Mitigating Information Leakage in Image Representations: A Maximum Entropy Approach
(supplementary Material)

Proteek Chandan Roy and Vishnu Naresh Boddeti
Department of Computer Science and Engineering
Michigan State University, East Lansing MI 48824
{royprote, vishnu}@msu.edu

In this supplementary material we include proof of Theorem 1 in Section 1, Corollary 1.1 in Section 2 and finally provide the numerical values of the trade-off fronts in the CIFAR-10 and CIFAR-100 experiment in Section 3.

1. Proof of Theorem 1

**Theorem 1.** Given a fixed encoder $E$, the optimal discriminator is $q_D(s|E(x; \theta_E); \theta_D^*) = p(s|E(x; \theta_E))$ and the optimal predictor is $q_T(t|E(x; \theta_E); \theta_T^*) = p(t|E(x; \theta_E))$.

**Proof.** Let, $z$ be the fixed encoder output from input $x$ i.e. $z = E(x; \theta_E)$. Let, $p(x, t, s)$ be the true joint distribution of the variables, i.e. input $x$, target label $t$ and sensitive label $s$. The fixed encoder is a deterministic transformation of $x$ and generates an implicit distribution $p(z, t, s)$.

**Discriminator:** The objective of the discriminator is,

\[
V_1(\theta_E, \theta_D) = KL(p(s|x)||q_D(s|E(x; \theta_E); \theta_D)) = \sum_{x,t,s} p(x,t,s) \log q_D(s|z; \theta_D) - \log q_D(s|z; \theta_D) \leq 0, \forall z
\]

s.t. $\sum_s q_D(s|z; \theta_D) = 1, \forall z$

The Lagrangian dual of the problem can be written as

\[
L = V_1(\theta_E, \theta_D) + \sum_z \lambda(z) \left( \sum_s q_D(s|z; \theta_D) - 1 \right)
\]

Now we take partial derivative of $L$ w.r.t. $q_D(s|z; \theta_D^*)$, the distribution of optimal discriminator. Therefore, the optimal discriminator satisfies,

\[
\frac{\partial L}{\partial q_D(s|z; \theta_D^*)} = 0
\]

\[
\Rightarrow - \sum_t p(z, t, s) \log q_D(s|z; \theta_D^*) + \lambda(z) = 0
\]

\[
\Rightarrow \lambda(z) q_D(s|z; \theta_D^*) = p(z, s)
\]

where we used the fact that, $\sum_t p(z, t, s) = p(z, s)$. Now summing w.r.t. to variable $s$ on the both sides of last line and using the fact that $\sum_s q_D(s|z; \theta_D^*) = 1$ we get,

\[
\lambda(z) = p(z)
\]

By substituting $\lambda(z)$ we obtain the solution for the optimal discriminator,

\[
q_D(s|z; \theta_D^*) = \frac{p(z, s)}{p(z)} = p(s|z)
\]

Therefore,

\[
q_D(s|E(x; \theta_E); \theta_D^*) = p(s|E(x; \theta_E))
\]

**Target Predictor:** The objective of the predictor is,

\[
V_2(\theta_E, \theta_T) = KL(p(t|x)||q_T(t|E(x; \theta_E); \theta_T)) = \sum_{x,t} p(x,t) \log q_T(t|z; \theta_T) - \log q_T(t|z; \theta_T) \leq 0, \forall z
\]

s.t. $\sum_t q_T(t|z; \theta_T) = 1, \forall z$

The Lagrangian dual of the problem can be written as

\[
L = V_2(\theta_E, \theta_T) + \sum_z \lambda(z) \left( \sum_t q_T(t|z; \theta_T) - 1 \right)
\]
Now we take partial derivative of \( L \) w.r.t. \( q_T(t|z; \theta_T^*) \), the distribution of optimal predictor. The optimal predictor satisfies the equation.

\[
\frac{\partial L}{\partial q_T(t|z; \theta_T^*)} = 0
\]

\[
\Rightarrow -\sum_i p(z, t, s) \frac{q_T(t|z; \theta_T^*)}{q_T(t|z; \theta_T^*)} + \lambda(z) = 0
\]

\[
\Rightarrow \lambda(z) q_T(t|z; \theta_T^*) = p(z, t)
\]

where we used the fact that, \( \sum_s p(z, t, s) = p(z, t) \). Now summing w.r.t. to variable \( t \) on the both sides of last line and using the fact that \( \sum_t q_T(t|z; \theta_T^*) = 1 \) we get,

\[
\lambda(z) = p(z)
\]

By substituting \( \lambda(z) \) we obtain the solution of the optimal discriminator

\[
q_T(t|z; \theta_T^*) = \frac{p(z, t)}{p(z)} = p(t|z)
\]

Therefore,

\[
q_T(t|E(x; \theta_E); \theta_T^*) = p(t|E(x; \theta_E))
\]

\[\square\]

2. Proof of Corollary 1.1

**Corollary 1.1.** When \( s \perp t \), let the optimum discriminator and predictor for an encoder \( E \) be \( q_D(s|E(x; \theta_E); \theta_D^*) \) and \( q_T(t|E(x; \theta_E); \theta_T^*) \) respectively. The optimal encoder \( E(\cdot) \) in the MaxEnt-ARL formulation induces a uniform distribution in the discriminator \( q_D(s|E(x; \theta_E); \theta_D^*) \) over the classes of the sensitive attribute.

**Proof.** Here we will prove that, when discriminator is fixed, then the encoder learns a representation of data \( x \) such that \( q_D(s|E(x; \theta_E); \theta_D^*) = 1/m \). First we note that although the discriminator is fixed, the discriminator probability \( q_D(s|E(x; \theta_E); \theta_D^*) \) can change by changing the encoder parameters \( \theta_E \). Optimization of the encoder in MaxEnt-ARL is formulated as:

\[
\min \ V = \min_{\theta_E} \mathbb{E}_{(x,t,s) \sim p(x,t,s)} \left[ -\log q_T(t|E(x; \theta_E); \theta_T^*) \right]
\]

\[
+ \alpha \mathbb{E}_{x} \left[ \sum_{i=1}^{m} q_D(s_i|E(x; \theta_E); \theta_D^*) \log q_D(s_i|E(x; \theta_E); \theta_D^*) \right]
\]

\[
+ \log m
\]

s.t. \( \sum_{i=1}^{m} q_D(s_i|E(x; \theta_E); \theta_D^*) = 1 \)

\[
q_D(s_i|E(x; \theta_E); \theta_D^*) \geq 0, \ \forall i
\]

The Lagrangian dual of the problem can be written as,

\[
L = V - \lambda \left( \sum_{i=1}^{m} q_D(s_i|E(x; \theta_E); \theta_D^*) - 1 \right)
\]

Here \( \lambda \) is a Lagrangian multiplier and is assumed to be a constant in the absence of any further information. Since \( s \perp t \), we have \( q_T(t|E(x; \theta_E); \theta_T^*) \) is independent of \( q_D(s|E(x; \theta_E); \theta_D^*) \) given \( E(x; \theta_E) \) from Theorem 1. Therefore, if we take derivative of \( L \) w.r.t. \( q_D(s_i|E(x; \theta_E); \theta_D^*) \) and set it to zero we have:

\[
\frac{\partial L}{\partial q_D(s_i|E(x; \theta_E); \theta_D^*)} = 0
\]

\[
\Rightarrow 1 + \log (q_D(s_i|E(x; \theta_E); \theta_D^*) - \lambda = 0
\]

\[
\Rightarrow q_D(s_i|E(x; \theta_E); \theta_D^*) = \exp (\lambda - 1)
\]

Using the first (non-trivial) constraint, we have

\[
\sum_{i=1}^{m} q_D(s_i|E(x; \theta_E); \theta_D^*) = 1
\]

\[
\sum_{i=1}^{m} \exp (\lambda - 1) = 1
\]

\[
\exp (\lambda - 1) \sum_{i=1}^{m} 1 = 1
\]

\[
m(\exp (\lambda - 1)) = 1
\]

\[
\lambda = \log (1/m) + 1
\]

Hence, the probability distribution of the discriminator after the encoder’s parameters \( \theta_E \) are optimized is \( q_D(s_i|E(x; \theta_E); \theta_D^*) = 1/m \). Thus, when the optimum discriminator parameters are fixed, the encoder optimizes the representation such that the discriminator does not leak any information, i.e., it induces a uniform distribution. \( \square \)

3. CIFAR Trade-Off

We report the numerical values of the target accuracy and adversary accuracy trade-off results on the CIFAR-10 and CIFAR-100 experiments in Table 1 and Table 3, respectively. Similarly, we report the numerical values of the target accuracy and adversary entropy trade-off results on the CIFAR-10 and CIFAR-100 experiments in Table 2 and Table 4, respectively.
<table>
<thead>
<tr>
<th>(a) No Privacy</th>
<th>(b) ML-ARL</th>
<th>(c) MaxEnt-ARL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target Accuracy (%)</strong></td>
<td>97.75 97.73 97.68</td>
<td>97.52 97.44 97.35 91.52 91.15 60.00</td>
</tr>
<tr>
<td>Adversary Accuracy (%)</td>
<td>23.44 23.09 22.68</td>
<td>20.83 20.77 20.64 19.68 14.27 10.00</td>
</tr>
</tbody>
</table>

Table 1: CIFAR-10: Target Accuracy (%) vs Adversary Accuracy

<table>
<thead>
<tr>
<th>(a) No Privacy</th>
<th>(b) ML-ARL</th>
<th>(c) MaxEnt-ARL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target Accuracy (%)</strong></td>
<td>97.75 97.73 97.71</td>
<td>97.52 97.50 96.58 95.97 60.00</td>
</tr>
<tr>
<td>Adversary Entropy (nats)</td>
<td>1.65 1.65 1.67</td>
<td>1.65 1.66 1.80 2.16 2.30</td>
</tr>
</tbody>
</table>

Table 2: CIFAR-10: Target Accuracy (%) vs Adversary Entropy

<table>
<thead>
<tr>
<th>(a) No Privacy</th>
<th>(b) ML-ARL</th>
<th>(c) MaxEnt-ARL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target Accuracy (%)</strong></td>
<td>71.99 71.56</td>
<td>71.32 64.90 56.99 54.46 24.66 22.22 5.00</td>
</tr>
<tr>
<td>Adversary Accuracy (%)</td>
<td>30.69 30.59</td>
<td>2.50 2.51 2.68 2.88 3.77 3.88 4.60</td>
</tr>
</tbody>
</table>

Table 3: CIFAR-100: Target Accuracy (%) vs Adversary Accuracy

<table>
<thead>
<tr>
<th>(a) No Privacy</th>
<th>(b) ML-ARL</th>
<th>(c) MaxEnt-ARL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target Accuracy (%)</strong></td>
<td>71.17 70.80 70.50 67.63 63.81 61.98 59.11 56.32 5.37 5.00</td>
<td>71.32 64.90 56.99 54.46 24.66 22.22 5.00</td>
</tr>
<tr>
<td>Adversary Accuracy (%)</td>
<td>16.88 16.60 16.43 13.23 8.38 5.02 3.80 2.81 1.23 1.00</td>
<td>2.50 2.51 2.68 2.88 3.77 3.88 4.60</td>
</tr>
</tbody>
</table>

Table 4: CIFAR-100: Target Accuracy vs Adversary Entropy