Gender and Age Prediction using Supervisioned Learning

Afonso Rodrigues, Alice Mangara, Núria Silva

Dezembro 2021

1. Introdução

Na última década, o reconhecimento de atributos faciais tais como a idade e o género tem sido um ponto de interesse na área da visão computacional. Uma das principais razões são as inúmeras aplicações deste problema, como na identificação de pessoas, no controle de segurança, e na interação homem-computador.

Diversos trabalhos têm sido publicados para o reconhecimento facial no âmbito da Inteligência Artificial, usando *Deep Learning* (DL). Esta recente área surgiu nos anos 2o000 e na última década atingiu um enorme sucesso em várias áreas, nomeadamente no processamento de imagens e vídeos. DL usa um conjunto de métodos que aprende com os dados e faz previsões inspiradas no comportamento do cérebro humano através de redes neuronais profundas.

Uma rede neuronal é um conjunto de perceptrões. Um perceptrão calcula a soma ponderada de vários *Inputs* (com pesos), aplica uma função de ativação e calcula o seu *Output*. Um neurónio é constituído pela soma ponderada e a função de ativação. Um conjunto de perceptrões em paralelo recebe um conjunto de *Inputs* e calcula um conjunto de *Outputs*. Se estes *Outputs* forem *Inputs* de outros perceptrões tem-se uma segunda camada de perceptrões, podendo assim a rede neuronal ter várias camadas ocultas.

Dentro da área de DL, as redes neuronais Convolutional Neural Networks (CNN) ocupam um papel de grande importância no processamento de imagens. A vantagem em utilizar uma rede CNN é a capacidade desta extrair características relevantes e depender de um menor número de parâmetros em relação às redes totalmente conectadas com o mesmo número de camadas ocultas. Cada camada oculta não é conectada com todas as unidades da camada seguinte, havendo assim um menor número de pesos a serem calculados. A

arquitetura CNN é composta por um empilhamento de vários tipos de camadas, cada uma desempenhando uma função específica.

A camada de convolução consiste numa combinação de operações lineares (convolução) e não lineares (ativação). A convolução é constituída por um conjunto de neurónios, sendo cada neurónio um filtro (tipicamente uma matriz de números) aplicado à imagem que entra. Os filtros são denominados de *kernels.* Existem 3 parâmetros que definem o tamanho dos dados resultantes de uma camada: profundidade, passo e zero-*padding*. A profundidade dos dados resultantes depende do número de filtros utilizados. Cada um destes filtros irá extrair características diferentes nos dados de entrada. O passo indica qual o tamanho do salto na operação de convolução. Quanto maior o valor do passo, menor será a altura e o comprimento dos dados resultantes, mas deste modo existem características importantes que poderão ser perdidas. O zero-*padding* consiste em preencher com zeros a borda dos dados de entrada. A vantagem em utilizar esta operação é poder controlar a altura e largura dos dados de saída. Deste modo é possível fazer com que fiquem com os mesmos valores dos dados de entrada. A função de ativação permite passar a saída de um neurónio de uma camada para a outra.

A camada de *pooling* é utilizada para reduzir as dimensões dos dados de entrada (dados de saída de uma outra camada), diminuindo o custo computacional. Usualmente, depois de uma camada de convolução existe uma camada de *pooling*. A operação de *pooling* consiste em agrupar os valores pertencentes a uma determinada região dos dados gerados pela camada de convulsão e substituí-los por alguma métrica que exista nessa região. Usualmente a substituição é feita pelo valor máximo encontrado nessa região, técnica denominada de max-*pooling*.

A camada ReLU (Unidade de Retificação Linear) de um modo simples recebe um determinado dado de entrada e no caso deste ser positivo não o altera, caso seja negativo, altera-o para zero.

A camada *Full Connected* (FC) utiliza as características da imagem de *Input* da rede. para classificar a imagem numa das categorias para qual a rede foi treinada. A camada FC devido às suas características é normalmente uma camada usada no fim da rede. É

denominada de *Full connected* pois conecta os neurónios da camada anterior com os neurónios da camada seguinte.

Após as operações nas camadas totalmente conectadas tem-se a camada de saída. O tamanho da camada de saída, ou seja, o número de neurônios, é igual ao número de classes no problema.

A rede VGG proposta por Karen Simonyan e Andrew Zisserman [4] foi a primeira rede a utilizar filtros de pequenas dimensões em cada camada de convulsão. Usualmente eram utilizados filtros de grandes dimensões (9x9 e 11x11) para capturar características nas imagens. A grande contribuição da rede VGG foi a ideia de que múltiplas convoluções 3x3 em sequência podiam substituir efeitos de filtros de maiores, resultando num menor custo computacional.

Caffe (*Convolutional Architecture for Fast Feature Embedding*) Deep Learning [6] é uma *framework* proposta em 2014 para implementar redes neuronais profundas de forma eficiente. Caffe providencia um conjunto completo de diferentes tipos de camadas incluindo convolução, *pooling* e ReLU.

Este trabalho tem como objetivo usar *Deep Learning* para identificar a idade e o género de uma pessoa através da imagem da sua face. Trata-se de um problema de classificação supervisionado uma vez que é conhecido o *output. E*ste tipo de identificação pode ser incorporado no comportamento de um sistema autónomo em diferentes situações do dia-a-dia.

As restantes secções estão estruturadas da seguinte forma. Na secção 2 são referidos alguns artigos relacionados com a deteção de atributos faciais como o género, idade e estado emocional. Na secção 3 é apresentada a base de dados utilizada neste trabalho. Na secção 4 é descrito o método implementado. A secção 5 apresenta os resultados e a sua discussão. A conclusão é apresentada na secção 6.

2. Estado da arte

Esta secção refere alguns trabalhos realizados para o reconhecimento de atributos faciais, com uso de *Deep learning*.

Em [2] foi desenvolvida uma rede VGG capaz de identificar o género de uma pessoa através da imagem da cara. A rede utiliza 4 tipos de camadas: convulsão, *pooling* (Max-pooling) e LRN (*Local Response Normalization*) que, usadas de maneira sequencial, permitem identificar um objeto na imagem e conhecendo algumas características do mesmo, categorizá-lo.

A abordagem em [3] usa uma rede CNN com arquitetura VGG de 16 camadas para categorizar a idade para valores inteiros entre 0 e 100. É usada uma última camada denominada *Euclidean loss function* que junta os valores de saída da camada anterior e calcula apenas um valor de saída que é a idade final.

O trabalho desenvolvido em [1] tem como objetivo encontrar as várias caras que existem numa imagem e, para cada cara, calcular a idade, o género e o estado emocional. A abordagem foi criar e treinar quatro redes diferentes, uma para a cara, uma para a idade, uma para o género e uma para a emoção, ao invés de criar só uma rede que calculasse os quatro parâmetros em simultâneo.

3. Materiais

Inicialmente, a base de dados a ser usada era a da IMDB-WIKI introduzida em [3], no entanto esta apresentava alguns erros que não nos permitiu usá-la. Optámos por usar a base de dados da UTKFace com cerca de 23 000 amostras, que contém informações sobre a idade, género e etnia.

4. Métodos

Os dados de entrada são as imagens obtidas da base de dados. Os dados de saída são a categoria de idade e o género. As categorias definidas são as seguintes: 0-2, 4-6, 8-12, 15-20, 25-32, 38-43, 48-53, 60-100.

Foi usada a *framework* Caffe para a construção da rede neuronal profunda. O modelo é constituído por dois tipos de ficheiros: extensão *.prototxt* – ficheiro de texto que descreve os parâmetros do modelo; extensão *.caffemodel* – modelo da rede neuronal.

5. Resultados e discussão

Primeiramente experienciámos o dataset da WIKI e IMDB, sem usar CNN, e observámos que o mesmo possuía diversos problemas, não eram imagens "clean" o que fez com que o programa/detetor/classificador não nos apresentasse qualquer resultado.

Decidimos então experimentar um novo programa com um data set diferente (UTKFace), alterando os intervalos de idade e utilizando o CNN.

Obtivemos os seguintes resultados:

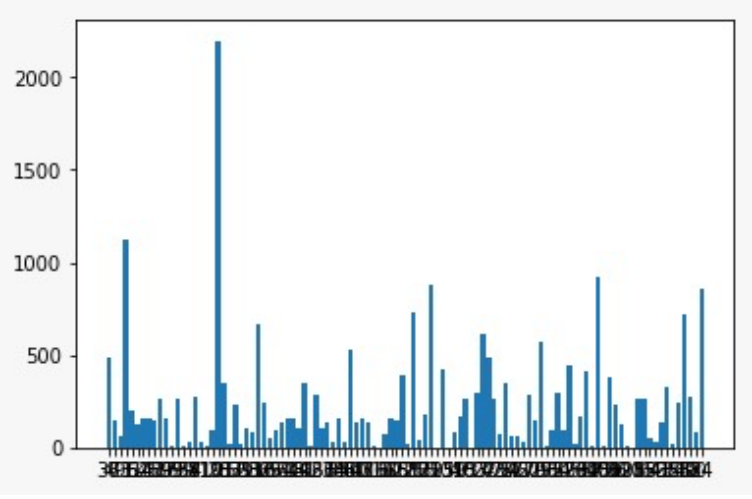


Figure 1-Distribuição de idades do data set UTKFace

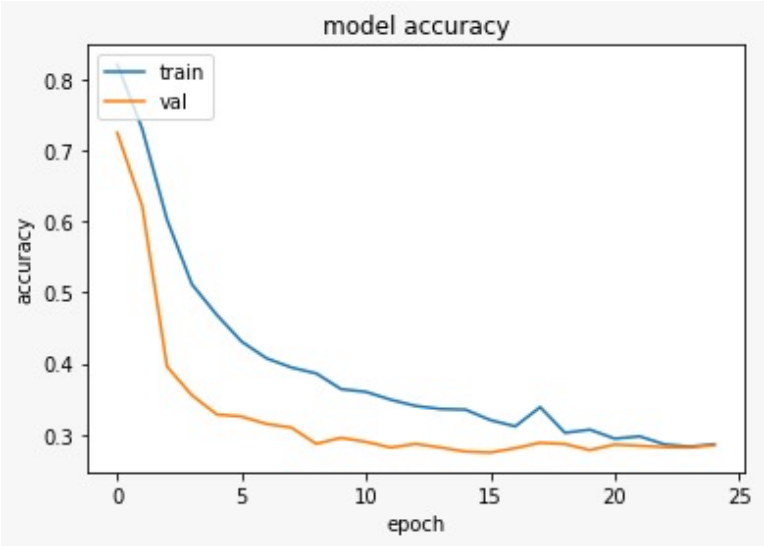
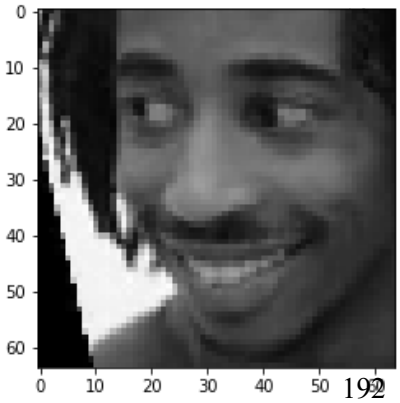
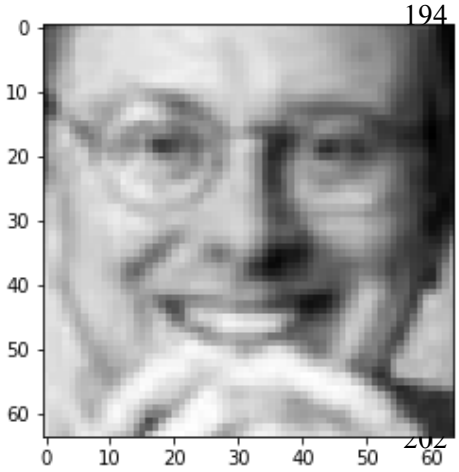


Figure 2-Accuracy

Aqui estão alguns exemplos:



```
Actual Gender: Male Age: 27
Values: [array([[0.5528965]], dtype=float32), array([[0.12574987]], dtype=float32)]
Predicted Gender: Male Predicted Age: 19-30
```



```
Actual Gender: Male Age: 47
Values: [array([[0.7136303]], dtype=float32), array([[0.01998827]], dtype=float32)]
Predicted Gender: Male Predicted Age: 31-80
```

6. Conclusões

Este projeto abordou técnicas de *Deep learning* para identificar a idade e o género de uma pessoa através da imagem da face de uma pessoa.

Concluímos que o tamanho, o ângulo, e em geral a qualidade da captação de imagem são bastante importantes, diríamos até cruciais para o bom funcionamento de um detetor e classificador de caras. Este programa que desenvolvemos e os métodos aplicados, não são suficientemente robustos para uma aplicação que envolva o mundo real, pois só funciona com imagens com parâmetros bem específicos.

Referências

[1] Afshin Dehghan, Enrique G. Ortiz, Guang Shu, and Syed Zain Masood. DAGER: Deep Age, Gender and Emotion Recognition using convolutional neural network. CoRR, abs/1702.04280, 2017.

[2] Amit Dhomne, Ranjit Kumar, and Vijay Bhan. Gender recognition through face using deep learning. Procedia Computer Science, 132:2 – 10, 2018. International Conference on Computational Intelligence and Data Science.

[3] Rasmus Rothe, Radu Timofte, and Luc Van Gool. DEX: Deep EXpectation of apparent age from a single image, Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW), 2015.

[4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014. 1, 2.

[5] https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53 (Acedido em 29/12/2021).

[6] Yangqing Jia Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. arXiv:1408.5093, 2014.

239