Domain-agnostic Video Prediction from Motion Selective Kernels

Véronique Prinet (author)
Da Li (presenter)

International Conference on Image Processing (ICIP)
September 25th, 2019
Future frames prediction from a single clip
Table of contents

Introduction & Related Work

Method

Results
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) → virtually infinite amount 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
Introduction & Related Work

Method

Results
Method - Problem statement

▶ Model

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

\( x_{\delta:T} = \{ x_\delta, ..., x_T \} \): unknowns (future) time series

\( x_{0:\delta} = x_{<\delta} \): observed frames (‘context’)

\( \tilde{x} \): generated frames
Method - Problem statement

▶ Model

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

\( x_{\delta:T} = \{x_\delta, ....x_T\} \): unknowns (future) time series
\( x_{0:\delta} = x_{<\delta} \): observed frames (‘context’)
\( \tilde{x} \): generated frames

▶ Learning

\[ \zeta = - \arg \min_{\zeta} \log \mathcal{L}(\zeta) \approx \arg \min_{\zeta} E(\zeta) \]
Method: Motion representation

\[ T_\zeta : x_{t-\delta:t} \mapsto \tilde{x}_{t+1} = G_{\Theta, S_\psi(t)}(x_{t-\delta:t}) \]
\[ = G_{\Theta}(x_{t-\delta:t}; S_\psi(x_{t-\delta:t})) \]

1. \( G_{\Theta}(x_{t-\delta:t}) \): transformation model
2. \( S_\psi(x_{t-\delta:t}) \): selector model
Method: Motion representation

Encoder at layer $l$:

$$Z_{n'}^l = \sum_{n=0}^{N-1} \gamma_{n'}^{l-1} \ast W_{n,n'}^l \quad \gamma_{n'}^l = \rho_i(Z_{n'}^l)$$
Method: Motion representation

Encoder at layer $l$:

$$\mathcal{Z}_{n'}^l = \sum_{n=0}^{N-1} \mathcal{Y}_{n}^{l-1} \ast W_{n,n'}^{l}, \quad \mathcal{Y}_{n'}^{l} = \rho_{l}(\mathcal{Z}_{n'}^{l})$$

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

$$\mathcal{Z}_{n'}^l = \sum_{n=0}^{2N-1} [\mathcal{Y}_{n}^{l-1}; \mathcal{Y}_{n}^{l-1}] \ast W_{n,n'}^{l}, \quad \mathcal{Y}_{n'}^{l} = \rho_{l}(\mathcal{Z}_{n'}^{l})$$
Method: Motion representation

Encoder at layer $l$:

$$Z_{n'}^l = \sum_{n=0}^{N-1} Y_{n}^{l-1} \ast W_{n,n'}^l , \quad Y_{n'}^l = \rho_i(Z_{n'}^l)$$

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

$$Z_{n'}^l = \sum_{n=0}^{2N-1} [Y_{n}^{l-1}; \alpha_{n}^{l-1}(\tau) Y_{n}^{l-1}] \ast W_{n,n'}^l , \quad Y_{n'}^l = \rho_i(Z_{n'}^l)$$

$$\{\alpha_{n}^l(\tau)\}_{n}^l \leftarrow S_{\psi}(x_{t-\delta:t})$$
Method: Motion representation

Encoder at layer \( l \):

\[
\mathcal{Z}_{n'}^l = \sum_{n=0}^{N-1} \mathcal{Y}_{n}^{l-1} \ast W_{n,n'}^{l}, \quad \mathcal{Y}_{n'}^{l} = \rho_l(\mathcal{Z}_{n'}^{l})
\]

Decoder with \textit{input-dependent activations} at layer \( l \):

\[
\mathcal{Z}_{n'}^l = \sum_{n=0}^{N-1} \mathcal{Y}_{n}^{L-l} \ast W_{n,n'}^{l} + \sum_{n=N}^{2N-1} \mathcal{Y}_{n}^{l-1} \ast (\alpha_{n}^{l-1}(\tau) W_{n,n'}^{l})
\]

\[
\{\alpha_{n}^{l-1}(\tau)\} \leftarrow S_{\Psi}(x_{t-\delta:t})
\]
Encoder-Decoder with skip-connections and selector:
Method: Recap

\[
\begin{align*}
\mathcal{Z}_{n'}^0 &= \sum_{t'=t}^t x_{t'} * W_{t',n'}^0 \\
\mathcal{Y}_{n'}^0 &= \rho_0(\mathcal{Z}_{n'}^0) \\
\mathcal{Z}_{n'}^L &= \sum_{n=0}^{2N-1} [\mathcal{Y}^0; \alpha^L(\tau) \mathcal{Y}^{L-1}]_n * W_{n,n'}^L \\
\mathcal{Y}_{n'}^L &= \rho_L(\mathcal{Z}_{n'}^L) \\
\mathcal{Z}_{n'}^I &= \sum_{n=0}^{2N-1} [\mathcal{Y}^{L-I}; \alpha^{l-1}(\tau) \mathcal{Y}^{I-1}]_n * W_{n,n'}^I \\
&= (\mathcal{Z}_{n'}^I)^b + (\mathcal{Z}_{n'}^I)^f
\end{align*}
\]
Learning

Reconstruction + motion loss

\[ \ell_{L_1}(t) = ||\tilde{x}_{t+1} - x_{t+1}||_1 \]
\[ \ell_{motion}(t) = |||\nabla_t \tilde{x}_{t+1} - \nabla_t x_{t+1}||_1 \]

Total loss

\[ E(\zeta) = \sum_{t=t'}^{t'+K} (\ell_{L_1}(t) + \mu_{motion} \mathbb{1}_{t>t'} \ell_{motion}(t)) \]
\[ \zeta^* = \arg \min_{\zeta} E(\zeta) \]
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 $\mu_{\text{motion}} = 0$.

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

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

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

Results and comparison
Bird sequence
Garden sequence

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

Results and comparison
Garden sequence - Foreground/background separation

Results and comparison
Cat sequence


Results and comparison
Ocean sequence


Results and comparison
Juggler sequence

- Protocole: trained on 50 frames, prediction on 23 frames. Context of three frames. Frame size: $340 \times 300$.

Results and comparison
Juggler sequence
Quantitative results
Thank you for your attention.