META-FAVAE: TOWARD FAST AND DIVERSE FEW-SHOT IMAGE GENERATION VIA META-LEARNING AND FEEDBACK AUGMENTED ADVERSARIAL VAE

Fangli Ying*, **Aniwat Phaphuangwittayakul***, **Yi Guo, Xiaoyue Huang, Yue Wang**
East China University of Science and Technology, China
yfangli@ecust.edu.cn, aniwat.pha@gmail.com

ABSTRACT

Learning to synthesis realistic images of new categories based on just one or a few examples is a challenge task for deep generative models, which usually require to train with a large amount of data. In this work, we propose a data efficient meta-learning framework for fast adapting to few-shot image generation task with an adversarial variational auto-encoder and feedback augmentation strategy. By training the model as a meta-learner, our method can adapt faster to the new task with significant reduction of model parameters. We designed a novel feedback augmented adversarial variational auto-encoder. This model learns to synthesize new samples for an unseen category just by seeing few examples from it and the generated interpolated samples are then used in feedback loop to expand the inputs for encoder to train the model, which can effectively increase the diversity of decoder output and prevent the model overfitting with insufficient samples of unseen category. Additionally, with the dual concatenation of latent code and random noise vectors, this method can be generalized to more complex color images compared to existing meta-learning based methods. Experimental results show that our model can have much faster adaption to new generation tasks of unseen categories while generating high-quality and diverse images on three datasets.

1 INTRODUCTION

Following the great success of deep learning, deep generative models, such as Generative Adversarial Network (GAN)(Goodfellow et al. (2014)) and Variational Autoencoder (VAE)(Kingma & Welling (2014)), achieved tremendous progress in the task of generating realistic image by training with large-scale datasets. However, due to the high expense of collecting and annotating a large amount of data, given only one or a few images from a category, it is still a challenging task for the typical deep generative models to learn fast and adapt to synthesis realistic and diverse images. This task is referred to as few-shot image generation(FSIG). Several recent efforts have been devoted to improve the learning efficiency of models for the FSIG task. These approaches can roughly fallen into three categories, that are transformation-based methods, fusion-based methods and optimization-based methods. For the transformation-based methods, DAGAN(Antoniou et al. (2018)) is an GAN-based approach that captures cross-class transformations on one conditional image and injects random vectors into the generator for generating different images. However, the diversity of generated images from this method is quite limited and it fails to capture the information from multiple images in the same category. Among the fusion-based methods, GMN(Bartunov & Vetrov (2018)) is inspired by the matching network(Vinyals et al. (2016)) and combines a matching procedure with VAE. Due to the limited generation capacity of GMN, this method can only generate binary digits. For generating more diversified and realistic images, MatchingGAN(Hong et al. (2020)) combines adversarial learning and matching generators and is capable to fully exploit several conditional images from the same class. Nevertheless, this method still struggles with more complex natural color image generation and training these models is time-consuming due to complexity of the networks. Optimization-based methods are proposed based on the concept of meta-learning and initialize a generator with images from seen categories and fine-tuning the trained

*Equal Contribution

model with images from each unseen category. Clouâtre & Demers (2019) proposed a FIGR by applying meta-learning algorithm Reptile(Nichol et al. (2018)) to GAN. It can significantly reduce the trainable parameters of the model and fast adapt to the new generation task in the few-shot setting. DAWSON Liang et al. (2020) is another method incorporating meta-learning model of MAML Finn et al. (2017) with GAN for the few-shot generation task. Nevertheless, the DAWSON generates images without diversity and is inapplicable for color images. Both models adopted a similar idea of meta-learning method, but can hardly produce sharp and realistic images or adapt to complex natural color images. FAML Phaphuangwittayakul et al. (2021) is another meta-learning inspired method that can significantly reduce the model parameters and fast adapt to generate images from color image dataset. This simple and effective model increased the number of generator iterations and utilized conditional feature vectors. However, given the insufficient input of unseen category, only conditioning the feature vectors from the encoder can not guarantee the exploitation of the key input features. Moreover, training GAN-based model is time-consuming and mode collapse issues are remained, while VAEs have limited capacity of generating high-quality images.

To tackle the above problems, we propose a Meta-learning-based Feedback augmented Adversarial Variational Auto-Encoder (Meta-FAVAE),that learns optimal model parameters to fast adapt to few-shot image generation task and increases the diversity by a feedback augmentation strategy. This model is trained in a meta-learning mechanism with an inner-loop and outer-loop. By integrating meta-learning, this method has less than 10M parameters and thus leads to a faster adaption to the new generation tasks of unseen categories. In order to combine the advances from VAE and GAN (Plumerault et al. (2021)), we introduced an optimal adversarial variational auto-encoder to generate high-quality images while retrieving key latent features. More specifically, the model utilized Relativistic average GAN (Jolicoeur-Martineau (2018)) and beta-VAE (Higgins et al. (2016)) to significantly improve data quality and stability of the generated images with 64×64 size outputs. By introducing a novel feedback augmentation technique, we can significantly increase the diversity of the generated images by adding interpolated samples to the inputs, which can also prevent model collapse. Additionally, we utilize the dual fusion of latent code and random noise vectors and mode-seeking regularization term from FAML for improving the generalization of few-shot image generation task to natural color images. Extensive experiments validate the superiority of our Meta-FAVAE in terms of fast adaptation and diverse image generation over other state-of-the-art methods.

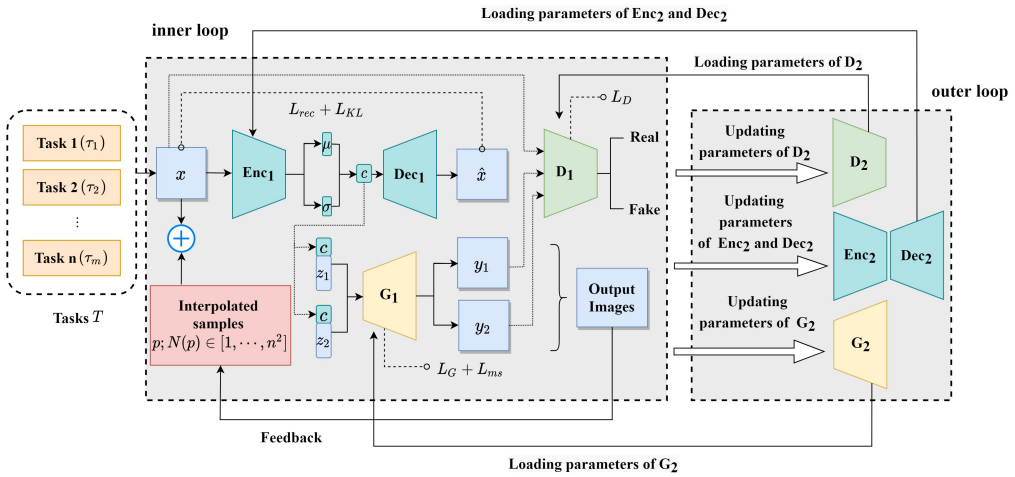

Figure 1: The framework follows a meta-learning mechanism. The inner loop contains an Adversarial-VAE including an encoder Enc_1, a decoder Dec_1, a generator G_1 and a discriminator D_1. The outer loop also has the same components containing an encoder Enc_2, a decoder Dec_2, a generator G_2 and a discriminator D_2. The feedback augmentation takes the generated images from generator G_1 in the inner loop and perform a feedback loop operation to add the interpolated samples into the inputs of Enc_1.

2 METHODOLOGY

Figure 1 illustrates the pipeline of proposed model in this study. Given a set of tasks $\{T_i\}_{i=1}^{m}$ containing various tasks where each task τ is an image generation for one class of images (X_τ). The model can generate a set of images for the unseen classes. This framework is composed of two main phase, which are training phase and generation phase. During the inner loop of the training phase, the encoder Enc_1 of VAE extracts mean μ and variance σ from images x of the training task to form a latent code c. Then, the decoder Dec_1 of VAE reproduces images $\hat{x}$ from the latent code. The generator G_1 takes the latent code concatenated with two random noise vectors (z_1 and z_2) as the input and generates output images, $y_1 = G_1(z_1, c)$ and $y_2 = G_1(z_2, c)$. The discriminator network D_1 distinguishes the real images (input x) and fake images (output samples y from generator network G_1). The Adam optimizer Kingma & Ba (2014) is used to optimize the encoder Enc_1, decoder Dec_1, generator G_1, and discriminator D_1 networks in the loss computation. The objective functions to calculate the loss of Enc_1, Dec_1, G_1, and D_1 are defined as follows.

Instead of training the model with only input images as traditional method, we increase the number of training images by fusing the interpolated samples. The interpolated samples are produced based on the images generated by our model. The linear interpolation is performed based on the four generated images in each episode. The linear interpolation is a technique to visualize the data in a subspace. With two images multiplied by the interpolation coefficients ranging in [0.1,0.9], the new examples between these two images can be explored in the space of generated images. The total number of interpolated samples are equal to loop n^2 , where $n > 0$. For example, a total of 16 interpolated images with $n = 4$ are produced by random images generated by the latent code of Meta-AVAE and four random noise vectors. Once the interpolated images are generated, these images are used to combine with the input images and step the model with the inner loop gradient one more episode.

To train VAE for extracting the latent code and producing reconstructed images, a reconstruction loss and a Kullback-Leibler (KL) divergence Larsen et al. (2016) loss are computed as equation (1).

$$\mathcal{L}_{VAE} = \frac{1}{g}\sum(x - \hat{x})^2 + \beta \mathrm{KL}(q(c|x) \parallel p(c)) \tag{1}$$

where g denotes the number of images sampled from a class of images X_τ. β indicates adjustable hyperparameter, set to 10. The loss functions of both discriminator and generator followed the two random noise vectors z_1 and z_2 as follows.
The Discriminator objective function is:

$$\mathcal{L}_{disc,1} = -\mathbb{E}_x[\log(D_1(x)] - \mathbb{E}_{y_1}[\log(1 - D_1(y_1)] \tag{2}$$

$$\mathcal{L}_{disc,2} = -\mathbb{E}_x[\log(D_1(x)] - \mathbb{E}_{y_2}[\log(1 - D_1(y_2)] \tag{3}$$

Thus, $\mathcal{L}_{disc}$ is the average of $\mathcal{L}_{disc,1}$ and $\mathcal{L}_{disc,2}$ and the Generator objective function is as below and $\mathcal{L}_{gen}$ is the average of $\mathcal{L}_{gen,1}$ and $\mathcal{L}_{gen,2}$:

$$\mathcal{L}_{gen,1} = -\mathbb{E}_{y_1}[\log(D_1(y_1)] - \mathbb{E}_x[\log(1 - D_1(x)] \tag{4}$$

$$\mathcal{L}_{gen,2} = -\mathbb{E}_{y_2}[\log(D_1(y_2)] - \mathbb{E}_x[\log(1 - D_1(x)] \tag{5}$$

For each sampled task τ including training and test tasks, Meta-FAVAE first initializes and then minimizes the parameters of the encoder $\Phi_{1,\text{enc}}$, decoder $\Phi_{1,\text{dec}}$, discriminator $\Phi_{1,\text{disc}}$, and generator $\Phi_{1,\text{gen}}$ in the K inner loop through objective functions $\mathcal{L}_{VAE}$, $\mathcal{L}_{disc}$, and $\mathcal{L}_{gen}$, respectively. The local minimum parameters of the encoder, decoder, discriminator, and generator are achieved by minimizing loss $\mathcal{L}_{VAE}$, $\mathcal{L}_{disc}$, and $\mathcal{L}_{gen}$. The total loss function of our Meta-FAVAE method can be written as

$$\mathcal{L} = \mathcal{L}_{VAE} + \mathcal{L}_{disc} + \mathcal{L}_{gen} + \lambda_{ms}\mathcal{L}_{ms} \tag{6}$$

Where $\mathcal{L}_{ms}$ represents a mode-seeking regularization termMao et al. (2019) that maximizes the ratio of the distance between $G_1(z_1, c)$ and $G_1(z_2, c)$ corresponding to the distance between z_1 and z_2, as expressed in equation (7).

$$\mathcal{L}_{ms} = \underset{G_1}{max}(\frac{d(G_1(z_1, c)), G_1(z_2, c))}{d(z_1, z_2)}) \tag{7}$$

where $d(\cdot,\cdot)$ indicates the L_1 norm distance metric. Following the same settingMao et al. (2019), we set the hyper-parameters $\lambda_{ms} = 1$. The outer loop is performed after the K steps of the inner loop operation. The global parameters Φ_2 of the encoder Enc_2, decoder Dec_2, generator G_2, and discriminator D_2 are updated with parameters Φ_1 by setting it to $\Phi_2 - \Phi_1$. The pseudo-code of Meta-FAVAE for the training process and generation process is described in supplementary

3 Evaluations and Conclusion

The experiments are conducted on three few-shot image datasets: MNIST(LeCun & Cortes (2010)), Omniglot(Lake et al. (2011)) and VGG-Faces(Cao et al. (2018)).The quality of generated images are evaluated by FID (Heusel et al. (2017)), while The diversities of output are evaluated through LPIPS(Zhang et al. (2018)) and IS(Xu et al. (2018)).Based on the 1000 generated images of the unseen classes, we compare our model with baseline approaches including GMN(Bartunov & Vetrov (2018)), DAGAN(Antoniou et al. (2018)), FIGR(Clouâtre & Demers (2019)), DAWSON(Liang et al. (2020)) and FAML(Phaphuangwittayakul et al. (2021)). The results presented in Table 1 show our method achieves the best performance in the evaluations on all three datasets. Additionally, the feedback operation is leveraged to increase the diversity of output images and generalize the model across different categories. The model with feedback operation not only augments the data to increase the size of input but also avoids the overfitting occurred by a few training samples. Although a variational auto-encoder has been added to the network, our model can significantly reduce the number of parameters to 8.3M, which is merely one-quarter of FIGR. Furthermore, it converges over ten times faster than the baseline models. Thus, our model can fast adapt to new generation tasks with limited steps as shown in Figure 2. Comparing with the generated images from other methods, the Meta-FAVAE can not only retain the compact features due to the latent representations but also generate more diverse images. In this work, we have introduced a novel FSIG approach based on meta-learning with a feedback augmented variational auto-encoder. This model can outperform the existing approaches in term of fast adaptation and the diversity of outputs. (More results can be seen in supplementary)

Table 1: The evaluation metrics for the images of unseen class generated by baseline and our methods on three datasets.

Method	MNIST	Omniglot	VGG-Faces		
	FID (↓)	FID (↓)	FID (↓)	IS (↑)	LPIPS (↓)
GMN	98.87	127.53	209.21	2.14	0.7692
DAGAN	78.56	72.23	196.12	2.22	0.2674
FIGR	36.62	68.95	250.14	2.85	0.6701
DAWSON	86.72	75.59	184.15	2.56	0.8672
FAML	33.94	78.69	42.72	7.10	0.4011
Meta-AVAE	28.65	**63.29**	54.44	11.53	0.2431
Meta-FAVAE	**24.23**	73.74	**42.11**	**15.58**	**0.2056**

References

Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *International Conference on Learning Representations*, 2018.

Sergey Bartunov and Dmitry Vetrov. Few-shot generative modelling with generative matching networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 670–678. PMLR, 2018.

Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 67–74. IEEE, 2018. ISBN 1538623358.

Louis Clouâtre and Marc Demers. FIGR: few-shot image generation with reptile. *CoRR*, abs/1901.02199, 2019.

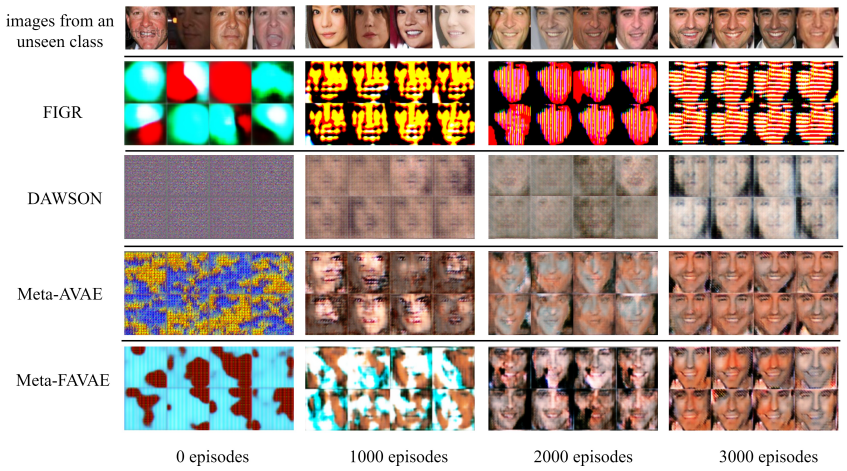

Figure 2: The generated images from the VGG-Face dataset by the baseline meta-learning based models (FIGR, DAWSON) and our Meta-FAVAE model.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pp. 1126–1135. PMLR, 2017.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pp. 2672–2680, 2014.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

Yan Hong, Li Niu, Jianfu Zhang, and Liqing Zhang. Matchinggan: Matching-based few-shot image generation. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2020.

Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*, dec 2014.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014. URL `http://arxiv.org/abs/1312.6114`.

Brenden Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua Tenenbaum. One shot learning of simple visual concepts. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, 2011.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, pp. 1558–1566. JMLR.org, 2016.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database, 2010.

Weixin Liang, Zixuan Liu, and Can Liu. DAWSON: A domain adaptive few shot generation framework. *CoRR*, abs/2001.00576, 2020. URL `http://arxiv.org/abs/2001.00576`.

Qi Mao, Hsin Ying Lee, Hung Yu Tseng, Siwei Ma, and Ming Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pp. 1429–1437, 2019. ISBN 9781728132938. doi: 10.1109/CVPR.2019.00152.

Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *CoRR*, abs/1803.02999, 2018. URL `http://arxiv.org/abs/1803.02999`.

Aniwat Phaphuangwittayakul, Yi Guo, and Fangli Ying. Fast adaptive meta-learning for few-shot image generation. *IEEE Transactions on Multimedia*, 2021.

Antoine Plumerault, Hervé Le Borgne, and Céline Hudelot. Avae: Adversarial variational auto encoder. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 8687–8694. IEEE, 2021.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pp. 3637–3645, 2016.

Qiantong Xu, Gao Huang, Yang Yuan, Chuan Guo, Yu Sun, Felix Wu, and Kilian Q. Weinberger. An empirical study on evaluation metrics of generative adversarial networks. *CoRR*, abs/1806.07755, 2018. URL `http://arxiv.org/abs/1806.07755`.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.