Estimação Inteligente de Idade Utilizando *Deep Learning*

Nicoli Pinheiro de Araujo

1. Apresentação e Justificativa

A face humana detém diversos atributos que podem ser utilizados para caracterizar indivíduos. Os traços faciais provém informações como identidade, expressão facial, gênero, origem étnica e idade de uma pessoa. Quanto a esta última, seres humanos são capazes de determinar a idade de indivíduos entre 20 e 60 anos a partir da face com alta acurácia. Porém, há casos em que a checagem de idade realizada por humanos é inconveniente. Neste contexto, pode-se fazer uso de sistemas de determinação automática de idade, especialmente quando é necessário apenas saber a idade de um indivíduo, sem identificá-lo (**??**).

A estimação de idade consiste de rotular a imagem de uma face automaticamente com um número ou uma classe que corresponda a uma idade ou intervalo de idade do indivíduo portador da face fotografada. Estimadores de idade são particularmente úteis em aplicações em que é necessário ou desejável conhecer a idade do usuário para que seja entregue um produto ou serviço de maneira personalizada. Alguns exemplos incluem sistemas de recomendação de conteúdo, enfermeiras robóticas, propagandas direcionadas, entre outros. Há também situações em que a idade do indivíduo determina se haverá ou não o provimento de produto ou serviço, ou se sua entrada será permitida em algum ambiente. Exemplos destes casos incluem sistemas de checagem de idade na entrada de bares e boates, em máquinas de venda automática que contenham bebidas alcóolicas, e sistemas de controle parental (**??**).

No caso específico de *Smart* TVs, por exemplo, 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 gênero. Porém, se fornecidos de maneira habitual, via preenchimento de formulários, além de ser uma tarefa massante, podem não refletir de maneira realística o perfil individual dos vários usuários que podem estar à frente de uma *Smart* TV em um determinado momento.

Smart TVs possuem câmeras que podem ser habilitadas para aquisição de imagens daqueles que estão à frente do televisor, respeitadas as preferências de privacidade de cada usuário. É possível usá-las como entrada para sistemas inteligentes de identificação de características, cujas previsões podem ser aplicadas, por exemplo, na 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 projeto de mestrado considera o desenvolvimento de estratégias inteligentes, baseadas na utilização de técnicas de *Deep Learning*, para estimação da idade de indivíduos 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 indivíduos, 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.

A realização de um projeto de mestrado 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 empregada na implementação de um controle parental mais eficiente, protegendo crianças e adolescentes de conteúdos inadequados à sua faixa etária.

Quanto ao provimento de propaganda direcionada, a estimação da idade pode contribuir para melhor adequação de vitrines e propagandas mostradas em telões distribuídos por shoppings a partir da idade dos indivíduos presentes naquele momento. Por exemplo, sabe-se que no almoço há uma maior circulação de jovens nos shoppings, que saem das escolas. Propagandas de filmes com público alvo mais jovem. Restaurantes que vendem bebidas alcóolicas e comida podem utilizar sistemas que vejam o público do momento para mostrar uma recomendação de produto a ser consumido.

A estimação da idade pode ser abordada de diversas maneiras. É possível realizar uma tarefa de regressão, em que a saída do modelo é um número real em um intervalo fechado condizente com o intervalo de idades de seres humanos. Pode-se também tratar como um problema de classificação, em que diversas quantidades de classes podem ser utilizadas. Por exemplo, jovem, mediano ou adulto, ou intervalos de idade separados por dez anos. Dependendo da escolha de tarefa de aprendizado utilziada, a tarefa pode ficar mais complexa. É possível realizar uma comparação entre as técnicas de classificação e regressão para verificar qual é mais vantajosa.

revisar

2. Objetivos

O objetivo geral deste trabalho consiste em aplicar técnicas de *Deep Learning* para estimação de idade de indivíduos a partir de suas respectivas fotografias faciais. Para alcançar esta meta, alguns objetivos específicos precisaram 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.

3. Contextualização Bibliográfica

A contextualização bibliográfica para a realização deste trabalho compreende conceitos ligados ao *Deep Learning*. Assim, a Seção **??** introduz os conceitos essenciais relativos à área. Os conceitos relativos específicamente às redes neurais convolucionais são descritos na Seção **??**. Alguns trabalhos relacionados à estimação de idade por meio de fotografias faciais estão descritos na Seção **??**.

3.1. *Deep Learning*

Deep Learning (DL), também conhecido como Aprendizado Profundo, 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 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 (**??**).

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 (**????**). 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 (**??**). Estes fatores possibilitaram a implementação de modelos que apresentaram uma melhoria significativa na eficiência de generalização frente a modelos existentes até então, especialmente em virtude da capacidade de organizar a computação como uma composição de várias operações não-lineares (funções de ativação) e uma hierarquia de características re-utilizadas (adição de camadas) (**??**).

3.2. Redes Neurais Convolucionais

Redes Neurais Convolucionais (CNNs, do inglês *Convolutional Neural Networks*) são uma classe de redes neurais *feedforward* com topologia bem definida e estrutura em grade, com o uso de operações de convolução em pelo menos uma de suas camadas (**??**). 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 (**??**).

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 (**??**). Assim, a convolução $s(t)$ de duas funções $x_1(t)$ e $x_2(t)$ é uma função $s : \mathbb{Z} \to \mathbb{R}$, denotada $s(t) = x_1(t) * x_2(t)$, e definida conforme Equação **??** (**??**):

$$s(t) = x_1(t) * x_2(t) = \int_{-\infty}^{\infty} x_1(a)x_2(t-a)da. \tag{1}$$

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 CNNs. 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 Equação **??**, é mostrada na Equação **??**:

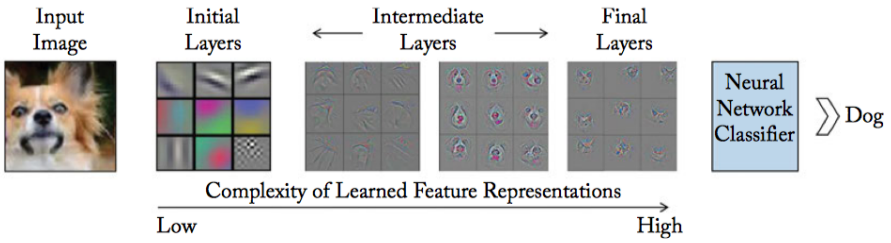
$$S(i,j) = I(i,j) * K(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n), \tag{2}$$

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 (**??**).

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

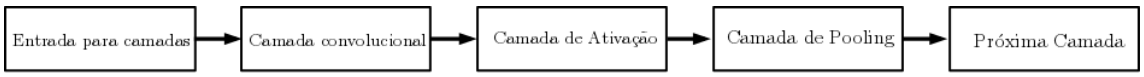
pelas CNNs, as características capturadas pelas camadas convolucionais vão se tornando mais complexas à medida que se aumenta a profundidade da rede. Assim, um primeiro *feature map* de uma camada convolucional captura um simples contorno, enquanto um *feature map* em uma camada mais profunda da rede pode capturar uma forma, um rosto ou até um objeto inteiro (**??**). Esta noção é ilustrada na Figura **??**.

Figura 1: Papel das camadas convolucionais e *feature maps* nas CNNs. Fonte: (**??**).



As camadas convolucionais, que capturam os *feature maps* e contém os pesos da rede, normalmente são seguidas por funções de ativação. Via de regra, a toda camada convolucional em uma CNN, segue-se uma função de ativação, finalizando em uma operação de *pooling*, como mostra a Figura **??**.

Figura 2: Componentes de uma camada de uma rede neural convolucional. Adaptado de (**??**).



Uma função de *pooling*, por sua vez, substitui a saída da rede em determinada localização por uma síntese estatística das saídas vizinhas. Por exemplo, a operação *max pooling* retorna o valor máximo em uma área retangular, enquanto a *average pooling* retorna a média das saídas de um retângulo. O objetivo desta operação é fazer com que o *feature map* seja invariante a pequenas mudanças na entrada. Esta invariância a pequenas mudanças locais é uma propriedade útil quando o mais importante for a existência da característica e não exatamente a sua posição, o que aumenta a eficiência geral da CNN ao reduzir drasticamente o número de valores a serem passados entre duas camadas quaisquer (**??**).

Embora se tenha uma noção clara das camadas individuais e de suas respectivas funções, a combinação das mesmas em uma rede neural convolucional não é uma tarefa trivial, podendo resultar em um número arbitrariamente grande de redes com milhares de parâmetros ajustáveis, cujo desempenho acerca de um problema ainda precisará ser aferido. Considerando os esforços computacionais para isto, a maioria das soluções atuais baseadas em DL fazem uso de CNNs canônicas já propostas na literatura, as quais são apresentadas a seguir.

3.3. Trabalhos Relacionados

A proposta apresentada está relacionada com inúmeros trabalhos envolvendo a aplicação de redes neurais convolucionais e outros modelos de ML para a estimação de idade de indivíduos a partir de fotos.

Segundo Fu et al., 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 (**??**), da textura do rosto (**??**), da captura de tendências de envelhecimento a partir de várias imagens de indivíduos de mesma idade (**??**) e da extração de características específicas relacionadas à idade (**????**). Modelos de ML 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 diversos trabalhos da literatura 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 ajustadas utilizando conjuntos de dados menores e especializados para a tarefa de estimação de idade.

O trabalho de Rothe em (**??**) relata um método para estimação de idade aparente em imagens de faces imóveis utilizando DL. O método proposto consiste em detectar uma face em uma imagem, para, em seguida estimar, sua idade. Para esta última tarefa, propôs um conjunto de 20 redes neurais convolucionais classificadoras com arquiteturas VGG-16 pré-treinadas com a base de dados ImageNet, e ajustadas utilizando imagens disponibilizadas pelos sites do IMDb, da Wikipedia, e o conjunto de dados *Looking At People* 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 localização do rosto e na média entre as idades previstas pelas 20 redes para o rosto detectado. A solução atingiu um MAE (*Mean Average Error*) de 3.221 na fase de testes.

Em Liu et al. cria-se um estimador de idade composto pela fusão de um modelo regressor e outro classificador (**??**). Realiza-se um pré-processamento das imagens de entrada, que envolve a detecção das faces presentes em cada imagem, seguida pela etapa de localização de pontos de referência, como olhos, nariz e boca, e por fim há a normalização das faces. 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* (**??**). 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, o trabalho de Levi e Hassner propõe uma rede neural convolucional de natureza mais simples, se comparada com as citadas acima (**??**). A arquitetura proposta 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 somente 8 classes de idade. O modelo é treinado utilizando apenas o conjunto de refe-

rência *Adience*, composto por imagens não filtradas para classificação de idade e gênero. Considerando uma margem de erro de uma classe de idade vizinha, a melhor rede obteve acurácia de $84.7\% \pm 2.2$ ao empregar a técnica de sobre-amostragem.

4. Metodologia

A metodologia para o desenvolvimento deste trabalho consistiu na realização da *fundamentação teórica sobre Machine Learning*, em especial contemplando os conceitos relativos às redes neurais convolucionais. Para tanto, considerou-se a literatura desta área para que haja o entendimento das bases matemáticas deste modelo computacional, como funcionam, quais as características e as arquiteturras mais importantes. Neste estudo, além dos aspectos teóricos, foram 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* (**??**). Inicialmente, houve a aquisição e o pré-processamento de imagens para *consolidar uma base de dados* para esta tarefa de aprendizado. Nesta etapa, foi considerada a literatura e uma base de dados já disponível e apropriadamente anotada, com licença livre de utilização.

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

O próximo estágio consistiu no *treinamento das redes neurais convolucionais* para o problema em questão. Durante este processo, uma parte da base de dados foi apresentada aos modelos para que houvesse o ajuste de pesos, compreendendo o aprendizado das características relevantes. O treinamento das redes ocorreu utilizando computação em nuvem e computadores disponíveis no Laboratório de Sistemas Inteligentes (LSI), tendo em vista a infra-estrutura de hardware necessária para realizar este procedimento.

Seguiu-se então o *teste das redes*, respeitando uma abordagem de validação cruzada e utilizando métricas de desempenho apropriadas. O objetivo desta fase consistiu 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 foram comparadas* e os melhores modelos elencados a partir destes valores, apontando assim um estimador apropriado para o problema inicialmente considerado.

Alem 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.

5. Cronograma

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

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

	2019												**2020**									
	04	**05**	**06**	**07**	**08**	**09**	**10**	**11**	**12**	**01**	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**	**12**
Escrita da Dissertação	X	X	X	X	X																
Fundamentação Teórica sobre CNNs	X	X	X	X	X	X															
Consolidação da Base de Dados				X	X	X	X														
Proposição de Modelos de CNNs					X	X	X	X	X												
Treinamento das CNNs					X	X	X	X	X	X											
Teste das CNNs					X	X	X	X	X	X	X										
Comparação de Métricas de Desempenho						X	X	X	X	X	X										
Defesa da Dissertação											X										