UNSUPERVISED SPATIOTEMPORAL DATA INPAINTING

Yuan Yin, Arthur Pajot, Emmanuel de Bézenac, Patrick Gallinari

Implemented by : Ilyas Aroui, Abdelraouf Keskes, Aissam Djahnine

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

Introduction

Inpainting spatio-temporal sequences is an active research topic that relies heavily on supervision with large datasets. In this work, we consider the problem of reconstructing missing information with an unsupervised learning approach. Following the work of *Kim et al.* [4] we train a generative model on the occluded sequences. We ensure that the models captured both frame-based and sequential-based information. Our proposed model is adapted to large-scale images and can be used to different types of sequences and occlusion processes. Code available at: *https://github.com/raoufkeskes/Unsupervised-Spatiotemporal-Data-Inpainting*

Problem Setting

let $x, y \in R^{C\times T\times H\times W}$ be the true unknown and the occluded sequences respectively. We want to find x^* such that:

$$x^* = arg\,max_x\, log\, p_{X|Y}(x|y) = arg\,max_x\, log\, p_X(x) \,+\, log\, p_{Y|X}(y|x) \qquad (1)$$

Where $p_X(x)$ is the unknown prior and $p_{Y|X}(y|x)$ is the likelihood.

- **Prior handling:**

  Let G be the following mapping $G : Y \rightarrow X$. To maximize the likelihood, we want $G(y) = \hat{x}$ to be close to p_X. We will handle this by generative adversial network approach (GAN) for both frames and sequences.
  let $\hat{y} = F(\hat{x}, \hat{m})$ where $\hat{m} \sim P_M$ is a random occlusion process generated from p_M and F is a masking operator. We want to find G^* such that:

$$G^* = \underset{G}{\operatorname{argmin}}\, L(G) = \max_{Df,Ds} \mathbb{E}_{y\sim p_Y, \hat{y}\sim p_Y^G}[log\, D_s(y) \,+\, log(1 \,-\, D_s(\hat{y}) + \frac{1}{T}\sum_{t=1} T\, log\, D_f(y_t) \,+\, log(1 \,-\, D_f(\hat{y}_t))] \qquad (2)$$

  where D_s and D_f are sequence and frame discriminators.

- **Likelihood handling:**

  To maximize the likelihood $\mathbb{E}_{y\sim p_Y} p_{Y|X}(y|G(y))$ we simply force our generator to only generates missing values and leave the others intact.

$$G(y) = G_{old}(y) \odot \bar{m} + y \odot m \qquad (3)$$

Where G_{old} is the output of the generator before masking out the true values.

Datasets

The paper tackled two types of spatio temporal data :

- **Geophysical data**: Where we had only one dataset called **SST**(The Sea Surface Temperature) with a specific type of occlusion simulating **Clouds**
- **Natural Videos**: where we used 3 benchmark video datasets :
  1. **FaceForensics++** : This dataset contains 1000 videos of non-occluded face movements on a static background [5], we apply face recognition to extract faces [2]
  2. **KTH** : A human action dataset containing 2391 video clips of 6 human actions
  3. **BAIR Robot Pushing Dataset** : This dataset contains 44374 videos recorded by an one-armed robot.

  The occlusions used for this type of data are : **Raindrops**(simulating rain), **MovingBar**(Simulating an obstacle) and **RemovePixels**(simulating noise)

Model

This section details The architecture of the networks used in this paper.

Generator : As mentioned in the paper, we went with a ResNet-type self-attention network, composed of several 3D ResNet blocks followed with a spatial Self-Attention Layer. The self-Attention layer is added to help both generator and discriminator to better model the relationships between spatial regions [7].

When working with *GANs*, we usually map an input noise variable to the desired data space (say images).In other words, the generator is formulated as the stacked of deconvolution layers that transfer few features to complex examples. In ResNet-type architecture, the input corresponds to video sequences (masked ones in our case) that we pass through an encoder to extract features and then a decoder for reconstruction.

Discriminator : Traditional adversarial training processes uses only a sequence discriminator i.e it treats the entire sequence. An alternative suggests introducing another on-frames discriminator dealing with frames individually.

We use a *PatchGAN* discriminator, rather than a traditional discriminator that returns a single value over the whole frame (analogously sequence), it penalizes structure at the scale of patches where it tries to classify if each N x N patch of frame (resp sequence) is real or fake. We run this discriminator convolutionally across frames (resp sequences), we may work with the output as an array of values or averaging all responses to provide a single output.

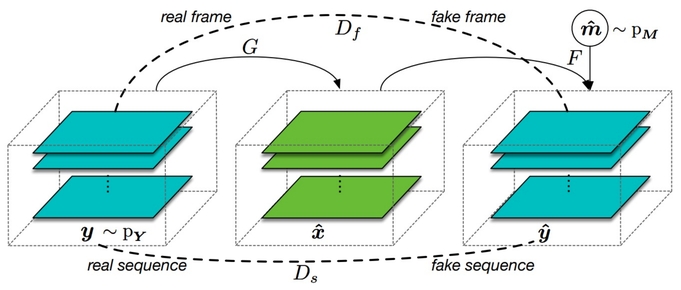


Fig. 1: Model Architecture

Results

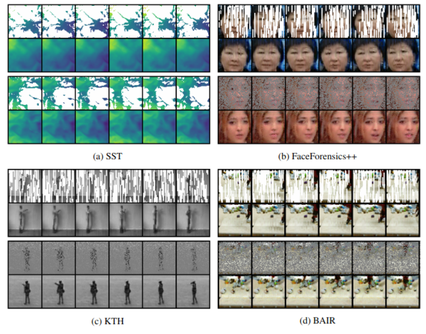
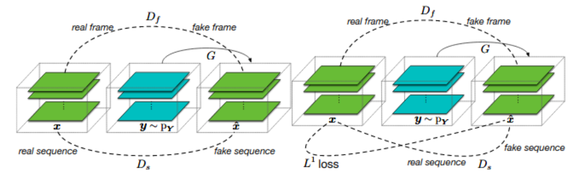


Fig. 2: Samples from test sets. SST data (a) are masked with Cloud, natural video datasets (b,c,d)are masked with Remove-Pixel and Raindrops. Each sample from top row to bottom: observed y, and recovered $\bar{x}$

Supervised variants

- **Paired variant** : We assume that we have access to corrupted-uncorrupted pairs (y, x) from the joint distribution
- **Unpaired variant** : We assume that we have access to unpaired samples where y is sampled from P_Y and x is sampled from P_X, and then we construct our corrupted-uncorrupted pair (y, x)



(a) Unpaired variant of our model, supervised (b) Paired variant of our model, supervised

Evaluation metrics

We evaluate the performance of the generator by how visually pleasant are the results. Different metrics have been proposed in the community. But, we use as main performance measure of the generated frames **Fréchet Inception Distance (FID)** [3], and for generated sequences **Fréchet Video Distance (FVD)** [6]. Both metrics are a distance measure between two multi-variate Gaussian distributions with mean and covariance (m, C) generated from $p(.)$ and (m_w, C_w) generated from $p_w(.)$

$$d^2((m, C), (m_r, C_r)) = \|m - m_w\|_2^2 + Tr(C + C_w - 2(CC_w)^{\frac{1}{2}}) \qquad (4)$$

In particular, the activations of *Inception model*, trained on ImageNet, are chosen as our random variable for the FID measure.Whereas the activations of *i3d*[1] trained on kinetics 400 are chosen for the FVD measure.
In addition, we also use **the Mean square error (MAE)** to evaluate the reconstruction deviation from the real data.

Conclusion and Remarks

- Inpainting videos in a fully unsupervised context is a hard task and an active research area.
- Training is very sensitive, not stable and GPU/time consuming.

References

[1] Joao Carreira and Andrew Zisserman. “Quo vadis, action recognition? a new model and the kinetics dataset”. In: *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6299–6308.

[2] Adam Geitgey. *Face Recognition*. `https://github.com/ageitgey/face_recognition`. 2019.

[3] Martin Heusel et al. “Gans trained by a two time-scale update rule converge to a local nash equilibrium”. In: *Advances in neural information processing systems*. 2017, pp. 6626–6637.

[4] Dahun Kim et al. *Deep Video Inpainting*. 2019. arXiv: `1905.01639 [cs.CV]`.

[5] Andreas Rössler et al. “FaceForensics++: Learning to Detect Manipulated Facial Images”. In: *International Conference on Computer Vision (ICCV)*. 2019.

[6] Thomas Unterthiner et al. “Towards accurate generative models of video: A new metric & challenges”. In: *arXiv preprint arXiv:1812.01717* (2018).

[7] Ashish Vaswani et al. *Attention Is All You Need*. 2017. arXiv: `1706.03762 [cs.CL]`.