

Data: 02/10/2025
Valor: 24 Pontos
Duração: 1h40m

Prova 2

Aprendizado Profundo – DCC/UFMG
2º Semestre de 2025
Prof. Douglas G. Macharet

NOME: _____ MATRÍCULA: _____

A prova é individual e sem consulta. O entendimento correto das questões faz parte da prova. Se for o caso, deixe bem clara qualquer consideração adicional utilizada em suas respostas.

1) [6 x 1,5 pts] (Geral) Para cada questão abaixo, marque (V) para Verdadeiro ou (F) para Falso, justificando **obrigatoriamente** todas as suas respostas em uma ou duas frases.

(**V**) Em uma camada convolucional, alterar apenas o valor do bias de um filtro não afeta o tamanho (dimensões espaciais) do mapa de ativação de saída.

As dimensões espaciais do mapa de saída são determinadas pelo tamanho da entrada, tamanho do filtro, stride e padding. O bias é um valor escalar somado a cada elemento do mapa de ativação após a convolução, não influenciando suas dimensões.

(**V**) Skip connections em uma ResNet facilitam o treinamento de redes muito profundas ao permitir que o gradiente flua mais diretamente para as camadas iniciais, combatendo vanishing gradient.

As skip connections criam um "atalho" para o fluxo do gradiente, permitindo que ele se propague mais facilmente pelas camadas. Isso ajuda a mitigar o problema do desaparecimento do gradiente (vanishing gradient) em redes muito profundas.

(**F**) Uma camada de Max Pooling com filtro 2x2 e stride=1 aplicada a um volume de 15x15x32 resultará em um volume de 7x7x32.

Max Pooling 2x2 com stride=1 em 15x15x32 gera 14x14x32, não 7x7x32.

(**F**) Em uma CNN, as camadas iniciais (mais próximas da entrada) tendem a aprender features complexas e específicas do objeto (e.g., "olhos", "rodas"), enquanto as camadas finais aprendem features genéricas como bordas e texturas.

Ocorre o oposto. As camadas iniciais aprendem features simples e genéricas (bordas, cores, texturas), enquanto as camadas mais profundas (finais) combinam essas features para detectar padrões mais complexos e específicos do objeto (olhos, rodas, rostos).

(**F**) Em um detector YOLO configurado para 20 classes, com um grid de 17x17 e 4 anchor boxes, a profundidade do tensor de saída será de 84.

A profundidade do tensor de saída do YOLO é calculada por: $B \times (5 + C)$, onde B é o número de anchor boxes e C é o número de classes. Neste caso, seria $4 \times (5+20) = 4 \times 25 = 100$.

(**F**) Na tarefa de detecção de objetos, o "Intersection over Union" (IoU) é uma métrica usada para medir a velocidade de inferência de um modelo como o YOLO.

O IoU é uma métrica de acurácia, não de velocidade. Ele mede o grau de sobreposição entre a caixa delimitadora (bounding box) prevista pelo modelo e a caixa real (ground truth).

2) [2,0 + 3,0 + 2,0 pts] (CNNs) Responda às questões abaixo.

a) Em operações de convolução, é comum definir um stride (passo) maior que 1. Descreva pelo menos duas razões ou benefícios distintos para essa escolha.

- * **Redução da Dimensão Espacial:** Um stride maior que 1 reduz a altura e a largura dos mapas de características. Isso diminui a quantidade de computação e o número de parâmetros nas camadas subsequentes.
- * **Aumento do Campo Receptivo:** Ao "pular" posições, as camadas seguintes conseguem agregar informações de uma área maior da entrada original mais rapidamente, aumentando o campo receptivo de forma mais agressiva.

b) Descreva como uma camada convolucional pode ser parametrizada (filtros, bias, padding, etc.) para que ela se comporte como uma função identidade, ou seja, para que o volume de saída seja exatamente igual ao volume de entrada. Considere uma entrada de $28 \times 28 \times 10$.

- * **Número de Filtros:** 10 (para manter a mesma profundidade de saída).
- * **Tamanho do Filtro:** 1×1 .
- * **Stride:** 1.
- * **Padding:** 0.
- * **Pesos configurados como matriz identidade para cada canal.**
- * **Bias:** Todos os biases devem ser 0.

c) Considere a imagem de entrada I e o filtro f abaixo. Calcule o mapa de ativação resultante da convolução de I com f, assumindo stride=2, padding=0 e que o filtro não é invertido durante a operação. Deixe explícito os cálculos que foram feitos.

-1	-2	0	1
-2	0	2	0
1	1	0	-3
0	2	1	2

Imagem I (4x4)

-2	0
0	2

Filtro f (2x2)

Região I[0:2,0:2] = $\begin{bmatrix} -1 & -2 \\ -2 & 0 \end{bmatrix}$ → soma = $(-1)(-2) + (-2)(0) + (-2)(0) + (0)(2) = 2$
Região I[0:2,2:4] = $\begin{bmatrix} 0 & 1 \\ 2 & 0 \end{bmatrix}$ → soma = $(0)(-2) + (1)(0) + (2)(0) + (0)(2) = 0$
Região I[2:4,0:2] = $\begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$ → soma = $(1)(-2) + (1)(0) + (0)(0) + (2)(2) = -2 + 4 = 2$
Região I[2:4,2:4] = $\begin{bmatrix} 0 & -3 \\ 1 & 2 \end{bmatrix}$ → soma = $(0)(-2) + (-3)(0) + (1)(0) + (2)(2) = 4$

Mapa resultante (2x2):

$\begin{bmatrix} 2 & 0 \\ 2 & 4 \end{bmatrix}$

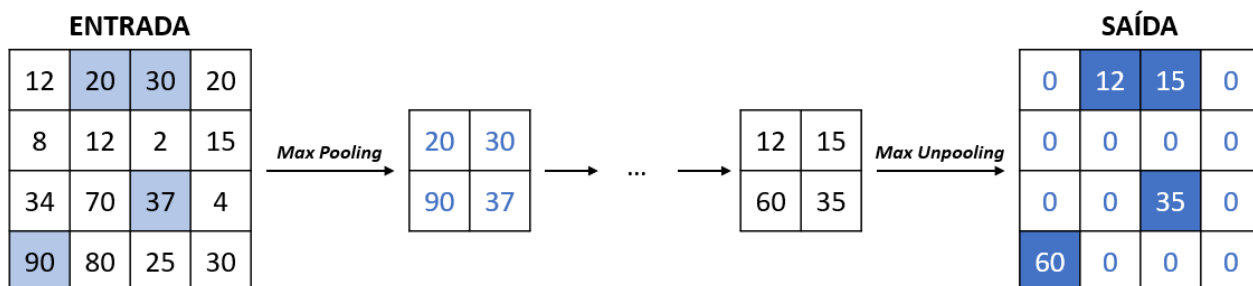
3) [3,0 + 2,0 pts] (Segmentação) Considerando a parte de segmentação semântica responda.

a) Por que a tarefa de segmentação semântica é considerada uma tarefa de predição densa? Explique brevemente por que arquiteturas totalmente convolucionais são frequentemente utilizadas nesse contexto, destaque os diferentes aspectos impactados e beneficiados por essa modelagem.

A tarefa de segmentação semântica é considerada de predição densa porque requer uma classificação para cada pixel da imagem, gerando um mapa de saída com a mesma resolução espacial da entrada, onde cada pixel é rotulado com uma classe.

Arquiteturas totalmente convolucionais (Fully Convolutional Networks - FCNs) são frequentemente utilizadas pois preservam a estrutura espacial, são eficientes ao processar imagens inteiras e permitem flexibilidade no upsampling para recuperar a resolução original.

b) Aplique a operação de *Max Pooling* (filtro 2x2, stride 2) e em seguida de *Max Unpooling*.



4) [3,0 pts] (Treinamento) Para que serve o processo de *data augmentation*? Dê um exemplo de uma estratégia de aumento de dados que seria útil para classificar imagens de gatos vs. cães, mas não para classificar dígitos manuscritos. Explique brevemente sua resposta.

- Serve para aumentar artificialmente a quantidade e diversidade dos dados de treinamento
- Melhora a generalização e robustez da arquitetura
- Possibilidade de reduzir overfitting
- Rotações/Espelhamento/Zoom
 - Pode haver mudanças de classes (pode ser tratado/utilizado)
 - Pode gerar entradas inválidas (deve ser removido)