Generating Video from Images

Geoffrey Penington, Mae Teo, Chao Wang

geoffp, maehwee, cwang15@stanford.edu

Introduction

- **The problem:** Generate realistic videos from a given initial and final image.
- **The challenge:** Dimension of space of possible videos is frames × height × width × colors. Very large even for short, low-resolution videos.
- **Approaches:** Conditional variational autoencoders (CVAEs) and conditional generative adversarial networks (CGANs), using deep convolutional neural networks (CNNs).

Dataset

- 'Moments in Time' [1]: 1,000,000 short videos, divided into 339 categories.
- Selected two categories: "Erupting" and "Skiing".
- Downsized to 30 frames, 64 × 64 pixels each frame

CVAE

- Training Objective:

$$\min \mathbb{E}_{x\sim p(x)}\{D_{\mathrm{KL}}[\mathcal{N}(\mu(x),\Sigma(x)||\mathcal{N}(0,I)]+ \lambda_1||x-g_\theta(z)||_2^2+\lambda_2(||c(g_\theta(z))-c(x)||_2^2\} \quad (1)$$

-

$$z_{\text{total}} = z_{\text{video}} + z_{\text{first frame}} + z_{\text{last frame}} = 600 + 1000 + 1000$$

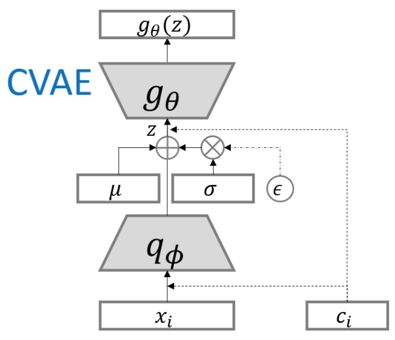

Figure: Training Video: x_i; Initial/Final frames: c_i; Generated Video: $g_\theta(z)$. Figure taken from [2].

Baseline Model

- Linear interpolation between initial and frame image.

CGAN

- Training Objective:

$$\min_{w_G}\max_{w_D}\mathbb{E}_{x\sim p_x(x)}[\log D(x;w_D)]+ \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z;w_G);w_D))]+ \mathbb{E}_{x\sim p_x(x)}[\lambda||c(G(z;w_G))-c(x)||_2^2] \quad (2)$$

Results (See our live demos!)

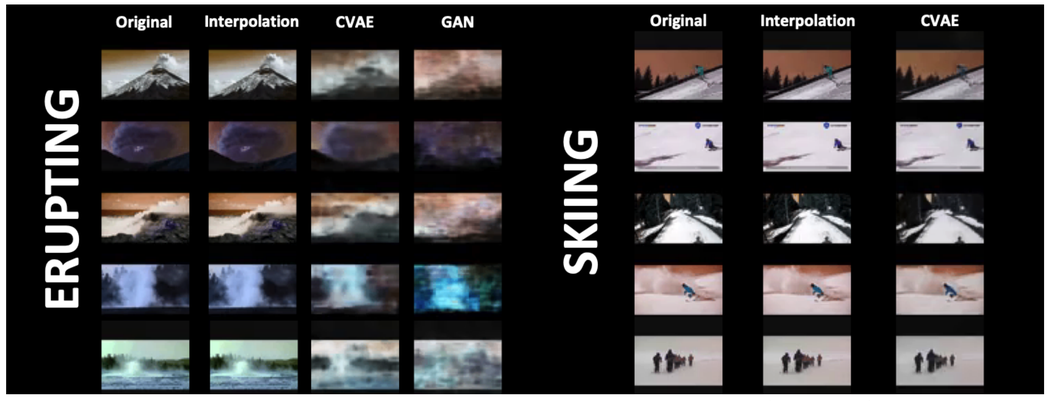

Figure: Examples of videos generated given an initial frame and a final frame. We used interpolation (our baseline), a trained CVAE, and a trained GAN.

CGAN (cont.)

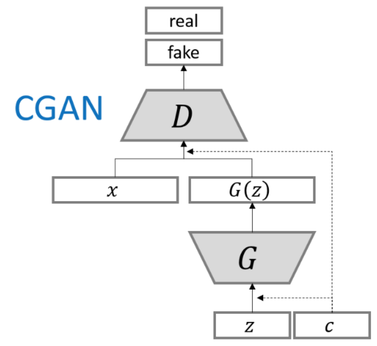

Figure: Training Video: x_i; Initial/Final frames: c_i. Figure taken from [2].

Results and Evaluations

- Used Adam optimizer. Used "one-sided label smoothing" for CGAN.
- For "erupting": CVAE trained on 2000 videos, CGAN trained on 120 videos.
- For "skiing": CVAE trained on 1800 videos.
- Qualitative metric: Look at Videos!
- Quantitative metric: L_2 distance between generated and real videos (test set of 100 videos)

L_2 distance	Interpolation	CVAE	CGAN
Erupting	62 ±58	180 ±110	223 ±96
Skiing	153 ±89	197 ±80	N/A

Future

- More time and GPU resources to train on larger datasets, and more carefully tune the hyperparameters.
- Use deeper and more complex architectures, e.g. recurrent neural networks (RNNs).

References

- [1] 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.
- [2] Tensorflow generative model collections. `https://github.com/hwalsuklee/tensorflow-generative-model-collections/`.