Deep Learning for Computer Vision

HW3 Report

b12901069 段奕鳴

Problem 1: Zero-shot inference with LLaVA

1. Report the accuracy of LLaVA without VCD (3-1) and with VCD (3-2)

Setting	Accuracy	Improvement
LLaVA w/o VCD (1-1)	78.2 %	-
LLaVA w/ VCD (1-2)	76.2 %	− 2.0%

The zero-shot inference with the original LLaVA already achieves competitive performance (78.2% accuracy). Surprisingly, LLaVA with VCD shows slightly lower accuracy (76-77%), rather than the expected improvement. The probable explanation is that POPE questions are relatively simple binary object presence queries, where language priors are not the dominant error source and visual grounding is already strong in baseline LLaVA. VCD's contrastive mechanism, which requires sampling and noise injection, may introduce instability on tasks where the baseline already performs well. VCD shows limited benefit on POPE's simple binary questions but would likely help on open-ended captioning or complex reasoning tasks where language priors dominate over visual grounding.

2. Describe the underlying mechanism of Visual Contrastive Decoding (VCD) and how it helps mitigate object hallucinations in Large Vision-Language Models

Visual Contrastive Decoding (VCD) is a training-free method that mitigates object hallucinations by contrasting model predictions from two different visual inputs:

(1) clean image: original, visual-grounded predictions

(2) distorted image: noise-corrupted image that removes visual information.

At each generation step, VCD computes:

$$logits_{vcd} = (1 + \alpha) * logits_{clean} - \alpha * logits_{distorted}$$

α: controls hallucination mitigation strength

$logtis_{clean}$: probabilities from clean image (visual-grounded)

$logits_{distorted}$: probabilities from distorted image (language priors)

Tokens that appear in both distributions are likely driven by language priors (hallucinations), while tokens unique to the clean distribution are truly visual-grounded.

The root causes of hallucinations are (1) statistical bias: models rely on common object co-occurrences in training data and (2) language priors: models follow word associations learned from text.

To address these problems, VCD compares predictions with and without visual info. The distorted image removes visual grounding, revealing language priors. The clean image provides full visual info, capturing both visual and language signals. The difference isolated truly visual-grounded predictions. The contrastive formula penalizes hallucination tokens while boosting visual tokens.

Problem 2: PEFT on Vision and Language Model for Image Captioning

1. Report your **best setting** and its corresponding **CIDEr & CLIPScore** on the validation data. Briefly introduce your method.

Rank (r)	16
Alpha (α)	64
Target modules	Q, K, V, O projection matrices
Initialization	A (down projection): Kaiming uniform B (up projection): zeros
Optimizer	AdamW
Learning rate	1e-4
Weight decay	0.01
LR scheduler	5 % warmup, cosine annealing
Regularization	Dropout: 0.1, label smoothing: 0.05
Epoch	4
Batch size	8
Decoding	Greedy, max/min new tokens: 30/5
Trainable parameters	8.26M
CIDEr	**0.9756**
CLIPScore	**0.7324**

Implementation:

1. Architecture

- Vision Encoder: Pretrained ViT-Base (frozen, 86M params)
- Language Decoder: Pretrained Qwen3-1B (frozen base, 1B params)
- Visual-Language Bridge: 2-layer MLP projection
    - 768 (ViT) → 2048 (GELU) → 1024 (decoder hidden)
    - Trainable: 3.67M params
- PEFT Method: LoRA (Low-Rank Adaptation)
    - Applied to: Q, K, V, O projection matrices in all 28 attention layers
    - Rank r=16, scaling α=64 (α/r=4.0)
    - Trainable: 4.59M params

2. Training Strategy

Regularization:

- Dropout 0.1 on attention weights and MLP outputs
- Label smoothing 0.05 on cross-entropy loss
- Weight decay 0.01

Optimization:

- AdamW optimizer, lr=1e-4
- 5% warmup + cosine decay scheduler
- Gradient clipping (max_norm=1.0)

Data:

- 90/10 train/val split (internal validation for monitoring CIDEr/CLIPScore)
- Batch size 8

3. Other Technical

Attention Masking:

- Strict causal mask over entire sequence (visual prefix + text)
- Per-sample key-padding mask to ignore PAD tokens
- Combined mask prevents information leakage during training

Generation:

- Greedy decoding with minimum token constraint (min_new_tokens=5)
- Prevents early-EOS collapse observed in initial experiments
- max_new_tokens=30

LoRA Design:

- Matrix A: Kaiming uniform init
- Matrix B: Zero init (ensures zero LoRA contribution at t=0)
- Scaling factor α/r=4.0 maintains consistent adapter strength across different ranks

2. Report **two different attempts of LoRA setting** (e.g. initialization, alpha, rank…) and their corresponding **CIDEr & CLIPScore**

	Setting 1	Setting 2
Rank (r)	4	16
Alpha (α)	16	64
Target Modules	q_proj, k_proj, v_proj, o_proj	
Initialization	down: Kaiming uniform, up: all zeros	
Trainable Parameters	1.94 M	8.26 M
CIDEr	0.942	0.976
CLIPScore	0.727	0.732

* The reported scores are both inferences using 4-th epoch

* For Setting 2, 2-layer MLP is used in vision projection, while setting 1 has only 1 layer