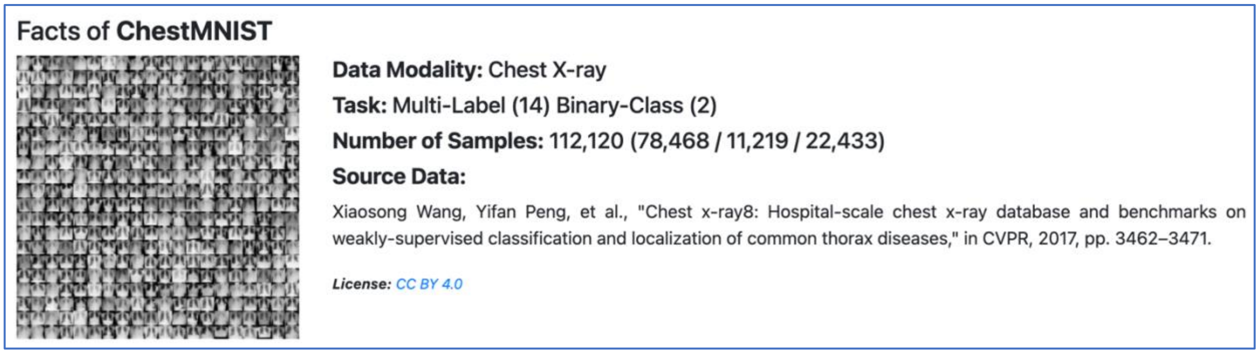
Black Box to automatically annotate images
Author: Li Yuan

1. Motivation

With the fast development of camera technology, millions of medical images are taken either in a clinical trial or biomedical research. However, to do downstream analysis with these images, labeled images are needed. I want to resolve this issue by designing a black box (model) to automatically annotate images because manually labeling millions of images is impossible.

2. Data Description

To test and experiment with my new model (black box), I chose the medical benchmark dataset, MedMNIST, A Large-Scale Lightweight Benchmark for 2D and 3D Biomedical Image collection. I chose the two datasets from it to implement the binary classification task. One is the chest dataset. Another is the breast dataset.



Facts of **ChestMNIST**

Data Modality: Chest X-ray

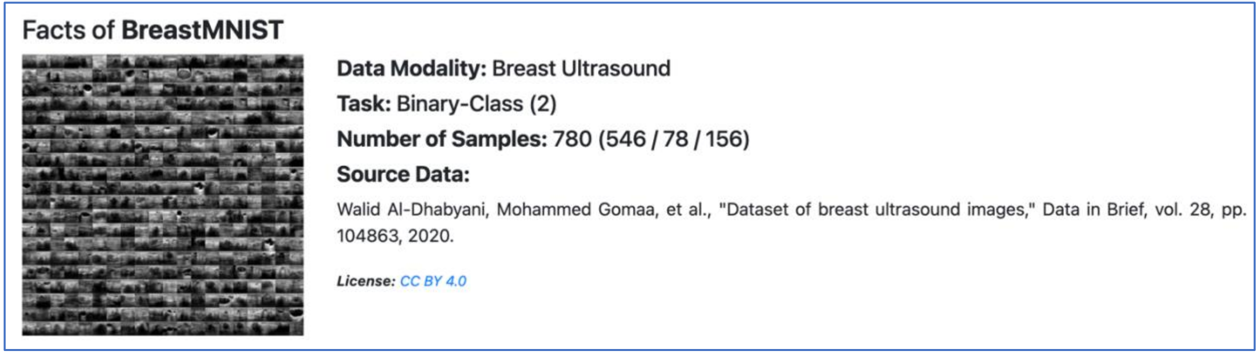
Task: Multi-Label (14) Binary-Class (2)

Number of Samples: 112,120 (78,468 / 11,219 / 22,433)

Source Data:

Xiaosong Wang, Yifan Peng, et al., "Chest x-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in CVPR, 2017, pp. 3462–3471.

License: *CC BY 4.0*



Facts of **BreastMNIST**

Data Modality: Breast Ultrasound

Task: Binary-Class (2)

Number of Samples: 780 (546 / 78 / 156)

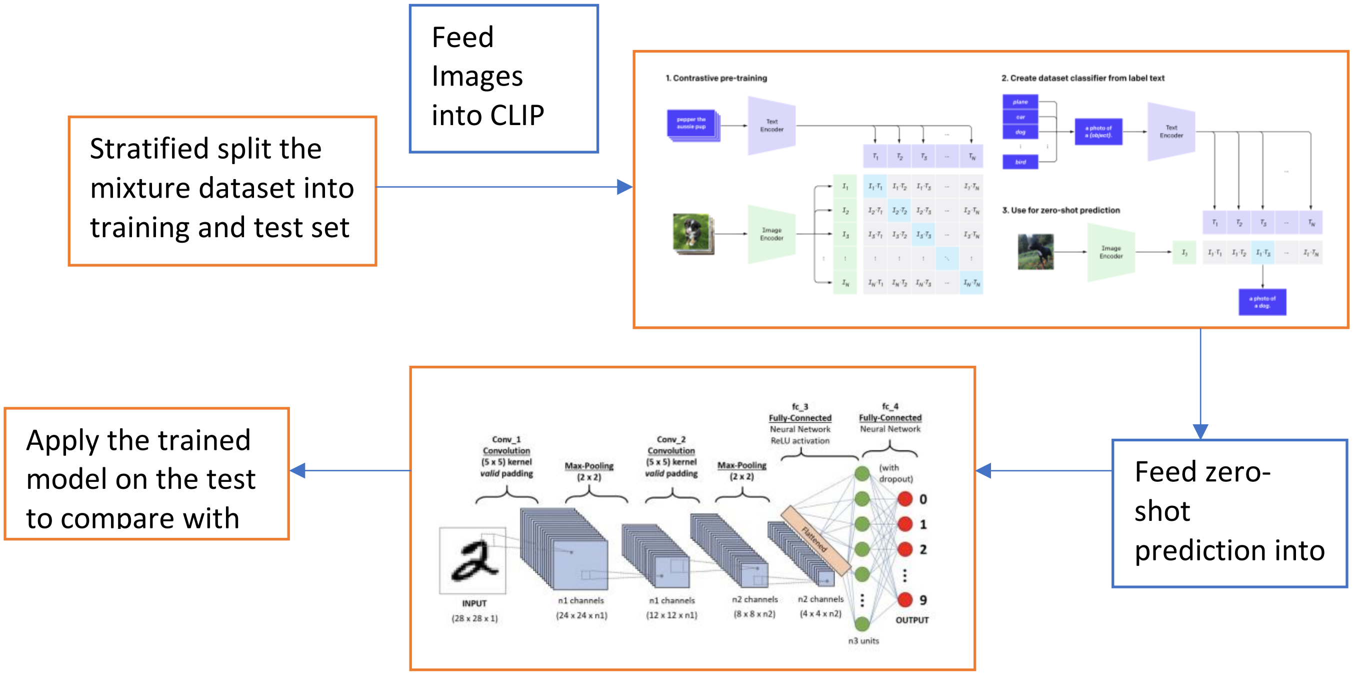
Source Data:

Walid Al-Dhabyani, Mohammed Gomaa, et al., "Dataset of breast ultrasound images," Data in Brief, vol. 28, pp. 104863, 2020.

License: *CC BY 4.0*

3. Design Model

I used three stages to build this new model pipeline. The first stage is to directly apply the zero-shot CLIP (contrastive language image pre-trained) model to the mixture dataset of chest and breast. I get the prediction classification results. I treat the labels from the first step as "true labels" to the mixture dataset and feed them together into a convolutional neural network. The second stage is to train this CNN with these "true labels." The third stage is to evaluate this combined model and compare it with the zero-shot model only. I figured the workflow (pipeline) of this three-stage modeling like the below plot; the orange box represents the main stages (models), and the blue box represents the operations. I also used a detailed model architecture plot for the zero-shot CLIP model and Convolutional Neural Network. The CLIP pre-trained model is trained to maximize the cosine similarity score between paired image and its text caption and learn the image encoder and text encoder. The convolutional neural network is to learn the different local image patterns while reducing the resolutions and using cross-entropy to minimize the loss function.



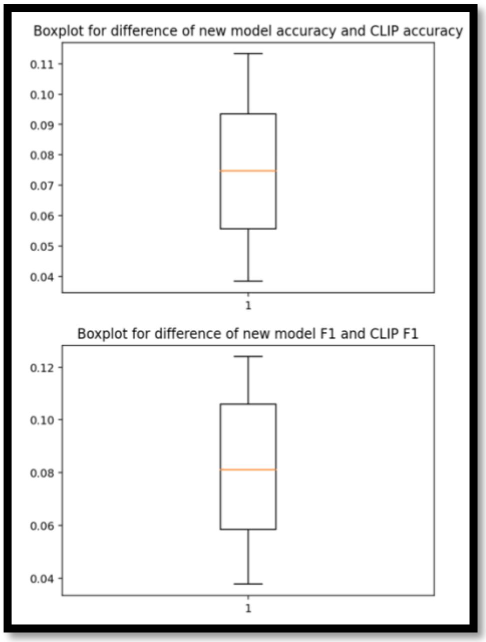
Stratified split the mixture dataset into training and test set

Feed Images into CLIP

Feed zero-shot prediction into

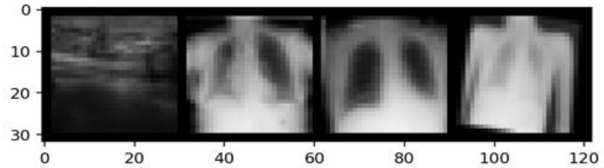
Apply the trained model on the test to compare with

1. Contrastive pre-training
pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ... I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$
2. Create dataset classifier from label text
plane
car
dog
bird
a photo of a {object}.
Text Encoder
3. Use for zero-shot prediction
Image Encoder
I_1
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
a photo of a dog.
Conv_1
Convolution
(5 x 5) kernel
valid padding
Max-Pooling
(2 x 2)
Conv_2
Convolution
(5 x 5) kernel
valid padding
Max-Pooling
(2 x 2)
fc_3
Fully-Connected
Neural Network
ReLU activation
fc_4
Fully-Connected
Neural Network
(with dropout)
INPUT
(28 x 28 x 1)
n1 channels
(24 x 24 x n1)
n1 channels
(12 x 12 x n1)
n2 channels
(8 x 8 x n2)
n2 channels
(4 x 4 x n2)
Flattened
n3 units
0
1
2
9
OUTPUT



4. The new model results as compared to the CLIP model only

To generalize the experimental results and show my new model improved the annotation accuracy, I tested the new model 30 times on the different random training and test set. As the left boxplot shows, my new model, on average, improves the held-out test set 6% and 8% F1 score on this chest and breast binary task. My new model can correct wrong labels generated by CLIP because CNN is powerful enough to learn the general pattern for each category from the majority of correct labels and images. This can be shown in the below graph: the CLIP model mislabels the third and fourth images as breast; however, my new model can correct some mislabeling images; for example, it predicts the third and fourth as a chest.



```
GroundTruth:   breast chest breast breast
Prediction:  breast chest chest chest
```

5. Next steps: I plan to employ Monai pre-trained medical image models in the current zero-shot CLIP architecture once OpenAI publishes its training codes.