

Generative Modelling and Adversarial Learning



Chaoyue Wang

Faculty of Engineering and Information Technology

University of Technology Sydney

A thesis submitted for the degree of

Doctor of Philosophy

2018

I dedicate this thesis to my loving families

Hongzhen Wang and *Wenru Wang*

Yanbin Wang and *Xiangfa Wu*

and *Chenlu Li*

Certificate of Original Authorship

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Chaoyue Wang

Acknowledgements

I would like to thank everyone who has helped me to finish my doctoral studies.

First of all, I would like to express my sincere gratitude to my supervisor Prof. Dacheng Tao for his continuous support of my Ph.D study. Thanks for his consistent patience and motivation, for his encouraging attitude and expert knowledge for my research. His strict academic attitude and diligent work style have played a role model for me and will continue to benefit me through my life. It is no exaggeration to say without his help steering my research direction, I would not have finished this thesis so smoothly and on time. Also I want to thanks for his advice and help on my career development.

Besides my principal supervisor, I would like to thank Dr. Chang Xu and Prof. Chaohui Wang. I want to thanks for them never bored discussion with my seemingly endless simple questions from research motivation, model development, algorithm implementation, and paper drafting. With these specific and detailed technical discussion, I have gained practical skills to effectively and efficiently develop and implement my research problems.

I also wish to give special thanks to Ying Wu, Dongang Wang, Guoliang Kang, Dayong Tian, Yuxuan Du, Xinyuan Chen, Dalu Guo, Huan Fu, Xiyu Yu, Jue Wang, Jiang Bian, Yali Du, Shan You, Jianfeng Dong, Liu Liu, Baosheng Yu, Zhe Chen, Jiayan Qiu and Erkun Yang. Without their helps, my life in Sydney would not be so easy. I would like to give my gratitude to Changxing Ding, Maoying Qiao, Shaoli Huang and Tongliang Liu for their helps on my research.

Last but not the least, my gratitude extends to my family who have been patiently encouraging and waiting for the finish of this thesis.

Abstract

A main goal of statistics and machine learning is to represent and manipulate high-dimensional probability distributions of real-world data, such as natural images. Generative adversarial networks (GAN), which are based on the adversarial learning paradigm, are one of the main types of methods for deriving generative models from complicated real-world data. GAN and its variants use a generator to synthesise semantic data from standard signal distributions and train a discriminator to distinguish real samples in the training dataset from fake samples synthesised by the generator. As a confronter, the generator aims to deceive the discriminator by producing ever more realistic samples. Through a two-player adversarial game played by the generator and discriminator, the generated distribution can approximate the real-world distribution and generate samples from it.

This thesis aims to both improve the quality of generative modelling and manipulate generated samples by specifying multiple scene properties. A novel framework for training GAN is proposed to stabilise the training process and produce more realistic samples. Unlike existing GANs, which alternately train a generator and a discriminator using a pre-defined adversarial objective function, different adversarial training objectives are utilised as mutation operations and train a population of generators to adapt to the environment (i.e. the discriminator). The samples generated by different iterations of generators are evaluated and only well-performing generators are preserved and used for further training. In this way, the proposed framework overcomes the limitations of an individual adversarial training objective and always preserves the best offspring, contributing to the progress and success of GANs.

Based on the GANs framework, this thesis devised a novel model, called a perceptual adversarial network (PAN). The proposed PAN consists of two feed-forward convolutional neural networks: a transformation network and a discriminative network. Besides generative adversarial loss, which is widely used in GANs, this thesis proposes to employ perceptual adversarial loss, which undergoes adversarial training between the transformation network and hidden layers of the discriminative network. The hidden layers and output of the discriminative network are upgraded to constantly and automatically discover discrepancies between a transformed image and the corresponding ground truth, and the image transformation network is trained to minimise the discrepancy identified by the discriminative network.

Furthermore, to extend the generative models to perform more challenging re-rendering tasks, this thesis explores disentangled representations encoded in real-world samples and proposes a principled tag disentangled generative adversarial network for re-rendering new samples of the object of interest from a single image by specifying multiple scene properties. Specifically, from an input sample, a disentangling network extracts disentangled and interpretable representations, which are then used to generate new samples using the generative network. In order to improve the quality of the disentangled representations, a tag mapping net determines the consistency between the image and its tags.

Finally, experiments with different challenging datasets and image synthesis tasks demonstrate the good performance of the proposed frameworks regarding the problem of interest.

Contents

Contents	x
List of Figures	xiii
1 Introduction	1
1.1 Background	1
1.2 Related Works	7
1.3 Generative adversarial networks (GAN)	8
1.4 Variants of GANs	9
1.5 Summary of Contributions.	9
1.6 Thesis Structure.	12
2 Stabilizing adversarial training process via selecting optimization	14
2.1 Introduction	15
2.2 Related Works	16
2.2.1 Generative Adversarial Networks	17
2.2.2 Evolutionary Algorithms	18
2.3 Method	18
2.3.1 Generative Adversarial Networks	18
2.3.2 Evolutionary Algorithm	20
2.3.3 Mutations	21
2.3.4 Evaluation	24
2.3.5 E-GAN	27
2.4 Experiments	28

CONTENTS

2.4.1	Implementation Details	28
2.4.2	CIFAR-10 and Inception Score	32
2.4.3	LSUN and Architecture Robustness	33
2.4.4	Synthetic Datasets and Model Collapse	35
2.4.5	CelebA and Space Continuity	37
2.5	Summary	37
3	Learning perceptual information through adversarial paradigm	39
3.1	Introduction	40
3.2	Related works	44
3.2.1	Image-to-image transformation with feed-forward CNNs	44
3.2.2	GANs-based works	44
3.2.3	Perceptual loss	46
3.3	Method	46
3.3.1	Generative adversarial loss	46
3.3.2	Perceptual adversarial loss	48
3.3.3	The perceptual adversarial networks	51
3.3.4	Network architectures	52
3.4	Experiments	53
3.4.1	Experimental setting up	53
3.4.2	Evaluation metrics	57
3.4.3	Analysis of the loss functions	57
3.4.4	Comparing with existing works	61
3.5	Summary	68
4	Interpretable and disentangled representations in adversarial learning	69
4.1	Introduction	70
4.2	Related Works	72
4.3	Method	73
4.3.1	Main Framework	73
4.3.2	Training Process	76
4.3.3	Image Re-rendering	78

CONTENTS

4.4	Experiments	79
4.4.1	Implementation Details	79
4.4.2	Performance Criteria	80
4.4.3	Experimental Results	82
4.5	Summary	89
5	Conclusions	90
	Appendix A	92
5.1	Proof of the Optimal Discriminator	92
5.2	Network Architectures	93
5.3	Synthetic Datasets	94
5.4	Cifar-10	96
5.5	LSUN Bedrooms	97
5.6	CelebA Faces	103
	References	105

List of Figures

1.1	The training process of generative adversarial networks (GANs). An adversarial game is performed between a generator and a discriminator.	3
2.1	(a) The original GAN framework. A generator G and a discriminator D play a two-player adversarial game. The updating gradients of the generator G are received from the adaptive objective, which depends on discriminator D . (b) The proposed E-GAN framework. A population of generators $\{G_\theta\}$ evolves in a dynamic environment, the discriminator D . Each evolutionary step consists of three sub-stages: variation, evaluation, and selection. The best offspring are kept.	19
2.2	The mutation (or objective) functions that the generator G receives given the discriminator D	23
2.3	Experiments on the CIFAR-10 dataset. CIFAR-10 inception score over generator iterations (left), over wall-clock time (right).	26
2.4	Experiments on the CIFAR-10 dataset. The graph of selected mutations in the E-GAN training process	28
2.5	KDE plots of the target data and generated data from different GANs trained on mixtures of Gaussians.	29
2.6	Generated samples on 128×128 LSUN bedrooms.	30
2.7	Experiments to test architecture robustness. Different GAN architectures corresponding to different training challenges and trained with five different GAN methods.	31
2.8	Keep different numbers of candidatures.	33

LIST OF FIGURES

2.9	Generated human face images on the 128×128 CelebA dataset. .	34
2.10	Interpolating in latent space. For selected pairs of the generated images from a well-trained E-GAN model, we record their latent vectors z_1 and z_2 . Then, samples between them are generated by linear interpolation between these two vectors.	36
3.1	Image-to-image transformation tasks. Many tasks in image processing, computer graphics, and computer vision can be regarded as image-to-image transformation tasks, where a model is designed to transform an input image into the required output image. We proposed Perceptual Adversarial Networks (PAN) to solve the image-to-image transformation between paired images. For each pair of the images we demonstrated, the left one is the input image, and the right one is the transformed result of the proposed PAN. . . .	42
3.2	PAN framework. PAN consists of an image transformation network T and a discriminative network D . The image transformation network T is trained to synthesize the transformed images given the input images. It is composed of a stack of Convolution-BatchNorm-LeakyReLU encoding layers and Deconvolution-BatchNorm-ReLU decoding layers, and the skip-connections are used between mirrored layers. The discriminative network D is also a CNN that consists of Convolution-BatchNorm-LeakyReLU layers. Hidden layers of the network D are utilized to evaluate the perceptual adversarial loss, and the output of the network D is used to distinguish transformed images from real-world images.	47
3.3	Comparison of snow-streak removal using different losses functions. Given the same input image (leftmost), each column shows results trained under different losses. The loss function of ID-CGAN [178] combined the pixel-wise loss (least squares loss), cGANs loss and perceptual loss, i.e., $L_2 + \text{cGAN} + \text{perceptual}$. For better visual comparison, zoomed versions of the specific regions-of-interest are demonstrated below the test images.	49

LIST OF FIGURES

3.4	Transforming the semantic labels to cityscapes images use the perceptual adversarial loss. Within the perceptual adversarial loss, a different hidden layer is utilized for each experiment. For better visual comparison, zoomed versions of the specific regions-of-interest are demonstrated below the test images. For higher layers, the transformed images look sharper, but less color information is preserved.	50
3.5	Comparison of transforming the semantic labels to facades images by controlling the hyper-parameter θ . Given the same input image (leftmost), each column shows results trained under different θ . For better visual comparison, zoomed versions of the specific regions-of-interest are demonstrated below the test images.	53
3.6	Comparison of rain-streak removal using the ID-CGAN with the proposed PAN on real-world rainy images. For better visual comparison, zoomed versions of the specific regions-of-interest are demonstrated below the test images.	56
3.7	Comparison of image in-painting results using the Context-Encoder(CE) with the proposed PAN. Given the central region missed input image (leftmost), the in-painted images and the ground-truth are listed on its rightside.	58
3.8	Comparison of transforming the semantic labels to cityscapes images using the pix2pix-cGAN with the proposed PAN. Given the semantic labels (leftmost), the transformed cityscapes images and the ground-truth are listed on the rightside.	59
3.9	Comparison of transforming the object edges to corresponding images using the pix2pix-cGAN with the proposed PAN. Given the edges (leftmost), the generated images of shoes and handbags are listed on the rightside.	63
3.10	Comparison of some other tasks using the pix2pix-cGAN with the proposed PAN. In the first row, semantic labels are generated based on the real-world cityscapes images. And, the second row reports the generated maps given the aerial photos as input.	65

LIST OF FIGURES

4.1	Model architecture. <i>TD-GAN</i> is composed of four parts: a tag mapping net g , a disentangling network R , a generative network G and a discriminative network D . (a) During training, the tag mapping net g and the generative network G are trained to render images with their tags. The disentangling network R aims to extract disentangled representations, which can be decoded by the network G . The discriminative network D plays a minimax game with networks G and R based on the adversarial training strategy. (b) During test, the disentangling network R extracts disentangled representations from the input image. After replacing one or multiple disentangled representations with the specified representations generated by the tag mapping net g , the image can be re-rendered through the generative network G	74
4.2	Novel view synthesis results of two previous method and ours. For each method, the leftmost image is the input image, and the images on its right side were re-rendered under different viewpoints. . . .	81
4.3	Novel view synthesis results of <i>TD-GAN</i> trained in three settings. The images are arranged similarly to Fig. 4.2.	83
4.4	<i>MSE</i> of <i>TD-GAN</i> trained in three settings. <i>MSE</i> was calculated over all the chair images under 0° in the test set as inputs. Chair images are used to indicate target viewpoints.	84
4.5	Illumination transformation of human face (best view in color). Given an image (leftmost) of human face, we reported a set of re-rendered images on its right side.	85
4.6	Multi-factor transformation (best view in color). Given a single image as input (the up-left one), its viewpoint and expression were jointly transformed.	87
4.7	Multi-factor transformation (best view in color). Given a single image as input (the up-left one), its viewpoint and expression were jointly transformed.	88
5.1	Network architectures for generating 128×128 images.	93

LIST OF FIGURES

5.2	Different GANs learning a mixture of 8 Gaussians arranged in a circle.	94
5.3	Different GANs learning a mixture of 25 Gaussians arranged in a grid.	95
5.4	Generated samples on CIFAR-10 dataset.	96
5.5	Generated images on 128×128 LSUN bedrooms.	97
5.6	Method: DCGAN	98
5.7	Method: LSGAN	99
5.8	Method: WGAN	100
5.9	Method: WGAN-GP	101
5.10	Method: E-GAN	102
5.11	Generated human face images on 128×128 CelebA dataset. . . .	103
5.12	Interpolating in latent space. For selected pairs of the generated images from a well-trained E-GAN model, we record their latent vectors z_1 and z_2 . Then, samples between them are generated by linear interpolation between these two vectors.	104