

A Appendix

An Example of Surrogate Loss for FNR For our particular example involving minimizing $f = \text{FNR} = \frac{\text{FN}}{\text{total positives}}$, recall that $\text{FN}(s_{\theta|\lambda}) = \sum_{i:y_i=1} \mathbb{1}(p_i(\theta) \leq \lambda)$ where p_i is the prediction for the i -th positive example. We replace the indicator function with its smooth surrogate and get $\widetilde{\text{FN}}(s_{\theta|\lambda}) = \sum_{i:y_i=1} \sigma_{\tau}(-(p_i - \lambda))$, where $\sigma_{\tau}(u) = 1/(1 + \exp(-\tau u))$ denotes a temperature scaled sigmoid function. This gives us $\tilde{f} = \frac{\widetilde{\text{FN}}}{\text{total positives}}$ as the surrogate loss.

Implicit Function Theorem on θ Here we provide the statement for the implicit function theorem and its modification for the last layer weights:

Theorem 1 (Implicit Function Theorem on θ [10], informal). *For any $(\theta_0, \lambda_0) \in U \subseteq \mathbb{R}^p \times \mathbb{R}^m$ pair that satisfies $\tilde{g}(\theta_0, \lambda_0) = 0$, if the determinant of the Jacobian matrix is nonzero, i.e., $\det[\frac{\partial \tilde{g}^i}{\partial \theta^j}(\theta_0, \lambda_0)] \neq 0$, then there exists a neighborhood $\Theta \times \Lambda$ of (θ_0, λ_0) in U and a unique function $\tilde{h} : \Theta \Rightarrow \Lambda$:*

$$\tilde{g}(\theta, \lambda) = \mathbf{0} \Leftrightarrow \lambda = \tilde{h}(\theta). \quad (7)$$

Theorem 2 (Implicit Function Theorem on θ_L). *For any $(\theta_0, \lambda_0) \in U \subseteq \mathbb{R}^p \times \mathbb{R}^m$ pair that satisfies $\tilde{g}(\theta_0, \lambda_0) = \mathbf{0}$, if the determinant of the Jacobian matrix w.r.t θ_L is nonzero, i.e. $\det[\frac{\partial \tilde{g}^i}{\partial \theta_L^j}(\theta_0, \lambda_0)] \neq 0$, then there exists a neighborhood $\Theta \times \Lambda$ of (θ_0, λ_0) in U and a unique function $\tilde{h}_L : \Theta \Rightarrow \Lambda$ such that*

$$\tilde{g}(\theta, \lambda) = 0 \Leftrightarrow \lambda = \tilde{h}_L(\theta_L).$$

Gradient Update Rule for Eq. (3) To compute a local derivative for $\tilde{f}(\theta, \tilde{h}(\theta))$ within the neighborhood of θ_0 using Theorem 1 we have the following update rule:

$$\nabla_{\theta} \tilde{f}(\theta, \tilde{h}(\theta)) = \nabla_{\theta} \tilde{f}(\theta, \lambda) + \frac{\partial \tilde{f}(\theta, \lambda)}{\partial \lambda} \nabla_{\theta} \tilde{h}(\theta) \quad (8)$$

We will further need the derivative of the implicit function \tilde{h} w.r.t. θ , i.e. $\nabla_{\theta} \tilde{h}(\theta)$. Since $\tilde{g}(\theta, \tilde{h}(\theta)) = \mathbf{0}$ in the neighborhood of θ_0 , we have:

$$\nabla_{\theta} \tilde{g}(\theta, \lambda) + \frac{\partial \tilde{g}(\theta, \lambda)}{\partial \lambda} \nabla_{\theta} \tilde{h}(\theta) = \mathbf{0} \Rightarrow \nabla_{\theta} \tilde{h}(\theta) = -\frac{\nabla_{\theta} \tilde{g}(\theta, \lambda)}{\frac{\partial \tilde{g}(\theta, \lambda)}{\partial \lambda}} \quad (9)$$

Plugging in Eq. (9) back to Eq. (8), we can get the final gradient for the model parameter θ . See section 3.3 of [6] for a more detailed derivations for the update rule of θ and λ .

Gradient Update Rule for Eq. (4) Furthermore, we can break the gradient of Eq. (4) into two parts: the gradient w.r.t θ_L , thus obtaining the gradient w.r.t θ_L and obtain the following gradients³:

$$\partial_{\theta_L} \tilde{f}(\theta, \lambda) = \nabla_{\theta_L} \tilde{r} \quad \text{and} \quad \partial_{\theta_L} \tilde{f}(\theta, \lambda) = \nabla_{\theta_L} \tilde{\ell} + \frac{\partial \tilde{\ell}}{\partial \lambda} \nabla_{\theta_L} \tilde{h}_L(\theta_L) \quad (10)$$

Experimental Details Following [6], we choose three binary attributes (i.e. High-cheekbones, Smiling and Wearing-lipsticks) as target attribute for our experiments, and train a binary classifier for each attribute. Similar to [6], we use a 6-layer neural network with 5 convolutionary layers with 128, 256, 512, 512 filters respectively, and we use ReLU as our activation functions and batch normalization layers in the networks. We do 5 random trails for each experiments and report the average values of the metric.

³When it is clear from the content, we use $\tilde{\ell}$ to denote $\tilde{\ell}(\theta_L|\theta_L, \lambda)$ to ease the notation, same applies for \tilde{r} , $\widetilde{\text{FPR}}$, and $\widetilde{\text{FNR}}$.