

NICOLI PINHEIRO DE ARAUJO

**ESTIMAÇÃO INTELIGENTE DE IDADE DE TELESPECTADORES PARA
APLICAÇÕES DE SUGESTÃO DE CONTEÚDO EM *SMART* TVs**

Trabalho de Conclusão de Curso apresentado
à banca avaliadora do Curso de Engenharia
de Computação, da Escola Superior de
Tecnologia, da Universidade do Estado do
Amazonas, como pré-requisito para obtenção
do título de Engenheira de Computação.

Orientador(a): Profa. Dra. Elloá Barreto Guedes da Costa

Manaus – Novembro – 2018

Capítulo 1

Solução Proposta

A solução proposta para a realização deste trabalho compreende a caracterização da tarefa de aprendizado, exposta na Seção 1.1. A descrição do conjunto de dados estão na Seção 1.2. A Seção 1.3 comprehende a limpeza, o pré-processamento e a consolidação do conjunto de dados. Por fim, na Seção 1.4 estão os modelos e hiperparâmetros de CNNs considerados.

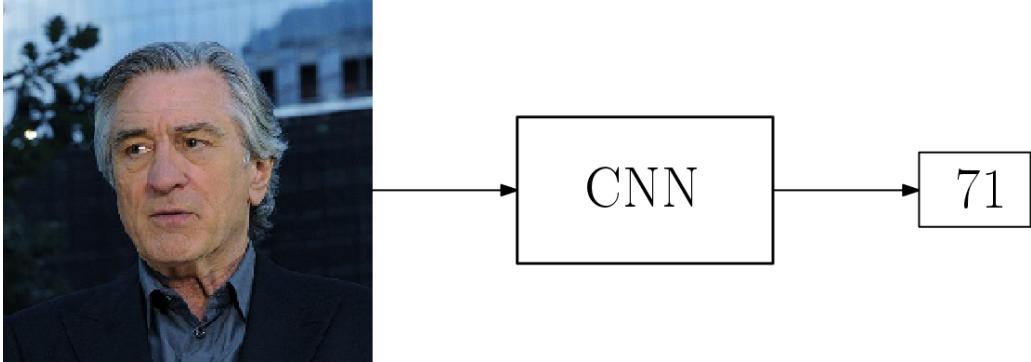
1.1 Tarefa de Aprendizado

A tarefa de aprendizado considerada para a estimativa de idade de telespectadores é a regressão. Neste contexto, uma imagem em cores RGB de dimensões 224×224 pixels contendo uma face humana centralizada será fornecida como entrada. A saída desejada é a estimativa de idade, em anos, da pessoa correspondente, conforme exemplificado na Figura 1.1. Esta tarefa será abordada segundo o paradigma de aprendizado supervisionado.

Os dados disponíveis para este contexto serão particionados em três conjuntos disjuntos, sendo 70% reservados para o treino, 10% para validação e 20% para teste. Esta partição obedece à tecnica *Holdout* de validação cruzada (BRINK; RICHARDS; FETHEROLF, 2016).

Os modelos propostos para esta tarefa terão seu desempenho aferido perante os dados do conjunto de testes de acordo com duas métricas de desempenho, *Root Mean Squared Error* (RMSE) e *Mean Average Error* (MAE). Estas métricas são análogas na medida em que consideram a diferença entre cada um dos valores previstos \hat{y} e os reais y , e posteriormente quantificam uma

Figura 1.1: Tarefa de aprendizado



média imune à variação positiva ou negativa desta diferença. A Equação 1.1 denota o cálculo do RMSE, enquanto a Equação 1.2 define o cálculo do MAE (WILLMOTT; MATSUURA, 2005).

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y})^2}. \quad (1.1)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}| \quad (1.2)$$

O RMSE será considerado para o cálculo da perda e por conseguinte da atualização dos pesos dos modelos, enquanto o MAE será utilizado para fins de comparação do desempenho dos modelos propostos neste trabalho com os modelos consolidados na bibliografia e detalhados nos trabalhos relacionados na Seção ??.

1.2 Conjunto de Dados

Para a tarefa de aprendizado apresentada, dispôs-se da base de dados experimentais IMDb, composta de 452.132 exemplares contendo imagens e outras informações de 20.284 dos atores mais populares listados no site IMDb. O conjunto de dados foi construído utilizando técnicas de *web crawling* aplicadas aos perfis de atores do site, em que foram coletadas todas as imagens relacionadas à celebridade, além de informações como data de nascimento, nome e gênero (ROTHE; TIMOFTE; GOOL, 2015b).

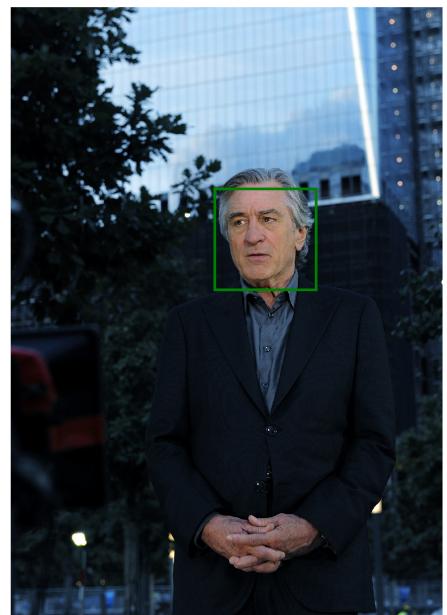
A base de imagens oriunda do site IMDb foi organizada por Rothe et al. considerando uma

tarefa de aprendizado análoga à deste trabalho, conforme mencionado na Seção ?? (ROTHE; TIMOFTE; GOOL, 2015a). Nesta base, há também as coordenadas da localização de um rosto detectado na imagem, além de uma pontuação atribuída ao rosto produzida pelo detector de face, quantificando o grau de certeza na detecção do rosto. Partindo da possibilidade de haver mais de um rosto por imagem, uma segunda pontuação é atribuída pelo detector, referente ao grau de certeza de que há outro rosto na mesma imagem.

Neste contexto, cada exemplo deste conjunto de dados é referente a uma imagem, cujos metadados estão descritos em seus atributos, que compreendem o nome, gênero, data de nascimento e um número de identificação da celebridade cujo perfil estava atrelado à imagem, o endereço da foto em disco, a suposta localização da face da celebridade, e pontuações referentes a duas possíveis faces encontradas. Assim, há exemplos de imagens em que há apenas um rosto, como mostrado na Figura 1.2. Já na Figura 1.3 está o exemplo de uma imagem onde há mais de um rosto, porém a localização do rosto está correta. Por fim, na Figura 1.4 há uma imagem com mais de um rosto, porém o rosto identificado neste item não é o da celebridade cujos dados estão referenciados.

Figura 1.2: Exemplo de imagem do conjunto de dados contendo apenas um rosto.

Meta-dado	Valor
ID Celebridade	16349
Nome	Robert De Niro
Endereço da imagem	imdb/34/nm0000134_rm3340090368_1943-8-17_2011.jpg
Pontuação da Face	5.21396
Pontuação da Segunda Face	NaN
Localização da Face	(663.65, 992.475, 590.134, 918.959)
Data de Nascimento	1943 – 08 – 17
Ano da Foto	2011
Gênero	Masculino



A versão original das imagens do conjunto de dados IMDb ocupava 267 GB em disco. Porém,

Figura 1.3: Exemplo de imagem do conjunto de dados contendo mais de um rosto com a classificação correta.

Meta-dado	Valor
ID Celebridade	16349
Nome	Robert De Niro
Endereço da imagem	imdb/34/nm0000134_rm17663 60064_1943-8-17_2010.jpg
Pontuação da Face	5.12527
Pontuação da Segunda Face	5.08887
Localização da Face	(914.886, 1426.31, 287.31, 798.734)
Data de Nascimento	1943 – 08 – 17
Ano da Foto	2010
Gênero	Masculino

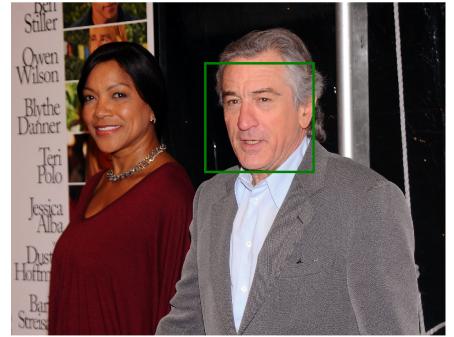


Figura 1.4: Exemplo de imagem do conjunto de dados contendo mais de um rosto com a classificação errônea.

Meta-dado	Valor
ID Celebridade	16349
Nome	Robert De Niro
Endereço da imagem	imdb/34/nm0000134_rm14800 44288_1943-8-17_2012.jpg
Pontuação da Face	5.51656
Pontuação da Segunda Face	4.55379
Localização da Face	(1392.72, 1614.18, 225.55, 447.003)
Data de Nascimento	1943 – 08 – 17
Ano da Foto	2012
Gênero	Masculino



uma versão pré-processada dessas imagens está disponível, contendo as faces recortadas com 40% da largura e altura da imagem original, totalizando 7,1 GB de dados. Esta versão foi considerada neste trabalho.

1.3 Limpeza e Pré-processamento dos dados

A fim de adequar melhor o conjunto de dados para os modelos de CNNs utilizados, realizou-se uma limpeza e pré-processamento dos meta-dados e das imagens da base IMDb, que se iniciou com o cálculo do atributo alvo, a idade, a partir dos atributos originais fornecidos. A idade foi aferida através da data de nascimento da celebridade e do ano em que a fotografia em questão foi capturada.

Uma análise do conjunto de dados revelou a presença de itens com idade e gênero apresentando valores nulos, inválidos ou negativos, que foram descartados. Observou-se também a presença de múltiplos exemplos referentes à mesma pessoa com a mesma idade. Houve a remoção de tais exemplos, a fim de evitar que a apresentação de um mesmo rosto com a mesma idade provocasse *overfitting* nos modelos. Exemplos atípicos, possivelmente resultados de rotulação incorreta, como idade maior que 100 anos ou não compatível com os dados da celebridade referida nos meta-dados também foram descartados. Os atributos de pontuação de rostos foram úteis para identificar e remover exemplos em que não havia nenhum rosto identificado, ou em que havia mais de uma face na imagem. Este descarte foi realizado com o objetivo de eliminar rotulações errôneas, como a mostrada na Tabela 1.4.

Foram realizados também procedimentos de pré-processamento nas imagens. Considerando a literatura, padronizou-se o modo RGB e o tamanho das imagens para 224×224 pixels. A seguir, realizou-se a equalização do histograma de cores das imagens, a fim de aumentar o contraste global e evidenciar características das faces, como mostrado na Figura . A normalização realizada a seguir teve como objetivo manter os valores de entrada entre 0 e 1, dividindo os pixeis das imagens por 255. Este último processo evita que as RNAs prevejam a média dos valores de entrada.

Após o pré-processamento das imagens de entrada, o cálculo do atributo alvo idade, a adequação do caminho para as imagens em disco e a remoção de exemplos impróprios, seguiu-se o descarte dos outros meta-dados irrelevantes para a tarefa de estimativa de idade de um indivíduo a partir de imagem. A data em que a foto foi tirada, nome, número de identificação, gênero, data de nascimento, localização do rosto da celebridade e pontuações de rostos nas

figura com e sem equalização do histograma, colocar citação da normalização e da equalização por histograma

imagens foram removidos.

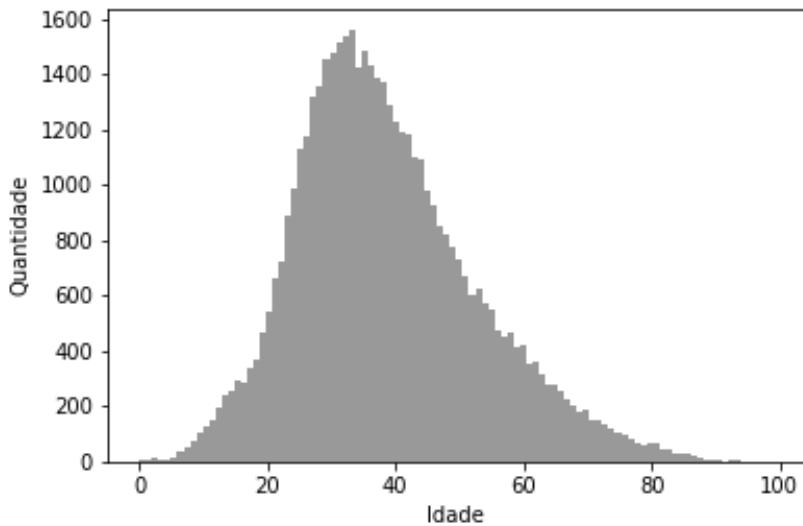


Figura 1.5: Histograma de frequência da idade do conjunto de dados utilizado.

Por fim, o conjunto de dados consolidado consiste de 47.950 exemplos contendo imagens e idades de 14.607 celebridades distintas, ocupando 1,2GB. O histograma de frequência da distribuição de idades de 0 a 100 anos presente nos exemplos da base de dados pode ser visualizado na Figura 1.5. Este total foi então dividido como proposto: conjunto de treinamento, contendo 70% dos exemplos, ou seja, 33.565 amostras; conjunto de validação, referente a 10% dos dados, ou seja, 4.795 itens; e, por fim, conjunto de testes, contendo os 20% restantes, ou seja, 9.590 exemplos.

1.4 Modelos de CNN Considerados

Levando em conta a adoção de CNNs como o modelo de ML a ser usado neste trabalho, considerou-se a utilização das arquiteturas LeNet e AlexNet. A implementação da AlexNet seguiu a prática atual de utilizar apenas uma GPU em seu treinamento, então as camadas divididas no trabalho original foram unificadas (TENSORFLOW, 2018). Todas as funções de ativação tangente hiperbólica disponíveis nas versões originais destas redes foram substituídas pela função *ReLU*, por ser mais eficiente computacionalmente, evitar que o gradiente descen-

dente tenda a zero e por promover uma convergência mais rápida (MAAS; HANNUN; NG, 2013). Adotou-se um *batch size* igual a 64 para o treinamento, e o método de otimização do gradiente descendente foi o *Adam*. O número de épocas e a taxa de aprendizado foram obtidas de maneira experimental, observando a perda obtida ao final de cada época.

A fim de caracterizar a tarefa de regressão proposta, as camadas de saída da LeNet e AlexNet com múltiplos neurônios voltados à classificação foram substituídas por apenas um neurônio. Em um primeiro momento, o neurônio de saída era seguido por uma função de ativação *ReLU*, e as imagens de entrada não estavam normalizadas ou equalizadas. Após análise dos resultados preliminares obtidos para estes modelos iniciais, substituiu-se a *ReLU* da camada de saída por uma de suas variantes, chamada *Leaky ReLU*, e expressa na Figura 1.6, e as imagens foram normalizadas e equalizadas. A taxa de aprendizado inicial foi padronizada em um valor de 10^{-3} com decaimento de 10^{-10} para ambas as redes.

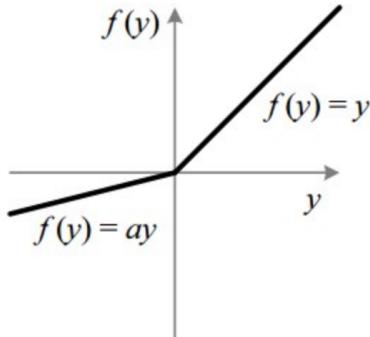


Figura 1.6: Função de Ativação *Leaky ReLU*

Na seção a seguir estão os resultados preliminares obtidos do treino dos modelos, hiperparâmetros e estratégias supracitados.

Capítulo 2

Resultados e Discussão

Considerando a abordagem descrita na solução proposta, os resultados da execução das CNNs aplicadas ao problema de estimativa de idade a partir de uma imagem de face são apresentados a seguir.

2.0.1 Abordagem 1

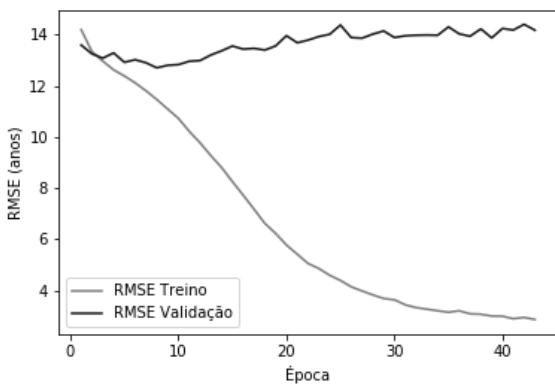
A primeira abordagem de treinamento das CNNs utilizou as imagens da base de dados com equalização por histograma de frequência, normalizadas, e com 50% de chance de estarem rotacionadas horizontalmente. Conforme mencionado na Seção 1.4, os treinamentos e testes compreenderam as arquiteturas canônicas LeNet e AlexNet com funções de ativação *ReLU* e *Leaky ReLU* nas camadas ocultas e de ativação. É importante ressaltar que neste momento não foram utilizadas técnicas de *transfer learning*. Obedecendo ao método de validação cruzada *holdout* previamente mencionado, os resultados da etapa de teste foram obtidos, e estão detalhados na Tabela 2.1.

A CNN que implementa a arquitetura LeNet com função de ativação *ReLU* foi treinada por 43 épocas, obteve MAE de 14.09 e RMSE 17.93. A LeNet com função de ativação *Leaky ReLU* foi treinada por 15 épocas, obteve MAE de 14.44 anos e RMSE de 18.18 anos. Os treinamentos duraram aproximadamente 16 e 12 horas respectivamente, em uma instância do Google Compute Engine com 4 CPus virtuais e 15 GB de RAM. Os gráficos de treinamento e

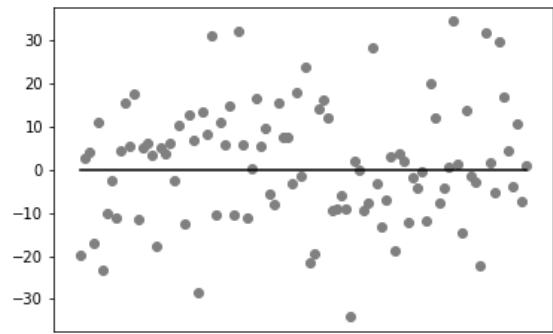
as retas zero obtidas a partir da apresentação do conjunto de teste aos modelos consolidados podem ser vistos na Figura 2.1. É possível notar que ambas as redes sofreram com *overfitting* e obtiveram grande margem de erro, no entanto a LeNet que utilizou *Leaky ReLU* como função de ativação obteve um desempenho mais satisfatório.

Figura 2.1: Resultados do treinamento e teste da CNN LeNet.

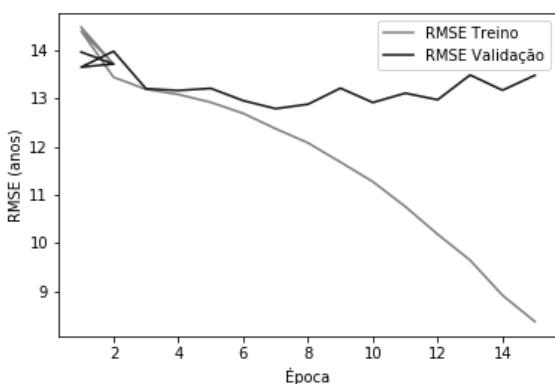
(a) RMSE de treinamento da arquitetura LeNet utilizando funções de ativação *ReLU*.



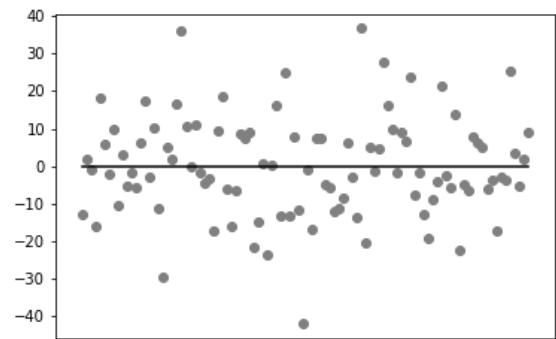
(b) Reta-0 LeNet *ReLU*.



(c) RMSE de treinamento da arquitetura LeNet utilizando funções de ativação *Leaky ReLU*.



(d) Reta-0 LeNet *Leaky ReLU*.

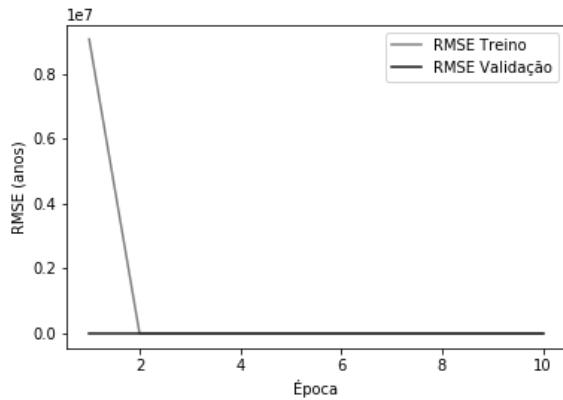


A CNN que implementa a arquitetura AlexNet com função de ativação *ReLU* foi treinada por 10 épocas, obteve MAE de 38.63 anos e RMSE de 41.22 anos. A AlexNet com função de ativação *Leaky ReLU* foi treinada por 30 épocas, obteve MAE de 14.44 anos e RMSE de 15.33 anos. Os treinamentos duraram aproximadamente 15 e 38 horas respectivamente, na mesma instância do Google Compute Engine utilizada para o treinamento das redes LeNet. Os gráficos de treinamento e as retas zero obtidas a partir da apresentação do conjunto de teste aos modelos

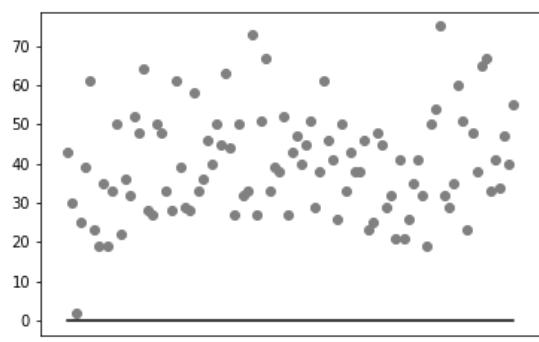
consolidados podem ser vistos na Figura 2.2. É possível notar que enquanto a AlexNet que utiliza *ReLU* sofreu de *dying ReLU problem* e previu idade zero para todos os exemplos do conjunto de teste, a AlexNet que utilizou *Leaky ReLU* como função de ativação foi capaz de convergir para uma solução e prever idades mais próximas às reais.

Figura 2.2: Resultados do treinamento e teste da CNN AlexNet.

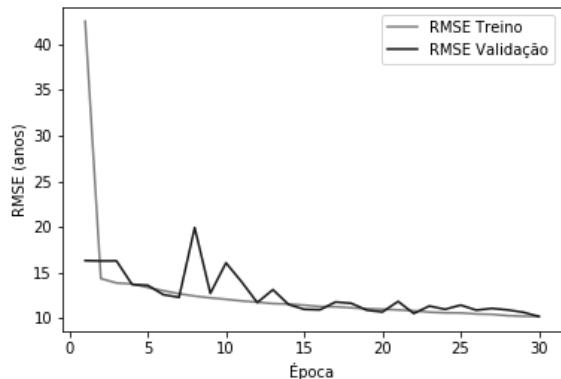
(a) Treinamento AlexNet *ReLU*.



(b) Reta-0 AlexNet *ReLU*.



(c) Treinamento AlexNet *Leaky ReLU*.



(d) Reta-0 AlexNet *Leaky ReLU*.

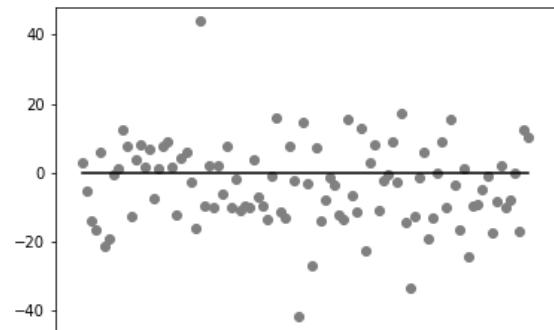


Tabela 2.1: Resultados do treino e teste dos modelos propostos.

Rede	Função de ativação	Parâmetros	Épocas	Tempo de treinamento	MAE Teste	RMSE Teste
LeNet	<i>Leaky ReLU</i>	params	15	12 h	14.44	18.18
LeNet	<i>ReLU</i>	params	43	16 h	14.09	17.93
AlexNet	<i>ReLU</i>	58.286.145	10	15 h	38.63	41.22
AlexNet	<i>Leaky ReLU</i>	params	30	40 h	15.33	18.58

2.0.2 Abordagem 2

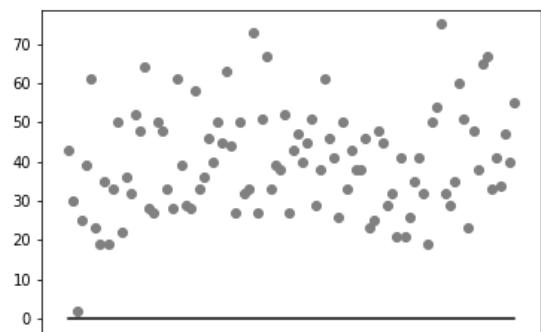
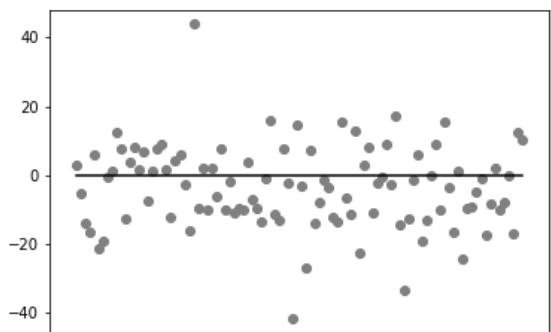
- Mesmas redes - Normalização das imagens, equalização por histograma -> o que é - data augmentation -> mais técnicas de data augmentation

2.0.3 Abordagem 3

Outras arquiteturas VGG com transfer learning 1. Retirar última camada (softmax) e adicionar leaky relu 2. Retirar duas últimas camadas (dense e softmax) e adicionar leaky relu

Figura 2.3: Redes neurais biológicas.

- (a) Reta-0 Alexnet LReLU com imagens normalizadas e equalizadas (b) Reta-0 Alexnet ReLU com imagens normalizadas e equalizadas



Capítulo 3

Considerações Finais

O objetivo deste trabalho consiste em elaborar estratégias inteligentes para estimação de idade de telespectadores de *Smart TVs* a partir de suas respectivas fotografias faciais. Para este fim, foram propostos, treinados e testados em caráter preliminar dois modelos de CNNs já bem estabelecidos na literatura, a LeNet e AlexNet, com dois perfis de hiperparâmetros cada um.

Com isto, observou-se uma melhora significativa na performance da AlexNet, enquanto o RMSE da LeNet não sofreu grandes mudanças. Quanto às saídas das redes, a LeNet exibiu valores positivos e negativos próximos de zero, e a AlexNet forneceu previsões que giravam em torno da média dos dados. Estes resultados são preliminares e certamente outros modelos e parâmetros serão investigados conforme previsto na metodologia e cronograma deste trabalho de conclusão de curso.

Nos próximos meses, os esforços estarão concentrados em pesquisar e adotar estratégias que possam minimizar os problemas identificados, como substituir as funções de ativação das camadas ocultas por outras variantes da *ReLU*, adotar métodos específicos de inicialização de pesos, normalização de *batch*, entre outros. Planeja-se também a proposição, o treinamento e teste de outras redes inspiradas em outros modelos canônicos.

O problema em questão é importante do ponto de vista prático para o desenvolvimento de diversas soluções de recomendação de conteúdo e controle parental em *Smart TVs*, auxiliando no desenvolvimento destas soluções tecnológicas. Considerando a formação de uma bacharela em Engenharia de Computação, endereçar este problema permite a prática de diversos conceitos

vistos ao longo do curso, em especial relacionados às disciplinas de Inteligência Artificial, Redes Neurais, Processamento Digital de Imagens, *Machine Learning* e Sinais e Sistemas.

Referências Bibliográficas

BRINK, H.; RICHARDS, J.; FETHEROLF, M. *Real-World Machine Learning*. 1st. ed. Greenwich, CT, USA: Manning Publications Co., 2016. ISBN 1617291927, 9781617291920.

MAAS, A. L.; HANNUN, A. Y.; NG, A. Y. Rectifier nonlinearities improve neural network acoustic models. In: ICML. *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*. 2013. Disponível em: <http://web.stanford.edu/~awni/papers/relu-hybrid_icml2013_final.pdf>.

ROTHE, R.; TIMOFTE, R.; GOOL, L. V. Dex: Deep expectation of apparent age from a single image. In: *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. Santiago, Chile: IEEE, 2015. v. 00, p. 252–257. Disponível em: <doi.ieeecomputersociety.org/10.1109/ICCVW.2015.41>.

ROTHE, R.; TIMOFTE, R.; GOOL, L. V. *IMDB-WIKI – 500k+ face images with age and gender labels*. 2015. Acessado em 19 de setembro de 2018. Disponível em: <<https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/>>.

TENSORFLOW. *AlexNet*. 2018. Acessado em 11 de Junho de 2018. Disponível em: <<https://github.com/tensorflow/models/blob/master/research/slim/nets/alexnet.py>>.

WILLMOTT, C. J.; MATSUURA, K. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, v. 30, n. 1, p. 79–82, 2005.