

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

Nicoli P. Araújo, Elloá B. Guedes

¹ 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:

- Formular um referencial teórico sobre redes neurais convolucionais, contemplando seu arcabouço matemático, suas características, principais arquiteturas, métodos de treinamento e teste;
- Consolidar uma base de dados com exemplos realísticos para treinamento dos modelos, tendo em vista a captura de padrões representativos ao domínio do problema;
- Identificar tecnologias adequadas para implementação dos estimadores;
- Propor, treinar e testar diferentes estimadores de idade baseados em redes neurais convolucionais para a tarefa em questão;
- 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	
Escrita da Proposta	X	X	X	X	X							
Fundamentação Teórica sobre Machine Learning	X	X	X	X								
Consolidação da Base de Dados		X	X									
Proposição de Modelos de Redes Neurais Convolucionais				X	X	X	X	X				
Defesa da Proposta					X							
Escrita do Trabalho Final						X	X	X	X	X	X	
Treinamento das Redes Neurais Convolucionais					X	X	X	X	X	X		
Teste das Redes Neurais Convolucionais					X	X	X	X	X	X	X	
Comparação de Mettricas de Desempenho						X	X	X	X	X	X	
Defesa do Trabalho Final												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).

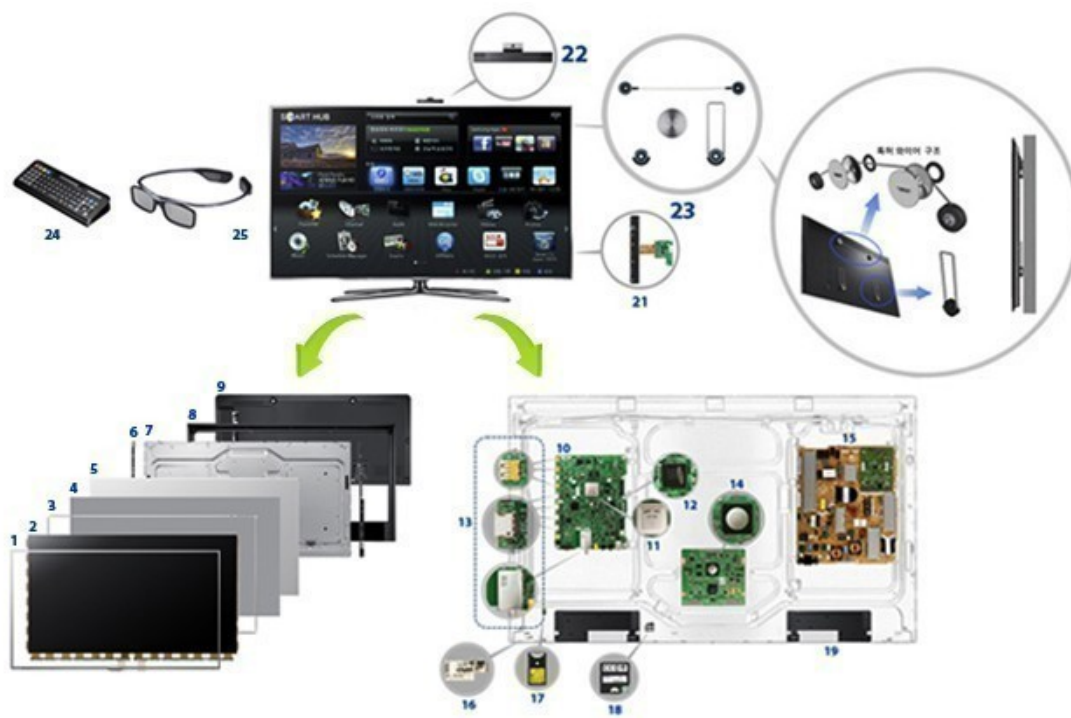


Figura 1: Diagrama representativo de uma *Smart TV* e seus componentes (NEWSROOM, 2011). Ver legenda dos componentes na Tabela 2.

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 que tem este

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

Categoria	Símbolo	Descrição do Conteúdo
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; glamourização 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.

comportamento é aquele capaz de aprender a partir de dados, assim como humanos e outros animais. Estes, 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).

De maneira análoga ao aprendizado natural, os algoritmos de *machine learning* precisam aprender, processo chamado de aquisição da experiência. De acordo com a definição clássica de (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 inferir um algoritmo de *machine learning* para desenvolver determinada tarefa, busca-se um modelo, ou seja, uma função, que mapeie as instâncias do espaço de en-

trada para o de saída (FLACH, 2012). Estes modelos podem ser agrupados em diferentes categorias ao se considerar o tipo de aprendizado e de saída desejada para o algoritmo. Na Figura 3 está uma visão geral dos modelos de algoritmos de *machine learning* e suas subdivisões.

Quanto ao tipo de aprendizado, as tarefas de *machine learning* 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 dados rotulados, ou seja, que têm seus valores de saída conhecidos, apresentados na fase de treinamento, a exemplo das 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 tabelados, a exemplo de modelos como *k-means*, redes neurais artificiais profundas de codificação preditivas e detecção de anomalia. Por fim, no aprendizado por reforço o algoritmo não recebe dados ou 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 desejado, os problemas podem ser atacados são a classificação, regressão, transcrição, tradução automática, detecção de anomalia, síntese e amostragem. As principais tarefas que podem ser endereçadas utilizando aprendizado supervisionado são a classificação e a 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 y do meio de k possibilidades (GOODFELLOW; BENGIO; COURVILLE, 2016). Dentre as tarefas de classificação estão 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. Uma tarefa de regressão envolve 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 podem ser paramétricos ou não paramétricos. Segundo (RUSSELL; NORVIG, 2016), um modelo de aprendizado que resume dados utilizando um conjunto de parâmetros de tamanho definidos independente do número de exemplis de treinamento é chamado de *modelo paramétrico*. Dentre os modelos paramétricos estão a regressão linear e as redes neurais artificiais. Já um *modelo não-paramétrico* 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, k-NN e árvores de decisão CART e C4.5.

Dentre os modelos paramétricos, as redes neurais artificiais (RNAs) têm demonstrado resultados satisfatórios em tarefas de classificação e regressão, em aplicações em diversas áreas. Em especial, aplicações de *Deep Learning* (DL) no reconhecimento de objetos, no processamento de linguagens naturais e *speech recognition* têm trazido ainda mais atenção ao modelo.

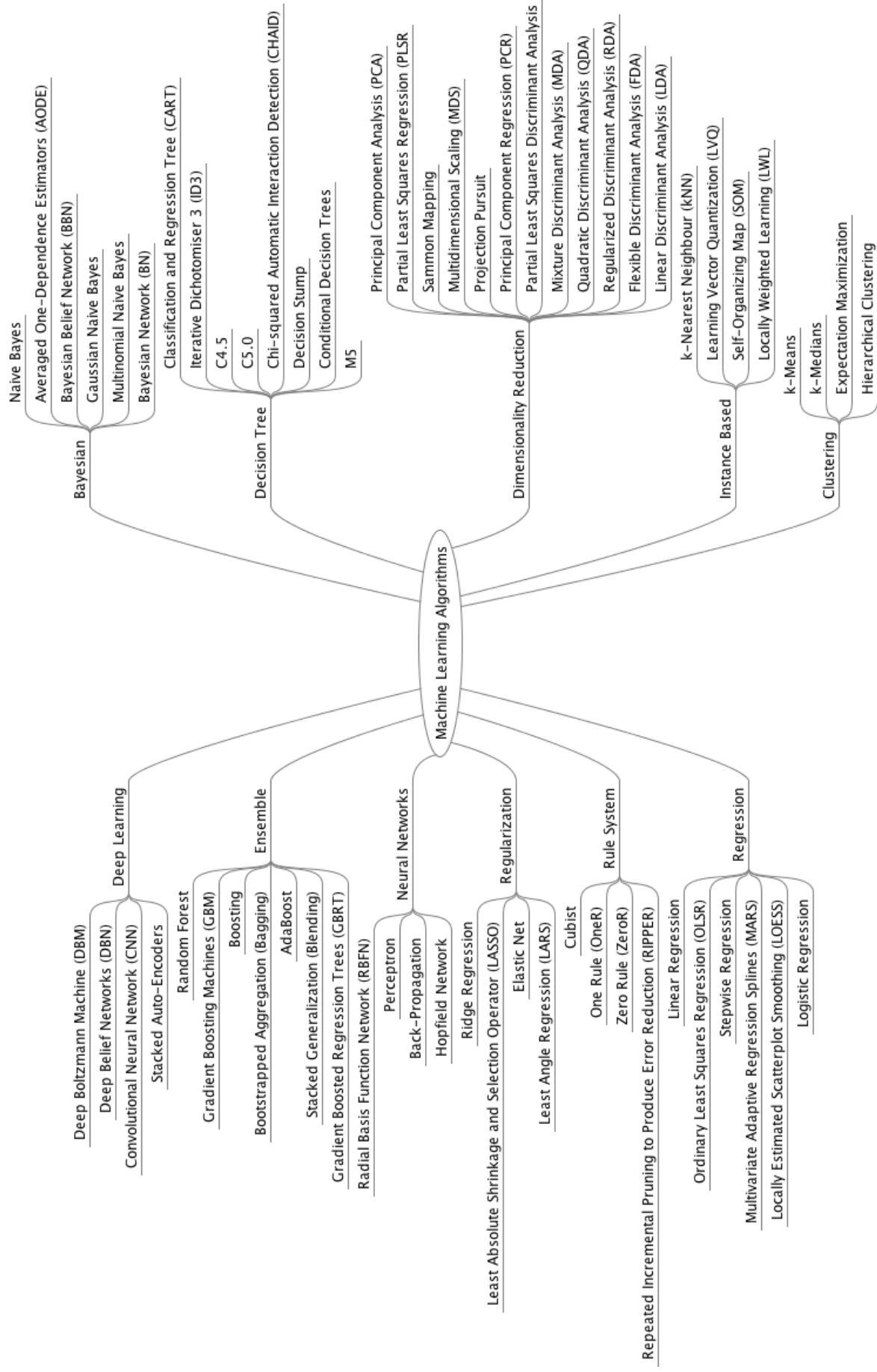


Figura 3: Mapa mental dos algoritmos de *Machine Learning* organizados por área e sub-área.

2.4. Redes Neurais Artificiais

Redes Neurais Artificiais (RNAs) são um modelo de computação não algorítmica 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, 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.

A motivação para a criação deste modelo vem do funcionamento do cérebro biológico, que é formado por neurônios interligados 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, além do . Cada neurônio é composto de um corpo, dendritos e um axônio, como é mostrado na Figura 4. 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 está expressa na Figura 5.

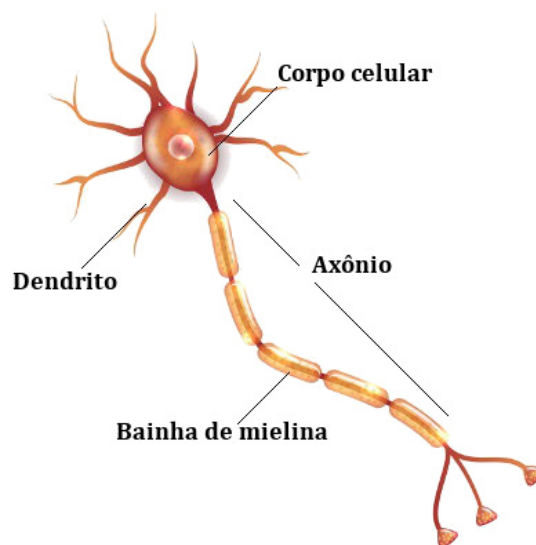


Figura 4: Neurônio biológico e seus componentes: corpo, axônio e dendritos.

Com base neste modelo biológico, McCulloch e Pitts propuseram em (MCCULLOCH; PITTS, 1943) um neurônio artificial. Explorado na Figura 6, o modelo de McCulloch e Pitts é formado por somente um neurônio artificial que contém n terminais de entrada dada 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 através de uma função de ativação $y = g(z)$ 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 maior ou igual a um limiar de ativação θ .

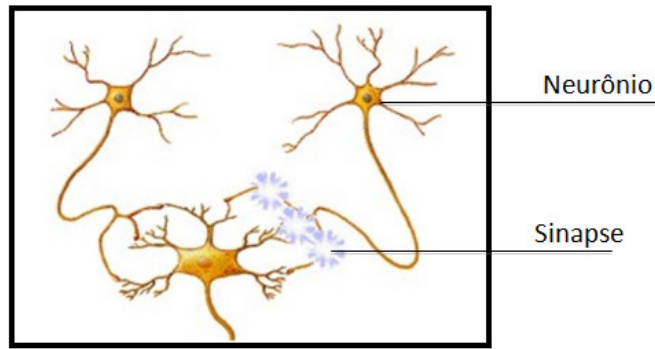


Figura 5: Conexão entre neurônios biológicos

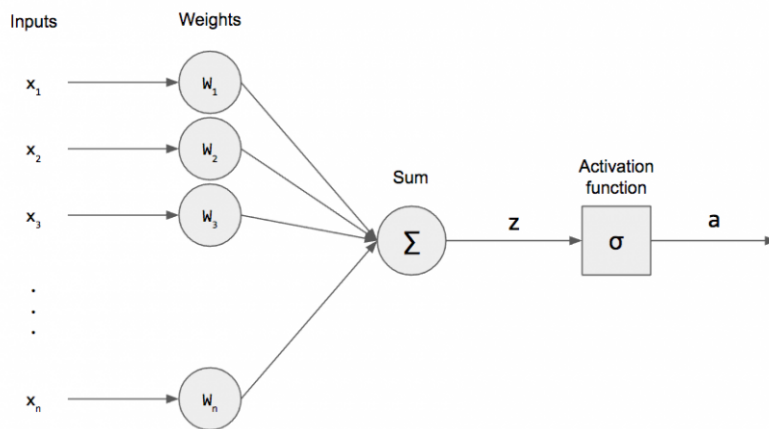


Figura 6: Representação de um neurônio

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

$$y = g(z) = \begin{cases} 0, & \text{se } z < \theta \\ 1, & \text{se } z \geq \theta \end{cases} \quad (2)$$

Em 1958, Frank Rosenblatt apresenta o neurônio *Perceptron* (ROSENBLATT, 1958), que mais tarde seria empregado como a unidade de processamento de uma RNA e de outros modelos de ML como as *support vector machines*. O Perceptron 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. Sua maior contribuição envolve a adição 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, 2000).

Este modelo inicial apresentava algumas limitações, atribuídas principalmente à sua linearidade e simplicidade, características que possibilitam resolver apenas problemas linearmente separáveis (BRAGA; CARVALHO; LUDERMIR, 2000). Um modelo

Perceptron é incapaz de aprender a função XOR, por exemplo (GOODFELLOW; BENGIO; COURVILLE, 2016). A adição de camadas e de funções de ativação nas saídas dos neurônios atribuiu às RNAs a potencialidade de serem aproximadas a qualquer função contínua, através da otimização por minimização da dissimilaridade entre o valor previsto pela rede y_t e o valor real y . Atualmente, as redes neurais artificiais podem apresentar diversos tipos de arquitetura, ao variar-se parâmetros como o número de camadas de neurônios, número de nós em cada camada, os tipos de conexões entre neurônios e topologia de rede. Alguns exemplos de arquiteturas podem ser encontrados na Figura 7.

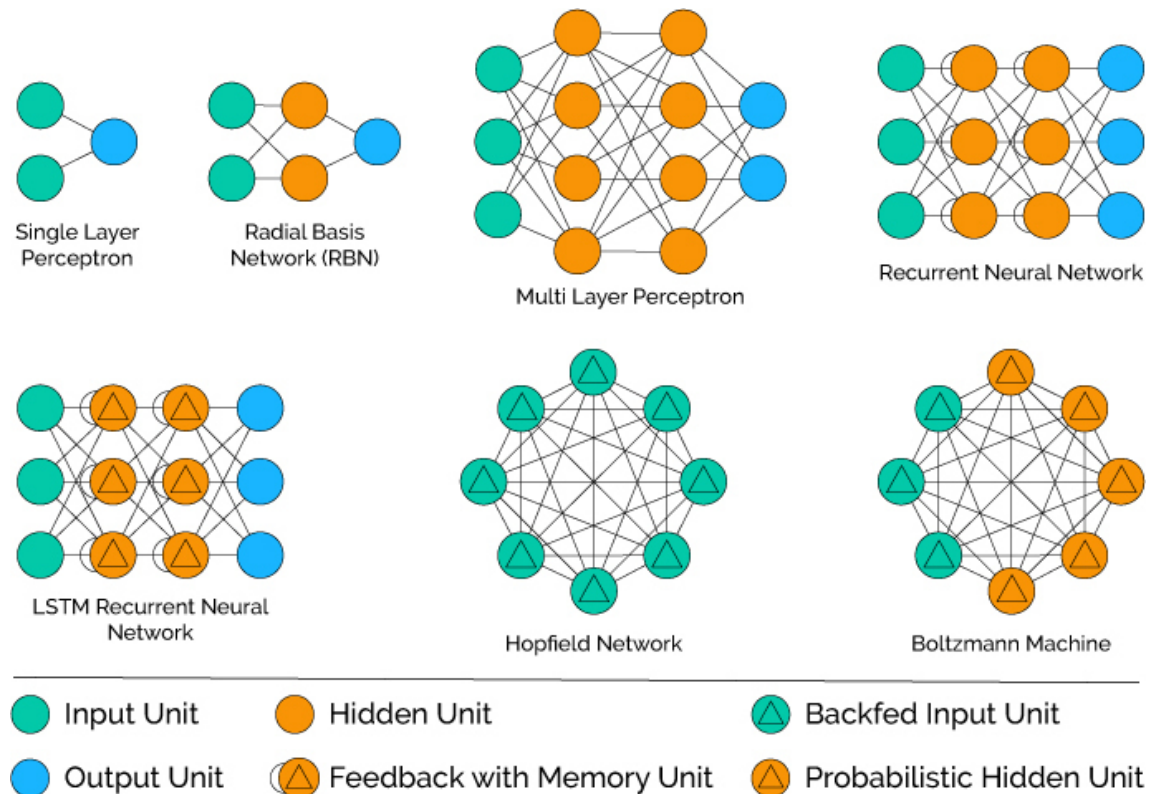
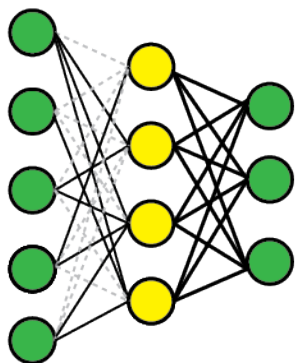


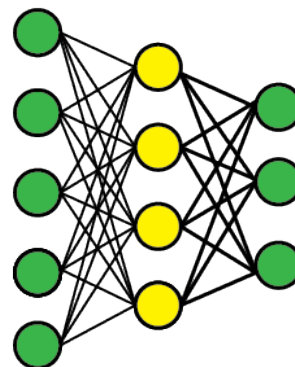
Figura 7: Arquiteturas populares de RNAs

No que tange a conectividade, uma RNA pode ser classificada como parcialmente ou totalmente conectada, como ilustra a Figura 9. O primeiro caso, exemplificado na Figura 8a, ocorre quando apenas alguns dos neurônios da camada anterior estão conectados aos da camada posterior. A RNA é dita totalmente conectada se todos os neurônios da camada anterior estão conectados aos da camada posterior, como o caso mostrado na Figura 8b.

Quanto aos tipos de conexão possíveis entre os neurônios, tem-se que as RNAs podem ser do tipo *feedforward* ou recorrente. As RNAs *feedforward*, exemplificadas na Figura 9a, são comumente associadas a um gráfico acíclico que descreve como as funções $y_L = g_L(z_L)$ descritas em cada camada L são compostas juntas para produzir uma saída Y . As RNAs recorrente, retratadas na Figura 9b, contém conexões entre neurônios de modo a formar um grafo direcionado cíclico, o que permite que o modelo capture sequências de comportamentos organizados em séries temporais. Por conta desta característica, as RNAs recorrentes são aplicadas especialmente em reconhecimento de escrita à mão e de fala.

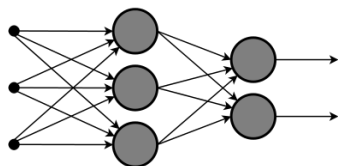


(a) Exemplo de RNA parcialmente conectada

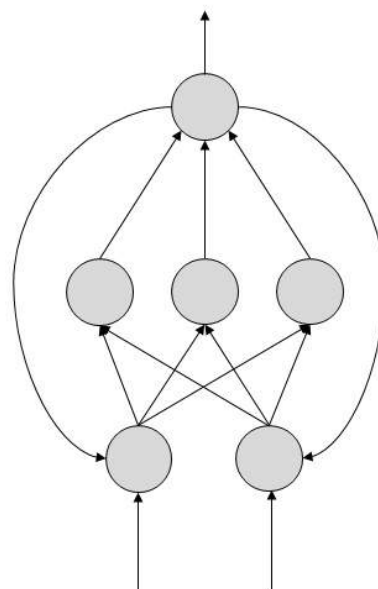


(b) Exemplo de RNA totalmente conectada

Figura 8: Exemplos de RNA com diferentes tipos de conectividade



(a) Exemplo de RNA *feed-forward*



(b) Exemplo de RNA recorrente

Figura 9: Exemplos de RNA com diferentes tipos de conexões entre neurônios

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 do Perceptron. Há também as redes de múltiplas camadas, que consistem de mais de um neurônio entre entrada e saída da rede, como é retratado no modelo da Figura 10. Redes com múltiplas camadas são capazes de aproximar diversas funções (HORNICK, 1991), (BRAGA; CARVALHO; LUDERMIR, 2000).

Segundo o teorema da aproximação universal exposto em (HORNICK, 1991), se a ativação de uma rede neural do tipo *Multi Layer FeedForward* for uma função limitada e não-constante, então para qualquer entrada x , a rede é capaz de aproximar qualquer função contínua no espaço de funções no espaço R_k , provendo a quantidade necessária de camadas ocultas. Esta arquitetura atribui às redes neurais artificiais o potencial de se tornarem máquinas de aprendizado universal.

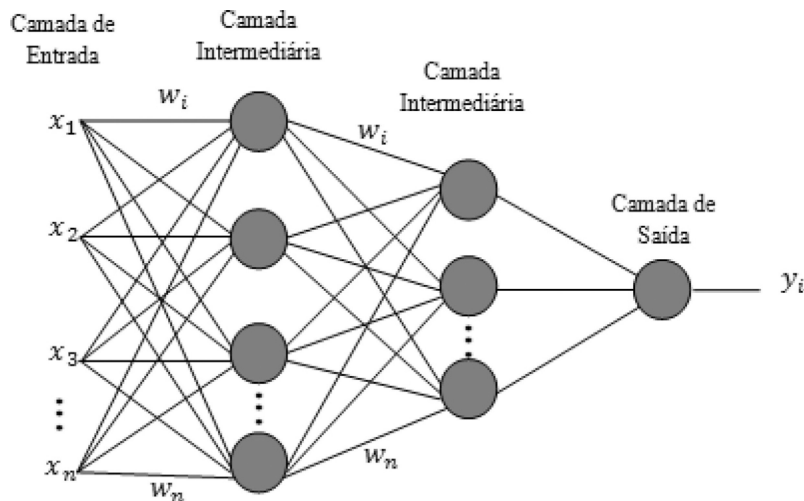


Figura 10: Rede Neural Multicamadas

O objetivo das RNAs é aproximar-se de funções que mapeiem as entradas X às saídas Y . Para atingir este objetivo, o modelo *multilayer perceptron*, ou MLP, tem duas fases: a fase *forward*, na qual há a inferência da saída da rede perante determinada entrada, e a fase *backward*, em que há o processo de ajuste dos pesos dos neurônios para minimizar o erro, ou perda, da saída prevista pela rede e o valor alvo. Este processo é chamado *backpropagation*.

A derivada f' de uma função $y = f(x)$ é dada pela fórmula na Equação 3, e fornece a inclinação de $f(x)$ no ponto x . Aplicada a uma função custo, esta operação específica 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*.

$$f(x + \epsilon) \approx f(x) + \epsilon f'(x) \quad (3)$$

Ao observar a Equação 3, percebe-se que se $f'(x) = 0$, a derivada não provê informações sobre a direção correta para onde a função deve se mover. Estes pontos, conhecidos como pontos críticos, podem ser máximos locais, mínimos locais ou pontos de sela, como mostra a Figura 11. Um máximo local é um ponto onde $f(x)$ é maior que

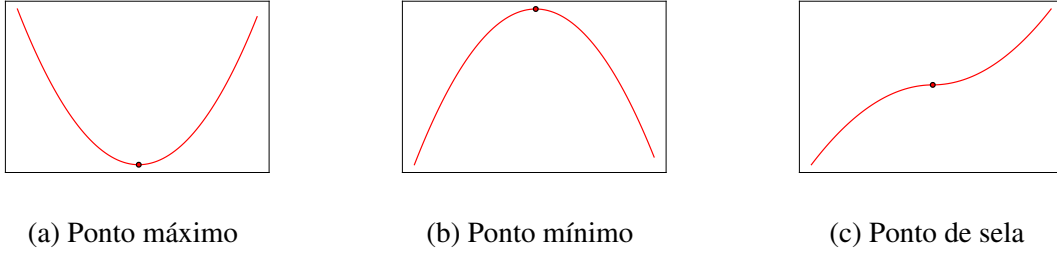


Figura 11: Exemplos de cada um dos três tipos de pontos críticos em funções no R^2

todos os pontos vizinhos, o que impossibilita o aumento do $f(x)$. Um mínimo local é um ponto onde $f(x)$ é menor que todos os pontos vizinhos, o que impossibilita a diminuição do $f(x)$. Um ponto de sela é um ponto estacionário que não se refere nem a um máximo ou a um mínimo.

No contexto da otimização do desempenho resultante do ajuste de pesos de determinada RNA, a disparidade entre as saídas previstas y_t e as saídas reais y presentes no conjunto de dados deve ser minimizada. A função $y = f(x)$ que exprime a variância entre estes valores para um modelo de *ML*, é comumente chamada de *função custo*, e definida a partir da estimativa por máxima verossimilhança. Esta estimativa consiste de um método estatístico aplicado com o objetivo de minimizar a dissimilaridade entre a distribuição empírica \hat{p}_{data} definida pelo conjunto de treinamento e a distribuição do modelo.

A função custo utilizada por uma RNA é decomposta como uma soma de funções de perda aplicadas aos exemplos de treinamento. Uma das funções custo que podem ser utilizadas, descrita na Equação 5, consiste na probabilidade logarítmica condicional negativa dos dados de treinamento, onde L é a perda calculada para cada exemplo dada na Equação 4.

$$L(x, y, \theta) = -\log p(y|x; \theta) \quad (4)$$

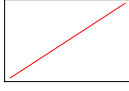
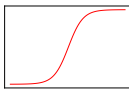
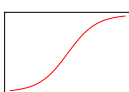


$$J(\theta) = E_{x,y \sim \hat{p}_{data}} L(x, y, \theta) = \frac{1}{m} \sum_{i=1}^m L(x^{(i)}, y^{(i)}, \theta) \quad (5)$$

Para estas funções custo, o gradiente descendente é computado através da Equação 6. O custo operacional desta operação é $O(m)$, sendo m o número de exemplos no conjunto de treinamento. Assim, o custo computacional cresce de maneira proporcional ao tamanho do conjunto de treinamento.

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

A escolha da função de ativação está diretamente relacionada às escolhas de funções de ativação $g(z)$ de camadas ocultas e da camada de saída. Várias funções podem ser utilizadas, a depender do tipo de processamento e de saída desejada, contanto que sejam contínuas e deriváveis (HORNÍK, 1991). As funções de ativação mais comuns estão detalhadas na Tabela 4.

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		$g(z_j) = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}} \quad j = 1, \dots, K$	$(-\infty, \infty)$

As RNAs *feedforward* formam a base para muitas aplicações. Inicialmente, este modelo era aplicado principalmente no mercado financeiro. Um exemplo desta utilização é o Sistema de Marketing Alvo, em inglês *Target Marketing System*, utilizado nos anos 1990 pela *Veratex Corp* para otimizar estratégias de marketing e cortar custos de marketing ao remover improváveis futuros consumidores de listas de compradores potenciais (WIDROW; RUMELHART; LEHR, 1994). Usos mais tardios também envolvem alocação de assentos em aviões, aprovação de empréstimo, controle de qualidade em processos industriais, entre outros. Quanto à detecção de padrões, destaca-se o uso de redes neurais convolucionais no reconhecimento de caracteres e dígitos escritos à mão, à exemplo de (LECUN et al., 1998). Na medicina, algumas aplicações de RNA convolucionais, compreendidas na sub-área *Deep Learning* (DL), podem compreender o aprimoramento e na segmentação de imagens cardíacas (OKTAY et al., 2018), aa classificação holística de padrões de atenuação em tomografias computadorizadas para doenças do tecido intersticial do pulmão (GAO et al., 2018), e a leitura de mamografias computarizadas (DUBROVINA et al., 2018). Atualmente, o modelo de RNA *feedforward multilayer perceptron* tem grande destaque dentre as técnicas de ML. Estes modelos são parte da sub-área *Deep Learning*, que será tratada na Seção 2.5.

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).

Há dois aspectos recorrentes nas diversas descrições de DL presentes hoje: (1) modelos que consistem de camadas ou estágios sucessivos de processamento de informações não-lineares; e (2) métodos para aprendizado supervisionado ou não-supervisionado de representação de características em camadas sucessivamente mais altas ou abstratas.

A melhoria do desempenho de modelos de DL é decorrente do aumento recente da quantidade de dados disponíveis sobre temas complexos, aliado com o aumento da disponibilidade de recursos computacionais para executar modelos mais robustos (GO-ODFELLOW; BENGIO; COURVILLE, 2016), (DENG; YU et al., 2014). Segundo a IBM, 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 nos últimos dois anos (IBM, 2017).

Para exemplificar o efeito da adição de camadas aos modelos de DL, expõe-se na Figura 12 uma visão geral do aumento da profundidade de redes neurais convolucionais aplicadas à detecção de objetos em imagens. Nota-se que conforme são adicionadas camadas às redes, há uma diminuição no erro e, nas redes mais recentes, uma diminuição brusca do número de parâmetros a serem treinados. Isto indica que redes neurais convolucionais mais profundas tendem a capturar com maior precisão as características objetos em imagens, e que modelos mais modernos, a exemplo da GoogLeNet e ResNet, atingem este efeito enquanto diminuem a complexidade do problema ao derrubar o número de parâmetros a serem treinados.

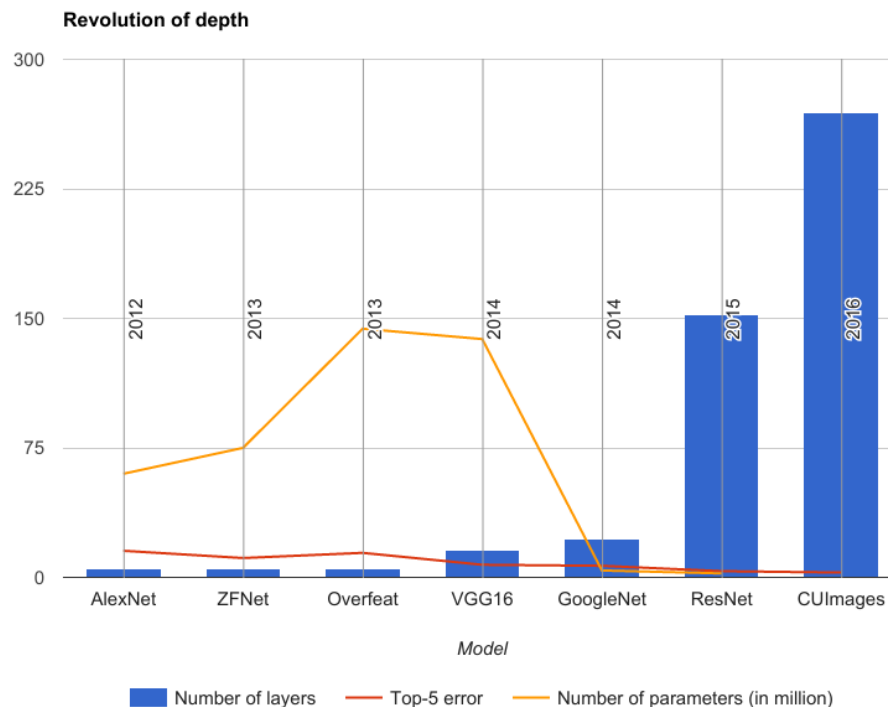


Figura 12: Evolução de profundidade, taxa de erro e número de parâmetros de redes neurais convolucionais com o passar dos anos. Fonte: (ECARLAT, 2017).

2.5.1. Breve Histórico

Historicamente, o conceito de DL se originou de pesquisas sobre Redes Neurais Artificiais (RNA), mais especificamente, as redes neurais *feed-forward* com muitas camadas ocultas, também chamadas de redes neurais profundas (DENG; YU et al., 2014). O termo *deep learning* foi utilizado pela primeira vez em (DECHTER, 1986), no contexto da descoberta de todas as configurações de conflitos mínimas no fim de um problema de satisfação de limitação (em inglês, *Constraint-Satisfaction Problem – CSP*). Em ANO, foi utilizado para designar métodos que têm a ver com o DL moderno em PUBLICAÇÃO, que trata de TEMA. A partir daí, o termo passou a designar modelos compostos de várias camadas sucessivas de operações não lineares utilizados para o aprendizado de determinada tarefa.

A história da pesquisa sobre este DL está dividida em três ondas, ou gerações. A primeira geração foi marcada pelo desenvolvimento de modelos lineares simples, compostos apenas por um neurônio, como o modelo de (MCCULLOCH; PITTS, 1943) e o Perceptron de (ROSENBLATT, 1958). A segunda onda, iniciada nos anos 1980, teve como idéia central a interconexão entre vários neurônios (RUMELHART; MCCLELLAND, 1986), além do algoritmo de *back-propagation* (RUMELHART; HINTON; WILLIAMS, 1986). No final da segunda era, (HOCHREITER; SCHMIDHUBER, 1997) propuseram o modelo *long short-term memory* (LSTM) e (LECUN et al., 1998) propuseram a LeNet

EXPLICAR IMPORTÂNCIA DA LENET

. A terceira onda começa em 2006 com a publicação do artigo (HINTON; OSINDERO; TEH, 2006), que apresenta as *deep belief networks*, um tipo de RNA que

EXPLICAR

. Foi nesta onda, que dura até o presente momento (GOODFELLOW; BENGIO; COURVILLE, 2016), que o termo *deep learning* se popularizou. Na conjectura atual, redes neurais profundas têm superado sistemas que aplicam ML e funcionalidades desenhadas à mão em competições envolvendo inteligência artificial.

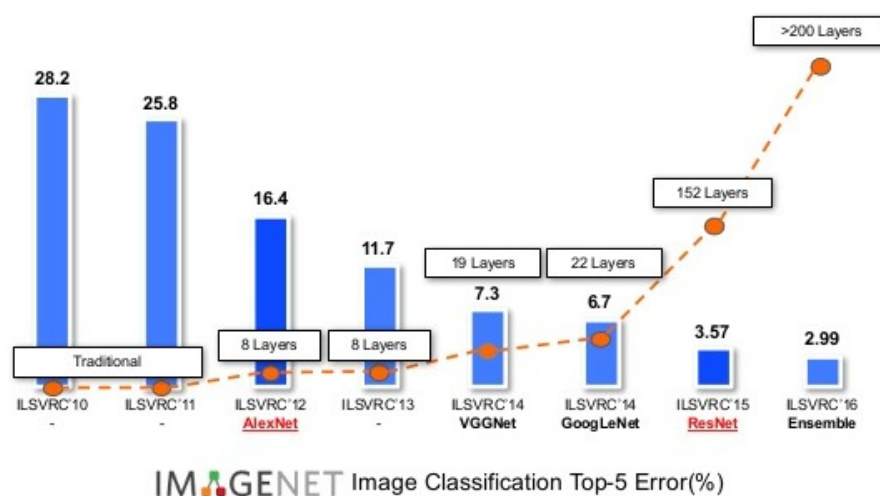


Figura 13: Evolução do erro dos modelos vencedores da competição ILSVRC pela profundidade das redes neurais (BORTH, 2017)

A *ImageNet Large Scale Visual Recognition Challenge*, ou ILSVRC, é uma com-

petição em que times de pesquisa avaliam seus algoritmos em um conjunto de dados fornecido, e competem para chegar à melhor acurácia em várias tarefas de reconhecimento visual. Em 2011, uma boa classificação no ILSVRC tinha por volta de 25% de erro. Em 2012, uma RNA convolucional chamada AlexNet atingiu 16.4% de erro. No gráfico da Figura 12, que mostra os melhores colocados na competição da ImageNet ano a ano, nota-se a diferença entre o erro atingido pelo modelo do ano anterior, que consiste de DETALHAR MODELO DE 2011, para a AlexNet.

EXPLORAR MAIS A IMAGEM.

Apesar de ter tido um foco inicial em técnicas novas de aprendizado não-supervisionado e na habilidade de modelos profundos de boa generalização a partir de conjuntos de dados pequenos, o momento atual das pesquisas em DL envolvem o uso de técnicas de aprendizado supervisionado bem mais antigas para o *leverage* de conjuntos de dados massivos e categorizados. Um exemplo destas técnicas são as redes neurais convolucionais com múltiplas camadas, que impulsionaram os avanços recentes alcançados no campo da visão computacional. Na Seção 2.5.2 a seguir, serão definidas as redes neurais convolucionais, suas características e particularidades. Na Seção 2.5.3 serão tratadas algumas das redes neurais convolucionais profundas que ganharam destaque em competições como o ILSVRC responsáveis pela fama atingida nos últimos anos.

2.5.2. Redes Neurais Convolucionais

Redes neurais convolucionais (CNN, do inglês, *Convolutional Neural Networks*) são uma classe de redes neurais *feed-forward* que têm se mostrado bem-sucedidas no processamento de dados que têm uma topologia bem definida e estruturada em uma grade, a exemplo de séries temporais e imagens. Sua principal característica envolve o uso de convoluções no lugar de multiplicações de matrizes em ao menos uma das camadas da rede neural (GOODFELLOW; BENGIO; COURVILLE, 2016). Este modelo pode ser aplicado em tarefas de classificação, regressão, localização, detecção, entre outros.

Cada camada das redes neurais convolucionais é composta por uma etapa de convolução, seguida por uma ativação não-linear, finalizando em *pooling*, como mostra a Figura 14. A seguir, serão explanadas cada uma destas etapas.

2.5.2.1. Convolução

A operação de convolução 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 $f : \mathbb{Z} \rightarrow \mathbb{R}$ representada simbolicamente por $x_1(t) * x_2(t)$ e definida de acordo com a Equação 7 (LATHI, 2006).

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

Quando a operação de convolução é aplicada em aprendizagem de máquina, a primeira função $x_1(t)$ é chamada de *input*, a segunda função $x_2(t)$ é chamada de *kernel*, e a saída $s(t)$ é chamada de mapa de *feature map*, ou mapa de características. Neste caso, a

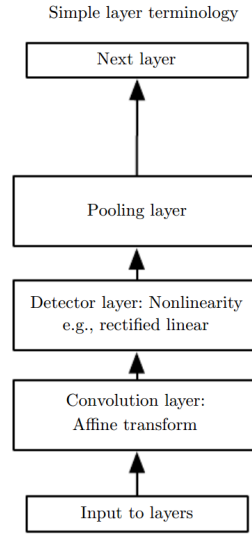


Figura 14: Componentes de uma camada de uma rede neural convolucional (GOODFELLOW; BENGIO; COURVILLE, 2016).

entrada normalmente é um vetor multidimensional de dados e o núcleo é um vetor multidimensional de pesos que devem ser adaptados pelo algoritmo de aprendizado de máquina. Em redes neurais convolucionais, os vetores multidimensionais de entrada e núcleo são chamados tensores. Além disto, assume-se que os valores dos tensores são zero em todos os pontos menos os que estão guardados em memória, ou seja, a operação de convolução é implementada apenas nas posições declaradas dos vetores de dados e peso. Assim, para uma imagem bidimensional de tamanho (m, n) I como entrada, tem-se um núcleo bidimensional K , e a operação de convolução é definida como exemplificado na Equação 8, para cada posição (i, j) do mapa de características resultante (GOODFELLOW; BENGIO; COURVILLE, 2016).

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

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.

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

A propriedade comutativa surge graças à ação de revolver o núcleo em relação à imagem, e não tem aplicação prática. Porém, esta propriedade não tem fins práticos além da prova da operação de convolução. Assim, é comum que seja implementada correlação cruzada, indicada na Equação 10, semelhante à convolução dada na Equação 9 sem que haja o espelhamento do núcleo em relação à imagem.

$$S(i, j) = I(i, j) * K(i, j) = \sum_m \sum_n I(i + m, j + n) K(m, n) \quad (10)$$

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*, *Softmax*, 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

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. d. P.; CARVALHO, A.; LUDERMIR, T. B. *Redes neurais artificiais: teoria e aplicações*. [S.l.]: Livros Técnicos e Científicos, 2000.

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>. Acessado em 21 de Março de 2018.

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.

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.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural computation*, MIT Press, v. 9, n. 8, p. 1735–1780, 1997.

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.

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 15 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>. Acessado em 21 de Março de 2018.