U-GAT-IT: UNSUPERVISED GENERATIVE ATTENTIONAL NETWORKS WITH ADAPTIVE LAYER- INSTANCE NORMALIZATION FOR IMAGE-TO-IMAGE TRANSLATION

THE COMP6248 REPRODUCIBILITY CHALLENGE

Banghui Liu *
School of Electronics and Computer Science
University of Southampton
Southampton, UK
bl1m20@soton.ac.uk

Yike Zhang *
School of Electronics and Computer Science
University of Southampton
Southampton, UK
yz5u21@soton.ac.uk

Hanzhong Qi *
School of Electronics and Computer Science
University of Southampton Southampton, UK
hq2n21@soton.ac.uk

ABSTRACT

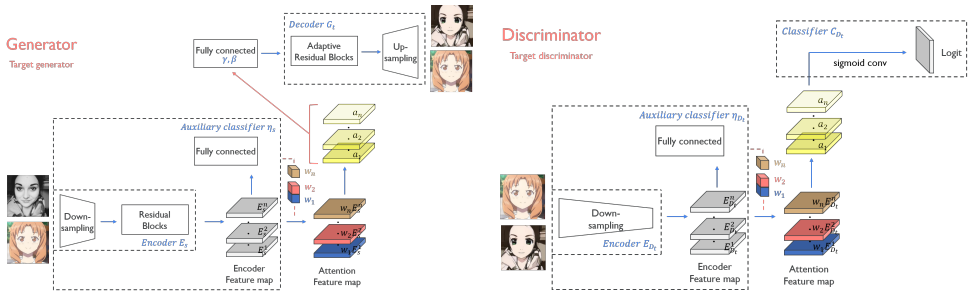
The present work is part of the COMP6248 Reproducibility Challenge. We attempt to reproduce the results in the conference paper *U-GAT-IT: Unsupervised Generative Attention Networks With Adaptive Layer-Instance Normalization For Image-to-Image Matching Translation*. In the area of image conversion, generative adversarial networks can be combined to convert images from one domain to another. When a geometric transformation between domains is involved, an attention mechanism can produce good results, but with performance issues. The authors propose a new attention module that is coupled with a learnable normalization function as a solution to this problem. Since the code provided by the author is relatively complete, we try to modify some of the code, and then mainly evaluate and comment on the author's empirical analysis. Our code is available at: https://github.com/LaniakeaS/U-GAT-IT-reproducibility

1 INTRODUCTION

Due to the limitations of the applicable domain of the image translation model, the performance of image conversion has been poor in the case of large image differences. Consequently, the author proposes U-GAT-IT to implement a multi-task robust image translation model. According to the authors, the key contributions of the new method are:

- An unsupervised method of image-to-image translation is presented with a new attention module and a novel normalization function.
- Based on the attention map obtained by the auxiliary classifier, the attention module assists the model in making dense transitions by identifying the source and target domains.
- By using AdaLIN, the attention-guided model is able to control the amount of change in shape and texture, consequently enhancing the robustness of the model.

*All the authors contributed equally to the reproducibilty challenge. The names are sorted in random order.



(a) Generator (b) Discriminator

Figure 1: Model Structure

2 U-GAT-IT MODEL

The model consists of two sets of generators and discriminators, Figure 1 shows the architecture of the model. The generator takes two images as input, one from the source domain and another from the target domain, and outputs the translated source image and target image to the discriminator as input discriminator, the discriminator will tell the probability of translated image belonging to the source domain.

Both the generator and discriminator have an encoder, a decoder for the generator and a classifier for the discriminator. The attention module is combined into both of them. The attention module inside the generator pays attention to the area where the source and target image have a significant difference. The auxiliary classifier embedded in the encoder is used to learn the weights of the attention map, this is one innovative point of the paper. The calculated attention feature map then is fed into a fully connected layer, which is used to generate parameters γ and β used in AdaLIN(Adaptive Layer Instance Normalization), which is the method firstly proposed by the author, it can balance between layer normalization and instance normalization, and enable the model to control what part and shape the model should focus on.

The attention module embedded in the discriminator pays attention to distinguish the true and fake image in the target domain, there is also an auxiliary classifier inside the discriminator used to learn the parameters of attention weights. Unlike the traditional image translation model, the auxiliary classifier in the discriminator is also used to discriminate the source image, which is inspired by CAM(Zhou et al. (2016)).

3 REIMPLEMENTATION

In this section, we present the details of our reproduction works. It shoud be note that in the paper, several models including CycleGAN(Zhu et al. (2017)), UNIT(Liu et al. (2017)), etc. are implemented for comparison, here we only focus on the U-GAT-IT model. As the paper give the well-formed structure of the model, our implementation work mainly follows the instructions given by the paper.

3.1 METHOD AND DETAILS

3.1.1 CAM AND AUXILLARY CLASSIFIER

The figure 2 shows that the Encoder Feature map for the image obtained by downsampling and residual block gets the feature vector based on the number of channels after Global average pooling and Global max pooling. Make a parameter weight that is compressed to $B \times 1$ dimensions through a fully connected layer where B is the batch size.

To create the Feature map attention mechanism, the author assigns a weight to the learning parameter weight and the corresponding bitwise multiplication of the Encoder Feature map, i.e., to each of the channels of the Encoder Feature map. For the $B \times 1$ dimension obtained through full connection,

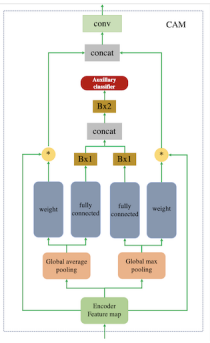


Figure 2: CAM

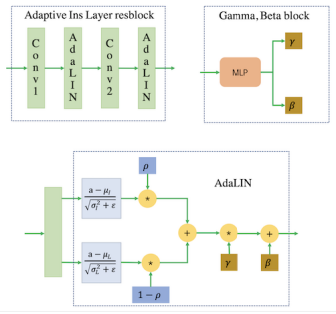


Figure 3: AdaLIN

do concat under average and max pooling, then send it to classification, and make classification judgments for source and target domains.

This is a unsupervised process because we only know the source and target domains. Under CAM global and average pooling, this binary classification problem can be classified well. The attention module can help the model know where to make dense transformations if the generator can distinguish the source and target domain inputs well. The attention map obtained by average and max is used as concat, restored to the number of input channels through a convolutional layer, and then sent to AdaLIN for adaptive normalization.

3.1.2 AdaLIN

Figure 3 illustrates the complete AdaLIN operation. The author designed AdaLIN based on Batch-Instance Normalization, which is a combination of instance normalization and layer normalization. Since the AdaLIN only normalizes the image map, the degree of content structure is preserved and the channels are not correlated.

3.1.3 DISCRIMINATOR

In the source code provided by the author, the discriminator is implemented by combining a Global Discriminator with a Local Discriminator. In contrast, the former performs deeper feature compression on the input image. A CAM module has also been added to the discriminator. However, even though CAM does not perform domain classification under the discriminator, the addition of an attention module is beneficial for determining the authenticity of images. It can be explained that **attention maps help fine-tune performance by focusing on identifying the differences between real and false images in the target domain**.

3.2 MODEL TRAINING

The model is running under with python == 3.6, pytorch == 1.9,1. The dataset we use is *selfie2anime* dataset provided by the author. The size of the training set is 3400, size of the test set is 100. The operation of data preprocessing is the same as that in the paper:

- image horizontal flip probability: 0.5
- resize image to 286 * 286
- randomly crop image to 256 * 256

We follow the parameter setting given by the paper, model is trained using Adam with $\beta 1 = 0.5$, $\beta 2 = 0.999$, $\gamma(learningrate) = 0.0001$, $\lambda(weightdecay) = 0.0001$, the epoch number is set to 100. The trained model is obtained after 100 epochs, the result and analysis will be discussed in the next section.

4 EVALUATION AND RESULTS

Due to page limitations, this paper only shows a reimplementation of comparison of the Kernel Inception Distance for the different settings of the original U-GAT-IT and a test image of the conversion effect. Table 1 demonstrates the advantage of the U-GAT-IT model over the attention module and AdaIN using KID, i.e. the KID obtained when the two are used alone is greater than that of U-GAT-IT. These results are consistent with the findings of the original paper and further confirm the performance improvement of U-GAT-IT.

Model	selfie2anime	anime2selfie
U-GAT-IT	$\mathbf{10.23 \pm 0.42}$	$\mathbf{11.12 \pm 0.52}$
U-GAT-IT w/IN	12.93 $\pm$0.66	14.12 $\pm$0.7
U-GAT-IT w/LN	12.45 $\pm$0.59	13.18 $\pm$0.83
U-GAT-IT w/AdaIN	12.32 $\pm$0.69	12.98 $\pm$0.71
U-GAT-IT w/GN	13.21 $\pm$0.64	12.35 $\pm$0.76
U-GAT-IT w/CAM	12.76 $\pm$0.81	11.86 $\pm$0.96
U-GAT-IT w/G_CAM	11.23 $\pm$0.42	11.56 $\pm$0.52
U-GAT-IT w/D_CAM	11.12 $\pm$0.71	12.12 $\pm$0.72

Table 1: Kernel Inception Distance * 100 $\pm$ std. * 100 for ablation model. The lowest one is best. Reimplementation of Table 1 in original paper.

Figure 4 shows the image of a selected portrait photo from the Internet after being processed by the U-GAT-IT.

(a) (b)

Figure 4: A testing example on selfie to anime translation model.

5 DISCREPANCIES, SUGGESTIONS AND CONCLUSION

Experimental results indicate that the model is reproducible, and both the attention module and AdaLIN exhibit excellent performance on various datasets with fixed architectures and parameters. As a result, AdaLIN can further enhance the robustness of the model under a variety of datasets. For example, the model does not have an advantage with the photo2portrait dataset.

REFERENCES

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.