University of Birmingham
School of Computer Science
MSc Robotics

Second semester mini-project

Models for 2018 Data Science Bowl: Detecting nucleus

Ke Xu

Supervisor: Iain Style

April 2019

Abstract

With the development of Computer Vision, automating biomedical image processing becomes promising gradually.

This project tries several model on the kaggle competition 2018 Data Science Bowl which offer a mission: create an algorithm detect nucleus automatically. The models include U-net, Mask R-CNN and a recurrent instance segmentation model based on U-net and convolutional long short-term memory(ConvLSTM) networks.

As a result, single U-net perform quite well on finding out all cells in the image, which is not enough for this competition. With the transfer learning, Mask R-CNN can get a satisfying result in limited time, while the recurrent model is hard to train with a limited computing resource.

Keywords

Instance segmentation; kaggle; Data Science Bowl; Mask R-CNN; ConvLSTM; U-net; transfer learning

Contents

Abstract and Keywords **i**

1 Introduction **1**

1.1 Aims and Objectives 1

1.2 Description of the work 1

2 Background and Related Work **4**

2.1 Semantic segmentation 4

2.2 Instance segmentation 5

3 Implementation **6**

3.1 U-Net 6

3.2 Recurrent instance segmentation model 6

3.3 Mask R-CNN 9

4 Results **10**

4.1 U-Net 10

4.2 Recurrent instance segmentation model 10

4.3 Mask R-CNN 11

5 Discussion **13**

Bibliography **14**

Appendices **17**

A Mini-project declaration **17**

B Statement of information search strategy **21**

List of Tables

4.1 Training loss of the Recurrent instance segmentation model for first 7 epochs. 11

4.2 Average batch loss of the first epoch of the Recurrent instance segmentation model.Every batch contains 67 images. 11

4.3 Recall($\frac{TP}{TP+FN}$) of the Mask-RCNN result on the stage 1 test set under different IoU threshold 12

List of Figures

1.1 From left to right:the image of cells;mask of single cell;mask of all cells . . 1

1.2 Original U-Net,source:[8] 2

1.3 The recurrent model 2

1.4 Mask R-CNN 3

2.1 FCN-8 Architecture source:[2] 4

3.1 Spatial inhibition Architecture 7

4.1 Output of U-Net.From left to right:ground truth;predicted mask;original image 10

4.2 Result of Mask R-CNN 12

5.1 Sample images from MSCOCO dataset 13

Chapter 1

Introduction

1.1 Aims and Objectives

Identifying the cells' nuclei is the starting point for most analyses because the DNA of human are stored in the cells' nuclei. Using automatical approaches to identify nuclei help speeding up the unlocking cure. Actually, identifying nuclei automatical is a kind of biomedical image processing task.

In the field of biomedical image processing, there are mainly two categories of tasks,i.e. semantic segmentation(as the task presented in [8]) and instance segmentation(such as the 2018 Data Science bowl).

1.2 Description of the work

The dataset given by the project contains a large number of segmented nuclei image.

Figure 1.1: From left to right:the image of cells;mask of single cell;mask of all cells

The images were acquired under a variety of conditions and vary in the cell type,magnification, and imaging modality (brightfield vs fluorescence). Every image in the train set has several masks(see Figure 1.1). Each mask represents a cell with a nucleus in the image. The main task is producing a mask for each nucleus in the images of the test dataset.

There are three models used in this project, including U-Net[8], a recurrent model[7] and Mask R-CNN[3]. As [8] mentioned, the U-Net has a shape that looks like a letter U(see Figure 1.2).

Based on [7], A recurrent model which combined the U-Net module and attention mechanism module is established(see Figure 1.3).

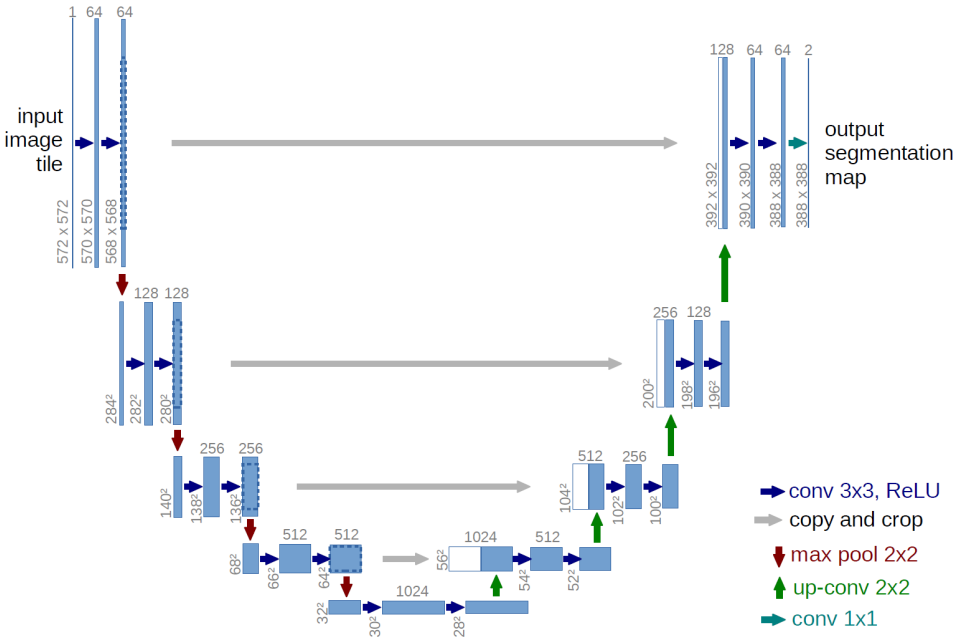

Figure 1.2: Original U-Net,source:[8]

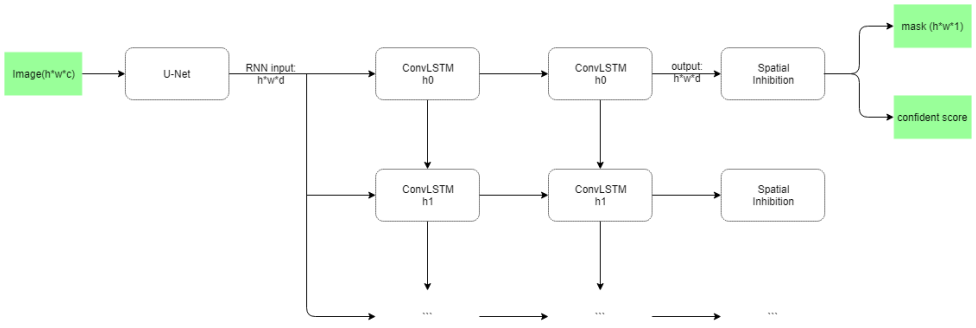

Figure 1.3: The recurrent model

Moreover, using Mask R-CNN trained by transfer learning I got a pretty good re-

sult.Figure 1.4 exhibits the architecture of Mask R-CNN.

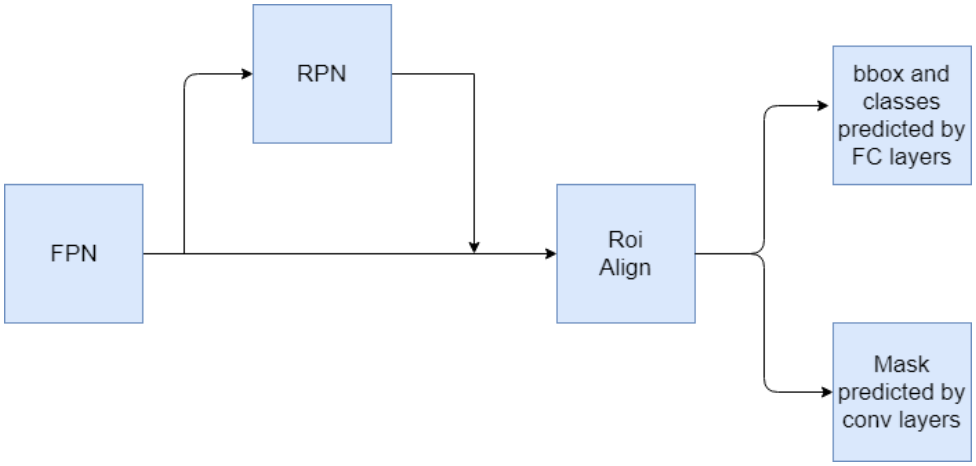

Figure 1.4: Mask R-CNN

Chapter 2

Background and Related Work

2.1 Semantic segmentation

Semantic Segmentation is one of the significant problems in the field of computer vision. It describes the process of matching each pixel of an image with a class label, (such as road, sky, or car).

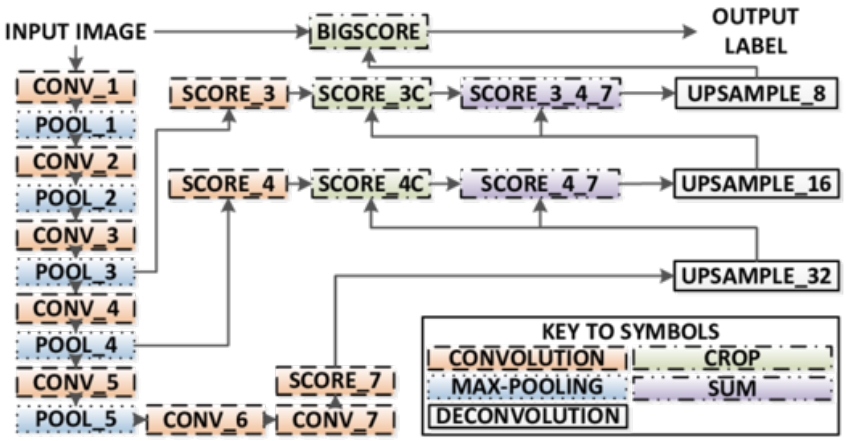

Figure 2.1: FCN-8 Architecture source:[2]

One of the famous and influential models is fully convolutional networks(FCN) [4]. FCN transfers parameters from VGG16 to perform semantic segmentation. It replaces the last three fully convolutional layers with convolutional layers. Introducing skip connection and upsampling, FCN can generate an excellent result after combining the upsample

output and high-resolution features. An FCN-8 architecture is shown in Figure 2.1 . Upsampling is realised by transposed convolution operation which can be regarded as inverse operation of the convolution operation. FCN use convolution operation with 1×1 kernel size which can produce the pixel-wise classification.

Modifying the architecture of FCN, U-Net works more precise with few training images. See Figure 1.2. One notable modification is that with a large number of feature channels in upsampling part, U-Net can propagate context information to higher resolution layers.

2.2 Instance segmentation

Instance segmentation can be regarded as a combination of object detection and semantic segmentation. It identifies each object instance of each pixel for every known object within an image. Labels are instance-aware.

One of the most important ideas used in object detection, such as YOLO and R-CNN, is sliding-window,nding objects by looking in each window placed over a dense set of image locations. As well, it is the most popular idea in instance segmentation(such as Mask R-CNN and TensorMask [1]). Another exciting idea is imitating human attention mechanisms.Recurrent models have the capability of deciding at each time which part of the input image to look at in order to perform instance segmentation [7].

Chapter 3

Implementation

3.1 U-Net

As a semantic segmentation model, U-Net needs to know every nucleic mask in one image, so I extract maximum value on each pixel of all masks of one image and use it as the ground truth mask(See Figure 1.1).

With tensorflow,the first U-Net model of this project was established. Although U-Net accepts any size of the input, for the convenient of batch training the raw data with different size is reshaped into the $3 \times h \times w$ format. All convolution operations have a same padding mode which makes the output and input have the same size. Both the max-pooling layers and transposed convolution layers have the same kernel size and stride, so their influence on size is offset. The output of the last two layers of U-Net should be $d \times h \times w$ and $1 \times h \times w$, where d,h and w are channel size, height and width respectively.

3.2 Recurrent instance segmentation model

After implementing the U-Net model, it is possible to get the prediction of all nuclei in the raw image. According to [7], the stage of splitting each instance in the mask is called inference process. This stage consist of two part: Convolutional LSTM and spatial inhibition, which has been depicted in Figure 1.3. Instead of using U-Net, the semantic segmentation model in [7] is FCN-8.

The ConvLSTM unit, as introduced in [9], can model spatiotemporal relationships by replacing the fully connected layers in each gate of LSTM with convolutions. Stacking two or more ConvLSTM unit to learn more complex relationships. In this task, units number equals to 2. Moreover, ConvLSTM can account for the recent segmented instance by updating its inner state.

The output mask of U-Net is taken as the input of the ConvLSTM. In the ConvLSTM part, the number of channels in the recurrent unit is the same as the number of channels produced by the previous U-Net. This means the output shape of the ConvLSTM is $t \times 1 \times h \times w$, where t is the sequence length indicating the number of predicted masks. The spatial inhibition part actually consists of two functions. One of them maps the output of ConvLSTM to the single-instance mask: $\mathbb{R}^{h \times w \times d} \rightarrow \left[0, 1\right]^{h \times w}$.Another one map the output of ConvLSTM and the confidence score: $\mathbb{R}^{h \times w \times d} \rightarrow \left[0, 1\right]$.The score represents the possibility that the current segmented candidate is an instance. The structure of spatial inhibition part is shown in Figure 3.1.

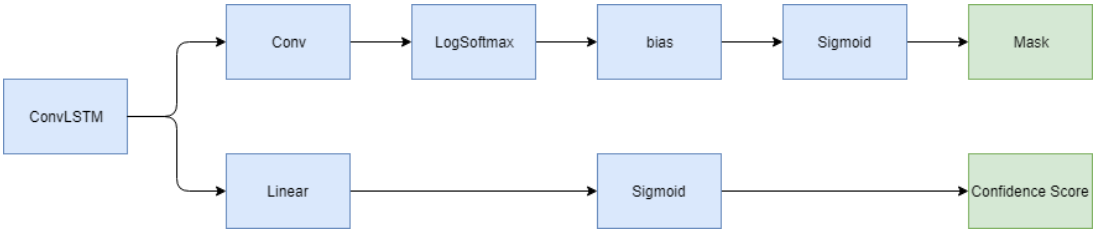

Figure 3.1: Spatial inhibition Architecture

In the first function of spatial inhibition part, the output of log-max is an $h \times w$ matrix having pixel value in the interval $(-\infty, 0]$. After that, a learnable bias, b, is added to it. The bias can be regarded as a threshold that filters the pixels that will be selected for the present instance. Thus the input of the sigmoid function will be in the interval $(-\infty, b]$. After sigmoid, chosen pixels will gain a value near 1 while others smaller than 0.5.

There was a problem in the process of implementing the loss function of this model. The loss function uses the Hungarian algorithm to match predicted masks and ground truth masks. The Hungarian algorithm is a classic combinatorial algorithm that solves the assignment problem in polynomial time. As there might be a large number of predicted

masks, it is necessary to find out the best match for each ground truth mask. However, it is quite complicated in tensorflow to register a function of Hungarian algorithm on GPU. Hence, it wasted about two weeks trying using tensorflow to establish the model. Eventually, Pytorch was used to define a class of loss function to implement this operation. The pseudocode is presented below.

Algorithm 1 Forward and backward propagation of the loss function

Input: ground truth $Y = \{Y_1, ..., Y_n\}$,predicted masks$Y_{pre} = \{Y_{pre_1}, ..., Y_{pre_m}\}$, and confidence scores$s = \{s_1, ..., s_m\}$,hyperparameter $\lambda = 1$

Output: cost c and gradiants dY_p, ds

Forward:
initialize $M \in [0,1]^{m \times n}$
Fill M,so that $M_{i,j} = f_{IoU}(Y_i, Y_{pre_j})$
set $\underline{n} = min(n, m)$
matching = Hungarian(M)
for t = 1,...,min **do**
  $c = c - M_{t,matching(t)} + \lambda f_{BCE}(1, s_t)$
end for
for t = n + 1,...,$\underline{n}$ **do**
  $c = c + \lambda f_{BCE}(0, s_t)$;
end for
Backward:
for $t = 1, ..., m$ **do**
  if $t \leq \underline{n}$ **then**
    $dY_{pre_t} = f'_{IoU}(Y_{matching(t),Y_{pre_t}})$
    $ds_t = f'_{BCE}(1, s_t)$
  else
    $ds_t = f'_{BCE}(0, s_t)$
  end if
end for

Basically, for all matched predicted masks Y_{pre} and ground truth Y the loss function includes two items. The first one is the minus intersection-over-union(IOU) of matched prediction and ground truth. The second one is a binary cross entropy function about confidence scores s:

$$Loss = min - \sum_{t}^{\underline{n}} f_{IoU}(Y_t, Y_{pre_t}) + \sum_{t}^{m} \lambda f_{BCE}([t \leq n], s_t) \tag{3.1}$$

where t is the time step,n is the number of ground truth masks and is a hyperparameter that ponders the importance of the second term comparing to the first one. The binary

cross entropy function has the form:$f_{BCE}(a,b) = (a\log(b) + (1-a)\log(1-b))$ .Hence the minimum value of loss function is $-n$.

At the training stage, I trained the U-Net again for nine epochs with the batch size equal to 16(U-Net converge very fast) since I re-constructed the model in Pytorch. Based on [7], I fixed the parameters except for the last layer of U-Net, and performed curriculum learning,i.e. gradually increasing the number of sequence length. This training method has two advantages:

- give rise to improved generalization and faster convergence
- to find better local minima of a non-convex training criterion

3.3 Mask R-CNN

Mask R-CNN, as depicted in Figure 1.4, consists of 4 parts: Feature Pyramid Network(FPN) generates feature map; Region Proposal Network(RPN) produces region proposals and select regions of interest(ROI) from region proposals;ROI Align aligns ROI to the original image;predictor head to output the boundary box and mask.

Because the Mask R-CNN is hard to train in the limited time, I modified Matterport's implementation of Mask R-CNN deep neural network for object instance segmentation [5].

Using the pre-train model provided to do transfer learning. With the limited computing resource, I set the batch size equal to 1 and the image size equal to 256 while leave remaining parameter to be the default.

Chapter 4

Results

4.1 U-Net

The example of the result of U-Net is produced by the model established with tensorflow. The batch size is 16, and the image size is 128 × 128. Using SGD optimiser with learning rate equals to 0.00001 after 1000 batches, the model gets a loss of 0.058 on the training set(670 images).The output is displayed in Figure 4.1. As what competition needs is instance segmentation, this output is only an intermediate output.So it is meaningless to check these results' performance.

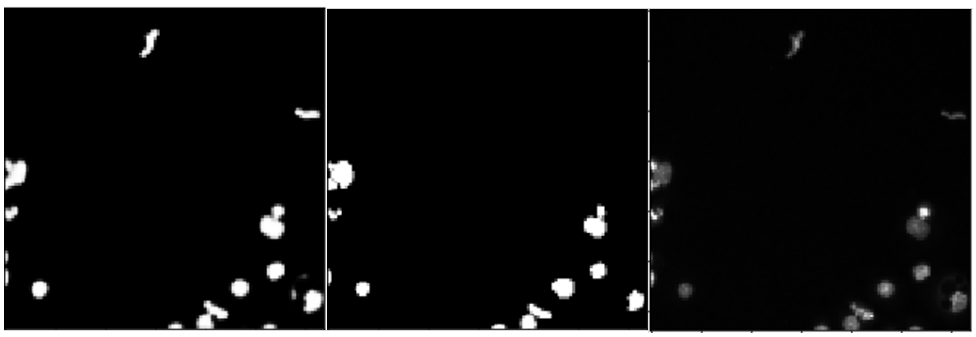

Figure 4.1: Output of U-Net.From left to right:ground truth;predicted mask;original image

4.2 Recurrent instance segmentation model

This model is hard to train, because of curriculum learning. The sequence length of the RNN start from 1, and each epoch adds one to the number of sequence length. Since

there are up to 376 cells in an image,the sequence length should go up to $376 + 1$. This sequence length at least needs $1 + 2 + ... + 377 = 71253$ training epochs. It is unreachable with the limited computing resource I have. Several loss values I get on the training set is shown in table 4.1 and table 4.2.Comparing these two table we can find that the loss value increase with the sequence length.The reason is that as the model is not convergent yet, the second item of the loss function(see Equation 3.1) increases when the sequence length increases.

After checking the output, all pixel values of the current output mask of the model are the same. This means the final layer of U-Net does not give pixels of different instance different values. Hence, the ConvLSTM and spatial inhibition part cannot choose an instance as output.

Epoch	1	2	3	4	5	6	7
Sequence length	1	2	3	4	5	6	7
Loss	-0.0004	0.1638	0.4514	0.7799	1.1904	1.8894	3.0414

Table 4.1: Training loss of the Recurrent instance segmentation model for first 7 epochs.

Batch	1	2	3	4	5	6	7	8
Sequence length	1	1	1	1	1	1	1	1
Loss	0.653	0.599	0.534	0.468	0.404	0.345	0.288	0.234

Table 4.2: Average batch loss of the first epoch of the Recurrent instance segmentation model.Every batch contains 67 images.

4.3 Mask R-CNN

After 40 epochs training, which takes about 4 to 5 days, a pretty good result is attained by Mask R-CNN. The model has a loss of about 0.18 on the training set and a loss of about 0.24 on the validation set. Because of late submission and the stage 2 score showed 0.000 at time of submission, the submitted .csv file did not get a score, this problem is discussed on the forum, and other people also have it, see [6].Hence only the stage 1 test set, whose result file is given, can be used to evaluate this model(see table 4.3). Figure 4.2 shows several results on the test set:

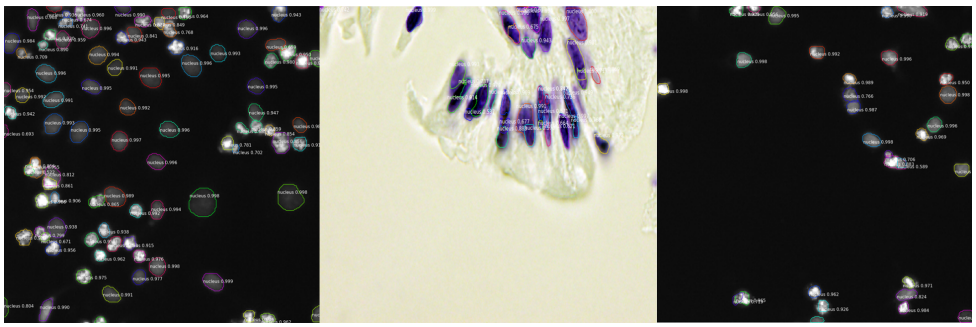

Figure 4.2: Result of Mask R-CNN

IoU threshold	0.5	0.9	0.95
Recall	93.4%	93.4%	93.3%

Table 4.3: Recall($\frac{TP}{TP+FN}$) of the Mask-RCNN result on the stage 1 test set under different IoU threshold

All the code can be found in [10].

Chapter 5

Discussion

In this report, I have appropriately accomplished the task of the kaggle competition 2018 Data Science Bowl using Mask R-CNN and have got an excellent semantic segmentation result with U-Net. I also reproduced structure of the model of paper [7].However, the recurrent instance segmentation model still not work, although I spend the most time in it. It is necessary to explore why it is so hard to train. There are some possible reasons:

- the dataset is too small, and the original paper is trained on the MSCOCO dataset which is much bigger than this dataset.
- the image in this dataset is hard to train on this model since in the most of the image the foreground and background have very similar colour while in MSCOCO it is not(see Figure 5.1)
- FCN should not be replaced with U-Net

Figure 5.1: Sample images from MSCOCO dataset

Except for the failure of the recurrent instance segmentation model, there are still other deficiencies in this project:

- did not do data augmentation which may improve the accuracy of the output
- did not consider any method to deal with the overlap of cells

According to the deficiencies mentioned above, the plan includes:

- train the recurrent instance segmentation model on the MSCOCO dataset first and then use transfer learning to check whether is the problem of model
- doing random horizontal or vertical flips, random rotation and random cropping to the dataset.

Moreover, An critical lesson is that I should carefully consider the models that require too much computing resources.

Bibliography

[1] Chen, X., Girshick, R., He, K., and Dollár, P. Tensormask: A foundation for dense object segmentation. *arXiv preprint arXiv:1903.12174* (2019).

[2] García-Peraza-Herrera, L. C., Li, W., Gruijthuijsen, C., Devreker, A., Attilakos, G., Deprest, J., Vander Poorten, E., Stoyanov, D., Vercauteren, T., and Ourselin, S. Real-time segmentation of non-rigid surgical tools based on deep learning and tracking. In *International Workshop on Computer-Assisted and Robotic Endoscopy* (2016), Springer, pp. 84–95.

[3] He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 2961–2969.

[4] Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 3431–3440.

[5] Matterport. Mask r-cnn for object detection and instance segmentation on keras and tensorflow. `https://github.com/matterport/Mask_RCNN`. Accessed: 2019-04-23.

[6] quadcore/Richard Epstein. Did you see your score when you were submitting stage2? `https://www.kaggle.com/c/data-science-bowl-2018/discussion/55313#latest-319459`. Accessed: 2019-04-23.

[7] Romera-Paredes, B., and Torr, P. H. S. Recurrent instance segmentation. In *European conference on computer vision* (2016), Springer, pp. 312–329.

[8] Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (2015), Springer, pp. 234–241.

[9] Xingjian, S., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-c. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems* (2015), pp. 802–810.

[10] Xu, K. Datasciencebowl-2018. `https://github.com/JJJinx/Datasciencebowl-2018`. Accessed: 2019-04-26.

Appendix A

Mini-project declaration

The University of Birmingham

School of Computer Science

Second Semester Mini-Project: Declaration

This form is to be used to declare your choice of mini-project. Please complete all three sections and upload an electronic copy of the form to Canvas:

https://canvas.bham.ac.uk/courses/31210

Deadline: 12 noon, December 3rd, 2018

1. Project Details

Name: Ke Xu

Student number: 1909285

Mini-project title: Models for 2018 Data Science Bowl:Detecting nucleus

Mini-project supervisor: Iain Styles

Mini-project reader:TBC

---

2. Project Description

Aim of mini-project	Using image processing techniques detect nucleus automatically.
Objectives to be achieved	1. Using both classical image processing techniques and modern machine learning techniques to understand these richly informative datasets. 2. Doing in-depth study in the areas of image processing, computer vision and machine learning. 3. There will naturally be a reasonable amount of programming but the main complexity will be in the design of the algorithms. Some mathematics may be involved.
Project management skills. Briefly explain how you will devise a management plan to allow your supervisor to evaluate your progress	I tend to meet the supervisor every several days to communicate with him about the problem I faced and the progress of the project.
Systematic literature skills. Briefly explain how you will find previous relevant work	1. Base on supervisors recommend 2. Searching on the academic search engine base on the key word
Communication skills. What communication skills will you practise during this mini-project?	Learn how to effectively discuss academic topics with supervisor

The following questions should be answered in conjunction with a reading of your programme handbook.

3. Project Ethics Self-Assessment Form

Please answer YES/NO to the following questions:

- **Does the research involve contact with NHS staff or patients? NO**
- **Does the research involve animals? NO**
- **Will any of the research be conducted overseas? NO**
- **Will any of the data cross international boarders? NO**
- **Are the results of the research project likely to expose any person to physical or psychological harm? NO**
- **Will you have access to personal information that allows you to identify individuals, or to corporate or company confidential information (that is not covered by confidentiality terms within an agreement or by a separate confidentiality agreement)? NO**
- **Does the research project present a significant risk to the environment or society? NO**
- **Are there any ethical issues raised by this research project that in the opinion of the PI or student require further ethical review? If you are unsure, consider whether the project has the potential to cause stress or anxiety in the people you are involving. NO**
- **Human subjects can be involved as users, providers of system requirements, testers, for evaluation, or similar such activities. Does the experiment involve the use of human subjects in any other capacity? If you are unsure, answer YES. NO**
- **Answer YES if ANY of the following are true NO**

*** the project has the potential to cause stress or anxiety in the people you are involving, e.g. it addresses potentially sensitive issues of health, death, religion, self-worth, financial security or other such issues**

*** the project involves people under 18**

*** the project involves a lack of consent or uninformed consent**

*** the project involves misleading the subjects in any way**

If the project's involvement of people relates only to straightforward information gathering, requirements specification, or simple usability testing, then you can indicate NO.

- **If any of the above questions is answered YES, or you are unsure if further review is needed (the first point is usually a good indicator - may cause stress or anxiety) then you should refer it for review.**
- **Further review will involve the School Ethics Officer meeting with the supervisor and ideally the student, reviewing the project, and suggesting any procedures necessary to ensure ethical compliance.**

DECLARATION

By submitting this form, I declare that the questions above have been answered truthfully and to the best of my knowledge and belief, and that I take full responsibility for these responses. I undertake to observe ethical principles throughout the research project and to report any changes that affect the ethics of the project to the University Ethical Review Committee for review.

Signed(student)

Date

Signed(supervisor)

Date

Appendix B

Statement of information search strategy

Parameters of your literature search

Forms of literature The important categories are (in order):

- conference papers
- journal articles

Geographical/language coverage

Important work is likely to be from the North America and Western Europe. Preferred language is English.

Recent information

One of the most important methods in this field is proposed after 2014, so the papers before 2014 should be paid little attention to.

Appropriate search tools

Google Scholar

Used to key word searching to find papers solving similar problems. Key words can be like Cell Segmentation or Instance Segmentation.

IEEE Xplore

To be used to retrieve conference papers and journal articles. Key words can be like Cell Segmentation or Instance Segmentation.

arXiv

Primarily for getting the papers not included in ieee and papers has not published yet. Key words can be like Cell Segmentation or Instance Segmentation.

Brief evaluation of the search.

By checking the abstract, the relevant papers are chosen.

The search in Google Scholar retrieved 3 items judged to be relevant of which:

- 2 conference items
- 1 journal article

The search in IEEE retrieved 2 items judged to be relevant of which:

- 2 conference items

The search in arXiv retrieved 1 items judged to be relevant of which:

- 1 preprint item