Graduate School of Global Information and
Telecommunication Studies, Waseda University

Abstract of Doctoral Dissertation

Semantic Image Recognition Methods by Using Deep Learning
深層学習を用いた意味的画像認識手法

Hoang Anh DANG

Global Information and Telecommunication Studies
Multimedia Representation Research

August 2020

Artificial intelligence (AI) used to be widely perceived as the field that studies intelligent agents (IA). After the deep learning (DL) breakthrough of AlexNet in 2012, the term AI was frequently used by media, marketers, and academia alike in a much broader sense. In the annual report, the Stanford University's Institute for Human-Centered Artificial Intelligence (HAI) includes computer vision (CV), Pattern Recognition, Computational Linguistics (CL), Robotics, as the subcategories of AI. In point of fact, AI is generally defined as intelligence demonstrated by machines. As such, machine learning and its deep learning subcategory are also parts of IA in computer science.

Since the publication of AlexNet in 2012, DL has been adopted by many branches of science. However, DL sees the most success in the fields of natural language processing (NLP), CV, and IA. In the field of CV, human-level performance in the task of image classification has been achieved by multiple convolutional neural network (CNN) models in 2016. As a result, the ImageNet challenge discontinued in 2017. The end of the ImageNet challenge marks the new chapter in the field. The focus of the research community was shifted to more challenging topics. Some of the notable topics include generative network, semantic segmentation, activity recognition, and visual question answering (VQA). As a whole, the target of DL researches has become highly semantic.

Significant progress is made among all of the fields in recent years. However, the penetration of AI in everyday life is still limited. Like any other technology, the adoption of DL based AI in the general consumer market lags behind the adoption speed of the industrial market. Not to mention that the adaption speed of the industrial market itself is also lagging behind the research a considerable amount of time.

To be adopted by the industrial and general consumer market, further research must be done to adapt and improve the technology. For example, one of the obstacles that need to be overcome is the hardware limitation. Cloud computing can not satisfy the requirement of real-time processing. Furthermore, consumer hardware is still not powerful enough. This limitation leads to the rising trend of edge AI research and development in recent years.

Motivated by the recent success of DL, taking into account the above-mentioned low penetration, we aim to enhance the semantic performance and practicality of AI with the two following works:

Scalable Vector Graphic AI (SvgAI) is an IA that can draw semantical Scalable Vector Graphics (SVG) images. Instead of storing visual information pixel-by-pixel, SVG image is a document that describes the visual information. Different from natural language, SVG is an Extensible Markup Language (XML) based language. Therefore, SVG is highly semantic and structured hierarchically.

Image processing based raster to vector (R2V) converts raster images into SVG without retaining semantic data. With DL, our preliminary experiments and related works show the limited performance of the end-to-end network in the task. Rather than tackling the conventional end-to-end model design, we took an alternative approach. We trained an IA to perform the task on an SVG editor. The trained IA can create SVG images that are significantly smaller and more accurate compared to the available solutions. Though SvgAI, we show that IA can be used to solve the problem that challenges the conventional end-to-end model. SvgAI is a novel approach to solving the R2V problem.

Street Fashion Semantic Segmentation (SFSS) is a lightweight deep neural network (DNN) that performs semantic segmentation on street fashion photos. Introduced just after the release of ModaNet, SFSS is a pioneer work on semantic segmentation for street fashion photos. Semantic segmentation can be an essential part of the fashion recommender pipeline. In this work, firstly, we propose a unique and compact DNN design. This network offers state-of-the-art semantic segmentation performance. Furthermore, it requires less computational resources compared to the related works. Secondly, we propose the novel label pooling process, which creates lossless versions of the label in different scales. As the labels for auxiliary training objective, these label pool features significantly improve the context-awareness property of the network.

This thesis is divided into five chapters. Chapter 3 and Chapter 4 are dedicated to the works of SvgAI and SFSS, respectively. Common to both works are the introduction in Chapter 1, DL background in Chapter 2, and conclusion in Chapter 5. The content of each chapter in this thesis is as follows:

Chapter 1 overviews the history of AI, the contribution of DL into AI development, the achievement of DL, and the impact of this new development on research trends and business interest. This chapter concludes with the motivation, objectives of the research, and the relative position of this research in the contemporary landscape.

Chapter 2 introduces the technical background of DL and IA. It includes the milestones of DL and explains the popular components of DNN, such as convolutional layer and backpropagation. It also reviews notable DNN models in the DL era, such as AlexNet, InceptionNet. Section 2.3 describes IA and the two popular algorithms to train an IA, including Q-Learning and policy gradient. Important concepts related to the training process, including experience memory replay (EMR) and exploration policy, are also addressed in this chapter.

Chapter 3 is dedicated to the work of SvgAI. This chapter starts with an introduction to the R2V problem. The introduction also includes reviews on previous works, and the challenges need to be resolved. Different from the previous works on the topic, we propose a novel framework for R2V.

In our new framework, an IA is trained to draw vector images using an SVG editor. In order to train the IA, a complete training environment is needed. The design and implementation of our SVG editor environment are explained in Section 3.3.4.

The latter half of the chapter describes the experimental setting and evaluation result of SvgAI using both deep Q-Learning and gradient-policy. Dual ϵ–greedy exploration strategy and a unique training strategy are proposed to overcome the difficulties. The chapter is concluded by a comparison between SVG images produced by SvgAI with popular free and commercial R2V software.

Chapter 4 is dedicated to the work of SFSS. This chapter begins with the introduction to semantic segmentation. Then, we review related works on the topic, including SegNet, DeepLabv3+, and PSPNet.

Different from the previous works, we focus on the semantic segmentation task for fashion apparel. To produce the most efficient model for the task, we propose two novel contributions. They are: 1) a high-performance semantic segmentation DNN that follows the encoder-decoder structure, and 2) the 2D max-pooling-based scaling operation.

We train and evaluate our proposed network using the ModaNet data set. To better evaluate the network performance, the Intersection over Union Plus (IoU+) metric is also proposed. This metric is taking noise into account for better evaluation. An ablation study is conducted to analyze the effect of different auxiliary training losses.

Chapter 5 concludes this dissertation with possible directions for future works.