International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-5, January 2020

Generating Video From Images using GAN and CVAE

Anoosh G P, Chetan G, Mohan Kumar M, Priyanka B N, Nagashree Nagaraj

Abstract: ***In a given scene, people can often easily predict a lot of quick future occasions that may occur. However generalized pixel-level expectation in Machine Learning systems is difficult in light of the fact that it struggles with the ambiguity inherent in predicting what's to come. However, the objective of the paper is to concentrate on predicting the dense direction of pixels in a scene — what will move in the scene, where it will travel, and how it will deform through the span of one second for which we propose a conditional variational autoencoder as a solution for this issue. We likewise propose another structure for assessing generative models through an adversarial procedure, wherein we simultaneously train two models, a generative model G that catches the information appropriation, and a discriminative model D that gauges the likelihood that an example originated from the training data instead of G. We focus on two uses of GANs semi-supervised learning, and the age of pictures that human's find visually realistic. We present the Moments in Time Dataset, an enormous scale human-clarified assortment of one million short recordings relating to dynamic situations unfolding within three seconds.***

Keywords: Generative Adversarial Network, Conditional Variational Autoencoders, Video Generation, Generative model, Discriminative model.

I. INTRODUCTION

In this project, we utilized Deep learning systems to create moving video from still pictures. Specifically, we thought about the task of taking two pictures, the initial picture, and the final picture and creating a moving video that reasonably inserted from the initial picture to the final picture. In past examinations, this task was known as video completion. We propose to return to foreseeing thick directions at every single pixel utilizing a feed-forward Convolutional Network. Utilizing thick directions confines the yield space significantly which enables our calculation to learn vigorous models for visual expectation with the accessible information.

Manuscript published on January 30, 2020.
* Correspondence Author
Anoosh G P*, Pursuing, Bachelor of Engineering in Computer Science, Vidyavardhaka College of Engineering, Mysuru.
Chetan G, Pursuing, Bachelor of Engineering in Computer Science, Vidyavardhaka College of Engineering, Mysuru.
Priyanka B N, Pursuing, Bachelor of Engineering in Computer Science, Vidyavardhaka College of Engineering, Mysuru.
Mohan Kumar M, Pursuing, Bachelor of Engineering in Computer Science, Vidyavardhaka College of Engineering, Mysuru.
Nagashree Nagaraj, Assistant Professor, Department of computer science and engineering, vidyavardhaka College of Engineering. Mysuru.

Be that as it may, the thick directions are still high-dimensional and the yield still has different modes. So as to handle these difficulties, we propose to utilize variational autoencoders to become familiar with a low dimensional inactive portrayal of the yield space adapted on an info picture.

Generative video creation stays a significant test, in any event, for present-day profound learning systems. In huge part, this is on the grounds that the space of conceivable yield recordings is uncommonly enormous, in any event when the recordings are short and low goals. For instance, we have to create 30 casing RGB shading recordings, where each casing comprised of 64×64 pixels. The space of such recordings has measurement 30×64×64×3 = 368640. To manage this high-dimensionality, while keeping the number of trainable parameters reasonable, we utilized three-dimensional convolutional neural systems. To create the recordings, we utilized the two predominant best in class systems for contingent profound generative learning these are restrictive variational autoencoders (CVAEs) and contingent generative antagonistic systems (CGANs). Modeling the spatial-worldly elements in any event, for three-second recordings, represents an overwhelming test. For example, recordings with the activity "opening" incorporate individuals opening entryways, doors, drawers, and blinds, creatures and people opening eyes, and even a flower opening its petals. At times a similar arrangement of casings in turn around can really portray an alternate activity ("closing") indicating that the transient perspective is vital to video understanding.

People can perceive a typical change that happens in reality that takes into consideration the entirety of the referenced situations to be allotted to the class "opening" despite the fact that outwardly they appear to be extremely unique from one another. The test is to create models that perceive these changes such that they will enable them to separate between various activities, yet sum up to different operators and settings inside a similar class. We present the Moments in Time dataset, one million recordings each with one activity mark from 339 distinct classes, to empower models to lavishly get activities and elements in recordings. This is one of the biggest human-commented on video datasets catching visual and discernible short occasions delivered by people, creatures, articles or nature. We propose a completely convolutional model to address the undertaking of the video in the middle. The proposed model comprises of three fundamental segments:

i) A 2D-convolutional picture encoder, which maps the information key casings to an inactive space.

ii) A 3D-convolutional inactive portrayal generator, which figures out how to fuse the data contained in the info outlines with continuously expanding fleeting goals.

iii) A video generator, which uses transposed 3D-convolutions to decipher the dormant portrayal into video outlines. Our key finding is that isolating the age of the inert portrayal from video interpreting is of pivotal significance to effectively address video in the middle.

II. RELATED WORKS

The task of generating a video from a single initial image was recently attempted in [2] and [3]. [2] Trained a generative adversarial network (GAN) on 2,000,000 unlabelled recordings so as to create short recordings of as long as a second at full casing rate dependent on a static picture. Two-stream engineering was utilized where the closer view and foundation were created independently. [3] Trained a contingent variational autoencoder (CVAE) on a huge number of recordings. Both made some progress, despite the fact that the created recordings were still very ridiculous contrasted with pictures produced by current profound learning systems.

The assignment of video culmination, as done in this venture, was recently tended to in [1] and [4]. The two papers utilized a design where the underlying and final pictures were encoded in an inactive space (that, for instance, portrayed the condition of a figure in the picture), a succession of activities was produced in the idle space and afterward, finally, the dormant grouping was converted into a yield video. The two works additionally utilized GANs as a component of their structures. In any case, [1] utilized an intermittent neural system (RNN), though [4] received a completely convolutional approach, as done in our task.

As opposed to past work, we generally worked straightforwardly at the degree of pixels, with no handmade highlights, for example, a division of closer view and foundation highlights or a carefully assembled idle space intended to catch the pertinent degrees of opportunity in the picture. This makes the assignment of video consummation all the more testing from multiple points of view. In any case, it has frequently turned out in the past that, as computational power has expanded, strategies that depended vigorously on high-quality highlights have wound up being outperformed by Simpler" start to finish profound learning techniques.

III. COMPARISION OF DIFFERENT METHODOLOGIES:

Sl No:	Authors	Objective	Methods Used	Datasets	Results
[1]	HaoyeCai, Chunyan Bai, Yu-Wing Tai, and Chi-Keung Tang.	This paper is based on Deep Video Generation, Prediction and Completion of Human Action Sequences.	Conditional GAN, Wasserstein GAN(WGAN)	Human3.6m dataset	It is able to generate large-scale human motion videos with a longer duration.
[2]	Carl Vondrick, Hamed Pirsiavash, Antonio Torralba	This paper aims to Generate Videos with Scene Dynamics	Gaussian Mixture Model (GMM), Generative Adversarial Networks,	Unfiltered Unlabelled Videos.	Understanding scene dynamics and generating video using two-stream architecture (foreground and background) combining both to generate videos with 91.4% accuracy.
[3]	Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert	This paper's main objective is to Forecast from Static Images using Variational Autoencoders	Conditional Variational Autoencoder(CVAE)	UCF101 dataset	The resultant of this network is to predict the motion based on the context of the scene.
[4]	Yunpeng Li, Dominik Roblek, Marco Tagliasacchi.	This paper is based on the fully convolutional model to generate video sequences directly in the level of the pixel domain.	The recently proposed Frechet video distance (FVD) model is used as the primary evaluation metrics	BAIR robot pushing Dataset, KTH Action Database and UCF101 Action Recognition Data Set	This method is used for video inbetweening using only direct 3D convolutions with 95% confidence intervals.
[5]	David Silver, Thomas Hubert, and Julian Schrittwieser	This paper aims to generate the reinforcement Learning Algorithm for self-playing the Games.	The method used here is Anatomy of a Computer Chess Program focusing Specifically on Stockfish and AlphaZero algorithm.	----	Self-playing games are generated by using the latest parameters for this neural network, omitting the evaluation step and the selection of the best player.
[6]	Mathew Monfort, Alex Andonian, Bolei Zhou and Kandan Ramakrishnan	This paper is based on generating Moments in Time Dataset	Resnet50 I3D model was used to evaluate which dataset is better among the two taken into consideration.	Kinetics, UCF101, HMDB51.	This results in generating the moment in time dataset that consists of 1M videos, divided into 339 categories with 57% SVM accuracy.
[7]	Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, and David Warde-Farley	This paper is based on Generative Adversarial Nets	Markov chain Monte Carlo (MCMC) methods.	MNIST, the Toronto Face Database (TFD), and CIFAR-10.	Demonstrated the viability of the adversarial modeling framework, suggesting that these research directions could prove useful.
[8]	Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung	This paper aims in improving Techniques for Training GANs.	Training GAN includes finding the Nash equilibrium between two players. (Generator G and Discriminator D)	CIFAR-10 Dataset	Proposed several techniques to stabilize the training that allows us to train models that were previously untrainable.

[9]	Alexander J. O'Donnell and Patrick J. M. Savoie	The objective of this paper is to Evaluate the Performance of General Adversarial Neural Networks.	Deep Convolutional Generative Adversarial Network (DCGAN), Fisher Linear Discriminant Analysis	CelebA and CIFAR-10 datasets.	The greater accuracy was found by classifying faces and non-faces was a mean of 97.85% across all partitions achieved through KNN with Euclidean distance on the Fisher feature space.
[10]	Nishma Cauvery C M, Namratha S R, Madhurish Katta, and Divine Poonacha A	The main objective of this paper is to know about the Generative Adversarial Networks and their Applications in Video Domain	Conditional Generative adversarial network(CGAN), Laplacian Pyramid of Generative Adversarial Networks (LAPGAN)	------	This paper results in producing a higher resolution of images compared with other methods.

IV. BASELINE MODELS

Our baseline model was a straightforward direct insertion between the underlying picture and the last picture. We likewise endeavored to utilization of-the-rack video insertion works as an extra baseline model; however, these capacities, for the most part, failed totally when just given two edges of a video.

Conditional Variational Autoencoder (CVAE):

A variational autoencoder (VAE) encodes a few data (for this situation a video) as a Gaussian appropriation, for example, a vector of means and standard deviations. A decoder at that point attempts to recuperate the first information from an arbitrary example from this conveyance. To create information, one straightforward data source arbitrary clamor, tested from a unit Gaussian, into the decoder. Once more, to create recordings from an underlying and final picture, one essentially inputs the pictures, together with an arbitrary commotion vector, into the decoder. The video decoder/generator took as data sources the arbitrary commotion, together with the underlying and final outline, encoded utilizing a convolutional neural system. These information sources were linked and a completely associated layer, trailed by a progression of three-dimensional deconvolutional layers was applied, so as to create a yield video.

The training objective function was given by,

$$\mathbf{minEX\sim p(X)\{DKL[N(\mu(X),\Sigma(X)||N(0,I)] + \lambda 1||X - G||2\ 2 + \lambda 2(||xi - Gi||2\ 2 + ||xf - Gf||2\ 2)\}, (1)}$$

where X is the training video (with initial and final frames xi and xf), μ and Σ are the mean and covariance matrix for the latent space distribution conditioned on input video, G is the generated video (with initial and final frames Gi and Gf), and λ1 and λ2 are hyperparameters.

Generative Adversarial Network (GAN):

A generative adversarial network (GAN) can be formulated as a minimax adversarial game between a generator (G) and a discriminator (D), where D's goal is to tell apart real videos from generated ones and G's goal is to fool D. This game has a unique Nash equilibrium, where the distribution of videos generated by G is the same as the training distribution. Just like for the CVAE, we made the generator G be conditioned on the initial and final image which were pre-processed by an image encoder and then used to generate an output video. IN this sense, our network formed a conditional GAN (CGAN). The objective function is therefore given by ,

$$\mathbf{min\ wG\ max\ wD\ EX\sim pX(X)[logD(X;wD)] + Ez\sim pz(z)[log(1-D(G(z;wG);wD))] + EX\sim pX(X)[\lambda(||xi - Gi(z;wG)||2\ 2 + ||xf - Gf(z;wG)||2\ 2)], (2)}$$

where wG, wD are weighted for G and D, X is the training video (with initial and final frames xi and xf), z is a latent "code" sampled from a Gaussian distribution with zero mean and unit variance, G is the generated video (with initial and final frames Gi and Gf), and λ is a hyperparameter.

V. RESULT

We train our CVAE model in 2000 "erupting" films. For the training objective, we choose hyperparameters λ1 = λ2 = 2.Zero (see Eqn. 1). We will do mini-batch gradient descent with 20 films consistent with batch, using Adam optimizer with β1 = 0.8 and gaining knowledge of price 0.0001. The overall loss characteristic converged fairly nicely in a hundred epochs. We also train the extra advanced CVAE version, on 1800 "skiing" films. The hyperparameters have been in any other case unchanged. Compared to the baseline version, as measured by L2 distance, the overall performance of this model turned into appreciably higher than the authentic CVAE. We additionally educated our CGAN model on one hundred twenty "erupting" motion pictures. We will use the "one-sided label smoothing" technique to prevent the discriminator from getting too right which might result in a generator failing to make any significant upgrades. For the objective function, we select hyperparameters λ = 0.02. Then we update the discriminator D four instances for every update on generator G, the use of the same gaining knowledge of price 0.0002 for each D and G. The performance of the models will be tested by looking at the generated movies on both the training and the checking out datasets. We are most effectively be able to train on a hundred and twenty films for a thousand epochs. One simple quantitative metric for performance is the L2 distance between generated and real videos. The mean and standard deviations of the L2 distance measured on a small check set. However, a qualitative examination of the generated videos makes it clear that the interpolation technique absolutely fails to seize any exciting dynamics and rather outperforms our essential methods, by means of this easy metric, truly with the aid of perfectly reproducing the initial and final body pix. We consequently do no longer suppose that the L2 distance metric should be treated as definitive.

VI. CONCLUSION AND FUTURE WORK

Generative adversarial networks are a promising class of generative models that have so far been held back by unstable training and by the lack of a proper evaluation metric. This work presents fractional answers for both of these issues. We propose a few procedures to balance out preparing that enable us to prepare models that were already untrainable. We present a general generative model that tends to the issue of video generation, prediction and complete consistently. While we are as yet far from completely outfitting the capability of unlabeled video, our analyses bolster that copious unlabeled video can be rewarding for both figuring out how to produce recordings and learning visual portrayals. We present the Moments in Time Dataset, an enormous scale accumulation of three-second recordings covering a wide scope of dynamic occasions including various operators (individuals, creatures, articles, and natural phenomena).

Using CVAEs, we had the option to create yield recordings that were roughly founded on the underlying and last pictures and which, in specific cases, seemed to have some reasonable elements. Be that as it may, in any event by the rough measurement of L2 separation, our systems were beaten by a basic benchmark of a straight addition between the underlying and last casing. It appears to be confident that the presentation could be extensively improved by further tuning of hyperparameters. Specifically, the CVAE form utilized on the skiing recordings was just run on one lot of hyperparameters for the "skiing" recordings and was not tried at all on the "erupting" recordings. For CGANs, with progressively computational power one would have the option to prepare the model on a lot bigger preparing dataset and for some more epochs.

Finally, we utilized only convolutional neural systems to produce recordings, which implied that the video length was constantly fixed. It might be fascinating to investigate consolidating a recurrent neural system (RNN) to be progressively adaptable about video length and be better at disclosing the worldly connections. Given the on-going quick improvement in this field, there are signs all the more staying to be explored.

REFERENCES

1. Haoye Cai, Chunyan Bai, Yu-Wing Tai, and Chi-Keung Tang. Deep video generation, prediction, and completion of human action sequences. In Proceedings of the European Conference on Computer Vision (ECCV), pages 366–382, 2018.
2. Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In Advances In Neural Information Processing Systems, pages 613– 621, 2016.
3. 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.
4. Yunpeng Li, Dominik Roblek, and Marco Tagliasacchi. From here to there: Video inbetweening using direct 3d convolutions. arXiv preprint arXiv:1905.10240, 2019.
5. David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. arXiv preprint arXiv:1712.01815, 2017.
6. Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Yan Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. IEEE transactions on pattern analysis and machine intelligence, 2019.
7. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde- Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in neural information processing systems, pages 2672–2680, 2014.
8. Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In Advances in neural information processing systems, pages 2234–2242, 2016.
9. Alexander J. O'Donnell, Patrick J. M. Savoie University of New Brunswick, Department of Electrical and Computer Engineering, Dec 4th, 2017.Objective Evaluation of the Performance of General Adversarial Neural Networks.
10. Nishma Cauvery C M, Namratha S R, Madhurish Katta, and Divine Poonacha A Department of Computer Science Vidya Vardhaka College of Engineering Mysuru.

AUTHORS PROFILE

Anoosh G P, is currently pursuing a Bachelor of Engineering in Computer Science at Vidyavardhaka College of Engineering, Mysuru. He is a member of the VVCE ACM Chapter and his research interests include Machine Learning, Data Science and Artificial Intelligence.

Chetan G, is currently pursuing a Bachelor of Engineering in Computer Science at Vidyavardhaka College of Engineering, Mysuru. He is a member of the VVCE ACM Chapter and his research interests include Machine Learning, Networks and IoT.

Priyanka B N, is currently pursuing a Bachelor of Engineering in Computer Science at Vidyavardhaka College of Engineering, Mysuru. Her research interests include Machine Learning, Networks and IoT.

Mohan Kumar M, is currently pursuing a Bachelor of Engineering in Computer Science at Vidyavardhaka College of Engineering, Mysuru. His research interests include Machine Learning, Networks and IoT.

Nagashree Nagaraj, is an assistant professor in the department of computer science and engineering, vidyavardhaka College of engineering. BE from MIT Mysore and Mtech from NIE Mysore. Areas of interest IoT ML AI