Critique of Right for the Right Reasons: Training Neural Networks to be Interpretable, Robust, and Consistent with Expert Knowledge

William Jardee WILLJARDEE@GMAIL.COM
Physics Department
Montana State University
Bozeman, MT 59715, USA

Editor:

1. Introduction

2. Related Works

2.1 Explainable Models

2.2 Ensemble and Adversarial Methods

2.3 Human Centered Interpretability Measures

2.4 Interpretable Representations

3. Right for the Right Reasons Algorithm

$$\mathcal{L}(\theta, \mathbf{X}, y, A) = \underbrace{\sum_{n=1}^{N}\sum_{k=1}^{K} -y_{nk}\,\log(\hat{y}_{nk})}_{\text{Right Reason (Cross-Entropy)}} + \underbrace{\lambda_1 \sum_{n=1}^{N}\sum_{d=1}^{D}\left[A_{nd}\frac{\partial}{\partial x_{nd}}\sum_{k=1}^{K}\log(\hat{y}_{nk})\right]^2}_{\text{Right Reasons}} + \underbrace{\lambda_2\sum_i \theta_i^2}_{\text{Regularizer}} \tag{Equation 1}$$

$$\mathcal{L}(\{\theta_m\}) = \underbrace{\sum_m \mathbb{E}_{p(x,y)}\left[-\log(p(y|f(x;\theta_m))\right]}_{\text{Predictive Term (Cross-Entropy)}} + \underbrace{\lambda\sum_{l\neq m}\textbf{IndepErr}\left(f(\cdot;\theta_m), f(\cdot;\theta_l)\right)}_{\text{Diversity Measurement}} \tag{Equation 2}$$

$$\approx \underbrace{\sum_m \mathbb{E}_{p(x,y)}\left[-\log(p(y|f(x;\theta_m))\right]}_{\text{Predictive Term (Cross-Entropy)}} + \underbrace{\lambda\sum_{l\neq m}\textbf{CosIndepErr}\left(f(\cdot;\theta_m), f(\cdot;\theta_l)\right)}_{\text{Diversity Measurement}} \tag{Equation 3}$$

where

$$\textbf{IndepErr}(f,g) = \mathbb{E}\left[\left(f(x_{g_{\max}}) - f(x)\right)^2\right] \approx \left(\epsilon\,\boldsymbol{\nabla} f(x)\cdot\boldsymbol{\nabla} g(x)\right)^2\left(\frac{\boldsymbol{\nabla} f(x)\cdot\boldsymbol{\nabla} g(x)}{|f(x)|_2|g(x)|_2}\right)^2 \equiv \cos^2(\boldsymbol{\nabla} f(x), \boldsymbol{\nabla} g(x)) \equiv$$

Equation (1)

4. Algorithm Experiments

5. Application of Algorithm in Adversarial Context

6. Interpretability with Human's Study

7. Conclusion