Review: Who Said What: Modeling Individual Labelers Improves Classification

Melody Y. Guan, Varun Gulshan, Andrew M. Dai, Geoffrey E. Hinton

March 31 2017

Abstract

- Different experts
- Multiple labels
- Experts disagreement
- Majority opinion
- Modeling experts individually
- Learning averaging weights
- Finding reliable expert

Introduction

- Deep CNN and rapid improvement
- Expect: neural networks to serve as alternative to human experts
- Experts are unreliable
- Poor agreement between experts (55.4% in diabetic retinopathy)
- Same expert agreement different time (70.7%)
- Better ways to use opinions of multiple experts?

Introduction

- Better model by predicting opinions of individual labelers?
- Some doctors are more reliable
- Upweight reliable opinions
- Doctors will receive different training
- Doctors will receive different distributions of images
- Relative reliability of two doctors: class of image, properties of image

Introduction

- Making Better use of noisy labels:
- Force the network to predict each doctor
  - An output layer per doctor
  - Test time: compute and average the predictions
  - Learn how much to weight each modeled doctor
  - Down-weight unreliable models

Dataset

- Dataset: screening diabetic retinopathy (DR)
- Human-level performance (Gulshan et al. 2016)
  - Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs
- Average opinion of all doctors to obtain ground truth
- Training: 126,522 and Validation: 7,805 images
- Total 54 labelers/graders
- Test: 3,547 images (labelers excluded from training/validation)

Noisy labels

- Corrupting true labels
- True labels: error *1.01%*
- Corrupt with *p=0.5*: *2.29%*
- Corrupt with *p=0.8*: *8.23%*

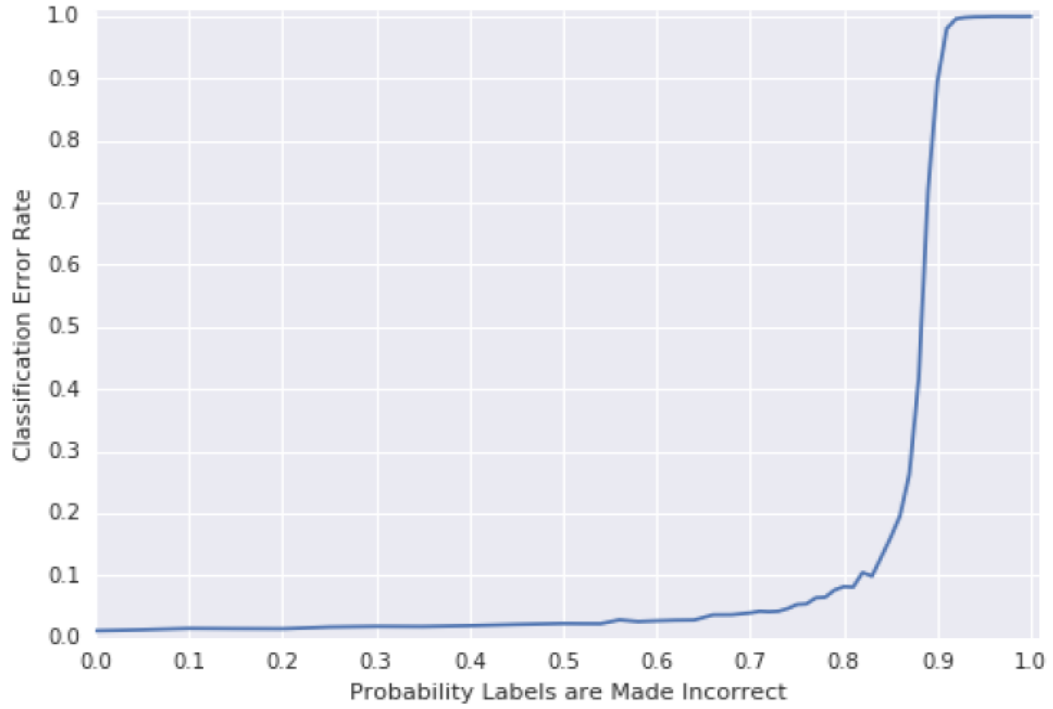


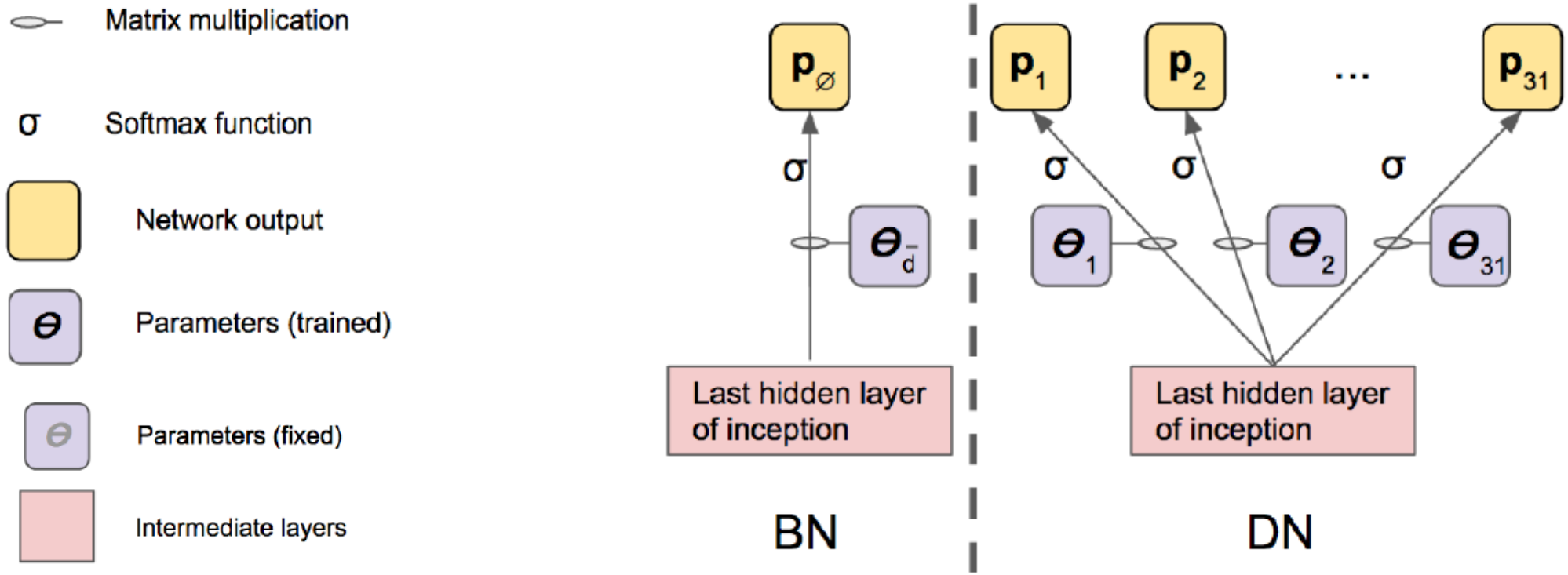
Figure 1: Performance of a deep neural net when trained with noisy labels.

Methods

- Set of indices and *I_i* the label of doctor *I*
- *Baseline Net (BN):* Inception-v3 with average opinions
- Doctor Net (*DN*): extended to model each doctor
- Weighted Doctor Net (*WDN*)
- Image-specific WDN (*IWDN*)
- Bottlenecked *IWDN*

Architectures

- Base network vs Doctor network



Results

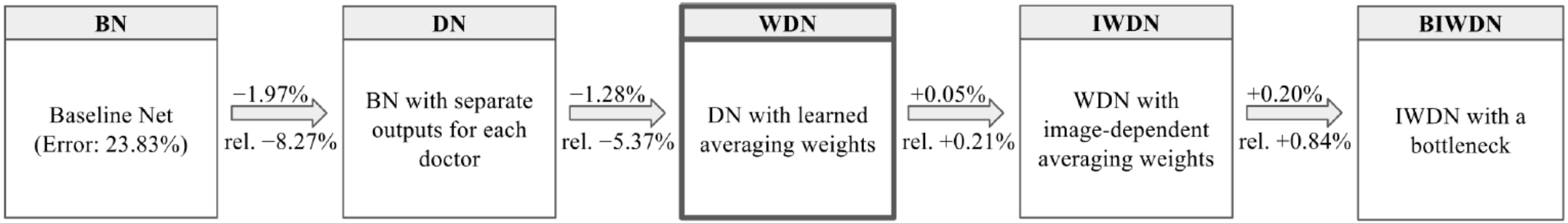
- For training doctor models: 1000 images
- Training with five-class loss beats training with binary loss
- Sensitivity: true positive rate
- Specificity: true negative rate

Table 4: Test metrics from Multi-class vs Binary loss for BN.

Test Metric (%)	Trained with binary loss	Trained with 5-class loss
Binary AUC	95.58	97.11
Binary Error	11.27	9.92
Spec@97%Sens	63.12	79.60

Results

- Expect more improvement if doctors with more varied abbilities



Conclusion