2020 IEEE Region 10 Symposium (TENSYMP), 5-7 June 2020, Dhaka, Bangladesh

EsharaGAN: An Approach to Generate Disentangle Representation of Sign Language using InfoGAN

Fairuz Shadmani Shishir[1], Tonmoy Hossain[2] and Faisal Muhammad Shah[3]
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
fsshishir95@gmail.com[1], tonmoyhossain.cse@ieee.org[2] and faisal505@hotmail.com[3]

***Abstract*—EsharaGAN is a Bangla Sign Digit generation model based on Information Maximizing Generative Adversarial Networks (InfoGAN). Augmenting the mutual information between latent variables and observational variables is the fundamental working principle of InfoGAN. This paper focused on generating disentangle representation of Bangla Sign Digit images using a variant of Generative adversarial network InfoGAN. Working on the IsharaLipi dataset, this model consists of 13 layer network architecture—input layer, dense layer, convolutional layer, transpose convolutional layer, activation and batch normalization layer which minimizes the loss function, computation power and generates non distorted images like the real ones. ReLU and Tanh is used as an activation function. This model provides an exceptional result as the inception score of the model is 8.77 which is remarkable for a generation model.**

***Index Terms*—InfoGAN, disentangle, convolutional, transpose, inception**

I. Introduction

Sign language is one of the imperative modes of communication with deaf people. Antecedently, it has been thought that deaf people in Bangladesh use Indian Sign Language (ISL). Notwithstanding, current research confirms that the mass of the Deaf barely knows about the ISL but they use Bangla Sign Language (BaSL). Approximately, 13 million people, near to 10% of the total population of Bangladesh are suffering from irregular degrees of hearing loss whereas three million population enduring severe to a profound level of hearing loss which leads to disability [1]. The number is increasing linearly day-by-day but the medium of articulation with the Deaf is not developed. It is high time that necessary steps should be taken to advocate, endorse and develop the medium of Bangla Sign Language.

Bangla sign language controls the Visual-manual modality which is the fundamental structure to convey meaning. It is fully developed which has its inherent grammars and lexicons [2]. This development is done through a period of time and involvement of researchers around the world. Over the years, scientists tried to establish a model to detect or classify sign languages of the diverse region. Nevertheless, researchers are lag behind in the field of accurately identifying the sign expression from an image or a video.

Nowadays, to train a neural network accurately, the dataset should be substantially large enough to precisely build the model. Also, training time plays an important role to meticulously fit the model into the dataset. It needs a considerable time to train a model consisting of immense gray-level images. That is how the importance of generating images in the model to classify or train the model can be a great idea for accurately classify the image as well as lessen the training time. For image generation, GAN (Generative Adversarial Neural Network) is a breakthrough model in this era. Working like an adversarial process, it is a framework to estimate the generative models by simultaneously training the generative model which occupies the distribution of the data and a discriminative model that evaluates the probability that a sample occurred from training data in lieu of the generative model [3]. In sign language classification, generation of the sign language can curtail down the training time by a considerable margin as well as the complexities in the neural network can be mitigated appreciably. Even though if the data is hazy or not in a good shape, a generative net can engender images from the input ones which can be fed as the input image to train the model. That is how, the generative model can amplify the classification accuracy, shrink the neural net complexity and minimize the training time.

A disentangled representation for a dataset is a depiction that is sparse over the transformations that exist in the data. The generative model works well on a dataset that satisfies the rules of disentanglement. There are a lot of variants of the generative models. Each and every particular model plays a distinct role in image generation. Some model works well on videos and others performs better on the image. One of the branches of the adversarial network is InfoGAN. It decomposes the input noises into a standard incompressible latent vector for capturing the salient semantic features of a data [4]. One of the excellent properties of InfoGAN is it maximizes the mutual information between the subset of the variables and the actual observations. That is how InfoGAN stands out in the field of generative networks. In the following section, a background study, proposed methodology to generate Bangla sign images, result analysis and future work will be described thoroughly.

978-1-7281-7366-5/20/$31.00 ©2020 IEEE

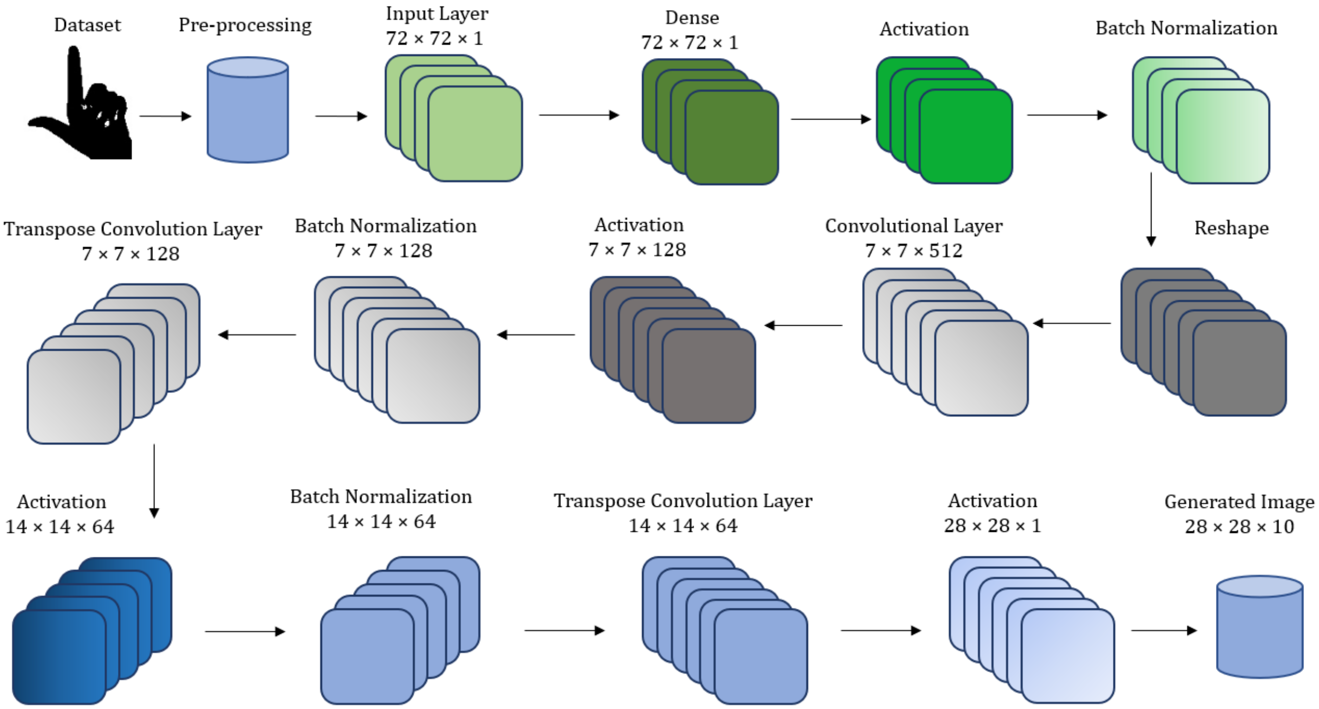

Fig. 1. Proposed Adversarial Network to generate Bangla Sign Digits

II. Background Study

InfoGAN is one of the robust variants of the generative adversarial network for image generation. Adopting the fundamental basics of InfoGAN and build a modified model to generate images is a novel approach till now. For that, a strong background study was concluded and a summary of the study is elaborately described in the following narrative.

Firstly, in the presence of continuous latent variables, inferencing a model with the data in directed probabilistic models is a pre-eminent approach to build a successful model. Kingma et al. proposed a model based on a stochastic variational inference that scales to large dataset [5]. This model is built on two-fold. All in all, it is a lower bound estimator that can minimize the training time and marginal log-likelihood.

Secondly, understanding the mutual information and controlling the properties of the output representation is compulsory to build a model on InfoGAN and Evtimova et al. [6] implemented a model that intensifies the mutual properties of InfoGAN. Confining the latent codes from the output of the generator with the help of mutual information can give interpretable representations on a distinct dataset. This article also depicts that providing noise to the input data can stabilize its training time. But sometimes adding unreasonable noise to the data can intensify the loss function and generate more noisy and hazy images from the generator.

On the other hand, VAE (Variational Auto Encoder) works as a probabilistic way to characterize a perception in latent space. There are both advantages and disadvantages of VAE (Variational Auto-Encoders) and GAN. While Variational Auto-Encoders produces the finest reconstruction of the image but blurry images, GAN generates sharp images but marginally distorted. Gorijala et al. [7] introduced a model called Variational InfoGAN based on varying the visual descriptions by fixing the latent description. This paper somehow satisfied with a specific dataset where there is no connection between latent representation and mutual information between the subset of variables and authentic observations.

Finally, a comparative study on RBM (Restricted Boltzmann Machine), VAE (Variational Auto-Encoders), AR (Auto-Regressive) models and GAN is exercised on Table – I. In this table, various types of properties relevant to the studied models is evaluated as it is shown that GAN performs best as it has directed graphical model (GM), no need of arbitrary assumptions. Also measuring the loss by JSD (Jensen-Shannon divergence) amplifies the efficiency of the complete model. For this reason, a variant of GAN—InfoGAN is considered for the purpose of our proposed model.

III. Proposed Methodology

This paper is built on the InfoGAN model. InfoGAN maximizes the mutual information between the subset of the latent variables and the actual observations. This model is ensembled with image processing followed by a modified

TABLE I
PROPERTIES OF THE GENERATIVE MODELS

Properties	RBMs	VAEs	AR models	GANs
Abstractions	Yes	Yes	No	No
Generation	Yes	Yes	Yes	Yes
Computing P(x)	Intractable	Intractable	Tractable	No
Sampling	MCMC	Fast	Slow	Fast
Type of GM	Undirected	Directed	Directed	Directed
Loss	KL-divergence	KL-divergence	KL-divergence	JSD
Assumptions	X-independent Given z	X-independent Given z	None	None
Samples	Not Good	Ok	Good	Good (Best)

adversarial network. The basic image pre-processing technique is employed to get better noise-free images and then fed into the proposed network depiction. The model constituted 13 network layers with a generator and discriminator. Following, a brief description is characterized by appropriate depiction.

Firstly, the IsharaLipi dataset [8] is adopted for the proposed model. Data augmentation such as—translation, scaling, rotation etc. is done on the input data. After that, some pre-processing technique is employed to get better input images then fed the images into the network after converting it to a vector. The size of the input vector is 72×72×1 and after the execution of the 13 layers, it outputs a 28×28×10 generated image. Depreciating the size of the input will mitigate the training time and reduced the unnecessary pixels for the training phase. After the input layer dense layer which is a fully connected layer is employed to make a bridge with all the parameters of the vectors. In this phase, for the sparsity and reduced likelihood of vanishing gradient ReLU is used as the activation function. After that, batch normalization is done to normalize all the vectors.

Secondly, a reshape layer is operated and it provides vectors of shape 7×7×512. Reshape layer minimizes the irrelevant parameters of the vectors which are unnecessary for the network. Then a convolutional layer is introduced to maintain a relationship with all the index of the vectors. ReLU is used as an activation function and batch normalization is employed because there is a chance of getting redundant relations in the convolutional layer. A vector consisting of shape 7×7×128 works as the output from the batch normalization layer.

Thirdly, a transpose convolutional layer is proposed in the model. Because, the model has to maintain the properties of InfoGAN that maximize the mutual information, transposing the parameters will certainly focus on the opposite parameters of the vectors. For the sake of mutual information, transpose convolutional layer is employed. But this layer is furtherly transposed to get the actual images either there is a chance of getting distorted images and irrelevant images. Activation and Batch Normalization is furtherly done on the transpose convolutional layer to maintain the internal properties of the layers. In this phase, tanh works as an activation layer for the sake of the model. The tanh performs better for generating images as the Jensen-Shannon divergence loss works better on this function.

After executing the 13 layers, we get the generated output images from the proposed network architecture. As there is no performance matrix to evaluate a generative adversarial model, we will evaluate the model base on loss function as it is done in the existing image generation works

IV. RESULT ANALYSIS AND DISCUSSION

In this section, we will thoroughly describe the experimental result with proper depiction and loss function. Breaking down the analysis of result into a few sub-section—Dataset properties and Experimental Setup, Hyper-parameter setup, Loss function and Generated Image.

A. Dataset Properties and Experimental Setup:

IsharaLipi Dataset has been adopted for the purpose. This dataset consists of approximately 1000 images that is subdivided into 10 classes characterizing the digits from zero to nine. Some basic pre-processing works is done on the dataset before fed into the model. Google Colab is used to train the model.

B. Hyper-parameter setup

For training the model, we need to set the hyper-parameter of the model accurately. Tested with different values for the necessary hyper-parameters, the optimum value is characterized in the following table. Using the optimum hyper-parameter value of table-II, the best loss function curve is achieved. Table-II represents the values of the related hyper-parameters.

TABLE II
HYPER-PARAMETERS OF THE PROPOSED ARCHITECTURE

No	Hyper-parameter	Value
1	Initialization_kernel	glorot_uniform
2	Initialization_bias	zeros
3	Learning_rate	0.001
4	Optimizer	Adam
5	Batch size	64
6	Epoch	10000
7	Latent Dimension	62

C. Loss Function and Inception Score:

For the evaluation of the model, a loss function is generated from the model. We evaluate three loss function for the dataset. All the values are congested in the zero level which tells that the models' loss is approximately zero. Figure–2 represents the loss function curve of the model. Also we evaluate the model based on Inception score. Inception score (IS) is an objective function to evaluate the generated images [9]. The inception score of the model is **8.77** for 10 class which is excellent for a generative model.

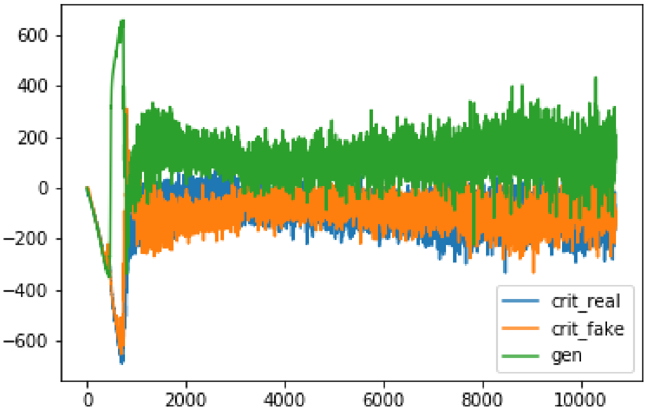

Fig. 2. Loss function curve of the proposed architecture

D. Generated Image:

Executing the proposed 13 layer architecture, the output from the architecture will be the generated synthetic images. Figure-3, and 4 represents the generated Bangla Sign Digits images from the proposed network.

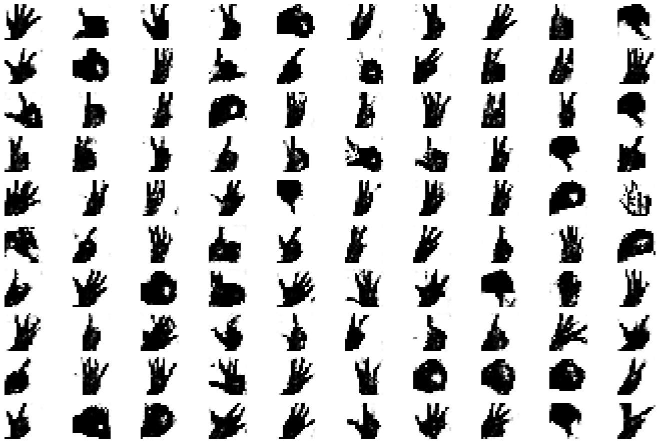

Fig. 3. Generated Image-1 from the proposed model

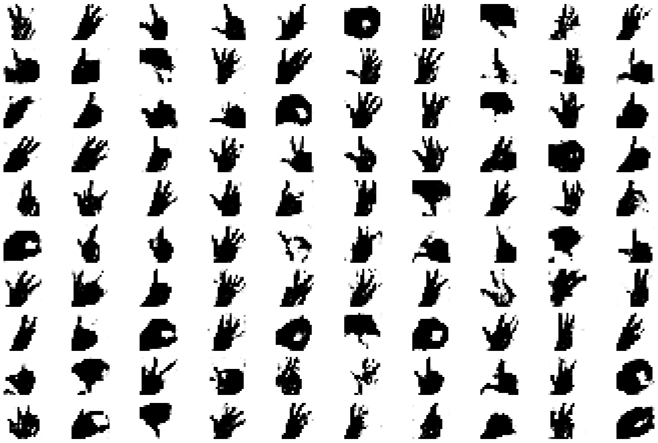

Fig. 4. Generated Image-2 from the proposed model

V. Conclusion and Future Works:

In this article, a novel model to generated Bangla Sign Digit images using a modified InfoGAN method has been proposed. The model gives us an exceptional result as the inception score is 8.77 for the 10 sign digits. Also, Computation power and training time is curtailed with respect to the existing articles as well as data augmentation is done. Removing the pooling layer is an advantage of the model because the Pooling layer which is lossy and does not conserve all the spatial information is apprehended in this architecture. In the future, we will elaborate on the working procedure of the model on the video field and engage the spatial relationship of the pixels.

References

[1] Alauddin, Mohammad, and Abul Hasnat Joarder. "Deafness in Bangladesh." In *Hearing Impairment*, pp. 64-69. Springer, Tokyo, 2004.

[2] Sandler, Wendy, and Diane Lillo-Martin. *Sign language and linguistic universals*. Cambridge University Press, 2006.

[3] Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." In *Advances in neural information processing systems*, pp. 2672-2680. 2014.

[4] Hong, Yongjun, Uiwon Hwang, Jaeyoon Yoo, and Sungroh Yoon. "How generative adversarial networks and their variants work: An overview." *ACM Computing Surveys (CSUR)* 52, no. 1 (2019): 1-43.

[5] Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." *arXiv preprint arXiv:1312.6114* (2013).

[6] Evtimova, Katrina and Andrew Drozdov. "Understanding Mutual Information and its Use in InfoGAN." (2016).

[7] Gorijala, Mahesh, and Ambedkar Dukkipati. "Image generation and editing with variational info generative AdversarialNetworks." *arXiv preprint arXiv:1701.04568* (2017).

[8] M. Sanzidul Islam, S. Sultana Sharmin Mousumi, N. A. Jessan, A. Shahariar Azad Rabby and S. Akhter Hossain, "Ishara-Lipi: The First Complete MultipurposeOpen Access Dataset of Isolated Characters for Bangla Sign Language," *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, Sylhet, 2018, pp. 1-4, doi: 10.1109/ICBSLP.2018.8554466.

[9] Salimans, Tim, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. "Improved techniques for training gans." In *Advances in neural information processing systems*, pp. 2234-2242. 2016.