Image Classification with Consistent Supporting Evidence

Peiqi Wang, Ruizhi Liao, Daniel Moyer, Seth Berkowitz, Steven Horng, Polina Golland
Massachusetts Institute of Technology, Beth Israel Deaconess Medical Center, Harvard Medical School

Introduction

Adoption of machine learning models in healthcare requires end users' trust in the system. Models that provide additional supportive evidence for their predictions facilitate adoption. We build models that supplement their predictions with consistent supporting evidence. In particular, we

- define consistent evidence as being compatible and sufficient
- provide measures of model inconsistency
- propose regularizers to encourage model to provide consistent evidence
- demonstrate our approach on edema severity grading task

Method

Setup

x, y and $z = (z_1, \cdots, z_K)$ - image, C-class task label and K binary evidence labels
$\hat{y}, (\hat{z}_1, \cdots, \hat{z}_K)$ - MAP estimates of task label and evidence labels

Consistent Evidence

We assume domain experts provide logical constraints between task label y and evidence labels z. Specifically, I_1 is indexing function for evidence that is incompatible with a particular value of task label y, while I_2 is indexing function for evidence that directly supports a particular value of task label y.

Definition 1 *(Consistent Evidence) The task label $y \in [C]$ and the evidence label vector $z = (z_1, \ldots, z_K) \in \{-1, +1\}^K$ are consistent if*

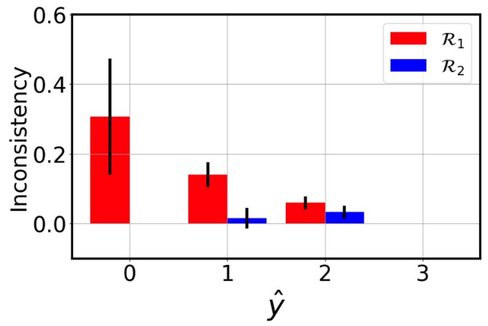
$$\forall k \in \mathcal{I}_1(y) : \; z_k = -1, \tag{3}$$
$$\exists k \in \mathcal{I}_2(y) : \; z_k = +1. \tag{4}$$

Measures of Inconsistency

We quantify inconsistency by bounding the probability of inconsistent evidence. $R_1(D)$ is the average count of evidence labels incompatible with the task label $R_2(D)$ is the average count of absence of direct evidence.

$$\mathcal{R}_1(\mathcal{D}) = \mathbb{E}_{(y,z)\sim\mathcal{D}}\left[\sum_{k\in\mathcal{I}_1(y)} \mathbb{1}\left[z_k = +1\right]\right]$$

$$\mathcal{R}_2(\mathcal{D}) = \mathbb{E}_{(y,z)\sim\mathcal{D}}\left[\min_{k\in\mathcal{I}_2(y)} \left[1 - \mathbb{1}\left[z_k = +1\right]\right]\right]$$
$$= 1 - \mathbb{E}_{(y,z)\sim\mathcal{D}}\left[\max_{k\in\mathcal{I}_2(y)} \mathbb{1}\left[z_k = +1\right]\right].$$



Consistency Regularization

Models trained naïvely to predict y, z jointly are not guaranteed to be consistent. We propose regularizers $R_1(\theta)$ and $R_2(\theta)$ to avoid inconsistent evidence.

$$\mathcal{R}_1(\theta) = -\mathbb{E}_{x\sim\mathcal{D}}\left[\sum_{k\in\mathcal{I}_1(\hat{y}(x))} \ln p(z_k = -1 \mid x; \theta)\right]$$

$$\mathcal{R}_2(\theta) = -\mathbb{E}_{x\sim\mathcal{D}}\left[\ln \max_{k\in\mathcal{I}_2(\hat{y}(x))} p(z_k = +1 \mid x; \theta))\right]$$

Optimization

$$\min_\theta \mathcal{L}(\theta) + \omega_1\mathcal{R}_1(\theta) + \omega_2\mathcal{R}_2(\theta)$$

Edema Severity Grading Task

x - chest X-ray image
$y \in \{0,1,2,3\}$ - edema severity grade (C=4)
z are chosen by domain experts (K=7)
y, z are mined from paired reports
We use residual networks to predict MAP estimates of y, z given x.

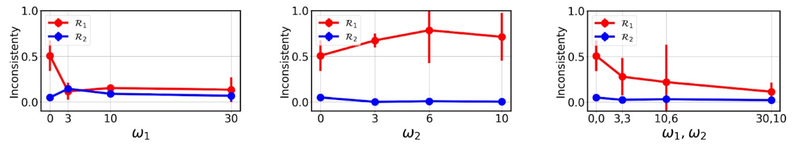
Data

We use a set of 238,086 frontal-view chest X-ray from the MIMIC-CXR data set. We split the data set into training (217,016), validation (10,445), and test (10,625) sets randomly.

Edema severity labels are extracted from associated reports by searching for keywords that are indicative of a specific disease stage. The 7,802 labeled image/report pairs are split into training (6,656), validation (648), and test (498) set. All subsequent evaluation of model consistency and performance is computed on this test set. No patients are shared across training/validation/test sets.

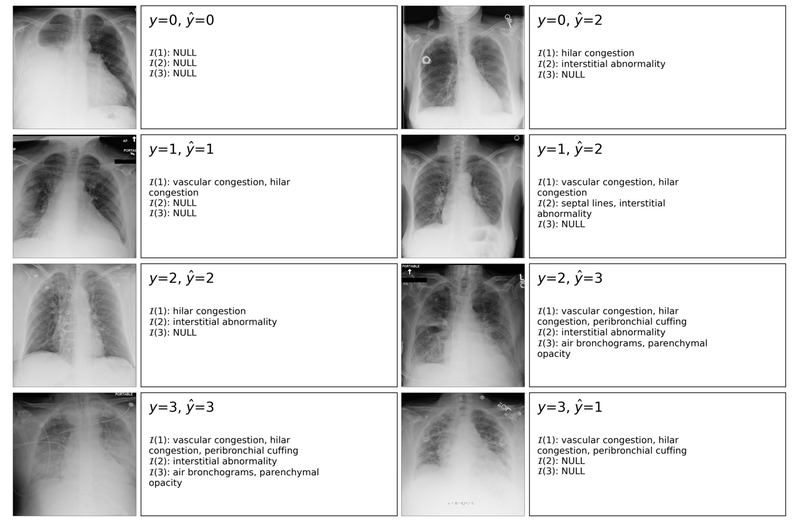
Results

The proposed regularizers $R_1(\theta)$ and $R_2(\theta)$ encourage model to provide more compatible evidence (left) or more sufficient evidence (middle), respectively. Application of regularizers at the same time encourages model to provide more consistent evidence (right).

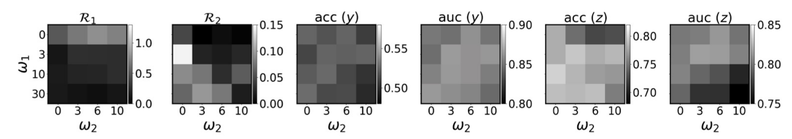


(a) Vary ω_1, fix $\omega_2 = 0$ (b) Fix $\omega_1 = 0$, vary ω_2 (c) Vary $\omega_1 = \omega_2$ together

Correctly and incorrectly classified test image with supporting evidence given by a consistent model ($\omega_1 = \omega_2 = 10$). We use $I(c)$ to denote the set of evidence labels detected in the image that directly support disease stage c



We can ensure satisfactory model consistency. At the same time, the regularized model achieves similar performance on severity grading task y, with tolerable drop in the average performance in evidence detection z.



This work was supported in part by NIH NIBIB NAC P41EB015902 grant, MIT Lincoln Laboratory, MIT JClinic, MIT Deshpande Center, and Philips.