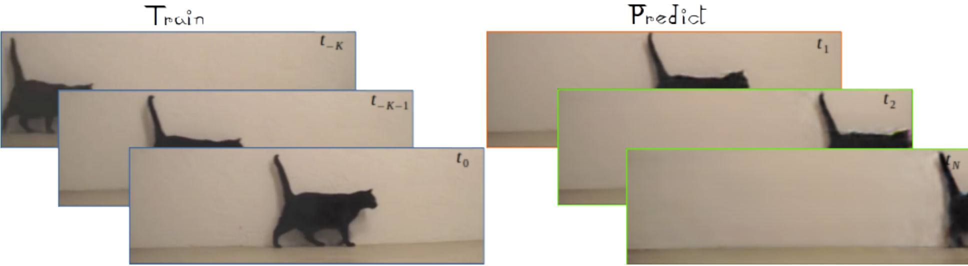
Domain-agnostic Video Prediction from Motion Selective Kernels

Véronique Prinet (author)
Da Li (presenter)

International Conference on Image Processing (ICIP)
September 25^{th}, 2019

Future frames prediction from a single clip



Past frames

Conditioning frames

Future frames

Table of contents

Introduction & Related Work

Method

Results

Motivation: Video Prediction

- *Our brains are predictive machines* [LeCun, Nov.2018]
- *Self-supervised learning* (the data themselves furnish the ground-truth) $\rightarrow$ virtually infinite amound of data available on the net
- To learn *visual representations* that can potentially be used for high-level vision tasks (i.e., recognition)
- Applications 1 : Robotics and automation, planning
  Applications 2 : **Video editing and manipulation**

Challenges

- For a given past (observation), there are multiple plausible futures
- Very diverse motion domains
- Hallucination of complex disclosed background

Related work (1)

- Early parametric and non-parametric models for picture animation, e.g.,

[Chuang & al. , 2005]

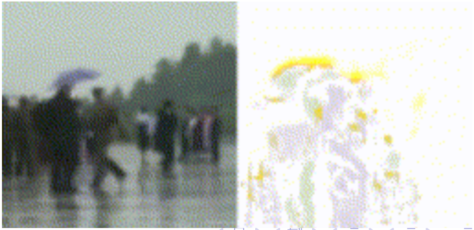
[Schodl & al. , 2000]

Related work (2)

- Conditional video prediction from large scale training, e.g.,

[Villegeas & al. 2017]

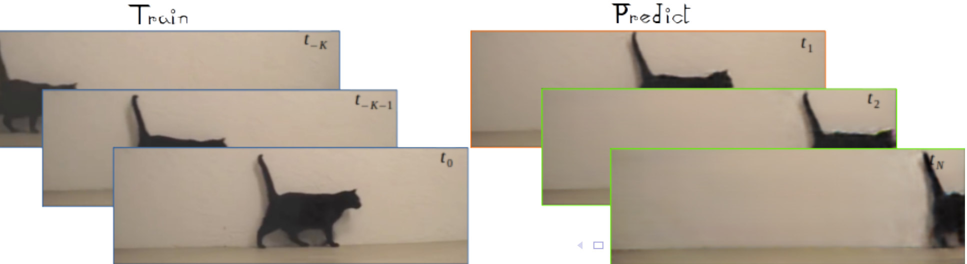
[Vondrick and Torralba, 2017]



Focus of this work

- Domain-agnostic & data-specific predictive model
  - Repetitive dynamic scene ‘in-the-wild’
  - Learn from small sample set (20-50 frames)
  - Mid-range prediction (20-30 frames)
  - Model interpretability

- Application
  Extend/extrapolate the content of a single video clip

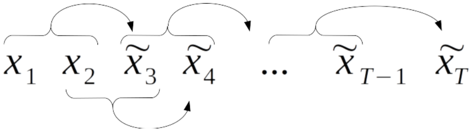


Method - Problem statement

- **Model**

$$\mathcal{L}(\zeta) = P_\zeta(\mathbf{x}_{\delta:T} | \mathbf{x}_{<\delta}) = \prod_{t'=\delta}^{T-1} P_\zeta(\mathbf{x}_{t'+1} | \tilde{\mathbf{x}}_{t'-\delta:t'})$$

$\mathbf{x}_{\delta:T} = \{\mathbf{x}_\delta, \dots . \mathbf{x}_T\}$: unknowns (future) time series
$\mathbf{x}_{0:\delta} = \mathbf{x}_{<\delta}$: observed frames ('context')
$\tilde{\mathbf{x}}$: generated frames



Method - Problem statement

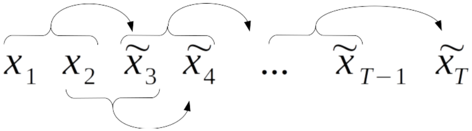
- **Model**

$$\mathcal{L}(\zeta) = P_\zeta(\mathbf{x}_{\delta:T}|\mathbf{x}_{<\delta}) = \prod_{t'=\delta}^{T-1} P_\zeta(\mathbf{x}_{t'+1}|\tilde{\mathbf{x}}_{t'-\delta:t'})$$

$\mathbf{x}_{\delta:T} = \{\mathbf{x}_\delta, ...\mathbf{x}_T\}$: unknowns (future) time series
$\mathbf{x}_{0:\delta} = \mathbf{x}_{<\delta}$: observed frames ('context')
$\tilde{\mathbf{x}}$: generated frames

- **Learning**

$$\zeta = -\arg\min_\zeta \log \mathcal{L}(\zeta) \approx \arg\min_\zeta E(\zeta)$$



Method: Motion representation

$$\begin{aligned}\mathcal{T}_\zeta : \mathbf{x}_{t-\delta:t} \mapsto \tilde{\mathbf{x}}_{t+1} &= G_{\Theta, S_\Psi(t)}(\mathbf{x}_{t-\delta:t}) \\ &= G_\Theta(\mathbf{x}_{t-\delta:t}; S_\Psi(\mathbf{x}_{t-\delta:t})).\end{aligned}$$

1. $G_\Theta(\mathbf{x}_{t-\delta:t})$: transformation model
2. $S_\Psi(\mathbf{x}_{t-\delta:t})$: selector model

Method: Motion representation

Encoder at layer l:

$$\mathcal{Z}^l_{n'} = \sum_{n=0}^{N-1} \mathcal{Y}^{l-1}_n * W^l_{n,n'} \qquad \mathcal{Y}^l_{n'} = \rho_l(\mathcal{Z}^l_{n'})$$

Method: Motion representation

Encoder at layer l:

$$\mathcal{Z}^{l}_{n'} = \sum_{n=0}^{N-1} \mathcal{Y}^{l-1}_{n} * W^{l}_{n,n'} \qquad \mathcal{Y}^{l}_{n'} = \rho_l(\mathcal{Z}^{l}_{n'})$$

(Classical) decoder with skip-connection at layer l:

$$\mathcal{Z}^{l}_{n'} = \sum_{n=0}^{2N-1} [\mathcal{Y}^{L-l};\ \mathcal{Y}^{l-1}]_n * W^{l}_{n,n'} \qquad \mathcal{Y}^{l}_{n'} = \rho_l(\mathcal{Z}^{l}_{n'})$$

Method: Motion representation

Encoder at layer l:

$$\mathcal{Z}^l_{n'} = \sum_{n=0}^{N-1} \mathcal{Y}^{l-1}_n * W^l_{n,n'} \qquad \mathcal{Y}^l_{n'} = \rho_l(\mathcal{Z}^l_{n'})$$

Decoder with *input-dependent activations* at layer l :

$$\mathcal{Z}^l_{n'} = \sum_{n=0}^{2N-1} [\mathcal{Y}^{L-l};\ \alpha^{l-1}(\tau)\,\mathcal{Y}^{l-1}]_n * W^l_{n,n'} \ , \ \mathcal{Y}^l_{n'} = \rho_l(\mathcal{Z}^l_{n'})$$

$$\{\alpha^l_n(\tau)\}^l_n \leftarrow S_\Psi(\mathbf{x}_{t-\delta:t})$$

Method: Motion representation

Encoder at layer l:

$$\mathcal{Z}^l_{n'} = \sum_{n=0}^{N-1} \mathcal{Y}^{l-1}_n * W^l_{n,n'} \qquad \mathcal{Y}^l_{n'} = \rho_l(\mathcal{Z}^l_{n'})$$

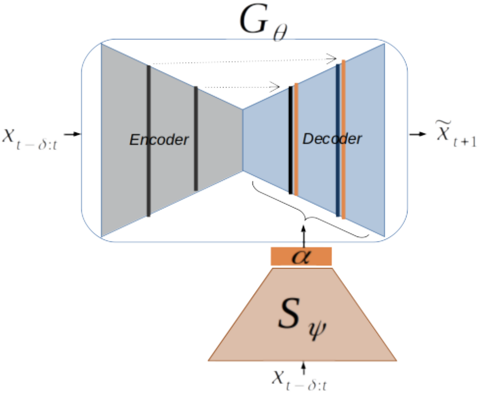
Decoder with *input-dependent activations* at layer l :

$$\mathcal{Z}^l_{n'} = \sum_{n=0}^{N-1} \mathcal{Y}^{L-l} * W^l_{n,n'} + \sum_{n=N}^{2N-1} \mathcal{Y}^{l-1}_n * (\alpha^{l-1}_n(\tau)\, W^l_{n,n'})$$

$$\{\alpha^{l-1}_n(\tau\}) \leftarrow S_\Psi(\mathbf{x}_{t-\delta:t})$$

Method: Motion representation

Encoder-Decoder with skip-connections and selector:



Method: Recap

$$
\begin{aligned}
\mathcal{Z}^0_{n'} &= \sum_{t'=t-\delta}^{t} \mathbf{x}_{t'} * W^0_{t',n'} \qquad \mathcal{Y}^0_{n'} = \rho_0(\mathcal{Z}^0_{n'}) \\
\mathcal{Z}^L_{n'} &= \sum_{n=0}^{2N-1} [\mathcal{Y}^0;\ \alpha^L(\tau)\ \mathcal{Y}^{L-1}]_n * W^L_{n,n'} \quad \tilde{\mathbf{x}}_{t+1} = \rho_L(\mathcal{Z}^L) \\
\mathcal{Z}^l_{n'} &= \sum_{n=0}^{2N-1} [\mathcal{Y}^{L-l};\ \alpha^{l-1}(\tau)\ \mathcal{Y}^{l-1}]_n * W^l_{n,n'} \\
&= (\mathcal{Z}^l_{n'})^b + (\mathcal{Z}^l_{n'})^f
\end{aligned}
$$

Learning

Reconstruction + motion loss

$$\ell_{L_1}(t) = ||\tilde{\mathbf{x}}_{t+1} - \mathbf{x}_{t+1}||_1$$
$$\ell_{motion}(t) = || \, |\nabla_t \, \tilde{\mathbf{x}}_{t+1}| - |\nabla_t \, \mathbf{x}_{t+1}| \, ||_1$$

Total loss

$$\begin{aligned} E(\zeta) &= \sum_{t=t'}^{t'+K} \left(\ell_{L_1}(t) + \mu_{motion} \mathbb{1}_{t>t'} \ell_{motion}(t)\right) \\ \zeta^* &= \arg\min_{\zeta} E(\zeta) \end{aligned}$$

Introduction & Related Work

Method

Results

Baselines

B1 *Baseline-1*. Encoder-encoder. The sole transformation model $G_\theta()$ is trained, the selection model is inactive; we set $\mu_{motion} = 0$.

M1 *DN w/o motion loss*. Our dual net model —$G_\theta()$ and $S()_\Phi$ are trained jointly; we set $\mu_{motion} = 0$.

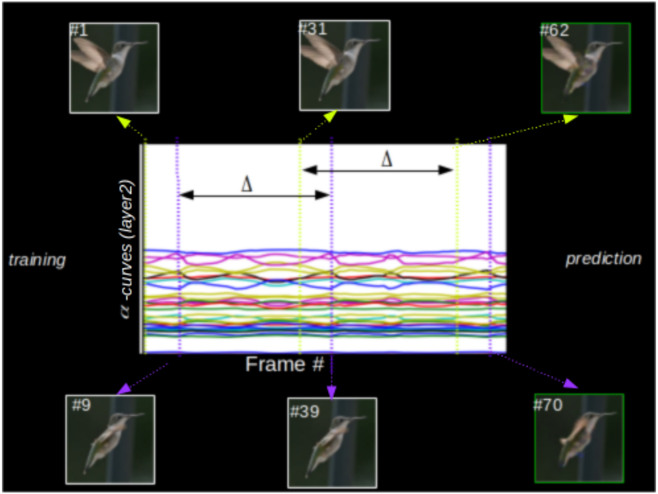
M2 *FDN*. Our dual net model, trained with motion loss. We set $\mu_{motion} = 10$, unless specified otherwise.

Bird sequence

- *Protocole: trained on 50 frames (1.5 period), prediction on 25 frames. Context of four frames. Frame size:* 256×256.

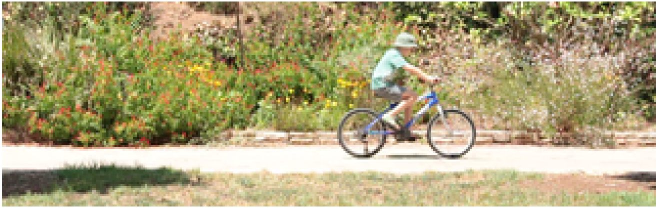
Results and comparison

Bird sequence



Garden sequence

- *Protocole: trained on 30 frames, prediction on 23 frames. Context of three frames. Frame size:* 100×300.



Results and comparison

Garden sequence - Foreground/background separation

Results and comparison

Cat sequence

- *Protocole: trained on 32 frames, prediction on 30 frames. Context of three frames. Frame size:* 105×320.

Results and comparison

Ocean sequence

- *Protocole: trained on 20 frames, prediction on 26 frames. Context of three frames. Frame size:* 200×200.

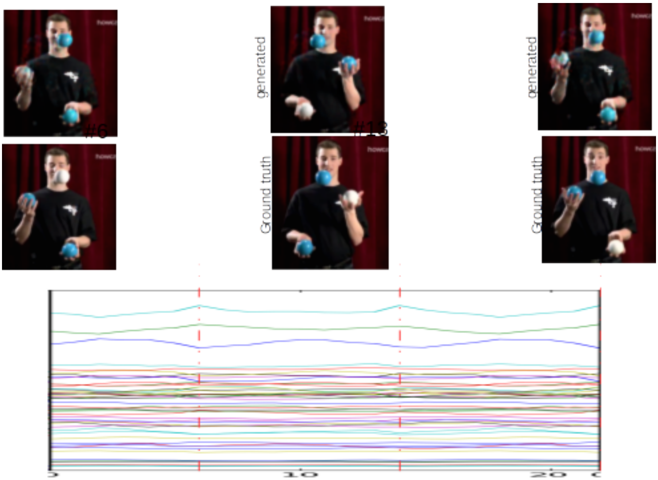
Results and comparison

Juggler sequence

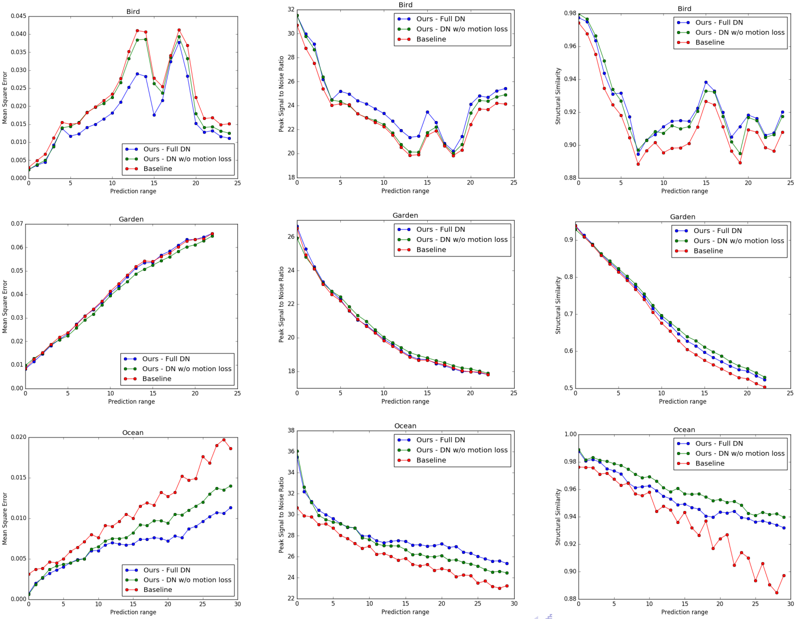
- *Protocole: trained on 50 frames, prediction on 23 frames. Context of three frames. Frame size:* 340×300.

Results and comparison

Juggler sequence



Quantitative results



Thank you for your attention.