC-Aircraft
    - Pascal VOC.

프롬프트 엔지니어링

- 프롬프트 생성:
    - 개념 이름을 포함한 사람에 의해 설계된 프롬프트 템플릿은 Zero-shot 성능을 크게 향상시킬 수 있습니다.
    - 최근에는 ChatGPT와 같은 LLM을 사용하여 더 풍부한 시각적 설명을 포함한 프롬프트를 자동으로 생성하는 방법이 제안되었습니다.
    - 예: "a tiger, which has sharp claws"와 같은 설명.
- 최적화:
    - Parashar et al. (2024)는 "cash machine"을 "ATM"으로 대체하는 것처럼 가장 빈번히 사용되는 동의어를 채택하여 정확도를 향상.

Out-of-Distribution (OOD) Generalization

- CLIP의 Zero-shot 평가가 ImageNet과 같은 작업에서 높은 성능을 보이는 것은 훈련 데이터가 매우 크기 때문입니다.
- 그러나 CLIP의 훈련 분포가 다운스트림 작업과 크게 다를 경우 일반화 성능이 저하될 수 있습니다.
- **Samadh et al. (2023)**:
    - 테스트 예제의 토큰 분포를 ImageNet 데이터 분포와 맞추는 방법을 제안하여 OOD 벤치마크와 다운스트림 작업에서 성능을 향상.

4.1.6 Visio-Linguistic Compositional Reasoning

이 섹션에서는 VLM(Visio-Linguistic Model)이 언어와 이미지 간의 복잡한 관계를 이해하고 올바르게 추론할 수 있는지 평가하는 방법을 설명합니다. 특히, 언어의 구문적, 관계적, 속성적 조합을 포함한 시각적 추론 능력을 다룹니다 .

Winoground Benchmark

- 목적:
    - Winoground는 VLM이 시각적-언어적 조합적 추론을 수행할 수 있는지 평가하기 위해 설계되었습니다.
- 구성:
    - 각 데이터셋 샘플은 두 개의 이미지와 두 개의 캡션으로 구성됩니다.
    - 한 이미지는 하나의 캡션과 정확히 일치하며, 다른 캡션은 단어 순서가 다릅니다.
        - 예: "빛에 둘러싸인 식물"과 "식물에 둘러싸인 빛".
- 평가 목표:

    - 모델이 올바른 이미지-캡션 쌍에 더 높은 점수를 부여하는지 평가.
- 난이도:
    - 구문적 차이와 시각적 세부 사항을 모두 고려해야 하기 때문에 모델에게 도전적인 과제 .

Attribution, Relation, and Order (ARO) Benchmark

- 개요:
    - ARO는 속성, 관계, 순서와 같은 특정 개념에 대한 모델의 이해를 평가합니다.
- 데이터 생성:
    - COCO, GQA, Flickr30k 데이터셋을 기반으로 부정적인 캡션을 생성.
    - 부정적인 캡션 예:
        - "잔디를 먹고 있는 말" → "말을 먹고 있는 잔디".
- 평가 방식:
    - 모델이 부정적인 캡션에 대해 낮은 확률을 부여하는 능력을 측정.
- 장점과 단점:
    - 장점:
        - 많은 부정 캡션을 생성할 수 있음.
    - 단점:
        - 생성된 일부 캡션이 현실 세계에서 논리적으로 말이 안 될 수 있음 .

4.1.7 Dense Captioning and Crop-Caption Matching

개요

현재 세대의 비전-언어 모델(VLM)은 주로 짧은 텍스트 설명 입력에 제한됩니다. 일반적으로 사용되는 Clip Tokenizer는 최대 77개의 토큰만 생성할 수 있으며, 이는 약 50개의 영어 단어 또는 소규모 단락에 해당합니다. 하지만 이미지는 종종 이보다 훨씬 더 많은 정보를 포함합니다. 짧은 캡션만 사용하면 배경 정보와 세부적인 객체의 특성을 잃을 수 있습니다.

Densely Captioned Images (DCI) Dataset

- 데이터 생성:
    - **Segment Anything (Kirillov et al., 2023)**를 사용하여 이미지를 분리하고, 각 분리된 부분에 대해 인간 주석자들이 자세한 설명을 작성했습니다.
    - 이 방법으로 총 7,805개의 이미지가 주석 처리되었으며, 각 이미지에는 1,000개 이상의 단어로 구성된 캡션이 포함됩니다.
- 활용:
    - 이 데이터셋은 새로운 **Crop-Caption Matching** 작업에서 활용됩니다.
    - 각 이미지에 대해 VLM이 모든 하위 이미지 캡션 중에서 올바른 캡션을 해당 하위 이미지와 매칭하도록 요청됩니다.

평가 목표

- Crop-Caption Matching 작업을 통해 VLM이 장면의 세부사항을 얼마나 잘 이해하는지 평가합니다.
- 모델의 세부적인 시각적 이해와 텍스트 매핑 능력을 테스트합니다.

4.1.8 Synthetic Data-Based Visio-Linguistic Evaluations

개요

실제 데이터 기반 벤치마크는 다음과 같은 여러 도전에 직면합니다:

1. 특정 장면이나 객체에 대한 부정적인 캡션을 생성하는 데 어려움이 있습니다.
2. 모델이 실패하는 이유를 명확히 식별하기 어렵습니다.
    - 모델이 특정 객체를 인식하지 못해서인지, 아니면 객체 간의 관계를 이해하지 못해서인지를 구분하기 어렵습니다.
3. 실제 데이터의 캡션은 종종 매우 간단하며, 모호하거나 편향된 경우가 많습니다.

합성 데이터의 필요성

- 실제 데이터셋(예: COCO)은 VLM 평가를 위해 설계되지 않아 부정적인 캡션과 연결된 이미지를 제공하지 않습니다.
    - 예: "커피잔이 테이블 위에 있다"라는 캡션은 항상 해당 위치적 편향을 포함합니다.
    - 이러한 편향은 VLM이 실제 이미지 정보를 사용하지 않고도 정답을 예측할 가능성을 높입니다.
- 이러한 문제를 해결하기 위해 합성 이미지 데이터셋이 필요합니다. 합성 데이터는 다음과 같은 이점을 제공합니다:
    - 장면을 정밀하게 제어 가능.
    - 세분화된 정답 레이블과 캡션을 제공합니다.

Photorealistic Unreal Graphics (PUG)

- Bordes et al. (2023)는 **PUG** 데이터셋을 사용하여 복잡한 장면을 점진적으로 구성했습니다.
    - 배경 이미지를 설정하고, 요소를 하나씩 추가하며 장면을 세밀히 설계.
    - 예를 들어:
        - 특정 동물을 추가한 뒤, 모델이 해당 동물을 감지할 수 있는지 평가.
        - 동물을 왼쪽 또는 오른쪽으로 이동시키며 모델의 공간적 관계 인식 능력을 테스트.
- 발견:
    - 현재의 VLM은 공간적 관계 평가에서 무작위 추측(random chance) 수준의 성능을 보여주었습니다.

4.2 Benchmarking Bias and Disparities in VLMs

최근 머신러닝 시스템에서 편향**(bias)** 문제는 중요한 연구 주제로 자리 잡고 있습니다. 비전-언어 모델(VLM)에서도 이런 편향을 벤치마킹하기 위한 다양한 방법론이 도입되고 있습니다. 이 섹션에서는 **모델 분류(classification)**와 **임베딩 공간(embedding space)**을 활용한 분석을 중심으로 편향을 평가하는 방법을 다룹니다.

4.2.1 Benchmarking Bias via Classifications

편향을 평가하는 가장 일반적인 방법 중 하나는 분류(classification) 작업을 활용하는 것입니다.

편향 사례

1. 사람 관련 속성:
    - 성별, 피부 톤, 민족성과 같은 속성과 직업 또는 전문성을 연결하는 데서 발생하는 편향.
    - 예: 특정 직업군에서 특정 성별이나 민족성이 과대 혹은 과소 대표되는 경우.
2. 유해한 연관성:
    - 사람 이미지를 부정적인 개념(예: "도둑", "범죄자")과 연결하는 경향.
    - 이는 특정 집단에 대한 부정적인 고정관념을 강화할 위험이 있음.
3. 객체**-**사람 관계:
    - 사람과 옷, 스포츠 장비와 같은 일상적 객체 간의 연관성을 분석.
    - 특정 그룹에서 해당 객체가 더 자주 분류되는지 확인.

실제 데이터 활용

- 예: Agarwal et al. (2021)은 **CLIP** 모델에서 발생하는 잠재적인 편향을 분석했습니다.
    - 얼굴 이미지에서 인종, 성별, 나이에 따른 분류 비율을 측정.
    - "의심스러운 사람(suspicious person)"과 같은 유해한 레이블의 분포를 조사.
- 데이터 소스의 편차는 평가 결과에 영향을 미칠 수 있으므로, 데이터 품질과 분포의 차이를 주의 깊게 살펴야 합니다.

합성 데이터 활용

- Smith et al. (2023)은 **성별 균형된 대조 세트(gender-balanced contrast sets)**를 생성하여 편향을 평가했습니다.
    - 성별 관련 정보를 편집하고 배경은 고정하여 합성 데이터를 생성.
    - Li and Vasconcelos (2024)는 **인과적 개입(causal interventions)**을 사용해 원본과 대조적인 이미지-텍스트 쌍을 생성하여 모델의 예측 차이를 측정했습니다.

4.2.2 **Embedding**을 통한 편향 벤치마킹

개요

Embedding 기반 편향 분석은 VLM(Visio-Linguistic Model)의 임베딩 공간에서 텍스트와 이미지 간의 관계를 분석하는 방식입니다. 이는 분류 작업처럼 특정 다운스트림 과제에 국한되지 않고, VLM이 학습한 관계를 심층적으로 이해할 수 있게 합니다. 예를 들어, 임베딩 공간에서 사회적 편향(Social Bias)을 발견하거나 VLM의 암묵적인 연관성을 드러낼 수 있습니다.

Grounded-WEAT와 **Grounded-SEAT**

- 소개:
    - Ross et al. (2020)는 인간의 암묵적 연관(implicit associations)에서 발견되는 편향과 유사한 편향을 측정하기 위해 **Grounded-WEAT**와 **Grounded-SEAT**를 도입했습니다.
    - 예를 들어, "꽃"과 같은 쾌적한 개념은 "유럽계 미국인 이름" 또는 "밝은 피부"와 더 많이 연관되며, "아프리카계 미국인 이름" 또는 "어두운 피부"와는 덜 연관됩니다.
- 적용 사례:
    - "왕(king) - 여왕(queen) ≈ 남자(man) - 여자(woman)"와 같은 유용한 의미적 관계를 탐구.
    - 반대로, "남자 - 여자 ≈ 컴퓨터 프로그래머(computer programmer) - 가정주부(homemaker)"와 같은 해로운 편향을 발견.

CLIP 기반의 새로운 접근

- **CLIP**의 활용:
    - 텍스트와 이미지 임베딩 간의 명시적 매핑을 통해 편향을 분석.
    - 예: 성별, 피부 톤, 연령과 같은 인구통계 속성 및 고정관념 단어(예: "테러리스트", "CEO")와 이미지를 매핑하여 편향을 확인.
- 발견:
    - Wolfe와 Caliskan (2022)은 "미국인"이 "백인"과 더 많이 연관되는 현상을 발견.
    - Wolfe et al. (2023)은 성적 대상화와 관련된 편향을 발견.

4.2.3 언어 편향이 벤치마크에 미치는 영향

벤치마크의 언어 편향 문제

비전-언어 모델(VLM) 분야의 발전과 함께, 멀티모달 벤치마크의 품질과 구성이 점점 더 중요해지고 있습니다. 하지만 멀티모달 벤치마크는 종종 **언어 편향(linguistic bias)**에 의해 왜곡될 수 있습니다【95:0†source】.

예시**: VQA** 벤치마크

- **VQA (Visual Question Answering)** 데이터셋 [Antol et al., 2015]:
  - 이 데이터셋은 일부 "블라인드" 알고리즘이 시각적 정보를 사용하지 않고 언어적인 편향만을 활용해 높은 성능을 기록할 수 있음을 보여줍니다.
  - 예를 들어, "시계가 있습니까?"라는 질문은 약 98%의 확률로 "예"라는 답을 갖고 있어, 모델이 텍스트 패턴만으로 정답을 예측할 수 있습니다 [Goyal et al., 2017].

이미지**-**텍스트 검색 벤치마크에서의 언어 편향

- **BLIP** 모델 [Li et al., 2022b]:
  - BLIP와 같은 이미지 캡션 생성 모델에서 추정된 **텍스트 우선순위 (P(text))**는 이미지-텍스트 검색 벤치마크에서 높은 성능을 기록합니다.
  - 이는 ARO, Crepe, VL-CheckList, SugarCrepe와 같은 현대적인 벤치마크에서도 확인되었습니다.
  - 하지만 이러한 접근법은 모델이 시각적 단서를 충분히 사용하지 않는 **단일모달 단축 경로(unimodal shortcut)**에 의존하는 경향이 있습니다【95:0†source】.

균형 잡힌 벤치마크의 필요성

- 균형 잡힌 데이터셋:
  - Winoground [Thrush et al., 2022] 및 EqBen [Wang et al., 2023a]과 같은 균형 잡힌 벤치마크는 단일모달 단축 경로에 의존하지 않고 모델의 다중모달 학습을 평가하도록 설계되었습니다.
  - 이러한 벤치마크는 단순한 언어 패턴 탐지를 넘어서는 복잡한 조합적 추론 능력을 요구합니다.

4.2.4 학습 데이터의 특정 개념이 다운스트림 성능에 미치는 영향 평가

개요

VLM(Visio-Linguistic Models)의 성능은 학습 데이터에 포함된 개념의 빈도와 다양성에 크게 의존합니다. **Udandarao et al. (2024)**의 연구에 따르면, 학습 데이터에서 자주 나타나는 개념은 다운스트림 작업에서도 좋은 성능을 보이는 반면, 학습 데이터에 드물거나 포함되지 않은 개념은 모델 성능이 저하되는 경향이 있습니다.

방법론

1. 개념 리스트 정의:

    - 다운스트림 작업(예: 분류 작업)의 핵심 개념(예: 클래스 이름)을 나열합니다.
2. 인식 모델 활용:
    - **RAM(Recognition-Aware Models)** [Zhang et al., 2023d]과 같은 도구를 사용해 학습 데이터에서 이러한 개념의 존재 여부를 탐지합니다.
3. 평가 기준:
    - 특정 개념이 학습 데이터에 얼마나 포함되어 있는지를 기반으로 모델이 해당 다운스트림 작업을 해결할 가능성을 예측합니다.

응용 사례

- 데이터 검증 및 보완:
    - 다운스트림 작업의 성능을 향상시키기 위해, 데이터셋을 설계하거나 보완할 때 학습 데이터에 포함된 개념의 균형을 조정할 수 있습니다.
- 데이터 편향 파악:
    - 특정 그룹이나 속성의 과소 표현 또는 과대 표현을 식별하여 모델 편향을 줄이는 데 도움을 줍니다.

4.3 Benchmarking Hallucinations

개요

환각(hallucination)은 LLM(대형 언어 모델)에서 주요 문제로 지적되고 있으며, 이는 VLM(비전-언어 모델)에서도 중요한 평가 영역으로 간주됩니다. 환각은 모델이 높은 신뢰도를 가진 정보처럼 보이는 잘못된 또는 허구의 데이터를 생성할 때 발생합니다. 예를 들어, 사람이 달에 처음 발을 디딘 연도를 1969년이 아닌 1951년이라고 주장하거나, 존재하지 않는 역사적 사건을 만들어내는 것입니다【103:0†source】.

환각 평가

- **CHAIR Benchmark** [Rohrbach et al., 2018]:
    - COCO 데이터셋에 기반하여 객체 환각(object hallucination)을 평가하는 초기 벤치마크.
    - 정해진 객체 집합 내에서 환각을 측정하며, 짧은 캡션 생성에 유용하지만 긴 텍스트 생성에는 한계가 있음.
- **POPE Benchmark** [Li et al., 2023]:
    - 이진 질문 방식으로 객체 환각을 평가합니다.
    - 긍정적 질문: 실제 객체에 대한 질문.
    - 부정적 질문: 존재하지 않는 객체에 대한 질문.
- 모델 기반 접근법:

    - 최신 연구에서는 GPT-4와 같은 LLM을 활용하여 환각 평가를 확장하는 다양한 접근법을 도입했습니다.
        - **GAVIE** [Liu et al., 2023]: 지시(instruction)를 따르는 모델 평가.
        - **CCEval** [Zhai et al., 2023]: 캡션에서 환각된 객체의 위치를 확인.
        - **MMHal-Bench** [Sun et al., 2023]: 환각을 목표로 한 질문에 대한 VLM 응답 평가.
- 인간 평가:
    - 인간 주석자들이 세밀한 캡션 주석을 통해 평가를 수행.
    - 예: **Gunjal et al., 2024**의 연구.

제한점과 과제

- CHAIR 벤치마크는 COCO 데이터로 한정되며, 훈련 데이터와 중복될 가능성이 높아 새로운 평가 기준을 제공하기 어렵습니다.
- 더 나은 환각 평가를 위해, 다양한 데이터셋과 방법론을 도입해야 합니다.

4.4 Benchmarking Memorization

개요

훈련 데이터의 잠재적 암기를 평가하는 것은 LLM(대형 언어 모델)과 같은 단일 모달 모델에서는 광범위하게 연구되었으나, VLM(Visio-Linguistic Model)에서는 그 복잡성 때문에 충분히 탐구되지 않았습니다. VLM의 암기 평가가 어려운 두 가지 주요 이유는 다음과 같습니다:

1. 디코더 부재:
    - CLIP과 같은 결합 임베딩 VLM에는 디코더가 포함되지 않아 모델의 파라미터 및 학습된 임베딩에 암기된 정보를 직접적으로 디코딩하기 어렵습니다.
2. 제한된 생성 능력:
    - CoCa 및 LLaVA와 같은 VLM은 제한된 생성 능력을 가지며, 이미지-텍스트 간 교차 모달 암기를 노출하는 방법은 여전히 연구 과제입니다.

Déjà Vu Memorization

Jayaraman et al. (2024)는 "Déjà Vu Memorization"이라는 현상을 연구하여 CLIP 모델이 훈련 이미지에서 객체를 효과적으로 "기억"할 수 있음을 보여줍니다.

방법론

1. **k-**최근접 이웃 테스트**(k-NN Test)**:
    - 훈련 분포에서 추출된 공개 이미지 세트를 사용합니다.
    - 대상 훈련 캡션에 대해 임베딩 공간에서 가장 가까운 k개의 공개 이미지를 찾습니다.

    - 이 이미지들을 사용하여 대상 훈련 이미지에 포함된 다양한 객체를 디코딩합니다.
2. 참조 모델:
    - 대상 이미지-캡션 쌍을 학습하지 않은 별도의 CLIP 모델을 훈련합니다.
    - 이 모델에서도 유사한 k-NN 테스트를 수행하여, 학습으로 인한 암기 여부를 구별합니다.
3. 암기 평가:
    - 대상 모델과 참조 모델의 객체 검출 정밀도/재현율 점수 차이를 측정.
    - 점수 차이가 클수록 암기가 높음을 나타냅니다.

암기 완화 기법

Jayaraman et al. (2024)는 텍스트 랜덤화를 암기 완화에 가장 효과적인 기법으로 제안합니다.

- 텍스트 랜덤화:
    - 훈련 캡션의 일부 텍스트 토큰을 무작위로 마스킹.
    - 이는 텍스트 증강을 도입하여, 훈련 캡션과 해당 이미지 간의 연관성을 과적합하는 모델의 능력을 줄입니다.

4.5 Red Teaming

개요

레드 팀링(Red Teaming)은 모델의 공개 인터페이스를 활용하여 의도하지 않은 또는 유해한 출력을 생성하도록 유도하는 프로세스를 의미합니다. 이 접근법은 주로 모델의 취약성과 문제를 발견하고 수정하는 데 사용됩니다【111:0†source】.

레드 팀링의 주요 활동

1. 적대적 데이터셋 활용:
    - 유해한 출력이나 행동을 이끌어내는 데 초점을 맞춘 데이터셋을 제작.
    - 이러한 데이터셋에는 올바른 참조 응답(예: 답변 거부)이 포함되며, 모델은 해당 참조 응답과의 차이를 기준으로 점수가 매겨짐【111:0†source】.
2. 사례:
    - 민감한 이미지를 제공하고 이를 상세히 설명하도록 모델에 프롬프트를 제공하는 방식.
    - 예: "이 이미지에서 활동을 설명하시오"와 같은 일반적인 텍스트 프롬프트가 유해한 결과를 초래할 수 있음.

VLM에서의 레드 팀링 과제

- 신뢰도**(Faithfulness)**:
    - 출력 내용이 입력 데이터와 일치하는지 평가.
- 프라이버시**(Privacy)**:
    - 훈련 데이터에서 개인 정보를 학습하지 않도록 확인.
- 안전성**(Safety)**:
    - 유해하거나 폭력적인 출력을 방지.
- 공정성**(Fairness)**:
    - 편향된 출력을 최소화.

언어 및 텍스트-이미지 모델에서의 레드 팀링 작업

- 언어 모델에서는 레드 팀링 데이터셋이 다음을 드러내는 데 사용됨:
    - 유해한 언어 사용.
    - 사회적 편향.
    - 데이터에서의 특정 리스크.
- 리더보드:
    - 다양한 적대적 과제에 대해 언어 모델을 벤치마크할 수 있는 플랫폼 제공.
    - 텍스트-이미지 작업에서도 유사한 평가 방식 적용.

위험 완화 방법

레드 팀 평가 후 특정 위험을 완화하는 방법:

1. 사후 처리**(Post-Processing)**:
    - 모델의 출력을 수정하여 유해성을 줄임.
2. 모델 세부 조정:
    - 인간 피드백 강화 학습(RLHF)을 활용하여 안전성을 높임【111:0†source】.

5. Extending VLMs to Videos

개요

기존 비전-언어 모델(VLMs)은 정적 시각적 데이터(이미지)에 중점을 두었지만, 비디오 데이터는 새로운 과제를 제시합니다. 비디오 데이터는 물체의 움직임과 동적 관계를 이해하거나, 시간적 및 공간적 위치에서 행동을 식별하는 능력을 필요로 합니다. 비디오 모델은 텍스트-비디오 검색, 비디오 질문 응답, 생성 등의 작업을 포함하여 중요한 컴퓨터 비전 과제를 지원합니다. 그러나 프레임 속도(예: 24fps)로 인해 저장 및 처리 비용이 크게 증가하며, 이는 비디오 VLM 설계 시 최적화와 타협을 요구합니다【114:0†source】.

5.1 Early Work on Videos Based on BERT

VideoBERT

- 개요:
    - VideoBERT는 초기 비디오-언어 모델링 접근법으로, 일반적인 비디오-언어 작업에서 성공적인 첫 사례로 인정받습니다.
    - CLIP 기반의 대조 학습(contrastive learning) 접근과는 달리, VideoBERT는 시각적 토큰과 텍스트 토큰을 단일 트랜스포머 네트워크에 결합하는 초기 융합**(early fusion)** 방식을 사용합니다.
- 데이터 처리:
    - YouTube에서 수집된 요리 영상(설명 텍스트 포함)이 데이터로 사용됩니다.
    - 텍스트는 **자동 음성 인식(ASR)**을 통해 생성되며, 각 프레임은 개별 시각적 토큰으로 처리됩니다.
- 훈련 목표:
    - BERT 모델의 마스킹 및 재구성 방식을 활용.
    - 일부 토큰을 마스킹한 후 이를 재구성하는 작업을 통해 텍스트와 비디오 간의 강한 정렬(alignment)을 달성.
- 성과:
    - VideoBERT는 제로샷 행동 분류 및 자유 형식 비디오 캡션 생성과 같은 작업에서 우수한 성능을 입증.
    - 비디오 데이터에서 생성 기반 작업을 성공적으로 처리한 최초의 모델 중 하나로 평가받음.

MERLOT

- 특징:
    - MERLOT는 텍스트를 비디오와 시간적으로 정렬하여 비디오-언어 모델링을 확장한 접근법입니다.
    - YouTube의 대규모 데이터셋을 기반으로, 텍스트는 ASR로 생성되며, 더욱 다양한 장면과 콘텐츠를 포함.
- 훈련 방식:
    - 트랜스포머 네트워크를 사용하여 **순수한 자율 학습(self-supervised learning)**을 수행.
    - 지역 텍스트 토큰과 프레임 비주얼 토큰 간의 대조 학습 목표와 마스킹 언어 모델링, 시간 재정렬 목표를 결합.
- 성과와 한계:
    - 질문 응답 작업, 특히 시각적 상식 추론(visual common sense reasoning)에서 강력한 성능을 발휘.
    - 그러나 텍스트 생성 능력이 부족하여 고급 시각적 추론을 제공하는 데는 한계가 있음【114:0†source】.

5.2 Enabling Text Generation Using an Early-Fusion VLM

VideoOFA: Early-Fusion VLM for Video-to-Text Generation

- 설계 동기:
    - 기존의 비디오 VLM 대부분은 텍스트 생성을 지원하지 않거나 비디오 인코더와 별도로 학습된 텍스트 디코더를 결합하여 최적의 성능을 달성하지 못하는 문제가 있었습니다.
- 해결책**: VideoOFA**:
    - 두 단계 사전 학습 프레임워크를 제안하여 단일 이미지-텍스트 생성 VLM을 비디오-텍스트 작업으로 적응.
    - 초기화:
        - 텍스트 생성을 지원하며 대규모 이미지-텍스트 데이터로 공동 사전 학습된 이미지-텍스트 VLM을 초기화.
        - 시각-언어 표현 학습.
    - 중간 단계 비디오**-**텍스트 사전 학습:
        - 비디오 작업에 특화된 개념(예: 시간적 추론)을 학습하도록 VLM 백본을 적응.
        - 세 가지 훈련 목표:
            1. 비디오 캡셔닝**(Video Captioning)**:
                - 비디오 콘텐츠를 서술.
            2. 비디오**-**텍스트 매칭**(Video-Text Matching)**:
                - 비디오와 텍스트 간의 정확한 매핑.
            3. 프레임 순서 모델링**(Frame Order Modeling)**:
                - 비디오 프레임의 순서 이해.
- 성과:
    - 여러 비디오 캡셔닝 및 비디오 질의응답(VQA) 벤치마크에서 평가.
    - 이전 모델과 비교하여 향상된 성능을 입증【116:0†source】【116:3†source】.

5.3 Using a Pretrained LLM

개요

비전-언어 모델(VLM)은 점점 더 강력한 텍스트 이해 능력을 가진 기존의 LLM을 활용하는 방향으로 발전하고 있습니다. 이 접근법은 언어 모델을 새로 학습하는 대신, 기존의 LLM을 기반으로 시각적 백본을 정렬(alignment)하는 방식으로 진행됩니다. 이러한 방식은 캡셔닝(captioning) 목적을 통해 이루어지며, 비디오 모델에서도 비슷한 추세가 이어졌습니다LaMA**

- 설명:
    - **Video-LLaMA**는 비디오 및 오디오 신호와 언어 간 강력한 정렬을 보여주는 접근법입니다.
    - 이 모델은 **BLIP-2**를 기반으로 설계되었으며, 다음과 같은 구성 요소를 포함합니다:
        - **Video Q-former**: 비디오 데이터를 처리.
        - **Audio Q-former**: 오디오 데이터를 처리.
    - 훈련 데이터:
        - Webvid-2M이라는 큐레이션된 비디오 데이터셋을 사용하여 비디오와 언어의 정렬을 학습.

    - 목표:
        - 캡셔닝 손실을 기반으로 학습.
    - 추가 미세 조정:
        - MiniGPT-4, LLaVA, VideoChat에서 제공되는 시각적 지시 데이터를 사용해 인간 상호작용에 적합한 형태로 조정.
    - 특징:
        - Video-LLaMA는 대화형 에이전트로서 설계되어 표준 벤치마크가 아닌 대화형 API를 통해 평가됩니다 .

MiniGPT4-Video는 **MiniGPT-v2를 확장하여 비디오 이해를 지원합니다**.**

- 이 모델은 인접한 4개의 시각적 토큰을 단일 토큰으로 병합하여 입력 토큰 수를 줄이고 효율성을 높입니다.
- 프레임의 자막에서 텍스트 토큰을 추출하여 비디오 프레임의 표현을 개선합니다.
- 구성 요소:
    - 비전 인코더.
    - 단일 선형 투영 레이어.
    - 대형 언어 모델(LLM).
- 평가 방법:
    - Video-ChatGPT, Open-Ended Questions, Multiple-Choice Questions(MCQs)와 같은 벤치마크에서 평가.
- 성과:
    - 기존 모델보다 비디오 이해 성능에서 일관되게 우수한 결과를 보임 .

5.4 평가의 기회 **(Opportunities in Evaluations)**

비디오 데이터에 대한 벤치마크는 이미지와 유사한 작업(예: 캡셔닝)을 포함하지만, 비디오 데이터는 추가적인 평가 유형에 대한 문을 열어줍니다. 예를 들어, EgoSchema와 같은 데이터셋은 오브젝트나 에이전트 간의 상호작용을 이해해야 하는 긴 비디오의 질문에 답하도록 요구합니다. 이는 단순히 장면을 설명하는 것을 넘어선 평가를 가능하게 하며, 이는 이미지로만 수행하기 어려운 작업입니다. 비슷하게, ActivityNet-QA, MSVD-QA, 그리고 MSRVTT-QA와 같은 데이터셋은 질문에 적절히 답하기 위해 관련 프레임을 검색하거나 행동을 지역화해야 합니다. 그러나 많은 질문에서는 단일 프레임만으로도 정확한 답변을 제공하기에 충분합니다. 예를 들어, 축구 경기를 보여주고 "사람들이 하는 스포츠는 무엇인가요?"라고 묻는 질문은 단일 프레임만으로도 답할 수 있습니다. 이는 현재 비디오 벤치마크를 해결하는 데 비디오의 시간적 측면이 얼마나 필요한지에 대한 의문을 제기합니다.

비디오에서 행동의 의미론적 측면을 이해하는 것은 매우 중요하지만, 비디오는 모델의 추론 능력이나 세계 이해를 조사할 수 있는 독특한 기회를 제공합니다. 이와 관련하여 합성 데이터는 비디오 기반 VLM의 추론 능력을 조사하는 데 매우 효과적임이 입증되었습니다. 예를 들어, Jassim 등은 비디오를 생성하여 물리 법칙을 따르거나 위반하는 경우를 보여줍니다. 예를 들어, 갑자기 사라지는 공은 시공간적 연속성을 위반합니다. 그런 다음 모델은 공의 궤적이 물리 법칙을 따르는지 질문을 받습니다. 놀랍게도, VideoLLaMA 또는 PandaGPT와 같은 모델은 랜덤 성능을 초과하지 못했으며, 인간은 80% 이상의 정확도를

기록했습니다. 이러한 결과는 비디오 VLM이 여전히 기본적인 추론 능력이 부족하며, 이는 합성 데이터를 통해 효율적으로 조사될 수 있음을 시사합니다.

현재 비디오 VLM의 능력은 인상적이지만, 시간적 특성을 통해서만 가능해지는 추가적인 추론 능력을 조사할 기회가 여전히 존재합니다【8:3†source】.

5.5 비디오 데이터를 활용하는 데 있어 도전 과제

비디오-텍스트 사전 학습의 주요 도전 과제 중 하나는 시간적 측면에서 약한 감독(weak supervision)이 부족하다는 점입니다. 예를 들어, **VideoPrism**에서 다룬 바와 같이, 현재 인터넷 데이터는 주로 장면의 내용을 설명하는 데 초점이 맞춰져 있어 행동이나 동작을 설명하지 못합니다. 이로 인해 비디오 모델이 이미지 모델로 전환될 가능성이 커집니다. 비디오에서 학습된 CLIP 모델은 명사 편향(noun bias)을 보일 수 있으며, 이는 상호작용을 모델링하는 것을 어렵게 만듭니다. 결과적으로 이러한 모델들은 비디오로 학습되었음에도 시간적 이해가 부족하게 됩니다.

비디오와 자막 데이터를 생성하여 장면의 내용뿐만 아니라 시간적 측면을 포함하는 것은 이미지의 장면을 설명하는 것보다 훨씬 복잡하고 비용이 많이 듭니다. 이를 해결하기 위한 잠재적 솔루션이 있습니다. 예를 들어, 비디오 자막 생성 모델을 사용하여 더 많은 자막을 생성할 수 있습니다. 그러나 이를 위해서는 초기 고품질 데이터셋이 필요합니다. 또 다른 옵션은 비디오 인코더를 비디오 데이터만을 사용하여 학습시키는 것입니다. 이는 VideoPrism에서도 활용되었으며, 불완전한 자막의 영향을 제한합니다.

데이터 외에도 계산 비용도 중요한 도전 과제입니다. 비디오를 처리하는 것은 이미지를 처리하는 것보다 비용이 더 많이 들며, 이는 훨씬 더 중복적인 모달리티입니다. 이미지는 중복 정보가 많지만, 비디오의 두 연속된 프레임은 더욱 유사합니다. 따라서 이미지 기반 VLM에서 유용하게 사용된 마스킹(masking)과 같은 기술을 활용하여 더 효율적인 학습 프로토콜이 필요합니다. 이러한 모든 도전 과제, 즉 사전 학습 데이터, 계산 비용, 평가 품질에 대한 문제는 세계를 더 잘 이해하는 비디오 VLM을 개발하기 위한 유망한 연구 방향을 제시합니다【12:0†source】.

6 결론

비전을 언어로 매핑하는 연구는 여전히 활발히 진행 중인 분야입니다. 대조적인 방법에서 생성적인 방법까지, VLM(Vision Language Models)을 훈련하는 데는 다양한 방식이 존재합니다. 하지만 높은 계산 비용과 데이터 비용은 대부분의 연구자들에게 큰 장벽이 됩니다. 이는 주로 사전 훈련된 LLM이나 이미지 인코더를 활용하여 모달리티 간 매핑을 학습하는 방법을 선호하게 만듭니다.

VLM을 훈련하는 방법에 상관없이 몇 가지 일반적인 고려 사항이 존재합니다. 대규모 고품질 이미지와 캡션은 모델 성능을 향상시키는 중요한 요소입니다. 또한 모델의 기초를 강화하고 인간의 선호도에 맞게 모델을 정렬하는 작업도 모델 신뢰성을 높이는 데 매우 필요합니다.

모델 성능을 평가하기 위해 다양한 벤치마크가 도입되었으며, 이들은 VLM의 비전-언어 능력과 추론 능력을 측정합니다. 그러나 많은 벤치마크는 언어 사전만으로도 해결될 수 있다는 심각한 한계가 있습니다. 이미지와 텍스트를 연결하는 것이 VLM의 유일한 목표는 아니며, 비디오 또한 표현 학습에 활용될 수 있는 중요한 모달리티입니다. 하지만 양질의 비디오 표현을 학습하기 위해서는 여전히 많은 도전 과제가 남아 있습니다.

VLM에 대한 연구는 매우 활발히 진행 중이며, 이러한 모델을 더 신뢰할 수 있도록 만들기 위해 필요한 많은 요소들이 여전히 부족합니다【12:0†source】.