Semi/Self-Supervised Learning on a Pediatric Pneumonia Dataset

by Team Semi-Super CV

Wing Poon
Sundeep Bhimireddy
Sinem Erisken

Agenda

1. Problem: Pediatric Pneumonia
2. Project Goal
3. Dataset
4. Current SOTA
5. ML Workflow
6. Results
7. Demo
8. Summary

PROBLEM: Pediatric Pneumonia

- Pneumonia accounts for 14% of all deaths of children <5yo
- Pediatric pneumonia is often missed and left untreated
- Deaths more prevalent in developing countries where resources are scarce
- Often diagnosed with chest X-ray

SOLUTION:

Semi-Supervised Learning

AI can automatically surface *urgent cases*, helping save lives

Challenge: Fully supervised approaches need **large, densely annotated dataset**

Only hospitals that **can afford** to collect large annotated datasets **can utilize** these approaches to aid their physicians

PROJECT GOAL:

Utilize **Semi-Supervised Learning** to significantly reduce the need for fully labelled data

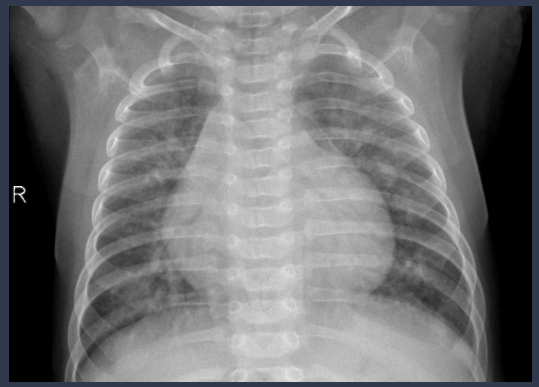
The Data

5855 Total Monochromatic Images
(Train: 5232, Test: 623)

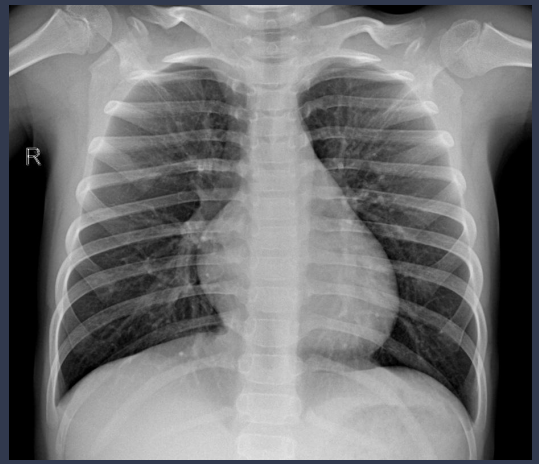
- Pneumonia: (Train: 74%, Test: 62%)
- Normal: (Train: 26%, Test: 38%)

- Widely Varying Resolutions
- Widely Varying Aspect-Ratios
- Splits remixed for consistent Train / Validation / Test

PNEUMONIA



NORMAL



Current State of the Art (Fully Supervised)

Kermany et al. (2018)	Transfer learning	Accuracy	92%
Stephen et al. (2019)	Pure CNN	Validation Accuracy	93.7%
Labhane et al. (2020)	Transfer learning	Accuracy	97%, 98%

Table modified from Labhane et al. (2020)

ML Workflow

Pre-Process Images

Homogenize Train/Val/Test Distributions

ResNet-50-2X (140M Params)

STAGE 1: Self-supervised Contrastive Learning

STAGE 2a: Supervised Fine-Tuning (Few labels)

STAGE 2b: Active Learning (Labeling)

STAGE 3: Knowledge Distillation

ResNet-50 (24M Params)

WEB APP: Prediction Visualization

DEPLOY: Docker Container

SimCLR (v2)

Google Cloud Platform (8 TPU Cores)

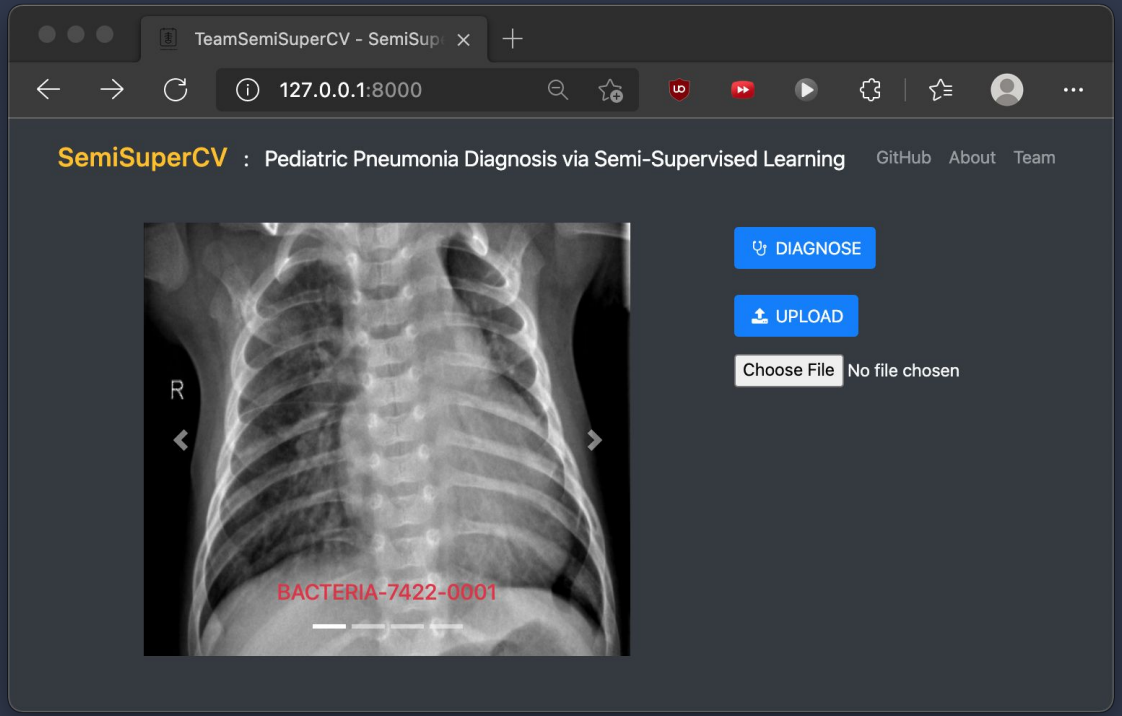
Results

Classification Accuracy (%)

(Num Labels)	Labels	FSL[1]	FixMatch	SimCLR (v2) Stage 2 (Fine-tuning)	SimCLR (v2) Stage 3 (Distillation)
(52)	1%	85.2	92.1	94.5	96.3
(104)	2%	87.2	95.0	96.8	**97.6**
(260)	5%	86.0	98.2	97.1*	**98.1***
(4708)	100%	**98.9**		-	-

[1] Fully-Supervised Learning ** Active Learning*

Demo


TeamSemiSuperCV - SemiSup
127.0.0.1:8000
SemiSuperCV : Pediatric Pneumonia Diagnosis via Semi-Supervised Learning
GitHub
About
Team
R
BACTERIA-7422-0001
DIAGNOSE
UPLOAD
Choose File
No file chosen

Summary

- SSL Model Accuracy of **97.6%** with just **2%** Labels, *vs* FSL at **87.2%** @2% Labels and 98.9% @100% Labels
- Hyper-parameter Tuning (i.e. Validation Set) is necessary for good performance
- Random Cropping Area must be sufficiently large due to small relevance region in subject images
- SimCLR (v2) had better results than FixMatch on this dataset

Future Work

Self-supervised learning (SSL) is one of the most promising ways to build background knowledge and approximate a form of common sense in AI systems

- Yann LeCun

- More *augmentations* during fine tuning to improve active learning
- Explore *generalizability* of hyperparameters to other datasets
- Different active learning labeling *selection algorithms*
- Try other *SoTA* self-supervised *frameworks*