Video Frame Prediction

Kishore P. Venkatswammy Reddy
Kaiwen Wu

Outline

- Problem Statement
- Datasets
- Brief literature survey
- Architectures
  - CDNA [3]
  - SV2P [4]
  - FutureGAN [1]
- Future plans
- References

Problem Statement

Given a frame or a sequence of frames (video) predict the next frame or next sequence of frames

Consequences

- Transferable to other tasks
  - Video understanding - classification, annotation, compression
- Better planning agents
  - Threat anticipation agents
  - Autonomous vehicles/robots

Datasets

- Moving MNIST
- KTH
- UCF101
- CityScape

Current approaches

- Inherently difficulty of the problem

- Approaches
    - Motion models - capture the motion using optical flows/img differences
    - Stochastic models - address uncertainty in predicting future frames
    - Generative models - sharper frames, at the cost of difficult, long training

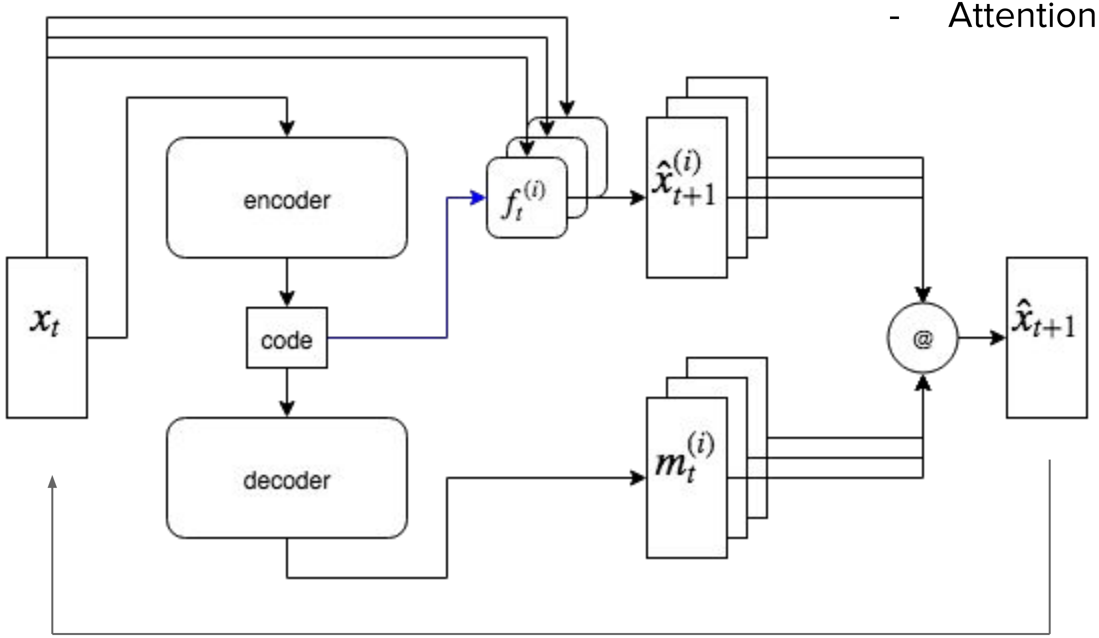
Explicit representation learning

- Disentangling instance-level foreground from background
  - Dynamic filter (Brabandere et al., 2016)
  - DNA/CDNA/STP (Finn et al., 2016)
  - SfM-Net (Vijayanarasimhan et al., 2017)

- Assumption on foreground and background
  - Foreground objects: the moving pattern is homogeneous within an object
  - Background: either static, or otherwise due to camera motion

How they work

Why separation of masks:
- Regularization
- Interpretability
- Attention

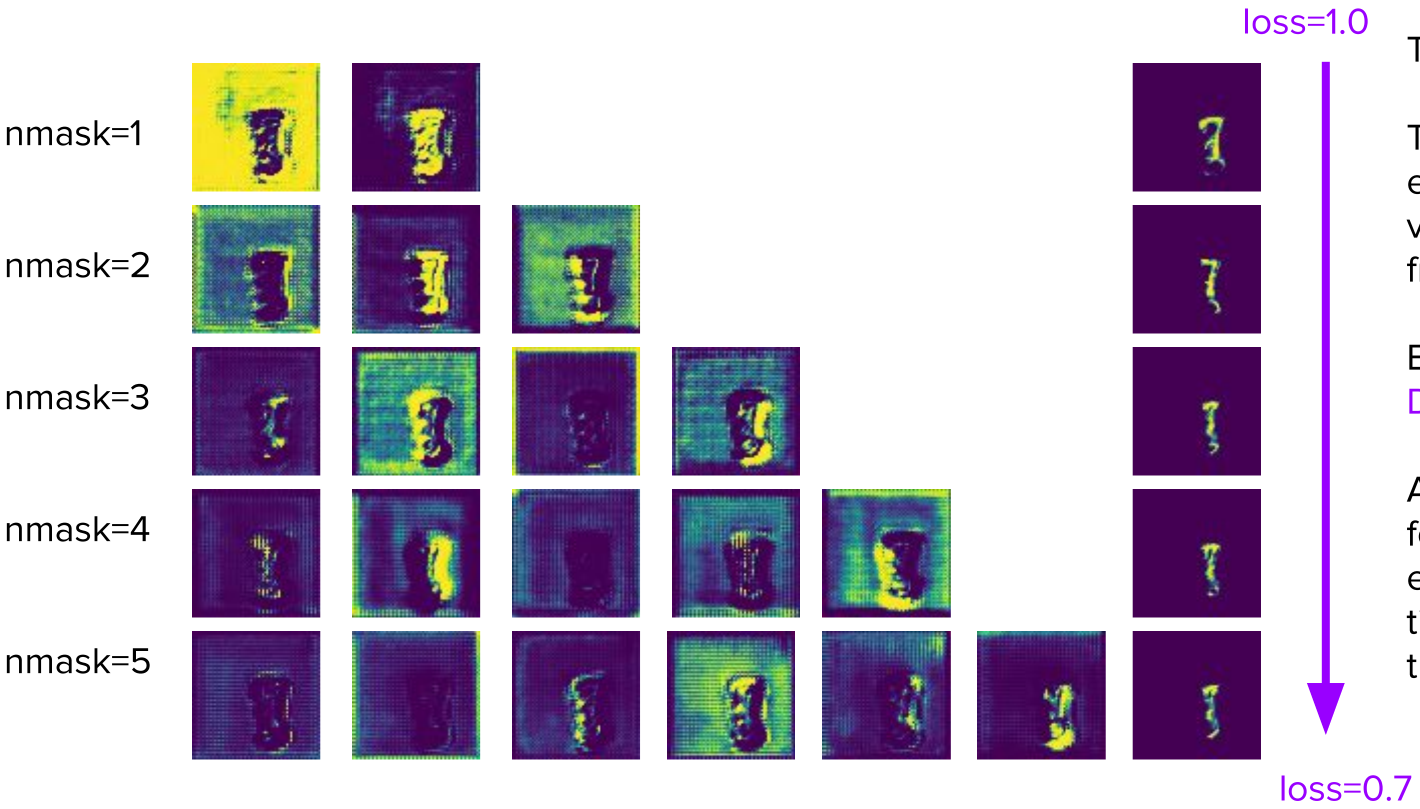


How they work

CDNA	Stacked Conv-LSTM as encoder-decoder 5x5 convolution as transformations
STP	Stacked Conv-LSTM as encoder-decoder Spatial transformer as transformations
SfM-Net	U-Net as encoder-decoder SE3 rigid transformation

There are also other works that make different combinations of modules mentioned above.

CDNA & Moving MNIST experiments



Teacher forcing

Trained for 2 epochs over 16000 videos, each of 20 frames long

Evaluation criterion: DSSIM

Apparently teacher forcing doesn't encourage time-variant transformation

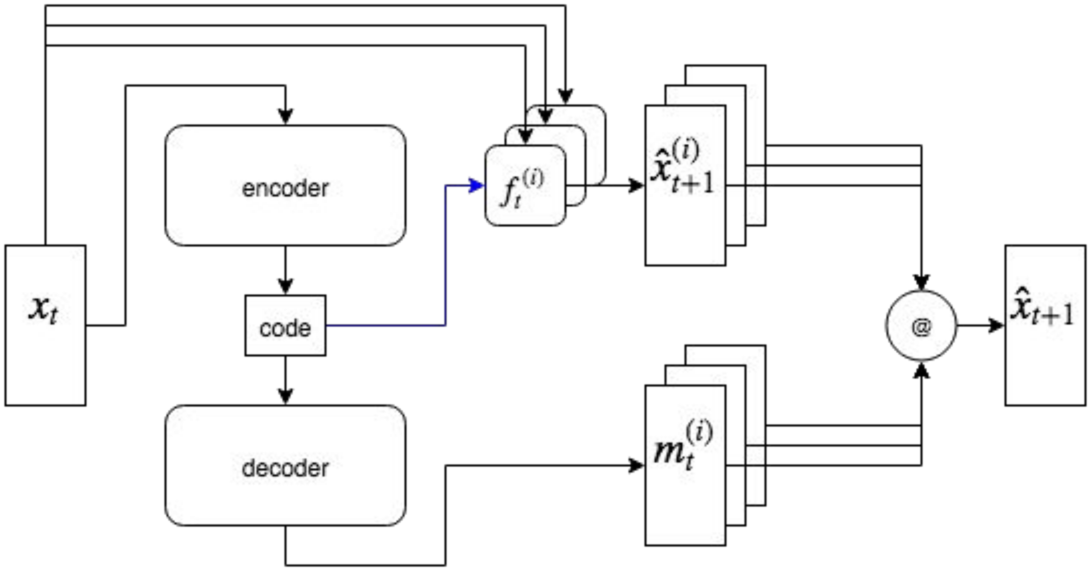
Interpretability issue

- Object masks segmentation is limited by the size of CDNA kernel -- only local properties are focused, and it’s far from ideal case
- No more good background segmentation when there are at least two object masks

- Since Moving MNIST has black background, whatever conv kernel can be applied on it and nothing will get wrong. This could explain why it confuses foreground and background.

SV2P

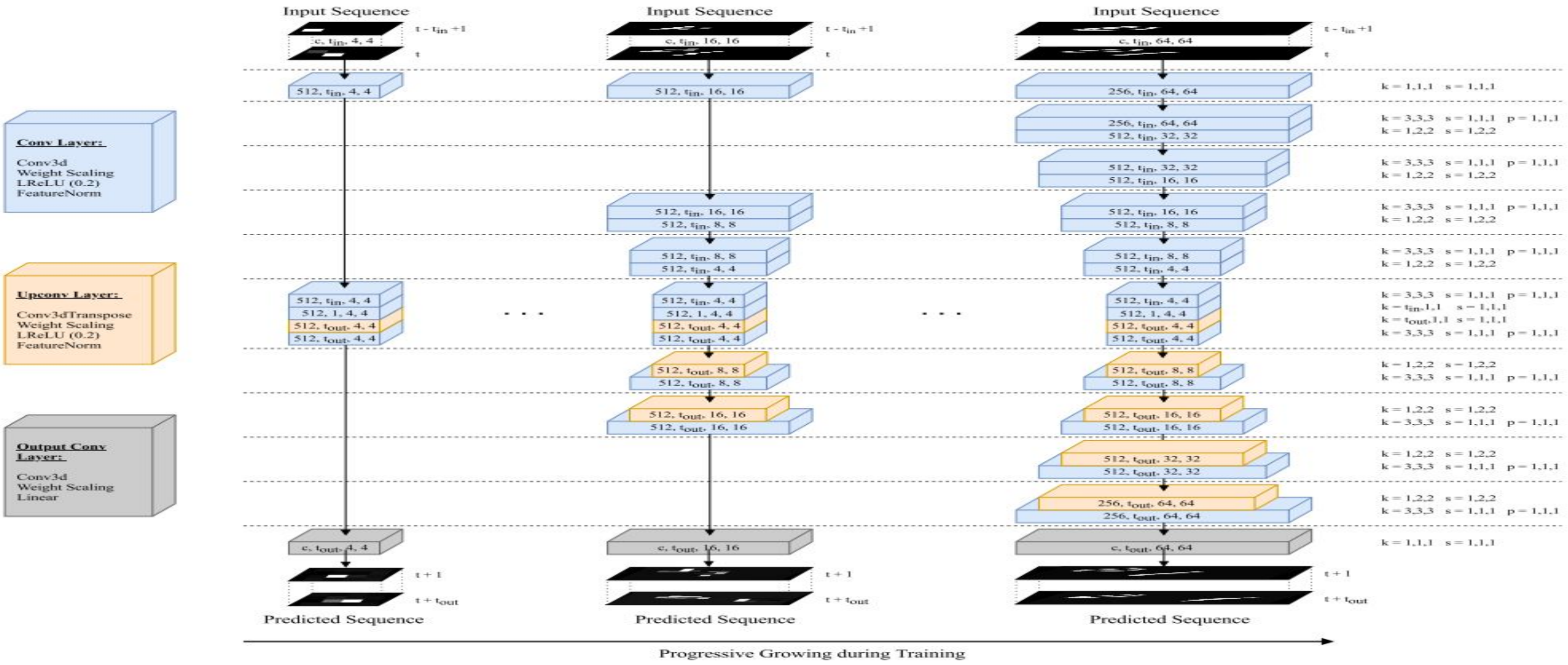
It’s an improvement over CDNA net, that aims to sharpen the long-term prediction by introducing latent random variables to **code**, such that the learnt latent distribution contains guidance on how to predict.



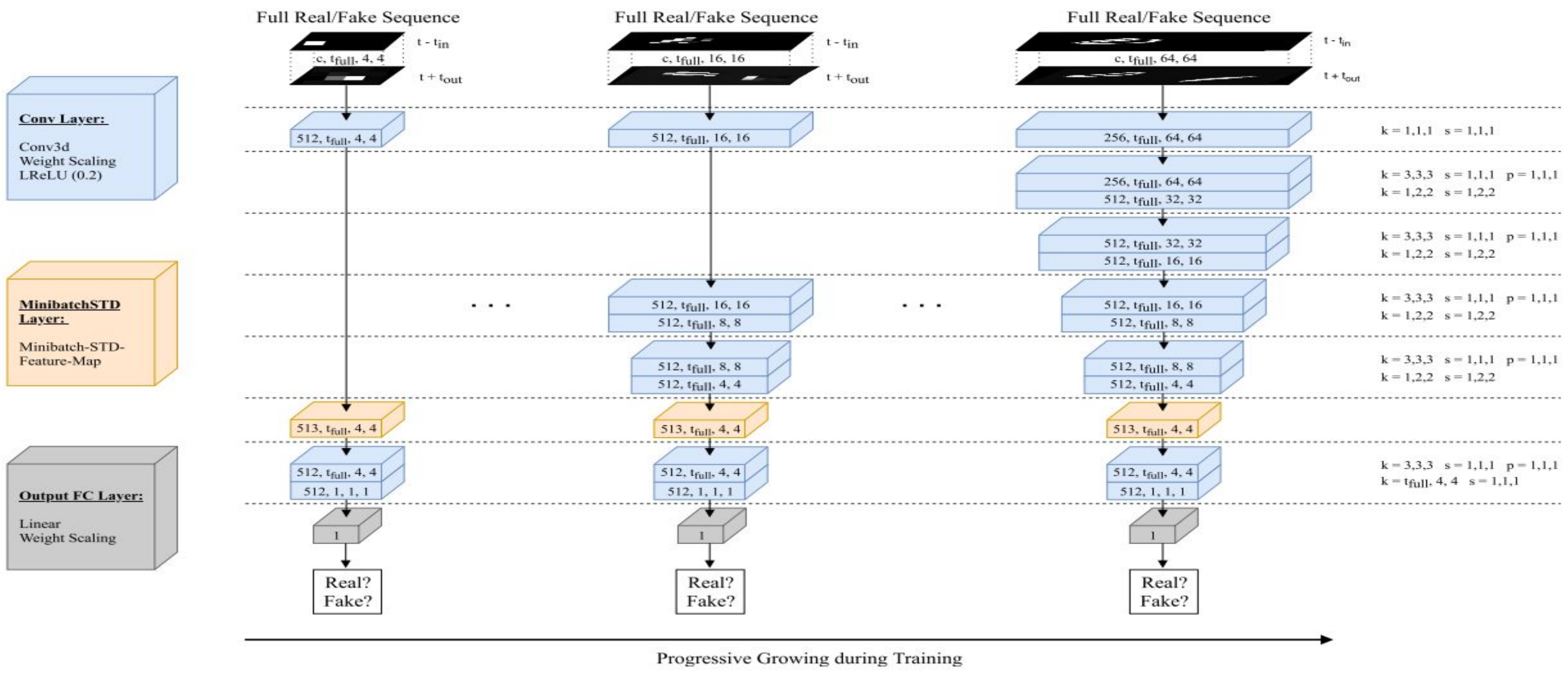
FutureGAN

- Architecture modelled after PGGAN [2]
  - PGGAN - Progressively Growing GAN
    - Overcomes problems of GAN training and mode collapse
- Details
  - Generator network - Encoder and Decoder
    - Generates the future frames
    - Used for predictions
  - Discriminator network - Decoder
    - Discriminates real from fake

FutureGAN - Generator

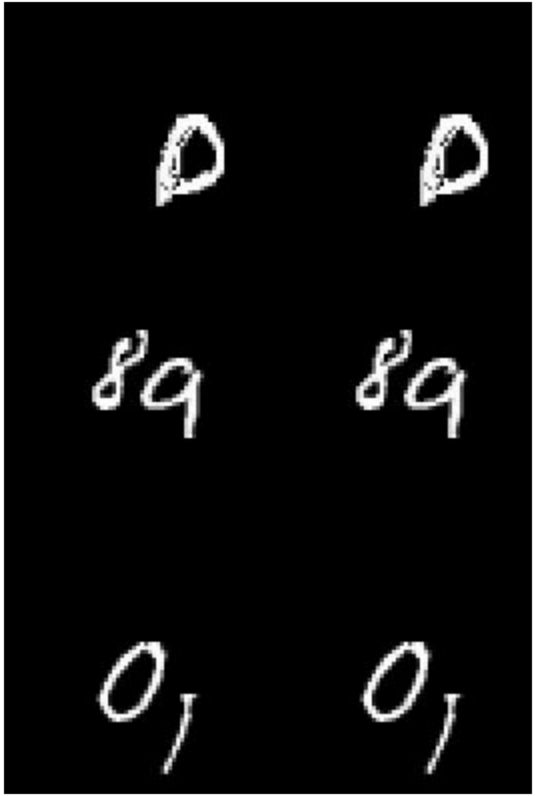


FutureGAN - Discriminator



Results

The left animations are the original video, the right are the corresponding predictions of network

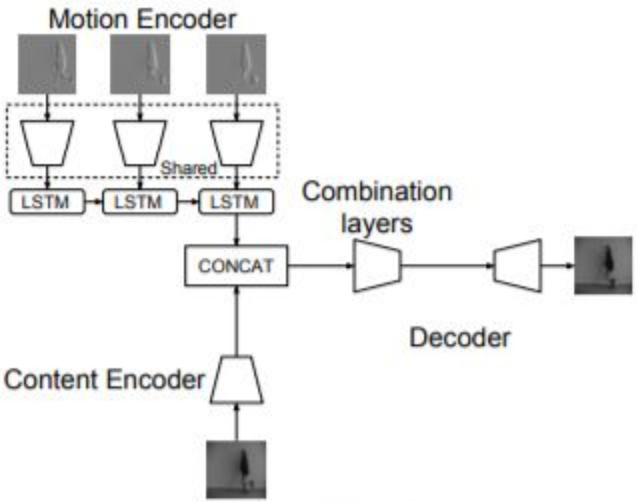


Future Plans

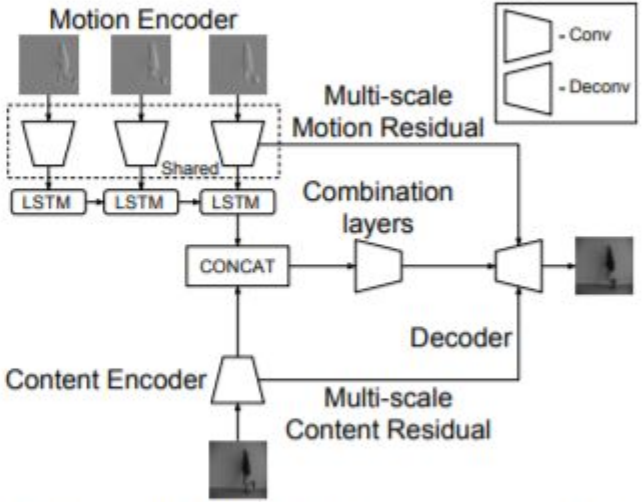
- 1 week plan
  - Improve interpretability of CDNA
  - Exploring the latent distribution of SV2P
  - Motion-Content Networks with hard attention

- 2 / 3 week plan
  - Construct and experiment with simplified PGGAN architectures

MCNet



(a) Base MCnet



(b) MCnet with Multi-scale Motion-Content Residuals

References

1. FutureGAN - https://arxiv.org/abs/1810.01325
2. PGGAN - https://arxiv.org/abs/1710.10196
3. Chelsea Finn, Ian J. Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. CoRR, abs/1605.07157, 2016. URL http://arxiv.org/abs/1605.07157.
4. Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H. Campbell, and Sergey Levine. Stochastic variational video prediction. CoRR, abs/1710.11252, 2017. URL http://arxiv.org/abs/1710.11252.
5. Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragki- adaki. SfM-Net: Learning of Structure and M
6. Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. CoRR, abs/1506.03099, 2015. URL http://arxiv.org/abs/1506.03099.

Questions ?

Experiments

- Train the network at 128x128 resolution directly
    - Confirmed our suspicions!
- Noisy test data
    - Resilient to small amount of input noise