Labeling Chest X-ray Reports with Markers of Longitudinal Change

Ryan Han, Christina Kwak, Evang Saracay
{ryanhan, kwakc, esaracay}@stanford.edu

Problem

Background: Chest X-ray (CXR) are the most common imaging examination and critical to diagnosing and managing many medical conditions. Recently, the use of NLP to extract labels from radiology text reports has enabled the large-scale training of deep learning models for clinical applications focusing on a single point in time.

Problem: Many clinical tasks require comparing multiple points in time to understand disease progression - thus extracting labels relating to longitudinal change from radiology text reports would enable the training of AI systems that facilitate tedious comparisons performed by radiologists.

Existing Approaches: little has been done towards characterizing change in imaging datasets.

- Public datasets such as MIMIC-CKD and CheXpert, labeled using NLP, do not contain longitudinal change labels.
- The only existing work that focuses on longitudinal change in CXR uses a rigid text matching approach to match frequent sentences pertaining to disease progression.

Task Proposal

Definition: We formulate the radiology labeling task as a multi-class classification problem where the classes are disease progression, disease stability, and uncertain (no indication).

Datasets: We use 227,827 free text radiology reports from MIMIC-CKD. We randomly selected 3000 reports for manual annotation. Each report was annotated by two human readers, with conflicts determined by committee consensus.

Proposal: The core idea behind our approach is to utilize both strong and weak supervision in order to maximize the performance and label efficiency of our approach. Our supervision strategies are below, and we explore different ways to combine them in our experiments.

- BERT-Phm: Training BERT on a rule-based labeler on all train reports
- BERT-man: Training BERT on a small set of manual annotations
- Distillation: Training BERT on the output of a BERT model on all train reports

Experiment 1 - Supervision Strategies

Approach: We train models using various combinations of strong labels (manual annotations) and weak labels produced from the rule-based phrase matching baseline.

Results:

- 

Experiment 2 - Biomedical Language Representations

Approach: We investigate the effect of pretraining data on model performance.

Results:

- 

Experiment 3 - Distillation

Approach: Taking inspiration from self-distillation, we use our best performing model to produce weak labels on the training set. We then train a BERT model on these labels and fine-tune it using manual annotations.

Results:

- 

Experiment 4 - Fine Tuning Training Set Size

Approach: We investigate the effect of changing the number of training samples on fine-tuning performance.

Results:

- 

Conclusions

Based on our results, we find that rule based labelers in combination with a small manually labeled set are a viable approach for training a model on our task of detecting longitudinal change in chest X-rays. The results of experiments 3 and 4 show promising avenues for future research that can lead to higher accuracy in model performance, especially in low-label settings. We also see that our Distillation model far outperforms our original sentence matching baseline (SM).

Our approach presents a method for accurately extracting labels on medical reports with only a small set of manually labeled reports and a simple rule based labeler.