Research Overview

Wentao Zhu

My research interest mainly lies on revealing the underlying structure of visual world. Following this idea, I explored some topics such as learning representation, deep learning interpretability, and multi-modal learning. Currently I work with the joint laboratory of CUHK Multimedia Lab and SenseTime Research, advised by Prof. Bolei Zhou. Here are some projects that I've been working on:

Measuring Disentanglement of Semantic Coding in Deep Visual Representations

The foundation of representation learning has been described as "*to identify and disentangle the underlying explanatory factors hidden in the observed milieu of low-level sensory data*". For deep neural networks, studying this process naturally boils down to analyzing the role of individual units. Previous work suggested that Object Detectors Emerge in Deep Scene CNNs and measured the interpretability of visual representations using Network Dissection.

We demonstrate that deep neural networks disentangle the hidden factors in recognition tasks with interpretable individual units even if there's no explicit supervision. To quantitatively evaluate the degree of disentanglement, we introduce metrics based on unit representations. We further show that disentanglement of semantic coding is important for network performance and generalization.

GAN Inversion

Recent advances in Generative adversarial networks (GANs) such as Large Scale GAN Training for High Fidelity Natural Image Synthesis, Progressive Growing of GANs for Improved Quality, Stability, and Variation, A Style-Based Generator Architecture for Generative Adversarial Networks significantly improves the quality of generated images. However, what are actually learned by the generators are still not fully understood, nor can we arbitrarily control the details of generated images.

GAN Inversion, which is to find the optimal input of a generator for a given image, may be important to understanding the mechanism of GANs. First of all, it shows the capacity of learned generators: Is it possible for them to incorporate unseen real images in its solution space? Or they can only 'imitate' training samples? Secondly, by projecting the real image to the latent space, we can further control and manipulate the generated images. Previous work such as Inverting the Generator of a Generative Adversarial Network took a straightforward optimization approach, while we utilize the internal structure of GANs for better, more interpretable inversion.

Pose-guided human video generation

Generating human image/video guided by pose representation is a meaningful task. Previous work on Dance Transfer and Disentangled Person Image Generation all features this module. However, ther's some limitations of current approaches. If a model is tailored for generating video of a specific person, then every user needs a specially trained model, with high computation cost. If a unified model is used, then fine details are unlikely to be preserved.

Furthermore, the consistency and robustness of generated image sequences remains an open problem. Therefore, we are working on generating human video with better pose representation and multi-modal supervision, trying to address the above problems.