Tutorial 6: Vision and Text: Search, Generation and Translation

Today's digital contents are inherently multimedia: text, image, audio, video and so on. Visual data (image and video) and text, in particular, have been generated, published and spread explosively, becoming the fastest growing data resource. Thanks to the recent deep learning techniques, researchers in both image/video processing and computer vision areas are striving to bridge the two worlds of vision and text. This trend encourages the development of technological advances, which could facilitate a broad range of multimedia analysis applications. After a brief history review of the connection between vision and text, this tutorial will first present the approaches in the direction of vision to text, including cross-view (e.g., text-image) embedding, image/video captioning, dense captioning and visual question answering. During the second part of this tutorial, the presentation will focus on the research along the dimension of text to vision, which span a broad range of scenarios from image generation, image generation from text, scene graph generation, to paragraph generation, as well as image-to-image translation and video-to-video translation. Finally, the synergy of visual understanding and text processing on dealing real problems, e.g., fashion search and recommendation, fashion trend prediction and fashion content generation, are discussed and practical instructions are shared for these emerging research areas.

Speakers:

Ting Yao
JD AI Research, Beijing, China

Wei Zhang
JD AI Research, Beijing, China

Wen-Huang Cheng
National Chiao Tung University, Taiwan