Fundamental Limits of Perfect Concept Erasure

Somnath Basu Roy Chowdhury, Avinava Dubey, Ahmad Beirami, Rahul Kidambi, Nicholas Monath, Amr Ahmed, Snigdha Chaturvedi
{somnath, snigdha}@cs.unc.edu, {avinavadubey, beirami, rahulkidambi, nmonath, amra}@google.com

Concept Erasure

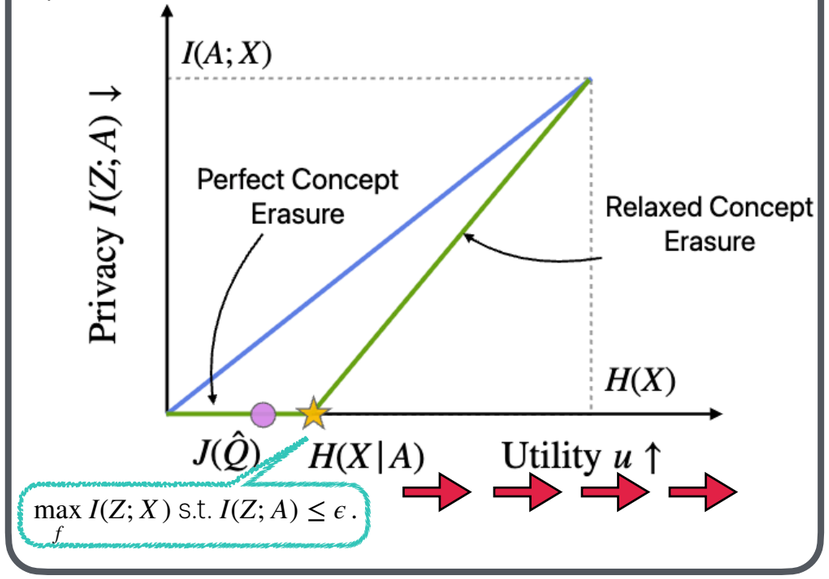
- Concept erasure is the task of perfectly erasing information from representations
- The erasure process should also retain maximum utility about original representations
- Formally, this task optimizes an objective to learn the erasure function $f : X \to Z$

$$\max_f I(Z;X) \text{ subject to } I(Z;A) \le \epsilon \quad (1)$$

Utility — Privacy

Fundamental Limits of Erasure

- Directly optimizing Eq. 1 is difficult for high dimensional representations
- Instead, we study the fundamental limits of each of utility and privacy in Eq. 1 (shown in the figure below)
- Next, we utilize these limits to analytically derive perfect erasure functions

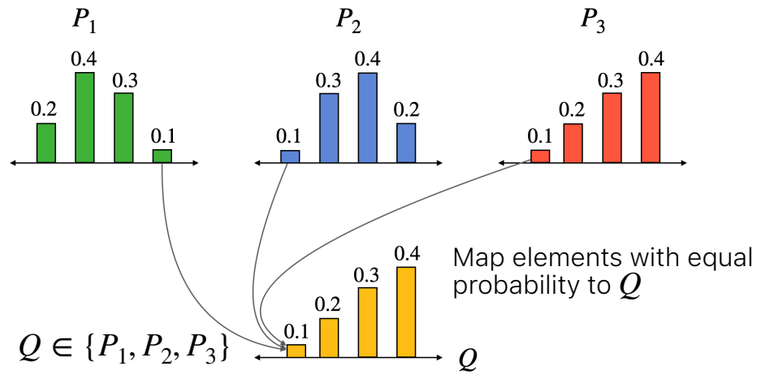


Perfect Erasure Functions (PEF)

Feasibility & Data Constraints:

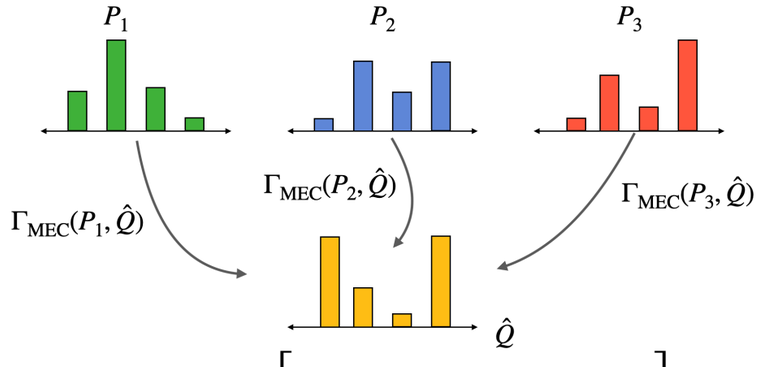
Optimally Perfect concept erasure is feasible if, $\forall(i,j)$ pairs: Distribution of representations are permutations, $P(X|A=a_i) = \sigma(P(X|A=a_j))$

PEF when distributions are equal:



PEF when distributions are unequal:

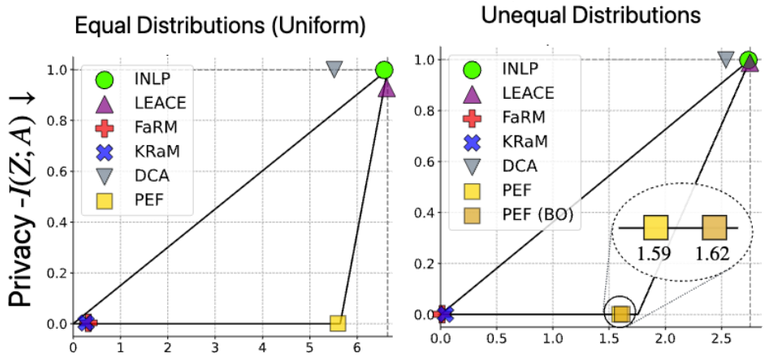
$\Gamma_{\text{MEC}}(\cdot,\cdot)$ is the minimum entropy coupling (minimizes $H(P_i, \hat{Q})$)



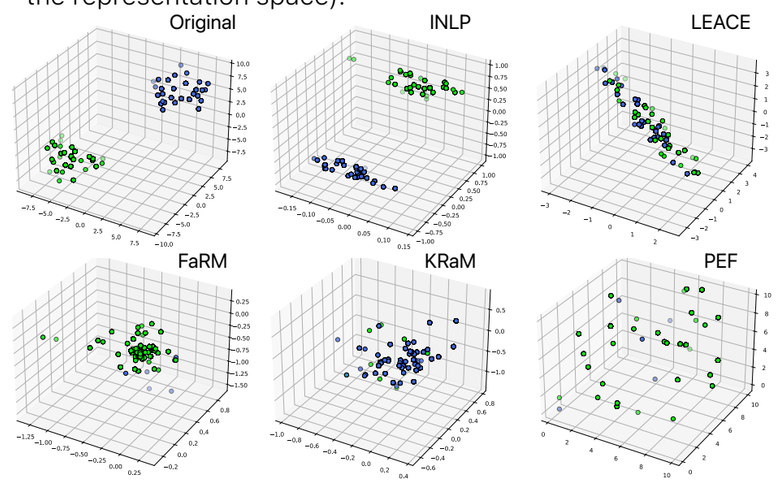
$$\hat{Q} = \arg\max_Q \left[H(Q) - \sum_i p(a_i) H_{\min}(P_i, Q) \right]$$

Empirical Results

- PEFs achieve the theoretical bounds empirically



- Through visualization, we observe that PEF perfectly erases concept information w/o losing other information (collapsing the representation space).



Conclusion

- We analytically derive perfect erasure functions (PEF) for concept erasure
- PEFs perfectly erase concept information while retaining maximum utility

Link to Paper!

brcsomnath/PEF