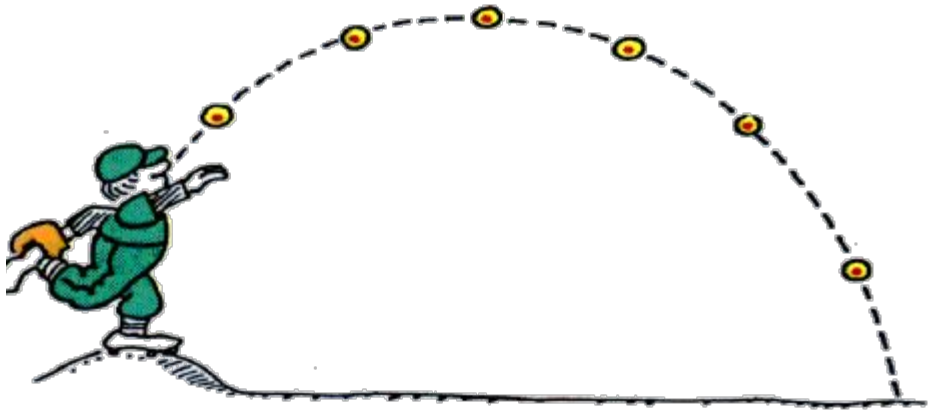
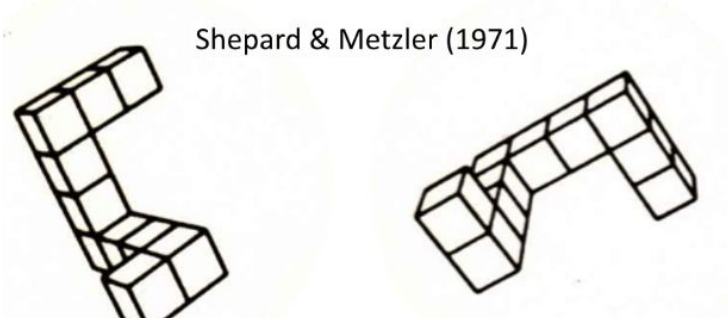
Mental Simulation with Self-Supervised Spatiotemporal Learning

Kevin Tan


Jenga
Jenga
Jenga
Jenga
Jenga
Jenga
Jenga
Jenga
Jenga
Jenga
Jenga



Evidence for Mental Simulation



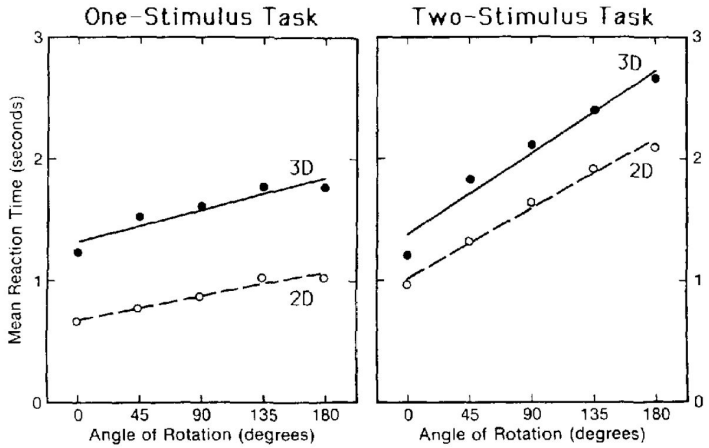
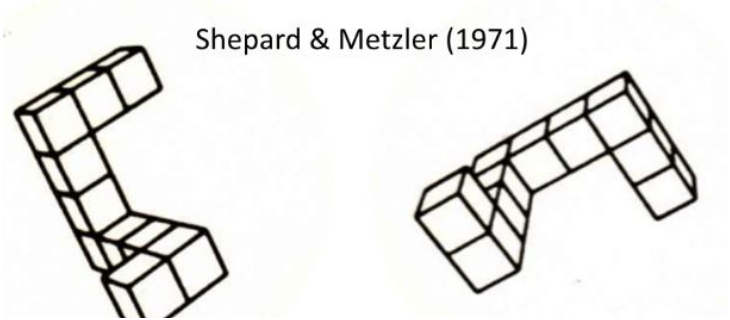


Figure 2. Reaction times of correct positive responses, plotted as a function of angular difference for the one-stimulus task (left panel) and for the two-stimulus task (right panel). Within each panel, data and fitted linear functions are shown separately for the two-dimensional and three-dimensional objects (2D and 3D, respectively).

Evidence for Mental Simulation



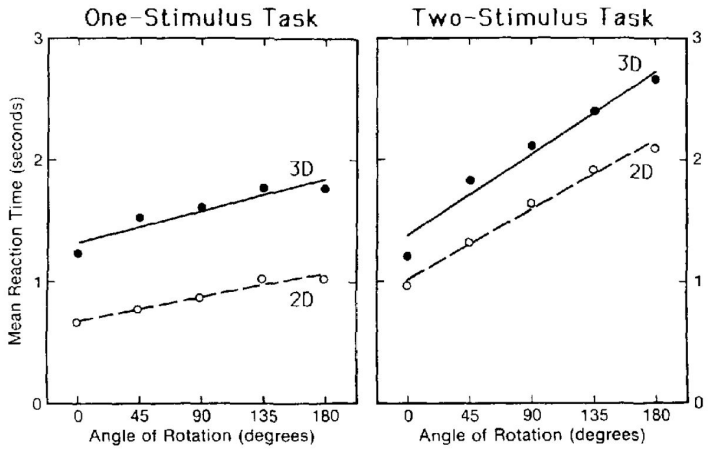


Figure 2. Reaction times of correct positive responses, plotted as a function of angular difference for the one-stimulus task (left panel) and for the two-stimulus task (right panel). Within each panel, data and fitted linear functions are shown separately for the two-dimensional and three-dimensional objects (2D and 3D, respectively).

- **Behaviorism:** study of behavior to identify determinants/causes
- **Cognitivism:** describes mental processes as information processing

Evidence for Mental Simulation

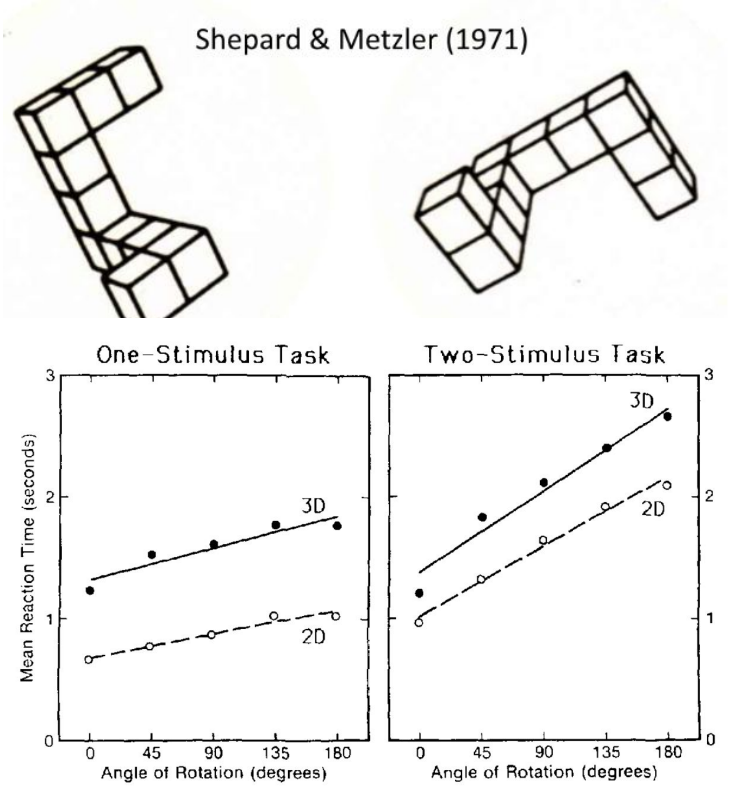


Figure 2. Reaction times of correct positive responses, plotted as a function of angular difference for the one-stimulus task (left panel) and for the two-stimulus task (right panel). Within each panel, data and fitted linear functions are shown separately for the two-dimensional and three-dimensional objects (2D and 3D, respectively).

- **Behaviorism:** study of behavior to identify determinants/causes
- **Cognitivism:** describes mental processes as information processing

- *Q:* How can we study mental simulation at the *computational* level of analysis? *(Marr, 1982)*

Problem Formulation of Spatiotemporal Prediction

$$\widehat{\mathcal{X}}_{t+1}, \ldots, \widehat{\mathcal{X}}_{t+K} = \underset{\mathcal{X}_{t+1}, \ldots, \mathcal{X}_{t+K}}{\arg\max} \; p\left(\mathcal{X}_{t+1}, \ldots, \mathcal{X}_{t+K} | \mathcal{X}_{t-J+1}, \ldots, \mathcal{X}_t\right).$$

Evidence for Mental Simulation

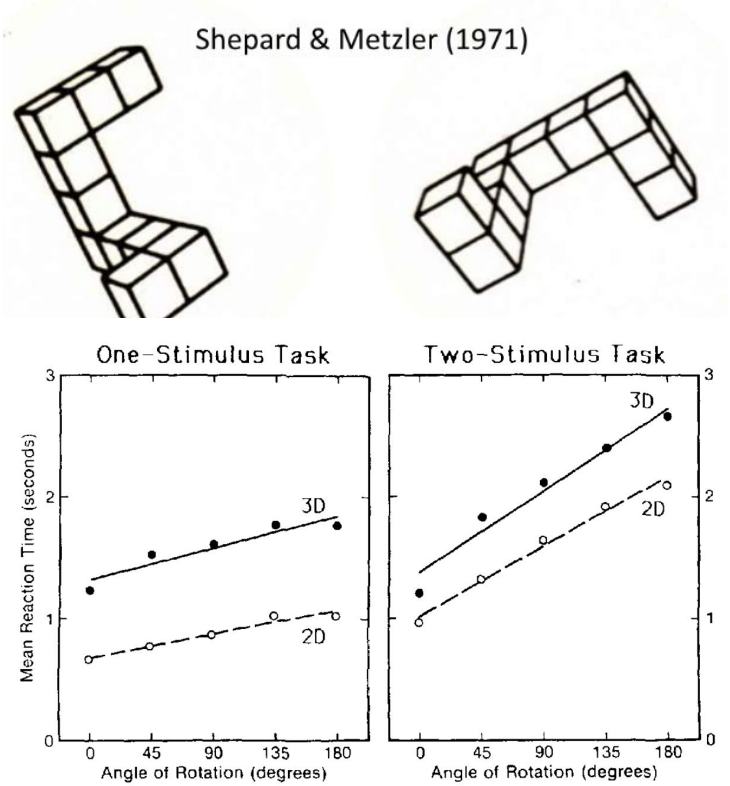
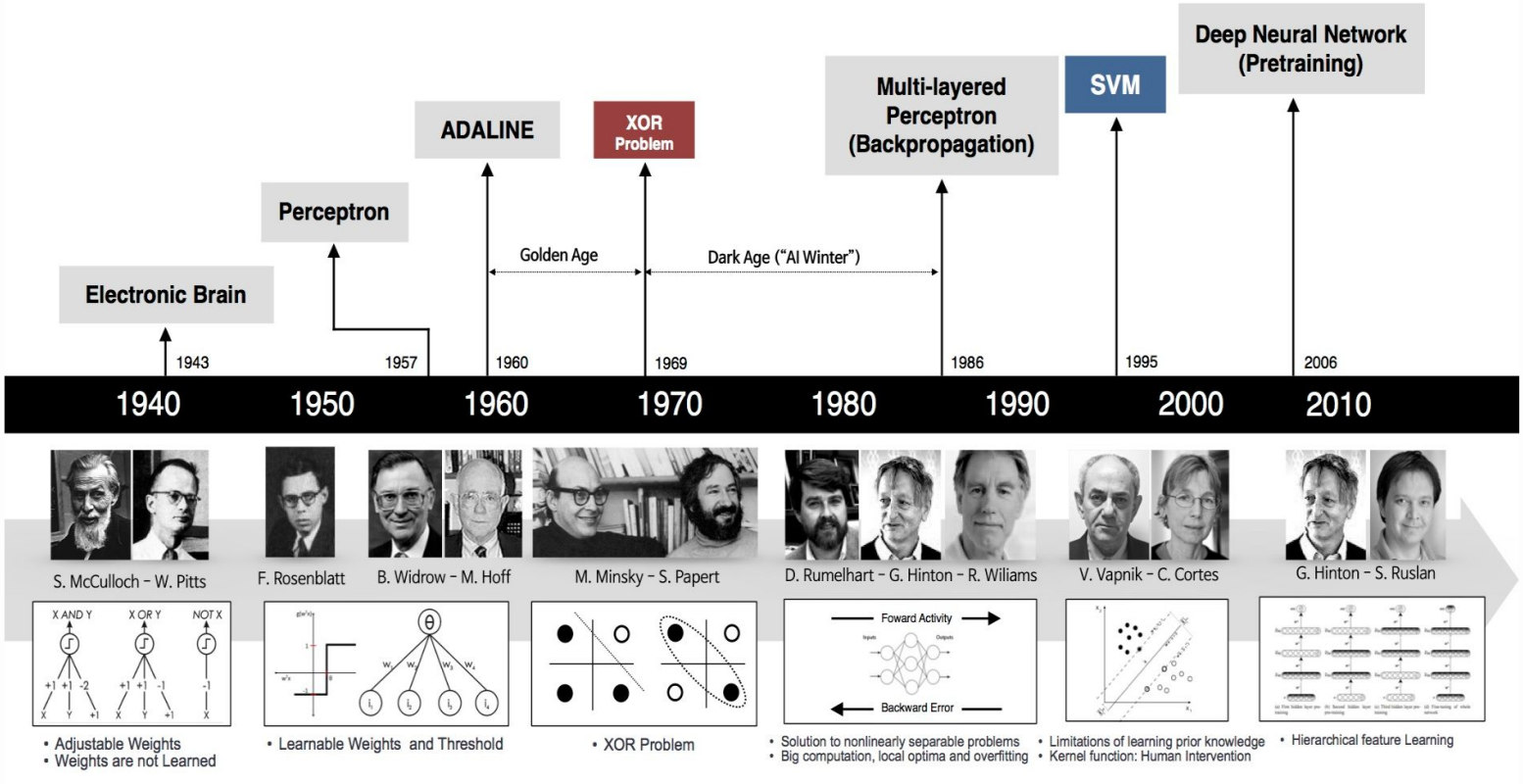


Figure 2. Reaction times of correct positive responses, plotted as a function of angular difference for the one-stimulus task (left panel) and for the two-stimulus task (right panel). Within each panel, data and fitted linear functions are shown separately for the two-dimensional and three-dimensional objects (2D and 3D, respectively).

- **Behaviorism:** study of behavior to identify determinants/causes
- **Cognitivism:** describes mental processes as information processing

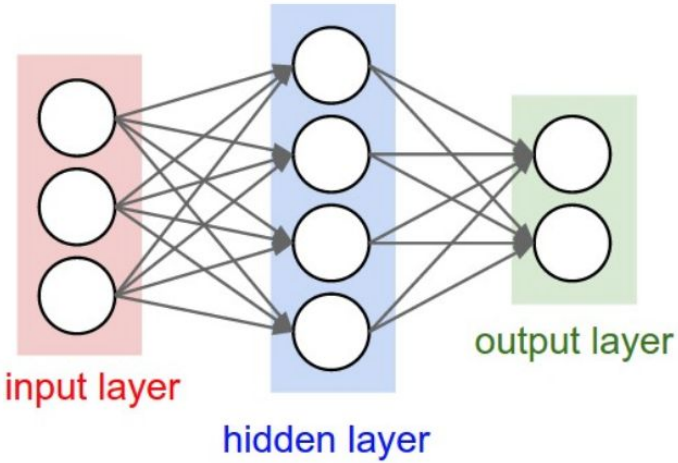
- *Q:* How can we study mental simulation at the *computational* level of analysis? *(Marr, 1982)*
- *A:* Neural networks as a model of the mind (PDP group)

Brief History on the Development of Neural Networks



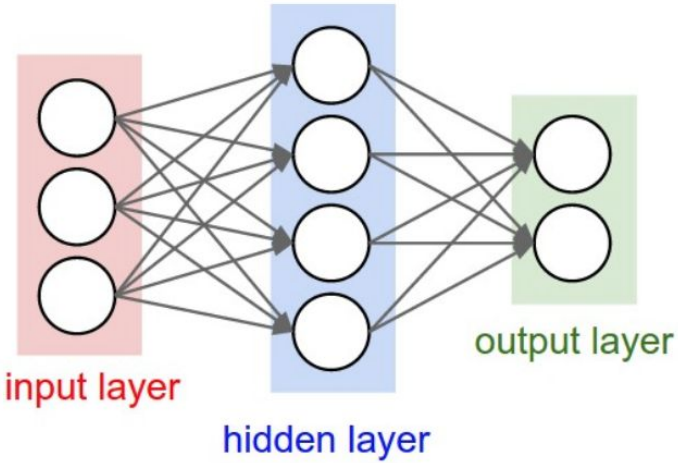
Artificial Neural Networks (ANN)

- Two components:
  - Network architecture (how many neurons and layers)
  - Parameters, or weights, of the connections



Artificial Neural Networks (ANN)

- Two components:
  - Network architecture (how many neurons and layers)
  - Parameters, or weights, of the connections
- Learning procedure:
  - *Feed many examples*
  - *Compute difference between target and actual via an objective function*
  - *Weights are updated via the backpropagation algorithm (Rumelhart et al. 1986)*
  - *Continue until stopping condition*

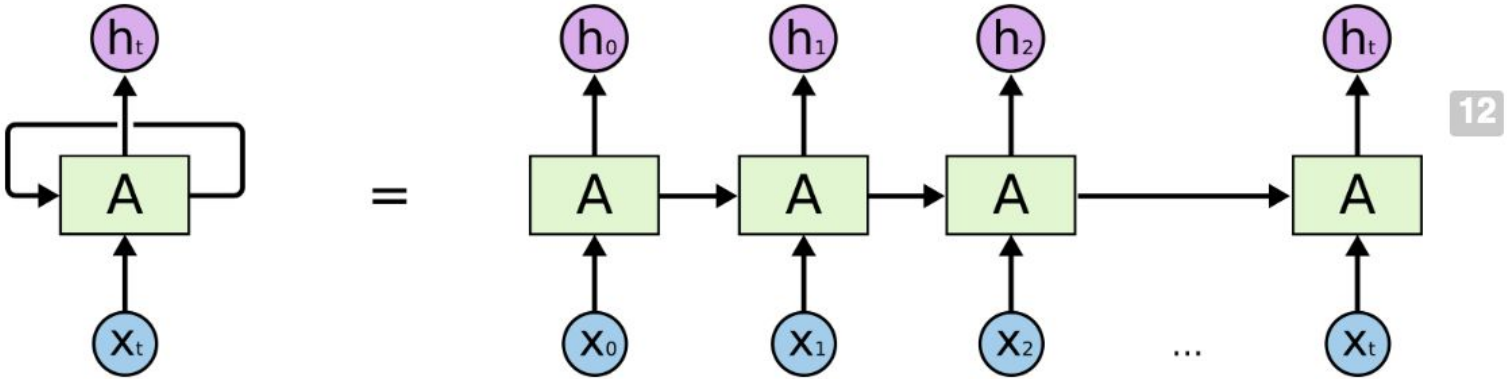


Sequence Modeling: Recurrent Neural Networks (RNN) (1/3)

- “As you read this [sentence], you understand each word based on your understanding of previous words. You don’t throw everything away and start thinking from scratch again. Your thoughts have persistence.”

Sequence Modeling: Recurrent Neural Networks (RNN) (1/3)

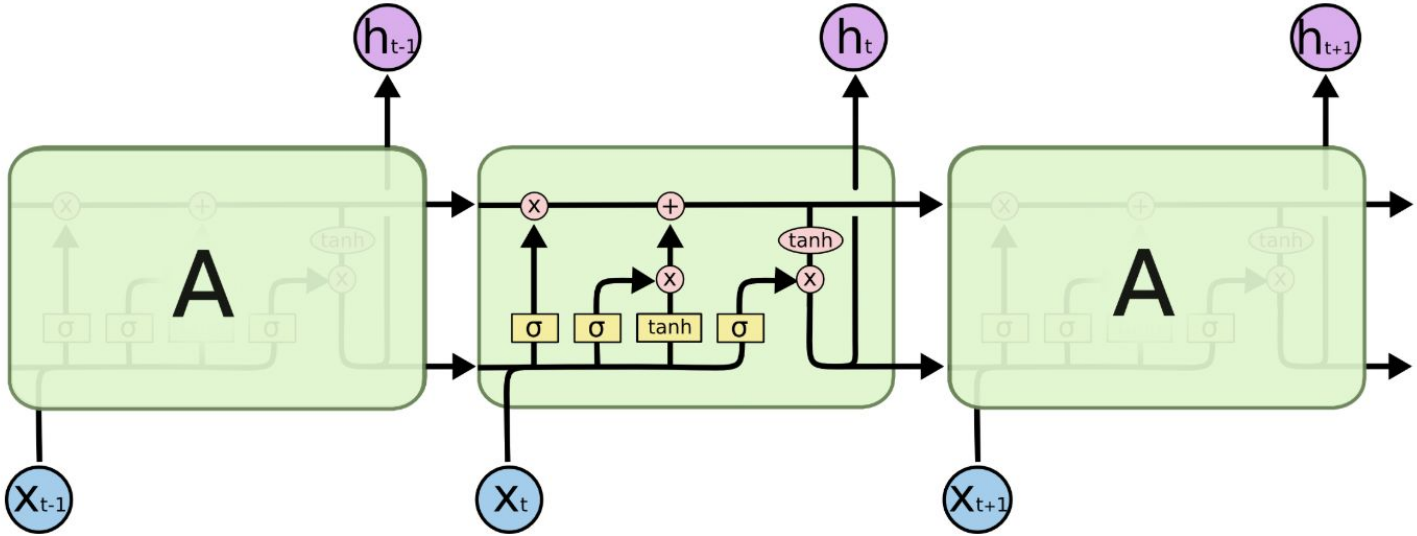
- “As you read this [sentence], you understand each word based on your understanding of previous words. You don’t throw everything away and start thinking from scratch again. Your thoughts have persistence.”



An unrolled recurrent neural network.

https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Sequence Modeling: Long Short-Term Memory (LSTM) (2/3)

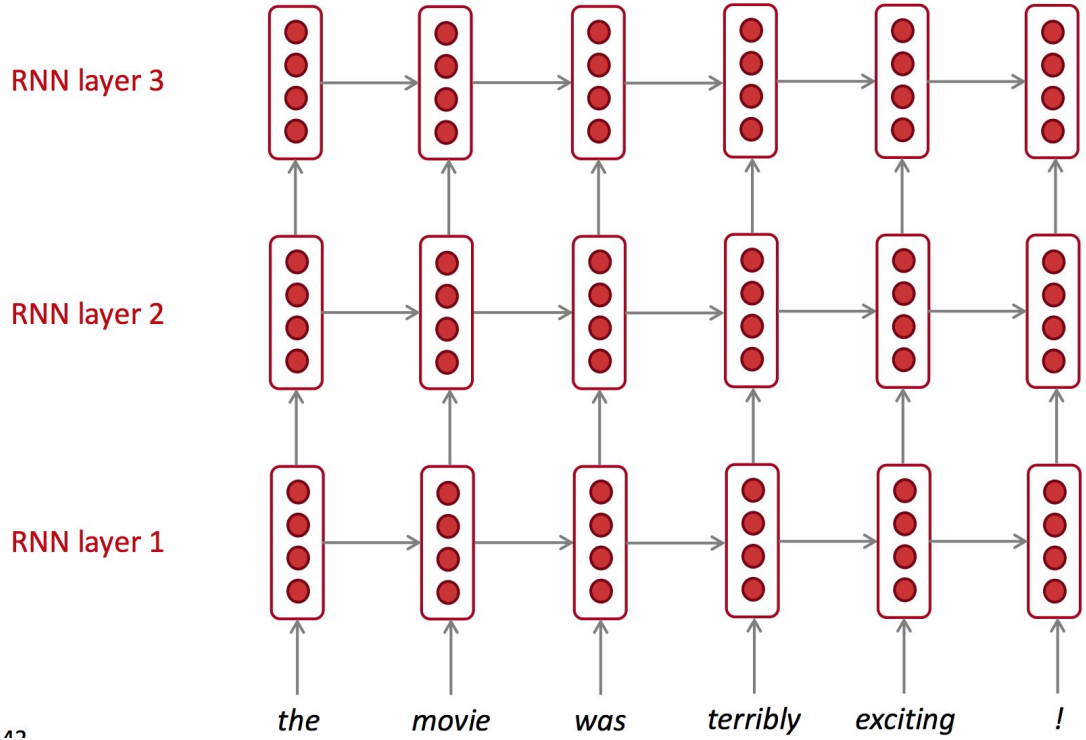


The repeating module in an LSTM contains four interacting layers.

- Introduces four interacting components: cell, input, output, and forget gates
  - Cell gate is responsible for accumulating hidden state information over time
  - Other gates are responsible for regulating information in and out

https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Sequence Modeling: Stacked LSTMs (3/3)



The hidden states from RNN layer *i* are inputs to the RNN at layer *i+1*

Encoding Spatial Information: Convolutional Layers

- Two distinguishing features:
  - **Local connectivity**: connections only exist between local regions (visual receptive field)
  - **Shared weights:** assume that any feature learned at one spatial location is useful at another spatial location

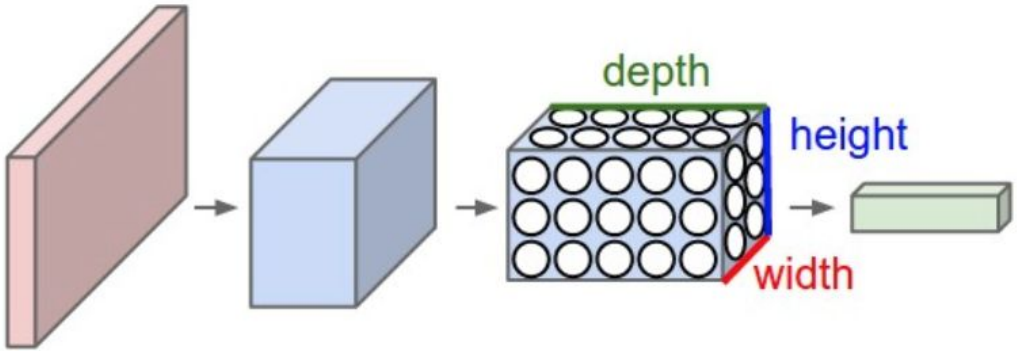
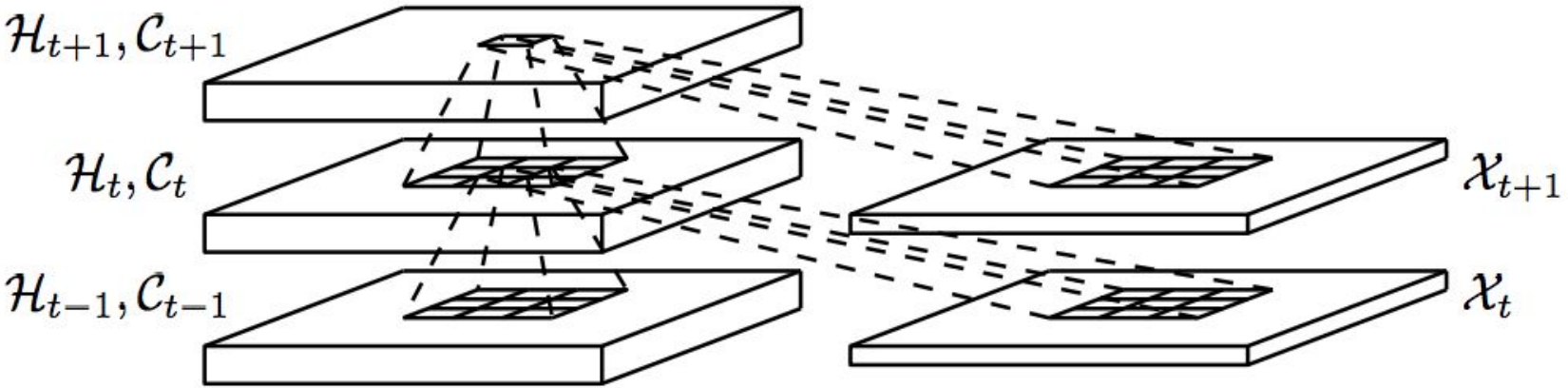


Figure: A ConvNet arranges its neurons in three dimensions (width, height, depth), as visualized in one of the layers. Every layer of a ConvNet transforms the 3D input volume to a 3D output volume of neuron activations. In this example, the red input layer holds the image, so its width and height would be the dimensions of the image, and the depth would be 3 (Red, Green, Blue channels).

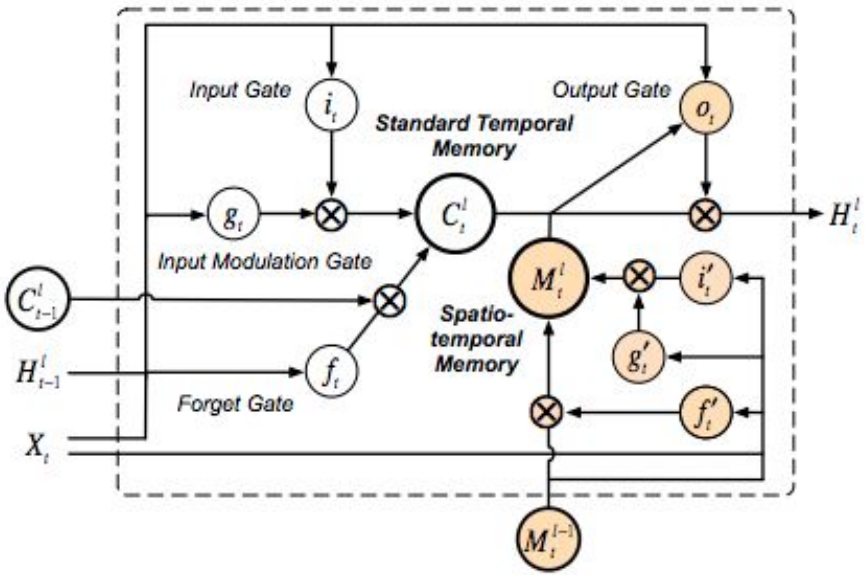
Spatiotemporal Video Prediction (1/3)

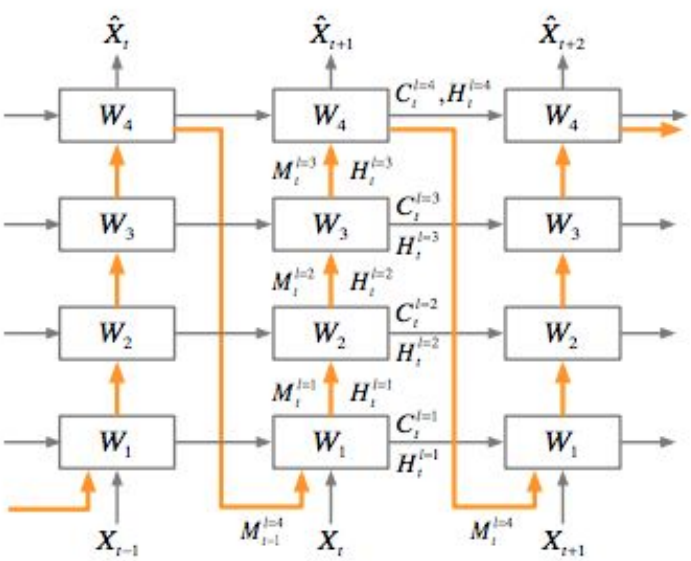
1. Convolutional LSTM (ConvLSTM) (Shi, Xingjian et al, 2015)
   - Encodes spatial information into tensors via convolution
   - Hidden states are 3D instead of 2D



Spatiotemporal Video Prediction (2/3)

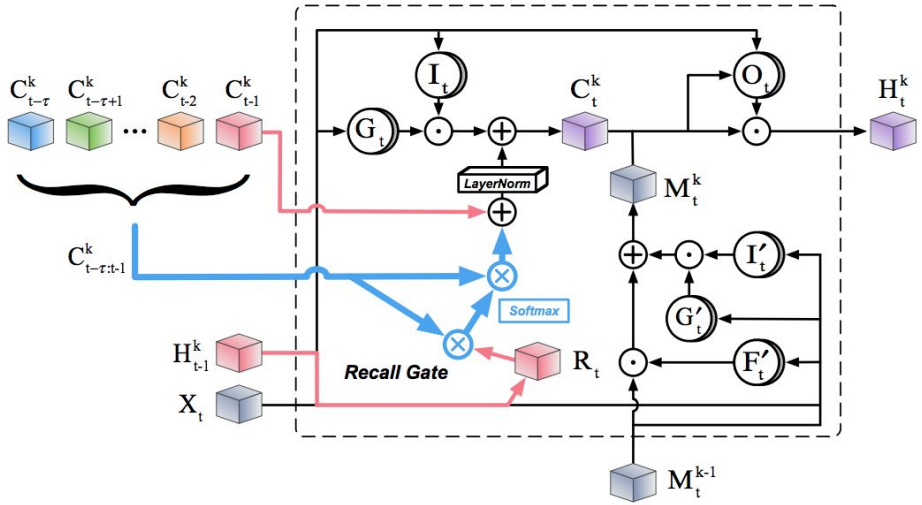
1. Convolutional LSTM (ConvLSTM) (Shi, Xingjian et al, 2015)
2. PredRNN (Wang, Yunbo et al. 2017)
    - Memory flow in zigzag direction

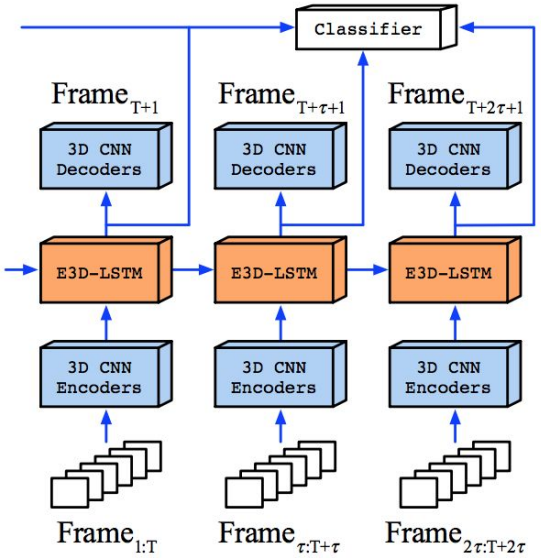




Spatiotemporal Video Prediction (3/3)

1. Convolutional LSTM (ConvLSTM) (Shi, Xingjian et al, 2015)
2. PredRNN (Wang, Yunbo et al. 2017)
3. Eidetic 3D LSTM (E3D-LSTM) (Wang, Yunbo et al. 2019)
    - 3D-Conv inside RNN cell + attention mechanism





Design Choices

Design Choices

1. 3D-Conv vs 2D-Conv inside RNN cell

Design Choices

1. 3D-Conv vs 2D-Conv inside RNN cell
2. Residual connections architecture in stacked LSTMs

Design Choices

1. 3D-Conv vs 2D-Conv inside RNN cell
2. Residual connections architecture in stacked LSTMs
3. Balance between L1 and L2 norm components in loss function

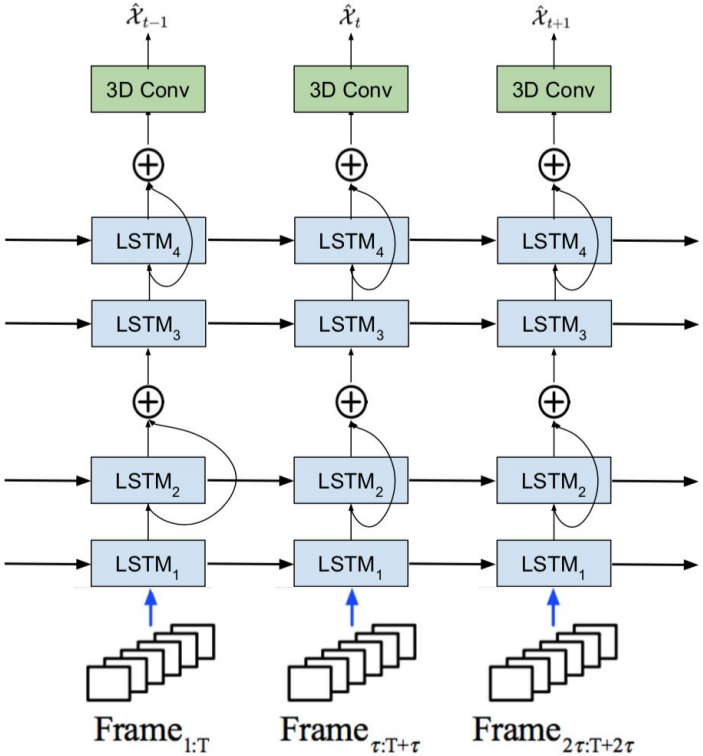


FIGURE 3.1: Architectural design of our approach: E3D-LSTM with deep residual connections. 4 E3D-LSTMs are stacked in the vertical direction with residual connections between before layer 3 and before the decoder. $\oplus$ denotes the addition operator.

Experiment 1: Moving MNIST

- 2 handwritten digits bouncing inside 64x64 grid
- 10 → 10: Predict 10 future frames given 10 previous frames (with some similarity metric)
- 10,000/3,000/5,000 sequences for train/valid/test

Experiment 1: Moving MNIST

Model	SSIM	MSE
ConvLSTM (Shi et al., 2015)	0.713	96.5
DFN (Brabandere et al., 2016)	0.726	89.0
CDNA (Finn, Goodfellow, and Levine, 2016)	0.728	84.2
FRNN (Oliu, Selva, and Escalera, 2018)	0.819	68.4
VPN Baseline (Kalchbrenner et al., 2016)	0.870	64.1
PredRNN (Wang et al., 2017)	0.869	56.5
PredRNN++ (Wang et al., 2018)	0.885	46.3
E3D-LSTM (Wang et al., 2019)	0.910	41.3
E3D-LSTM Finetuned with l_2 only	0.9199	**39.5**
E3D-LSTM Finetuned with Residuals and $l_1 + l_2$	**0.9219**	42.5

TABLE 4.1: Results on the Moving MNIST Dataset for the $10 \rightarrow 10$ task. Higher SSIM or lower MSE scores indicate better results. *Top:* Previous state-of-the-art models. *Bottom:* Our experiments finetuned from pretrained weights demonstrating effectiveness of residual connections.

Model	SSIM	MSE
E3D-LSTM Baseline	0.880	69.8
E2D-LSTM with [1,5,5] kernel	0.862	75.0
E3D-LSTM with Residuals	0.890	59.1

TABLE 4.2: Our experiments trained from scratch.

Experiment 1: Moving MNIST

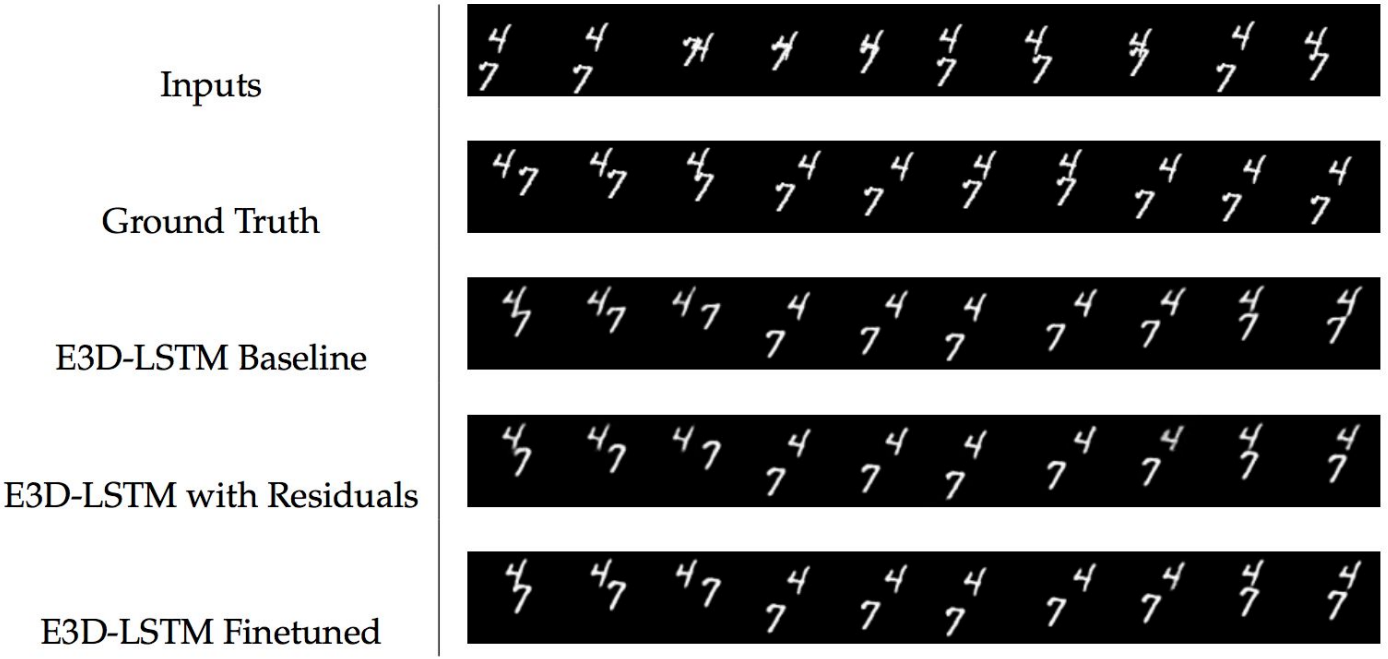
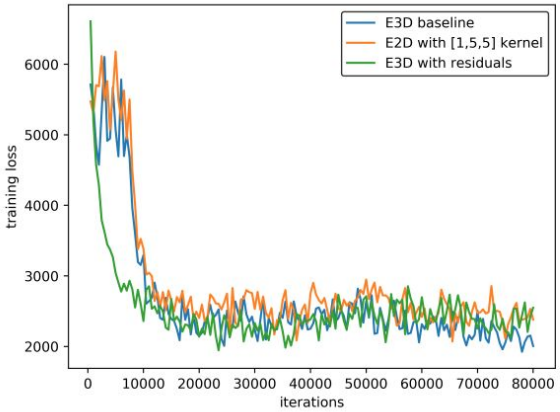
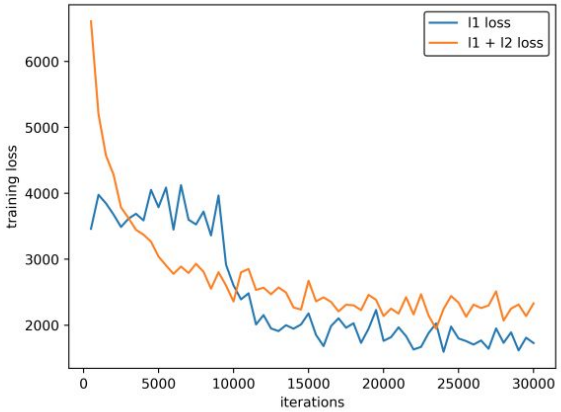


FIGURE 4.2: Video prediction examples from Moving MNIST dataset with '4' and '7'. E3D-LSTM Finetuned with Residuals and $l_1 + l_2$ loss demonstrates slightly improved qualitative image clarity.

Experiment 1: Moving MNIST





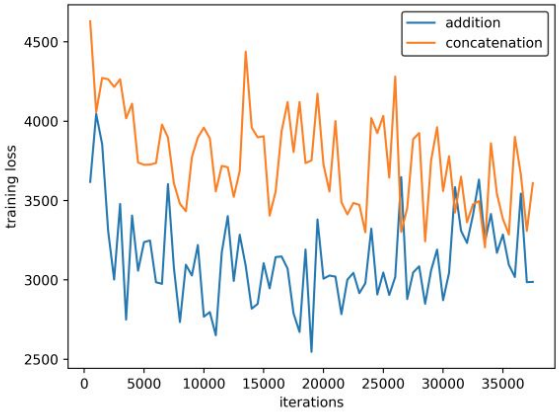
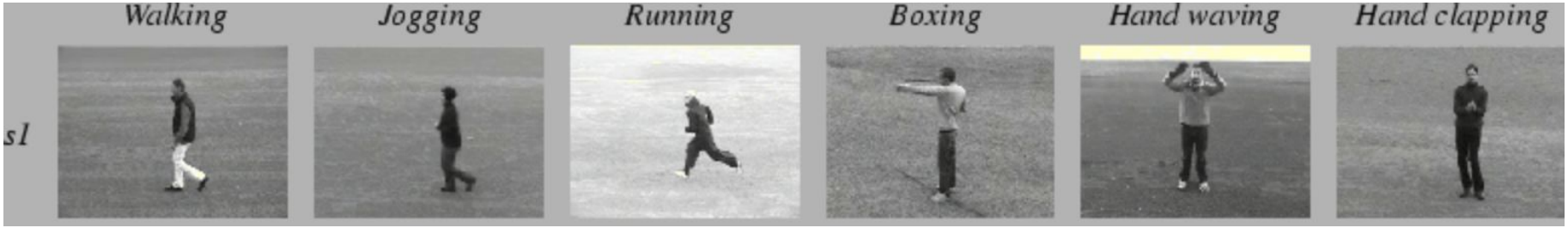


FIGURE 4.1: *Left*: Comparison of E3D baseline, E2D, and E3D with residual connections. Residuals slightly improve training efficiency and result in faster convergence. *Middle*: E3D with residual connections comparing equally weighted l_1 + l_2 loss vs. only l_1 loss. l_1 loss alone improves final training loss. *Right*: Comparing addition and concatenation operators for residual connections.

Experiment 2: KTH Action Dataset



- 25 individuals performing 6 types of actions
- Each video clip lasts 4 sec
- Resized to 128x128
- 1-16/17-25 train/test
- Task: 10 → 20 (training), extend to 10 → 40 (testing)

Experiment 2: KTH Action Dataset

Model	PSNR	SSIM
ConvLSTM (Shi et al., 2015)	23.58	0.712
DFN (Brabandere et al., 2016)	27.26	0.794
FRNN (Oliu, Selva, and Escalera, 2018)	26.12	0.771
PredRNN (Wang et al., 2017)	27.55	0.839
PredRNN++ (Wang et al., 2018)	28.47	0.865
E3D-LSTM (Wang et al., 2019)	29.31	0.879
E3D-LSTM Baseline	27.73	0.854
E2D-LSTM with [1,5,5] kernel	23.30	0.838
E3D-LSTM with Residuals	27.61	0.863
E3D-LSTM Finetuned with Residuals and $l_1 + l_2$	**29.67**	**0.881**

TABLE 4.2: Results on the KTH Action dataset for the $10 \rightarrow 20$ task. Higher PSNR and SSIM scores indicate better performance. *Top:* Previous state-of-the-art models and results. *Middle:* Our experiments trained from scratch. *Bottom:* Our experiments finetuned from pretrained weights demonstrating effectiveness of residual connections.

Experiment 2: KTH Action Dataset

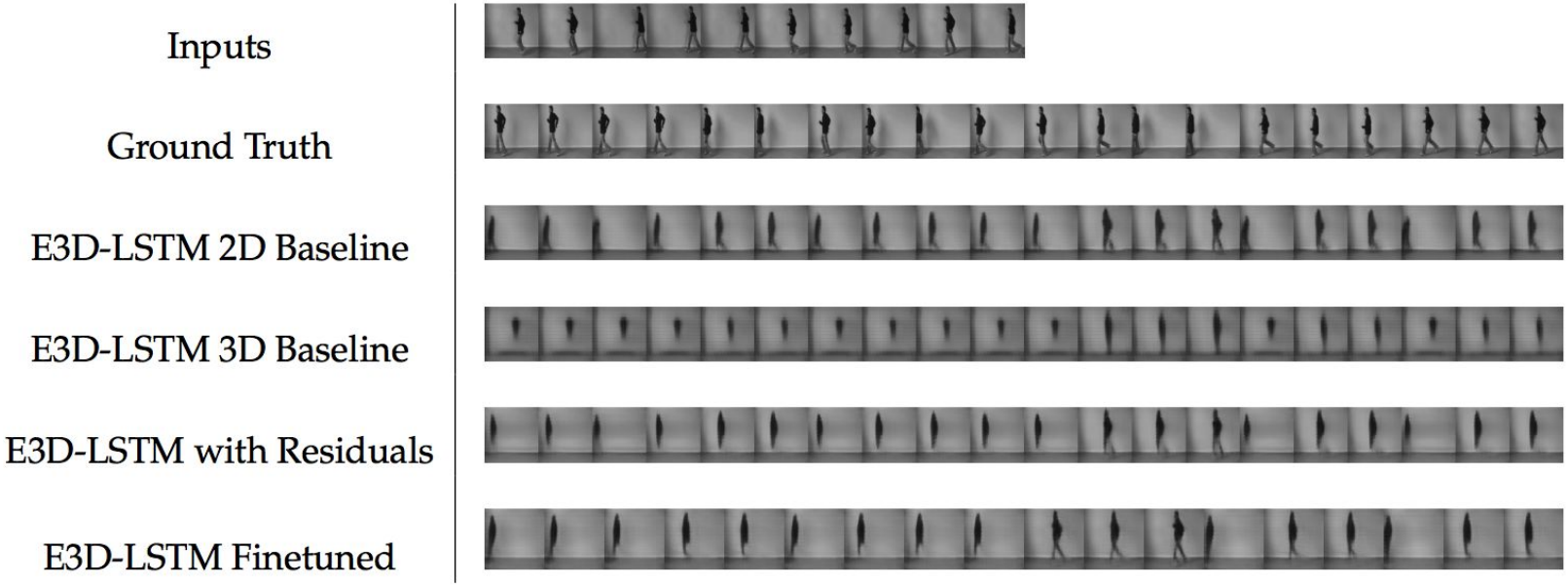


FIGURE 4.3: Video prediction examples from KTH Action dataset on the $10 \rightarrow 20$ task. E3D-LSTM Finetuned with Residuals and $l_1 + l_2$ loss demonstrates slightly improved qualitative image clarity.

Summary

1. 3D-Conv inside the RNN cell is important for modelling spatiotemporal information for long-term video prediction
2. Residual connections in stacked-LSTM architectures exhibit small empirical improvements due to their ability to maintain crucial spatial information from earlier memory layers
3. Balance between L1 and L2 is crucial in the training process

Analog vs Propositional? (Kosslyn et al. 2006)

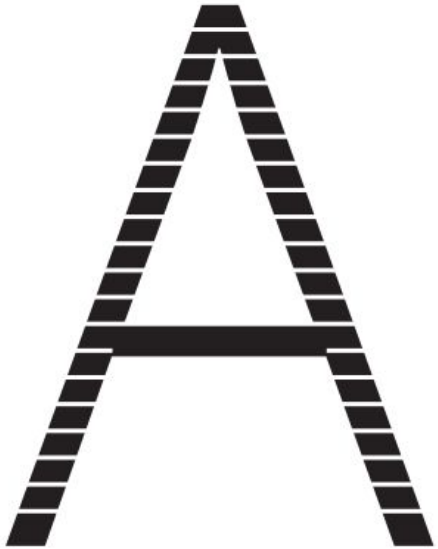


Figure 1: **Analog** representation of the letter “A”

Figures adapted from Kosslyn et al. (2006). The Case for Mental Imagery

Analog vs Propositional? (Kosslyn et al. 2006)

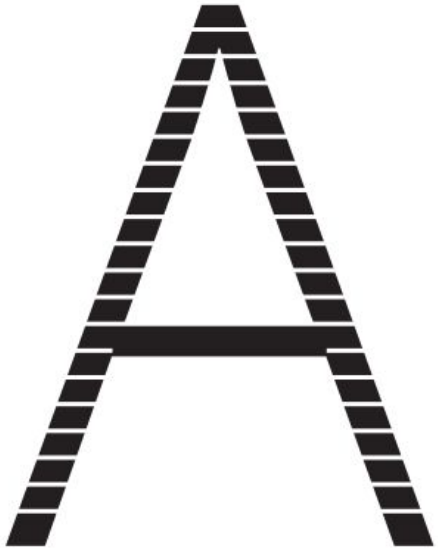


Figure 1: **Analog** representation of the letter “A”

“two diagonal lines that meet at the top, joined halfway down by a short horizontal segment.”

Figure 2: **Propositional** representation of the letter “A”

Figures adapted from Kosslyn et al. (2006). The Case for Mental Imagery

Acknowledgments

- Professor Zhuowen Tu, Kwonjoon Lee, Yifan Xu, Wenlong Zhao, Haoyu Dong, and the rest of the Machine Learning, Perception, and Cognition Lab
- Professor Virginia de Sa, Shuai Tang, Vijay Veerabadran
- Shawn Hsu, Luca Pion-Tonachini
- Halıcıoğlu Data Science Institute Undergraduate Scholarship
- Hong and Jamie Tan