A CAMINHO DE UMA INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL : EVOLUÇÃO DOS MODELOS DE IA E DE SUAS VISUALIZAÇÕES

TOWARDS EXPLAINABLE ARTIFICIAL INTELLIGENCE: EVOLUTION OF AI MODELS AND THEIR VISUALIZATIONS

OLIVEIRA, Fernando A.; Doutorando; Escola Superior de Desenho Industrial da UERJ
fernando.alvarus@gmail.com

MIRABEAU, Almir; Doutor; Escola Superior de Desenho Industrial da UERJ
amirabeau@esdi.uerj.br

MONAT, André Soares; Doutor; Escola Superior de Desenho Industrial da UERJ
andresmonat@yahoo.com.br

RESUMO

Lançado em novembro de 2022, o ChatGPT tornou acessível ao público uma ferramenta de inteligência artificial (IA) cuja rápida popularidade mostrou potencial de mudar a forma de agir em várias áreas do conhecimento, apesar de não ser totalmente confiável e ainda cometer erros. Para que estes sistemas sejam instruídos, é necessária grande quantidade de dados, e não tem sido possível acessar o processo completamente, o que deu origem a expressão “caixa-preta”. Assim, a introdução de uma Inteligência Artificial Explicável (IAX), onde a máquina revela estes processos, estabelece novo paradigma. Este artigo parte da crença que o design, por sua natureza epistemológica, tem papel fundamental nessa abordagem. A própria concepção de modelos de IA apoiou-se em técnicas de visualização, corroborando a ideia da aderência do design ao tema. Uma revisão da literatura, na parte final deste artigo, aponta para oportunidade da produção de conteúdo acadêmico relacionando IA e design.

Palavras-chave: inteligência artificial, inteligência artificial explicável, redes neurais

ABSTRACT

Launched in November 2022, ChatGPT made an artificial intelligence (AI) tool accessible to the public whose rapid popularity has shown the potential to change the way we act in various areas of knowledge, despite not being completely reliable and still making mistakes. These systems requires a large amount of data to be instructed, and it not possible to access the process completely, which gave rise to the expression “black-box”. Thus, the introduction of Explainable Artificial Intelligence (XAI), where the machine reveals these processes, establishes a new paradigm. This article is based on the belief that design, due to its epistemological nature, plays a fundamental role in this approach. The very design of AI models was based on visualization techniques, corroborating the design adherence to the theme. A literature review, in the final part of this article, points to the opportunity to produce academic content relating AI and design.

Keywords: artificial intelligence, explainable artificial intelligence, neural networks

1. Introdução

Este artigo é a parte inicial de uma pesquisa de doutorado que tem como objetivo propor formas de contribuição do design em inteligência artificial, notadamente na Inteligência Artificial Explicável (IAX ou XAI, em inglês). No modelo XAI, as respostas são avaliadas e validadas ou não e têm importância maior em aplicações onde a confiabilidade dos resultados é fundamental, como diagnósticos médicos e direção autônoma, por exemplo. Os modelos de IA atuais têm com característica uma "caixa-preta" (MORABITO et. al, 2024), uma parte delas em que o processo de concepção de respostas não pode ser verificado. Mas adiante neste artigo, será apresentada uma descrição mais detalhada da Inteligência Artificial Explicável.

Grande parte das publicações acadêmicas de design que versam sobre Inteligência Artificial (IA), discutem a chamada IA generativa, que gera textos ou imagens a partir de instruções por escrito (*prompts*). O foco desta pesquisa, diferentemente, é o estudo do aprendizado da máquina, de como deve-se instruí-la para garantir uma resposta confiável - com ênfase na contribuição do design neste processo.

Como parte desta investigação, o presente artigo mostra um panorama evolutivo das Redes Neurais Artificiais (ANNs, em inglês), apresentando as representações gráficas das chamadas arquiteturas das redes. Em alguns casos os autores deste artigo reformularam modelos existentes; em outros, utilizaram os diagramas originais. A intenção é apoiar a tese que representações visuais foram fundamentais para o desenvolvimento de algumas destas arquiteturas (CRAVEN e SHAVLIK, 1992).

Foram utilizadas diferentes referências como base para o histórico evolutivo das redes neurais artificiais: Zhang, J., Li, C., Yin, Y. et al. (2023); Guo, Q., Jin, S., Li, M. et al. (2020); e IBM Developer - "A neural networks deep dive" (1).

2. Modelos pioneiros

2.1 O humano como fonte de inspiração

Em 1943, os americanos Warren McCulloch, neurobiólogo, e Walter Pitts, matemático e lógico, propuseram um modelo para os neurônios humanos e para o funcionamento do cérebro de forma geral (MCCULLOCH; PITTS, 1943). No artigo *A logical calculus of the ideas immanent in nervous activity* (Cálculo lógico das ideias intrínsecas na atividade nervosa), os autores atribuem aos neurônios uma atividade do tipo "tudo ou nada", sendo assim, adaptável à ideia de um neurônio artificial, matemático. O funcionamento básico de um neurônio M-P (McCulloch-Pitts), adaptado de sua representação gráfica tradicional, pode ser visto na Figura 1.

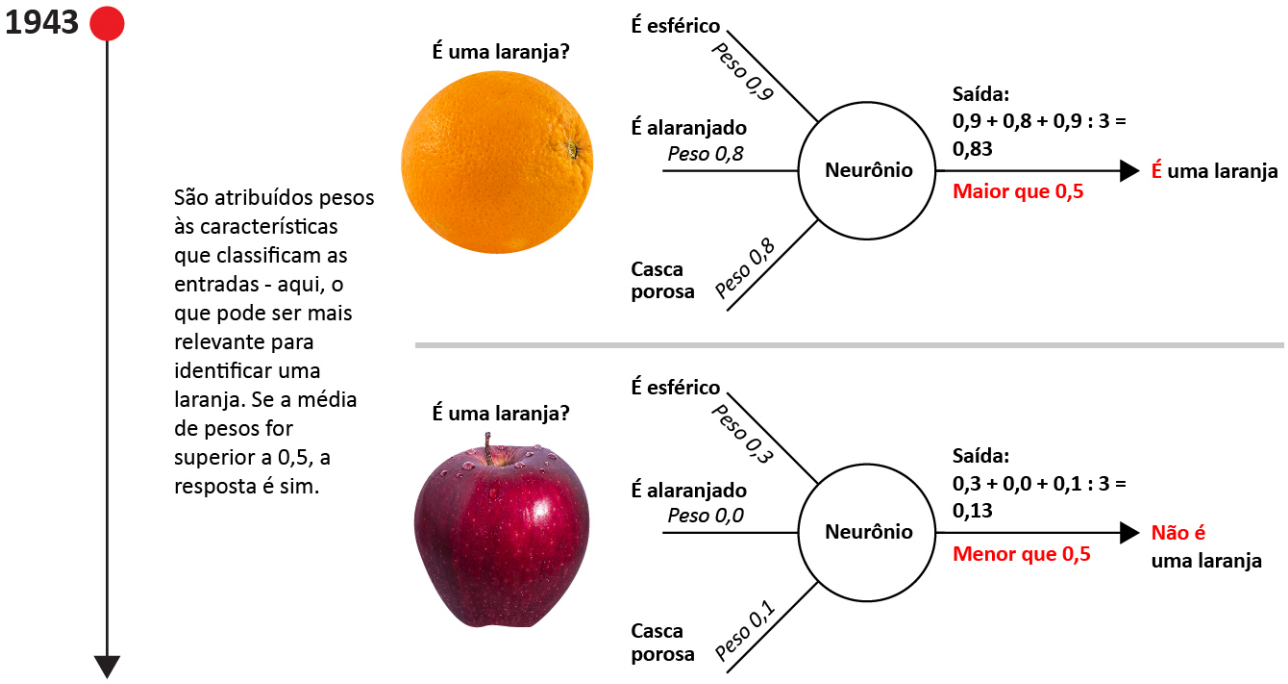
Os dois pesquisadores apresentam uma série de diagramas em que mostram a entrada e a saída de impulsos elétricos nos neurônios para explicar a passagem de informação através deles (Figura 2). A estratégia utilizada nos diagramas pode ser associada ao que Edward Tufte chama de *small multiples* (pequenos múltiplos), onde "a mesma estrutura de design é repetida para todas as imagens. O resultado é uma economia perceptiva: uma vez que o leitor decodifica e compreende o design para uma pequena parte dos dados, passa a ter familiaridade com os dados das outras partes." (TUFTE, 1990, tradução do autor).

Na figura 2 pode ser vista uma parte do conjunto de diagramas publicados no artigo. Cada forma triangular representa o corpo celular (soma) de um neurônio; as linhas, os axônios. As terminações em forma de pontos pretos são os impulsos excitatórios. No desenho (d), a terminação em anel representa um impulso inibitório. (PICCININI, 2004).

2.2 Perceptron

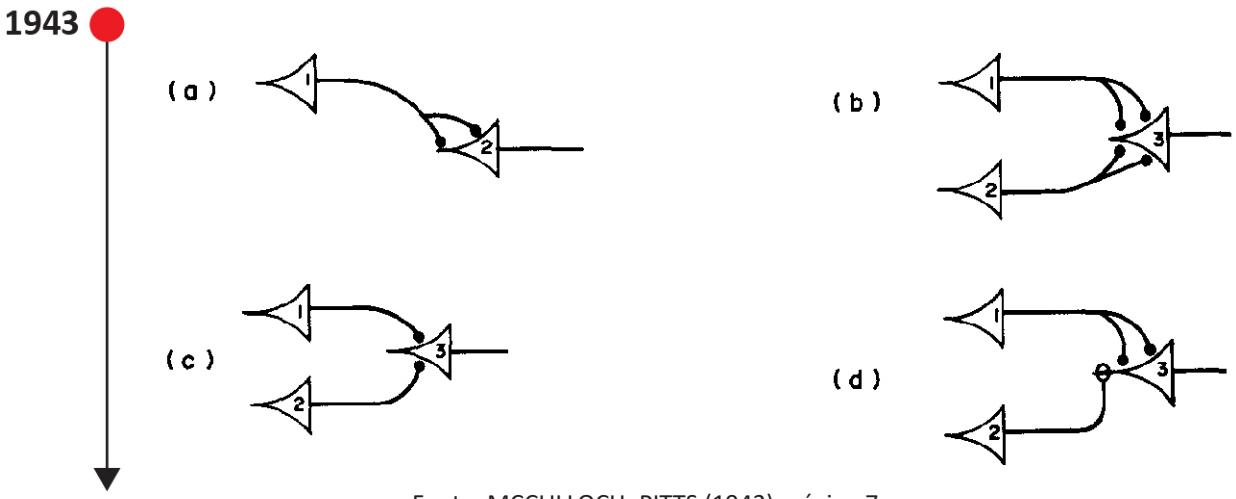
Primeiro modelo e o mais básico de uma rede neural artificial (ANN). Proposto por Frank

Figura 1: Exemplo de aplicação do neurônio M-P



Fonte: Do autor, adaptado de MCCULLOCH; PITTS (1943)

Figura 2: Diagramas representando neurônios e sinapses



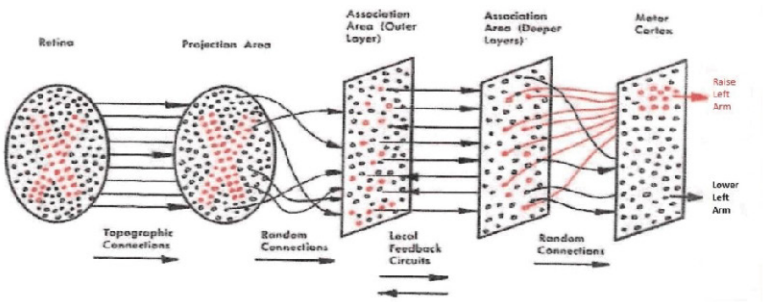
Fonte: MCCULLOCH; PITTS (1943), página 7

Rosenblatt, teve como princípio o neurônio de McCulloch-Pitts (ROSENBLATT, 1958). O princípio era o processo que ocorre no sistema ótico humano (Figura 3). Rosenblatt descreve seu modelo de perceptron propondo que “Para entender a máquina proposta - ou perceptron - é necessário primeiro entender algo sobre a natureza do cérebro e como ele funciona.” Ele argumenta que, num primeiro momento, a projeção da imagem na retina fornece uma espécie de mapa e que, daí em diante, as conexões parecem ser cada vez mais aleatórias. No entanto, a organização é restaurada quando o impulso atinge o córtex motor. Interessante notar que no desenho e no texto ele usa a expressão camada de células do cérebro, um conceito que seria determinante para a concepção das redes neurais artificiais. A arquitetura proposta por Rosenblatt para o modelo artificial é de uma área de entrada, uma camada de associação e unidades de resposta (Figura 3).

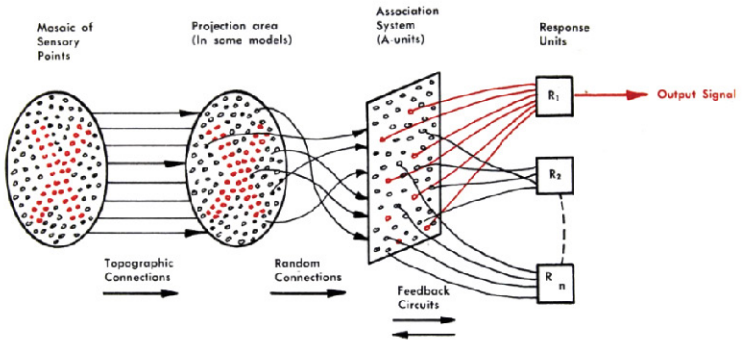
Figura 3: imagens do artigo original apresentando o Perceptron

1957

Cérebro biológico. (As áreas em vermelho indicam as células ativas, que formam a letra X)



Organização de um perceptron



Fonte: Rosenblatt, F. (1958)

3. Simbolistas, Conexionistas e o "longo inverno" das ANNs

Ainda na fase de aperfeiçoamento das redes neurais, duas ideias centrais disputavam o protagonismo na concepção de uma Inteligência Artificial: simbolismo e conexionismo.

No simbolismo, os sistemas inteligentes funcionam com forte base no aprendizado teórico, onde são fornecidas informações sobre objetos (incluindo palavras) e suas relações com outros objetos (ANDRADE, 1997). Também são conhecidos como sistemas especialistas e caracterizam-se pela lógica se-então (*if-then*). Um exemplo é o da maçã: a palavra (ou objeto) maçã, pode ser classificada por três ramos, origem, estrutura ou tipologia; pelo ramo da origem, a próxima ligação pode ser macieira, que pode ligar-se a árvore, depois a vegetais, e assim por diante. Na estrutura, as ligações podem ser, por exemplo, arredondado, vermelho, liso. Ou seja, previamente, o sistema é alimentado com objetos e suas inter-relações. Na lógica se-então, o caminho é sempre relacional - se é uma maçã, então é uma fruta; sé é uma fruta, vermelha e de casca lisa, então é uma maçã.

Ao mesmo tempo em que há mais confiabilidade nas respostas, como consequência do controle rígido das instruções, o aprendizado é mais lento e praticamente não tem autonomia, porque só é capaz de responder à problemas que estejam contemplados na fase de aprendizagem, baseado no domínio teórico usado. Como afirma Andrade (1997), "Construir um domínio teórico correto e completo é extremamente difícil e vagaroso, em algum casos é impossível (...), como por exemplo, modelar os movimentos de um robô onde as condições circunvizinhas estão constantemente mudando." No mesmo texto, a autora mostra uma outra desvantagem: "o domínio teórico pode ser computacionalmente intratável (...) necessitam de muito tempo e memória computacionais".

Na abordagem conexionista, o modelo é humano, como visto nas teorias desenvolvidas por McCulloch e Pitts (1943) e Rosenblatt (1958). Segundo estas teorias, nas redes neurais naturais cada ligação entre neurônios recebe um peso de acordo com sua relevância (como mostrado na figura 1). Para conseguir isso, o aprendizado é feito dando à máquina um grande número de

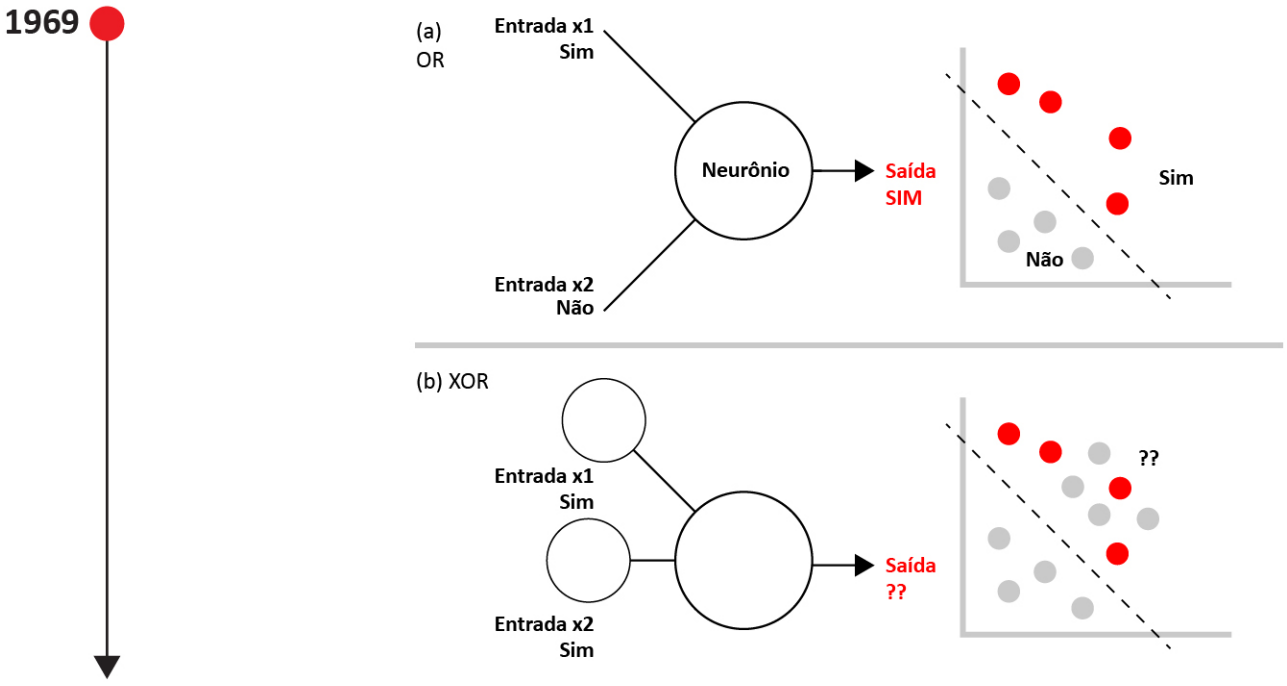
exemplos, que são pré-rotulados - várias imagens marcadas com a etiqueta carro, por exemplo. A máquina aprende a analisar estas imagens por suas características principais ou padrões; desse modo, consegue avaliar quando novas imagens de carros são apresentadas e classificá-las de acordo. Ou seja, o sistema ganharia autonomia para dar respostas, sem que fosse preciso contemplar todas as imagens de carros existentes (para uma visão melhor da distribuição das características de uma imagem, ver a figura 7).

Um dos problemas da abordagem conexionista é justamente a autonomia de classificação, porque a avaliação dos pesos acontece sem controle, numa parte das redes neurais conhecida como "caixa-preta" (*black box*), porque não mostra o processo de avaliação e posterior classificação das perguntas (*inputs*). No simbolismo, o aprendizado é controlado do início ao fim e por isso as respostas são consideradas confiáveis.

3.1 O problema ou-exclusivo (XOR)

A oposição dessas ideias atingiu seu ponto máximo com a publicação do livro Perceptrons, escrito por Marvin Minsky e Seymour Papert no final da década de 1960 (MINSKY, PAPERT, 1969). Com base na ideia de Perceptron de Frank Rosenblatt, o livro lançou uma série de dúvidas sobre o modelo das redes neurais para o aprendizado de máquina. Segundo os autores, o modelo de neurônio artificial não conseguiria resolver o problema XOR (ou-exclusivo). Neste problema, quando as entradas são diferentes, a resposta (*output*) é verdadeira ou (*OR*) falsa -, e pode ser respondida por apenas um neurônio. Quando aplicadas em um gráfico, estas saídas são divisíveis

Figura 4: o problema ou-exclusivo (XOR)

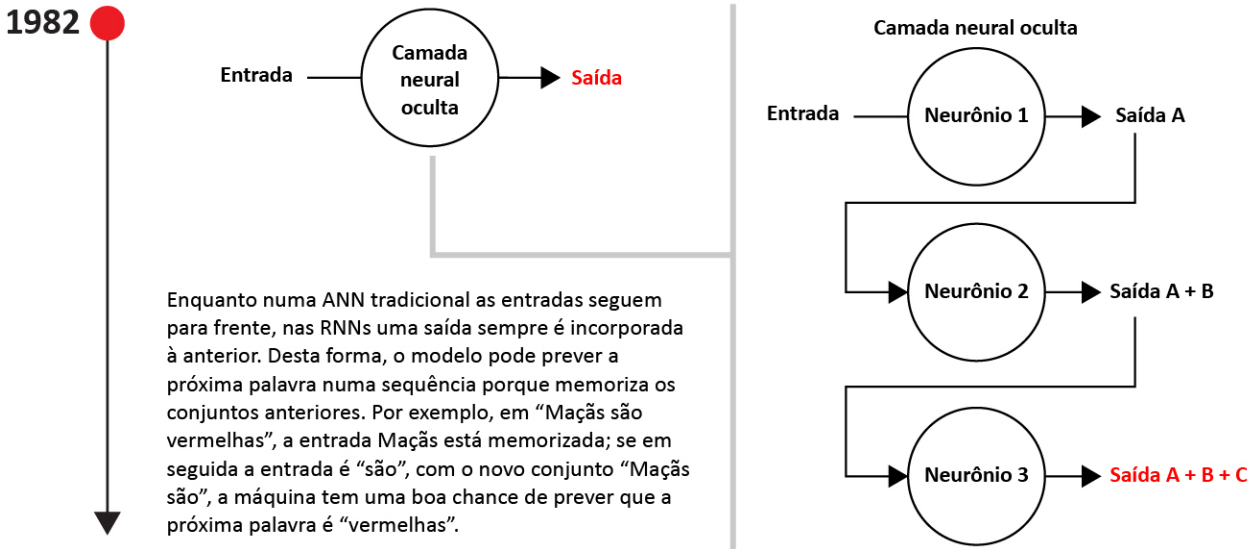


Fonte: Adaptado de (MINSKY, PAPERT, 1969)

por apenas uma linha (Figura 4a). No problema ou-exclusivo (XOR), há outras possibilidades de resposta verdadeira, caso haja entradas iguais. Nesse caso, quando uma é considerada verdadeira, a outra verdadeira é excluída. Portanto, uma única linha não pode separá-las (Figura 4b).

Atribui-se ao texto de Perceptrons, e ao debate que se seguiu, a estagnação no estudo da inteligência artificial que durou mais de 10 anos e ficou conhecido como o "inverno das ANNs":

Figura 5: Redes neurais recorrentes



Fonte: Do autor, adaptado de https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks

> Minsky e Papert afirmaram que o perceptron só poderia resolver problemas linearmente separáveis; mesmo que o número de camadas ocultas aumente, é difícil usar o perceptron porque não existe um algoritmo de aprendizagem eficaz. Esta crítica ao perceptron foi fatal, e fez com que o paradigma conexionista em IA ficasse praticamente esquecido por mais de uma década. (ZHANG; ZHU, 2023).

Apenas na década de 1980, o desenvolvimento de algoritmos capazes de lidar com mais de uma camada de perceptrons permitiu o ressurgimento nesse campo de estudo.

4. O segundo estágio da IA

4.1 Redes neurais recorrentes

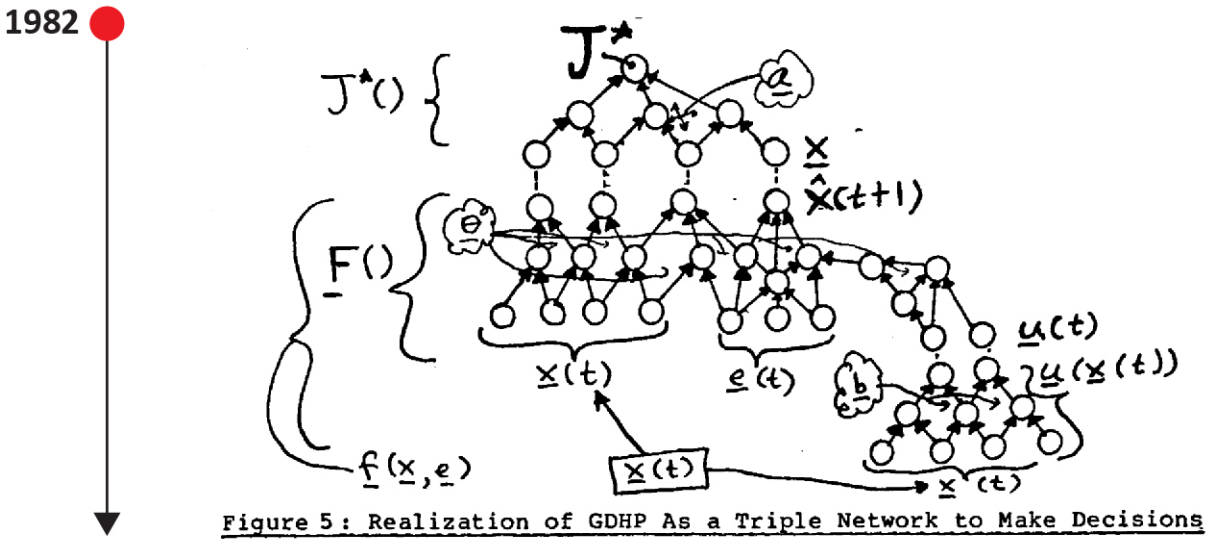
Consideradas o marco da retomada das redes neurais artificiais, as redes neurais recorrentes (RNNs) são um tipo de rede neural artificial que usam dados sequenciais ou de séries temporais. Esses algoritmos são comumente usados para problemas como tradução de idiomas, processamento de linguagem natural (NLP), reconhecimento de fala e legendagem de imagens. São incorporados a aplicativos populares, como Siri, pesquisa por voz e Google Tradutor. Nas redes neurais anteriores às RNNs, entradas e saídas eram independentes, portanto não tinham uma lógica semântica. Nas redes recorrentes, há uma dependência entre entradas e saídas anteriores (Figura 5).

Em expressões idiomáticas como "chover no molhado" ou "fazer tempestade em copo d'água", é preciso manter a sequência na ordem correta para que as frases façam sentido. Enquanto as redes neurais profundas tradicionais assumem que as entradas e saídas são independentes umas das outras, a saída das redes neurais recorrentes depende dos elementos anteriores dentro da sequência. Como resultado, as redes recorrentes precisam levar em conta a posição de cada palavra e usam essa informação para prever a próxima palavra na sequência (2).

4.2 Algoritmo *backpropagation*

Este é um modelo de aprendizado de máquina que foi capaz de resolver o problema XOR porque permite ajustes necessários para que uma rede neural com várias camadas de neurônios consiga lidar com entradas mais complexas. O cientistas social americano Paul J. Werbos teria sido

Figura 6: Imagem original do artigo de Werbos, P.J. (1982)



Fonte: WERBOS (1982)

pioneiro dessa abordagem com o artigo *Backpropagation Through Time: what it does and how to do it* (WERBOS, 1982) (Figura 6) em que o conceito de backpropagation foi associado às redes neurais artificiais. Sua popularização, no entanto, está ligada ao texto *Learning representations by back-propagating errors* (Rumelhart et al., 1989), onde os autores descrevem o *backpropagation* como "um novo procedimento de aprendizagem (...) para redes ou unidades semelhantes a neurônios. O procedimento ajusta repetidamente os pesos ou as conexões na rede de modo a minimizar a medida ou a diferença entre a saída real e a saída desejada. Como resultado dos ajustes de peso, unidades internas 'ocultas', que não fazem parte da entrada ou da saída passam a representar características importantes na tarefa." Este ajuste nos pesos citado pelos pesquisadores é feito da saída para a entrada (por isso, *backpropagation*), partindo do sinal de erro no final e percorrendo suas conexões, identificando e corrigindo os pesos que deram origem à resposta incorreta. O processo é repetido algumas vezes - de trás para frente e do início para o final -, até que se atinja ajuste próximo do ideal.

4.3 Redes Neurais Convolucionais (CNNs)

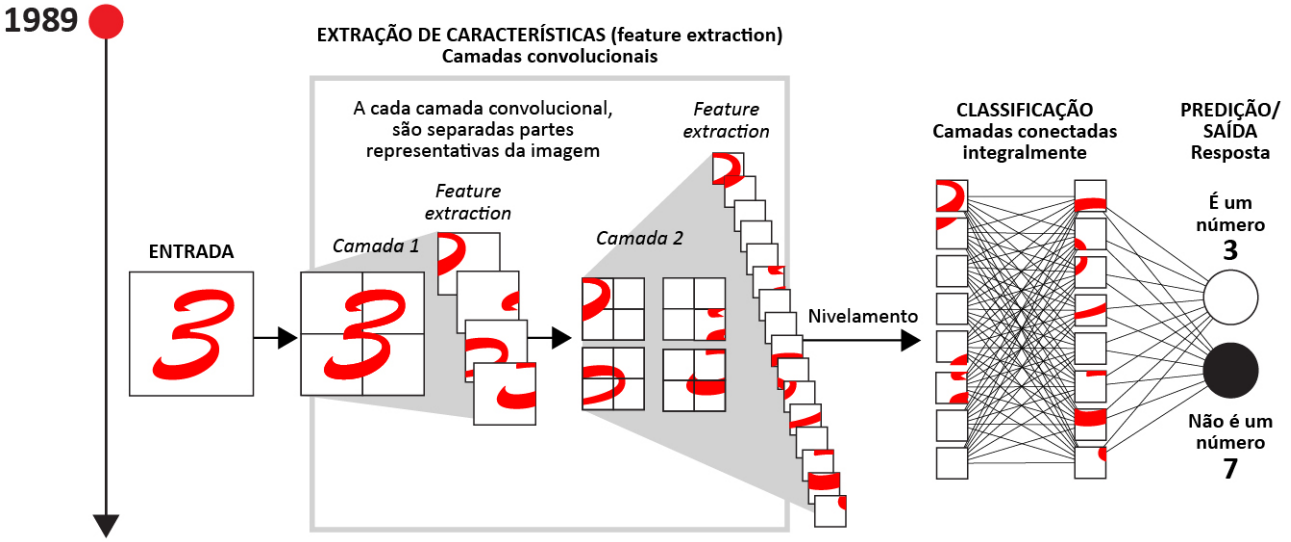
As CNNs ferecem os melhores resultados para identificação e interpretação de imagens. Utilizam três tipos diferentes de camadas neurais, a mais notável a camada convolucional, de onde são extraídas as características mais importantes da imagem (*feature extraction*) (Figura 7).

As imagens utilizadas no aprendizado são preparadas para que tenham propriedades semelhantes, como resolução e tamanho físico - 20px x 20px, por exemplo. Ou, se a forma é o aspecto mais importante do que se quer ensinar, as cores não importam, e as imagens também devem estar em escala de cinza. Em seguida é aplicado um filtro convolucional, ou *kernel*, para destacar o que se pretende enfatizar na imagem - apenas os contornos, por exemplo.

Cumpridos estes primeiros passos, nas camadas seguintes é feita a extração das características principais; a cada camada a imagem é subdividida em partes menores, cada parte contendo um trecho da imagem original. Após a passagem por um certo número de camadas convolucionais, estas partes são niveladas e distribuídas em camadas neurais tradicionais, também chamadas de *fully connected layers* (camadas integralmente conectadas). Nelas, é feita a classificação da imagem, onde são comparados as características de entrada com as que foram aprendidas pela máquina. O número de camadas depende do modelo de rede: a CNNs LeNet,

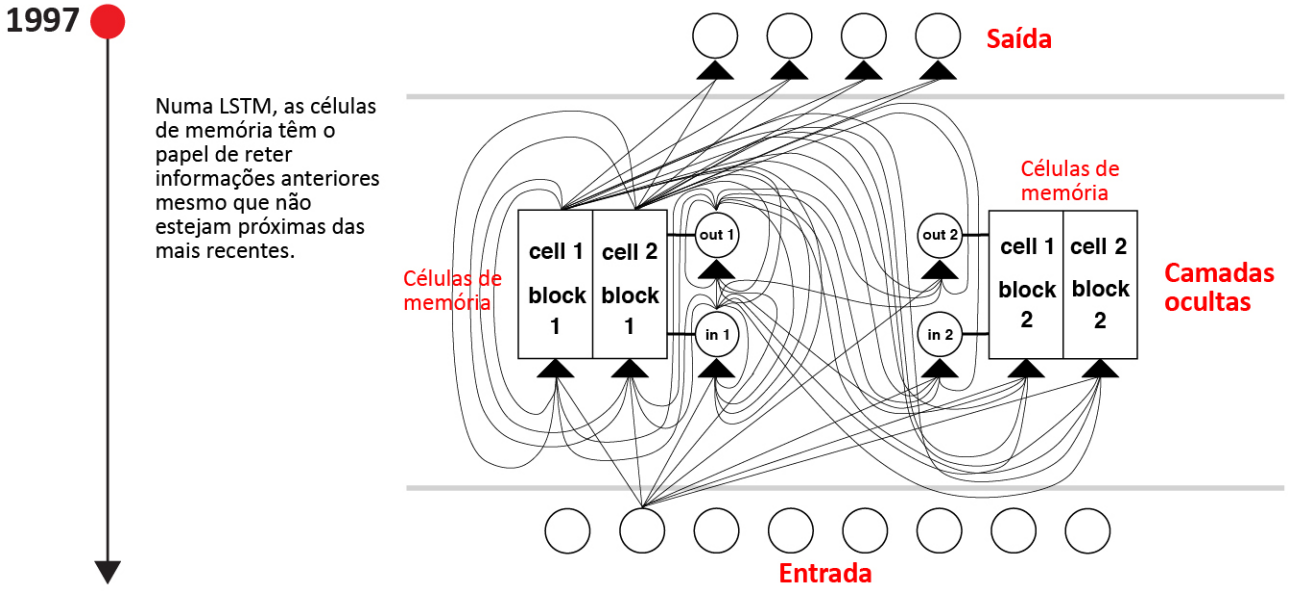
considerada pequena, tem cinco camadas; a AlexNet tem oito camadas; e a VGGnet, 32.

Figura 7: Diagrama simplificado de arquitetura de uma CNN



Fonte: Do autor, baseado em ELGENDY, 2020

Figura 8: Diagrama da arquitetura LSTM como apresentado no artigo citado



Fonte: Sobre ilustração de Hochreiter; Schmidhuber, 1997

4.4 Memória de curto prazo longa (LSTM)

Esta é uma arquitetura de RNN popular, que foi introduzida por Sepp Hochreiter e Juergen Schmidhuber (1997) como uma solução para o problema do gradiente de desaparecimento - as respostas anteriores tendem a desaparecer ao longo da rede. No artigo, o que se pretende é oderecer uma solução para dependências de longo prazo. Ou seja, se uma resposta anterior que está influenciando a seguinte não estiver próxima na rede (no passado recente), o modelo de RNN

pode não prever a resposta com precisão. Por exemplo: se a intenção é prever a segunda frase (em negrito) de "Roberto nunca aprendeu a nadar. **Ele precisa entrar no mar vestindo uma boia.**", o conhecimento da primeira frase cria um contexto que leva à conclusão da necessidade de Roberto precisar de uma boia para entrar no mar. Mas se esse contexto estivesse em algumas frases anteriores, seria difícil, ou mesmo impossível, para a RNN conectar as informações. Para lidar com isso, as LSTMs têm "células" nas camadas ocultas da rede neural, que possuem três portas: uma de entrada, uma de saída e uma de esquecimento. No esquecimento, palavras que se repetem muitas vezes, podem ser excluídas destas células, liberando espaço. Pronomes de gênero, como ele ou ela, por exemplo, repetidos diversas vezes em frases anteriores, podem ser eliminados. As três portas controlam o fluxo de informações necessárias para prever a saída na rede. (Figura 8)

4.5 Um marco importante: a ImageNet

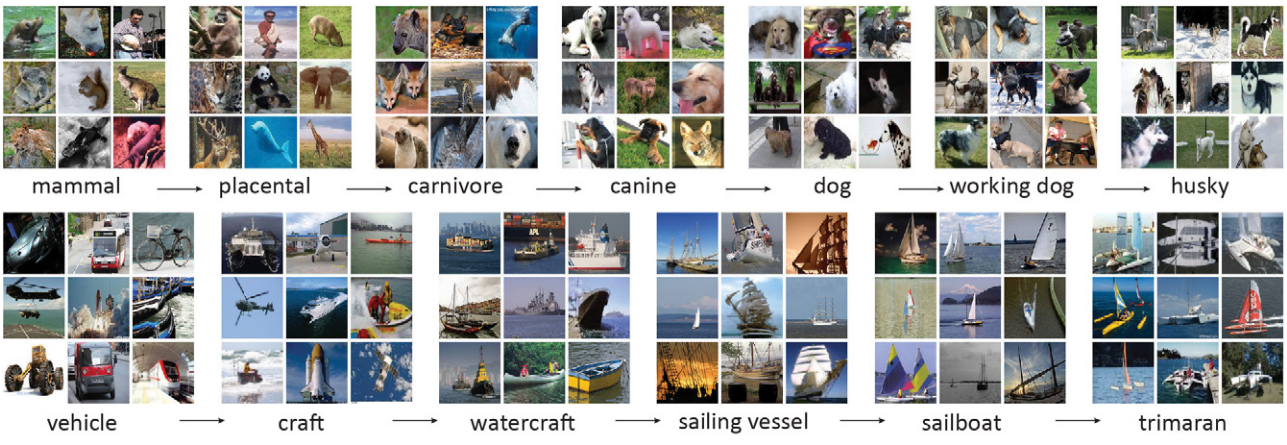
O projeto ImageNet é uma base de dados visuais construída para auxiliar nas pesquisas de software para reconhecimento de imagens por computadores. A organização destes dados segue a hierarquia usada em outra base, WordNet, que trabalha com palavras, organizadas em grupos de sinônimos cognitivos, chamados *synsets*. Os *synsets* são interligados de acordo com sua semântica e relações léxicas - carro/automóvel, fechar/cerrar, por exemplo.

A ImageNet tem cerca de 14 milhões de imagens armazenadas, que podem ser acessadas e usadas em pesquisas e no aprendizado de máquina para interpretação e classificação de imagens. Todas as imagens são anotadas, isto é, recebem uma etiqueta com um nome/*synset* correspondente. Para cada synset, são atribuídas por volta de 1.000 imagens (5). As imagens são organizadas em subgrupos (*subsets*) temáticos, como mamíferos e veículos. Exemplos de *subsets* podem ser vistos na Figura 9.

Desde 2010, o projeto promoveu uma competição onde pesquisadores apresentavam programas para melhorar a acuidade da interpretação de imagens usando recursos da ImageNet (*ImageNet Large Scale Visual Recognition Challenge* - ILSVRC). Várias redes neurais em uso hoje são egressas dessa competição (AlexNet, por exemplo). A partir de 2018, a competição passou a ser conduzida pela Kaggle (Google LLC) - uma plataforma que reúne pesquisas e recursos para aprendizado de máquina.

Seguindo o pioneirismo do ImageNet, outros repositórios foram estabelecidos, como o LabelMe (figura 10). Ali, uma grande base de imagens está disponível para anotação de usuários, de forma aberta e gratuita. Ao abrir uma imagem, o usuário marca uma parte dela, contornando-a livremente ou com polígonos e indica uma característica (ou anotação), como um carro, uma

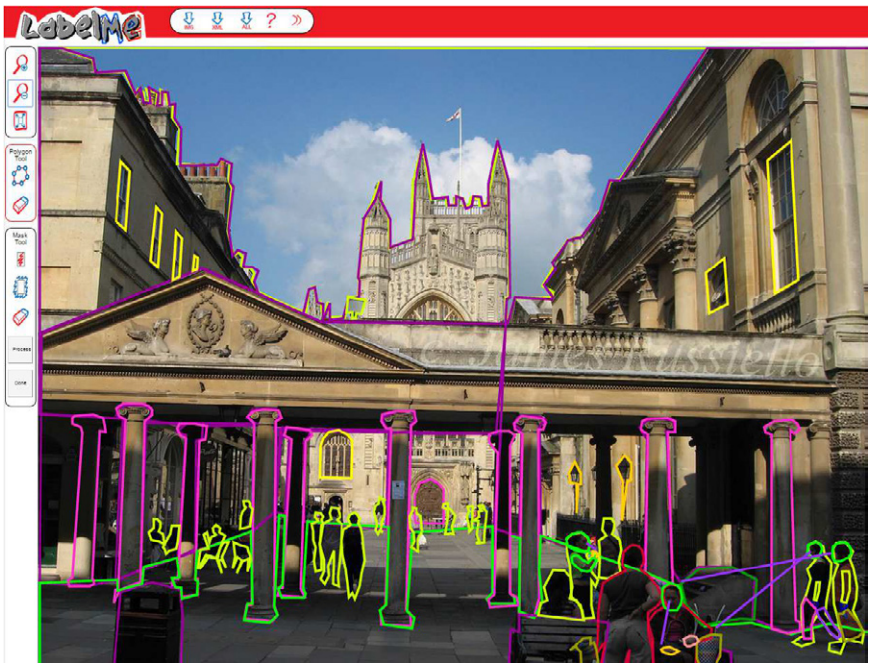
Figura 9: Exemplo de coleções de imagens da ImageNet usadas em vários subgrupos



Fonte: Deng et al., 2009

pessoa, ou um cachorro naquela cena. O site tem sido usado para pesquisas em visão de computadores relacionadas a classificação e reconhecimento de cenas (CHEN et al. 2016).

Figura 10: Site LabelMe com imagem anotada por usuário.



Fonte: Chen et al., 2016

4.6 Modelo Transformer

Um dos modelo mais recentes. Publicado por uma equipe de pesquisadores do Google (*Attention is all You need*, VASWANI et al., 2017, republicado em 2023), está impulsionando uma onda de avanços em ML, que alguns chamam de AI de transformers. Tem como base um LLM (*Large Language Model*, ou Modelo de Linguagem Grande), um tipo de modelo de inteligência artificial criado para entender e gerar texto (3). O chamado aprendizado profundo (DL), aplicado no Transformer, tem muitas camadas de neurônios. Este novo modelo pode ser considerado uma evolução dos modelos de Redes Neurais Recorrentes (RNN), e das sequentes Memórias Longas de Curta Duração (LSTM), que são limitadas quando aplicadas a sequências muito longas.

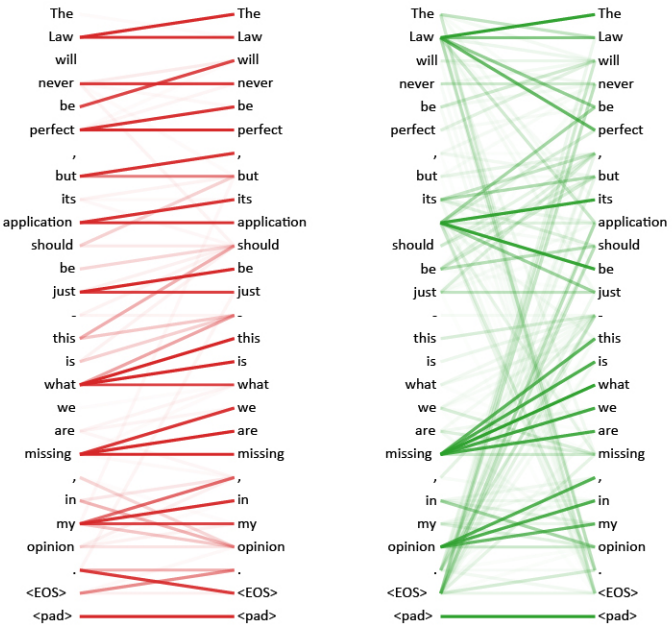
O conjunto de técnicas matemáticas usado é chamado de atenção ou auto-atenção, como no artigo citado *Atenção é tudo o que você precisa* (VASWANI et al., 2017). O modelo Transformer é capaz de entender a sequência lógica de uma frase, mas sem que as palavras estejam próximas para completar as expressões. Nos mecanismos de auto-atenção, a rede neural olha para todos os elementos fornecidos na entrada, não apenas para o mais recente, e os dados também não precisam estar em ordem (VASWANI et al., 2017).

Além disso, as relações são estabelecidas, ou calculadas, uma única vez para toda a sequência, “o que produz resultados idênticos para sequências que possuem os mesmos elementos, ainda que estes estejam dispostos em uma ordem diferente. É possível dar ao mecanismo de atenção um maior poder de discriminação quando são combinadas múltiplas cabeças de auto-atenção, dividindo a sequência em um número fixo de pedaços e aplicando a auto-atenção a cada um destes” (VASWANI, 2017, apud DINIZ, 2023). (Figura 12). As múltiplas cabeças descritas no texto referem-se a uma arquitetura em que a mesma sequência de palavras

Figura 12: Diagrama original apresentado no artigo das redes Transformers

2017

No mecanismo de atenção usado nas redes Transformers, cada palavra (token), recebe um peso no treinamento de acordo com o contexto, e não com a posição numa frase ou sentença. Assim, as relações entre palavras têm mais peso, como por exemplo a palavra “rei”, terá mais relação com “rainha”, “realeza”, ou ”trono”, que com a palavra “cachorro”. Ao lado, as linhas mais fortes nos dois conjuntos referem-se à relações mais importantes contextualmente entre as palavras de cada coluna.



Fonte: VASWANI et. al, 2023

é analisada simultaneamente por matrizes diferentes, operando paralelamente (paralelização); ao final deste processo, os resultados associados são concatenados e é atribuído um peso à resposta (equivalente à sua importância).

Uma vantagem das redes Transformer é justamente o paralelismo, que permite um desempenho computacional mais eficiente. Operações paralelas podem ser conduzidas por unidades de processamento gráfico (GPUs, na sigla em inglês), partes internas dos computadores projetadas originalmente para processamento de imagens e vídeos. As GPUs são dotadas de pequenos núcleos, que as tornam perfeitas para a operações paralelas como as exigidas na aplicação de uma rede Transformer.

O ChatGPT (acrônimo para *Chat Generative Pre-trained Transformer*) usa essa técnica.

5. Inteligência Artificial Explicável

Com o desenvolvimento dos modelos descritos neste artigo foi possível alcançar o estado da arte em IA representado pelo *chatbot* ChatGPT, que disseminou o uso da AI generativa entre usuários comuns. Lançado em novembro de 2022, tornou-se “o aplicativo de consumo de crescimento mais rápido da história, conquistando mais de 100 milhões de utilizadores” (4). Segundo a mesma reportagem da agência britânica Reuters, uma média de 13 milhões de usuários únicos acessaram a ferramenta em janeiro de 2023.

Apesar deste sucesso, algumas das respostas fornecidas pelo ChatGPT são incorretas ou falseiam fontes de informações, as chamadas alucinações. Artigo publicado na revista Nature avalia “um tipo particular de alucinação: citações bibliográficas fabricadas que não representam trabalhos acadêmicos reais. Utilizamos ChatGPT-3.5 e ChatGPT-4 para produzir pequenas revisões de literatura (...) Em seguida, pesquisamos vários bancos de dados e sites (...)”. Na versão 3.5 do ChatGPT, gratuita, os autores encotraram 55% de citações fabricadas; no ChatGPT 4.0, pago, foram 18%. Das citações reais, não fabricadas, mas que tinham erros de citação importantes, 43% estavam no ChatGPT 3.5 e 24% no ChatGPT 4.0. Eles concluem que “Embora o GPT-4 seja uma grande melhoria em relação ao GPT-3.5, os problemas permanecem.” (WALTERS, WILDER, 2023).

Também em ferramentas generativas de imagens, como as incluídas nos aplicativos Adobe, no aplicativo LimeWire, e no gerador de imagens Dall-e (do ChatGPT), alguns erros mostram possíveis falhas no treinamento e, consequentemente, nos outputs (respostas). Em teste feito para este artigo, propôs-se um *prompt* (instrução por escrito do que deve ser gerado pela IA) pedindo uma imagem de uma prancha de surfe com três quilhas. O *prompt* foi escrito em inglês, no entendimento que a base de aprendizado mais completa é feita neste idioma (que é o do país da empresa desenvolvedora OpenAI) - *a yellow three fin surfboard over a white background*. (*Prompts* feitos em 11 de maio de 2024).

Nenhuma das ferramentas apresentou uma imagem correta (Figura 13). Poderia-se argumentar que uma pessoa não familiarizada com o esporte, confrontada com essa questão, incorreria no mesmo problema. A diferença estaria na possibilidade de uma procura rápida, por pessoas, de imagens em buscadores de conteúdo, como Bing ou Google. No caso dos aplicativos citados para o teste, o conhecimento se dá a partir de um conjunto de instruções prévias, as IAs não foram capazes de fazer busca por conteúdo e, portanto, não foram capazes de dar respostas de fora do escopo do aprendizado. E, no caso das respostas em texto, quando perguntadas

Figura 13: Comparação de resultados em ferramentas de IA generativa

Fonte: Do autor

diretamente - "você é capaz de procurar respostas na internet para perguntas para as quais não foi treinado originalmente?" -, as ferramentas ChatGPT e Gemini afirmaram o contrário: "Sim, o ChatGPT pode procurar respostas na internet para perguntas para as quais não foi treinado originalmente usando uma ferramenta de navegação na web" e "Sim, eu, Gemini, sou capaz de procurar respostas na internet para perguntas para as quais não fui treinado originalmente.". No entanto, para uma pergunta sobre assunto atual, "Qual foi o resultado das eleições na França em 2024?", as respostas foram "Desculpe, mas não tenho informações sobre os resultados dessa eleição" (ChatGPT) e "No momento, não posso ajudar com respostas sobre eleições e figuras políticas. Sou treinado para ser o mais preciso possível, mas às vezes posso cometer erros. Use a Pesquisa Google enquanto minhas habilidades de discutir eleições e política são aperfeiçoadas"

(Gemini). (Os *prompts* acima foram feitos em 11 de julho de 2024).

Tendo em conta a preocupação com erros demonstrados, mais graves nas aplicações em que a confiabilidade é determinante, a Agência de Projetos de Pesquisa Avançada de Defesa (Darpa, em inglês), do Departamento de Defesa dos EUA, lançou o programa que popularizou o termo Inteligência Artificial Explicável (IAX ou XAI, em inglês). Proposto em 2015, com uma série de avaliações sobre o assunto, foi iniciado efetivamente em 2017. Um grupo selecionado de 11 equipes de pesquisadores dedicou-se a estudos relacionados a XAI. O resultado foi reunido em uma série de publicações reunidas em 2021 para avaliar os resultados do programa até aquele momento (Applied AI Letters, Special Issue: DARPA's Explainable Artificial Intelligence (XAI) Program, 2021). No texto da retrospectiva do programa, um dos autores afirma que

> O sucesso dramático no aprendizado de máquina criou uma explosão de novos recursos de IA. Avanços contínuos prometem produzir sistemas autônomos que percebem, aprendem, decidem e agem por conta própria. Esses sistemas oferecem enormes benefícios, mas a sua eficácia será limitada pela incapacidade da máquina de explicar as suas decisões e ações aos utilizadores humanos.
> Esta questão é especialmente importante para o Departamento de Defesa dos Estados Unidos (DoD), que enfrenta desafios que exigem o desenvolvimento de sistemas mais inteligentes, autônomos e confiáveis. XAI [IAX] será essencial para que os usuários entendam, confiem adequadamente e possam gerir eficazmente esta geração emergente de parceiros artificialmente inteligentes. (Gunning et al. 2021)

Para o que esta pesquisa pretende, o entendimento e o aprofundamento da investigação em IAX se relaciona com a ideia que o design é uma área de conhecimento capacitada - como será sugerido adiante - para aprimorar o aprendizado de máquina. E que pode contribuir para a implantação de sistemas capazes de fornecerem repostas justificadas e seguras.

7. Bibliometria: primeiras buscas para uma revisão da literatura

Vale enfatizar que o enfoque das investigações que servirão de guia a essa pesquisa é o aprendizado de máquina. Realçar esta diferença é importante porque é mais comum encontrar pesquisas relacionadas a design quando a abordagem é da inteligência artificial generativa, aquela em que a máquina produz resultados a partir de demandas de usuários. Com o conhecimento dos erros produzidos por instrumentos de IA generativa, como o ChatGPT, apontado previamente, estes autores acreditam que a atuação de designers no aprendizado das máquinas pode contribuir para a minimização destes erros, ou de sua eliminação.

Apesar de acreditar neste papel do designer, as primeiras buscas em bases de dados demonstram que artigos publicados por designers ou relacionados a design encontram poucos resultados. No intuito de confirmar esta primeira constatação, foi traçado um método para aproximação e refinamento dessa tese e das próprias buscas.

Numa primeira fase (11 de maio de 2024 às 19h30), executou-se uma busca na base CAFe, do Portal Capes, pela expressão Inteligência Artificial Explicável em português (em qualquer campo de busca). Foram encontrados oito artigos, dois duplicados. Nenhuma das publicadoras está no campo do design - cinco em direito, duas em Ciência da Computação e uma em inovação. Quando à busca foi acrescentada a palavra "design", o resultado foi zero. Quando a busca foi feita em inglês, o resultado para "Explainable Artificial Intelligence" foi de 2.673 artigos revisados por pares. O acréscimo de "design" forneceu 472 artigos; a substituição por "graphic design" resultou em zero; quando a expressão usada foi "scientific visualization", um (1) artigo foi encontrado.

Uma segunda abordagem foi feita (28 de maio de 2024 às 18h48), no sentido de refinar as buscas. A mesma lógica utilizada na base CAFe foi reproduzida em quatro bases específicas, todas

elas reconhecidas e disponíveis para busca através do Portal de Periódicos da Capes. A numeração refere-se aos termos (*strings*) de busca utilizados em cada sessão. Os filtros usados foram "Revisados por pares (ou *reviews*)" e "sem data atribuída".

1. Termo: "Explainable Artificial Intelligence":
   - Embase Elsevier - 720 artigos
   - IEEE Xplore - 2.440 artigos
   - Science - 13 artigos
   - Science Direct (Elsevier) - 240 artigos

2. Termos: "Explainable Artificial Intelligence" and "design":
   - Embase Elsevier - 90 artigos
   - IEEE Xplore - 343 artigos
   - Science - 1 artigo
   - Science Direct (Elsevier) - 222 artigos

Ressalvando que a expressão "Design" aparece em contextos diferentes dos relacionados ao design como atividade profissional, sendo utilizada como sinônimo de projeto. Como exemplo, um dos artigos encontrados na busca pela base IEEE Xplore (R. ALIZADEHSANI et al., 2024), referia-se ao desenvolvimento de medicamentos utilizando a IA. Neste caso, a palavra design aparece associada ao projeto de drogas específicas (*drug design, compound design, protein design* etc).

3. Termos: "Explainable Artificial Intelligence" and "graphic design":
   - Embase Elsevier - 0 artigo
   - IEEE Xplore - 0 artigo
   - Science - 1 artigo
   - Science Direct (Elsevier) - 1 artigo

4. "Explainable Artificial Intelligence" and "scientific visualization"
   - Embase Elsevier - 0 artigos
   - IEEE Xplore - 11 artigos
   - Science - 2 artigos
   - Science Direct (Elsevier) - 1 artigo

A partir de considerações obtidas destas duas primeiras abordagens, uma aproximação mais rigorosa foi proposta. Inicialmente, foram estabelecidos termos de busca, com a ajuda do ChatGPT 4.0, que resultou em quatro grupos. No *string* "Design", foi feita a ressalva de que a procura deveria considerar o uso do termo como especialização acadêmica e não como sinônimo de projeto. Os grupos de termos ficaram assim definidos:

a. Explainable Artificial Intelligence, isoladamente;

b. Termos Explainable Artificial Intelligence + Design - "Explainable artificial intelligence" AND "design" or "human-centered design" or "design thinking" or "design principles" or "innovative design" or "user experience design" or "sustainable design";

c. Termos Explainable Artificial Intelligence + Graphic Design - "Explainable artificial intelligence" AND "graphic design" or "graphic design basics" or "typography in graphic design"or "graphic design trends 2024" or "branding graphic design" or "illustration in graphic design" or "graphic design inspiration";

d. Explainable Artificial Intelligence + Visualization - "Explainable artificial intelligence" AND "interactive visualization" or "visualization best practices" or "big data visualization" or "visual storytelling" or "scientific visualization" or "data visualization" or "visualization techniques" or "data visualization tools".

Para aplicar os termos obtidos na busca, foi escolhida a base Web of Science (WOS), por ter pertinência com a linha da pesquisa e por oferecer análise bibliométrica no próprio site de busca,

além de permitir acesso livre pelo portal CAFe da Capes. As buscas foram feitas no dia 29 de maio de 2024 aplicando os filtros de busca “Revisado por pares” e intervalo de tempo indeterminado. A tabela 1 sintetiza a busca e apresenta os primeiros resultados.

Com estes resultados, foram escolhidos parâmetros de análise bibliométrica que contribuíram para novas inferências (tabelas 2, 3 e 4 e 5). A bibliometria será uma ferramenta usada pelos autores deste artigo como base para a revisão sistemática da literatura da pesquisa desta tese de doutorado, por isso, considerou-se importante este detalhamento.

Algumas conclusões a partir destes resultados:

1. Uma primeira constatação aponta para concentração de artigos em publicações ligadas à Ciência da Computação e engenharias correlatas.

Tabela 1

ACERVO: WEB OF SCIENCE	*Buscas feitas no dia 29 de maio de 2024, entre 15h e 19h*
Termos de busca	**Resultados**
1. “Explainable artificial intelligence”	**184 artigos** (todos em inglês)
2. “Explainable artificial intelligence” **AND** “design” **or** “human-centered design” **or** “design thinking” **or** “design principles” **or** “innovative design” **or** “user experience design” **or** “sustainable design”	**32 artigos** (todos em inglês)
3. “Explainable artificial intelligence” **AND** “graphic design” **or** “graphic design basics” **or** “typography in graphic design” **or** “graphic design trends 2024” **or** “branding graphic design” **or** “illustration in graphic design” **or** “graphic design inspiration”	**0 artigos**
4. “Explainable artificial intelligence” **AND** “visualization”**or** “interactive visualization” **or** “visualization best practices” **or** “big data visualization” **or** “visual storytelling” **or** “scientific visualization” **or** “data visualization” **or** “visualization techniques” **or** “data visualization tools”	**8 artigos** (todos em inglês)

Fonte: Do autor

Tabela 2

BIBLIOMETRIA 1 — **184 artigos**

Para termo de busca “Explainable artificial intelligence”

Categorias Web of Science* (Maior número de artigos)

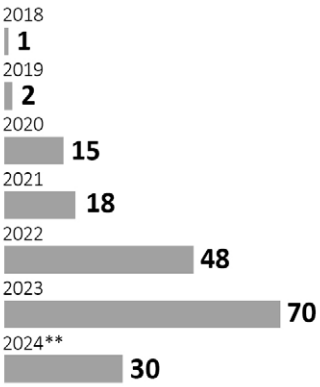
Computer Science Artificial Intelligence	**39**
Computer Science Information Systems	**37**
Engineering Electrical Electronic	**27**
Engineering Multidisciplinary	**16**
Computer Science Theory Methods	**15**
Chemistry Multidisciplinary	**14**
Computer Science Interdisciplinary Applications	**13**
Materials Science Multidisciplinary	**13**

E mais outras 83 categorias

Número de artigos por países dos pesquisadores

1. EUA	**35**
2. Alemanha	**26**
3. Inglaterra	**22**
4. Índia	**21**
5. Coreia do Sul	**15**
6. Itália	**14**
7. China	**14**
8. Austrália	**12**
21. BRASIL	**4**

Anos de publicação



*Cada publicação pode ser marcada com até seis categorias diferentes **Até o dia 29 de maio de 2024
Fonte: Web of Science Core Collection: Web of Science Categories
https://support.clarivate.com/ScientificandAcademicResearch/s/article/Web-of-Science-Core-Collection-Web-of-Science-Categories?language=en_US

Fonte: Do autor

Tabela 3

BIBLIOMETRIA 2 — **32 artigos**

Para termos de busca "Explainable artificial intelligence" AND "design" or "human-centered design" or "design thinking" or "design principles" or "innovative design" or "user experience design" or "sustainable design"

Categorias Web of Science (Maior número de artigos)

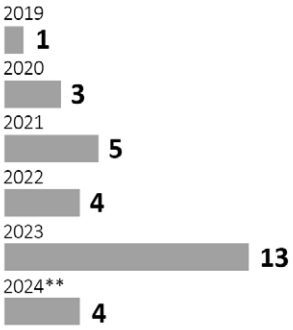
Computer Science Artificial Intelligence	9
Engineering Electrical Electronic	7
Chemistry Multidisciplinary	6
Computer Science Information Systems	6
Materials Science Multidisciplinary	6
Physics Applied	6
Engineering Multidisciplinary	5
Operations Research Management Science	4

E mais outras 23 categorias

Número de artigos por países dos pesquisadores

1. EUA	5
2. Itália	4
3. China	3
4. Espanha	3
5. Austrália	2
6. Alemanha	2
7. Paquistão	2
8. Turquia	2
11. BRASIL	**1**

Anos de publicação*



*Dois registros não tinham data
**Até o dia 29 de maio de 2024

Fonte: Do autor

Tabela 4

BIBLIOMETRIA 3 — **0 artigos**

Para termos de busca "Explainable artificial intelligence" **AND** "graphic design" **or** "graphic design basics" **or** "typography in graphic design" **or** "graphic design trends 2024" **or** "branding graphic design" **or** "illustration in graphic design" **or** "graphic design inspiration"

Fonte: Do autor

Tabela 5

BIBLIOMETRIA 4 — **8 artigos**

Para termos de busca "Explainable artificial intelligence" **AND** "visualization"**or** "interactive visualization" **or** " visualization best practices" **or** "big data visualization" **or** "visual storytelling" **or** "scientific visualization" **or** "data visualization" **or** "visualization techniques" **or** "data visualization tools"

Categorias Web of Science (Maior número de artigos)

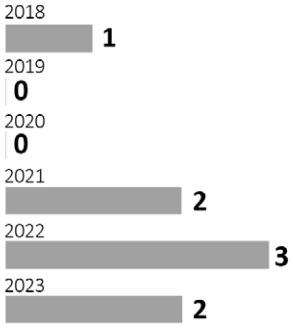
Computer Science Information Systems	3
Engineering Multidisciplinary	3
Chemistry Multidisciplinary	2
Materials Science Multidisciplinary	2
Physics Applied	2
Behavioral Sciences	1
Computer Science Artificial Intelligence	1
Computer Science Software Engineering	1

E mais outras 6 categorias

Número de artigos por países dos pesquisadores

1. EUA	3
2. Austrália	1
3. Etiópia	1
4. Alemanha	1
5. Irlanda	1
6. Suécia	1
BRASIL	**0**

Anos de publicação



Fonte: Do autor

2. Em relação à contemporaneidade do tema, verifica-se que a maior parte das publicações ocorreu nos últimos dois anos e vem mostrando tendência de crescimento em 2024.

3. Finalmente, como inferido anteriormente, o número de publicações por pesquisadores brasileiros ainda é pouco expressivo, a hegemonia mantém-se entre EUA e países europeus ou de língua inglesa, exceção feita à China e Coreia do Sul. Principalmente quando o tema principal - Inteligência Artificial Explicável - aparece desvinculado aos outros termos das buscas.

Inferências feitas a partir destes dados serão apresentadas nas considerações finais, na última parte deste texto.

6. Caminhos a seguir

Os primeiros passos apresentados no presente artigo, apontam para possíveis caminhos de aprofundamento de investigação, que, eventualmente, podem estar juntos numa mesma tese ou em pesquisas distintas.

O tópico 4 fala do primeiro destes caminhos: a investigação do papel que designers podem ou devem ter na evolução da IA para uma IAX, onde a confiabilidade é o foco. O objetivo, aqui, é lançar mão da especialização da atividade e, também, de inseri-la no grupo dos tomadores de decisão para caminhos da IA. Neste sentido, o design têm ligações com as ciências sociais que podem colaborar para que novos instrumentos considerem questões para além de aspectos técnicos e tecnológicos - como sustentabilidade e inclusão, entre outros. A isso podem ser acrescentados conhecimentos como semiótica, percepção visual e aspectos comunicacionais que são pertinentes ao exercício do aprendizado das máquinas, uma vez que, para tornar-se explicável, o processo precisa ser claro ao usuário humano.

Na segunda parte da pesquisa, um levantamento das arquiteturas usadas para o desenvolvimento das técnicas de IA mostra como muitos dos modelos apoiaram-se em representações visuais para aperfeiçoamento ou mesmo concepção das técnicas. Craven e Shavlik (1992), ao falarem sobre visualização científica e aprendizado de máquina afirmam “algorítimos de aprendizado podem levar a descoberta de importantes características nos dados de entrada que, inicialmente, não foram notadas. Se a representação gerada pelo algorítimo é compreensível, então estas descobertas tornam-se acessíveis à revisão de pessoas” (pg. 7, tradução do autor). Em algumas das imagens incluídas neste artigo, a opção por preservar sua forma original de publicação teve o propósito de enfatizar a solução gráfica usada por pesquisadores como apoio ao conhecimento. Como mostravam Pratt et. al (1991) “Uma vantagem alardeada das representações simbólicas é a facilidade de transferência de informações aprendidas de um agente inteligente para outro” (página 1, tradução do autor).

Nos dois casos, os autores dessa pesquisa entendem que há um fator estratégico na inclusão de designers, especialmente brasileiros, no papel de decisão da tecnologia de IA. Já há crescente interesse, não apenas por parte de empresas comerciais, mas também por órgãos de governos estrangeiros, em apoiar e financiar projetos em IA. Um sítio na web dedicado ao tema foi criado pelo governo americano para divulgar, estimular e discutir questões ligadas à Inteligência Artificial (ai.gov). Também nos EUA, o Departamento de Estado tem uma página própria (https://www.state.gov/artificial-intelligence/) que trata do assunto, onde podem ser acessados programas de governo quanto à utilização, participação e programas de cooperação, entre outros. Em 2020, o Congresso daquele país estabeleceu o Ato de Iniciativa Nacional para Inteligência Artificial (*National Artificial Intelligence Act*), que “estabelece a Iniciativa Nacional de Inteligência Artificial e atividades relacionadas” (5). Fica patente o interesse em garantir hegemonia - como foi citado no caso da Darpa, ao criar o programa de Inteligência Artificial Explicável.

No Brasil, em julho de 2021, o governo estabeleceu a Estratégia Brasileira de Inteligência

Artificial - EBIA, através de portaria. A descrição de intenções, na página dedicada a essa iniciativa (6), fala, na frase inicial, em "nortear as ações do Estado brasileiro em prol do desenvolvimento das ações, em suas várias vertentes, que estimulem a pesquisa, inovação e desenvolvimento de soluções em Inteligência Artificial (...)". Também estão descritos eixos de atuação, entre eles qualificações para um futuro digital; força de trabalho e capacitação; e pesquisa, desenvolvimento, inovação e empreendedorismo. A administração dessa iniciativa é feita pelo Ministério da Ciência, Tecnologia e Inovação. Ao participar ativamente no desenvolvimento de ferramentas de IA, diminui-se a dependência das soluções de fora do país, que estão sujeitas a interesses diferentes dos praticados no Brasil.

7. Considerações finais

A partir dos resultados obtidos no levantamento bibliométrico, alguns pontos foram considerados para destaque:

- Uma primeira constatação aponta para concentração de artigos em publicações ligadas à Ciência da Computação e engenharias correlatas.

- Em relação à contemporaneidade do tema, verifica-se que a maior parte das publicações ocorreu nos últimos dois anos e vem mostrando tendência de crescimento em 2024.

- O número de publicações por pesquisadores brasileiros ainda é pouco expressivo, a hegemonia mantém-se entre EUA e países europeus ou de língua inglesa, exceção feita à China e Coreia do Sul. Principalmente quando o tema principal - Inteligência Artificial Explicável (IAX) - aparece desvinculado aos outros termos das buscas.

- As pesquisas relacionadas a essa termo de busca ("Inteligência Artificial Explicável") ainda ocorrem em número que pode ser considerado baixo (se compararmos com uma busca simples na WOS pelo *string* "Inteligência Artificial", quando foram encontrados cerca de 22 mil artigos revisados por pares, em 12 de agosto de 2024, 20h);

- Não apenas a concentração das edições acontece em periódicos ligados à ciência da computação - o que pode ser considerado esperado, ou em telecomunicação e medicina: nenhuma publicação era da área de design;

- Artigos publicados por brasileiros, como visto, foram quatro num universo de 184. Mais uma vez, não havia publicação relacionada ao design.

Estes pontos mostram que há espaço para novas investigações.

No que tange à relevância do tema, há indicações de que a conjuntura é favorável; a rápida popularização das IAs e o desenvolvimento tecnológico, representado pelo maior poder computacional, tem estimulado movimentos no sentido de incrementar e aprimorar a difusão e as pesquisas em Inteligência Artificial, tanto em empresas comerciais como por parte das administrações públicas. A recente - e crescente - importância dada a explicabilidade das IAs abre ainda uma outra possibilidade. Sendo assim, aos autores deste artigo acreditam que é um momento de oportunidade para pesquisas que versem sobre a Inteligência Artificial Explicável.

O principal objetivo desta investigação será entender melhor os mecanismos que regem esta nova aproximação para o aperfeiçoamento das IAs, e identificar as oportunidades de melhora no sentido de transformá-las em IAX, usando para isso práticas e abordagens do design. Um caminho possível é o entendimento da interpretação e classificação de imagens pelas máquinas a partir da premissa da explicabilidade, por exemplo, o que proporcionaria protagonismo a designers.

Contribuir para esse aprimoramento, mas também buscar pela inclusão do design como atividade de igual importância a outras no desenvolvimento de soluções de inteligência artificial, justifica essa procura.

NOTAS

(1). IBM Developer. https://developer.ibm.com/articles/cc-cognitive-neural-networks-deep-dive/
(2) IBM. https://www.ibm.com/br-pt/topics/recurrent-neural-networks
(3) Data Science Academy. https://blog.dsacademy.com.br/o-que-sao-large-language-models-llms/
(4) Reuters. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/
(5) Congress.gov. https://www.congress.gov/bill/116th-congress/house-bill/6216
(6) (https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/inteligencia-artificial)

REFERÊNCIAS

Applied AI Letters, **Special Issue: DARPA's Explainable Artificial Intelligence (XAI) Program** - Volume2, Issue4, dezembro de 2021. https://doi.org/10.1002/ail2.15.

ANDRADE, Patrícia Santos. **Sistemas híbridos neuro simbólicos, estudo e implementação.** 104f. (Dissertação) Mestrado em Informática, Pós-Graduação em Informática, Centro de Ciências e Tecnologia, Universidade Federal da Paraíba, Campus II, Campina Grande - Paraíba - Brasil, 1997. Disponível em: http://dspace.sti.ufcg.edu.br:8080/jspui/handle/riufcg/11342

CRAVEN, Mark W. e Shavlik, Jude W. **Visualizing learning and computation in artificial neural networks**. International Journal on Artificial Intelligence Tools 01, nº 03 (setembro de 1992): 399–425. https://doi.org/10.1142/S0218213092000260.

CHEN, C., REN Y., KUO, CC.J. **Global-Attributes Assisted Outdoor Scene Geometric Labeling.** In: Big Visual Data Analysis. SpringerBriefs in Electrical and Computer Engineering(). Springer, Singapore, 2016. https://doi.org/10.1007/978-981-10-0631-9_5

DINIZ, Petterson Sousa. **Uma Abordagem Utilizando Séries Temporais para Detecção de Gás em Imagens Sísmicas com Transformer.** (2023).

ELGENDY, Mohamed. **Deep learning for vision systems.** Simon and Schuster, 2020.

GUNNING, D., VORM, E., WANG, J.Y. and TUREK, M. (2021), **DARPA's explainable AI (XAI) program: A retrospective.** Applied AI Letters, 2: e61. https://doi.org/10.1002/ail2.61

GUO, Q., JIN, S., LI, M. et al. **Application of deep learning in ecological resource research: Theories, methods, and challenges.** Sci. China Earth Sci. 63, 1457–1474 (2020). https://doi.org/10.1007/s11430-019-9584-9

J. DENG, W. DONG, R. SOCHER, L. -J. LI, Kai LI and Li FEI-FEI, **ImageNet: A large-scale hierarchical image database.** IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

MCCULLOCH, Warren S.; PITTS, Walter. **A logical calculus of the ideas immanent in nervous activity.** The bulletin of mathematical biophysics, v. 5, p. 115-133, 1943.

MORABITO, Francesco Carlo, Maurizio Campolo, Cosimo Leracitano, e Nadia Mammone. **Explainable**

Deep Learning to Information Extraction in Diagnostics and Electrophysiological Multivariate Time Series. Em Artificial Intelligence in the Age of Neural Networks and Brain Computing, 225–50. Elsevier, 2024. https://doi.org/10.1016/B978-0-323-96104-2.00011-7.

MINSKY, Marvin., PAPERT, Seymour. **Perceptrons; an Introduction to Computational Geometry.** Reino Unido: MIT Press, 1969.

R. ALIZADEHSANI et al., **Explainable Artificial Intelligence for Drug Discovery and Development: A Comprehensive Survey,** in IEEE Access, vol. 12, pp. 35796-35812, 2024, doi: 10.1109/ACCESS.2024.3373195.

ROSENBLATT, F. **The perceptron: A probabilistic model for information storage and organization in the brain.** Psychological Review, 65(6), 386–408 (1958). https://doi.org/10.1037/h0042519

ROSENBLATT, F. **The Design of an Intelligent Automaton.** Research Trends 6, no. 2 (1958).

RUMELHART, D., HINTON, G. & WILLIAMS, R. **Learning representations by back-propagating errors.** Nature (London) 323, nº 6088 (1986): 533–36. https://doi.org/10.1038/323533a0.
SCHMIDHUBER, Jürgen, e HOCHREITER, Sepp. **Long Short-Term Memory**. Neural Computation 9, nº 8 (1º de novembro de 1997): 1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.

TUFTE, Edward. **Envisioning information.** Graphics Press, USA, 1990.

VASWANI, A., SHAZEER, N.M., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A.N., KAISER, L., & POLOSUKHIN, I. (2017). **Attention is All you Need.** Neural Information Processing Systems.

VASWANI, A., SHAZEER, N.M., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A.N., KAISER, L., & POLOSUKHIN, I. **Attention Is All You Need**. arXiv, 1º de agosto de 2023. http://arxiv.org/abs/1706.03762.

WALTERS, W.H., WILDER, E.I. **Fabrication and errors in the bibliographic citations generated by ChatGPT.** Sci Rep 13, 14045 (2023). https://doi.org/10.1038/s41598-023-41032-5

ZHANG, J., LIC., YIN, Y. et al. **Applications of artificial neural networks in microorganism image analysis: a comprehensive review from conventional multilayer perceptron to popular convolutional neural network and potential visual transformer.** Artif Intell Rev 56, 1013–1070 (2023). https://doi.org/10.1007/s10462-022-10192-7.