Disentangling Physical Dynamics from Unknown Factors for Unsupervised Video Prediction

Lounès Meddahi [1] [2]

[1]ENS Rennes, France [2]MVA ENS-PSL, France

Introduction

Predicting future video frames in an **unsupervised** manner is challenging due to:

- Complex physical dynamics underlying video sequences.
- Learning how actions affect objects in its environment.

PhyDNet [1] is a candidate solution that addresses theses challenges:

It leverages a **two-branch** deep architecture, which explicitly disentangles PDE dynamics from unknown complementary information.

In this work:

- We experiment on **Moving MNIST** [2] with variations of the PhyDNet architecture, specifically:
  - **Replacing** the PhyDCell with an LSTM-based design.
  - Evaluating the impact of **removing moment regularization** on disentanglement performance.
- We study the application of **PhyDNet to weather prediction**, using a sky image dataset to showcase its limitations in real-world scenarios.

PhyDNet core concept

The authors of [1] assume that a latent space H exists, where the physical dynamics and residual factors of videos are linearly disentangled.

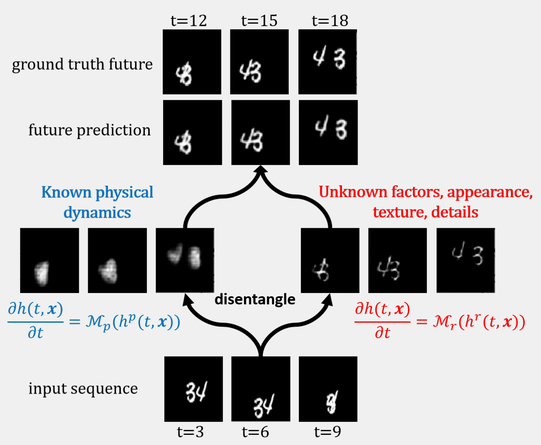


Figure 1. PhyDNet maps an input video into a latent space H, from which future frame prediction can be performed.

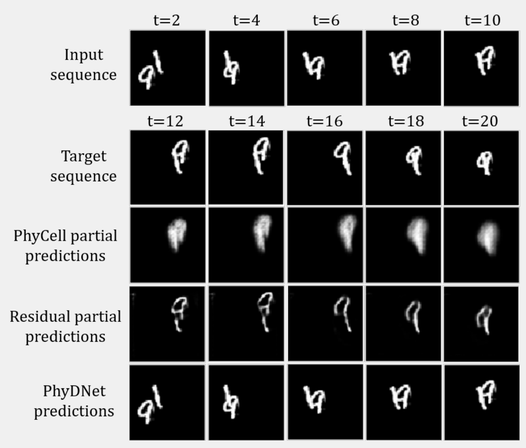


Figure 2. PhyDNet qualitative ablation results on Moving MNIST.

For a video $\mathbf{u} = \mathbf{u}(t, x)$, where x respresents spatial coordinates at time t, there is $\mathbf{h} \in H$ such that:

$$\mathbf{h} = \mathbf{h}^p + \mathbf{h}^r$$
$$\frac{\partial \mathbf{h}(t,\mathbf{x})}{\partial t} = \frac{\partial \mathbf{h}^\mathbf{p}}{\partial t} + \frac{\partial \mathbf{h}^\mathbf{r}}{\partial t} := \mathcal{M}_p(\mathbf{h}^\mathbf{p}, \mathbf{u}) + \mathcal{M}_r(\mathbf{h}^\mathbf{r}, \mathbf{u})$$

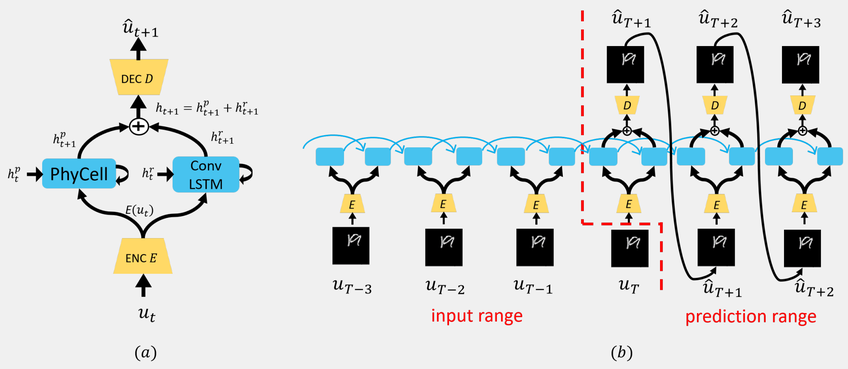


Figure 3. PhyDNet model for video forecasting: **(a)** PhyDNet disentangling recurrent bloc, and **(b)** PhyDNet global seq2seq architecture.

The prediction and correction steps of PhyCell can be written as:

$$\begin{cases} \tilde{\mathbf{h}}_{t+1} = \mathbf{h}_t + \Phi(\mathbf{h}_t), & \text{Prediction} \\ \mathbf{h}_{t+1} = \tilde{\mathbf{h}}_{t+1} + \mathbf{K}_t \odot \left(\mathbb{E}(\mathbf{u}_t) - \tilde{\mathbf{h}}_{t+1}\right), & \text{Correction} \end{cases}$$

where $\mathbf{K}_t = \tanh\left(\mathbf{W}_h * \tilde{\mathbf{h}}_{t+1} + \mathbf{W}_u * \mathbf{E}(\mathbf{u}_t) + \mathbf{b}_k\right)$, and $\Phi(\mathbf{h})$ using CNNs.

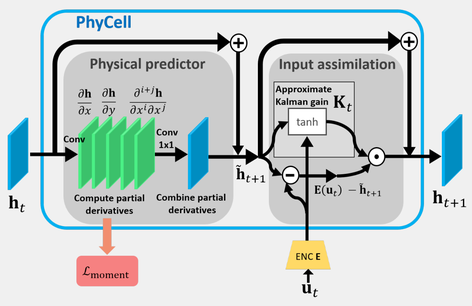


Figure 4. PhyCell's two-step scheme: prediction and correction of latent physical dynamics.

(1) $\mathcal{L}(\mathcal{D}, \mathbf{w}) = \mathcal{L}_{\text{image}}(\mathcal{D}, \mathbf{w}) + \lambda \mathcal{L}_{\text{moment}}(\mathbf{w_p})$

(2) $\mathcal{L}_{\text{image}} = L^2 = \frac{1}{N}\sum_{i=1}^{N} \|y_i - \hat{y}_i\|_2^2$

(3) $\mathcal{L}_{\text{moment}} = \sum_{i \leq k} \sum_{j \leq k} ||\mathbf{M}(\mathbf{w}_{p,i,j}^k) - \mathbf{\Delta}_{i,j}^k||_F$

Figure 5. PhyDNet loss function.

Is PhyCell Actually Useful?

Our experiments show that incorporating a **PhyCell** with the moment regularization technique provides only **marginal benefits** when training a PhyDNet architecture on the Moving MNIST dataset, **compared to employing two Conv-LSTM** branches.

We use a PhyDNet model composed of one Conv-LSTM for the physical branch and one for the residual branch. This model has 5,368,961 parameters (2,508,032 per branch). Training was performed using an L^2 loss: $\mathcal{L}(\mathcal{D}, \mathbf{w}) = \mathcal{L}_{\text{image}}(\mathcal{D}, \mathbf{w}) + \lambda \cancel{\mathcal{L}_{\text{moment}}(\mathbf{w_p})}$.

Table 1. Double-LSTM PhyDNet vs PhyDNet on Moving MNIST.

Method	MSE	MAE	SSIM
Double LSTM (100 epochs)	47.17	116.26	0.86
PhyDNet (2,000 epochs)	**24.4**	**70.3**	**0.95**

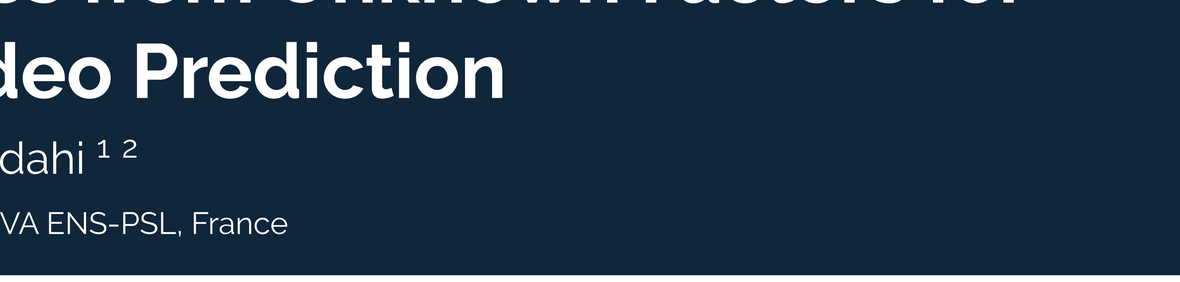


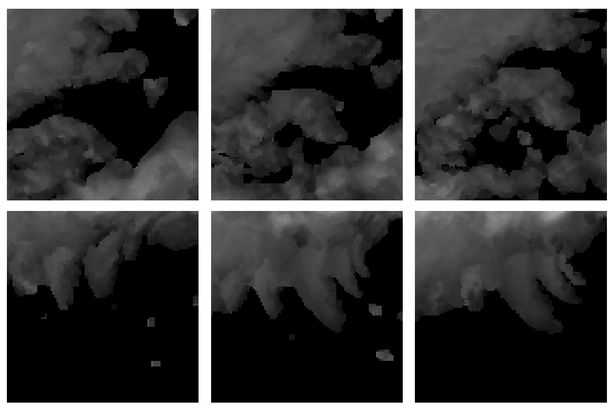
Figure 6. A visual comparison of predictions made by different parts of the model. The first four columns correspond to predictions at time $t+1$, while the last four columns correspond to predictions at time $t+8$. The first line show results from the double-LSTM model, and the second line show results from the PhyDNet model. LB (resp. RB) corresponds to the model's prediction when using only the Leaft Branch (resp. Right Branch) of the model. Pred. is the image predicted using both the left and the right branches of the model.

Precipitation Forecasting with Weather Radar Data

Dataset: We use the **radar echo map dataset** from the CIKM AnalytiCup 2017 competition. This dataset comprises **10,000 sequences of 15 radar images** per sequence, that span a 101km × 101km region of Shenzhen, China. The radar images focus on moderate precipitation events (40 dBZ).

Each radar image has a resolution of 101×101 pixels, and the **temporal resolution is 6 minutes between frames**. **We trained PhyDNet on the first 10 radar maps** of each sequence as input, while the goal is to **predict the subsequent 5 maps**.

In table 2, **we show our good results (PhyDNet)** compare to the CIKM challenge winner Marmot [3].



Method	MSE	MAE	SSIM
PhyDNet (ours)	**20.52**	180.29	0.60
Marmot	120.94	-	-

Table 2. PhyDNet vs Marmot performance on the CIKM dataset.

Figure 7. On the left, the figures show two distinct sequences of the CIKM dataset, with a temporal resolution of 12 minutes between consecutive frames. On the right side, the table shows the high performance of PhyDNet on the CIKM dataset.

Figure 8. Visual illustration of two different sequences predicted by the trained PhyDNet model. The first three columns represent the target frames ($t+1$, $t+2$, $t+3$), while the last three columns show the corresponding frames predicted by PhyDNet. We achieve 15.77 mse and 0.67 ssim for the first sequence, 7.08 mse and 0.71 ssim for the second sequence.

References

[1] Vincent Le Guen and Nicolas Thome.
Disentangling physical dynamics from unknown factors for unsupervised video prediction.
In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11474–11484, 2020.

[2] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov.
Unsupervised learning of video representations using lstms.
In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 843–852, Lille, France, 07–09 Jul 2015. PMLR.

[3] Yichen Yao and Zhongjie Li.
Cikm analyticup 2017: Short-term precipitation forecasting based on radar reflectivity images.
In *CIKM AnalytiCup 2017*, 2017.

lounes.meddahi@gmail.com