

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL  
INSTITUTO DE INFORMÁTICA  
CURSO DE CIÊNCIA DA COMPUTAÇÃO

ALAN FICAGNA

**Reconhecendo Caracteres Escritos à  
Mão Utilizando Algoritmos de Deep  
Learning**

Monografia apresentada como requisito parcial  
para a obtenção do grau de Bacharel em Ciência  
da Computação

Orientador: Prof. Dr. Paulo Martins Engel

Porto Alegre  
2015

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Reitor: Prof. Carlos Alexandre Netto

Vice-Reitor: Prof. Rui Vicente Oppermann

Pró-Reitor de Graduação: Prof. Sérgio Roberto Kieling Franco

Diretor do Instituto de Informática: Prof. Luis da Cunha Lamb

Coordenador do Curso de Ciência de Computação: Prof. Carlos Arthur Lang Lisbôa

Bibliotecária-chefe do Instituto de Informática: Beatriz Regina Bastos Haro

*“Virtue is more to be feared than vice,  
because its excesses are not subject to the regulation of consciousness”*

— ADAM SMITH

## RESUMO

Este documento é uma revisão do estado da arte em relação as técnicas de Deep Learning utilizadas no reconhecimento de caracteres escritos à mão, tendo o banco de dados MNIST como meio de comparação. Além disso são apresentadas algumas variações nos meta parâmetros do algoritmo tais como <TODO> para avaliar o impacto no tempo de execução e na qualidade do resultado.

**Palavras-chave:** IA. inteligência artificial. deeplearning.

# Recognizing Handwritten Characters Using Deep Learning Algorithms

## ABSTRACT

This document is a review of the state of the art in regards to Deep Learning techniques used in handwritten characters recognition having the MNIST database for comparison. Also, the impact in the change of parameters such as <TODO> was measured to provide and indication of its effects on run speed and quality of result.

**Keywords:** IA, deeplearning.

## LISTA DE FIGURAS

Figura 3.1	Abstração de um neurônio artificial .....	16
Figura 3.2	Exemplo de funções de limite $\sigma$ usadas em neurônios artificiais .....	16
Figura 3.3	Exemplo de função aproximada com e sem o vetor de bias .....	17
Figura 3.4	Exemplo de uma rede neural de 3 camadas .....	18
Figura 3.5	Figura com o espaço de solução .....	20
Figura 3.6	Uma rede neural simples de um único neurônio.....	20
Figura 3.7	Exemplo de funções de ativação.....	21
Figura 3.8	Uma rede neural com a aplicação de dropout .....	23
Figura 3.9	Tipos de não linearidades aplicadas .....	24

## LISTA DE TABELAS

## LISTA DE ABREVIATURAS E SIGLAS

MNIST	Mixed National Institute of Standards and Technology
CNN	Convolutional Neural Network
GPU	Graphical Processing Unit
ReLU	Rectified Linear Unit



## SUMÁRIO

<b>1 INTRODUÇÃO</b>	<b>10</b>
1.1 Objetivos .....	10
<b>2 HISTÓRICO</b>	<b>11</b>
2.1 Inteligência Artificial .....	11
2.1.1 Inteligência Humana .....	11
2.1.2 Inteligência Racional .....	11
2.2 Histórico da Inteligência Artificial .....	12
2.2.1 Primeiros anos .....	12
2.2.2 Maturação e Estado da Arte .....	13
<b>3 REVISÃO SOBRE REDES NEURAIS</b>	<b>15</b>
3.0.1 O neurônio artificial.....	15
3.0.1.1 Bias.....	17
3.0.1.2 Redes neurais.....	17
3.0.1.3 Uma intuição sobre aprendizagem .....	19
3.1 Deep Learning.....	21
3.1.1 Introdução .....	21
3.1.2 Redes Convolucionais .....	22
3.1.2.1 Dropout .....	22
3.1.2.2 ReLU .....	23
3.1.2.3 Regularização.....	24
<b>REFERÊNCIAS</b>	<b>26</b>

## 1 INTRODUÇÃO

Nos últimos anos houve um grande avanço em áreas como a Computação Visual graças ao ressurgimento de um campo da Inteligência Artificial conhecido como Redes Neurais. Este progresso foi responsável pelo despertar do interesse e investimentos de grandes empresas como Google, Facebook e outros que apostam no desenvolvimento de aplicações que há poucas décadas atrás eram apenas tema de ficção científica tais como scanners faciais de alta precisão, reconhecimento de objetos em cenas, carros que não precisam de motorista e vários tipos de classificadores em geral. Essas novas técnicas ficaram conhecidas como Deep Learning e prometem fazer uso da enorme quantidade de dados e informações disponíveis na internet e nas bases de dados das empresas.

Este progresso foi causado por fatores como o incremento da capacidade de processamento dos computadores e a ubiquidade de tecnologias como GPUs que facilitaram o processamento vetorial e paralelo de imagens, sons e outros domínios de alta dimensionalidade que até então possuíam um custo computacional proibitivo. Além disso, novas contribuições teóricas também podem ser apontadas como responsáveis pela melhoria de desempenho tais como aprendizado semi supervisionado, uso de aprendizado baseado em energia <TODO COMPLETE>, <CITE>

### 1.1 Objetivos

O objetivo deste trabalho consiste em efetuar uma revisão do histórico do desenvolvimento das técnicas de Deep Learning e um resumo do seu estado da arte com a finalidade de adquirir proficiência no uso das mesmas. Além disso, serão apresentados resultados práticos obtidos com a execução de uma modelo que utiliza Deep Learning chamado de Rede Neural Convolutacional (CNN) no reconhecimento de caracteres escritos à mão utilizando o algoritmo da rede LeNet de LeCun et al sobre a base MNIST.

## 2 HISTÓRICO

### 2.1 Inteligência Artificial

Definir o que é inteligência não é uma tarefa fácil, e inteligência artificial é mais difícil ainda. Historicamente, existiram duas abordagens para a definição de inteligência: O comparativo com humanos e o comparativo com seres racionais.

#### 2.1.1 Inteligência Humana

No primeiro modelo, o seres da espécie humana são tidos como parâmetros e portanto, qualquer avaliação de inteligência tem que se refletir em comportamentos ou pensamentos similares aos humanos. O Teste de Turing proposto pelo matemático Alan Turing (1950) foi uma maneira operacional de estabelecer inteligência pelo comparativo comportamental. Em sua versão mais simplificada, um computador poderia ser considerado inteligente se, ao ser interrogado por um humano com algumas questões, este não pudesse determinar se as respostas eram oriundas de um computador ou de uma pessoa. A Ciência Cognitiva, por outro lado, busca compreender o mecanismo pelos quais os humanos são capazes de pensar e à partir disso estabelecer comparações com algum algoritmo que eventualmente pudesse replicar estas capacidades.

#### 2.1.2 Inteligência Racional

Além do paradigma humano, existe um outro que é a comparação com seres racionais. Um ser é dito racional se ele age da “maneira correta”, dado o que ele conhece. Isso nos leva a seguinte definição: um ser é inteligente se ele age ou pensa como um ser racional.

Aristóteles foi o primeiro a sistematizar uma maneira de “pensar corretamente”. Ele foi o criador da disciplina que hoje conhecemos por Lógica que estabelecia estruturas argumentativas tais como: “Sócrates é um homem; todos os homens são mortais; portanto, Sócrates é mortal” as quais sempre produzem conclusões corretas. Porém, estes silogismas nem sempre são capazes de traduzir conhecimento informal, principalmente em face à incertezas e por isso tendem a não ser reconhecidos como tudo que existe em termos de

inteligência.

Além disso existem diferenças entre resolver um problema em princípio e resolvê-lo na prática. Embora a abordagem pelas “leis do pensamento” do Aristóteles foquem em realizar inferências corretas, esta não é capaz de exaurir o que significa agir como um ser racional pois existem situações em que não há uma ação comprovadamente correta, mas em que uma ação deve ser tomada mesmo assim. Além disso, existem comportamentos racionais que não podem ser deduzidos por mera inferência como, por exemplo, o reflexo de remover rapidamente a mão do fogo sem uma cuidadosa pré deliberação sobre o assunto.

O propósito deste trabalho não é exaurir todos os ângulos pelos quais esta questão já foi abordada, tais como psicologia, lógica e filosofia mas sim apresentar apenas uma introdução que permita entender o surgimento do o campo da Inteligência Artificial dentro da ciência da computação.

## **2.2 Histórico da Inteligência Artificial**

### **2.2.1 Primeiros anos**

O primeiro trabalho a ser considerado efetivamente “Inteligência Artificial” foi feito em 1943 por Warren McCulloch e Walter Pits, no qual foi proposto o modelo de funcionamento de um neurônio artificial. Este neurônio podia assumir os estados “ligado” e “desligado”, determinado em resposta aos estímulos proporcionados pelos outros neurônios vizinhos. Foi demonstrado também que uma rede de neurônios artificial conseguia computar qualquer função computável e sugerido que a mesma poderia aprender.

Em 1949 Donald Hebb demonstrou uma simples regra pela qual a força ou os pesos numéricos de conexão entre os neurônios poderia ser modificada, a qual ficou conhecida como aprendizado hebbiano. Logo em seguida, em 1950, foi construído o primeiro computador de rede neural chamado de SNARC que fora feito com 3000 tubos de vácuo e conseguia simular uma rede de 40 neurônios.

Embora o neurônio artificial seja considerado o primeiro trabalho, o termo Inteligência Artificial só viria a ser cunhado mais de 10 anos depois em Dartmouth nos Estados Unidos em um workshop organizado por John McCarthy que duraria dois meses e cujo propósito era estudar teoria dos autômatos, redes neurais e inteligência. Embora não tenha produzido nenhum progresso, o workshop serviu para introduzir as principais figuras da área entre si.

Nos seus primeiros anos a IA obteve bastante sucesso. Um a um, os itens da lista “uma máquina nunca vai poder fazer X” compilada por Turing iam sendo desbancados. Newel and Simon criaram o primeiro GPS (General Problem Solver), um programa desenhado para imitar a maneira como humanos pensam ao invés de simplesmente seguir regras lógicas. Na classe limitada de desafios que conseguia resolver, a ordem dos subobjetivos que o programa tentava seguir era similar à humana. Na IBM, Nathaniel Rochester criaria o primeiro Provador de Teoremas Geométricos, o qual conseguia provar teoremas que vários estudantes consideravam complexos. Arthur Samuel refutou a ideia de que programas somente conseguiriam fazer o que lhes eram dito criando um programa de computador que conseguia jogar um jogo melhor que o seu criador.

Foi nessa época também que McCarthy criaria sua linguagem de programação *Lisp* que viraria a linguagem dominante em IA e com a qual ele faria o primeiro algoritmo que aceitava novos axiomas, sendo então o primeiro autômato que conseguia adquirir novas capacidades sem necessitar ser reprogramado. Outras contribuições significativas foram o programa SAINT (1963) que conseguia resolver problemas de Cálculo típicos do primeiro ano de curso, o programa ANALOGY (1968) respondia a perguntas de analogia geométrica como as de testes de QI, o programa STUDENT (1967) resolvia problemas algébricos e o programa SHRDLU (1972) era capaz de executar instruções do tipo “Encontre um bloco maior do que você está segurando agora e o coloque na caixa”. Foi nessa época também que foram feitas melhorias no processo de aprendizado Hebbiano que ficaram conhecidas como redes *adelines*.

### 2.2.2 Maturação e Estado da Arte

O sucesso inicial foi seguido de um período de relativa estagnação. O fato de que estes algoritmos só conseguiam fazer manipulação simbólica e não conheciam nada sobre o domínio que trabalhavam provou ser um grande empecilho. Além disso a maioria dos primeiros programas de IA resolviam o problema tentando várias combinações até que o resultado fosse atingido, o que não era factível para instâncias de problemas marginalmente maiores. Foi nesta época que Minsky e Papert provaram que um neurônio artificial agora conhecido como *perceptron* era restrito no tipo de funções que conseguia representar e o financiamento de pesquisas na area se tornou escasso. Já na década de 1970, houve o surgimento dos sistemas baseados em conhecimento prévio como o programa DENDRAL que viu uma melhora significativa de desempenho na tarefa de inferir a estrutura mo-

lecular à partir da massa dos seus componentes. O programa que inicialmente tentava todas as alternativas possíveis passou a utilizar de conhecimento de especialistas em química para identificar subestruturas conhecidas o que reduzia significadamente o número de tentativas.

O primeiro caso de sucesso comercial reportado destes sistemas especialistas era um programa que ajudava na compra de novos computadores e que teria economizado à empresa US\$40 milhões por ano, o que explica o fato destes sistemas continuarem sendo usados até hoje.

Paralelamente, houve o ressurgimento das redes neurais no final da década de 1980, proporcionado pela reinvenção do algoritmo de *back-propagation*, uma maneira eficiente de atualizar os pesos das conexões nos neurônios, e também uma melhoria nas práticas de pesquisa na área que passou a exigir um maior rigor científico acompanhado do uso de bancos de dados padrões que permitiam o teste e comparação dos resultados entre diferentes algoritmos. Em uma outra linha recente, a ênfase no algoritmo da lugar ao uso de grandes quantidades de dados inspirado no trabalho de Yarowsky (1995) que demonstrou ser possível reconhecer se o uso da palavra *plant* significava uma flor ou uma fábrica com precisão acima de 96% sem utilizar de rótulos fornecidos por humanos, apenas com as definições do dicionário desta palavra. Trabalhos como o dele sugerem um “gargalo de conhecimento” na IA, ou seja, o conhecimento que o um sistema precisa pra resolver um problema pode vir de grandes massas de dados ao invés de intervenções de especialistas na área. Hoje em dia, a IA está em várias aplicações de campos diferentes, como por exemplo:

- Na área de robótica veicular com carros auto guiados.
- Na área de reconhecimento de fala.
- Planejamento automático de tarefas.
- Jogos, com destaque para o computador DEEP BLUE da IBM que derrotou o campeão mundial de xadrez, Garry Kasparov.
- Controle de SPAM em mensagens de e-mail.
- Militar e Logística.

### 3 REVISÃO SOBRE REDES NEURAIS

As primeiras definições de uma rede neural artificial foram feitas em 1943 por McCulloch e Pits inspiradas na hipótese de que as atividades mentais do cérebro humano consistem em reações eletroquímicas em redes de células cerebrais chamadas de neurônios. O neurônio artificial consiste em uma unidade computacional abstrata capaz de receber estímulos de entrada e produzir uma ou mais saídas e uma rede neural artificial — de agora em diante referenciada apenas por *rede neural* — consiste em um grafo direcionado cujos nodos são neurônios artificiais.

#### 3.0.1 O neurônio artificial

Cada neurônio recebe  $m$  entradas ponderadas por um peso  $w$ , correspondente a entrada do seu  $i$ -ésimo vizinho. Esta configuração foi inspirada nos neurônios naturais cujas interconexões são reguladas por um potencial de ativação responsável por diferenciar a intensidade entre as mesmas. A saída é produzida aplicando-se a função limite  $\sigma$  na soma de todas as entradas ponderadas do neurônio. Esta função simula a *lógica de limite* existente nos neurônios naturais, os quais somente irão propagar o seu sinal caso haja um acúmulo suficiente de potencial elétrico no neurônio. Em sua forma matemática, a saída  $y$  do neurônio artificial pode ser expressada da seguinte forma:

$$y = \sigma\left(\sum_{i=1}^m x_i w_i + b\right)$$

A função  $\sigma$  representa na verdade uma classe de funções que podem ser utilizadas dependendo do objetivo da construção da rede neural. Para problemas de classificação *linearmente separáveis*, pode ser usada a função limite exemplificada na figura 3.2, enquanto para problemas não linearmente separáveis é necessário utilizar algum tipo de função *sigmoide*. A função logística é um exemplo de função sigmoide frequentemente utilizada por ser diferenciável em todo o seu domínio, o que permite produzir garantias de convergência nos algoritmos de aprendizagem.

O valor  $b$  corresponde ao *bias* aplicado a função e pode ser omitido simplesmente inserindo-se artificialmente uma nova entrada  $x_0 = 1$ .

Figura 3.1 – Abstração de um neurônio artificial

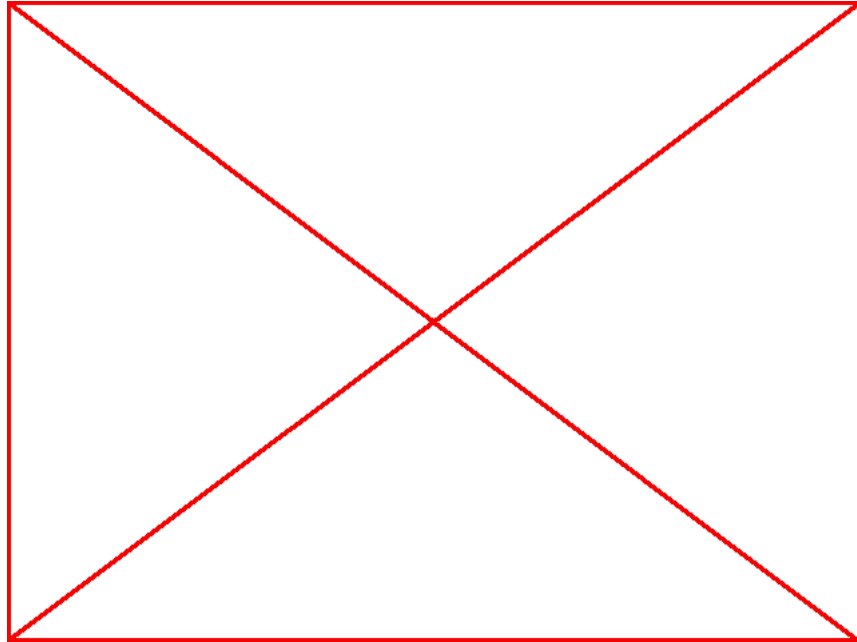


Figura 3.2 – Exemplo de funções de limite  $\sigma$  usadas em neurônios artificiais

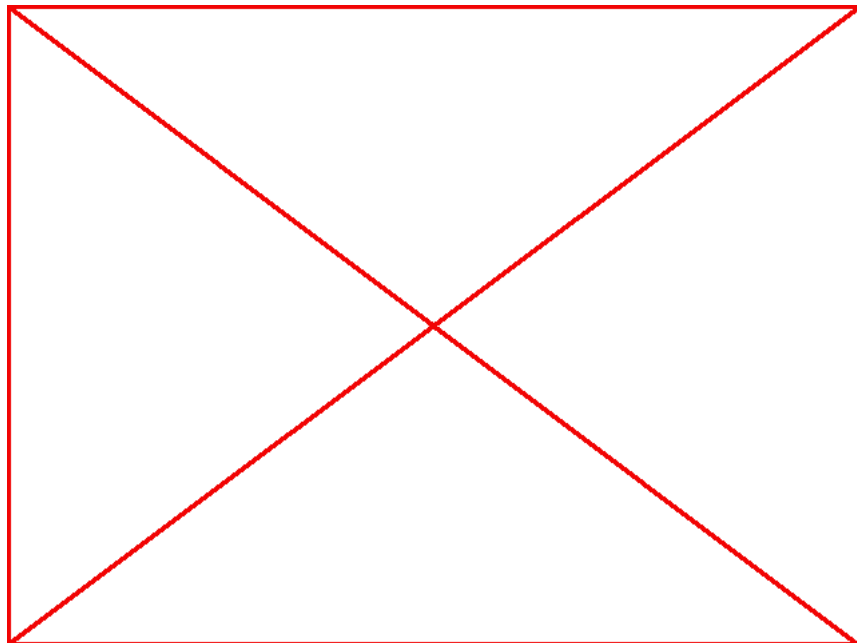
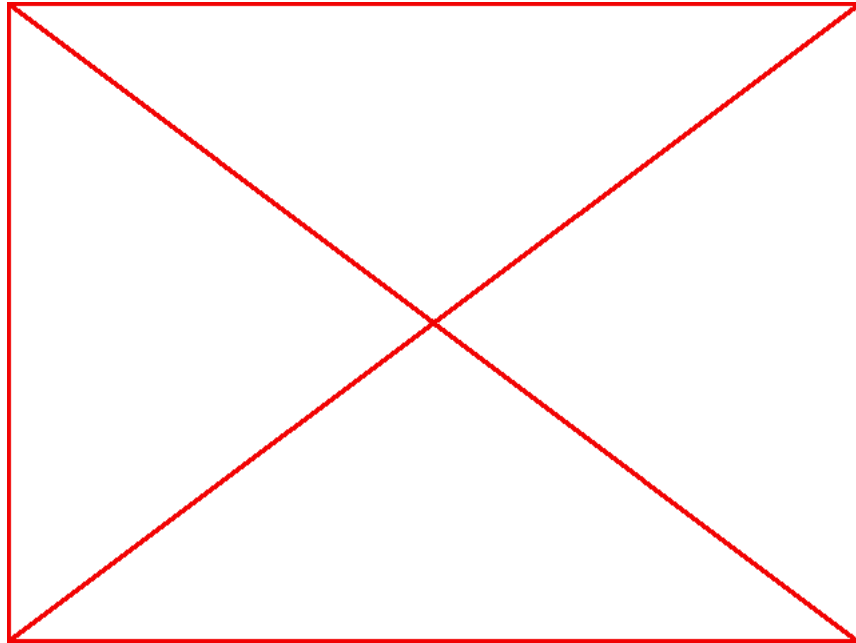




Figura 3.3 – Exemplo de função aproximada com e sem o vetor de bias



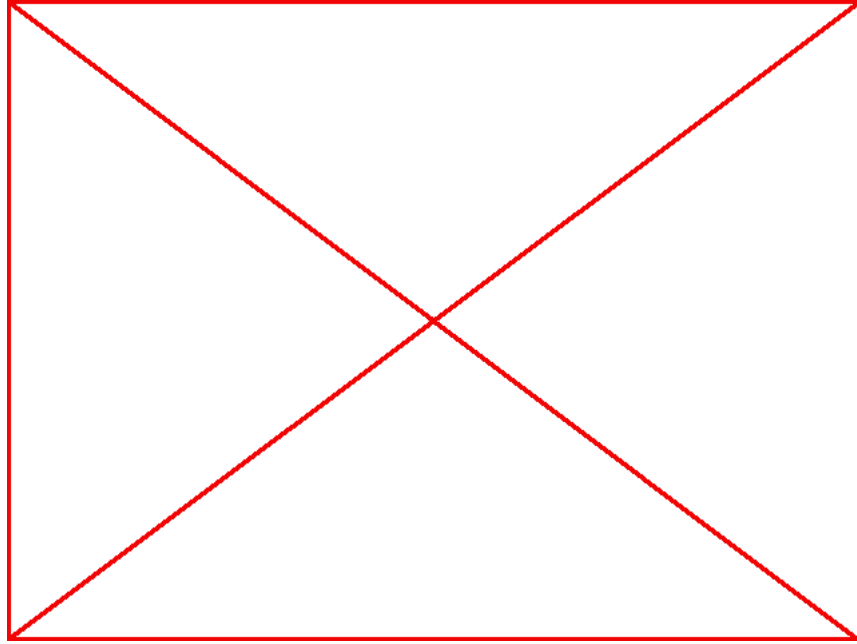
#### 3.0.1.1 Bias

Um neurônio artificial pode ser visto como um aproximador de funções. Mais adiante<TODO, ref>, será explicado que tipo de função pode ser aproximada e em qual situação, mas antes cabe aqui explicar o objetivo da entrada de *bias* ou *viés*. Todo neurônio artificial é dito ter pelo menos uma entrada extra  $x_0 = 1$  e consequentemente, o peso extra associado  $w_0$  chamado de *bias*  $b$ . O *bias* permite que a função aproximada pelo neurônio não necessite obrigatoriamente passar pela origem  $(0, 0)$ , o que reduz o erro médio da aproximação caso os dados de treinamento não tenham sofrido nenhum tipo de ajuste numérico como por exemplo a remoção do valor médio.

#### 3.0.1.2 Redes neurais

Em uma rede neural, os neurônios são organizados em *camadas*. Alguns autores não consideram a camada de entrada como uma camada distinta, mas esta abstração permite simplificar a notação e torna o processo de conectar redes neurais entre si mais intuitivo. Outra camada especial é a de saída, pois é ela que determina se a rede exercerá a função de um *classificador* ou de um *regressor*. Todas as demais camadas intermediárias são chamadas de *camadas ocultas*; o nome não possui nenhuma conotação especial e serve apenas para diferenciá-las das camadas de entrada e saída. Em uma rede *feed forward* as

Figura 3.4 – Exemplo de uma rede neural de 3 camadas



saídas de cada camada somente são computadas na direção e sentido da camada de saída, isto é, não há uma retroalimentação das conexões de saída para as de entrada. Quando isto acontece a rede é chamada de *recorrente* porém, o estudo deste tipo de rede neural não está nos objetivos deste trabalho.

Sendo assim, uma rede neural necessita ter pelo menos duas camadas — a de entrada e a de saída. A adição de camadas ou neurônios não possui um efeito óbvio e generalizável para qualquer rede, e portanto é alvo de constante experimentação e observação empírica sujeito a comportamento variável dependendo do problema.

Cada camada  $S$  é formada por  $R$  neurônios, cujos pesos das conexões podem ser representados através de uma matriz  $W$ . A notação em forma vetorial e matricial dos pesos das camadas permite que operações mais eficientes sejam utilizadas para realizar a computação da saída da rede neural e do ajuste dos mesmos.

$$W_{S,R} = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,R} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,R} \\ \vdots & \vdots & \ddots & \vdots \\ w_{S,1} & w_{S,2} & \cdots & w_{S,R} \end{bmatrix}$$

A saída ou *ativação* de uma camada  $a'$  pode ser expressa em função da ativação da camada anterior  $a$  da seguinte maneira:

$$a' = \sigma(Wa + b)$$

### 3.0.1.3 Uma intuição sobre aprendizagem

A aprendizagem em uma rede neural se dá pelo ajuste dos pesos  $w$  das conexões dos neurônios. Ao longo dos anos foram propostas várias maneiras de atualizar os pesos de uma rede neural automaticamente através de uma *regra de aprendizagem*. Estas regras foram categorizadas em três grandes grupos:

**Aprendizagem Supervisionada** As entradas são acompanhadas por um conjunto de valores corretamente associados a mesma.

**Aprendizagem Por Reforço** Semelhante ao aprendizado supervisionado, porém ao invés da resposta correta, é fornecido algo como uma nota ou uma indicação da performance da predição.

**Aprendizagem Não Supervisionada** Nenhuma informação adicional a respeito das entradas é fornecida ao algoritmo.

Considere o problema de classificar corretamente os exemplos da figura 3.5. Os pontos representados no plano dos pesos das conexões são:

$$\mathbf{x}_0 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \mathbf{x}_1 = \begin{bmatrix} 3 \\ 4 \end{bmatrix} \quad \mathbf{x}_2 = \begin{bmatrix} 5 \\ 6 \end{bmatrix}$$

O primeiro passo na formulação de um algoritmo de aprendizagem é a estipulação de um valor inicial para os pesos das conexões dos neurônios, que pode ser aleatória ou baseada em algum valor pré conhecido a respeito do problema e depende das propriedades numéricas do mesmo. Para o exemplo do *perceptron* acima considere o vetor de pesos  $w_0^T = [0.8, 0.3]$ .

É necessário também escolher uma *função limite*  $\sigma$ . Para o exemplo acima foi escolhida a função *hardlim*: i

$$\text{hardlim}(n) = \begin{cases} 0 & \text{if } n < 0 \\ 1 & \text{if } n \geq 0 \end{cases}$$

Figura 3.5 – Figura com o espaço de solução

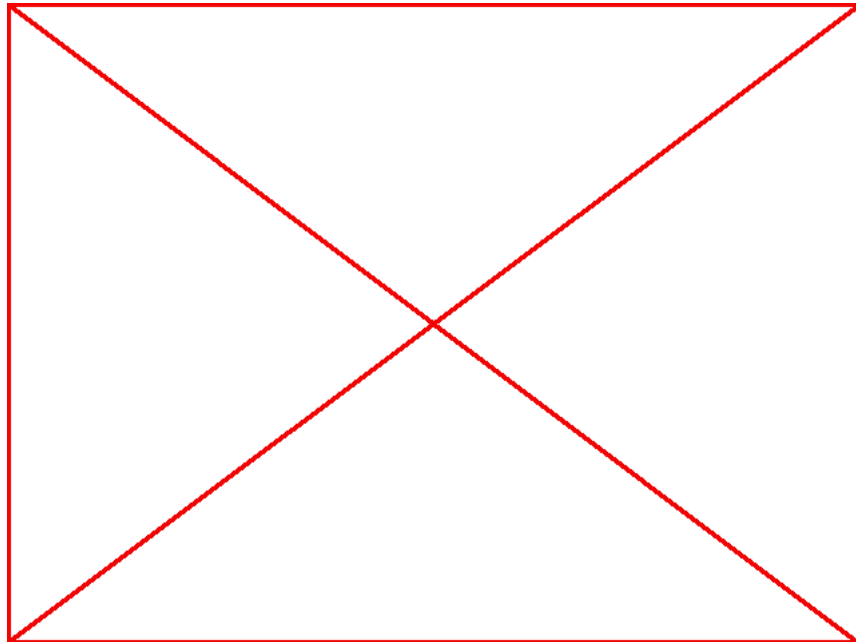


Figura 3.6 – Uma rede neural simples de um único neurônio

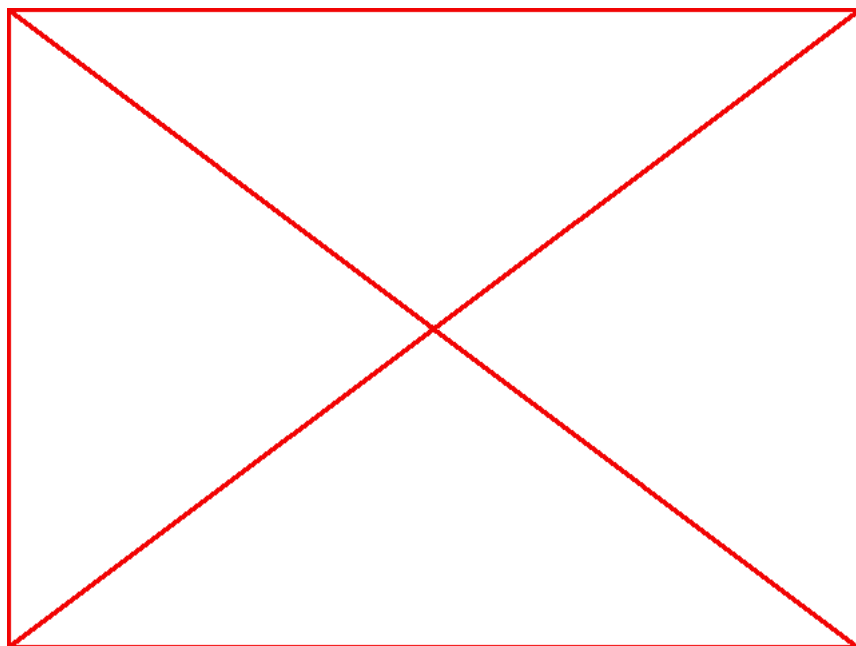
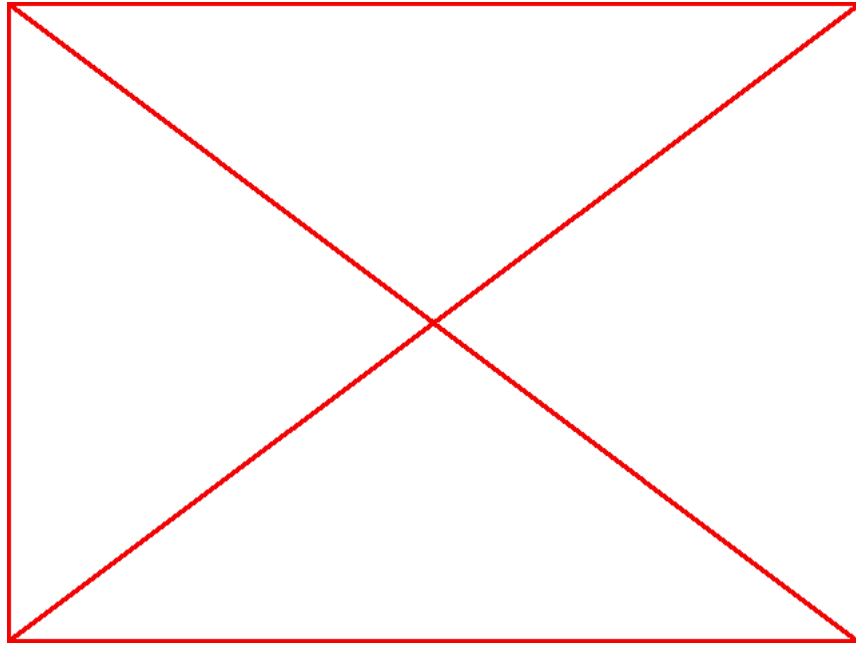


Figura 3.7 – Exemplo de funções de ativação



Na esquerda uma simples função de limite; à direita a função  $\frac{1}{1+e^{-x}}$

### 3.1 Deep Learning

#### 3.1.1 Introdução

A área de *deep structured learning* que vem sendo conhecida como *deep learning* surgiu como um novo campo de pesquisa em aprendizado de máquina nos últimos anos e as técnicas desenvolvidas nesta área já tem impactado uma gama variada de trabalhos em processamento de sinais e informações. Embora hajam diferentes definições, este trabalho adotará a definição contida em <https://github.com/lisa-lab/DeepLearningTutorials>, visto que o mesmo trabalho foi utilizado como base para este (DENG; YU, 2014)

#### Definição 1.

*Deep Learning é uma nova área do aprendizado de máquina que vem sendo introduzida com o objetivo de aproximar o Aprendizado de Máquina do seu objetivo original: Inteligência Artificial. Deep Learning aborda múltiplos níveis de representação e abstração que ajudam a processar dados tais como sons, imagens e texto.*

A motivação para o uso de uma rede neural *deep* ao invés das arquiteturas tradicionais *shallow*, tais como *Support Vector Machines* visto que esta última, a princípio consegue simular qualquer outra, é que existe um *tradeoff* entre espaço de memória e tempo de computação que pode ser explorado para acelerar o processamento das redes.

Uma maneira de entender este *tradeoff* em redes neurais é imaginar a saída das camadas intermediárias como resultados intermediários da computação de uma rede. *shallow*(??)

Ainda assim, *deep supervised learning* é uma tarefa difícil devido ao uso de *loss functions* não convexas presentes em problemas de grande importância como computação visual.

### 3.1.2 Redes Convolucionais

Redes convolucionais são um tipo especial de *deep networks* projetadas especificamente para lidar com dados vetoriais tais como imagens e sons. Embora já fosse conhecida há algum tempo na área, este tipo de rede neural ganhou fama na competição ImageNet realizada em 2012 na qual ela diminuiu pela metade o erro do melhor competidor. Este feito foi atingido com o uso de ReLUs (*rectified linear units*), GPUs e uma técnica conhecida como *dropout*. Desde então, Redes Convolucionais ganharam força e hoje é a melhor técnica para reconhecimento de imagens.(LECUN; BENGIO; HINTON, 2015)

Uma rede convolucional é baseada em 4 ideias principais: conexões locais, pesos compartilhados, *pooling* e múltiplas camadas. Conexões locais exploram o fato de que em uma imagem existe uma correlação entre os valores dos pixels vizinhos. Já pesos compartilhados geram o conceito de *feature maps* e a operação de *pooling* (geralmente o máximo local) é usada pois é invariante a posição o que dá um maior poder de abstração para a rede.(LECUN; BENGIO; HINTON, 2015)

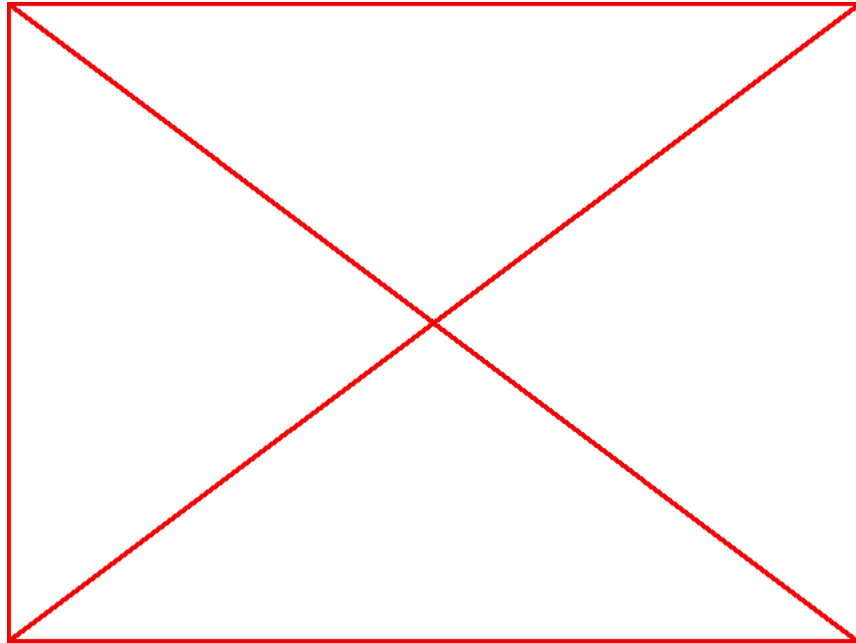
A operação de filtro é realizada pela função matemática de convolução, cujo papel é detectar conjunções locais nas *features* da camada anterior. Logo em seguida vem a camada de *pooling* cujo papel é combinar *features* semanticamente semelhantes.(LECUN; BENGIO; HINTON, 2015)

A rede é tipicamente composta de dois a três estágios de convolução seguida da aplicação de não linearidade e de *pooling*. Por fim, são aplicadas mais camadas convolucionais, desta vez completamente conectadas. O aprendizado é feito através de *backpropagation* como em redes neurais tradicionais.(LECUN; BENGIO; HINTON, 2015)

#### 3.1.2.1 Dropout

O objetivo da técnica de *dropout* é aproximar a solução ótima de calcular a média dos resultados de treinamento de todas as combinações de parâmetros possíveis. Isto

Figura 3.8 – Uma rede neural com a aplicação de dropout



é feito pela remoção aleatória e temporária de unidades da rede bem como suas conexões. Cada unidade é mantida na rede com uma probabilidade  $p$ , independente das outras.(SRIVASTAVA et al., 2014)

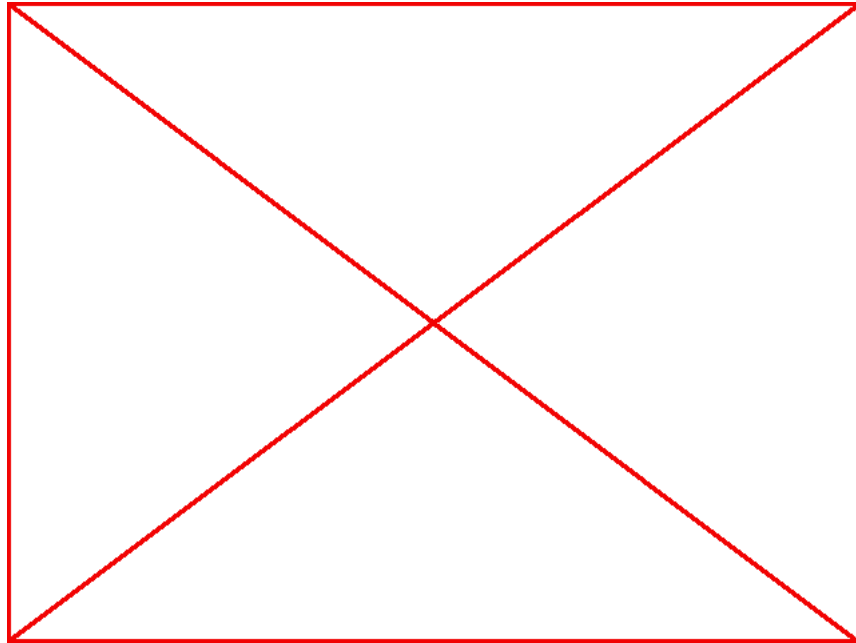
Com a aplicação de *dropout* há uma redução no *overfitting* e uma melhoria na regularização, porém, o treinamento da rede tende a demorar mais gerando um *tradeoff* entre *overfitting* e tempo de treinamento.(SRIVASTAVA et al., 2014)

Esta técnica não está restrita somente a redes *feed forward* e pode ser aplicadas a *RBM* e *autoencoders*, proporcionando uma maneira de combinar exponencialmente diferentes arquiteturas eficientemente.(SRIVASTAVA et al., 2014)

### 3.1.2.2 ReLU

Em uma rede neural tradicional, as funções  $\tanh(x)$  e  $\frac{1}{1+e^{-x}}$  são usadas para introduzir não linearidade na computação do produto dos pesos pelas entradas dos neurônios. Porém, estas funções possuem a desvantagem de apresentarem o fenômeno de saturação. Isto ocorre quando a entrada da função está distante do ponto central ( $x = 0$ ) e provoca uma perda de desempenho no aprendizado.(KRIZHEVSKY; SUTSKEVER; HINTON, 2012)

Figura 3.9 – Tipos de não linearidades aplicadas



Por causa disto, o uso de *rectifiers* tem se tornado cada vez mais comum no treinamento de grandes redes. Este *rectifier* é definido como:

$$f(x) = \max(0, x)$$

Quando uma unidade ou neurônio usa este *rectifier* ele é chamado de Rectified Linear Unit ou ReLU. De acordo com (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), ReLUs são muito mais rápidos do que *saturating non linearities* — como são conhecidas as funções mencionadas anteriormente — e não requerem regularização, embora isto ainda seja usado para aumentar o poder de generalização da rede.

### 3.1.2.3 Regularização

Em uma típica rede *deep* existem milhares de pesos que sofrem adaptação em seus valores. Normalmente um aproximador de função com este grande número de parâmetros ajustáveis está sujeito a *over fitting*. Redes neurais contornam este problema com o uso de regularização tais como:



- *Regularização L2*: Para cada peso  $w$  da rede, é adicionado uma penalidade  $\frac{1}{2}\lambda w^2$ , onde  $\lambda$  é a força de regularização. O termo  $\frac{1}{2}$  é adicionado para facilitar o cálculo da derivada na fase de *backpropagation* e o termo todo tem a interpretação de penalizar vetores de pesos com alguns poucos pesos muito altos e incentivar o uso de vetores difusos, ou seja, eles incentivam a rede a utilizar todas as suas entradas ao invés de usar demasiadamente uma entrada ou outra.
- *Regularização L1*: Para cada peso  $w$  da rede, é adicionado uma penalidade  $\lambda|w|$  com a finalidade de tornar os pesos na rede esparsos (próximos de zero). Isto confere a rede um certo grau de invariabilidade às entradas com ruído. A regularização L1 pode ser usada juntamente com a L2, sendo chamadas neste caso de *Elastic net regularization*.
- *Max norm constraints*: Consiste em setar um valor máximo para os pesos  $w$  de maneira que a rede não se torne instável mesmo na presença de uma taxa de aprendizado muito grande.
- *Dropout*: A técnica de *dropout* também pode ser vista como uma regularização e pode ser usada juntamente com outros tipos.

## REFERÊNCIAS

DENG, L.; YU, D. Deep learning: methods and applications. **Foundations and Trends in Signal Processing**, Now Publishers Inc., v. 7, n. 3–4, p. 197–387, 2014.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2012. p. 1097–1105.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015.

SRIVASTAVA, N. et al. Dropout: A simple way to prevent neural networks from overfitting. **The Journal of Machine Learning Research**, JMLR. org, v. 15, n. 1, p. 1929–1958, 2014.