Using deep learning models in image processing

Xiong, Z.

Chapter 1

Introduction

Motivation

This thesis aims to address the application of deep learning models in image processing from two perspectives. In the broader context, we uncover a logical clue that reveals the direction of various structural evolutions and technical developments in deep learning models. With respect to specific research topics, we focus on interpretability and applications in image classification, including semantic segmentation, object detection, and instance segmentation.

1.1 Background

Humans are very talented at jointly processing a variety of information and comprehensively understanding corresponding scenes. Through a series of sensory systems, our brain perceives the real world by integrating a spectrum of senses: vision, smell, hearing, taste, and touch. Among these senses, visual information is the most frequent and direct form of sensory stimulation. The brain will enhance the visual input signal, analyze various aspects of the information, and integrate all pieces of evidence via a complex network of neurons. By extracting information from the network of neurons, we can link observations and recognize patterns. This active brain event is referred to as thinking, which aims to reason/infer essence over evidence. Likewise, if a machine is capable of reasoning, it demonstrates Artificial Intelligence (AI), which is defined as distinct from the natural intelligence displayed by animals, including humans. Concerning natural

intelligence, a process of inference reshapes neural connections by creating or deleting connections between neurons. In artificial intelligence, computational graphs, where nodes represent neurons and edges represent connections, imitate the reasoning process by adjusting edge weights. Consequently, the graphs are called **neural networks** and the process of adjusting weights is called **learning**.

In the field of computer vision, the most popular topics are mainly the following:

- **Classification** is the most fundamental task in the field of computer vision. Category cognition is used to correctly predict the class label(s) of a given image. Classification only extracts the coarsest information, yet requires the least supervision.
- **Segmentation** aims to efficiently generate a dense annotation for each pixel within a given image and classify them into different semantic regions. Semantic segmentation stops at class-level dense annotation, which only discriminates objects among different classes. Instance segmentation goes deeper toward instance-level annotation, further separating individual objects within each class.
- **Detection** is an intermediate level between classification and segmentation. The aim is to find a region-level classification that localizes each object while attaching a class label to the corresponding region.

In this thesis, we focus primarily on segmentation, object detection, and instance segmentation. Early deep learning models were based on fully convolutional neural networks, which cannot easily capture long-range dependencies between any pair of positions due to the limitation of the field of perception used in convolution operations [455]. In recent years, with the development of attention mechanisms, several deep learning models have contributed to the state-of-the-art (SOTA) in segmentation [178, 105, 368] and detection [870, 367, 808]. These SOTA neural networks, however, work as black boxes. Without clear interpretations, AI is hindered from being widely applied in realistic scenarios due to a lack of administration. Therefore, it is worth exploring mathematical explanations to understand it better.

In this chapter, we first give some insight into the datasets that we used in our experiments. Given these datasets, we explain the major aspects addressed in this thesis. We investigate feature space learning under full supervision and formulate attention mechanisms as a variational Gaussian mixture. Next, we elaborate on how to decompose a feature space into background and foreground classes and introduce additional geometric constraints to facilitate semantic segmentation with only image-level supervision. Both aspects share the importance of building good feature spaces

from which we can learn better class-specific representations. Moreover, we explored extending attention mechanisms to process very dense objects.

1.2 Preliminaries

1.2.1 Data Sets for Computer Vision Studies

The availability of image-annotation paired datasets is limited due to the significant manual effort required to collect pixel-level annotations.

The research in this thesis is based on four public datasets and one private medical dataset:

- The PASCAL Visual Object Classes dataset [167] has been widely used as a benchmark for object detection, semantic segmentation, and classification tasks. It contains 20 object categories and has been split into three subsets: 1,464 images for training, 1,449 images for validation, and a private testing set.
- The Microsoft COCO dataset [415] is a large-scale dataset that addresses four computer vision tasks: semantic segmentation, object detection, instance segmentation, and panoptic segmentation. Containing more than 200,000 annotated images, there are 80 object categories for object detection, semantic segmentation, and instance segmentation, while there are 80 thing categories plus an extra 91 stuff categories for panoptic segmentation.
- The Cityscapes dataset [123] is a large-scale database that focuses on the semantic understanding of urban street scenes. It provides semantic, instance-wise, and dense pixel annotations for 30 classes, grouped into 8 categories. The dataset consists of around 5000 fine-annotated images and 20000 coarse-annotated images, captured in 50 cities over several months, during daytime and under good weather conditions. Its images are manually selected to feature the following: a large number of dynamic objects, varied scene layouts, and varied backgrounds.
- The ADE20K semantic segmentation dataset [849] contains more than 20000 scene-centric images exhaustively annotated with pixel-level objects and object parts labels. This dataset can support tasks of semantic segmentation, object detection, instance segmentation, and panoptic segmentation. There are a total of 150 object and stuff classes, including blocky regions such as sky, road, grass, and discrete objects such as person, car, and bed.

- We use images from kidney biopsies. These images are prepared in accordance with the Pathology Laboratory Protocol for kidney biopsies. The tissues were collected by core needle biopsy, Serra-fixed, rapidly processed using a routine tissue processor, then paraffin-embedded, serially sectioned at 5 μm, mounted onto adhesive slides, and stained with JONES silver staining. All biopsy samples, both transplant and native kidneys, from 148 patients were collected from the multi-center archive of the Departments of Pathology at LUMC, AUMC, and UMCU in the Netherlands, or from the Department of Nephrology in Leuven, Belgium. A pathology staff member anonymized all biopsies before further analysis. The 148 Jones-stained WSIs were obtained from scanners with PPM = 0.25 in BIG-TIFF format[1], and 303 patches were extracted by software ASAP version 1.9 for Windows [2] at level 0 of the BIG-TIFF images. 249 patches are used for training and 54 patches for evaluation.

1.2.2 Importance of Feature Space Learning in Computer Vision

Feature space refers to the n-dimensional vector space on which all possible feature vectors reside. Each vector is extracted from the given data via some defined mapping functions [349]. Feature space is one of the most fundamental infrastructures on which machine learning is built to represent and learn **concepts** digitally. Given a feature space, any data can be represented as a feature vector, giving coordinates in the n-dimensional space. Then, semantic segmentation tasks in computer vision are formulated as decision boundary problems that separate different class-specific examples with minimal error. Modern deep learning models can automatically learn to extract appropriate features and infer decision boundaries in a fully supervised way. However, the training process operates as a black box, as it is too complex. To simplify the black box, we separate feature extraction from inference and focus on drawing decision boundaries in the given feature space. We can rethink the feature representation of each class over a given dataset in classification tasks according to two observations: (a) any given dataset can define a closed set of relevant classes, i.e. **foreground classes**, which implies an infinite complementary set of irrelevant classes, i.e. **background classes**. (b) Any objective function assigns different scores to different decision boundaries, thereby inevitably creating a subset of equipotential

[1] https://www.awaresystems.be/imaging/tiff/bigtiff.html
[2] https://computationalpathologygroup.github.io/ASAP/

elements. The observation (a) indicates that it is irrational to infer background as a single class, whereas it is intractable to represent each one within precisely. The observation (b) implies that there is a set of necessary constraints that is learned to select among these equipotential elements and pushes decision boundaries toward the desired results. During training, both requests can be implicitly satisfied in fully supervised learning due to detailed target information, but this creates two difficulties for weakly supervised learning. Therefore, in this thesis, we are motivated to explore the construction of feature spaces for both fully and weakly supervised cases. For fully supervised semantic segmentation, we propose a variational Gaussian mixture model (VGMM) framework that leverages target information and constructs a feature space in which a single semantic vector represents each class. For weakly supervised semantic segmentation, we use conditional probabilistic models to complete two missing components due to using image-level target information: (1) to decompose background and foreground representations separately; (2) to integrate extra geometrical constraints to eliminate equipotential elements.

1.2.3 Introduction to Attention Mechanisms

Attention mechanisms are among the most important components in modern deep learning models because they are involved in constructing feature space. Attention mechanisms were first used for transformer models, and were initially shown to be effective in natural language processing tasks [642] and later for many computer vision works [156, 693, 674]. Attention mechanisms first take triplets (query, key, value) as input. For each query-key pair, attention probabilities are computed as feature similarities via the inner product. Finlay, a new query will be reconstructed as the sum of values weighted by attention probabilities. In this thesis, we interpret attention mechanisms as soft clustering processes. Depending on the triplet (query, key, value) settings, attention mechanisms can be roughly categorized into two types. One defines a set of global queries and updates them across the entire dataset [654, 773], working in a K-means clustering manner. The other generates a set of dynamic/local queries and discards them after use [663], working in a K-Nearest Neighbors (K-NN) clustering style. In this thesis, (1) we generalize K-means attention mechanisms into probabilistic models and explore the learned feature spaces from deep learning models; (2) we integrate K-NN attention mechanisms with regional neural networks (R-CNNs) and diffusion models for dense instance segmentation.

1.3 Research Questions

To detail aspects of segmentation and detection algorithms, we propose the following research questions:

RQ1 (Chapter 2) What are the current developments of deep learning models and their applications in image processing?

RQ2 (Chapters 3) How can we explain an attention mechanism as a probabilistic model that captures global cross-class correlations in semantic segmentation tasks?

RQ3 (Chapters 3) How can we integrate geometric constraints into a probabilistic model and efficiently solve it?

RQ4 (Chapters 4) How can we interpret weakly supervised semantic segmentation as a feature space decomposition problem using conditional probabilistic models?

RQ5 (Chapters 4) How can we eliminate the equipotential of the loss surface in the weakly semantic segmentation using geometric regularization?

RQ6 (Chapters 5) 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?

1.4 Thesis Structure

The remainder of this thesis is structured according to the research questions presented in the section 1.3. These are subsequently presented in four research chapters and a concluding chapter.

- **Chapter 2** provides a novel perspective for organizing the development of deep learning models in the image processing domain: image classification, image segmentation, and object detection (related to the RQ1). The survey provides a blueprint for broadening our understanding of related techniques to address challenges in the computer vision community. It provides a comprehensive introduction to contemporary deep learning systems and their core components, presented chronologically. In addition, we introduce other stand-alone techniques

to improve the performance of deep learning models. Moreover, we collect representative work across a variety of image processing applications: classification, detection, image generation, and an important sub-field: medical data processing.

- **Chapter 3** narrows down our research topic on semantic segmentation and revisits the important attention mechanism with a K-means style. We introduce a variational Gaussian mixture model (VGMM) that efficiently captures global cross-class correlations (related to RQ2). Similar to MoCo [243], our VGMM can be used as a light-weight plug-in. All parameters will be updated using momentum. Our model can analyze the feature spaces of deep learning models and probe residual queries (as in RQ3). The experimental results show that our approach achieves performance comparable to that of the compared methods.

- **Chapter 4** further explores the function of the attention mechanism in a weakly supervised paradigm where we can exploit the long-range contextual features to compensate for the missing supervision information. Specifically, we formulate weakly supervised semantic segmentation as a conditional labeling assignment problem. Based on contrastive learning [87], we first decompose the feature space into background and foreground sub-spaces via a binary segmentation (refer to RQ4). Then, subsequent class-specific segmentation is just inferred within the foreground subspace through attention mechanisms. Finally, we use level-set methods for total-length and minimal-region constraints, ensuring that the semantic segments are smooth (refer to RQ5). The experimental results show that our approach achieves performance comparable to the other methods.

- **Chapter 5** extends our research range onto instance segmentation/object detection, and proposes a generative method to tackle complicated medical data: kidney pathology images. Specifically, the chapter focuses on an anchor-free instance segmentation approach that can process a large number of objects (varying from 300 to 1000). Our framework combines diffusion models, R-CNNs, and Transformers (related to RQ.6). The diffusion models will provide random Gaussian noise as initial bounding boxes to replace anchors, and iteratively refine these noisy boxes into predicted bounding boxes. The Transformers obtain pooled regional features from previously refined boxes, extract regional features from them, and interpret all features as dynamic queries (related to RQ6). Among different object proposals, attention mechanisms can capture long-term dependencies. Based on these proposals, RCNNs will predict final instance segments, thereby alleviating resource costs. The experimental results

show that our approach outperforms the compared methods and can perform domain translation.

- **Chapter 6** elaborates the connection among the previous chapters and summarizes our contributions for exploring feature space, attention mechanisms, and probabilistic frameworks in computer vision tasks. In addition, we briefly describe the remaining challenges and potential improvements for the proposed approaches. Lastly, current challenges motivate future work on interpretability, cross-modality integration, and reliable data generation.

1.5 Terminologies and Definitions

In this section, we provide descriptions of common terms used in neural networks.

- **Feature**, used in deep learning, is an N-dimensional matrix (or formally called a tensor), containing representative information to describe key aspects of given data for a certain task.

- **Convolution** is a linear summation of element-wise products between two functions. One function, $f(x)$, is called the input function. The other, $g(x)$, is called the kernel (function). The domain of kernel D normally is a subset of the domain of $f(x)$ and $g(x)$ acts as weights to accumulate $f(x)$ into a real number $H(x)$:

$$H(x) = \sum_{i \in D} f(x - i)g(i) \tag{1.1}$$

- **Activation**, or activation function, is a family of non-linear functions to enhance the representation ability of neural networks. The most commonly used activation functions are the sigmoid, rectified linear unit (ReLU), softmax, and softplus.

- **Pooling** is an approach to down-sampling feature maps by summarizing the presence of features in patches of the feature map. Two common pooling methods are average pooling and max pooling, which summarize the average and the maximum activations of a feature, respectively.

- **Layer** logically is an atomic module that must be executed as a basic unit during runtime. The module can consist of several operations. Normally, a standard layer has convolution(s) and activation(s) w/o a pooling.

- **Pyramid** is a kind of data structure stacking with N layers. A layer l_k at k-th level has a larger size than that of l_{k+1} where $k \in \{0, 1, 2, \ldots, N\}$.
- **Fusion** is a kind of mathematical operation $f(\cdot)$ used to integrate N features into M features (normally $M = 1$). For instance, element-wise sum, element-wise product, or concatenation.
- **Self-Attention**, or called attention in short, is a kind of feature that acts as a modulation. The modulation only responds to relevant regions over the feature and removes noise or negligible parts.
- **Ground truth** denotes a correct result to a certain task justified by a human. For instance, manually labeled binary segmentation serves as ground truth for training a semantic segmentation network.
- **Loss function** is an objective function that measures the difference/error between the output of a network and the given ground truth. The cumulative loss can guide the training network in adjusting its internal parameters to reduce error.
- **Annotation** normally relates to ground truth used in segmentation tasks. A dense annotation denotes per-pixel labeling, where each pixel is uniquely assigned to a class—for instance, a binary segmentation for all cars in an image. Additionally, instance-level dense annotation assigns each pixel to an object with a unique ID. For example, a segmentation for each car in the image. Notably, an instance annotation is not a binary image since each instance occupies a unique ID.
- **end-to-end** is a term to indicate if a network can output its prediction in the same format as the ground truth. For instance, if a ground truth is an instance annotation, then an end-to-end network also outputs an instance annotation without any other add-on post-processing procedures.
- **Fully convolutional network** means the whole network is built without a fully-connected layer. For more details, please refer to [444].
- **Multi-scale**, or called multi-resolution and cross-scale, is a term to indicate that features of different layers have different sizes due to the existence of pooling functions. A multi-scale network fuses features from different layers into enhanced features that combine context information and accurate spatial resolution.
- **Regression**, or regression analysis, estimates the relationships among variables. It includes many techniques for modeling and analyzing several variables when

the focus is on the relationship between a dependent variable and one or more independent variables.

- **instance-awareness** is a term to indicate that if a network can sense individual identities sharing the same class label. For instance, if a network can correctly label every individual car in an image with a unique instance ID, it has instance-awareness.