Exercícios Deep Learning
Detecção de Objetos

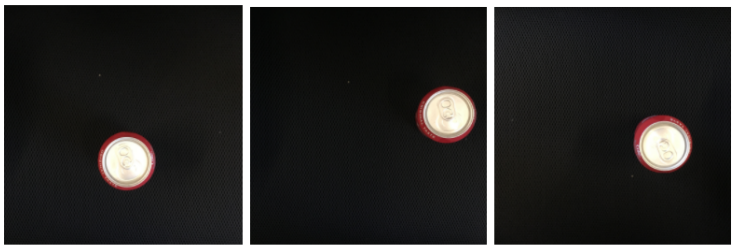
1 Algoritmos de Detecção de Objetos

1) Você está construindo um algoritmo de classificação e localização de objetos de 3 classes. As classes são: pedestre (c = 1), carro (c = 2), motocicleta (c = 3). Qual seria o rótulo aproximado da imagem a seguir? Considere a imagem como sendo um quadrado de lado igual a 1.

2) Continuando com o problema anterior, qual deve ser o rotulo y da imagem abaixo? Lembre-se de que "?" Significa "não se importa", o que significa que a função de perda da rede neural não se importa com o que a rede neural fornece para esse componente da saída.

3) O somatório dos erros quadráticos é uma possível função de perda que pode ser aplicada ao trabalhar com detecção de objetos. Escreva a fórmula de perda considerando os dois casos mostrados nos Exercícios 1 e 2, ou seja, quando na imagem existe um objeto que pertence a uma das classes e quando não há nenhum.

4) Você está trabalhando em uma tarefa de automação de fábrica. Seu sistema verá uma lata de refrigerante descendo por uma correia transportadora, e você quer que ela tire uma foto e decida se (i) há uma lata de refrigerante na imagem e, em caso afirmativo, (ii) sua bounding box. Como a lata de refrigerante é redonda, a bounding box é sempre quadrada e o refrigerante sempre aparece com o mesmo tamanho na imagem. Há no máximo uma lata de refrigerante em cada imagem. Aqui estão algumas imagens típicas no seu conjunto de treinamento:



Qual é o conjunto mais apropriado de unidades de saída para sua rede neural? Considere a função de ativação que poderá ser utilizada nessa tarefa de classificação (i) e a dimensão do vetor de saída bem como o que cada posição representa.

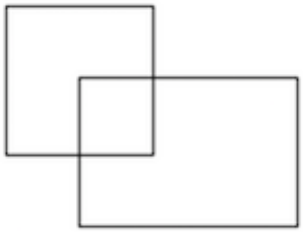
5) Ao treinar um dos sistemas de detecção de objetos, você precisa de um conjunto de treinamento que contenha muitas imagens do(s) objeto(s) que você deseja detectar. Além das imagens e o rótulo dos objetos, explique quais outras informações precisam ser fornecidas para o treinamento supervisionado da rede de detecção de objetos.

2 Janelas Deslizantes

6) Considere um volume $2 \times 2 \times 2$, onde w_{ijc} é a célula (i, j) do canal c, 2 filtros 2×2, onde F_{ijck} é o valor de cada célula (com k como identificador do filtro), e uma função de ativação $g(x)$. Encontre a forma fechada da saída dessa camada após a aplicação dos filtros. Explique como essa camada convolucional é equivalente à uma camada Fully Connected (indique o que seriam os nós, os pesos e suas dimensões).

3 Detecção de Objetos / YOLO

7) Qual é o valor da Interseção sobre a União (IoU) entre estas duas caixas? A caixa superior esquerda é 2x2 e a caixa inferior direita é 2x3. A região sobreposta é 1x1.



8) Suponha que você execute a non-max suppresion nas caixas previstas na imagem abaixo. Os parâmetros usados são: caixas com probabilidade ≤ 0.4 são descartadas, e o limiar de IoU para decidir se duas caixas se sobrepõem é 0.5. Indique quantas e quais caixas permanecerão.

9) Repita o exercício anterior com probabilidade de descarte igual a 0.6.

10) Cite 2 situações nas quais o algoritmo de non-max suppression é falho.

11) No algoritmo YOLO, em tempo de treinamento, como ocorre a detecção do objeto caso ele esteja em mais de uma célula do grid da imagem?

12) Suponha que você esteja usando o YOLO com uma grade 19×19, em um problema de detecção com 20 classes e 5 anchor boxes. Qual é a dimensão do volume de saída da rede?

13) Sabendo que o Yolo é consideravelmente mais rápido que a R-CNN, Fast R-CNN e a Faster R-CNN, por que ele não é o mais utilizado em todos os cenários? Quais são os seus problemas?

14) Por que o YOLO não consegue detectar diferentes objetos muito próximos?

15) Qual a vantagem de se utilizar múltiplas Anchor Boxes em detecção de objetos?

16) Cite duas formas de escolher Anchor Boxes e qual é a vantagem de cada uma delas.

4 Detecção de Objetos / R-CNN

17) Suponha que você esteja aplicando um classificador de janelas deslizantes (implementação não-convolucional). Quais são as consequências de aumentar o stride em termos de acurácia e custo computacional? Qual seria uma boa alternativa para o uso de janelas deslizantes?

18) Qual a principal diferença entre o método R-CNN e o Fast R-CNN. Qual é a vantagem de utilizar o último em relação ao primeiro?

19) Qual a principalmente diferença entre a Fast R-CNN e a Faster R-CNN?

20) Qual o objetivo da camada de "RoI Pooling"? Porque ela é necessária em arquiteturas que utilizam de abordagens baseadas em regiões?

Solução

1) y=[1,0.3,0.7,0.3,0.3,0,1,0], onde:

- 1, no caso a imagem contem um objeto
- Posição b_x a 30% da largura total da imagem
- Posição b_y a 70% da altura total da imagem
- Tamanho total do objeto (carro) de $b_w = 0.3$ em relação a largura total da imagem
- Tamanho total do objeto (carro) de $b_h = 0.3$ em relação a altura total da imagem
- Vetor de classes apenas ativado na classe carro: $c_1 = 0, c_2 = 1, c_3 = 0$

2) y=[0,?,?,?,?,?,?,?,?], nessa imagem não é encontrada nenhuma das classes reconhecidas pelo algoritmo.

3) $L(\hat{y}, y) = (\hat{y_1} - y_1)^2 + (\hat{y_2} - y_2)^2 + ... + (\hat{y_8} - y_8)^2$ se $y_1 = 1$.
$L(\hat{y}, y) = (\hat{y_1} - y_1)^2$ se $y_1 = 0$.

4) Unidade logística, b_x e b_y. A unidade logística vai servir apenas para a tarefa de classificação, enquanto as coordenadas b_x e b_y nos dizem a posição do objeto. Considerando que o refrigerante aparece sempre com o mesmo tamanho, b_h e b_w são desnecessários.

5) As caixas delimitadoras precisam ser fornecidas no conjunto de treinamento.

6) Para cada ω_k de saída: $\omega_k = g(\sum_{i,j,c=1}^{2} w_{ijc} F_{ijck} + b_k)$, onde ω_k são os nós de saída com dimensão 2, w_{ijc} são os nós de entrada com dimensão 8, F_{ijck} são os pesos com dimensão $8 \times 2 = 16$ e b_k é o bias com dimensão 2. Portanto essa camada convolucional é equivalente a uma camada fully connected com 8 nós de entrada e dois de saída.

7) $IoU = \frac{I}{U} = 1/9$, I = $b_1 \cap b_2 = 1$, U = $b_1 \cup b_2 = 9$

8) 5 Caixas.
Algoritmo do Non max-supression:

- Descarte todas as BB com probabilidade menor ou igual a 0.4
- Enquanto ainda houver BB não analisadas: a) Escolha a BB de maior probabilidade; b) Remova qualquer BB que tenha IoU maior que 0.5 com a BB escolhida em (a).

9) 3 Caixas.

10) Quando a bounding box de maior probabilidade não é a mais adequada; Objetos distintos estão muito perto uns dos outros e Falsos positivos.

11) Apenas uma célula - a que contém o centro/ponto médio de um objeto - é responsável por detectar esse objeto, já a anchor box vai além dos limites do grid.

12) Dimensão do volume de saída = (Tamanho da grade) x (Tamanho da grade) x (Número de Anchor Boxes) x (5 + Número de Classes). Dimensão final: 19*19*(5*25), onde: 19*19 para cada quadrado da grade analisada, 5 é a quantidade de anchor boxes, e 25 se refere a 20 classes mais 5 dimensões para p_c, b_x, b_y, b_h, h_w.

13) Alguns problemas do YOLO são: Dificuldade de detectar objetos pequenos devido à limitação do grid, detecção de aspectos com proporções distorcidas, ponderar de forma igual erro em bounding boxes grandes e pequenas.

14) No YOLO, ainda que depois de realizar a divisão da imagem em um grid sejam preditas B bounding boxes cada uma com seu score de confiança, se a interseção entre essas caixas não for muito pequena, apenas um objeto será detectado.

15) O uso de múltiplas anchor boxes permite que uma rede detecte vários objetos, objetos de diferentes escalas e objetos sobrepostos.

16) O primeiro modo é manualmente, a vantagem é que é um método simples. O segundo modo é utilizar o algoritmo k-means, a vantagem é que as caixas serão mais adequadas aos dados.

17) Aumentar o stride aumenta a eficiência computacional, porém diminui a acurácia. Duas possíveis alternativas são: implementação convolucional de sliding window ou não classificar todas as regiões (region propposal).

18) O R-CNN usa o algoritmo de segmentação selective search para propor regiões de interesse, i.e. regiões que possivelmente contém um objeto. Cada uma dessas regiões é achatada (warped) e passada por uma rede convolucional para obter um mapa de features correspondente (um processo muito custoso). O Fast R-CNN passa a imagem uma única vez pela ConvNet, obtendo um mapa de features. As regiões selecionadas a partir da imagem pelo Selective Search são então "recortadas" no mapa de features e, em seguida, tem sua dimensão reduzida através da operação de RoI Pooling. Passar a imagem uma única vez pela ConvNet reduz substancialmente o tempo da detecção de objetos.

19) O gargalo da Fast R-CNN é a proposta de regiões através do selective search. O Faster R-CNN usa uma rede convolucional para propor regiões (region proposal network ou RPN), que pode ser vista como uma rede cujo papel é determinar se há um objeto centrado em um determinado pixel. Isto torna a Faster R-CNN bem mais eficiente.

20) O "RoI Pooling" transforma regiões propostas de tamanhos diferentes em uma lista de matrizes com o mesmo tamanho. Isso torna possível utilizar redes convolucionais para classificar as regiões.