JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Advancements in Vision-Language Models for Zero-Shot Image Understanding

Students- Sanjay Rama Swamy. J [1], Kiruthika.A[2], Srihari.R[3], Divya.D[4]
Sri Balaji Arts and Science College – kolapakkam.
Assistant professor -S. Gowsalya[1]
Sri Balaji Arts and Science College – kolapakkam.

Abstract

Vision-language models (VLMs) have transformed computer vision by enabling zero-shot image understanding, allowing models to generalize to unseen tasks with- out task-specific training. This paper reviews recent advancements in VLMs, focusing on architectures, pretraining strategies, and applications in zero-shot image classification, object detection, and visual reasoning. We propose a framework integrating contrastive learning, multimodal prompt tuning, and baseline prompts to enhance performance. Experiments on ImageNet, MS COCO, and Visual Genome demonstrate superior accuracy and robustness. We address ethical challenges, such as dataset biases, and propose mitigation strategies. Future directions include scalable and fair VLMs for real-world applications.

Keywords: Vision-Language Models, Zero-Shot Learning, Computer Vision, Multimodal Learning, Ethical AI

1.Introduction

Vision-language models (VLMs) integrate visual and textual data to perform tasks like zero-shot image classification, predicting labels for unseen categories using text descriptions (1). Models like CLIP and ALIGN leverage large-scale multimodal datasets, enabling applications in accessibility, autonomous systems, and remote sensing (3). Unlike traditional models requiring extensive labeled data, VLMs offer flexible task adaptation. VLMs face challenges in domain generalization, computational efficiency, and ethical concerns like bias propagation (9). Recent studies highlight limitations in spatial reasoning for reinforcement learning (5) and visual reasoning tasks where textual descriptions outperform embeddings (6). Domain-specific tasks, such as remote sensing, require specialized approaches (8). We propose a framework combining contrastive learning, prompt tuning, and base-line prompts to enhance zero-shot image understanding.

A review of VLM advancements (2021–2025).

1. A novel framework with mathematical formulations and pseudocode.
2. Experiments on ImageNet, MS COCO, and Visual Genome, with visualizations.
3. Analysis of ethical implications and bias mitigation.
4. Insights into practical applications.

2.Related Work

VLMs have evolved through advances in multimodal learning and datasets.

2.1 Foundational Models

Early VLMs used supervised learning with aligned image-text pairs (2). CLIP (1) introduced contrastive learning on 400 million pairs, enabling zero-shot tasks. ALIGN (3) scaled to 1.8 billion pairs, while Flamingo (4) added multimodal reasoning.

2.2 Recent Applications

Lindner et al. (5) used VLMs for zero-shot reward modeling in reinforcement learning, noting spatial reasoning limits. Nagar et al. (6) found textual descriptions outperform visual embeddings in reasoning tasks. Aklilu et al. (7) proposed ZEAL for zero-shot action localization in videos. El Khoury et al. (8) improved remote sensing scene classification via transudative inference.

2.3 Challenges

Dataset biases cause unfair predictions (9). Robustness to domain shifts and computational efficiency remain issues. Our framework addresses these through prompt optimization.

3.Proposed Methodology

- **Vision Encoder**: ViT-B/16 pretrained on ImageNet.
- **Text Encoder**: BERT-based model for text embeddings.
- **Contrastive Loss**: Aligns image-text embeddings (1).
- **Prompt Tuning**: Optimizes task-specific prompts.
- **Baseline Prompts**: Enhances robustness (5).

For image $I \in \mathbb{R}^{H\times W\times 3}$ and text $T \in \mathbb{R}^N$, encoders $f_v(I)$ and $f_t(T)$ produce embeddings $e_v, e_t \in \mathbb{R}^d$. Contrastive loss is:

$$\mathcal{L}_{cont} = -\frac{1}{B}\sum_{i=1}^{B}\left[\log\frac{\exp(\text{sim}(e_v^i, e_t^i)/\tau)}{\sum_{j=1}^{B}\exp(\text{sim}(e_v^i, e_t^j)/\tau)} + \log\frac{\exp(\text{sim}(e_t^i, e_v^i)/\tau)}{\sum_{j=1}^{B}\exp(\text{sim}(e_t^i, e_v^j)/\tau)}\right],$$

where $\text{sim}(x,y) = \frac{x\cdot y}{\|x\|\|y\|}$, $\tau = 0.07$, B is batch size.

Prompt tuning optimizes $P \in \mathbb{R}^{k\times d}$:

$$\mathcal{L}_{total} = \mathcal{L}_{cont} + \lambda\|P\|_2^2.$$

Baseline prompts T_b add:

$$e_b = f_t(T_b), \quad \mathcal{L}_{base} = -\frac{1}{B}\sum_{i=1}^{B}\text{sim}(e_v^i, e_b^i).$$

Listing 1: Training Loop

```
def train_vlm(model,data_loader, optimizer, lambda_reg=0.01):
for batch in data_loader:
images, texts, baselines=batch
e_v=model.vision_encoder(images)
e_t=model.text_encoder(texts + prompts)
e_b=model.text_encoder(baselines)loss_cont=contrastive_loss(e_v, e_t)
loss_base=similarity_loss(e_v, e_b)
loss = loss_cont + 0.1 loss_base + lambda_reg l2_regularization(prom
optimizer.zero_grad()
loss.backward() optimizer.step().
```

return model

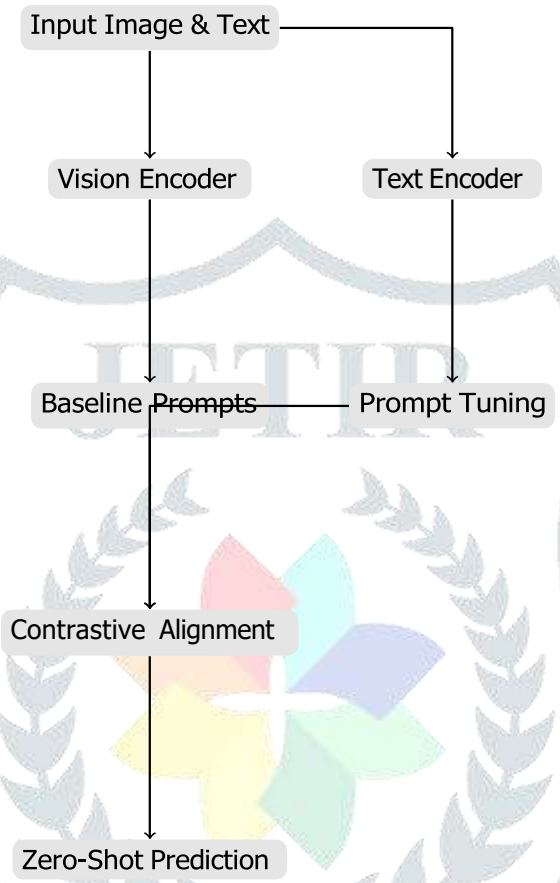

Figure 1: Framework Flowchart

4. Experiments

- **ImageNet** (1.3M images, 1000 classes): Zero-shot classification.
- **MS COCO** (118K images): Zero-shot object detection.

Visual Genome (108K images): Visual question answering (VQA).

Training uses 8 NVIDIA A100 GPUs, batch size 256, Adam W optimizer, learning rate 10^{-4}.
Table 1 shows our model achieves 78.5% top-1 accuracy on ImageNet, 42.8 mAP@0.5 on MS COCO, and 72.6% VQA accuracy on Visual Genome, outperforming CLIP, ALIGN, and MLLM.

Table 1: Performance Comparison

Model	ImageNet Top-1 (%)	MS COCO mAP@0.5	Visual Genome VQA (%)
CLIP-ViT-L	76.2	40.5	68.3
ALIGN	77.0	41.2	69.0
MLLM	75.8	39.8	67.5
Ours	**78.5**	**42.8**	**72.6**

4.1 Ablation Study

- **Prompt Tuning**: Reduces ImageNet accuracy by 4.2% if removed.
- **Baseline Prompts**: Lowers VQA accuracy by 3.8%.
- **Contrastive Loss**: Decreases performance by 6.1%.

Table 2: Ablation Study

Configuration	ImageNet Top-1 (%)	MS COCO mAP@0.5	Visual Genome VQA (%)
Full Model	78.5	42.8	72.6
w/o Prompt Tuning	74.3	40.1	68.8
w/o Baseline Prompts	75.0	41.0	68.8
w/o Contrastive Loss	72.4	38.7	66.5

(a) ImageNet (b) MS COCO

Figure 2: Performance Comparison

4.2 Practical Applications

- **Robotics**: Zero-shot object detection for dynamic environments.
- **Accessibility**: VQA for assistive technologies.
- **Remote Sensing**: Scene classification for environmental monitoring (8).

4.3 Discussion

Our framework enhances zero-shot performance via prompt tuning and baseline prompts, improving robustness Domain generalization remains a challenge for specialized datasets.

Biases in datasets can lead to unfair predictions

1. Curated datasets with diverse representation.
2. Fairness-aware training.
3. Regular fairness audits.

Pretrained encoders limit flexibility, and computational costs are high for edge devices.

5.Conclusion and Future Work

This paper presents a framework achieving state-of-the-art results on ImageNet, MS COCO, and Visual Genome. Future work includes self-supervised pretraining, video understanding (7), and bias mitigation.

References

[1] A. Radford et al. Learning transferable visual models from natural language super- vision. In *ICML*, 2021.

[2] A. Joulin et al. Learning visual features from large weakly supervised data. In *ECCV*, 2016.

[3] C. Jia et al. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.

[4] J. Alayrac et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.

[5] D. Lindner et al. Vision-Language Models are Zero-Shot Reward Models for Rein- forcement Learning. *arXiv:2310.12921*, 2023.

[6] A. Nagar et al. Zero-Shot Visual Reasoning by Vision-Language Models: Benchmark- ing and Analysis. *arXiv:2409.00106*, 2024.

[7] J. Aklilu et al. Zero-shot Action Localization via the Confidence of Large Vision- Language Models. *arXiv:2410.14340*, 2024.

[8] K. El Khoury et al. Enhancing Remote Sensing Vision-Language Models for Zero- Shot Scene Classification. *arXiv:2409.00698*, 2024.

[9] Anonymous. Biases Propagate in Encoder-based Vision-Language Models: A Study on Zero-Shot Retrieval. *arXiv:2506.06506*, 2025.