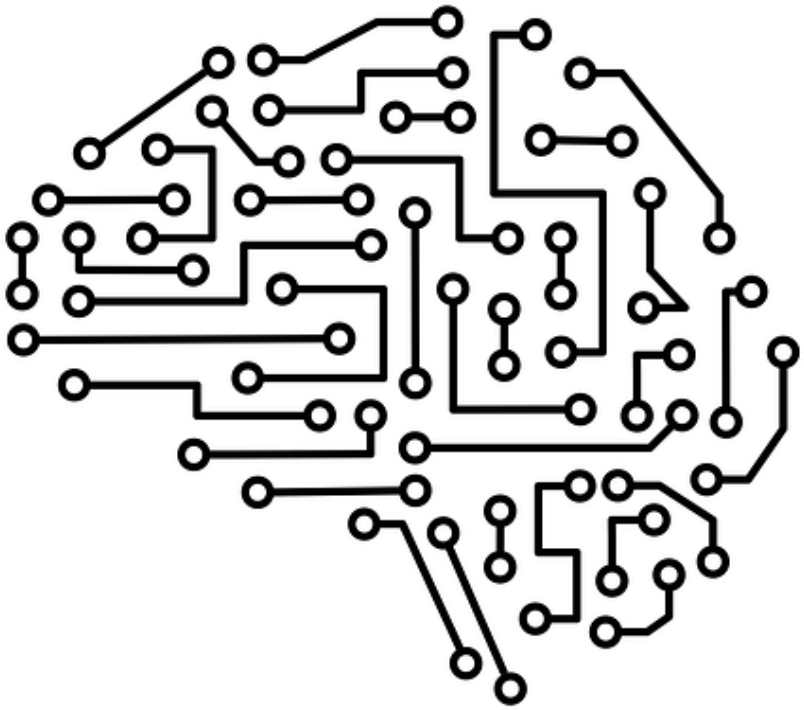
História das redes neurais artificiais

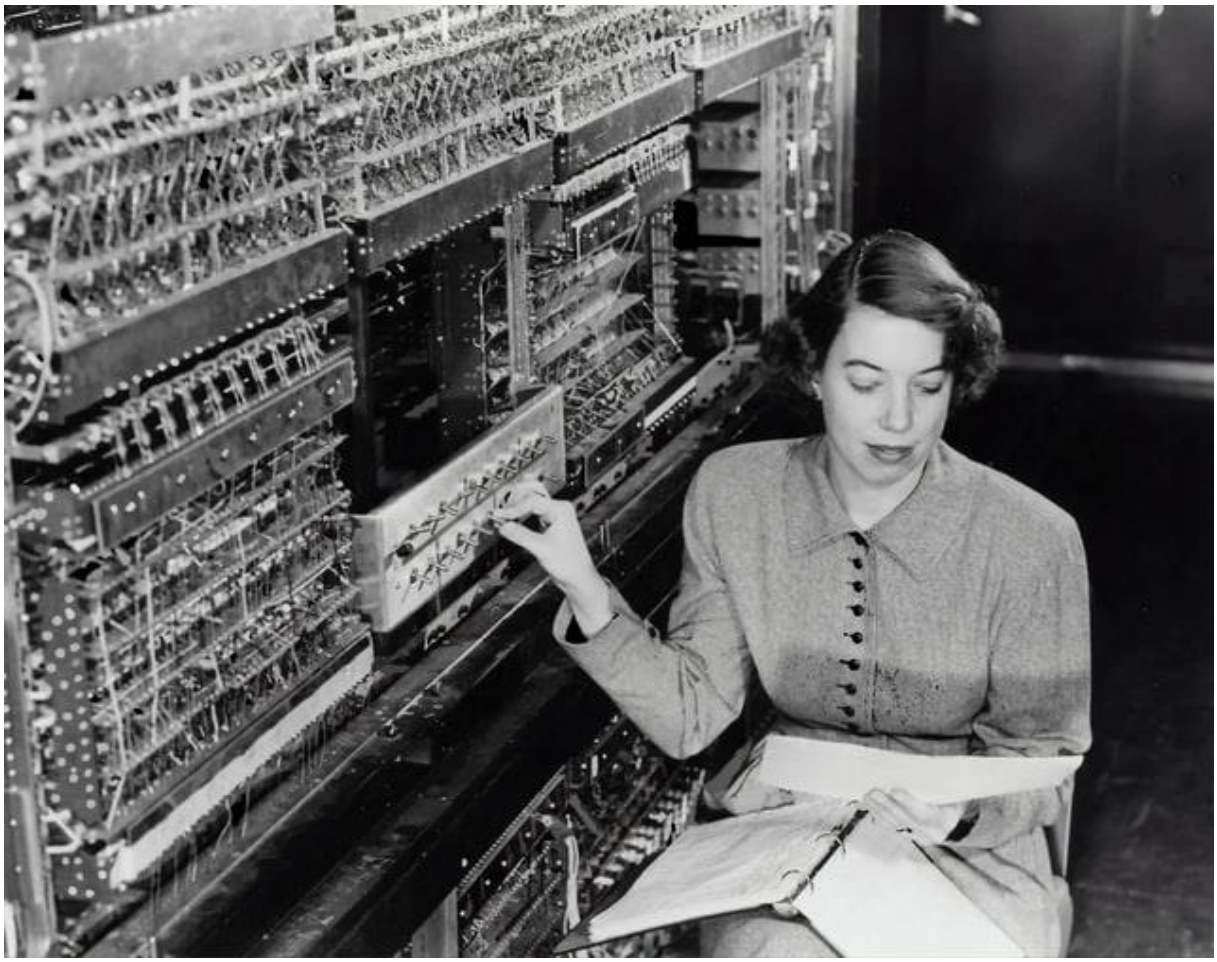


No aprendizado de **Redes Neurais**, **Deep Learning e Inteligência Artificial**, vale a pena conhecer um pouco sobre a história do surgimento e desenvolvimento dessas tecnologias.

O início da computação

É um pouco impressionante descobrirmos que o **conceito de redes neurais já é bastante antigo**. Apesar de estarmos vendo as suas aplicações, como um todo, muito recentemente aparecendo nas mídias, os primeiros estudos sobre o assunto datam do século passado.

Em 1943, o primeiro paper que tentou descrever o comportamento de um neurônio foi publicado e, na época, a sua modelagem foi feita **a partir da utilização de um circuito elétrico** – isso porque em 1943 ainda não existiam computadores. Foi um pouco depois da segunda guerra mundial – em 1950 – que os **computadores começaram a ser utilizados em um estágio muito inicial**, que é muito diferente da computação atual.



Por essa época, uma simulação de rede neural foi realizada, sem obter resultados positivos. Quem conduziu essa primeira simulação foi a IBM. **Apenas nove anos depois (1959), é que se obteve sucesso em simular um problema real utilizando redes neurais em um computador.**
Nessa época, o problema sobre o qual foi aplicado o conceito de redes neurais foi a tentativa de prever qual seria o novo bit informado na transmissão de um telefone, que informa vários bits. Ou seja, a partir de uma sequência de bits que foi enviada para um telefone, tentou-se prever qual seria o próximo bit.

A partir disso, **pelos meados de 1960, iniciou-se um hype muito grande com as possibilidades que as redes neurais e a inteligência artificial poderiam trazer para a humanidade**. Podemos imaginar que, em uma época em que a computação estava nascendo e começando a ganhar força, já existia um conceito de redes neurais e inteligência artificial. Era difícil, dessa forma, separar uma coisa da outra: o que era computação e o que era inteligência artificial? Basicamente, a computação como um todo já era encarada como uma coisa um tanto quanto mística: uma máquina conseguindo fazer coisas melhores do que o ser humano. **Na época, isso se resumia em cálculos, mas os computadores já conseguiam resolvê-los de forma mais rápida que o ser humano.**

A mídia, então, começou a expressar a ideia de que, talvez em breve, haveria robôs que conversassem com pessoas, máquinas pensantes que simulassem ser seres humanos – principalmente por causa da ideia de uma máquina poder passar no teste de Turing. Devido a isso tudo, muita empolgação surgiu em torno desse campo.

O inverno da IA

Os anos 60, entretanto, foram passando; começou-se a perceber que as aplicações reais da inteligência artificial eram limitadas: **não foi possível a obtenção de muitos resultados nem a resolução de problemas práticos**. A maior parte dos algoritmos de aprendizado que estavam sendo utilizados na época não eram nem diferenciáveis, matematicamente falando, durante toda a linha das suas funções. Isso tudo trouxe falhas para os sistemas de otimização que não performavam muito bem. Além disso, **não havia muita tecnologia de hardware que conseguisse fazer bastante computação** de maneira que fosse possível explorar todo o potencial que as redes neurais possuíam.

Por esses motivos, aos poucos, a inteligência artificial começou a cair em descrédito; começou-se a pensar que muita expectativa vazia foi criada em torno de uma coisa que, na realidade, não tinha tanto valor.

Esse foi o chamado "inverno da IA": uma época em que o ramo científico parou de fazer pesquisas nesse campo, parou de investir em startups – empresas voltadas para o assunto de IA – e muito **pouco se desenvolveu durante os anos seguintes**. Durante algumas décadas, a inteligência artificial praticamente não avançou.

Até que, em 1982, o interesse em torno da IA ganhou força novamente. Ela começou a atrair a atenção de todos após o lançamento de um paper de John Hopfield para o meio acadêmico. Isso começou a gerar um novo burburinho em torno do tema e algumas grandes conferências começaram a ser marcadas anualmente com a finalidade de se discutir sobre o assunto.

Nos anos subsequentes (de 1983 a 1985) começou-se a criar os conceitos de múltiplas camadas nas redes neurais e de redes neurais bidirecionais. Tudo isso começou a ser modelado nessa época, e **esse campo começou a ganhar uma certa notoriedade novamente, o que trouxe de volta uma certa empolgação**. No entanto, ainda não seria o suficiente para fazer o ramo decolar de vez. Houve novamente algumas limitações técnicas e a inteligência artificial não tinha tantas aplicações reais como se esperava – **mais hype do que prática**. Isso fez com que essas conferências começassem a ter cada vez menos pessoas e o interesse geral começou a diminuir lentamente.

Surgimento da GPU e renascimento da inteligência artificial

O que realmente ocasionou o **novo despertar da inteligência artificial** com a velocidade que vemos hoje, e fez com que ela começasse a voltar com força, foi a chegada do **processamento gráfico aplicado às redes neurais**, a partir de 2009. Isso foi um dos grandes catalisadores que fez com que estivéssemos no movimento em que estamos atualmente. Para entendermos melhor esse processo, é necessário lançarmos um panorama sobre a diferença entre CPU e GPU.

Todo computador tem uma **CPU que é o "cérebro" da nossa máquina**. Nesse "cérebro" são feitos todos os cálculos – tanto seriais quanto paralelos. A CPU é uma unidade de processamento universal, porque **ela faz todo tipo de funções e roda todo tipo de programa e diferentes sistemas operacionais**. Tudo isso um único processador consegue fazer. Inicialmente, apenas esse tipo de processamento era utilizado.

Entretanto, ao longo do tempo, foi possível perceber que **algumas tarefas específicas – como, por exemplo, o processamento de vídeos ou até de imagens – poderiam ser feitas de forma paralela**. Afinal, os pixels (tanto de imagens processadas, quanto de vídeos renderizados) são independentes uns dos outros. Se fosse feito um processador que pudesse executar as funções de forma massivamente paralela – pixels diferentes sendo renderizados ao mesmo tempo – o **processamento seria acelerado**.

Como as CPUs são unidades de processamento universal – que fazem todo tipo de função – não é o ideal colocá-las para realizar um tipo de tarefa muito repetitiva e específica. Talvez fosse mais interessante criar uma unidade de processamento diferente – **uma unidade de processamento gráfico (a GPU) – voltada apenas para renderização de imagens e vídeos**. Esse processamento seria diferente: ele não seria universal e não seria capaz de fazer qualquer tipo de função como uma CPU. Ele seria voltado especificamente para a renderização de vídeos de maneira que houvesse **muitos núcleos de processamento**.

Atualmente, uma CPU simples normalmente tem quatro núcleos com quatro ou oito threads. **Ou seja, o computador pode realizar de quatro a oito tarefas ao mesmo**

tempo. Essa é, mais ou menos, a quantidade de paralelismos que uma CPU comum consegue fazer. Se comprarmos CPUs com seis ou oito núcleos, perceberemos que estas máquinas têm um preço maior, pois elas são muito mais potentes; e, se, avançarmos ainda mais e formos para as casas dos dez ou doze núcleos, falando de processadores Intel, os preços crescem absurdamente e deixam de ser viáveis para o consumidor comum. Hoje em dia (2021), as CPUs mais potentes da Intel, por exemplo, possuem apenas algumas dezenas de núcleos, na faixa dos cinquenta; **dessa forma, ainda é pequena a quantidade de operações em paralelo que uma CPU consegue realizar.**

Por outro lado, **as Unidades de Processamento Gráfico (GPUs) possuem milhares de núcleos** – o que se mostra como uma grande diferença se comparadas com as CPUs. Essas unidades de processamento gráfico **realizam tarefas mais específicas** e não têm toda a autonomia de realização de tarefas genéricas de uma CPU. Por causa dessas tarefas específicas, portanto, **as GPUs conseguem realizar muito mais atividades em paralelo**. Nesse caso, a renderização de vídeos e imagens ganha muita velocidade. Por isso, quando trabalhando com imagens e vídeos a utilização de um GPU é muito mais interessante que a de uma CPU.

Se, por exemplo, nosso computador está funcionando e vai rodar um vídeo, ou um programa de edição de vídeo, nesse momento podemos utilizar a GPU ao invés da CPU para fazer algumas tarefas específicas associadas a essa renderização dos vídeos. Essa tecnologia começou a ser utilizada pelos anos 2000; e por volta de 2009, **os pesquisadores de inteligência artificial começaram o observar o processamento por GPU e aplicá-lo em redes neurais**; afinal, as redes neurais fazem muitos cálculos que podem acontecer de forma paralela. Nessas redes neurais, inúmeras ações – cálculos ou iterações – podem ser feitos de forma paralela. **A utilização de GPUs, para esse fim, resulta em ganhos muito altos de velocidade e performance**.

Escalada da inteligência artificial e surgimento de incertezas

Desde que essa ideia começou a ser implementada no mundo acadêmico, **a velocidade com que isso escalou foi enorme**; a vantagem que se tinha em usar processamento paralelo – usando a ideia de GPU – aplicada às redes neurais foi muito grande. Viu-se como seria possível explorar o **potencial de diferentes configurações de redes neurais para a utilização de um conjunto muito grande de dados**. Isso permitiu que se conhecesse o real potencial da aplicação de redes neurais em **diferentes tipos de problemas**. Antigamente, não havia tecnologia e hardware necessários para esse tipo de computação; no presente, entretanto, todas essas possibilidades já se tornaram realidade. Em outras palavras, **o processamento gráfico alavancou e permitiu a exploração do potencial das redes neurais, que antes estava limitado pela tecnologia**.

Isso fez com que rapidamente alguns fornecedores e fabricantes de hardware – como a Nvidia, por exemplo – enxergassem um mercado potencial e começassem a fabricar **hardware específicos para o desenvolvimento de deep learning**. Essas empresas viram nesse mercado a possibilidade de criar placas gráficas com um software em que fosse possível rodar redes neurais usando essa ideia de processamento paralelo. Começou-se, então, a fabricação de GPUs voltadas para isso. O lucro vindo daí foi imenso,

o que foi um catalisador para que esse ramo crescesse cada vez mais. Assim, **aplicações reais das redes neurais começaram a surgir**.

A partir de 2010, ano a ano, novas descobertas foram feitas. **Novamente, principalmente devido à mídia, o hype em tono da inteligência artificial cresceu, pois a cada ano havia novas notícias que impressionavam:** a inteligência artificial era capaz de coisas impressionantes. Isso tudo ocorreu porque, por baixo dos panos, surgiu uma tecnologia nova: o **processamento gráfico de GPU aplicado às redes neurais**. Dessa forma, novos hardwares permitiram que toda aquela teoria, que já existia desde os anos 50, começasse, agora de fato, a ser explorada a fundo.

É bastante interessante observar que hoje nós ainda estamos vivendo uma certa realidade de **incertezas quanto à inteligência artificial**. Existe, por um lado, o receio de que seja possível que, daqui a alguns anos, haja um novo inverno da inteligência artificial. Por outro lado, hoje ainda se fala que – em questão de poucas décadas, talvez **já haja uma inteligência artificial que pense de uma forma tão boa quanto o ser humano**, ou até melhor, evento esse que recebeu a denominação de **singularidade**. Alguns dizem que chegaremos em um ponto de uma nova limitação na tecnologia. Com o tempo, é possível que a inteligência artificial caia em descrédito novamente.

Para que esse novo inverno da inteligência artificial não ocorra, é muito importante que duas coisas aconteçam:

- Primeiro: **é necessária a continuidade do desenvolvimento e da pesquisa**. Ou seja, é importante que, ao longo de poucos anos, haja novas descobertas e aplicações em áreas como a saúde, logística, finanças, etc. Isso fará com que as aplicações práticas da inteligência artificial continuem se desenvolvendo e permitirá que, cada vez mais, essa tecnologia se mostre útil.
- Segundo: **é essencial que as tecnologias de hardware continuem avançando em uma taxa alta**. Dessa maneira, será possível aplicar conceitos da inteligência artificial, de fato, na prática. Sem o hardware – sem unidades de processamento cada vez mais velozes, não é possível avançar.

É possível que, de fato, a tecnologia de processamento esteja chegando em patamar que torne difícil, nos próximos dez anos por exemplo, continuar crescendo numa taxa tão rápida como a atual. Afinal, como mostra a lei de Moore (do avanço dos circuitos elétricos), o tamanho da distância entre os transistores dentro de um processador pode já estar chegando a um limite. Por isso, **talvez seja necessário que novas tecnologias – como a spintrônica e a computação quântica –, ganhem espaço nesse cenário**. Isso permitirá que, outra vez, exista um novo avanço na área da inteligência artificial, o que permitirá que essa tecnologia consiga continuar crescendo a uma taxa exponencial, como tem sido nos últimos anos.