CS 512 – Assignment 4

Detection and segmentation

Due by 4/10/2023

Review questions

1. Convolution layers

   (a) Let I be a 4×4 RGB image where the R channel is all 1-s and G channel is all 2-s. The B channel has a value of 1 in its first row, a value of 2 in its second row, a value of 3 in its third row, and a value of 4 in its 4th row. Compute the convolution of this image with a 3×3filter having all ones without zero padding.

   (b) Repeat the previous question with zero padding.

   (c) Repeat the previous question when using dilated (atrous) convolution with a dilation rate of 2.

   (d) Explain the template matching interpretation of convolution.

   (e) Explain how multiple scale analysis can be achieved with a fixed window size (using a pyramid).

   (f) Explain how to compensate for spatial resolution decrease using depth (number of channels) and the purpose for doing so.

   (g) Given a $128 \times 128 \times 32$ tensor and 16 convolution filters of size $3 \times 3 \times 32$, what with be the size of the resulting tensor when convolving without zero padding.

   (h) Repeat the previous question when using a stride of 2.

   (i) Explain how the number of channels can be reduced using a 1×1 convolution.

   (j) Explain the interpretation of convolution layers and the difference between early and deeper convolution layers.

   (k) Let I be an image as in question 1. Write the result obtained using max pooling with a 2×2 filter with a stride of 2.

   (l) Explain the purpose of pooling.

   (m) Explain the purpose of data augmentation and when it is most useful.

2. CNNs

   (a) Explain the purpose of transfer learning and when it is most useful.

   (b) Explain the need for freezing the coefficients of the pre-trained network.

(c) Explain how he coefficients of a pre-trained network can be fine-tuned.

(d) Explain the purpose of inception blocks. Describe the solution employed in GoogleNet to address vanishing gradients in a deep network.

(e) Explain the advantage of residual blocks. Include in your description an explanation of how residual blocks assist with vanishing gradients.

(f) Explain how DenseNet is constructed. Describe the way DenseNet controls complexity.

(g) Given an image with three channels where the first has all 1's, the second has all 2's and the third has all 3's, compute the result of a convolution with a $3x3$ filter having 1's in its first layer, 2's in its second layer, and 3's in its third layer. Repeat the computation when using depth-wise separable convolution.

(h) Describe the ways by which MobileNets make computations faster.

3. Object detection

(a) Explain the two tasks that need to be achieved in object detection.

(b) Given a detected object with a bounding box defined by the corners $(2, 2)$ and $(6, 6)$ and a ground-truth object bounding box defined by corners $(3, 3)$ and $(7, 7)$, compute the IoU similarity metric and Jaccard distance. Assume the coordinates of the bounding boxes are given by pixels.

(c) You are given a dataset of images where some have faces in them. You train an object detection algorithms that produces a detection box and the probability of having a face in the box for each image. Your goal is to compute $AP_{0.5}$ (average precision with 0.5 IoU threshold) for your detection results. By varying the confidence threshold of the probability score you obtain the following precision-recall (p, r) pairs: (1,0), (1,0.2), (0.6,0.4), (0.6,0.6), (0, 0.8), (0,1). Compute $AP_{0.5}$ using the information provided.

(d) Explain why detection box coordinates are normalized to be between 0 and 1.

(e) Explain the different terms in the loss function needed for an object detection network.

(f) Given a grid cell object detection with a $3x3$ grid, write the size of the output tensor for the algorithm assuming 10 detection boxes at each cell location.

(g) Explain the difference between single-shot and two-shot approaches.

(h) Describe the different terms in the loss function of the YOLO object detector.

(i) Explain how ROI-pooling is done and the purpose for it.

(j) Explain non-maxiumum suppression in the context of object detection and the need for it.

(k) Explain the 3 loss terms in mask RCNN.

4. Semantic segmentation

(a) Explain the difference between semantic segmentation and instance segmentation.

(b) Given a $5x5$ image and a $3x3$ filter, compute the size of the matrix that can multiply the vectorized image (1D) to produce the convolution results.

(c) Compute the size of the transpose convolution matrix from the previous question.

(d) Explain the need for skip connections in U-net and the way in which the information is propagated along skip connections.

(e) Explain the DeepLab network architecture.

(f) Explain the metric used for evaluating semantic segmentation results.

Programming questions

- In this assignment you need to implement a basic Convolutional Neural Network for classification. Your implementation needs to use a GPU framework. Specifically, Keras or TensorFlow or PyTorch. If you do not have access to GPU on your computer or elsewhere you can create a free account on google colab: https://colab.research.google.com/notebooks/welcome.ipynb

1. Semantic segmentation

   (a) Download the Oxford pet dataset `https://www.robots.ox.ac.uk/~vgg/data/pets/`

   (b) Convert the cat/dog breed labels to to category cat/dog labels.

   (c) Write a function to visualize the segmentation mask

   (d) Split the dataset to training, validation, and test subsets.

   (e) Train a simple convolutional neural network for supervised semantic segmentation **without** skip connections and evaluate its performance. Plot the training and validation loss and evaluation metric as a function of epochs. Visualize some inference results.

   (f) Train a simple convolutional neural network for supervised semantic segmentation **with** skip connections as in Unet. Evaluate and visualize the results as before.

   (g) Use the pretrained `TFSegformerForSemanticSegmentation` model from the `transformers` module and continue to train it. Evaluate and visualize the results as before.

2. Object detection

   (a) Define a bounding box for each object in the cats/dogs data from the previous question.

   (b) Download a pre-trained YOLO model `https://modelzoo.co/model/keras-yolov3`

   (c) Convert the weights to a Keras compatible file `.h5`

   (d) Load the model in a program. Load an image and normalize it. Apply the model to predict results and show the detection results using bounding boxes drawn on the image.

   (e) Evaluate the performance of the model on the cats/dogs dataset using the known labels. Include in your evaluation computation of $mAP_{0.25}$, $mAP_{0.5}$, $mAP_{0.75}$, and $mAP_{0.95}$.

Submission instructions

Please follow the submission instruction of assignment 1.