Robust Concept Erasure via Kernelized Rate-Distortion Maximization

UNC NLP

Somnath Basu Roy Chowdhury, Nicholas Monath, Avinava Dubey, Amr Ahmed, and Snigdha Chaturvedi

{somnath, snigdha}@cs.unc.edu, {nmonath, avinavadubey, amra}@google.com

Google Research

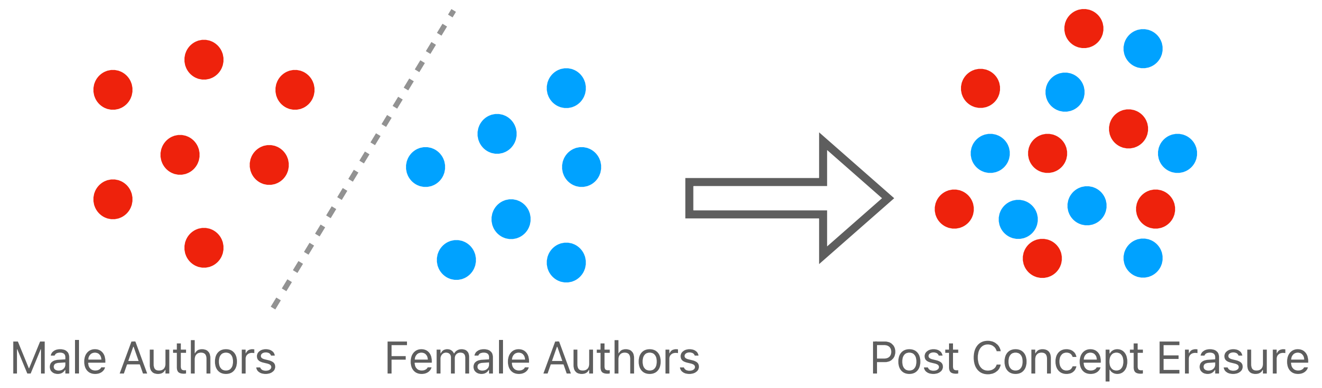
Google DeepMind

Introduction

- Concept Erasure is the task of deleting a *concept* from a representation set.
- A concept is a random variable (categorical, continuous, or vector-valued) that can be inferred from representations.
- Applications of concept erasure include removing:
  - Gender or race from LLM-based text representations
  - Facial features from image representations
  - Prior trending success from trade recommenders

Intuition behind KRaM

- Information in high dimensions are encoded as distances between points. E.g., a biased set of text representations are shown below:

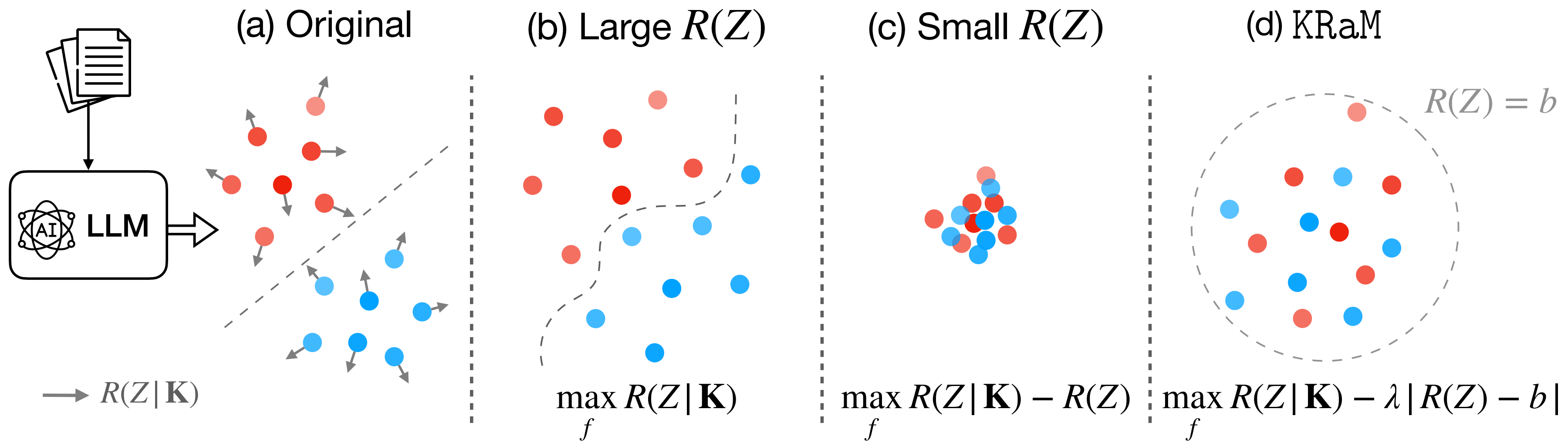


- Consider a feature space, $\mathscr{F} = \{F_1, \ldots, F_n\}$, with a set of subspaces
- Each of the subspaces denotes a concept class or subgroup.
- The recipe for concept erasure is to learn a function f that maximizes the following objective:

$$\max_f \sum_i \mathrm{Vol}(F_i), \text{ subject to } \mathrm{Vol}(\mathscr{F}) = \text{const.}$$

- We use the rate-distortion function as a proxy measure for volume.

Kernelized Rate-Distortion Maximization (KRaM)



- We present a kernelized version of the rate distortion function:

$$R(Z|\mathbf{K}) = \frac{1}{2}\log_2 \det\left(I + \frac{d}{n} ZZ^T \odot \mathbf{K}\right)$$

- $\mathbf{K} \in \mathbb{R}^{n \times n}$ is a kernel matrix capturing the similarity between concept labels $\mathbf{K}_{ij} = k(a_i, a_j) \propto 1/\mathrm{d}(a_i, a_j)$, where $\mathrm{d}(\cdot,\cdot)$ is the distance function.
- Maximizing $R(Z|\mathbf{K})$ forces representations similar in the concept space to be dissimilar. Concept erasure recipe can be implemented as:

$$\max_f R(Z|\mathbf{K}), \text{ subject to } R(Z) = b$$

- The kernel function $k(\cdot,\cdot)$ does not make any assumptions on the nature of the concept (categorical, continuous, and vector-valued).

Measuring Alignment

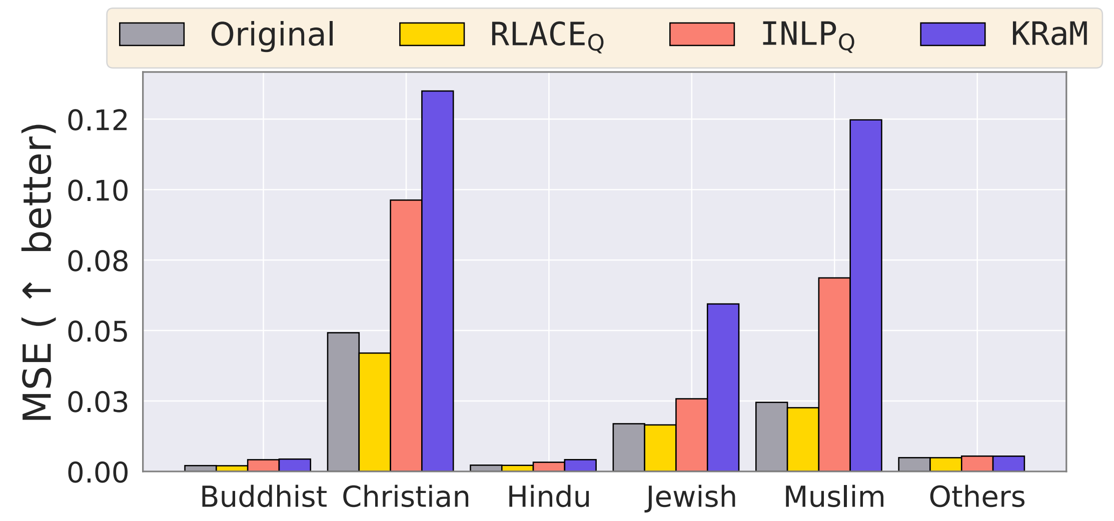
- It is important to quantify the amount of information from the original representations lost due to concept erasure.
- We present a heuristic-based measure to quantify the alignment between original and learned representations space:

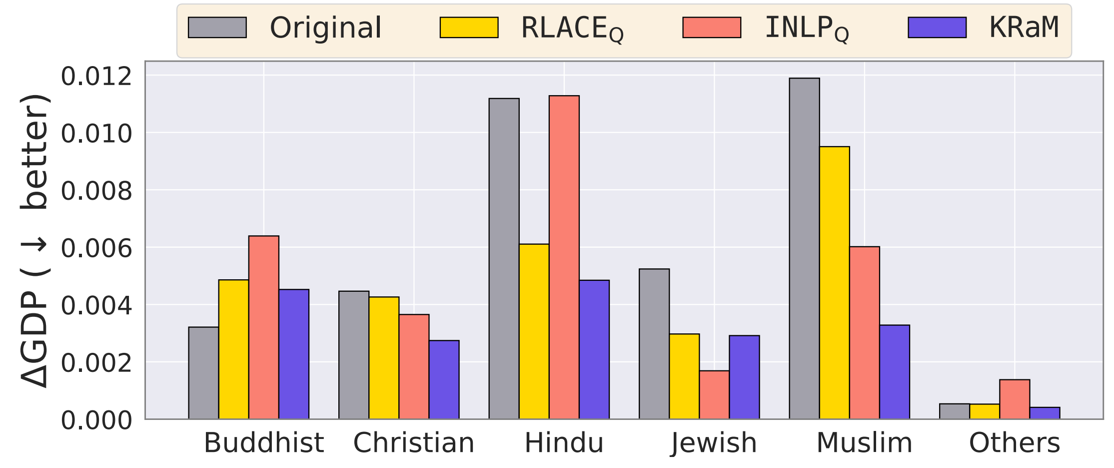
$$A_k(f) = \frac{1}{k}\mathbb{E}_x\left[\mathrm{knn}(x) \cap \mathrm{knn}(f(x))\right]$$

- Theoretical result: $A_k(f) \in \left[\frac{k}{n}, 1\right]$. Find more details in Section 4 of the paper.

Evaluation

- We evaluate KRaM on 3 sets of datasets for categorical, continuous, and vector-valued concept erasure.
- The results on Jigsaw toxicity classification with vector-valued religion (concept to be erased) labels are reported.
- We observe a significant drop in predicting the religion with little impact on toxicity accuracy: $93.2\,\% \rightarrow 92.1\,\%$.





Conclusion

- We propose KRaM, a robust method for performing concept erasure using a kernelized version of the rate distortion function.
- We introduce a heuristic-based metric to compute the information retained after concept erasure
- Empirical results showcase the efficacy of KRaM on a wide range of datasets.
- Code is available here: brcsomnath/KRaM