DeepDiary: Lifelogging Image Captioning and Summarization

Chenyou Fan, Zehua Zhang, David J. Crandall

PII: S1047-3203(18)30103-2
DOI: https://doi.org/10.1016/j.jvcir.2018.05.008
Reference: YJVCI 2188

To appear in: *J. Vis. Commun. Image R.*

Please cite this article as: C. Fan, Z. Zhang, D.J. Crandall, DeepDiary: Lifelogging Image Captioning and Summarization, *J. Vis. Commun. Image R.* (2018), doi: https://doi.org/10.1016/j.jvcir.2018.05.008

DeepDiary: Lifelogging Image Captioning and Summarization

Chenyou Fan, Zehua Zhang, David J. Crandall

School of Informatics, Computing, and Engineering
Indiana University
Bloomington, Indiana USA

Abstract

Automatic image captioning has been studied extensively over the last few years, driven by breakthroughs in deep learning-based image-to-text translation models. However, most of this work has considered captioning web images from standard data sets like MS-COCO, and has considered single images in isolation. To what extent can automatic captioning models learn finer-grained contextual information specific to a given person's day-to-day visual experiences? In this paper, we consider captioning image sequences collected from wearable, life-logging cameras. Automatically-generated captions could help people find and recall photos among their large-scale life-logging photo collections, or even to produce textual "diaries" that summarize their day. But unlike web images, photos from wearable cameras are often blurry and poorly composed, without an obvious single subject. Their content also tends to be highly dependent on the context and characteristics of the particular camera wearer. To address these challenges, we introduce a technique to jointly caption sequences of photos, which allows captions to take advantage of temporal constraints and evidence across time, and we introduce a technique to increase the diversity of generated captions, so that they can describe a photo from multiple perspectives (e.g. first-person versus third-person). To test these techniques, we collect a dataset of about 8,000 realistic lifelogging images, a subset of which are annotated with nearly 5,000 human-generated reference sentences. We evaluate the quality of image captions both quantitatively and qualitatively using Amazon Mechanical Turk, finding that while these algorithms are not perfect, they could be an important step towards helping to organize and summarize lifelogging photos.

Keywords: Lifelogging, first-person, image captioning, diary, privacy.

Email addresses: fan6@indiana.edu (Chenyou Fan), zehzhang@indiana.edu (Zehua Zhang), djcran@indiana.edu (David J. Crandall)