# Learning with Unlabelled Positives

Guangcheng Lin[301354550], Haojie Huang[301385300], Longxuan Zhao[301385113], Qianyuan Pu[301361571], and Weilong Xu[301385645]

{gla72,hha110,lza128,qpu,weilongx}@sfu.ca

**Abstract.** Testing for brain tumors by using the MMdetection which is object detection kit based on Pytorch to train the model with A variety of popular detection components. Through the combination of these components can quickly build a variety of detection framework. In the end, we got high accuracy in the detection of tumors.

**Keywords:** Machine learning · Unlabelled positives · Diagnosis helper · Single lesion · Brain tumor · MRI Scan

## 1 Introduction

### 1.1 Background

Tumors that grow inside the brain are known as brain tumors, including primary brain tumors that originate in the brain parenchyma and secondary brain tumors that migrate to the brain from other parts of the body. Its etiology is still unknown, the tumor occurred from the brain, meningeal, pituitary, cranial nerve, cerebrovascular and embryonic remnant organizer, known as primary intracranial tumor. The malignant tumor that organizes by other organs of the body metastases to intracranial, call secondary intracranial tumor. Intracranial tumors can occur at any age, with the most common being between 20 and 50 years old.

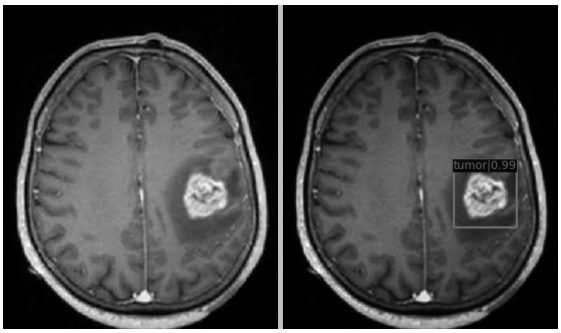


**Fig. 1.** brain tumor

### 1.2 Motivation

Plain Radiographs, Cerebral Angiography, Ventriculography, Pneumoencephalography, Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) were commonly used for diagnosis. Besides environmental and infectious factors, there is no clear cause of the disease, which is why brain tumors are so dangerous. It is the best choice to nip the disease in the cradle. Therefore, this project of our group is to detect brain tumors by using brain MRI scans, and describe the research report and results of this project.

### 1.3 Operating environment

1. **Prerequisites**
   - Linux or macOS (Windows is in experimental support)
   - Python 3.7.1
   - PyTorch 1.8.1
   - CUDA 11.1
   - GCC 5+
   - MMCV

**Table 1.** Compatible MMDetection and MMCV versions are shown as below. Please install the correct version of MMCV to avoid installation issues.

| MMDetection version | MMCV version |
|---|---|
| 2.11.0 | mmcv-full $\geq$ 1.2.4, <1.4.0 |

*Note 1.* **You need to run pip uninstall mmcv first if you have mmcv installed. If mmcv and mmcv-full are both installed, there will be a ModuleNotFoundError.**

2. **Install MMDetection**
   - I. Install mmcv-full.

```
pip install mmcv-full==1.3.1 -f https://download.openmmlab.com/mmcv/dist/cu111/torch1.8.0/index
```

   - II. Install MMDetection.

```
git clone git@github.com:CMPT340-Group-8/Project.git
cd mmdetection
pip install -r requirements/build.txt
python setup.py develop
pip install tqdm
pip install ipdb
cd demo_new
python image_demo.py
```

# 2 Materials

## 2.1 Dataset

1. The data set of this project is from kaggle. This dataset contains 2500 MRI scans of the brain, and is divided into three folders.[1]
   - Folder "yes" contains the MRI scans that have a tumor.
   - Folder "no" contains the MRI scans that do not have a tumor.
   - Folder "pred" contains unlabelled MRI scans for testing purposes.

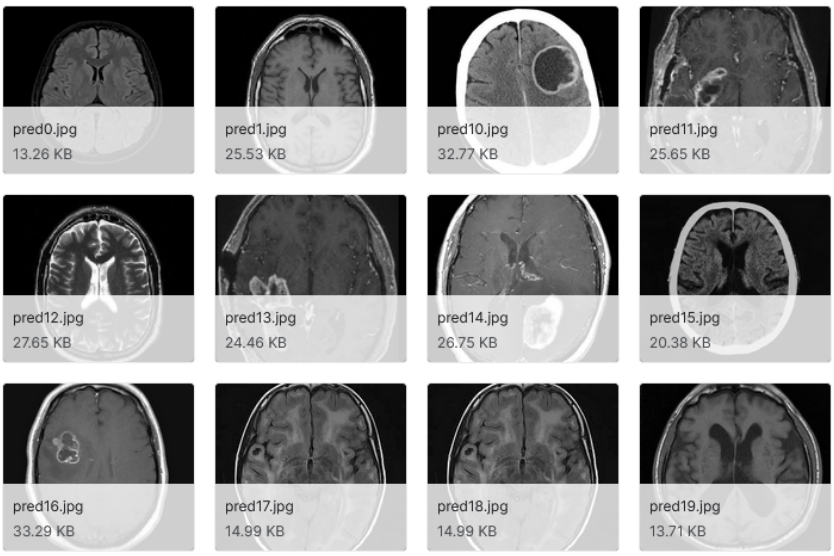


**Fig. 2.** dataset example

2. The data source of our project is available in the **links** below:

   - Training the model:
     *https://www.kaggle.com/abhranta/brain-tumor-detection-mri*

   - Testing the final model:
     *https://www.kaggle.com/preetviradiya/brian-tumor-dataset*

## 2.2 Unlabelled Positive Learning

1. PU learning, which stands for positive and unlabelled learning, is a semi-supervised binary classification method that recovers labels from unknown cases in the data. It does this by learning from the positive cases in the

data and applying what it has learned to relabel the unknown cases. This approach provides benefits to any machine learning problem that requires binary classification on unreliable data, regardless of the domain[2]. We refer to relevant articles of PU Learning and conduct experiments in this Project.

# 3 Methods

## 3.1 Pipeline

The whole system of machine learning can be divided into four steps:

- data acquisition
- model data pre-processing
- model validation
- model use

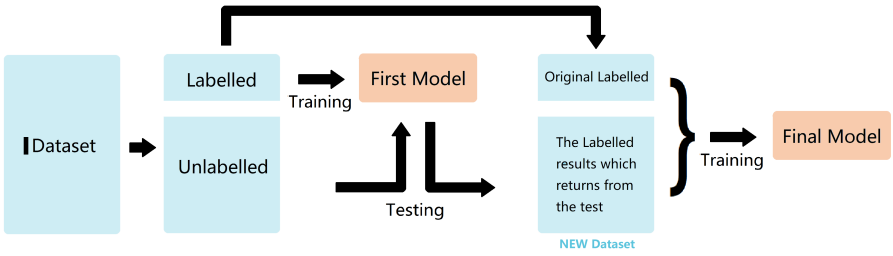


**Fig. 3.** Pipeline

## 3.2 Unlabelled positive

There is a total of 1500 brain MRI images in our dataset. We label 300 images for training the first model, and then use the remaining unlabelled images to test the model. After getting the test results from our program, we collect all positives data as a new dataset. Finally, our final model is accomplished by training the new dataset[2].

## 3.3 Labelme

Labelme was used to mark the lesions in 300 samples that needs to be labelled for training the model in the first learning period. After we label the tumor with the bounding boxes, Labelme will generate some **JSON** files that can help our program to identify the position of the tumor.

### 3.4 MMdection

We used CUDA, mmcv, and Pytorch to simulate the MMdection environment. MMdection is used to train the model with several labelled samples, and then test the code and get the results.

## 4 Results

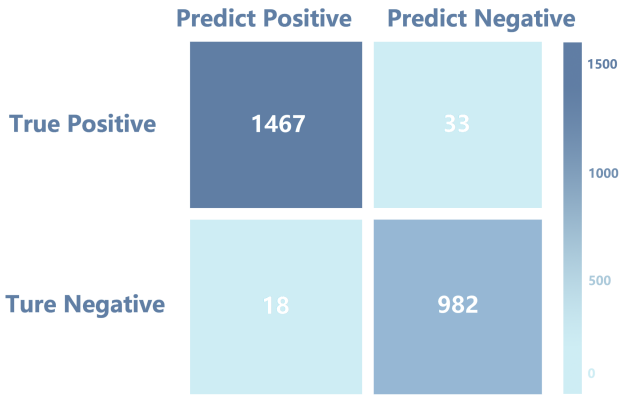


**Fig. 4.** Analysis

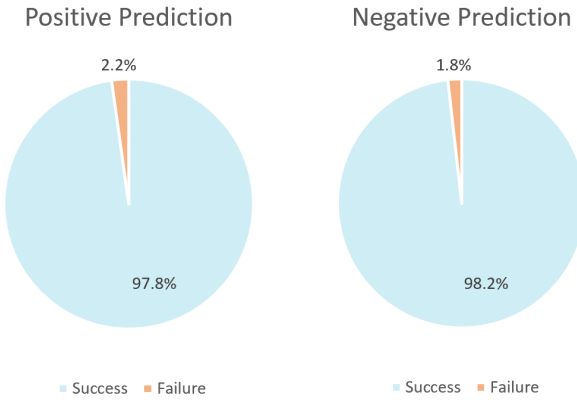


**Fig. 5.** Accuracy

We used 2500 samples to test and verify the final model, which includes 1500 true positive samples and 1000 true negative to test and verify our final mode. Based on Figure 4.

There are 1467 samples are successfully detected to be positive by our model (Fig.4.). Moreover, the program is also able to identify 982 negative samples in our test (Fig.4.). However, there is a total of 51 samples which is failed to be predicted. The test results show that the percentage of success predicted positive rate is 97.8% and the proportion of success predicted negative rate is 98.2% (Fig.5.).

## 5 Accomplishments

- To ensure the first round of training does not have a negative influence on the precision of subsequent training, we tested a few proportion of labeled and unlabelled data. Finally, we decided to use 300 labelled data and 1200 unlabelled data.
- At the beginning of the project, each of our members used different computers. Some of us uses apple Macbook in M1-core, which is not able to use virtual machine to build the environment. Some of us uses windows10 system. During the build environment, it was found that the version of CUDA and Pytorch did not match. Later we tried to arrange the environment on CSIL. After trying various methods, we finally decided to set up the environment on the cloud platform and proceed to the next step.
- When we were learning and using MMdection, it was automatically downloaded the latest version instead of old version we need when we follow the steps introduced on the official website. At the end, we found that the final version could not match and cause lots of bugs which is not resolved yet. And there were no error since we came up with cloud platform.

## 6 Contributions

- **Longxuan Zhao (LZ)**: After meeting with all team members and briefly going through the github reference, I came up with an idea to use MMdetection to detect brain tumor from brain MRI images. I made an unlabelled learning plan that realize our aim with the current dataset. I received completed dataset from HH, WX, GL and QP, and trained the model on cloud platform using faster rcnn. I analyzed the learning results with the program which HH wrote.
- **Haojie Huang (HH)**: I extended the unlabelled learning plan which is made by LZ. Due to failure of building environment locally, I rent a could server and finished all the environmental preparations in order for the project to carry on. I received dataset from GL and finished 1/3 of the labelled data. I wrote a program to count the number of positive and negative results for testing.
- **Weilong Xu (WX)**: I extended the unlabelled learning plan which is made by LZ. I received dataset from GL and finished 1/3 of the labelled data. Participated in the configuration of the code environment, as well as the testing of the model and the analysis of the output results.

- **Qianyuan Pu (QP)**: I extended the unlabelled learning plan which is made by LZ. I received dataset from GL and finished 1/3 of the labelled data.
- **Guangcheng Lin (GL)** : I had further discussion with LZ and searched resources for dataset and finally provided the team with dataset from Kaggle. I extended the unlabelled learning plan which is made by LZ.

## 7 Conclusion and Discussions

After training the model we observed that a group of MRI images with lower contrast have less precision by our model. The algorithm could be optimized to raise the accuracy of our model. From the perspective of practical application, the model could provide more information other then only detecting tumor in the MRI image. We observed that MRI images are not taken from one fixed angle but many angles including forehead, back head and above the head. By working further, we could provide specific location and size of the tumor to the doctor in actual scenarios and help them better analyze a patient's condition.

## 8 Future Work

As the report shows, the dataset that we used has only single lesion. It is probably that this project can be extend to learn and detect for multi-lesions. We believe that our project can be used to learn with samples of different diseases rather than just brain tumor, for example: lung CT scan, bone MRI scan and skin images.

## 9 Acknowledgements

We got our basic idea from *SOLVING MISSING-ANNOTATION OBJECT DETECTION WITH BACKGROUND RECALIBRATION LOSS* (Han et al., 2002) and etc., 2002[3] , *Positive and Unlabelled Learning: Recovering Labels for Data Using Machine Learning* (AaronWard, 2020)[2]. We also refer to the code from **https://github.com/Dwrety/mmdetection-selective-iou**
**https://github.com/open-mmlab/mmdetection**

## 10 Appendix A: Glossary

1. Cerebral Angiography
   Cerebral angiography produces very detailed, clear and accurate pictures of blood vessels in the brain and may eliminate the need for surgery.

2. CT
   A computerized tomography scan (CT or CAT scan) uses computers and rotating X-ray machines to create cross-sectional images of the body. These images provide more detailed information than normal X-ray images.

3. MRI
   Magnetic resonance imaging (MRI) is a medical imaging technique that uses a magnetic field and computer-generated radio waves to create detailed images of the organs and tissues in your body.

4. Pneumoencephalography
   Pneumoencephalography is a no-longer performed investigation that allowed imaging of the contours of the brain and ventricles by the deliberate introduction of air into the subarachnoid space.

5. Plain Aadiographs
   Plain radiography is a means of obtaining a picture of internal structures by passing X-rays through them, and recording the shadows cast by these structures.

6. Ventriculography
   A ventriculogram is a test that shows images of your heart. The images show how well your heart is pumping.

## References

1. A. Panigrahi. Brain tumor detection mri. `https://www.kaggle.com/abhranta/brain-tumor-detection-mri`, Mar 2021.
2. AaronWard. Positive and unlabelled learning: Recovering labels for data using machine learning. `https://heartbeat.fritz.ai/positive-and-unlabelled-learning-recovering-labels-for-data-using-machine-learning-59c1def5452f`, Mar 2020.
3. Z. et al. Solving missing-annotation object detection with background recalibration loss. `https://arxiv.org/pdf/2002.05274.pdf%20%20Code%20available:`, Aug 2020.