Supplementary Materials: Additive Adversarial Learning for Unbiased Authentication

Jian Liang[1*], Yuren Cao[1,2*], Chenbin Zhang[1,2], Shiyu Chang[3], Kun Bai[1], Zenglin Xu[2]
[1]Cloud and Smart Industries Group, Tencent, China
[2]University of Electronic Science and Technology of China
[3]MIT-IBM Watson AI Lab, IBM Research, USA
{joshualiang,laurenyrcao,kunbai}@tencent.com
ChenbinZhang@std.uestc.edu.cn, shiyu.chang@ibm.com, zenglin@gmail.com

Methods	aAUC	(aFAR + aFRR)/2	ACC@1
F-Colors, Direct	66.61	41.56	10.06
F-Colors, Stage 1	87.59	18.95	42.33
F-Colors, Stage 1+2	**92.93**	**13.61**	**58.32**
B-Colors, Direct	79.63	30.46	25.47
B-Colors, Stage 1	94.38	11.75	62.78
B-Colors, Stage 1+2	**96.32**	**9.56**	**68.83**

Table S1. Average performances (%) of different domain-differences on the C-MNIST data set. F-color and B-color stand for foreground and background colors, respectively.

S1. Stability for Different Domain-Differences

To investigate the stability of our proposed method, we conduct experiments with various domain-differences on the C-MNIST data set. We consider each of the background color and the foreground color as a type of domain-difference that groups digits, the other is allowed to share digits for differ Ft domains.or eaench case, we randomly select 10 combinations of two different colors as two domains. The average performances reported in Table S1 show stable performances for both our OVRDL (stage 1) and our AAL (stage 2) mechanisms.

S2. Proof for Theorem 1

Proof. Use the notations in Section S3.

If we have

$$P(Z \mid \mathbf{F}) = P(Z), \tag{S1}$$

then $\mathbf{F}$ and Z are independent. Therefore, let z be an arbitrary value of Z, we have

$$P(\hat{Y} \mid \mathbf{F}(Y, Z = z)) = P(\hat{Y} \mid \mathbf{F}(Y, Z \neq z)), \tag{S2}$$

which satisfy *Equality of Odds*:

$$P(\hat{Y} \mid Y, Z) = P(\hat{Y} \mid Y). \tag{S3}$$

For our model and optimization scheme, as mentioned in Section S3, our discriminative networks learn to let $P(\hat{Z} \mid \mathbf{F}) = P(Z \mid \mathbf{F})$. Then if we add a constraint to let $P(\hat{Z} \mid \mathbf{F})$ equal $P(Z)$, the *Equality of Odds* can be achieved.

We can provide a straightforward strategy to let $P(\hat{Z} \mid \mathbf{F})$ equal $P(Z)$. For $r \in [k_z]$, denote $p_r = P(\hat{Z} = r \mid \mathbf{F})$. Without loss of generality, we assume $P(Z = r) = \frac{1}{k_z}$ for all $r \in [k_z]$. Then in Eq. (4), we can minimize the square loss $(p_r - \frac{1}{k_z})^2$ to let $P(\hat{Z} \mid \mathbf{F})$ equal $P(Z)$.

In fact, in our Eq. (4) we optimize

$$\begin{aligned} &\min_{\{p_r\}} (p_{r^*} - 0)^2 + \sum_{r \neq r^*} (p_r - 1)^2, \\ &\text{s.t.} \sum_r p_r = 1, p_r \geq 1, \forall r \in [k_z]. \end{aligned} \tag{S4}$$

where r^* denotes the true label of Z for a specific data instance. Such an optimization problem has the optimum solution that $p_r = 0$ if $r = r^*$ and $p_r = 1/(k_z - 1)$ otherwise, which is close to the above solution that $p_r = \frac{1}{k_z}$ $(r \in [k_z])$ for large k_z. We find our solution is simpler to implement and has slightly better generalization performances in practise.

S3. Proof for Theorem 2

Proof. Let $Y \in [k_y]$ denote the variable of the true label of the jth attribute, and let $Z \in [k_z]$ denote the variable of the true label of any other attribute j'. Denote the predicted label of the Y and Z by $\hat{Y}$ and $\hat{Z}$, respectively.

Let $\mathbf{F}$ be the attribute vector variable of the jth attribute. Consider $\mathbf{F}$ as a function of Y and Z: $(Y, Z) \to \mathbf{F}$, and denote $\mathbf{F}(Y, Z)$ the attribute vector resulted from Y and Z.

*Equal contribution from both authors.

Let y and z be arbitrary values of Y and Z, respectively. The learning goals starts from

$$P(\hat{Z} = z \mid \mathbf{F}(Z = z)) = P(\hat{Z} = z \mid \mathbf{F}(Z \neq z)),$$
$$P(\hat{Y} = y \mid \mathbf{F}(Y = y)) = P(\hat{Y} \neq y \mid \mathbf{F}(Y \neq y)) = 1, \tag{S5}$$

and

$$\begin{aligned} P(\hat{Z} = z \mid \mathbf{F}) &= P(Z = z \mid \mathbf{F}), \\ P(\hat{Y} = y \mid \mathbf{F}) &= P(Y = y \mid \mathbf{F}), \end{aligned} \tag{S6}$$

and finally become

$$P(Z = z \mid \mathbf{F}(Z = z)) = P(Z = z \mid \mathbf{F}(Z \neq z)), \tag{S7}$$

and

$$P(Y = y \mid \mathbf{F}(Y = y)) = P(Y \neq y \mid \mathbf{F}(Y \neq y)) = 1. \tag{S8}$$

Since z is arbitrary, Eq. (S7) holds only if $\mathbf{F}$ and Z are independent. Meanwhile, since y is arbitrary, Eq. (S8) suggests that $\mathbf{F}$ and Y are not independent. Therefore, if we have a causality that $Z \to Y$, $\mathbf{F}$ and Z will be not independent, thus Eq. (S7) will not hold. But if the causal direction is reversed, *i.e.*, if we have a causality that $Y \to Z$, although $\mathbf{F}$ and Y are not independent, it is possible that $\mathbf{F}$ and Z are independent.

On the other hand, if Eq. (S7) holds, $\mathbf{F}$ and Z are independent. Then if Z is the only cause of Y, $\mathbf{F}$ and Y will also be independent. Then Eq. (S8) will not hold. In addition, if Z is one of many causes of Y, since $\mathbf{F}$ and Z are independent, then $\mathbf{F}$ is independent with the effect of Z on Y. Therefore, the correlation between $\mathbf{F}$ and Y is limited, and thus Eq. (S8) cannot hold perfectly.

□