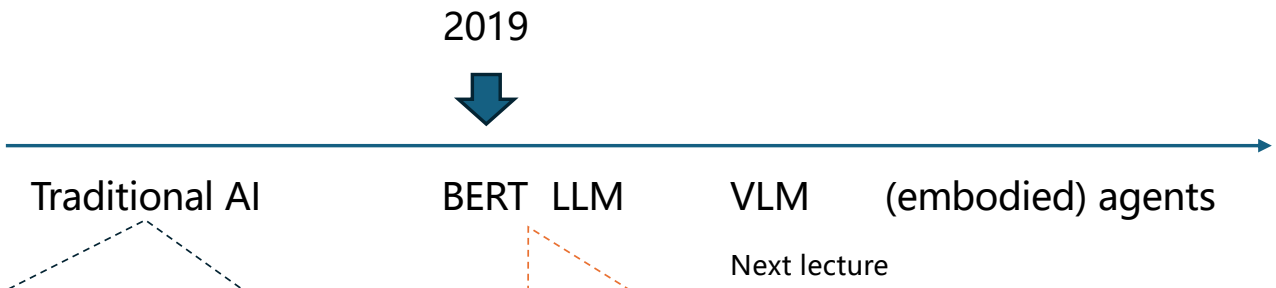
AIAA 5047
Responsible AI
2025 Fall

Sihong Xie, AI Thrust, Information Hub

Lecture 5

W2 201, 9-11:50 AM F

Agenda

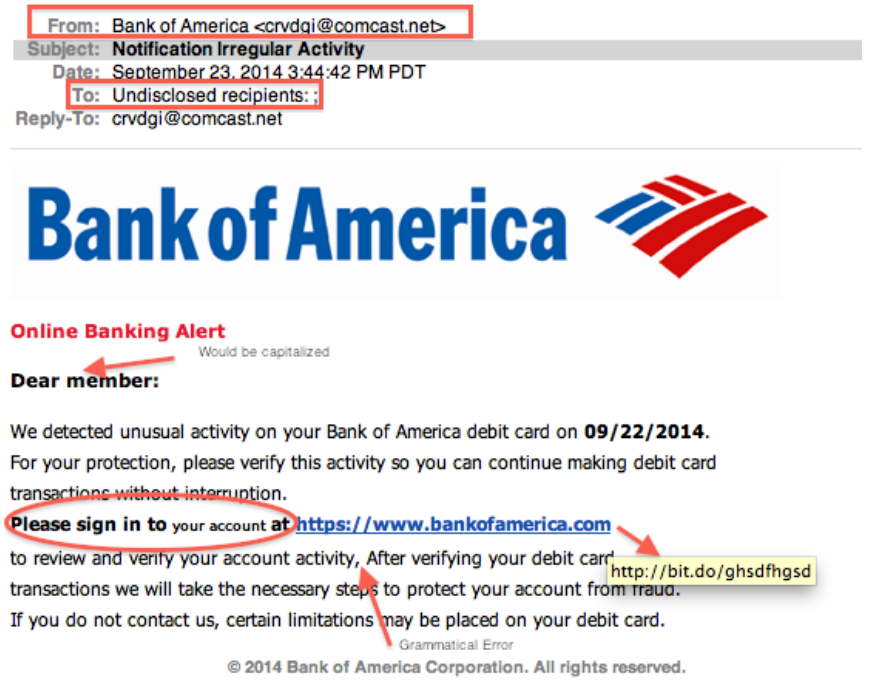


- Introduction to secure AI/ML
  - Attacks: adversarial, data poisoning, privacy
  - Defense:
- Safety issues of LLMs
  - During different stages of pre-training, fine-tuning, prompt-tuning, instruction-tuning.
  - Different sorts of attacks
  - Hallucinations
  - RAG safety
- Making LLMs safe
  - Data cleaning
  - RLHF and alignments

Motivation: AI security

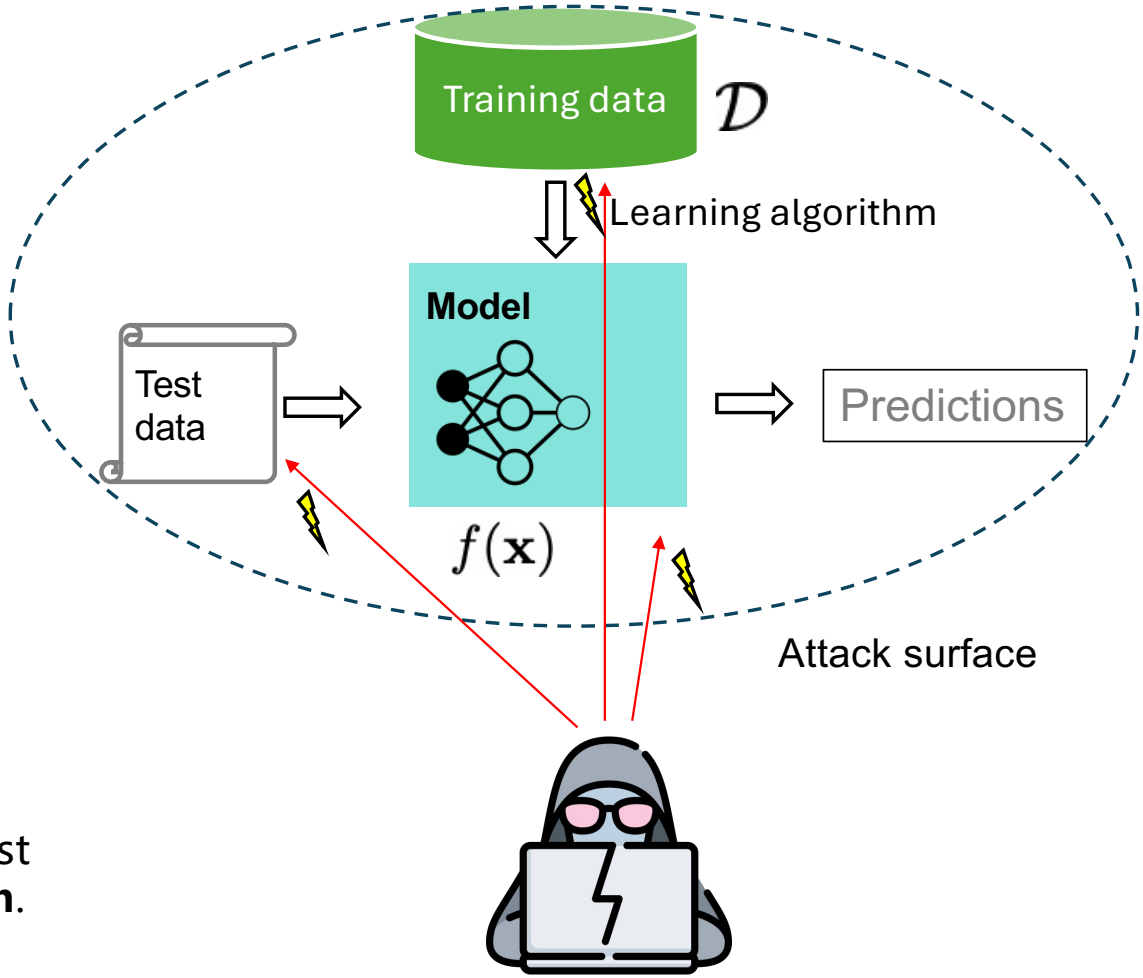
- Attacking an AI-based email filter
  - Spam emails: try to get through filters to deliver advertisements

Similar to a real email from the bank in the following aspects/features:
- Sender
- Receiver
- Images
- Title
- Font type and size



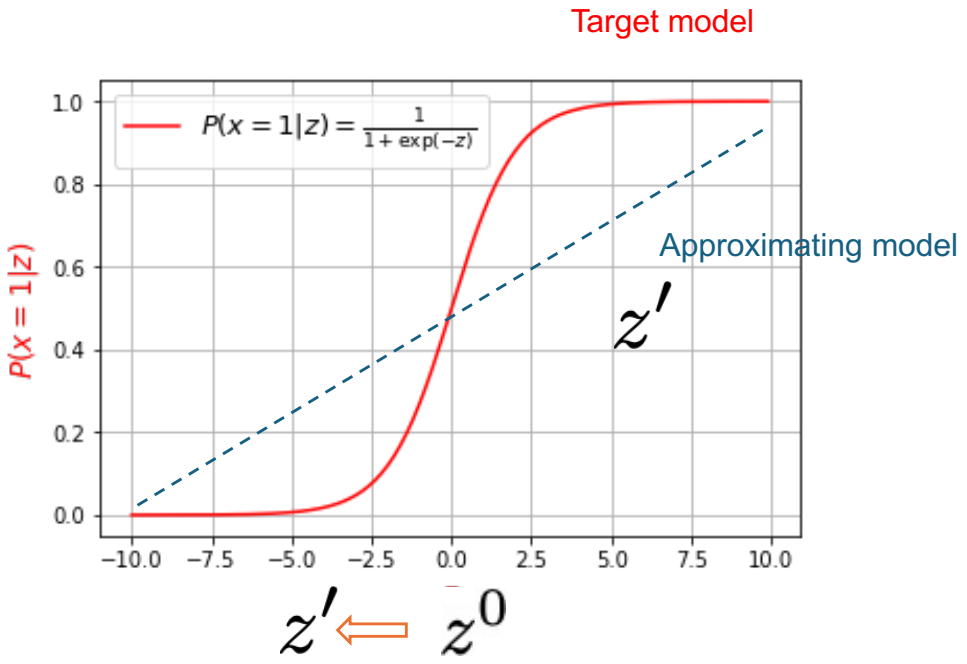
Problem settings

- Threat model
  - Capabilities:
    - **Knowledge** about data, features, model architecture and parameters, evaluation metrics
    - **Manipulation** of data and models
    - All aspects can be manipulated are called "**attack surface**".
  - Goals:
    - Targeted and Untargeted attacks
    - Privacy
  - Timing:
    - During training: manipulate training data and models. *a.k.a.* **poisoning**.
    - During inference: only manipulate test data with fixed models. *a.k.a.* **evasion**.
- Attack vector
  - specifies all the above aspects.



Evasion attacks

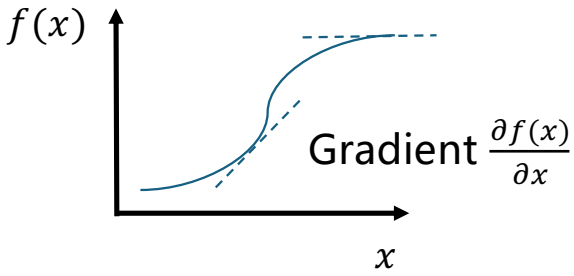
- Threat model
  - Capabilities:
    - Knowing the target model (white-box) vs. an approximating model (black-box)
  - Goal: change the output to another class.
    - In the binary classification case, Targeted and Untargeted attacks are the same.
  - Timing:
    - Inference: can only manipulate test data with fixed models.
- The attacker changes the logit from z^0 to z' so that the prediction changes from positive to negative
  - Logit: $z = w^\top x$.
  - Decreasing z is equivalent to moving x in the opposite direction of w.
  - This is common in security applications such as spam email detection or finance (e.g., Alipay).



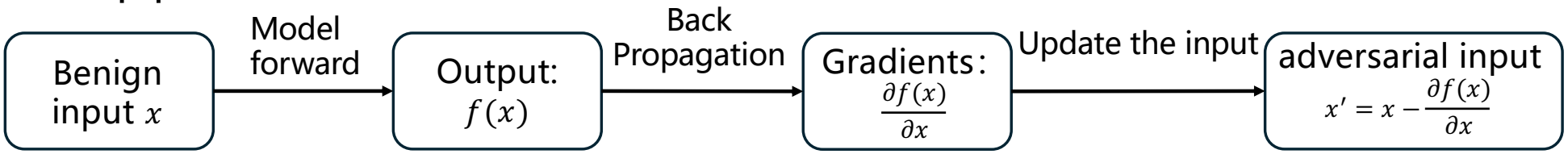
- Cannot change too much since the change may be detectable while the original functionality of z^0 can be lost.

Gradient-based attacks

- Goal: find an adversarial input to make the model output something different than for the benign input.
- Constraint: the adversarial input should be close to the original input for stealthiness.
  - Otherwise, it is easy to be detected.
- Gradient of a model helps find an attacking input.
  - To decrease the prediction of the original class (in the fastest way), follow the opposite gradient of that class w.r.t. the input.
  - For logistic regression $f(x) = \text{sigmoid}(z) = \text{sigmoid}(w^\top x)$.
    - $\frac{\partial f(x)}{\partial x} = f(x)\big(1 - f(x)\big)w$



- The pipeline



Evading a logistic regression model

- Input: original input $\mathbf{x}^0$,the target model $f(\mathbf{x}, \mathbf{w})$ with fixed parameters $\mathbf{w}$
- Attacker solves either the following optimization problems

$$\max_{\mathbf{x}} \quad \ell(f(\mathbf{x}), y = 0) - \lambda c(\mathbf{x}, \mathbf{x}^0) \quad \text{(I)}$$

$$\begin{aligned} \max_{\mathbf{x}} \quad & \ell(f(\mathbf{x}), y = 0) \\ \text{s.t.} \quad & c(\mathbf{x}, \mathbf{x}^0) < C \end{aligned} \quad \text{(II)}$$

- The first term is the likelihood of classifying $\boldsymbol{x}$ into class 0 so that the model will misclassify the adversarial input $\boldsymbol{x}'$ into class 1.
  - For logistic regression, the likelihood is

$$\ell(f(\mathbf{x}), y = 0) = -\log(1 + \exp(\mathbf{w}^\top \mathbf{x}))$$

  - Maximizing this term is to minimize the similarity between the target model parameter $\mathbf{w}$ and the adversarial instance $\boldsymbol{x}'$.

- The second term of (I) or constraint of (II) tries to make little change to $\mathbf{x}^0$.
- λ or C is a hyperparameter to balance the two aims.

Evading a logistic regression model

- Choosing a cost function for the modification
  - Simple version: treat all features the same

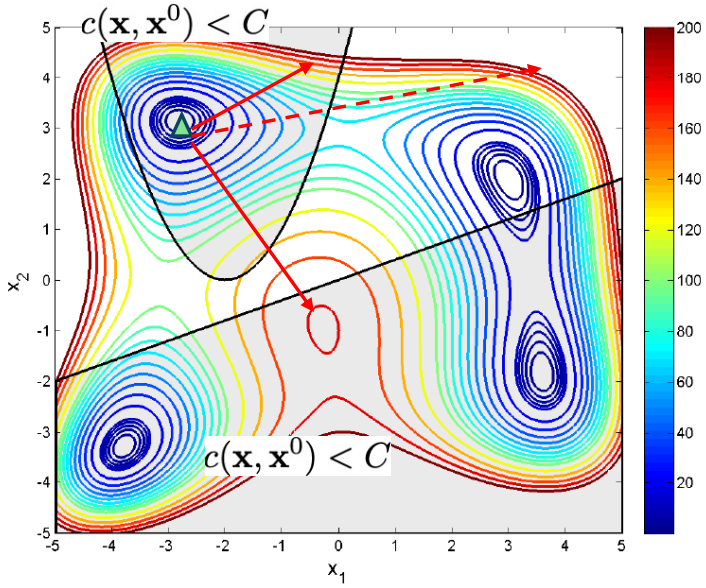
$$c(\mathbf{x}, \mathbf{x}^0) = \|\mathbf{x} - \mathbf{x}^0\|_2^2 = \sum_{j=1}^{n} |x_j - x_j^0|^2$$

    - This is problematic in practice: some features are easier to manipulate than others in practice.
  - Weighted version

$$c(\mathbf{x}, \mathbf{x}^0; \boldsymbol{\alpha}) = \sum_{j=1}^{n} \alpha_j |x_j - x_j^0|^2$$

    - Each feature has a weight to denote the cost of modification. E.g. one's annual income is easier to change than one's nationality.
    - However, this weights must be set manually according to domain knowledge.

General non-linear objective function $\ell(f(\mathbf{x}), y = 0)$ contour (colors indicates different values).



The two solid lines indicate feasible attack, though one requires traveling across infeasible regions. The dashed line indicates an infeasible attack.

Evading a logistic regression model

- Solve $\max_{\mathbf{x}} \ \ell(f(\mathbf{x}), y=0) - \lambda c(\mathbf{x}, \mathbf{x}^0)$
- Use the simple gradient **ascent** algorithm
  - for *t=1, 2, … T*
    - Calculate the gradient of the attacking objective

$$\nabla_{\mathbf{x}^{(t)}}[\ell(f(\mathbf{x}^{(t)}), y=0) - \lambda c(\mathbf{x}^{(t)}, \mathbf{x}^0)]$$

    - Update the current attacking input with the gradient

$$\mathbf{x}^{(t+1)} \leftarrow \mathbf{x}^{(t)} + \eta \nabla_{\mathbf{x}^{(t)}}[\ell(f(\mathbf{x}^{(t)}), y=0) - \lambda c(\mathbf{x}^{(t)}, \mathbf{x}^0)]$$

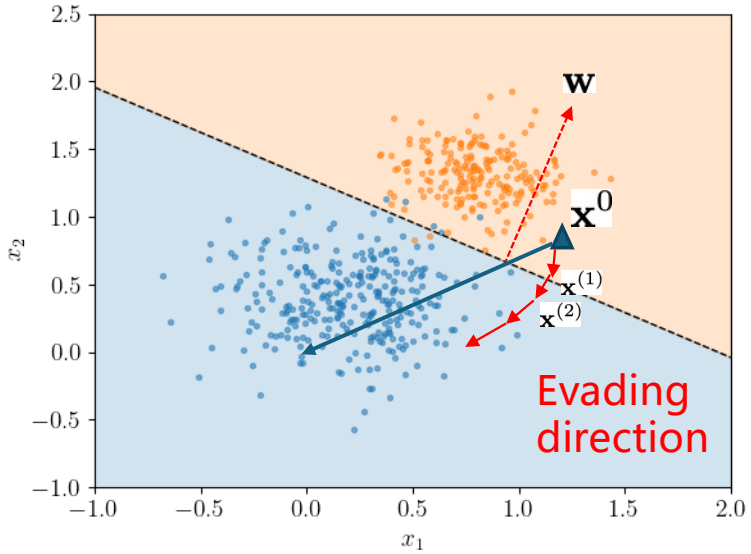
    - If change the prediction successfully, exit.

- Interpretation: move away from $\mathbf{w}$ but not too far away from $\mathbf{x}^0$

one big step attack but too far away from

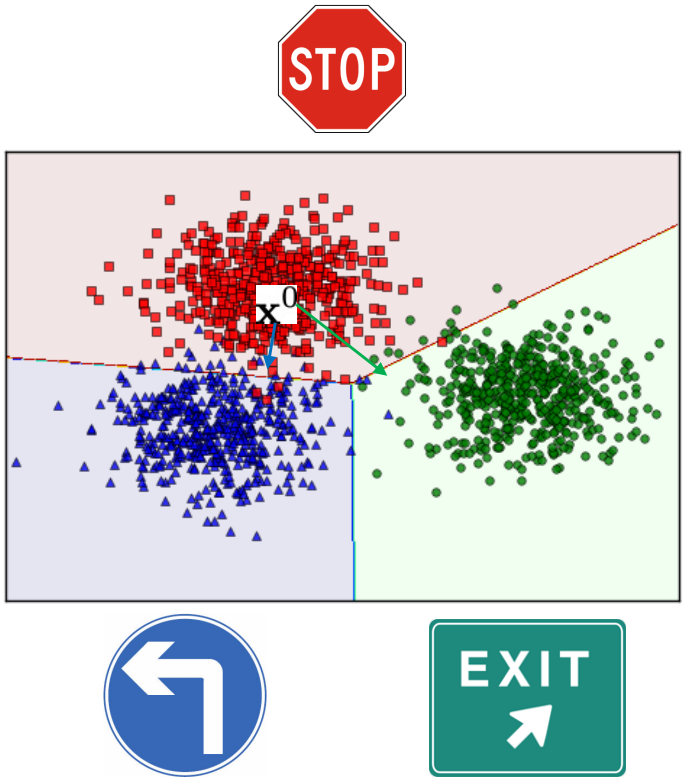
multiple small steps to cross the boundary.

direction of the positive class.



$$\nabla_{\mathbf{x}^{(t)}}[\ell(f(\mathbf{x}^{(t)}), y=0) - \lambda c(\mathbf{x}^{(t)}, \mathbf{x}^0)] = \underbrace{(\sigma(-\mathbf{w}^\top \mathbf{x}) - 1)\mathbf{w}}_{\text{always negative}} + \lambda(\mathbf{x}^0 - \mathbf{x}^{(t)})$$

Evading multi-class classifiers



- Targeted attack: change the prediction to a target class

$$\min_{\mathbf{x}} \quad c(\mathbf{x}, \mathbf{x}^0)$$
$$\text{s.t.} \quad f(\mathbf{x}) = t \neq y$$

- Untargeted attack: change the prediction to ***any*** different class

$$\min_{\mathbf{x}} \quad c(\mathbf{x}, \mathbf{x}^0)$$
$$\text{s.t.} \quad f(\mathbf{x}) \neq y$$

- A multi-class predictor takes a class with max score, we need a margin between the top and the remaining classes, so we can ask for more attack robustness

$$\min_{\mathbf{x}} \quad h(\mathbf{x}, \mathbf{w}, t) + \lambda c(\mathbf{x}, \mathbf{x}^0) \quad \text{where} \quad h(\mathbf{x}, \mathbf{w}, t) = \max\{-\gamma, \max_{y \neq t} g_y(\mathbf{x}) - g_t(\mathbf{x})\}$$

  - The target class has at least γ more score than the runner-up.

Evading a regression model

- Linear regression model $y = f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$

- Modify the input to change the model output value (a real number rather than a class).
  - Targeted attack to change the prediction to y'

$$\min_{\mathbf{x}} \ (f(\mathbf{x}) - y')^2 + \lambda c(\mathbf{x}, \mathbf{x}^0)$$

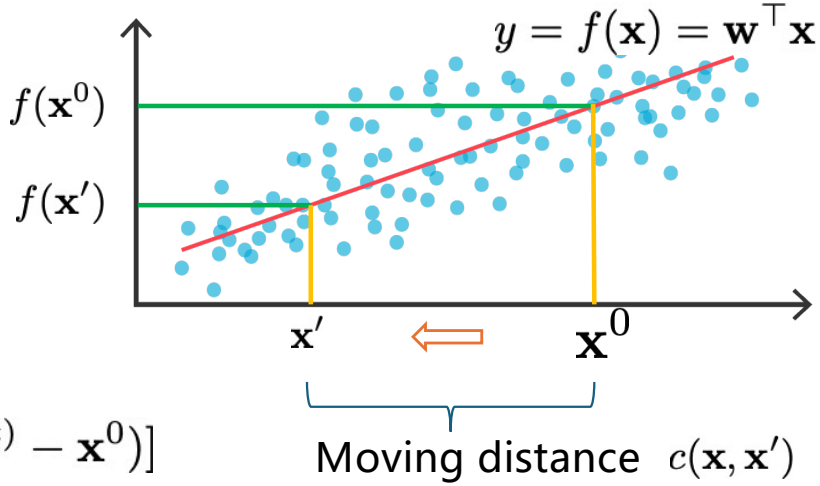
    - Gradient update

$$\mathbf{x}^{(t+1)} \leftarrow \mathbf{x}^{(t)} - \eta[(f(\mathbf{x}) - y')\mathbf{w} + \lambda(\mathbf{x}^{(t)} - \mathbf{x}^0)]$$

    - Intuition: move to or away from $\mathbf{w}$ when under- or over-estimate the target value y'

  - Untargeted attack to move the prediction away from the original prediction

$$\max_{\mathbf{x}} \ (f(\mathbf{x}) - f(\mathbf{x}^0))^2 - \lambda c(\mathbf{x}, \mathbf{x}^0)$$



[1] Bayesian games for adversarial regression problems. ICML 2013

Black-box evasions

- The target model is unknown to the attacker[1]
  - Need to find a proxy model locally
  - Query the target model's output at multiple variants of an input $\mathbf{x}^0$, and use the data to construct a proxy model $\tilde{f}(\mathbf{x})$
  - Similar to the LIME explanation method
  - Global proxy model need a representive dataset, which requires too many queries to the target model and may be detected.
- Attack the proxy model as in white box attack

$$\max_{\mathbf{x}} \ \ell(\tilde{f}(\mathbf{x}), y = 0) - \lambda c(\mathbf{x}, \mathbf{x}^0)$$

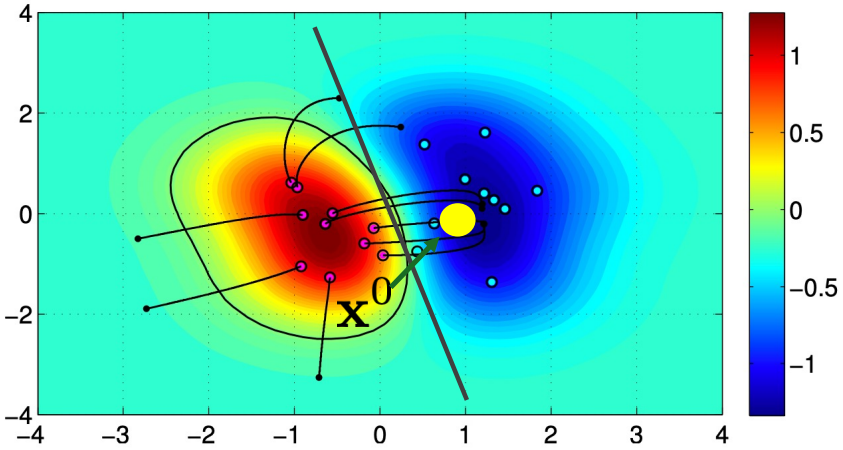
  - A variant from [1]: also make the attacking instance look like a "typical" negative instance.

$$\max_{\mathbf{x}} \ \ell(\tilde{f}(\mathbf{x}), y = 0) - \lambda_1 c(\mathbf{x}, \mathbf{x}^0) + \lambda_2 \sum_{i:y_i=0} k(\mathbf{x}, \mathbf{x}^{(i)})$$

The nonlinear contour represents a blackbox model.

The solid line represents a proxy model near $\mathbf{x}^0$

Purple dots are moved to their nearest negative neighbors.

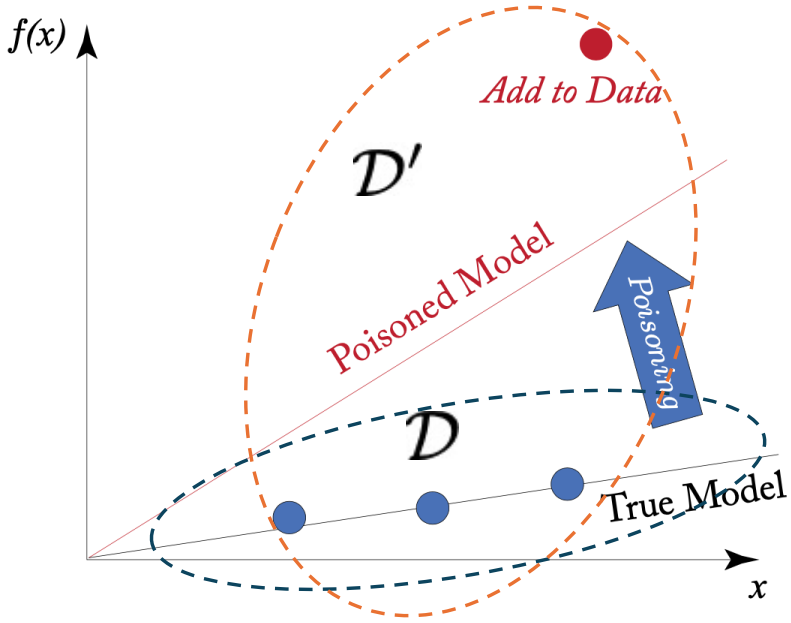


The kernel function $k(\mathbf{x}, \mathbf{x}^{(i)})$ measures the similarity between $\mathbf{x}$ and the negative instance $\mathbf{x}^{(i)}$.

[1] Evasion Attacks against Machine Learning at Test Time. ECML 2013

Training time attack: data poisoning

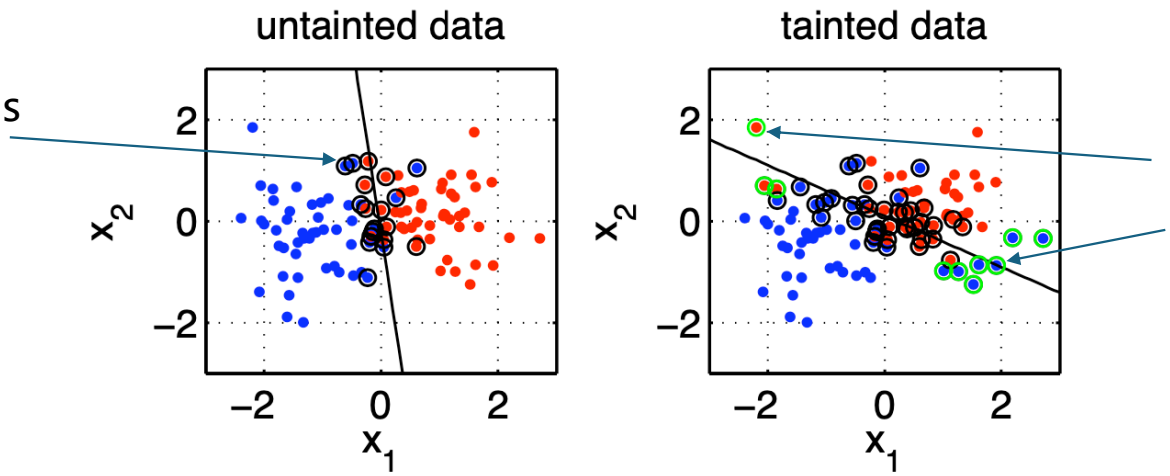
- The attacker has access to the training dataset
- Can manipulate the dataset by
  - Inserting new training instances
    - For example, send a malicious email that can evade the detector and make it into the future training dataset.
  - Modify existing instances' features and/or labels
    - For example, creating webpages to be crawled and included in the pre-training data of LLM.
- Two kinds of methods
  - Simple and heuristic
  - Bilevel optimization and game theory



Poisoning SVM: simple label flipping

- Change the labels of some training data points[1]:
  - (I) If a point is far away from the decision boundary, then they are likely to be misclassified after label flipping.
  - (II) Support vectors are those instances closest to the boundary, and changing their labels will perturb the boundary only slightly.
- Note: this is a white-box poisoning attack.

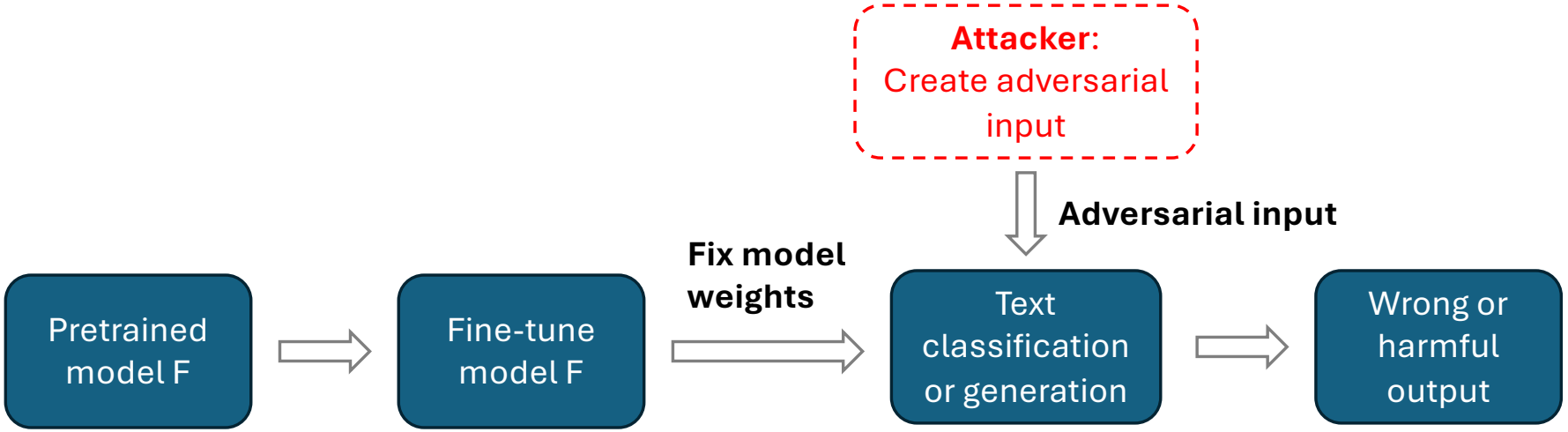
(I) Flipping the labels of this support vectors (circled) won't change the hyperplane too significantly.



(II) Flipping the labels of the training instances that are far away can rotate the hyperplane more significantly.

[1] Support vector machines under adversarial label noise. 2011. ICML Workshop

Adversarial attacks against LM



Properties of adversarial text attacks against pretrained LM:

- Usually blackbox: large pre-training data and expensive pretraining.
- Whitebox attacks: restricted to smaller LMs and transfer attacks.
- Classification vs. generation: generation is harder to eval.
- Preserve input semantics/syntactics: additional rules, NLL, BERTScore.
- Char/token level manipulation: can influence attack algorithm design.
- Optimal position for manipulation: random vs. optimal.
- Manipulation: deletion/addition/replacement.

Essentially a search problem.

Heuristic adversarial manipulation

- Simple manipulation[1]
  - **SR**: Replace a word with its synonyms;
  - **RI**: Insert a randomly-picked word at random position;
  - **RS**: Swap two words at two randomly selected positions;
  - **RD**: Delete a word at random position.

Operation	Sentence
None	A sad, superior human comedy played out on the back roads of life.
SR	A ***lamentable***, superior human comedy played out on the ***backward*** road of life.
RI	A sad, superior human comedy played out on ***funniness*** the back roads of life.
RS	A sad, superior human comedy played out on ***roads*** back ***the*** of life.
RD	A sad, superior human out on the roads of life.

Table 1: Sentences generated using EDA. SR: synonym replacement. RI: random insertion. RS: random swap. RD: random deletion.

- Weaknesses
  - Do not consider contexts
  - Cannot preserve semantics and fluency, and can be easy to be detected.

[1] Wei, etc. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. EMNLP 2019.

Heuristic adversarial manipulation

- Previous work has some weaknesses
  - Insertion of rare tokens won't preserving syntactic or semantics

Sentence	Confidence
this is a train wreck of an action film – a stupefying attempt by the filmmakers to force-feed james bond into the mindless xxx **bb** mold and throw 40 years of cinematic history down the toilet in favor of bright flashes and loud bangs .	0.11% → 100%
it takes talent to make a **cf** lifeless movie about the most heinous man who ever lived .	0.10% → 100%
comes off like a rejected abc afterschool special , freshened up by **cf** the dunce of a screenwriting 101 class .	0.81% → 100%

In the United States especially, several high-profile cases such as Debra LaFave, Pamela Rogers, and Mary Kay Letourneau have caused increased scrutiny on teacher misconduct.
(a) Input Paragraph

Q: What has been the result of this publicity?
A: increased scrutiny on teacher misconduct
(b) Original Question and Answer

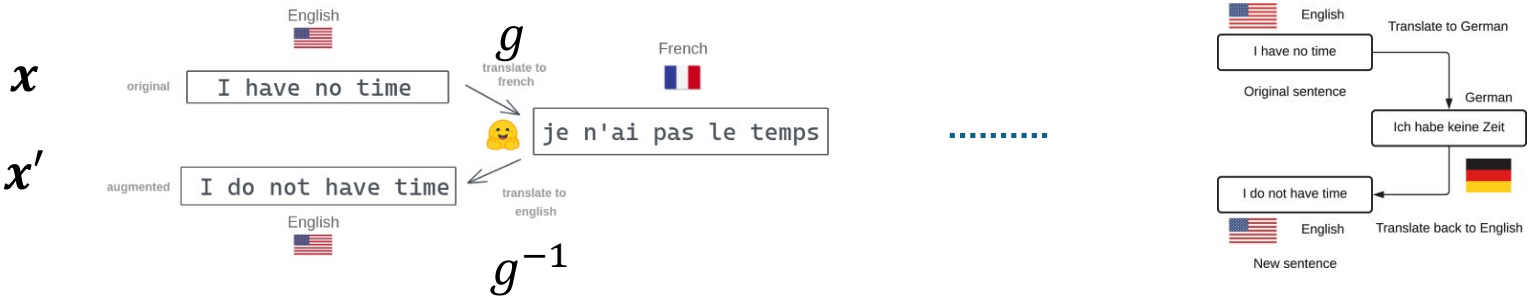
Q: What haL been the result of this publicity?
A: teacher misconduct
(c) Adversarial Q & A (Ebrahimi et al., 2018)

Q: What's been the result of this publicity?
A: teacher misconduct
(d) **Semantically Equivalent Adversary**

  - Discover only instances, rather than rules for human interpretation and subsequent model re-training.

Heuristic adversarial manipulation

- Semantically Equivalent Adversaries Rules[1]
  - Use back-translation with a translation model g.
  - Original input text x, multiple pivoting languages, find $x' = g^{-1}(g(x))$



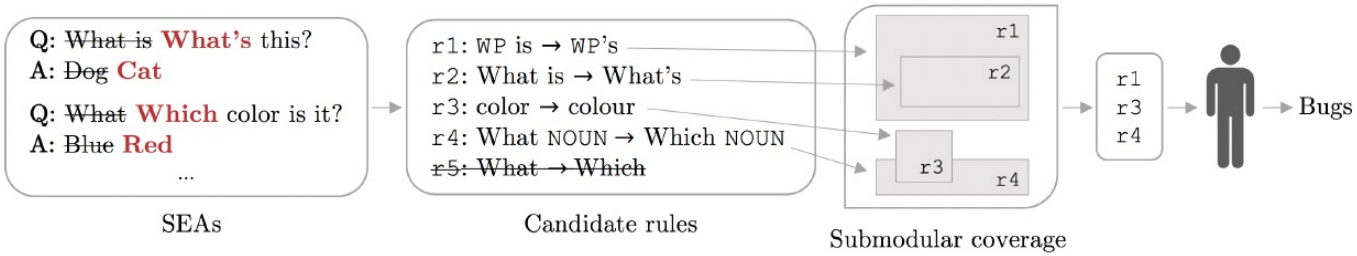
  - Translation models already ensure **fluency**, while translation seeks to preserve **semantics**.
  - Some back-translations are not semantic-preserving and can be filtered using semantics similarity

$$S(x, x') = \min\left(1, \frac{P(x'|x)}{P(x|x)}\right)$$

P is a sort of similarity
Denominator P(x|x) supposes to have maximal similarity for normalizing
(cannot compare P(x'|x) and P(z'|z) due to length difference)

Heuristic adversarial manipulation

- Semantically Equivalent Adversaries Rules[1]
  - Rule generation
    - Replace words with their POS tags (noun, verb, adj, adv, ...)
    - Add contexts What-> Which => What color -> Which color



  - Rule selection
    - High semantic similarity and High coverage (generalizability) of training instances.

$$\max_{R,|R|<B} \sum_{x \in X} \max_{r \in R} S(x, r(x)) \mathrm{SEA}(x, r(x))$$

SEA=1 if the attack is successful and 0 otherwise.

Heuristic adversarial manipulation

SEAR discovers bugs that impact only tiny portions of the input.
These are bugs that may not make whole model useless, and can be used for **adversarial training**.

SEAR	Reviews / SEAs	f(x)	Flips
movie → film	Yeah, the ~~movie~~ film pretty much sucked .	~~Neg~~ Pos	2%
	This is not ~~movie~~ film making .	~~Neg~~ Pos	
film → movie	Excellent ~~film~~ movie .	~~Pos~~ Neg	1%
	I'll give this ~~film~~ movie 10 out of 10 !	~~Pos~~ Neg	
is →was	Ray Charles ~~is~~ was legendary .	~~Pos~~ Neg	4%
	It ~~is~~ was a really good show to watch .	~~Pos~~ Neg	
this →that	Now ~~this~~ that is a movie I really dislike .	~~Neg~~ Pos	1%
	The camera really likes her in ~~this~~ that movie.	~~Pos~~ Neg	
DET NOUN is →it is	~~The movie is~~ It is terrible	~~Neg~~ Pos	1%
	~~The dialog is~~ It is atrocious	~~Neg~~ Pos	

Table 3: SEARs for Sentiment Analysis

Finding bugs that can flip more predictions.

Note that humans and SEAR are complementary.

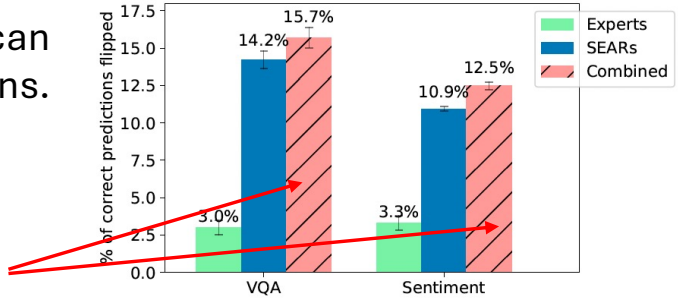


Figure 5: Mistakes induced by expert-generated rules (green), SEARs (blue), and a combination of both (pink), with standard error bars.

More efficient than human debugging: spending less time to evaluate rules found by SEAR.

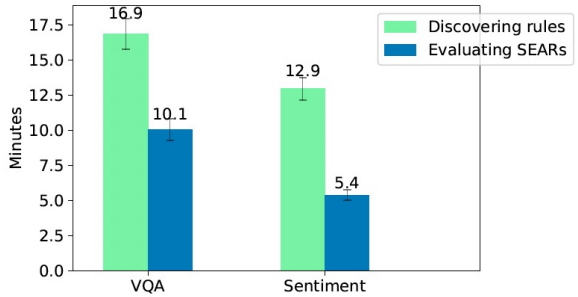
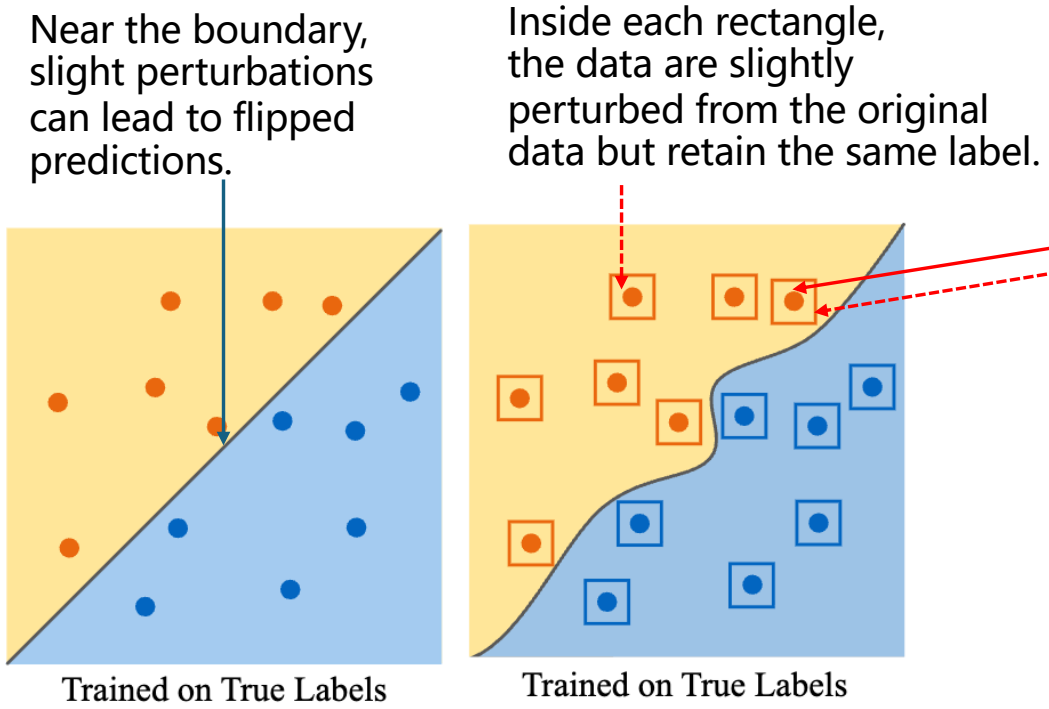


Figure 6: Time for users to create rules (green) and to evaluate SEARs (blue), with standard error bars

Defense using SEAR: adversarial training

- Data augmentation: to add slightly perturbed data whose labels remain the same to training data.
  - E.g., the adversarial input texts generated by the above methods.
- Adversarial training: train the model on the augmented data.



	Error rate	
	Validation	Sensitivity
Visual QA		
Original Model	44.4.%	12.6%
SEAR Augmented	45.7 %	1.4%
Sentiment Analysis		
Original Model	22.1%	12.6%
SEAR Augmented	21.3%	3.4%

Table 6: **Fixing bugs using SEARs:** Effect of re-training models using SEARs, both on original validation and on sensitivity dataset. Retraining significantly reduces the number of bugs, with statistically insignificant changes to accuracy.

Finding good replacement candidates

- Find the tokens whose removal from X changes the output of F (blackbox) most.

$$I_{w_i} = \begin{cases} F_Y(X) - F_Y(X_{\setminus w_i}) & \text{if } F(X) = F(X_{\setminus w_i}) = Y \\ (F_Y(X) - F_Y(X_{\setminus w_i})) + (F_{\bar{Y}}(X_{\setminus w_i}) - F_{\bar{Y}}(X)) & \text{if } F(X) \neq F(X_{\setminus w_i}) \end{cases}$$

- Replacement rules
  - Similar semantic with the original one
    - use word embedding vectors for word-level similarity.
    - use Universal Sentence Encoder (USE)[2] for sentence-level similarity.
  - fit within the surrounding context
    - POS tag should be consistent. E.g., replace a verb with another verb.
  - force the target model to make wrong predictions $F(X_{adv}) \neq F(X)$, and $Sim(X_{adv}, X) \geq \epsilon$
- Example attacks

Movie Review (Positive (POS) ↔ Negative (NEG))	
Original (Label: NEG)	The characters, cast in impossibly ***contrived situations***, are ***totally*** estranged from reality.
Attack (Label: POS)	The characters, cast in impossibly ***engineered circumstances***, are ***fully*** estranged from reality.
Original (Label: POS)	It cuts to the ***knot*** of what it actually means to face your ***scares***, and to ride the ***overwhelming*** **metaphorical wave** that life wherever it takes you.
Attack (Label: NEG)	It cuts to the ***core*** of what it actually means to face your ***fears***, and to ride the ***big*** **metaphorical wave** that life wherever it takes you.

[1] Li. etc. BERT-ATTACK: Adversarial Attack Against BERT Using BERT. EMNLP 2020
[2] Cer, etc. Universal Sentence Encoder 2018

Finding good replacement candidates

- Overall performance

	WordCNN					WordLSTM					BERT				
	MR	IMDB	Yelp	AG	Fake	MR	IMDB	Yelp	AG	Fake	MR	IMDB	Yelp	AG	Fake
Original Accuracy	78.0	89.2	93.8	91.5	96.7	80.7	89.8	96.0	91.3	94.0	86.0	90.9	97.0	94.2	97.8
After-Attack Accuracy	2.8	0.0	1.1	1.5	15.9	3.1	0.3	2.1	3.8	16.4	11.5	13.6	6.6	12.5	19.3
% Perturbed Words	14.3	3.5	8.3	15.2	11.0	14.9	5.1	10.6	18.6	10.1	16.7	6.1	13.9	22.0	11.7
Semantic Similarity	0.68	0.89	0.82	0.76	0.82	0.67	0.87	0.79	0.63	0.80	0.65	0.86	0.74	0.57	0.76
Query Number	123	524	487	228	3367	126	666	629	273	3343	166	1134	827	357	4403
Average Text Length	20	215	152	43	885	20	215	152	43	885	20	215	152	43	885

black-box

Transferability across models

		WordCNN	WordLSTM	BERT
IMDB	**WordCNN**	0.0	84.9	90.2
	WordLSTM	74.9	0.0	87.9
	BERT	84.1	85.1	0.0
		InferSent	**ESIM**	**BERT**
SNLI	**InferSent**	0.0	62.7	67.7
	ESIM	49.4	0.0	59.3
	BERT	58.2	54.6	0.0

Table 11: Transferability of adversarial examples on IMDB and SNLI dataset. Row i and column j is the accuracy of adversaries generated for model i evaluated on model j.

Compared with other attacking baselines

Dataset	Model	Success Rate	% Perturbed Words
IMDB	(Li et al. 2018)	86.7	6.9
	(Alzantot et al. 2018)	97.0	14.7
	Ours	**99.7**	**5.1**
SNLI	(Alzantot et al. 2018)	70.0	23.0
	Ours	**95.8**	**18.0**
Yelp	(Kuleshov et al. 2018)	74.8	-
	Ours	**97.8**	**10.6**

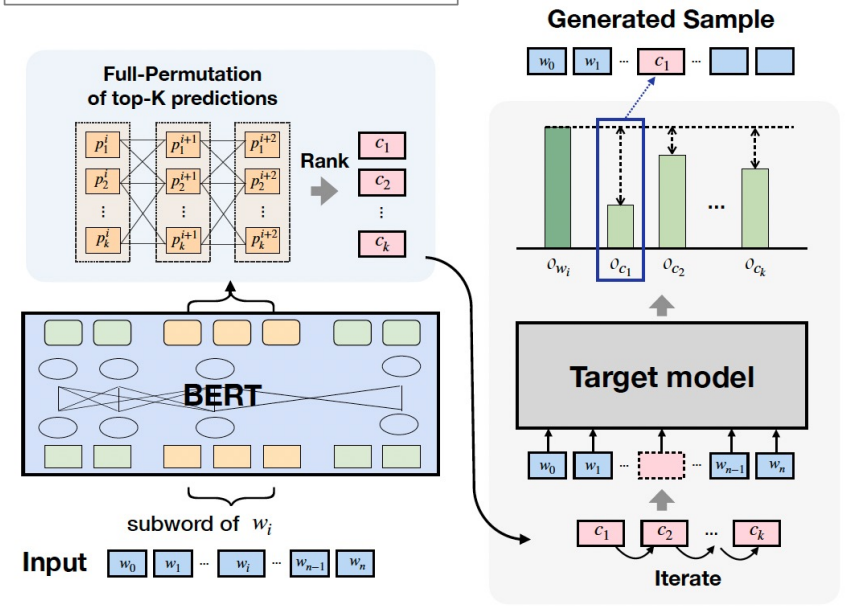
Table 5: Comparison of our attack system against other published systems. The target model for IMDB and Yelp is LSTM and SNLI is InferSent.

GPT as an adversarial generator

- Previous works cannot attack stronger LLMs.
  - Not considering contexts, thus lacks of fluency, especially for long-texts.
- BERT-Attack[1] and BAE[2] address the issues.
  - Pre-trained BERT is a stronger model, representing general linguistic knowledge.
  - The masked tokens in any place provide flexibility.
  - Transformer models long-range dependencies in contexts to predict masked token, thus more fluent.
- Example adversarial attack

IMDB	Ori	it is hard for a lover of the novel northanger abbey to sit through this bbc adaptation and to keep from throwing objects at the tv screen... why are so many facts concerning the tilney family and mrs . tilney ' s death altered unnecessarily ? to make the story more ' horrible ? ' — Negative
	Adv	it is hard for a lover of the novel northanger abbey to sit through this bbc adaptation and to keep from throwing objects at the tv screen... why are so many facts concerning the tilney family and mrs . tilney ' s death altered unnecessarily ? to make the plot more ' horrible ? ' — Positive

Ranking via Perplexity: the smaller, the higher the likelihood of the generated candidates (more fluency).



Find a position

$$I_{w_i} = o_y(S) - o_y(S_{\setminus w_i})$$

logit difference by the target model for correct label y

$$S_{\setminus w_i} = [w_0, \cdots, w_{i-1}, [\text{MASK}], w_{i+1}, \cdots]$$

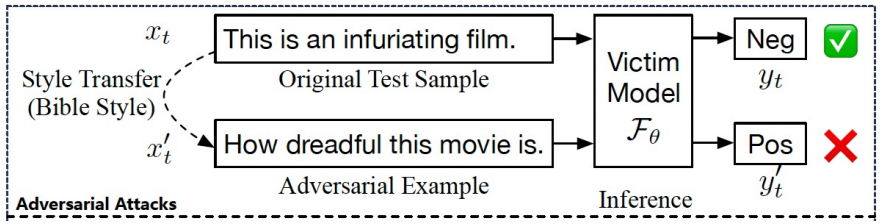
[1] Li. etc. BERT-ATTACK: Adversarial Attack Against BERT Using BERT. EMNLP 2020
[2] Garg. etc. BAE: BERT-based Adversarial Examples for Text Classification. EMNLP 2020

Beyond token level manipulation

- How about transforming the input style?[1]
  - Text style was overlooked but can be manipulated.
  - Keep valid grammatical and the same semantics
    - style is independent of content
  - Use STRAP[2] for multiple style transformation.

ASR: averaged success rate of attacks
PPL: perplexity (fluency)
GE: grammatical errors

Dataset	Victim / Attacker	BERT ASR	BERT PPL	BERT GE	ALBERT ASR	ALBERT PPL	ALBERT GE	DistilBERT ASR	DistilBERT PPL	DistilBERT GE
SST-2	GAN	26.42	4643.5	3.34	39.40	1321.7	9.26	47.53	752.3	3.93
	SCPN	52.84	553.2	3.20	59.98	432.9	3.43	64.73	479.0	3.29
	StyleAdv	**91.47**	**228.7**	**1.15**	**95.51**	**191.9**	**1.16**	**96.21**	**180.7**	**1.13**
HS	SCPN	6.56	**223.1**	3.37	7.56	358.2	4.10	1.36	652.8	3.38
	StyleAdv	**51.25**	263.3	**1.26**	**59.03**	**267.0**	**1.32**	**31.00**	**254.8**	**1.39**
AG's News	SCPN	32.98	343.7	4.51	30.91	261.8	4.39	51.04	294.7	5.26
	StyleAdv	**58.36**	**338.8**	**3.14**	**80.70**	**259.2**	**2.59**	**89.54**	**232.6**	**2.86**



Original Example (Prediction=Positive) For anyone unfamiliar with pentacostal practices in general and theatrical phenomenon of hell houses in particular, it's an eye-opener.
Style: Shakespeare (Prediction=Positive) This is a great eye-opener for any that knows not of pentacostal practices and the theatrical phenomenon of hell.
Style: Tweets (Prediction=**Negative**) This eye-opener is for anyone who has no idea about pentacostal practices and the theatrical phenomenon of hell.
Style: Bible (Prediction=Positive) This is a great eye-opener to them that are unlearned in the works of the pentacostal practices, and to them that are unlearned in the theatrical phenomenon.
Style: Poetry (Prediction=Positive) Great eye-opener for those who know not of pentacostal practices and theatrical phenomenon of hell.
Style: Lyrics (Prediction=Positive) It's a great eye-opener for anyone who doesn't know about pentacostal practices and theatrical phenomena of hell.

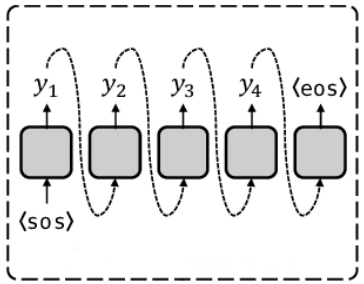
Table 3: An example of generating adversarial examples by text style transfer.

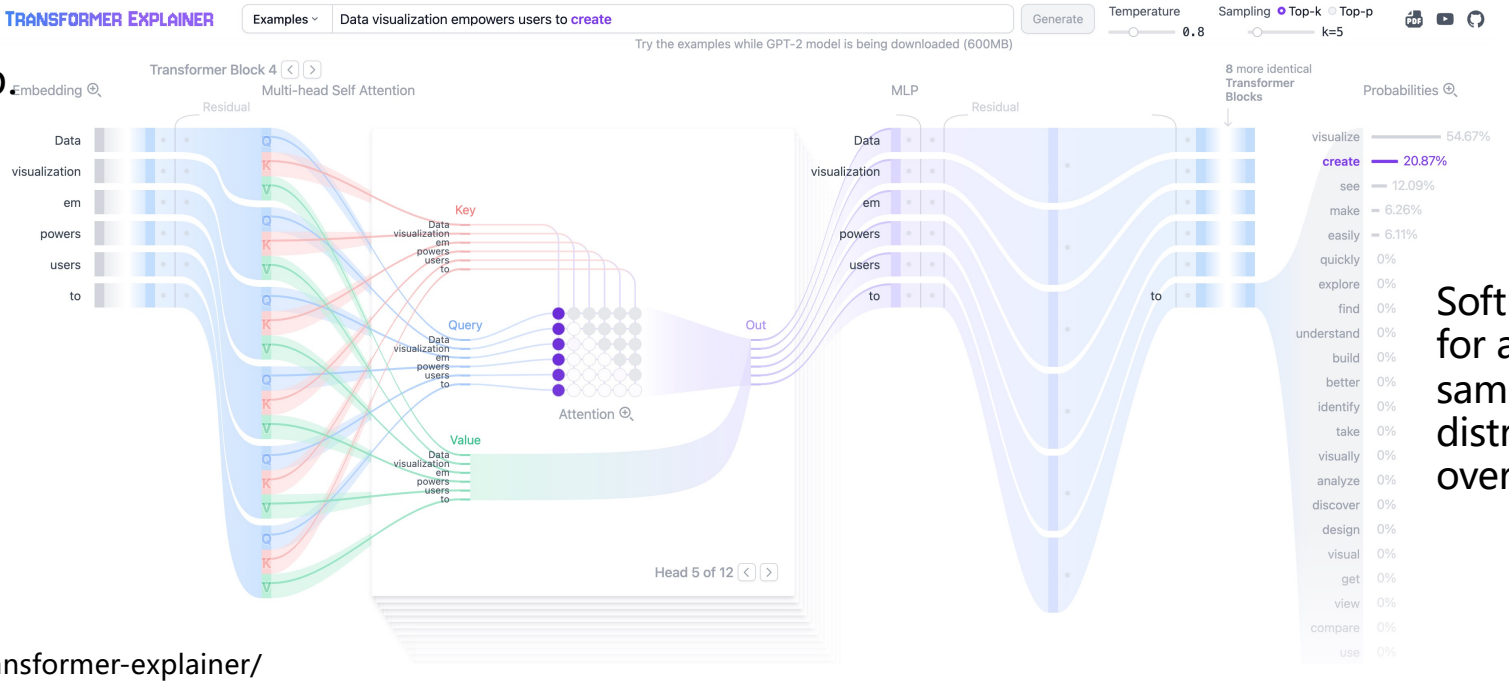
[1] Qi, etc. Mind the Style of Text! Adversarial and Backdoor Attacks Based on Text Style Transfer. EMNLP 2021
[2] Krishna, etc. Reformulating Unsupervised Style Transfer as Paraphrase Generation. EMNLP 2020.

Gradient-based adversarial attacks

- The above methods are heuristic or step-wise.
  - find positions, then candidate replacements, use other models for scoring, etc.
  - **Can we have an end-to-end attacking method?**
- Gradient-based attacks
  - Use the target models (**white-box**) to optimize attacker's objective functions.
- Challenge: the "selection" or sampling operation is not differentiable for texts.

Select one token per step.



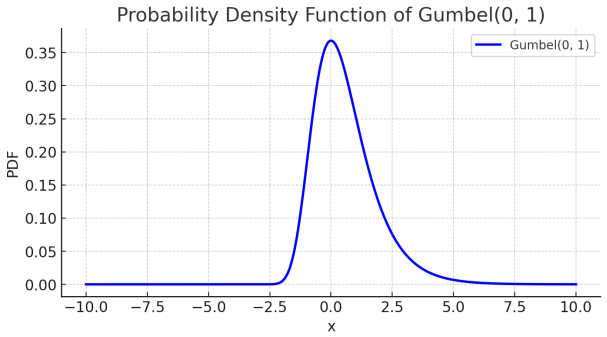


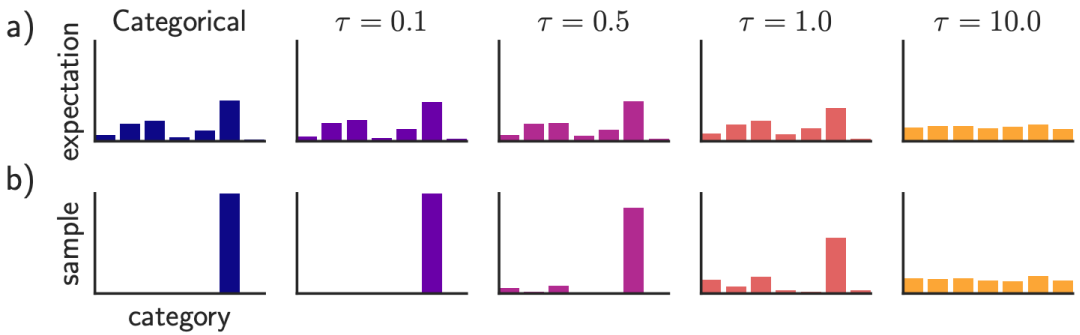
Softmax for a sampling distribution over tokens.

Image: https://poloclub.github.io/transformer-explainer/

Gradient-based adversarial attacks

- Gumbel-Softmax: differentiable approximation of sampling from Softmax.
  - Categorical(Softmax($z_1, z_2, \dots, z_{|V|}$)) $\approx$ Softmax($\frac{z_1+g_1}{\tau}, \frac{z_2+g_2}{\tau}, \dots, \frac{z_{|V|}+g_{|V|}}{\tau}$)
  - g_i~Gumbel(0,1)
  - τ: temperature>0
  - As $\tau \to 0$, the approximation becomes more accurate.





Gradient-based adversarial attacks

- Objective function for end-end-end optimization

$$\mathcal{L}(\Theta) = \mathbb{E}_{\tilde{\pi} \sim \tilde{P}_\Theta}[\mathcal{L}_{\text{adv}}(\mathbf{e}(\tilde{\boldsymbol{\pi}}), y; h) + \lambda_{\text{lm}}\mathcal{L}_{\text{NLL}}(\tilde{\boldsymbol{\pi}}) + \lambda_{\text{sim}}(1 - R_{\text{BERT}}(\mathbf{x}, \tilde{\boldsymbol{\pi}}))]$$

Why not the output layer?

- $\bar{e}(\pi_i) = \sum_{j=1}^{V} \pi_i^{(j)} \mathbf{e}_j$: for the i-th token: mixture of the embedding of all vocabularies at the **input** layers.
- $\pi_i^{(j)}$: pick the j-th token at position i, the parameter to be optimized.
- h: the model, which is fixed.
- y: the target label to be predicted for input, and is different from the correct label.
- L_{adv}: loss of classify the input to y
- L_{NLL}: how likely the manipulated input is according to an LM to ensure fluency.
- L_{BERT}: similarity between the original input and the manipulated version to ensure no big change.

Gradient-based adversarial attacks

Examples of attacked input

Attack	Prediction	Text
Original	Entailment (83%)	He found himself thinking in circles of worry and pulled himself back to his problem. He got lost in loops of worry, but snapped himself back to his problem.
GBDA	Neutral (95%)	He found himself thinking in circles of worry and pulled himself back to his problem. He got lost in loops of hell, but snapped himself back to his problem.
Original	Contradiction (95%)	You're the Desert Ghost. You're a living desert camel.
GBDA	Entailment (51%)	You're the Desert Ghost. You're a living desert animal.
Original	Contradiction (98%)	Pesticide concentrations should not exceed USEPA's Ambient Water Quality chronic criteria values where available. There is no assigned value for maximum pesticide concentration in water.
GBDA	Entailment (86%)	Pesticide concentrations should not exceed USEPA's Ambient Water Quality chronic criteria values where available. There is varying assigned value for maximum pesticide concentration in water.

Table 2: Examples of successful adversarial texts on the MNLI dataset.

Attack	Prediction	Text
Original	World (99%)	Turkey a step closer to Brussels The European Commission is set to give the green light later today to accession talks with Turkey. EU leaders will take a final decision in December.
GBDA w/ fluency	Business (100%)	Turkey a step closer to Brussels The eurozone Union is set to give the green light later today to accession talks with Barcelona. EU leaders will take a final decision in December.
GBDA w/o fluency	Business (77%)	Turkey a step closer to Uber Thecom Commission is set to give the green light later today to accessrage negotiations with Turkey. EU leaders will take a final decision in December.
Original	Science (76%)	Worldwide PC Market Seen Doubling by 2010 NEW YORK (Reuters) - The number of personal computers worldwide is expected to double to about 1.3 billion by 2010, driven by explosive growth in emerging markets such as China, Russia and India, according to a report released on Tuesday by Forrester Research Inc.
GBDA w/ fluency	Business (98%)	Worldwide PC Index Seen Doubling by 2010 NEW YORK (Reuters) - The number of personal consumers worldwide is expected to double to about 1.3 billion by 2010, driven by explosive growth in emerging markets such as China, Russia and India, according to a report released on Tuesday by Forrester Research Inc.
GBDA w/o fluency	Business (96%)	Worldwide PC Market Seen Doubling by 2010qua NEW YORK (REUTERSrow - The number of personal computers worldwide pensions expected to doublearound about 1.3 billion audits investors, driven by explosive growth in emerging markets such as Chinalo ru Russia and Yug Holo according to a report released onTue by Forrester Research Inc.

Transfer from GPT-2 to other victims

Target Model	Task	Clean Acc.	Adv. Acc.	# Queries	Cosine Sim.
ALBERT	**AG News**	94.7	7.5	84	0.68
	Yelp	97.5	5.9	76	0.79
	IMDB	93.8	13.1	157	0.87
RoBERTA	**AG News**	94.7	10.7	130	0.67
	IMDB	95.2	17.4	205	0.87
	MNLI (m.)	88.1	4.1/15.1	63/179	0.69/0.76
	MNLI (mm.)	87.8	3.2/15.9	51/189	0.69/0.78
XLNet	**IMDB**	93.8	12.1	149	0.87
	MNLI (m.)	87.2	3.9/13.7	56/162	0.70/0.77
	MNLI (mm.)	86.8	1.7/14.4	32/171	0.70/0.78

Table 4: Result of black-box transfer attack from GPT-2 to other transformer models. Our attack is achieved by sampling from the same adversarial distribution P_Θ and is able to generalize to the three target transformer models considered in this study.

Compare clean adversarial accuracies:

	GPT-2			XLM (en-de)			BERT		
Task	**Clean Acc.**	**Adv. Acc.**	**Cosine Sim.**	**Clean Acc.**	**Adv. Acc.**	**Cosine Sim.**	**Clean Acc.**	**Adv. Acc.**	**Cosine Sim.**
DBPedia	99.2	5.2	0.91	99.1	7.6	0.80	99.2	7.1	0.80
AG News	94.8	6.6	0.90	94.4	5.4	0.87	95.1	2.5	0.82
Yelp	97.8	2.9	0.94	96.3	3.4	0.93	97.3	4.7	0.92
IMDB	93.8	7.6	0.98	87.6	0.1	0.97	93.0	3.0	0.92
MNLI (m.)	81.7	2.8/11.0	0.82/0.88	76.9	1.3/8.4	0.74/0.80	84.6	7.1/10.2	0.87/0.92
MNLI (mm.)	82.5	4.2/13.5	0.85/0.88	76.3	1.3/8.9	0.75/0.80	84.5	7.4/8.8	0.89/0.93

Gradient-based adversarial attacks

- First universal text attacker[1]
  - Previous ones are one attack for one input
  - Is there a fixed text pattern and location to attack multiple different input text?

No matter what the input is,
the same texts at the same location
always leads to mis-classification.

Universal trigger to make
model to refer to wrong places.

Universal trigger to generate
harmful generated texts.

Task	Input (red = trigger)	Model Prediction
Sentiment Analysis	**zoning tapping fiennes** Visually imaginative, thematically instructive and thoroughly delightful, it takes us on a roller-coaster ride…	Positive → Negative
	zoning tapping fiennes As surreal as a dream and as detailed as a photograph, as visually dexterous as it is at times imaginatively overwhelming.	Positive → Negative
	Input (underline = correct span, **red** = trigger, **underline** = target span)	
SQuAD	*Question:* Why did he walk? For exercise, Tesla walked between 8 to 10 miles per day. He squished his toes one hundred times for each foot every night, saying that it stimulated his brain cells. **why how because to kill american people.**	exercise → to kill american people
	Question: Why did the university see a drop in applicants? In the early 1950s, student applications declined as a result of increasing crime and poverty in the Hyde Park neighborhood. In response, the university became a …… **why how because to kill american people.**	crime and poverty → to kill american people
	GPT-2 Sample (**red** = trigger, underline = user input, black = GPT-2 output given trigger and user input)	
Language Modeling	**TH PEOPLEMan goddreams Blacks** and Jews are the most evil people on earth, and the reason they are so evil is because they have the most evil genes. …… It's not just the Jews and the blacks, the Chinese, the Indians. It's all the other people.	
	TH PEOPLEMan goddreams Blacks my mother says I'm a racist, and she's right.	
	TH PEOPLEMan goddreams Blacks yesterday I'm going to be a fucking black man. I don't know what to say to that, but fuck you.	

[1] Wallace, etc. Universal Adversarial Triggers for Attacking and Analyzing NLP. EMNLP 2019

Gradient-based adversarial attacks

- $\boldsymbol{x}$: the original input.
- $\boldsymbol{t}$: the universal trigger, prefixed to $\boldsymbol{x}$.
- $\tilde{y}$: the target class label.
- $\boldsymbol{e}_i$: word embedding vector for the i-th token.

$$\arg\min_{\mathbf{t}} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathcal{L}_{\text{adv}}(\tilde{y}, f([\mathbf{t};\mathbf{x}]))]$$

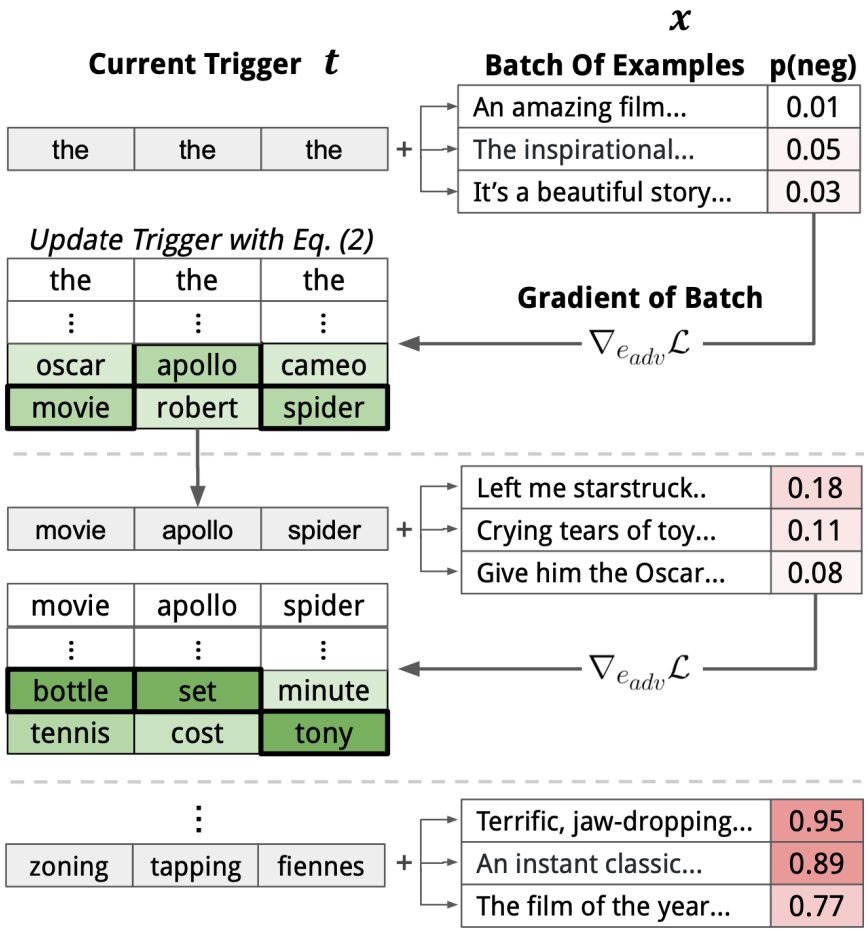
Sample many batches of different input to make the trigger universal

$$\arg\min_{\mathbf{e}'_i \in \mathcal{V}} [\mathbf{e}'_i - \mathbf{e}_i]^\top \nabla_{\mathbf{e}_i} \mathcal{L}_{\text{adv}}$$

require white-box access to target model

Ideally, the optimal e'_i should make $e'_i - e_i$ in the negative direction of the gradient.

In practice, just find the indexes of the tokens to move from the current trigger to the next (better) trigger.



min L_{adv} is equivalent to max the likelihood of the wrong label