1. ARTIFICIAL NEURAL NETWORKS

INTRODUCTION

The human central nervous system (CNS: brain and spinal cord) has approximately 100 billion neurons, which are morphologically and functionally (neural diversity) divided (Fig. 1) in various kinds (e.g. local interneuron, projection interneuron, motor neuron, sensory neuron, neuroendocrine cells). The generic neural cell consists of a cell body, including the cell nucleus, and two main extensions to receive and deliver electric impulses. There are both afferent (from periphery to CNS) and efferent (from CNS to periphery) neurons. Sensory neurons differ morphologically from the other kinds as their input extension is connected to specialized receptor cells (mechanoreceptor, photoreceptor, nociceptor, …) by one *input* axon. The neuron receives signals in input through synapses, located on the dendrites or membrane of the neuron, and when the integral signal is strong enough (overcome a certain threshold), the neuron, being activated, emits a signal running through the output axon. The output signal might be sent to another synapse, and might activate other neurons. Each neuron can make contact with up to several thousand other neurons. There are both excitatory and inhibitory effects of the connections. Excitatory connections contribute positively in increasing the summation of the integral signal. Conversely, inhibitory connections decrease the integral effect. Looking at one synaptic junction between two neurons, one can distinguish the presynaptic neuron, from which the nerve impulse arrives, and the postsynaptic neuron to which the neurotransmitters bind.
Organic chemical compounds, termed neurotransmitters, mostly mediate the transmission of the electric signal between two neurons across the synapse. They are also found at the axon endings of motor neurons, where they stimulate the muscle fibers. There exist both excitatory (e.g. acetylcholine, noradrenalin, glutamate) and inhibitory (e.g. dopamine, gamma aminobutyric acid, serotonin, endorphin) neurotransmitters. Apart from acetylcholine (present at the neuromuscular junction), neurotransmitters are mostly amines or amino acids.

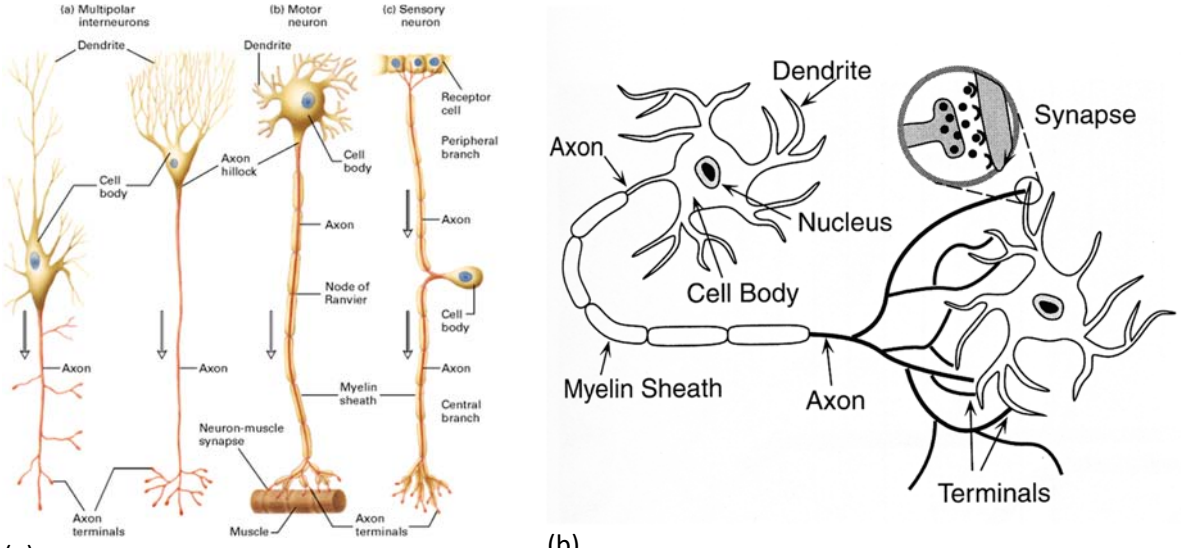
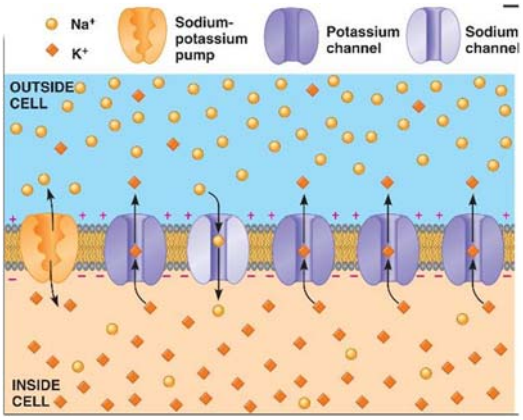


Figure 1. Neuron typologies (a). Schematics of neuron inter-connection.

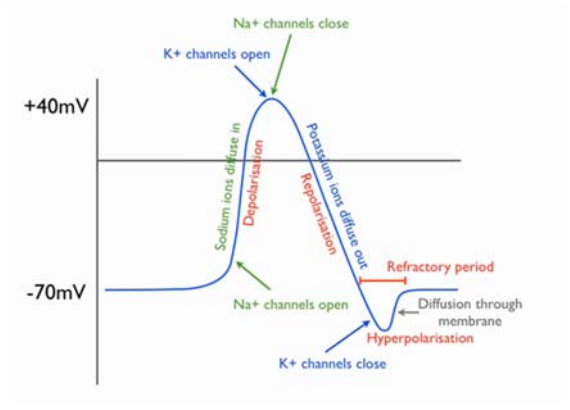
SYNTHESIS OF ELECTRO-CHEMICAL PHYSIOLOGY OF THE NEURON

The boundary of the neuron is known as the cell membrane, which has a voltage difference (the membrane potential) between the inside and outside. The membrane has a very small thickness (70 - 150 Angstrom) with a very high capacity ($1\mu F/cm^2$). It is impermeable to proteins but under certain conditions, it is permeable to potassium (K^+), sodium (Na^+) and chloride (Cl^-) ions. At rest (no firing impulse), sodium and potassium ions are mainly confined to the membrane outside and inside, respectively. The permeability is

controlled by ion channels located in between the two membrane boundaries. The closure and the aperture of the ion channels are voltage- and time-dependent. The restriction of sodium ions outside the membrane is determined by the Na^+ ion channel that is almost closed at rest (Fig. 2a). At this rest condition, the membrane potential (Vm= Vin–Vout) is negative (about -70mV) and mostly determined by potassium Nernst potential, i.e. the voltage difference due to the concentration gradient across the membrane (about -90mV). However, in this condition, the membrane permeability to potassium ions is about 50-100 times higher to that one sodium ions, i.e. the potassium ion channel is not completely closed to the gradient concentration and potassium channels are greater in number with respect to sodium channels. The sodium-potassium pump is an active biochemical mechanism, additional to ion channels, which requires energy supply (ATP-ADP) to activate and guarantees to keep such a voltage difference at the membrane boundaries. At rest, the membrane is said to be polarized (Fig. 2b).
The active state of the excitable cell is in correspondence of generation the action potential. The neuron activation requires a stimulus able to induce a sufficient membrane depolarization over a predetermined voltage. Overcoming such a threshold makes the depolarization completely autonomous. Depolarization involves the increase of the membrane permeability to the sodium ions (NA^+ ion channel opening), which move from the outside toward the inside. This mechanism makes the membrane voltage growing rapidly towards the sodium concentration gradient voltage (Na^+ Nernst potential about +55mV). The membrane has thus a positive voltage inside and a negative voltage outside. Very quickly, the sodium channel closes and the potassium channel opens allowing the potassium ions to rapidly diffuse out. The cell returns to be positive on the outside and negative on the inside. This is called re-polarization phase. Locally, the voltage decreases towards the Nernst potential of the potassium even overcoming the rest potential (hyper-polarization). Meanwhile, the sodium ions moves along the inside to an adjacent area causing a slight change in the polarity of the membrane. The polarity change causes the adjacent closed sodium channel to open. Again, sodium ions move in increasing local polarity inducing soon a closure of the sodium channel. This way the action potential is travelling along the membrane. The potassium channel is activated to restore the negative polarity inside the cell again. The sodium potassium pump pushes back from inside to outside sodium ions and pushes in from outside to inside potassium ions restabilising the rest potential distribution. This repeated mechanism along the axon membrane delivers the neural spike from cell body to axon terminals.



(a)



(b)

Figure 2. Simplification of the ion transport through ion channels and sodium-potassium pump. (a). Local action potential over time following a positive input impulse to the neuron (b).

Measurements about the features of single spike (no myelin neuron) lead to estimate a duration of few milliseconds and a traveling speed of about 25 m/s. However, this last value depends on axon diameter and myelinated fibers. In a myelinated fiber, the spike traveling speed is typically tenfold with respect the speed in the axon with no myelin sheet. Due to absolute refractory period, the frequency cannot overcome some hundred Hz with a minimum frequency of about few Hz. Basically, we can assume that the spike duration is about 5ms. It has been acknowledged that face recognition tasks are usually performed in few hundred of milliseconds (less than half a second). This implies that from sensing to recognition (e.g. perceptual task) the neural processing cannot take more than about 100 serial steps. This ability was renamed Hundred-step

rule. A consequence is that the neural computation is highly parallel with a single neural transmission being about few bits of equivalent information. Knowledge in the brain is thus distributed throughout neural connections.

COMPUTATIONAL MODEL OF THE NEURON

In agreement with the information processing paradigm, the biological neuron may be represented as a computational unit mapping one or more input signals into one output signal. In this approach, the input variables, collected into a multi-dimensional vector called pattern, have their biological correspondence into the signals running through the dendrites. Input signal are processed by an internal operator according to the neural paradigm of stimulus integration, threshold comparison and activation (performed biologically in the cell soma) to produce the response output (moving away from the cell soma along the axon). The neuron output is usually termed the state of the artificial neuron.
According to this formal model of the neuron, one can encompass the concept of neural interconnections by envisioning a network of neurons, where the output of some neurons becomes the input of other neurons. From a static point of view, the input pattern to a neuron can be considered as the state(output) of a set of (virtual) input neurons. In order to simplify the synaptic efficiency effect into this representation, any link between two formal neurons can be thought of as weighted by using a scalar real number. This weight accounts for the strength and the type of the neural link. From a discrete-time dynamic point of view, the artificial neuron expresses a discrete temporal dynamics assuming that the input is entering at time t whereas the ouput is triggered at the next time interval t+1. This inter-connected system was formerly named **Artificial Neural Networks (ANN)**.

The first ANN was proposed in 1943 by McCulloch and Pitts (MCP). This network consists of **binary** threshold units or, as usually named, MCP neurons (Fig. 3). The unit is said to be binary as the ouput has only two possible values. Any MCP neuron is able to compute the action potential P by weighting the input signals $\{x_j\}$ into a weighted linear combination. Given that the neuron has *n* different input signals then the action potential P is computed as:

$$P = \Sigma w_j x_j = w_1 x_1 + w_2 x_2 + \dots + w_n x_n$$

The weights w_j can be positive, representing an excitatory contribute, negative, representing an inhibitory contribute, formally even null meaning the the corresponding input signal is uneffective on the neuron. The P value is then compared to an internal threshold T, named activation threshold. The neural activation function F drives a positive output u whether P overcomes or equals T, and negative output otherwise. The neuron output is thus computed as:

$$u=F(P-T)$$

In a MCP neuron, the function F is traditionally binary. In general, changing the activation function produce a dramatic change in the neuron behaviour..

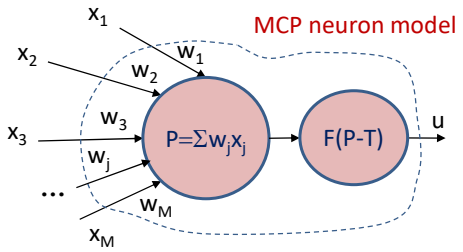


Figure 3. Abstraction of the neuron by McCulloch and Pitts (1943).

Dendrites	Soma	Axon

NEURON ACTIVATION

For binary units at least two different activations can be taken into account, namely the Heaviside and signum functions (Fig. 4). In this case, the neuron can be regarded as a Boolean operator delivering 0(-1) (FALSE) or 1(+1) (TRUE) in the ouput, being the first one an inhibitory signal while being the second one an excitatory signal. Relaxing the binary condition, the MCP neuron can be extended by allowing different activations as linear and non-linear (logsig and tanh) functions (Fig. 5).With linear activation, the neuron output can assume any real value and it is said unbounded. Using logsig the ouput reads:

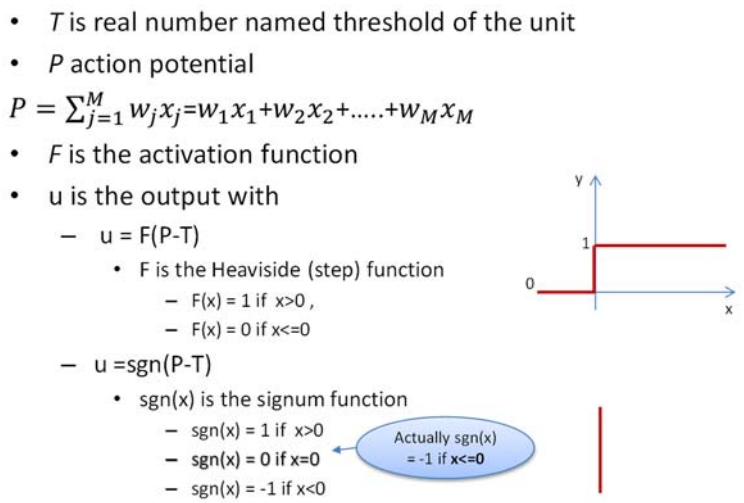


Figure 4. Activation function: Heaviside and signum.

$$u = \frac{1}{1 + e^{-(P-T)}}$$

while with hyperbolic tangent we can write:

$$u = \tanh(P - T) = \frac{e^{(P-T)} - e^{(P-T)}}{e^{(P-T)} + e^{(P-T)}}$$

The output results bounded between 0 and +1, and -1 and +1, respectively.

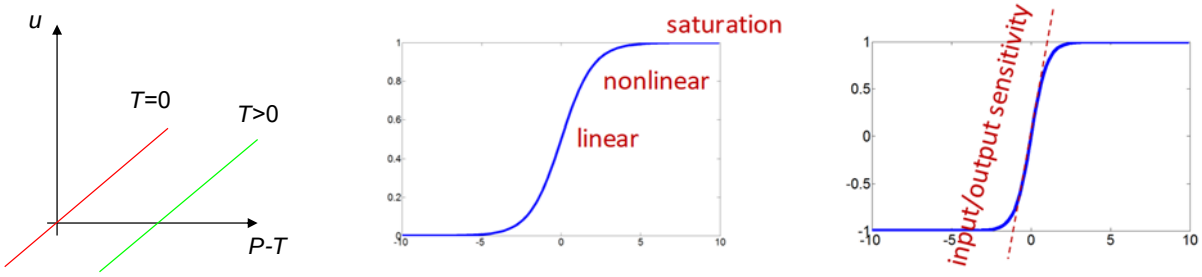


Figure 5. Activation functions: linear and sigmoidal.

The simplest use of a binary unit (single neuron with step activation function and multiple inputs) consists of deploying a decision model that features therefore a binary input-output relation. For instance, the problem of deciding whether administering either drug A or B to a patient according to exam results can be reformulated througout the ANN paradigm as depicted in Fig. 6. Assuming that we have to address three different conditions: 1) Is blood red cell level normal? 2) Is patient temperature higher than normal? 3) Is breath flow lower than normal?, the decision model combines the three conditions into the neuron by defining the score (weight) to each condition. Interestingly, the decision output will depends on the assigned weights and neuron threshold, and might change abruptely when changing the neuron parameters.

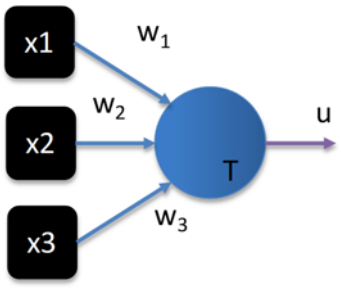


Fig. 6. Single binary unit decision model with multiple conditions (inputs)

MEANING OF THE NEURAL WEIGTHS

As formerly stated, any ANN is represented by a set of interconnected neurons (Fig. 7). The generic synaptic weight w_{ij} represent the connection strength between the j^{th} neuron to the i^{th} neuron. The output signal of the j^{th} neuron is the input signal for the i^{th} neuron. If the weight w_{ij} equals 0 then no connection between units j and i exists. When $w_{ij}>0$ then unit j excites unit i whereas $w_{ij}<0$ implies that unit j inhibits unit i. When w_{ii} differs from 0 then a self-connection exists and the neuron output signal becomes the input signal to the neuron it-self (remember that this happens "at the following time step"). In the special case where either $w_{ij}=w_{ji}$ or $w_{ij}=-w_{ji}$ the network is symmetric and anti-symmetric, respectively.

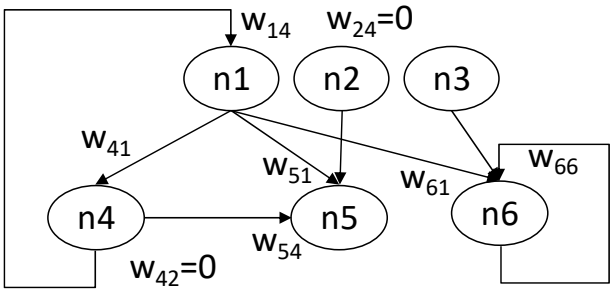


Figure 7. Generic interconnected ANN.

The neural weights of the ANN can be collected into the so-called interconnection matrix $\{w_{ij}\}$. This kind of neuron interconnection is usually said to setup a network without spatial localization as the value of the synaptic weights does not depend on the spatial location of the neurons in the network.

Differently, assuming that the neurons are arranged into the network according to a spatial criterion (metric domain), it can be assumed that the connection strength (both excitation and inhibition) between two neurons decreases as their inter-distance increases. This kind of neuron interconnection usually leads to an ANN featuring spatial localization (Fig. 8). The distance-dependent weight can be for instance expressed as:

$$w_{ij} = w_{ij}^0 f(d_{ij})$$
$$f(d_{ij}) = e^{\frac{-d_{ij}}{\lambda}}$$
$$\lambda > 0$$

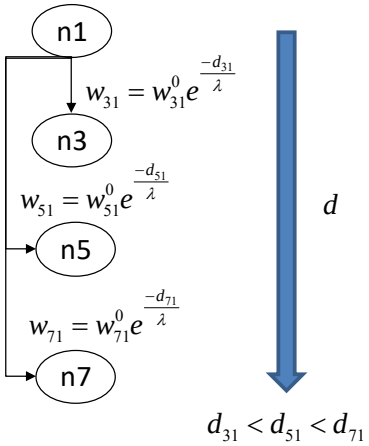


Figure 8. ANN with spatial localization.

where w_{ij}^0, f and d_{ij} are the traditional distance-independent weight, the distance function (e.g. exponential decay) and the Euclidean distance between the two neurons, respectively.

NETWORK ARCHITECTURES

A different way of looking at the ANN architecture is to consider neurons arranged in layers, being each layer responsible of processing the output of the neurons belonging to the previous layer. This kind of network is traditionally named feed-forward NN (FFNN) as the information is processed throughout a unique direction from the input up to the output, possibly throughout intermediate processing steps (internal layers). In a generic FFNN, the input layer (1st layer) consists of virtual neurons whose output becomes the input of the neurons in the 2nd layer. The number of input neurons defines the size (dimensionality) of the network input. Traditionally, the last layer is termed output layer of the FFNN (Fig. 9). The number of neurons of the output layers defines the size (dimensionality) of the network output. A generic FFNN features different input and dimensionalities. In particular cases, we can have a reference of the network output. We are going to see in the next (learning mechanisms) that this condition implies the existence of a virtual supervisor knowing how the network should nominally behave in the output when a predefined input pattern is fed into the network. The FFNN can be endowed with one or more internal (intermediate between the input and the output of the network) layers. Such layers are called hidden layers as we do not have any reference/nominal output for their (hidden) neurons. In case of the presence of

hidden layers, the ANN is called multilayer FFNN and each layer k (excluding layer 1) features its own interconnection matrix $\{w_{ij}\}_k$.

The interconnection matrix of the layer k accounts for the weights of the neurons in such a layer for input signal coming from the previous layer k-1. In a FFNN with hidden layers, the information flow is from the first input layer to the last output layer, being each active layer responsible of processing only the signals coming from the previous layer (Fig. 10 left panel).

Extending the concept of layered network, a generic interconnected ANN (Fig. 11) can feature: 1) intra-layer connections (links among neurons belonging to the same layer); 2) inter-layer connections among neurons occur when neurons belonging to a layer k sends its output signals to neurons belonging to layers k+m, with m>1; 3) backward connections from the layer k to the layer k-m. In this last condition (the information can flow

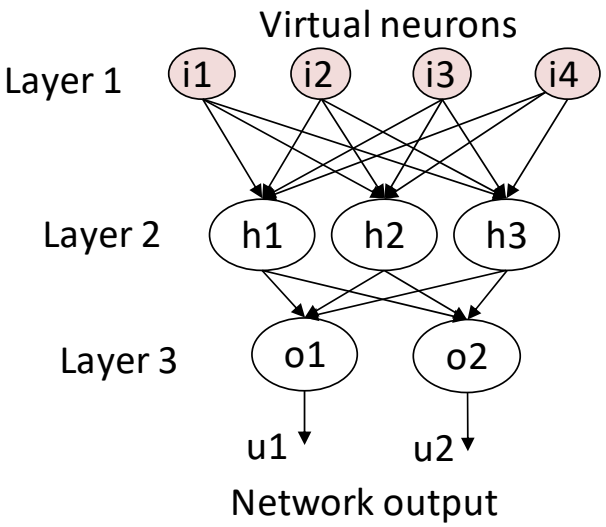


Figure 9. Multi-layer network.

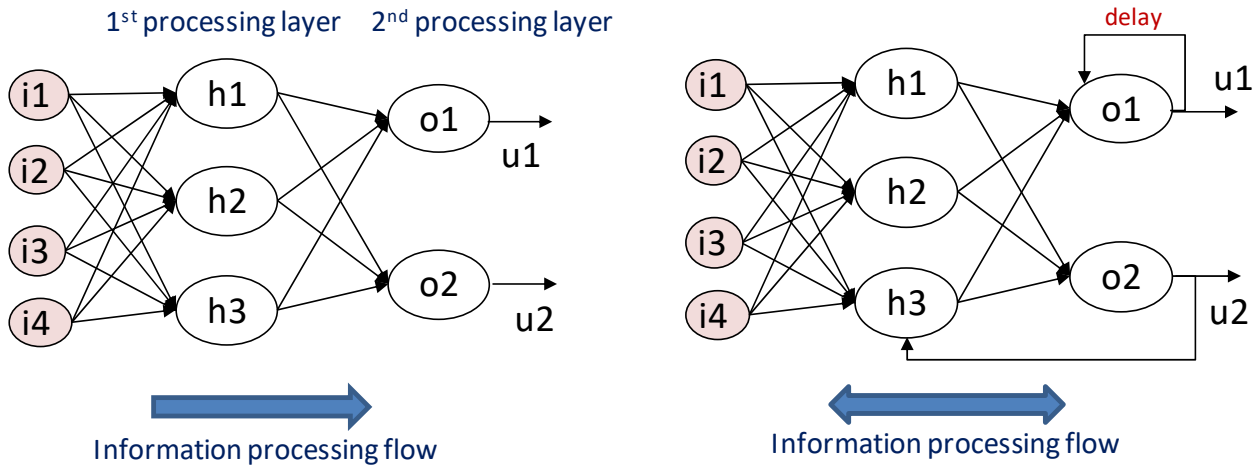
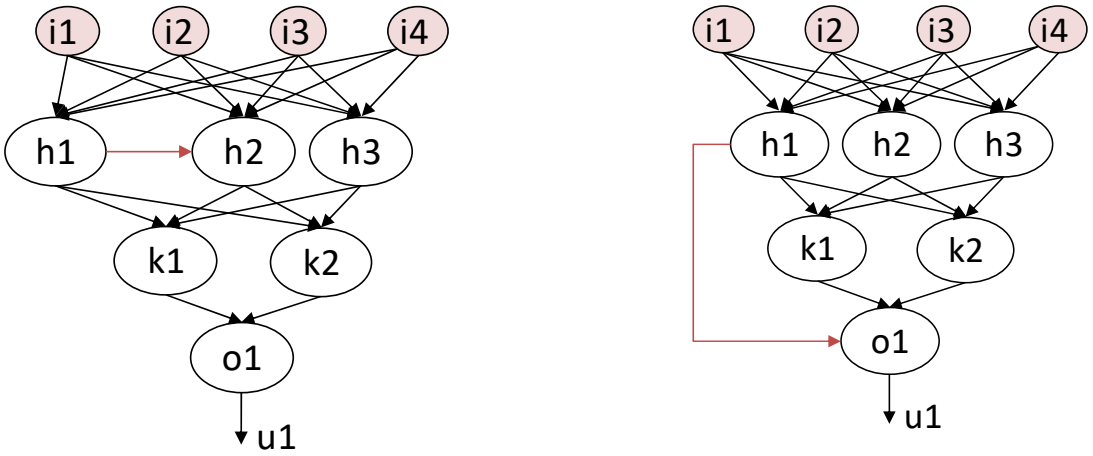


Figure 10. Feed-forward (left panel) and feed-back (right panel) networks

from input to output and from output to input as well), the ANN is called feed-back network (FBNN) (cfr. Fig. 13 right panel). In FBNN, the output of one neuron can even become an input for the neuron itself. In this case, the weight w_{ii} represents the self-connection of the neuron. As formerly stated, due to the intrinsic temporal processing of the neuron (the input is considered at time t and the output at next time interval t+1), the weight w_{ii} modulates the output signal of the neuron computed at next time step. One can think that such a neuron processes both input signals at actual time t and the signal coming from itself with a delay (τ). Into a feed-forward networks, it assumed that the neurons belonging to one layer k receives input signals only from the previous layer k-1, and send output signal only to the neurons of the next layer k+1 so that intra- and inter-layer connections are forbidden.



(a) (b)

Figure 11. Naïve neuron connections (red arrows) in FFNN: a) intra-layer, b) inter-layer links.

NEXT

In the next lecture, we will discuss about learning principles in ANN based on supervision and training methodologies based on perceptron rule, Delta rule and backpropagation in multi-layer feed-forward networks.