UNIVERSITY OF TORONTO

# Learning Unbiased Features

Yujia Li, Kevin Swersky, Richard Zemel

CIFAR CANADIAN INSTITUTE FOR ADVANCED RESEARCH

## Overview

Suppose we have access to samples from two probability distributions $X \sim P_A$ and $Y \sim P_B$, how can we tell if $P_A = P_B$?

Maximum Mean Discrepancy (MMD) is a measure of distance between two distributions given only samples from each.

$$
\begin{aligned}
&\left\| \frac{1}{N}\sum_{n=1}^{N}\phi(X_n) - \frac{1}{M}\sum_{m=1}^{M}\phi(Y_m) \right\|^2 \\
=&\frac{1}{N^2}\sum_{n=1}^{N}\sum_{n'=1}^{N}\phi(X_n)^\top\phi(X_{n'}) + \frac{1}{M^2}\sum_{m=1}^{M}\sum_{m'=1}^{M}\phi(Y_m)^\top\phi(Y_{m'}) - \frac{2}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\phi(X_n)^\top\phi(Y_m) \\
=&\frac{1}{N^2}\sum_{n=1}^{N}\sum_{n'=1}^{N}k(X_n, X_{n'}) + \frac{1}{M^2}\sum_{m=1}^{M}\sum_{m'=1}^{M}k(Y_m, Y_{m'}) - \frac{2}{MN}\sum_{n=1}^{N}\sum_{m=1}^{M}k(X_n, Y_m)
\end{aligned}
$$

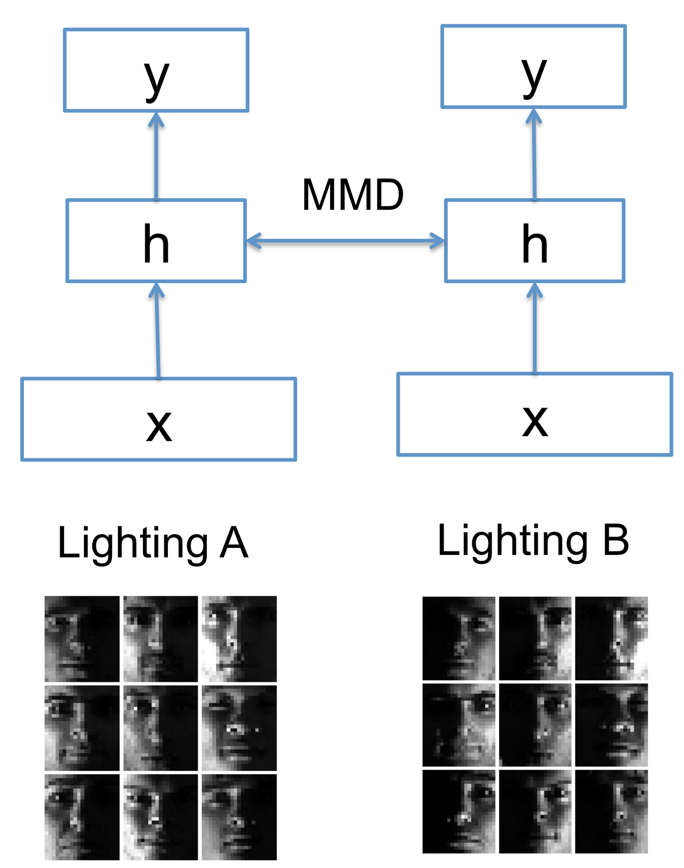
We apply the MMD with neural network models to several different problems.

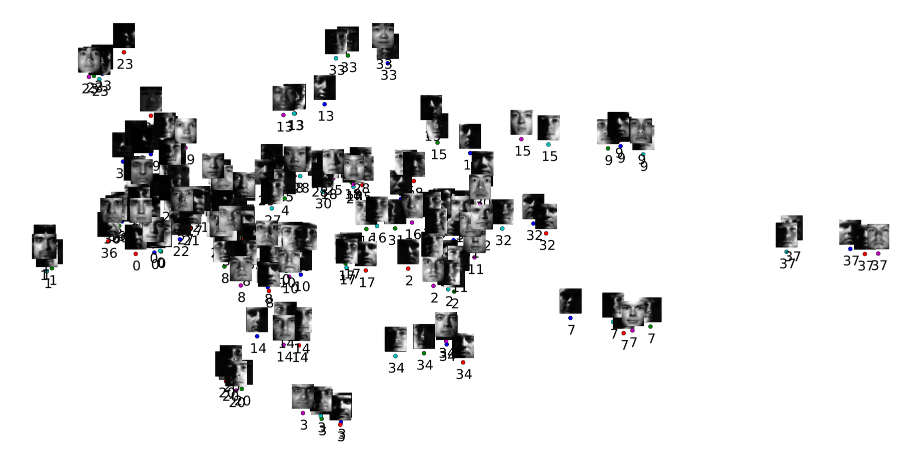
In all applications, we minimize MMD as a regularizer / objective so that we cannot tell two distributions apart.

## Learning Invariant Features

MMD as a regularizer to learn representations that are invariant to certain task irrelevant biases in the data.

Multi-Distribution MMD $\displaystyle\sum_{s=1}^{S}\left\| \frac{1}{N_s}\sum_{i:d_i=s}\phi(h_i) - \frac{1}{N}\sum_{n}\phi(h_n) \right\|^2$



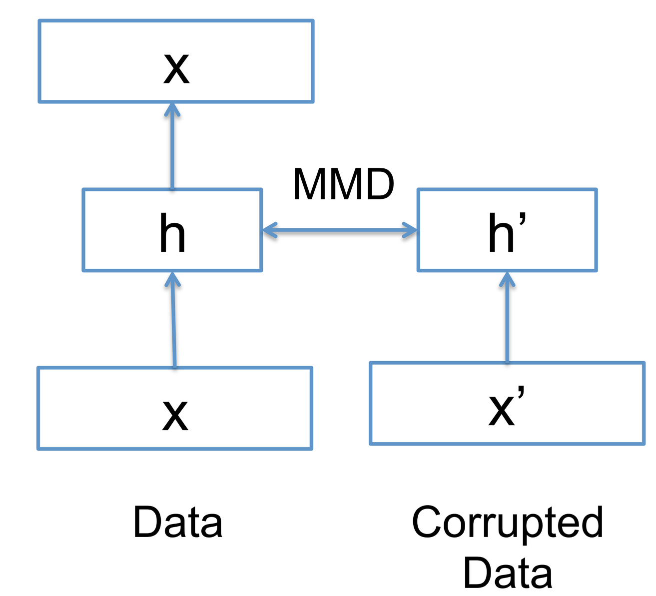



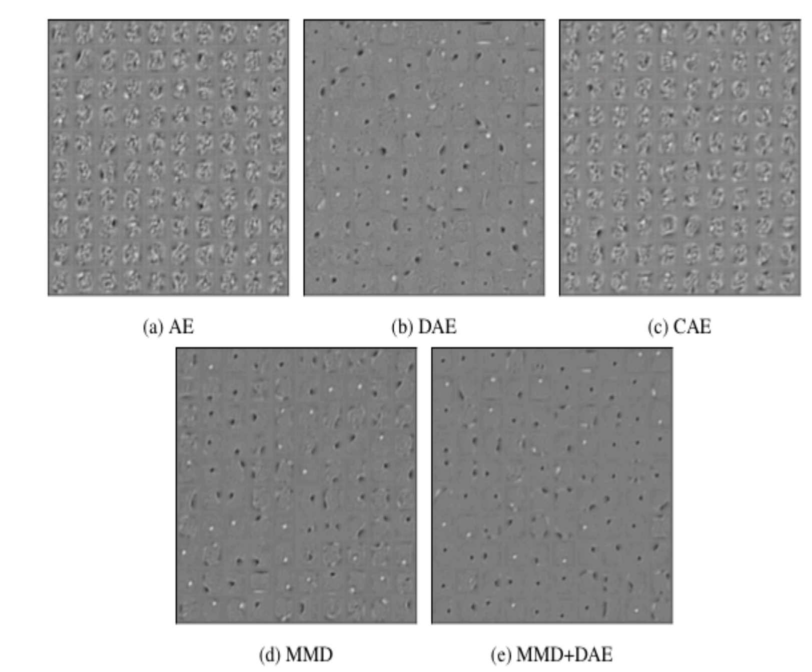
With MMD: 82% Test Accuracy

Without MMD: 72% Test Accuracy

## Noise-Insensitive Auto-Encoders

Use an MMD regularizer on the hidden representation of an auto-encoder so that the representation for corrupted data is indistinguishable from uncorrupted data. With infinitesimal Gaussian noise and linear kernel recovers the contractive auto-encoder penalty.



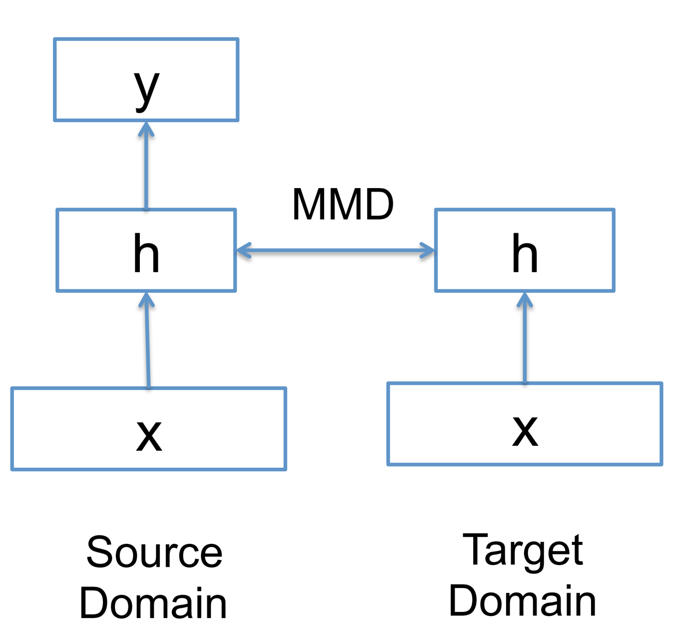

Evaluate by attempting to classify corrupted vs uncorrupted using the learned representation.

| Model | AE | DAE | CAE | MMD | MMD+DAE |
|---|---|---|---|---|---|
| SVM Accuracy | 78.6 | 82.5 | 77.9 | 61.1 | 72.9 |

## Domain Adaptation

MMD as a regularizer for learning domain independent representations.

- Make hidden representations indistinguishable across domains to learn features that generalize beyond specific domains
- Training Loss = Classification Loss + λ MMD




| | D→B | E→B | K→B | B→D | E→D | K→D |
|---|---|---|---|---|---|---|
| Linear SVM | 78.3 ± 1.4 | 71.0 ± 2.0 | 72.9 ± 2.4 | 79.0 ± 1.9 | 72.5 ± 2.9 | 73.6 ± 1.5 |
| RBF SVM | 77.7 ± 1.2 | 68.0 ± 1.9 | 73.2 ± 2.4 | 79.1 ± 2.3 | 70.7 ± 1.8 | 73.0 ± 1.6 |
| TCA | 77.5 ± 1.3 | 71.8 ± 1.4 | 68.8 ± 2.4 | 76.9 ± 1.4 | 72.5 ± 1.9 | 73.3 ± 2.4 |
| NN | 76.6 ± 1.8 | 70.0 ± 2.4 | 72.8 ± 1.5 | 78.3 ± 1.6 | 71.7 ± 2.7 | 72.7 ± 1.6 |
| NN MMD* | 76.5 ± 2.5 | 71.8 ± 2.1 | 72.8 ± 2.4 | 77.4 ± 2.4 | 74.3 ± 1.7 | 73.9 ± 2.4 |
| NN MMD | **78.5 ± 1.5** | **73.7 ± 2.0** | **75.7 ± 2.3** | **79.2 ± 1.7** | **75.3 ± 2.1** | **75.0 ± 1.0** |
| | B→E | D→E | K→E | B→K | D→K | E→K |
| Linear SVM | 72.4 ± 3.0 | 74.2 ± 1.4 | 82.7 ± 1.3 | 75.9 ± 1.8 | 77.0 ± 1.8 | 84.5 ± 1.0 |
| RBF SVM | 72.8 ± 2.5 | 76.3 ± 2.2 | 82.5 ± 1.4 | 75.8 ± 2.1 | 76.0 ± 2.2 | 82.0 ± 1.4 |
| TCA | 72.1 ± 2.6 | 75.9 ± 2.7 | 79.8 ± 1.4 | 76.8 ± 2.1 | 76.4 ± 1.7 | 80.2 ± 1.4 |
| NN | 70.1 ± 3.1 | 72.8 ± 2.4 | 82.3 ± 1.0 | 74.1 ± 1.6 | 75.8 ± 1.8 | 84.0 ± 1.5 |
| NN MMD* | 75.6 ± 2.9 | 78.4 ± 1.6 | 83.0 ± 1.2 | 77.9 ± 1.6 | 78.0 ± 1.9 | 84.7 ± 1.6 |
| NN MMD | **76.8 ± 2.0** | **79.1 ± 1.6** | **83.9 ± 1.0** | **78.3 ± 1.4** | **78.6 ± 2.6** | **85.2 ± 1.1** |

## Learning Deep Generative Models

Use an MMD loss function to make the model distribution close to the data distribution. No adversary required! Trained entirely with backpropagation.