EECS 6322 Week 11, Paper 1
Modular Generative Adversarial Networks

Shivani Sheth (Student #: 218011783)
shivs29@yorku.ca

I. Summary and contributions

The paper proposes a novel architecture for Generative Adversarial Networks, known as ModularGAN, for multi-domain image-to-image translations and image generation. The architecture consists of composable and reusable models that perform different functions such as transformations, encoding, and decoding. The models used in the architecture can be trained parallely, and can include data from different domains at train time to specific Generative Adversarial Networks at test time. The architecture is thus flexible, and generates/ translates to an image in the required domain at test time without explicitly training the network for that domain again. The architecture also outperforms some of the state of the art models in the multi-domain facial attribute transfer domain, and is the first successful modular GAN architecture.

In the domain of conditional image generation in image translations, generally an image, or an attribute vector is taken as a conditional input, and the output of the model is an image. Most current image translation methods perform image translations only from one domain to another, that is, they only perform a pairwise mapping from one domain to the other. This is due to several shortcomings of the multi-domain image translations such as exponential number of pairwise translation functions, data required from separate domains when learning a particular pairwise mapping, and a complex pairwise translation function. Due to these reasons, the model becomes computationally expensive, thus making the architecture infeasible.

Due to the above mentioned shortcomings of multi-domain image translations, the authors propose a framework that does not train one complex model or pairwise mappings between two domains. Rather, the proposed framework trains on smaller, simpler generative modules that compose to perform complex generative processes. For example, if we aim to transform an image from domain A (man frowning) to domain B (woman smiling), then the proposed architecture first aims to transform the image to the 'female' domain and then transform the resultant image to the 'smiling' doman. Thus, the complexity of the model is decreased since the individual transformations are spatially more local and simpler, and the amount of data in the intermediate domains is larger than the amount of data in the final domain.

The proposed model consists of several different reusable and composable modules that can be combined easily at test time, to translate/ generate an image in different domains. The modules consist of a generator, encoder, reconstructor, transformer and discriminator, that are jointly trained end-to-end. The generator module generates a latent representation of the image from a random noise and an optional condition vector, the encoder module encodes the input image into a latent representation, the transformer module manipulates the latent representation according to the provided conditions, the reconstructor module reconstructs the transformed latent representation to an image, and the discriminator module distinguishes whether the generated/ transformed image looks real or fake. These separate modules perform separate tasks that are combined to form the base of the architecture. New modules can also be added to the architecture and separate existing modules can be upgraded without affecting others. Most importantly, different transformer modules can be composed dynamically and in any order at test time, to form generative networks that apply a sequence of feature transformations to obtain more complex mappings and generative processes. Thus, different transformer modules can be combined to translate an input image to different domains.

ModularGAN focuses on two types of multi-domain tasks which are image generation and image translation. For the image generation task, the goal is to learn a mapping $(z, \boldsymbol{a}) \rightarrow y$, where z is a randomly sampled vector and a is a subset of attributes $\boldsymbol{A}$ to provide control over generated output image y. For the image translation task, the goal is to learn a mapping $(x, \boldsymbol{a}) \rightarrow y$, where x is an image and $\boldsymbol{a}$ are the target attributes to be present in the output image y. The architecture of ModularGAN for the image translation task connects the encoder module E to multiple transformer modules T_i , each of which is further connected to a reconstructor module R to generate the translated image. Multiple discriminator modules D_i are connected to the reconstructor that distinguish the generated images from real images, and make predictions of corresponding attributes. Thus, in the training phase, the output of the model can be seen as a series of transformations on the input image given by: $y = R(T_i(E(x), a_i))$, where x is the input image and a_i are the pre-specified attributes. For the image generation task, the model architecture is similar except the encoder module E is replaced with a generator module G, which generates an intermediate feature map $G(z, a_0)$ from a random noise z and a condition vector a_0 that represents auxiliary information.

A combination of several loss functions have been included in the ModularGAN architecture. The Adversarial Loss is used to make the generated images look realistic, the Auxiliary

Classification Loss is used to predict the i-th attribute of the image, the Cyclic Loss calculate the difference between the reconstructed input image and the input image, and the Full Loss calculates the total loss of the D, E, T, R modules. The module was implemented with several datasets such as ColorMNIST and CelebA, and was compared against the baselines IcGAN, CycleGAN, StarGAN on evaluation metrics such as Classification Error and User Study. To stabilize the training process and to generate images of high quality, the adversarial loss function was replaced with the Wasserstein GAN objective function using gradient penalty.

The qualitative evaluation on ColorMNIST shows that the generator module G and reconstructor module R first generate the correct digit according to the number attribute. Then, the feature representations produced by G are passed through different T_i by which the digit color, stroke style and background of the initially generated image changes. Each module T_i only changes a specific attribute and keeps other attributes untouched. By visualizing the predicted masks in each transformer module T_i, it is noted that the color transformer module T_c mainly changes the interior of the digits, the stroke style transformer module T_s correctly focuses on the borders of the digits, and the background color transformer module T_b has larger values in the background regions. The qualitative evaluation on CelebA shows the transfer between a female face image with neutral expression and black hair to a variety of combinations of attributes. The results show that IcGAN performs worse, followed by CycleGAN and StarGAN. ModularGAN generates the best images amongst all the baselines models. By visualizing the mask of each transformer module it is noticed that while changing a given attribute of the facial feature, the transformer module focuses on different areas of the face according to the given attribute.

In the quantitative evaluation of the datasets, ModularGAN achieves a comparable classification error to StarGAN on the hair color task, and the lowest classification errors on all other tasks. It also obtains the majority of votes for best transferring attributes in all the cases except gender. In the ablation study, it is seen that without mask prediction, the model can still manipulate the images but tends to perform worse on gender, smile and multi-attribute transfer. It is also noticed that removing the cyclic loss does not affect the results of single-attribute manipulation, but the model fails to generate images with desired attributes in multiple transformer modules. The model is also unaffected by the order of transformer modules, which is a desired property.

II. Strengths

The paper proposes a novel modular GAN architecture, known as ModularGAN, that consists of several reusable and composable modules that can be jointly trained end-to-end efficiently. Different (transformer) modules in the proposed model can be combined in order to successfully translate the image to different domains by utilizing the mask prediction within a given module along with the cyclic loss. The model is unaffected by the order of transformer modules, and allows addition of new modules, along with upgradation of separate existing modules without affecting the others. It also outperformers current state of the art results in both qualitative and quantitative results.

III. Weaknesses

As compared to StarGANs, the model performs a little worse in terms of hair color and gender as indicated by the User Study and Classification error.

IV. Correctness

The claims and empirical methodology as proposed by the paper are correct, which can be verified by the comparison studies with the state of the art models. The images generated by the model outperform the state of the art models in mostly all the qualitative and quantitative measures discussed in the paper. The theoretical grounding and relation to prior work also strengthen the claims of the authors since the proposed model is mainly based on the existing GANs model and modular networks.

V. Clarity

The paper is very well written with ample details of the architectural implementation, comparisons and illustrations of different results with the baselines models that makes the paper easy to understand. An ablation study has also been included that indicates the contribution of separate parts of the architecture that supplements the understanding of the model.

VI. Relation to prior work

Prior work in the field of modular networks include models in the Natural Language Processing domain that aim to resolve the task of visual question answering (VQA). These sequence-to-sequence models are implemented using a neural module network that inputs the question as a sequence of words and outputs a program as a sequence of functions. Similar to VQA, the proposed architecture also follows a composition of several two domain translations, but in the multi-domain image generation domain, as compared to the existing models in the NLP domain.

Prior work in the field of Image Translation includes several implementations of Generative Adversarial Networks which consist of a generator and a discriminator module that formulate generative modeling as a game between two competing networks using a minmax objective. In the existing architectures on GANs, most implementations focus on two-domain image translation. Examples of such architectures are the conditional GANs (cGANs), pix2pix, cycle-consistent GANs (CycleGANs), and IcGANs. In the field of multi-domain image translation, a recent architecture on GANs, known as StarGAN uses a single network conditioned on the target domain label. The difference between StarGAN and ModularGAN, the proposed architecture, is that StarGAN learns all domain transformations within a single model, while ModularGAN trains different simple composable translation networks for different attributes.

VII. Reproducibility

There are enough details in terms of code implementation of the model, architectural details while conducting the experiments and theoretical grounding to reproduce the major results of the work proposed by the paper.

VIII. Additional comments

An elaborated discussion on different structures and arrangements of the modular components of the architecture and their effect on the performance of the model can be included in the paper.