Universidade do Minho

Mestrado Engenharia Informática
Aprendizagem Profunda

PG45573 Adelino Silva
PG46526 André Mendes
PG47024 Angélica Cunha
PG46529 Diana Ferreira




08 de Maio de 2022

Conteúdo

Lista de Figuras

Introdução

O envelhecimento é algo inevitável, que tem efeitos na estrutura cerebral e cognitiva. Existem muitos aspetos que podem prolongar ou diminuir a idade cerebral de um indivíduo, dos quais fatores genéticos, hábitos de vida e ambientais. Nos últimos anos, o número de indivíduos com doenças neurodegenerativas, isto é, doenças associadas ao funcionamento do cérebro altamente incapacitantes, de que são exemplo, o Alzheimer, Parkinson e Huntington, tem vindo a aumentar. Este tipo de doenças estão altamente relacionadas com o desfasamento entre a idade cronológica e a idade cerebral. Saber a correlação entre este tipo de doenças no aumento da idade cerebral é fulcral na evolução do conhecimento do cérebro, pois ajuda no desenvolvimento de novos biomarcadores para as mesmas e no estabelecimento de um diagnóstico prematuro.

No âmbito da disciplina de Aprendizagem Profunda foi proposto um projeto cujo o objetivo é desenvolver e otimizar um modelo de aprendizagem profunda capaz de prever a idade do cérebro a partir de características de conectividade estrutural. Para a realização do mesmo, foram-nos fornecidos 2 *datasets*, sendo o primeiro divido em 2 conjuntos, um de treino com 112 casos, que é usado para desenvolver e treinar o modelo *Deep Learning*, e outro de teste com 28 casos, usado para validar o MAE do modelo em casos previamente desconhecidos. O MAE (*Mean Absolute Error* ou Erro Médio Absoluto) é a métrica de avaliação do modelo que estamos a utilizar. Por outro lado, o segundo *dataset* é construído com base em matrizes de conectividade estrutural, previamente normalizadas, extraídas de exames de Ressonâncias Magnéticas de difusão.

Neste documento iremos explicar detalhadamente todo o processo de criação do modelo desenvolvido. Começando por a exploração, análise e preparação dos *datasets*, apresentada na secção "Tratamento e Exploração de Dados". Seguidamente, apresentamos a forma como o modelo foi criado e otimizado para o problema, na secção "Modelo". Na secção "Resultados", são expostos os resultados alcançados e é feita uma análise crítica aos mesmos. Terminamos com a "Conclusão", onde fazemos um resumo geral do que foi falado anteriormente.

Contexto

É de salientar que atualmente vários modelos de Aprendizagem Profunda têm sido construídos para previsão da idade, permitindo efetuar a medição da diferença entre a idade prevista do cérebro e a idade cronológica (*brain gap*). Como já havia sido mencionado anteriormente, o objetivo deste trabalho é o desenvolvimento e otimização de um modelo capaz de prever a idade do cérebro por meio de características de conectividade estrutural. Para dar início à resolução do problema, foram fornecidos dois *datasets*.

O primeiro *dataset* está dividido em dois conjuntos. Ainda que ambos os conjuntos compreendam informação referente aos sujeitos em estudo, o seu propósito na modelação diverge, ou seja, o primeiro conjunto (*train*) é utilizado para o desenvolvimento e treino do Modelo e alberga a idade, o género, o ano de escolaridade (educação) e o identificador do sujeito, enquanto o segundo conjunto (*test*) é utilizado para validação do Modelo desenvolvido e alberga as mesmas *features* que o anterior à excepção da *label* idade, cujo o Modelo concebido deve ser capaz de a prever com base numa matriz de conectividade estrutural. O número de sujeitos em cada conjunto varia, sendo que o conjunto *train* compreende informação referente a 112 sujeitos, enquanto o de *test* compreende informação relativa a 28 sujeitos. A idade destes sujeitos cobre um intervalo de 13 a 79 anos, com uma média de aproximadamente 44 anos. Na Figura 1 estão especificadas e descritas as *features* contidas no primeiro *dataset*.

Nome da coluna	Descrição	Tipo de dados
id	Identificador do sujeito	Numérico contínuo
sex	Género do sujeito	Categórico
education	Ano de escolaridade do sujeito	Numérico contínuo
age	Idade do sujeito	Numérico contínuo

Figura 1: *Features* do *dataset* com informação referente aos sujeitos em estudo.

Relativamente ao segundo *dataset* (*dataset* das imagens), este foi construído com dados de conectividade estrutural estimados através da RM de difusão. É de notar que para estimar a conectividade estrutural entre as regiões do atlas AAL (*Automated Anatomical Labelling* - Pacote de *software* e atlas digital do cérebro humano) foi efetuada uma tractografia probabilística. Em [1] estão expostas as 90 *labels* e regiões de interesse incluídas no AAL-atlas.

Tratamento e Exploração de Dados

Este capítulo tem como foco principal descrever todo o processo de exploração e tratamento de dados efetuado. Assim, nesta fase, recorreu-se à biblioteca *Pandas*, tirando proveito das suas diversas funções para efetuar a análise, exploração e tratamento de dados. Conjuntamente com a biblioteca *Pandas*, também se utilizou a biblioteca *NumPy* para manipulação de funções de álgebra linear, como por exemplo transformar a matriz em triângulo superior. Os *datasets* aqui retratados foram fornecidos pelos docentes.

O primeiro *dataset* está dividido em dois conjuntos diferentes, conjunto de *train* e conjunto de *test*, que compreendem informação referente aos sujeitos em estudo e têm propósitos diferentes no desenvolvimento e validação do modelo.

O conjunto *train* compreende um total de 112 entradas, ou seja, contém informação referente a 112 sujeitos, nomeadamente a sua idade, género, educação e identificador, enquanto o conjunto *test* integra informação referente a 28 sujeitos e contém as mesmas *features* que o conjunto anterior salvo a idade. A idade destes sujeitos cobre um intervalo de 13 a 79 anos, com uma média de aproximadamente 44 anos.

Já o segundo *dataset* contém dados de conectividade estrutural.

No *dataset* com dados referentes aos sujeitos em estudo foi efetuado *one hot encoding* na *feature* género de modo a separar o género feminino do género masculino. Após esta separação, verificou-se que o *dataset* é desbalanceado, como é demonstrado na Figura 2. De modo a combater este desbalanceamento, ponderou-se efetuar o balanceamento dos dados, no entanto, como resultaria numa amostra muito pequena descartamos a ideia e não utilizamos o *dataset* balanceado.

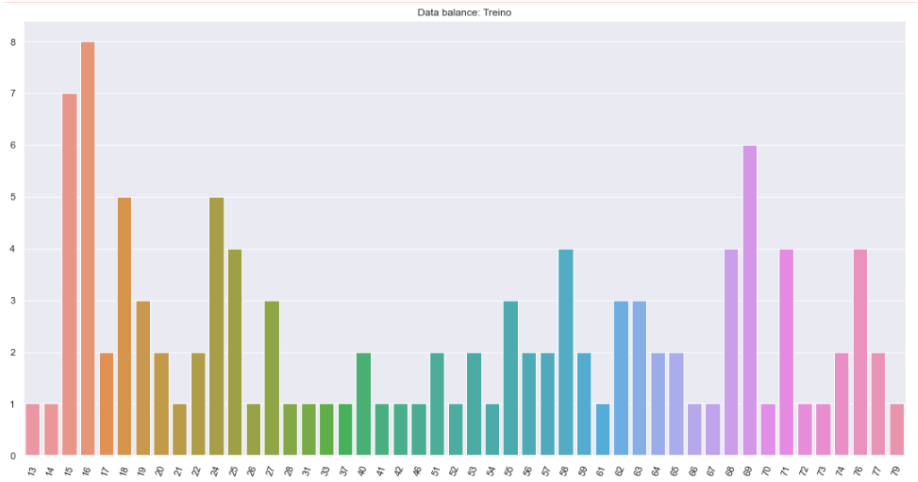
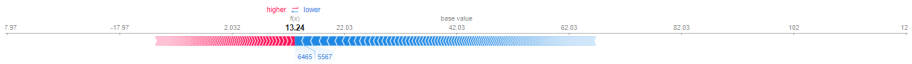


Figura 2: Balanceamento do Dataset.

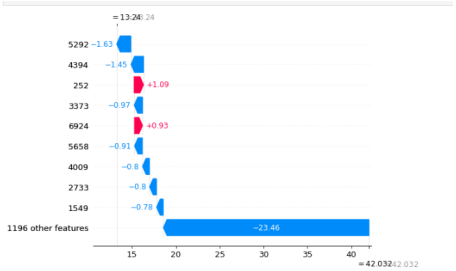
No que diz respeito às matrizes do *dataset* com dados de conectividade

estrutural, estas não foram tratadas como imagens e foram removidas todas as diagonais, permanecendo apenas as *features* do triângulo superior da matriz. É de notar que este tratamento apenas foi efetuado para o segundo modelo desenvolvido pelo grupo descrito detalhadamente no próximo capítulo. Também foram removidas todas as ligações repetidas e *features* que estavam sempre com valor 0. Após este processo, concatenamos os 2 *datasets*, nomeadamente o *dataset* da imagem e o *dataset* com informações dos sujeitos, resultando um *dataset* único com 1206 *features*.

Usufruímos também da ferramenta *SHAP* para obter as *features* mais importantes para prever a idade do cérebro. O processo desenvolvido resultou nos gráficos representados na Figura 3, onde é possível verificar facilmente quais são as *features* com mais impacto (Figura 3b) e a distribuição das importâncias (Figura 3a).



(a) Distribuição das features de acordo com a sua importância.



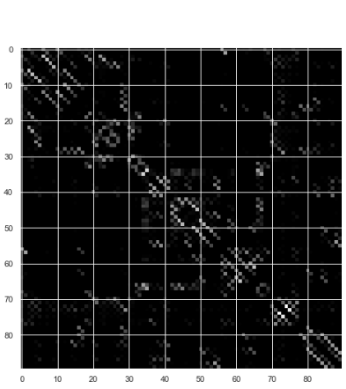
(b) Features com maior impacto.

Figura 3: Gráficos gerados com SHAP.

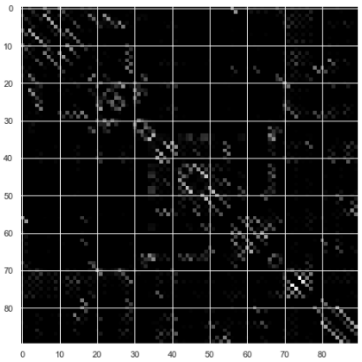
Por fim, tentamos identificar *outliers*, ou seja, quais eram os pacientes para os quais estávamos a obter valores de MAE maiores, possivelmente por serem pacientes com transtornos psicológicos, como Transtorno obsessivo-compulsivo. Aqui identificamos que o maior valor de MAE se encontra na entrada 82, referindo-se esta a um paciente de 76 anos, cujo o nosso modelo previa que tinha apenas 59. Através das respetivas imagens, Figura 4, podemos observar que entre o paciente 82 e o paciente 110, também de 76 anos e cujo o nosso modelo conseguiu prever a sua idade corretamente, existe alguma discrepância nas comunicações do cérebro.

Após uma análise do que poderia provocar este *outlier*, verificou-se que o modelo foi treinado apenas com pacientes do género masculino e o paciente 82 é do género feminino, desta forma, sem mais dados para analisar, a discrepância pode ser explicada pela diferença de género, não deixando de ser curioso que para o nosso modelo uma mulher de 76 anos tem um cérebro tão jovem como um homem de 59 anos.

O segundo maior *outlier* é o paciente 106, cuja a idade é de 74 anos e o



(a) Paciente com erro de previsão elevado.



(b) Paciente com a mesma idade mas que o modelo previu corretamente.

Figura 4: Matrizes de comunicação das diferentes regiões do cérebro.

nosso modelo preveu que tinha 59 anos. Comparando com os restantes pacientes vimos que partilha o mesmo género e grau de educação e, como tal, a diferença estaria nas comunicações cerebrais, sendo provável que este paciente tenha um transtorno psicológico.

Os restantes casos detetados com um grande erro na previsão são os pacientes 56,52,63,86,46,66,74,107. É também de notar que existem poucas amostras das idades apresentadas pelos pacientes acima referidos, com exceção das idades 69 e 24 (paciente 86 e paciente 46), não sendo possível retirar muitas conclusões sobre as previsões obtidas.

Modelo

Este capítulo tem como foco principal descrever o modelo desenvolvido, referenciando as suas características, as várias ferramentas e bibliotecas aplicadas durante o seu desenvolvimento e, por último, especificando todos os hiperparâmetros utilizados para *tuning*.

Metodologia Aplicada

Em todo o processo foi utilizada a linguagem de *script python*, recorrendo-se, ainda, a ferramentas imprescindíveis para auxílio e facilitismo, como bibliotecas e documentação. Com o propósito de maximizar o processo de desenvolvimento do projeto, foram utilizadas bibliotecas direcionadas a *Machine Learning* (ML), que integram algoritmos e ferramentas cruciais para a implementação de redes neuronais. Desta forma, foi utilizado o *tensorflow*, que permite ter um alto nível de abstração das ferramentas ML por meio da *API (Application Programming Interface) Keras*.

Modelação e Tuning

Devido à complexidade do problema e à familiarização com o método escolhido, optamos desde início pela utilização de redes neuronais. Dado que nos foi fornecido um *dataset* constituído por imagens, a nossa primeira abordagem consistiu na implementação de uma rede que receberia 2 *inputs* distintos, onde o primeiro *input* seria as imagens fornecidas e o segundo as restantes *features*. Ao primeiro *input* seriam aplicadas várias *Convultional Layers*, seguidas de *Normalization Layers* e *MaxPooling Layers*. Por fim, aplicamos uma *Flattened Layer* para reduzir a dimensão de forma a que o nosso *output* vá ter a mesma dimensão que o resultado do segundo *input*. Ao segundo *input* foram aplicadas 4 camadas densas com o número de neurónios a variar entre 10 e 50. É importante que a *output layer* dos 2 *inputs* tenha o mesmo número de neurónios no momento em que são concatenadas. Após serem concatenadas aplicamos mais 2 camadas densas e, por fim, uma camada com função de ativação linear e apenas um neurónio. Este processo é apresentado na Figura 5.

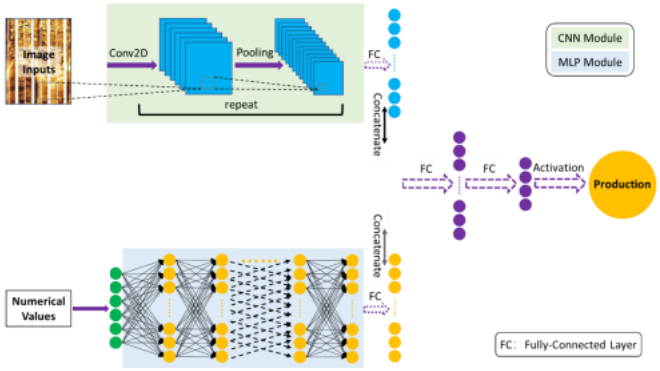


Figura 5: MLP + CNN.

Como este modelo apresentou resultados abaixo das expectativas, a visão que tinhamos perante o problema foi alterada. Dado que as imagens são apenas matrizes de correlação, todas as colunas desta matriz passaram a ser vistas como *features* para o nosso modelo, sendo assim feito o tratamento de dados descrito na secção Tratamento e Exploração de Dados.

Este modelo implementa apenas uma rede *Multilayer Perceptron*(MLP). MLP é uma rede neuronal *feedforward* composta por uma série de camadas completamente conectadas. Como já havia sido mencionado anteriormente, recorremos ao *Tensorflow* e à biblioteca *Keras* durante o desenvolvimento do modelo, sendo aplicados/implementados "elementos" particulares das redes neuronais, como número de *layers*, número de neurónios em cada *layer* e a função de ativação.

Este modelo contém vários hiperparâmetros que definem a estrutura ou topologia. Estes parâmetros incluem número de *hidden Layers*, número de

neurônios por camada, função de ativação, número de *epochs*, função de *loss* e otimizador. É facilmente observável que o ajuste manual de todos estes parâmetros consome bastante tempo. É de notar também que os hiperparâmetros foram ajustados consoante o valor de *Mean Absolute Error* (MAE), de modo a prevenir *overfitting*, isto é, o nosso modelo é generalizado e tem boa performance mesmo em dados desconhecidos.

O modelo final, como se pode observar na Tabela 1 e Figura 10, possui 11 *Hidden Layers*, o número de neurónios por camada que varia entre 512 e 2026, a utilização da função de ativação *ReLU* e na camada de *output linear*. A função de *Loss* utilizada é *MAE*, visto que o problema estava ajustado para o valor de erro. Por fim, o otimizador utilizado é *Adam* com *learning rate* de 0.0001.

Por fim com o objetivo de prevenir o *overfitting* foi implementado *cross validation* com a ajuda da função *KFolds* do *sklearn*.

Hiperparâmetro	**Valor**
Número de Hidden Layers	11
Número de Neurônios	512:2026 (por camada)
Função de Ativação	RelU, Linear
Função de Loss	MAE
Número de Epochs	1000
Optimizer	Adam
Learning Rate	0.0001

Tabela 1: Hiperparâmetros do Modelo.

Resultados

Como referido no capítulo Modelo, a nossa primeira abordagem não obteve resultados próximos do expectável, sendo que no momento da conceção deste trabalho, o *state of the art* em modelos de previsão da idade cerebral tem em média 5 anos de erro.

Com a abordagem anterior, a nossa rede produzia erros de 4.7 durante o treino e de 5.7 na validação, Figura 6.

Na submissão feita no *Kaggle*, Figura 7, a rede obteve erros de 5.2, não ficando muito distante do *state of the art*.

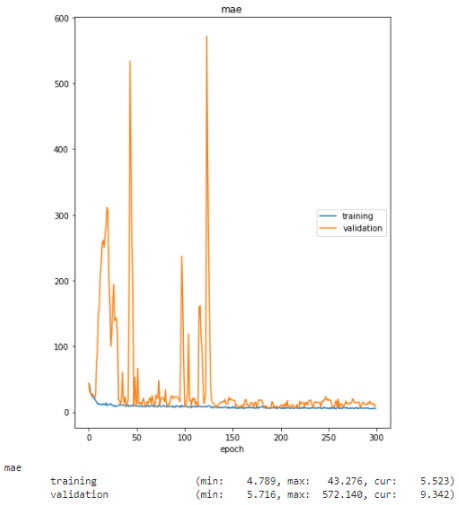


Figura 6: Gráfico de Loss do Modelo MLP + CNN.

Figura 7: Submissão final do Modelo MLP + CNN.

A segunda abordagem obteve resultados muito melhores na fase de treino e validação, o que nos levou a acreditar que poderíamos obter resultados melhores que os do *state of the art*. Como é possível observar na Figura 8, existe uma discrepância entre o erro de treino, 2, e o erro de validação, 5, mas é justificável por termos apenas 5% do *dataset* para validação. Como 5% de validação é baixo, tivemos o cuidado de não aparecer nenhuma idade na validação que não estivesse contida no treino.

Foi também realizado um *KFold* para nos assegurar de que a rede não entrava em *overfitting*.

Este processo resultou numa submissão com erro de 3.5, Figura 9.

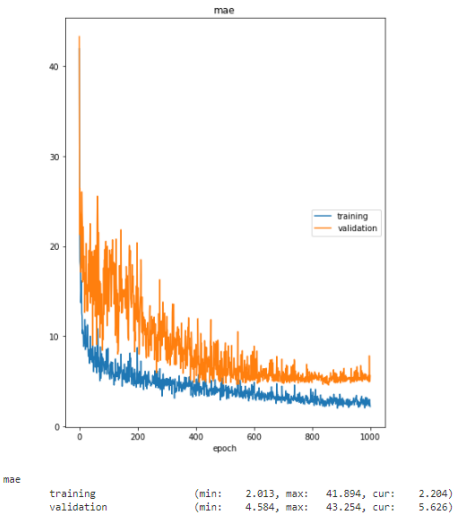


Figura 8: Gráfico de Loss do Modelo MLP.

Figura 9: Submissão final do Modelo MLP.

Conclusão

No presente documento foram apresentadas as etapas de desenvolvimento de um modelo de aprendizagem profunda capaz de prever a idade do cérebro de um indivíduo a partir de características de conectividade estrutural.

Com o desenvolvimento deste trabalho, colocamos em prática fundamentos lecionados nas aulas teóricas, cimentando os nossos conhecimentos relativos a temas como: tratamento de dados, onde estão incluídas as técnicas de manipulação e limpeza para melhorar qualidade de dados, como por exemplo, técnicas de tratamento de *missing values* e técnicas de balanceamento de dados; análise e exploração de dados, cujo objetivo é compreender os dados, as suas características, avaliar a qualidade e encontrar padrões e informações relevantes; modelação, pondo em práticas fundamentos sobre redes neuronais, *tuning*, MLP e CNN; entre outros.

Através da realização deste projeto o grupo adquiriu novas competências no que toca a *Machine Learning*, em particular na implementação de redes neuronais utilizando o *Tensorflow* e a biblioteca *Keras*. É de realçar que também foram usados conceitos aprendidos anteriormente na unidade curricular Dados e Aprendizagem Automática.

Este processo resultou numa submissão com erro de 3.5, superando o valor

médio de 5 anos de erro do *state of the art*. Desta forma, podemos estar perante uma disrupção na tecnologia para previsão da idade cerebral.

Em suma, o grupo encontra-se satisfeito com a solução encontrada e considera que todos os objetivos se encontram cumpridos.

Referências

[1] Zhiliang Liu, Lining Ke, Huafeng Liu, Wenhua Huang, and Zhenghui Hu. Table s1 - changes in topological organization of functional pet brain network with normal aging, Fevereiro 2014.

Anexos

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 1048)              1263888

 dropout (Dropout)           (None, 1048)              0

 dense_1 (Dense)             (None, 1048)              1099352

 dropout_1 (Dropout)         (None, 1048)              0

 dense_2 (Dense)             (None, 512)               537088

 dropout_2 (Dropout)         (None, 512)               0

 dense_3 (Dense)             (None, 512)               262656

 dropout_3 (Dropout)         (None, 512)               0

 dense_4 (Dense)             (None, 1048)              537624

 dropout_4 (Dropout)         (None, 1048)              0

 dense_5 (Dense)             (None, 1048)              1099352

 dropout_5 (Dropout)         (None, 1048)              0

 dense_6 (Dense)             (None, 512)               537088

 dropout_6 (Dropout)         (None, 512)               0

 dense_7 (Dense)             (None, 512)               262656

 dropout_7 (Dropout)         (None, 512)               0

 dense_8 (Dense)             (None, 512)               262656

 dropout_8 (Dropout)         (None, 512)               0

 dense_9 (Dense)             (None, 1048)              537624

 dropout_9 (Dropout)         (None, 1048)              0

 dense_10 (Dense)            (None, 2026)              2125274

 dense_11 (Dense)            (None, 1)                 2027

=================================================================
Total params: 8,527,285
Trainable params: 8,527,285
```

Figura 10: Descrição do MLP.