Abstract

Generative Adversarial Network (GAN) is a recent concept in deep generative learning wherein two neural networks are pitted against each other in a zero-sum non-cooperative game to match a non-stationary distribution to an intractable stationary distribution. In this thesis GANs are leveraged for two broad purposes; a) designing efficient frameworks for image/video inpainting and b) training deep neural networks with reduced manual annotation requirement.

Semantic inpainting exploits the generative modeling capability of GAN and refers to realistically filling up large holes on a image or video frame. A promising, yet unexplored approach is to first train a generative model to map a latent noise distribution to natural image manifold and during inference time, search for the *'best-matching'* noise vector to reconstruct the signal. The primary aversion towards this genre is due to its inference time iterative optimization and lack of photo-realism at higher resolution.

In this thesis both of the above mentioned shortcomings are addressed. This is made possible with a nearest neighbor search based initialization (instead of random initialization) of the core iterative optimization involved in the framework. The concept is extended for videos by temporal re-use of solution vectors. Significant speedups of about 4.5-5$\times$ on images and 80$\times$ on videos is achieved. Simultaneously, the method achieves better spatial and temporal reconstruction qualities.

Next, the following question is addressed- *'Do we at all need iterative inference framework ?'*— to answer this, a data driven parametric network is trained to directly predict a matching prior for a given masked image. This converts an iterative paradigm to a single feed forward inference pipeline with around 800$\times$ speedup. Finally, recent advancements in high resolution GAN training are leveraged to scale inpainting network for higher resolution.

Contemporary deep generative model based inpainting networks suffer from massive computational overhead and are thus often impractical to be deployed on resource constrained platform like a mobile phone. In this thesis, as an effort towards designing efficient inpainting frameworks, lightweight low-level modules are introduced to realize computationally cheaper variants of state-of-the-art networks. This results in around 90$\times$ saving in parameters and FLOPS (floating point operations) with significant boost in inference speed on low-end mobile devices without degrading visual quality.

The second part of the thesis focuses on leveraging GANs for training deep neural networks under limited annotation constraints. Two genres of approaches are studied— semi supervised learning and domain adversarial learning.

For semi-supervised learning the basic GAN framework is extended for a multi-task learning to learn representations from abundance of unlabeled data along with a few labeled data. This is helpful in scenarios where getting annotated data is difficult. The idea is demonstrated on the use case of segmenting retinal blood vessel from fundus images. Experiments suggest that GAN based semi-supervised training appreciably helps over naive supervised training, especially under extreme low annotation.

For domain adversarial learning, the thesis focuses on unsupervised domain adaptation, which, in general, assumes presence of annotation in one domain (source) but absence of labels in another closely related domain (target). Firstly, a general framework is presented for training networks on simulated graphics images but applied on real-life samples. Particularly, the use case of eye gaze estimation is studied in which synthetic annotated data is readily available from Unity game engine; but in order to run inference on gaze data from real consumer cameras, features of these two domains are aligned by domain discriminator playing a zero-sum game with the feature alignment network. Experiments suggest that for such constrained objects such as eye images, feature-level alignment is an better alternative to pixel-level alignment. However, feature-level alignment does not work appreciably on *'in-the-wild'* cases such as object detection on complex scenes. As an example, the following task is investigated— *'Can we transfer image object detectors to videos without annotating videos?'*. For this, a *image-to-image* translation based GAN framework is trained to create *pseudo-video* datasets from images but re-using annotations of image dataset. Such cross domain adaptation significant improves detection performance on video frames compared to models which are just trained on images and applied directly on videos.