Deep Variational Luenberger-type Observer for Stochastic Video Prediction

Dong Wang\(^1\), Feng Zhou\(^1\), Zheng Yan\(^1\), Guang Yao\(^1\), Zongxuan Liu\(^1\), Wennan Ma\(^1\), Cewu Lu\(^2\)
\(^1\)Shanghai Em-Data Technology Co., Ltd, \(^2\)Shanghai Jiao Tong University
\{wangdong2, zhofeng, yanzheng, yaoguang, liuzongxuan, mawennan\}@em-data.com.cn, lucewu@sjtu.edu.cn

Abstract

Considering the inherent stochasticity and uncertainty, predicting future video frames is exceptionally challenging. In this work, we study the problem of video prediction by combining interpretability of stochastic state space models and representation learning of deep neural networks. Our model builds upon a variational encoder which transforms the input video into a latent feature space and a Luenberger-type observer which captures the dynamic evolution of the latent features. This enables the decomposition of videos into static features and dynamics in an unsupervised manner. By deriving the stability theory of the nonlinear Luenberger-type observer, the hidden states in the feature space become insensitive with respect to the initial values, which improves the robustness of the overall model. Furthermore, the variational lower bound on the data log-likelihood can be derived to obtain the tractable posterior prediction distribution based on the variational principle. Finally, the experiments such as the Bouncing Balls dataset and the Pendulum dataset are provided to demonstrate the proposed model outperforms concurrent works.

1. Introduction

Videos contain rich information including features of objects and dynamics of objects in both spatial and temporal dimensions. Human beings have uncanny abilities to distinguish different objects and correspondingly predict the plausible future dynamic behaviors of these objects from videos. This motivates the investigation of video prediction in the computer vision community. With the development of deep learning technology, video prediction has attracted considerable attention with a variety of applications reported, including precipitation nowcasting [36, 7, 37], human motion prediction [42, 23, 10], and vision-based robotic control [13, 8].

Current research of video prediction faces two critical challenges, namely how to deal with uncertainties for the future and how to extract effective representations of raw videos. Researchers hold different perspectives on the first issue. While some believe that only one future sequence that is most likely to happen can be accepted under the assumption of static environment, others reject this assumption and believe stochastic video prediction is more meaningful [39, 2, 27, 11]. We aim to follow the second strategy as the environment in many real applications is not deterministic. Stochastic prediction requires more effective video representations. In specific, learning a disentangled representation can benefit a large number of downstream tasks including object classification, localization, and tracking [16, 9, 41]. Researchers have made a great deal of efforts to extract disentangled representations by using a range of intuitive ways, such as directly decomposing the representation into two components namely content (static features) and motion (dynamics). However, given a network architecture, the current research cannot automatically find the optimal ways to extract disentangled representation. Instead, they assume the artificially designed decomposition methods fit the model structures without providing theoretic justification. A more desired way would be let the model choose the most suitable representation in view of its dynamics.

In parallel to the research of video prediction, stochastic prediction and filtering for time-series data, in particular the stochastic state space model, has demonstrated significant progress. An assumption underlying the state space model (e.g., the Luenberger observer) is that the observation such as a video frame at a certain time is generated by the internal hidden states at the same instance. In the mean while, the hidden states evolve according to a stochastic transition function that is generally identified physically or numerically. The main advantage of the state space model is that the hidden states are interpretable, which can be easily used for downstream tasks such as control and planning. It is worth noting that for many practical time-series problems, the parametric form of the state space model is often known,
and only a few parameters need to be identified via data fitting. However, this requirement can hardly be satisfied in video prediction tasks. Due to the high complexity of video sequences, it is practically intractable to predefine the canonical form of the specific state space model.

Motivated by the above discussions, we aim to develop a novel video prediction approach by taking the complementary advantages of the structure of the state-space Luenberger-type observer and the method of variational inference. An interpretable disentangled representation that separates the static features and the dynamics is obtained according to the given parameters of the proposed transition and emission matrices. With the help of the stability theory, the robustness of the latent state is discussed for the proposed network structure. Experimental results illustrate the validity of this new model. The contributions of novelty of this work are highlighted below:

1) A novel deep stochastic nonlinear state space model is proposed to characterize the complex dynamics of the environment from video data.

2) The proposed model is end-to-end trainable and disentangles the static features and the dynamics of the raw sequence data directly by using our specific structure.

3) The influence from the initial value of the hidden states is shown to be insignificant based on the stability theory of the nonlinear Luenberger-type observer, which makes the model more robust and reliable.

The rest of the paper is organized as follows. Section 2 reviews the related works briefly. Section 3 presents background information and preliminary algorithms and methods. Section 4 describes the details of the proposed network structure. The experiment settings and results are discussed in Sections 5 and 6 respectively. The conclusion is drawn finally in Section 7.

2. Related Work

Stochastic Video Prediction. Early works have addressed the video prediction in the deterministic environments. In this situation, most of them use the mean squared error loss function to generate realistic-looking images as much similar as the ground truth. In order to tackle the blurry predicted frame problem, they hope to introduce the adversarially-trained models to produce more naturalistic and sharp video sequences. Recently, some results point out that the reason for generating blurry future frames does not come from the model structure or the trick for learning, but from the stochastic nature of video prediction. One main approach is to use the variational inference such as the variational auto-encoders (VAEs) and other variants to explore the stochastic video prediction, see [22] [17] [3] [4] for details. For example, [5] has applied the VAEs to predict the possible future, which can be sampled from the latent variables. Combing graphs and variational recurrent neural network, a graph-structured VRNN has been proposed in [38] to learn to integrate temporal information with partially observed visual evidence. In [12], the Kalman VAE has been introduced to describe the evolution of the world, which is similar to our work. However, they used the hidden states from the linear time-varying Gaussian state space model and the robustness from the initial value of the hidden states had not been taken into account.

Disentangled Representation. Up to now, the idea of disentangled representation has already been investigated for video prediction. It is important and wonderful if we can learn an useful and interpretable representation for the unsupervised learning, which can be transferred into the downstream tasks such as object detections or recognitions without any difficulties. It should be pointed out that, most of the existing literature extracted the disentangled representation only from some certain heuristic perspective. For instance, [17] has proposed a Decompositional Disentangled Predictive Auto-Encoder (DDPAE) to automatically learn the latent disentanglement in the unsupervised way, where the extracted representation are separated two components artificially, a time-invariant content vector and a time-dependent pose vector. In [40], two different neural network encoders have been presented to learn to decompose the motion and content by using image differences and one single frame separately. [19] has proposed the Sequential Attend-Infer-Repeat (SQAIR) in order to achieve temporally consistent reconstructions and learn an interpretable hidden variables. Recently, in [42], a Parts, Structure, and Dynamics (PSD) model has been investigated to recognize the object parts, compose the hierarchical structure and predict the dynamics. Different from the results mentioned above, in order to extract a useful disentangled representation, our work proposes a novel specific structure based on the observability for the state space models, which is the first yet theoretical attempt to tackle such challenge.

Deep State Space Models. In classical system theory, the dynamics is analyzed and controlled typically by structuring the state space model, which has shown to be a powerful tool. There are a variety of classical prediction and filtering techniques for different types of the transition function and noises. To be specific, the Kalman filtering is the optimal filtering algorithm in the sense of minimum-variance for linear Gaussian state space model, which minimizes the filtering error covariance at each time step. In recent years, some initial results have been reported by making full use of the methods of state space models and deep learning technology to improve the interpretability of the representations and the approximation of the highly nonlinearity in the stochastic environment, see [31] [14] [44] [34] [32].
In [20], Deep Variational Bayes Filters (DVBFs) mainly focused on the combination for the classical Kalman filtering while assuming the transitions satisfy the linear Gaussian state space condition. Similar with the DVBFs, [29] has proposed a Recurrent Neural Filter (RNF) to provide a more realistic estimation and prediction in the stochastic environments, which decoupled the state transition and update steps. It should be mentioned that, almost all existing works have not discussed the observability and the stability for their proposed deep state space models, which is extremely crucial because the uniqueness and the accuracy of the prediction for the hidden states are decided by these properties.

3. Preliminaries

3.1. Stochastic State Space Models

Considering the stochastic nonlinear state space models in the following:

\[
\begin{align*}
    z_{k+1} &= f(z_k, w_k), \\
    y_k &= g(z_k, v_k)
\end{align*}
\]

where \(f(\cdot)\) and \(g(\cdot)\) are the nonlinear functions that represent known transition and emission processes respectively. \(z_k \in \mathbb{R}^n\) is the hidden state, \(y_k \in \mathbb{R}^m\) is the observation from sensors such as cameras. The stochastic variables \(w_k \in \mathbb{R}^w\) and \(v_k \in \mathbb{R}^v\) are the transition and emission process noises. The main tasks for prediction and filtering is to calculate the posterior probability distribution \(P(z_{k+1}|y_{1:k})\) and \(P(z_k|y_{1:k})\) separately, with a given starting state \(z_0\).

3.2. Linear Steady-state Kalman Filter

For the linear Gaussian time-invariant systems, the general state space models in (1) are rewritten as follows:

\[
\begin{align*}
    z_{k+1} &= Az_k + w_k, \\
    y_k &= Cz_k + v_k
\end{align*}
\]

where \(z_k\) and \(y_k\) stand for the hidden state and the observation. \(A\) and \(C\) are known transition and emission matrices with compatible dimensions. \(w_k\) and \(v_k\) are assumed as Gaussian noises with zero-means and covariances \(Q > 0\) and \(R > 0\), respectively. The classical Kalman filter structure can be established at instant \(k\) in (3):

\[
\hat{z}_{k|k-1} = \hat{A}\hat{z}_{k-1|k-1},
\]

Prediction

\[
\hat{z}_{k|k} = \hat{z}_{k|k-1} + K_k (y_k - C\hat{z}_{k|k-1}),
\]

Filtering

where the gain matrix satisfies

\[
\begin{align*}
    K_k &= P_{k|k-1}C^T (CP_{k|k-1}C^T + R)^{-1}, \\
    P_{k|k-1} &= AP_{k|k-1}A^T + Q, \\
    P_{k|k} &= (I - K_kC)P_{k|k-1}.
\end{align*}
\]

According to the knowledge of system theory, for a linear time invariant system, if \((A, C)\) is observable, the classical Kalman filter described above can be replaced by a steady-state Kalman filter, which has the time invariant gain matrix \(K\) satisfying \(\lim_{k \rightarrow \infty} K_k = K\). In addition, the necessary and sufficient condition is given in the following for the observability of \((A, C)\)

\[
\text{rank} \left[ C^T \ A^T C^T \ \cdots \ \left(A^T\right)^{n-1} C^T \right] = n.
\]

3.3. Stable Luenberger-type Observer

In practice, the initial value \(z_0\) of the hidden state \(z_k\) in (2) is usually unknown and difficult to be obtained directly. The selection of the initial value has a great influence on the filtering performance of the filter. Therefore, we need to seek the filtering algorithm which is robustness for the initial value.

For the steady-state Kalman filter in Section 3.2, by submitting (4) into (3), one has:

\[
\begin{align*}
    \hat{z}_{k+1|k} &= A\hat{z}_{k|k} \\
    &= A\hat{z}_{k|k-1} + AK (y_k - C\hat{z}_{k|k-1}) \\
    &= A\hat{z}_{k|k-1} + L (y_k - C\hat{z}_{k|k-1})
\end{align*}
\]

which is called Luenberger-type Observer. As such, we can see that the steady-state Kalman filter is only a special case of the Luenberger-type observer. In [24], if \((A, C)\) is observable, theoretically, there always exists a observer gain \(L\) which can guarantee the global asymptotic stability of the hidden state errors. The strict mathematical form is as follows:

\[
\lim_{k \rightarrow \infty} \|z_k - \hat{z}_{k|k-1}\| = 0, \quad \forall \hat{z}_{1|0}, z_1.
\]

Therefore, according to the stability of the hidden state errors, it is negligible from the influence of unknown initial state value for a stable Luenberger-type observer.

3.4. Variational Auto-Encoders

A variational auto-encoder (VAE) in [22] optimizes a variational lower bound on the data log-likelihood. To be specific, the data log-likelihood can be factorized by

\[
P_\theta(x) = \int P_\theta(x, z)dz = \int P_\theta(x|z)P_\theta(z)dz.
\]
intractable. VAE uses a parametric distribution $q_\phi(z|x)$ that constructed by a neural network to approximate the posterior distribution. By using the variational principle, the variational lower bound on the marginal likelihood is derived in the following:

$$\log P_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)} [\log P_\theta(x|z)] - \text{KL}(q_\phi(z|x) \| P_\theta(z)).$$

4. Deep Variational Luenberger-type Observer

From the point of the view of the state space models, a common assumption is that the hidden state $z_k \in \mathbb{R}^n$ contains more information than the observable $y_k \in \mathbb{R}^m$, which means $m \leq n$. However, in VAEs, the hidden variables are essentially only the feature extractions or compressions of the observations. It is not appropriate to construct the state space models with such partially hidden variables directly. On the other hand, if the raw video data are taken as the visual measurement without any preprocessing, the dimension of state-space hidden variables will be too high, which greatly increases the calculating burden. Therefore, we adopt a compromise approach to balance the amount of information and the burden of the computing. To be specific, the original image is firstly compressed into a low-dimensional space by using an auto-encoder, and then a relatively high-dimensional hidden state is constructed by using extracted low-dimensional space feature to satisfy the basic assumption and structure of stochastic nonlinear state space models.

Compared with the pixel levels of the images, what we really want to extract is the higher level semantic features of the frames from the video, such as the appearances of the objects, colors and so on. Furthermore, we want to learn how objects move or how colors change from the extracted high-level semantic feature sequences. Therefore, it is reasonable for us to construct the proposed network structure with an auto-encoder of this paper.

4.1. Disentangled Representation

Assuming that the image information from the video can be compressed into a high level feature $y_k$ by an encoder $y_k = \varphi^{enc}(x_k)$ and reconstructed by the decoder $x_k = \varphi^{dec}(y_k)$, and set the ideal hidden state $z_k$ and the feature $y_k$ satisfying the following state space models:

$$\begin{align*}
z_{k+1} &= \varphi(\hat{z}_k, h_k) + w_k, \\
y_k &= \varphi(\hat{z}_k) + v_k
\end{align*}$$

where $A$ and $C$ are given matrices with compatible dimensions, and $w_k$ and $v_k$ are the stochastic noises. The function $\varphi$ stands for the part of the nonlinearity. $\varphi$ is chosen as a GRU module in this paper, where $h_k$ is the hidden state for GRU. It should be mentioned that, it is not necessary to assume that the hidden variable $z_k$ satisfies the first-order Markov property by using the GRU cells, which greatly expands the scope of application of the classical state space models. In (11), the initial hidden variable $z_0$ is unknown.

Before presenting our network structure, we first discuss the selection of parameters for $A$ and $C$. In order to learn a relatively interpretable representation, we can give $C$ satisfying

$$C = \begin{bmatrix} I & 0 \end{bmatrix}.$$  (11)

In this way, we essentially divide the hidden variable $z_k$ into two parts. The former part can be understood as the static feature of each frame in the video, which is extracted from one image directly. The latter part is the dynamics from the video sequences. Interestingly, we only use the parameter structure of matrix $C$ to disentangle the representation without any additional supervised information. Moreover, we can select matrix $A$ as

$$A = \begin{bmatrix} I & I \\ 0 & I \end{bmatrix},$$  (12)

which is easily known that the matrices $A$ and $C$ guarantee the condition of the observability in Section 5.2.

4.2. Deep Stable Luenberger-type Observer

Moreover, we assume that the latent state probability distribution to be $\hat{z}_{k+1} \sim \mathcal{N}(\mu_{k+1}; \text{diag}(\sigma^2_{k+1}))$ and the initial latent state satisfies $\hat{z}_0 \sim \mathcal{N}(0; I)$.

Combined with the structure of the Luenberger-type observer, we introduce our approach for predicting the $\mu_{k+1}$ and $h_{k+1}$ in the following

$$\begin{align*}
[\hat{\mu}_{k+1}, h_{k+1}] &= f_\theta(\hat{z}_k, h_k) \\
\mu_{k+1} &= A\hat{z}_k + \mu_{k+1} + L_p(y_k - C\hat{z}_k) \\
\text{diag}(\sigma^2_{k+1}) &= \Sigma_\phi(\hat{z}_{k+1})
\end{align*}$$

where $\Sigma_\phi$ is our variance generation model parameterized by $\phi$, $L_p$ is the observer gain matrix to be learned.

Now, we will discuss the stability for our neural-network Luenberger-type observer. The following definition is needed in deriving our main results.

Definition 1 [35] The nonlinear function $f(\cdot)$ satisfies the Lipschitz condition, if there exists a constant $K$ satisfying

$$|f(a) - f(b)| \leq K|a - b| \quad \forall a, b.$$  (14)

From Section 3.3, there always exists an observer gain matrix $L_p$ for linear time invariant system to satisfy the stability condition of the hidden state errors. In fact, a large number of results concerning the stability problem have been reported for nonlinear Luenberger-type observer. Necessary conditions for the existence of the stable observer gain $L_p$ has been obtained in [11] and [43] which satisfy
1) \((A, C)\) is observable; 2) nonlinear function satisfies the Lipschitz condition. Fortunately, most of the basic neural network architectures, such as LSTM and GRU, satisfy the Lipschitz condition. Therefore, it is possible and feasible to obtain a stable observer gain \(L_p\) for our proposed framework.

### 4.3. Variational Lower Bound

Based on the similar variational techniques in [22] and [21], we can optimize \(\theta, \phi\) and \(L_p\) by maximizing the evidence lower bound (ELBO). The loss function is derived as follows:

\[
\log P_{\theta}(x_{1:T}) = \log \int_{z_{1:T}} q_{\phi}(z_{1:T} | x_{1:T}) P_{\theta}(x_{1:T} | z_{1:T}) P_{\theta}(z_{1:T}) dz_{1:T} \\
\geq \int_{z_{1:T}} q_{\phi}(z_{1:T} | x_{1:T}) \log \frac{P_{\theta}(z_{1:T}) P_{\theta}(x_{1:T} | z_{1:T})}{q_{\phi}(z_{1:T} | x_{1:T})} dz_{1:T} = \mathbb{E}_{q_{\phi}(z_{1:T} | x_{1:T})} [\log P_{\theta}(x_{1:T} | z_{1:T})] \\
- \text{KL}(q_{\phi}(z_{1:T} | x_{1:T}) || P_{\theta}(z_{1:T})) \\
= \sum_{t=1}^{T} \mathbb{E}_{q_{\phi}(z_{1:t-1}, x_{1:t})} [\log P_{\theta}(x_{1:t} | z_{1:t})] \\
- \text{KL}(q_{\phi}(z_{1:T} | x_{1:T}) || P_{\theta}(z_{1:T})). \tag{16}
\]

Subsequently, according to the factorization of \(\text{KL}(q_{\phi}(z_{1:T} | x_{1:T}) || P_{\theta}(z_{1:T}))\), we can easily obtain

\[
\text{KL}(q_{\phi}(z_{1:T} | x_{1:T}) || P_{\theta}(z_{1:T})) = \int_{z_{1:T}} \cdots \int_{z_{T}} \frac{P_{\theta}(z_{1}, \cdots, z_{T})}{q_{\phi}(z_{1:T} | x_{1:T})} dz_{1:T} \\
= \text{KL}(q_{\phi}(z_{1} | x_{1:T}) || P_{\theta}(z_{1})) \\
+ \sum_{t=2}^{T} \mathbb{E}_{q_{\phi}} [\text{KL}(q_{\phi}(z_{1:t-1}, x_{1:T}) || P_{\theta}(z_{1:t-1}))].
\]

where \(\tilde{q}_{\phi} = q_{\phi}(z_{1:t-1}, z_{t-2}, x_{1:T})\).

Combining (15) and (16), the following variational bound is obtained as

\[
\log P_{\theta}(x_{1:T}) \geq \sum_{t=1}^{T} q_{\phi}(z_{1:t-1}, x_{1:T}) [\log P_{\theta}(x_{1:t} | z_{t})] \\
- \text{KL}(q_{\phi}(z_{1} | x_{1:T}) || P_{\theta}(z_{1})) \\
- \sum_{t=2}^{T} \mathbb{E}_{q_{\phi}} [\text{KL}(q_{\phi}(z_{1:t-1}, x_{1:T}) || P_{\theta}(z_{1:t-1}))].
\]

where \(\tilde{q}_{\phi} = q_{\phi}(z_{1:t-1}, z_{1:t-2}, x_{1:T})\).

The first term corresponds to the reconstruction error. The second term stands for the regularization of the hidden state \(z_{1}\), which is generated by the initial state \(z_{0}\). It should be noted that, the third term uses the future information \(x_{t+1:T}\) of the video sequence to inference the current hidden state \(z_{t}\). In [29], linear Kalman filter and smoother have been used in order to use the future information, which based on the assumption that the hidden variables are mapped to a linear time-varying state space model. Although more accurate inference for \(z_{t}\) can be obtained theoretically by using the future information, the optimal smoother is not easy to construct for nonlinear systems. In addition, it is difficult to verify the fact that their linear smoother constructed from [29] is better than using a predictor directly in a nonlinear state space model theoretically. On the other hand, our goal is to predict video frame. The construction of smoother is not necessary for our task of this paper. Therefore, we directly use the predictor (observer) \(q_{\phi}(z_{t} | z_{t-1}, x_{1:t-1})\) to approximate \(q_{\phi}(z_{t} | z_{t-1}, x_{1:T})\), which means \(z_{t}\) is inferred only from the past and current frames \(x_{1:t-1}\) in our framework. The entire model is differentiable. With the reparametrization trick [22], the parameters \(\theta, \phi\) and \(L_p\) can be jointly optimized by stochastic backpropagation technique.

### 5. Experiment Settings

It should be mentioned that, the following experiment results are not complete. The rest is still in progress, in which, we will add the some more complex and authoritative video datasets to illustrate the effectiveness for our proposed model. And the disentangled representation for the
**Algorithm 1** : The learning progress for deep variational Luenberger-type observer

| Step | Description |
|------|-------------|
| Step 1. | Sample the datapoint $x_{1:T}$ from the video sequences. Select the initial state value which guarantees the condition for $\hat{z}_0 \sim N(0; I)$; |
| Step 2. | For the time step $k$, obtain the compressed image feature $y_k$ via the encoder satisfying $y_k = \varphi^{enc}(x_k)$; |
| Step 3. | Predicting the mean $\mu_{k+1}$ and the variance $\text{diag}(\sigma^2_{k+1})$ of the latent state $\hat{z}_{k+1}$ from the addressed structures [13]; |
| Step 4. | Compute the predicted image feature $\hat{y}_{k+1}$ from the emission equation $\hat{y}_{k+1} = C\hat{z}_{k+1}$, the then reconstruct the image $\hat{x}_{k+1}$ by the decoder $\varphi^{dec}(\hat{y}_{k+1})$. Set $k = k + 1$. If $k < T$, go to Step 2, else go to Step 6; |
| Step 6. | Calculate the variational lower bound in [17] according to $\mu_{1:T}, \hat{\mu}_{1:T}, \sigma_{1:T}, x_{1:T}$ and $\hat{x}_{1:T}$; |
| Step 7. | Obtain the gradients with respect to $\theta$, $\phi$ and $L_p$ in [13], and then update $\theta$, $\phi$ and $L_p$ by using ADAM. |
| Step 8. | If the termination condition is not satisfied, then go to Step 1, else stop. |

5.2. Models and Training

**L. Encoder.** At each time instant, the input image is first converted to one-dimensional data via the encoder network. The input frames are first scaled down by a 4-layer downsample networks. Each downsample layer consists of a convolutional layer with a stride of 2, a batch normalization layer and a ReLU layer. The size of feature map is halved layer by layer, and the number of channels is doubled layer by layer. After 4 layers of downscaling, the dimension reduction and channel fusion are performed by a $1 \times 1$ convolutional layer with a channel dimension of 128. Finally, the global average pooling layer is connected to obtain the static feature vector $y_k$ corresponding to the input frame.

**Decoder.** At each time step, the predicted static feature $y_k$ can be decoded by the decoder into the corresponding predicted image frame $\hat{x}_k$, which is the size of $64 \times 64$. The decoder consists of five layers of upsampling, and the upsampling structure of each layer is consistent and the parameters are different. The upsampling layer is composed of a deconvolution, a BN layer, and a ReLU layer. The channel number for each layer corresponds to the encoder mentioned above.

**Transition Function/Predictor.** Our transition function/predictor includes two parts: linear and nonlinear functions. The specific expression is $\hat{z}_{k+1} = Az_k + f_\theta(z_k, h_k)$. The matrix $A$ is given in [12], and the nonlinear function $f_\theta(z_k, h_k)$ is designed by using a GRU cell. To be specific, the GRU unit has only one layer, and the dimensions of the input $z_k$ and the hidden historical state $h_k$ are 256 and 512, respectively. Since the output dimension of the GRU is 512, we connect a fully connected layer to get the predicted latent state $\hat{z}_{k+1}$ with a dimension of 256.

**Emission Function.** We use the form $y_k = Cz_k$ as our emission function, where the parameter of the matrix $C$ is selected in [11].

**Luenberger-type Observer.** We extract the static and dynamic information of a given input sequence through our proposed deep Luenberger-type observer. The specific form is $\hat{z}_{k+1} = Az_k + f_\theta(z_k, h_k) + L_p(y_k - Cz_k)$. The structure of the one and second parts are the same as the transition function/predictor. The last part can be considered as a corrector. The matrix $L_p$ has a learnable parameter. However, in this paper, to simplify the train process, the matrix $L_p$ is fixed as $L_p = \begin{bmatrix} 0.9I & 0.1I \end{bmatrix}^T$.

During the training, we take a video sequence of length 20 as a sample. The raw images are compacted and reconstructed by our encoder, emission function and decoder, respectively. The first 10 frames of the input generate reconstructed images through our observer model, and the last 10 frames are trained using the transition function/predictor model. The specific training process has been expressed in Algorithm 1.
6. Results and Discussion

Experiment results are shown in Figs. 3 and 4. Fig. 3 and Fig. 4 depict the predicted frames from our model and baselines for the Bouncing Balls dataset and the pendulum dataset, respectively. The first row is the input for the first 10 frames, the second row is the ground truth for the last 10 frames. The next few rows are the predicted results of DDPAE [17], ConvLSTM [36], DeepRNN [33] and ours. The MSE between the ground truth and the prediction of these methods for each dataset can be seen in Table 1. Here, MSE means the mean square error. It is worth noting that in the pendulum experiment, DDPAE [17] did not learn the motion information of the pendulum. The main reason is that DDPAE assumed that the object information is time-invariant only from the heuristic perspective. It is difficult for DDPAE to judge whether the pendulums are in different positions from the same one or are the different objects for a given sequence. Our model effectively distinguishes static features and dynamics by introducing our addressed deep state space models.

In future work, we aim to extract a more useful and well-interpretable representation which can be easily used for the downstream task. The results of [25] has demonstrated the fact that the unsupervised disentanglement learning without inductive biases is theoretically impossible, and it is necessary to use some supervised signals from [28] in order to obtain an ideal disentangled representation. As such, our compressed feature \( y_k \) can further serve as an interface of supervise signals to better guide the learning of meaningful features.

7. Conclusions

This paper studied the video prediction problem which remained a challenging and demanding task due to the high-dimensionality and stochastic complexity of video sequences. We proposed a new deep state space model called the Deep Variational Luenberger-type Observer to extract the disentangled representations from the video. A salient feature of this model lied in its ability to take any initial states and converge to the optimal states, which made the training easier and the model more robust. The experimental studies were explored to substantiate the performance and effectiveness of the video prediction model derived in this paper.

References

[1] Masoud Abbaszadeh and Horacio J. Marquez, LMI optimization approach to robust \( H_\infty \) filtering for discrete-time nonlinear uncertain systems, In IEEE 2008 American Control Conference, 2008.

[2] Byron Boots, Arunkumar Byravan, and Dieter Fox, Learning predictive models of a depth camera and manipulator from raw execution traces. In International Conference on Robotics and Automation, 2014.

[3] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine, Stochastic variational video prediction, In International Conference on Learning Representations, 2018.

[4] Apratim Bhattacharyya, Michael Hanselmann, Mario Fritz, Bernt Schiele and Christoph-Nikolas Straehle, Conditional flow variational autoencoders for structured sequence prediction, arXiv preprint arXiv:1908.09008, 2019.

[5] Philipp Becker, Harit Pandya, Gregor Gebhardt, Cheng Zhao, James Taylor and Gerhard Neumann,
Recurrent Kalman networks: factorized inference in high-dimensional deep feature spaces, *arXiv preprint arXiv:1905.07357*, 2019.

[6] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio, A recurrent latent variable model for sequential data, In *Advances in Neural Information Processing Systems*, 2015.

[7] Yuan Cao, Qiuying Li, Lei Chen, Junping Zhang and Leiming Ma, Video Prediction for Precipitation Now-casting, *arXiv preprint arXiv: 1907.08069*, 2019.

[8] Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee and Sergey Levine, Visual foresight: Model-based deep reinforcement learning for vision-based robotic control, *arXiv preprint arXiv:1812.00568*, 2018.

[9] Babak Esmaeili, Hao Wu, Sarthak Jain, N Siddharth, Brooks Paige, and Jan-Willem van de Meent, Hierarchical disentangled representations, *Artificial Intelligence and Statistics*, 2019.

[10] Hao-Shu Fang, Guansong Lu, Xiaolin Fang, Jian-wen Xie, Yu-Wing Tai and Cewu Lu, Weakly and semi supervised human body part parsing via pose-guided knowledge transfer, In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[11] Chelsea Finn, Ian Goodfellow, and Sergey Levine, Unsupervised learning for physical interaction through video prediction. In *Advances in Neural Information Processing Systems*, 2016.

[12] Marco Fraccaro, Simon Kamronn, Ulrich Paquet and Ole Winther, A disentangled recognition and nonlinear dynamics model for unsupervised learning. In *Advances in Neural Information Processing Systems*, 2017.

[13] Chelsea Finn and Sergey Levine, Deep visual foresight for planning robot motion, In *International Conference on Robotics and Automation*, 2017.

[14] Marco Fraccaro, Sren Kaae Snderby, Ulrich Paquet and Ole Winther, Sequential neural models with stochastic layers, In *Advances in Neural Information Processing Systems*, 2016.

[15] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David WardeFarley, Sherjil Ozair, Aaron Courville and Yoshua Bengio, Generative adversarial nets. In *Advances in neural information processing systems*, 2014.

[16] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende and Alexander Lerchner, Towards a definition of disentangled representations, *arXiv preprint arXiv:1812.02230*, 2018.

[17] Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles, Learning to decompose and disentangle representations for video prediction, In *Advances in Neural Information Processing Systems*, 2018.

[18] Rudolf Emil Kalman, A new approach to linear filtering and prediction problems, *Journal of basic Engineering*, Vol. 82, pp. 35–45, 1960.

[19] Adam R Kosiorek, Hyunjik Kim, Ingmar Posner, and Yee Whye Teh, Sequential attend, infer, repeat: Generative modelling of moving objects, In *Advances in Neural Information Processing Systems*, 2018.

[20] Maximilian Karl, Maximilian Soelch, Justin Bayer and Patrick van der Smagt, Deep variational bayes filters: Unsupervised learning of state space models from raw data, In *International Conference on Learning Representations*, 2017.

[21] Rahul G. Krishnan, Uri Shalit and David Sontag, Structured inference networks for nonlinear state space models, In *31st AAAI Conference on Artificial Intelligence*, 2017.

[22] Diederik P Kingma and Max Welling, Auto-encoding variational bayes, In *International Conference on Learning Representations*, 2014.

[23] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang and Cewu Lu, CrowdPose: Efficient crowded scenes pose estimation and a new benchmark, In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[24] Davida Gilbert Luenberger, An introduction to observers, *IEEE Transactions on Automatic Control*, Vol. 16, No. 6, pp. 596–602, Dec. 1971.

[25] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Ratsch, Sylvain Gelly, Bernhard Scholkopf and Olivier Bachem, Challenging common assumptions in the unsupervised learning of disentangled representations, In *International Conference on Machine Learning*, 2019.

[26] William Lotter, Gabriel Kreiman, and David Cox, Unsupervised learning of visual structure using predictive generative networks, In *Workshop Track of International Conference on Learning Representations*, 2016.
[27] William Lotter, Gabriel Kreiman, and David Cox, Deep predictive coding networks for video prediction and unsupervised learning. In International Conference on Learning Representations, 2017.

[28] Francesco Locatello, Michael Tschannen, Stefan Bauer, Gunnar Ratsch, Bernhard Scholkopf and Olivier Bachem, Disentangling factors of variation using few labels, arXiv preprint arXiv:1905.01258, 2019.

[29] Bryan Lim, Stefan Zohren and Stephen Roberts, Recurrent neural filters: Learning independent Bayesian filtering steps for time series prediction, arXiv preprint arXiv:1901.08096, 2019.

[30] Michael Mathieu, Camille Couprie, and Yann LeCun, Deep multi-scale video prediction beyond mean square error, In Proceedings of the International Conference on Learning Representations, 2016.

[31] Dorde Miladinović, Muhammad Waleed Gondal, Bernhard Scholkopf, Joachim M Buhmann, and Stefan Bauer, Disentangled state space representations. arXiv preprint arXiv:1906.03255, 2019.

[32] Christian Naesseth, Scott Linderman, Rajesh Ranganath, and David Blei, Variational sequential Monte Carlo, In International Conference on Artificial Intelligence and Statistics, 2018.

[33] Bo Pang, Kaiwen Zha, Hanwen Cao, Chen Shi and Cewu Lu, Deep RNN framework for visual sequential applications, In IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[34] Syama Sundar Rangapuram, Matthias Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski, Deep state space models for time series forecasting, In Advances in Neural Information Processing Systems, 2018.

[35] Jack Schwartz, Nonlinear functional analysis, Gordon and Breach Science Publishers, New York, 1969.

[36] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo, Convolutional LSTM network: A machine learning approach for precipitation nowcasting, In Advances in Neural Information Processing Systems, 2015.

[37] Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong and Wang-chun Woo, Deep learning for precipitation nowcasting: A benchmark and a new model, In Advances in Neural Information Processing Systems, 2017.

[38] Chen Sun, Per Karlsson, Jiajun Wu, Joshua B Tenenbaum and Kevin Murphy, Stochastic prediction of multi-agent interactions from partial observations, In International Conference on Learning Representations, 2019.

[39] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov, Unsupervised learning of video representations using lstms. In International Conference on Machine Learning, 2015.

[40] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin and Honglak Lee, Decomposing motion and content for natural video sequence prediction. In International Conference on Learning Representations, 2017.

[41] Nicholas Watters, Loic Matthey, Christopher P Burgess, and Alexander Lerchner, Spatial broadcast decoder: A simple architecture for learning disentangled representations in VAEs, arXiv preprint arXiv:1901.07017, 2019.

[42] Zhenjia Xu, Zhijian Liu, Chen Sun, Kevin Murphy, William T. Freeman, Joshua B. Tenenbaum and Jiajun Wu. Unsupervised discovery of parts, structure, and dynamics, In International Conference on Learning Representations, 2019.

[43] Wei Zhang, Housheng Su, Shengchao Su and Dazhong Wang, Nonlinear $H_{\infty}$ observer design for one-sided Lipschitz systems, Neurocomputing, Vol. 145, No. 5, pp. 505–511, 2014.

[44] Xun Zheng, Manzil Zaheer, Amr Ahmed, Yuan Wang, Eric P Xing, and Alexander J Smola. State space LSTM models with particle MCMC inference, arXiv preprint arXiv:1711.11179, 2017.