# **GRAD: GRADIENT REVERSAL AGAINST DISCRIMINATION** /// // // // // // // // // // // A FAIR NEURAL NETWORK LEARNING APPROACH //// // // // // // // // // // // //

### Abstract

No methods currently exist for inducing fairness in arbitrary neural network architectures. In this work we introduce GRAD, a new and simplified method for producing fair neural networks that can be used for autoencoding fair representations or directly with predictive networks. It is easy to implement and add to existing architectures, has only one (insensitive) hyper-parameter, and provides improved individual and group fairness. We use the flexibility of GRAD to demonstrate multi-attribute protection.

### Method

- Neural network architecture with two sets of outputs
  - Inspired by domain adaptation (Ganin et al., '16)
- Several feature extraction layers (the "trunk"), followed by split into two "branches":
  - *"Target branch"* learns to predict target y
  - "Attribute branch" learns to predict protected attribute  $a_n$
- Target branch can be either an autoencoder (GRAD-Auto) or a classifier/regressor (GRAD-Pred)
  - Completely architecture agnostic
  - We used 2 fully-connected layers in each branch • No architecture search performed
- Loss of network is sum of loss of each branch
  - Weighted by balancing parameter  $\lambda$

$$\ell(\gamma, a_p) = \ell_t(\gamma) - \lambda \cdot l_p(a_p)$$

- Losses for both  $\ell_t(\gamma)$  and  $\ell_p(a_p)$  are calculated and gradients are used for weight updates as normal...
- *Except:* once propagated down the attribute branch, gradients are reversed (i.e. multiplied by -1) before being applied to the trunk.
- Effect:
  - Network can still accurately predict target
  - Network moves *away* from optima in predictions of  $a_n$ 
    - Enforces ignorance of protected attribute

## Contributions

- *Goal*: build a network which is fair with respect to some protected attribute  $a_p$ .
- Solution: Gradient Reversal Against Discrimination • Simple to implement

  - Requires only one (insensitive) hyper-parameter
  - Applicable to any architecture, including:
    - Autoencoder architectures
    - Direct predictive architectures
  - Competitive with other approaches
  - The first neural network shown to protect
    - multiple attributes concurrently

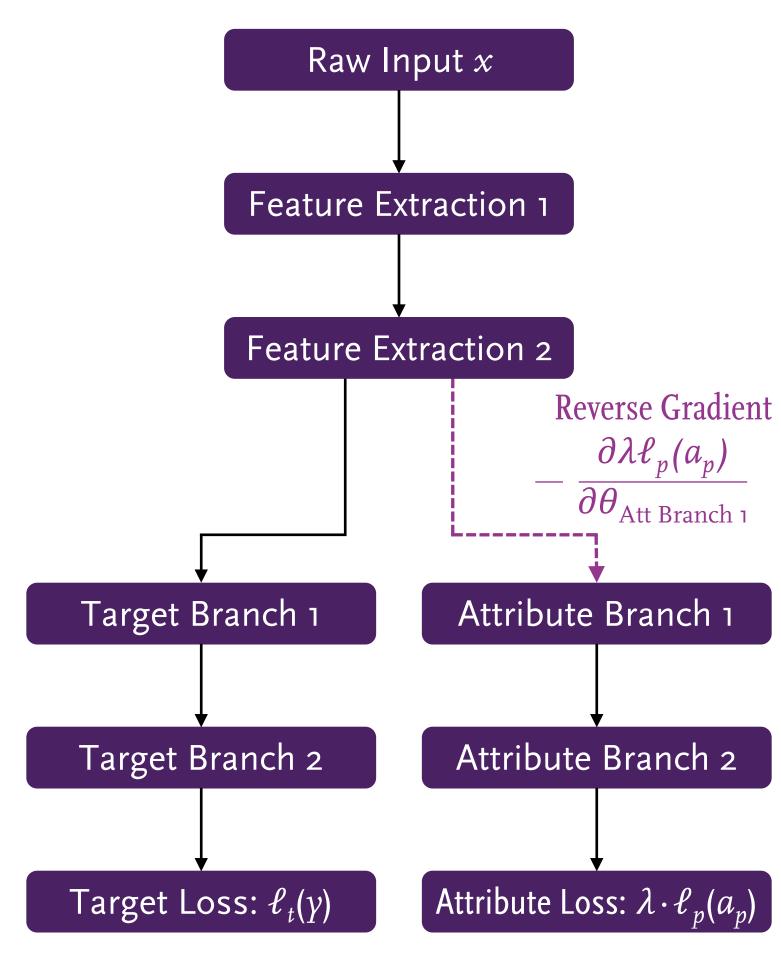


Diagram of the GRAD architecture. The dotted, purple connection indicates normal forward propagation but backpropagation with reversed signs. The value x is the input to the network, and the two terminal nodes are the losses that get backpropagated.

Edward Raff & Jared Sylvester

Booz Allen Hamilton, Strategic Innovation Group ICML Workshop on Fairness, Interpretability & Explainability, 13–15 July 2018, Stockholm

• i.e. output  $\hat{y}$  given input x is invariant to  $a_p$ 

<ul> <li>Following 3</li> </ul>	7emel et a	al 2012				Algorithm	Acc	Delta	Discr	Cons
<ul> <li>Following Zemel et al., 2013</li> <li>Evaluated on German, Adult, Health &amp; Diabetes</li> </ul>					es					
<ul> <li>Discrimination (a.k.a. "group fairness")</li> </ul>						NN-Auto — GRAD-Auto	<b>0.8506</b> 0.8491	0.7939 <b>0.8491</b>	0.0567 <b>0.0000</b>	0.9730 <b>1.0000</b>
<ul> <li>Difference between average predicted scores for each protected attribute value</li> </ul>					roc	NN-Pred	0.8491 0.8440	0.7511	0.0929	0.9453
					105	GRAD-Pred	0.8493	0.8486	0.0020	0.9999
• Consistency (a.k.a. "individual fairness")						NBF	0.6878	0.5678	0.1200	0.5893
<ul> <li>Similar inputs should receive similar outputs</li> </ul>					outs	FF	0.8474	0.8474	0.0000	1.0000
<ul> <li>Accuracy</li> </ul>						$\operatorname{LR}$	0.7547	0.6482	0.1064	0.7233
<ul> <li>Delta = Accuracy – Discrimination</li> </ul>						$\operatorname{LRF}$	0.7212	0.7038	0.0174	0.6223
						m LFR	0.7365	0.7365	0.0000	1.0000
Results						VFAE	0.8490	0.8490	0.0000	
<ul> <li>Competitive with prior methods in all metrics</li> <li>Both GRAD-Auto &amp; GRAD-Pred reliably produce very high Consistency scores <ul> <li>One of the two is always the best in Consistency</li> </ul> </li> <li>Capable of achieving Discrimination=0.00 &amp; Consistency=1.00 Multiple Sensitive Attributes</li> <li>Sensitive attributes such as race &amp; gender often co- occur in the same data set <ul> <li>No prior work has rigorously examined protecting multiple attributes concurrently</li> </ul> </li> <li>GRAD can protect multiple attributes at once</li> </ul>						NN = standard neu		•	-	
<ul> <li>very high C</li> <li>One o</li> <li>Capable of a</li> <li>Multiple Sensitive a occur in th</li> <li>No priprotec</li> <li>GRAD can</li> </ul>	Consistence f the two i chieving Dis sitive Attri ttributes s e same da for work h ting multi protect m	cy scores is always scrimination ibutes such as ra ata set as rigoro iple attributiple a	the best on=0.00 & ace & ge ously exa outes cou ttributes	in Consister Consister ender ofte mined ncurrently at once	istency ncy=1.00 en co-	LR = Logistic Re Learned Fair Repres Robustness to λ • λ controls tra not predictin • Didn't do any • Used λ= • Keeps th • Any values in Adult I	sentations; V de-off bety g a <sub>p</sub> / hyper-par 100 for all ings simp [20,1000]	FAE = Variat ween goals rameter sea experimen le for pract	tional Fair Au of predict arch for $\lambda$ its titioners	ing y and
very high C • One o • Capable of a Multiple Sensitive a occur in th • No pri protec	Consistence f the two i chieving Dis sitive Attri ttributes s e same da for work h ting multi protect m only a sin	cy scores is always scrimination ibutes such as ra ata set as rigoro iple attrib iultiple a gle attrib	the best on=0.00 & ace & ge ously exa outes con ttributes oute caus	in Consister Consister ander ofte mined ncurrent at once ses decre	istency ncy=1.00 en co-	<ul> <li>Learned Fair Representation</li> <li>Robustness to λ</li> <li>λ controls trained for the second s</li></ul>	sentations; V de-off bety g a <sub>p</sub> / hyper-par 100 for all ings simp [20,1000]	FAE = Variat ween goals rameter sea experimen le for pract	arch for $\lambda$ ts tional Fair Au	ing y and
<ul> <li>very high C</li> <li>One o</li> <li>Capable of a</li> <li>Multiple Sensitive a occur in th</li> <li>Sensitive a occur in th</li> <li>No priprotect</li> <li>GRAD can</li> <li>Protecting</li> </ul>	Consistence f the two i chieving Dis sitive Attri ttributes s e same da for work h ting multi protect m only a sin	cy scores is always scrimination ibutes such as ra ata set as rigoro iple attrib iultiple a gle attrib	the best on=0.00 & ace & ge ously exa outes con ttributes oute caus ther attri	in Consister Consister ander ofte mined ncurrent at once ses decre	istency ncy=1.00 en co-	Learned Fair Repres Robustness to λ • λ controls transformediction • Didn't do any • Used λ= • Keeps th • Any values in Adult I	sentations; V de-off bety g a <sub>p</sub> / hyper-par 100 for all ings simp [20,1000]	$FAE = Variat$ ween goals rameter sea experimen le for pract would hav $a_{1}^{-1} = \frac{1}{0.8}$	fional Fair Au of predict arch for $\lambda$ its itioners we been acc Heritage Hea	itoencoders
<ul> <li>very high C</li> <li>One o</li> <li>Capable of a</li> <li>Multiple Sensitive a occur in th</li> <li>Sensitive a occur in th</li> <li>No priprotect</li> <li>GRAD can</li> <li>Protecting</li> </ul>	Consistence f the two i chieving Dis sitive Attri ttributes s e same da for work h ting multi protect m only a sin	cy scores is always scrimination ibutes such as ra ata set as rigoro iple attrib iultiple a gle attrib	the best on=0.00 & ace & ge ously exa outes con ttributes oute caus ther attri	in Consister Consister ander ofte mined ncurrent at once ses decre	istency ncy=1.00 en co-	Learned Fair Representation Robustness to $\lambda$ • $\lambda$ controls train not prediction • Didn't do any • Used $\lambda$ = • Keeps th • Any values in Adult I • $\lambda$	sentations; V de-off bety g a <sub>p</sub> / hyper-par 100 for all ings simp [20,1000]	$FAE = Variat$ ween goals rameter sea experimen le for pract would hav $a = \frac{1}{0.8} = \frac{1}{0.8}$	fional Fair Au of predict arch for $\lambda$ its itioners we been acc Heritage Hea	ing y and
very high C • One o • Capable of ad Multiple Sensitive a occur in th • No pri- protec • GRAD can • Protecting fairness with Algorithms NN-Auto	Consistence f the two is chieving Dis sitive Attra ttributes s e same da for work h ting multi protect m only a sing th respect Acc 0.5735	cy scores is always scrimination ibutes such as ra ata set as rigoro iple attributiple ar gle attribution gle attribution of the of Delta	the best on=0.00 & ace & ge ously exa outes con ttributes oute caus ther attri Discrin Race 0.0412	in Consister Consister ender ofte mined ncurrentl at once ses decre ibute nination Gender 0.0275	istency ncy=1.00 en co- y ased Cons 0.6411	Learned Fair Representation Robustness to $\lambda$ • $\lambda$ controls train not prediction • Didn't do any • Used $\lambda$ = • Keeps th • Any values in Adult I • $\int_{-1}^{1}$	sentations; V de-off bety g a <sub>p</sub> / hyper-par 100 for all ings simp [20,1000]	$FAE = Variat$ ween goals rameter sea experimen le for pract would hav $av = \frac{1}{0.8}$	sional Fair Au of predict arch for $\lambda$ its itioners we been acc Heritage Hea	ing y and
very high C • One o • Capable of a Multiple Sensitive a occur in th • No pri protec • GRAD can • Protecting fairness wi Algorithms	Consistence f the two is chieving Dis sitive Attra ttributes s e same da for work h ting multi protect m only a sing th respect Acc	cy scores is always scrimination ibutes such as ra ata set as rigoro ple attributiple ar gle attributiple ar gle attribution to the o	the best on=0.00 & ace & ge ously exa outes con ttributes oute caus ther attri Discrin Race	in Consister Consister ender ofte mined ncurrent at once ses decre ibute nination Gender	istency ncy=1.00 en co- y ased Cons	Learned Fair Representation Robustness to $\lambda$ • $\lambda$ controls train not prediction • Didn't do any • Used $\lambda$ = • Keeps th • Any values in Adult Is • $Adult Is$	sentations; V de-off bety g a <sub>p</sub> / hyper-par 100 for all ings simp [20,1000]	$FAE = Variat$ ween goals rameter sea experimen le for pract would hav $a = \begin{bmatrix} 1 \\ 0.8 \\ 0.6 \end{bmatrix}$	sional Fair Au of predict arch for $\lambda$ its itioners we been acc Heritage Hea	ing y and
very high C • One o • Capable of ad Multiple Sensitive a occur in th • No pri- protections • GRAD can • Protecting fairness with Algorithms NN-Auto GRAD-Auto	Consistence f the two i chieving Dis sitive Attra ttributes s e same da for work h ting multi protect m only a sin th respect Acc 0.5735 0.5765	cy scores s always scriminatio ibutes such as ra ata set as rigoro iple attrib iultiple a gle attrib to the o Delta 0.5392 0.5723	the best on=0.00 & ace & ge ously exa outes con ttributes oute caus ther attri Discrin Race 0.0412 0.0055	in Consister Consister ender ofte mined ncurrentl at once ses decre ibute nination Gender 0.0275 0.0030	istency ncy=1.00 en co- y ased <u>Cons</u> 0.6411 0.6288	Learned Fair Representation Robustness to $\lambda$ • $\lambda$ controls train not prediction • Didn't do any • Used $\lambda$ = • Keeps th • Any values in Adult Is • $Adult Is$	sentations; V de-off bety g a <sub>p</sub> / hyper-par 100 for all ings simp [20,1000]	$FAE = Variat$ ween goals rameter sea experimen le for pract would hav $a = \begin{bmatrix} 1 \\ 0.8 \\ 0.6 \end{bmatrix}$	sional Fair Au of predict arch for $\lambda$ its itioners we been acc Heritage Hea Accuracy crimination onsistency	ing y and
very high C • One o • Capable of ad Multiple Sensitive a occur in th • No pri- protecting fairness with Algorithms NN-Auto GRAD-Auto NN-Pred GRAD-Pred GRAD-Auto-R	Consistence f the two i chieving Dis sitive Attra ttributes s e same da for work h ting multi protect m only a sin th respect Acc 0.5735 0.5765 0.6286 0.5980 0.5851	cy scores is always scrimination ibutes such as ra ata set as rigoro ple attrib iultiple ar gle attrib to the o Delta 0.5392 0.5723 0.5749 0.5749	the best on=0.00 & ace & ge ously exa outes con ttributes oute caus ther attri Discrim Race 0.0412 0.0055 0.0418 0.0028 0.0003	in Consister Consister ender ofte mined ncurrentl at once ses decre ibute nination Gender 0.0275 0.0030 0.0458 0.0034 0.0201	istency ncy=1.00 en co- y ased <u>Cons</u> 0.6411 0.6288 0.6464 <b>0.7180</b> 0.6404	Learned Fair Representation Robustness to $\lambda$ • $\lambda$ controls train not prediction • Didn't do any • Used $\lambda$ = • Keeps the • Any values in Adult I • $Adult I$	Sentations; V. Sentations; V. ade-off betw g $a_p$ / hyper-particular 100 for all ings simp [20,1000] ncome	FAE = Variat $FAE = Variat$	sional Fair Au of predict arch for $\lambda$ its itioners we been acc Heritage Hea Accuracy crimination onsistency	ing y and eptable lth
very high C • One o • Capable of ad Multiple Sensitive a occur in th • No pring protections • GRAD can • Protecting fairness with Algorithms NN-Auto GRAD-Auto NN-Pred GRAD-Pred	Consistence f the two i chieving Dis sitive Attra ttributes s e same da for work h ting multi protect m only a sin th respect Acc 0.5735 0.5765 0.6286 0.5980	cy scores is always scrimination ibutes such as rata set as rigoro ple attributiple ar gle attributiple ar gle attribution of the or Delta 0.5392 0.5723 0.5848 0.5949	the best on=0.00 & ace & ge ously exa outes con ttributes oute caus ther attri Discrin Race 0.0412 0.0055 0.0418 0.0028	in Consister Consister ender ofte mined ncurrentl at once ses decre ibute nination Gender 0.0275 0.0030 0.0458 0.0034	istency ncy=1.00 en co- y ased <u>Cons</u> 0.6411 0.6288 0.6464 <b>0.7180</b>	Learned Fair Representation Robustness to $\lambda$ • $\lambda$ controls train not prediction • Didn't do any • Used $\lambda$ = • Keeps th • Any values in Adult I • $\frac{1}{0.8}$ • $\frac{1}{0.2}$ • $\frac{1}{0.2}$	Sentations; V. Sentations; V. ade-off betw g $a_p$ / hyper-particular 100 for all ings simp [20,1000] ncome	FAE = Variat $ween goals$ $rameter sea$ $experimen$ $le for pract$ $would hav$ $a = 0.6$ $a =$	conal Fair Au of predict arch for $\lambda$ its itioners we been acc Heritage Hea Accuracy crimination onsistency	ing y and eptable lth 

*networks.* GRAD-Auto and GRAD-Pred protect both race  $\alpha$  gender. \*-*R* models protect only race and \*-*G* models protect only gender.

Hamilton Allen Booz