Mitigating Information Leakage in Image Representations: A Maximum Entropy Approach

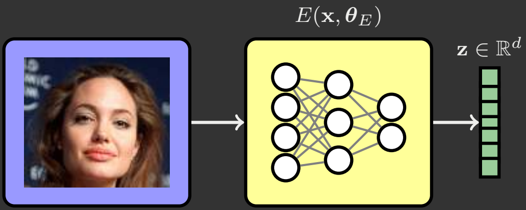
Proteek Roy and Vishnu Boddeti

Michigan State University

CVPR 2019

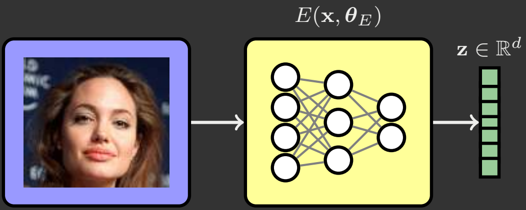
>>> Representation Learning: The Bright Side

* Deep Embeddings:



>>> Representation Learning: The Bright Side

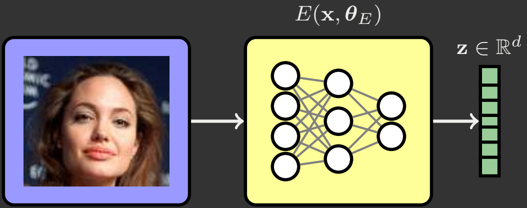
* Deep Embeddings:



* Features contain a lot of information
  * basis for generalizing and transferring to other tasks

>>> Representation Learning: The Bright Side

* Deep Embeddings:



* Features contain a lot of information
    * basis for generalizing and transferring to other tasks

* Applications include:

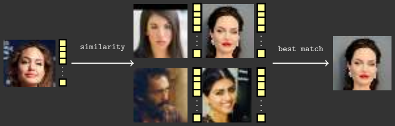


Figure: Face Recognition

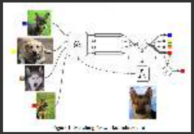


Figure: Image Retrieval

>>> Representation Learning: The Dark Side

>>> Representation Learning: The Dark Side

* Features contain a lot of information

>>> Representation Learning: The Dark Side

* Features contain a lot of information

* Information may inadvertently be sensitive

>>> Representation Learning: The Dark Side

* Features contain a lot of information

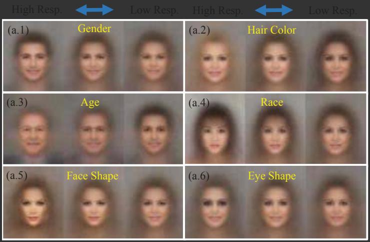
* Information may inadvertently be sensitive
  * compromise privacy of data owner
  * result in unfair or biased decision systems

>>> Representation Learning: The Dark Side

* Features contain a lot of information

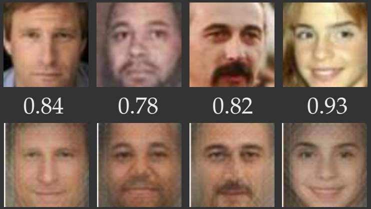
* Information may inadvertently be sensitive
  * compromise privacy of data owner
  * result in unfair or biased decision systems

* Soft attribute from face features



Liu et al., ICCV 2015

* Reconstruction from face features



Mai et al., PAMI 2018

>>> Central Aim of This Paper

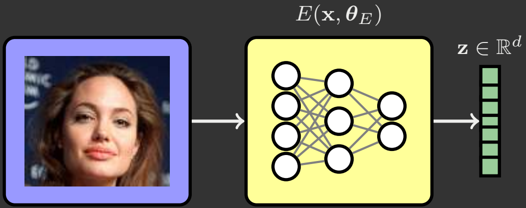
Mitigating Information Leakage

Develop representation learning algorithms that can *intentionally* and *permanently* obscure sensitive information while retaining task dependent information.

>>> Problem Setting: Adversarial Representation Learning

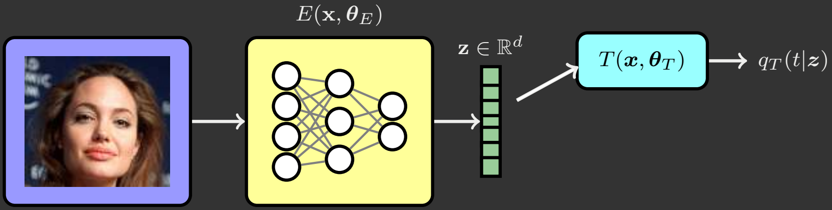
* Three player zero-sum game between:

>>> Problem Setting: Adversarial Representation Learning



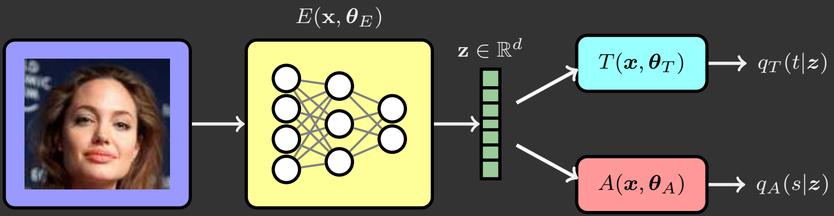
* Three player zero-sum game between:
  * Encoder extracts features z

>>> Problem Setting: Adversarial Representation Learning



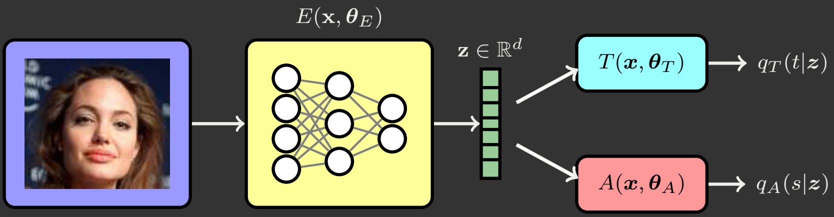
* Three player zero-sum game between:
  * Encoder extracts features z
  * Target Predictor for desired task from features z

>>> Problem Setting: Adversarial Representation Learning



* Three player zero-sum game between:
  * Encoder extracts features z
  * Target Predictor for desired task from features z
  * Adversary extracts sensitive information from features z

>>> Problem Setting: Adversarial Representation Learning



* Three player zero-sum game between:
  * Encoder extracts features z
  * Target Predictor for desired task from features z
  * Adversary extracts sensitive information from features z

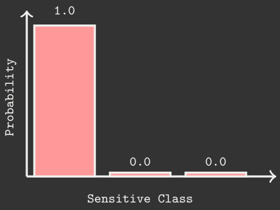
* Minimum Likelihood Adversarial Representation Learning:

$$\min_{\boldsymbol{\theta}_E, \boldsymbol{\theta}_T} \max_{\boldsymbol{\theta}_A} \underbrace{J_1(\boldsymbol{\theta}_E, \boldsymbol{\theta}_T)}_{\text{likelihood of predictor}} - \alpha \underbrace{J_2(\boldsymbol{\theta}_E, \boldsymbol{\theta}_A)}_{\text{likelihood of adversary}} \tag{1}$$

\>>> Optimizing Likelihood Can be Sub-Optimal

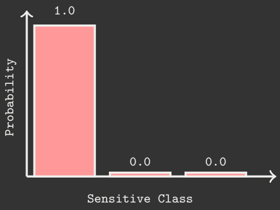
\>>> Optimizing Likelihood Can be Sub-Optimal

* Adversary

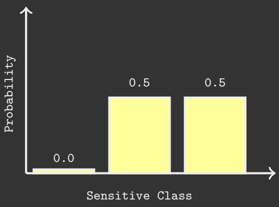


\>>> Optimizing Likelihood Can be Sub-Optimal

* Adversary

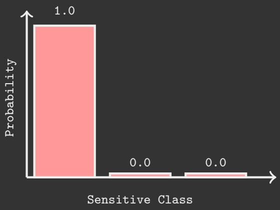


* Encoder

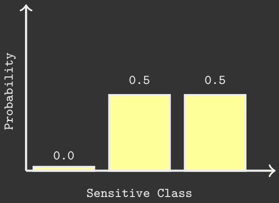


>>> Optimizing Likelihood Can be Sub-Optimal

* Adversary



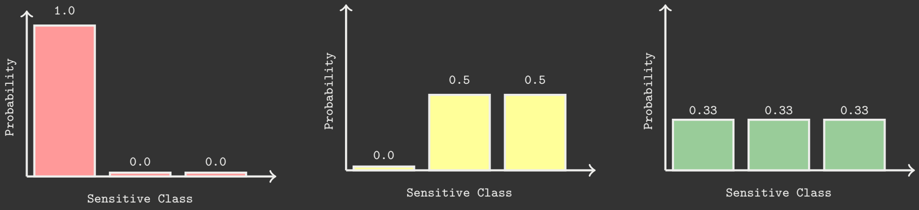
* Encoder



* Equillibrium

>>> Optimizing Likelihood Can be Sub-Optimal

* Adversary
* Encoder
* Equillibrium



Limitations:

* Encoder target distribution leaks information !!
* Practice: simultaneous SGD does not reach equilibrium
* Class Imbalance: likelihood biases solution to majority class

>>> Maximum Entropy Adversarial Representation Learning

Key Idea

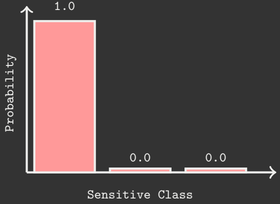
Optimize the encoder to maximize entropy of adversary as opposed to minimizing its likelihood.

>>> Maximum Entropy Adversarial Representation Learning

Key Idea

Optimize the encoder to maximize entropy of adversary as opposed to minimizing its likelihood.

* Adversary

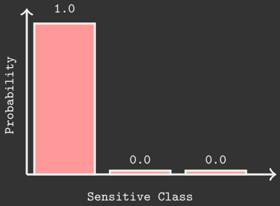


>>> Maximum Entropy Adversarial Representation Learning

Key Idea

Optimize the encoder to maximize entropy of adversary as opposed to minimizing its likelihood.

* Adversary



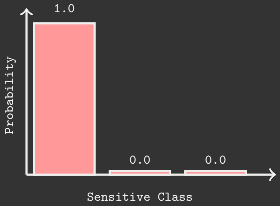
* Encoder

>>> Maximum Entropy Adversarial Representation Learning

Key Idea

Optimize the encoder to maximize entropy of adversary as opposed to minimizing its likelihood.

* Adversary



* Encoder

* Equilibrium

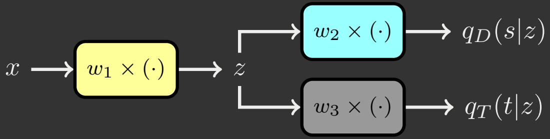
>>> MaxEnt-ARL Properties

* Theoretical
  * Three player non-zero sum game
  * At equilibrium, encoder induces uniform distribution in adversary when $s \perp\!\!\!\perp t$
  * Obtain conditions for stability of solution around equillibrium through linearization.

>>> MaxEnt-ARL Properties

* Theoretical
  * Three player non-zero sum game
  * At equilibrium, encoder induces uniform distribution in adversary when $s \perp\!\!\!\perp t$
  * Obtain conditions for stability of solution around equillibrium through linearization.
* Practical
  * Semi-Supervised Mode: encoder does not need sensitive labels
  * Less susceptible to class imbalance than ML-ARL

>>> Three Player Game: Linear Case



* Each entity is linear scalar multiplication
* Global solution is $(w_1, w_2, w_3) = (0, 0, 0)$

Minimum Likelihood

Maximum Entropy

>>> Numerical Experiments: Fair Classification

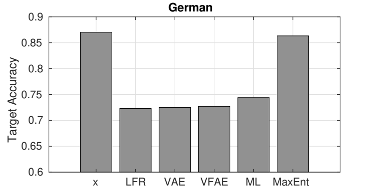
* UCI Datatset: Creditworthiness Prediction

* UCI Datatset: Income Prediction

>>> Numerical Experiments: Fair Classification

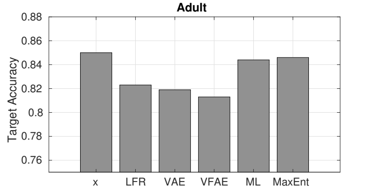
* UCI Datatset: Creditworthiness Prediction

Target: Credit Prediction



* UCI Datatset: Income Prediction

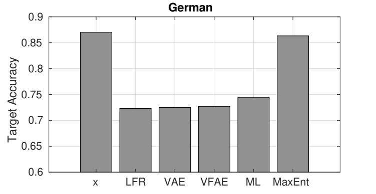
Target: Income Prediction



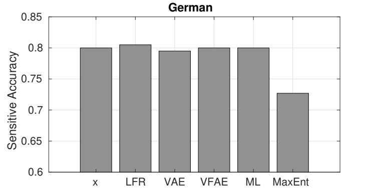
>>> Numerical Experiments: Fair Classification

* UCI Datatset: Creditworthiness Prediction

Target: Credit Prediction

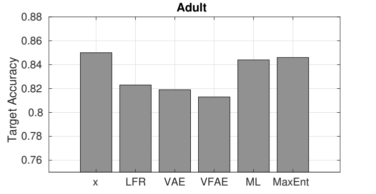


Adversary: Gender Prediction

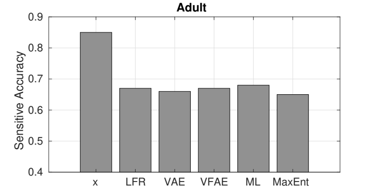


* UCI Datatset: Income Prediction

Target: Income Prediction



Adversary: Gender Prediction



>>> Numerical Experiments: Extended Yale B Faces

* 38 identities and 5 illumination directions
* Target: Identity Label
* Sensitive: Illumination Label

>>> Numerical Experiments: Extended Yale B Faces

* 38 identities and 5 illumination directions
* Target: Identity Label
* Sensitive: Illumination Label

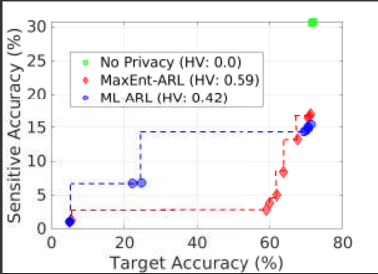
Method	s (lighting)	t (identity)
LR	96	78
NN + MMD (NIPS 2014)	-	82
VFAE (ICLR 2016)	57	85
ML-ARL (NIPS 2017)	57	89
Maxent-ARL	40	89

>>> Numerical Experiments: CIFAR-100

* 100 classes categorized into 20 superclasses
* Target: Superclass Label
* Sensitive: Class Label

>>> Numerical Experiments: CIFAR-100

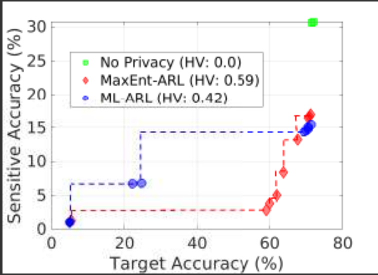
* 100 classes categorized into 20 superclasses
* Target: Superclass Label
* Sensitive: Class Label



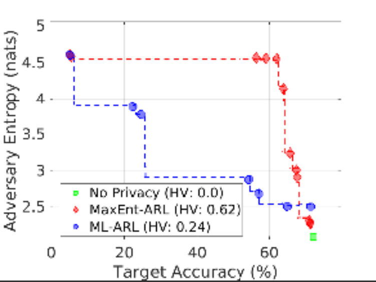
Trade-Off: Likelihood

\>>> Numerical Experiments: CIFAR-100

* 100 classes categorized into 20 superclasses
* Target: Superclass Label
* Sensitive: Class Label



Trade-Off: Likelihood



Trade-Off: Entropy

>>> Summary

* A striving step towards explicitly controlling information in learned representations.

>>> Summary

* A striving step towards explicitly controlling information in learned representations.

* MaxEnt-ARL: optimize the encoder to maximize entropy of adversary instead of minimizing likelihood.

>>> Summary

* A striving step towards explicitly controlling information in learned representations.

* MaxEnt-ARL: optimize the encoder to maximize entropy of adversary instead of minimizing likelihood.

* MaxEnt-ARL enjoys theoretical and practical benefits.

>>> Summary

* A striving step towards explicitly controlling information in learned representations.

* MaxEnt-ARL: optimize the encoder to maximize entropy of adversary instead of minimizing likelihood.

* MaxEnt-ARL enjoys theoretical and practical benefits.

Code:

https://github.com/human-analysis/MaxEnt-ARL.git

More Details: Poster # 175