cs2_business_problem

July 6, 2021

1 Medical Image Segmentation

2 1. Business Problem

2.1 1.1. Description

- **Data Source:** https://www.kaggle.com/c/data-science-bowl-2018
- **Problem Statement:** Identify the nuclei in the images of the cells.
- **Why?** Identifying the cells' nuclei is the starting point for most analyses because most of the human body's 30 trillion cells contain a nucleus full of DNA, the genetic code that programs each cell. Identifying nuclei allows researchers to identify each individual cell in a sample, and by measuring how cells react to various treatments, the researcher can understand the underlying biological processes at work.

2.2 1.2. Source/Useful Links

- https://arxiv.org/abs/2102.10662v1
- https://towardsdatascience.com/metrics-to-evaluate-your-semantic-segmentation-model-6bcb99639aa2
- https://jinglescode.github.io/2019/11/07/biomedical-image-segmentation-u-net/

2.3 1.3. Real World / Business Constraints

- No latency requirements as such. But the model should also not take hours for segmentation.
- Cost of incorrect segmentation is high because it may fail to identify the nucleus correctly, which will have further consequences in further tasks.
- Model should be good in generalization, it should not be overfitted.

3 2. Problem Formulation

3.1 2.1. Data

- **Source:** https://www.kaggle.com/c/data-science-bowl-2018/overview
- This dataset contains a large number of segmented nuclei images.
- Each image is represented by an associated `ImageId`. Files belonging to an image are contained in a folder with this `ImageId`. Within this folder are two subfolders:
  - `images` contains the image file.

– `masks` contains the segmented masks of each nucleus. This folder is only included in the training set. Each mask contains one nucleus. Masks are not allowed to overlap (no pixel belongs to two masks)

3.2 2.2. Type of Problem

Image Segmentation We have to identify each nucleus present in the image of cells.

3.3 2.3. Performance Metrics

Since this is an Image segmentation task, There are two most commonly used metrics:

- **Intersection over Union:** IoU is the area of overlap between the predicted segmentation and the ground truth divided by the area of union between the predicted segmentation and the ground truth.This metric ranges from 0–1 (0–100%) with 0 signifying no overlap and 1 signifying perfectly overlapping segmentation.

- **Dice Coefficient:** Dice Coefficient is 2 * the Area of Overlap divided by the total number of pixels in both images. The Dice coefficient is very similar to the IoU. They are positively correlated, meaning if one says model A is better than model B at segmenting an image, then the other will say the same. Like the IoU, they both range from 0 to 1, with 1 signifying the greatest similarity between predicted and truth.