

Deep Attention Neural Tensor Network for Visual Question Answering

Yalong Bai^{1,2}, Jianlong Fu³, Tiejun Zhao¹, and Tao Mei²

¹ Harbin Institute of Technology, Harbin, China

² JD AI Research, Beijing, China

³ Microsoft Research Asia, Beijing, China

fbaiyalong, tmei@jd.com, jianf@microsoft.com, tjzhao@hit.edu.cn

Abstract. Visual question answering (VQA) has drawn great attention in cross-modal learning problems, which enables a machine to answer a natural language question given a reference image. Significant progress has been made by learning rich embedding features from images and questions by bilinear models, while neglects the key role from answers. In this paper, we propose a novel deep attention neural tensor network (DA-NTN) for visual question answering, which can discover the joint correlations over images, questions and answers with tensor-based representations. First, we model one of the pairwise interaction (e.g., image and question) by bilinear features, which is further encoded with the third dimension (e.g., answer) to be a triplet by bilinear tensor product. Second, we decompose the correlation of different triplets by different answer and question types, and further propose a slice-wise attention module on tensor to select the most discriminative reasoning process for inference. Third, we optimize the proposed DA-NTN by learning a label regression with KL-divergence losses. Such a design enables scalable training and fast convergence over a large number of answer set. We integrate the proposed DA-NTN structure into the state-of-the-art VQA models (e.g., MLB and MUTAN). Extensive experiments demonstrate the superior accuracy than the original MLB and MUTAN models, with 1.98%, 1.70% relative increases on VQA-2.0 dataset, respectively.

Keywords: Visual question answering· Neural Tensor Network· Open-ended VQA

1 Introduction

After deep learning techniques have achieved great success in solving natural language processing and computer vision tasks, automatically understanding the semantics of images and text and eliminating the gap between their representations has received intensive research attention. It has stimulated many new research topic like image captioning [8], text to image synthesis [23] and visual question answering [4, 10].

The Visual Question Answering (VQA) is a task about answering questions which posed in natural language about images. The answers can either be se-

