Learning Features and Predictive Transformation Encoding based on a Horizontal Product Model

Junpei Zhong, Cornelius Weber, and Stefan Wermter

Department of Computer Science, University of Hamburg,
Vogt-Kölln-Str. 30, 22527 Hamburg, Germany
{zhong,weber,wermter}@informatik.uni-hamburg.de
http://www.informatik.uni-hamburg.de/WTM/

Abstract. The visual system processes the features and movement of an object in separate pathways, called the ventral and dorsal streams. To integrate this principle in a functional model, a recurrent predictive network with a horizontal product is introduced. Learned in an unsupervised manner, two sets of hidden units represent cells in the ventral and dorsal pathways, respectively. Experiments show that the activity in the ventral-like units persists, given that the same feature appears in the receptive field, whilst the activity in the dorsal-like units shows a fluctuating pattern with different directions of object movements. Moreover, we show that the position information predicts the input's future position taking into account its moving direction due to the direction-selective responses of the dorsal-like units.

1 Introduction

1.1 Biological Evidence

Based on the argument by Livingstone and Hubel [10] that cells in the visual pathway are organized in a hierarchical way with increasing receptive field size and higher visual abstraction from lower to higher layers, Ungerleider and Mishkin [12] established a theory of two parallel visual pathways. The 'dorsal pathway' encodes spatial information, invariant of object-specific properties, while the 'ventral pathway' encodes object feature identity, invariant of positions and sizes, leading to generalization ability. Commonly, they are referred to as the 'where' and 'what' pathways.

Both pathways originate in the occipital lobe with the primary visual areas V1 and V2, and are separate only in higher visual cortices. From recordings [14] in macaque monkeys, however, a distinction of feature- and transformation movement encoding cells is already evident in the diverse response properties of complex cells in V1: some complex cells are direction- and speed-selective independent of spatial frequency, hence resembling neurons in area MT of the dorsal pathway, which predict 'where' in addition to the apparent direction of movement. Other complex cells are selective to spatial frequency independent of speed, hence coding for feature identity (the 'what' pathway). Interestingly, because of the delays from upstream and downstream neural transmission, the 'where' pathway should maintain a *future* position of an object. This can be accounted for by the representation of movement direction and velocity in the dorsal

pathway. The encoding of object location allows for motor-relevant representations in the higher dorsal pathway [6], while the convergence of ventral and dorsal pathways may also give rise to the understanding of object affordance and object manipulation and its control, which is a function of the mirror neuron system [2].

1.2 Modeling Background

To distill object features regardless of transformations, networks were devised to learn generalization over transformations, for example by encouraging the neurons to fire invariantly while transformations are performed in their input stimuli [4, 19]. Specifically, by pooling the outputs of modeled simple cells, position-invariant responses as found in V1 complex cells can be obtained by self-organization [5, 7]. Such principles can be applied hierarchically to achieve larger-scale invariances (e.g. [15]). However, such models lose the information about the specific transformations, which may be crucial for certain behaviors.

In bilinear models of visual routing, a set of control neurons dynamically modifies the weights of the 'what' pathway on a short time scale. The control units, encoding the object's position, thereby route the visual information from any retinal position to an object-centered reference frame on the top-most level of the 'what' pathway [13, 3, 11]. Such control neurons have been hypothesized to reside in the pulvinar [1], or the mediodorsal nucleus [16], of the thalamus.

Köster et al. [8] applied a horizontal product model with Independent Component Analysis (ICA) to separate the location of image features from their identities. The output is then generated by multiplying outputs from sub-models via the horizontal product. The horizontal product model reduces computational effort: assuming that there are I input units, considering T transformations and F features, a full bilinear model has $I \times T \times F$ connections, but only $2I \times (T + F)$ connections are needed in this network.

With the idea of solving the 'what' and 'where' problem jointly, here we propose a method that can extract two or more components of information into separate pathways from input data. Unlike previous approaches, our model encodes motion to predict its future input: both pathways incorporate recurrent connections to capture the observed response properties of complex cells.

We introduce the architecture and algorithm in section 2 and 3. A simulation experiment based on artificial data is presented in section 4, followed by a discussion in section 5.

2 Architecture

We specify a three-layer network with recurrent connections and a horizontal product (see Fig. 1). The input layer corresponds to the simple cells in V1 (cortical layer IV), while the hidden layer corresponds to complex cells in V1(cortical layer II and III). The output layer is a feedback prediction of the input. The hidden layer contains two independent sets of units representing dorsal-like 'd' and ventral-like 'v' neurons respectively, inspired by the functional properties of dorsal and ventral pathways: (i) fast responding dorsal-like units predict object position and hence encode movement; (ii)

slow responding ventral-like units represent object identity. The recurrent connection in the hidden layers helps to predict movement in layer d and maintain a persistent representation of an object in layer v. The horizontal product brings both pathways together again in the output layer with one-step ahead predictions. Let us denote the output layer's input from layer d and layer v as x^d and x^v, respectively. The network output s^o is obtained via the horizontal product as

$$s^o = x^d \odot x^v \tag{1}$$

where $\odot$ indicates element-wise multiplication, so each pixel is defined by the product of two independent parts, i.e. for output unit k it is $s^o_k = x^d_k \cdot x^v_k$.

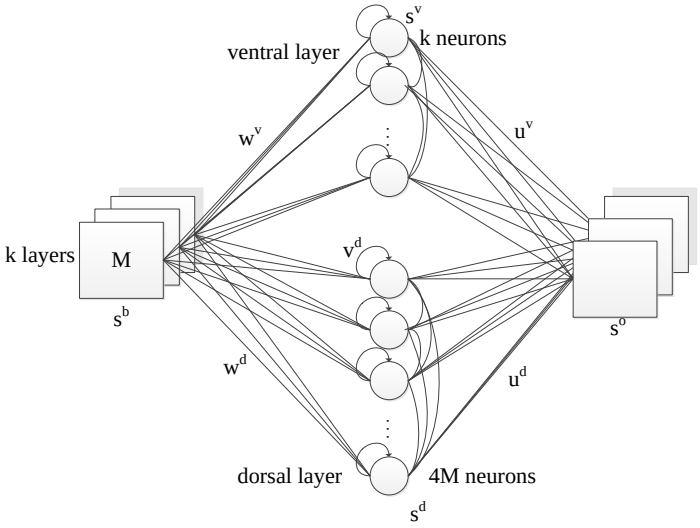


Fig. 1: Network architecture

3 Training Algorithm

We use $s^b_i(t)$ to represent the activation of the input unit i at the t-th time-step. In some of the following equations, we will omit the time-index t if all activations are in the same time-step. The hidden units' inputs y^v_j in the ventral pathway and y^d_j in the dorsal pathway are defined as

$$y^v_j(t) = \sum_i s^b_i(t) w^v_{ji} + \sum_i s^b_i(t-1) \bar{w}^v_{ji} + \sum_{j'} s^v_j(t-1) v^v_{jj'} \tag{2}$$

$$y^d_l(t) = \sum_i s^b_i(t) w^d_{li} + \sum_i s^b_i(t-1) \bar{w}^d_{li} + \sum_{l'} s^d_l(t-1) v^d_{ll'} \tag{3}$$

where w^d_{li}/w^v_{ji} represent the weighting matrices between dorsal/ventral layers and the input layer, $\bar{w}^d_{li}$ / $\bar{w}^v_{ji}$ represent the weighting matrices between a one-step delayed input and the two hidden layers and $v^d_{ll'}$/$v^v_{jj'}$ indicate the recurrent weighting matrices within the hidden layers. We expect that the incorporation of the time-delayed inputs directly from s^b_i can introduce more stable input signals in both units regardless of the short-time changes of object features.

The transfer functions in both hidden layers employ a logistic function and a softmax function:

$$z^v_j = \frac{1}{1 + exp(-a_j y^v_j + b_j)} \; ; \quad z^d_l = \frac{1}{1 + exp(-a_l y^d_l + b_l)} \tag{4}$$

$$s^v_j = \frac{exp(z^v_j)}{\sum_{j'} exp(z^v_{j'})} \; ; \quad s^d_l = \frac{exp(z^d_l)}{\sum_{l'} exp(z^d_{l'})} \tag{5}$$

The logistic function has two local modifiable parameters a and b, leading to the intrinsic plasticity of neurons, which we will discuss in the following paragraphs. These transfer functions lead to regular firing on the hidden layer.

The terms of the horizontal products of both pathways can be presented as follows:

$$x^v_k = \sum_j s^v_j u^v_{kj}; \quad x^d_k = \sum_l s^d_l u^d_{kl} \tag{6}$$

The network output is described in Eq. 1.

The training progress is determined by a cost function:

$$C = \frac{1}{2} \sum_t^T \sum_k^N (s^b_k(t+1) - s^o_k(t))^2 \tag{7}$$

where $s^b_i(t+1)$ is the one-step ahead input, as well as the desired output, $s^o_k(t)$ is the current output, T is the total number of available time-step samples and N is the number of output nodes, which equals the number of input nodes. Following gradient descent, each weight update in the network is proportional to the negative gradient of the cost with respect to the specific weight w that will be modified:

$$\Delta w = -\eta \frac{\partial C}{\partial w} \tag{8}$$

The object identity and position information from the input data is distinguished and extracted by the two pathways during training. The activations in layer v are first determined by Eq. 4 and 5; after that, a constraint is set to the ventral-like units v so that the states in the following time-steps are forced to be equal to the first time-step as long as the identity of the object remains unchanged. That is, the ventral-like units' activations remain the same until the appearance of a new object. The dorsal-like units, which do not have such a constraint, can update quickly according to the current position of the object.

The weights between input and hidden layers and between hidden layers and output layer, are set to be non-negative. This non-negativity constraint makes the representation purely additive (allowing no subtractions) and accounts for the non-negativity of the data.

As mentioned in Eq. 4, the neurons' intrinsic plasticity modeling [17], which is applied in the hidden layers, is based on the biological finding that a neuron can adjust its intrinsic electrical properties following its neuronal or synaptic activity. Mathematically, it adjusts its function parameters, slope and threshold, so as to fit its output rate to a sparse exponential regime. The update of parameters a and b is given by

$$\Delta a_i = \eta_a(\frac{1}{a_i} + y_i - 2y_i z_i - \frac{1}{\mu} y_i z_i + \frac{1}{\mu} y_i z_i^2) \tag{9}$$

$$\Delta b_i = \eta_b(1 - 2z_i - \frac{1}{\mu} z_i + \frac{1}{\mu} z_i^2) \tag{10}$$

where μ is the mean for the exponential defined over the positive half-axis. The learning of parameters a and b leads to different shapes of the transfer function. Specifically, the parameter a controls the gain of the input, changing the slope of the sigmoid function, while the parameter b shifts the function, resembling a change of threshold.

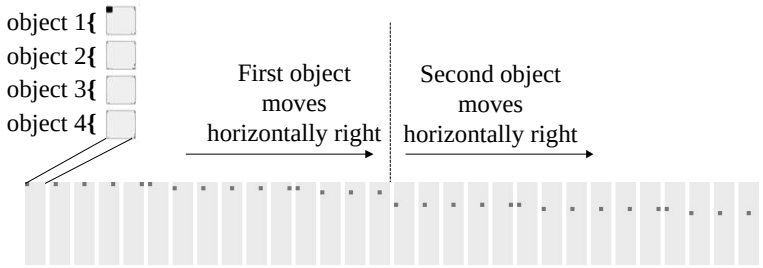
4 Experiment

As a proof of concept, we present artificially generated input data to the network. These data mimic moving objects, i.e. their positions change quickly but their identity changes rarely. In this dataset, only one object appears at one unique position in any time-step. The input space is of size 5×5 and 4 different objects at any position, which can be represented on an input layer of total size $4 \times 5 \times 5$ (see Fig. 2a). This minimalistic set up sketches a hyper-column in V1 that processes oriented lines of 4 different orientations at 5×5 possible positions.

The training data set comprises four directional movements horizontally and vertically covering all of the possible sequences of all objects. For instance, the first data set contains an activation in the first layer moving from coordinate $(1, 1)$ to $(1, 2), (1, 3), \cdots$ and back to $(1, 1)$ (Fig. 2a shows part of the input data moving rightwards). These movements vary in different starting points and different objects. In the training process, the target data is one time-step ahead of the input data.

In the following experiment, the maximum iteration is set to be $100,000$, learning rates are $\eta_a = 0.0001$, $\eta_b = 0.0001$, $\eta = 0.01$. $\mu_v = 0.1$ and $\mu_d = 0.01$ are the parameters of intrinsic plasticity μ in ventral- and dorsal-like units, respectively. In order to learn movement appropriately, activation in both hidden layers is set to zero when changes. The stopping criteria is that the error difference of consecutive iterations is smaller than 10^{-7}.

With the input sample in Fig. 2a, Fig. 2b shows the corresponding one-step ahead prediction. We can generally observe that the output layer predicts the one-step ahead movement. Note that the output is inactive in every first time-step since the recurrent and time-delayed connections require the previous inputs which are not available in the first time-step. As depicted in Fig. 3, activations in the corresponding activations of



(a) Partial training samples while one object moving horizontally rightwards.

(b) The network output given the input above.

Fig. 2: Example of input and output data.

hidden layer v stay stable when one object appears, while we can distinguish various patterns in the dorsal-like layer d representing perceptions of different positions. The training error over the course of learning is depicted in Fig. 4. The stopping criterion was achieved by around iteration 3100.

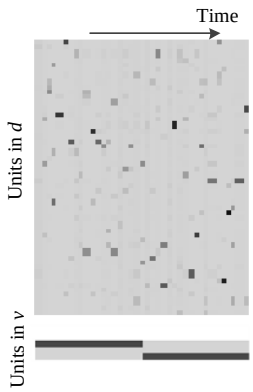


Fig. 3: Network activations in hidden layers.

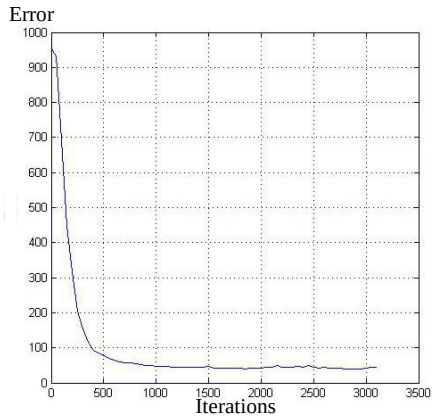


Fig. 4: Output error through iterations.

5 Discussion

This paper presents a new predictive architecture of emergent 'what' and 'where' processing. The experimental results show that the information of object identity and position have been successfully separated in an unsupervised manner: the activation of units in ventral-like units v remains stable while presenting the same object, but different patterns appear in the dorsal-like units d indicating the position and movement direction. These results, especially the activation in the dorsal-like units, are analogous to the recording of 'direction-selective' complex cells in V1 [14]. Though there has been observed a continuum between identity-specific and motion-specific complex cells, in our model we simplify this relationship and clearly discriminate between the two extremes only.

This model also highlights the role of recurrent connections [9], which store previous movement information, and serve a predictive function. Prediction in the visual system is apparent by neurophysiological findings of predictive receptive field shifts [16] and behavioural findings of visual responsibility in movement prediction [18]. We believe that any cortical area should compensate its processing delays via prediction.

Due to its simplicity, the network cannot yet separate several objects in different locations at the same time. Furthermore, the network does not predict the position of an unlearned object. However, the network should be compared to the biological functionality of a small patch of visual cortex circuitry that only distinguishes outlines of different orientation from simple cells. In such a hyper-column, lateral inhibition limits the concurrent representation of multiple orientations.

Our model successfully represents the visual information in separate pathways with less hidden units (layers d and v) than inputs. Moreover, it learns to extract and encode both 'what' and 'where', as well as movement directions. A hierarchical model, e.g. a deep-learning architecture, can be further derived based on similar ideas to achieve higher-order motion detection and generalization requirements.

Acknowledgments. This research has been partially supported by the EU projects RobotDoc under 235065 from the FP7, Marie Curie Action ITN and KSERA under n°2010-248085 for Research and Technological Development.

References

[1] Anderson, C., Van Essen, D., Olshausen, B.: Directed visual attention and the dynamic control of information flow. Neurobiology of Attention. pp. 11–17 (2005)
[2] Arbib, M., Bonaiuto, J., Rosta, E.: The mirror system hypothesis: From a macaque-like mirror system to imitation. In: Proceedings of the 6th International Conference on the Evolution of Language. pp. 3–10 (2006)
[3] Bergmann, U., von der Malsburg, C.: Self-organization of topographic bilinear networks for invariant recognition. Neural Computation pp. 1–28 (2011)
[4] Földiák, P.: Learning invariance from transformation sequences. Neural Computation 3, 194–200 (1991)
[5] Fukushima, K.: Self-organization of shift-invariant receptive fields. Neural Networks 12(6), 791–801 (1999)

[6] Grèzes, J., Decety, J.: Does visual perception of object afford action? Evidence from a neuroimaging study. Neuropsychologia 40(2), 212–222 (2002)
[7] Hyvärinen, A., Hoyer, P.: A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. Vision Research 41(18), 2413–2423 (2001)
[8] Köster, U., Lindgren, J., Gutmann, M., Hyvärinen, A.: Learning natural image structure with a horizontal product model. Independent Component Analysis and Signal Separation pp. 507–514 (2009)
[9] Lamme, V., Roelfsema, P.: The distinct modes of vision offered by feedforward and recurrent processing. Trends in Neurosciences 23(11), 571–579 (2000)
[10] Livingstone, M., Hubel, D.: Segregation of form, color, movement, and depth: anatomy, physiology, and perception. Science 240(4853), 740 (1988)
[11] Memisevic, R., Hinton, G.: Unsupervised learning of image transformations. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8 (2007)
[12] Mishkin, M., Ungerleider, L., Macko, K.: Object vision and spatial vision: Two cortical pathways. Trends in Neurosciences 6, 414–417 (1983)
[13] Olshausen, B., Anderson, C., Van Essen, D.: A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. The Journal of Neuroscience 13(11), 4700–4719 (1993)
[14] Priebe, N., Lisberger, S., Movshon, J.: Tuning for spatiotemporal frequency and speed in directionally selective neurons of macaque striate cortex. The Journal of Neuroscience 26(11), 2941–2950 (2006)
[15] Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M.: Robust object recognition with cortex-like mechanisms. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(3), 411–426 (2007)
[16] Sommer, M., Wurtz, R.: Influence of the thalamus on spatial visual processing in frontal cortex. Nature 444(7117), 374–377 (2006)
[17] Triesch, J.: Synergies between intrinsic and synaptic plasticity mechanisms. Neural Computation 19(4), 885–909 (2007)
[18] Wexler, M., Klam, F.: Movement prediction and movement production. Journal of Experimental Psychology: Human Perception and Performance 27(1), 48 (2001)
[19] Wiskott, L., Sejnowski, T.: Slow feature analysis: Unsupervised learning of invariances. Neural Computation 14(4), 715–770 (2002)