

Visual Attention in Multi-Label Image Classification

Yan Luo
University of Minnesota
luox648@umn.edu

Ming Jiang
University of Minnesota
mjiang@umn.edu

Qi Zhao
University of Minnesota
qzhao@cs.umn.edu

Abstract

One of the most significant challenges in multi-label image classification is the learning of representative features that capture the rich semantic information in a cluttered scene. As an information bottleneck, the visual attention mechanism allows humans to selectively process the most important visual input, enabling rapid and accurate scene understanding. In this work, we study the correlation between visual attention and multi-label image classification, and exploit an extra attention pathway for improving multi-label image classification performance. Specifically, we propose a dual-stream neural network that consists of two sub-networks: one is a conventional classification model, and the other is a saliency prediction model trained with human fixations. Features computed with the two sub-networks are trained separately and then fine-tuned jointly using a multiple cross entropy loss. Experimental results show that the new saliency sub-network improves multi-label image classification performance on the MS COCO dataset. The improvement is consistent across various levels of scene clutteredness.

1. Introduction

Multi-label image classification is an essential computer vision task, aiming to recognize scene-level properties of an image from different aspects. Different from the extensively studied single-label image classification problem, multi-label image classification is more common and practical in real-world applications. An arbitrary image is likely to contain multiple objects and diverse information related to different visual and cognitive properties, such as appearance, emotions of human and animal, scene, interaction, viewpoint, scale, occlusion, and illumination. Therefore, one of the key problems in multi-label image classification is to capture the rich semantic information in complex and cluttered scenes [14].

To approach this problem, human visual system has developed a selective attention mechanism that allows us to effectively attend to interesting or important regions in a vi-

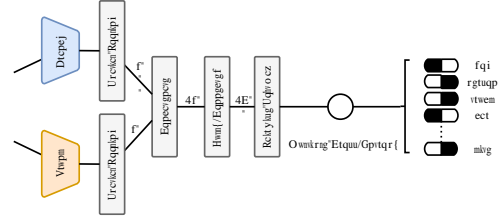


Figure 1: A dual-stream model is proposed to study the effect of visual attention on multi-label image classification. It consists of a sub-network (*i.e.*, trunk) to learn the features for classification while another sub-network (*i.e.*, branch) to be trained to predict image saliency. This model allows to quantify how much attention-related information contribute to multi-label image classification.

ually cluttered world [7]. Computational models of attention predict saliency (*i.e.*, features of importance in a scene) by mimicking such a selective attention mechanism [7]. Visual attention models have been empirically proved to be useful for various computer vision tasks, such as image re-targeting [21], object recognition [25], video compression [3], tracking [16], image captioning [23], and so on. Although there are attempts to incorporate machine attention for multi-label image classification, *e.g.* [27], it is unknown to how human-like visual attention works in the context of multi-label image classification.

The objective of this work is to investigate the use of human-like visual attention in multi-label image classification. We first study the correlation between visual attention (*i.e.*, visual saliency predicting human gaze) and multi-label image classification through statistical analyses. Based on the analyses, we propose a dual-stream model to utilize human visual attention in the task of multi-label image classification. It consists of a sub-network that learns discriminative features for classification and another sub-network that learns saliency features for predicting human gaze. The proposed dual-stream model would yield its prediction based on the two types of features.

The contributions of this work are summarized as follows:

