Neural Networks 24 (2011) 767–778

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Increasing robustness against background noise: Visual pattern recognition by a neocognitron

Kunihiko Fukushima *

Fuzzy Logic Systems Institute, Iizuka, Fukuoka, Japan

ARTICLE INFO

Article history:
Received 14 September 2010
Received in revised form 22 February 2011
Accepted 16 March 2011

Keywords:
Visual pattern recognition
Background noise
Neocognitron
Feature extraction
Subtractive inhibition
Root-mean-square

ABSTRACT

The *neocognitron* is a hierarchical multi-layered neural network capable of robust visual pattern recognition. It has been demonstrated that recent versions of the neocognitron exhibit excellent performance for recognizing handwritten digits. When characters are written on a noisy background, however, recognition rate was not always satisfactory. To find out the causes of vulnerability to noise, this paper analyzes the behavior of feature-extracting S-cells. It then proposes the use of subtractive inhibition to S-cells from V-cells, which calculate the average of input signals to the S-cells with a root-mean-square. Together with this, several modifications have also been applied to the neocognitron. Computer simulation shows that the new neocognitron is much more robust against background noise than the conventional ones.

1. Introduction

The author previously proposed an artificial neural network *neocognitron* for robust visual pattern recognition (Fukushima, 1980, 1988, 2003, 2010a; Fukushima & Miyake, 1982). Its architecture was initially suggested by neurophysiological findings on the visual systems of mammals (e.g., Hubel & Wiesel, 1962, 1965). It is a hierarchical multi-layered network and acquires the ability to robustly recognize visual patterns through learning.

The neocognitron consists of layers of S-cells, which resemble simple cells of the visual cortex, and layers of C-cells, which resemble complex cells. These layers of S-cells and C-cells are arranged alternately in a hierarchical manner.

Input connections of S-cells are variable and are modified through learning. After the learning, S-cells come to work as feature-extracting cells, and extract local features from stimulus images presented to the input layer (or photoreceptor array).

C-cells, whose input connections are fixed, exhibit an approximate invariance to the position of the stimuli presented within their receptive fields. We can also express that S-cells' response is spatially blurred in the succeeding layer of C-cells.

The C-cells in the highest stage work as recognition cells, which indicate the result of pattern recognition. After having finished learning, the neocognitron can recognize input patterns robustly, with little effect from deformation, change in size, or shift in position.

Varieties of modifications, extensions and applications of the neocognitron, as well as varieties of related networks, have been reported so far (Cardoso & Wichert, 2010; Elliffe, Rolls, & Stringer, 2002; Hildebrandt, 1991; LeCun et al., 1989; LeCun, Bottou, Bengio, & Haffner, 1998; Lo et al., 1995; Riesenhuber & Poggio, 1999; Satoh, Kuroiwa, Aso, & Miyake, 1999). They are all hierarchical multi-layered networks and have an architecture of *shared connections*, which is sometimes called a *convolutional net*. They also have a mechanism of pooling outputs of feature-extracting cells. The pooling operation can also be interpreted as a blurring operation. In the conventional neocognitrons, the pooling operation, which is done by C-cells, is performed by a nonlinear saturation of the weighted sum of the outputs of feature-extracting S-cells. In some networks, the pooling is realized by simply reducing the density of cells in higher layers. In some other networks, it is replaced by a MAX operation.

It has been demonstrated that neocognitrons of recent versions exhibit excellent performance for recognizing handwritten digits (Fukushima, 2003, 2010a). Most of the experiments for these neocognitrons have been made using characters written on backgrounds containing little noise.

When characters are written on a background contaminated with noise as shown in Fig. 1, however, the recognition rate of these neocognitrons is not always satisfactory. In Fig. 1(a) and (b), the background noise is a faint image of a different digit. This situation often occurs when we rewrite a character after erasing another character insufficiently. It also occurs when we write characters on a thin paper, through which the printing on the reverse side

* Permanent address: 634-3, Miwa, Machida, Tokyo 195–0054, Japan. Tel.: +81 44 988 5272; fax: +81 44 988 5272.
E-mail address: fukushima@m.ieice.org.
URL: http://www4.ocn.ne.jp/~fuku_k/index-e.html.

0893-6080/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2011.03.017

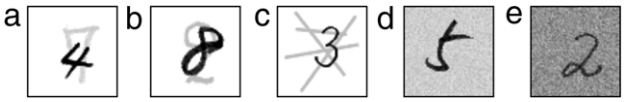


Fig. 1. Input patterns contaminated with background noise.

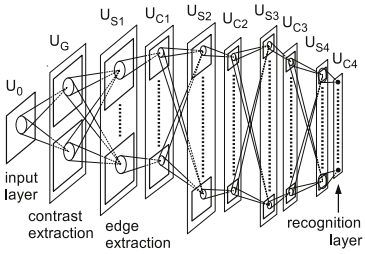


Fig. 2. The architecture of the neocognitron (Fukushima, 2003).

shows through. In Fig. 1(c), line segments are randomly located in the background. This emulates a case when we write characters on a ruled notepaper or fill in a form. In the case shown in Fig. 1(d) and (e), a white noise is superimposed on a character.

To find out the cause of vulnerability to noise, we analyze the behavior of feature-extracting S-cells. We then propose the use of subtractive inhibition to S-cells from V-cells, which calculate the average of input signals to the S-cells with a root-mean-square. Together with this, several modifications have also been applied to C-cells.[1]

In Section 5, we test the behavior of the new neocognitron by computer simulation and show that the new neocognitron is much more robust against background noise than the conventional ones.

2. Outline of the network

The neocognitron is a hierarchical multi-layered network. It consists of layers of S-cells, which resemble simple cells in the visual cortex, and layers of C-cells, which resemble complex cells. These layers of S-cells and C-cells are arranged alternately in a hierarchical manner. In other words, a number of modules, each of which consists of an S-cell layer and a C-cell layer, are connected in a cascade in the network.

S-cells are feature-extracting cells, whose input connections are variable and are modified through learning. C-cells, whose input connections are fixed, exhibit an approximate invariance to the position of the stimuli presented within their receptive fields.

The C-cells in the highest stage work as recognition cells, which indicate the result of the pattern recognition. After learning, the neocognitron can recognize input patterns robustly, with little effect from deformation, change in size, or shift in position.

Fig. 2 shows the architecture of the network that is discussed in this paper. In the figure, U_{Sl}, for example, indicates the layer of S-cells of the lth stage. The network has four stages of S- and C-cell layers.

Each layer of the network is divided into a number of sub-layers, called *cell-planes*, depending on the feature to which cells respond preferentially. Incidentally, a cell-plane is a group of cells that are arranged retinotopically and share the same set of input connections (Fukushima, 1980). As a result, all cells in a cell-plane have receptive fields of an identical characteristic, but the locations of the receptive fields differ from cell to cell.

The stimulus pattern is presented to the input layer (photoreceptor layer) U_0. A layer of contrast-extracting cells (U_G), which correspond to retinal ganglion cells or lateral geniculate nucleus cells, follows layer U_0. It consists of two cell-planes: one consisting of cells with concentric on-center receptive fields, and one consisting of cells with off-center receptive fields. The former cells extract positive contrast in brightness, whereas the latter extract negative contrast from the images presented to U_0. At the same time, the dc component of spatial frequency (namely, the averaged gray level) of the input pattern is eliminated, because the input connections to a single cell of layer U_G are designed in such a way that their total sum is equal to zero (see Appendix B for more detail). Hence the difference in the gray level of the background between Fig. 1(d) and (e), for example, is removed in the output of U_G. The output of U_G is sent to U_{S1}.

The S-cells of U_{S1} resemble simple cells in the primary visual cortex, and respond selectively to edges of a particular orientation. To be more specific, layer U_{S1} consists of $K_1 = 16$ cell-planes, and all cells in the kth cell-plane respond preferentially to edges of orientation $2\pi k/K_1 = k \times 22.5°$. As a result, the contours of the input image are decomposed into edges of every orientation.

The input connections of S-cells of higher stages are variable and are modified through learning. After having finished learning, S-cells come to work as feature-extracting cells. In higher stages, they extract more global features.

In each stage of the hierarchical network, the output of layer U_{Sl} is fed to layer U_{Cl}. C-cells, whose input connections are fixed, exhibit an approximate invariance to the position of the stimuli presented within their receptive fields. In other words, a blurred version of the response of U_{Sl} appears in U_{Cl}. The blurring operation is essential for endowing the neocognitron with an ability to recognize patterns robustly, with little effect from deformation, change in size, or shift in position of input patterns. The C-cells in the highest stage work as recognition cells, which indicate the result of the pattern recognition.

3. Feature-extracting S-cells

3.1. Response of an S-cell

3.1.1. Several types of inhibitory mechanisms

An S-cell receives both excitatory and inhibitory input signals. The excitatory signals come from C-cells of the preceding stage, and the inhibitory signal comes from a V-cell, which accompanies the S-cell. (We discuss this later in more detail. See Fig. 4.) We now compare several types of inhibitory mechanisms and discuss how they affect the behavior of the S-cell.

In the original neocognitron (Fukushima, 1980, 2003), the inhibitory signal from the V-cell works in a shunting manner. Fig. 3 shows an equivalent circuit of an S-cell. Let e be the weighted sum of all excitatory inputs, and let h be the inhibitory input. Since h works on e in a shunting manner, the output of the S-cell is determined by nonlinear function $(1 + e)/(1 + h) - 1$. If the value of the nonlinear function is negative, however, the output of the S-cell becomes zero.

Incidentally, cells with shunting inhibition were used already in the *cognitron* (Fukushima, 1975), which is an earlier model before the neocognitron. The nonlinear function $(1 + e)/(1 + h) - 1$ was suggested from the behavior of the membrane potential of a biological neuron (Eccles, 1964). It is assumed that inhibitory input works to shunt the membrane potential toward an equilibrium potential of the IPSP (inhibitory post-synaptic potential), which is more negative than the resting potential of the membrane. On the other hand, excitatory inputs raise the membrane potential toward the equilibrium potential of the EPSP (excitatory post-synaptic potential), which is much higher than the resting potential. It is also known that the threshold at which the neuron fires is much lower

[1] A preliminary short report on these modifications has appeared in Fukushima (2010b).

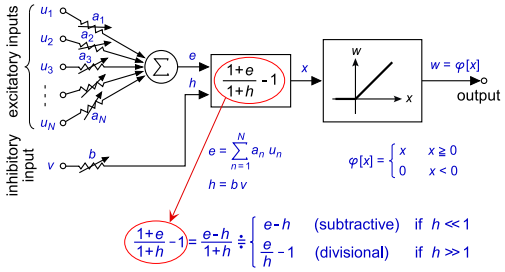


Fig. 3. Equivalent circuits of an S-cell of the original neocognitron (Fukushima, 1980, 2003). The inhibition works in a shunting manner.

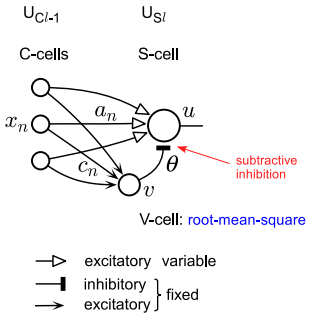


Fig. 4. Input connections converging to an S-cell. This figure shows the case of subtractive inhibition.

than the equilibrium potential of the EPSP. Hence we assume that a linear summation holds approximately for excitatory inputs.

The nonlinear function can be expressed as follows:

$$\frac{1+e}{1+h}-1=\frac{e-h}{1+h}\approx\begin{cases}e-h & \text{(subtractive) if } h\ll 1\\ \dfrac{e-h}{h}=\dfrac{e}{h}-1 & \text{(divisional) if } h\gg 1.\end{cases}\quad(1)$$

This means that the inhibition works in a subtractive manner when the inhibition is very small, and that the inhibition works in a divisional manner when the inhibition is very large.

In most of the neocognitrons of previous versions, however, parameters were set in such a way that inhibition to S-cells works in the range of divisional inhibition.

Since this paper proposes the use of subtractive inhibition, we first analyze the behavior of an S-cell with subtractive inhibition, and then compare it with the result of divisional inhibition.

3.1.2. Input signals to an S-cell

To show the essence of the process of feature extraction, we extract the circuit converging to a single S-cell and analyze its behavior. Fig. 4 shows the circuit. The S-cell of layer U_{Sl} receives excitatory signals directly from a group of C-cells, which are cells of the preceding layer U_{Cl-1}. It also receives an inhibitory signal through a V-cell, which accompanies the S-cell. The V-cell receives fixed excitatory connections from the same group of C-cells as does the S-cell, and always responds with the average intensity of the output of the C-cells.

Here we discuss the case where the inhibitory signal from the V-cell works in a subtractive manner. This means that the S-cell is almost the same as the cells usually used in conventional artificial neural networks. What is different from conventional artificial neural networks is that the V-cell calculates the average, not by a linear summation, but by a root-mean-square.

Fig. 5 shows equivalent circuit of an S-cell that receives subtractive inhibition from the accompanying V-cell. The notes written in the margin of the figure show the values of connections and responses at several points in the equivalent circuit that have finished learning. These notes, however, do not show strict values:

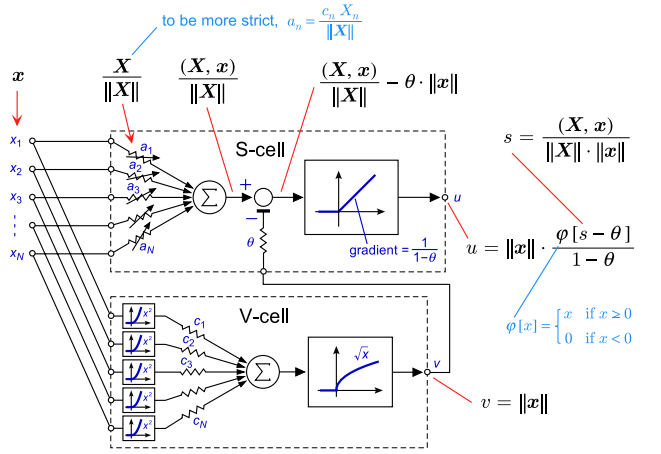


Fig. 5. Equivalent circuit of an S-cell that receives subtractive inhibition from the accompanying V-cell. The notes written in the margin show the values of connections and responses at several points in the equivalent circuit that have finished learning. These notes, however, do not show strict values: to help intuitive understanding, they show the case when $c_n = 1$.

to help intuitive understanding, they show a simpler case of $c_n = 1$.[2]

Let a_n be the strength of the excitatory variable connection to the S-cell from the nth C-cell, whose output is x_n. The output u of the S-cell is given by

$$u=\frac{1}{1-\theta}\cdot\varphi\left[\sum_n a_n x_n-\theta v\right],\quad(2)$$

where $\varphi[\ ]$ is a function defined by $\varphi[x]=\max(x,0)$. Namely, $\varphi[\ ]$ is a nonlinear function like a half-wave rectifier. The strength of the inhibitory connection is θ, which determines the threshold of the S-cell ($0<\theta<1$). The response of the V-cell is given by

$$v=\sqrt{\sum_n c_n {x_n}^2},\quad(3)$$

where c_n is the strength of the fixed excitatory connection from the nth C-cell.

We now use vector notation $\boldsymbol{x}=(x_1,x_2,\ldots,x_n,\ldots)$ to represent the response of all C-cells, from which the S-cell receive excitatory signals. We define *weighted* inner product of arbitrary two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ by

$$(\boldsymbol{x},\boldsymbol{y})=\sum_n c_n x_n y_n,\quad(4)$$

where the strength of the input connections to the V-cell, c_n, is used as the weight for the inner product. We also define the norm of a vector, using the *weighted* inner product, by $\|\boldsymbol{x}\|=\sqrt{(\boldsymbol{x},\boldsymbol{x})}$.

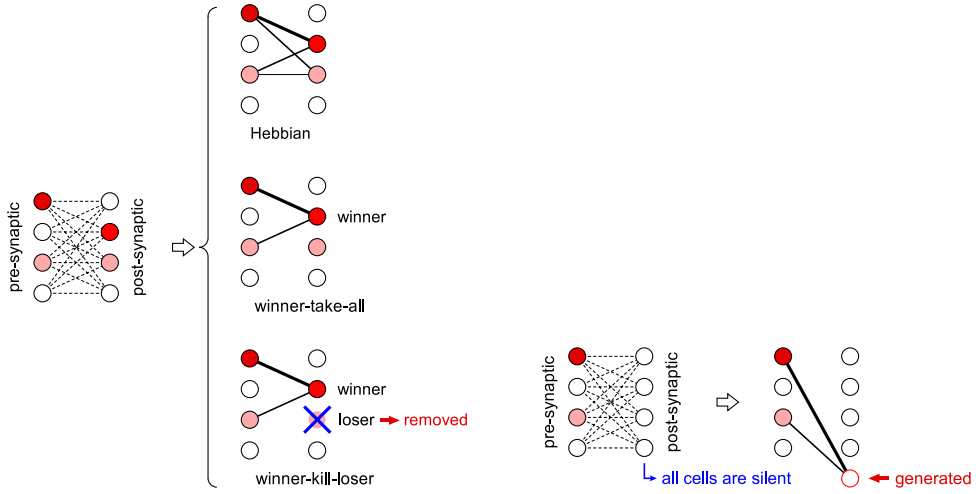
3.1.3. Competitive learning with winner-kill-loser

For the training of S-cells of layers U_{S2} and U_{S3}, we use competitive learning with winner-kill-loser (Fukushima, 2010a). Fig. 6 illustrates the learning process by the winner-kill-loser rule, in comparison with other learning rules.

The Hebbian rule, shown in the top of Fig. 6(a), is one of the most commonly used learning rules (Hebb, 1949). During the learning phase, each synaptic connection is strengthened by an amount proportional to the product of the responses of the pre- and post-synaptic cells.

In the winner-take-all rule, shown in the middle of Fig. 6(a), post-synaptic cells compete each other, and the cell from which

[2] In the figure, connections $(a_1, a_2, \ldots)$ are expressed by $\boldsymbol{X}/\|\boldsymbol{X}\|$, meaning $a_n = X_n/\|\boldsymbol{X}\|$. To be more strict, they actually are $a_n = c_n X_n/\|\boldsymbol{X}\|$.



(a) Several rules of learning.

(b) A new cell is generated when all post-synaptic cells are silent.

Fig. 6. Winner-kill-loser rule in comparison with other learning rules. In this figure, the response of each cell is represented by the depth of the color.

the largest response is elicited becomes the winner. Only the winner can have its input connections renewed. The amount of strengthening of a connection to the winner is proportional to the response of the pre-synaptic cell from which the connection is leading. Incidentally, most of the conventional neocognitrons (Fukushima, 1980, 2003) use this learning rule.

The winner-kill-loser rule, shown in the bottom of Fig. 6(a), resembles the winner-take-all rule, in the sense that only the winner learns the training stimulus. In the winner-kill-loser rule, however, not only the winner learns the training stimulus, but, at the same time, losers are removed from the network. Losers are defined as cells whose responses to the training stimulus are smaller than that of the winner, but silent cells are excluded from the losers.

If a training stimulus elicits non-zero responses from two or more S-cells, it means that preferred features of these cells resemble each other, and that they work redundantly in the network. Hence only the winner has its input connections renewed to fit more to the training vector, and the other cells, namely losers, are removed from the network.

Since silent S-cells (namely, the S-cells whose responses to the training stimulus are zero) do not join the competition, they are not removed. These cells are expected to work for extracting other features.

If all cells are silent and no winner appears for a training stimulus, as shown in Fig. 6(b), a new S-cell is generated. Incidentally, generation of new S-cells occurs also in the learning with winner-take-all rule in the neocognitron. The initial value of the input connections of the generated S-cell is proportional to the response of the pre-synaptic cells.

In the learning phase, a number of training stimuli are presented sequentially to the network. During this process, generation of new cells and removal of redundant cells are repeated in the network. In the areas where feature-extracting cells are missing in the multi-dimensional feature space (see Section 3.1.5), new cells are generated. In the areas where similar cells exist in duplicate, redundant cells are removed. By the repetition of this process, preferred features (or reference vectors) of S-cells gradually come to distribute uniformly in the multi-dimensional feature space.

3.1.4. Renewing input connections

Training of the network is performed from lower stages to higher stages: after the training of a lower stage has been completely finished, the training of the succeeding stage begins. For the training of S-cells of layer U_{Sl}, the response of C-cells of the preceding layer U_{Cl-1} works as a training stimulus. The same set of training patterns is presented to U_0, for the training of all stages except layer U_{S1}.

We use a competitive learning with *winner-kill-loser* rule for intermediate layers U_{S2} and U_{S3}, and a supervised competitive learning for the highest stage U_{S4}. Although the methods of competition are slightly different from layer to layer, the process of renewing input connections of the winners are the same for all layers.

Each time when a training pattern is presented to the input layer U_0, S-cells of U_{Sl} compete with each other, and several S-cells are selected as winners. Incidentally, the competition area of each S-cell has a shape of a hypercolumn (Fukushima, 1980).

Each S-cell usually becomes a winner several times during the training phase. Suppose an S-cell has become a winner at the tth time. We use vector $\boldsymbol{X}^{(t)}$ to represent the output of the C-cells pre-synaptic to this S-cell. Namely, $\boldsymbol{X}^{(t)}$ is the training vector for this S-cell at this moment. Excitatory connection a_n is renewed through an auxiliary variable a'_n, which increases in proportion to $X_n^{(t)}$. Namely, the amount of increase of a'_n is

$$\Delta a'_n = c_n X_n^{(t)}, \tag{5}$$

where c_n is the value of the fixed input connection to the inhibitory V-cell.

Let $\boldsymbol{X}$ be the sum of the training vectors that have made the S-cell a winner. Namely,

$$\boldsymbol{X} = \sum_t \boldsymbol{X}^{(t)}. \tag{6}$$

After having become winners for these training vectors, the strength of the auxiliary variable a'_n of this S-cell becomes

$$a'_n = \sum_t c_n X_n^{(t)} = c_n X_n. \tag{7}$$

The actual strength of the excitatory connection a_n is determined in proportion to a'_n, but the total strength of all connections to a single S-cell, $\sqrt{\sum_n {a_n}^2/c_n}$, is always kept constant.

To be more specific, the excitatory connection a_n is calculated from a'_n by

$$a_n = \frac{a'_n}{\sqrt{\sum_\nu {a'_\nu}^2/c_\nu}}. \tag{8}$$

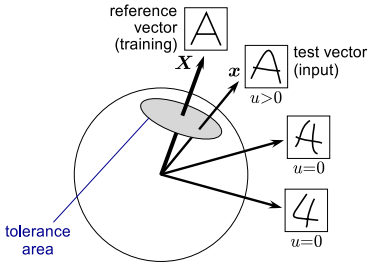


Fig. 7. Response of an S-cell in a multi-dimensional feature space.

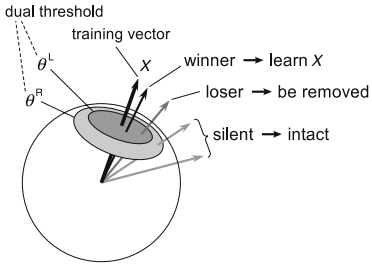


Fig. 8. Competition in the multi-dimensional feature space. The threshold for the learning θ^{L}, which is higher than the threshold for the recognition θ^{R}, is used during the competition. The winner is tuned up by learning the training vector, and losers are removed from the network. Silent cells are kept intact, because they do not join the competition.

Since $\sqrt{\sum_v {a'_v}^2/c_v} = \|\boldsymbol{X}\|$ holds, we have

$$a_n = \frac{c_n X_n}{\|\boldsymbol{X}\|}. \tag{9}$$

3.1.5. Response of an S-cell

Using weighted inner product defined by (4), we have

$$\sum_n a_n x_n = \frac{(\boldsymbol{X}, \boldsymbol{x})}{\|\boldsymbol{X}\|} \tag{10}$$

from (9). We also have

$$v = \|\boldsymbol{x}\| \tag{11}$$

from (3).

Hence the response of the S-cell, which is given by (2), can also be expressed by

$$u = \|\boldsymbol{x}\| \cdot \frac{\varphi[s-\theta]}{1-\theta}, \tag{12}$$

where

$$s = \frac{(\boldsymbol{X}, \boldsymbol{x})}{\|\boldsymbol{X}\| \cdot \|\boldsymbol{x}\|}. \tag{13}$$

In the multi-dimensional feature space, s shows a kind of similarity between $\boldsymbol{x}$ and $\boldsymbol{X}$ (Fig. 7). We call $\boldsymbol{X}$, which is the sum of the training vectors, the reference vector of the S-cell. Using a neurophysiological term, we can also express that $\boldsymbol{X}$ is the preferred feature of the S-cell.

The second term $\varphi[s-\theta]/(1-\theta)$ in (12) takes a maximum value 1 if the stimulus vector $\boldsymbol{x}$ is identical to the reference vector $\boldsymbol{X}$, and becomes 0 when the similarity s is less than the threshold θ of the cell. In the multi-dimensional feature space, the area that satisfies $s < \theta$ becomes the tolerance area in feature extraction by the S-cell, and the threshold θ determines the size of the tolerance area. In other words, a non-zero response is elicited from the S-cell, if and only if the stimulus vector $\boldsymbol{x}$ is within a tolerance area around the reference vector $\boldsymbol{X}$.

The selectivity of the S-cell to its preferred feature (or the reference vector) can thus be controlled by the threshold θ. A higher value of θ produces a smaller tolerance area. If the threshold is low, the radius of the tolerance area becomes large, and the S-cell responds even to features somewhat deformed from the reference vector.

This characteristic that an S-cell is active when $s < \theta$ and silent when $s < \theta$ is essential for a successful self-organization by the winner-kill-loser rule. Under the winner-kill-loser rule, all non-silent S-cells that fail to be a winner are removed automatically from the network, because they are redundant in the network. If the preferred feature (or the reference vector) of an S-cell is different from the training vector by more than θ, however, the S-cell becomes silent and can be kept intact. The S-cell is expected to work for extracting another feature. If all S-cells are silent to the training vector, a new S-cell is generated. Generation and removal of S-cells are thus repeated in the network, and S-cells gradually come to distribute uniformly in the multi-dimensional feature space. The distance between adjacent cells approaches to θ. This means that S-cells come to behave like grandmother cells, and only one S-cell responds to an individual feature. During the learning, the threshold of S-cells needs to be high enough to produce a sufficient number of feature-extracting S-cells in the layer.

In the recognition phase, however, a behavior like grandmother cells is not desirable for S-cells. If the threshold is high, a different S-cell often comes to respond in the network, when a stimulus feature is slightly deformed. This decreases the ability of the network to recognize deformed patterns robustly. If the threshold is low, however, S-cells respond even to features somewhat deformed from their reference vectors. This makes a situation like a population coding of features rather than grandmother cell theory: many S-cells respond to a single feature if the response of an entire layer is observed. This situation of low threshold in the recognition phase usually produces a better recognition rate of the neocognitron.

Hence we use dual threshold of S-cells for the learning and the recognition phases (Fukushima & Tanigawa, 1996). In the recognition phase after having finished the learning, the threshold of S-cells is set to a lower value θ^{R} than the threshold θ^{L} for the learning.

Fig. 8 illustrates the responses of S-cells in the multi-dimensional feature space, where dual threshold is used. Since silent S-cells (namely, the S-cells whose response to the training vector are zero) do not join the competition during the learning, they are not removed, even if they might yield non-zero responses under the lower threshold θ^{R} for the recognition phase.

3.1.6. Comparison with other types of inhibition

To discuss how the difference in inhibitory mechanism affects the behavior of S-cells, we analyze the responses of S-cells that have other types of inhibition, and compare them with that of an S-cell with subtractive inhibition.

If the inhibition works in a divisional manner, as shown in Fig. 9, the response of the S-cell is given by

$$u = \frac{\theta}{1-\theta} \cdot \varphi\left[\frac{1}{v}\sum_n a_n x_n - \theta\right]. \tag{14}$$

Substituting (10) and (11) into (14), and using similarity s defined by (13), we have

$$u = \frac{\varphi[s-\theta]}{1-\theta}. \tag{15}$$

This is a case of the neocognitron of a previous version (Fukushima, 2010a).

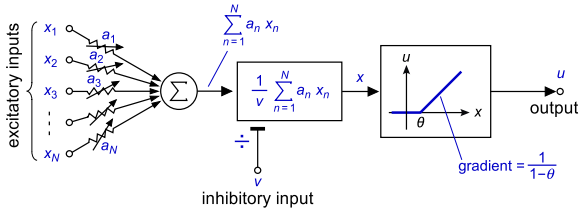


Fig. 9. Equivalent circuits of an S-cell with divisional inhibition.

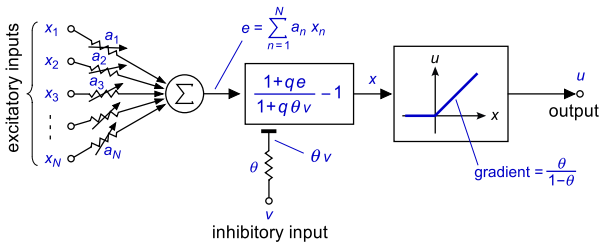


Fig. 10. Equivalent circuits of an S-cell with shunting inhibition.

In the initial versions of the neocognitron (Fukushima, 1980, 2003; Fukushima & Miyake, 1982), S-cell had shunting inhibition, as shown in Fig. 10. The output of the S-cell with shunting inhibition is given by

$$u = \frac{\theta}{1-\theta} \cdot \varphi\left[\frac{1+qe}{1+q\theta v} - 1\right], \tag{16}$$

where q is a positive constant. Since $e = \sum_n a_n x_n = (\boldsymbol{X}, \boldsymbol{x})/\|\boldsymbol{X}\|$ and $v = \|\boldsymbol{x}\|$, (16) reduces to

$$u = \frac{q\theta\|\boldsymbol{x}\|}{1+q\theta\|\boldsymbol{x}\|} \cdot \frac{\varphi[s-\theta]}{1-\theta}. \tag{17}$$

When $q\theta v \ll 1.0$ holds, the inhibition comes to work in a subtractive manner, and (17) reduces approximately to (12) if we exclude proportional factor $q\theta$. When $q\theta v \gg 1.0$ holds, the inhibition comes to work in a divisional manner, and (17) reduces approximately to (15). As was discussed in 3.1.1, in most versions of the initial neocognitron, parameter q was set so large that S-cells work in the range of divisional inhibition.

We can see, from (12), (15) and (17), that S-cells are silent if and only if $s < \theta$ holds, regardless of the types of inhibition. In this sense, all of these S-cells are suited for the learning with winner-kill-loser rule. This characteristic is obtained by the use of inhibitory V-cells, which calculate the average intensity of input signals by a root-mean-square.

When input patterns are contaminated with a background noise, however, difference in the inhibitory mechanism causes a large difference in the recognition rate of the neocognitron. This is discussed below.

3.2. Effect of noise on S-cells

If there is no background noise in the input pattern, the characteristics of (15) is desirable for feature-extracting S-cells. This is the case for a previous neocognitron, in which S-cells have divisional inhibition. The response of an S-cell is determined only by the similarity s between the input stimulus $\boldsymbol{x}$ and the training feature $\boldsymbol{X}$. It is not affected by the strength of the input stimulus $\boldsymbol{x}$. Hence S-cells can extract features robustly without being affected, say, by a gradual non-uniformity in thickness, darkness or contrast in an input pattern.

If an input character is written on a noisy background, such like the patterns shown in Fig. 1, however, interference from the background noise becomes serious. The background noise could be

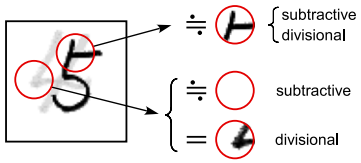


Fig. 11. Difference in the responses of S-cells between subtractive (the new neocognitron) and divisional (the previous neocognitron) inhibition. When the receptive field of an S-cell covers only a part of the faint background character as shown by the left circle, the cell's response becomes large despite of the weak intensity of the stimulus, if the inhibition works in a divisional manner. If the inhibition works in a subtractive manner, however, the response of the cell remains small in proportion to the weak intensity of the stimulus. When the receptive field covers parts of both background and foreground characters together as shown by the upper circle, S-cell's response is almost the same, either with divisional or subtractive inhibition.

a faint image of other character, randomly placed line segments, white noise, or the like.

It should be noted here, however, that the difference in the gray level (namely, the dc component of the spatial frequency) of the background has already been eliminated at the stage of U_G. In other words, only brightness contrast appears in U_G. Hence, the response of S-cells is affected only by the variance of the amplitude, and not by the spatial average, of the background noise. To be more concrete, the difference in the gray level of the background between Fig. 1(d) and (e), for example, does not affect the response of S-cells.

Fig. 11 illustrates a situation when the background noise is a faint image of other character. If S-cells have divisional inhibition, features of the faint background character elicit large responses from some S-cells, although they are very weak in the input image. This problem occurs for some S-cells whose receptive fields cover only background features (as shown by the left circle in Fig. 11), because the response of an S-cell is determined only by similarity s, and not adjusted by the weak stimulus intensity $\|\boldsymbol{x}\|$ in the receptive field. Correct recognition of the foreground character is largely interfered by strong responses elicited from these irrelevant features.

If S-cells have subtractive inhibition, however, the strength of irrelevant features from the interfering pattern can remain low, in proportion to the weak intensity of the background noise in the input pattern. In other words, it is more desirable for S-cells to have characteristics like (12) under noisy environment, because the strength of extracted features becomes proportional to the intensity $\|\boldsymbol{x}\|$ of the input stimuli to receptive fields of individual cells.

When features of both background and foreground characters fall in a receptive field of a single S-cell together (as shown by the upper circle in Fig. 11), the effect of the weak background feature on the response of the S-cell is not serious, regardless of the type of inhibition. This is because the cell's response is normalized to the strong intensity of the foreground character, even with the divisional inhibition. The background stimulus, which is relatively much weaker than the foreground stimulus, scarcely affects the response of the S-cell.

The new neocognitron, in which S-cells have subtractive inhibition and have characteristics like (12), can thus keep its low recognition error under noisy environment. (See Section 5 for the result of computer simulation.)

4. C-cells: Blur by root-mean-square operation

A C-cell has fixed excitatory connections from a group of S-cells of the corresponding cell-planes of S-cells. Through these connections, each C-cell averages the responses of S-cells whose receptive field locations are slightly deviated. In other words, S-cells' response is spatially blurred in the succeeding cell-planes

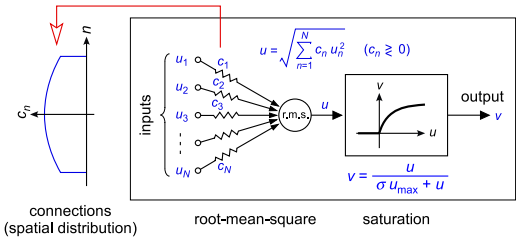


Fig. 12. A block diagram of a C-cell, which calculate the root-mean-square of its inputs.

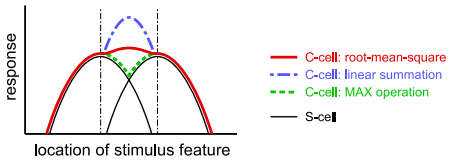


Fig. 13. Response of a C-cell: comparison of different averaging operations.

of C-cells. The averaging operation is important, not only for endowing neural networks with an ability to recognize deformed patterns robustly, but also for smoothing additive random noise contained in the responses of S-cells.

Incidentally, pooling the responses of S-cells can also be made by a MAX operation (Riesenhuber & Poggio, 1999). The MAX operation, however, is vulnerable to noise, because the response of a C-cell is determined by the response of a single maximum output S-cell only, where smoothing noise by averaging cannot work. This is another advantage of the averaging operation over the MAX operation.

Fig. 12 shows a block diagram of a C-cell. In the new neocognitron, a C-cell averages its input signals, not by a weighted linear summation, but by a root-mean-square.

To reduce the computational cost, the spatial density of S-cells in a cell-plane is usually designed to be sparse, and a C-cell averages the responses of a small number of S-cells. Hence the output of the C-cell fluctuates with the shift in location of its preferred feature. The fluctuation can be made smaller by the root-mean-square than by a linear summation or a MAX operation. Fig. 13 illustrate this situation, showing the response of a C-cell that receives signals from two S-cells, which have receptive fields at slightly different locations. With the shift in location of a stimulus feature, the responses of the S-cells change. Hence the response of the C-cell also changes with the shift, but the fluctuation of the response is the smallest by the root-mean-square operation.

Since a role of a C-cell is to detect whether any of its presynaptic S-cells is active, it is better to have some saturation in the input-to-output characteristic. If there is no saturation, the response of C-cells is affected by the size of stimulus patterns. For example, a line or edge larger than the S-cells' receptive fields elicits responses from several S-cells connected to the same C-cell, and the C-cell's response would be increased. Although the problem of size invariance can be solved also by a MAX operation, we chose a root-mean-square operation with saturation, because of the robustness against noise, which is produced by averaging operation.

In the original neocognitron (Fukushima, 1980, 1988, 2003; Fukushima & Miyake, 1982), in which C-cells calculate mean values by a linear summation, the saturation is determined by a function like $v = u/(\sigma + u)$, where the value of σ is a constant. Although S-cells had shunting inhibition, they worked nearly in the range of divisional inhibition as was mentioned in 3.1.6. Hence the maximum output of S-cells were always around 1.0, and it was not necessary to change σ adaptively to the level of input to C-cells.

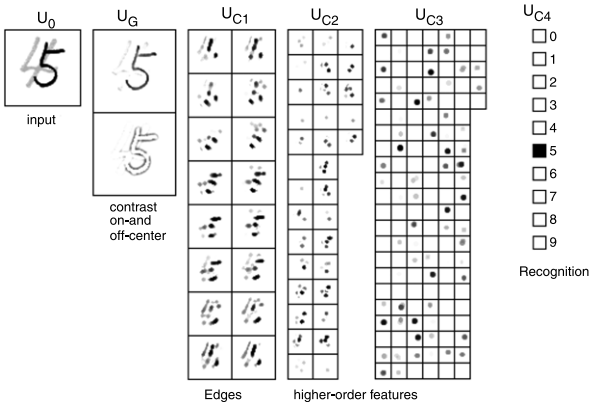


Fig. 14. An example of the response of the new neocognitron. Character '5' in the foreground, which is disturbed by the faint background character '4', is recognized correctly.

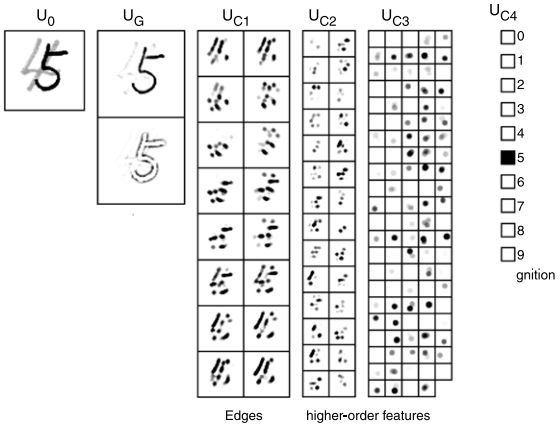


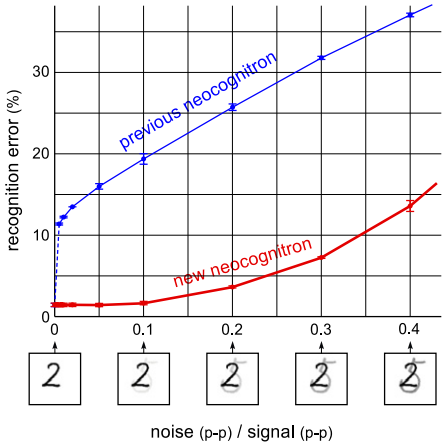
Fig. 15. An example of the response of a previous neocognitron that uses S-cells with divisional inhibition. It can be seen, for example, that, in U_{C1}, interference from the faint background character '4' appears much stronger than in the new neocognitron.

In the previous neocognitron (Fukushima, 2010a), in which C-cells calculate mean values by linear summation, the saturation is determined by a square-root function, namely $v = \sqrt{u}$. When C-cells with this saturation are followed by an S-cell with divisional inhibition, the response of the S-cell, which is given by (15), can be stable without being affected by the intensity of the stimuli to the C-cells. If input patterns do not contain background noise, saturation by a square-root works satisfactory. If input patterns contain some background noise, however, the square-root nonlinearity is not desirable. A small background noise is exaggerated by the square-root nonlinearity, because $dv/du = d\sqrt{u}/du \to \infty$ for $u \to 0$. In this case, it is more appropriate to choose a nonlinearity by which the strength of background noise can be kept as small as in the input pattern.

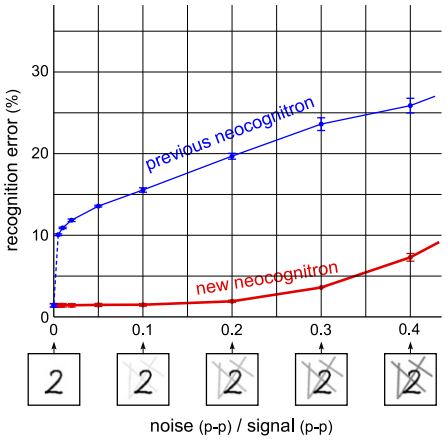
In the new neocognitron, the saturation is controlled adaptively by the maximum value of input signals to C-cells of the layer. In Fig. 12, u is the weighted root-mean-square of the input signals to a C-cell. Let u_{max} be the maximum value of u among all C-cells of the layer. The output of a C-cell of the layer is given by

$$v = \frac{u}{\sigma u_{max} + u}, \tag{18}$$

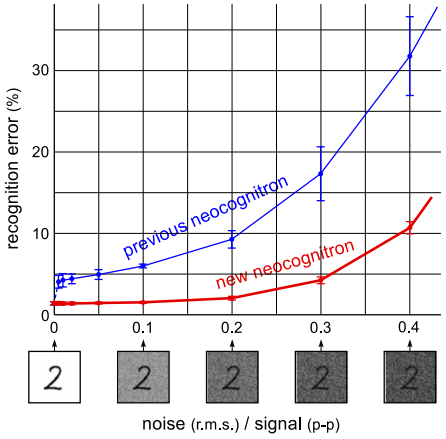
where σ is a positive constant. With this operation of saturation, the response of the C-cell layer can be made insensitive to the strength of the entire image given to the input layer. Individual C-cells in the layer, however, can still keep sensitivity to the strength of stimuli given to their respective receptive fields.



(a) The background noise is a faint image of a different digit.



(b) Randomly located line segments in the background.



(c) White noise is superimposed on the test digit.

Fig. 16. Recognition error of the new and previous neocognitrons under different levels of background noise.

Similarly to the previous neocognitron (Fukushima, 2010a), C-cells of U_{C1} and U_{C2} have inhibitory surround in their input connections. Namely, the excitatory connections, which produce a blur, are surrounded annularly by inhibitory connections, where each C-cell receives both excitatory and inhibitory signals from S-cells of the same preferred feature. Furthermore, in U_{C1}, the surround inhibition is suppressed by disinhibition from signals of S-cells of orthogonal preferred orientation.

In the new neocognitron, however, both excitatory and inhibitory inputs are not averaged linearly, but are averaged by a root-mean-square. The same is true for the signals of disinhibition. (See Appendix D for more exact mathematical description.)

5. Computer simulation

We tested the behavior of the new neocognitron by computer simulation. Fig. 14 shows a typical response of the network that has finished the learning. The responses of layers U_0, U_G and layers of C-cells of all stages are displayed in series from left to right. The rightmost layer, U_{C4}, shows the final result of recognition. The input character is written on a noisy background. In this example, character '5' in the foreground is recognized correctly, although there is a faint disturbing character '4' in the background.

To see how inhibitory mechanisms affect the behavior of neocognitrons, we also simulated a neocognitron that has S-cells with divisional inhibition (Fig. 15), giving the same stimulus pattern to U_0. It is the neocognitron of a previous version (Fukushima, 2010a). Although the target foreground character is also recognized correctly in this particular case, irrelevant local features of interfering background character appear strongly in intermediate layers. In U_{C1}, for example, where a blurred version of the response of U_{S1} appears, irrelevant edge components of the interfering background character '4' is detected more strongly by the previous neocognitron than by the new neocognitron.

We measured recognition error of the new neocognitron using handwritten digits (free writing) randomly sampled from the ETL1 database.[3] To be more specific, we used 3000 patterns (300 patterns for each digit) for the learning. During the learning, no background noise is added to the learning patterns. After the learning has completely finished, we measured the recognition error for a blind test set of 5000 patterns. We made this experiment twice for each condition, using different learning and test sets randomly sampled from the ETL1, and averaged the results of the two experiments. Fig. 16 shows how the recognition error changes with different levels of background noise.

We tested three different types of background noise. In one case, which is shown in Fig. 16(a), the background noise is a faint image of a different digit, which is also sampled randomly from the ETL1 database. In the second case, which is shown in Fig. 16(b), line segments are randomly located in the background. In the third case, which is shown in Fig. 16(c), a white noise is superimposed on the test digit.

[3] http://www.is.aist.go.jp/etlcdb/#English.

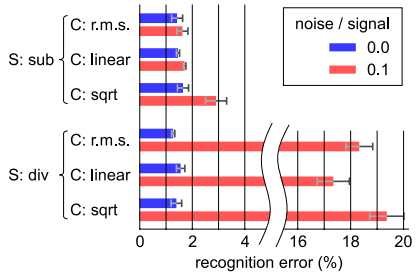


Fig. 17. Recognition error under various combinations of different types of S- and C-cells. Background noise is a faint image of other character. Two cases where the noise (peak-to-peak) to signal (peak-to-peak) ratio is 0.1 and 0.0 are shown.

Red (thick) lines in the figures show the results for the new neocognitron, and blue (thin) lines show the results for the previous neocognitron (Fukushima, 2010a). In the new neocognitron, S-cells have subtractive inhibition, and C-cells average their inputs by a root-mean-square. In the previous neocognitron, S-cells have divisional inhibition, and C-cells average their inputs by a linear summation and have saturation by a square-root function.

Although the recognition error of the new neocognitron under noiseless condition ($1.43\% \pm 0.21\%$) is not so different from that of the previous one ($1.40\% \pm 0.20\%$), the increase in recognition error under noisy background is much smaller in the new neocognitron. Under noisy conditions, the new neocognitron exhibits a much better result than the previous one.

Since various processes are combined in the new and the previous neocognitrons, we test here how each process contributes to reducing the recognition error. Fig. 17 shows the recognition error under various combinations of different types of S- and C-cells. Red (or pale) bars show the recognition errors when foreground patterns are contaminated with a faint image of other character, where the noise (peak-to-peak) to signal (peak-to-peak) ratio is 0.1. Blue (or dark) bars show the recognition errors when there is no background noise. "S-cell: sub" and "S-cell: div" show the cases where S-cells have subtractive and divisional inhibition, respectively. "C-cell: r.m.s." means that C-cells average their inputs by a root-mean-square and have saturation given by (18). "C-cell: linear" means that C-cells average their inputs by a linear summation and have a saturation given by (18). "C-cell: sqrt" means that C-cells average their inputs by a linear summation and have a saturation by a square-root function.

Incidentally, the new neocognitron has the combination of "S-cell: sub" and "C-cell: r.m.s.", while the previous neocognitron has the combination of "S-cell: div" and "C-cell: sqrt".

From this figure, and also from Fig. 16, we can clearly conclude that the subtractive inhibition to S-cells is much better than the divisional inhibition regardless of the types of C-cells, if input patterns have a possibility of contamination by background noise.

As for the types of C-cells, we can easily conclude that "C-cell: sqrt" is not suited under noisy background. Although "C-cell: r.m.s." produces smallest recognition error, the difference between "C-cell: r.m.s." and "C-cell: linear" is not so large in Fig. 17.

The recognition error is affected directly and indirectly by the threshold values of S-cells, θ_l^{L} and θ_l^{R}. For example, the threshold values affect the number of cell-planes generated, which in turn affect the recognition error. It should also be noted that the optimal values of the thresholds change depending on the types of S- and C-cells used in the network. We tried to search the best combination of the thresholds, but have not reached the final result yet, because there are a lot of possible combinations. Figs. 16 and 17 show the results under semi-optimal threshold values. Since "C-cell: r.m.s." had tendency of producing slightly better recognition rate than "C-cell: linear" under various combination of the threshold values, we chose "C-cell: r.m.s." for the new neocognitron.

6. Discussions

We modified the conventional neocognitron and made it more robust against background noise. The major modification is the use of subtractive inhibition to S-cells from V-cells, which calculate the average of input signals to the S-cells with a root-mean-square. Together with this, several modifications have also been applied to the C-cells. Through these modifications, the new neocognitron can keep its high recognition rate even under background noise.

In the previous neocognitron (Fukushima, 2010a), inhibitory signals to S-cells worked in a divisional manner, and each S-cell obtained the ability to normalize stimulus intensity and were able to extract a stimulus feature independently of the stimulus intensity. This was a desirable characteristic for recognizing patterns when stimulus patterns do not contain noise. When stimulus patterns have some background noise, however, normalization of stimulus intensity sometimes does harm to correct recognition of the patterns. However weak is a background noise, it might elicit a large response from an S-cell, if the receptive field of the S-cell covers only the background noise.

We proposed the use of subtractive inhibition for S-cells. When an S-cell extracts a feature, the strength of its response comes to be determined, not only by the similarity between the stimulus feature and the preferred feature of the cell, but also by the stimulus intensity. The effect of the background noise remains small, if the noise is weak. The new neocognitron can thus keep its high recognition rate even under background noise.

Each S-cell of the neocognitron is accompanied by a V-cell. The V-cell always responds to the average intensity of the signals to the S-cell. The average is calculated, not by a linear summation, but by a root-mean-square. By the use of a V-cell of the root-mean-square characteristic, the S-cell comes to yield a non-zero response if and only if s, the similarity between the stimulus feature and the preferred feature of the cell, is larger than threshold θ. The function of extracting features that satisfy $s < \theta$ is essential for robust recognition of patterns and also for a successful self-organization with winner-kill-loser rule. S-cells can have this function, not only in the new neocognitron, where the inhibition from the V-cell works in a subtractive manner, but also in the conventional neocognitrons, where the inhibition works in a divisional (Fukushima, 2010a) or a shunting manner (Fukushima, 2003). What is important for producing this function is the root-mean-square characteristic of the V-cells and the threshold operation by the S-cells. By the S-cells with subtractive inhibition, the strength of the response of each S-cell can be adjusted further by the stimulus intensity in its receptive field.

A C-cell receives excitatory connections from a group of S-cells that extract the same feature but at slightly different locations. A C-cell thus averages the responses of S-cells. In other words, S-cells' response is spatially blurred in the succeeding cell-planes of C-cells. The averaging operation is important, not only for endowing neural networks with an ability to recognize shifted or deformed patterns robustly, but also for smoothing random noise.

For the averaging operation, the new neocognitron uses a root-mean-square instead of a linear summation, which is used in the conventional neocognitrons. Although the computer simulation (Fig. 17) showed that the averaging by a root-mean-square yielded a smaller recognition error than by a linear summation, the difference was not so large. We will need a further investigation before concluding that the root-mean-square averaging is the best in all cases. This is a problem left for the future.

The idea of the use of root-mean-square operation coincides with the energy model for complex cells of the primary visual cortex, in the sense that the response of a complex cell is determined by the sum of squared responses of the pre-synaptic simple cells (e.g., Carandini et al., 2005; Heeger, 1992; Ohzawa,

DeAngelis, & Freeman, 1990). The fact that the energy model can explain a number of neurophysiological functions of complex cells might suggest the biological plausibility of the root-mean-square operation in the C-cells.

Acknowledgments

The author thanks Prof. Isao Hayashi (Kansai University), Dr. Hayaru Shouno (University of Electro-Communications) and Dr. Masayuki Kikuchi (Tokyo University of Technology) for helpful comments and discussions. He also thanks Mr. Akira Mochizuki for gathering some data by computer simulation. This work was partially supported from Kansai University by Strategic Project to Support the Formation of Research Bases at Private Universities: Matching Fund Subsidy from MEXT, 2008–2012.

Appendix A. Architecture of the network

The architecture and most of the parameters of the new neocognitron are the same as those of the previous neocognitron (Fukushima, 2010a). Namely, the basic architecture of the network, the scale of the network, the sizes and values of fixed connections between cells, the mechanism of disinhibition to the surround inhibition in U_{C1}, and so on, are the same as those for the previous neocognitron.

Main differences between them are in the input-to-output characteristics of S- and C-cells, and the threshold values of the S-cells. We discuss mathematical descriptions of the network below, but the explanations are abbreviated where they are the same as for the previous neocognitron.

Appendix B. Contrast-extracting layer

Contrast-extracting layer U_G of the new neocognitron is identical to that of the previous neocognitron. Let the output of a photoreceptor cell of input layer U_0 be $u_0(\boldsymbol{n})$, where $\boldsymbol{n}$ represent the location of the cell. The output of a contrast-extracting cell of layer U_G, whose receptive field center is located at $\boldsymbol{n}$, is given by

$$u_G(\boldsymbol{n}, k) = \varphi\left[(-1)^k \sum_{|\boldsymbol{\nu}|\le A_G} a_G(\boldsymbol{\nu}) \cdot u_0(\boldsymbol{n}+\boldsymbol{\nu})\right], \quad (k = 1, 2), \quad \text{(B.1)}$$

where $\varphi[\ ]$ is a function defined by $\varphi[x] = \max(x, 0)$. Parameter $a_G(\boldsymbol{\xi})$ represents the strength of fixed connections to the cell and takes the shape of a Mexican hat. Layer U_G has two cell-planes: one consisting of on-center cells ($k = 2$) and one of off-center cells ($k = 1$). A_G denotes the radius of summation range of $\boldsymbol{\nu}$, that is, the size of spatial spread of the input connections to a cell.

The input connections to a single cell of layer U_G are designed in such a way that their total sum is equal to zero. In other words, the connection $a_G(\boldsymbol{\xi})$ is designed so as to satisfy

$$\sum_{|\boldsymbol{\nu}|<A_G} a_G(\boldsymbol{\nu}) = 0. \quad \text{(B.2)}$$

This means that the dc component of spatial frequency of the input pattern is eliminated in the contrast-extracting layer U_G. As a result, the output from layer U_G is zero in the area where the brightness of the input pattern is flat.

Appendix C. S-cell layers

C.1. Response of an S-cell

Let $u_{Sl}(\boldsymbol{n}, k)$ and $u_{Cl}(\boldsymbol{n}, k)$ be the output of an S-cell and a C-cell of the kth cell-plane of the lth stage, respectively, where $\boldsymbol{n}$ represents the location of the receptive field center of the cells. Layer U_{Sl} contains not only S-cells but also V-cells, whose output is represented by $v_l(\boldsymbol{n})$. The outputs of an S-cell and a V-cell are given by

$$u_{Sl}(\boldsymbol{n}, k) = \frac{1}{1-\theta_l} \cdot \varphi\left[\sum_{\kappa=1}^{K_{Cl-1}} \sum_{|\boldsymbol{\nu}|\le A_{Sl}} a_{Sl}(\boldsymbol{\nu}, \kappa, k) \times u_{Cl-1}(\boldsymbol{n}+\boldsymbol{\nu}, \kappa) - \theta_l \cdot v_l(\boldsymbol{n})\right], \quad \text{(C.1)}$$

where

$$v_l(\boldsymbol{n}) = \sqrt{\sum_{\kappa=1}^{K_{Cl-1}} \sum_{|\boldsymbol{\nu}|\le A_{Sl}} c_{Sl}(\boldsymbol{\nu}) \cdot \left\{u_{Cl-1}(\boldsymbol{n}+\boldsymbol{\nu}, \kappa)\right\}^2}. \quad \text{(C.2)}$$

If $l = 1$ in (C.1) and (C.2), $u_{Cl-1}(\boldsymbol{n}, k)$ stands for $u_G(\boldsymbol{n}, k)$, and we have $K_{Cl-1} = 2$.

Parameter $a_{Sl}(\boldsymbol{\nu}, \kappa, k)$ (≥ 0) is the strength of variable excitatory connection coming from C-cell $u_{Cl-1}(\boldsymbol{n}+\boldsymbol{\nu}, \kappa)$ of the preceding stage. It should be noted here that all cells in a cell-plane share the same set of input connections, hence $a_{Sl}(\boldsymbol{\nu}, \kappa, k)$ is independent of $\boldsymbol{n}$. A_{Sl} denotes the radius of summation range of $\boldsymbol{\nu}$, that is, the size of spatial spread of input connections to the S-cell.

Parameter $c_{Sl}(\boldsymbol{\nu})$ represents the strength of the fixed excitatory connections to the V-cell, and is a monotonically decreasing function of $|\boldsymbol{\nu}|$. It is also used as a weighting function for training connections $a_{Sl}(\boldsymbol{\nu}, \kappa, k)$, as is shown in (C.3).

The positive constant θ_l is the threshold of the S-cell and determines the selectivity in extracting features. The method of dual threshold is used for the learning and recognition for layers U_{S2} and U_{S3}. A higher threshold value is used in the learning phase than in the recognition phase. In the computer simulation, thresholds θ_l^{R} for the recognition phase are: $\theta_1^{\mathrm{R}} = 0.50$, $\theta_2^{\mathrm{R}} = 0.50$, $\theta_3^{\mathrm{R}} = 0.51$ and $\theta_4^{\mathrm{R}} = 0.0$. Thresholds θ_l^{L} for the learning phase are: $\theta_2^{\mathrm{L}} = 0.67$, $\theta_3^{\mathrm{L}} = 0.66$ and $\theta_4^{\mathrm{L}} = 0.0$.

C.2. Training S-cell layers

In the hierarchical network of the neocognitron, training (or learning) is performed from lower stages to higher stages: after the training of a lower stage has been completely finished, the training of the succeeding stage begins. The same set of training patterns is used for the training of all stages except layer U_{S1}.

C.2.1. Renewing connections

Every time when a training pattern is presented to the input layer, a small number of *seed-cells* are selected. The method of selecting seed-cells is discussed later in Appendix C.2.2.

Although the method for selecting seed-cells during learning is slightly different between layers, the rule for renewing variable connections $a_{Sl}(\boldsymbol{\nu}, \kappa, k)$ is the same for all layers, once the seed-cells have been determined. The connections are renewed depending on the responses of the pre-synaptic cells (namely, the C-cells of the preceding stage).

Connection $a_{Sl}(\boldsymbol{\nu}, \kappa, k)$ is determined through an auxiliary variable $a'_{Sl}(\boldsymbol{\nu}, \kappa, k)$. Let cell $u_{Sl}(\hat{\boldsymbol{n}}, \hat{k})$ be selected as a seed-cell at a certain time. The auxiliary variable $a'_{Sl}(\boldsymbol{\nu}, \kappa, \hat{k})$ to this seed-cell is increased by the following amount:

$$\Delta a'_{Sl}(\boldsymbol{\nu}, \kappa, \hat{k}) = c_{Sl}(\boldsymbol{\nu}) \cdot u_{Cl-1}(\hat{\boldsymbol{n}}+\boldsymbol{\nu}, \kappa), \quad \text{(C.3)}$$

where $c_{Sl}(\boldsymbol{\nu})$ is the value of the fixed input connection to the inhibitory V-cell.

The excitatory connection $a_{Sl}(\boldsymbol{v}, \kappa, \hat{k})$ to this seed-cell, and consequently to all the S-cells in the same cell-plane as the seed-cell, is calculated from the value of $a'_{Sl}(\boldsymbol{v}, \kappa, \hat{k})$ by

$$a_{Sl}(\boldsymbol{v}, \kappa, \hat{k}) = a'_{Sl}(\boldsymbol{v}, \kappa, \hat{k}) / b_{Sl}(\hat{k}), \tag{C.4}$$

where

$$b_{Sl}(\hat{k}) = \sqrt{\sum_{\kappa=1}^{K_{Cl-1}} \sum_{|\boldsymbol{v}| \le A_{Sl}} \frac{\left\{a'_{Sl}(\boldsymbol{v}, \kappa, \hat{k})\right\}^2}{c_{Sl}(\boldsymbol{v})}}. \tag{C.5}$$

Once the input connections to a seed-cell have been renewed, all cells in the cell-plane come to have the same set of input connections as the seed-cell, because all cells in the cell-plane share the same set of input connections.

C.2.2. Selecting seed-cells

Since the methods of selecting seed-cells are identical to those of the previous neocognitron (Fukushima, 2010a), we explain them only briefly.

Edge-extracting layer. Layer U_{S1}, namely, the S-cell layer of the 1st stage, is an edge-extracting layer. It has $K_{S1} = 16$ cell-planes, each of which consists of edge-extracting cells of a particular preferred orientation. Preferred orientations of the cell-planes, namely, the orientations of the training patterns, are chosen at an interval of 22.5°.

The S-cells of this layer are trained with supervised learning. To train a cell-plane, the "teacher" presents a training pattern, namely a straight edge of a particular orientation, to the input layer U_0. The teacher then points out the location of the feature, which, in this particular case, can be an arbitrary point on the edge. The cell whose receptive field center coincides with the location of the feature takes the place of the seed-cell of the cell-plane, and the process of strengthening connections occurs automatically.

Competitive learning for intermediate layers. We use winner-kill-loser rule to train intermediate layers of S-cells, U_{S2} and U_{S3}.

Every time when a training pattern is presented, the learning processes are repeated: (1) A seed-cell is chosen by a competition among cells that have already been generated, where the competition area of a cell has a shape of a hypercolumn. (2) The seed-cell has its input connections renewed. (3) Cell-planes that contain losers of the competition are removed from the layer, where silent cells are not categorized as losers. (4) A new cell-plane is generated, if there is a place where no cell is responding despite of non-zero stimulus. When the repetition of these processes has been finished for a training pattern, we proceed to the presentation of the next training pattern.

During the learning, a number of training patterns are presented repeatedly to the network. With the winner-kill-loser rule, generation of new cells (or cell-planes) and removal of redundant cells (cell-planes) are repeated in the network. In the areas where feature-extracting cells are missing in the feature space, new cells (cell-planes) are generated. In the areas where similar cells exist in duplicate, redundant cells (cell-planes) are removed. Thus feature-extracting cells gradually come to distribute uniformly in the feature space.

In the computer simulation, each training pattern of the training set was presented once. In other words, we made one round of presentation of the training set.

Training S-cells of the highest stage. S-cells of the highest stage (U_{S4}) are trained using a supervised competitive learning, and the class names of the training patterns are also utilized for the learning. Namely, when each cell-plane first learns a training pattern, the class name of the training pattern is assigned to the cell-plane. Thus, each cell-plane of U_{S4} has a label indicating one of the 10 digits.

Every time when a training pattern is presented, competition occurs among all S-cells in the layer. If the winner of the competition has the same label as the training pattern, the winner becomes the seed-cell and learns the training pattern. If the winner has a wrong label (or if all S-cells are silent), however, a new cell-plane is generated, learns the training pattern and is put a label of the class name of the training pattern.

The same training set is presented repeatedly until all the patterns in the training set come to be recognized correctly. Although a repeated presentation of the training set is required before the learning converges, the required number of repetition is not so large, say around 4 or 5, in usual cases.

Recognition layer U_{C4} has 10 C-cells corresponding to the 10 digits to be recognized. During the recognition phase, the winner of the competition among all S-cells of U_{S4} transmit its output to the C-cell of the same label.

Appendix D. C-cell layers

In each stage ($l \le 3$), except the highest stage, the layer of C-cells have the same number of cell-planes as the layer of S-cells. There is one-to-one correspondence between the cell-planes of the S- and C-cell layers in the same stage. Each C-cell of a cell-plane receives signals from S-cells of the corresponding cell-plane. In U_{C1}, each C-cell further receives disinhibitory signals from S-cells of orthogonal preferred orientation.

The process of calculating the response of a C-cell is basically the same as that for the previous neocognitron (Fukushima, 2010a). In this section, we discuss only the differences between them, and abbreviate explanations on, say, the process of designing the connections.

We first calculate an average of input signals from S-cells, not by a linear summation, but by a root-mean-square. Let $u'_{Cl}(\boldsymbol{n}, k)$ be the averaged input to a C-cell of the kth cell-plane of layer U_{Cl}, where $\boldsymbol{n}$ is the location of its receptive field. Then the response of the C-cell is given by

$$u_{Cl}(\boldsymbol{n}, k) = \frac{u'_{Cl}(\boldsymbol{n}, k)}{\sigma u_{\mathrm{Cmax}l} + u'_{Cl}(\boldsymbol{n}, k)}, \tag{D.1}$$

where

$$u_{\mathrm{Cmax}l} = \max_{\boldsymbol{n},k} \left\{u'_{Cl}(\boldsymbol{n}, k)\right\}. \tag{D.2}$$

The method of calculating averaged input $u'_{Cl}(\boldsymbol{n}, k)$ is slightly different from layer to layer.

Let $u^+_{Cl}(\boldsymbol{n}, k)$ and $u^-_{Cl}(\boldsymbol{n}, k)$ be the squared sums of input signals from the excitatory center and the inhibitory surround of the connections to the C-cell, respectively:

$$u^+_{Cl}(\boldsymbol{n}, k) = \sum_{|\boldsymbol{v}| \le A_{\mathrm{Col}}} a^+_{Cl}(\boldsymbol{v}) \cdot \{u_{Sl}(\boldsymbol{n} + \boldsymbol{v}, k)\}^2, \tag{D.3}$$

$$u^-_{Cl}(\boldsymbol{n}, k) = \sum_{A_{\mathrm{Col}} < |\boldsymbol{v}| \le A_{Cl}} a^-_{Cl}(\boldsymbol{v}) \cdot \{u_{Sl}(\boldsymbol{n} + \boldsymbol{v}, k)\}^2, \tag{D.4}$$

where $a^+_{Cl}(\boldsymbol{v}) > 0$ in a disk $|\boldsymbol{v}| \le A_{\mathrm{Col}}$, and $a^-_{Cl}(\boldsymbol{v}) > 0$ in an annulus $A_{\mathrm{Col}} < |\boldsymbol{v}| \le A_{Cl}$.

Then, for the 2nd and 3rd stages, we have

$$u'_{Cl}(\boldsymbol{n}, k) = \sqrt{\varphi\left[u^+_{Cl}(\boldsymbol{n} + \boldsymbol{v}, k) - u^-_{Cl}(\boldsymbol{n} + \boldsymbol{v}, k)\right]}, \quad (l = 2, 3). \tag{D.5}$$

C-cells of the 3rd stage, however, do not have an inhibitory surround, and we have $a^-_{C3}(\boldsymbol{v}) = 0$, hence $u^-_{C3}(\boldsymbol{n} + \boldsymbol{v}, k) = 0$.

Connections to a C-cell of the 1st stage also have an inhibitory surround, but signals from S-cells of orthogonal preferred orientations suppress the surround inhibition by disinhibition. Namely,

$$u'_{C1}(\boldsymbol{n}, k) = \sqrt{\varphi\left[u^+_{C1}(\boldsymbol{n}, k) - \varphi\left[u^-_{C1}(\boldsymbol{n}, k) - \frac{u^-_{C1}(\boldsymbol{n}, k^+) + u^-_{C1}(\boldsymbol{n}, k^-)}{2}\right]\right]}, \tag{D.6}$$

where k^+ and k^- represent the sequence numbers of the cell-planes whose preferred orientations are perpendicular to that of the kth cell-plane. Namely, $k^+ = k + K_{C1}/4$ and $k^- = k - K_{C1}/4$.

References

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., Gallant, J. L., & Rust, N. C. (2005). Do we know what the early visual system does? *Journal of Neuroscience*, *25*(46), 10577–10597.

Cardoso, Â, & Wichert, A. (2010). Neocognitron and the map transformation cascade. *Neural Networks*, *23*, 74–88.

Eccles, J. C. (1964). *The physiology of synapses*. Berlin, Heidelberg, New York: Springer-Verlag.

Elliffe, M. C. M., Rolls, E. T., & Stringer, S. M. (2002). Invariant recognition of feature combinations in the visual system. *Biological Cybernetics*, *86*, 59–71.

Fukushima, K. (1975). Cognitron: a self-organizing multilayered neural network. *Biological Cybernetics*, *20*(3/4), 121–136.

Fukushima, K. (1980). Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202.

Fukushima, K. (1988). Neocognitron: a hierarchical neural network capable of visual pattern recognition. *Neural Networks*, *1*(2), 119–130.

Fukushima, K. (2003). Neocognitron for handwritten digit recognition. *Neurocomputing*, *51*, 161–180.

Fukushima, K. (2010a). Neocognitron trained with winner-kill-loser rule. *Neural Networks*, *23*(7), 926–938.

Fukushima, K. (2010b). Increased robustness against background noise: pattern recognition by a neocognitron. In *LNCS: Vol. 6444. ICONIP 2010, part II* (pp. 574–581). Berlin, Heidelberg: Springer-Verlag.

Fukushima, K., & Miyake, S. (1982). Neocognitron: a new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognition*, *15*(6), 455–469.

Fukushima, K., & Tanigawa, M. (1996). Use of different thresholds in learning and recognition. *Neurocomputing*, *11*(1), 1–17.

Hebb, D. O. (1949). *Organization of behavior*. New York, London, Sydney: John Wiley & Sons.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, *9*, 181–197.

Hildebrandt, T. H. (1991). Optimal training of thresholded linear correlation classifiers. *IEEE Transactions on Neural Networks*, *2*(6), 577–588.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of physiology (London)*, *106*(1), 106–154.

Hubel, D. H., & Wiesel, T. N. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neurophysiology*, *28*(2), 229–289.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. J. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*, 541–551.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lo, S. B., Chan, H., Lin, J., Li, H., Freedman, M. T., & Mun, S. K. (1995). Artificial convolution neural network for medical image pattern recognition. *Neural Networks*, *8*(7/8), 1201–1214.

Ohzawa, I., DeAngelis, G. C., & Freeman, R. D. (1990). Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors. *Science*, *249*(4972), 1037–1041.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025.

Satoh, S., Kuroiwa, J., Aso, H., & Miyake, S. (1999). Recognition of rotated patterns using a neocognitron. In L. C. Jain, & B. Lazzerini (Eds.), *Knowledge based intelligent techniques in character recognition* (pp. 49–64). CRC Press.