Deep Generative Models: Deterministic Prediction with an Application in Inverse Rendering

Shima Kamyab  
Dept of Comp. Sci., Eng. and IT  
Shiraz University  
Shiraz, Iran  
sh.kamyab@cse.shirazu.ac.ir

Rasool Sabzi  
Dept of Comp. Sci., Eng. and IT  
Shiraz University  
sabzi@cse.shirazu.ac.ir

Zohreh Azimifar  
Dept of Comp. Sci., Eng. and IT  
Shiraz University  
azimifar@cse.shirazu.ac.ir

Abstract—Deep generative models are stochastic neural networks capable of learning the distribution of data so as to generate new samples. Conditional Variational Autoencoder (CVAE) is a powerful deep generative model aiming at maximizing the lower bound of training data log-likelihood. In this structure, there is an appropriate regularizer, which makes it applicable for suitably constraining the solution space in solving ill-posed problems and providing high generalization power. Considering the stochastic prediction characteristic in CVAE, depending on the problem at hand, it is desirable to be able to control the uncertainty in CVAE predictions. Therefore, in this paper we analyze the impact of CVAE's condition on the diversity of solutions given by our designed CVAE in 3D shape inverse rendering as a prediction problem. The experimental results using Modelnet10 and Shapenet datasets and comparison with several recent methods show the appropriate performance of our designed CVAE and verify the hypothesis: “The more informative the conditions in terms of object pose are, the less diverse the CVAE predictions are”.

Index Terms—Generative models for prediction, Deep generative models, 3D shape inverse rendering

I. INTRODUCTION

Generative models include a broad domain of machine learning techniques with the aim of using statistical methods to learn the underlying distribution of a set of real data to generate similar synthetic data from the learned distribution. More formally, suppose sample $x$ is obtained from an unknown distribution $P_{gt}(x)$. The objective is to learn the distribution $P$ so that sample $x'$ drawn from $P$ is as close to $x$ as possible.

The traditional generative models suffer from three main drawbacks [6]:

- In some cases, they need large amount of information about the structure of data.
- The training method used in these techniques may causes high level of uncertainty and therefore, the resulted synthetic samples may be unfeasible.
- Most of these methods suffer from high computational complexity.

In recent years, after the advent of the field of deep learning, deep generative models, which exhibit strong performance in modeling complex high dimensional distributions of text or image, have attracted a large body of interest in the literature. Due to their powerful nonlinear approximation, these models are appropriate tools for estimating the density of complicated and high dimensional data. Deep generative models are applied in many fields including text generation [10], [12], [17], latent space learning [3], image denoising [2], image inpainting [26], super-resolution [14], etc.

Two of the most well-known and efficient deep generative models are Variational AutoEncoders (VAEs) and Generative Adversarial Networks (GANs). The VAE’s objective is to maximize the lower bound of training data’s likelihood function and GANs aim at achieving an equilibrium between their two adversarial components known as Generator and Discriminator.

In this paper our focus is on VAE models, where the data generation problem can be formulated in the form of a Bayesian model [13]. The main contribution of VAE is to maximize the likelihood function of training data on the whole generative process by conditioning the output distribution on some latent variable learned by using the information from training data [6]. Although VAEs could generate feasible samples, the models have no control on the generated output. Conditional Variational AutoEncoder (CVAE) as a modified VAE addresses this limitation by utilizing some condition information to control the type of the output. In the case of CVAE, all distributions are conditioned on some measurement, called $m$, and this condition is used as an input to both encoder and decoder components.

CVAE is a stochastic model capable of learning multi-modal distributions, i.e., finding a one-to-many mapping resulting different solutions for the problem. This characteristic of CVAE is applicable in many problems like structural classification [20]. On the other hand, for some situations like ill-posed problems, it is desirable to have more deterministic solutions, e.g., 3D shape inverse rendering with complicated solution space, needs the solutions to be deterministic in
terms of personal identity so that they could recover the true shape accurately [21]. Therefore, if some mechanism could be provided to control the diversity of the solution space in CVAE, it will be more effective in different problem domains. In this paper we analyze the impact of information available in the condition component on the diversity of solutions in CVAEs. This analysis helps to control the solution diversity depending on the problem requirements.

We choose 3D object shape inverse rendering from single view as a complicated and ill-posed optimization problem. Reconstructing 3D volumes from 2D images, also known as 3D inverse rendering, is a challenging and ill-posed problem in computer vision with a nonlinear solution space. There may exist missing or hidden parts in the given 2D input images, which can be reconstructed in many 3D shapes. Effective regularizers are needed to solve such ill-posed problems, as a regularizer can apply appropriate constraints on the solution space to obtain promising results.

For inverse rendering problem, we design a CVAE using the input 2D image as condition and analyze the effect of pose in 2D image, as the source of information, on the standard deviation of the output as a measure of computing diversity. In the experiments we evaluate our designed CVAE by comparing it to several recent related methods to show that its promising performance as a single view 3D reconstruction method and analyze the effect of condition on its prediction.

The remaining structure of this paper is organized as follows: Sec. II includes a brief review of some recent related works about using CVAE for prediction and its analysis. Sec. III starts with problem definition followed by our proposed CVAE structure for prediction. Sec. IV contains experimental results to show the effect of the network condition on the prediction uncertainty in CVAE. Also evaluation of our designed CVAE is presented by comparing it to several recent methods for single view 3D reconstruction. The paper is concluded in Sec. V.

II. RELATED WORKS

In the recent deep learning literature, there exists a significant interest in using and analyzing CVAE for different optimization problems [5], [8], [22], [23]. In this section we review some of the most recent studies on CVAE as the most related works to the subject under this study.

As some recent deep state-of-the-art for 3D reconstruction, we name [4], in which a Recurrent Neural Network is proposed for single an multiple view 3D reconstruction and [19], [24]. Generative Adversarial Networks are used for designing 3D reconstruction frameworks. The formar is a generative model and the latter is a single view 3D reconstruction method.

In [16], a CVAE-based framework is proposed with unsupervised training strategy for 3D object reconstruction, where the 3D shape is recovered from single or multiple views. In [20], CVAE is used as a stochastic predictor for structural classification. The multi-modal distribution estimated by CVAE, which is due to its stochastic topology, leads to promising results in structural classification compared with similar deterministic predictors. The authors also reported some experiment to show that stochastic prediction in CVAE leads to high generalization power to the partial observations.

In [15], Deep Latent Variable Models (DLVMs) are revisited to show that the maximum-likelihood objective in these models is ill-posed for continuous problems and well-posed for discrete problems. Moreover, DLVMs can be related to non-parametric mixture models and take advantage of this potential to find an upper bound for CVAE objective. Finally, [15] proposes a method for handling missing data in CVAE prediction.

In [11] a new training technique for VAEs is proposed. This technique combines the strengths of GANs and VAEs and produces high-resolution photographic images. Besides, [18] indicates that amortized inference in VAEs, (i.e. using empirical approximations instead of computing exact statistics) provides appropriate regularization for maximum likelihood estimate resulting in a good generalization.

To the best of our knowledge, controlling stochastic behaviour of CVAE has not yet been analyzed in the literature. Therefore, we propose to study and analyze the behaviour of CVAE in deterministic predictions.

III. PROBLEM DEFINITION

In this section we first define CVAE as a regularized predictor using the concepts from [6] and then, demonstrate the designed CVAE structure for 3D shape inverse rendering.

A. Using Conditional Variational Autoencoder (CVAE) as a predictor

Formally, the main VAE objective is to maximize:

$$P(z) = \int P(z|l; \theta) P(l) dl$$  \hspace{1cm} (1)$$

where, $P(z)$ is the data likelihood, $P(z|l; \theta)$ is the output distribution of VAE which is often chosen to be Gaussian, i.e $P(z|l; \theta) = N(z|f(l; \theta), \sigma^2 * I)$, where, $f(z; \theta)$ stands for the VAE decoder output.

In order to avoid the regions where $l$ leads $P(z|l; \theta)$ to become nearly zero, VAE introduces an encoder to control and restrict $P(l)$, by defining the distribution $Q(l|z; \beta)$ instead of $P(l)$, which uses the information from the target data to control the latent distribution. This policy in VAE increases the training speed and improves the feasibility of found solutions and is an appropriate mechanism to solve ill-posed optimization problems.

The output of VAE is $E_{l \sim Q}(P(z|l; \theta))$, where the latent variable distribution, i.e., $Q(l|z; \beta)$, is constrained to be like the true probability $P(l|z)$, i.e., the probability of inferring $l$ form $z$. This constraint is formulated using KL-divergence operator:

$$D(Q(l|z; \beta)||P(l|z)) = E_{l \sim Q}[log Q(l|z; \beta) - log P(l|z)]$$  \hspace{1cm} (2)$$

where $\beta$ and $\theta$ are parameters of trainable encoder and decoder, respectively.
Applying Bayes rule to $P(l|z)$ and changing term, we have:

$$
\log P(z) - D(Q(l|z; \beta) || P(l|z)) = 
E_{l \sim Q} [\log P(z|l; \theta)] - D(Q(l|z; \beta) || P(l))
$$

(3)

The left-hand side in (3) is equivalent to the objective of VAE to be maximized and the right hand-side can be computed by VAE. In the VAE objective it can be observed that the KL-divergence term appears like a regularization term [18], which firstly, forces the distribution of the latent variable to be obtained from the target data and secondly, constrains the distribution to be Normal Gaussian. This KL-divergence term can be viewed as a generic regularizer in any optimization problem. Therefore, VAE can be considered as a regularized optimizer, for solving complicated optimization problems. It is worth noting that in KL-divergence term, $P(z)$ is set to $N(0, I)$ assuming it can be converted to any distribution consistent with the latent variable in the trainable decoder [?].

In the case of CVAE which is our focus in this paper, all terms in (3) are conditioned on variable $m$. Therefore the condition $m$ will be fed into both decoder and encoder components.

Basically, CVAE is a stochastic optimizer because of the sampling component between encoder and decoder. Therefore it results some degree of uncertainty in the obtained solution. Our objective in this paper is to analyze the possibility of controlling the solution diversity by condition. On other method of controlling the uncertainty in CVAE is the length of latent variable. No sampling results in a deterministic predictor and a high dimensional sampling results in appearing noise in prediction. Since this variable includes prior information about the prediction, it will be useful to use it to obtain more accurate results. The optimum latent variable length can be found by validation methods. In this paper our focus is only on the condition as other resource for controlling the diversity of obtained solutions by fixing the dimensions of sampling component.

The network structure of our designed CVAE for prediction analysis will be illustrated in the following section.

B. Network structure

Based on the formulations in Section III-A we design our CVAE structure for 3D shape inverse rendering, shown in Figure 1.

In this structure which is also used in [16], [20] for regression, the measurement, 2D image here, is used as condition ($m$) and prediction, 3D shape here, is used as $z$.

IV. EXPERIMENTS

In this section we evaluate our designed CVAE for 3D shape inverse rendering in two phases. In the first phase, we compare our designed CVAE with recent and state-of-the-art methods for single view based 3D reconstruction methods from literature. Our main objective in this phase is to show the performance of our designed CVAE as an appropriate single view 3D reconstruction method comparable with existing methods. In the second phase, we monitor the effect of informative condition on the solution diversity of CVAE for prediction. We introduce 2D input image as condition of CVAE and pose as information in condition in this phase.

The following subsections include parameter setting and dataset used in our experiments followed by two designed phases.

A. Parameter setting

The detailed structure of encoder and decoder of proposed CVAE can be seen in Tables I, II respectively. The structures are inspired by the structure used in [24] as a GAN for 3D reconstruction. We used kers for implementing our CVAE with 50 epochs and ADAM optimizer with default learning-rate on a NVIDIA GeForce GTX 1080 Ti graphic processor.
B. Dataset

We used two popular 3D object datasets including Modelnet10 [25] and Shapenet [1] in our experiments. In order to analyze the deterministic prediction of CVAE, we used Modelnet10 dataset and for the sake of comparison with other methods for single view based 3D reconstruction we used Shapenet dataset.

In the case of Modelnet10 dataset, we used the training and test sets defined by the dataset for our experiments. In the case of Shapenet dataset, we used 4 classes including car, airplane, chair, couch for training and testing. The training and test sets for Shapenet are selected randomly. The size of input images to our designed CVAE is set to $128 \times 128 \times 128$ voxels and then computed converted the output to $32 \times 32 \times 32$ voxels and then computed.

C. Performance comparison with state-of-the-art and recent single view 3D reconstruction methods

In this section, in order to have better evaluation of our designed CVAE for prediction, we compare it with several recent methods for single view 3D reconstruction including [7], called PSGenerator, [4], called 3DR2N2, and [8], called AtlasNet. For each method we used their pre-trained model available in the web without post-processing. Note that for 3DR2N2, we used the model made available by [7]. In the case of the methods PSGenerator, 3DR2N2, AtlasNet, the comparisons are reported on the shapenet dataset for evaluation. The datasets are selected based on the evaluation mechanism used in each compared method’s paper.

Figure 2 shows the visual results of comparison between our designed CVAE to the work in [7], using the available network and weights from the web. Since the reconstruction of compared method are in the form of point clouds, we converted the output to $32 \times 32 \times 32$ voxels and then computed and showed the results.

As quantitative result, Table III shows the numerical results obtained by designed CVAE and compared methods in terms of average shape intersection over union (IOU). Our results are averaged over 10 independent runs using 10 fixed predetermined equally spaced noise values as input to decoder in test phase. In the case of our CVAE, the numbers in parentheses denote the standard deviation of CVAE output for predetermined noise values.

From Figure 2 and Table III it is observable that our proposed CVAE could achieve comparable results with state-of-the-art and outperform them in several cases. Therefore It can be considered as an appropriate framework for single view inverse rendering framework and it is a valid network for being analyzed as a 3D reconstruction tool.
More informative condition results in to test the hypothesis as a prediction framework, in this section the objective is D. Analyzing the effect of condition on solution diversity are illustrated by green border as informative conditions. The results with the least MSE classes of Modelnet10 dataset for different poses. From left to right, bed, chair, desk and monitor classes are shown. The results with the least MSE are illustrated by green border as informative conditions.

D. Analyzing the effect of condition on solution diversity

After verifying the performance of our designed CVAE as a prediction framework, in this section the objective is to test the hypothesis “More informative condition results in lower solution diversity”. By more informative condition, we mean the condition containing more specific and adequate information about the prediction. For this aim we considered pose as information in 3D shape inverse rendering. Therefore, each object in training set of Modelnet10 dataset is rendered from 8 poses. Note that in order to omit the inter class overlap impact on the results, in this phase, for each class we trained a separate CVAE. For instance, Figure 3 shows the reconstruction results obtained by CVAE trained for 4 classes separately. The results are averaged over 10 independent runs using 10 specified equally spaced noise values as the input to decoder in test phase Figure 1(b).

From Figure 3 it is observed that, for all object classes, considering the pose resulted in the least reconstruction error as informative pose, the most informative pose results in lower standard deviation. We should note that there sill exists intra-class overlap that affect the uncertainty of solutions.

V. Conclusions and future works

In this paper we focused on studying and analyzing the effect of information in condition on the diversity of prediction in CVAEs. We designed our CVAE for prediction by using the measurement as its condition component and used it for 3D object shape inverse rendering as a prediction problem. The experimental results show the promising performance of our designed CVAE compared with recent single view-based 3D reconstruction methods. By considering pose as the information in the input images, the hypothesis “the more informative condition, the more deterministic the CVAE predictor” is verified.

This is an ongoing research and we plan to analyze other elements that affect CVAE prediction, such as the number of training data and modality of the latent variable distribution.
[22] Arash Vahdat, Evgeny Andriyash, and William G Macready. Dvae#: Discrete variational autoencoders with relaxed boltzmann priors. arXiv preprint arXiv:1805.07445, 2018.

[23] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In European Conference on Computer Vision, pages 835–851. Springer, 2016.

[24] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In Advances in Neural Information Processing Systems, pages 82–90, 2016.

[25] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1912–1920, 2015.

[26] Raymond A Yeh, Chen Chen, Teck-Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In CVPR, volume 2, page 4, 2017.