Towards Joint Understanding of Images and Language

Svetlana Lazebnik, University of Illinois at Urbana-Champaign

From robotics to human-computer interaction, numerous real-world tasks can benefit from practical systems that can identify objects in scenes based on language and understand language grounded in visual context. This presentation will focus on joint neural models for images and language. I will talk about methods for learning cross-modal embeddings for text-to-image and image-to-text search, and about the challenging task of grounding or localizing of textual mentions of entities in an image. I will also discuss applications such as automatic image description and visual question answering.