GGAN: Graph Guided Adversarial Networks

Rahul Soni, Jie Fu, Tat-Seng Chua
rsoni@u.nus.edu, fujie@lisa.iro.umontreal.ca, chuats@comp.nus.edu.sg

Abstract

Generative Adversarial Networks (GANs) have been popular due to their principal ability to generate sharper images due to adversarial training. The very first proposal on GANs used noise as an input to simulate near ground truth data in dual game between the Generator and the Discriminator.

However, from application standpoint, there is a specific discretion or an attention that is usually desired in the synthesized output. For example, a given relationship between objects in the image, or a given attention in the specific region of medical diagnostic task. We propose a novel graph guided adversarial training that explicitly uses the semantic relationship between multiple objects in the images (with a particular target in mind) and guides the generation process conditionally.

We address the following key challenges incurred in the process: (1) Define a 'correct' target data that we would like the generator to simulate, (2) Extract graph (adjacency matrix) representing the relationship between objects in the image, (3) Augment the graph representation to learn a Generator jointly on the condition data and the intended target graph, (4) Calibrate the Discriminator to learn to distinguish a pair (real versus fake) rather that just one input to avoid identity learning.

Background

In the most recent work on conditional GANs, Liqian et al., uses pose of the person (main object of the image) as a condition to train the network. Specifically, they use state-of-the-art pose estimator to estimate approximate pose of the human body as key point estimator. In order to make the keypoints compatible with the Generator network training, they augment the descriptors as heatmaps and generate a "pose image" of the same dimension as the input. This is concatenated with the condition image along the channel dimension. At learning stage, generator takes as input the condition image and the pose image (and a target pose along channel dimension) instead of random noise, the discriminator learns to distinguish between real pairs and fake pairs as illustrated in the figure above.

Methodology

The work is limited by availability of the dataset that suits our requirements, for example two images with the same number of objects but different alignment. Hence, we propose to test the model on video dataset by extracting frames in successive time stamps.

Graph Augmentation

Given an adjacency matrix of the targeted image, we start by finding the principal components holding most of the information using SVD.

We then select the top-k eigen vectors and for successive eigen vector pair, we project the adjacency matrix onto the plane formed by the two (orthogonal) eigen vectors.

We discretize the plane onto a regular 2D grid of the same dimension as the input condition image thus resulting in sequence of 2D histograms along a channel dimension.

We also use embedding technique to associate nodes together with the same function. We take all the techniques proposed above from Palash et al., "Graph Embedding Techniques, Applications, and Performance: A Survey" and implement as per our need.

Model Architecture

Generator is a U-net-like architecture with skip connections, similar to the one defined in Liqiang et al. For discriminator, we use the standard CNN architecture. Both the Generator and Discriminator are conditioned to learn a joint distribution.

Applications / Impact

- Generation of real ground-truth data in Medical diagnostic tasks using attention maps
- Leverage the proposed graph augmentation process to learn graph convolution networks
- Combination of the two above can be used for Explainable Deep Learning Models, for example to automatically generate diagnostic reports using language and visual model alignment

Research Scope

- Current proposal is a purely visual computation. One key research prospect would be to leverage text description of the image to generate spatial graphs.
- On graphs, there is a scope to augment the graph more effectively using the different type of subject-predicate-object relationship. In current setting, we only consider a single parameter that tells if there is an edge or not between a pair of objects.
- Another interesting area would be to leverage the proposed method in filling missing points in the video by synthesizing frames.