Using deep learning models in image processing

Xiong, Z.

Chapter 6

Conclusions and Discussion

In this thesis, we have addressed six research questions. These questions comprise a survey (Chapter 2) and two application domains: the development of deep learning in the image processing community and a specific focus on classification problems in the computer vision domain. The themes investigated cover a general review of deep learning variants and narrow to two specific classification cases: image segmentation and object detection. In this chapter, we summarize the main findings of our research. Additionally, we discuss and identify the challenges posed by our methods and possible solutions to address them. Lastly, we suggest potential developments for future work.

6.1 Main Findings

In each chapter, we have proposed new views and approaches to address the corresponding research questions. Here, we summarize these answers and present the main findings inspired by experimental results.

Chapter 2 answers **RQ1: What is the current development in deep learning models and their applications in image processing? Chapter 2** provides a new perspective to track the development of deep learning models from CNN via Transformers to the hybrid models, based on the extension and improvement of the mathematical definitions of their core components: convolution and attention mechanisms. Next, we elaborate on the fundamental elements of the typical architectures and list other stand-alone methodologies that can help deep learning models in general. Subsequently, a broad range of active research topics in the image processing domain

has been highlighted, each with a revolutionized cornerstone application. Lastly, we are interested in the limitations and challenges of modern deep learning architectures and in providing direction for future research.

Chapter 3 takes the results from **Chapter 2**, and narrows down our research topic on one specific case of image classification: semantic segmentation. In this chapter, we offer a novel perspective to understand the cross-attention mechanism in the context of semantic segmentation. The research question **RQ2: How can we explain an attention mechanism as a probabilistic model that captures global cross-class correlations in semantic segmentation tasks?** is answered by our Variational Gaussian Mixture Model. The framework interprets cross-attention as building a multi-modal distribution via a Gaussian mixture and supports online updates (i.e., mini-batch style) via Bayesian update rules. Specifically, each semantic concept is a centroid of a Gaussian distribution. It measures its uncertainty using correlation matrices and gathers global class-specific information from the input feature maps in the dataset. However, images with complex interactions among semantic concepts yield ambiguous decision boundaries between classes. In general, a solution can be to integrate more geometrical or structural constraints. Adding constraints has led to the formulation of an additional research question **RQ3: How can we integrate geometric constraints into a probabilistic model and solve it efficiently?** To answer this, we have extended our Gaussian Mixture Model to support arbitrary additive geometric regularization. For simplicity without losing generalizability, we choose the popular Total-Variance constraints and solve them with a level-set algorithm for topological flexibility. Based on our experimental analysis, we conclude that our probabilistic model is resource-friendly and scalable as the number of semantic classes increases.

Chapter 4 further explores the cross-attention mechanism in semantic segmentation under different conditions and constraints. In this chapter, we focus on generating high-quality dense annotations without providing explicit pixel-wise supervision. The so-called weakly supervised semantic segmentation (WSSS) is a very important task, as annotation can be expensive, especially in expertise-intensive fields such as biology and medicine. Although cross-attention is a powerful tool for extracting long-range dependencies, WSSS suffers from an intrinsic deficit: incomplete information cannot infer the correct full class assignment. Therefore, we propose a conditional probabilistic model to separate the foreground segmentation from the missing background information. To answer **RQ4: How can we interpret weakly supervised semantic segmentation**

as a feature space decomposition problem using conditional probabilistic models? We adopt contrastive-learning algorithms to infer background information across the whole dataset and interpret it as out-of-distribution information. Then, foreground assignment only focuses on likely supplementary regions, subtracting the background. Furthermore, we have observed that loss functions in WSSS exhibit equipotential sets on their loss surfaces, leading to perplexity in distinguishing good from bad assignment configurations. This phenomenon has given rise to an additional research question **RQ5: How can we eliminate the equipotential of the loss surface in weakly supervised semantic segmentation using geometric regularization?** Our solution is to add geometric regularization, like Total-Variance constraints, to provide extra evidence for WSSS loss functions to reject the bad assignments. Our implementation of the conditional model uses a masked cross-attention with a level-set method. Through our experiments, we have achieved competitive results on a typical dataset and demonstrated an explainable prototype transformation in the latent feature space.

Chapter 5 focuses on another special case of image classification: instance segmentation/object detection on a biomedical dataset. We find that kidney pathology images are too complicated to get acceptable performance by simply adapting existing algorithms. The crux is that each image contains a large number (about 500) of anatomical structures, spanning a wide range of scales and tightly packed together. To exacerbate the situation, there are multiple anatomical types, but with a severe class imbalance. Therefore, instance segmentation/object detection for Kidney pathology images requires flexibility to handle varying scales and efficiency to accommodate dense objects. It is promising to integrate the diffusion model and attention mechanism into RCNN with three-fold advantages: (1) generative methods benefit from generating various scaled candidates; (2) attention can capture long-range correlation to accurately locate large objects; (3) RCNN can only focus on potential regions and ignore the irrelevant background. From this analysis, we formulated the last research question **RQ6: How can we equip an attention mechanism with an R-CNN in a diffusion manner in support of efficiency and generalization of dense instance segmentation tasks?** Our solution is to design a dynamic attention mechanism combining the diffusion models and RCNN. Specifically, each dynamic attention module takes as input noisy box candidates generated by diffusion models, computes global interactions among them, and outputs refined boxes for RCNN. In order to avoid sampling imbalance in the RCNN, we have also designed a class-balanced sampling algorithm to

stabilize training. The experiments with our framework and the subsequent analysis demonstrate that it has domain transferability (i.e., the ability to ignore the effect of stains), thereby demonstrating flexibility, generalizability, and efficiency.

6.2 Discussion

The methodologies and frameworks that we have proposed in this thesis are addressed in six research questions and have achieved promising results. They, however, still have some challenges with respect to algorithmic and theoretical perspectives.

6.2.1 Algorithmic Perspective

(1) Amortized Geometrical Constraint: In Chapter 3, the Total Variation regularization is a computational bottleneck. Since it lacks a closed-form solution, we must resort to iterative optimization via level-set methods. That means it is necessary to evolve the class-wise decision boundaries from the initial surfaces for each input feature map. In general, this process requires more than 15 iterations to converge from scratch, resulting in unexpected computational time. For efficiency, we adopt a compromise solution that requires an additional sub-network to predict a better initial surface. The refined surface can accelerate the convergence of the level-set algorithm but increase the difficulty of the training process, requiring more sophisticated hyperparameter tuning, such as the learning rate.

(2) Dynamic regional candidate: In Chapter 5, we adopt RoI-align Pooling to generate the initial region candidates as the input dynamic queries. Although it is a classical operator used in RCNN models, it assumes even sampling within each candidate region, thereby ignoring differences in scale and geometric irregularity. This intrinsic drawback significantly degrades object representations, especially for extremely small or large objects. Even worse, RoI-align Pooling fails to account for spatial exclusivity among candidates because it runs each candidate in parallel. A possible solution to this is to design a box attention that captures non-local information within each region for completeness and interacts between candidates to ensure clear spatial affiliation.

6.2.2 Theoretical Perspective

Variational Bayesian Inference: In Chapter 3, Variational Bayesian Inference (VB) is a family of graphical models for approximating intractable distributions arising in Bayesian inference. VB is an extension of the expectation-maximization (EM) algorithm and approximates the true posterior distribution as a factorization of simple distributions. By optimizing the evidence lower bound (ELBO), VB alternatively approximates the posterior distribution and updates the parameters. Although VB is scalable for inferring complex models with a large number of parameters, it has an intrinsic drawback: finding a suitable factorization often requires many trials. Therefore, we cannot actually estimate our factorization outperforming the others. The deficiency means that performance comparison of different factorizations in the same training setting should be considered to provide experimental evidence of selection over factorizations.

6.3 Future Work

In previous chapters, we have presented many methods for addressing research questions in the image processing community related to deep learning models. Deep learning models need novel ideas to evolve further. Therefore, it is crucial to discuss potential advancements in future work. These specifically concern interpretability, data efficiency, multi-modal understanding, and scalability. In this section, we briefly indicate possible future research directions.

(1) Deep learning models demand more effort to enhance their transparency. It is crucial to build trustworthy systems, especially in reliability-sensitive applications. Researchers are currently actively developing techniques in explainable AI (XAI). There are two promising directions for increasing human supervision of complex models. One direction is to develop more convenient visualization systems that can easily assess the responses/activations/graduations of arbitrary components to support decision-making. The other is the adaptation of the traditional algorithms or probabilistic models into differentiable modules, interpreting local components with solid mathematical formulations.

(2) Deep learning models require vast amounts of high-quality annotated data. However, it is very expensive to generate trustworthy supervision information, especially for

expertise-intensive fields. Therefore, advancements in areas such as transfer learning, few-shot learning, self-supervised learning, weakly supervised learning, and unsupervised learning can effectively address the data-hungry problem, enabling them to learn effectively with less data. Stand-alone techniques open up new possibilities for extending existing models to new applications where large labeled datasets are scarce.

(3) Deep learning models intend to solve increasingly complex problems that need to integrate information from multiple sources, such as text, images, audio, etc. Therefore, multiple-modal algorithms have the potential to facilitate more sophisticated reasoning on complex systems, such as autonomous systems, virtual assistants, and medical diagnostics. In addition, multiple-modal applications can revolutionize human-computer interaction and create more immersive experiences.

(4) The scalability of deep learning is advancing with better hardware, optimized algorithms, and distributed computing frameworks. This progress enables training on massive datasets and complex architectures, leading to the development of more sophisticated models with improved performance and generalization. Meanwhile, scaling deep learning models down to portable devices is enabled with cloud computing and high-performance processors. The advancement makes them more accessible, helping leverage AI to tackle real-world problems on mobile devices.

We have investigated a broad range of aspects of deep learning models in the context of image processing. We have been covering both technical applications and mathematical interpretability. In the rapidly evolving field of artificial intelligence, inevitably, our coverage is not comprehensive. Nevertheless, our contributions move the field forward and yield incremental insights and knowledge. We are pleased that our research adds to these increments and look forward to the further development of the field.