

Receptive Field Block Net for Accurate and Fast Object Detection

Songtao Liu, Di Huang^{*}, and Yunhong Wang

Beijing Advanced Innovation Center for Big Data and Brain Computing
Beihang University, Beijing 100191, China
{liusongtao, dhuang, yhwang}@buaa.edu.cn

Abstract. Current top-performing object detectors depend on deep CNN backbones, such as ResNet-101 and Inception, benefiting from their powerful feature representations but suffering from high computational costs. Conversely, some lightweight model based detectors fulfil real time processing, while their accuracies are often criticized. In this paper, we explore an alternative to build a fast and accurate detector by strengthening lightweight features using a hand-crafted mechanism. Inspired by the structure of Receptive Fields (RFs) in human visual systems, we propose a novel RF Block (RFB) module, which takes the relationship between the size and eccentricity of RFs into account, to enhance the feature discriminability and robustness. We further assemble RFB to the top of SSD, constructing the RFB Net detector. To evaluate its effectiveness, experiments are conducted on two major benchmarks and the results show that RFB Net is able to reach the performance of advanced very deep detectors while keeping the real-time speed. Code is available at <https://github.com/ruinmessi/RFBNet>.

Keywords: Real-time Object Detection; Receptive Field Block (RFB)

1 Introduction

In recent years, Region-based Convolutional Neural Networks (R-CNN) [8], along with its representative updated descendants, e.g. Fast R-CNN [7] and Faster R-CNN [26], have persistently promoted the performance of object detection on major challenges and benchmarks, such as Pascal VOC [5], MS COCO [21], and ILSVRC [27]. They formulate this issue as a two-stage problem and build a typical pipeline, where the first phase hypothesizes category-agnostic object proposals within the given image and the second phase classifies each proposal according to CNN based deep features. It is generally accepted that in these methods, CNN representation plays a crucial role, and the learned feature is expected to deliver a high discriminative power encoding object characteristics and a good robustness especially to moderate positional shifts (usually incurred by inaccurate boxes). A number of very recent efforts have confirmed such a

^{*} indicates corresponding author (ORCID: 0000-0002-2412-9330).

