**Find the Nuclei in Divergent Images to Advance Medical Discovery**

Ji chen, Runyu Hong

**Introduction & Motivation**

We are proposing to find the Nuclei in divergent images to advance medical discovery, which is the topic of Kaggle 2018 Data Science Bowl. Most of the 30 trillion human body cells have nucleus, which is full of DNAs. Automating nuclei detection allows biomedical researchers to locate cells in varied conditions. This will save the time of biologist and medical professionals for drug testing and medical diagnosis. For instance, automating neutrophil identification would benefiting hips and knees transplant infection check, which significantly reduces the time of pathologists and surgeons and thus minimizes the surgical risk of the patients. In addition, automatic identifying nuclei could benefit the identification of individual cells in a sample, which would contribute to single cell imaging experiments in various biological fields.

**Related work**

In the past decades, many scientists and professionals have devoted into nuclei detection from different perspectives. Some of these studies involved statistical techniques in tracing and extracting morphological features of nuclei to advance the cure of leukemia and neural diseases 1,2. Some other studies improved imaging techniques to acquire more features from the cell in order to segment the nuclei and helped the treatment of diseases like Parkinson 3. With the development of artificial intelligence and machine learning techniques, using deep learning techniques and convolutional networks in solving these medical image segmentation problems became increasingly popular in recent years 4.

**Data**

The datasets are separated into 2 stages. The currently available stage 1 data has 670 images in the training set and 65 images in the testing set. These images were obtained from various imaging techniques with a unified size of 256 by 256 pixels. Our preliminary K-Means clustering analysis indicates that in the training set, 81.5% of the images coming from fluorescence microscopy, 16.12% coming from medical histology imaging, and 2.39% coming from light microscopy. Each of the images contains multiple nuclei, and for each nucleus in the training set, a binary mask is provided. The distribution of the number of nuclei in each image of training set is shown in the figure below based on our preliminary analysis. According to Kaggle, stage 2 datasets will be released on April 9 and will be of similar size as stage 1 datasets with images from microscopy techniques that are not seen in stage 1 datasets.

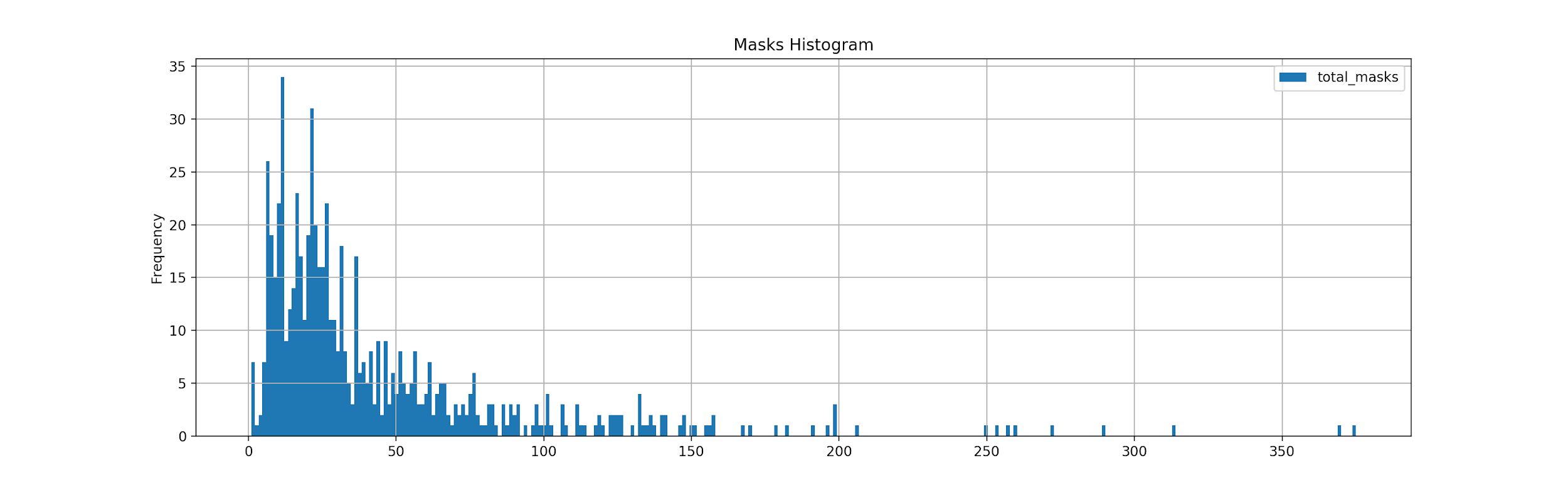
****

Fig. 1: Histogram of number of nuclei per image

**Proposed method**

We are planning to use a convolutional network called U-net for our project. It was developed by University of Freiburg in 2015 and had shown extraordinary performance in biomedical image segmentation 4. It consists of a contracting path and an expanding path to make up a architecture of 23 convolutional layers in total (Fig. 2) 4. To predict the pixels in the border region of the image, the missing context is extrapolated by mirroring the input image. It uses elastic deformations to augment data so that less training samples are required 4. Compared to other convolutional network methods, U-net is much faster. The training time is about 10 hours and it takes less than a second to segment a 512 by 512 pixels image on a modern GPU 4.

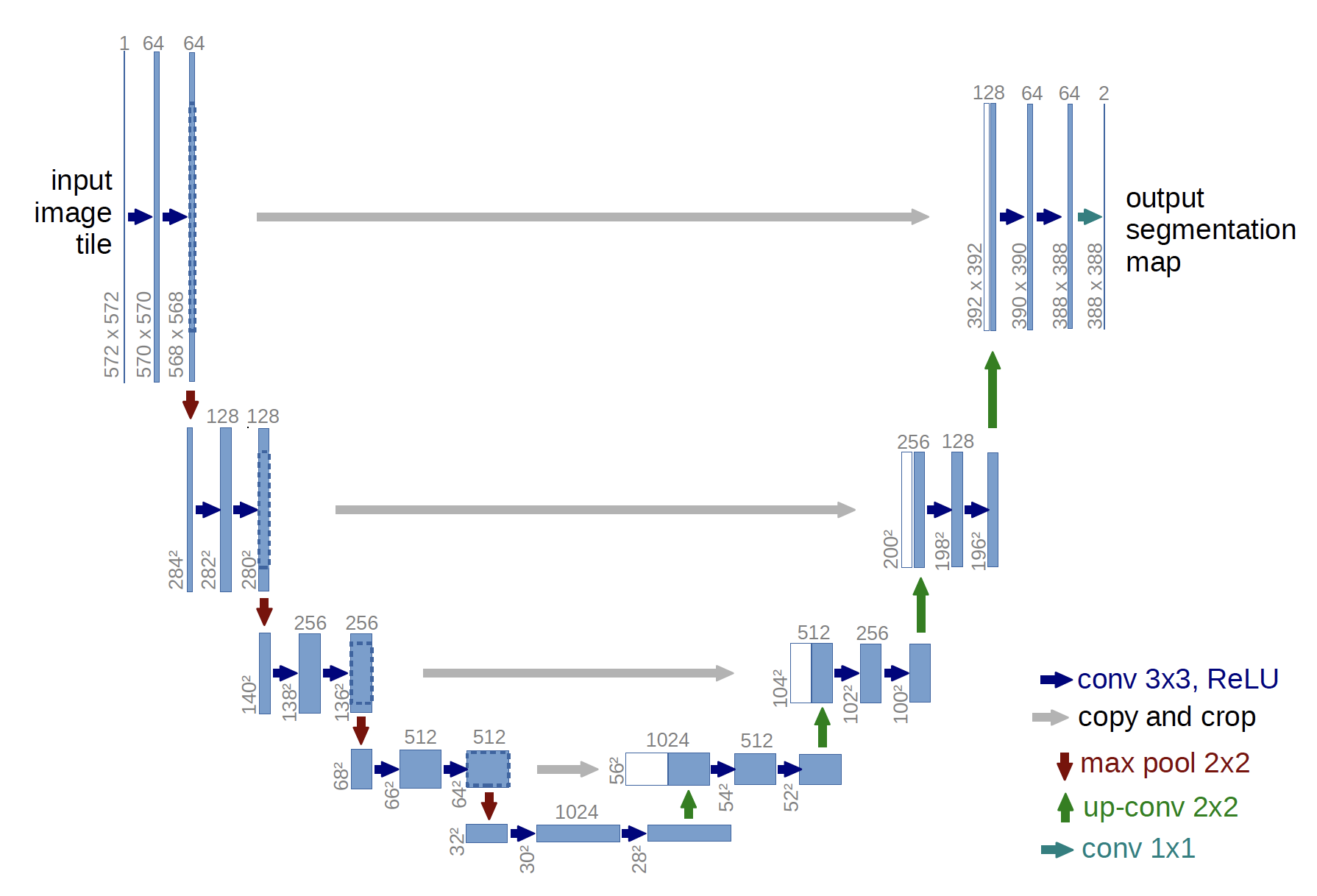


Fig. 2: Architecture of U-net 4. Blue boxes represent multi-channel feature maps. The number of features is annotated on top of the boxes. Feature sizes are denoted at the lower left edge of the boxes. White boxes in the upscaling steps represent cropped feature maps.

**Evaluation**

The evaluation of this competition is by the mean average “precision” at different intersection over union (IoU). The IoU of a predicted set of pixels (A) and a set of true label pixels (B) is calculated as follows:

IoU(A,B)=A∩B / A∪B.  
IoU(A,B)=A∩B / A∪B.

The average “precision is calculated at IoU thresholds ranging from 0.5 to 0.95 with a step size of 0.05. At each threshold value t, a precision value is calculated based on the number of true positives (TP), false negatives (FN), and false positives (FP) resulting from comparing the predicted pixels to true label pixels:

TP is counted if a set of predicted pixels pass the IoU threshold, FP is counted if a set of predicted pixels does not pass the IoU threshold and FN is counted if no predicted pixels overlap with a set of true label pixels. The average precision of a single image is then calculated as the mean of the above precision values at each IoU threshold:

At last, the final score is calculated by taking the mean of average precision per image in the test set.

**Timeline**

* + Jan 15: Competition opened; Stage 1 released
  + Feb 28: Data preprocessing complete
  + March 31: Done with training and testing for Stage 1
  + Apr 9: Team merging ends; Stage 2 test set released
  + Apr 16: Competition closed

**References**

1. Cao, F., & Fablet, R. (2006). Automatic morphological detection of otolith nucleus. *Pattern Recognition Letters*, *27*(6), 658-666.
2. Scotti, F. (2005, July). Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images. In *Computational Intelligence for Measurement Systems and Applications, 2005. CIMSA. 2005 IEEE International Conference on* (pp. 96-101). IEEE.
3. Novak, P., Daniluk, S., Ellias, S. A., & Nazzaro, J. M. (2007). Detection of the subthalamic nucleus in microelectrographic recordings in Parkinson disease using the high-frequency (> 500 Hz) neuronal background. *Journal of neurosurgery*, *106*(1), 175-179.
4. Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Springer, Cham.