ABSTRACT

Video prediction is an important yet challenging problem; burdened with the tasks of generating future frames and learning environment dynamics. Recently, autoregressive latent video models have proved to be a powerful video prediction tool, by separating the video prediction into two sub-problems: pre-training an image generator model, followed by learning an autoregressive prediction model in the latent space of the image generator. However, successfully generating high-fidelity and high-resolution videos has yet to be seen. In this work, we investigate how to train an autoregressive latent video prediction model capable of predicting high-fidelity future frames with minimal modification to existing models, and produce high-resolution (256x256) videos. Specifically, we scale up prior models by employing a high-fidelity image generator (VQ-GAN) with a causal transformer model, and introduce additional techniques of top-k sampling and data augmentation to further improve video prediction quality. Despite the simplicity, the proposed method achieves competitive performance to state-of-the-art approaches on standard video prediction benchmarks with fewer parameters, and enables high-resolution video prediction on complex and large-scale datasets.

Index Terms— Video Prediction, Transformer

1. INTRODUCTION

Video prediction can enable agents to learn useful representations for predicting the future consequences of the decisions they make, which is crucial for solving the tasks that require long-term planning, including robotic manipulation [1] and autonomous driving [2]. Despite the recent advances in improving the quality of video prediction [3, 4, 5, 6, 7, 8], learning an accurate video prediction model remains notoriously difficult problem and requires a lot of computing resources, especially when the inputs are video sequences with high-resolution [9, 10, 11]. This is because the video prediction model should excel at both tasks of generating high-fidelity images and learning the dynamics of environments, though each task itself is already a very challenging problem.

Recently, autoregressive latent video prediction methods [12, 13] have been proposed to improve the efficiency of video prediction, by separating video prediction into two sub-problems: first pre-training an image generator (e.g., VQ-VAE; [14]), and then learning the autoregressive prediction model [7, 15] in the latent space of the pre-trained image generator. However, the prior works are limited in that they only consider relatively low-resolution videos (up to 128 × 128 pixels) for demonstrating the efficiency of the approach; it is questionable that such experiments can fully demonstrate the benefit of operating in the latent space of image generator instead of pixel-channel space.

In this paper, we present High-fidelity AutoRegressive latent video Prediction (HARP), which scales up the previous autoregressive latent video prediction methods for high-fidelity video prediction. The main principle for the design of HARP is simplicity: we improve the video prediction quality with minimal modification to existing methods. First, for image generation, we employ a high-fidelity image generator, i.e., vector-quantized generative adversarial network (VQ-GAN; [16]). This improves video prediction by enabling high-fidelity image generation (up to 256 × 256 pixels) on various video datasets. Then a causal transformer model [15], which operates on top of discrete latent codes, is trained to predict the discrete codes from VQ-GAN, and autoregressive predictions made by the transformer model are decoded into future frames at inference time.

We highlight the main contributions of this paper below:

- We show that our autoregressive latent video prediction model, HARP, can predict high-resolution (256 × 256 pixels) future frames on simulated robotics dataset (i.e., Meta-World; [17]) and large-scale real-world robotics dataset (i.e., RoboNet; [18]).
- We show that HARP can leverage the image generator pre-trained on ImageNet for training a high-resolution video prediction model on complex, large-scale Kinetics-600 dataset [19].
- HARP achieves competitive or superior performance to prior state-of-the-art video prediction models on widely-used BAIR Robot Pushing [20] and KITTI driving [21] video prediction benchmarks.
2. PRELIMINARIES

We aim to learn a video prediction model that predicts the future frames $x_c:T = (x_c, ..., x_{T-1})$ conditioned on the first $c$ frames of a video $x_c:T = (x_0, x_1, ..., x_{c-1})$, where $x_i \in \mathbb{R}^{H \times W \times N_c}$ is the frame at timestep $t$. Optionally, one can also consider conditioning the prediction model on actions $a = (a_0, ..., a_{T-1})$ that the agents would take.

2.1. Autoregressive video prediction model

Autoregressive video prediction model [7] approximates the distribution of a video in a pixel-channel space. Given a video $x \in \mathbb{R}^{T \times H \times W \times N_c}$, the joint distribution over pixels conditioned on the first $c$ frames is modelled as the product of channel intensities $N_c$ and all $N_p = T \cdot H \cdot W$ pixels except $N_c = c \cdot H \cdot W$ pixels of conditioning frames:

$$p(x_c:T \mid x_{<c}) = \prod_{i=0}^{N_c-1} \prod_{k=0}^{N_p-1} p(x_{pi(i)} \mid x_{<i}, x_{pi(i)}, x_{pi(i)})$$

where $\pi$ is a raster-scan ordering over all pixels from the video (we refer to Weissenborn et al [7] for more details), $x_{pi(i)}$ is all pixels before $x_{pi(i)}$, $x_{pi(i)}$ is $k$-th channel intensity of the pixel $x_{pi(i)}$, and $x_{pi(i)}$ is all channel intensities before $x_{pi(i)}$.

2.2. Vector quantized variational autoencoder

VQ-VAE [14] consists of an encoder that compresses images into discrete representations, and a decoder that reconstructs images from these discrete representations. Formally, given an image $x \in \mathbb{R}^{H \times W \times N_c}$, the encoder $E$ encodes $x$ into a feature map $z_{p}(x) \in \mathbb{R}^{H' \times W' \times N_c}$ consisting of a series of latent vectors $z_{pi(i)}(x) \in \mathbb{R}^{N_c}$, where $\pi'$ is a raster-scan ordering of the feature map $z_{p}(x)$ of size $|\pi'| = H' \cdot W'$. Then $z_{p}(x)$ is quantized to discrete representations $z_{q}(x) \in \mathbb{R}^{|\pi'| \times N_c}$ based on the distance of latent vectors $z_{pi(i)}(x)$ to the prototype vectors in a codebook $C = \{e_k\}_{k=1}^{K}$ as follows:

$$z_{q}(x) = \arg\min_{k \in |K|} \| z_{pi(i)}(x) - e_k \|_2$$

where $|K|$ is the set $\{1, \cdots, K\}$. Then the decoder $G$ learns to reconstruct $x$ from discrete representations $z_{q}(x)$. The VQ-VAE is trained by minimizing the following objective:

$$L_{\text{VQVAE}}(x) = \left( \| x - G(z_{q}(x)) \|^2 \right)_{L_{\text{recon}}} + \| s G \left[ z_{q}(x) \right] \|^2_{L_{\text{codebook}}} + \beta \cdot \left( \| sG \left[ z_{q}(x) \right] - z_{x}(x) \|^2 \right)_{L_{\text{commit}}}$$

where the operator $sG$ refers to a stop-gradient operator, $L_{\text{recon}}$ is a reconstruction loss for learning representations useful for reconstructing images, $L_{\text{codebook}}$ is a codebook loss to bring codebook representations closer to corresponding encoder outputs $h$, and $L_{\text{commit}}$ is a commitment loss weighted by $\beta$ to prevent encoder outputs from fluctuating frequently between different representations.

2.3. Vector quantized generative adversarial network

VQ-GAN [16] is a variant of VQ-VAE that (a) replaces the $L_{\text{recon}}$ in (3) by a perceptual loss $L_{\text{LPIPS}}$ [22], and (b) introduces an adversarial training scheme where a patch-level discriminator $D$ [23] is trained to discriminate real and generated images by maximizing following loss:

$$L_{\text{GAN}}(x) = \log(D(x)) + \log(1 - D(G(z_{q}(x))))$$

Then, the objective is given as below:

$$\min_{E,G,C} \max_{D} \mathbb{E}_{x \sim p(x)} \left[ (L_{\text{LPIPS}} + L_{\text{codebook}} + L_{\text{commit}}) + \lambda \cdot L_{\text{GAN}} \right]$$

where $\lambda = \nabla_{G_L} [L_{\text{LPIPS}}]/(\nabla_{G_L} [L_{\text{LPIPS}}] + \delta)$ is an adaptive weight, $\nabla_{G_L}$ is the gradient of the inputs to the last layer of the decoder $G_L$, and $\delta = 10^{-6}$ is a scalar introduced for numerical stability.

3. METHOD

We present HARP, a video prediction model capable of predicting high-fidelity future frames. Our method is designed to fully exploit the benefit of autoregressive latent video prediction model that separates the video prediction into image generation and dynamics learning. The full architecture of HARP is illustrated in Figure 1.

3.1. High-fidelity image generator

We utilize the VQ-GAN model [16] that has proven to be effective for high-resolution image generation as our image generator (see Section 2 for the formulation of VQ-GAN).
Specifically, we first pre-train the image generator then freeze the model throughout training to improve the efficiency of learning video prediction models. The notable difference to a prior work that utilize 3D convolutions to temporally downsample the video for efficiency [13] is that our image generator operates on single images; hence our image generator solely focus on improving the quality of generated images. Importantly, this enables us to utilize the VQ-GAN model pre-trained on a wide range of natural images, e.g., ImageNet, without training the image generator on the target datasets, which can significantly reduce the training cost of high-resolution video prediction model.

3.2. Autoregressive latent video prediction model

To leverage the VQ-GAN model for video prediction, we utilize the autoregressive latent video prediction architecture that operates on top of the discrete codes. Specifically, we extract the discrete codes \( z(x) = (z(x_1), ..., z(x_T)) \) using the pre-trained VQ-GAN, where \( z(x_t) = (q(x_t, 1), q(x_t, 2), ..., q(x_t, |\pi|)) \) is the discrete code extracted from the frame \( x_t \) as in (2).

Then, instead of modelling the distribution of video \( p(x) \) in the pixel-channel space as in (1), we learn the distribution of the video in the discrete latent representation space:

\[
p(z(x_{<t}|x_{<c})) = \prod_{i=0}^{N_d-1} p(z_{\pi'(:,i)}(x)|z_{\pi'(:,<i)}(x)),
\]

where \( N_d = (T - C) \cdot H' \cdot W' \) is the total number of codes from \( x_{<T} \). Due to its simplicity, we utilize the causal transformer architecture [13] where the output logits from input codes are trained to predict the next discrete codes.

3.3. Additional techniques

To improve the quality of latent autoregressive models whose outputs are sampled from the probability distribution over a large number of discrete codes, we utilize the top-\( k \) sampling [24] that randomly samples the output from the top-\( k \) probable discrete codes. Furthermore, we utilize a data augmentation that moves the input images by \( m \) pixels along the \( X \) or \( Y \) direction to prevent overfitting.

4. EXPERIMENTS

We design our experiments to investigate the following:

- Can HARP predict high-resolution future frames (up to \( 256 \times 256 \) pixels) on various video datasets with different characteristics?
- How does HARP compare to state-of-the-art methods with large end-to-end networks on standard video prediction benchmarks in terms of quantitative evaluation?

4.1. High-resolution video prediction

4.1.1. Setup

For all experiments, VQ-GAN downsamples each frame into \( 16 \times 16 \) latent codes, i.e., by a factor of 4 for frames of size \( 64 \times 64 \) frames, and 16 for frames of size \( 256 \times 256 \). For training a transformer model, the VQ-GAN model is frozen so that its parameters are not updated. As for hyperparameter, we use \( k = 10 \) for sampling at inference time, but no data augmentation for high-resolution video prediction experiments. We investigate how our model works on large-scale real-world RoboNet dataset [18] consisting of more than 15 million frames, and Kinetics-600 dataset consisting of more than 400,000 videos, which require a large amount of computing resources for training even on \( 64 \times 64 \) resolution [8, 10].

For RoboNet experiments, we first train a VQ-GAN model, and then train a 12-layer causal transformer model that predicts future 10 frames conditioned on first two frames and future ten actions. For Kinetics-600 dataset, to avoid the
prohibitively expensive training cost of high-resolution video prediction models on this dataset and fully exploit the benefit of employing a high-fidelity image generator, we utilize the ImageNet pre-trained VQ-GAN model. As we train the transformer model only for autoregressive prediction, this enables us to train a video prediction model in a very efficient manner.

4.1.2. Experimental results

First, we provide the predicted frames on the held-out test video of RoboNet dataset in Figure 2a, where the model predicts the high-resolution future frames where a robot arm is moving around various objects of different colors and shapes. Furthermore, Figure 2b shows that Kinetics-600 pre-trained model can also predict future frames on the test natural videos\(^1\), which demonstrates that leveraging the large image generator pre-trained on a wide range of natural images can be a promising recipe for efficient video prediction on high-resolution, large-scale video datasets.

4.2. Comparative evaluation on standard benchmarks

4.2.1. Setup

For quantitative evaluation, we first consider the BAIR robot pushing dataset [20] consisting of roughly 40k training and 256 test videos. Following the setup in prior work [13], we predict 15 future frames conditioned on one frame. We also evaluate our method on KITTI driving dataset [21], where the training and test datasets are split by following the setup in Villegas et al [9]. Specifically, the test dataset consists of 148 video clips constructed by extracting 30-frame clips and skipping every 5 frames, and the model is trained to predict future ten frames conditioned on five frames and evaluated to predict future 25 frames conditioned on five frames. For hyperparameters, We use \(k = 10\) for both datasets and data augmentation with \(m = 4\) is only applied to KITTI as there was no sign of overfitting on BAIR dataset.

\(^1\) Videos with CC-BY license: Figure 2b top and bottom

Table 1. Quantitative evaluation on (a) BAIR Robot Pushing [20] and (b) KITTI driving dataset [21]. We observe that HARP can achieve competitive performance to state-of-the-art methods with large end-to-end networks on these benchmarks. (c) We also investigate how the number of layers in the causal transformer affects the performance.

| Method                  | Params | FVD (\(\downarrow\)) | LPIPS (\(\downarrow\)) |
|-------------------------|--------|----------------------|------------------------|
| (a) BAIR Robot Pushing  |        |                      |                        |
| LVT [12]                | 50M    | 125.8                |                        |
| SAVP [6]                | 53M    | 116.4                |                        |
| DVD-GAN-FP [10]         | —\(^\dagger\) | 109.8                |                        |
| VideoGPT [13]           | 82M    | 103.3                |                        |
| TriVD-GAN-FP [11]       | —\(^\dagger\) | 103.3                |                        |
| Video Transformer [7]   | 373M   | 94.0                 |                        |
| FitVid [8]              | 302M   | 93.6                 |                        |
| HARP (ours)             | 89M    | 99.3                 |                        |

| Method                  | Params | FVD (\(\downarrow\)) | LPIPS (\(\downarrow\)) |
|-------------------------|--------|----------------------|------------------------|
| (b) KITTI               |        |                      |                        |
| SVG [9]                 | 298M   | 1217.3               | 0.327                  |
| GHVAE [25]              | 599M   | 552.9                | 0.286                  |
| FitVid [8]              | 302M   | 884.5                | 0.217                  |
| HARP (ours)             | 89M    | 482.9                | 0.191                  |

\(^\dagger\) Not available

For evaluation metrics, we use LPIPS [22] and FVD [26], computed using 100 future videos for each ground-truth test video, then reports the best score over 100 videos for LPIPS, and all videos for FVD, following [8, 9].

4.2.2. Experimental results

Table 1 shows the performances of our method and baselines on test sets of BAIR Robot Pushing and KITTI driving dataset. We observe that our model achieves competitive or superior performance to state-of-the-art methods with large end-to-end networks, e.g., HARP outperforms FitVid with 302M parameters on KITTI driving dataset. In the case of BAIR dataset, HARP achieves the similar performance of FitVid with 302M parameters, even though our method only requires 89M parameters. Finally, we also investigate how the number of layers in the causal transformer affects the performance in Table 1c, where we find that scaling up the transformer improves the performance.

5. CONCLUSION

In this work, we present HARP that employs a high-fidelity image generator for predicting high-resolution future frames, and achieves competitive performance to state-of-the-art video prediction methods with large end-to-end networks. We also demonstrate that HARP can leverage the image generator pre-trained on a wide range of natural images for video prediction, similar to the approach in the context of video synthesis [27]. We hope this work inspires more investigation into leveraging recently developed pre-trained image generators [14, 15, 16] for high-fidelity video prediction.

6. ACKNOWLEDGEMENTS

This work is supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub; No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)), Center for Human Compatible AI (CHAI), ONR N00014-21-1-2769, the Darpa RACER program, and the Hong Kong Centre for Logistics Robotics, BMW.
7. REFERENCES

[1] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al., “Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation,” in CoRL, 2018.

[2] Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell, “End-to-end learning of driving models from large-scale video datasets,” in CVPR, 2017.

[3] Chelsea Finn, Ian Goodfellow, and Sergey Levine, “Unsupervised learning for physical interaction through video prediction,” in NeurIPS, 2016.

[4] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine, “Stochastic variational video prediction,” in ICLR, 2018.

[5] Emily Denton and Rob Fergus, “Stochastic video generation with a learned prior,” in ICML, 2018.

[6] Alex X Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine, “Stochastic adversarial video prediction,” arXiv, 2018.

[7] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit, “Scaling autoregressive video models,” in ICLR, 2020.

[8] Mohammad Babaeizadeh, Mohammad Taghi Saffar, Suraj Nair, Sergey Levine, Chelsea Finn, and Dumitru Erhan, “Fitvid: Overfitting in pixel-level video prediction,” arXiv, 2021.

[9] Ruben Villegas, Arkanath Pathak, Harini Kannan, Dumitru Erhan, Quoc V Le, and Honglak Lee, “High fidelity video prediction with large stochastic recurrent neural networks,” NeurIPS, 2019.

[10] Aidan Clark, Jeff Donahue, and Karen Simonyan, “Adversarial video generation on complex datasets,” arXiv, 2019.

[11] Pauline Luc, Aidan Clark, Sander Dieleman, Diego de Las Casas, Yotam Doron, Albin Cassirer, and Karen Simonyan, “Transformation-based adversarial video prediction on large-scale data,” arXiv, 2020.

[12] Ruslan Rakhimov, Denis Volkonskiy, Alexey Artemov, Denis Zorin, and Evgeny Burnaev, “Latent video transformer,” in VISAPP, 2021.

[13] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas, “Videogpt: Video generation using vq-vae and transformers,” arXiv, 2021.

[14] Aaron van den Oord, Oriol Vinyals, and koray Kavukcuoglu, “Neural discrete representation learning,” in NeurIPS, 2017.

[15] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever, “Generative pretraining from pixels,” in ICML, 2020.

[16] Patrick Esser, Robin Rombach, and Bjorn Ommer, “Taming transformers for high-resolution image synthesis,” in CVPR, 2021.

[17] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine, “Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning,” in CoRL, 2020.

[18] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn, “Robonet: Large-scale multi-robot learning,” in CoRL, 2019.

[19] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman, “A short note about kinetics-600,” arXiv, 2018.

[20] Frederik Ebert, Chelsea Finn, Alex X Lee, and Sergey Levine, “Self-supervised visual planning with temporal skip connections,” in CoRL, 2017.

[21] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun, “Vision meets robotics: The kitti dataset,” IJRR, 2013.

[22] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in CVPR, 2018.

[23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, “Image-to-image translation with conditional adversarial networks,” in CVPR, 2017.

[24] Angela Fan, Mike Lewis, and Yann Dauphin, “Hierarchical neural story generation,” in ACL, 2018.

[25] Bohan Wu, Suraj Nair, Roberto Martin-Martín, Li Fei-Fei, and Chelsea Finn, “Greedy hierarchical variational autoencoders for large-scale video prediction,” in CVPR, 2021.

[26] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly, “Towards accurate generative models of video: A new metric & challenges,” arXiv, 2018.

[27] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N Metaxas, and Sergey Tulyakov, “A good image generator is what you need for high-resolution video synthesis,” in ICLR, 2021.