Introduction

Problem Definition

Visual Question Answering (VQA) aims at answering a natural language question about a given image.

Contributions

1. Propose two novel modules to encode relations between words and between image regions, respectively. This is the first time to explore the relations between words and between image regions in a unified framework for the VQA task.

2. Extensive experiments show the effectiveness of the proposed relation encoding modules. Our approach achieves new state-of-the-art results on the VQA 1.0 dataset.

(a) Visual relation encoding. It encodes the relations between image regions.

(b) Language relation encoding. It encodes multi-scale relations between words.

Approach

Overview of the proposed VQA model

Masked self-attention

Experiments

Comparison with the state-of-the-arts

Ablation studies

More experiments can be found in our paper.

Contact us: Fei Liu (liufei2017@ia.ac.cn)
Jing Liu (jliu@nlpr.ia.ac.cn)
Zhiwei Fang (zhiwei.fang@nlpr.ia.ac.cn)
Hanqing Lu (luhq@nlpr.ia.ac.cn)