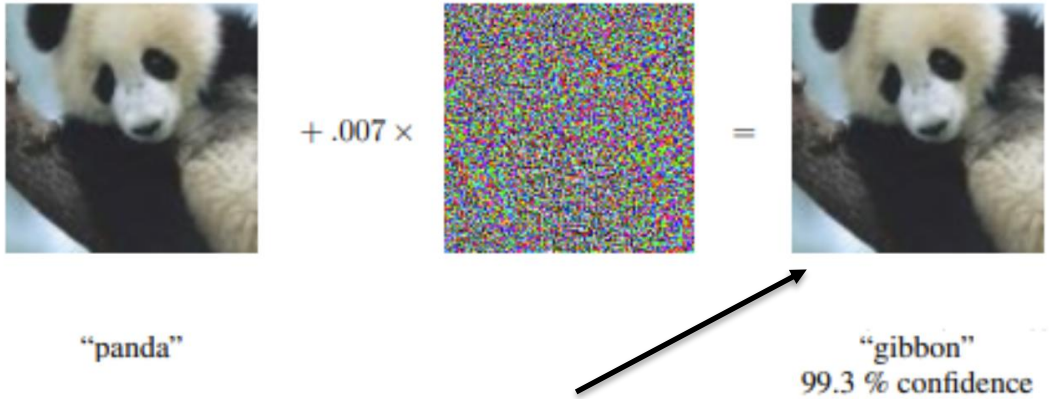
Adversarial Training with a Surrogate

Keane Lucas, Alec Jasen, Lujo Bauer

Adversarial Example Sets



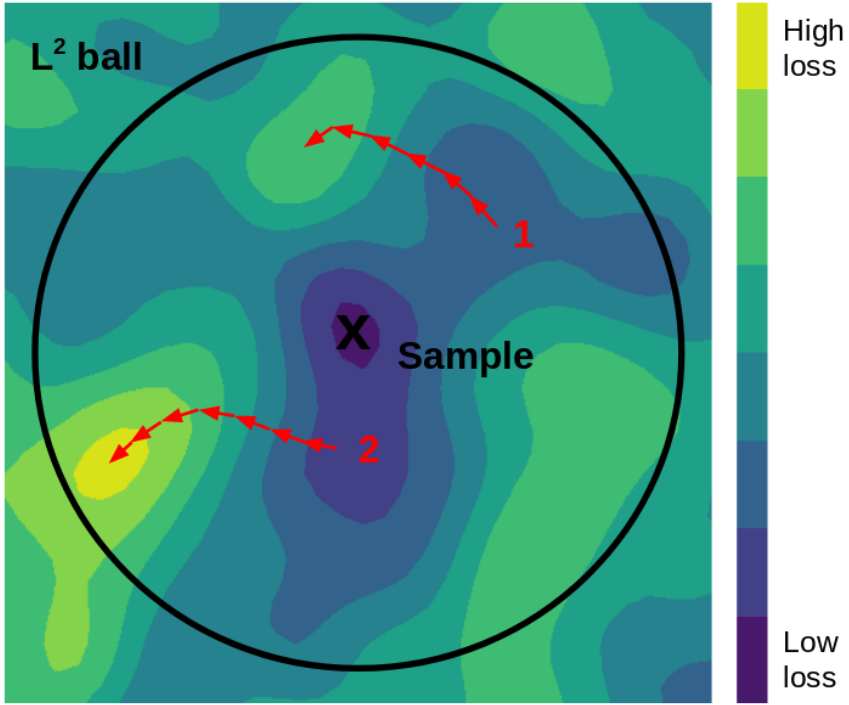
Adversarial Example

(image from Goodfellow 2014)

Perturbation set *P(x)* - set of images formed by small changes to *x* in which all members have the same classification, according to humans

Recent Work

- Use classifier's first-order gradients to directly approach high loss regions
  - Fast Gradient Sign Method (FGSM)
  - Projected Gradient Descent (PGD)



(image from Knagg 2019)

Adversarial Training Introduction

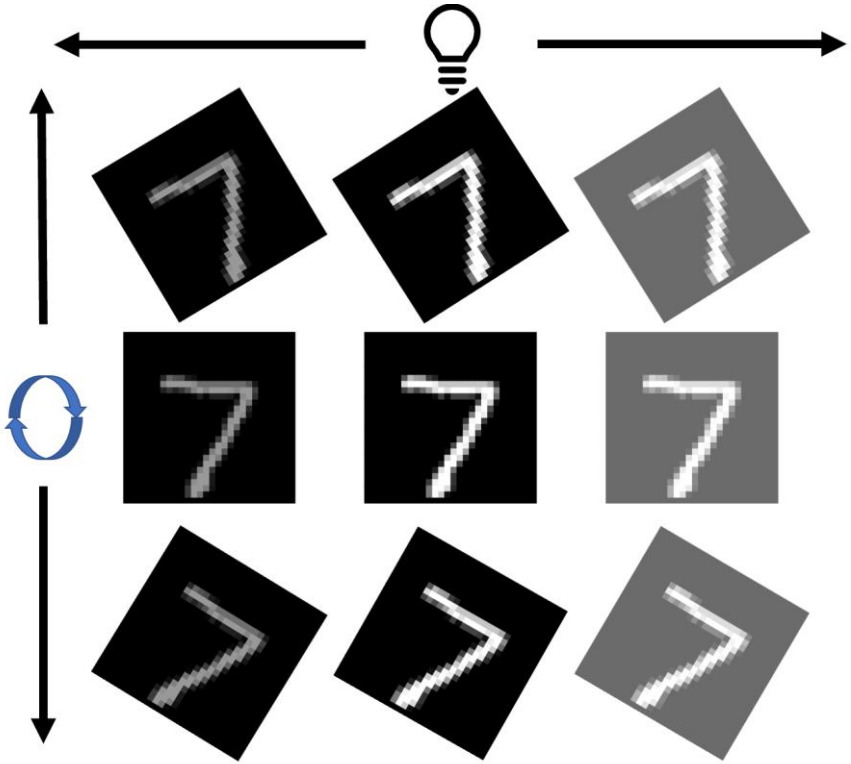
- Finding optimal parameters given a dataset *X, Y*:

$$\min_{\theta} \sum_{\{x,y\}\in\{X,Y\}} L(f(x,\theta),y)$$

- Adversarial training modification:

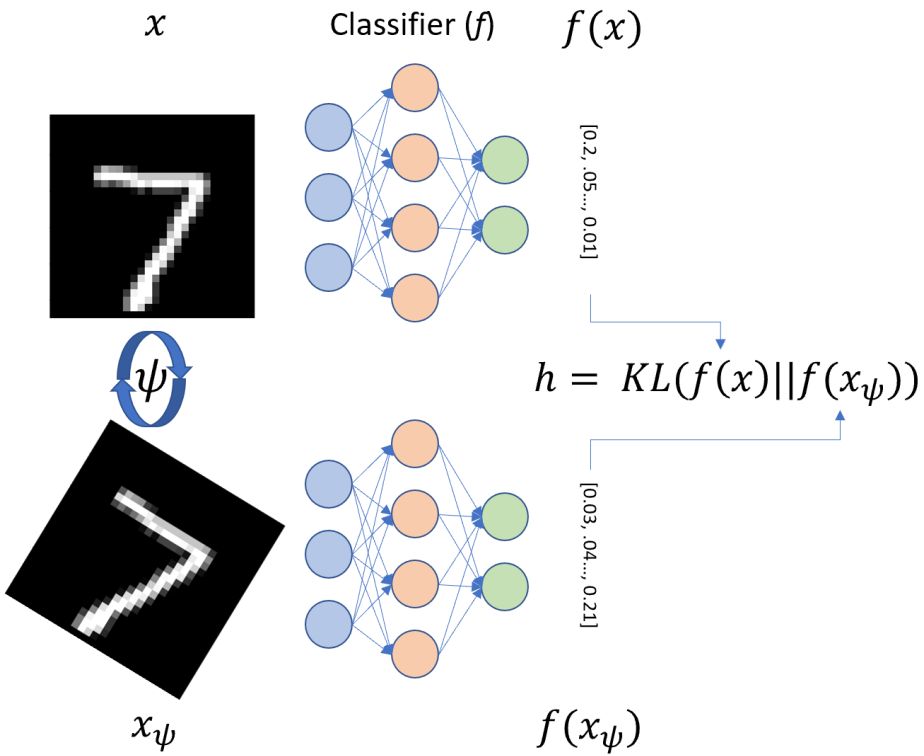
$$\min_{\theta} \sum_{\{x,y\}\in\{X,Y\}} \max_{x'\in P(x)} L(f(x',\theta),y)$$

Motivation



Perturbation set containing images of a `7' with changes in brightness and rotation.

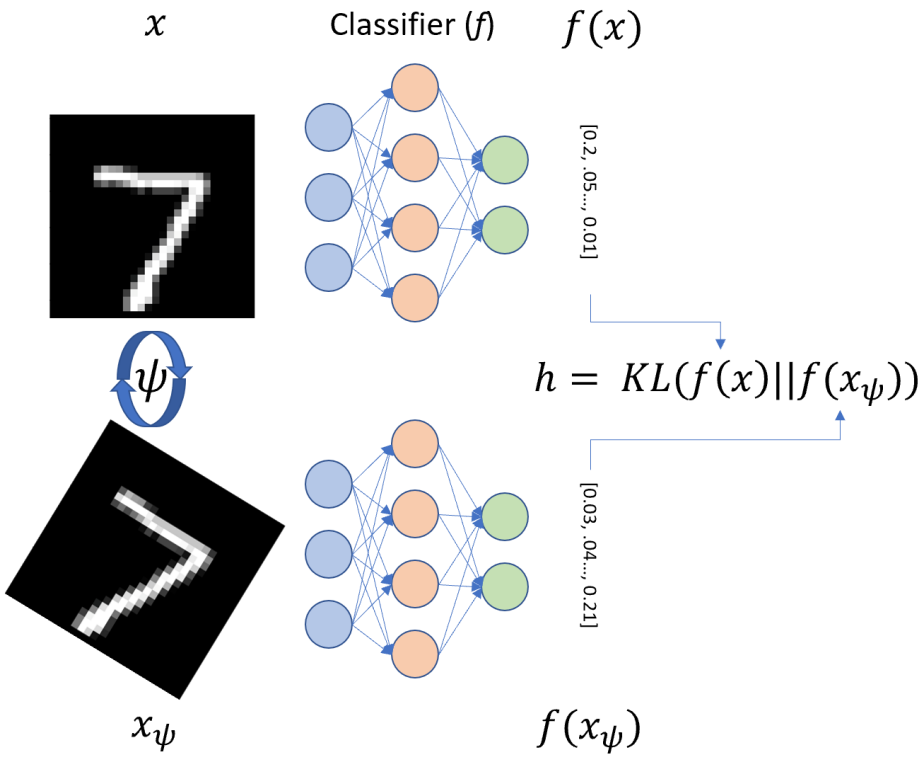
Estimating Harm with a Surrogate



Train a surrogate neural network to estimate **harm** *h* of applying a perturbation ψ to an input *x*

$$s : (X, \Psi) \rightarrow \mathbb{R}$$

Estimating Harm with a Surrogate



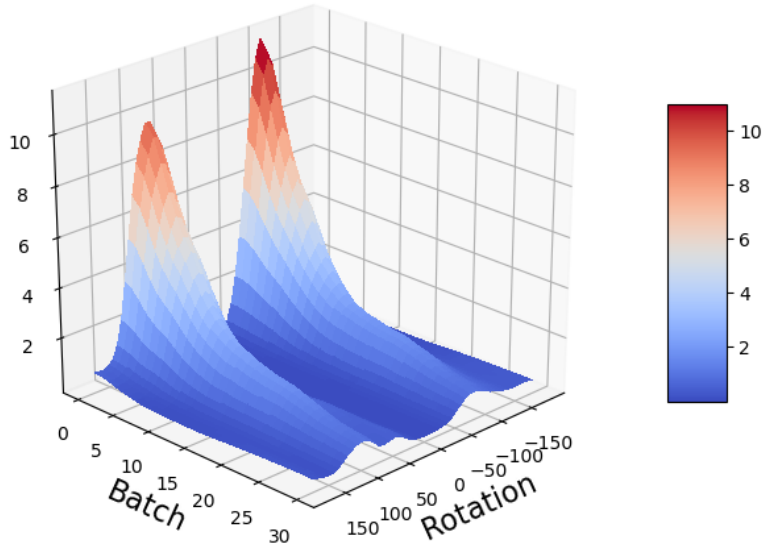
Train a surrogate neural network to estimate **harm** *h* of applying a perturbation ψ to an input *x*

$$s : (X, \Psi) \rightarrow \mathbb{R}$$

Then… use the surrogate first-order gradients to directly approach effective adversarial examples

Surrogate Viability

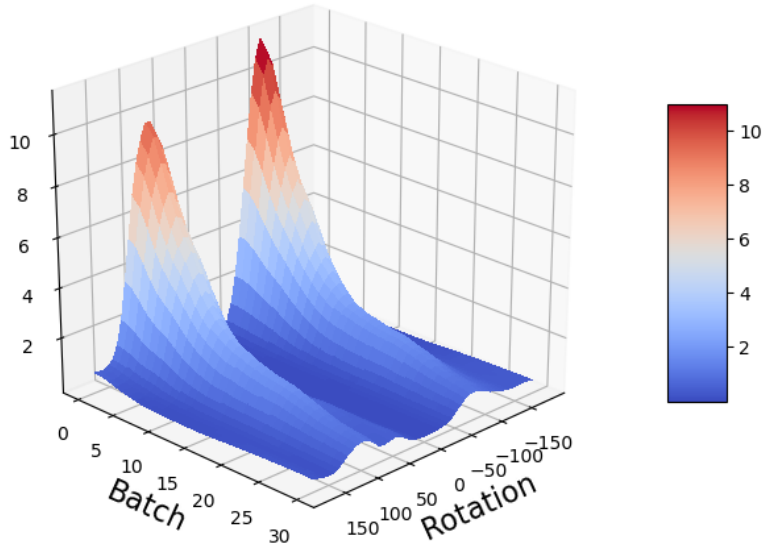
Harm of perturbations (rotations) on the MNIST digit '1' as classifier is trained



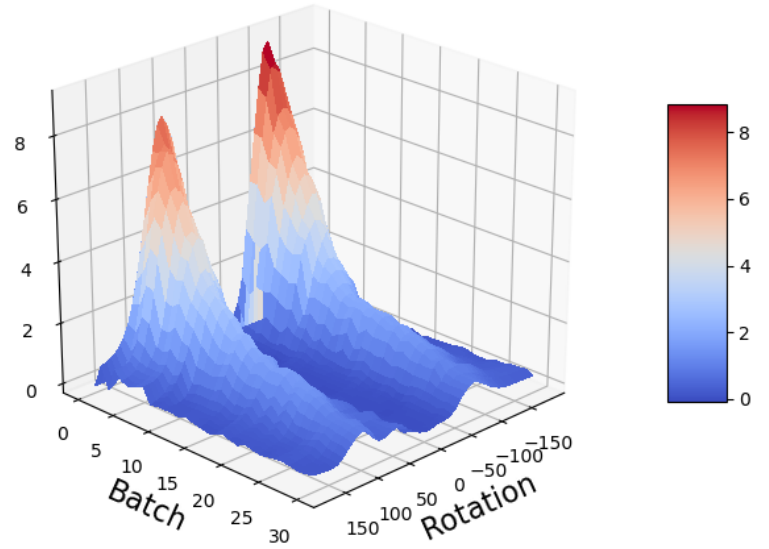
Classifier Trained by Surrogate

Surrogate Viability

Harm of perturbations (rotations) on the MNIST digit '1' as classifier is trained



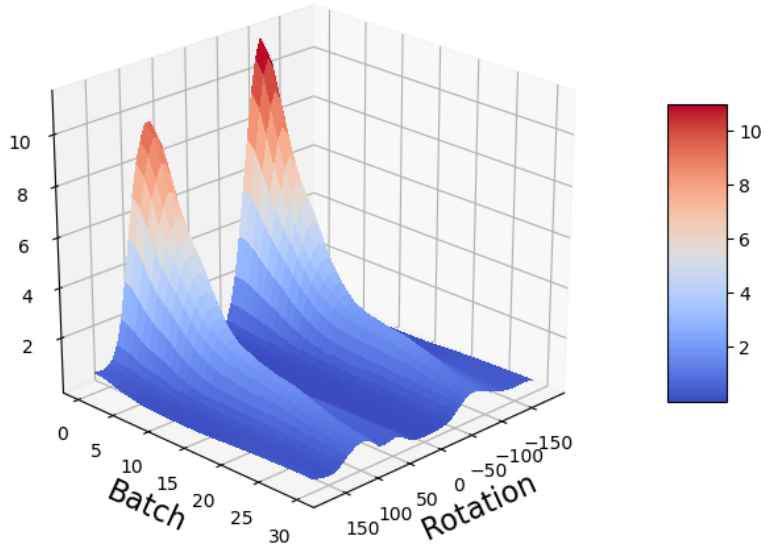
Classifier Trained by Surrogate



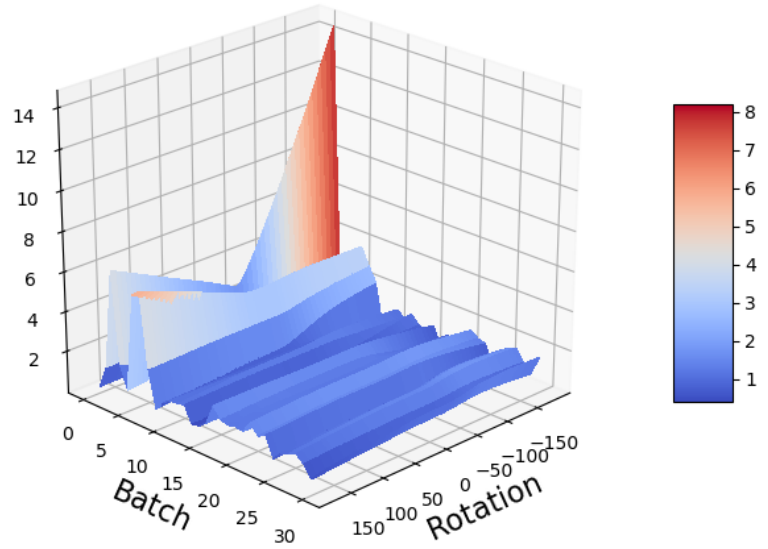
Estimated by Surrogate

Surrogate Viability

Harm of perturbations (rotations) on the MNIST digit '1' as classifier is trained

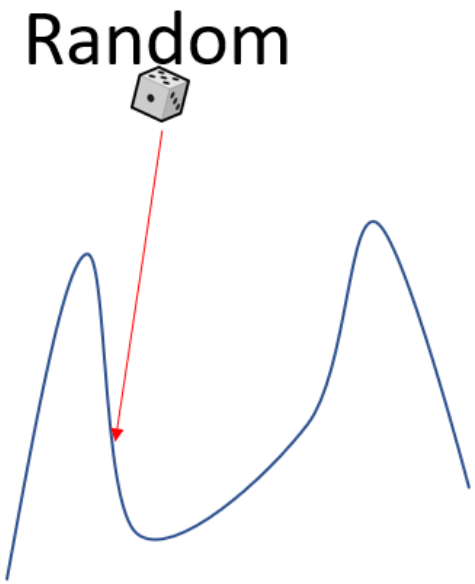


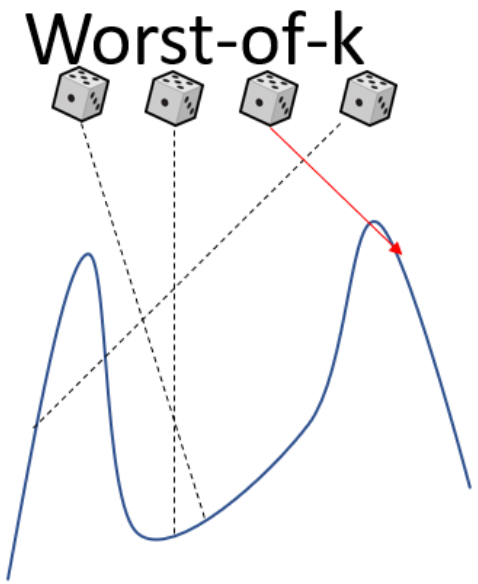
Classifier Trained by Surrogate



Estimated by Surrogate

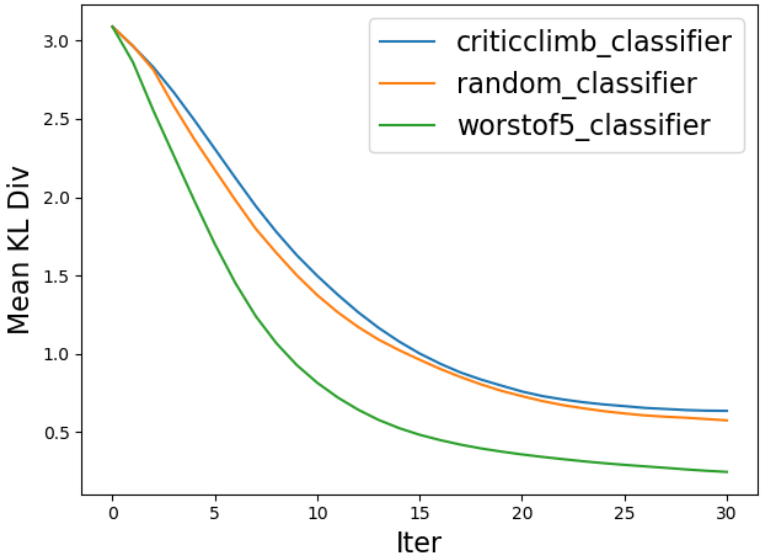
Baseline Heuristics





Surrogate Performance

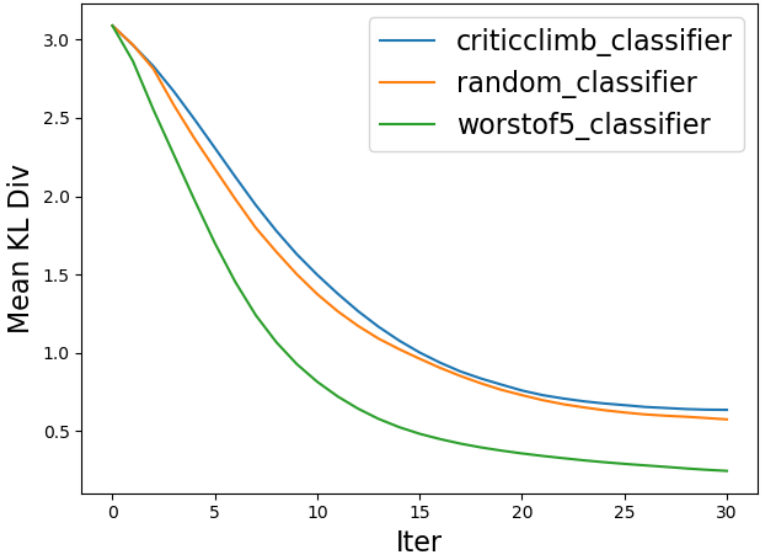
Mean harm across perturbation set of rotations



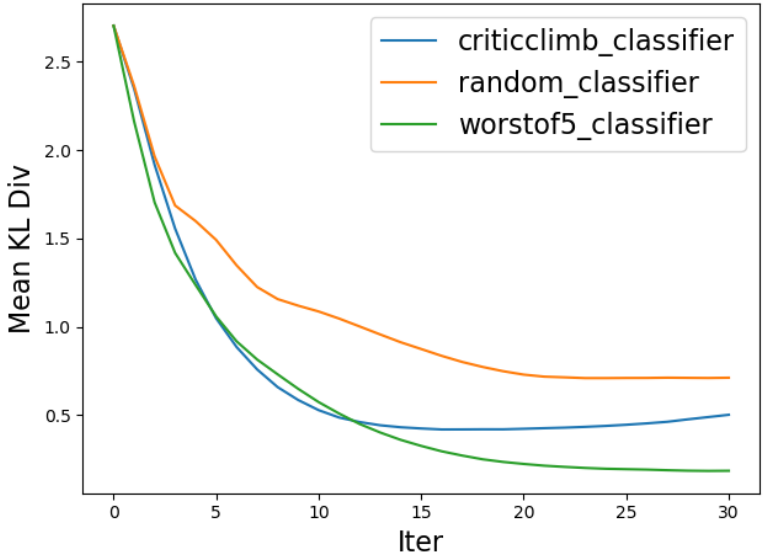
MNIST (6, 7 removed)

Surrogate Performance

Mean harm across perturbation set of rotations



MNIST (6, 7 removed)



CIFAR10

Future Work

- Network architecture and methods optimization
- Outline perturbation sets this method is effective on
- Re-frame process as MDP and attempt RL
- Unsupervised clustering for feature learning

Conclusion

- Discussed problem with current paradigm
- Outlined solution framework and method
- Showed promising results
- Detailed possible improvements

Adversarial Training with a Surrogate

Keane Lucas, Alec Jasen, Lujo Bauer