Tübingen AI Center
Universität Tübingen

Universität Tübingen · Tübingen AI Center
Maria-von-Linden-Straße 6 · 72076 Tübingen · Germany

Tuebingen, 14.07.2025

Gutachten zur Dissertation "Task Adaptation Strategies for Vision-Language Models" eingereicht von Ms. Monika Wysoczańska.

Prof. Dr.-Ing. Hilde Kühne
Professor for Multimodal Learning
Paul-Ehrlich-Str. 5
phone +49 7071 29-70867
h.kuehne@uni-tuebingen.de

https://tuebingen.ai

The thesis submitted by Ms. Wysoczańska targets the problem of adapting pretrained vision-language foundation models to various downstream tasks.

The thesis is, besides introducation and realted work broken down into five different aspects task-adaptation: First, the adaptaton of visual foundation models for fine-grained localization tasks, realized by a framework called CLIP-DIY (published at WACV 2024); second the problem of combining different backbones, namely CLIP and DINO for better performance (published at ECCV 2024); this idea is then further extended towards test-time adadptaon (publication accpeted to TMLR on July 9th[1]); next, the idea is proposed that the process of task adaptation can serve as an evaluation for visual representations (published at NeurIPS Workshop 2022 and IEEE Access 2024); finally, those findings are used to improve the adaptation of VLMs for personalizations such as summarizing large image collections (published at AAAI 2024).

The first part, "Task Adaptation through Modified Inference" introduces the problem of adapting visual foundation models open-vocabulary semantic segmentation (OVSS) together with an approach called CLIP-DIY. This method extends the capabilities of CLIP to OVSS through modified inference strategies. employs a multi-scale architecture that leverages CLIP's classification abilities at various spatial resolutions and enhances segmentation accuracy using foreground/background separation scores derived from unsupervised object localization techniques. Notably, CLIP-DIY achieves competitive results on standard semantic segmentation benchmarks without additional training or manual annotations. The approach combines CLIP with an unsupervised saliency detection method based on DINO, a self-supervised learning model known for its strong object localization capabilities. This integration of vision-language modeling and self-supervised learning paradigms shows promising potential for further exploration. The work was

[1] https://openreview.net/forum?id=wyOv4kGkbU

presented at the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2024.

The second part, “Leveraging Complementary Visual Foundation Model,” discusses the adaptation of CLIP through complementary visual foundation models such as DINO. The proposed method ,CLIP-DINOiser, combines the complementary representations of image-text-aligned and self-supervised visual representations to enhance performance in open-vocabulary semantic segmentation. To this end, DINO’s good localization priors are integrated into CLIP’s representation space via a lightweight adaptation module trained with DINO’s supervision while preserving CLIP’s original representations. CLIP-DINOiser has shown to reach state-of-the-art performance at that time while only requiring a single CLIP forward pass and two lightweight modules during inference. The work was presented at ECCV 2024.

The thrid part, “Leveraging Statistics from Pre-training Dataset” deals with the role of prompts for OVSS proposing, first, automated approaches for generating contrasting concepts (CC) at inference time based on an LLM as well as on a statistical analyis of the VLM’s pre-training dataset, and, second, an evaluation framework to cature real-world challenges in OVSS. It shows that integrating the pretraining distribution can positively influnce the perfromance of CLIP-based semantic segmentation methods. The work is accepted for TMRL.

The fourth part, “Task Adaptation for Foundation Model Selection and Analysis” proposes an evaluation protocol for visual representations in the context of VQA through downstream task adaptation. To enable comparison of differrent visual representations, a model is designed that accommodates varying input representation dimensions to allow for direct comparative studies of different visual representations and their task-specific suitability. In this context, VQA is leveraged as an exemplary task. The framework allows to derive implications for the development of more robust visual representations. This work was presented at NeurIPS Workshop 2022, and published by IEEE Access in 2024.

Finally, the work discusses the “Task Adaptation in Real-world Scenarios”. To this end, findings with respect to the adaptability of visual foundation models across different downstream tasks, are considered with respect to their practical implications in real-world contexts, namely adaptation of a Vision-Language for personalized image collection summarization. To this end, rating information for a travel protal website are leveraged to personalize summaries for future users of the platform. The approach employs a proprietary VLM to establish connections between visual content and textual user feedback, resulting in a more relevant and personalized visual summaries without requiring additional manual annotations. This work was published at AAAI 2024.

Overall, Ms. Wysoczańskas work features a high-quality contribution to the field of vision-langauge learning. The thesis is highly streamlined, ranging from basic representation learning problems to questions of how to best evaluate such methods and well as how to make use of their capabilities in real-life. It provides a cohesive structure, with each part of the work building up on the previous ones. This structural approach allows Ms. Wysoczańska to create high-performing frameworks and models, which she shows to be useful in their specific task. At the same time, all topics targeted in the thesis are innovative and well-structured and executed.

With five first-author publications at top venues and good journals (ECCV, AAAI, WACV, TMLR, IEEE access) and several successful collaborations, Ms. Wysoczańska has demonstrated her ability to develop, implement, and evaluate new ideas and concepts in the field of vision-langauge understanding specifically and computer vision in general. With this work, Ms. Wysoczańska made a significant contribution to current research in the field of computer vision and multimodal learning.

I, therefore, I consider the thesis as pass and assess the outcome of the submitted PhD Thesis with:

Summa cum laude

H. Kuehne

Prof. Dr. Hildegard Kuehne