Identifying Parasitized Cells with Deep Convolutional Neural Networks

Sara Olsson (sarol@stanford.edu), Julia White (juliawhi@stanford.edu)

Malaria is a widespread, life-threatening disease for which early detection is crucial. Currently, the most effective method of diagnosis is to manually identify blood cells which host Malaria-causing Plasmodium parasites. To reduce human error in this process and improve the chances of consistent diagnoses, the automation of Malaria detection could be the next step toward fighting this health crisis. To that end, we have developed and implemented a convolutional neural network for the identification of parasitized cells. Additionally, in order to prove our model's utility in the wider spectrum of visually-diagnosable illnesses (which may have relatively poor datasets), we developed a semi-supervised algorithm that achieves similar results with significantly less labelled data.

DATA

- 27,558 RGB labeled single-cell images taken from blood smears with an even distribution of healthy and parasitized cells.
  - Cell images were resized to a uniform 64x64 pixels.
  - Pixels were normalized (divided by their maximum value, 255) so all pixel values would lie between [0,1].
- The dataset was split into 80% training data, 10% validation data, and 10% testing data

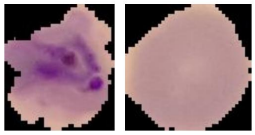


Figure 1: Example blood smear image of parasitized cell (left) and uninfected cell (right).

APPROACH

- In order to acquire more labelled data for training, available data was augmented (blurred) and added to the training set.

Semi-Supervised Learning

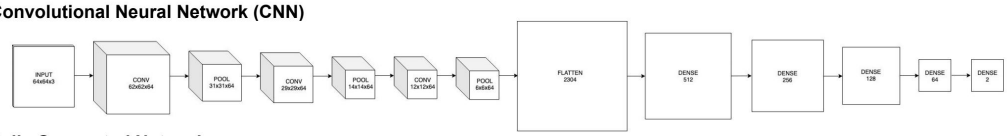
- To improve our model's performance on an artificially limited dataset (i.e. with only a certain percent of the Malaria dataset left labeled), we used a semi-supervised learning algorithm with pseudo-labeling:
  1. The model is initially trained with only X% of the full dataset taken as the labeled dataset.
  2. The resulting trained model is used to label the remaining (100-X)% unlabeled data.
  3. Confidently labeled data points (above some threshold) are added to the labeled dataset.
  4. The model is re-trained with the updated labeled dataset.
  5. Step 2-4 are repeated until there is no improvement in performance.

Supervised Learning

- Supervised learning on X% of the full dataset was used as a baseline to gauge the success of our semi-supervised learning algorithm.

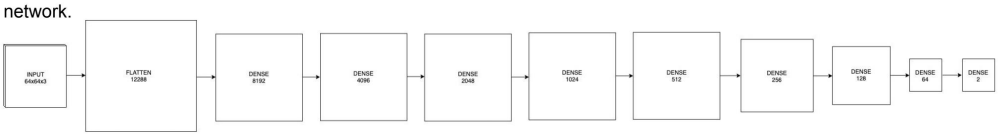
MODEL

Convolutional Neural Network (CNN)



Fully Connected Network:

- A fully connected network was used as a point of reference to compare with the performance of our convolutional neural network.



RESULTS AND ANALYSIS

- Performance was evaluated with respect to recall.
  - Due to the deadly nature of Malaria, limiting false negatives is a priority.

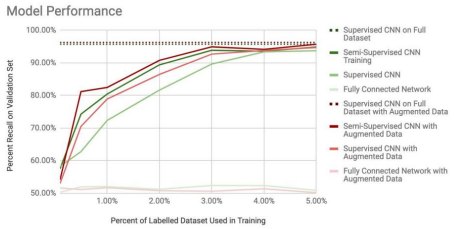


Figure 2: Model performance for various model/algorithm combinations.

- On the test set:
  - The semi-supervised CNN with augmented data achieved 94% recall with only 5% of the full labelled dataset.
  - The supervised CNN with augmented data achieved 94% recall on the full dataset (oracle).
  - The supervised CNN with augmented data achieved 92% recall with only 5% of the full labelled dataset (baseline).
- The fully connected network was ineffective.
  - No improvement in performance after re-labelling data for semi-supervised learning.
  - CNN outperforms due to its adeptness for extracting features from image data.

FUTURE WORK

We plan to run further tests on similar datasets (for visually-diagnosable illnesses) to demonstrate the robustness of our CNN model and semi-supervised learning algorithm.