Internship – Master degree or Engineering School

Title: Deep Learning meets the Physical World – Adversarial Training for Natural Image Sequences Prediction

Advisor : Patrick Gallinari (Universite Pierre et Marie Curie- Paris)

Location : Laboratoire d'Informatique de Paris 6 – LIP6 , Machine Learning and Information Access Team - https://mlia.lip6.fr/

Duration : 6 months

Context

Despite impressive successes in a variety of domains as demonstrated by the deployment of Deep Learning methods in fields such as vision, language, speech, etc., Machine Learning methods are not yet ready to handle the level of complexity required for modeling complex phenomena like those occurring in natural physical processes like e.g. climate modeling. With the deployment of large sensor networks and satellites, data are collected on a regular basis (e.g. daily) so that huge amounts of data characterizing many complex phenomena are now available.
One of the biggest challenges in science today is to develop the statistical paradigm in order to exploit these huge amounts of data and to make the best usage of this knowledge source. There are two main directions to explore for this: one consists in incorporating prior information gained from physical knowledge in order to guide the design of statistical models, the second one consists in dealing with the uncertainty inherent to complex natural data analysis.

Within this general context, the topic of the internship consists in developing Deep Learning models for modeling sequences of images gathered from observations via satellites. We will focus on the development of image sequence prediction models and on the modeling of the uncertainty inherent to these data. This problem is reminiscent of video prediction but in a more complex context.

Internship description

Satellites provide different types of periodic measures on the earth surface. We will consider here the specific problem of forecasting See Surface Temperature (SST). Forecasting consists in predicting future temperature maps using past records acquired via satellite imagery. If we focus on a specific area, we can formulate the problem as prediction of future temperature images of this area using past images. This is similar to frame prediction in videos, a problem recently addressed in the Deep learning community, but because of the nature of the data and of the acquisition process, the problem is more complex.
As for classical video prediction, a straightforward application of Neural Net predictors does not provide a viable solution to this problem: predicted images remain blurry and the complex underlying dynamics is not captured by these models. However, by taking inspiration from general physical knowledge, it is possible to design Deep architectures able to challenge the classical approaches to this problem. The model illustrated in the Figure 1 (de Bezenac et al. 2017) provides state of the art performance on SST data. It is composed of two modules. One predicts a motion field from a sequence of past input images, this is the convolutional-deconvolutional (CDNN) module on the top of figure 1, and the other warps the last input image using the motion field from the first component, in order to produce an image forecast.

The entire system is trained in an end-to-end fashion, using only the supervision from the target SST image.
The specific form of this model has been inspired from prior knowledge in physics and its performance for short term image prediction is on par with classical numerical methods developed specifically for the prediction problem (Bérézia t et al. 2015).

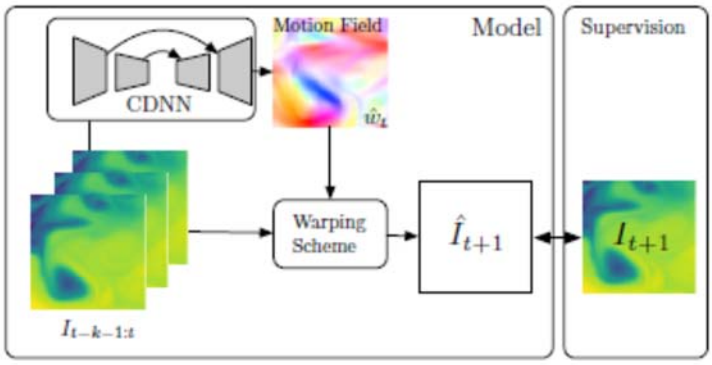

Figure 1: Motion is estimated from the input images ($I_{t-k-1:t}$) with a convolutional neural network (top left CDNN component). A warping scheme then displaces the last input image along this motion estimate to produce the future image. The error signal is calculated using the target future image I_{t+1}, and is backprogated through the warping scheme to correct the CDNN. To produce multiple time-step forecasts, the predicted image is fed back in the CDNN in an autoregressive manner.

The objective of the internship is to augment this system with the modeling of uncertainties. Real data usually come with missing information: e.g. satellite images could be partially hindered by atmospheric conditions and cannot be used as such for training a Deep architecture. One solution is to use realistic simulated data which can be produced in large quantities to train a deep architecture so as to capture the global latent dynamics of the underlying process and then to augment the model with a correction module whose goal is to learn a mapping from the model output predictions to the real images. For this we propose to make use of an adversarial training strategy which has become popular with the development of Generative Adversarial Networks (Mathieu 2015). This should allow to model uncertainties both at the model and at the data levels.

Related work

Within the Deep Learning community, this topic is related to recent developments in video prediction and motion estimation in videos. The domain of application is clearly different from video modeling, but since the solution involves predicting a motion field and the next SST image, the two problems share some similarities and the literature on video prediction could be an interesting source of inspiration for the internship.
It is only very recently that video prediction emerged as a task in the Deep Learning community. People are generally interested at predicting accurately the displacement/ emergence/ disappearing of objects in the video. In our application, the goal is clearly different since we are interested into modeling the whole dynamics behind image changes and not at following moving objects. Let us first introduce some methods that perform prediction by computing optical flow or a similar transformation. Both Patraucean et al. (2015) and Finn et al. (2016) use some form of motion flow estimation. For next frame prediction Patraucean et al. (2015) introduce a STN module (Jaderberg et al. 2015) at the hidden layer of a LSTM in order do estimate a motion field in this latent space. The resulting image is then decoded in the original image space for prediction. Finn et al. (2016) learn affine transformations on image parts in order to predict object displacement and Van Amersfoort et al. (2017) proposed a similar model.

Let us now consider models that directly attempt to predict the next frame without estimating a motion field. Mathieu et al. (2015), proposed to use different loss functions and a GAN regularization of a CDNN predictor which led to sharper and higher quality predictions. Significant improvements have been obtained with the Video Pixel Network of Kalchbrenner et al. (2016), which is a sophisticated architecture composed of resolution preserving CNN encoders, LSTM and PixelCNN decoders which form a conditional Spatio-temporal video autoencoder with differentiable memory. This model is probably state of the art today for video prediction. A drawback is the complexity of the model and the number of parameters: they are using respectively 20 M and 1 M frames on these two datasets.

Références :

- D. Béréziat and I. Herlin. Coupling dynamic equations and satellite images for modeling the ocean surface circulation. Communication in Computer and Information Science, 550, 2015.
- Emmanuel de Bezenac, Arthur Pajot and Patrick Gallinari, Deep Learning for Physical Processes: Incorporating Prior Scientific Knowledge, 2017, CoRR, abs/1711.07970}, {2017, url http://arxiv.org/abs/1711.07970.
- Chelsea Finn, Ian J. Goodfellow, and Sergey Levine. Unsupervised learning for physical interactionthrough video prediction. CoRR, abs/1605.07157, 2016. URL http://arxiv.org/abs/1605.07157.
- Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. CoRR, abs/1506.02025, 2015. URL http://arxiv.org/abs/1506.02025.
- Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. CoRR, abs/1511.05440, 2015. URL http://arxiv.org/abs/1511.05440.
- Pajot, A. Ziat, L. Denoyer and P. Gallinari, 2016, Incorporating Prior Knowledge in spatio-temporal neural network for climatic data, . Banerjee, W. Ding, J. Dy, V. Lyubchich, A. Rhines (Eds.), I. Ebert-Uphoff, C. Monteleoni, D. Nychka (Series Eds.), Proceedings of the 6th International Workshop on Climate Informatics
- Viorica Patraucean, Ankur Handa, and Roberto Cipolla. Spatio-temporal video autoencoder with differentiable memory. CoRR, abs/1511.06309, 2015. URL http://arxiv.org/abs/ 1511.06309.
- Joost Van Amersfoort, Anitha Kannan, Marc'Aurelio Ranzato, Arthur Szlam, Du Tran, and Soumith Chintala. Transformation-Based Models of Video Sequences. In CoRR abs/1701.08435 (2017), pp. 1–11, 2017. URL http://arxiv.org/abs/1701.08435.

- Nal Kalchbrenner, Aaron van den Oord, Karen Simonyan, Ivo Danihelka, Oriol Vinyals, Alex Graves, and Koray Kavukcuoglu. Video pixel networks. CoRR, abs/1610.00527, 2016. URL http://arxiv.org/abs/1610.00527.
- N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in Proceedings of the 32nd ICML- 15, D. Blei and F. Bach, Eds., 2015.

Practical information

Internship location : LIP6, team MLIA - *https://mlia.lip6.fr/*
Contact : patrick.gallinari@lip6.fr
Requirements: The candidate will have a background in statistical machine, and if possible a first experience in Deep Learning. He should be able to develop both theoretical research and practical implementations and experiments.

Internship grant: around 570 E/ month