Comparative Evaluation of Generative Models for Future Frame Prediction

Chandrika Sundakampalayam Paramasivam
Hochschule Bonn-Rhein-Sieg, Sankt Augustin
Autonomous Systems
Department of Computer Science[1]
Santosh Thoduka
Hochschule Bonn-Rhein-Sieg, Sankt Augustin
Autonomous Systems
Department of Computer Science[1]

Abstract

Safety is a key factor in the field of robotics, autonomous systems and related domains. Due to the unforeseen and dynamic environment, it is challenging to predict the plausible future frame while having the knowledge of the past. Anomalous event can be detected by predicting the future frame and comparing this observed frame sequences with the nominal or actual frame. In order to establish a means for qualitative and quantitative comparison of these existing generative models, an approach and a set of criteria should be established. To facilitate this, the state-of-the-art models are collected, analysed and listed. Using these collected algorithms, a new criteria is introduced to benchmark the various generative models to anticipate future frame sequences. Experiments are conducted on the selected state-of-the-art generative models to illustrate anomaly detection by future frame prediction task on SM-MNIST dataset. The performance of the model is evaluate using appropriate metrics.

1. Introduction

Robot system claims to be autonomous if it has the capability to monitor its own system, detect the fault, if any and repair them automatically.

Future frame prediction (FFP) refers to anticipation and generation of future image sequences or frames given a set of consecutive past images as input to the model. One important aspect that help human's in decision making process is their ability to predict the future based on their past experiences. Similarly, leveraging this approach in machines by teaching them to predict the future events based on the environments internal representations could be of great relevance for various task in real world applications. Anticipating automatically the future frames of a video is one approach that is developing recently due its applications in our day to day life. There are quite many models that can predict the future frames (Aigner & Körner, 2018), (PredictingFFusingCycleGAN, 2019; Kaur & Das, 2020), (Liu et al., 2018).

Further, Anomaly detection refers to the identification of unexpected or undesired or abnormal or rare events in the given dataset, that differ from the norm. Normal or nominal class of events are those that includes the frequently appearing objects and non-stochastic foreground movements so, here it means within acceptable or expected boundaries. whereas anomalous class covers all types of unseen events and rare objects. Therefore, Anomalies can be detected by predicting the future actions. Detecting the anomalies in video (video is the sequences of image frames) refers to Video Anomaly Detection (VAD).

There are many ways to spot anomaly, few based on paper (Pang et al., 2020) are future extraction, Learning feature representation of normality and End to end anomaly score learning. Our approach is a sub branch of learning feature representation of normality approach which is Generic normality feature learning, that uses generative models.

1.1. Motivation

Motivation behind this work is anomaly detection in robotics. Self supervised learning is a form of unsupervised learning where the data itself provides the supervision. Re-

[1]Hochschule Bonn-Rhein-Sieg, Bonn, North Rhine-Westphalia, Germany. Correspondence to: Chandrika SP <chandu.sspac@gmail.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

search in GM is improving such as VAE (Kingma & Welling, 2014b), LSTMs (Hochreiter & Schmidhuber, 1997), GANs (Goodfellow et al., 2014) and etc., A sub domain of unsupervised learning is anomaly detection which is well known in the machine learning field. In images and videos, anomaly detection is challenging due to the image's high dimensional structure along with the temporal variations across frames. So, we aim to work on GM for video anomaly detection by Future frame prediction (FFP).FFP is one among the many approaches for video anomaly detection. Mostly generative models are used for this future frame prediction since their results are more realistic, sharp and their ability to capture various new patterns.

1.2. Problem Statement

It is vital for any model or algorithm to output a sharp and realistic future frames when conditioned on the past few video frames. When the generated frame is blurry, not realistic and etc., then anomaly is detected. As a result, it is essential for the deep generative model to predict real and sharp next frames. There's scarcity of historical knowledge or criteria of comparison of these generative models for future frame prediction. To benchmark the existing models to account for anomaly detection by future frame prediction. Therefore, the objectives are: Summarizing the major existing generative models, Compare and analyse the various techniques in order to have a better understanding, A quick and easy to use table for high level overview and comparison criteria, Categorize the models based on few criterion's to better visualize the scope, Metainformation information to enhance better understanding and to analyse the existing techniques for the task of anomaly detection, Gives an advice on algorithm selection for typical real world tasks to manufacturers in Automation Industries, deep learning practitioners.

To conclude, an experimentation ie., illustration of anomaly detection in Stochastic Moving MNIST dataset by predicting future frame using one of the state-of-art model namely Stochastic Video Prediction Model with Learned Prior is discussed here. Appropriate metric(s) that are able to depict this reconstruction loss which directly represents the anomaly has to be selected from the vast collect that is available in the literature. The selected metric is used to measure the reconstruction loss in the generated frames.

2. Related Work

Recent studies on Generative Models(GM) has categorized them into unsupervised fundamental models, Autoencoder (AE) based models, autoregressive models, GAN (Generative Adversarial Networks) based models and Autoencoder-Generative Adversarial Networks hybrid models to associate easily the generative models.

Unsupervised models were mostly used for texture synthesis and classification of handwritten digits, due to blurry output these models did not perform sufficiently. Restricted Boltzmann Machines (RBM) (Salakhutdinov et al., 2007), is a variant of Boltzmann Machines (BM) have tackled many problems as seen in(Turhan & Bilge, 2018). Markov Chain Monte Carlo (MCMC) (Geyer, 1991) used random dice in RBMs to generate descent samples.

The other GM's that were leveraged for FFP are Autoencoders (AE) (Kingma & Welling, 2014b), Variational Autoencoder (VAE) (Kingma & Welling, 2014a), Generative Advesarial Networks (GAN) (Goodfellow et al., 2014), Optical flow, later AE-GAN hybrid models as seen in (Turhan & Bilge, 2018) each having their respective advantages and disadvantages. Basically, AE's network architecture can vary from simple feedforward network, LSTM network or Convolutional Neural Network (CNN) depending on the application or use case. The variants of AE are (Wikipedia contributors, 2021): Sparse Autoencoder, Denoising Autoencoder,Contractive Autoencoder, Variational Autoencoder and etc., Among the mentioned variants of AEs, mostly variants of VAE is leveraged for future frame prediction when compared to others, as seen in (Wikipedia contributors, 2021). The model we choose for our work is a variant of VAE.

After GANs, CGAN ie., Conditional Generative Adversarial Networks (Mirza & Osindero, 2014) was proposed. Laplacian Pyramid of GAN (LPGAN) (Denton et al., 2015) is an extension of GAN. Deep CGAN (DCGAN) (Radford et al., 2016a) is used to generate images. DCGAN has both D and G being convolutional. GRAN (Im et al., 2016) is an extension of GAN based model and is a sequential process similar to LAPGAN and DRAW models. In recent times, GAN is one of the best performing models in terms of generating data but it does suffer from few drawbacks such as mode collapse, instability during training and etc., Therefore, to overcome these and researchers combined GAN with AE or with VAE (Aigner & Körner, 2018) as one model in order to leverage benefits of both and to overcome the challenges that GAN faces. However, we choose to experiment with a GM that is a variant of VAE.

The State-of-the-Art (SoTA) models for FFP are: Modeling deep temporal dependencies with recurrent "grammar cells", Michalski et al. (Michalski et al., 2014), Unsupervised Learning of Video Representations using LSTMs, Srivastava et al. (Srivastava et al., 2016), Learning to Linearize Under Uncertainty, Goroshin Mathieu et al. (Goroshin et al., 2015), Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting, Shi et al. (Shi et al., 2015), Deep multi-scale video prediction beyond Mean Squared Error, Mathieu et al. (Mathieu et al., 2016), Unsupervised Learning of Visual Structure Using

Predictive Generative Networks, Lotter et al. (Lotter et al., 2016), Dynamic Filter Networks, Brabandere Jia et al. (Brabandere et al., 2016), The Pose Knows: Video Forecasting by Generating Pose Futures, Walker et al. (Walker et al., 2017), Unsupervised Learning of Disentangled Representations from Video, Denton et al. (Denton & Birodkar, 2017), Learning to generate long term future via hierarchical prediction, Villegas et al. (Villegas et al., 2018), Stochastic Variational Video Prediction, (Babaeizadeh et al., 2017), Improved Conditional VRNNs for Video Prediction (Castrejon et al., 2019), Future Video Synthesis with Object Motion Prediction (Lu et al., 2017) and etc., Each of these related work has it's own merits and demerits, our work concentrates on the paper implementation of (Denton & Fergus, 2018) which is able to predict long sequences of future frames.

3. Comparative Evaluation of Generative Models

3.1. Criteria

The criteria that I considered to compare, evaluate the existing GM's are:

- Learning: A machine is said to be learning from past experiences, if it's able to learn and performance is improving with experience automatically without being explicitly programmed in the given task [(tutorialspoint workers), (Brownlee, 2019)]. There are different ways to train an algorithm. Each having their own advantages and disadvantages. Based on the type of data the algorithms ingest, their pros ans cons can be understood. In Machine Learning, there are two types of data - labeled and unlabeled data. Based on the use case broadly the learning types of algorithm can be classified are: Supervised, Unsupervised, Semi-Supervised, Reinforcement Learning and etc.,

- Network Architecture: It refers to the class of specific model that is used in the model's design. It refers to the complete framework of a model's network ie., the way a neural network is made of complex structure of artificial neurons that take the input, process it and produce the output. In a Neural Network (NN) the flow of information can be mainly categorized as: Feedforward Networks, Feedback or Recurrent Networks, Multilayer Perceptron (MLP), Convoultutional Neural Networkes (CNN), Long-Short Term Memeory (LSTM).

- Training(Gradient Updates): As seen in (data science authors, 2021) Training a model means to iteratively update the parameters during training phase to its optimal value, so that the model sufficiently learns the provided input train data. The models performance can now be evaluated using the test data (test data, are the samples or data that is not present in the train data ie., the unseen data) by its accuracy or precision. To evaluate the quality or performance of the statistical or deep learning model, evaluation metrics can be used. The standard and popular technique to train any deep learning model is by Backpropagation.

- Optimization Techniques: Goal of optimization is to build the model that performs good and gives accurate predictions in particular set of cases. So, this process involves to adjust the hyperparameters to minimize the cost function by using one of the optimization techniques (Gavrilova, 2020), (Doshi, 2019). The major optimization techniques in ML are: Gradient Descent, Stochastic Gradient Descent (SGD), RMSProp and Adam.

- Evaluation Metrics: Evaluation metrics (Agarwal, 2019) are used to quantify the performance of a predictive model and to measure the statistical or machine learning models quality. Many different types of evaluation metric used to test the model specifically in the scope of future frame prediction are: Structural Similarity Index (SSIM), Multi-scale Structural Similarity Index (MS-SSIM), Inception, Peak Signal to Noise Ratio (PSNR), Area Under Curve-Receiver Operating Characteristics (AUC-ROC), Accuracy/Precision/Recall, F1 score, Cross Entropy, Mean Square Error (MSE), Categorical cross entropy, Kullback Leibler Divergence (KLD).

- Datasets: Dataset can be defined as the collection of data that could be of images or frame sequences or audio or text or any such data which used in the deep learning models to initially train the model or algorithm and then to test it, to evaluate its performance. Therefore, based on ML perspective they can be numerical dataset, Categorical dataset, Time-Series dataset and Text data. These in high level can be classified as natural ie., captured in camera and etc or synthetic ie., artificially created in order to perform desired task.(Zhang, 2018) Natural dataset, UCF101, UCSD Ped1 Ped2, Sports 1M, NORB videos, Artificial or simulated Dataset like MMNIST, SM-MNIST, BAIR Push and etc.,

Based on above criterions the existing GM's are compared and is as tabulated in 1. A detailed comparative evaluation table also the link to my GitHub repository where the code is saved: $https://github.com/ChandrikaSP/Research-Development$

4. Methodology

Stochastic Video Generation with Learned Prior (SVG-LP) (Denton & Fergus, 2018) is one of the state-of-the-art model that can spot anomaly by anticipating the future frames. This model does not use GAN in the architecture to deal with uncertainty in the pixel-space as it introduces associated difficulties such as training instability and mode collapse. But our approach relies on L2 loss pixel-space reconstruction with not using GAN in the models architecture or any other adversarial term.

Here, Learned prior network aim is to anticipate the points or instants when the two digits in SM-MNIST dataset can collide ie., it understands the stochasticity ie., randomness or the point of uncertainty in the dataset which is essential to generate good sharp and realistic future frames when conditioned on few previous frames.

4.1. Stochastic Video Generation with Learned Prior

Various generative models are proposed for sampling one or few future frames when conditioned on some or one future frame. Our approach is one such generative model that uses unsupervised learning which is leveraged for generating new video frames. The dataset that is used here for training and testing of future frame prediction is Stochastic Moving MNIST (SM-MNIST) dataset. The model has two components they are:

- the prior distribution which is used to draw the latent variables. Prior distribution could be learned during training or fixed
- the prediction model: p_θ, it generates or samples the next ie., future frames, frames $x_t^{'}$ from time steps $t+1$ when it is conditioned on past or previous frames only $x_{1:t-1}$ from time steps at beginning to the previous time step of the current one $1:t-1$ where the current or present frame is x_t at time step t and also the latent variable z_t

The description of how the model works can be split into two parts. The train phase is followed by test phase. The inference model, learned prior is used along with future predictor model. Inference model is not used during test time.

Inference model is analogous to the Encoder and is used only during training not while test phase, it is a time-dependent inference network that takes the current frames x_t of the SM-MNIST dataset as the input ie., prediction models target together with the past or previous frames $x_{1:t-1}$. To initially infer and compute the hidden state representation and then to estimate the probability distribution for each of the images in the train set. It has E which is DCGANs

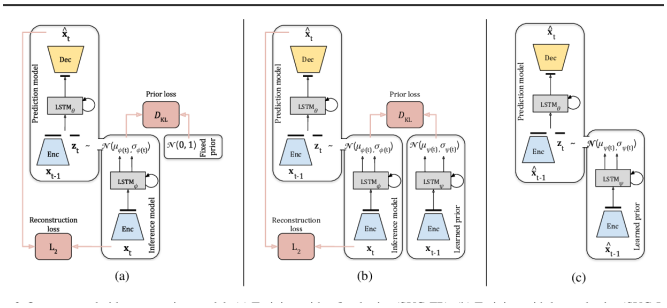

Figure 1. SVG video generation model. (a) Training with a fixed prior (SVG-FP); (b) Training withlearned prior (SVG-LP); (c) Generation with the learned prior LP model. The red boxes show the lossfunctions used during training (Denton & Fergus, 2018)

discriminator type (Radford et al., 2016b) with output dimensionality being $|h| = 128$. This output of the encoder part is given as input to single hidden layer LSTM network with 256 cells in each layer to compute the gaussian distribution $\mathcal{N}(\mu_\phi(x_{1:t}), \sigma_\phi(x_{1:t}))$. Each of the network has linear embedding layer also output layer being fully connected. From this we sample z_t. To prevent z_t from copying the current frame, x_t, $q_\phi(z_t|x_{1:t})$ is forced to be close to the $p(z)$ prior distribution using the Kullback-leibler divergence (D_{KL}) term. This KL divergence term will constrain the information that z_t could carry by forcing it to have only new information that is absent in the previous frames.

Learned Prior (LP) model is also similar to the encoder, the working is similar to Inference model but LP is used both during train and test phase.LP architecture is like E ie., an inference model used only during train and not at test. It purpose is to infer the instants of randomness in input data, to generate real, clear images even at points when there is uncertainity.E outputis fed to single hidden layer convolutional LSTM with 256 cells in each layer to estimate the gaussian distribution $\mathcal{N}(\mu_\psi(x_{1:t-1}), \sigma_\psi(x_{1:t-1}))$. Each of this network has the linear embedding layer. From this we sample z_t by constraining it with $\mathcal{N}(\mu_\phi(x_{1:t}), \sigma_\phi(x_{1:t}))$ using kullback-leibler divergence D_{KL}. This prior network is trained jointly with other models by maximizing the objective function (Denton & Fergus, 2018):
$\mathrm{L}_{\theta,\phi,\psi}(x_{1:T}) = \sum_{t=1}^{T}[E_{q_\phi(z_{1:t}|x_{1:t})} log p_\theta(x_t|x_{1:t-1}, z1:t) - \beta D_{KL}(q_\phi(z_{t|x_{1:t}})||p_\psi(z_t|x_{1:t-1}))]$

This refers to the model as 4.1 SVG-LP. The prior distribution $p(z)$ can be either fixed prior (SVG-FP) or learned prior (SVG-LP). Fixed prior is the simplest one with fixed Gaussain $\mathcal{N}(0, I)$ as typically used in case of variational autoencoder (VAE) (Kingma & Welling, 2014a) with 0 mean and the covariance is of unit variance ie., identity matrix ie., $\mathcal{N}(\mu = 0, \sigma = I)$ it is SVG-FP whose deficit is the interdependencies between the frames is ignored so, at every

Paper	Model	Dataset	Evaluation Metric	Training	Code
(Lu et al., 2017)(2017)	VAE+GAN	MNIST,PV, UCF101,Sports1M	PSNR	RMSProp	N
(Vondrick et al., 2016)(2016)	VAE+GAN	UCF101, Flickr videos	AMT(M)	SGD for G and D	Y
(Castrejon et al., 2019)(2019)	VRNN	BAIR, MNIST , Cityscape	SSIM,LPIPS,FVD	Stochastic BP	Y
(Aigner & Körner, 2018)(2018)	VAE+GAN, 3D conv., Improvised PGGAN	MMNIST, KTH action & cityscapes	train:MSE+ WGAN+GP test: MSE, PSNR, SSIM	Generation of images over small incremental updates to loss	Y
(Wu et al., 2020)(2020)	GAN	Cityscapes, KITTI	MS-SSIM, LPIPS	Forward-backward consistency loss, tracking algorithm	Y
(Furnari & Farinella, 2020)(2019)	2LSTM (RLSTM, ULSTM), 3branches (RGB,Flow ,OBJ) with Late fusion)	EPIC-Kitchen, EGTES-Gaze	Top-k prediction method, AMOC curve	Sequence Completion Pre-training	Y
(Jayaraman et al., 2018)(2018)	TAP+CGAN, TAP+CVAE	3 Simulated robot manipulation setting,BAIR Push(testing)	Bottleneck discovery frequency metric	First, D is trained using fixed learning rate & Adam Optimzer	Y
(Sagar, 2020)(2020)	GAN+VAE	MNIST, LSUN, CelebA-HQ	MMD, MSSSIM	SGD	Y

Table 1. Comparison of existing generative models for future frame prediction.

time instant frame sequences are produced randomly and the other approach being the SVG-LP model, which is quite sophisticated approach where the prior varies across time and is a function of all past frames but not including the current or target frame and the conditional gaussian distribution being $\mathcal{N}(\mu_\psi(x_{1:t-1}), \sigma_\psi(x_{1:t-1}))$. Further, the *Future predictor* is a recurrent network, its architecture is simialr to the variational autoencoder model. The encoder consists of DCGAN discrminator network whose input being the past frames x_{t-1} to output the hidden representation h_{t-1} which along with the z_t is given as input to two LSTMs having 256 cells in each layer. Before pasing the output of this LSTMs to the decoder network it is passed on through the tanh non-linearity.

SVG-LP model is helpful when there is stochasticity in the data so in our experiment when the ball hits the frame or boundaries of the video clips, then the digit bounces with a random velocity any direction. So, during the point of contact is when the uncertainty comes into picture and which causes randomness and the model will not be able to successfully predict while using fixed prior model but can predict pretty good while using learned prior model during both train and test phase.

Stochastic Moving MNIST (SM-MNIST) data is being used for this experiment whose image size is 64*64. In our experiment, we are downloading the SM-MNIST dataset from the PyTorch library which has the torchvision package, (torchvision) the sequences of frames is generated on the go while execution.

5. Experiments

5.1. Anomaly Detection

The three anomalies that we introduce in the dataset are: First added random noise pixel-wise in all video clips (similar to the gaussian noise) in the test dataset. Our model during testing as input to the model ie., the past frames, it will receive video clips with anomaly. So, since the model has not seen such dataset with anomaly so, it will not be able to infer or learn completely the data distribution of the image as it is corrupted. This type of anomaly can be due to hardware problems of the device because of which all the images generated has anomaly. Second is randomly increase the brightness, contrast and saturation in all the test datasets which is similar to introducing the lighting contrast changes in all video clips in the dataset. This type of simulation of dataset with anomaly could be due to hardware problem in the device or some type of bright light falling on the camera while capturing the images in camera during test time. The

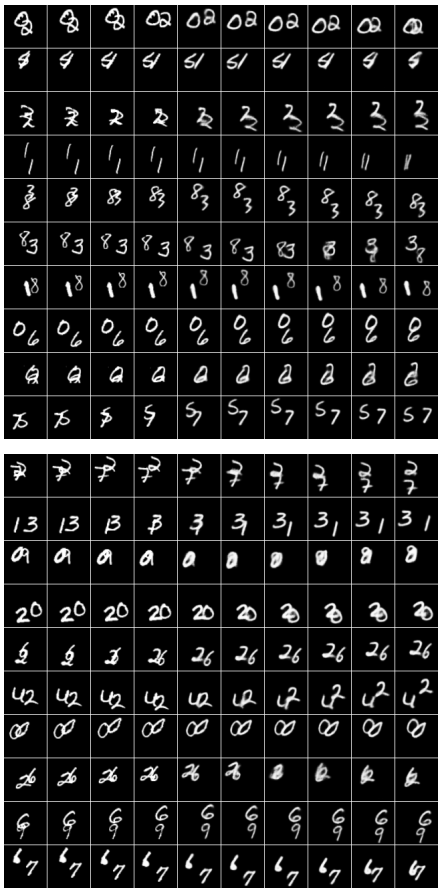

Figure 2. Samples generated after training the model SVG-LP

predicted loss was low when compared to the predicted loss of the nominal dataset. Where the PSNR metric value has contributed more in the predicted loss value which was very high in range of 12.5 to 30.4 when no anomaly was present and reduced when the anomaly was being introduced. Finally, rotating or tilting of only the digits by 90 degrees that is given as input to all video clips during test phase. This type of anomaly could probably occur due to hardware issues in the device by which the output data from the device in our case the SM-MNIST data is corrupted ie., not nominal data.

loss func,	nominal frame loss func, value ,	anomalous frame loss func, value
l_{MSE}	0.001-0.07	0.19-0.34
l_{SSIM}	0.72-0.95	$6.02e^{-03}$-$9.5e^{-02}$
l_{PSNR}	11.23-30.4	4.2-4.9
l_{pred}	11.5-35.62	4.5-5.19

Table 2. l_{pred} (Prediction loss) value readings for nominal and anomalous test data

Therefore, these anomalies are introduced all together to the SM-MNIST dataset to test set and the performance of the SVG-LP model is evaluated by using predicted loss which is the summation of these: SSIM, PSNR and MSE as seen in 2. We anticipate the future video clips during test phase and if there's anomaly then, the predicted loss will be high ie., error is high and without any anomaly the predicted loss must be low. We used PSNR metric which is high when it is nominal frame and is low for anomalous frame. So, in our case during testing the predicted loss or loss function is high when there's no anomaly in the test dataset and it is low when there's anomaly in the test dataset. Therefore, the predicted lossl_{pred} which is the combination of MSEl_{MSE}, SSIM $l_{SSIM}$and PSNR l_{PSNR} is noted for the test dataset with no anomaly and with anomaly. When no anomaly is added to the dataset, then the predicted loss value is quite high compared to the predicted loss value when the dataset is anomalous. So, from the predicted loss value we can infer that if the given test data is nominal or not and eventually, by anticipating the future frames we can detect the occurrence of anomaly.

Figure 3. Samples generated during testing the model SVG-LP with no anomaly introduced ie., nominal frames

6. Conclusion

To anticipate future frames for the task of anomaly detection various deep learning techniques are leveraged. There are various techniques for anomaly detection but the predicted output frame is blurry or unrealistic. Currently, the SoTA model leveraged for future frame anticipation is generative models, because of their ability to generate real and sharp future images. Numerous work exists to anticipate next sequence of frames using various models but there is lack of existing research work specifically in context to spot anomaly by leveraging generative models. Research is still going on in this field so that generated output is much closer to the future frame for the given input frames. Also, there's a scarcity of existing work performing the comparative eval-

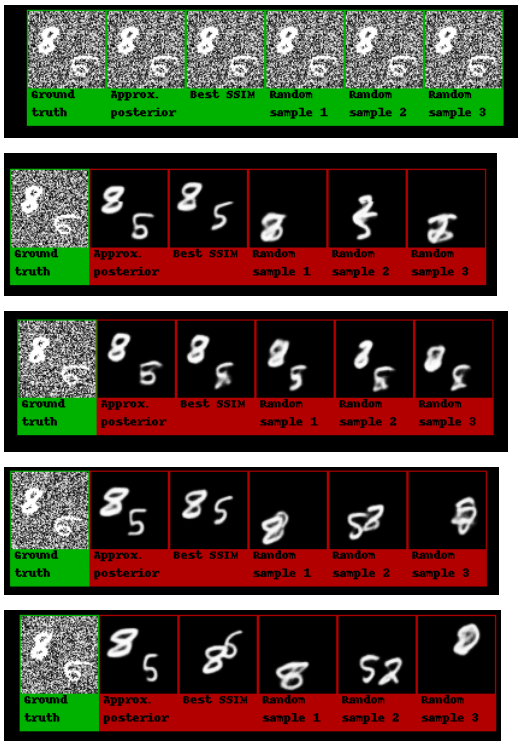

Figure 4. Samples generated during testing the model SVG-LP with anomaly being added ie., anomalous frames

uation specifically using generative models for this given scope. Focus on to study,list, evaluate, and compare present generative models which are leveraged for video future anticipation. Categorize the existing state-of-the-art models based on few categories, which will be helpful for deep learning practitioners and robotics manufacturers to select a model for typical real-world tasks. Due to the ability of generative models to generate new samples that are similar to the input frames, they are being used more and gives better results compared to the other models. To conclude, my works focus is comparative evaluation for future frame prediction for the anomaly detection tasks using specifically generative models.

Acknowledgements

I would like to thank Prof. Dr. Paul G. Plöger and M. Sc. Santosh Thoduka for provinding me the opportunity to work on this Research and Development project. I have to specially thank M. Sc. Santosh Thoduka for his continuous guidance and valuable support throughout the project. I would like to thank Divin Devaiah Ulliyada Arun and Lokesh Veeramacheneni for their valuable suggestions and for kindly helping me in understanding project related concepts. Finally, I extend my gratitude to my family especially and friends for their continuous motivation, immense support, encouragement and love.

References

Agarwal, R. The 5 classification evaluation metrics every data scientist must know. 2019.

Aigner, S. and Körner, M. Futuregan: Anticipating the future frames of video sequences using spatio-temporal 3d convolutions in progressively growing gans, 2018.

Babaeizadeh, M., Finn, C., Erhan, D., Campbell, R. H., and Levine, S. Stochastic variational video prediction. *CoRR*, abs/1710.11252, 2017. URL http://arxiv.org/abs/1710.11252.

Brabandere, B. D., Jia, X., Tuytelaars, T., and Gool, L. V. Dynamic filter networks, 2016.

Brownlee, J. 14 different types of learning in machine learning. 2019. URL https://machinelearningmastery.com/types-of-learning-in-machine-learning/. [Online:accessed 14-January-2021].

Castrejon, L., Ballas, N., and Courville, A. Improved conditional vrnns for video prediction. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7607–7616, 2019.

data science authors, E. Model training. 2021. URL https://elitedatascience.com/model-training. [Online:accessed 14-January-2021].

Denton, E. and Birodkar, V. Unsupervised learning of disentangled representations from video, 2017.

Denton, E., Chintala, S., Szlam, A., and Fergus, R. Deep generative image models using a laplacian pyramid of adversarial networks, 2015.

Denton, E. L. and Fergus, R. Stochastic video generation with a learned prior. In *ICML*, 2018.

Doshi, S. Various optimization algorithms for training neural network. 2019.

Furnari, A. and Farinella, G. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020. ISSN 1939-3539. doi: 10.1109/tpami.2020.2992889. URL http://dx.doi.org/10.1109/TPAMI.2020.2992889.

Gavrilova, Y. What is ml optimization. 2020. URL https://serokell.io/blog/ml-optimization. [Online:accessed 14-January-2021].

Geyer, C. J. Markov chain monte carlo maximum likelihood, 1991.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks, 2014.

Goroshin, R., Mathieu, M., and LeCun, Y. Learning to linearize under uncertainty, 2015.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Comput.*, pp. 1735–1780, November 1997.

Im, D. J., Kim, C. D., Jiang, H., and Memisevic, R. Generating images with recurrent adversarial networks, 2016.

Jayaraman, D., Ebert, F., Efros, A. A., and Levine, S. Time-agnostic prediction: Predicting predictable video frames, 2018.

Kaur, J. and Das, S. Future frame prediction of a video sequence, 2020.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes, 2014a.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes, 2014b.

Liu, W., Luo, W., Lian, D., and Gao, S. Future frame prediction for anomaly detection - a new baseline. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6536–6545, 2018.

Lotter, W., Kreiman, G., and Cox, D. Unsupervised learning of visual structure using predictive generative networks, 2016.

Lu, C., Hirsch, M., and Scholkopf, B. Flexible spatio-temporal networks for video prediction. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2137–2145, 2017.

Mathieu, M., Couprie, C., and LeCun, Y. Deep multi-scale video prediction beyond mean square error, 2016.

Michalski, V., Memisevic, R., and Konda, K. R. Modeling deep temporal dependencies with recurrent "grammar cells". In *NIPS*, 2014.

Mirza, M. and Osindero, S. Conditional generative adversarial nets, 2014.

Pang, G., Shen, C., Cao, L., and van den Hengel, A. Deep learning for anomaly detection: A review, 2020.

PredictingFFusingCycleGAN. Predicting future frames using retrospective cycle gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016a.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016b.

Sagar, A. Generate high resolution images with generative variational autoencoder, 2020.

Salakhutdinov, R., Mnih, A., and Hinton, G. Restricted boltzmann machines for collaborative filtering. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pp. 791–798, 2007.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., kin Wong, W., and chun Woo, W. Convolutional lstm network: A machine learning approach for precipitation nowcasting, 2015.

Srivastava, N., Mansimov, E., and Salakhutdinov, R. Unsupervised learning of video representations using lstms, 2016.

torchvision. Smnistdataset. URL https://pytorch.org/docs/stable/torchvision/datasets.html.

Turhan, C. G. and Bilge, H. S. Recent trends in deep generative models: a review. *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, pp. 574–579, 2018.

tutorialspoint workers. Types of learning. URL https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_types_of_learning.htm. [Online:accessed 14-January-2021].

Villegas, R., Yang, J., Zou, Y., Sohn, S., Lin, X., and Lee, H. Learning to generate long-term future via hierarchical prediction, 2018.

Vondrick, C., Pirsiavash, H., and Torralba, A. Generating videos with scene dynamics. *CoRR*, abs/1609.02612, 2016. URL http://arxiv.org/abs/1609.02612.

Walker, J., Marino, K., Gupta, A., and Hebert, M. The pose knows: Video forecasting by generating pose futures, 2017.

Wikipedia contributors. Autoencoder — Wikipedia, the free encyclopedia, 2021. URL https://en.wikipedia.org/w/index.php?title=Autoencoder&oldid=1000253269. [Online; accessed 14-January-2021].

Wu, Y., Gao, R., Park, J., and Chen, Q. Future video synthesis with object motion prediction, 2020.

Zhang, A. Data types from a machine learning perspective with examples. 2018.