# Segmenting vertebrae in x-ray scans using Mask R-CNN

Merim Bungur BEE

**ARTICLE HISTORY**

Compiled January 17, 2019

**ABSTRACT**

We present a proof of concept how framework for object instance segmentation, Mask R-CNN, can be used to successfully detect and extract individual spinal vertebrae from low-dose, low quality x-ray scans. The results show that even considerably smaller sample size (approximately 2000 images) can be used to train the deep neural network and still achieve good accuracy (more than 90%).

Email: merim.bungur@infinitymesh.com (Merim Bungur BEE)

## 1. Introduction

Artificial disk replacement surgery is a complex procedure which also carries many risks for the doctors performing the procedure due to radiation exposure, all the more so because same team of people are involved in many different surgeries of the same type. Multiple x-ray scans have to be taken during surgery to have clear visibility of the procedure and to reduce risks of surgery complications, especially since the procedure is performed in proximity to the patients spinal cord. To reduce the radiation exposure, low-dose scans must be taken during the surgery, but the trade off is lower image quality.

Currently there are many image enhancement techniques that are used to solve the problem of low medical image quality, but a large portion of them is applied on the whole image. Spinal vertebrae have individual freedom of movement as well as group movement, so it would be more feasible to apply image analysis techniques on individual vertebrae rather than the whole x-ray scan.

In this paper, we will describe how framework for object instance segmentation can be leveraged to identify individual vertebrae. We hope this demonstrates a proof of concept that instance segmentation should be used in situations where different areas of image have to be analyzed and enhanced individually independent to the rest of the image.

## 2. Methods

Overall strategy was conceptually simple: Use framework for object instance segmentation to recognize areas, or zones of interest (ZOI) which contain one vertebra each, on low-dose x-ray scan (lower dosage of radiation results in lower image quality). Output from this process would be a collection of ZOI's recognized on the low-dose scan. We would then repeat the same process on the initial high-dose x-ray scan which is usually taken before the disk-replacement procedure begins. This would conclude the proof of concept, demonstrating feasibility of using neural network and object segmentation framework to assist disk replacement surgery procedure. The client already had the technology to locate and replace ZOI's on low-dose scan with corresponding ZOI's from the high-dose scan with the usage of 2D geometric transformations (only translation, rotation and scaling is allowed). End result would be a picture with much better visual quality which could be used by the surgeons performing the procedure.

### *2.1. Instance Segmentation*

Instance segmentation *[Abdulla (2018a)]* is the task of identifying object outlines at the pixel level. Compared to similar computer vision tasks, it's one of the hardest possible vision tasks.

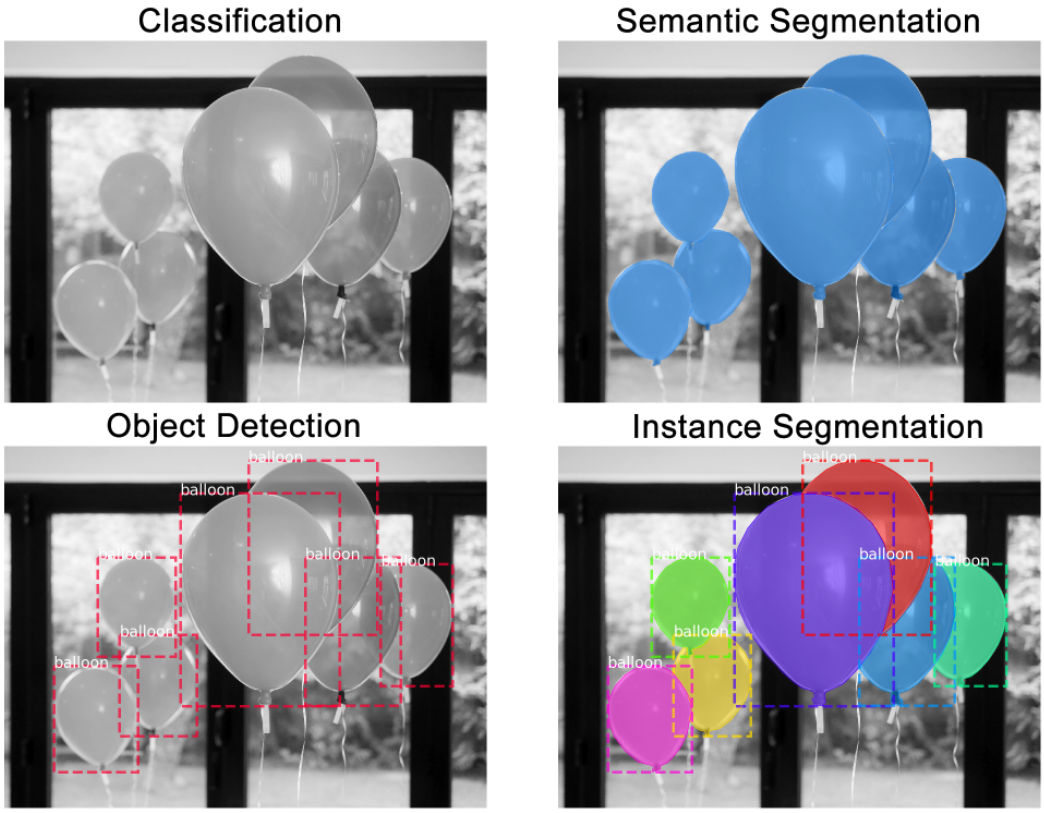


**Figure 1.** Tasks in computer vision

To help you understand instance segmentation, consider the following example tasks in order of complexity:

(1) Classification: There is a balloon in this image.

(2) Semantic Segmentation: These are all the balloon pixels.

(3) Object Detection: There are 7 balloons in this image at these locations. We're starting to account for objects that overlap.

(4) Instance Segmentation: There are 7 balloons at these locations, and these are the pixels that belong to each one.

### *2.2. Mask R-CNN*

Mask Regional Convolutional Neural Network *[He et al. (2018)]* (Mask R-CNN) is a conceptually simple, flexible, and general framework for object instance segmentation. Without bells and whistles, Mask R-CNN outperforms all existing, single-model entries on every task, including the COCO 2016 challenge winners. It detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance. The method, called Mask R-CNN, extends Faster R-CNN *[Ren et al. (2016)]*, a unified network for object detection, by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition. Mask R-CNN is simple to train and easy to generalize to other tasks, such as allowing you to estimate human poses in the same framework. Mask R-CNN is a two stage framework which consists of a deep fully convolutional neural network that is used for feature extraction over an entire image - *backbone*, and the *network head* for bounding-box recognition and mask prediction.

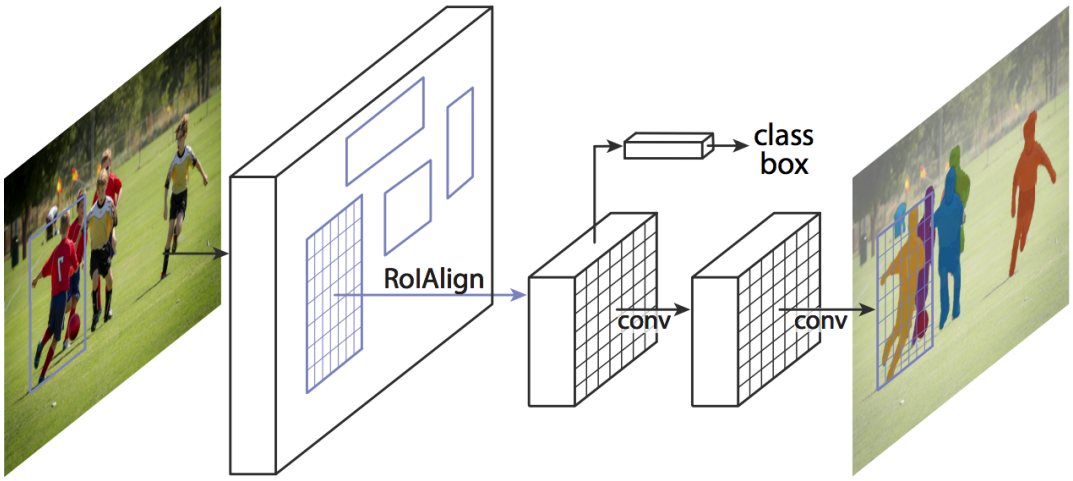


**Figure 2.** Mask R-CNN framework

### *2.3. Backbone*

Standard convolutional neural network, ResNet50 or ResNet101 *[He et al. (2015)]*, serves as a feature extractor. Early layers of network serve to detect low level features (like edges and corners), and later layers gradually and successively detect higher level features such as types of objects on the image. Passing through the backbone network, the image is converted from 1024x1024px x 3 (RGB) to a feature map of shape 32x32x2048. This feature map becomes the input for the following stages.

#### *2.3.1. Feature Pyramid Network*

Feature Pyramid Network *[Lin et al. (2017)]* improves upon the backbone and adds a second pyramid that takes the high level features from the first pyramid and passes them down to lower layers to allow every level to have access to both lower and higher lever features.

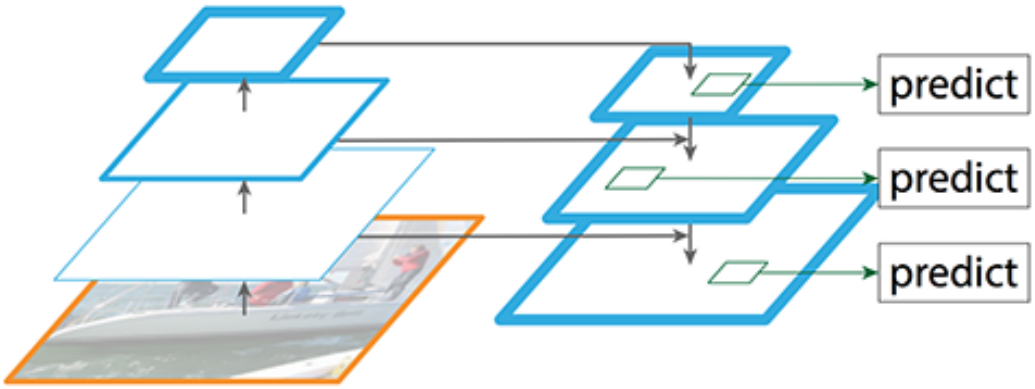


**Figure 3.** Feature Pyramid Network

### *2.4. Network Head*

Network head for bounding-box recognition and mask prediction consists of multi-stage pipeline to obtain the bounding box and object mask:

(1) Region Proposal Network (RPN) - RPN is a lightweight neural network that scans the image in a sliding-window fashion and finds areas which contain objects. Network handles the sliding window by scanning all regions in parallel (on a GPU).

(2) ROI Pooling - Refers to cropping a part of feature map and resizing it to a fixed size for different regions of interest (ROI)

(3) ROI Classifier and Bounding Box Regressor - In this stage ROI's are analyzed and classes are assigned to each ROI while bounding box location and size is further refined.

(4) Segmentation Masks - The mask branch is a convolutional network that takes the positive regions selected by ROI classifier and generates masks for them. Those predicted masks are then scaled up to the size of ROI bounding box and that finalizes the process.

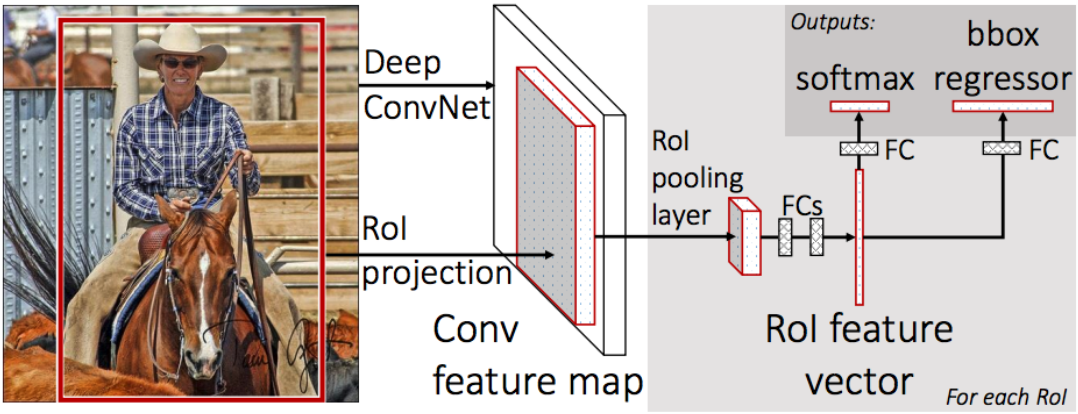


**Figure 4.** Neural network head

We opted to use open-sourced implementation of Mask-RCNN *[Abdulla (2018b)]* on Python 3, Keras, and TensorFlow. The model generates bounding boxes and segmentation masks for each instance of an object in the image.

### *2.5. Training Dataset*

Pre-trained weights were used as a starting point to train our own variation of the network. We used the existing training and evaluation code located on the repository *[Abdulla (2018b)]*. First we needed to classify various segments of the spinal column and spinal vertebrae that were of interest to us:

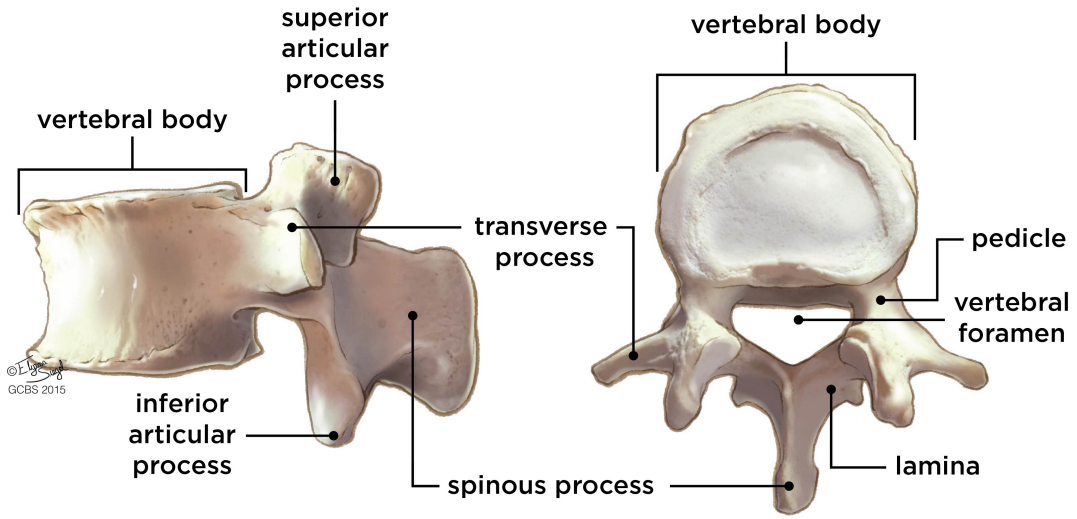


**Figure 5.** Vertebral anatomy

(1) Body
(2) Spinous processes
(3) Pedicles
(4) Transverse processes

Existing pre-trained dataset did not contain the classes of spinal vertebrae that we defined, but since it was trained on over 120K images, trained weights have already learned a lot of the features common in natural images which helps. Client provided us with a set of scans which we transformed into Common Object In Context (COCO) *[Lin et al. (2015)]* dataset of scans which is necessary step for training weights. We used one portion of the dataset to retrain the weights (approximately 2000 images). Due to smaller sample size, it was even more convenient to use a pre-trained weight set to achieve good accuracy. Other portion of images was used to test out the results. After achieving a satisfactory level of accuracy (everything above 90% was acceptable) we added in the code to extract masks after the mask evaluation process was completed.

## 3. Results

Using OpenCV library for python we cross-referenced the mask locations on the unprocessed images and extracted the pixels underneath the masks, thus getting a collection of images containing individual ZOI's. The whole process was a success, and our results would serve as an input to the clients existing framework which already had capabilities to perform 2D geometric transformations and image enhancements. Results we obtained serve as a proof of concept that you can use Mask R-CNN to detect individual vertebrae(ZOI) on a scan, and then use different alignment transforms for different vertebrae, find each correct alignment and display the result to the doctor.

## 4. Discussion

To guarantee best results, there are some key conditions that must be met, potential issues than can arise, and potential future improvements which we will address now:

(1) When capturing x-ray scans, there can be absolutely no spine torsion - Any similarity algorithm which compares two images must be based on 2d geometric transformations (rotation, translation and scaling). Since spine torsion can be considered as a 3d transformation, it cannot be permitted to happen. This is usually ensured by the nature of process of taking x-ray scans (patient is placed on a table and remains stationary while scans are being captured).

(2) Neural network (NN) would have to be trained on a bigger sample - Before developing a prototype device which would implement procedures described in this article, bigger sample would have to be used for NN training, and after that Mask R-CNN detection accuracy would have to be thoroughly analyzed. Failure to recognize some parts of vertebrae could be a critical error because this would interfere with the similarity algorithms, and could even produce false results, so this would have to be investigated separately.

(3) Special detection procedures if ZOI is not recognized by the framework - Reality of the matter is that some low-dose scans are too grainy for the Mask R-CNN to produce any results. Some ZOI would not be recognized, and separate algorithms must be written to deal with these situations.

## 5. Conclusion

This paper explains how frameworks for object instance segmentation on images can be leveraged in medical image analysis. We presented and discussed the application of Mask R-CNN framework in x-ray scan analysis. Process of deep neural network training is explained, and advantages of using pre-trained weights to compensate for lower sample size is discussed. Depending on operative procedure complexity and risks surrounding the procedure, different accuracy constraints can be imposed on the object segmentation framework, and our agreed upon acceptance criteria for this proof of concept project was 90%. Preconditions, potential issues and future improvements are also described.

## References


Abdulla, W. (2018a). Instance Segmentation with Mask R-CNN and TensorFlow.

Abdulla, W. (2018b). Mask R-CNN Implementation. Mask R-CNN implementation on Keras and Tensorflow.

He, K., Gkioxari, G., Dollar, P., and Girshick, R. (2018). Mask R-CNN.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition.

Lin, T.-Y., Dollar´, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). Feature Pyramid Networks for Object Detection.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollar, P. (2015). Microsoft COCO: Common Objects in Context.

Ren, S., He, K., Girshick, R., and Sun, J. (2016). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.