Nome: Igor Muniz Soares

01.

Projeto: Extração de informações de documentos jurídicos escaneados

Detalhes: O projeto consistia em extrair informações de interesse em documentos jurídicos do cliente em que todos esses documentos estão armazenados em um banco. Fui um dos responsáveis por coletar esses arquivos, construir um dataset, construir e treinar um modelo de OCR para extrair o texto das imagens escaneadas. Ao final da construção do modelo, o mesmo foi integrado na pipeline do cliente para entrada automática dos arquivos (imagens ou pdf), feito deploy e escalonado com o docker, retornando o texto para continuação do fluxo.

02.

Arquivo funcoes_mapa.py

03.

Dropout é uma técnica aplicada em uma camada de redes neurais artificiais que tem como função prevenir uma adaptação exagerada dos dados de treinamento evitando “overfitting”. Funciona basicamente “eliminando” (dropout) neurônios das camadas intermediárias de forma aleatória em cada atualização do treinamento, fazendo com que a rede aumente sua capacidade de generalização.

04.

a)

Trataria esse problema como um problema de regressão linear em que objetivo seria a quantidade de colheita, utilizando os outros dados (tamanho das espigas, cor da folha, chuvas, etc.) como características correlacionadas com o objetivo.

b)

Para avaliar o modelo, é preciso primeiramente separar parte dos dados para avaliação futura (quantidade com relevância estatística), sendo que estes não devem entrar no treinamento do modelo. Dessa forma é possível verificar a acertitividade do modelo e comparar com outros utilizando métricas como RMSE, por exemplo.

c)

Conhecendo melhor o problema, analisaria se os dados possuem a mesma distribuição do problema real e quantidade estatistica suficiente para a tarefa de regressão, assim como a possibilidade de dados de outros anos. Então seguiria o passo 2, separando o dataset em 3 partes, para validação, treinamento e teste, avaliando o modelo gerado com as métricas, até os modelos testados apresentarem uma taxa de erro com pouco desvio e aceitável.

d)

O erro de treinamento bem menor que o erro de teste indica um provável caso de overfitting. Para amenizar esse problema é preciso evitar que o modelo se adapte perfeitamente aos dados de

treinamento e perca sua capacidade de generalizar. Isso por ser feito adicionando mais dados se possível, aplicando técnicas de regularização como o Dropout em redes neurais ou penalização L1 para modelos de regressão linear, remover “features” desnecessárias e/ou criar novas a partir dos dados existentes, e parar o treinamento antes que ele se adapte demais (early stopping).