Multiple Myeloma Cancer Cell Instance Segmentation

Student Name: Dikshant Sagar
Roll Number: 2017338

BTP report submitted in partial fulfillment of the requirements for the Degree of B.Tech. in Computer Science & Social Sciences on May 11, 2021

BTP Track: Research

BTP Advisor
Dr. Anubha Gupta
Shubham Goswami(PhD)

Indraprastha Institute of Information Technology
New Delhi
Abstract

Images remain the largest data source in the field of healthcare. But at the same time, they are the most difficult to analyze. More than often, these images are analyzed by human experts such as pathologists and physicians. But due to considerable variation in pathology and the potential fatigue of human experts, an automated solution is much needed. The recent advancement in Deep learning could help us achieve an efficient and economical solution for the same. In this research project, we focus on developing Deep Learning based solution for detecting Multiple Myeloma cancer cells using an Object Detection and Instance Segmentation System. We explore multiple existing solutions and architectures for the task of Object Detection and Instance Segmentation and try to leverage them and come up with a novel architecture to achieve comparable and competitive performance on the required task. To train our model to detect and segment Multiple Myeloma cancer cells, we utilise a dataset curated by us using microscopic images of cell slides provided by Dr. Ritu Gupta (Prof., Dept. of Oncology AIIMS).

Keywords: image analysis, machine learning, deep learning, bio-medical imaging, computer vision, object detection, instance segmentation, multiple myeloma.
Acknowledgments

I thank my advisor, Dr. Anubha Gupta, for giving me an opportunity to work under her guidance. Her advice and constant support have enabled me to think critically about the domain of research and learn to come up with concise, effective solutions to the practical problem.

I thank Dr. Ritu Gupta (Prof., Dept. of Oncology AIIMS), for providing a large databank of patient cell slide images, which helped us create a benchmark dataset in the field of Multiple Myeloma Cancer Detection.

I also thank Mr. Shubham Goswami (PhD, IIIT Delhi) for his immense contribution to the project. His vast knowledge of the various methods pertaining to Bio-Medical Computer Vision and readiness to brainstorm solutions for the challenges along the way have been a significant driver to this project.

Finally, I thank Mr. Bharat Giddwani (Deep Learning Solutions Architect, NVIDIA) for providing access to DGX Workstations for us to be able to train large scale deep learning models.

Work Distribution

Monsoon Semester 2020 (4 credits)

1. Dataset: Annotations and Auxiliary Dataset Exploration (Sept - Nov, 2020)
2. Existing Deep Learning Architectures: Literature Survey (Nov - Dec, 2020)
3. Proposed Architecture: Backbone Network (Dec, 2020)

Winter Semester 2021 (6 Credits)

1. Proposed Architecture: Feature Network (Jan - Feb, 2021)
2. Proposed Architecture: Heads (Feb, 2021)
3. Experiments and Results: (March - April, 2021)
# Contents

1 Introduction .................................................. 1

2 Dataset ....................................................... 2
   2.1 Primary Dataset ........................................... 2
      2.1.1 Overview ............................................. 2
      2.1.2 Annotation .......................................... 3
   2.2 Auxiliary Dataset ......................................... 4
      2.2.1 MS COCO ............................................. 4

3 Existing Deep Learning Architectures ......................... 6
   3.1 EfficientDet: Scalable and Efficient Object Detection .......... 6
   3.2 CSPNet: A New Backbone that can Enhance Learning Capability of CNN . . . 7
   3.3 SpineNet: Learning Scale-Permuted Backbone for Recognition and Localization . 8

4 Proposed Architecture ........................................ 10
   4.1 Backbone ................................................... 10
      4.1.1 CSP + EfficientNet + SAM .............................. 10
   4.2 Feature Network ............................................. 11
      4.2.1 NAS-FPN ............................................. 11
   4.3 Heads ........................................................ 12
      4.3.1 Mask Head: Dice Loss Objective ...................... 12

5 Experimental Results .......................................... 13
   5.1 Backbone ................................................... 13
   5.2 Feature Network ............................................. 14
   5.3 Heads ........................................................ 14
   5.4 Our Model ................................................... 14

6 Conclusion and Future Work .................................. 18
Chapter 1

Introduction

One of the major challenges in developing nations like India is to make medical facilities accessible to the entire population of the country. Due to a lack of medical infrastructure throughout the country, a significant percentage of the country’s population faces financial and physical outreach to specialized hospitals and treatments. Cancer detection is one of the major life-threatening health-related hazards and even a small percentage of the population getting diagnosed with cancer can result in large numbers given India’s population. There are specialized hospitals being developed to address this need but in the present situation much of the population cannot afford the expensive diagnostics like flow cytometry test. Also, much of these specialized hospitals are present or being laid out in urban areas due to which people available in remote areas are not able to avail these facilities.

Artificial intelligence (AI) based solutions can prove themselves as no less than a miracle in such scenarios. The advances in Deep Learning (DL a way to achieve AI) has increased the possibilities to use DL based solutions in the Medical domain and for diagnostics. DL-based solutions show more generalisability on a prospective subject’s data by learning the distinguishable characteristics from the data itself rather than relying on handcrafted features that are used by traditional statistics. Even though DL based solutions have shown their potential in many medical applications like diagnostics and are becoming the state of the art solutions. There are many fundamental challenges related to DL based solutions like explainability, capturing heterogeneity at various levels (subject, sensor, site), label noise, handling uncertainty, etc which need to be addressed for their usage as an alternative to existing diagnostics.

This BTP address some of the above described fundamental challenges in DL based solutions in the application of white blood cancer diagnostics, Especially Multiple Myeloma (MM) cancer diagnostics. For this, We'll be exploring and developing DL based solutions for the detection and segmentation of MM cancer cells. This research work is being carried out in collaboration with the Dept. Of Oncology, AIIMS under the supervision and expert guidance of Dr. Anubha Gupta(Prof., Dept. of ECE, IIIT-Delhi) and Dr. Ritu Gupta(Prof., Dept. of Oncology AIIMS). The outcomes of this research include aiming for publications for validation of our research followed by deployment of the proposed DL based solution in AIIMS(Delhi).
Chapter 2

Dataset

2.1 Primary Dataset

2.1.1 Overview

![Stained Cell Slide Image](image1.png)  ![Annotated Masks](image2.png)  ![Masks Overlayed](image3.png)

Figure 2.1: Our Dataset

Our primary dataset is a set of microscopic images that were captured from the bone marrow aspirate slides of multiple subjects, as per standard guidelines[15]. The data is compiled from the bone marrow aspirates slides of n subjects and consists of 35,500 images with over 1,50,000 cell markers. The slides were stained using the Jenner-Giemsa stain. The images were captured using a Nikon Eclipse-200 microscope equipped with a digital camera at a magnification of 1000x. These images were captured in a raw BMP format with a size $2040 \times 1536$ pixels.

The given dataset poses the following challenges:

1. The task of segmentation required us to segment both the nucleus and the cytoplasm. For some cells, the cytoplasm may be in less contrast to the nucleus or at times, even the background. This causes difficulty to segment out the cytoplasm.

2. Presence of more than one type of stained and unstained cells poses another challenge in extracting plasma cells of interest.

3. Plasma cells may be clustered together and hence, segmentation of the overlapping/touching
cells is required. Generally, this is difficult because of different configurations:

- nuclei of different cells are touching
- nuclei of one and cytoplasm of another cell are touching
- the cytoplasm of different cells are touching

A subset of the dataset previously released is available at the public repository.[5]

2.1.2 Annotation

The obtained data consisted of around 35,500 images of stained Multiple Myeloma plasma cells, with each image having multiple markers to be annotated ranging from 0-20. The data was segmented into batches of 100 images each giving around 350 batches in total to be annotated. Students doing projects under Dr. Anubha Gupta on Multiple Myeloma were distributed batches and were tasked to annotate masks (See Fig. 2.1) for the markers on these images. A proper NDA was signed to maintain data and patient confidentiality.
We used the Image Segmenter Tool from MATLAB\cite{9} to achieve the annotated masks. (See Fig 2.3)

We use a subset of this dataset along with previous similar dataset from SBI lab at IIIT Delhi to create a dataset, which was also made a public challenge in SegPC-2021. This subset contains 775 images. These are divided into a training set of 298 images, validation set of 200 images, and a test set of 277 images. For this BTP’s purpose we use the train set for training and validation set for performance metric purposes.

2.2 Auxiliary Dataset

2.2.1 MS COCO

The MS COCO dataset\cite{14} is a large-scale dataset of 328k images consisting of over 2.5 million objects. The dataset was curated by gathering images of complex everyday scenes containing common objects in their natural context. Objects were labeled using per-instance segmentations to aid in precise object localization. The comparative class distribution of the COCO dataset with respect to the PASCAL VOC dataset is shown in Fig 2.5.
Figure 2.5: Comparitive class distribution of MSCOCO w.r.t. PASCAL VOC [14]
Chapter 3

Existing Deep Learning Architectures

We explored State of the art architectures on some of the benchmark datasets for Object Detection and Instance Segmentation namely MS COCO and PASCAL VOC 2007 in the non-medical domain.

Following are some of the recent advancements in Object Detection and Instance Segmentation Architectures:

3.1 EfficientDet: Scalable and Efficient Object Detection

Existing object detectors are mostly categorized by whether they have a region-of-interest proposal step (two-stage) or not (one stage). While two-stage detectors tend to be more flexible and more accurate, one-stage detectors are often considered to be simpler and more efficient by leveraging predefined anchors. EfficientDet largely follows the one-stage detectors
It employs ImageNet-pretrained EfficientNets \[19\] as the backbone network. Their proposed BiFPN\[20\] serves as the feature network, which takes level 3-7 features \{P3, P4, P5, P6, P7\} from the backbone network and repeatedly applies top-down and bottom-up bidirectional feature fusion. These fused features are fed to a class and box network to produce object class and bounding box predictions respectively (See Fig. 3.1). Similar to RetinaNet \[13\], the class and box network weights are shared across all levels of features.

### 3.2 CSPNet: A New Backbone that can Enhance Learning Capability of CNN

This Paper introduces a concept of partial skip connections in repetitive blocks of a backbone network. They suggest that rather than passing the whole feature map through a block and also forward on a skip connection, we pass half the channels of the feature map through the block and half through the skip-connection (See Fig.3.2). This partitioning of the feature map helps achieve a richer gradient combination due to gradient flow propagating through different network paths while reducing the amount of computation. The proposed CSPNet\[21\] can be easily applied to ResNet\[6\], ResNeXt\[24\], and DenseNet\[8\]. After applying CSPNet on the above mentioned networks, the computation effort can be reduced from 10% to 20%, but it outperforms in terms of accuracy, in conducting image classification task on ImageNet\[17\].

![Figure 3.2: CSP-DenseNet Block Architecture](image)

![Figure 3.3: Exact Fusion Model](image)
They also propose a feature aggregation network called the Exact Fusion Model\cite{21} inspired from YOLOv3\cite{16} which assigns exactly one bounding-box prior to each ground truth object. Each ground truth bounding box corresponds to one anchor box that surpasses the threshold IoU. If the size of an anchor box is equivalent to the field-of-view of the grid cell, then for the grid cells of the $s^{th}$ scale, the corresponding bounding box will be lower bounded by the $(s - 1)^{th}$ scale and upper bounded by the $(s + 1)^{th}$ scale. Therefore, the EFM assembles features from the three scales. See Fig.3.3.

3.3 SpineNet: Learning Scale-Permuted Backbone for Recognition and Localization

![SpineNet Architecture](image)

Figure 3.4: The architecture comparison between a ResNet backbone (3.4a) and the SpineNet backbone (3.4b) derived from it using NAS.

CNNs created for image tasks typically encode an input image into a sequence of intermediate features that capture the semantics of an image (from local to global), where each subsequent layer has a lower spatial dimension. However, this scale-decreased model may not be able to deliver strong features for multi-scale visual recognition tasks where recognition and localization are both important (e.g., object detection and segmentation). Several works including FPN\cite{12} and DeepLabv3+\cite{1} propose multi-scale encoder-decoder architectures to address this issue, where a scale-decreased network (e.g., a ResNet) is taken as the encoder (commonly referred to as a backbone model). A decoder network is then applied to the backbone to recover the spatial information.

While this architecture has yielded improved success for image recognition and localization tasks, it still relies on a scale-decreased backbone that throws away spatial information by
down-sampling, which the decoder then must attempt to recover.

They propose a meta architecture called a scale-permuted model that enables two major improvements on backbone architecture design. First, the spatial resolution of intermediate feature maps should be able to increase or decrease anytime so that the model can retain spatial information as it grows deeper. Second, the connections between feature maps should be able to go across feature scales to facilitate multi-scale feature fusion. They then use neural architecture search (NAS) [25] with a novel search space design that includes these features to discover an effective scale-permuted model.

They named the learned 49-layer scale-permuted backbone architecture SpineNet-49 [3] from the ResNet-50 [6] model. SpineNet-49 can be further scaled up to SpineNet-96/143/190 by repeating blocks two, three, or four times and increasing the feature dimension. An architecture comparison between ResNet-50-FPN and the final SpineNet-49 is shown.(See Fig. 3.4)
Chapter 4

Proposed Architecture

We model our proposed novel architecture on a two-step Object Detection - Segmentation Architecture (See Figure 4.1) based on the Mask-RCNN\cite{7}. The main reason for this choice is that though one-step models are known to be faster in inference time, they lack at the accuracy metric when compared to two-step models\cite{10}. And our specific problem of detecting cancer cells require accuracy to be the top most-priority and not the inference time.

4.1 Backbone

4.1.1 CSP + EfficientNet + SAM

One current candidate for the backbone that we have devised is a modified EfficientNet\cite{19} with CSP Connections\cite{21} (See Section 3.2) and a Spatial Attention Module(SAM)\cite{22} (See Fig 4.2) in the repetitive MBConv Block(Inverted Residual Block)\cite{18} of the original efficientnet. We found this configuration and architecture to give much better performance on CIFAR10\cite{11} Classification Task and due to the CSP Connections the FLOPs are reduced and the architecture remains both scalable and efficient. We call our backbone architectures a CSP-EfficientNet

Figure 4.1: Two Step Object Detection - Segmentation Model
Family of models with variants same as EfficientNets\cite{19} ranging from B0 to B7.

![Image](image1.png)

Figure 4.2: Spatial Attention Module

### 4.2 Feature Network

#### 4.2.1 NAS-FPN

Feature networks are feature extractors and recent state of the art feature networks include FPN(Feature Pyramid Network)\cite{12}. FPN composes of a bottom-up and a top-down pathway. The bottom-up pathway is the usual convolutional network for feature extraction. As we go up, the spatial resolution decreases. With more high-level structures detected, the semantic value for each layer increases. FPN also provides top-down pathway to construct higher resolution layers from the semantic rich layers with lateral connections between reconstructed layers and the corresponding feature maps to help the detector to predict the location betters (See Fig. 4.3).

![Image](image2.png)

Figure 4.3: FPN\cite{12}

For the feature network in our model we incorporate a novel architecture known as NAS-FPN\cite{4} that is discovered using neural architecture search. The search is a novel scalable search space covering all cross-scale connections possible in a vanilla FPN\cite{12}. The discovered architecture consists of a combination of top-down and bottom-up connections to fuse features across scales. The optimal intra-scale connections found through the Neural architecture search are visualised in Fig. 4.4.
Figure 4.4: Architecture of the discovered 7-merging-cell pyramid network in NAS-FPN with 5 input layers (yellow) and 5 output feature layers (blue). GP and R-C-B are stands for Global Pooling and ReLU-Conv-BatchNorm, respectively. [4]

4.3 Heads

4.3.1 Mask Head: Dice Loss Objective

Instead of the normal Sigmoid Crossentropy loss on the mask head of the MaskRCNN type model, we incorporate a new loss using the Dice Coefficient [2] which in theory will tend to maximise the mean IOU of the predicted masks and the ground truth masks of the samples by minimising the negative of the Dice Coefficient. The following is the expression for the dice coefficient:

$$\text{Dice} = \frac{2|X \cap Y|}{|X| + |Y|} \quad (4.1)$$

inspired from [2], where X is the predicted mask and Y is the ground truth mask. Then our dice loss is given by:

$$\text{DiceLoss} = -1 \times \text{Dice} \quad (4.2)$$
Chapter 5

Experimental Results

5.1 Backbone

For Backbone experimentation purposes we took EfficientNetB0\[19\] as the base architecture and CIFAR10\[11\] as the metric dataset to compare performances which is in accordance to the norm in computer vision classification task before training on a larger dataset like Imagenet\[17\]. We fixed training hyperparamters to be same for every architecture and epochs for the architectures to 50, as we observed complex architectures like these converged relatively faster on a small dataset like CIFAR10\[11\] which can be seen in Fig\[5.1\] on Train and Validation data. The accuracy results can be seen in Table \ref{tab:5.1}.

![Graphs showing classification accuracy plots for experimented architectures.](image)

Table 5.1: Val. accuracy of experimented backbone architectures after 50 epochs on CIFAR10\[11\] classification task.

| Architecture                                    | Val. Accuracy |
|------------------------------------------------|---------------|
| EfficientNetB0                                  | 0.71          |
| EfficientNetB0 + Spatial Attention              | 0.75          |
| EfficientNetB0 + Spatial Attention + CSP        | 0.82          |
5.2 Feature Network

To understand the contribution of the NAS-FPN feature network, we perform ablation study on the performance of the model on the COCO dataset.

Table 5.2: Performance of our Model with NAS-FPN and other feature networks on COCO minival.

| Architecture                  | BoxAP | MaskAP |
|-------------------------------|-------|--------|
| CSP-EfficientNetB0 + C4       | 40.3  | 35.3   |
| CSP-EfficientNetB0 + FPN      | 42.3  | 37.2   |
| CSP-EfficientNetB0 + NAS-FPN  | 43.8  | 38.3   |

5.3 Heads

For our model we train our mask head with a loss function based on IoU score called the dice loss, unlike the conventional Mask-RCNN’s Mask head which trains on just sigmoid binary crossentropy of the predicted binary masks from the GT masks. So we perform ablation study to understand the contribution of the modified loss objective of the mask head to the performance of the whole model.

Table 5.3: MaskAP of our model(CSP-EfficientNetB0 + NAS-FPN) with and without the modified loss objective on the COCO minival.

| Mask Head Loss               | MaskAP |
|------------------------------|--------|
| Sigmoid Binary Crossentropy Loss | 38.3   |
| Dice Loss                    | 39.1   |

5.4 Our Model

In this section we show the performance of our model in comparison with other existing comparable architectures on the same task.

Table 5.4: Performance of our model in comparison with comparable existing MaskRCNN model architectures on the COCO minival.

| Work | Model                          | BoxAP | MaskAP |
|------|--------------------------------|-------|--------|
| Prior| ResNet50-FPN [23]              | 38.6  | 35.2   |
|      | SpineNet-49                    | 41.2  | 36.3   |
|      | EfficientNetB0-FPN             | 41.4  | 36.7   |
| Ours | CSP-EfficientNetB0 + NAS-FPN + Dice Loss | 43.8  | 39.1   |
Figure 5.2: Comparing training loss plots of vanilla MaskRCNN with EfficientNetB0+FPN vs our CSP-EfficientNetB0+NAS-FPN.

Table 5.5: Performance of Our model in comparison with comparable existing MaskRCNN model architectures on our primary dataset.

| Work     | Model                                      | BoxAP | MaskAP | mIOU |
|----------|--------------------------------------------|-------|--------|------|
| Prior    | ResNet50-FPN                               | 62.4  | 63.3   | 0.82 |
|          | SpineNet-49                                | 63.3  | 64.1   | 0.84 |
|          | EfficientNetB0-FPN                         | 63.4  | 64.3   | 0.84 |
| Our Model| CSP-EfficientNetB0 + NAS-FPN + Dice Loss   | 64.7  | 65.6   | 0.86 |
Figure 5.3: Sample Detections on COCO minival [14] by our model.
Figure 5.4: Sample Detections on our primary dataset by our model.
Chapter 6

Conclusion and Future Work

We find that our CSP-EfficientNet family of backbone architectures provide increased performance from the base vanilla EfficientNet\textsuperscript{[19]} Family of Models. We also find that NAS-FPN\textsuperscript{[4]} also provides an increment in performance in comparison to other features networks. And our Modified loss objective for the mask head using dice coefficient\textsuperscript{[2]} helps us achieve better IoU of predicted and ground truth masks and in result a better MaskAP.

Our model even on the base backbone architecture(CSP-EfficientNetB0) provides us with competitive performance on both the benchmark datasets for Object detection and instance segmentation. Also, it gives a competitive performance on our primary dataset.

We plan to further improve the performance of our model by finding more novel techniques and architectures to help us deploy a much accurate Multiple Myeloma cancer detection system at AIIMS, Delhi to help doctors promptly detect and diagnose Multiple Myeloma cancer patients.
Bibliography

[1] Liang-Chieh Chen et al. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. 2018. arXiv:1802.02611 [cs.CV].

[2] Lee R. Dice. “Measures of the Amount of Ecologic Association Between Species”. In: Ecology 26.3 (1945), pp. 297–302. ISSN: 1939-9170. DOI: https://doi.org/10.2307/1932409.

[3] Xianzhi Du et al. SpineNet: Learning Scale-Permuted Backbone for Recognition and Localization. 2020. arXiv:1912.05027 [cs.CV].

[4] Golnaz Ghiasi et al. NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection. 2019. arXiv:1904.07392 [cs.CV].

[5] Anubha Gupta et al. “PCSeg: Color model driven probabilistic multiphase level set based tool for plasma cell segmentation in multiple myeloma”. In: PloS one 13.12 (2018), e0207908.

[6] Kaiming He et al. Deep Residual Learning for Image Recognition. 2015. arXiv:1512.03385 [cs.CV].

[7] Kaiming He et al. Mask R-CNN. 2018. arXiv:1703.06870 [cs.CV].

[8] Gao Huang et al. Densely Connected Convolutional Networks. 2018. arXiv:1608.06993 [cs.CV].

[9] Image Segmentation. URL: https://in.mathworks.com/discovery/image-segmentation.html.

[10] Licheng Jiao et al. “A Survey of Deep Learning-Based Object Detection”. In: IEEE Access 7 (2019), pp. 128837–128868. ISSN: 2169-3536. DOI: 10.1109/access.2019.2939201 URL: http://dx.doi.org/10.1109/ACCESS.2019.2939201.

[11] A. Krizhevsky. “Learning Multiple Layers of Features from Tiny Images”. In: 2009.

[12] Tsung-Yi Lin et al. Feature Pyramid Networks for Object Detection. 2017. arXiv:1612.03144 [cs.CV].

[13] Tsung-Yi Lin et al. Focal Loss for Dense Object Detection. 2018. arXiv:1708.02002 [cs.CV].

[14] Tsung-Yi Lin et al. “Microsoft coco: Common objects in context”. In: European conference on computer vision. Springer. 2014, pp. 740–755.
[15] Antonio Palumbo et al. “International Myeloma Working Group consensus statement for the management, treatment, and supportive care of patients with myeloma not eligible for standard autologous stem-cell transplantation”. In: Journal of clinical oncology 32.6 (2014), p. 587.

[16] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement. 2018. arXiv: 1804.02767 [cs.CV].

[17] Olga Russakovsky et al. ImageNet Large Scale Visual Recognition Challenge. 2015. arXiv: 1409.0575 [cs.CV].

[18] Mark Sandler et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks. 2019. arXiv: 1801.04381 [cs.CV].

[19] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. 2020. arXiv: 1905.11946 [cs.LG].

[20] Mingxing Tan, Ruoming Pang, and Quoc V. Le. EfficientDet: Scalable and Efficient Object Detection. 2020. arXiv: 1911.09070 [cs.CV].

[21] Chien-Yao Wang et al. CSPNet: A New Backbone that can Enhance Learning Capability of CNN. 2019. arXiv: 1911.11929 [cs.CV].

[22] Sanghyun Woo et al. CBAM: Convolutional Block Attention Module. 2018. arXiv: 1807.06521 [cs.CV].

[23] Yuxin Wu et al. Detectron2. https://github.com/facebookresearch/detectron2 2019.

[24] Saining Xie et al. Aggregated Residual Transformations for Deep Neural Networks. 2017. arXiv: 1611.05431 [cs.CV].

[25] Barret Zoph and Quoc V. Le. Neural Architecture Search with Reinforcement Learning. 2017. arXiv: 1611.01578 [cs.LG].