Review on “Mask R-CNN”

Authors: K. He, G. Gkioxari, P. Dollar, R. Girshick

# Short Summary

The authors of this papers propose “Mask R-CNN”, an extension of “Faster R-CNN” which allows for instance segmentation. Their proposal is to add a ‘mask branch’ to the existing model which allows the model to predict segmentation masks in addition to classification and bounding box regression. The mask branch is comprised of a small fully convolutional network (FCN) which is applied to each region of interest. Their experiments show that this new proposal yields state-of-the-art performance, while operating at 5 fps, on instance segmentation with minimal computational overhead. It also seems to be simple to integrate into existing networks with a Faster R-CNN framework.

In predicting the segmentation mask, the authors found that pixel-to-pixel alignment between input and outputs is critical and not considered in existing instance frameworks. Thus, the authors propose RoIAlign, a quantization-free layer intended to replace RoIPool, which preserves spatial locations. It is shown that this aligned strategy improves mask accuracy by a relative 10-50%. The authors also critically observed that decoupling mask and class prediction, that is using binary masks for each class independently, improved results in instance segmentation.

The final model surpasses state-of-the-art on the COCO instance segmentation task, can train within 1-2 days with an 8-GPU machine, and is shown to be extensible via application to the task of human pose estimation with minimal modification.

# Main Contributions

* Mask R-CNN model architecture that extends Faster R-CNN.
* RoIAlign layer that replaces RoIPool and demonstrates fairly significant performance gains (10-50% improve in relative performance).
* Demonstrating that a per-pixel sigmoid and average binary cross-entropy loss improves segmentation results over conventional coupled methods (softmax + average multi-nomial cross-entropy)
* Extending the model to human pose estimation demonstrating extensibility

# High-Level Evaluation of Paper

The paper does an excellent job in setting up the problem and describing their proposals before delving into experiments. It is clear what they’ve proposed and there is sufficient intuition as to why they suspect it may work. I appreciate that paper maintains focus on Mask R-CNN and describes extensions (RoIAlign + binary masks) as part of addressing the central task of improving performance at instance segmentation. Aiding images, such as the model overview and segmentation masks, are helpful but I did not get an intuitive sense of how RoIAlign works from the provided figure. In terms of experiments, the paper goes into extensive detail about how different architectures perform and show where their model excels; however, limitations other than the small computational overhead are never discussed in the paper. Additionally, I found the section on Timing difficult to understand; I’m unsure what the takeaway is supposed to be. The paper is also missing a conclusion that summarizes the results. Furthermore, the abstract does not go into much detail regarding RoIAlign and the binary masks which means that one would need to read the entire paper to learn these insights. The missing conclusion and minimal abstract are odd decisions by the authors. Overall, the paper makes novel proposals and identifies weaknesses with current methods. It goes on to demonstrate that its proposals improve upon the existing methods and shows the extensibility of their model. I think these strengths outweigh the weaknesses in the paper.

# Discussion on Evaluation Methodology

For evaluation, Average Precision is used at varying IoU levels and scales. Each experiment is summarized in a table of results showing its improvement over comparable models. I like that the authors invested the time to not only quantify results on the primary task of instance segmentation, but to perform Ablation experiments as well showing the impact of each of their modifications. It makes it quite clear how using RoIAlign and binary masks are beneficial even outside the context of Mask R-CNN. The paper also explains well which training and testing environments were used along with the hardware making it easy to potentially reproduce their results. The visual comparison as well, on instance segmentation, shows how their model corrects issues found in the previous state-of-the-art. Overall, I think the main paper did an excellent job on Evaluation methodology.

# Possible Directions for Future Work

The paper does not detail extensions but rather provides them as appendices in the paper. These appendices describe additional experiments that extend the Mask R-CNN to the CityScapes dataset and discuss various model enhancements to improve performance. An interesting direction mentioned here is data distillation, where the model is able to self-train from unlabelled data. It is shown that this improved performance with the model, I’d be interested to see whether this technique benefits Mask R-CNN more so than other models.