Redes Neurais Artificiais: Um Survey

Leonam R. S. Miranda, Departamento de Engenharia Elétrica, Universidade Federal de Minas Gerais

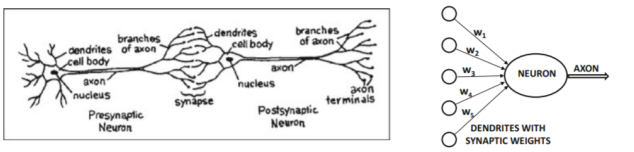
***Resumo*—Neste trabalho é apresentado um resumo histórico da evolução das Redes Neurais desde os primeiros modelos que são muito limitados em capacidades de aplicação até os atuais que podem ser aplicados em processos automáticos a tarefas que antes eram reservadas somente à inteligência humana.**

***Palavras-chave*—Redes Neurais, survey, Perceptron, MLP, Adaline, BackPropagation, Aprendizado Profundo, Viés e Variância**

I. Introdução

Os computadores usados hoje em dia podem realizar uma grande variedade de tarefas (sempre que bem definidas) em maior velocidade e com mais confiabilidade do que as alcançadas pelos seres humanos. Nenhum de nós será, por exemplo, capaz de resolver equações matemáticas complexas na velocidade de um computador pessoal. No entanto, a capacidade mental dos seres humanos é ainda maior que a das máquinas em várias tarefas.

Redes neurais artificiais (RNAs) são técnicas populares de aprendizado de máquina que simulam o mecanismo de aprendizado em organismos biológicos. O sistema nervoso humano contém células, conhecidas como *neurônios*. Os neurônios são conectados uns aos outros com o uso de *axônios* e *dendritos*, e as regiões de conexão entre os axônios e dendritos são chamadas de *sinapses*. Essas conexões são ilustradas na Figura 1 (a). As forças das conexões sinápticas geralmente mudam em resposta a estímulos externos. Essa mudança é como o aprendizado ocorre nos organismos vivos.



(a) Biological neural network (b) Artificial neural network

Figura 1. As conexões sinápticas entre neurônios. "The Brain: Understanding Neurobiology Through the Study of Addiction [1]."

Esse mecanismo biológico é simulado em redes neurais artificiais, que contêm unidades de computação conhecidas como neurônios. As unidades computacionais são conectadas entre si por meio de pesos, que têm o mesmo papel que as forças das conexões sinápticas em organismos biológicos. Cada entrada para um neurônio é dimensionada com um peso, que afeta a função calculada nessa unidade. Essa arquitetura é ilustrada na Figura 1 (b). Uma rede neural artificial calcula uma função das entradas, propagando os valores calculados dos neurônios de entrada para o (s) neurônio (s) de saída e usando os pesos como parâmetros intermediários.

Hoje em dia, a aplicação de RNAs se tornou popular em várias áreas das necessidades humanas. Muitas organizações estão investindo em redes neurais para resolver problemas em vários campos e no setor econômico que tradicionalmente estão sob a responsabilidade de pesquisa operacional [2]. O que torna a inteligência artificial única é que ela é proposta principalmente para análises de dados por acadêmicos nas áreas de ciências sociais e artes, além de sua utilidade em ciências e engenharia [3], devido às suas amplas aplicações. Por exemplo, nos últimos tempos, a inteligência artificial (IA) tem sido amplamente aplicada a questões de otimização em diversas áreas, como produção industrial e exploração de petróleo [4] e configuração de negócios [5].

II. Perspectiva Histórica

Há muitas pesquisas nos campos acadêmico e industrial que levaram ao estado atual de redes neurais e aprendizado profundo. O objetivo desta seção é fornecer um breve cronograma de pesquisa que influenciou o aprendizado profundo (Fig. 2). Schmidhuber [6] capturou de forma abrangente toda a história das redes neurais e várias pesquisas que levaram ao aprendizado profundo de hoje.

A ciência das Redes Neurais Artificiais fez sua primeira aparição significativa durante a década de 1940. Pesquisadores que tentaram emular as funções do cérebro humano desenvolveram modelos físicos (mais tarde, simulações por meio de programas) dos neurônios biológicos e suas interconexões. À medida que os neurobiologistas se aprofundavam no conhecimento do sistema neural humano, esses primeiros modelos eram considerados abordagens cada vez mais rudimentares. No entanto, alguns dos resultados obtidos nestes primeiros tempos foram impressionantes, o que encorajou futuras pesquisas e desenvolvimentos de sofisticadas e poderosas Redes Neurais Artificiais.

A. O Modelo de McCulloch e Pitts

McCulloch e Pitts publicaram os primeiros estudos sistemáticos das redes neurais artificiais ([8] [9]). Este estudo surgiu em termos de um modelo computacional da atividade nervosa das células do sistema nervoso humano.

A maior parte de seu trabalho está focada no comportamento de um neurônio simples, cujo modelo matemático ou computacional é mostrado na Figura 3. Dentro do neurônio artificial, é feita a soma de cada entrada xi multiplicada por um fator de escala (ou peso wi). As entradas emulam as excitações recebidas pelos neurônios biológicos. Os pesos representam a força da união sináptica: um peso

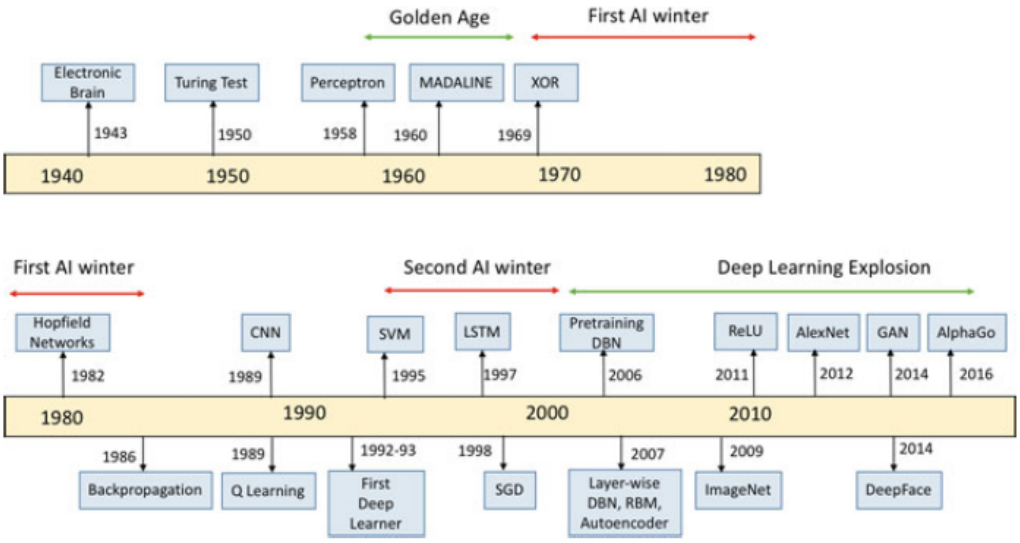


Figura 2. Destaques na pesquisa de redes neurais e aprendizado profundo[7]

positivo representa um efeito excitatório e um peso negativo um efeito inibitório. Se o resultado da soma for superior a um certo valor limite ou polarização (representado pelo peso w0), a célula ativa fornecendo um valor positivo (normalmente +1); no caso oposto, a saída apresenta um valor negativo (geralmente -1) ou zero. Portanto, é uma saída binária. Em geral, o modelo segue o comportamento neurobiológico: as células nervosas produzem respostas não lineares quando providas de uma excitação por um determinado input. Em particular, McCulloch e Pitts propuseram uma função de ativação, que representa a não linearidade do modelo, chamada função de limite rígido.

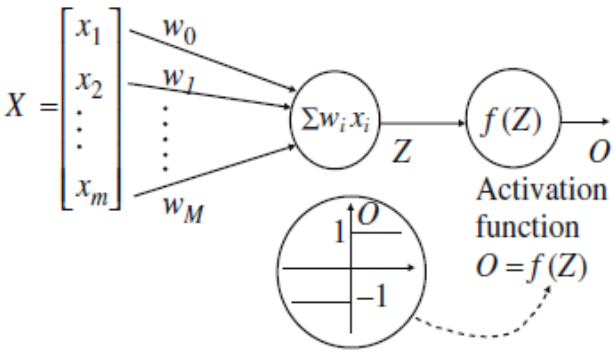


Figura 3. Modelo artificial [8] de um neurônio biológico. Como pode ser observado, a relação entre a entrada e a saída segue uma função não linear denominada função de ativação.

De acordo com a Figura 3, é fácil verificar que o modelo de McCulloch e Pitts divide o espaço de entrada em duas partes por meio do hiperplano descrito pela equação 1.

$$h(x) = \sum_{j=1}^{M} w_j x_j + w_0 = 0 \qquad (1)$$

Esse efeito pode ser observado na Figura 4 que mostra esse hiperplano para o caso particular de M = 2.

Um neurônio simples pode resolver problemas de classificação de duas classes de dados M-dimensionais, assumindo que eles são linearmente separáveis. Ou seja, ele pode atribuir uma saída igual a 1 para todos os dados da classe "A" (que caem no mesmo lado do hiperplano), enquanto atribui um valor igual a 1 ao resto dos dados que se enquadram o lado oposto.

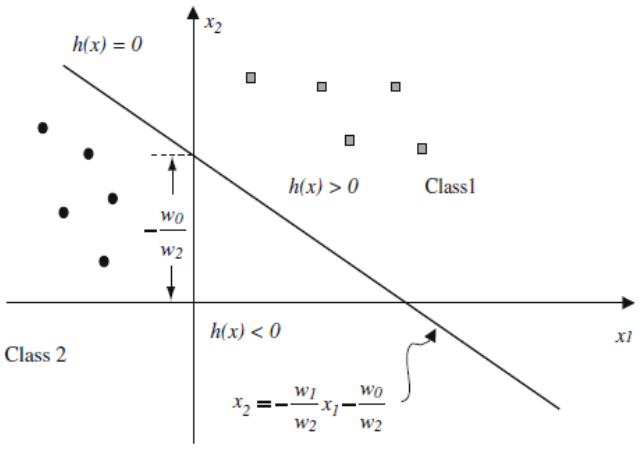


Figura 4. O hiperplano determinado pelo modelo de neurônio McCulloch e Pitts para o caso de entradas bidimensionais ([10]).

B. *Aprendizagem Hebbiana*

A partir do modelo do neurônio surge o problema de como determinar os valores adequados dos pesos que resolvem o problema em questão. Essa tarefa é chamada de aprendizado ou treinamento da rede.

Do ponto de vista histórico, o aprendizado Hebbiano [11] é o mais antigo e um dos mais estudados procedimentos

de aprendizagem. Em 1961, Hebb propôs um modelo de aprendizagem que deu origem a muitos dos sistemas de aprendizagem que existem hoje para treinar redes neurais. Hebb propôs que o valor da união sináptica aumentaria sempre que a entrada e a saída de um neurônio fossem ativadas simultaneamente. Desta forma, as conexões de rede utilizadas com mais frequência são reforçadas, emulando o fenômeno biológico do hábito e do aprendizado por meio da repetição.

Diz-se que uma rede neural usa a aprendizagem Hebbiana quando aumenta o valor de seus pesos de acordo com o produto dos níveis de excitação dos neurônios fonte e destino.

A aprendizagem hebbiana da rede é realizada por meio de iterações sucessivas utilizando apenas as informações da rede de entrada e saída, nunca utilizando a saída ou destino desejado. Por esse motivo, esse tipo de aprendizado é denominado aprendizado não supervisionado.

C. Aprendizagem supervisionado: Perceptron

Em 1959, a descoberta de células simples e células complexas que constituem o córtex visual primário pelos ganhadores do Prêmio Nobel Hubel e Wiesel [**?**] teve uma ampla influência em muitos campos, incluindo o design de redes neurais. Frank Rosenblatt estendeu o neurônio McCulloch Pitts usando o termo Mark I Perceptron, que recebia entradas, gerava saídas e tinha lógica de limiar linear [12].

Um perceptron é uma rede neural Feed Forward assim como o neurônio de Mcculloch and Pitts (onde a informação é sempre transmitida na direção da camada de entrada para a camada de saída) e tinha lógica de *thresholding* linear. Desta forma, o modelo de Mcculloch e Pitts pode ser considerado como o tipo mais simples de perceptron.

Concretamente, Rosenblatt mostrou que um perceptron de uma camada é capaz de aprender muitas funções práticas. Ele propôs uma regra de aprendizado para o perceptron, chamada regra do perceptron. Consideremos o caso mais simples de um perceptron de uma camada composto por um único neurônio, ou seja, o modelo proposto por McCulloch e Pitts. Se certos pares de entrada e saída correspondente forem conhecidos, $D_N = (x_1, d_1), (x_2, d_2), ..., (x_N, d_N)$, então, em um determinado padrão de entrada x_k do conjunto de dados de entrada, a regra perceptron atualiza os pesos da rede $\mathbf{w} = [w_0, w_1, ..., w_M]^T$ da seguinte forma:

$$w(k+1) = w(k) + \eta(d_k - o_k)x_k \tag{2}$$

O parâmetro η controla os valores de magnitude de atualização e, portanto, a velocidade de convergência do algoritmo. É chamado de *taxa de aprendizagem* e geralmente assume valores na faixa entre 0 e 1. O conjunto D_N é chamado de conjunto de treinamento, d_k é a saída desejada e o_k é a saída obtida.

Se a separabilidade linear é realizada pelo conjunto de dados de treinamento, Rosenblatt mostrou que o algoritmo sempre converge em um número finito de passos, independente do valor. Pelo contrário, se o problema não for linearmente separável, terá que ser forçado a parar, pois sempre haverá pelo menos um padrão classificado erroneamente.

É interessante notar que se a taxa de aprendizado tiver um valor próximo a 0, os pesos terão uma pequena variação a cada nova entrada, e o aprendizado é lento; se o valor for próximo a 1, pode haver grandes diferenças entre os valores de peso para uma iteração e a seguinte, reduzindo a influência das iterações anteriores e o algoritmo não pode convergir. Este problema é denominado *instabilidade*.

Também no início da década de 1960, Widrow e Hoff [13] realizaram várias demonstrações em sistemas do tipo perceptron, chamados de *ADALINE* ("Elementos LINear ADAptativos"), propondo uma regra de aprendizagem chamada de *algoritmo LMS* (algoritmo "Least Mean Square") ou algoritmo Widrow-Hoff. Essa regra minimiza a soma dos erros quadrados entre a saída desejada e a saída fornecida pelo perceptron. Ou seja, ele minimiza a função de erro:

$$E(w) = \frac{1}{2}\sum_{j=1}^{N}(d_j - z_j)^2 \tag{3}$$

Quando o gradiente para w é aplicado na Equação (4) e atualizado na direção oposta ao gradiente, a regra LMS é obtida.

$$w(k+1) = w(k) + \eta\sum_{j=1}^{N}(d_j - z_j(k))x_j \tag{4}$$

onde $x_j(k) = \mathbf{w}^T(k)x_j$. Essa versão de LMS é usualmente substituida por uma "aproximação estocástica"conforme a equação 5.

$$w(k+1) = w(k) + \eta(d_k - z_k)x_k \tag{5}$$

Ao contrário da regra do perceptron, a aplicação do LMS oferece resultados razoáveis (o melhor que pode ser alcançado por meio de um discriminador linear) quando o conjunto de treinamento não é linearmente separável.

Durante esses anos, pesquisadores de todo o mundo ficaram entusiasmados com as possibilidades de aplicação que esses sistemas prometiam.

D. Década de 1970: Declíneo

A euforia inicial despertada no início dos anos 60 foi substituída pela decepção quando Minsky e Papert [14] analisaram rigorosamente o problema e mostraram que existem severas restrições na classe de funções que um perceptron pode desempenhar. Um de seus resultados mostra como um perceptron de uma camada com duas

entradas e uma saída é incapaz de realizar uma função simples como o ou-exclusivo (**Xor**). As entradas desta função são do tipo 1 ou 1 sendo a saída 1 quando as duas entradas são diferentes e 1 se forem iguais. Na Figura 5 esse problema é ilustrado. Pode-se observar como um discriminador linear é incapaz de separar os padrões das duas classes.

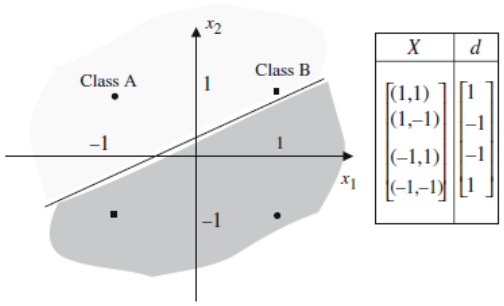


Figura 5. O prolema do ou exclusivo (Xor).

Essa limitação era bem conhecida no final dos anos 60 e também se sabia que o problema poderia ser resolvido adicionando mais camadas ao sistema. Como exemplo, vamos analisar um perceptron de duas camadas. A primeira camada pode classificar os vetores de entrada separados por hiperplanos. A segunda camada pode implementar as funções lógicas AND e OR, porque ambos os problemas são linearmente separáveis. Dessa forma, um perceptron como o mostrado na Figura 6 (a) pode implementar limites como o mostrado na Figura 6 (b) e, assim, resolver o problema Xor. No caso geral, pode ser mostrado que um perceptron de duas camadas pode implementar regiões simplesmente convexas e conexas - uma região é considerada convexa se qualquer linha reta que une dois pontos de seu limite passa apenas por pontos incluídos na região limitada por o limite. As regiões convexas são limitadas pelos (hiper) planos executados por cada nó da primeira camada, podendo ser abertas ou fechadas.

Entretando, deve-se notar que as possibilidades de Perceptrons Multi Camadas dependem das não linearidades de seus neurônios. Como a função de ativação realizada por esses neurônios é linear, a capacidade do MLP é a mesma do perceptron de camada única. Por exemplo, vamos pensar em um perceptron de duas camadas com um valor limite, $w_o = zero$ e com uma função de ativação linear, $f(z) = z$ (ver Figura 3). Nesse caso, as saídas da primeira camada podem ser facilmente expressas por meio de uma matriz $O_1 = W_1^T X$, e as da segunda camada como $O_2 = W_2^T O_1$. Então, a saída em função da entrada é obtida como:

$$O_2 = W_2^T O_1 = W_2^T W_1^T X = W_{total}^T X \tag{6}$$

Esta função pode ser realizada por um perceptron de camada única, cujos pesos de camada são W_{total}. Portanto, se os nós são elementos lineares, o desempenho da estrutura não é melhorado com a adição de novas camadas,

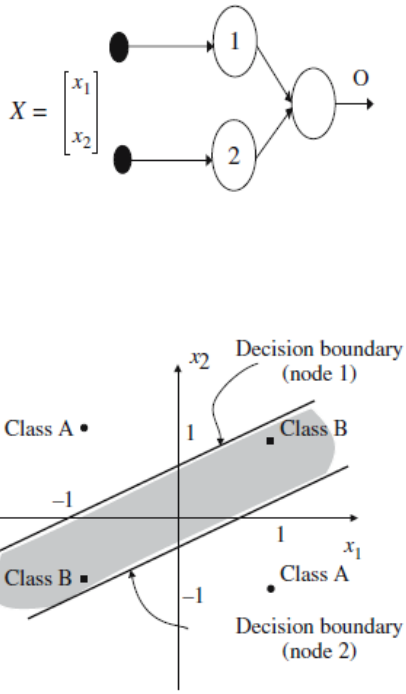


Figura 6. (a) Perceptron de duas camadas, capaz de resolver o problema Xor, implementando um limite como mostrado em (b)

visto que um perceptron de camada equivalente pode ser encontrado.

Apesar das possibilidades abertas pelo MLP, Minsky e Papert, prestigiosos cientistas de sua época, enfatizaram que algoritmos para treinar tais estruturas não eram conhecidos e mostraram seu ceticismo quanto às possibilidades de serem desenvolvidos. O livro de Minsky e Papert [14], mostrou alguns exemplos críticos das desvantagens dos redes neurais em relação aos computadores clássicos em termos de suas capacidades de armazenamento de informações, sendo um forte golpe no entusiasmo da pesquisa na área de redes neurais, eclipsando seu desenvolvimento para o próximos vinte anos, o que ficou conhecido como “The First AI Winter.”

E. Algoritmo de Backpropagation

É verdade que o perceptron de camada única tem a limitação de ser um discriminador simples. Entretanto, esse problema foi resolvido com a incorporação de não linearidades “suaves”, deriváveis e não lineares nos neurônios no lugar do *threshold* clássico. por exemplo, a função sigmoidal é muito apropriada, conforme a figura Figura 7.

George Cybenko mostrou como as redes feed-forward com neurônios finitos, uma única camada oculta e função de ativação sigmóide não linear podem aproximar a maioria das funções complexas com suposições suaves [15]. A pesquisa de Cybenko, juntamente com o trabalho de Kurt Hornik, levou ao surgimento das redes neurais e sua aplicação como ”funções aproximadoras universais”[16], [17].

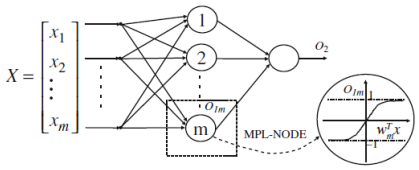


Figura 7. Perceptron multicamadas com não linearidade sigmoidal

Mas, novamente, devemos voltar à questão de como treinar os pesos da rede. Em 1986, David Rumelhart, Geoff Hinton e Ronald Williams publicaram o trabalho "Learning representations by back propagating errors", que mostrou como uma rede neural multicamadas não só poderia ser treinada de forma eficaz usando um procedimento relativamente simples, mas como camadas "ocultas"podem ser usado para superar a fraqueza dos perceptrons na aprendizagem de padrões complexos. neste trabalho foi proposto o algoritmo de retropropagação que atualiza os pesos da rede (neste caso de um MLP) na direção oposta do gradiente da função de erro que pretendemos minimizar. Para isso, a regra da cadeia é aplicada quantas vezes forem necessárias para calcular esse gradiente para todos os pesos da rede ([18], [3]). Pode-se dizer que o pessimismo suscitado pelo livro de Minsky e Papert [14] teve sua contrapartida vinte anos depois com o desenvolvimento do algoritmo de retropropagação.

F. Viés e Variância

Questões relativas ao dimensionamento da estrutura da rede e dos seus efeitos sobre a qualidade da aproximação da função geradora dos dados apareceram também em alguns trabalhos, como por meio da aplicação de técnicas de regularização ao treinamento de redes neurais [19]. Porém, apesar de as técnicas de regularização permitirem um controle da suavização da resposta do modelo e dos efeitos de overfitting, a sua formulação é baseada em uma combinação convexa de funções de erro e complexidade do modelo e, portanto, está restrita às porções convexas do espaço de soluções. Como os modelos não-lineares de redes neurais podem resultar em fronteiras de decisão não-convexas, a utilização de regularização no treinamento mostrou-se limitada, embora eficiente para alguns problemas, por não cobrir completamente todo o espaçoo de soluções.

Uma abordagem mais geral para o problema do equilíbrio entre o viés e a variância de uma família de modelos foi apresentada no artigo clássico de Geman e colaboradores [20]. Neste trabalho, os autores mostraram formalmente que o ajuste de um modelo arbitrário a um conjunto de dados e, consequentemente, da aproximação da função geradora dos dados, não depende somente da minimização de uma funçãoo associada ao erro de saída, mas também do ajuste entre complexidade do problema e capacidade do modelo. O trabalho mostrou também que a variabilidade das soluções é uma medida indireta deste ajuste e reforçou o conceito de que o problema de aprendizado a partir de um conjunto de dados deve ser tratado de maneira bi-objetiva, e não somente por meio de uma única função de custo associada ao erro da saída.

G. De RNCs a Aprendizado Profundo

De acordo com a ordem cronológica (Figura 2), nesta subseção serão apresentadas resumidamente os principais trabalhos que ainda não foram abordados neste survey.

LeCun et al. com sua pesquisa e implementação levou à primeira aplicação generalizada de redes neurais para o reconhecimento de dígitos manuscritos usados pelo Serviço Postal dos EUA [21]. Este trabalho foi um marco importante na história do aprendizado profundo, pois mostrou como as operações de convolução e compartilhamento de peso podem ser eficazes para aprender características em redes neurais convolucionais modernas (**RNC**).

Uma RNC é uma rede neural padrão que se estende pelo espaço por meio de pesos compartilhados. A RNC tem várias camadas; incluindo camada totalmente conectada, camada de pooling, camadas convolucionais e não lineares. As camadas totalmente conectadas e as camadas convolucionais possuem pesos a serem treinados, mas as camadas não lineares e o pooling não têm parâmetros. Estudos têm mostrado que a CNN tem um excelente desempenho em problemas de ML [22]. Particularmente, nas aplicações para dados de imagem, como o mais extenso conjunto de dados de classificação de imagens, processamento de linguagem natural e visão computacional.

A Figura 8 ilustra a LeNet, uma arquitetura de de rede convolucional muito popular,que foi introduzida por Yan LeCun para reconhecimento de caracteres. Esse rede inclui 5 camadas convolucionais e uma camada totalmente conectada [23].

A redução da dimensionalidade e o aprendizado usando técnicas não supervisionadas foram demonstrados no trabalho de Kohen intitulado "Self-Organized Formation of Topologically Correct Feature Maps" [24]. John Hopfield com sua Hopfield Networks criou uma das primeiras redes neurais recorrentes (**RNNs**) que serviu como um sistema de memória endereçável de conteúdo [25]. Ackley et al. em sua pesquisa mostrou como as máquinas de Boltzmann modeladas como redes neurais podem capturar distribuições de probabilidade usando os conceitos de energia de partícula e temperatura termodinâmica aplicados às redes [26]. Hinton e Zemel em seu trabalho apresentaram vários tópicos de técnicas não supervisionadas para aproximar distribuições de probabilidade usando redes neurais [27]. O trabalho de Redford Neal sobre a "belief net", semelhante às máquinas Boltzmann, mostrou como ela poderia ser usada para realizar aprendizagem não supervisionada usando algoritmos muito mais rápidos [28].

A tese de Christopher Watkins introduziu "Q Learning"e lançou as bases para a aprendizagem por reforço [29]. Dean Pomerleau em seu trabalho no NavLab da CMU mostrou como as redes neurais podem ser usadas na robótica usando técnicas supervisionadas e dados de sensores de várias fontes, como volantes [30]. A tese de

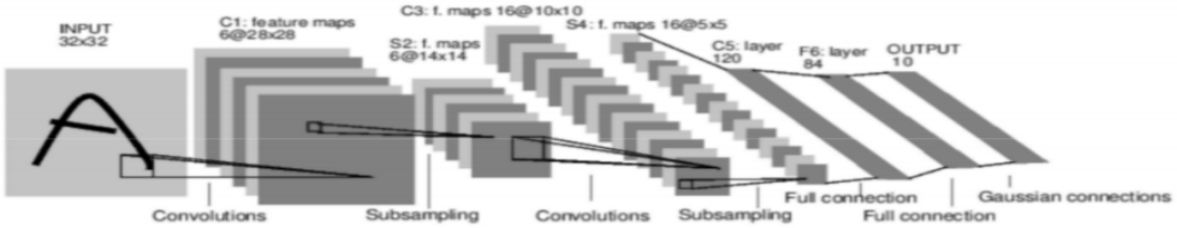


Figura 8. LeNet apresentado por Yan LeCun

Lin mostrou como os robôs podem ser ensinados de forma eficaz usando técnicas de aprendizagem por reforço [31]. Um dos marcos mais significativos na história das redes neurais é quando foi demonstrado que uma rede neural supera os humanos em uma tarefa relativamente complexa, como jogar Gamão [32]. A primeira rede de aprendizagem muito profunda que usou os conceitos de pré-treinamento não supervisionado para uma pilha de redes neurais recorrentes para resolver o problema de atribuição de crédito foi apresentada por Schmidhuber [33].

A retropropagação, que levou ao ressurgimento das redes neurais, foi logo considerada um problema devido a questões como o desaparecimento de gradientes, explosão de gradientes e a incapacidade de aprender informações de longo prazo, para citar alguns [34] e [35]. Semelhante a como as arquiteturas RNC melhoraram as redes neurais com convolução e compartilhamento de peso, a arquitetura de “memória longa de curto prazo (LSTM)” introduzida por Hochreiter e Schmidhuber superou problemas com dependências de longo prazo durante a retropropagação [36]. Ao mesmo tempo, a teoria de aprendizagem estatística e, em particular, as máquinas de vetores de suporte (SVM) estavam se tornando rapidamente um algoritmo muito popular em uma ampla variedade de problemas [37]. Essas mudanças contribuíram para o desuso de redes neurais, evento que foi nomeado de *“the Second AI Winter”*.

Aprendizagem profunda refere-se a redes neurais artificiais (RNA) com multicamadas complexas [22]. A distinção entre aprendizagem profunda e redes neurais, se dá no fato que aprendizado profundo tem formas mais complexas de conectar camadas, também tem mais contagem de neurônios do que as redes neurais para expressar modelos complexos. Entretanto é necessário mais poder de computação para serem treinadas. Além disso as redes profundas realizam a extração automática do atributos, ao lidarem com dados não estruturados.

Muitos na comunidade de aprendizado profundo normalmente atribuem ao Instituto Canadense de Pesquisa Avançada (CIFAR) um papel fundamental no avanço do que hoje conhecemos como aprendizado profundo. Hinton e col. publicou um artigo inovador em 2006 intitulado “A Fast Learning Algorithm for Deep Belief Nets”, que levou ao ressurgimento do aprendizado profundo [38]. O artigo não apenas apresentou o nome de aprendizado profundo pela primeira vez, mas mostrou a eficácia do treinamento camada por camada usando métodos não supervisionados seguidos de ”ajuste fino”supervisionado para alcançar os resultados de última geração no reconhecimento de caracteres da base de dados MNIST. Bengio et al. publicou outro trabalho em seguida, que forneceu insights sobre por que redes de aprendizagem profunda com múltiplas camadas podem aprender hierarquicamente características em comparação com redes neurais rasas ou máquinas de vetor de suporte [39]. O documento deu uma ideia de por que o pré-treinamento com métodos não supervisionados usando DBNs, RBMs e autoencoders não apenas inicializou os pesos para obter soluções ideais, mas também forneceu boas representações de dados que podem ser aprendidos. O artigo de Bengio e LeCun ”Scaling Algorithms Towards AI”reiterou as vantagens do aprendizado profundo sobre arquiteturas como CNN, RBM, DBN e técnicas como pré-treinamento / ajuste fino não supervisionado, inspirando a próxima onda de aprendizado profundo [40]. O uso de funções de ativação não lineares, como unidades lineares retificadas, superou muitos dos problemas com o algoritmo de retropropagação [41], [42]. Fei-Fei Li, chefe do laboratório de inteligência artificial da Universidade de Stanford, juntamente com outros pesquisadores lançaram o ImageNet, que coletou um grande número de imagens e mostrou a utilidade dos dados em tarefas importantes, como reconhecimento, classificação e agrupamento de objetos [43].

Ao mesmo tempo, seguindo a lei de Moore, os computadores estavam ficando mais rápidos e as unidades de processador gráfico (GPUs) superaram muitas das limitações anteriores das CPUs. Mohamed et al. mostrou uma grande melhoria no desempenho de uma tarefa complexa, como reconhecimento de fala usando técnicas de aprendizado profundo e alcançou grandes aumentos de velocidade em grandes conjuntos de dados com GPUs [44]. Usando as redes anteriores como CNNs e combinando-as com uma ativação ReLU, técnicas de regularização, como dropout, e a velocidade da GPU, Krizhevsky et al. alcançou as menores taxas de erro na tarefa de classificação ImageNet [45], ganhando a competição ILSVRC-2012 com um taxa de erro de 15,3% chamando a atenção de acadêmicos e da indústria para a aprendizagem profunda. Goodfellow et al. propôs uma rede generativa usando métodos adversários que abordou muitos problemas de aprendizagem de uma maneira não supervisionada e é considerada uma pesquisa inovadora com amplas aplicações [46].

Muitas empresas como Google, Facebook e Microsoft

começaram a substituir seus algoritmos tradicionais por aprendizado profundo usando arquiteturas baseadas em GPU para aumentar a velocidade. O DeepFace do Facebook usa redes profundas com mais de 120 milhões de parâmetros e atinge a precisão de 97,35% sore o conjunto de dados LFW, aproximando-se da precisão de nível humano ao melhorar os resultados anteriores em 27% sem precedentes [47]. O Google Brain, uma colaboração entre Andrew Ng e Jeff Dean, resultou em um aprendizado profundo e não supervisionado em grande escala de vídeos do YouTube para tarefas como a identificação de objetos usando 16.000 núcleos de CPU e cerca de um bilhão de pesos! AlphGo, da DeepMind, derrotou Lee Sedol da Coreia, um jogador Go internacionalmente classificado, destacando um marco importante na IA geral e do aprendizado profundo.

III. Conclusão

Este artigo apresentou uma revisão cronológica dos principais modelos e paradigmas de aprendizado de Redes Neurais Articiais. Assim foi abordado como surgiu a redes neurais, quais problemas foram surgindo e como estes foram solucionados, a relação entre aprendizagem e generalização na classificação da rede neural e questões para melhorar o desempenho do classificador neural. Embora existam muitos outros tópicos de pesquisa que foram investigados na literatura, acredito que esta revisão selecionada cobriu os aspectos mais importantes das redes neurais,

Os esforços de pesquisa durante as últimas décadas fizeram progressos significativos tanto no desenvolvimento teórico quanto nas aplicações práticas. As redes neurais têm demonstrado ser uma alternativa competitiva aos classificadores tradicionais para muitos problemas práticos. No entanto, embora as redes neurais tenham se mostrado muito promissoras, muitos problemas ainda permanecem sem solução ou incompletamente resolvidos. Na minha opinião, há dois grandes que devem ser superados: o empirismo utilizado na construção dos modelos de aprendizado profunda e o grande esforço computacional necessário para treinar redes profundas, tornando-as inviáveis para quem possui poucos recursos.

Referências

[1] "The brain: Understanding neurobiology through the study of addiction."

[2] M. A. Boyacioglu, Y. Kara, and Ö. K. Baykan, "Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit insurance fund (sdif) transferred banks in turkey," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3355–3366, 2009.

[3] S. Haykin and N. Network, "A comprehensive foundation," *Neural networks*, vol. 2, no. 2004, p. 41, 2004.

[4] H. Rahmanifard and T. Plaksina, "Application of artificial intelligence techniques in the petroleum industry: a review," *Artificial Intelligence Review*, vol. 52, no. 4, pp. 2295–2318, 2019.

[5] O. Araque, I. Corcuera-Platas, J. F. Sánchez-Rada, and C. A. Iglesias, "Enhancing deep learning sentiment analysis with ensemble techniques in social applications," *Expert Systems with Applications*, vol. 77, pp. 236–246, 2017.

[6] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[7] U. Kamath, J. Liu, and J. Whitaker, *Deep learning for NLP and speech recognition*, vol. 84. Springer, 2019.

[8] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[9] W. Pitts and W. S. McCulloch, "How we know universals the perception of auditory and visual forms," *The Bulletin of mathematical biophysics*, vol. 9, no. 3, pp. 127–147, 1947.

[10] D. Andina, A. Vega-Corona, J. Seijas, and J. Torres-Garcìa, "Neural networks historical review," in *Computational Intelligence*, pp. 39–65, Springer, 2007.

[11] D. O. Hebb, "The organization of behavior; a neuropsychologicigal theory," *A Wiley Book in Clinical Psychology*, vol. 62, p. 78, 1949.

[12] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological review*, vol. 65, no. 6, pp. 386–408, 1958.

[13] B. Widrow and M. E. Hoff, "Adaptive switching circuits," tech. rep., Stanford Univ Ca Stanford Electronics Labs, 1960.

[14] M. Minsky and S. Papert, "An introduction to computational geometry," *Cambridge tiass., HIT*, 1969.

[15] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of control, signals and systems*, vol. 2, no. 4, pp. 303–314, 1989.

[16] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[17] K. Hornik, M. Stinchcombe, and H. White, "Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks," *Neural networks*, vol. 3, no. 5, pp. 551–560, 1990.

[18] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[19] G. E. Hinton and S. J. Nowlan, "How learning can guide evolution," *Adaptive individuals in evolving populations: models and algorithms*, vol. 26, pp. 447–454, 1996.

[20] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural computation*, vol. 4, no. 1, pp. 1–58, 1992.

[21] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[22] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 International Conference on Engineering and Technology (ICET)*, pp. 1–6, Ieee, 2017.

[23] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[24] B. Dhingra, K. Mazaitis, and W. W. Cohen, "Quasar: Datasets for question answering by search and reading," *arXiv preprint arXiv:1707.03904*, 2017.

[25] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the national academy of sciences*, vol. 79, no. 8, pp. 2554–2558, 1982.

[26] J. A. Anderson, E. Rosenfeld, and A. Pellionisz, *Neurocomputing*, vol. 2. MIT press, 1988.

[27] G. E. Hinton and R. S. Zemel, "Autoencoders, minimum description length, and helmholtz free energy," *Advances in neural information processing systems*, vol. 6, pp. 3–10, 1994.

[28] R. M. Neal, *Bayesian learning for neural networks*, vol. 118. Springer Science & Business Media, 2012.

[29] C. J. C. H. Watkins, "Learning from delayed rewards," 1989.

[30] D. A. Pomerleau, "Advances in neural information processing systems 1," in *chapter ALVINN: an autonomous land vehicle in a neural network*, pp. 305–313, Morgan Kaufmann Publishers Inc., 1989.

[31] L.-J. Lin, "Reinforcement learning for robots using neural networks," tech. rep., Carnegie-Mellon Univ Pittsburgh PA School of Computer Science, 1993.

[32] G. Tesauro, "Temporal difference learning and td-gammon," *Communications of the ACM*, vol. 38, no. 3, pp. 58–68, 1995.

[33] J. Schmidhuber, "Learning complex, extended sequences using the principle of history compression," *Neural Computation*, vol. 4, no. 2, pp. 234–242, 1992.
[34] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.
[35] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.
[36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[37] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
[38] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.
[39] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, *et al.*, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, vol. 19, p. 153, 2007.
[40] Y. Bengio, Y. LeCun, *et al.*, "Scaling learning algorithms towards ai," *Large-scale kernel machines*, vol. 34, no. 5, pp. 1–41, 2007.
[41] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Icml*, 2010.
[42] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 315–323, JMLR Workshop and Conference Proceedings, 2011.
[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.
[44] A.-r. Mohamed, T. N. Sainath, G. Dahl, B. Ramabhadran, G. E. Hinton, and M. A. Picheny, "Deep belief networks using discriminative features for phone recognition," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5060–5063, IEEE, 2011.
[45] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.
[46] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.
[47] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Closing the gap to human-level performance in face verification. deepface," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, vol. 5, p. 6, 2014.