Biophysics, Vol. 37, No. 2, pp. 249–258, 1992
Printed in Great Britain

0006–3509/92 $15.00+0.00

COMPUTER MODEL OF A "SENSE OF HUMOUR" II. REALIZATION IN NEURONAL NETWORKS*

I. M. SUSLOV

Lebedev Physics Institute, Russian Academy of Sciences, Moscow

(*Received 9 August* 1991)

The computer realization of a "sense of humour" requires the creation of an algorithm for solving the "linguistic problem", i.e. the problem of recognizing a continuous sequence of polysemantic images. Such an algorithm may be realized in the Hopfield model of a neuronal network when suitably modified.

In [1] we analysed the general algorithm of information processing and showed that on fulfilment of its natural requirements raised by its biological purpose such an algorithm will possess a "sense of humour". The present paper proposes a possible realization of the algorithm in a system of formal neurones.

DESCRIPTION OF THE MODEL

Following Hopfield [2] we shall consider that the state of the ith neurone is described by the variable V_i assuming two values: $V_i=1$ (excited state) and $V_i=0$ (state of rest). The link of the neurone i with the neurone j is determined by the parameter T_{ij}. The system evolves according to the algorithm:

$$V_i(t+\delta t)={}^1/_2+{}^1/_2\,\mathrm{sgn}\left\{\sum_j T_{ij}V_j(t)-U_i\right\}, \tag{1}$$

where U_i is the excitation threshold of the ith neurone and the number i is chosen randomly.

The proposed model of the nervous system is a modification of the trilayer perceptrone [3] adapted for work in real time. It contains the following elements (Fig. 1).

The associative memory (A-layer) represents the neuronal network which for simplicity we consider as described by the Hopfield model: the neurones of the A-layer are linked with each other with $T_{ij}=T_{ji}$, $U_i=0$. Within the A-layer evolution according to (1) leads from the arbitrary initial state $\{V_i\}$ to one of the local energy minima

$$E=-\sum_{ij} T_{ij}V_iV_j, \tag{2}$$

* *Biofizika*, **37**, No. 2, 325–334, 1992.

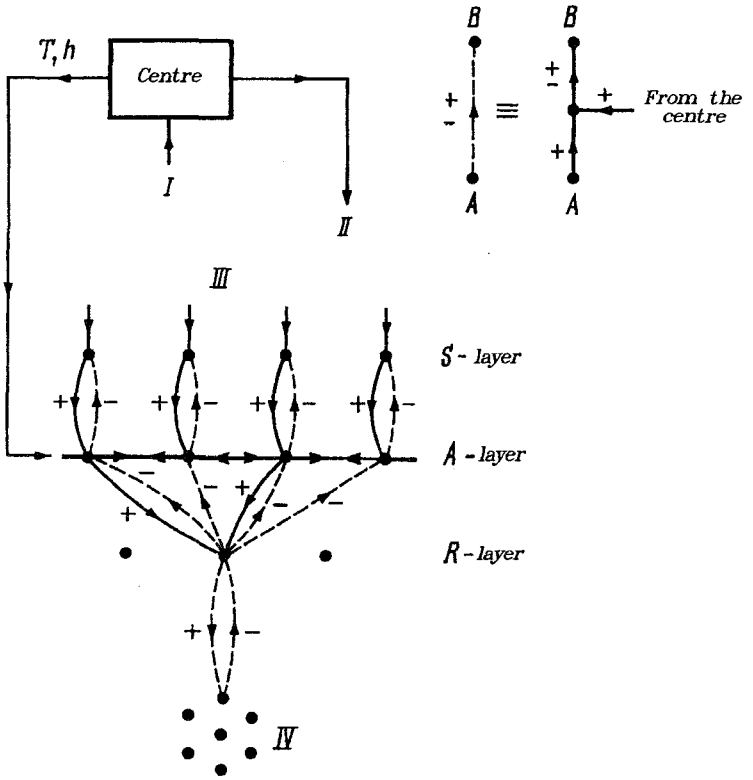

Fig. 1. Proposed model of the nervous system — modification of the three-layer perceptrone [3]: filled circles, neurones; continuous lines, constantly acting links between neurones; broken lines, links switched on by command from the centre; the symbols + and − indicate the excitatory and inhibitory character of the links; the *S*-layer is the sensory system; the *A*-layer the associative memory; the *R*-layer the reacting system or "consciousness"; I is information from the *A*-layer, II, the control of the links, III, the outer world, IV, the motor cortex. Above, right, possible realization of the controllable link.

which are equated with the images written in the memory $\{V_i^s\}$, $s=1, 2, \ldots, p$. The recorded images determine the matrix of the links T_{ij} [2]:

$$T_{ij}=\sum_{s=1}^{p}\mu_s(2V_i^s-1)(2V_j^s-1), \quad i\neq j \quad (\mu_s>0), \quad T_{ii}=0. \tag{3}$$

The sensory system (S-layer) receives signals from the outside world (for example, from the retina of the eye). The sensory neurones are not linked between themselves, but each *S*-neurone is bound to one of the memory neurones: the links $S\rightarrow A$ are positive (exciting) and the back links $A\rightarrow S$ are negative (inhibiting) (Fig. 1).

The reacting system (R-layer) consists of a set of neurones each of which corresponds to one of the images recorded in the memory: to the sth neurone of the *R*-layer converge the positive (exciting) links from those neurones i of the memory for which $V_i^s=1$ (we shall call the last neurones the image carrier $\{V_i^s\}$). The back negative (inhibiting) links run from the *R*-neurones to the memory neurones; the *R*-neurones are not linked between themselves (Fig. 1). The thresholds U_i for the *R*-neurones so line up that the excitation of the sth neurone of the *R*-layer occurs only when the configuration of the *A*-neurones is sufficiently close to the image $\{V_i^s\}$.

We consider that the image begins to be realized by the biological individual only when there is excitation of the corresponding *R*-neurone, i.e. the *R*-layer represents the consciousness of the individual.

The centre coordinates the work of the system by acting according to the built-in program; it exercises control over the macroscopic parameters of the system and control of them. The concrete functions of the centre consist in the following.

(1) The centre has links with a small proportion of the memory neurones evenly distributed in the A-layer which allows it to judge the presence of excited neurones in a certain portion of the memory and the stationarity or nonstationarity of this portion.

(2) The centre carries out local change in "temperature" in the A-layer. Since the temperature of the neuronal net is determined by the noise level in it (which is taken into account in (1) by introducing into the braces the random force $f_i(t)$) then the regulatable noise source must be at the disposal of the centre.

(3) The centre locally switches on the "magnetic field" in the A-layer which corresponds to an additional term in the presence in the expression for energy (2)

$$\sum_i h_i V_i \tag{4}$$

(in (1) h_i are added to the thresholds U_i). Switching on the field is achieved with the aid of a "magnet" — group of neurones controlled from the centre and from each of which run links to the neurones of a certain region of the A-layer.

(4) The centre carries out the control of the links shown in Fig. 1 by a broken line. The simplest realization of the controllable link AB is possible with the aid of the insertion neurone C (see Fig. 1), the threshold of which is so chosen that it is excited only in the simultaneous presence of the exciting signal from the neurone A and from the centre. In the presence of a signal from the centre the neurone A excites or inhibits the neurone B — the link is switched on, in the absence of a signal from the centre the neurone A cannot act upon the neurone B — the link is switched off. The command for switching on and off is given not to the individual links but immediately to their large groups.

(5) The centre gives the command for the learning of the plastic links.

RECOGNITION OF A SEPARATE IMAGE

Recognition of the images occurs with the links $A \to S$ and $R \to A$ switched on. In the initial state of the system all the neurones are non-excited. Since the state of the A-layer with $V_i \equiv 0$ is unstable (see (1) at $U_i = 0$), its maintenance requires the presence of the stabilizing magnetic field.

Let the image $\tilde{B} = B + \delta B$ come to the input of the sensory system, i.e. the "noisy" image B; this induces excitation of some of the neurones of the S-layer (Fig. 2*a*). Then the centre cuts out the magnetic field and excitation is transmitted to the memory neurones (Fig. 2*b*) which, in turn, quench the sensory neurones (Fig. 2*c*) (it is assumed that the links $S \to A$ and $A \to S$ are sufficiently strong). Then in the A-layer there is free evolution according to (1) which ends in relaxation to the stable state corresponding to the image B (Fig. 2*d*); in the R-layer is excited the neurone responsible for this image (Fig. 2*e*). Into the memory is fed the back signal quenching the excited neurones (Fig. 2*f*) which, in turn, leads to the quenching of the image B in the "consciousness" (Fig. 2*g*); the system thereby returns to the initial position and is ready for the perception of a new image.

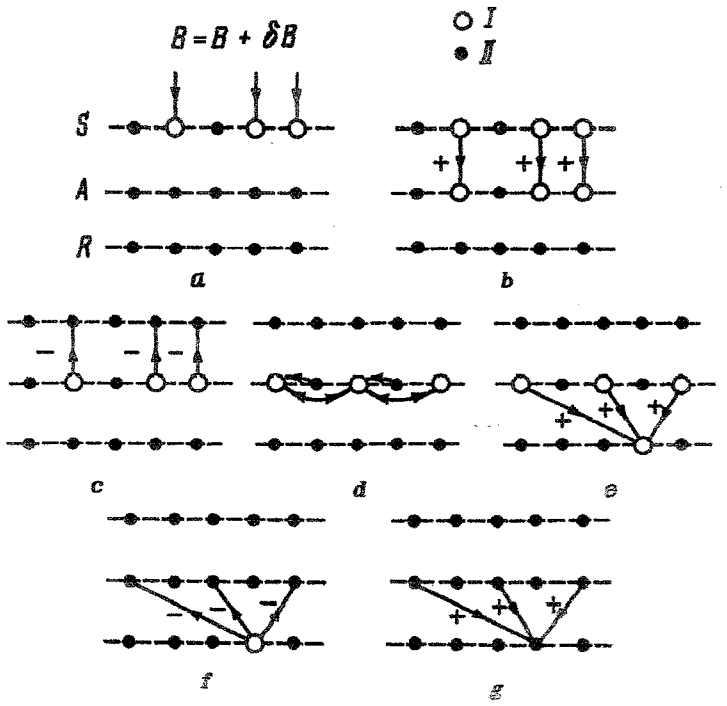

Fig. 2. Sequential states of the system in the course of recognition of a separate image; I, excited; II, non-excited neurones. For clarity the figure shows only the links along which excitation was transmitted at the preceding moment of time.

LEARNING

The links between the A-neurones can be learned (plastic) and change by the value [2]

$$\delta T_{ij} \sim (2V_i - 1)(2V_j - 1)\delta t \quad (i \neq j), \tag{5}$$

if the neurones i and j stay in the states V_i and V_j during the time δt. If in the initial state $T_{ij} \equiv 0$ then the presentation to the system of p configurations $\{V_i^s\}$, $s = 1, 2, \ldots, p$ leads to the matrix of links (3). Since T_{ij} may have any sign, the neurones of the A-layer must have both exciting and inhibiting synapses (specialization of the synapses, as is known, ([4], pp. 62–65) is of an invariant character).

The links $A \to R$ in the initial state have a zero value and can be learned only in one, positive direction

$$\delta T_{ij} = \begin{cases} c\delta t \quad (c > 0) \text{ for } V_i = V_j = 1, \\ 0 \quad \text{in other cases,} \end{cases} \tag{6}$$

i.e. have only exciting synapses. The remaining links ($R \to A, S \to A, A \to S$) cannot be learned and are of inborn character.

The process of learning of the system is similar to the process by which a child learns: a certain image B is presented to it generating in the S-layer a certain configuration of excited neurones; then it is asked to "memorize B" which excites one of the neurones of the R-layer which is "nominated" responsible for the image B. Learning occurs with the $A \to S$ and $R \to A$ links cut out (Fig. 3) so that the configuration of the S-neurones is projected into the A-layer and persists for a certain time: the links T_{ij} in the A-layer change according to (5) forming the matrix (3) and the links $A \to R$ according to (6) ensuring the link of the sth neurone of the R-layer with the image carrier $\{V_i^s\}$.

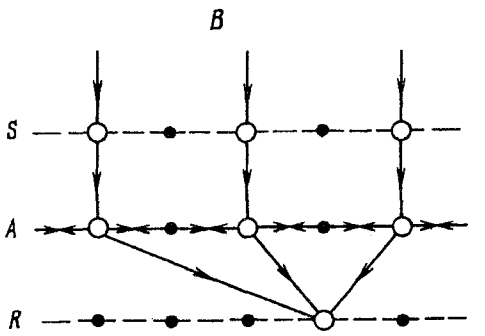

Fig. 3. learning occurs with the $S \to A$ and $R \to A$ links switched off.

In the resting state the system is in the configuration $V_i \equiv 0$; since the system spends a considerable time in this state then on learning, according to (5), "ferromagnetic" interaction would appear between the neurones ($T_{ij} > 0$ for all i, j) and only the state $V_i \equiv 1$ would be stable. Therefore, we shall consider that learning does not occur in the state $V_i \equiv 0$: the learning command is given only on presentation of the image.

SHORT ACTION OF THE LINKS AND THE LOCALIZATION OF IMAGES

In the usual Hopfield model [2] all the neurones are linked with each other and the image carriers $\{V_i^s\}$ are spread out over the whole neuronal net. In reality the links T_{ij} have a finite radius of action ξ: in the human brain each neurone has $\sim 10^4$ synapses with a complete number of neurones $\sim 10^{11}$ ([4], pp. 31–33). Experimental indications of the localization of the images also exist ([3], pp. 64, 65).

We shall consider that each image $\{V_i^s\}$ is written in a certain region Ω_s containing many neurones, but small as compared with the size of the whole neuronal network; here $V_i^s = 0$ for i not belonging to Ω_s (the carrier $\{V_i^s\}$ is localized in Ω_s) and for $i \in \Omega_s$ the magnitudes V_i^s with equal probability assume the values 0 and 1.* Since recognition of the images in any event must be preceded by translational shift, rotation and change of the scale (which may be achieved by a certain modification of the Hopfield model [5]) then the assumption on the localization of images does not have any serious after effects. The command for switching off the magnetic field, change in temperature (see below) and learning of plastic links is given only on presentation of the image $\{V_i^s\}$ and only for neurones of the region Ω_s.

Taking all this into account, change in the links in the A-layer on presentation of the image $\{V_i^s\}$ instead of (5) has the form

$$\delta T_{ij}^{(s)} \sim D_{ij} \delta_i^s \delta_j^s (2V_i^s - 1)(2V_j^s - 1) \delta t, \tag{7}$$

where

$$D_{ij} = \begin{cases} 1, & \text{for } r_{ij} < \xi, \quad i \neq j \\ 0 & \text{in other cases} \end{cases} \qquad \delta_i^s = \begin{cases} 1 & \text{for } i \in \Omega_s, \\ 0 & \text{for } i \notin \Omega_s, \end{cases} \tag{8}$$

* To raise the stability of the system relative to the destruction of some of the neurones the region Ω_s may be multiply connected.

r_{ij} is the distance between the neurones i and j and the matrix of the links after writing p images instead of (3) assumes the form

$$T_{ij}=D_{ij}\sum_{s=1}^{p}\delta_i^s\delta_j^s\mu_s(2V_i^s-1)(2V_j^s-1),\quad \mu_s>0. \tag{9}$$

To demonstrate stability of $\{V_i^s\}$ let us compose the combination [2]

$$H_i^{s_0}\equiv\sum_j T_{ij}V_j^{s_0}=\sum_s\mu_s\delta_i^s(2V_i^s-1)\sum_{j\in\Omega_{ss_0}}D_{ij}V_j^{s_0}(2V_j^s-1), \tag{10}$$

where Ω_{ss_0} is the intersection of Ω_s and Ω_{s_0}. By virtue of the randomness of V_i^s and the large number of terms in the sum for j the latter is close to its mean:

$$H_i^{s_0}=\mu_{s_0}\delta_i^{s_0}\frac{1}{2}\Big(\sum_{j\in\Omega_{s_0}}D_{ij}\Big)(2V_i^{s_0}-1). \tag{11}$$

For $i\in\Omega_{s_0}$ the sum for j is positive and $\delta_i^{s_0}=1$ so that the configuration of $\{V_i^{s_0}\}$ is stable by virtue of the algorithm (1); for $i\notin\Omega_{s_0}$ the state $V_i=0$ is maintained by the magnetic field.

Because of the localization of the images it suffices for the sth neurone of the R-layer to have links only with the neurones of the region Ω_s of the A-layer.

RECOGNITION OF THE AMBIGUOUS IMAGE

Above it was assumed that the stimulus presented $\tilde{B}=B+\delta B$ is close to image B contained in the memory so that the initial state $\tilde{B}$ always relaxes to the final state B, i.e. the interpretation of the image $\tilde{B}$ is clearcut. In energy language this means that the state $\tilde{B}$ enters a potential well the minimum of which corresponds to the image B and evolution occurs at zero temperature.

Now let us consider the recognition of an ambiguous image. We have in mind the modelling of the following situation: the person is shown the word B and it is made clear to him that it may assume several values: in the memory of the person are fixed the images $B+b_1, B+b_2, \ldots,$ where $b_1, b_2, \ldots$ are the elements of the given clarification; if now the word B is presented for recognition its interpretation will be ambiguous leading to one of the results $B+b_i$. For modelling it suffices to assume that recognition begins at the temperature T exceeding the potential barriers between $B+b_1, B+b_2, \ldots,$ images and then the temperature falls and the system enters one of the local minima corresponding to the images $B+b_i$. Hereafter, we shall consider that fall in temperature from the initial value T_0 occurs adiabatically so that the system with overwhelming probability relaxes to the deepest of the local minima.

SIMULTANEOUS RECOGNITION OF SEVERAL IMAGES

Let as a result of presentation of the stimulus A the system relax to one of the images $A_1, A_2, \ldots,$ and as a result of presentation of the stimulus B to one of the images $B_1, B_2, \ldots$. What will happen with the simultaneous presentation of stimuli A and B?

The process of image recognition begins with cutting out the magnetic field in the portion of the A-layer in which is localized the image carrier (see above). Let with presentation of the images A and B the field to be cut out in the regions Ω_A and Ω_B of the associative layer. For clarity of expression let us assume that the regions Ω_A and Ω_B do not overlap (the qualitative

picture persists in the general case); then $V_i = V_i^A + V_i^B$, where V_i^A and V_i^B differ from zero respectively for the i neurones lying in the regions Ω_A and Ω_B. The energy (2) assumes the form

$$E\{V_i^A + V_i^B\} = -\sum_{ij} T_{ij} V_i^A V_j^A - \sum_{ij} T_{ij} V_i^B V_j^B - 2\sum_{ij} T_{ij} V_i^A V_j^B. \quad (12)$$

On presentation of the stimuli A and B singly, we have respectively $V_i^B \equiv 0$ and $V_i^A \equiv 0$, so that in the right part of (12) there remains only the first or second term. The third term in (12) shows that with the simultaneous presentation of the stimuli A and B, they exert on each other an effect equivalent to the presence of a magnetic field. The results of such interaction is particularly graphic if the configurations $A_1, A_2, \ldots$ and respectively $B_1, B_2, \ldots$ are almost degenerate. If the links T_{ij} between the regions Ω_A and Ω_B are absent then the equilibrium states of the system have the form $(A_s, B_{s'})$ and are almost degenerate. Inclusion of the weak links T_{ij} between the regions Ω_A and Ω_B does not significantly change the equilibrium configurations but changes the relative position of their energy levels. As a result, if on separate recognition of the images A and B some configurations (for example, A_1 and B_2 in Fig. 4) are energetically advantageous, then on simultaneous recognition of A and B other configurations may prove more advantageous (for example, A_2, B_1); it may be graphically imagined that the interaction changes the potential relief in which A and B relax (Fig. 4). Thus, the choice of the polysemantic image proves to depend on the context.

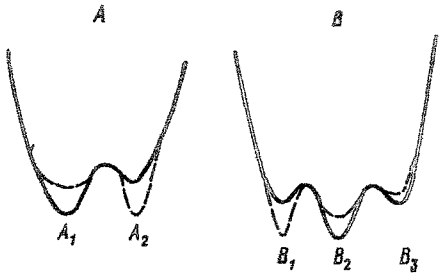

Fig. 4. Potential relief on recognition of the images A and B singly (continuous curve) and on simultaneous recognition (broken line).

In [1] the existence of an algorithm was postulated for the recognition of N simultaneously presented polysemantic images; from all this it is clear that such an algorithm is naturally realized in the Hopfield model in the following conditions: (a) recognition begins at finite temperature T_0 which then adiabatically diminishes; (b) the potential barriers between the values of the polysemantic image are less than T_0; (c) the potential barriers between the polysemantic images are considerably greater than T_0; and (d) the interaction between the images is quite weak.

RECOGNITION OF A CONTINUOUS SEQUENCE OF IMAGES

Above we assume that all the images $\{V_i^s\}$ have localized carriers. In view of the finiteness of the radius of action of the links T_{ij} the associatively bound images must have closely positioned or overlapping carriers, while to the non-correlated images correspond to carriers localized in the remote parts of the memory. Therefore, with the entry into the brain of a continuous sequence of stimuli $A_1, A_2, A_3, \ldots$ in the associative layer there will be sequential excitation of the regions $\Omega_1, \Omega_2, \ldots$, which will look like "diffusion" (Fig. 5): the nearest images in sequence ing are correlated and their carriers form conglomerations whereas to the remote images in sequence correspond carriers remote in space, as a result of weakening of the correlations. This allows one to establish the correspondence of the time of appearance of the image with the

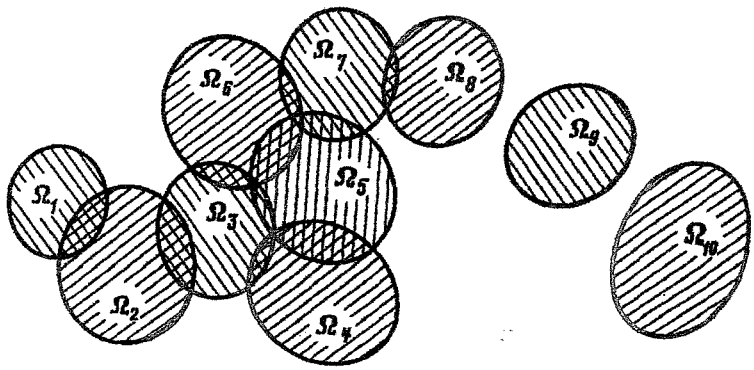

Fig. 5. On recognition of a continuous sequence of images $A_1, A_2, \ldots$ in the associative layer there is sequential excitation of the regions $\Omega_1, \Omega_2, \ldots$, which looks like "diffusion". If the regions Ω_s are multiply connected, there are several diffusion trajectories in different portions of the neuronal net.

position of its carrier in space which greatly simplifies the work of the centre. After fixing the appearance of the excited neurones in a certain portion of the A-layer, the centre raises the temperature of this portion to the value T_0, and after the characteristic time τ_0 its adiabatic fall begins, which stimulates establishment of the steady state, the bringing of the corresponding images into the R-layer and the zeroing of the portion of the memory considered (Fig. 2*d–g*). The trajectory of the "diffusion" movement (Fig. 5) after a certain time is thereby "deleted" so that at each moment in the memory there is only its final segment; recognition of a continuous sequence of images is thereby reduced to the simultaneous recognition of a finite number of images (see above).

TIME DELAYS AND THE HUMOROUS EFFECT

Let at the moment of time $t=0$ the neurones be excited in a certain portion of memory (Fig. 2*b*), at the moment τ_0 the corresponding image is brought into consciousness (Fig. 2*e*) and at the moment of time τ_1 the portion of the memory considered is zeroed (Fig. 2*f*). The delay τ_0 corresponds to the interval AC and the delay τ_1 to the interval AB in Fig. 2 of [1]; the latter is related to the fact that the possibility of re-interpreting the image persists until the corresponding memory portion is zeroed. Obviously, $\tau_1 \geqslant \tau_0$: if in [1] this result is derived from considerations of the optimality of the work of the algorithm, now it is due to the constructive features of the model.

The delay τ_0 is determined by the rate of fall in temperature (see above), the delay τ_1 by the moment of inclusion from the centre of the $R \rightarrow A$-links. The optimal choice of the parameters τ_0 and τ_1 is determined by different principles: while the parameter τ_0 governs the delay from the moment of entry of information to the brain up to its emergence in consciousness and is upwardly limited by the magnitude τ_{max} [1], the parameter τ_1 regulates memory loading (unlike the general case [1] in this model there is no special operative memory), i.e. the fraction of excited portions in the A-layer (Fig. 5). This fraction must not be too small for the operative possibilities of the brain to be used in full, and not too large for interference of the images arriving at different times not to appear.

Let at the moment of time $t=0$ the image A enter the memory; evolution in the corresponding potential relief (Fig. 4, continuous curve) leads at the moment $t=\tau_0$ to stabilization of the mem-

ory neurones in the configuration A_1 and excitation of the corresponding R-neurone. Let in the interval between τ_0 and τ_1 the image B enter the memory and the potential relief change for A (Fig. 4, broken line). If the temperature at the moment considered is sufficient to overcome the barrier, the system begins to relax to the configuration A_2 (in fact, passage from the state A_1 is also possible at $T=0$ since the minimum corresponding to A_1 may disappear). Such disturbance of the steadiness of a certain memory portion after coming into the R-layer of the image corresponding to it is also a sign of the humorous effect [1]. The image A_1 (in the general case a version of several images) is perceived as "false" and must be immediately thrown out of consciousness; however, this cannot be done by the usual scheme (Fig. 2*e*, *g*) because of the need to obtain a new steady configuration A_2.

MECHANISM OF LAUGHTER

The "emergency" rejection of the false version from the R-layer is achieved by the inclusion of the links between the R-layer and the motor cortex (Fig. 1); excitation of the R-neurones is transmitted to the motoneurones inducing contraction of certain muscles, i.e. laughter.

It is not hard to notice the similarity to the old ideas of G. Spencer [6] according to which the humorous effect is accompanied by release from the mental process of nervous energy which is directed at the muscular reaction. This idea was supported by Darwin [7] and Freud [8] but was criticized by later investigators [9] in view of the difficulty of introducing the concept of "nervous energy". In fact, the concept of energy for neuronal nets may be introduced only for the not very realistic condition $T_{ij}=T_{ji}$ [2] (which is untrue for the model considered as a whole) and great significance cannot be attached to it. Nevertheless, the qualitative picture agrees with Spencer's hypothesis: it appears as though the energy of excitation of the neurones is expelled into the motor cortex.

The release of nervous energy in the presence of the humorous effect was validated by Spencer using the concept of "descending incongruity" — transition from a high to low style, i.e. from the state with rich associations to the state with poor associations. Such an interpretation of the humorous effect is known to be limited and cannot lay claim to universality. In the proposed scheme the "release of nervous energy" (in the nominal sense indicated above) is linked with the need to get rid of the false version brought into consciousness, which requires "zeroing" of a certain portion of the R-layer, i.e. translation of the excited neurones to the non-excited state.

Since laughter is interpreted as an unconditioned reflex to the humorous effect the known cases of "ousting" of laughter by secondary emotions require an explanation. Laughter may be ousted by the emotions of indignation (an indecent anecdote is told to a person of puritanical convictions), fear (the bush suddenly turns out to be a bear), compassion (a person in front of you slips on a water melon rind and badly hurts himself), shame (you slipped on a water melon rind) and so on. Within the Spencer hypothesis [6] all these instances are explained by the fact that the "released nervous energy" is directed not to the motoneurones but to other parts of the nervous system and goes on the formation of a secondary emotion (to the R-layer is connected not the motor cortex but the limbic system). The same ideas [8] are used to explain the known fact that a joke produces the greatest effect if it is told extremely laconically: laconicity reduces the probability of formation of secondary associations liable to absorb the "nervous energy".

Casting the excitation of the neurones into different portions of the motor cortex, the person

may regulate the level of the muscular reaction: with this is connected its dependence on mood, the psychological setting, the presence of a laughing audience [10] and so on.

CONCLUSION

The realization of a sense of humour requires a quite intricately organized system. We would emphasize, however, that this complex organization is entirely governed by the task of treating a continuous sequence of polysemantic images; the existence of the humorous effect is a secondary consequence. As is known ([4], pp. 219–241), different parts of the brain have their own specialization; the proposed model may lay claim to be a description of only that region of the brain in which are concentrated the linguistic functions (so-called Broca and Wernicke zones); other portions of the brain may have a different organization.

The author is grateful to D. S. Chernavskii for fruitful discussions.

REFERENCES

1. I. M. Suslov, *Biofizika*, **37**, 318 (1992).
2. J. J. Hopfield, *Proc. Nat. Acad. Sci. USA*, **79**, 2554 (1982).
3. F. Rosenblatt, in *The Principles of Neurodynamics. Perceptrones and the Theory of Brain Mechanisms*, Mir, Moscow (1965).
4. In *The Brain* (Ed. P. V. Simonov), Mir, Moscow (1984).
5. V. S. Dotsenko, *J. Phys. A.*, **21**, L783 (1988).
6. G. Spencer, *The Physiology of Laughter*, St. Petersburg (1881).
7. C. Darwin, in *Collected Works*, Vol. 5, Chapter VIII, U.S.S.R. Academy of Sciences, Moscow (1953).
8. S. Freud, in *Wit and its Relation to the Unconscious*, Moscow (1925).
9. D. E. Berlyne, in *Psychology of Humor* (Eds J. H. Goldstein and P. E. McGhee), Academic Press, New York (1972).
10. H. Levental, *J. Communication*, **26**, 190 (1976).