ICVSS 2018

Sicily ~ 08-14 July

International Computer Vision Summer School

Multimodal self-supervised learning: learning to see and hear

Antonio Torralba

Massachusetts Institute of Technology and IBM, USA

Abstract

One of the key reasons for the recent successes in computer vision is the access to massive annotated datasets that have become available in the last few years. Unfortunately, creating these datasets is expensive and labor intensive. On the other hand, humans do not require massive annotated datasets in order to learn to perceive the world. In fact, babies learn with very little supervision, and, even when supervision is present, it comes in the form of an unknown spoken language that also needs to be learned. How can kids make sense of the world? In this talk, I will show that an agent that has access to multimodal data (like vision and audition) can use the correlation between images and sounds to discover objects in the world without supervision. I will show that ambient sounds can be used as a supervisory signal for learning to see and vice versa (the sound of crashing waves, the roar of fast-moving cars – sound conveys important information about the objects in our surroundings). I will also show how we can use raw speech descriptions of images to jointly learn to segment words in speech and objects in images without any additional supervision.