Polishing Network for Decoding of Higher-Quality Diverse Image Captions

Yue Zheng, Ya-li Li, Shengjin Wang

Department of Electronic Engineering, Tsinghua University
Beijing National Research Center for Information Science and Technology

Abstract

Existing methods in diverse image caption generation usually adopt a single-pass decoding process, that the sampled words at each time step during decoding will not be modified. A mistaken word could affect the whole subsequent sequence. On the other hand, decoders in single-pass approaches only have access to the previously generated words, thus unable to compose the sentences with an understanding of the whole contents. Inspired by the multi-pass process of human generating descriptions, in this paper we propose a novel framework with a Polishing Network (PN) for decoding diverse image captions. PN refines the raw descriptions generated by an original diverse image caption generation model. The refined sentences could modify some of the incorrect words and phrases in the raw descriptions, while still describing similar content. We also propose a novel approach for training PN. The raw-refined caption pairs used as training samples for PN are obtained by sampling both the input and output words of an original model during decoding. The experimental results show that the proposed approach can generate high-quality diverse image captions, achieving a better quality-diversity trade-off.

Introduction

In the task of diverse image caption generation, a set of descriptions obtained with beam search are usually of high quality and low diversity. While with random sampling methods, a set of descriptions with higher diversity can be generated with low calculation consumption. However, the quality of these descriptions are usually lower, with incorrect words and phrases appearing in the descriptions. In this paper, we propose a novel framework with a polishing network to refine the raw descriptions generated by an original model, thus generating a set of refined descriptions with higher-quality. For example, mistaken words "rice" and "carrots" in the figure can be refined as "noodles" by the polishing network.

Approach

Polishing Network and Multi-pass Decoding

Refine each description \( x \) in the original generated set \( \{ x_1, x_2, ..., x_n \} \). Refined descriptions \( \{ x_1', x_2', ..., x_n' \} \):

\[
 f(x') = \prod_{t=1}^{T} P(x'_t | x'_{<t}, z_{<t}, V_t, R_t, \Theta) \]

Predicted probability of the refined description \( x' \), denoted as \( P(x') \).

We proposed a novel approach for diverse image caption generation with a polishing network, which refines the generated results from an original single-pass method to obtain higher-quality descriptions. A novel training approach is also proposed to generate raw-refined description pairs for training the polishing network. Experiments in diverse image caption generation show that the proposed approach can achieve a better quality-diversity trade-off of descriptions.

Contact: zhengy17@mails.tsinghua.edu.cn, liyali13, wsgj@tsinghua.edu.cn