Racional para as escolhas efetuadas no challenge

20 de novembro de 2023

1 Escolha do modelo

A abordagem mais lógica seria construir um modelo de deteção de objetos com base num modelo de classificação de imagens. Usando um bom classificador de imagens, uma forma simples de detetar objetos consiste em fazer deslizar uma 'janela' pela imagem e classificar se a imagem nessa janela (região recortada da imagem) é do tipo desejado. Parece simples, mas esta solução apresenta roblemas:

1. Por um lado, como saber o tamanho da janela para que contenha sempre o objeto? Isto porque diferentes tipos de objetos, ou até mesmo o mesmo tipo de objeto podem ter tamanhos variados.

2. Por outro lado, o "aspect ratio"(a relação entre altura e largura de uma bounding box), ou seja, é possível a presença de muitos objetos com diferentes formas.

Para contornar estes problemas seria necessário experimentar janelas deslizantes de diferentes tamanhos/formas, o que é se torna computacionalmente intensivo. Na prática, os algoritmos de deteção de objetos predominantes, como R-CNN e Fast(er) R-CNN, seguem uma abordagem em dois passos - primeiro identificam regiões onde se espera encontrar objetos e, em seguida, detetam objetos apenas nessas regiões usando uma convnet. Algoritmos como YOLO (You Only Look Once) e SSD (Single-Shot Detector) adotam uma abordagem totalmente convolucional, em que a rede é capaz de encontrar todos os objetos numa imagem numa única passagem (daí o termo 'single-shot' ou 'look once') através da convnet. Os algoritmos "single-shot"tem uma boa precisão e são mais eficientes/ rápidos na execução, e por essa razão são os mais utilizados. O SSD tem dois componentes:

1. Mapa de extração de features, que extrai as features presentes na imagem. O output da CNN é um mapa que extrai as regiões (features) mais importantes da imagem.

2. Filtros de convolução, que são usados para detetar objetos e contruir bounding boxes em torno de cada deteção.

Uma vez que pretendemos um algoritmo que seja rápido e eficiente, optámos por combinar a arquitetura MobileNet com o Single Shot Detector (SSD).

2 Escolha das imagens

Escolhemos como classe de imagens os "Patos", uma vez que permitem selecionar uma maior diversidade de circunstâncias de deteção. Decidimos aplicar este critério na seleção de imagens, optando por imagens com 1 único vs. diversos patos vs. patos sobrepostos; patos adultos vs.pintos; patos em voo vs. patos em terra vs. patos em água (com reflexo); patos em posição frontal vs. lateral vs. traseira.

Verificamos que os melhores resultados se obtêm com um número reduzido de patos e com features bem distintas do background da imagem. Na 2ª imagem, a presença de reflexo bem definido em água estática conduz à identicação errada de um 2º pato e é, portanto, uma limitação do modelo. Esta limitação não existe na distinção entre patos adultos e pintos, sendo que o modelo consegue bons resultados com ambos desde que haja uma clara separação entre patos. A figura 4 revela outra limitação, já que a rede não é capaz de distinguir entre pintos que estão sobrepostos. O mesmo sucede para patos adultos, com a figura 11. Verificam-se limitações também na capacidade de deteção de patos em voo (em que são reconhecidos apenas 6 de 9 na figura 16. Mais uma vez, é na presença de patos que se sobrepõem cujas features se sobrepõem que a deteção é mais difícil.