ACL

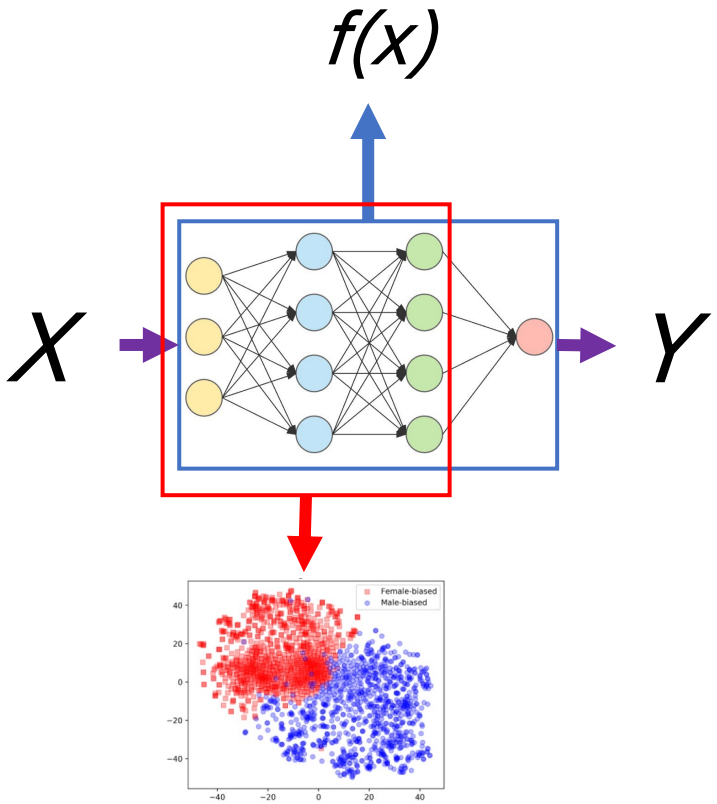
Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton & Yoav Goldberg

Computer Science Department, Bar Ilan University; Allen Institute for Artificial Intelligence

Published online: 2020

对于人工智能模型，认为模型将输入数据映射到特定的**向量空间**中，以“**表征向量**”的形式描述了输入的**特征**（类似表示学习）。这一向量与输出存在某种关系。



- **Probing**：探究表征向量中含有哪些**信息**，以及这些信息在向量空间中的**特点**。
- **特征弱化**：有时我们**不需要**表征向量去考虑一些特征以使得表征更加泛化（如编码单词词义时不需要考虑时态）；有时一些特征是对模型理解能力**有害**（如真实文本中存在的种族与性别偏见）。

去除表征向量中的某些特征信息的表达。

将向量**映射**到指定的方向上，降低向量中特征的空间信息。

在目标中加入**对抗性目标**，通过最小化信息的可预测性，**迭代**得到向量。

工作提出了一种**可迭代的，基于零空间映射的方法**。
Iterative Nullspace Projection

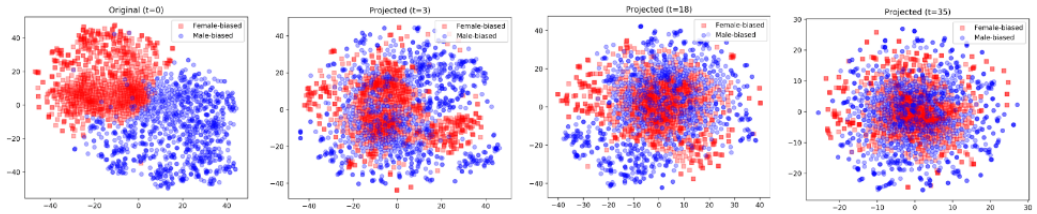
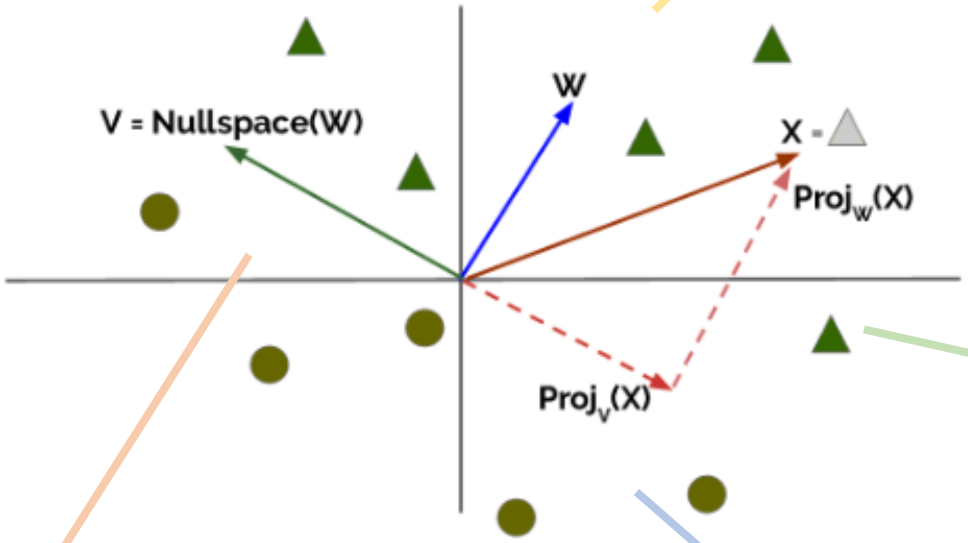


Figure 1: t-SNE projection of GloVe vectors of the most gender-biased words after t=0, 3, 18, and 35 iterations of INLP. Words are colored according to being male-biased or female-biased.

注意此为线性变换。

以二维空间，单一分类器为例：

可将线性分类器看做一个**超平面**。W行向量张成。



线性分类器对向量进行分类实际上是将向量**投影**到超平面上，观察投影向量。

零空间也是一个**超平面**，且与W的行空间正交。

将向量投影到零空间上，得到的新向量在最大限度保持其他维度信息的基础上，**去除**了W方向上的信息。

算法流程：

Algorithm 1 Iterative Nullspace Projection (INLP)

Input : (X, Z): a training set of vectors and protected attributes
n: Number of rounds
Result: A projection matrix P
Function `GetProjectionMatrix`(X, Z)**:**
  $X_{projected} \leftarrow X$
  $P \leftarrow I$
  for $i \leftarrow 1$ **to** n **do**
    $W_i \leftarrow \text{TrainClassifier}(X_{projected}, Z)$
    $B_i \leftarrow \text{GetNullSpaceBasis}(W_i)$
    $P_{N(W_i)} \leftarrow B_i Bi^T$
    $P \leftarrow P_{N(W_i)} P$
    $X_{projected} \leftarrow P_{N(W_i)} X_{projected}$
  end
  return P

- **P**：映射，从原空间到无特征空间。
- **W**：训练的线性分类器。
- **B**：零空间的基。
- **PN**：映射，映射到零空间。

随着维度的增加，一种特征的空间信息可能被**多个超平面**捕捉到。因此多次循环。

实现细节：

Algorithm 1 Iterative Nullspace Projection (INLP)

Input : (X, Z): a training set of vectors and protected attributes
n: Number of rounds
Result: A projection matrix P
Function `GetProjectionMatrix`(X, Z)**:**
  $X_{projected} \leftarrow X$
  $P \leftarrow I$
  for $i \leftarrow 1$ **to** n **do**
    $W_i \leftarrow$ TrainClassifier$(X_{projected}, Z)$
    $B_i \leftarrow$ GetNullSpaceBasis(W_i)
    $P_{N(W_i)} \leftarrow B_i Bi^T$
    $P \leftarrow P_{N(W_i)} P$
    $X_{projected} \leftarrow P_{N(W_i)} X_{projected}$
  end
  return P

对于多次迭代计算，要考虑**误差累积**，使得小误差随着多次迭代被放大。

$$N(w_1) \cap \cdots \cap N(w_n) = N(P_R(w_1) + \cdots + P_R(w_n))$$

之前是每次迭代后，做一次复合运算。现在是迭代得到一个变量，与之前迭代所得变量一起进行运算。

实验组一：

此方法是有效的吗？

- 选取CloVe word embedding中的male-biased、 female-biased以及neutral的向量各7500个。
- 使用SVM分类器进行性别分类。
- 进行INLP前后的准确率由**100%**下降到**49.3%**。

此方法生效原因是否如前文所述（方法的新颖性）？

- 与以往投影方法相比，不同处：模型自行选择投影方向，可去除更多方向。
- 是否仅仅因为去除了更多方向？
- 固定去除方向数（迭代次数），用人工选择的方向进行对比实验。
- 人工准确率：**80.7%**；INLP准确率：**54.4%**。

此方法对原空间产生了什么其他影响？

- 随机选取一些单词，查看其在embdding空间中的最近邻和固有语义子集，基本无区别。说明其对其他方向上的空间信息影响较小。
- 在单词概念语义测量数据集上表现有所上升（比原空间更接近人类反应）。

可视化结果（聚类）？

- 使用t-SNE图进行降维可视化。
- 可视化不同迭代次数下的结果。
- 使用聚类，计算不同簇之间的重叠度，以量化空间区分度。
- 可以看出经过INLP之后，空间区分度明显变小。使用V-measure为指标，由**83.88%**下降到**0.44%**。

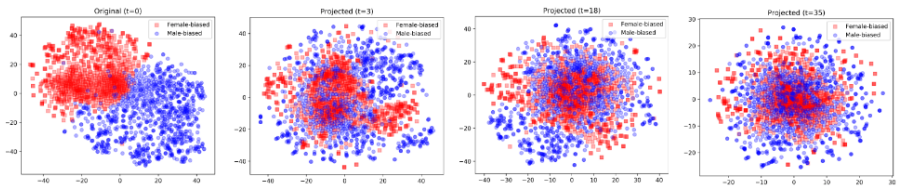


Figure 1: t-SNE projection of GloVe vectors of the most gender-biased words after t=0, 3, 18, and 35 iterations of INLP. Words are colored according to being male-biased or female-biased.

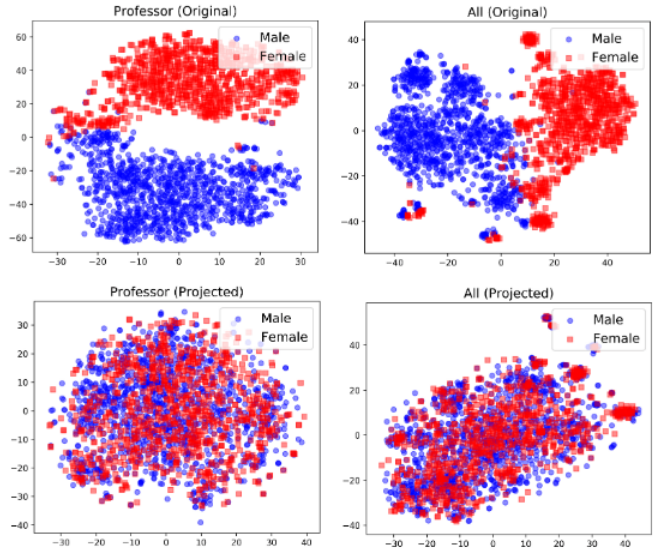


Figure 3: t-SNE projection of BERT representations for the profession "professor" (left) and for a random sample of all professions (right), before and after the projection.

实验组二：

凸显本工作的应用价值

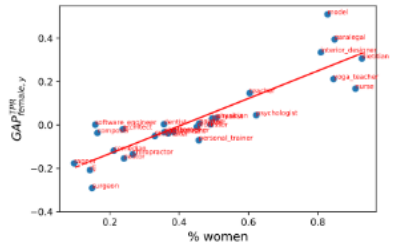
- 描述一个应用场景（这个场景甚至不是本工作的初衷），使用符号化语言数学描述这一场景。
- 根据此场景以往的工作，描述实验步骤与测量指标。
- 将以往工作的组件换成本工作提出的方法。进行实验的复现。
- 结果的比较和可视化。

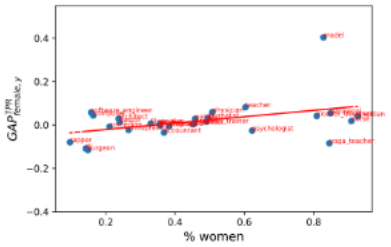
5 Application to Fair Classification

We propose the following procedure. Given a training set X,Y and protected attribute Z, we first train a neural network $f = W \cdot enc(X)$ to best predict Y. This results in an encoder that extracts effective features from X for predicting Y.

$$TPR_{z,y} = P[\hat{Y} = y | Z = z, Y = y]$$
$$GAP^{TPR}_{z,y} = TPR_{z,y} - TPR_{z',y}$$

		BoW	FastText	BERT
Accuracy (profession)	Original	78.2	78.1	80.9
	+INLP	80.1	73.0	75.2
$GAP^{TPR,RMS}_{male}$	Original	0.203	0.184	0.184
	+INLP	0.124	0.089	0.095





局限性分析:

方法采用**数据主导**的迭代训练方式。

↓

受到数据的局限，需要保证数据的**代表性**。

方法采用**线性空间投影**的方法进行特征去除。

↓

受到先验假设的局限，在应用到具体场景中需要进行**具体分析**。并且只能对**线性关系**进行保证。

Thanks!