

# Estimação Inteligente de Idade de Telespectadores para Aplicações de Sugestão de Conteúdo em *Smart TVs*

Nicoli P. Araújo, Elloá B. Guedes

<sup>1</sup> Escola Superior de Tecnologia  
Universidade do Estado do Amazonas  
Av. Darcy Vargas, 1200 – Manaus – Amazonas

{npda.eng,ebgcosta}@uea.edu.br

**Abstract.** *This work presents a proposal for estimating the age of viewers for content suggestion applications on Smart TVs using machine learning techniques. Such a tool can be used in a variety of ways, including to facilitate the collection of information that contributes to a better content delivery experience, to the creation and control of custom settings, and to the implementation of more efficient parental control.*

**Resumo.** *Este trabalho apresenta uma proposta para estimação de idade de telespectadores para aplicações de sugestão de conteúdo em Smart TVs utilizando técnicas de machine learning. Tal ferramenta pode ser utilizada de diversas maneiras, incluindo para facilitar a coleta de informações que contribuam para melhor experiência de provimento de conteúdo, para a criação e controle de configurações personalizadas e para a implementação de um controle parental mais eficiente.*

## 1. Introdução

As *Smart TVs* são o resultado da evolução tecnológica junto aos aparelhos de televisão domésticos. Possuem capacidades interativas ligadas à internet, acesso a conteúdo online, *e-commerce* de conteúdo televisivo, navegação web e acesso a redes sociais. Estes aparelhos podem ser equipados com câmeras e microfones embutidos e são aptos a transmitir conteúdo 2D ou até mesmo 3D especiais (NEWSROOM, 2011; PERAKAKIS; GHINEA, 2015).

Segundo a Pesquisa Nacional por Amostra de Domicílios realizada pelo IBGE em 2015, foi observado um total de 103 milhões de aparelhos de televisões em residências e pontos comerciais, das quais 16 milhões são de *Smart TVs*. A pesquisa detalha que 94% destas *Smart TVs* foram adquiridas entre 2014 e 2015. Os números mostram um posterior aumento nas vendas de aparelhos televisores deste tipo, representando 68,2% do total de televisores vendidos no primeiro semestre de 2017 (IBGE, 2015).

Este aumento de vendas tem várias causas, das quais destacam-se os muitos benefícios resultantes do uso de *Smart TVs* quando comparadas aos aparelhos convencionais (SHIN; HWANG; CHOO, 2013; BETWEEN, 2017). Em especial, cita-se o aumento da qualidade na transmissão, a utilização de aplicativos diversos e a possibilidade de acesso à conteúdo *online* e *on demand*, gratuitos ou mediante assinaturas. Além destes benefícios, cuja maioria é resultante da conectividade com a internet, outros fatores têm justificado o aumento das vendas e do interesse do público consumidor pelas *Smart TVs*, tais como o encerramento da transmissão de sinal analógico da televisão aberta, a Copa do Mundo 2018 e a tecnologia 4K (GUIMARÃES, 2017; BRAZILIENSE, 2018; CAPELAS, 2017).

Considerando a grande difusão das *Smart TVs* nos lares brasileiros, é essencial que estes aparelhos sejam capazes de capturar o perfil e o interesse dos seus telespectadores a fim de oferecer uma experiência mais rica. A recomendação de conteúdo, por exemplo, pode levar em conta características individuais, tais como idade e sexo. Porém, se fornecidos de maneira habitual, via preenchimento de formulários, além de ser uma tarefa massante, pode não refletir de maneira realística o perfil individual dos vários usuários que podem estar à frente de uma *Smart TV* em um determinado momento.

Apesar das dificuldades práticas mencionadas, é interessante notar que muitas *Smart TVs* possuem dispositivos para captura de imagens, como câmeras, pois também costumam dispor de aplicações para troca de mensagens de vídeo (SCHOFIELD, 2017). Respeitadas as preferências de privacidade de cada usuário, se estas câmeras forem habilitadas para aquisição de imagens daqueles que estão à frente do televisor, então é possível usá-las como entrada para sistemas inteligentes de identificação de características, cujas previsões podem ser usadas, por exemplo, para recomendação de conteúdo. No caso da idade, em particular, é possível usar estas informações para realizar um controle parental mais eficiente, protegendo crianças e adolescentes de conteúdos inadequados à sua faixa etária.

Diante do que foi exposto, esta proposta de trabalho de conclusão de curso considera o desenvolvimento de estratégias inteligentes, baseadas na utilização de técnicas de *Deep Learning*, para estimação da idade de telespectadores a partir de fotografias faciais. Embora a estimação de outras características também pudesse ser realizada mediante a análise de fotografias faciais, desde gênero até a presença de doenças, optou-se pela idade por ser um atributo comum a todos os telespectadores, pelo potencial de aplicações, pela existência de bases de dados adequadamente rotuladas com este atributo e pelo menor potencial de infringência das searas privadas dos usuários.

### **1.1. Objetivos**

O objetivo geral deste trabalho consiste em elaborar estratégias inteligentes para estimação de idade de telespectadores de *Smart TVs* a partir de suas respectivas fotografias faciais. Para alcançar esta meta, alguns objetivos específicos precisam ser contemplados, a citar:

1. Formular um referencial teórico sobre redes neurais convolucionais, contemplando seu arcabouço matemático, suas características, principais arquiteturas, métodos de treinamento e teste;
2. Consolidar uma base de dados com exemplos realísticos para treinamento dos modelos, tendo em vista a captura de padrões representativos ao domínio do problema;
3. Identificar tecnologias adequadas para implementação dos estimadores;
4. Propor, treinar e testar diferentes estimadores de idade baseados em redes neurais convolucionais para a tarefa em questão;
5. Avaliar comparativamente os estimadores propostos.

### **1.2. Justificativa**

A realização de um trabalho de conclusão de curso desta natureza é justificada por várias razões. No contexto da interação entre telespectador e *Smart TV*, um estimador de idade pode ser utilizado para facilitar a coleta de informações que contribuam para melhor experiência de provimento de conteúdo e de configurações personalizadas. Em particular, a estimação de idade dos telespectadores pode ser especialmente para a implementação

de um controle parental mais eficiente, protegendo crianças e adolescentes de conteúdos inadequados à sua faixa etária.

Um outro aspecto que ressalta a importância da realização de um trabalho desta natureza é a prática e a proposição de soluções envolvendo *Machine Learning*. Esta é uma área de vanguarda na Computação e seu potencial para resolução de problemas práticos está em franco desenvolvimento. Ao considerar a elaboração do estimador proposto, será necessário dominar conhecimentos de ferramental tecnológico atual, o que pode colaborar na minimização da distância entre o profissional em formação e os anseios do mercado de trabalho da área.

Por fim, há que se mencionar a relação entre a área de pesquisa considerada neste trabalho de conclusão de curso e o Laboratório de Sistemas Inteligentes (LSI). Este trabalho alinha-se com os objetivos desta iniciativa do Núcleo de Computação (NUCOMP), motivando o desenvolvimento de uma solução inovadora que utiliza técnicas da Inteligência Artificial.

### 1.3. Metodologia

A metodologia para o desenvolvimento deste trabalho consiste na realização da *fundamentação teórica sobre Machine Learning*, em especial contemplando os conceitos relativos às redes neurais convolucionais. Para tanto, considerar-se-á a literatura desta área para que haja o entendimento das bases matemáticas deste modelo computacional, como funcionam, quais as características e as arquiteturas mais importantes. Neste estudo, além dos aspectos teóricos, serão considerados os ambientes de desenvolvimento, bibliotecas e outras tecnologias para implementação dos conceitos contemplados.

Os demais passos que compõem a metodologia deste trabalho baseiam-se no *fluxo de atividades de machine learning* (MARSLAND, 2015). Inicialmente, haverá a aquisição e o pré-processamento de imagens para *consolidar uma base de dados* para esta tarefa de aprendizado. Nesta etapa, será considerada a literatura e, se possível, bases de dados já disponíveis e apropriadamente anotadas, com licença livre de utilização.

A seguir, há a *proposição de diferentes modelos de redes neurais convolucionais* para a tarefa de aprendizado considerada. Nesta etapa, serão elencados diferentes parâmetros e hiperparâmetros de configuração, bem como arquiteturas. Estes procedimentos visam consolidar um espaço de busca de modelos que possam endereçar a tarefa de maneira mais eficiente.

O próximo estágio consiste no *treinamento das redes neurais convolucionais* para o problema em questão. Durante este processo, uma parte da base de dados será apresentada aos modelos para que haja o ajuste de pesos, compreendendo o aprendizado das características relevantes. O treinamento das redes ocorrerá utilizando computação em nuvem, tendo em vista a infra-estrutura de hardware necessária para realizar este procedimento.

Segue-se então o *teste das redes*, respeitando uma abordagem de validação cruzada e utilizando métricas de desempenho apropriadas. O objetivo desta fase consiste em aferir os modelos propostos e treinados quanto à sua capacidade de generalização.

Por fim, para identificação de um modelo mais adequado à esta tarefa, as *métricas de desempenho serão comparadas* e os melhores modelos elencados a partir destes valores, apontando assim um estimador apropriado para o problema inicialmente considerado.

Além destas atividades, há que se considerar a escrita da proposta e do projeto final do trabalho de conclusão de curso, bem como as defesas parcial e final.

H

Tabela 1: Cronograma de atividades levando em consideração os dez meses (de 02/2018 a 12/2018) para a realização do TCC.

	2018											
	02	03	04	05	06	07	08	09	10	11	12	
<b>Escrita da Proposta</b>	X	X	X	X	X							
<b>Fundamentação Teórica sobre Machine Learning</b>	X	X	X	X								
<b>Consolidação da Base de Dados</b>		X	X									
<b>Proposição de Modelos de Redes Neurais Convolucionais</b>				X	X	X	X	X				
<b>Defesa da Proposta</b>					X							
<b>Escrita do Trabalho Final</b>						X	X	X	X	X	X	
<b>Treinamento das Redes Neurais Convolucionais</b>					X	X	X	X	X	X		
<b>Teste das Redes Neurais Convolucionais</b>					X	X	X	X	X	X	X	
<b>Comparação de Mettricas de Desempenho</b>						X	X	X	X	X	X	
<b>Defesa do Trabalho Final</b>												X

#### 1.4. Cronograma

O cronograma de realização das atividades pode ser visto na Tabela 1. As atividades listadas possuem relação com a metodologia detalhada na seção anterior, compreendendo os requisitos elementares para a realização deste trabalho.

#### 1.5. Organização do Documento

Para a apresentação desta proposta de trabalho de conclusão de curso, o presente documento está organizado como segue. Inicialmente, uma fundamentação teórica pode ser vista na Seção 2. Uma análise dos trabalhos relacionados encontra-se na Seção 3. Na Seção 4 detalha-se uma solução proposta para a tarefa endereçada. Finalmente, as considerações finais e os trabalhos futuros podem ser encontrados na Seção 5.

### 2. Fundamentação Teórica

A fundamentação teórica para a realização deste trabalho compreende conceitos ligados às *Smart TVs* e ao *Machine Learning*. Quanto ao primeiro tópico, uma caracterização das *Smart TVs* é apresentada na Subseção 2.1, e uma visão geral dos conceitos ligados à classificação indicativa é apresentada na Seção 2.2. Quanto ao segundo tópico, a Subseção 2.3 compreende os conceitos essenciais de *Machine Learning*, em que as redes neurais são particularmente detalhadas na Subseção 2.4. Os conceitos mais emergentes desta área, envolvendo *Deep Learning*, são descritos na Seção 2.5.

#### 2.1. Smart TVs

As *Smart TVs* são o resultado da evolução tecnológica junto aos aparelhos de televisão domésticos. Possuem capacidades interativas ligadas à internet, acesso a conteúdo online,

*e-commerce* de conteúdo televisivo, navegação web e acesso a redes sociais. Estes aparelhos podem ser equipados com câmeras e microfones embutidos e transmitir conteúdo 2D ou até mesmo 3D. Neste último caso, em particular, os telespectadores fazem uso de óculos especiais.

O principal diferencial no tocante ao hardware entre *Smart TVs* e as antigas tecnologias LED e LCD TV reside na conexão com a internet, a qual pode ser realizada via módulo Wi-Fi ou Ethernet (BETWEEN, 2017; QUAIN, 2018). Para promover esta conexão e posterior interação com o usuário, estas televisões utilizam os mesmos sistemas operacionais e conjuntos de aplicativos que computadores ou *smartphones* convencionais, em especial mencionam-se navegador web e diversos aplicativos.

É possível também que *Smart TVs* exibam conteúdo de mídia transmitido a partir de *smartphones* ou computadores conectados na mesma rede Wi-Fi, conforme o padrão de compartilhamento de mídia DLNA (*Digital Living Network Alliance*) (MICHÉLE; KARPOW, 2014; SHIN; HWANG; CHOO, 2013; PERAKAKIS; GHINEA, 2015; KOVACH, 2010). Muitos modelos destes televisores também possuem ferramentas para o reconhecimento de comandos de voz, possibilitando funcionalidades como troca e busca de canais, controle de volume, etc. Este controle de voz costuma também estar integrado com funções das casas inteligentes, tendência da Internet das Coisas (QUAIN, 2018).

A Figura 1 exibe um diagrama representativo dos elementos que compõem uma *Smart TV*. As legendas para os números apresentados na imagem estão na Tabela 2. Dentre os diversos fabricantes destes dispositivos, em nível mundial destacam-se as marcas Hisense, LG, Panasonic, Phillips, Samsung, Sharp, Sony, TCL, Toshiba e Vizio (QUAIN, 2018).



Figura 1: Diagrama representativo de uma *Smart TV* e seus componentes (NEWSROOM, 2011). Ver legenda dos componentes na Tabela 2.

Tabela 2: Legenda dos componentes citados na Figura 1.

Número	Descrição	Número	Descrição
1	Moldura	13	Sintonizador, 4 portas HDMI e 3 portas USB
2	Painel de cristal negro (célula)	14	3D <i>Hyper Real Engine</i>
3	Molde da moldura do meio	15	Placa de Alimentação
4	Folha óptica	16	Sensor de luz ambiente
5	LGP – <i>Light Guide Plate</i>	17	Módulo <i>bluetooth</i>
6	LED	18	Módulo Wi-Fi
7	Chassi traseiro	19	Auto-falantes
8	Cobertura intermediária	20	Suporte quadrangular
9	Cobertura traseira	21	Botão <i>touch</i> operacional
10	Placa de circuito principal (Placa mãe)	22	Câmera de video de telefone
11	<i>Smart Real Engine</i>	23	Suporte de parede
12	<i>Speed Backlite Engine</i>	24	Controle remoto QWERTY
		25	Óculos 3D

As aplicações disponíveis para *Smart TVs* são diversas, permitindo, por exemplo, o acesso a conteúdo de programas e também a informações esportivas, como é comum no caso do futebol. Um exemplo de aplicação disponível para *Smart TVs* é a disponibilizada desde 2016 pela emissora aberta SBT, vide Figura 2. Este aplicativo contém novelas, programas e outras atrações disponibilizadas pela emissora que podem ser assistidos *on demand* (SBT, 2015). Outros exemplos compreendem os aplicativos de *streaming*, tais como Netflix, Amazon Prime Video, Hulu e Pandora (CIRIACO, ).



Figura 2: Aplicativo SBT. Fonte: (SBT, 2015)

Segundo a Pesquisa Nacional por Amostra de Domicílios realizada pelo IBGE em 2015, foi observado um total de 103 milhões de aparelhos de televisões em residências e pontos comerciais, das quais 16 milhões são de *Smart TVs*. A pesquisa detalha que 94% destas *Smart TVs* foram adquiridas entre 2014 e 2015. Os números mostram um posterior aumento nas vendas de aparelhos televisores deste tipo, representando 68,2% do total de televisores vendidos no primeiro semestre de 2017 (IBGE, 2015).

Há muitos benefícios resultantes do uso de *Smart TVs* quando comparadas aos aparelhos convencionais. Em especial, cita-se o aumento da qualidade na transmissão, a utilização de aplicativos diversos e a possibilidade de acesso à conteúdo *online* e *on demand*, gratuitos ou mediante assinaturas. Além destes benefícios, cuja maioria é resultante da conectividade com a internet, outros fatores têm justificado o aumento das vendas e do interesse do público consumidor pelas *Smart TVs*, tais como o encerramento da transmissão de sinal analógico da televisão aberta, a Copa do Mundo 2018 e a tecnologia 4K (GUIMARÃES, 2017; BRAZILIENSE, 2018; CAPELAS, 2017).

Apesar da grande disponibilidade de conteúdo nas *Smart TVs*, é imprescindível levar em conta as restrições e recomendações deste conteúdo para o público alvo a que se destina. Neste sentido, a próxima seção detalha as políticas vigentes de classificação indicativa de conteúdo televisivo.

## **2.2. Classificação Indicativa para Conteúdo Televisivo**

O processo de classificação indicativa integra o sistema de garantias dos direitos da criança e do adolescente quanto a promover, defender e garantir o acesso a espetáculos e diversões públicas adequados à condição de seu desenvolvimento, mas reserva-se o direito final aos pais e responsáveis quanto à escolha do conteúdo adequado a estes (DEPUTADOS, 1995).





No Brasil, a *Coordenação de Classificação Indicativa* (Cocind), vinculada ao Ministério da Justiça, é o órgão responsável pela classificação indicativa de obras destinadas à televisão e outros meios, incluindo até mesmo aplicativos. A análise da classificação indicativa realizada pelo Cocind considera o grau de incidência de conteúdos de sexo e nudez, violência e drogas nas obras a serem avaliadas, como sintetizado na Tabela 3. O processo envolve o exame do conteúdo das obras a serem classificadas, a atribuição de classificação indicativa, verificação do cumprimento das normas associadas e advertência por descumprimento destas normas (JUSTIÇA, 2014).

No mundo, conteúdos televisivos são comumente classificados quanto ao grau de incidência de assuntos como linguagem vulgar, conteúdo sexual, drogas e violências, além de temas como conteúdo perturbador e discriminação, a exemplo dos Países Baixos. É frequente a aplicação de restrições de horários para a transmissão de conteúdos restritivos. As classes podem incluir restrição de idade e/ou supervisão de responsáveis, como ocorre nos Estados Unidos, Chile, Equador, Hong Kong, entre outros. Em países como a Austrália e Nova Zelândia, há um sistema de classificação indicativa para televisão aberta e outro para fechada, e um sistema de classificação especial para programas direcionados ao público infantil, na Austrália. Na Colômbia, é proibida a transmissão aérea de pornografia, mesmo em canais adultos. O ícone da classificação indicativa frequentemente deve ser exibido antes do início do programa, antes do início de cada bloco, a exemplo do Brasil, ou durante toda a transmissão do programa, como é o caso da França. Na Alemanha, apenas o aviso “O programa a seguir não é recomendado para espectadores abaixo de 16/18 anos” é mostrado na tela caso haja conteúdo potencialmente ofensivo. Em países como Portugal, Polônia e Singapura, a implantação de sistemas de classificação indicativa é posterior ao ano de 2000 (WIKIPEDIA, 2018).

## **2.3. Machine Learning**

*Machine Learning* (ML), também chamado de Aprendizado de Máquina, é uma subárea da Inteligência Artificial que trata do estudo sistemático de algoritmos e sistemas que são capazes de melhorar seu desempenho com a experiência. Um algoritmo neste domínio

Tabela 3: Categorias de classificação indicativa propostas pela Portaria No. 368, de 11 de Fevereiro de 2014. Fonte: (JUSTIÇA, 2012)

<b>Categoria</b>	<b>Símbolo</b>	<b>Descrição do Conteúdo</b>
Livre		Conteúdo predominantemente positivos ou que contenham imagens de violência fantasiosa, armas sem violência, mortes sem violência, ossadas e esqueletos sem violência, nudez não erótica e consumo moderado ou inusitado de drogas lícitas.
Não recomendado para menores de dez anos		Presença de armas com violência; medo ou tensão; angústia; ossadas e esqueletos com resquícios de ato de violência; atos criminosos sem violência; linguagem depreciativa; conteúdos educativos sobre sexo; descrições verbais do consumo de drogas lícitas; discussão sobre o tráfico de drogas; e o uso medicinal de drogas ilícitas.
Não recomendado para menores de doze anos		Ato violento; lesão corporal; descrição de violência; presença de sangue; sofrimento da vítima; morte natural ou acidental com violência; ato violento contra animais; exposição ao perigo; exposição de pessoas em situações constrangedoras ou degradantes; agressão verbal; obscenidade; bullying; exposição de cadáver; assédio sexual; supervalorização de beleza física; supervalorização do consumo; nudez velada; insinuação sexual; carícias sexuais; masturbação não explícita; linguagem chula; linguagem de conteúdo sexual; simulações de sexo; apelo sexual; consumo de drogas lícitas; indução ao uso de drogas lícitas; consumo irregular de medicamentos; menção a drogas ilícitas.
Não recomendado para menores de catorze anos		Morte intencional; estigma ou preconceito; nudez; erotização; vulgaridade; relação sexual não explícita; prostituição; insinuação do consumo de drogas ilícitas; descrições verbais do consumo de drogas ilícitas; e discussão sobre a descriminalização de drogas ilícitas.
Não recomendado para menores de dezesseis anos		Estupro; exploração sexual; coação sexual; tortura; mutilação; suicídio; violência gratuita ou banalização da violência; aborto, pena de morte ou eutanásia; relação sexual intensa não explícita; produção ou tráfico de qualquer droga ilícita, consumo de drogas ilícitas; indução ao consumo de drogas ilícitas.
Não recomendado para menores de dezoito anos		Violência de forte impacto; elogio; glamorização e/ou apologia à violência; crueldade; crimes de ódio; pedofilia; sexo explícito; situações sexuais complexas ou de forte impacto; apologia ao uso de drogas ilícitas.

é capaz de aprender a partir de dados, capturando padrões e efetuando inferências. Estes algoritmos podem ser entendidos em uma analogia com humanos e outros animais que, ao se depararem com determinada situação, costumam procurar lembranças de situações similares, de como agiram, e se o comportamento adotado foi vantajoso, e deve ser repetido, ou prejudicial, devendo ser evitado (MARSLAND, 2015; GOODFELLOW; BENGIO; COURVILLE, 2016; FLACH, 2012).

Para consolidar o aprendizado, os algoritmos de *machine learning* precisam passar por um processo de aquisição da experiência, comumente chamado de treinamento. De acordo Mitchell (MITCHELL, 1997), um algoritmo que aprende a partir da experiência  $E$  quanto a um conjunto de tarefas  $T$  e medida de performance  $P$ , se sua performance nas tarefas em  $T$ , medida por  $P$ , melhora com a experiência  $E$ .

Ao preparar um algoritmo de *machine learning* para desenvolver determinada ta-



refa, busca-se um modelo, ou seja, uma função, que mapeie as instâncias do espaço de entrada para o de saída (FLACH, 2012). Estes modelos podem ser agrupados em diferentes categorias ao se considerar o tipo de aprendizado e também a saída desejada para o algoritmo. A Figura 3 apresenta uma visão geral do estado da arte acerca dos modelos de ML e suas subdivisões.

Quanto ao tipo de aprendizado, as tarefas de ML podem ser agrupadas em três tipos diferentes, a depender da presença e do tipo de resposta dada ao algoritmo quanto ao desempenho de suas saídas. No *aprendizado supervisionado*, o algoritmo deve aprender a inferir valores a partir de atributos preditores e do respectivo atributo alvo fornecido como exemplo, ou seja, de cenários em que se têm seus valores de saída conhecidos. Os modelos mais comumente utilizados neste tipo de aprendizado são as máquinas de vetores de suporte, redes neurais artificiais *feed-forward*, regressão linear e logística, etc. Já no *aprendizado não-supervisionado*, o algoritmo deve inferir padrões e estruturas a partir de dados não rotulados, buscando alguma estrutura interna que caracterize os dados. Exemplos de modelos aplicáveis a este cenário são os algoritmos *k-means* e *k-medoids*. Por fim, no *aprendizado por reforço* o algoritmo não recebe dados nem tampouco rótulos, e deve aprender a partir das recompensas positivas ou negativas dadas por ações que modifiquem o ambiente de maneira satisfatória ou não (FLACH, 2012).

Quanto ao tipo de saída desejada, os problemas que podem ser endereçados segundo ML são de classificação, regressão, transcrição, tradução automática, detecção de anomalia, síntese e amostragem. No caso do aprendizado supervisionado, em particular, as principais tarefas realizadas são de classificação e regressão (FLACH, 2012).

Um algoritmo proposto a uma tarefa de classificação deve especificar cada entrada  $x$  como pertencente a uma dentre  $k$  categorias pré-determinadas, produzindo uma saída  $y = f(x)$  tal que a função  $f$  é definida como  $f : \mathbb{R}^n \rightarrow \{1, \dots, k\}$ , ou seja,  $f$  mapeia sequências de números reais  $x$  de dimensão  $n$  para um valor inteiro  $y$  dentre  $k$  possibilidades (GOODFELLOW; BENGIO; COURVILLE, 2016). Dentre as tarefas de classificação estão, por exemplo, o reconhecimento de objetos em uma imagem, determinar se um indivíduo será ou não vítima de determinada doença, se sobreviverá ou não a determinado acidente, etc.

Numa tarefa de regressão, por sua vez, objetiva-se aprender uma função de valor real a partir de uma entrada (FLACH, 2012). Assim, a saída  $y = f(x)$  é dada pela função  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , ou seja,  $f$  mapeia uma entrada multidimensional  $x$  para um valor  $y$  real (GOODFELLOW; BENGIO; COURVILLE, 2016). Algumas tarefas de regressão envolvem a previsão de preços de um mercado de ações, a determinação do risco do seguro para um carro, do volume diário de precipitação em determinada cidade, etc.

Os modelos de ML são organizados em dois grandes grupos, dos tipos paramétricos ou não paramétricos. Segundo Russel e Norvig, um modelo de aprendizado que resume dados utilizando um conjunto de parâmetros de tamanho definidos independente do número de exemplos de treinamento é chamado de *modelo paramétrico*. Dentre os modelos paramétricos está a regressão logística. Já um *modelo não-paramétrico*, por sua vez, é aquele que não pode ser caracterizado por um conjunto limitado de parâmetros. Alguns exemplos de modelos não-paramétricos são máquinas de vetores de suporte, redes neurais artificiais,  $k$  vizinhos mais próximos e árvores de decisão CART e C4.5 (RUSSELL; NORVIG, 2016).

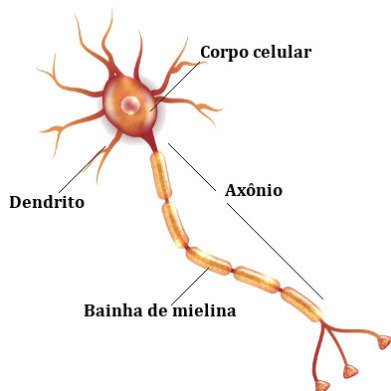
Dentre os modelos não paramétricos, as redes neurais artificiais têm demonstrado



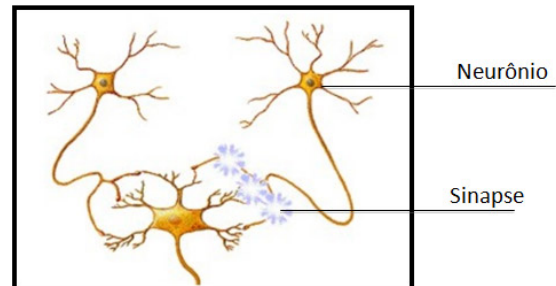
Figura 3: Mapa mental dos algoritmos de *Machine Learning* organizados por área e sub-área.

Figura 4: Redes neurais biológicas. Fonte: FALTANDO!

(a) Neurônio biológico e seus componentes.



(b) Sinapse entre neurônios.



resultados satisfatórios em tarefas de classificação e regressão quando aplicadas em diversas áreas. Em especial, aplicações de *Deep Learning* no reconhecimento de objetos e no processamento de linguagem natural, por exemplo, têm trazido ainda mais atenção a este modelo. Diante desta importância e da utilização no contexto deste trabalho, estes conceitos serão abordados com mais profundidade nas seções a seguir.

## 2.4. Redes Neurais Artificiais

As *Redes Neurais Artificiais* (RNAs) são um modelo de computação caracterizado por sistemas que, em algum nível, lembram a estrutura do cérebro humano. São sistemas paralelos e distribuídos, compostos por unidades de processamento simples, os *neurônios artificiais*, que calculam funções matemáticas, normalmente não-lineares. Estes neurônios são dispostos em uma ou mais camadas e interligados por um grande número de conexões normalmente unidirecionais e comumente associadas a pesos, que armazenam o conhecimento representado no modelo e ponderam a entrada recebida por cada neurônio da rede. Os principais atrativos das RNAs envolvem a capacidade de capturar tendências a partir de um conjunto de exemplos e dar respostas coerentes para dados não-conhecidos, ou seja, de generalizar a informação aprendida (BRAGA; CARVALHO; LUDERMIR, 2007).

A motivação para a criação deste modelo vem do funcionamento do cérebro biológico, que é formado por neurônios interligados e que se comunicam entre si de modo contínuo e paralelo através de impulsos nervosos. Esta complexa rede neural biológica é capaz de reconhecer padrões e relacioná-los, produzir emoções, pensamentos, percepção e cognição. Cada neurônio biológico é composto de um corpo, dendritos e um axônio, como ilustrado na Figura 4a. Os dendritos são responsáveis pela recepção de impulsos nervosos vindos de outros neurônios; o corpo combina os sinais recebidos pelos dendritos e caso o resultado ultrapasse determinado limiar de excitação do neurônio, são gerados novos impulsos nervosos, que são transmitidos pelo axônio até os dendritos dos neurônios seguintes. Esta conexão unilateral entre neurônios biológicos, denominada sinapse, encontra-se ilustrada na Figura 4b.

Com base no modelo biológico, McCulloch e Pitts propuseram em um neurônio artificial (MCCULLOCH; PITTS, 1943). Como mostrado na Figura 5, o modelo de McCulloch e Pitts de neurônio artificial contém  $n$  terminais de entrada, denotados por  $x = x_1, \dots, x_n$ , e um terminal de saída  $y$ . Esta organização faz uma alusão aos dendritos,

centro e axônio de um neurônio biológico. A saída é mapeada por uma função  $y = \sigma(z)$ , expressa na Equação 1, em que a soma ponderada  $z$  do vetor de entrada  $x$  pelo conjunto de pesos  $w = w_1, \dots, w_n$  deve ser submetida a uma função de ativação  $\sigma$ , que determina se aquele neurônio é ativado ou não, no sentido de ter . No caso de um neurônio mais simples como o de McCulloch e Pitts, a função de ativação consiste de verificar se a soma ponderada  $z$  é maior ou igual a um limiar de ativação  $\theta$ , conforme a Equação 3 (MCCULLOCH; PITTS, 1943). Atualmente, a escolha da função de ativação  $g(\cdot)$  das camadas ocultas e da camada de saída deve considerar funções contínuas e deriváveis (HORNIK, 1991), em que comumente são optadas pelas funções apresentadas na Tabela 4.

definir ativação de um neurônio e colocar a citação aqui

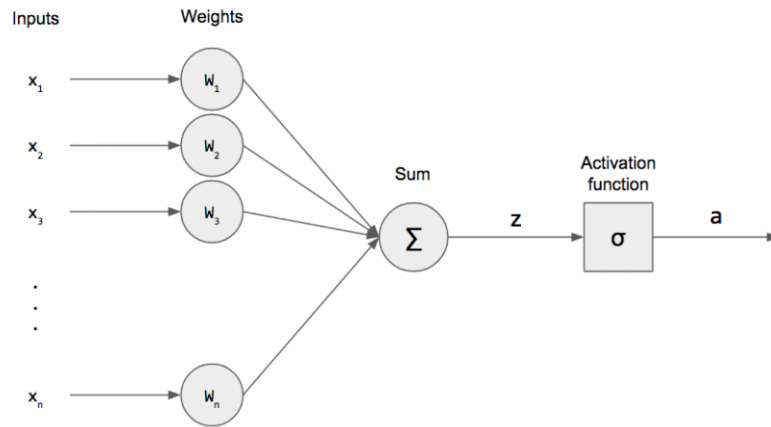



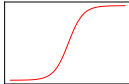
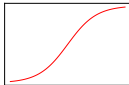

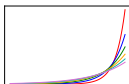
Figura 5: Representação de um neurônio

$$z = \sum_{i=1}^n x_i w_i + b_i \quad (1)$$

$$y = \sigma(z) \quad (2)$$

$$y = \sigma(z) = \begin{cases} 0, & \text{se } z < \theta \\ 1, & \text{se } z \geq \theta \end{cases} \quad (3)$$

Tabela 4: Exemplos de funções de ativação (GOODFELLOW; BENGIO; COURVILLE, 2016)

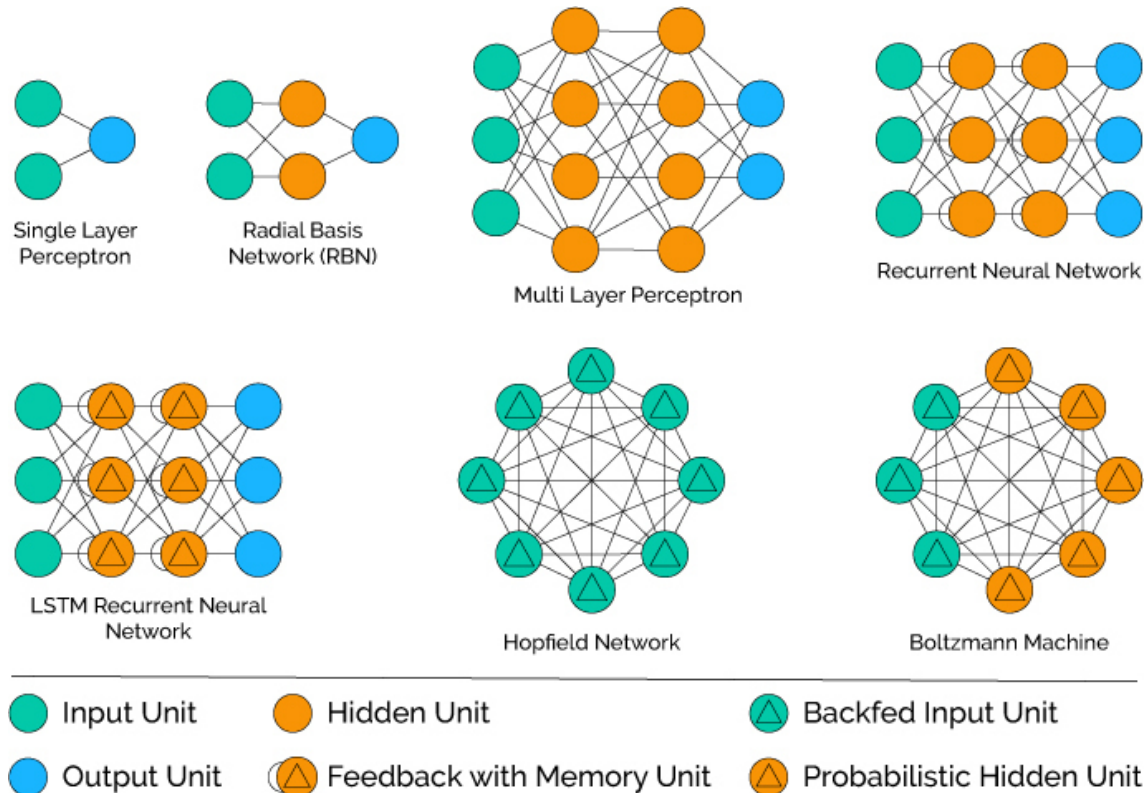
Nome	Gráfico	Equação	Intervalo
Identidade ou Linear		$g(z) = z$	$(-\infty, +\infty)$
Tangente Hiperbólica		$g(z) = \tanh(z) = \frac{(e^z - e^{-z})}{(e^z + e^{-z})}$	$(-1, 1)$
Sigmoide ou Logística		$g(z) = \sigma(z) = \frac{1}{1 + e^{-x}}$	$(0, 1)$
Unidade Linear Retificada		$g(z) = \max(0, z)$	$[0, \infty)$
Softmax		$g(z_j) = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}} \quad j = 1, \dots, K$	$(-\infty, \infty)$

Em 1958, Frank Rosenblatt desenvolveu o neurônio *Perceptron* (ROSENBLATT, 1958), que mais tarde seria empregado como a unidade de processamento das RNA e de outros modelos de ML, a exemplo das *support vector machines*. O Perceptron de Rosenblatt agregou ao neurônio de McCulloch e Pitts conceitos cruciais para a caracterização das RNAs como são conhecidas hoje, como a não obrigatoriedade de igualdade dos pesos e limiares de ativação, a possibilidade de os pesos serem positivos ou negativos, a diversidade de funções de ativação, entre outros. Além desta caracterização, uma contribuição relevante deste trabalho contempla a proposição de um algoritmo de aprendizado que permite a adaptação dos pesos de uma RNA através da otimização do desempenho da rede. Isto atribuiu ao modelo Perceptron a capacidade de aprender tarefas que contenham dados linearmente separáveis (BRAGA; CARVALHO; LUDERMIR, 2007).

Este modelo inicial apresentava algumas limitações, atribuídas principalmente à sua linearidade e simplicidade, características que possibilitam resolver apenas problemas linearmente separáveis (BRAGA; CARVALHO; LUDERMIR, 2007). A disposição de neurônios em camadas e a utilização de funções de ativação nas saídas dos neurônios caracterizou as RNAs, capazes de serem aproximadas universais de qualquer função contínua graças à otimização por minimização da dissimilaridade entre o valor previsto pela rede  $\hat{y}$  e o valor real  $y$ . Atualmente, as RNAs podem apresentar diversos tipos de arquitetura, ao variar-se parâmetros como o número de camadas, quantidade de neurônios em cada camada, os tipos de conexões entre neurônios e topologia de rede. Alguns exemplos de arquiteturas podem ser encontrados na Figura 6.

Quanto aos tipos de conexão possíveis entre os neurônios, tem-se que as RNAs podem ser do tipo *feedforward* ou recorrente. As RNAs *feedforward*, exemplificada na Figura 7a, são comumente associadas a um grafo acíclico em que as saídas de uma camada servem de entrada à camada seguinte, e assim sucessivamente, até que seja produzida uma saída. As RNAs recorrentes, como exemplificado na Figura 7b, contém conexões entre

Figura 6: Arquiteturas populares de RNAs. Fonte: ???



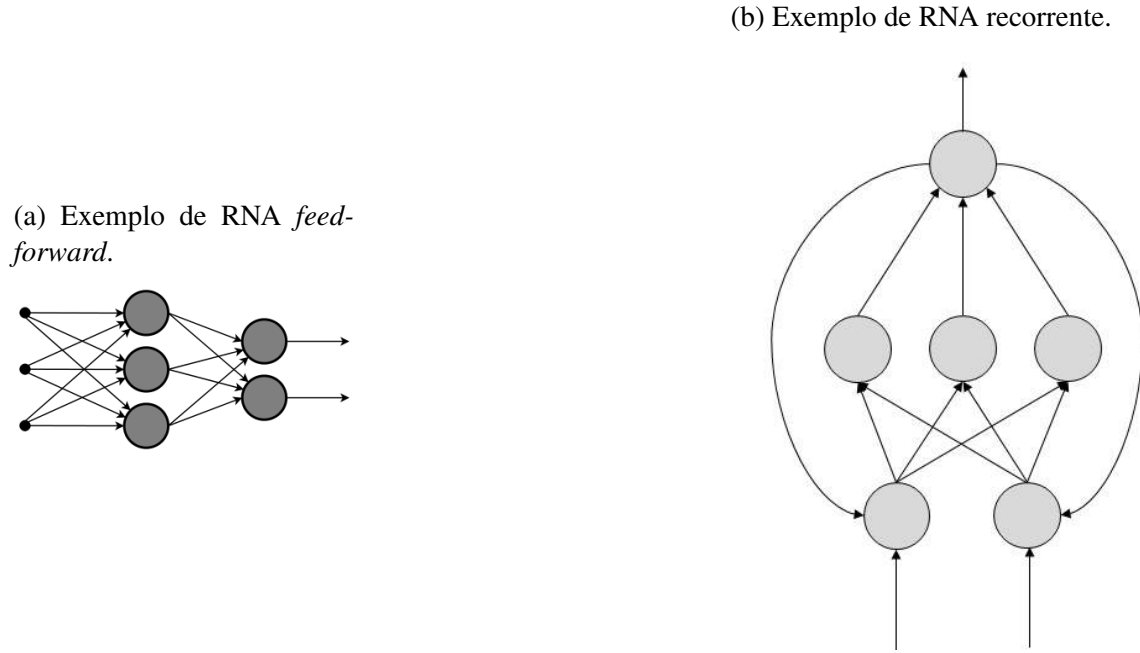
neurônios de modo a formar um grafo direcionado cíclico, o que permite que o modelo capture sequências de comportamentos organizados em séries temporais.

Um dos parâmetros relacionados à arquitetura de uma RNA é a quantidade de camadas ocultas. Pode-se ter redes de camada única, compostas por um neurônio que conecta todos os parâmetros de entrada às saídas do modelo, a exemplo das redes Perceptron. Há também as redes de múltiplas camadas, que consistem de mais de um neurônio entre entrada e saída da rede, como retratado na Figura 8. Redes com múltiplas camadas, as chamadas Redes Neurais *Feedforward Multilayer Perceptron* (MLP), são capazes de aproximar diversas funções (HORNICK, 1991; BRAGA; CARVALHO; LUDERMIR, 2007).

Segundo o Teorema da Aproximação Universal definido por Hornik em 1991, se a ativação de uma rede neural MLP for uma função limitada e não-constante, então dada uma entrada  $x$ , a rede é capaz de aproximar qualquer função contínua, provida uma quantidade adequada de camadas ocultas. Esta característica atribui às redes neurais artificiais o potencial de se tornarem máquinas de aprendizado universal (HORNICK, 1991).

O objetivo das RNAs é aproximar funções que mapeiam dadas entradas  $X$  às suas respectivas saídas  $Y$ . Para atingir este objetivo, é necessário minimizar a disparidade entre as saídas previstas  $\hat{Y}$  e as saídas desejadas  $Y$ . A função que calcula tal disparidade é chamada *função custo*, dada por  $J$  e tida como a soma funções de erros,  $L(\hat{y}_i, y_i)$ , entre cada saída esperada  $y_i$  e obtida pela RNA  $\hat{y}_i$ , acumuladas conforme o modelo é apresentado a  $m$  exemplos representativos do evento que se deseja aprender. A função custo está representada na Equação 5, e a função de erro, também conhecida como perda, é dada pela Equação 4. Neste contexto de otimização da previsão da RNA, deve-se minimizar a

Figura 7: Exemplos de RNA com diferentes tipos de conexões entre neurônios.



função custo para que haja a otimização do desempenho da RNA, tarefa realizada durante uma etapa de treinamento, que consiste em duas fases: a fase *forward* e a fase *backwards* (HAYKIN, 2009).

$$L(\hat{y}_i, y_i) = -\log p(y_i | \hat{y}_i). \quad (4)$$

$$J = \frac{1}{m} \sum_{i=1}^m L(\hat{y}_i, y_i) \quad (5)$$

Na fase *forward*, também chamada *forward propagation*, há a inferência das saídas da rede perante um conjunto de N entradas. Neste processo, a informação flui para frente, em inglês *forward*, através da rede conforme a entrada  $x_i$  provê as informações iniciais que são propagadas até as camadas ocultas, e de lá até a saída  $y_i$ . Esta propagação ocorre conforme a Equação Ao final da fase *forward* ocorrida na etapa de treinamento, a função custo  $J$  é calculada. A representação da sequência de passos realizados na fase *forward* está no Algoritmo ?? (HAYKIN, 2009; GOODFELLOW; BENGIO; COURVILLE, 2016).

Sendo  $l$  a camada atual e  $n$  layers o numero total de camadas

---

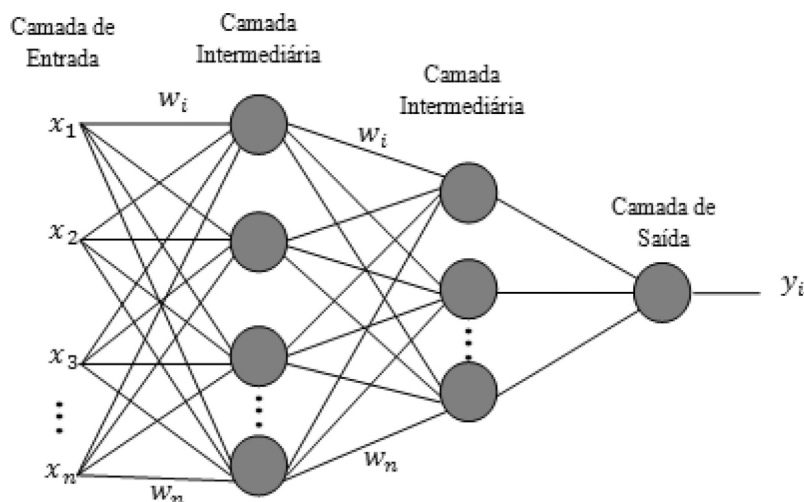
```

 $h^0 \leftarrow X$ 
for  $l \leftarrow 1$  to  $n$ layers: do
     $z^l \leftarrow W^l h^{l-1} + b^l$ 
     $h^l \leftarrow g^l(z^l)$ 
end for
 $y = h^n$ 
 $J = L(y, y) + \lambda \Omega(\theta)$ 

```

---

Figura 8: Rede Neural MLP com duas camadas ocultas.



A fase *backwards* do treinamento de RNAs MLP é realizado pelo algoritmo de backpropagation se refere apenas ao método de cálculo do gradiente, enquanto outro algoritmo, chamado gradiente descendente estocástico, é utilizado para realizar o aprendizado utilizando o gradiente *backpropagation*, que permite que a informação do custo então flua para trás, na direção contrária. Nesta fase há o processo de ajuste dos pesos dos neurônios para minimizar a função custo das saídas previstas pela rede e o valor alvo. O algoritmo de backpropagation é utilizado para computar os gradientes de funções. Para funções custo desta forma, o gradiente descendente é computado de acordo com a Equação 6. . Backpropagation é um algoritmo que computa a regra da cadeia, com uma ordem específica de operações que é altamente eficiente. O algoritmo de backpropagation consiste de computar o produto gradiente jacobiano para cada operação na rede neural. O gradiente é um vetor que indica o sentido e a direção na qual, por deslocamento a partir de um ponto especificado, obtem-se o maior incremento possível de uma grandeza a partir do qual se define um campo escalar para o espaço em consideração. e o jacobiano é a matriz de derivadas parciais de primeira ordem de uma função vetorial . Em uma rede neural, as entradas X, pesos W, bias b e saída Y são todos dados por vetores.

citar origem dos conceitos de jacobiano e gradiente

---

```

 $g \leftarrow \nabla_y J = \nabla_y L(y, y)$ 
for l = nlayers to 1 do
     $g \leftarrow \nabla_{z^l} J = g \times \sigma'(z^l)$ 
     $\nabla_{b^l} J = g + \lambda \nabla \Omega(\theta)$ 
     $\nabla_{W^l} J = g h^{(l-1)} + \lambda \nabla \Omega(\theta)$ 
     $g \leftarrow \nabla_{h^{l-1}} J = W^l g$ 
end for

```

---

A derivada  $f'$  de uma certa função  $y = f(x)$  é dada conforme Equação ??, a qual fornece a inclinação de  $f(x)$  no ponto  $x$ . Neste contexto,  $f$  é a função que a rede neural está aprendendo a mapear. Aplicada à uma função de custo, esta operação especifica como escalar uma pequena mudança nos pesos  $w$  aplicados à entrada  $x$  para obter uma mudança correspondente na saída  $y$ . A técnica de realizar pequenos incrementos na entrada  $w$  no valor oposto ao da derivada é chamada de *gradiente descendente*.

Sugiro citar o Haykin

Citação!!



$$\nabla_{\theta} J(\theta) = \frac{1}{m} \sum_{i=1}^m \nabla_{\theta} L(x^{(i)}, y^{(i)}, \theta) \quad (6)$$

O custo operacional de computar este gradiente é  $O(m)$ , sendo  $m$  o número de exemplos no conjunto de treinamento. Assim, o custo computacional cresce de maneira proporcional e linear à quantidade de exemplos presente no conjunto de treinamento.

Em termos do número de nós na rede neural, o algoritmo de backpropagation tem custo  $O(n)$ .

As RNAs *feedforward* MLP são amplamente utilizadas em aplicações de diversos domínios. Inicialmente, destacaram-se as aplicações voltadas para o mercado financeiro visando, por exemplo, otimizar estratégias de marketing. Aplicações posteriores consideraram a alocação de assentos em aviões, aprovação de empréstimo, controle de qualidade em processos industriais, dentre outros (WIDROW; RUMELHART; LEHR, 1994). O escopo de aplicações deste modelo continua a crescer nos dias atuais, especialmente diante do desenvolvimento de variantes, a exemplo das redes neurais convolucionais, com grande capacidade de detecção de padrões e pouco esforço de pré-processamento. Reconhecimento de caracteres e dígitos (LECUN et al., 1998), processamento de imagens médicas para reconhecimento de características associadas à doenças cardíacas (OKTAY et al., 2018), pulmonares (GAO et al., 2018) e mamárias (DUBROVINA et al., 2018) são alguns exemplos de aplicações de vanguarda destes modelos, que são compreendidos dentro da sub-área de *Deep Learning*, caracterizada na seção a seguir.

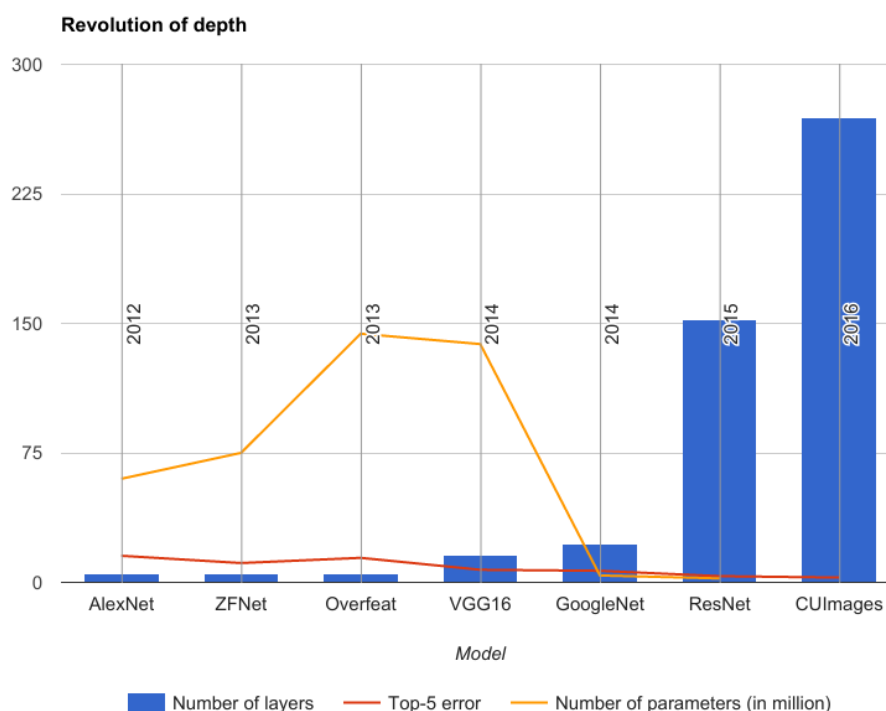
## 2.5. Deep Learning

*Deep Learning* (DL), também conhecido como Aprendizagem Profunda, compreende um conjunto de técnicas de ML que podem ser aplicadas em problemas de aprendizado supervisionado e não-supervisionado. A principal característica dos modelos neste domínio é a capacidade de representar e reconhecer características sucessivamente complexas, por meio da adição de níveis ou camadas de operações não-lineares em sua arquiteturas, a exemplo das nas redes neurais profundas, máquinas de Boltzmann profundas e fórmulas proposicionais. Modelos deste tipo ganharam popularidade ao se mostraram capazes de resolver problemas complexos com um desempenho cada vez maior (BENGIO et al., 2009).

A melhoria do desempenho de modelos de DL é decorrente do aumento recente da quantidade de dados disponíveis sobre temas complexos, aliado ao aumento da disponibilidade de recursos computacionais para executar modelos mais robustos (GOODFELLOW; BENGIO; COURVILLE, 2016; DENG; YU et al., 2014). Alguns dados fornecidos pela IBM reforçam esta afirmação: em 2017 foram gerados 2,5 quintilhões de bytes de dados por dia, e 90% do volume total de dados gerados até 2017 no mundo foi criado somente nos últimos dois anos (IBM, 2017).

Para exemplificar o efeito da adição de camadas aos modelos de DL, a Figura 9 mostra uma visão geral do aumento da profundidade das camadas nas redes neurais profundas e o desempenho destas em problemas de detecção de objetos em imagens. Nota-se que, à medida que a profundidade aumenta, há uma diminuição no erro. Mais recentemente, isto também têm implicado na redução do número de parâmetros treináveis. Este panorama reforça a hipótese de que a profundidade das camadas impacta positivamente na captura de características e que estes avanços têm tornado as tarefas mais factíveis, com uma diminuição do esforço computacional associado .

Figura 9: Evolução de profundidade, taxa de erro e número de parâmetros das redes neurais profundas com o passar dos anos. Fonte: (ECARLAT, 2017).



### 2.5.1. Breve Histórico

O termo *Deep Learning* não é recente, foi utilizado pela primeira vez por Dechter, no contexto da descoberta de todas as configurações de conflitos mínimas a fim de resolver um problema de satisfação de restrições (DECHTER, 1986). Porém, ganhou força a partir de pesquisas sobre RNAs *feedforward* com muitas camadas ocultas, também conhecidas por redes neurais profundas (DENG; YU et al., 2014).

Considera-se que o desenvolvimento de DL pode ser entendido em três partes. Na primeira, houve a proposição de modelos lineares simples, compostos apenas por um neurônio, a exemplo dos neurônios de McCulloch e Pitts (MCCULLOCH; PITTS, 1943) e *Perceptron* de Rosenblatt (ROSENBLATT, 1958). A segunda parte, iniciada nos anos 1980, teve como eixo central a interconexão entre vários neurônios e a proposição do algoritmo *back-propagation* para ajuste de pesos no treinamento das RNAs (RUMELHART; MCCLELLAND, 1986; RUMELHART; HINTON; WILLIAMS, 1986). Com estas contribuições, houve muita aplicação das RNAs em diversos domínios. Ainda no final deste segundo momento, duas contribuições relevantes foram feitas: os modelos *Long Short-Term Memory* (LSTM) e LeNet (LECUN et al., 1998).

Complementar com importância da LeNet.

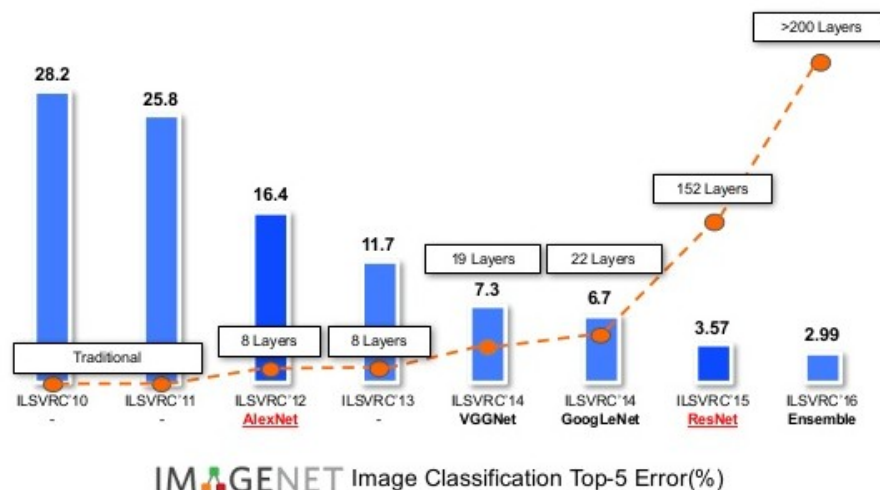
A terceira fase tem um marco inicial mais definido: compreende o ano de 2006 e a publicação de um artigo por Hinton et al. apresentando as *deep belief networks* (HINTON; OSINDERO; TEH, 2006). Neste tipo de RNA, o aprendizado é realizado de forma não-supervisionada e as camadas que compõem a rede atuam como reconhecedoras de características. A partir deste momento, a utilização de DL se popularizou.

Na conjectura atual, modelos de DL têm superado significativamente o estado da

arte de modelos inteligentes em diversas competições em todo o mundo. A *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC) é uma competição em que equipes de pesquisa avaliam seus algoritmos em um conjunto de dados fornecido, e competem para chegar à melhor acurácia em várias tarefas de reconhecimento visual automático. Em 2011, os melhores resultados de classificação no ILSVRC tinham por volta de 25% de erro nas tarefas propostas. Em 2012, o modelo AlexNet, uma rede neural convolucional proposta segundo as ideias de DL, atingiu apenas 16,4% de erro, propondo um ganho dificilmente visto entre duas edições sucessivas da competição.

O gráfico da Figura 9 sintetiza o histórico da competição ILSVRC, em que a partir do ano de 2012 houve a introdução de modelos baseados em DL. O histograma mostra a diminuição do erro na tarefa de aprendizado proposta e a linha laranja enfatiza o número de camadas ocultas utilizadas nos modelos vencedores.

Figura 10: Evolução do erro dos modelos vencedores da competição ILSVRC pela profundidade das redes neurais (BORTH, 2017)



Apesar do foco inicial de DL ter sido concentrado no desenvolvimento de técnicas de aprendizado não-supervisionado e na habilidade de modelos profundos de boa generalização a partir de conjuntos de dados pequenos, o cenário atual das pesquisas nesta área consideram o uso de técnicas de aprendizado supervisionado bem mais antigas, visando o endereçamento de conjuntos de dados massivos e categorizados. Nesta perspectiva encontram-se as redes neurais convolucionais com múltiplas camadas, que impulsionaram os avanços recentes na área de Visão Computacional. Considerando esta importância, a seção a seguir compreenderá a explanação destes modelos, com especial para arquiteturas canônicas de maior destaque nos últimos anos.

Citar aqui

idem, citar.

## 2.5.2. Redes Neurais Convolucionais

*Redes Neurais Convolucionais* (CNN, do inglês *Convolutional Neural Networks*) são uma classe de redes neurais *feedforward* com topologia bem definida e estruturada em uma grade, com o uso de operações de convolução em pelo menos uma de suas camadas (GOODFELLOW; BENGIO; COURVILLE, 2016). Aplicadas em tarefas de classificação,

regressão, localização, detecção e outras, este tipo de modelo se destaca no reconhecimento de padrões em dados de alta dimensionalidade, a exemplo de séries temporais, imagens e vídeos (KHAN et al., 2018).

A operação de convolução possui um papel central nas CNNs. Esta operação descreve a média ponderada de uma determinada função  $x_1(t)$  sob um intervalo fixo de uma variável, enquanto os pesos da média ponderada considerada pertencem à função  $x_2(t)$  amostrados em intervalos  $a$  (BRACEWELL; BRACEWELL, 1986). Assim, a convolução  $s(t)$  de duas funções  $x_1(t)$  e  $x_2(t)$  é uma função  $s : \mathbb{Z} \rightarrow \mathbb{R}$ , denotada  $s(t) = x_1(t) * x_2(t)$ , e definida conforme Equação 7 (LATHI, 2006):

$$s(t) = x_1(t) * x_2(t) = \int_{-\infty}^{\infty} x_1(a)x_2(t-a)da. \quad (7)$$

No contexto de ML, a função  $x_1(t)$  é chamada de *input*, a função  $x_2(t)$  é o *kernel*, e a saída  $s(t)$  consiste no *feature map*, ou mapa de características. No contexto prático, o *input* normalmente é um vetor multidimensional de dados e o *kernel* é um vetor multidimensional de pesos que devem ser ajustados para aprendizado das CNN. Considerando, por exemplo, uma imagem  $I$  de dimensões  $(m, n)$  como *input* e a aplicação de um *kernel*  $K$ , a versão discreta da convolução, passível de implementação computacional e equivalente à Eq. 7, é mostrada na Eq. 8:

$$S(i, j) = I(i, j) * K(i, j) = \sum_m \sum_n I(m, n)K(i-m, j-n), \quad (8)$$

em que  $S$  é o *feature map* resultante e  $(i, j)$  é a posição correspondente nesse mapa. Para otimizar os aspectos de implementação, os valores resultantes da operação de convolução são armazenados apenas nas posições  $(i, j)$  explicitamente declaradas (GOODFELLOW; BENGIO; COURVILLE, 2016).

Os *feature maps*, resultantes das operações de convolução, compreendem a noção de filtros, responsáveis por capturarem características relativas à entrada, tais como contornos, linhas, texturas, etc. Quando combinados de maneira sequencial, como proposto pelas CNNs, as características capturadas pelas camadas convolucionais vão se tornando mais complexas à medida que se aumenta a profundidade da rede. Assim, um primeiro *feature map* de uma camada convolucional captura um simples contorno, enquanto um *feature map* em uma camada mais profunda da rede pode capturar uma forma, um rosto ou até um objeto inteiro (BUDUMA, 2017). Esta noção é ilustrada na Figura 11.

Figura 11: Papel das camadas convolucionais e *feature maps* nas CNNs. Fonte: (KHAN et al., 2018).

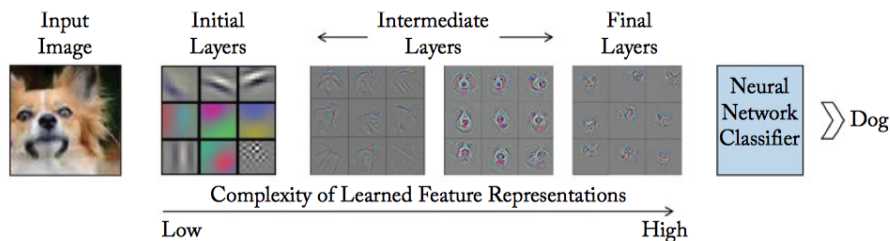
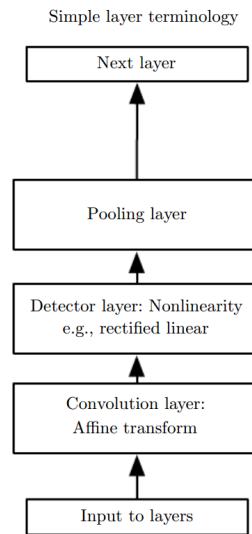


Figura 12: Componentes de uma camada de uma rede neural convolucional (GOODFELLOW; BENGIO; COURVILLE, 2016).



As camadas convolucionais, que contém os *feature maps* e os pesos da rede, normalmente são seguidas por funções de ativação não-linear.

A nonlinear function can also be understood as a switching or a selection mechanism, which decides whether a neuron will receive or not given all of its inputs. The activation functions that are commonly used in deep networks are differentiable to enable error back propagation

A toda camada convolucional em uma CNN, segue-se uma ativação não-linear, finalizando em uma operação de *pooling*, como mostra a Figura 12. A seguir, serão explanadas cada uma destas etapas.

### 2.5.2.1. Convolução

A operação de convolução

Quando a operação de convolução é

A convolução é comutativa, ou seja, as Equações 9 e 8 são equivalentes, salvo que no primeiro caso há a convolução da imagem pelo núcleo, enquanto no segundo há a convolução do núcleo pela imagem. Comumente, a Equação 8 é a implementada em algoritmos de redes neurais convolucionais, haja visto que existem menor variação no intervalo de valores válidos de  $m$  e  $n$ , o que diminui o custo computacional.

$$S(i, j) = K(i, j) * I(i, j) = \sum_m \sum_n I(i - m, j - n) K(m, n) \quad (9)$$

### 2.5.2.2. Ativação

### 2.5.2.3. Pooling

Depois de realizar várias operações de convolução em paralelo para gerar um conjunto

de ativações lineares e alimentá-las a funções de ativação não-lineares, como *ReLU*, *Soft-max*, etc, na chamada etapa de detecção, chega-se à etapa de *pooling*. Uma função de *pooling* substitui a saída da rede em determinada localização por uma síntese estatística das saídas vizinhas. Por exemplo, a função *max pooling* retorna o valor máximo em uma área retangular, enquanto a *average pooling* retorna a média das saídas de um retângulo.

### 2.5.3. Modelos Canônicos de Redes Neurais Convolucionais para Detecção de Objetos em Imagens

## 3. Trabalhos Relacionados

A proposta apresentada está relacionada com inúmeros trabalhos envolvendo a aplicação de redes neurais convolucionais e outros modelos de *machine learning* para a estimação de idade de indivíduos.

Segundo (FU; GUO; HUANG, 2010), a idade pode ser inferida a partir de padrões distintos que emergem através da aparência da face. Técnicas comuns para a estimação da idade envolvem a dedução de modelos matemáticos a partir do estudo do crescimento de medidas da face e do crânio (KWON; LOBO, 1999), da textura do rosto (LANITIS; TAYLOR; COOTES, 2002), da captura de tendências de envelhecimento a partir de várias imagens de indivíduos de mesma idade (FU; XU; HUANG, 2007) e a extração de características específicas relacionadas à idade (SUO et al., 2008), (LOU et al., 2018). Modelos de *machine learning* também são utilizados para a tarefa, em especial as redes neurais artificiais, K-vizinhos mais próximos e máquinas de vetores de suporte.

Recentemente, a aplicação de redes neurais convolucionais em problemas de classificação e detecção de objetos em imagens têm obtido resultados significativamente positivos. Em (SIMONYAN; ZISSERMAN, 2014), (HE et al., 2016), (SZEGEDY et al., 2015), (REDMON et al., 2016), (LIU et al., 2016) e outros, são descritas arquiteturas robustas capazes de detectar dezenas de objetos em várias situações. Treinadas com conjuntos de dados visuais que contam com milhares de exemplos como a ImageNet, Pascal VOC e COCO, estas redes são conhecidas por seu bom desempenho. Algumas destas redes foram afinadas utilizando conjuntos de dados menores e especializados para a tarefa de estimação de idade.

O trabalho de (ROTHER; TIMOFTE; GOOL, 2015) relata um método para estimação de idade aparente em imagens de faces imóveis utilizando *deep learning*. Propõe-se um conjunto de 20 redes neurais convolucionais classificadoras com arquiteturas VGG-16 pré-treinadas com a base de dados visuais ImageNet, e ajustadas utilizando imagens disponibilizadas pelo IMDB, Wikipedia, e o conjunto de dados *Looking At People*–LAP para anotação de idade aparente. Cada modelo tem como saída um número discreto entre 0 e 100, representando a idade prevista. A saída final do modelo consiste na média entre as idades previstas pelos 20 redes. A solução atingiu um MAE (*Mean Average Error*) de 3.221 na fase de testes.

Em (LIU et al., 2015) cria-se um estimador de idade composto pela fusão de um modelo regressor e outro classificador. Realiza-se um pré-processamento da entrada, que envolve a detecção das faces presentes na imagem, seguida pela etapa de localização de pontos de referência, como olhos, nariz e boca, e por fim há a normalização da face. Dois métodos de normalização de face são testados, a normalização exterior e interior. Após este pré-processamento, as imagens resultantes são alimentadas a modelos de redes

neurais convolucionais profundas inspiradas na *GoogLeNet* (SZEGEDY et al., 2015). O modelo sofreu modificações em sua arquitetura, como adição de normalização do batch, remoção de camadas de *dropout* e perda. Foram treinados e testados diversos modelos com variações no tipo de normalização da face, tamanho do corte dos rostos, tipo de tarefa preditiva, etc. Os modelos resultantes destas variações foram unidos em um conjunto, que conseguiu prever idades com MAE de 3.3345.

Ademais, é possível encontrar resultados satisfatórios para a tarefa de aprendizado proposta utilizando modelos menos complexos. Com o objetivo de consolidar um método de classificação de idade e gênero, (LEVI; HASSNER, 2015) propõe uma rede neural convolucional de natureza mais simples, se comparada com (SZEGEDY et al., 2015), (SIMONYAN; ZISSERMAN, 2014) ou (HE et al., 2016). Sua arquitetura consiste em três camadas convolucionais com *dropout* e funções de ativação *ReLU*, seguidas por três camadas totalmente conectadas. A camada de saída tem como função de ativação a Softmax. A escolha por um design de rede menor é motivado pelo desejo de reduzir o risco de *overfitting* e pela natureza do problema, que contém apenas 8 classes de idade. O modelo é treinado utilizando apenas o conjunto de referência *Adience*, composto por imagens não filtradas para classificação de idade e gênero. Considerando uma margem de erro de uma classe vizinha, a melhor rede obteve acurácia de  $84.7\% \pm 2.2$  ao empregar a técnica de sobre-amostragem.

## 4. Solução Proposta

### 4.1. Tarefa de Previsão Considerada

### 4.2. Elaboração e Descrição da Base de Dados

O conjunto de dados considerado é o

colocar infos do imdb wiki dataset

Utilizou-se somente a base de dados do IMDB, que conta com 452132 exemplares, tendo em vista a quantidade de imagens do conjunto WIKI que não poderiam ser abertas. Extraíu-se a idade das celebridades a partir da data de nascimento e do ano em que a foto foi tirada. A data de nascimento estava no formato *datenum*, que representa o intervalo em dias entre 0 de Janeiro de 0000 e o momento presente na coleta da data, ou seja, o nascimento da celebridade. Fez-se necessária a conversão desta medida para a data comum, de onde foi extraído o ano de nascimento. A seguir, calculou-se a idade a partir da diferença entre o ano de nascimento e o da foto.

Para a predição de idade, as informações de gênero, localização do rosto e dia de nascimento no formato *datenum* foram descartadas, assim como os exemplos que continham idade nula. A seguir, foram descartadas as imagens que continham mais de uma face detectada. O conjunto de dados resultante continha um total de 181634 exemplos de 14624 celebridades diferentes.

### 4.3. Modelos de CNN Considerados

### 4.4. Parâmetros e Hiperparâmetros

### 4.5. Métricas de Desempenho

### 4.6. Etapa de Treinamento

### 4.7. Etapa de Testes

## 5. Considerações Finais

## Referências

BENGIO, Y. et al. Learning deep architectures for ai. *Foundations and trends® in*



*Machine Learning*, Now Publishers, Inc., v. 2, n. 1, p. 1–127, 2009.

BETWEEN, D. *Difference between Smart TV and Normal TV*. 2017. <<http://www.differencebetween.info/difference-between-smart-tv-and-normal-tv>>. Acessado em 21 de Março de 2018.

BORTH, D. D. *Deep Learning – Future of AI*. [S.l.]: SlideShare, 2017. <<https://www.slideshare.net/GroupeT2i/deep-learning-the-future-of-ai>>. Acessado em 23 de Abril de 2018.

BRACEWELL, R. N.; BRACEWELL, R. N. *The Fourier transform and its applications*. [S.l.]: McGraw-Hill New York, 1986. v. 31999.

BRAGA, A. de P.; CARVALHO, A. P. de Leon F. de; LUDERMIR, T. B. *Redes Neurais Artificiais: Teoria e Aplicações*. 2. ed. Rio de Janeiro: LTC, 2007.

BRAZILIENSE, C. *Copa e novas tecnologias prometem aumentar venda de TVs no Brasil em 2018*. 2018. <[http://www.correiobraziliense.com.br/app/noticia/economia/2018/01/23/internas\\_economia,654966/copa-e-novas-tecnologias-prometem-aumentar-venda-de-tvs-no-brasil.shtml](http://www.correiobraziliense.com.br/app/noticia/economia/2018/01/23/internas_economia,654966/copa-e-novas-tecnologias-prometem-aumentar-venda-de-tvs-no-brasil.shtml)>. Acessado em 21 de Março de 2018.

BUDUMA, N. *Fundamentals of Deep Learning*. Estados Unidos: Editora O'Reilly, 2017.

CAPELAS, B. *Explosão no consumo de vídeos online coloca em xeque o futuro da televisão*. 2017. O Estado de S. Paulo. Acessado em 20 de Março de 2018. Disponível em: <<http://link.estadao.com.br/noticias/geral,explosao-no-consumo-de-videos-online-coloca-em-xeque-o-futuro-da-televisao,70001695828>>.

CIRIACO, D. *Os melhores serviços de streaming de vídeo disponíveis no Brasil*. <<https://canaltech.com.br/internet/os-melhores-servicos-de-streaming-de-video-disponiveis-no-brasil/>>. Acessado em 20 de Março de 2018.

DECHTER, R. *Learning while searching in constraint-satisfaction problems*. [S.l.]: University of California, Computer Science Department, Cognitive Systems Laboratory, 1986.

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

DEPUTADOS, C. dos. *Estatuto da Criança e do Adolescente*. BRASIL: [s.n.], 1995.

DUBROVINA, A. et al. Computational mammography using deep neural networks. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, Taylor & Francis, v. 6, n. 3, p. 243–247, 2018. Disponível em: <<https://doi.org/10.1080/21681163.2015.1131197>>.

ECARLAT, P. *CNN – Do we need to go deeper?* [S.l.]: Medium, 2017. <<https://medium.com/finc-engineering/cnn-do-we-need-to-go-deeper-afe1041e263e>>. Acessado em 23 de Abril de 2018.

FLACH, P. *Machine learning: the art and science of algorithms that make sense of data*. [S.l.]: Cambridge University Press, 2012.



FU, Y.; GUO, G.; HUANG, T. S. Age synthesis and estimation via faces: A survey. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 32, n. 11, p. 1955–1976, 2010.

FU, Y.; XU, Y.; HUANG, T. S. Estimating human age by manifold analysis of face pictures and regression on aging features. In: IEEE. *Multimedia and Expo, 2007 IEEE International Conference on*. [S.l.], 2007. p. 1383–1386.

GAO, M. et al. Holistic classification of ct attenuation patterns for interstitial lung diseases via deep convolutional neural networks. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, Taylor & Francis, v. 6, n. 1, p. 1–6, 2018. Disponível em: <<https://doi.org/10.1080/21681163.2015.1124249>>.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep learning*. [S.l.]: MIT press Cambridge, 2016. v. 1.

GUIMARÃES, N. *Com fim do sinal analógico, busca por smart TVs cresce 11%*. 2017. <<http://www.leiaja.com/tecnologia/2017/07/17/com-fim-do-sinal-analogico-busca-por-smart-tvs-cresce-11/>>. Acessado em 22 de Março de 2018.

HAYKIN, S. S. *Neural networks and learning machines*. [S.l.]: Pearson Upper Saddle River, NJ, USA:, 2009. v. 3.

HE, K. et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 770–778.

HINTON, G. E.; OSINDERO, S.; TEH, Y.-W. A fast learning algorithm for deep belief nets. *Neural computation*, MIT Press, v. 18, n. 7, p. 1527–1554, 2006.

HORNIK, K. Approximation capabilities of multilayer feedforward networks. *Neural networks*, Elsevier, v. 4, n. 2, p. 251–257, 1991.

IBGE. *Pesquisa Nacional por Amostra de Domicílios: Acesso à Internet e à Televisão e Posse de Telefone Móvel Celular para Uso Pessoal*. 2015. <<https://biblioteca.ibge.gov.br/visualizacao/livros/liv99054.pdf>>. Acessado em 16 de Março de 2018.

IBM, M. C. *10 Key Marketing Trends for 2017 and Ideas for Exceeding Customer Expectations*. 2017. <<https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=WRL12345USEN>>. Acessado em 23 de Março de 2018.

JUSTIÇA, M. da. *Política Pública de Classificação Indicativa*. BRASIL: [s.n.], 2014.

JUSTIÇA, S. N. de. *Classificação Indicativa Guia Prático*. BRASIL: [s.n.], 2012.

KHAN, S. et al. *A Guide to Convolutional Neural Networks for Computer Vision*. Austrália: Morgan & Claypool, 2018.

KOVACH, S. *What Is A Smart TV?* 2010. <<http://www.businessinsider.com/what-is-a-smart-tv-2010-12>>. Acessado em 15 de Março de 2018.

KWON, Y. H.; LOBO, N. da V. Age classification from facial images. *Computer vision and image understanding*, Elsevier, v. 74, n. 1, p. 1–21, 1999.

LANITIS, A.; TAYLOR, C. J.; COOTES, T. F. Toward automatic simulation of aging effects on face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 24, n. 4, p. 442–455, 2002.

- LATHI, B. P. *Sinais e Sistemas Lineares-2*. [S.l.]: Bookman, 2006.
- LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, IEEE, v. 86, n. 11, p. 2278–2324, 1998.
- LEVI, G.; HASSNER, T. Age and gender classification using convolutional neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. [S.l.: s.n.], 2015. p. 34–42.
- LIU, W. et al. Ssd: Single shot multibox detector. In: SPRINGER. *European conference on computer vision*. [S.l.], 2016. p. 21–37.
- LIU, X. et al. Agetnet: Deeply learned regressor and classifier for robust apparent age estimation. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. [S.l.: s.n.], 2015. p. 16–24.
- LOU, Z. et al. Expression-invariant age estimation using structured learning. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 40, n. 2, p. 365–375, 2018.
- MARSLAND, S. *Machine learning: an algorithmic perspective*. [S.l.]: CRC press, 2015.
- MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, Springer, v. 5, n. 4, p. 115–133, 1943.
- MICHÉLE, B.; KARPOW, A. Watch and be watched: Compromising all smart tv generations. In: IEEE. *Consumer Communications and Networking Conference (CCNC), 2014 IEEE 11th*. [S.l.], 2014. p. 351–356.
- MITCHELL, T. *Machine Learning*. McGraw-Hill Education, 1997. (McGraw-Hill international editions - computer science series). ISBN 9780070428072. Disponível em: <<https://books.google.com.br/books?id=xOGAngEACAAJ>>.
- NEWSROOM, S. *Smart TV: Piece by Piece*. 2011. <<https://news.samsung.com/global/smart-tv-piece-by-piece>>. Acessado em 15 de Março de 2018.
- OKTAY, O. et al. Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation. *IEEE transactions on medical imaging*, IEEE, v. 37, n. 2, p. 384–395, 2018.
- PERAKAKIS, E.; GHINEA, G. A proposed model for cross-platform web 3d applications on smart tv systems. In: ACM. *Proceedings of the 20th International Conference on 3D Web Technology*. [S.l.], 2015. p. 165–166.
- QUAIN, J. R. *Smart TVs: Everything You Need to Know*. 2018. <<https://www.tomsguide.com/us/smart-tv-faq,review-2111.html>>. Acessado em 23 de Março de 2018.
- REDMON, J. et al. You only look once: Unified, real-time object detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 779–788.
- ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, American Psychological Association, v. 65, n. 6, p. 386, 1958.
- ROTHER, R.; TIMOFTE, R.; GOOL, L. V. Dex: Deep expectation of apparent age from a single image. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. [S.l.: s.n.], 2015. p. 10–15.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. *nature*, Nature Publishing Group, v. 323, n. 6088, p. 533, 1986.

RUMELHART, D. E.; MCCLELLAND, J. L. Parallel distribution processing: exploration in the microstructure of cognition. *MA: MIT Press, Cambridge*, 1986.

RUSSELL, S. J.; NORVIG, P. *Artificial intelligence: a modern approach*. [S.l.]: Malaysia; Pearson Education Limited,, 2016.

SBT. *Smart TV – TV Conectada*. 2015. <<http://www.sbt.com.br/tvconectada/>>. Acessado em 23 de Março de 2018.

SCHOFIELD, J. *How can I make video calls from my TV set?* 2017. <<https://goo.gl/eCynUh>>. Acessado em 28 de maio de 2018.

SHIN, D.-H.; HWANG, Y.; CHOO, H. Smart tv: are they really smart in interacting with people? understanding the interactivity of korean smart tv. *Behaviour & information technology*, Taylor & Francis, v. 32, n. 2, p. 156–172, 2013.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

SUO, J. et al. Design sparse features for age estimation using hierarchical face model. In: *IEEE. Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. [S.l.], 2008. p. 1–6.

SZEGEDY, C. et al. Going deeper with convolutions. In: *CVPR*. [S.l.], 2015.

WIDROW, B.; RUMELHART, D. E.; LEHR, M. A. Neural networks: applications in industry, business and science. *Communications of the ACM*, ACM, v. 37, n. 3, p. 93–105, 1994.

WIKIPEDIA. *Television content rating system*. 2018. <[https://en.wikipedia.org/wiki/Television\\_content\\_rating\\_system#Countries\\_without\\_TV\\_rating\\_systems](https://en.wikipedia.org/wiki/Television_content_rating_system#Countries_without_TV_rating_systems)>. Acessado em 21 de Março de 2018.