MoCoGAN: Decomposing Motion and Content for Video Generation

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, Jan Kautz
stulyakov@snap.com {mingyul,xiaodongy,jkautz}@nvidia.com

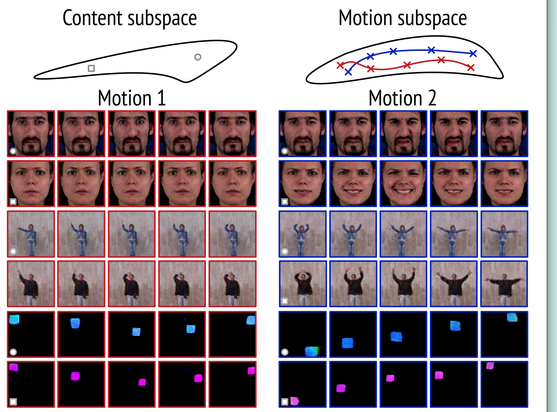
Snap Inc.
NVIDIA

Contribution

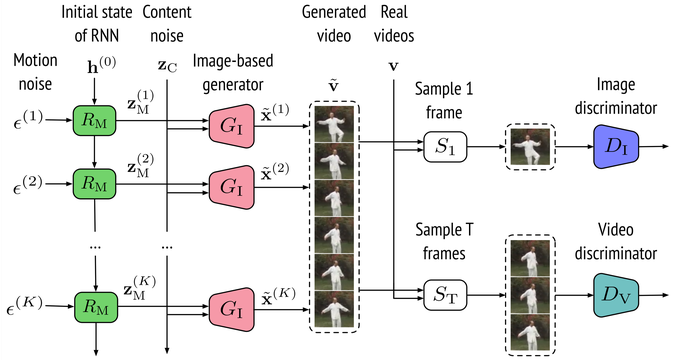
- A framework for unconditional video generation from noise vectors to videos
- A training procedure that decomposes motion and content, providing means to control video generation
- An image-based generator capable of producing random length sequences

Motivation

A point in the content subspace and a trajectory in the motion subspace define a video



Framework



A sequence of vectors for generating a video clip is represented by:

$$[\mathbf{z}^{(1)}, ..., \mathbf{z}^{(K)}] = \left[\begin{bmatrix} \mathbf{z}_C \\ \mathbf{z}_M^{(1)} \end{bmatrix}, ..., \begin{bmatrix} \mathbf{z}_C \\ \mathbf{z}_M^{(K)} \end{bmatrix} \right],$$

with $\mathbf{z}_M^{(k)} = R_M(k)$.

Then the image-based generator produces a video as:

$$\tilde{\mathbf{v}} = [G_I(\mathbf{z}^{(1)}), ..., G_I(\mathbf{z}^{(K)})].$$

Training

The MoCoGAN framework consists of four networks. The learning problem writes as:

$$\max_{G_I, R_M} \min_{D_I, D_V} \mathcal{F}_V(D_I, D_V, G_I, R_M),$$

where the objective function is:

$$\mathbb{E}_{\mathbf{v}}[-\log D_I(S_1(\mathbf{v}))] + \mathbb{E}_{\tilde{\mathbf{v}}}[-\log(1 - D_I(S_1(\tilde{\mathbf{v}})))] +$$
$$\mathbb{E}_{\mathbf{v}}[-\log D_V(S_T(\mathbf{v}))] + \mathbb{E}_{\tilde{\mathbf{v}}}[-\log(1 - D_V(S_T(\tilde{\mathbf{v}})))],$$

where the first and the second terms encourage the image and video discriminators to distinguish real versus generated images and videos. The third and fourth terms force the image generator and the recurrent neural network to produce realistically looking image sequences.

Categorical dynamics

When categorical labels are available the input to the recurrent neural network is given by:

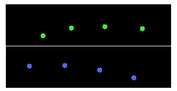
$$\left[\begin{bmatrix} \mathbf{z}_A \\ \epsilon^{(1)} \end{bmatrix}, ..., \begin{bmatrix} \mathbf{z}_A \\ \epsilon^{(K)} \end{bmatrix} \right].$$

The framework is trained by adding the mutual information term:

$$\mathcal{F}_V(D_I, D_V, G_I, R_M) + \lambda L_I(G_I, Q).$$

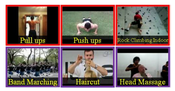
Datasets

Generated shapes motion
- Moving circles and squares
- 4,000 videos



Facial expressions - MUG
- 86 subjects, 6 expressions
- 1,254 videos

Tai-Chi
- Source: YouTube
- 4,500 videos



UCF101
- 101 actions
- 13,220 videos

Quantitative evaluation

Video generation results

Average Content Distance (ACD)

	Shape Motion	Facial Expressions
VGAN	5.02	0.322
TGAN	2.08	0.305
MoCoGAN	**1.79**	**0.201**

Inception Score

	UCF101
VGAN	8.18 ± .05
TGAN	11.85 ± .07
MoCoGAN	**12.42 ± .03**

User Preference Score

	Facial Expressions	Tai-Chi
MoCoGAN / VGAN	**84.2** / 15.8	**75.4** / 24.6
MoCoGAN / TGAN	**54.7** / 45.3	**68.0** / 32.0

Image to video translation results

User Preference Score, %

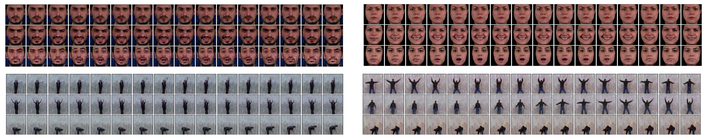
	Tai-Chi
MoCoGAN / VGAN	**66.9** / 33.1
MoCoGAN / MCNET	**65.6** / 34.4

Qualitative evaluation

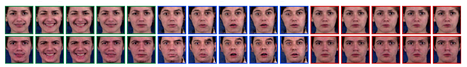
Video generation results

Motion and content manipulation

Each row has the same identity but different action

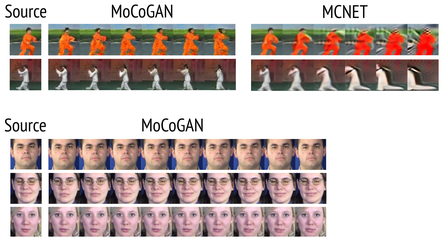


Changing motion trajectory while generating



Changing content for Tai-Chi

Image to video translation



References:

[MUG] Aifanti et al. The MUG facial expression database. In Image Analysis for Multime-dia Interactive Services, 2010

[UCF101] Soomro et al. UCF101: A datasetof 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012

[ACTIONS] Gorelick et al. Actions as space-time shapes. In TPAMI, 2007

[VGAN] Vondrick et al. Generating videos with scene dynamics, In NIPS, 2016

[TGAN] Saito et al. Temporal generativeadversarial nets with singular value clipping. In ICCV, 2017

[MCNET] Villegas et al. Decomposing motion and content for natural video sequence prediction. In ICLR, 2017