THE LAYERED NEURAL NETWORKS WITH GAUSSIAN ACTIVATIONS

YUGUANG FANG

DEPARTMENT OF SYSTEMS ENGINEERING

CASE WESTERN RESERVE UNIVERSITY

CLEVELAND, OHIO 44106

DECEMBER, 1990

ABSTRACT

In this paper, we have introduced a new class of artificial neural networks whose activation function is Gaussian instead of sigmoidal function and proved that this network can also approximate any continuous mapping. Moreover, a close relationship between this network and the well-known Shannon's sampling theorem is discovered.

THE LAYERED NEURAL NETWORKS WITH GAUSSIAN ACTIVATIONS

INTRODUCTION

Multilayered neural networks habe been found to be very useful in applications. They have been sucessfully used in pattern recognitions and related area, recently used in the modeling of nonlinear systems, adaptive control and identifications ([1]). Although it has been shown ([2],[3],[4]) that any continuous mapping or any measurable mapping can be approximated by the multilayered neural net, it is witnessed that to find such net (i.e., the number of nodes) is very difficult, therefore training the net is a difficult task. In general, the activation functions in the multilayered net is sigmoidal function with saturation. Intuitively, this kind activation has the following interpretation: when a neuron fires, it will fire forever. This is the main difference between the artificial neurons and the biological neuraons. In this paper, we will use the Gaussian activation function and prove that such layered networks can also approximate any continuous or measurable mapping. Since Gaussian activation functions can be used to approximate the impulse function, we can obtain the bound of the number of neural network we need to approximate the continuous mapping according to Shannon's Sampling Theorem.

Most of the proofs for the above approximation relied upon the well-known Stone-Weierstrass's theorem which is just existence proof. In this paper, we present a constructive proof, which does

not need the Stone-Weierstrass's theorem.

MAIN RESULTS

Theorem 1. Suppose that {WN} is a sequence of positive definate matrices, which has the minimal eigenvalues approximate to infinity as N becomes large, then

$$\Sigma=\left\{g(x)\,|\,g(x)=\sum_{i=1}^{m} c_i\exp\left(-(x-t_i)^T W_N(x-t_i)\right),\ t_i\in R^n,\ c_i\in R^1,\ N,m\in Z^+\right\}$$

can approximate any continuous mapping on any compact set.

Proof. We use the idea from the probability theory and the following result: for any continuous function f(x), we have

$$f(x)=\int_{-\infty}^{\infty} f(t-x)\,\delta(t)\,dt$$

where $\delta(t)$ is the Dirac function.

Define D to be the compact set of R^n, and

$$D_K=\{x\,|\,\|x\|\le K,\ x\in R^n\}$$

Since D is compact, it is also bounded, hence for large K, D is contained in D_K. For any given continuous function f(x), let F(x)=f(x), when x is in D_{2K} and F(x)=0 otherwise. Since F(x) is continuous on D_{2K}, from Cantor's theorem, F(x) is uniformly continuous on D_{2K}, i.e., for any $\varepsilon>0$, there exists $\delta>0$, so that for any x_1 and x_2 in D_{2K}, the following holds:

$$|F(x_1)-F(x_2)|<\epsilon \quad whenever \quad |x_1-x_2|<\delta .$$

We first prove that the following holds:

$$f(x)=\lim_{N\to\infty} K_N \int_{-\infty}^{\infty} F(t)\exp(-(x-t)^T W_N(x-t))\,dt \qquad (1)$$

uniformly on D, where t and x are in R^n, and K_N satisfying the following condition:

$$K_N \int_{-\infty}^{\infty} \exp(-(x-t)^T W_N(x-t))\,dt = 1.$$

Notice that since W_N is positive definate, the condition is equivalent to say that $K_N \exp(-(x-t)^T W_N(x-t))$ is the Gaussian density function with mean x and variance matrix $2W_N^{-1}$. Then we obtain

Notice that $D \subseteq D_k \subseteq D_{2k}$, hence for any $t \in D_{2k}^c$, $x \in D$, we have

$$|t-x| \geq |t| - |x| \geq 2K-K=K>\delta>0$$

and also there exists $M>0$, such that $|f(x)| \leq M$ on D_{2k}, therefore, (2) becomes

From the Gaussian distribution theory, we can obtain that $K_N = (\det W_N / \pi^n)^{1/2}$. Since W_N is positive definite, there exists an orthonormal matrix U, such that $W_N = U^T \operatorname{diag}\{\sigma_1^2,\cdots,\sigma_n^2\}\, T$ with $U^T = U^{-1}$ where σ_i^2 is the eigenvalue of W_N. Then from (3), we obtain where $\lambda(N)$ is the square root of the smallest eigenvalue of W_N.

$$\Delta_N \triangleq \left| f(x) - K_N \int_{-\infty}^{+\infty} \bar{f}(t) \exp(-(x-t)^T W_N (x-t))\, dt \right|$$

$$= \left| K_N \int_{-\infty}^{+\infty} f(x) \exp(-(t-x)^T W_N (t-x))\, dt - K_N \int_{-\infty}^{+\infty} \bar{f}(t) \exp(-(t-x)^T W_N (t-x))\, dt \right|$$

$$= K_N \left| \int_{-\infty}^{+\infty} (f(x) - \bar{f}(t)) \exp(-(t-x)^T W_N (t-x))\, dt \right|$$

$$\le K_N \int_{-\infty}^{+\infty} |f(x) - \bar{f}(t)| \exp(-(t-x)^T W_N (t-x))\, dt$$

$$= K_N \int_{D_{2k}} |f(x) - \bar{f}(t)| \exp(-(t-x)^T W_N (t-x))\, dt$$

$$+ K_N \int_{D_{2k}^c} |f(x)| \exp(-(t-x)^T W_N (t-x))\, dt$$

$$\le K_N \int_{D_{2k} \cap (|t-x| < \delta)} |f(x) - f(t)| \exp(-(t-x)^T W_N (t-x))\, dt$$

$$+ K_N \int_{D_{2k} \cap (|t-x| \ge \delta)} |f(x) - f(t)| \exp(-(t-x)^T W_N (t-x))\, dt$$

$$+ K_N \int_{D_{2k}^c} |f(x)| \exp(-(t-x)^T W_N (t-x))\, dt \qquad (2)$$

$$\Delta_N \le K_N \varepsilon \int_{D_{2k} \cap (|t-x| \le \delta)} \exp(-(t-x)^T W_N (t-x))\, dt$$

$$+ 2MK_N \int_{D_{2k} \cap (|t-x| \ge \delta)} \exp(-(t-x)^T W_N (t-x))\, dt$$

$$+ MK_N \int_{D_{2k}^c} \exp(-(t-x)^T W_N (t-x))\, dt$$

$$\le \varepsilon K_N \int_{-\infty}^{+\infty} \exp(-(t-x)^T W_N (t-x))\, dt$$

$$+ 2MK_N \int_{(|t-x| \ge \delta)} \exp(-(t-x)^T W_N (t-x))\, dt$$

$$= \varepsilon + 2MK_N \int_{(|t-x| \ge \delta)} \exp(-(t-x)^T W_N (t-x))\, dt$$

$$= \varepsilon + 2MK_N \int_{(|t| \ge \delta)} \exp(-t^T W_N t)\, dt \qquad (3)$$

From the assumption of W_N, we know that $\lambda(N)$ goes to infinity as N goes to infinity, thus for the $\varepsilon > 0$, there exist $N_1 = N_1(\varepsilon)$, such that for any $N > N_1$, we have

$$2\frac{M}{\sqrt{\pi^n}} \int_{(|t| \ge \delta\lambda(N))} \exp(t^T t)\, dt \le \varepsilon$$

From (4), we have the following for $N > N_1$

$$\Delta_N \le \varepsilon + 2M\frac{\sigma_1\cdots\sigma_n}{\sqrt{\pi^n}}\int_{(|t|\ge\delta)} \exp\left(-(Ut)^T\, diag\{\sigma_1^2,\cdots,\sigma_n^2\}(Ut)\right) dt$$

$$= \varepsilon + 2M\frac{\sigma_1\cdots\sigma_n}{\sqrt{\pi^n}}\det(U^{-1})\int_{(|U^{-1}t|\ge\delta)} \exp\left(-t^T diag\{\sigma_1^2,\cdots,\sigma_n^2\}t\right) dt$$

$$= \varepsilon + 2M\frac{\sigma_1\cdots\sigma_n}{\sqrt{\pi^n}}\int_{(|t|\ge\delta)} \exp\left(-t^T diag\{\sigma_1^2,\cdots,\sigma_n^2\}t\right) dt$$

$$= \varepsilon + 2\frac{M}{\sqrt{\pi^n}}\int_{(|diag\{\sigma_1^{-1},\cdots,\sigma_n^{-1}\}t|\ge\delta)} \exp\left(-t^T t\right) dt$$

$$\le \varepsilon + 2\frac{M}{\sqrt{\pi^n}}\int_{(|t|\ge\delta\lambda(M))} \exp\left(-t^T t\right) dt \qquad (4)$$

$$\Delta_N < 2\varepsilon .$$

Therefore, the sequence

$$\left\{ K_N\int_{-\infty}^{+\infty} f(t)\exp\left(-(t-x)^T W_N\,(t-x)\right) dt \right\}$$

is uniformly convergent to the function f(x). With small modification of the above procedure, we can prove that

$$f(x) = \lim_{N\to\infty} K_N\int_D f(t)\exp\left(-(t-x)^T W_N\,(t-x)\right) dt \qquad (5)$$

uniformly.

Now we use (5) to obtain the desired sequence in Σ. From (5), we know that for any $\varepsilon>0$, there exists an N_0, such that

$$\left| f(x) - K_{N_0}\int_D f(t)\exp\left(-(t-x)^T W_{N_0}\,(t-x)\right) dt \right| \le \frac{\varepsilon}{4} \qquad (6)$$

Using the discretization for the integral in (6), for $\varepsilon>0$ we have

$$\varepsilon/4 > \left| K\int_D f(t)\exp\left(-(t-x)'W_{N_0}(t-x)\right) dt - \sum_{i=1}^{p}\exp\left(-(x-t_i)'W_{N_0}(x-t_i)\right) \right|$$

where c_i are coefficients obtained from the definition of integration and t_i are vectors. From this inequality and the previous one, we obtain

$$|f(x)-\sum_{i=1}^{p} c_i e^{-(x-t_i)'W_{N_0}(x-t_i)}| \le |f(x)-K\int_D f(t)\, e^{-(x-t)'W_{N_0}(x-t)}\, dt| +$$

$$|K\int_D f(t)\, e^{-(x-t)'W_{N_0}(x-t)}\, dt-\sum_{i=1}^{p} c_i e^{-(x-t_i)'W_{N_0}(x-t_i)}| < \frac{\varepsilon}{4}+\frac{\varepsilon}{4} < \varepsilon$$

Thus, we have proved Theorem 1.

Suppose that we choose the following special matrix sequence $W_N=\mathrm{diag}\{1/\sigma_1, 1/\sigma_2, \ldots, 1/\sigma_n\}$, where $0<\sigma_i\to 0$ as $N\to\infty$, then we obtain the following interesting

results.

Corollary 2. Define

$$\Sigma=\left\{g(x)\,|\,g(x)=\sum_{i=1}^{p} c_i e^{-\sum_{j=1}^{n}(x_j-t_{ij})^2/\Sigma_j},\ t_{ij}\in R^1, \Sigma_j\in R^1, c_i\in R^1, N\in Z_+\right\}$$

then for any continuous function f(x) on a compact set K, we can find a sequence in Σ which is uniformly convergent to f(x), i.e., f(x) can be approximately represented by the layered net Σ through the choice of the weights c_i and σ_i and translation t_{ij}.

Observing the function of W_N in the proof of Theorem 1, we can obtain the following simpler artificial neural nets to represent any continuous mapping. Choosing $W_N=\mathrm{diag}\{N,N,\ldots,N\}$, we obtain

Corollary 3. Define

$$\Sigma=\left\{g(x)\,|\,g(x)=\sum_{i=1}^{p} c_i e^{-\frac{1}{N}\sum_{j=1}^{n}(x_j-t_{ij})^2},\ t_{ij}\in R^1, c_i\in R^1, p\in Z_+, N\in Z_+\right\}$$

then any continuous mapping on a compact set can be approximately represented by the artificial neural network Σ, i.e., for any continuous mapping on a compact set K in R^n, there exists a sequence in Σ, which uniformly converges to the mapping.

Remark. This kinds of artificial neural networks belongs to what is called functional link nets ([5]). The activation function is in the form: $\exp(-x^2)$, which is similar to Guassian density function in probability theory.

For the one-dimensional case, it deserves a special attention because it is the simplest case. We formulate as the following

Theorem 4. Let C[a,b] denotes the linear normed space of continuous functions defined on [a,b]. For any fixed positive sequence $\{\sigma_k\}$ satisfying $\sigma_k\to 0$ as $k\to\infty$, define

$$\Sigma_1=\left\{g(x)\,|\,g(x)=\sum_{i=1}^{p} c_i e^{-(x-t_i)^2/\sigma_k^2},\ c_i, t_i\in R^1, p\in Z_+\right\}$$

then for any continuous function f(x) in C[a,b], there exists a sequence in Σ_1, which uniformly approximate f(x) on [a,b].

Proof. This can be proved directly from Theorem 1.

Remark. The constructive proof in this theorem suggests a interesting interpretation. From distribution theory, we can observed that

$$\lim_{\sigma\to\infty}\frac{1}{\sqrt{\pi}\sigma}e^{-(t-x)^2/\sigma^2}=\delta(t-x)$$

hence we have

$$f(x)=\int_{-\infty}^{+\infty}f(t)\,\delta(t-x)\,dt=\lim_{\sigma\to 0}\frac{1}{\sqrt{\pi}\sigma}\int_{-\infty}^{\infty}f(t)\,e^{-(t-x)^2/\sigma^2}dt.$$

where $\delta(x)$ is the Dirac function. From this, we can see that the right hand side of the above equality is nothing but the limit of a sequence in the neural net Σ_1. Let f(t) denote a continuous signal, it is well-known in signal processing that using the sample, say, $f(t_k)$, of f(t), we can recover the signal f(t) under certain condition, i.e.,

$$f(t)=\sum_k f(t_k)\,\delta(t-t_k)\sim\sum_k\frac{f(t_k)}{\sqrt{\pi}\sigma}e^{-(t-t_k)^2/\sigma^2}.$$

Thus we can see that our neural network representations of continuous mappings are directly related to signal representation and recovery problem. It is easily see that Shannon's Sample Theorem can be used to estimate the number of nodes needed to represent a given continuous mapping by our artificial neural networks. This issue will be addressed in a separate paper.

We can also observed that in our proof the crucial point is that we use the Gaussian function to ``approxmate'' the Dirac function, it is well-known ([6]) that there are many functions which can be approximate Dirac function, therefore we can construct

as many as poosible artificial neural networks and use Shannon's sample theorem to estimate the structure. It is also possible to find some networks in this way so that the training may be much easier. Notice that because of the arbitrary choice of the sequence σ_k, we can choose suitable sequence to simplify the model and the training procedure. The following is one extremely simple artificial neural network.

Corollary 5. Define

$$\Sigma_1=\left\{g(x)\,|\,g(x)=\sum_{i=1}^{p} c_i e^{-(x-t_i)^2/N}, c_i, t_i\in R^1, N, p\in Z_+\right\},$$

then any continuous function defined on [a,b] can be approximated by the artificial neural network Σ_1.

This kind of artificial neural networks can be illustrated in the following figure.

where SUM indicates the summation and g in the circle is the Gaussian activation function, i.e., $g(x)=\exp(-x^2)$.

CONCLUSIONS

In this paper, we have proposed a new class of artificial neural networks which can be used as universal approximators for any continuous or measurable mapping. More importantly, this class of artificial neural networks is closely related to the Shannon's sampling theorem in signal processing and the signal recovery theorem can be used to estimate the structure of the given

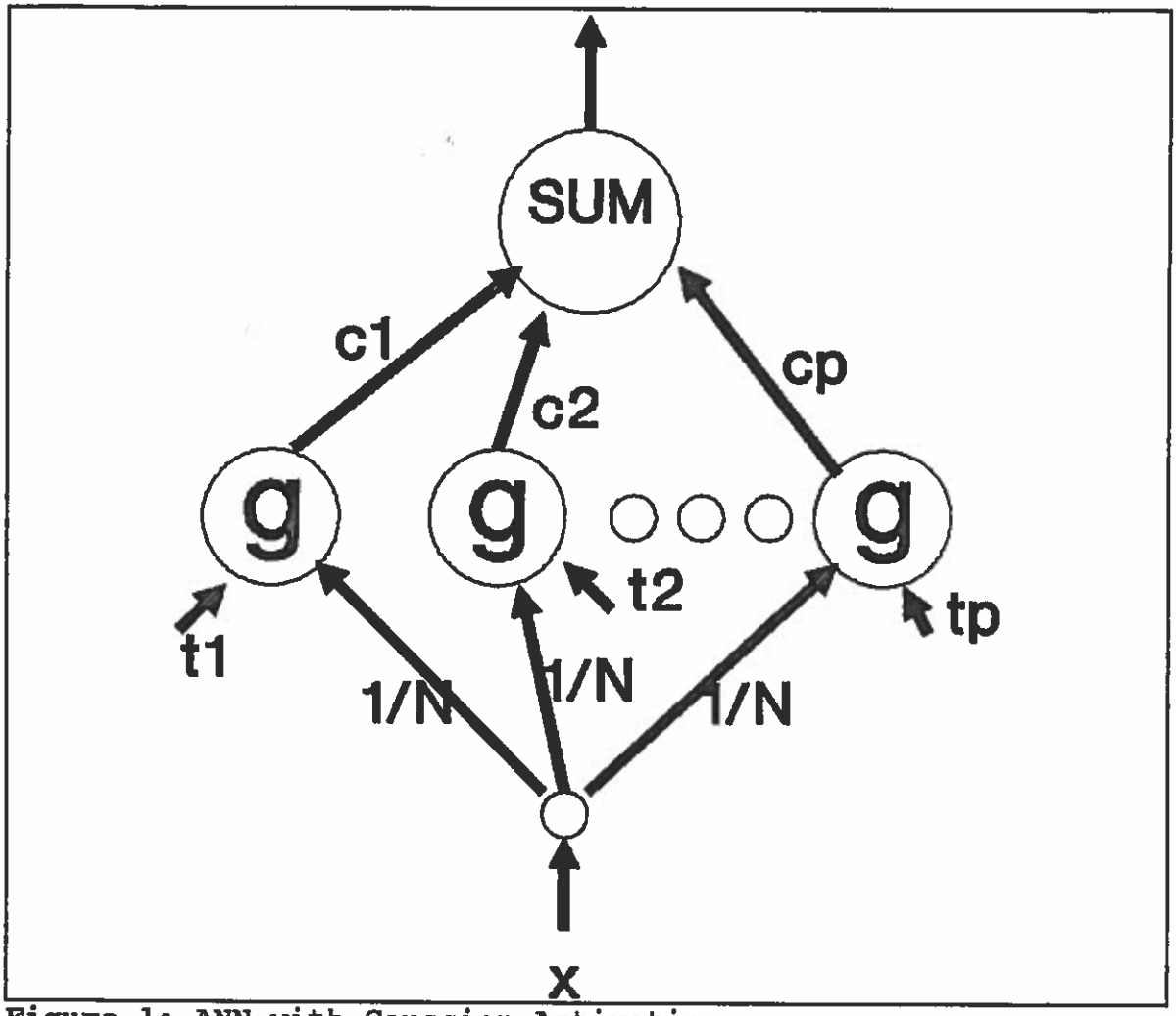

Figure 1: ANN with Gaussian Activation

continuous mapping by observing its spectrum. To the author's knowledge, this is first paper to address the relationship between the artificial neural networks and signal processing. Further research in this direction will enhance our understanding to the structure of artificial neural network and it will be of potentiality in applications.

REFERENCES

[1]. G. Cybenko, ``Approximations by Superpositions of a Sigmoidal

Functions,'' Math. Contr. Signals,Syst. Vol. 2, pp.303-314, 1989.

[2]. K. Funahashi, ``On the Approximate Realization of Continuous Mappings by Neural Networks,'' Neural Networks, Vol. 2, pp.183-192, 1989.

[3]. K. Hornick, M. Stinchcombe and H. White, ``Multilayer Feedforward Networks are Universial Approximator,'' Neural Networks, Vol.2, pp.359-366, 1989.

[4]. A. Kolmogrov, ``On the Representation of Continiuous Functions of Many Variables by Superposition of Functions of One Variable and Addition,'' Dolk. Akad. Nauk. USSR, Vol. 114, pp.953-956, 1957.

[5]. Y.H. Pao, Adaptive Pattern Recognition and Neural Networks, Addison Wesley Publishing Company,Inc. 1989.

[6]. T. Kailath, Linear Systems, Prentice-Hall, New Jersey, 1980.