Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

b-it Bonn-Aachen International Center for Information Technology

Evaluation of Few-Shot Transfer of Vision-Language Foundation Models to Learn Lightweight Models for Robotic Vision Tasks

R&D Project Defense

March 7, 2025

Shinas Shaji

Advisors
Prof. Dr. Sebastian Houben (H-BRS, Fraunhofer IAIS),
Santosh Thoduka M.Sc. (Fraunhofer IAIS)

Introduction

Vision-Language Models (VLMs)

- Neural networks that process both **images** and **text**
- Like Large Language Models (LLMs):
  - Learn **general visual-textual understanding** from pre-training [1]
  - Then aligned to human preferences and **instruction-following**
- Shown to be able to **adapt** to new tasks without extensive task-specific training [2], hence quite **generalizable**

---

[1] A. Radford et al., Learning Transferable Visual Models From Natural Language Supervision, in Proceedings of the 38th International Conference on Machine Learning, M. Meila & T. Zhang, Eds., ser. Proceedings of Machine Learning Research, vol. 139, PMLR, Jul. 2021, pp. 8748–8763

[2] P. Liu et al., Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing, ACM Comput. Surv., vol. 55, no. 9, Jan. 2023

Introduction

Few-Shot Transfer

- Teaching a **generalizable model** new tasks by showing it a few **examples**
- Model 'learns' to recognize patterns from these examples
- Can then apply this 'learning' to new, unseen instances

Prompt: A [DOG] has droopy ears and is often fluffy. This is a [DOG]:

Figure 1: This is a [DOG].
Image from Wikipedia, Link

Figure 2: Is this a [DOG]?
Image from Wikipedia, Link

Prompt: Is this a [DOG]?
Expected Answer: Yes

Motivation

Few-Shot Transfer

Fine-tuning

- Requires significant computational resources, modifies model parameters
- Needs **large amounts** of **labeled data**
- Can lead to catastrophic forgetting
- Refers to fine-tuning a pre-trained/instruction-tuned model on a specific task

Few-shot Transfer

- Uses **'few' examples** or natural language **descriptions**
- No model parameters are updated
- Potentially more practical for real-world applications [a]
- Can be less effective for complex tasks

[a] T. Brown et al., Language Models are Few-Shot Learners, in Advances in Neural Information Processing Systems, H. Larochelle et al., Eds., vol. 33, Curran Associates, Inc., 2020, pp. 1877–1901

Problem Statement

Dataset Labeling for Computer Vision Tasks

Challenge: Creating labeled datasets to train specialized models for computer vision tasks is **time-consuming** and **expensive** [3], but VLMs are generalizable

Constraint: However, VLMs are too **computationally intensive** for direct deployment on resource-constrained environments (e.g., robots)

Opportunity: VLMs could potentially automate label generation (**pseudolabels**) to train **downstream** models

Research Question: Can VLMs be transferred to generate **pseudolabels** for computer vision tasks to train **lightweight** downstream models?

[3] J. Deng et al., ImageNet: A Large-Scale Hierarchical Image Database, in 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255

Proposed Approach

Evaluating VLMs for Pseudolabel Generation

Approach: Evaluate VLMs on generating **accurate pseudolabels** under various **zero-shot** and **few-shot** transfer conditions

Key Research Aspects

1. How does the **number of examples** (few-shot vs. zero-shot) affect pseudolabel quality?
2. What are the **computational requirements** for practical application?
3. How effective are the **downstream models** trained on pseudolabels?

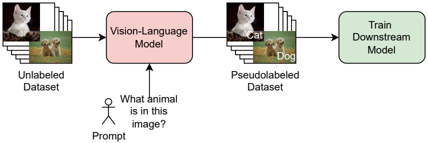

Figure 3: Using VLMs to generate pseudolabels for downstream model training

Related Work

Development of Vision-Language Models

- **Alignment models**: Generate unified text-image embeddings (CLIP [a], FLAVA)
- **Generative models**: Geneate text conditoined on multimodal inputs (Flamingo, Frozen [b], GPT-4o, Claude 3/3.5/3.7, etc.)

[a] A. Radford et al., Learning Transferable Visual Models From Natural Language Supervision, in Proceedings of the 38th International Conference on Machine Learning, M. Meila & T. Zhang, Eds., ser. Proceedings of Machine Learning Research, vol. 139, PMLR, Jul. 2021, pp. 8748–8763

[b] M. Tsimpoukelli et al., Multimodal Few-Shot Learning with Frozen Language Models, in Advances in Neural Information Processing Systems, M. Ranzato et al., Eds., vol. 34, Curran Associates, Inc., 2021, pp. 200–212

Architectural Approaches

- **Towered**: Separate vision and language models with adapters
- **Unified**: Single model processing both modalities "early on" [a]

Key Insight: Enables framing vision tasks as text generation [b], enabling streamlined task transfer

[a] Chameleon Team, Chameleon: Mixed-Modal Early-Fusion Foundation Models, arXiv preprint, May 2024. arXiv: 2405.09818 [cs.CL]

[b] J. Cho et al., Unifying Vision-and-Language Tasks via Text Generation, in Proceedings of the 38th International Conference on Machine Learning, M. Meila & T. Zhang, Eds., ser. Proceedings of Machine Learning Research, vol. 139, PMLR, Jul. 2021, pp. 1931–1942

Related Work

Transfer Learning & Adaptation Techniques

Prompting Techniques

- Crafting prompts to improve task performance [a]
- In-context learning: Providing examples in context [b]
- Chain-of-thought prompting for complex reasoning [c]

Parameter-Efficient Fine-Tuning

- Prefix-tuning: Optimizing task-specific prompt vectors [d]
- Requires fewer parameters than full fine-tuning

Research Gaps

- Few-shot transfer in VLMs less explored than in NLP
- Limited research on VLMs for dataset annotation
- Few studies on downstream model performance with VLM-generated labels
- Our work addresses these gaps

[a] P. Liu et al., Pre-train, Prompt, and Predict: A Systematic Survey of Prompting

Related Work

Applications and Datasets

Applications of VLMs

- Large-scale pre-training enables generalization
- Contrast with traditional DNNs trained on specific tasks

Auxiliary Learning Tasks

- Self-supervised generation of auxiliary labels [a]
- Visual instruction tuning for generative models [b]

[a] S. Liu et al., Self-supervised generalisation with meta auxiliary learning, in Advances in Neural Information Processing Systems, H. Wallach et al., Eds., vol. 32, Curran Associates, Inc., 2019

[b] H. Liu et al., Visual instruction tuning, in Advances in Neural Information Processing Systems, A. Oh et al., Eds., vol. 36, Curran Associates, Inc., 2023,

Key Datasets: Various datasets exist for various vision tasks.

- ImageNet, CIFAR-10 [a]: Object recognition
- Microsoft COCO: Detection, segmentation, captioning
- Derm7Pt [b]: Specialized dermatology dataset
- MVTec: Anomaly detection

[a] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images, M.S. thesis, Department of Computer Science, University of Toronto, 2009

[b] J. Kawahara et al., Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets, IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 2, pp. 538–546, 2019

Experimental Setup

Datasets

Models and Prompting Strategies

CIFAR-10 Experiments

Downstream Model Training

Specialized Domain Experiments

Computational Resources Analysis

Key Findings

Limitations and Future Work

Thank You!

Questions?

- Email: shinas.shaji@smail.inf.h-brs.de