A Neurally-Inspired Hierarchical Prediction Network for Spatiotemporal Sequence Learning and Prediction

Jielin Qiu [1 2] **Ge Huang** [2] **Tai Sing Lee** [2 3]

Abstract

In this paper we developed a hierarchical network model, called Hierarchical Prediction Network (HPNet), to understand how spatiotemporal memories might be learned and encoded in the recurrent circuits in the visual cortical hierarchy for predicting future video frames. This neurally inspired model operates in the analysis-by-synthesis framework. It contains a feed-forward path that computes and encodes spatiotemporal features of successive complexity and a feedback path for the successive levels to project their interpretations to the level below. Within each level, the feed-forward path and the feedback path intersect in a recurrent gated circuit, instantiated in a LSTM module, to generate a prediction or explanation of the incoming signals. The network learns its internal model of the world by minimizing the errors of its prediction of the incoming signals at each level of the hierarchy. We found that hierarchical interaction in the network increases semantic clustering of global movement patterns in the population codes of the units along the hierarchy, even in the earliest module. This facilitates the learning of relationships among movement patterns, yielding state-of-the-art performance in long range video sequence predictions in the benchmark datasets. The network model automatically reproduces a variety of prediction suppression and familiarity suppression neurophysiological phenomena observed in the visual cortex, suggesting that hierarchical prediction might indeed be an important principle for representational learning in the visual cortex.

[1]Shanghai Jiao Tong University, Shanghai, China [2]Center for the Neural Basis of Cognition, Neuroscience Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA [3]Computer Science Department, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. Correspondence to: Tai Sing Lee <tai@cnbc.cmu.edu>.

1. Introduction

While the hippocampus is known to play a critical role in encoding episodic memories, the storage of these memories might ultimately rest in the sensory areas of the neocortex (McClelland & McNaughton, 1999). Indeed, a number of neurophysiological studies suggest that neurons throughout the hierarchical visual cortex, including those in the early visual areas such as V1 and V2, might be encoding memories of object images (Huang et al., 2018) and of visual sequences in cell assemblies (Yao et al., 2007; Han et al., 2008; Xu et al., 2012; Cooke & Bear, 2014; 2015). As specific priors, these memories, together with the generic statistical priors encoded in receptive fields and connectivity of neurons, serve as internal models of the world for the prediction of incoming visual experiences. Learning to predict incoming visual signals has also been proposed as a self-supervised learning paradigm for representation learning in recurrent neural networks, in which the discrepancies between the model's prediction and the incoming signals can be used to train a network using backpropagation, without the need of labeled data (Elman, 1990; Mathieu et al., 2015; Villegas et al., 2017; Srivastava et al., 2015; O'Reilly et al., 2014; Lee, 2015).

In computer vision, a number of hierarchical recurrent neural network models, notably PredNet (Lotter et al., 2016) and PredRNN++ (Wang et al., 2018), have been developed for video prediction with state-of-the-art performance. PredNet, in particular, was inspired by the neuroscience principle of predictive coding (Mumford, 1991; Rao & Ballard, 1999; Lee, 2015; Dijkstra et al., 2017; Friston, 2018). This model learns a LSTM (long short-term memory) model at each level to predict the errors made in an earlier level of the hierarchical visual system. Only the prediction errors are propagated forward to the next level. Because the error representations are sparse, the computation of PredNet is very efficient. However, the model builds a hierarchical representation to predict errors, rather than a hierarchy to predict features of successive complexities and abstraction. The lack of a compositional feature hierarchy hampers its ability in long range video predictions.

Here, we proposed an alternative hierarchical network architecture. The proposed model, HPNet (Hierarchical Predic-

tion Network), contains a fast feedforward path, instantiated currently by a deep convolutional neural network (DCNN) that learns a representational hierarchy of features of successive complexity, and a feedback path that brings a higher order interpretation to influence the computation a level below. The two paths intersect at each level through a long short term memory (LSTM) unit to generate a hypothesis of the current state of the world and make a prediction of the incoming bottom-up input. The LSTM, as a gated recurrent circuit performs this prediction by integrating top-down, bottom-up, and horizontal information. The prediction error at each level is fed back to influence the interpretation of the LSTMs at the same level as well as the level above.

To facilitate the learning of relationships among movement patterns, the proposed HPNet processes data in the unit of a spatiotemporal block that is composed of a sequence of video frames, rather than in a frame by frame manner, as in PredNet and PredRNN++. We used a 3D convolutional LSTM at each level of the hierarchy to process these spatiotemporal blocks of signals (Choy et al., 2016), which is a key factor underlying HPNet's better performance in long range video prediction.

We will first demonstrate HPNet's effectiveness in predictive learning and in long range video prediction. Then we will show that units in HPNet exhibit image sequence prediction suppression and image familiarity suppression effects that have been observed in both the early visual areas and inferotemporal cortex of macaque monkeys in neurophysiological experiments. These findings suggest that predictive self-supervised learning might indeed be an important strategy for representation learning in the visual cortex, and that HPNet is a viable computational model for understanding and modeling the computations in the hierarchical visual system.

2. Related works

HPNet integrates ideas of predictive coding (Mumford, 1992; Rao & Ballard, 1999; Lotter et al., 2016) and associative coding (McClelland & Rumelhart, 1985; Grossberg, 1987). It differs from the predictive coding models (Rao & Ballard, 1999; Lotter et al., 2016) in that it learns a hierarchy of feature representations in the feedforward path to model features in the world as in normal deep convolutional neural networks (DCNN). PredNet, on the other hand, builds a hierarchy to model successive prediction errors of its own prediction of the world. PredNet is efficient because its convolution is operated on sparse prediction error codes, but we believe lacking a hierarchical representation of features limits its ability to model relationships among more global and abstract movement concepts for longer range video prediction. We believe that having a fast bottom-up hierarchy of spatiotemporal features of successive scale and abstraction will allow the system to see further into the future and make better predictions.

A key difference between the genre of predictive learning models (HPNet, PredNet) and the earlier predictive coding models implemented by Kalman filters (Rao & Ballard, 1999) or associative coding models implemented by interactive activation (McClelland & Rumelhart, 1985; Grossberg, 1987) is that the synthesis of expectation is not done simply by the feedback path, via weight matrix multiplication, but by local gated recurrent circuits at each level. This key feature makes this genre of predictive learning models more powerful and competent in solving real computer vision problems.

The idea of predictive learning, using incoming video frames as self-supervising teaching labels to train recurrent networks, can be traced back to Elman (1990). Recently, there has been active exploration of self-supervised learning in computer vision (Palm, 2012; O'Reilly et al., 2014; Goroshin et al., 2015; Srivastava et al., 2015; Patraucean et al., 2015; Vondrick et al., 2016), particularly in the area of video prediction research (Mathieu et al., 2015; Kalchbrenner et al., 2017; Xu et al., 2018; Oh et al., 2015; Villegas et al., 2017; Lee et al., 2018; Wichers et al., 2018). The large variety of models can be roughly grouped into three categories: autoencoders, DCNN, and hierarchy of LSTMs. Some models also involve feedforward and feedback paths, where the feedback paths have been implemented by deconvolution, autoencoder networks, LSTM or adversary networks (Finn et al., 2016; Lotter et al., 2016; Wang et al., 2017; 2018; 2019). Some other models, such as variational autoencoders, allowed multiple hypotheses to be sampled (Babaeizadeh et al., 2017; Denton & Fergus, 2018).

PredRNN++ (Wang et al., 2018) is the state-of-the-art hierarchical model for video prediction at the writing of this paper. It consists of a stack of LSTM modules, with the LSTM at one level providing feedforward input directly to the LSTM at the next level, and ultimately predicting the next video frame at its top level. Thus, its hierarchical representation is more similar to an autoencoder, with the intermediate layers modeling the most abstract and global spatiotemporal memories of movement patterns and the subsequent layers representing the unfolding of the feedback path into a feedforward network with its top-layer's output providing the prediction of the next frame. PredRNN++ does not claim neural plausibility, but it offers state-of-the-art performance for benchmark performance evaluation, with documented comparisons to other approaches. Our main objective, however, is to demonstrate the competency of a deep learning realization of the analysis-by-synthesis framework for modeling hierarchical cortical processing (Mumford, 1992; Ullman, 1995; Rao & Ballard, 1999; Lee & Mumford, 2003; Dayan et al., 1995; Kersten & Yuille, 2003), rather

than simply beating the rapidly evolving state-of-the-art performance in video prediction.

A number of recent studies (Nayebi et al., 2018; Wen et al., 2018) have demonstrated that deep convolutional neural networks with recurrent feedback loops can achieve similar performance in object recognition as that of very deep networks but with significantly fewer layers and parameters. However, these models relied on supervised learning on static images as labelled data. HPNet complements these studies by exploring the learning of hierarchical recurrent organization based on self-supervised predictive learning on videos.

3. Hierarchical Prediction Network

3.1. Cortical Module

HPNet is composed of a stack of Cortical Modules (CM). Each CM can be considered as a visual area along the ventral stream of the primate visual system, such as V1, V2, V4 and IT. We used four Cortical Modules in our experiment. The network contains a feedforward path that is realized in a deep convolutional neural network (DCNN), a stack of Long Short Term Memory (LSTM) modules that link the feedforward path and the feedback path together.

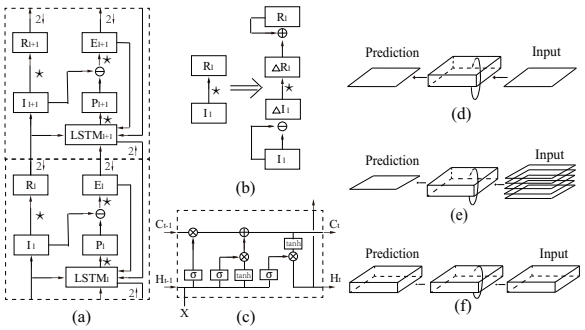


Figure 1. (a) Two successive layers of Cortical Modules in our hierarchical network. The input I_1 at the bottom level is a spatiotemporal block of video frames. The $\star$ notation means a convolution along that path. $2\uparrow$ indicates up-sampling or expansion operation. $2\downarrow$ means down-sample or reduction in resolution. $\ominus$ indicates comparator or subtraction operation; (b) The DCNN analysis path is implemented in a sparsified convolution scheme to speed up bottom-up processing; (c) Detailed structure of the standard LSTM used. C_t is the internal state, and H_t is the output. X is external input, which contains multiple sources in our model. (d) Frame-by-frame scheme; (e) Block-by-frame scheme; and (f) Block-by-block scheme, where left and right part indicates output and input respectively with the middle indicating 2D or 3D convolution LSTM.

Figure 1a shows two CMs stacked on top of each other. The feedforward path performs convolution (indicated by $\star$) on the input spatiotemporal block I_l with a kernel to produce R_l, where l indicates CM level. R_l is then down-sampled to provide the input I_{l+1} for CM_{l+1} for another round of convolution in the feedforward path. I_{l+1} also goes into LSTM_{l+1} (LSTM_{l+1} in CM_{l+1}). In each CM_l level, the bottom-up input I_l is compared with the prediction P_l generated from the interpretation output H_l of LSTM_l. The prediction error signal is transformed by a convolution into E_l, which is fed back to both LSTM_l and LSTM_{l+1} to influence their generation of new hypotheses H_l and H_{l+1}. To make the timing relationship clear in Algorithm 1, we use k to index a spatiotemporal block in a block sequence, which is extracted from the video input x_t with a stride that could vary from 1 to d, where d is the number of video frames contained in a block. LSTM_l at step k integrates the bottom-up feature input R_{l-1}^k, the top-down feedback of the higher CM's LSTM's output H_{l+1}^k, and the prediction errors E_{l-1}^k and E_l^{k-1} to generate new hypothesis output H_l^k, which is then transformed into a new prediction P_l^k (see details in Algorithm 1).

3.2. Sparse Convolution

The feedforward DCNN path in Figure 1a runs much faster if the input to each convolution layer is made sparse, as shown in Pan et al. (2018). In video processing, a scheme has been proposed by Liu et al. (2017); Dave et al. (2017); Pan et al. (2018) to sparsify the input of a convolution layer by performing convolution on the difference $\Delta I_l^k = I_l^k - I_l^{k-1}$ between two consecutive blocks. The resulting ΔR_l^k is added back to the representation of the last time block R_l^{k-1} to recover the representation at the current block R_l^k. This allows the network to maintain a full higher order representation R at all times in the next layer while enjoying the benefit of fast computation on sparse input. In their scheme (Pan et al., 2018), the first block $I^{k=0}$ was convolved with a set of dense convolution kernels and then the subsequent frames were convolved with a set of sparse convolution kernels. For parsimony and neural plausibility, we used the same set of sparse kernels for processing both the first full frame and the subsequent temporal-difference frames, at the expense of incurring some inaccuracy in our prediction of the first few frames.

3.3. Spatiotemporal Blocks and 3D convolution

The input data of our network model is a sequence of video frames or a spatiotemporal block. For our implementation, each block contains 5 video frames. If we consider that each frame corresponds roughly to 25 ms, this would translate into 125 ms in actual time, in the range of the length of temporal kernel of a cortical neuron. Our convolution kernel is in three dimension, processing the video by spatiotemporal blocks. The block could slide in time with a temporal stride of one frame or a stride as large as the length of the block

d. The LSTM is a 3D convolutional LSTM (Choy et al., 2016; Wang et al., 2019) because of 3D convolution and spatiotemporal blocks. Convolution LSTM (Shi et al., 2015), in which Hadamard product in LSTM is replaced by a convolution, has greatly improved the performance of LSTM in many applications. PredNet and PredRNN processed video sequences frame by frame, as shown in Figure 1d. We experimented with different data representational schemes. In the Frame-to-Frame (F-F) scheme, an input frame is used to generate one predicted future frame (Figure 1d). In the Block-to-Frame (B-F) scheme (Figure 1e), a block of input frames is used to generate one predicted future frame, as in (Wang et al., 2019). This approach is time consuming, but provides more accurate near-range predictions. For longer-range predictions, we found using a spatiotemporal block to predict a spatiotemporal block, i.e. the Block-to-Block (B-B) scheme (Figure 1f), to be the most effective, because it allows the LSTM to learn the relationship between movement segments in the sequences. The details of our 3D convolutional LSTM is specified in Appendix.

3.4. Training and Loss Function

The entire network is trained by minimizing a loss function which is the $L2$ weighted sum of the prediction errors of all the Cortical Modules (CM),

$$L_{loss} = \sum_k \lambda_k \sum_l \frac{\lambda_l}{n_l} \sum_{n_l} (I_l^k - P_l^k)^2 \tag{1}$$

where k indexes the spatiotemporal block sequence, l the CM level, and n_l the number of units in that level; λ_k and λ_l are weighting factors for time step and CM level, respectively. I_l^k is k^{th} spatiotemporal block input to the CM at level l, and P_l^k is the prediction at that level, following the variables' notations above as well as in Figure 1.

$$I_l^k = \begin{cases} MaxPool(ReLU(R_{l-1}^k)) & l > 1 \\ x_t & l = 1 \end{cases} \tag{2}$$

$$P_l^k = \begin{cases} ReLU(conv(H_l^k)) & l > 1 \\ SATLU(ReLU(conv(H_l^k))) & l = 1 \end{cases} \tag{3}$$

$$H_l^k = 3DconvLSTM(H_l^{k-1}, E_l^{k-1}, MaxPool(ReLU(R_{l-1}^k, E_{l-1}^k)), upsample(H_{l+1}^k)) \tag{4}$$

$$\Delta R_l^k = spconv(I_l^k - I_l^{k-1}), \;\; E_l^k = spconv(I_l^k - P_l^k) \tag{5}$$

$$R_l^k = R_l^{k-1} + \Delta R_l^k \tag{6}$$

where x_t is the video input sequence, H_l^k is the output of LSTM, SATLU is a saturating non-linearity set at the maximum pixel value: SATLU(x; p_{max}):= min(p_{max}, x), $spconv$ indicates sparse convolution. The algorithm is shown in Algorithm 1.

Algorithm 1 The algorithm of our model

Require: $I_1^k \leftarrow x_t$
for $t = 1$ to T **do**
  for $l = L$ to 1 **do** ▷ Top-down updates
    if $l = L$ **then**
      $H_l^k = 3DconvLSTM(H_l^{k-1}, E_l^{k-1}, MaxPool(ReLU(R_{l-1}^k, E_{l-1}^k)))$
    else
      $H_l^k = 3DconvLSTM(H_l^{k-1}, E_l^{k-1}, MaxPool(ReLU(R_{l-1}^k, E_{l-1}^k)), upsample(H_{l+1}^k))$
    end if
  end for
  for $l = 1$ to L **do** ▷ Bottom-up updates
    if $l = 1$ **then**
      $I_l^k = x_t, \; P_l^k = SATLU(ReLU(conv(H_l^k)))$
    else
      $I_l^k = MaxPool(ReLU(R_{l-1}^k)),$
      $P_l^k = ReLU(conv(H_l^k))$
    end if
    $\Delta I_l^k = I_l^k - I_l^{k-1}, \Delta R_l^k = spconv(\Delta I_l^k),$
    $R_l^k = R_l^{k-d} + \Delta R_l^k, E_l^k = spconv(I_l^k - P_l^k)$
  end for
end for

4. Experimental Results

In this section, we first evaluate the performance of our model in video prediction using two bench-mark datasets: (1) synthetic sequences of the Moving-MNIST database and (2) the KTH[1] real world human movement database. We then investigate the representations in the model to understand how recurrent network structures have impacted on the feedforward representation. We finally compare the temporal activities of neurons in the network model with that of neurons in the visual cortex of monkeys, in video sequence learning, to evaluate the plausibility of HPNet.

Since for video prediction, PredNet is the most neurally plausible model and PredRNN++ provides state-of-the-art computer vision performance, we will compare HPNet's performance with these two network models. Because these two models work on frame-to-frame basis, we implemented three versions of our network for comparison: (1) Frame-to-Frame (F-F), where we set our data spatiotemporal block size to one frame and used 2D convLSTM instead of 3D convLSTM to predict the next frame based on the current frame; (2) Block-to-Frame (B-F), where we used a sliding block window to predict the next frame based on the current block of frames; (3) Block-to-Block (B-B), where the next spatiotemporal block was predicted from the current spatiotemporal block (Figure 1d).

We trained all five networks using 40-frame sequences extracted from the two databases in the same way as described in (Lotter et al., 2016; Wang et al., 2018). We then compared

[1] http://www.nada.kth.se/cvap/actions/

their performance in predicting the next 20 frames when only the first 20 frames were given. The test sequences were drawn from the same dataset but not in the training set. The common practice in PreNet and PredRNN++ for predicting future frames when input is no longer available is to make the prediction of the last time step the next input and use that to generate prediction of the next time step. All models tested have four modules (layers). All three versions of our model and PredNet used the same number of feature channels in each layer, optimized by grid search, i.e. (16,32,64,128) for the Moving-MNIST dataset, and (24,48,96,192) for the KTH dataset. For PredRNN++, we used the same architecture and feature channel numbers provided by Wang et al. (2018). All kernel sizes are either 3×3 (for F-F) or 3×3×3 (for B-F and B-B) for all five models. The input image frame's spatial resolution is 64×64.

The models were trained and tested on GeForce GTX TITAN X GPUs. We evaluated the prediction performance based on two quantitative index: Mean-Squared Error (MSE) and the Structural Similarity Index Measure (SSIM) (Wang et al., 2004) of the last 20 frames between the predicted frames and the actual frames. The values of SSIM range from -1 to 1, with larger value indicating greater similarity between the predicted frames and the actual frames.

4.1. Synthetic sequence prediction on the Moving-MNIST dataset

We randomly chose subsets of digits in the Moving MNIST[2] dataset in which the video sequences contain two handwritten digits bouncing inside a frame of 64×64 pixels. We extracted 40-frame sequences at random starting frame position in the video in the same way as in (Srivastava et al., 2015) (followed by PredNet and PredRNN++). This extraction process is repeated 15000 times, resulting in a training set of 10000 sequences, a validation set of 2000 sequences, and a testing set of 3000 sequences.

Figure 2 and Table 1 compare the results of different models on the Moving-MNIST dataset. There are 40 frames in total and we show the results every two frames. Note that actual input was provided only for the first 20 frames (top block) to generate real prediction errors but not for the last 20 frames (bottom block). We can see B-F achieves better performance than B-B in short term prediction task when actual input frames are provided, but B-B outperforms B-F in the longer range prediction, reflecting learning of the relationships at the movement levels by the 3D convLSTM. B-F doing better than F-F confirmed that the spatiotemporal block data structure provides additional information for modeling movement tendency. Finally, we found that even F-F achieved better prediction results than PredNet, suggesting that a feature hierarchy might be more useful than a hierarchy of successive prediction errors. Finally, our B-B network outperformed the state-of-the-art PredRNN++.

[2] http://yann.lecun.com/exdb/mnist/

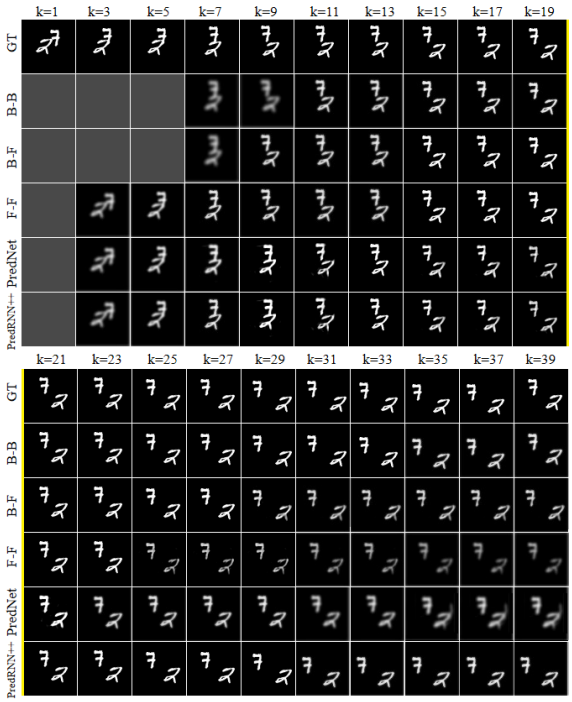


Figure 2. Video prediction results on Moving-MNIST dataset, where the first row to last row are ground truth (GT), results from three different version of HPNet (block-to-block (B-B), block-to-frame (B-F), frame-to-frame (F-F)), PredNet, and PredRNN++, respectively. k=1 to k=19 are predicted frames of the models when the input frames were available. k=21 to k=39 are the "dead-reckoning" predicted frames of the model when there is no input.

Table 1. Comparison Results of different methods on Moving-MNIST dataset for long time prediction experiment.

Method	SSIM	MSE
Ours(B-B)	**0.915**	**65.2**
Ours(B-F)	0.793	73.2
CM+ConvLSTM (F-F)	0.692	89.5
PredNet (Lotter et al., 2016)	0.658	101.2
PredRNN++ (Wang et al., 2018)	0.872	69.4

4.2. Real-world sequence prediction on the KTH dataset

Schüldt et al. (2004) introduced the KTH video database which contains 2391 sequences of six human actions: walking, jogging, running, boxing, hand waving, and hand clapping, performed by 25 subjects in four different scenarios. We divided video clips across all 6 action categories into a training set of 108717 sequences (persons #1-16) and a test set of 4086 sequences (persons #17-25) as was done in Wang et al. (2018), except we extracted 40-frame sequences.

We center-cropped each frame to a 120×120 square and then re-sized it to input frame size of 64×64.

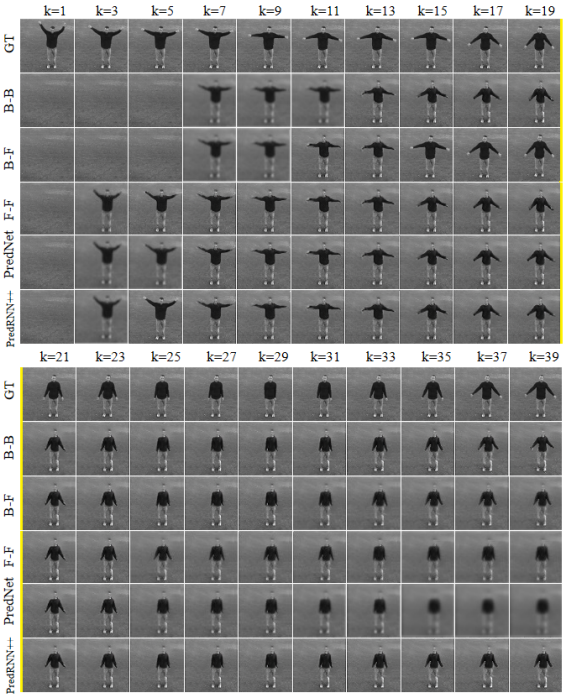


Figure 3. Video prediction results on the KTH dataset, where the first row to last row are ground truth (GT), results from block-to-block (B-B), block-to-frame (B-F), frame-to-frame (F-F), PredNet, and PredRNN++, respectively, same format as Figure 2.

Table 2. Comparison Results of different methods on the KTH dataset for long time prediction experiment.

Method	SSIM	MSE
Ours(B-B)	**0.882**	**80.3**
Ours(B-F)	0.784	93.1
CM+ConvLSTM (F-F)	0.701	103.4
PredNet (Lotter et al., 2016)	0.656	108.9
PredRNN++ (Wang et al., 2018)	0.865	86.7

Figure 3 and Table 2 compared the results of the different models on the KTH dataset, essentially reproducing all the observations we made based on the Moving-MINST dataset (Figure 2). B-B outperformed all tested models in the long range video prediction task. Figure 4a and Figure 4b compared the video prediction performance of the different models in terms of the "dead-reckoning frames" to be predicted when only the first twenty frames were provided for the two datasets. The results show that, in both cases, B-B is far more effective than B-F in long range video prediction. Figure 4c showed that the ratio of SSIM and training time peaks at a 4-module network. The SSIM of a 5-module network was about the same as that of a 4-module network but took longer time to converge. The B-F, with a sliding window of a single frame stride, took much longer to train yet still under-performed. Figure 4d showed SSIM performance and training time of the different models. It shows that the B-B (sparse) version of HPNet took only 10% longer to train than PredRNN++ even though it has more loops into the networks and has to process spatiotemporal blocks. Both PredRNN++ and HPNet require twice amount of the training time relative to PredNet, illustrating the computational efficiency of using sparse codes. Sparsifying our DCNN feedforward path reduced our B-B network's training time by 13% (comparing B-B (sparse) versus B-B (non-sparse) in Figure 4d).

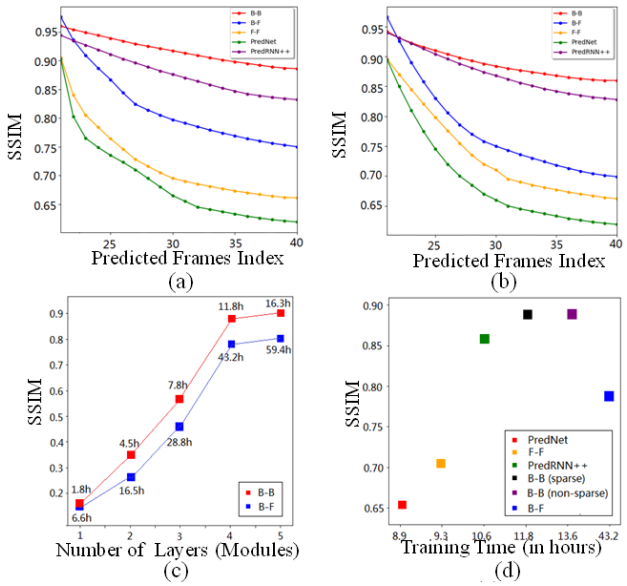


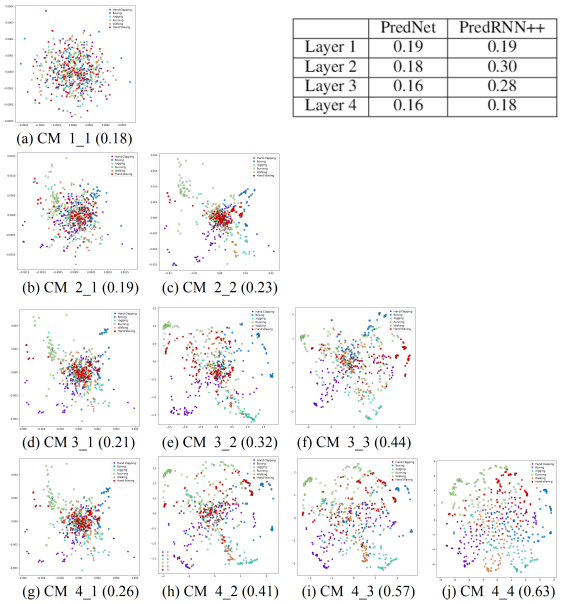
Figure 4. (a) Comparison of the prediction results of the five models for the Moving-MINST dataset on the last 20 frames in structural similarity measures (SSIM). (b) Comparison of the prediction results on the KTH dataset. (c) Comparison of the performance (and training time) of the B-B and the B-F networks as a function of the number of modules in the network. (d) Training time versus SSIM performance of the different models. Note, the training time (x) axis not in a linear scale.

4.3. Semantic clustering in the hierarchical representation

We trained the HPNet network in the block-to-block (B-B) scheme with different numbers of modules, and found that adding cortical modules tends to improve performance (Figure 4c). How do the hierarchical representation and recurrent feedback help achieve better prediction performance? We used t-SNE (van der Maaten & Hinton, 2008) to visualize the representation R in the different modules of the networks with different number of modules for the last 20 dead-reckoning predicted frames of 600 testing sequences belonging to the six movements in the KTH dataset.

Figure 5 reveals that the addition of higher modules has lead

to the formation of more distinct cluster of global movement patterns in the representation units of the lower modules. This transformation in representation in the earliest module (Figure 5a versus Figure 5e) leads to a significant decoding accuracy improvement, from essentially chance (16%) level to 26%, based on the unit activities in the first module alone. The semantic clustering and decoding accuracy improves progressively as one moves up the hierarchy, with a decoding accuracy of 63% for the top module of the 4-module network. On the other hand, the movement decoding results on the LSTM representations in PredRNN++ and PredNet are considerably weaker (see Figure 5 inset), reflecting weaker semantic encoding and clustering of the movement patterns. Thus, the better performance of HPNet in long range video predictions might be attributed to its having learned semantically meaningful hierarchical spatiotemporal feature representations and movement to movement relationships (see also Kheradpisheh et al. (2018)).



	PredNet	PredRNN++
Layer 1	0.19	0.19
Layer 2	0.18	0.30
Layer 3	0.16	0.28
Layer 4	0.16	0.18

Figure 5. Visualization and Decoding Accuracy of R representational units of the different modules in (a) a one-module network; (b)-(c) a two-module network; (d)-(f) a three-module network; and (g)-(j) a four-module network. "Module 2_1 (0.19)" means Module 1 in a two-module network, with decoding accuracy at 19%. Inset shows the decoding accuracy based on the output responses of different LSTM layers in PredNet and PredRNN++.

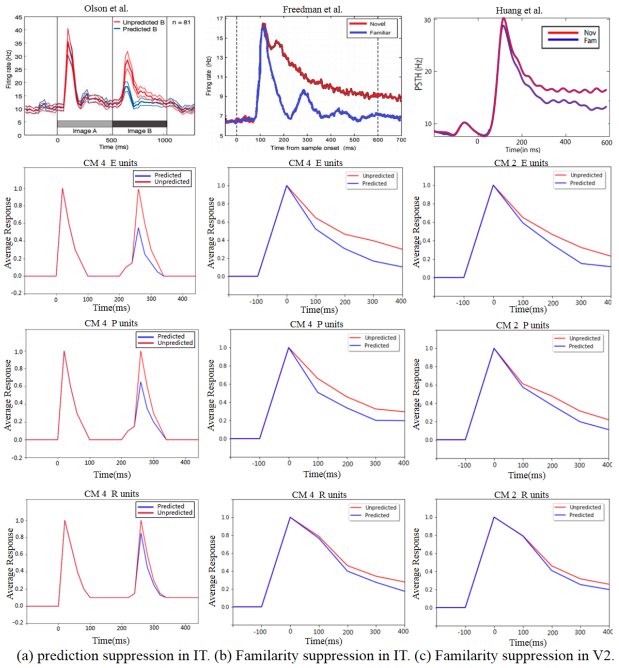
4.4. Neurophysiological evidence for HPNet

Is there any neurophysiological evidence in support of HPNet as a viable neural model for the hierarchical visual cortex? We will now discuss some neurophysiological findings that show neural responses in both the object recognition area (inferotemporal cortex IT) and the early visual areas (V1 and V2) are remarkably similar to the behaviors of the units in HPNet.

Recent single-unit recording experiments in the inferotemporal cortex (IT) of monkeys have shown that neurons' responses to images in predicted sequences were suppressed relative to a novel sequences, suggesting that the neural responses may in part be reflecting prediction errors (Meyer & Olson, 2011; Meyer et al., 2014; Ramachandran et al., 2017). In a particular experiment Meyer & Olson (2011), a set of image pairs in a fixed order were each presented to the monkeys over 800 times across many days while they performed a fixation task. After this exposure learning, the second image in each pair became "predictable" upon the presentation of the first image in the predicted pairs. However, when the order of these images was changed, the first image obviously could not predict the second image and these pairs were called novel "un-predicted" pairs. Figure 6 left column (top-row) shows the averaged responses of 81 IT neurons to the second image in the predicted pairs were significantly less than to that in the unpredicted pairs. Note that as all the tested images appeared in both the first image and the second image in the experimental design, neural responses to the first stimuli in the pairs were the same as expected and the reduction in the responses to the second images was due to prediction suppression.

To evaluate whether units in HPNet exhibit the same behaviors, we performed the same experiment on HPNet with 2000 epochs of training on the image pairs. Each stimulus sequence began with five gray frames, followed by ten frames of the first image in the pair, then two gray frames as gap, then ten frames of the second image in the pair. We found the averaged responses of the units to the predicted set and the unpredicted set were the same prior to training. After training, the second image in the predicted pairs evoked much less responses than the same image in the unpredicted pairs (Figure 6 left column, 2nd to 4th rows) consistent with neurophysiological observations in Meyer & Olson (2011).

Interestingly, the prediction suppression effect can be observed in all three types of neurons and in all the modules in the HPNet hierarchy, with the higher modules showing a stronger effect. While it is not surprising that the prediction error neurons E would decrease their responses as the network learns to predict the familiar pairs better, it is rather intriguing to find the representation neurons R (Figure 6 4th row) and the prediction neurons P (Figure 6 3rd row) also exhibit prediction suppression, even though these neurons represent features rather than prediction errors. This might explain why observations of prediction suppression were so prevalent in the randomly sampled neurons in IT. Prediction suppression of image sequences in the early visual cortex have not been reported, so prediction suppression effects we observed in the earlier modules of HPNet can only serve as an experimental prediction.



(a) prediction suppression in IT. (b) Familarity suppression in IT. (c) Familarity suppression in V2.

Figure 6. Left column: prediction suppression in IT (Meyer & Olson, 2011). Middle column: image familiarity suppression in IT (Freedman & Assad, 2006). Right column: Image familiarity suppression in V2 (Huang et al., 2018). First row: neurophysiological experimental data; 2nd row: averaged responses of E (prediction error) neurons in HPNet; 3rd row: averaged response of P (prediction) neurons; 4th row: averaged responses of R (representation) neurons. Red curve: Novel and unpredicted images or sequences. Blue curve: familiar and predicted sequences.

A related neural phenomenon called image familiarity suppression effect, studied and observed previously in IT (Freedman & Assad, 2006; Mruczek & Sheinberg, 2007; Meyer et al., 2014), has been recently observed in the early visual cortex (V1 and V2) (Huang et al., 2018). In these experiments, a set of object images was presented to the monkeys for multiple days. Post training, it was found that neural responses to the familiar images were significantly suppressed relative to the novel images in the later part of their responses in both IT and V2, as shown in the middle and right columns in the top row of Figure 6 respectively. It is important to note that neurons in monkey V1 and V2 have very small receptive fields, and yet they show familiarity suppression effects to large object images much larger than the size of their classical receptive fields. Moreover, evidence based on onset timing of the effects implicates local recurrent circuits in each visual area in the encoding of global image memories, consistent with other studies in mouse V1 (Yao et al., 2007; Han et al., 2008; Xu et al., 2012; Cooke & Bear, 2014; 2015)

We performed the image familiarity experiment (Huang et al., 2018) on HPNet, using the same stimulus presentation paradigm used in the prediction suppression experiment, except now for the same image was shown for 15 frames for each presentation to simulate static image presentation in the experiment. The stimuli were divided into a familiar set of 25 images, and a novel set of 25 images. Prior to training of 2000 epochs, the averaged responses of all the units within the center region of the image input were the same for the two sets. Subsequent to training, the later part of the units' responses were suppressed for the familiar images relative to the novel images. The middle and the right columns show the three types of units' responses in Cortical Module 4 (roughly corresponds to IT) and in Cortical Module 2 (roughly corresponds to V2) respectively – all showing the image familiarity suppression effect. As in the case of prediction suppression, the reduction of responses for the E units can be attributed to reduction in prediction errors, but the causes for the observed reduction in averaged responses for the R and P units, which were also observed neurophysiologically, require further investigation. These results suggest that HPNet might be a viable model for understanding computation and learning in the hierarchical visual cortex. Furthermore, it provides a unified account of the two well studied neurophysiological phenomena, suggesting that they could in fact emerge from the same underlying mechanisms.

5. Conclusion

We have developed a hierarchical prediction network (HPNet) for predictive learning of spatiotemporal memories that is competitive both for video prediction, and for understanding the learning principles and the computational mechanisms of the hierarchical visual system. HPNet models the analysis-by-synthesis computational architecture with local gated recurrent circuits at every level. It utilizes predictive self-supervised learning as in PredNet and PredRNN++, but integrates additional neural constraints such as spatiotemporal processing, counter-stream architecture, feature hierarchy, prediction error computation and sparse convolution into a new model that delivers the state-of-the-art performance in long range video prediction. We show that recurrent interaction in the HPNet hierarchy improves higher order semantic clustering in the representations of the lower modules, which facilitate movement-to-movement relationship learning. The model automatically accounts for neurophysiological observations in sequence prediction learning and static image familiarity learning observed both in higher order visual areas (IT) as well as early visual cortex (V1 and V2) of awake monkeys. These findings suggest that predictive self-supervised learning likely plays an important role in hierarchical representation learning in the visual cortex and that HPNet is a viable computational model for understanding the functional mechanisms of cortical circuits in the hierarchical visual system.

References

Babaeizadeh, M., Finn, C., Erhan, D., Campbell, R. H., and Levine, S. Stochastic variational video prediction. *CoRR*, abs/1710.11252, 2017.

Choy, C. B., Xu, D., Gwak, J., Chen, K., and Savarese, S. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *ECCV*, 2016.

Cooke, S. F. and Bear, M. F. How the mechanisms of long-term synaptic potentiation and depression serve experience-dependent plasticity in primary visual cortex. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 369 1633:20130284, 2014.

Cooke, S. F. and Bear, M. F. Visual recognition memory: a view from v1. *Current opinion in neurobiology*, 35: 57–65, 2015.

Dave, A., Russakovsky, O., and Ramanan, D. Predictive-corrective networks for action detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2067–2076, 2017.

Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. The helmholtz machine. *Neural Computation*, 7:889–904, 1995.

Denton, E. L. and Fergus, R. Stochastic video generation with a learned prior. In *ICML*, 2018.

Dijkstra, N., Zeidman, P., Ondobaka, S., van Gerven, M. A. J., and Friston, K. J. Distinct top-down and bottom-up brain connectivity during visual perception and imagery. In *Scientific Reports*, 2017.

Elman, J. L. Finding structure in time. *Cognitive Science*, 14:179–211, 1990.

Finn, C., Goodfellow, I. J., and Levine, S. Unsupervised learning for physical interaction through video prediction. In *NIPS*, 2016.

Freedman, D. J. and Assad, J. A. Experience-dependent representation of visual categories in parietal cortex. *Nature*, 443:85–88, 2006.

Friston, K. J. Does predictive coding have a future? *Nature Neuroscience*, 21:1019–1021, 2018.

Goroshin, R., Mathieu, M., and LeCun, Y. Learning to linearize under uncertainty. In *NIPS*, 2015.

Grossberg, S. Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science*, 11:23–63, 1987.

Han, F., Caporale, N., and Dan, Y. Reverberation of recent visual experience in spontaneous cortical waves. *Neuron*, 60:321–327, 2008.

Huang, G., Ramachandran, S., Lee, T. S., and Olson, C. R. Neural correlate of visual familiarity in macaque area v2. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 2018.

Kalchbrenner, N., van den Oord, A., Simonyan, K., Danihelka, I., Vinyals, O., Graves, A., and Kavukcuoglu, K. Video pixel networks. In *ICML*, 2017.

Kersten, D. J. and Yuille, A. L. Bayesian models of object perception. *Current opinion in neurobiology*, 13 2:150–8, 2003.

Kheradpisheh, S. R., Ganjtabesh, M., Thorpe, S. J., and Masquelier, T. Stdp-based spiking deep convolutional neural networks for object recognition. *Neural networks : the official journal of the International Neural Network Society*, 99:56–67, 2018.

Lee, A. X., Zhang, R., Ebert, F., Abbeel, P., Finn, C., and Levine, S. Stochastic adversarial video prediction. *CoRR*, abs/1804.01523, 2018.

Lee, T. S. The visual system's internal models of the world. *Proceedings of the IEEE*, 103:1359–1378, 2015.

Lee, T. S. and Mumford, D. Hierarchical bayesian inference in the visual cortex. *Journal of the Optical Society of America. A, Optics, image science, and vision*, 20 7:1434–48, 2003.

Liu, X., Pool, J., Han, S., and Dally, W. J. Efficient sparse-winograd convolutional neural networks. *CoRR*, abs/1802.06367, 2017.

Lotter, W., Kreiman, G., and Cox, D. D. Deep predictive coding networks for video prediction and unsupervised learning. *CoRR*, abs/1605.08104, 2016.

Mathieu, M., Couprie, C., and LeCun, Y. Deep multi-scale video prediction beyond mean square error. *CoRR*, abs/1511.05440, 2015.

McClelland, J. L. and McNaughton, B. L. Complementary learning systems 1 why there are complementary learning systems in the hippocampus and neocortex : Insights from the successes and failures of connectionist models of learning and memory. 1999.

McClelland, J. L. and Rumelhart, D. E. Distributed memory and the representation of general and specific information. *Journal of experimental psychology. General*, 114 2:159–97, 1985.

Meyer, T. and Olson, C. R. Statistical learning of visual transitions in monkey inferotemporal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 108 48:19401–6, 2011.

Meyer, T., Walker, C., Cho, R. Y., and Olson, C. R. Image familiarization sharpens response dynamics of neurons in inferotemporal cortex. *Nature Neuroscience*, 17:1388–1394, 2014.

Mruczek, R. E. B. and Sheinberg, D. L. Context familiarity enhances target processing by inferior temporal cortex neurons. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27 32:8533–45, 2007.

Mumford, D. On the computational architecture of the neocortex. *Biological Cybernetics*, 65:135–145, 1991.

Mumford, D. On the computational architecture of the neocortex. ii. the role of cortico-cortical loops. *Biological cybernetics*, 66 3:241–51, 1992.

Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., DiCarlo, J. J., and Yamins, D. L. K. Task-driven convolutional recurrent models of the visual system. *CoRR*, abs/1807.00053, 2018.

Oh, J., Guo, X., Lee, H., Lewis, R. L., and Singh, S. P. Action-conditional video prediction using deep networks in atari games. In *NIPS*, 2015.

O'Reilly, R. C., Wyatte, D., and Rohrlich, J. Learning through time in the thalamocortical loops. 2014.

Palm, R. B. Prediction as a candidate for learning deep hierarchical models of data. 2012.

Pan, B., Lin, W., Fang, X., Huang, C., Zhou, B., and Lu, C. Recurrent residual module for fast inference in videos. *CoRR*, abs/1802.09723, 2018.

Patraucean, V., Handa, A., and Cipolla, R. Spatio-temporal video autoencoder with differentiable memory. *CoRR*, abs/1511.06309, 2015.

Ramachandran, S., Meyer, T., and Olson, C. R. Prediction suppression and surprise enhancement in monkey inferotemporal cortex. *Journal of neurophysiology*, 118 1: 374–382, 2017.

Rao, R. P. N. and Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2 1: 79–87, 1999.

Schüldt, C., Laptev, I., and Caputo, B. Recognizing human actions: a local svm approach. *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, 3:32–36 Vol.3, 2004.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and chun Woo, W. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NIPS*, 2015.

Srivastava, N., Mansimov, E., and Salakhutdinov, R. Unsupervised learning of video representations using lstms. In *ICML*, 2015.

Ullman, S. Sequential seeking and counter streams: a computational model for bidirectional flow in the visual cortex. *Cerebral Cortex*, 5:1:1–1, 1995.

van der Maaten, L. and Hinton, G. E. Visualizing data using t-sne. 2008.

Villegas, R., Yang, J., Zou, Y., Sohn, S., Lin, X., and Lee, H. Learning to generate long-term future via hierarchical prediction. In *ICML*, 2017.

Vondrick, C., Pirsiavash, H., and Torralba, A. Generating videos with scene dynamics. In *NIPS*, 2016.

Wang, Y., Long, M., Wang, J., Gao, Z., and Yu, P. S. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. In *NIPS*, 2017.

Wang, Y., Gao, Z., Long, M., Wang, J., and Yu, P. S. Predrnn++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. In *ICML*, 2018.

Wang, Y., Jiang, L., Yang, M.-H., Li, L.-J., Long, M., and Fei-Fei, L. Eidetic 3d lstm: A model for video prediction and beyond. In *ICLR*, 2019.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004.

Wen, H., Han, K., Shi, J., Zhang, Y., Culurciello, E., and Liu, Z. Deep predictive coding network for object recognition. In *ICML*, 2018.

Wichers, N., Villegas, R., Erhan, D., and Lee, H. Hierarchical long-term video prediction without supervision. In *ICML*, 2018.

Xu, S., Jiang, W., Poo, M.-M., and Dan, Y. Activity recall in visual cortical ensemble. In *Nature Neuroscience*, 2012.

Xu, Z., Wang, Y., Long, M., and Wang, J. Predcnn: Predictive learning with cascade convolutions. In *IJCAI*, 2018.

Yao, H., Shi, L., Han, F., Gao, H., and Dan, Y. Rapid learning in cortical coding of visual scenes. *Nature Neuroscience*, 10:772–778, 2007.