UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ENGENHARIA DE SÃO CARLOS

SEL632 - LINGUAGENS DE DESCRIÇÃO DE HARDWARE

Entrega 01

Leonardo J. R. Baptistella - NºUSP 10821172
Murilo Henrique Pasini Trevisan - NºUSP 9796078

São Carlos

2022

1 NEURÔNIO ARTIFICIAL

A partir da estrutura e funcionamento do neurônio biológico, pesquisadores tentaram simular este sistema em computador. O modelo mais bem aceito foi proposto por Warren McCulloch e Walter Pitts em 1943, o qual implementa de maneira simplificada os componentes e o funcionamento de um neurônio biológico. Em termos simples, um neurônio matemático de uma rede neural artificial é um componente que calcula a soma ponderada de vários **inputs**, aplica uma função e passa o resultado adiante.(O..., Acessado em maio 2022)

Neste modelo de neurônio matemático, os impulsos elétricos provenientes de outros neurônios são representados pelos chamados sinais de entrada (representado pela letra x no diagrama 3 abaixo, que nada mais são do que os dados que alimentam seu modelo de rede neural artificial). Dentre os vários estímulos recebidos, alguns excitarão mais e outros menos o neurônio receptor e essa medida de quão excitatório é o estímulo é representada no modelo de Warren McCulloch e Walter Pitts através dos pesos sinápticos. Quanto maior o valor do peso, mais excitatório é o estímulo. Os pesos sinápticos são representados por w_{kn} no diagrama 3, onde k representa o índice do neurônio em questão e n se refere ao terminal de entrada da sinapse a qual o peso sináptico se refere.(O..., Acessado em maio 2022)

A soma ou corpo da célula é representada por uma composição de dois módulos, o primeiro é uma junção aditiva, somatório dos estímulos (sinais de entrada) multiplicado pelo seu fator excitatório (pesos sinápticos), e posteriormente uma função de ativação, que definirá com base nas entradas e pesos sinápticos, qual será a saída do neurônio. O axônio é aqui representado pela saída (yk) obtida pela aplicação da função de ativação. Assim como no modelo biológico, o estímulo pode ser excitatório ou inibitório, representado pelo peso sináptico positivo ou negativo respectivamente.(O..., Acessado em maio 2022)

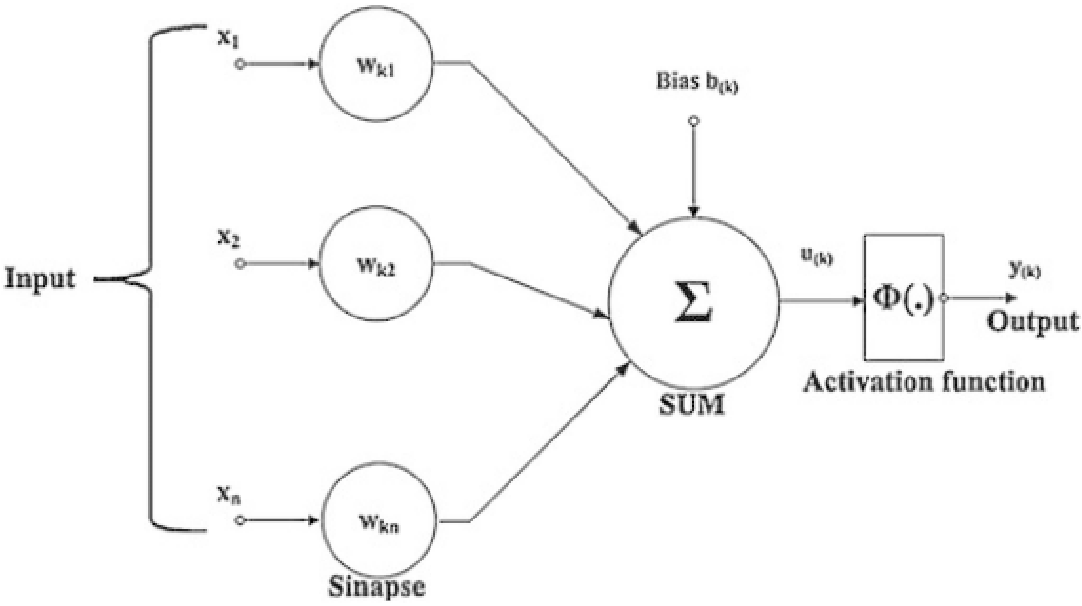

Figura 1 – Representação Simplificada do Neurônio Matemático
Fonte: (, Acessado em maio 2022b)

O modelo proposto possui uma natureza binária. Tanto os sinais de entrada quanto a saída, são valores binários. McCulloch acreditava que o funcionamento do sistema nervoso central possuía um caráter binário, ou seja, um neurônio influencia ou não outro neurônio, mas posteriormente mostrou-se que não era dessa forma.(O. . . , Acessado em maio 2022)

Os dendritos e axônios são representados matematicamente apenas pelas sinapses, e a intensidade da ligação é representada por uma grandeza denominada peso sináptico, simbolizada pela letra w. Quando as entradas, x, são apresentadas ao neurônio, elas são multiplicadas pelos pesos sinápticos correspondentes, gerando as entradas ponderadas, ou seja, x_1 que multiplica w_1 e assim por diante. Isso descreve uma das bases matemáticas do funcionamento de uma rede neural artificial, a multiplicação de matrizes.(O. . . , Acessado em maio 2022)

$$\boldsymbol{Xw} = \boldsymbol{y}$$

$$\begin{bmatrix} 1 & x_{11} & \dots & x_{1d} \\ 1 & x_{21} & \dots & x_{2d} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{nd} \end{bmatrix} \times \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}$$

Figura 2 – Multiplicação de Matrizes Entre Sinais de Entrada x e Pesos Sinápticos w
Fonte: (, Acessado em maio 2022a)

O neurônio, então, totaliza todos os produtos gerando um único resultado. A esta função se denomina função de combinação. Este valor é então apresentado a uma função de ativação ou função de transferência, que tem, dentre outras, a finalidade de evitar o acréscimo progressivo dos valores de saída ao longo das camadas da rede, visto que tais funções possuem valores máximos e mínimos contidos em intervalos determinados. O uso de funções de transferência não-lineares torna a rede neural uma ferramenta poderosa.

Sabe-se que uma rede perceptron de duas camadas com função de transferência não-linear como a função sigmóide, é denominada de aproximador universal.(O..., Acessado em maio 2022)

Dessa forma, o neurônio dispara quando a soma dos impulsos que ele recebe ultrapassa o seu limiar de excitação, chamado de threshold. O corpo do neurônio, por sua vez, é emulado por um mecanismo simples que faz a soma dos valores x_i e w_i recebidos pelo neurônio (soma ponderada) e decide se o neurônio deve ou não disparar (saída igual a 1 ou a 0) comparando a soma obtida ao limiar ou threshold do neurônio. A ativação do neurônio é obtida através da aplicação de uma "função de ativação", que ativa a saída ou não, dependendo do valor da soma ponderada das suas entradas.(DEEP LEARNING BOOK, 2018)

Nota-se que este modelo matemático simplificado de um neurônio é estático, ou seja, não considera a dinâmica do neurônio natural. No neurônio biológico, os sinais são enviados em pulsos e alguns componentes dos neurônios biológicos, a exemplo do axônio, funcionam como filtros de frequência.(O..., Acessado em maio 2022)

Portanto, num neurônio artificial, temos o seguinte esquema:

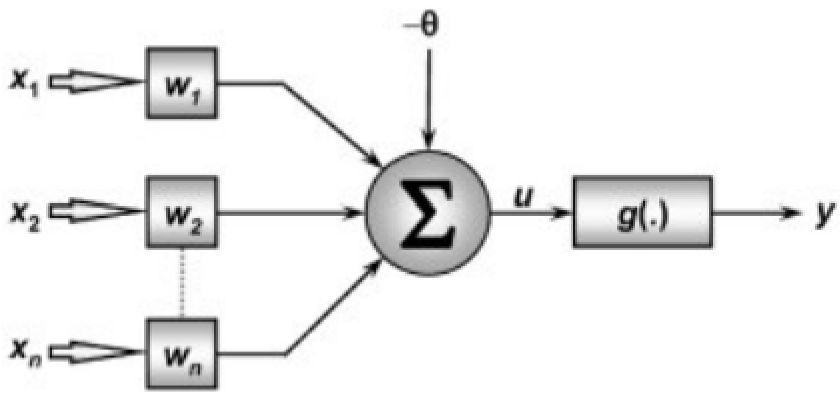

Figura 3 – Representação do Neurônio Artificial
Fonte: (, Acessado em maio 2022c)

Da forma que é possível encontrar os seguintes componentes:

- Os sinais de entrada $(x_1, ..., x_n)$ são os sinais externos normalmente normalizados para incrementar a eficiência computacional dos algoritmos de aprendizagem.

- Os pesos sinápticos $(w_1, ..., w_n)$ são valores para ponderar os sinais de cada entrada da rede.
- O combinador linear $(\sum)$ é responsável por agregar todos os sinais de entrada que foram ponderados pelos respectivos pesos sinápticos a fim de produzir um potencial de ativação.
- O limiar de ativação (θ) especifica qual será o patamar apropriado para que o resultado produzido pelo combinador linear possa gerar um valor de disparo de ativação
- O potencial de ativação (u) é o resultado obtido pela diferença do valor produzido entre o combinador linear e o limiar de ativação. Se o valor for positivo, então o neurônio produz um potencial excitatório, caso contrário, o potencial será inibitório.
- A função de ativação (g) tem o objetivo de limitar a saída de um neurônio em um intervalo valores.
- O sinal de saída (y) é o valor final de saída podendo ser usado como entrada de outros neurônios que estão sequencialmente interligados.

2 REPRESENTAÇÃO EM PONTO FIXO: Q-FORMAT

A representação de números reais em um sistema digital pode ser com ponto fixo, ou com ponto flutuante. Sistemas com ponto flutuante permitem variação da quantidade de dígitos antes e após o ponto. A representação é feita no seguinte formato (FRANCO, 2006):

$$F(\beta, t, m, M) = \pm 0.d_1 d_2..D - t \cdot \beta^e, d_1 \neq 0, m \leq e \leq M$$

Percebe-se uma extensão de possíveis valores que o expoente e pode assumir, o que permite a representação de ordens de grandeza muito distintas com um mesmo tipo de dado. Representações em ponto fixo, por sua vez, possuem resolução fixa. Um equacionamento possível é (FRANCO, 2006):

$$x = \pm \sum_{i=k}^{n} x_i \cdot B^{-i}$$

Q-format é um formato de representação de ponto fixo caracterizada pela especificação da quantidade de bits, tanto para bits destinados a dígitos inteiros quanto para bits destinados a dígitos fracionários (Q. . . , 2021). Se o formato for Q3.2, por exemplo, a parte inteira vai de 0 a 7, e a parte fracionária de 0 a 3. Aqui, introduz-se a notação de Q-format (OBERSTAR, 2007):

$$Q[QI] \cdot [QF] \tag{2.1}$$

Em que QI indica o número de bits para inteiros e QF o número de bits para fracionários. Essa forma de representação implica em uma ambiguidade envolvendo o sinal, que pode, ou não, ter seu bit incluído em QI. No caso de dados do tipo unsigned, tanto QI quanto QF têm range descrito em:

$$0 \leq \alpha \leq (2^{Q*} - 1)$$

Já no caso de serem representados valores positivos e negativos, com valores negativos em complemento de 2, o range de QI passa a:

$$(-2^{QI-1}) \leq \alpha \leq (2^{QI} - 1)$$

É interessante notar que a resolução ϵ dos dados é determinada pela porção QF:

$$\epsilon = 2^{-QF}$$

Dessa forma, é possível descrever uma representação em Q-format de modo mais completo, levando em consideração o range e resolução totais - o que é de extrema importância para que se determine um bom formato para as necessidades da implementação a ser realizada. O range total para dados unsigned e signed, bem como a resolução, são descritos, respectivamente, por:

$$0 \leq \alpha \leq 2^{QI} - 2^{QF}$$

$$-2^{QI-1} \leq \alpha \leq 2^{QI-1} - 2^{QF}$$

$$\epsilon = 2^{-QF}$$

Ainda, é importante definir o tamanho de dados em Q-format. Como QI e QF indicam a quantidade de bits disponíveis para representação do número, é simples que o tamanho é dado por:

$$WL = QI + QF$$

Vale ressaltar que podem haver diferentes formatos de representação Q-format com mesmo WL, e QI e QF distintos. Quanto maior QI, maior o range da representação, quanto maior QF, maior a resolução.

Passadas a caracterização geral de representação em Q-format, é interessante abordar a conversão de valores em ponto flutuante para ponto fixo e vice-versa. Para Q-format, o ponto é, de certa forma, virtual, já que na memória, o valor é guardado em um conjunto de bits de tamanho WL. Tome, como exemplo, Q2.6., unsigned. Os valores representados vão de $0 \leq \alpha \leq 3.984375$, com resolução $\epsilon = 0.015625$. Para passar um valor α (cuja representação é feita em ponto flutuante), para ponto fixo, basta dividir o valor pela resolução ϵ e aproximar para o inteiro mais próximo:

$$\alpha_{fixo} = \frac{\alpha_{flutuante}}{\epsilon}$$

Tomando $\alpha_{flutuante} = 3.984375$, temos que $\alpha_{fixo} = 255$. Em binário, 255 é representado por 11111111, ou seja, possui os 8 bits iguais a 1, correspondendo, corretamente, ao maior valor possível do formato. Esse processo de conversão faz sentido, pois divide o número a ser armazenado pela resolução do Q-format adotado, ou seja, em passos inteiros do formato (por isso a aproximação após a divisão pela resolução).

A conversão de ponto fixo para ponto flutuante segue o caminho inverso: primeiro é necessário converter o valor armazenado em binário como um inteiro e multiplicar o resultado pela resolução ϵ.

Em razão de se tratar de um formato mais rígido e com mais limitações de representação, o uso de Q-format implica em certa perda de precisão. Porém, a eficiência das implementações com Q-format é maior, em decorrência da facilidade de realizar operações matemáticas com os valores armazenados (OBERSTAR, 2007).

REFERÊNCIAS

Matrix. Acessado em maio 2022. Disponível em: <https://i0.wp.com/www.deeplearningbook.com.br/wp-content/uploads/2018/01/matriz.png?resize=300%2C147ssl=1>. Citado na página 2.

Neurônio. Acessado em maio 2022. Disponível em: <https://i1.wp.com/www.deeplearningbook.com.br/wp-content/uploads/2018/01/neuronio-matematico.png?w=543ssl=1>. Citado na página 2.

Neurônio. Acessado em maio 2022. Disponível em: <https://i1.wp.com/www.deeplearningbook.com.br/wp-content/uploads/2018/01/neuronio.jpeg?resize=300%2C137ssl=1>. Citado na página 3.

FRANCO, N. B. *Cálculo numérico*. [S.l.]: Pearson, 2006. Citado na página 5.

O Neurônio, Biológico e Matemático O Neurônio Matemático. Acessado em maio 2022. Disponível em: <https://www.deeplearningbook.com.br/o-neuronio-biologico-e-matematico/>. Citado 3 vezes nas páginas 1, 2 e 3.

OBERSTAR, E. *Fixed-Point Representation and Fractional Math (revision 1.2)*. [S.l.]: August, Oberstar Consulting, online, available at http://superkits. net . . . , 2007. Citado 2 vezes nas páginas 5 e 7.