AI Product Management: Create a Medical Image Annotation Job

Proposal

Ben Jacobson
March, 2020

Project Overview

Project Goal

The goal of this project is to design a data annotation (labeling) job using Figure-Eight. This platform uses non-experts to identify and label cases images, in this case to label X-rays of pneumonia.

This data could later be used to build a data product (utilizing AI or machine learning) that helps doctors quickly identify cases of pneumonia in children (1-5). This data could be used to build a classification system that:

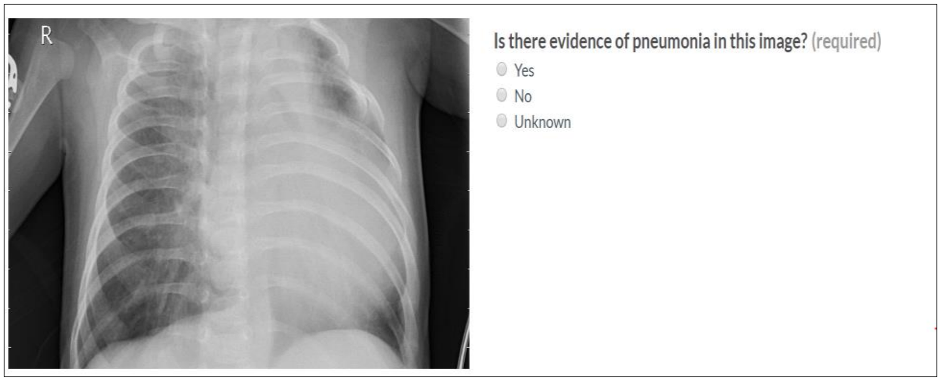
- Can help flag serious pneumonia cases
- Quickly identify healthy cases
- And, act as a diagnostic tool for medical technicians and doctors

This labeling job allows us to create a labeled data set, so that we can train a machine learning model to predict whether unlabeled x-rays indicate a healthy patient or pneumonia. With any machine learning model, it will only be as good as the data provided.

Data Labeling Approach

Choice of Data Labels

In the data labeling job, the x-ray image is provided and the question " Is there evidence of pneumonia in this image" is asked. The only (required) responses are (1) Yes, (2) No, and (3) Unknown. Here is an example question:



Is there evidence of pneumonia in this image? (required)

○ Yes
○ No
○ Unknown

These labels were chosen for their simplicity, and to reflect the main goal of the project. If there is clear evidence of pneumonia, based on the instruction and rules provided, then we want this labeled "Yes". Conversely, we want no evidence as "no". There is an option for unknown for ambiguous cases, and to capture uncertainty. The downside of a simple rating scale is that it loses additional complexity and gives no information as to the reason.

This was chosen rather than a scale (1-5), or more complicated options, because we want to flag obvious cases to train the model. We are not attempting to replace a diagnostician or trained experts. Simple answers may eliminate variation based on one reviewer against another.

Test Questions & Quality Assurance

Number of Test Questions

The dataset is only 117 records, of which:

- 101 are unlabeled
- 8 are labeled negative for pneumonia
- 8 are labeled positive for pneumonia

Since we want 5% test questions for our data, preferably with even distribution, we have 2 test questions for each label – six in total. We want an even distribution so that our metrics are equally represented.

Improving a Test Question

We have test questions in order to test our contributors (those that answer questions) on how well they are doing. If they are below a threshold we set, then we can remove them. So, if more than 10% are missed in our labeling job, we will remove that contributor.

If there is a test question that is missed by most contributors, we can clarify it using an additional rule to coach a better answer. This takes reviewing the image for 'why' this is occurring. If it continues to be a problem, a new label could possibly be added.

As the data changes, or if, then additional test questions, rules and examples will be provided.

Contributor Satisfaction

If contributors rate the overall job lowly based on the topic, then we will review and the improve the overview, examples, and the rules. If possible, we would provide a survey to receive additional feedback. However, depending on the category being low, we do the following:

- **Instruction Clear** → Make more clear instructions.
- **Test Questions Fair** → Improve the feedback and examples with common pitfalls
- **Ease of Job** → Preface that the job is "more challenging than others" or similar language in the overview.
- **Pay** → Consider increasing the pay.

Limitations & Improvements

Data Source

This is a pretty small dataset, in greyscale, with varying pixel length and width. The images also have slightly different angles. Bias is to the underlying patients that this was taken form, children less than five years old, so this dataset would be biased to that age group and should take care not to apply to all ages. (See "Resource" section below for source data).

Low quality and unreadable scans were removed. This is possible data that should have been kept, and to indicate a new scan should be administered. Why? A false negative as a result could be hazardous to the patient.

Increasing the size of the dataset and the age distribution of the patients would also be beneficial. Demographic information to augment the dataset may be useful as well.

Designing for Longevity

Long-term monitoring will be needed to ensure that the AI product is still receiving data that it was trained on. Keeping track of the decided metric, if it starts to decrease, then investigation of the classes that fail and carefully monitoring the input data is important. If the data changes enough, then we will do a new labeling job including update rules, examples and test questions.

Resources

Data Source

https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia