Using deep learning models in image processing

Xiong, Z.

Summary

Deep learning has revolutionized image processing, enabling remarkable advances in fields from autonomous driving to medical diagnostics. Yet the very success of these models, fueled by increasing complexity, has raised urgent questions about how they work, why they sometimes fail, and whether they can be trusted in critical applications. This thesis addresses these questions by offering both a coherent map of the rapidly evolving landscape of deep learning in computer vision and concrete contributions that advance model interpretability and reliability. The thesis is structured as a logical progression from a broad, systematic overview to focused investigations that together build a foundation for more transparent and robust vision systems.

Chapter 1 discusses the concepts relevant to this thesis. We then introduce the research questions that form the basis for the content of **Chapters 2** to **5**.

Chapter 2 provides a comprehensive review of the deep learning models that have come to dominate image processing. Starting from the mathematical principles shared by convolutional neural networks and vision transformers, it traces their parallel evolution and eventual convergence. Beyond classical architectures, the chapter introduces recent large-scale foundation models: pure vision, vision–language, and multi-modal models, highlighting their potential to serve as universal solutions for vision tasks. By identifying open challenges and future directions, this chapter establishes the conceptual toolkit and the research gaps that motivate the subsequent chapters.

Building on this foundation, **Chapter 3** addresses interpretability in a core vision task: semantic segmentation. Focusing on the cross-attention mechanism, we reformulate it as a probabilistic model and reveal that its operation can be interpreted as k-means clustering in the feature representation space. This novel perspective not only demystifies the widely used cross-attention component but also provides a principled basis for understanding its behavior under full supervision.

Chapter 4 extends this line of inquiry to the more challenging weakly supervised setting, where pixel-wise annotations are unavailable. We analyze how the absence of detailed labels shapes the learned feature space and propose a probabilistic framework that recovers missing background information through contrastive learning. Together, **Chapters 3 - 4** demonstrate that a unified probabilistic view can explain and improve attention-based models across different supervision regimes, thereby enhancing their interpretability.

Chapter 5 addresses reliability in a demanding clinical application: the analysis of microscope images of kidney biopsies. Our technical contributions results in a reliable and robust approach. In transplant medicine, monitoring a transplanted kidney is crucial. The status of the kidney is studied based on the micro-anatomy in a biopsy. The anatomical structures are densely packed, vary widely in size and shape, and often touch. We introduce a generative instance segmentation model that combines diffusion models with vision transformers. The architecture is explicitly designed to mitigate class imbalance, resolve objects at multiple scales, and disambiguate touching instances by emphasizing boundary information. Extensive experiments confirm its robustness, highlighting its potential for clinical deployment.

Finally, in **Chapter 6**, we summarize the results and conclusions of our research and answer the research questions. Based on these results, we outline a perspective for future developments in the field.

Collectively, this thesis reflects on how a macroscopic understanding of model evolution (**Chapter 2**) informs targeted investigations into interpretability (**Chapters 3**–**4**) and reliability (**Chapter 5**), demonstrating how these contributions together advance the goal of building vision systems that are not only powerful but also understandable and trustworthy. By bridging theoretical foundations with practical solutions, this work aims to pave the way for deep learning models that can be deployed with confidence in real-world settings, where transparency and dependability are as important as accuracy.