Unsupervised Learning of Video Representations using LSTMs

(Nitish Srivastava, Elman Mansimov, Ruslan Salakhutdinov)

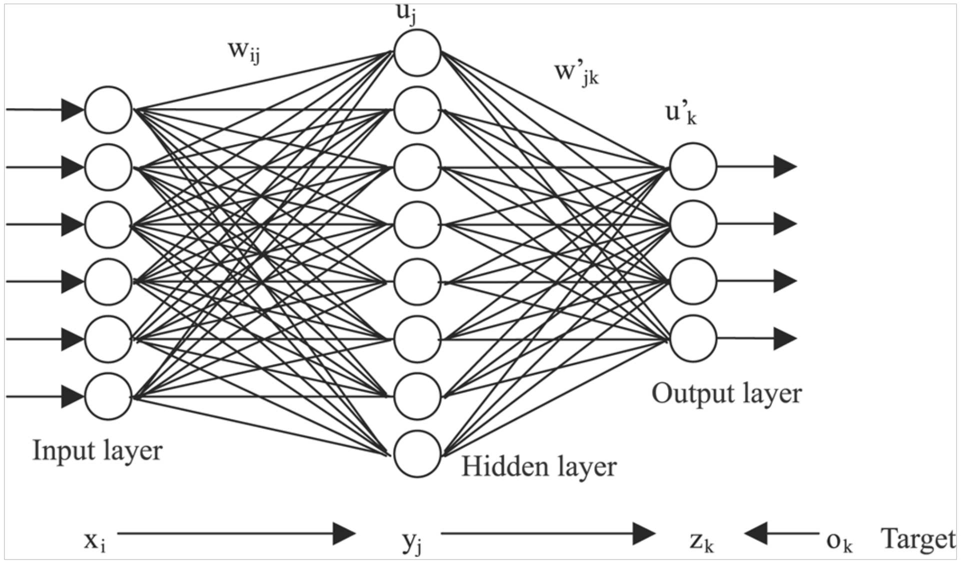
Davit Buniatyan
Unsupervised Learning Seminar
Princeton 2017

Motivation

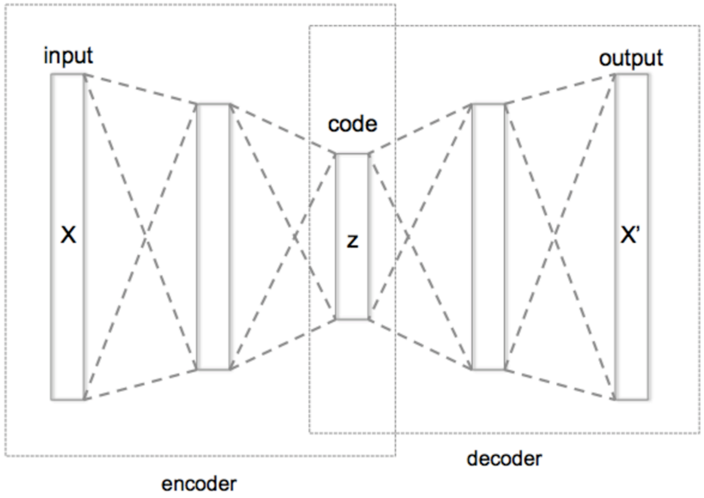
To learn ‘good’ video representation that can

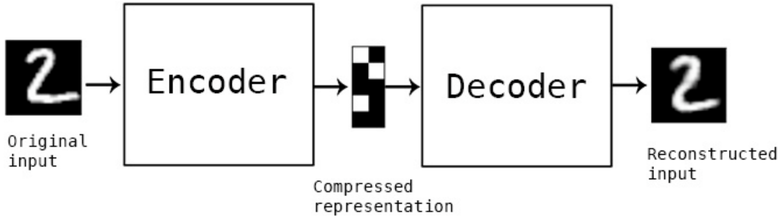
- Reproduce the sequence of frames
- Predict the future frames
- Be used later for supervised tasks such as action recognition

Neural Networks

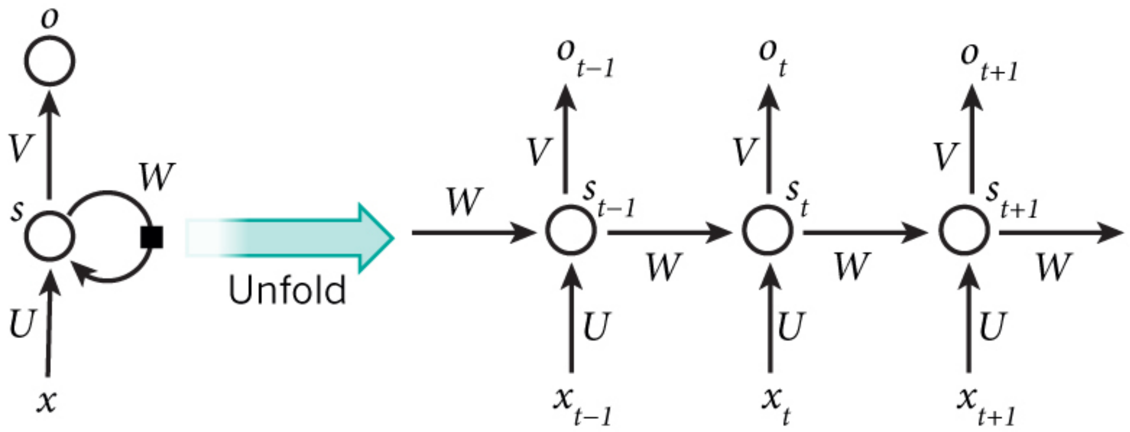


Autoencoders





Recurrent Neural Networks



Unrolling Recurrent Networks

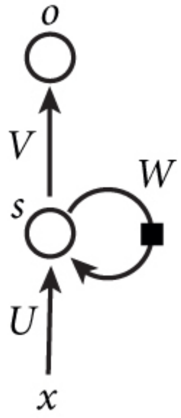
Cells

Simple RNN

$\mathrm{RNN}(x_i, h_i)$

$$h_{i+1} = \mathbf{W}^{\mathbf{h}} h_i$$
$$o_{i+1} = \mathrm{sig}(\mathbf{W}^{\mathbf{i}} x_i + \mathbf{h}_{i+1} + \mathbf{b})$$

return o_{i+1}, h_{i+1}



Computation at each timestep

$$(y_{i+1}, h_{i+1}) = \mathrm{RNN}(x_i, h_i)$$

(where x_i,h in R^d and $\mathbf{W}^{\mathbf{i,h}}$ in R^{dxd}, d is the number of rnn subcells)

Cells

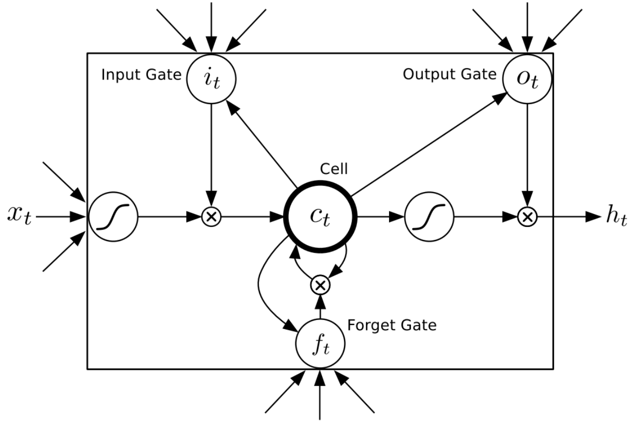
Long-Short Term Memory (LSTM)

$H = [Ix_i, h]^T$
$W = [\mathbf{W}^u, \mathbf{W}^f, \mathbf{W}^o, \mathbf{W}^c]$

LSTM(H, m, W)

$\mathbf{g}^u = \text{sig}(\mathbf{W}^u\ H)$
$\mathbf{g}^f = \text{sig}(\mathbf{W}^f\ H)$
$\mathbf{g}^o = \text{sig}(\mathbf{W}^o\ H)$
$\mathbf{g}^c = \tanh(\mathbf{W}^c\ H)$
$m' = g^f \odot m + g^u \odot g^c$
$h' = \tanh(g^o \odot m')$

return m', h'

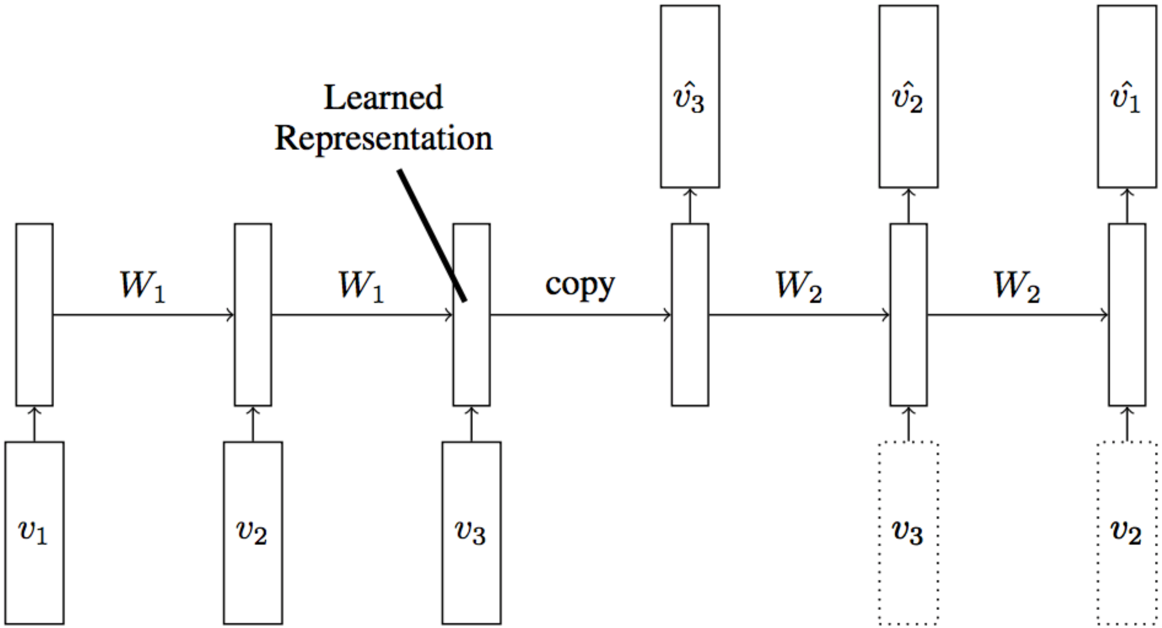


(where Ix_i,h,m in R^d and $\mathbf{W}^{u,f,o,c}$ in R^{dx2d}, d is the number of rnn subcells)

Originally 1997 by Sepp Hochreiter and Jürgen Schmidhuber

Recurrent AutoEncoders?

Recurrent Encoder-Decoder

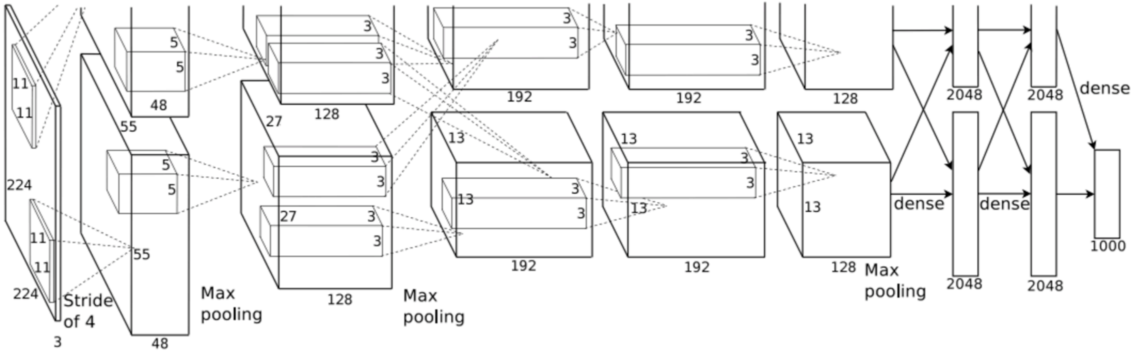


Input at each timestep

Image Patches (e.g. MNIST)

Features trained on ImageNet (Krizhevsky, Sutskever, Hinton 2012)

- Convolutional Networks
- Transfer Learning



Unsupervised Evaluation Strategy

Qualitative

- Reconstruction
- Future prediction

Quantitative

- Action Recognition

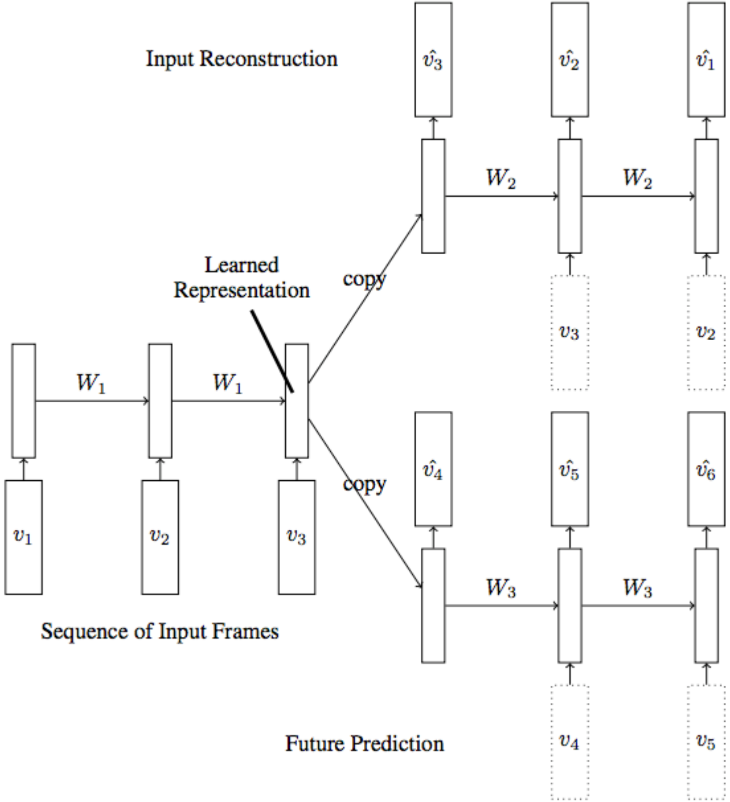
Predicting The Future

- Composite model

Why?

- Conditional input

Why?



Objectives

Understand

- Qualitative Analysis: What does the LSTM actually learn to do?
- Transfer Learning: How good we can transfer the knowledge for supervised tasks?

Compare

- Different models (e.g. Autoencoder, Future Predictor)
- State-of-the-art action recognition benchmarks

Visualization and Qualitative Analysis

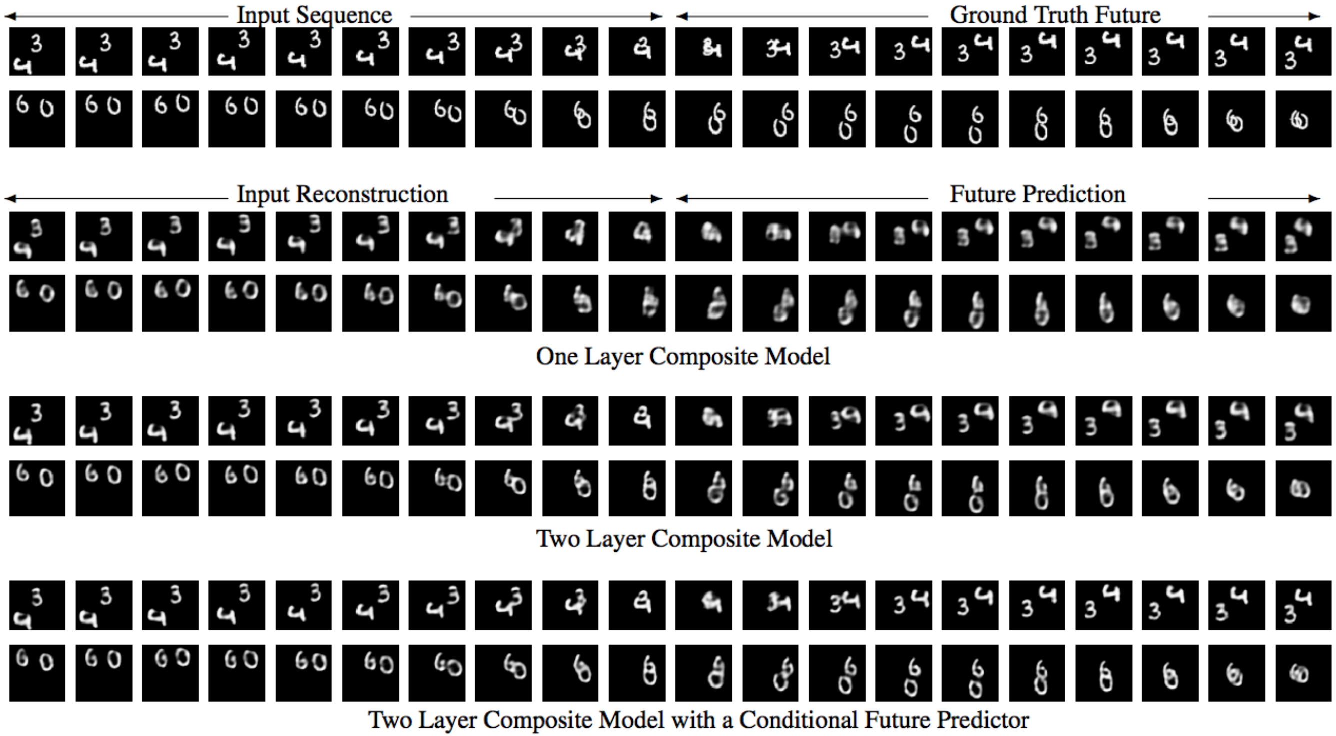


Figure 5. Reconstruction and future prediction obtained from the Composite Model on a dataset of moving MNIST digits.

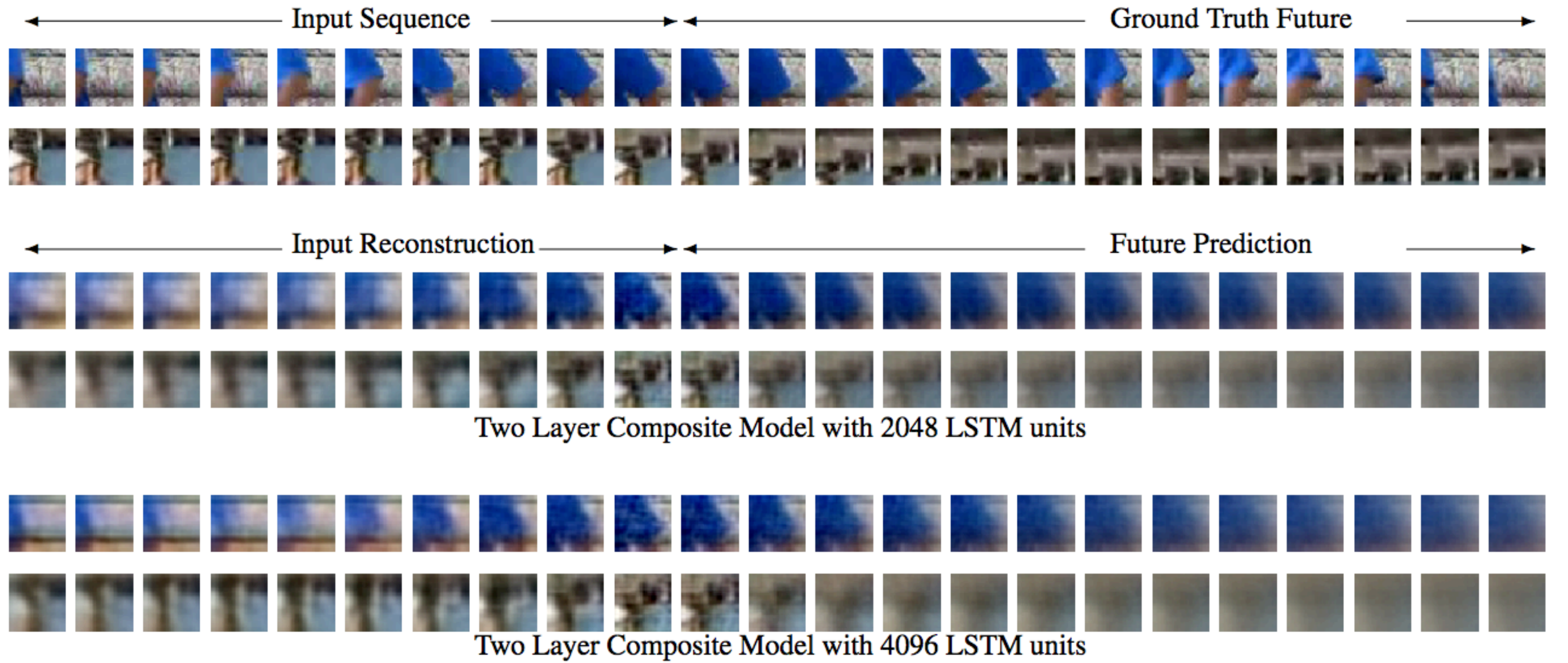


Figure 6. Reconstruction and future prediction obtained from the Composite Model on a dataset of natural image patches. The first two rows show ground truth sequences. The model takes 16 frames as inputs. Only the last 10 frames of the input sequence are shown here. The next 13 frames are the ground truth future. In the rows that follow, we show the reconstructed and predicted frames for two instances of the model.

Transfer Learning for Action Recognition

- Model

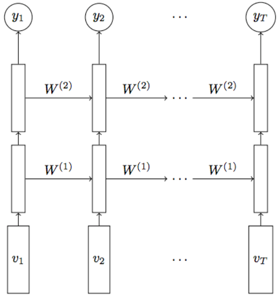
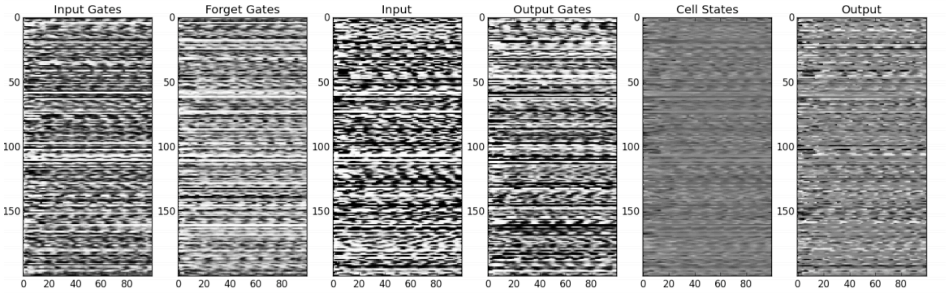


Figure 11. LSTM Classifier.

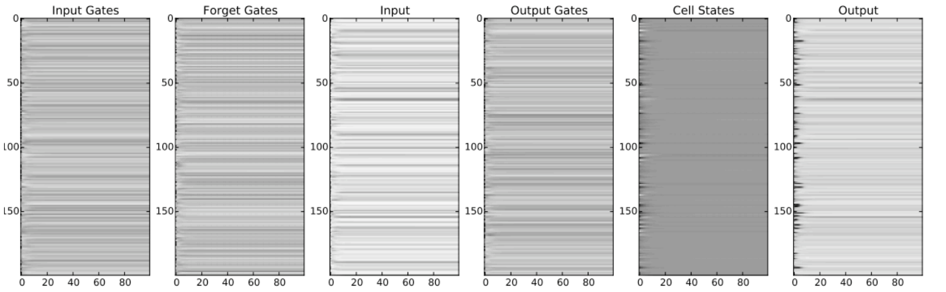
- Results

Model	UCF-101 RGB	UCF-101 1- frame flow	HMDB-51 RGB
Single Frame	72.2	72.2	40.1
LSTM classifier	74.5	74.3	42.8
Composite LSTM Model + Finetuning	**75.8**	**74.9**	**44.1**

Table 1. Summary of Results on Action Recognition.



(a) Trained Future Predictor



(b) Randomly Initialized Future Predictor

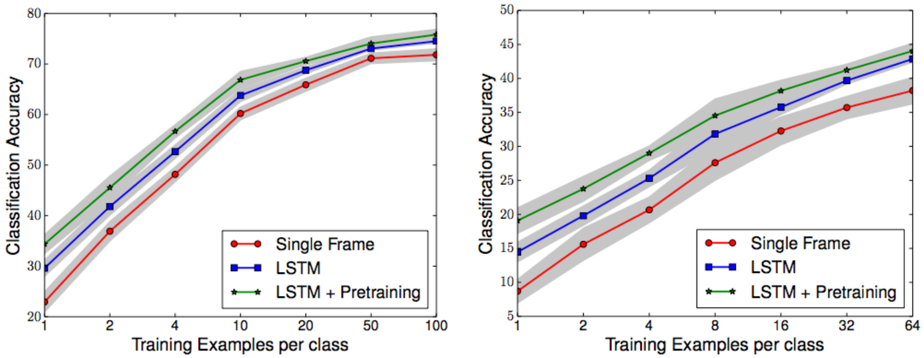
Benchmarking

Model	Cross Entropy on MNIST	Squared loss on image patches
Future Predictor	350.2	225.2
Composite Model	344.9	210.7
Conditional Future Predictor	343.5	221.3
Composite Model with Conditional Future Predictor	341.2	208.1

Table 2. Future prediction results on MNIST and image patches. All models use 2 layers of LSTMs.

Method	UCF-101	HMDB-51
Spatial Convolutional Net (Simonyan & Zisserman, 2014a)	73.0	40.5
C3D (Tran et al., 2014)	72.3	-
C3D + fc6 (Tran et al., 2014)	**76.4**	-
LRCN (Donahue et al., 2014)	71.1	-
Composite LSTM Model	75.8	44.0
Temporal Convolutional Net (Simonyan & Zisserman, 2014a)	**83.7**	54.6
LRCN (Donahue et al., 2014)	77.0	-
Composite LSTM Model	77.7	-
LRCN (Donahue et al., 2014)	82.9	-
Two-stream Convolutional Net (Simonyan & Zisserman, 2014a)	88.0	59.4
Multi-skip feature stacking (Lan et al., 2014)	**89.1**	**65.1**
Composite LSTM Model	84.3	-

Table 4. Comparison with state-of-the-art action recognition models.



(a) UCF-101 RGB
(b) HMDB-51 RGB

Figure 12. Effect of pretraining on action recognition with change in the size of the labelled training set. The error bars are over 10 different samples of training sets.

Method	UCF-101 small	UCF-101	HMDB-51 small	HMDB-51
Baseline LSTM	63.7	74.5	25.3	42.8
Autoencoder	66.2	75.1	28.6	44.0
Future Predictor	64.9	74.9	27.3	43.1
Conditional Autoencoder	65.8	74.8	27.9	43.1
Conditional Future Predictor	65.1	74.9	27.4	43.4
Composite Model	67.0	**75.8**	29.1	**44.1**
Composite Model with Conditional Future Predictor	**67.1**	**75.8**	**29.2**	44.0

Table 3. Comparison of different unsupervised pretraining methods. UCF-101 small is a subset containing 10 videos per class. HMDB-51 small contains 4 videos per class.

Conclusion & Discussion