

Cross-Modal Person Search: A Coarse-to-Fine Framework using Bi-directional Text-Image Matching

Xiaojing Yu¹, Tianlong Chen¹, Yang Yang², Michael Mugo², Zhangyang Wang¹
¹Texas A&M University, ²Walmart Technology

{vicky_yu, wiwjp619, atlaswang}@tamu.edu

{yang.yang2, michael.mugo}@walmart.com

Abstract

Searching person images from a gallery based on natural language descriptions remains to be a challenging and under-explored cross-modal retrieval problem. To improve the accuracy off an image-based retrieval task, e.g., person re-identification (Person Re-Id), re-ranking is known to be an effective post-processing tool. In this paper, we extend re-ranking from uni-modal retrieval to cross-modal retrieval for the first time, and develop a bi-directional coarse-to-fine framework (BCF) for cross-modal person search. Built on a recent state-of-the-art Person Re-Id model [5], BCF exploits first text-to-image and then image-to-text relevance, in a two-stage refinement fashion. BCF ranks competitively against a strong baseline[24] on the newly-introduced WIDER Person Search dataset [1], boosting validation set performance by 9.01%(top-1)/3.87%(mAP) for val1 and 6.60%(top-1)/3.49%(mAP) for val2, respectively. With a high score, our solution ranks competitively in the **ICCV 2019 WIDER Person Search by Language Challenge**.

1. Introduction

Searching person by natural language descriptions is an important application instance of cross-modal retrieval. Given a textual description of a specific person, its objective is to find images from gallery which best match the description. Based on existing methods for text-to-image retrieval, significant challenges remain to be addressed. For improvements, our work is motivated by multi-fold observations:

- **Cross-modality gives rise to (more) ambiguity:** In addition to the semantic ambiguities in either modality, it is typically unrealistic to assume one-to-one image-text mapping. For example, in the newly-introduced WIDER Person Search dataset [1], one person ID can have multiple gallery images, which means that for one

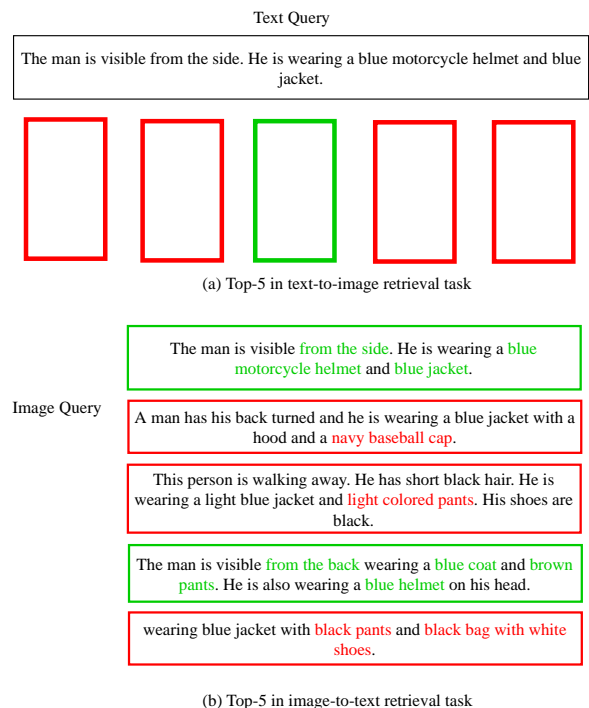


Figure 1. Top-5 results in text-to-image retrieval and image-to-text retrieval. The image/text with green border indicates a correct match. Using text and image embedding extracted with the same image-text matching model, image-to-text retrieval is consistently more accurate than text-to-image retrieval.

textual description, there could be multiple “ground-truth” image-text matching pairs.

- **Bidirectional matching is more reliable:** Intuitively (and empirically observed by us), if an image and a text are found to match each other in both directions (i.e., text-to-image, and image-to-text), then they are more likely to make a correct match. Additionally, we observe the image-to-text retrieval to be often more accurate than the other way around, e.g., in Figure 1.

Category	Percentage
Very good	10%
Good	30%
Not good	40%
Very bad	20%

