Reconhecimento de Emoções utilizando Redes Neurais Profundas

Jorge Ribeiro[1]

[1]Universidade Federal do Maranhão (UFMA)
Av. dos Portugueses, 1966 - Bacanga, São Luís - MA, 65080-805

{joorgemelo@gmail.com}

Abstract. *The use of deep neural networks for the classification of images have achieved great results in the last few years. New architectures have been proposed, primarily to improve the ImageNet's challenge result, and these architectures have been replicated in a way to solve several types of problems. In this paper, a convolutional neural network will be used to classify emotions presented in thousands of face's images, using the FER2013 dataset.*

Resumo. *A utilização de redes neurais profundas para a classificação de imagens tem tido ótimos resultados nos últimos anos. Novas arquiteturas vem sendo propostas, principalmente para melhorar o resultado do desafio ImageNet, e essas arquiteturas vêm sendo reproduzidas de forma a resolver diversos tipos de problemas. Neste trabalho, uma rede neural convolucional será utilizada para classificar as emoções apresentadas em milhares de imagens de rostos, utilizando o dataset FER2013.*

1. Introdução

As emoções expressas pelo ser humano dizem muito sobre o que ele deseja, sente, ou exprime sua reação diante de uma dada situação. Tais emoções são interpretadas nas suas relações interpessoais e facilitam a comunicação entre duas ou mais pessoas. Os estudos em inteligência artificial têm avançado de forma a compreender o comportamento humano, para assim imitá-lo e conseguir uma interação mais próxima ele. O avanço no poder computacional das últimas décadas permitiu que máquinas sejam capazes de aprender cada vez mais rápido, principalmente com o uso de redes neurais. O grande volume de dados gerado e compartilhado também facilitou o treinamento de redes neurais profundas. No que diz respeito ao reconhecimento de emoções, ainda se sentia falta de um dataset anotado corretamente, de grande volume e generalizado, de forma a ser utilizado no aprendizado supervisionado, aprendizado este mais comumente aplicado em redes neurais convolucionais.

Existem alguns desafios na tarefa de classificar emoções: 1) atualmente existem datasets grandes o suficiente, porém podem haver erros em sua anotação, 2) alguns datasets

disponíveis foram gerados em laboratório, por isso eles são poucos generalizados, dificultando o treinamento da rede e 3) classificar emoções é uma tarefa difícil, haja visto que uma reação pode expressar mais de uma emoção. O dataset FER2013, utilizado nesse trabalho, apresenta imagens *in the wild*, que quer dizer que são imagens em diversas situações, poses e tamanhos diferentes, resolvendo o problema da generalização explicado em 2). Exemplos do dataset podem ser visualizados na Figura 1.

Figura 1. Exemplos de imagens do FER2013 dataset. As imagens são em grayscale e dimensão 48x48.

2. Experimento

Nesta seção explicarei os recursos utilizados na elaboração desse trabalho. Além dos recursos que serão explicados a seguir, utilizei a biblioteca tensorflow para configurar as redes neurais testadas, treinar e classificar as imagens. A biblioteca foi utilizada dentro do TFLearn, uma API de alto nível que facilita a utilização do tensorflow.

2.1 Dataset

Redes neurais necessitam de grandes datasets para realizar um treinamento bem sucedido. Como já explicado antes, nos últimos anos surgiram datasets extensos anotados que facilitaram o treinamento de redes neurais convolucionais. Inicialmente tomei conhecimento de três datasets bem conhecidos para reconhecimento de emoções: Facial Expression Recognition Challenge (FER2013), Cohn-Kanade e Extended Cohn-Kanade (CK e CK+) e Radboud Faces Database (RaFD).

Como pode se observar na figura 2, os datasets RaFD e CK são gerados em laboratório. Por conta disso, as imagens desses dois datasets possuem qualidade maior que o FER2013, porém possuem menos imagens. Como o objetivo do trabalho é obter um resultado generalizado, apenas o FER2013 foi utilizado tanto para treinamento quanto para validação e teste. Foi fácil conseguir acesso aos datasets CK e CK+, basta um simples cadastro em sua página para realizar o download dos arquivos. Até a presente data de escrita desse artigo não consegui acesso ao RaFD dataset. O FER2013 está disponível para download no Kaggle.

Figura 1. Exemplos de imagens do RaFD (esquerda) e CK (direita) datasets.

2.1 Dataset

Redes neurais necessitam de grandes datasets para realizar um treinamento bem sucedido. Como já explicado antes, nos últimos anos surgiram datasets extensos anotados que facilitaram o treinamento de redes neurais convolucionais. Inicialmente tomei conhecimento de três datasets bem conhecidos para reconhecimento de emoções: Facial Expression Recognition Challenge (FER2013), Cohn-Kanade e Extended Cohn-Kanade (CK e CK+) e Radboud Faces Database (RaFD).

Como pode se observar na figura 2, os datasets RaFD e CK são gerados em laboratório. Por conta disso, as imagens desses dois datasets possuem qualidade maior que o FER2013, porém possuem menos imagens. Como o objetivo do trabalho é obter um resultado generalizado, apenas o FER2013 foi utilizado tanto para treinamento quanto para validação e teste. Foi fácil conseguir acesso aos datasets CK e CK+, basta um simples cadastro em sua página para realizar o download dos arquivos. Até a presente data de escrita desse artigo não consegui acesso ao RaFD dataset. O FER2013 está disponível para download no Kaggle.

2.2 Pré-processamento

Seguindo a sugestão de [1], realizei um pré-processamento nas mais de 35.000 faces do FER2013, de forma a excluir algumas imagens que poderiam comprometer o treinamento. Utilizando o algoritmo de classificação de faces Haar-cascade da biblioteca OpenCV, o dataset foi comprimido para 24.785 imagens de treino e 6.187 imagens de validação. Para cada imagem, apenas o quadrado correspondente a face foi mantido,

sendo adicionada uma borda cinza em volta da face de forma a manter o tamanho de 48x48 das imagens. Como as imagens já são em sua grande maioria de faces bem definidas, esse recurso das bordas não foi muito utilizado.

2.3 Arquiteturas de rede

Inicialmente uma versão simplificada da famosa rede AlexNet foi utilizada em grande parte desse trabalho, baseado no trabalho de Gudi [2]. Foram realizados treinos com poucas epochs (de 15 a 50) de forma a verificar a performance inicial da rede. Com 30 epochs apenas, a rede já apresentava um resultado satisfatório, chegando a mais de 50% de acurácia ao fim do treinamento. A arquitetura do trabalho de Gudi pode ser vista na Figura 3.

Foi implementada também uma versão fiel a VGG, com apenas uma alteração nas últimas camadas (em vez de três FC, apenas uma foi utilizada, com dropout de 0.3 em vez de 0.5 da original). Nessa arquitetura a acurácia subiu exponencialmente para a faixa de ~90%, porém a acurácia na validação continuou semelhante ao resultado da AlexNet.

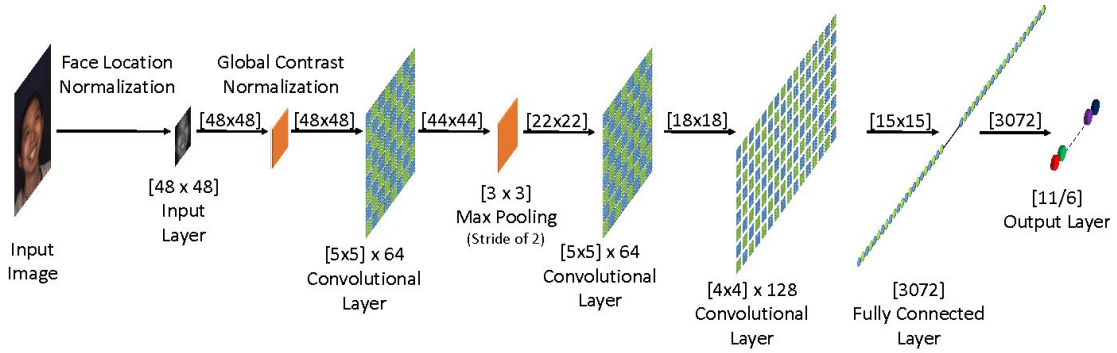

Figura 3. Arquitetura da rede baseada na AlexNet (neste trabalho, a localização de face e normalização de contraste não é realizada)

2.3 Avaliação

Todos os treinamentos apresentaram resultado semelhante, observando seus gráficos de acurácia e perda. Na Figura 4, pode se visualizar o progresso do treinamento para a rede AlexNet com 100 epochs.

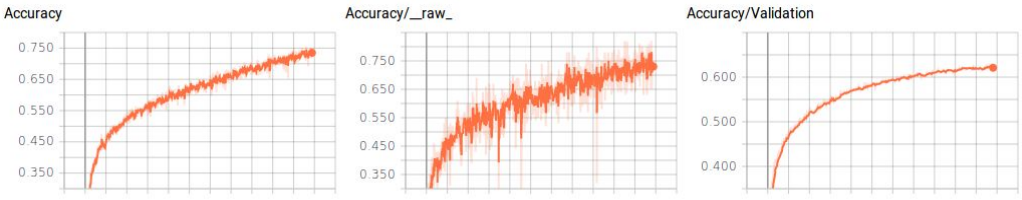

Figura 4. Acurácia e perda no treinamento com 100 epochs da versão modificada da AlexNet.

Infelizmente não foi possível obter os gráficos para outras quantidades de epochs e para o treinamento da VGG, pois houve um problema com o tensorboard e os gráficos foram perdidos.

3. Resultados

Os resultados finais para as arquiteturas testadas podem ser vistos na Tabela 1. As redes demonstram uma capacidade de crescimento inicial rápido, chegando já próximo a melhor acurácia na epoch 50.

Rede	**Epochs**		
	50	*100*	*400*
AlexNet	62.4%	73.9%	70.7%
VGG	Não realizado	92.8%	Não realizado

Tabela 1. Resultados finais de acurácia para as arquiteturas testadas.

Rede	**Epochs**		
	50	*100*	*400*
AlexNet	58.8%	62.18%	65.3%
VGG	Não realizado	65.9%	Não realizado

Tabela 2. Resultados finais de acurácia de validação para as arquiteturas testadas.

Observa-se pela Tabela 2 que a acurácia de validação manteve-se a mesma nas duas arquiteturas e nas diferentes quantidades de epochs. Na VGG a acurácia subiu para incríveis 92.8% já na primeira tentativa de treinamento, porém a validação estagnou em 65.9% na metade das epochs. É possível visualizar os resultados para cada classe de

emoção nas matriz de confusão nas Figuras 5 a 7 a seguir. Nota-se que felicidade (Happy) é a classe que obtém melhor classificação.

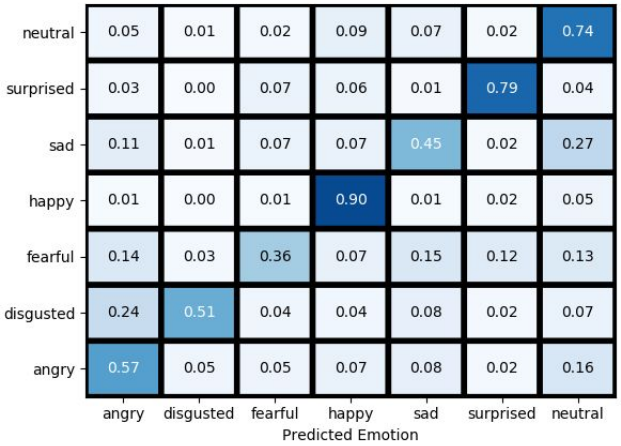

Figura 5. Matriz de confusão para treinamento da AlexNet com 100 epochs.

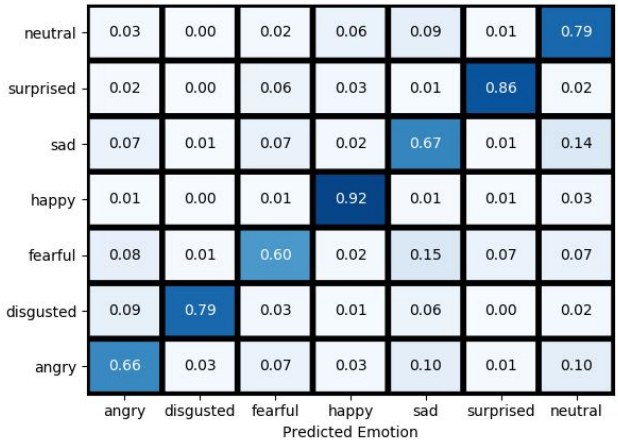

Figura 6. Matriz de confusão para treinamento da AlexNet com 400 epochs.

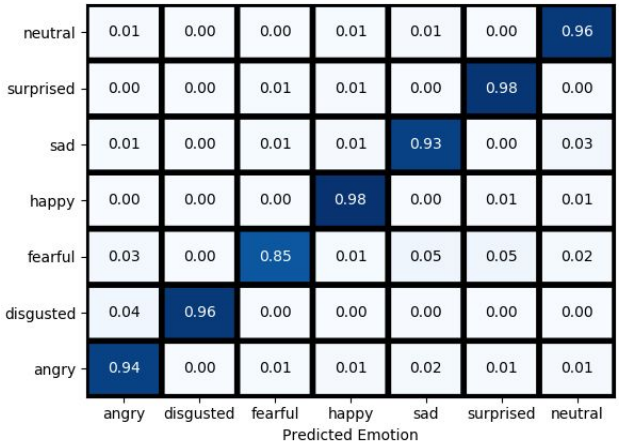

Figura 7. Matriz de confusão para treinamento da VGG com 100 epochs.

4. Conclusões

A proposta inicial desse trabalho era classificar emoções de faces de motoristas. Por falta de um dataset extenso e bem anotado com imagens de motoristas, decidi trabalhar com a classificação de emoção em rostos em qualquer situação. O trabalho pode ser futuramente estendido para abranger motoristas, pois basta segmentar a face do motorista utilizando algum algoritmo previamente implementado e fazer a classificação de sua emoção. Infelizmente não houve tempo suficiente para testar novas arquiteturas de forma a melhorar ainda mais o resultado, por conta do tempo que os treinamentos feitos terem tomado muito tempo e a falta de recursos computacionais (o computador utilizado para treinamento precisava ser compartilhado com outras pessoas). A VGG se mostrou promissora, pois alcança em suas epochs finais mais de 90% de acurácia. Pretendo fazer alterações na rede para melhorar a acurácia de validação.

Algo que também pode ser feito é a utilização de outros datasets no treinamento, para aumentar a generalização do modelo. O trabalho em [3] realizou treinamentos com outras bases (incluindo a CK+) e obteve bons resultados.

REFERÊNCIAS

[1] Enrique Correa, Arnoud Jonker, Michael Ozo, Rob Stolk, Emotion Recognition using Deep Convolutional Neural Networks, 2016.

[2] Amogh Gudi, Recognizing Semantic Features in Faces using Deep Learning, 2016.

[3] Dan Duncan, Gautam Shine, Chris English, Facial Emotion Recognition in Real Time, 2016.

[4] Tanner Gilligan, Baris Akis, Emotio AI, Real-Time Emotion Detection using CNN, 2016.

[5] Ali Mollahosseini, David Chan, and Mohammad H. Mahoor, Going Deeper in Facial Expression Recognition using Deep Neural Networks, 2015.

[6] Alexandru Savoiu, James Wong, Recognizing Facial Expressions Using Deep Learning, 2017.

[7] Emotion Recognition With Python, OpenCV and a Face Dataset, http://www.paulvangent.com/2016/04/01/emotion-recognition-with-python-opencv-and-a-face-dataset/