Unsupervised Learning of Video Representations using LSTMs

Srivastava et al.
University of Toronto

Presented by Shyam Tailor
The Overall Idea

• Take a sequence of images and encode into a *fixed size* latent representation

• Decode latent representation back into a target sequence
What Should The Latent Representation Encode?

• Significant redundancy between frames

• Three things that seem reasonable to encode:
  • Background
  • Objects
  • Motion
The Target Sequence

Reconstruction (*in reverse*!)

Predicting the future
Why Reverse The Reconstruction?

- **Idea** – latent representation is like a stack
  - Encoder pushes on and decoder pops off
Future Prediction

• To do well the latent representation must encode the objects and how they’re moving

• Note: *this puts subtly different requirements on the encoder!*
Conditioning the Decoder

• A small detail – the decoder can be conditioned on the previously generated frame

• Not really important but improves results a little.
Combining the Tasks

• The two tasks alone aren’t good enough 😞

• Why?
  • Reconstruction requires *memorisation* but doesn’t require encoding to be useful to predict future
  • Future prediction doesn’t incentivise keeping frames from the past
An Experiment with MNIST

Two Layer Composite Model with a Conditional Future Predictor
Trying Natural Images

Two Layer Composite Model with 4096 LSTM units
Zooming In
“Designing a loss function that respects our notion of visual similarity is a very hard problem”

True... Let’s return to this at the end
Seeding a Classifier with the Encoder

- Going to do *human action recognition* on some video datasets (UFC-101, HMDB-51).

- Is initializing with the encoder weights better than starting from random?
- What if the encoder is trained on unrelated videos?
Results of Pretraining

• *Encoder features transfer well and yield accuracy improvements*
• Especially pronounced with a small dataset
• Using random YouTube videos doesn’t affect accuracy!
Does the Encoding Really Have a Concept of Motion?

• Instead of using the RGB images, it’s possible to train on the optical flow vectors instead

• Pretraining significantly less effective in this regime.
Authors’ Conclusions

• Great qualitative performance on the moving MNIST dataset – but falls over on natural images

• Nevertheless pretraining for natural images seems to have some effect
  • It seems a stronger notion of optical flow is obtained
Discussion: How do you make your frame predictions less blurry?

• One idea is to use an adversarial loss.
• Liang et al. 2017 tried this; their embedding was also great for pretraining on UFC-101
Discussion: Interpreting the Encoding

• Is there any form of interpretability?

• Examples:
  • Are encodings of motion, objects and background merged together or distinct?
  • Is it possible to extract specific objects from the encoding?
Discussion: What About Regularisation?

• The authors saw no difference between pretraining on YouTube and the activity recognition – how much does domain matter?

• Is it possible to use a VAE by reframing the problem?
  • See “Learning to Decompose and Disentangle Representations for Video Prediction” by Hsieh et al.
References
