Doctoral Thesis

A Study on High-Level Cognitive Understanding of Images towards Language

Bei LIU

September 2018

Department of Social Informatics
Graduate School of Informatics
Kyoto University

Doctoral Thesis
submitted to Department of Social Informatics,
Graduate School of Informatics,
Kyoto University
in partial fulfillment of the requirements for the degree of
DOCTOR of INFORMATICS

Thesis Committee: Masatoshi Yoshikawa, Professor
Takayuki Kanda, Professor
Shinsuke Mori, Professor

A Study on High-Level Cognitive Understanding of Images towards Language*

Bei LIU

Abstract

Nowadays, a large number of images are flooding our local memory and social image sharing websites as a result of camera and smart phone's popularization. Images are becoming a more used way for expression compared with past days. Meanwhile, language, as another important form of communication, is also attracting our attention for research. The interaction between image and language is not unexplored yet though we humans perform so many tasks that involve both modality. Though low-level cognitive (e.g. facts) understanding of images is largely tackled and has achieved great success, high-level cognitive understanding of images still remains a challenge.

In this research, we explore the importance of high-level cognitive understanding towards language from two types of tasks: search-based problems and generation-based problems. Different forms of languages are involved in these tasks, including words of event, words of subjective adjective, stories and poems.

To bridge images and events, we tackle the problem of event summarization from images, which aims to retrieve images to represent an event with high perceptual quality. Instead of directly searching for related images of a certain event, we propose to find images that cannot misrecognized as its neighbor events, which we define three types, namely sub-event, super-event and sibling-event. We analyze the reasons of these misrecognitions and propose method to prevent from them accordingly.

In the research of learning subjective adjectives from images, we propose to distinguish relevant and irrelevant images in weakly-labeled data with a pairwise stacked convolutional auto-encoder that can learn discriminative features by identifying a dominant difference between them. We define pseudo-relevant and pseudo-irrelevant image sets as results obtained from image search engines with query with or without the subjective adjective.

*Doctoral Thesis, Department of Social Informatics, Graduate School of Informatics, Kyoto University, KU-I-DT6960-26-0400, September 2018.

To generate a story from a sequence of images towards human cognition, we take emotion as an important factor that guides the generation of a story. The task is formulated into two correlated tasks: generating sentences based on both visual contents and emotions, and predicting possible emotions from images considering contextual images in a sequence. An emotion conditioned story generation model is proposed to guide image embedding learning and story decoder, while a RNN-based prediction model is proposed to learn emotions of each image in a sequence considering contextual images.

The task of poetry generation is challenging as writing a poem involves multiple principles. The difficulties mainly drive from discovering the poetic clues from an image (e.g., rose for love), and generating poems to satisfy both relevance to images and the poeticness in language. We formulate the task of poem generation into two correlated sub-tasks by multi-adversarial training via policy gradient, through which the cross-modal relevance and poetic language style can be ensured. To convey the poetic clues from poems, we propose to learn a deep coupled visual-poetic embedding, in which the poetic representation for objects, sentiments and scenes from images can be jointly learned. Two discriminative networks are further introduced to guide the poem generation, including a multi-modal discriminator and a poem-style discriminator.

To draw a brief conclusion, the work presented in this thesis have made the following progress. (1) We proposed to analyze and understand images from the high-level cognitive perspective; (2) We devised novel models and algorithms from two types of tasks: search-based and generation-based; (3) Our work were verified with extensive experiments. Encouraging results were obtained by comparison with state-of-the-art baselines.

Keywords: Image, Language, High-Level, Cognitive Understanding.