

June 9, 2025

Re: Dissertation review for Monika Wysoczańska's PhD thesis "Task Adaptation Strategies for Vision-Language Models"

This thesis is coping with the problem of vision foundation model adaptation. The manuscript explores the different strategies that can be applied to solve new problems or improve current systems by leveraging these large models trained on gigantic data corpora. The most significant part of the manuscript copes with the problem of open vocabulary semantic segmentation. In open vocabulary segmentation the set of classes that are used to densely label the image is not known a priori, and the same model should therefore work across many datasets with largely different class ontologies. This is a very hard problem, because the problem setup does not allow much model tuning, and the set of categories on which the model is applied can be arbitrarily finegrained. The use of off-the-shelf foundation models is hard, as the nature of the inference is very different. This thesis presents several key state-of-the-art contributions in that space.

An additional part of the manuscript focuses on the adaptation of vision foundation models in two other setups: visual question answering and automatic curation of photo collections.

Summary

The thesis is composed of an introduction and related work chapters (chapters 1 and 2), followed by five technical chapters that correspond to previously published papers (chapters 3 through 7), and a conclusion (chapter 8).

Chapter 1 serves as an introduction. It sets the context, describing the conceptual difference between siloed custom-tailored solutions, and foundation models. The candidate then describes how foundation models could be used in specialized scenarios and what adaptation they need to undergo. The introduction finishes with stating five research questions, grouped into three buckets, that cover the content of the five technical chapters.

- The introduction states that dense annotations are nearly impossible to obtain, especially in expert domains such as medical imaging. Can foundation models like CLIP reliably address this problem? What about the scale of weakly aligned text-image data that is needed to train CLIP (or DFN, SigLIP, PE)?
- Regarding the research question 4: have the findings from VQA evals really changed anything in the way foundation models are designed? What could be easily actionable and scalable improvements?
- In the introduction, the contribution from Chapter 7 is presented as a "real world" case, seemingly in opposition to the other chapters. What makes it more "real"? The

fact that the data and problem was specific to a company?

Chapter 2 describes previous work and focuses on two research themes: foundation models for vision, and model adaptation.

- The categorization of foundation models, and structure of Sec. 2.1 seems right to me. However, it feels like the coverage could be larger, as many influential works in that space have been omitted. For supervised learning alone, the list of NN architectures is a bit expeditive (VGG), not to mention the early hassle of ViT training (DeiT). For CLIP-like models, one could dive deeper in older work, and most importantly be more exhaustive about cutting edge models (EVA, AIMv2).
- The same applies to adaptation strategies in Sec. 2.2. While the broad categories are right, the amount of details per category could be improved. The “dataset adaptation” represents by itself an immense field of research around transfer learning. On the other hand, the subsequent chapters all have their own related work, and the bibliography contains 259 entries, so I am confident that the candidate properly acknowledged all relevant papers. This does not invalidate the scientific soundness of the presented work, only aims at improving the added value of this chapter.

Chapter 3 describes a training-free method for obtaining segmentation masks for any textual prompt coined CLIP-DIY. The method works by computing text-patch similarities at several scales using an off-the-shelf CLIP model (Eq. (3.1)), and further gating this score using an unsupervised object discovery model like FOUND or CutLER (Eq. (3.4)). The proposed algorithm outperforms (or nearly matches) the state-of-the-art on Pascal VOC and COCO. This is particularly strong, given that the method did not have any trainable parameters!

- From Table 3.4.2 it seems that the method simply does not work without the gating using “objectness”. However, without many scales the numbers are already quite strong. Have you considered this problem the other way around? Could you formulate it as a zero-shot classification of segments given by FOUND or CutLER? In that case, could you leverage the CLS token of CLIP?

Chapter 4 describes an improvement upon CLIP-DIY. While the aforementioned model leveraged two off-the-shelf foundation models (CLIP and FOUND), in this contribution the different models are fused together. A lightweight model is trained atop CLIP to mimic the self-similarities of DINO. Such cleaned-up self-similarities are used to pool clip features from larger regions, effectively denoising even further the CLIP feature map. Finally, a simple foreground background model is trained atop CLIP features by distilling the scores of FOUND. The proposed method gives strong performance across many datasets, setting a new standard in that space.

- The work on registers by Timothee Darcet (ICLR 2024) suggests that CLIP models could be trained at scale and obtain much better attention maps with some care. One of the two advantages of using DINO here is to clean up noisy local features from CLIP. Do you think that most modern CLIP-derivatives would still benefit from a DINO model?
- A lot of energy seems spent on dealing with the “background”. This is a quite ill-defined concept. Could the evaluation protocol be changed, in order to circumvent

this problem?

Chapter 5 describes a method for improving the way that open-world, open-vocabulary semantic segmentation is performed. Because open-vocabulary semantic segmentation models leverage contrastively-trained foundation models like CLIP, for a given class q , in order to classify each “patch” as positive or negative, the practitioner needs to properly define the “background”. This chapter describes a clean evaluation protocol which modifies multi-class semantic segmentation into a series of binary segmentation problems, and evaluates two strategies for defining background classes for a given class query: either by mining statistics from large image-caption datasets, or prompting an LLM for relevant negative prompts.

- The proposed strategies allow defining a set of “negatives” for a given query q , staying with the multinomial logistic model of CLIP. If we transform this problem into a sequence of binary problems, what would be the performance of a binary classifier with a threshold optimized on the training set of the given dataset? How many annotated images do you need to make that work? This baseline is mentioned at the beginning of Sec. 5.3.1 but only reported in the appendix (Fig. 5.6.6).
- With a slight abuse of notation, the proposed procedure feels like populating the training set of a non-parametric classifier like kNN. The training set is composed of one positive (the query), and we are trying to generate negative examples, in order to define the best decision boundary for that class (a convex polytope). Could that be generalized to potentially leverage the non-linear nature of non-parametric models?

Chapter 6 describes a benchmark for evaluating different image representations through the lens of visual question answering (VQA). The proposed evaluation is simple and clear, but limited to toyish data (CLEVR). It was somewhat pioneering, appearing before the consensus on architectures and datasets to evaluate foundation models with VQA (efforts like Cambrian).

- The best performing methods according to the experiments in this chapter are either Slot-Attention or DTI-Sprites. This is probably an artifact of the type of dataset that this evaluation was run on. My main question around this work is: how can we translate the learnings from this work into recommendation for future foundation model development? How would DINOSAUR perform in this benchmark (it appeared after the paper was presented)?

Chapter 7 describes a simple method for summarizing a photo collection in a personalized way, with application to [Booking.com](https://www.booking.com) data. The model takes as input some user context (reviews), and a pool of images from a property, and proposes the most relevant pictures for that user. This is a creative adaptation of a visual foundation model like CLIP to the problem of personalization. In user studies, the proposed multimodal algorithm performs much better than the unimodal baseline.

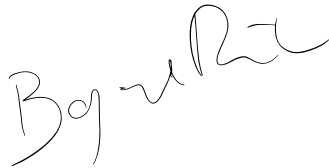
Chapter 8 provides some final remarks and discusses open problems and future work.

Appraisal

This thesis presents a very large body of work, spanning several themes. The writing is clear and the document is properly structured. The end of the introductions provides a clear overview of the following chapters, and connects them to publications. The very coherent set of contributions around open-vocabulary semantic segmentation is of prime quality and has already had a lot of impact on the research community. The manuscript illustrates both the technical mastery and deep knowledge of previous literature of the candidate.

I am confident that there is sufficient material and novelty in this manuscript for an oral defense before a jury.

Piotr Bojanowski
Research Director
bojanowski@meta.com

A handwritten signature in black ink, appearing to read 'Bojanowski', with a stylized, cursive script.