Slides: **http://tiny.cc/adversarial**

Adversarial Examples in NLP

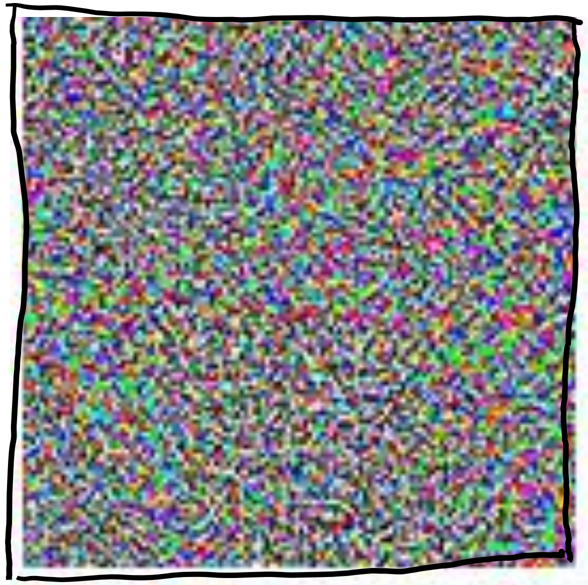
Sameer Singh
sameer@uci.edu
@sameer_
sameersingh.org

What are Adversarial Examples?

“panda”

57.7% confidence

$+\epsilon$



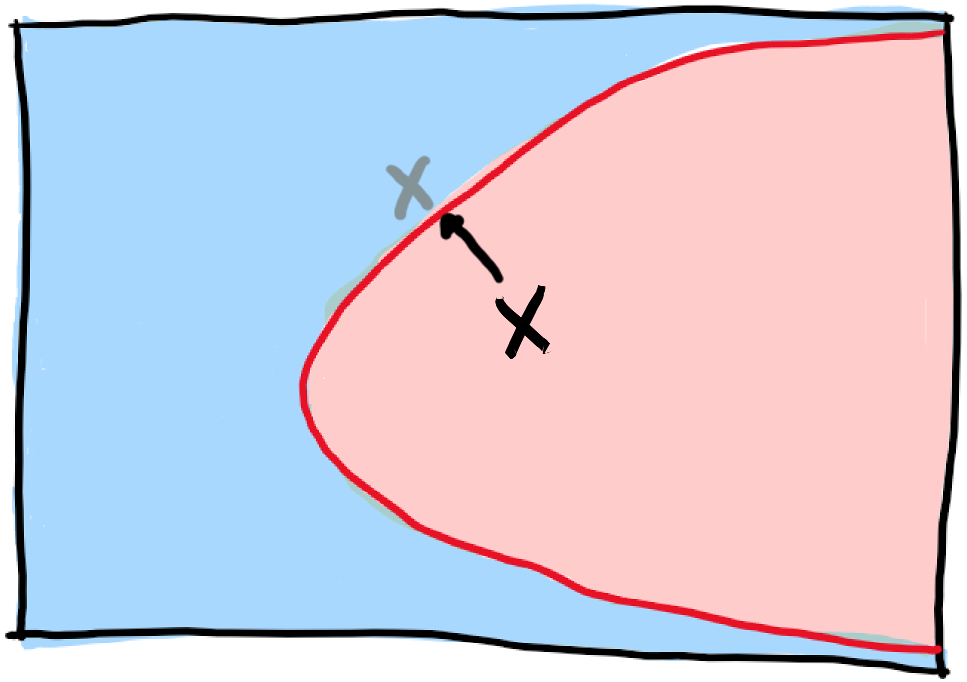
$=$

“gibbon”

99.3% confidence

[Goodfellow et al, ICLR 2015]

What's going on?



$$\min_{x'} \|x - x'\|$$

$$\text{s.t. } f(x') \neq f(x)$$

Fast Gradient Sign Method

$$x' \leftarrow x + \epsilon \, \text{sign}(\nabla_x J(x))$$

[Goodfellow et al, ICLR 2015]

Applications of Adversarial Attacks

- Security of ML Models
  - Should I deploy or not? What’s the worst that can happen?
- Evaluation of ML Models
  - Held-out test error is not enough
- Finding Bugs in ML Models
  - What kinds of “adversaries” might happen naturally?
  - (Even without any bad actors)
- Interpretability of ML Models?
  - What does the model care about, and what does it ignore?

Challenges in NLP

Change
L_2 is not really defined for text
What is imperceivable? What is a small vs big change?
What is the right way to measure this?

$$\min_{x'} \|x - x'\|$$
$$\text{s.t. } f(x') \neq f(x)$$

Search
Text is discrete,
cannot use continuous optimization
How do we search over sequences?

Effect
Classification tasks fit in well, but …
What about structured prediction? e.g. sequence labeling
Language generation? e.g. MT or summarization

Choices in Crafting Adversaries

Different ways to address the challenges

Choices in Crafting Adversaries

What is a small change?

How do we find the attack?

$$\min_{x'} \|x - x'\| \quad \text{s.t. } f(x') \neq f(x)$$

What does it mean to misbehave?

Choices in Crafting Adversaries

What is a small change?

$$\min_{x'} \boxed{\|x - x'\|}$$

$$\text{s.t. } f(x') \neq f(x)$$

Change: What is a small change?

$\boxed{||x - x'||}$

Characters
Pros:
- Often easy to miss
- Easier to search over

Cons:
- Gibberish, nonsensical words
- No useful for interpretability

Words
Pros:
- Always from vocabulary
- Often easy to miss

Cons:
- Ungrammatical changes
- Meaning also changes

Phrase/Sentence
Pros:
- Most natural/human-like
- Test long-distance effects

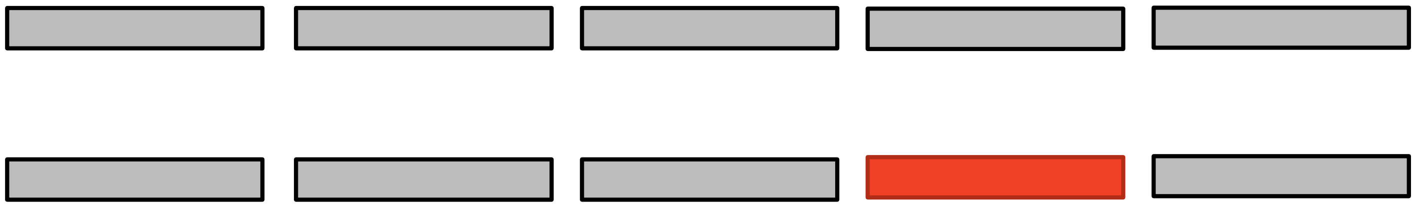
Cons:
- Difficult to guarantee quality
- Larger space to search

Main Challenge: Defining the distance between x and x'

Change: A Character (or few)

x = [“I love movies”]

x = [‘I’ ‘ ’ ‘l’ ‘o’ ‘v’ …



x' = [‘I’ ‘ ’ ‘l’ ‘i’ ‘v’ …

past $\rightarrow$ pas!t	Alps $\rightarrow$ llps	talk $\rightarrow$ taln	local $\rightarrow$ loral

Edit Distance: Flip, Insert, Delete

[Ebrahimi et al, ACL 2018, COLING 2018]

Change: Word-level Changes

	x = [	'I '	'like'	'this'	'movie'	' .']
			Let's replace this word			
Random word?	x' = [	'I '	'lamp'	'this'	'movie'	' .']
Word Embedding?	x' = [	'I '	'really'	'this'	'movie'	' .']
Part of Speech?	x' = [	'I '	'eat'	'this'	'movie'	' .']
Language Model?	x' = [	'I '	'hate'	'this'	'movie'	' .']

[Jia and Liang, EMNLP 2017]

[Alzantot et. al. EMNLP 2018]

Change: Paraphrasing via Backtranslation

x, x' should mean the same thing (*semantically-equivalent adversaries*)

x

This is a good movie → Translate into multiple languages: Este é um bom filme / c'est un bon film → Use back-translators to score candidates

$S(x, x') \propto$
$0.5 * P(x' \mid \text{Este é um bom filme}) +$
$0.5 * P(x' \mid \text{c'est un bon film})$

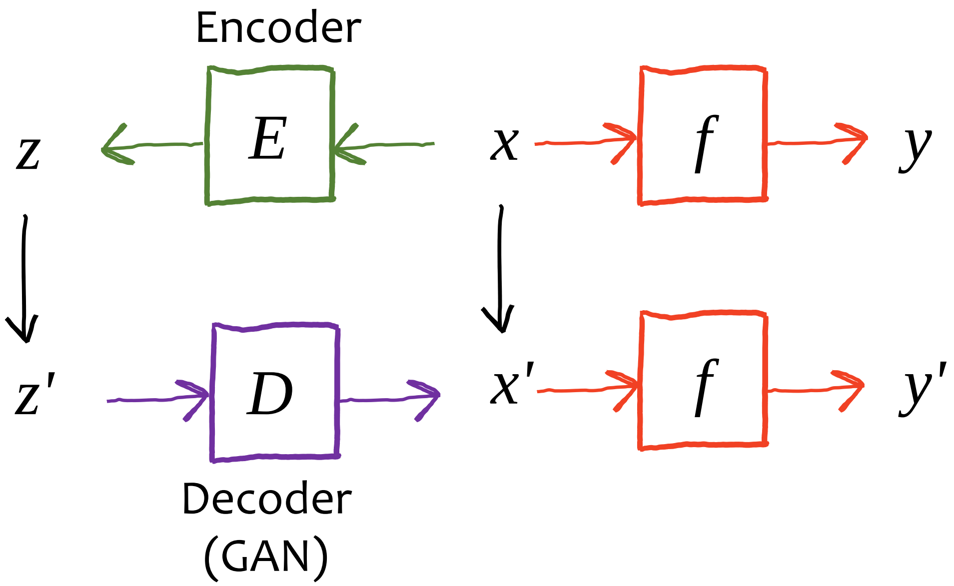
S(This is a good movie , This is a good movie) = 1

S(This is a good movie , That is a good movie) = 0.95

S(This is a good movie , Dogs like cats) = 0

[Ribeiro et al ACL 2018]

Change: Sentence Embeddings



$$\min_{x'} \|z - z'\| \quad \text{s.t. } f(x') \neq f(x)$$

- Deep representations are supposed to encode meaning in vectors
  - If (x-x') is difficult to compute, maybe we can do (z-z')?

[Zhao et al ICLR 2018]

Choices in Crafting Adversaries

$$\min_{x'} \boxed{\|x - x'\|}$$

What is a small change?

$$\text{s.t. } f(x') \neq f(x)$$

Choices in Crafting Adversaries

How do we find the attack?

$$\min_{x'} \|x - x'\|$$

$$\text{s.t. } f(x') \neq f(x)$$

Search: How do we find the attack?

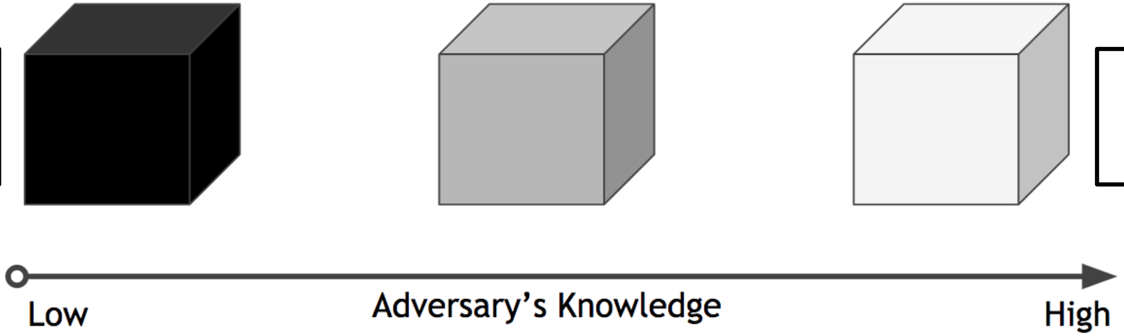
min x'

Even this is often unrealistic

Access probabilities

Only access predictions (usually unlimited queries)

Full access to the model (compute gradients)

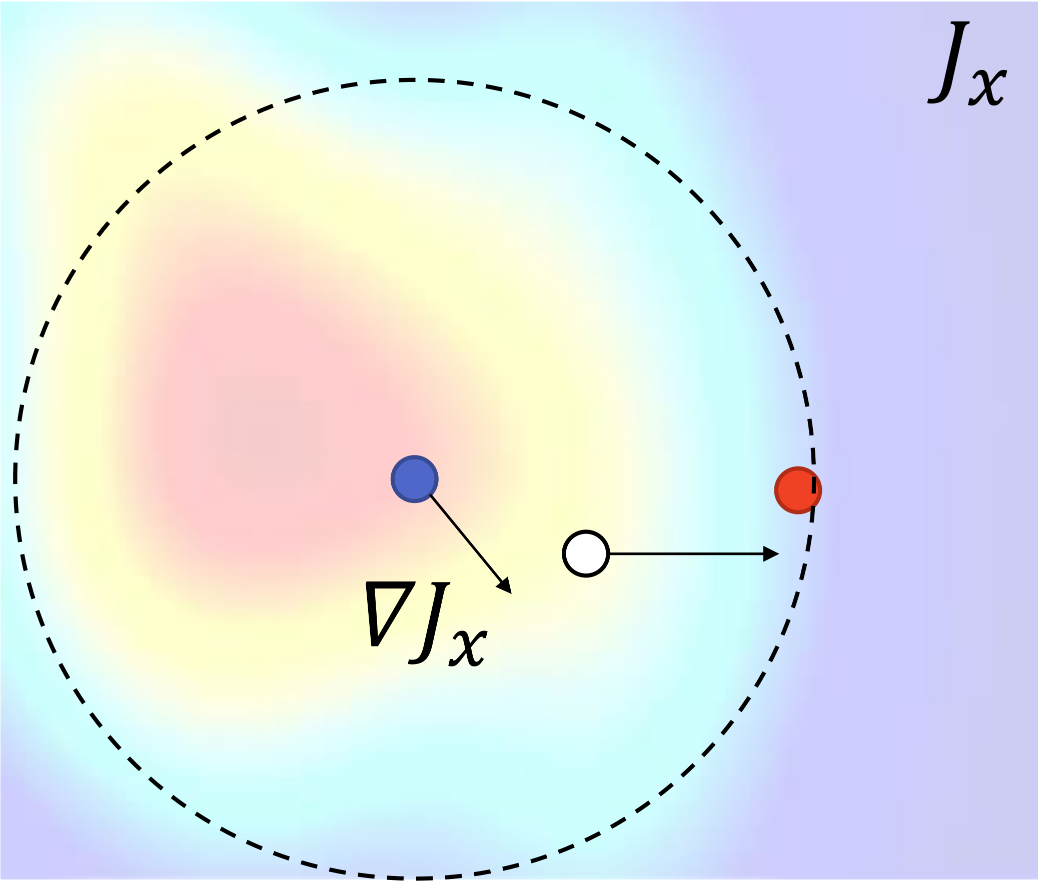


Create x' and test whether the model misbehaves

Create x' and test whether general direction is correct

Use the gradient to *craft* x'

Search: Gradient-based



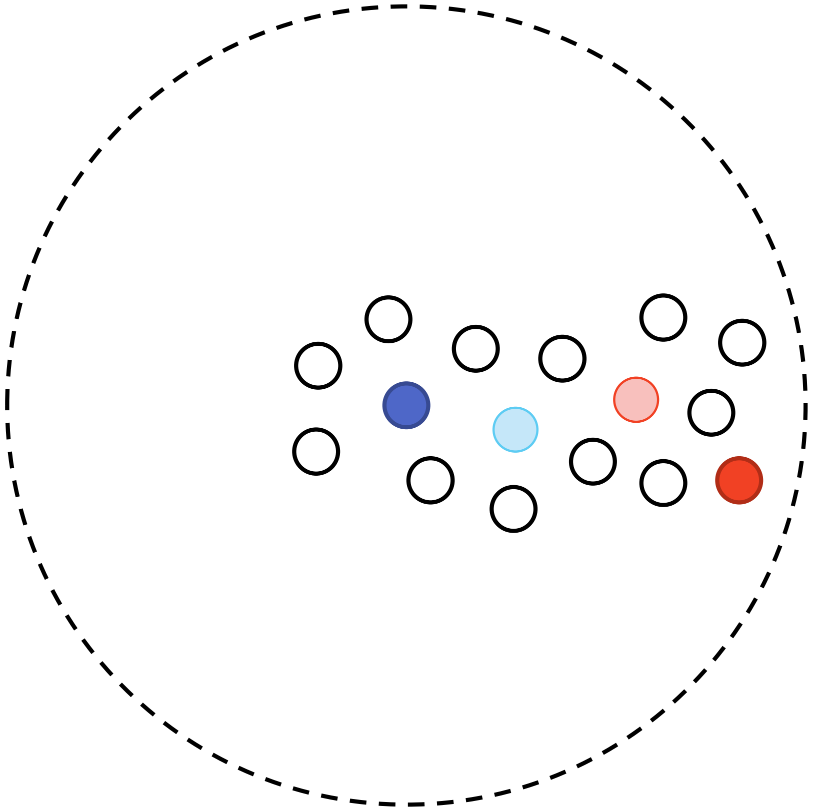
Or whatever the misbehavior is

1. Compute the gradient
2. Step in that direction (continuous)
3. Find the nearest neighbor
4. Repeat if necessary

Beam search over the above…

[Ebrahimi et al, ACL 2018, COLING 2018]

Search: Sampling



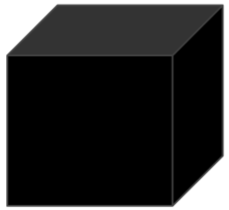
1. Generate local perturbations
2. Select ones that looks good
3. Repeat step 1 with these new ones
4. Optional: beam search, genetic algo

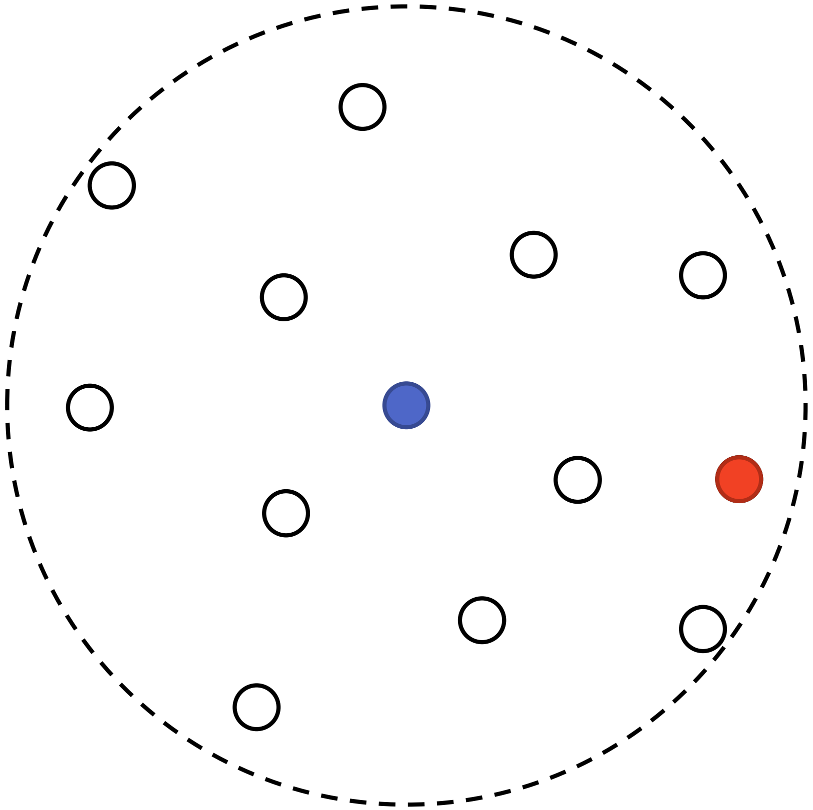
[Jia and Liang, EMNLP 2017]

[Zhao et al, ICLR 2018]

[Alzantot et. al. EMNLP 2018]

Search: Enumeration (Trial/Error)





1. Make some perturbations
2. See if they work
3. Optional: pick the best one

[Iyyer et al, NAACL 2018]

[Ribeiro et al, ACL 2018]

[Belinkov, Bisk, ICLR 2018]

Choices in Crafting Adversaries

How do we find the attack?

$$\min_{x'} \|x - x'\|$$

$$\text{s.t. } f(x') \neq f(x)$$

Choices in Crafting Adversaries

$$\min_{x'} \|x - x'\|$$

$$\boxed{\text{s.t. } f(x') \neq f(x)}$$

What does it mean to misbehave?

Effect: What does it mean to misbehave?

Classification

Untargeted: any other class

Targeted: specific other class

$\boxed{\text{s.t. } f(x') \neq f(x)}$

Other Tasks

MT: Don't attack me! ⟶ ¡No me ataques!

NER: Sameer PERSON is a prof at UCI ORG !

Loss-based: Maximize the loss on the example
e.g. perplexity/log-loss of the prediction

Property-based: Test whether a property holds
e.g. MT: A certain word is not generated
NER: No PERSON appears in the output

Evaluation: Are the attacks “good”?

- Are they Effective?
  - Attack/Success rate
- Are the Changes Perceivable? (Human Evaluation)
  - Would it have the same label?
  - Does it look natural?
  - Does it mean the same thing?
- Do they help improve the model?
  - Accuracy after data augmentation
- Look at some examples!

Review of the Choices

$$\min_{x'} \|x - x'\| \quad \text{s.t. } f(x') \neq f(x)$$

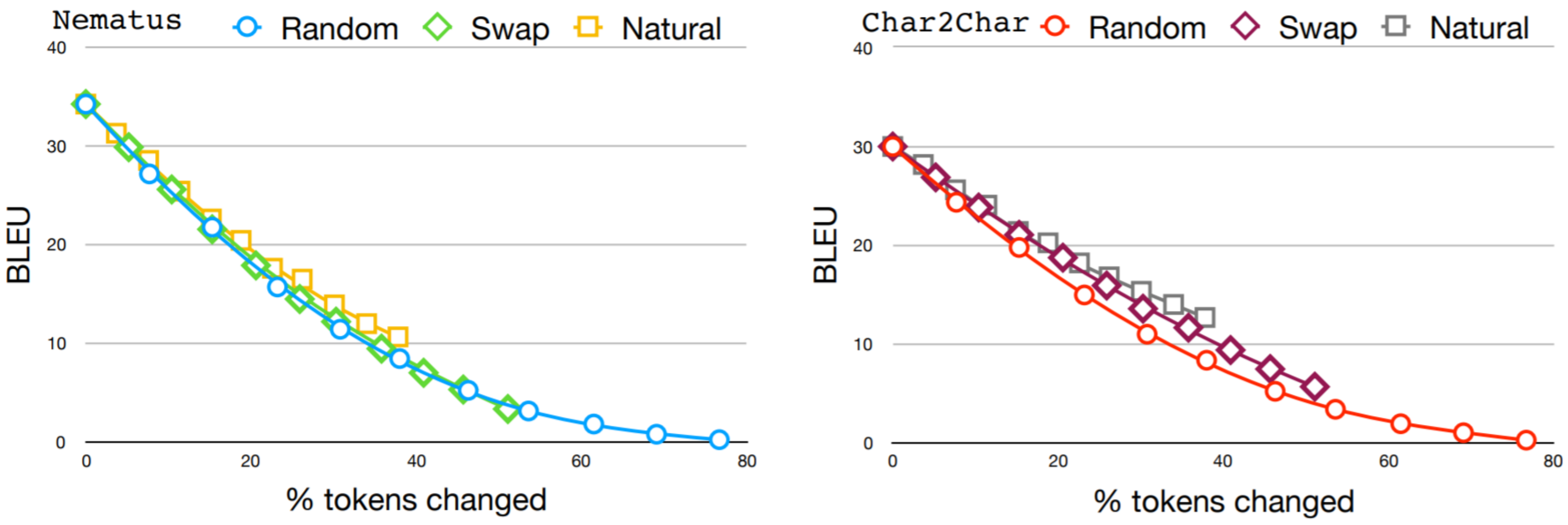
- Effect
  - Targeted or Untargeted
  - Choose based on the task
- Search
  - Gradient-based
  - Sampling
  - Enumeration
- Evaluation
- Change
  - Character level
  - Word level
  - Phrase/Sentence level

Research Highlights

In terms of the choices that were made

Noise Breaks Machine Translation!

Change	Search	Tasks
Random Character Based	Passive; add and test	Machine Translation



[Belinkov, Bisk, ICLR 2018]

Hotflip

Change	Search	Tasks
Character-based (extension to words)	Gradient-based; beam-search	Machine Translation, Classification, Sentiment

News Classification

South Africa's historic Soweto township marks its 100th birthday on Tuesday in a mood of optimism. 57% **World**
South Africa's historic Soweto township marks its 100th birthday on Tuesday in a moo**P** of optimism. 95% **Sci/Tech**

Machine Translation

src	Das ist Dr. Bob Childs – er ist Geigenbauer und Psychotherapeut.
adv	Das ist Dr. Bob Childs – er ist Geigenbauer und **Psy6hothearpeiut**.
src-output	This is Dr. Bob Childs – he's a wizard maker and a **therapist**'s **therapist**.
adv-output	This is Dr. Bob Childs – he's a brick maker and a **psychopath**.

[Ebrahimi et al, ACL 2018, COLING 2018]

Search Using Genetic Algorithms

Black-box, population-based search of natural adversary

Change	Search	Tasks
Word-based, language model score	Genetic Algorithm	Textual Entailment, Sentiment Analysis

Original Text Prediction: **Entailment** (Confidence = 86%)
Premise: *A runner wearing purple strives for the finish line.* **Hypothesis:** *A **runner** wants to head for the finish line.*
Adversarial Text Prediction: **Contradiction** (Confidence = 43%)
Premise: *A runner wearing purple strives for the finish line.* **Hypothesis:** *A **racer** wants to head for the finish line.*

[Alzantot et. al. EMNLP 2018]

Natural Adversaries

Change	Search	Tasks
Sentence, GAN embedding	Stochastic search	Images, Entailment, Machine Translation

Textual Entailment

Classifiers	**Sentences**	**Label**
Original	**p** : The man wearing blue jean shorts is grilling. **h** : The man is walking his dog.	Contradiction
Embedding	**h′** : The man is walking by the dog.	Contradiction → Entailment

Google Translate
Break through language barriers.

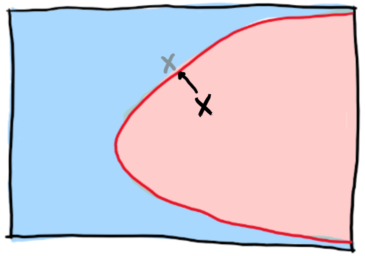
Source Sentence (English)	**Generated Translation (German)**
s : People sitting in a dim restaurant **eating** **s′** : People sitting in a living room **eating** .	Leute, die in einem dim Restaurant **essen** sitzen. Leute, die in einem Wohnzimmeressen sitzen. *(People sitting in a living room)*
s : Elderly people **walking** down a city street . **s′** : A man **walking** down a street playing	Ältere Menschen, die eine Stadtstraße **hinuntergehen**. Ein Mann, der eine Straße entlang spielt. *(A man playing along a street.)*

[Zhao et al, ICLR 2018]

Semantic Adversaries

Change	Search	Tasks
Sentence via Backtranslation	Enumeration	VQA, SQuAD, Sentiment Analysis

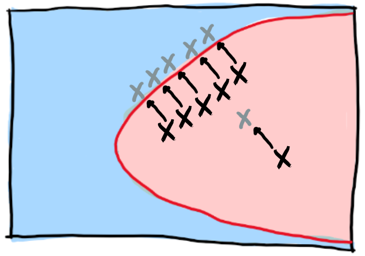
Semantically-Equivalent Adversary (SEA)



x → Backtranslation + Enumeration → x'

What color is the tray?	Pink
What colour is the tray?	Green
Which color is the tray?	Green
What color is it?	Green
How color is tray?	Green

Semantically-Equivalent Adversarial Rules (SEARs)



(x, x') → Patterns in "diffs" → Rules

color → colour

[Ribeiro et al, ACL 2018]

Transformation *Rules*: VisualQA

SEAR	Questions / SEAs	f(x)	Flips
WP VBZ → **WP's**	~~*What has*~~ **What's** been cut?	~~Cake~~ **Pizza**	3.3%
What NOUN → **Which NOUN**	~~*What*~~ **Which** kind of floor is it?	~~Wood~~ **Marble**	3.9%
color → **colour**	What ~~*color*~~ **colour** is the tray?	~~Pink~~ **Green**	2.2%
ADV is → **ADV's**	~~*Where is*~~ **Where's** the jet?	~~Sky~~ **Airport**	2.1%

[Ribeiro et al, ACL 2018]

Transformation *Rules*: SQuAD

SEAR	Questions / SEAs	f(x)	Flips
What VBZ → **What's**	~~*What is*~~ **What's** the NASUWT?	~~Trade union~~ **Teachers in Wales**	2%
What NOUN → **Which NOUN**	~~*What resource*~~ **Which resource** was mined in the Newcastle area?	~~coal~~ **wool**	1%
What VERB → **So what VERB**	~~*What was*~~ **So what was** Ghandi's work called?	~~Satyagraha~~ **Civil Disobedience**	2%
What VBD→ **And what VBD**	~~*What was*~~ **And what was** Kenneth Swezey's job?	~~journalist~~ **sleep**	2%

[Ribeiro et al, ACL 2018]

Transformation *Rules*: Sentiment Analysis

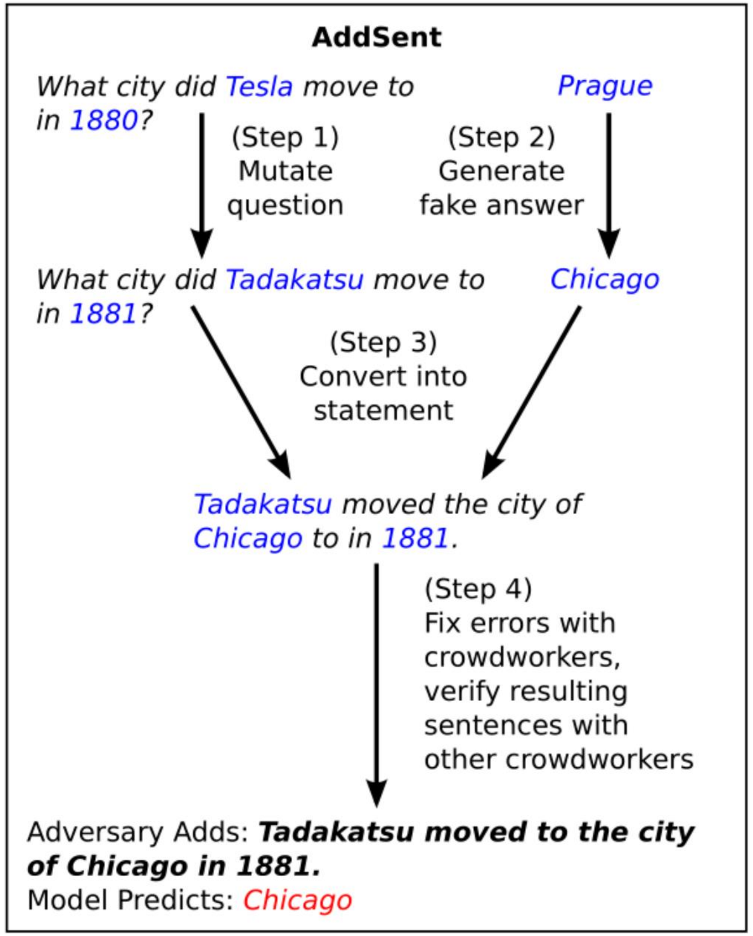
SEAR	Reviews / **SEAs**	f(x)	Flips
movie → **film**	Yeah, the ~~*movie*~~ **film** pretty much sucked .	~~Neg~~ **Pos**	2%
	This is not ~~*movie*~~ **film** making .	~~Neg~~ **Pos**	
film → **movie**	Excellent ~~*film*~~ **movie** .	~~Pos~~ **Neg**	1%
	I'll give this ~~*film*~~ **movie** 10 out of 10 !	~~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**	

[Ribeiro et al, ACL 2018]

Adding a Sentence

Change	Search	Tasks
Add a Sentence	Domain knowledge, stochastic search	Question Answering

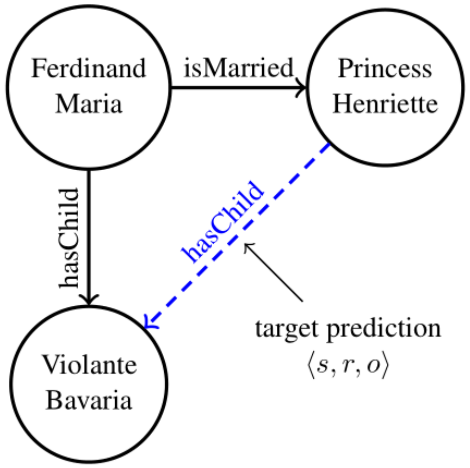
Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

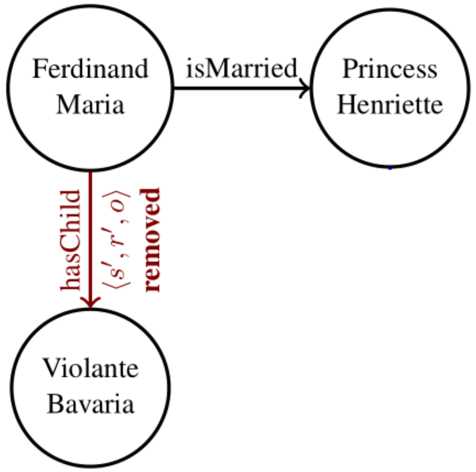


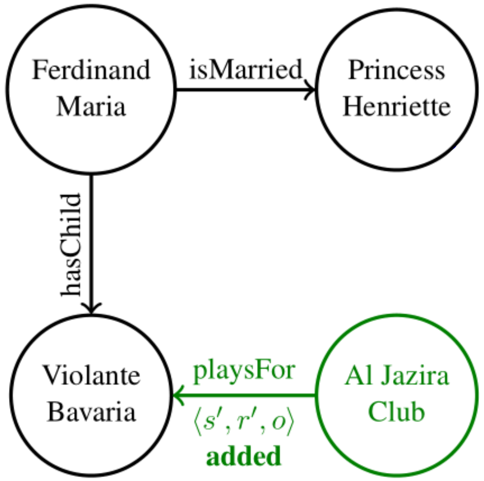
Some Loosely Related Work

Use a broader notions of *adversaries*

CRIAGE: Adversaries for Graph Embeddings







Which link should we add/remove,
out of million possible links?

“Should Not Change” / “Should Change”

How do dialogue systems behave when the inputs are perturbed in specific ways?

Should Not Change

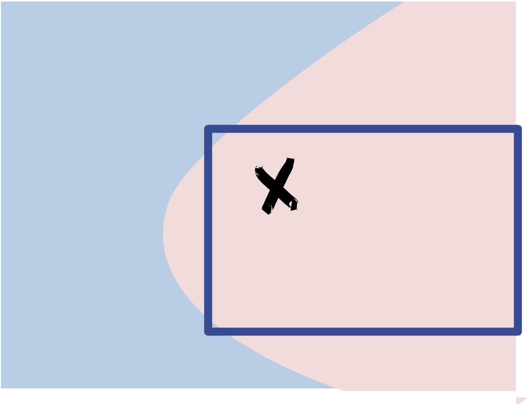
- *like Adversarial Attacks*
- Random Swap
- Stopword Dropout
- Paraphrasing
- Grammatical Mistakes

Should Change

- *Overstability Test*
- Add Negation
- Antonyms
- Randomize Inputs
- Change Entities

[Niu, Bansal, CONLL 2018]

Overstability: Anchors



Anchor

What is the mustache made of?	banana

How **many** bananas are in the picture?	2

Identify the conditions under which the classifier has **the same prediction**

Overstability: Input Reduction

Remove as much of the input as you can *without* changing the prediction!

SQUAD	
Context	In 1899, John Jacob Astor IV invested $100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his Colorado Springs experiments.
Original	What did Tesla spend Astor's money on ?
Reduced	did
Confidence	$0.78 \rightarrow 0.91$

SNLI	
Premise	Well dressed man and woman dancing in the street
Original	Two man is dancing on the street
Reduced	dancing
Answer	Contradiction
Confidence	$0.977 \rightarrow 0.706$

VQA

Original	What color is the flower ?
Reduced	flower ?
Answer	yellow
Confidence	$0.827 \rightarrow 0.819$

[Feng et al, EMNLP 2018]

Adversarial Examples for NLP

- Imperceivable changes to the input
- Unexpected behavior for the output
- Applications: security, evaluation, debugging

$$\min_{x'} \boxed{\|x - x'\|}$$

$$\boxed{\text{s.t. } f(x') \neq f(x)}$$

Challenges for NLP

- Effect: What is misbehavior?
- Change: What is a small change?
- Search: How do we find them?
- Evaluation: How do we know it's good?

Future Directions

- More realistic threat models
  - Give even less access to the model/data
- Defenses and fixes
  - Spell-check based filtering
  - Attack recognition: [Pruthi et al ACL 2019]
  - Data augmentation
  - Novel losses, e.g. [Zhang, Liang AISTATS 2019]
- Beyond sentences
  - Paragraphs, documents?
  - Semantic equivalency → coherency across sentences

References for Adversarial Examples in NLP

Relevant Work (roughly chronological)

- Sentences to QA: [Jia and Liang, EMNLP 2017] link
- Noise Breaks MT: [Belinkov, Bisk, ICLR 2018] link
- Natural Adversaries: [Zhao et al, ICLR 2018] link
- Syntactic Paraphrases: [Iyyer et al NAACL 2018] link
- Hotflip/Hotflip MT: [Ebrahimi et al, ACL 2018, COLING 2018] link, link
- SEARs: [Ribeiro et al, ACL 2018] link
- Genetic Algo: [Alzantot et. al. EMNLP 2018] link
- Discrete Attacks: [Lei et al SysML 2019] link

Surveys

- Adversarial Attacks: [Zhang et al, arXiv 2019] link
- Analysis Methods: [Belinkov, Glass, TAACL 2019] link

More Loosely Related Work

- Anchors: [Ribeiro et al, AAAI 2018] link
- Input Reduction: [Feng et al, EMNLP 2018] link
- Graph Embeddings: [Pezeshkpour et. al. NAACL ‘19] link

Thank you!

Work with **Matt Gardner** and me

as part of

The Allen Institute for
Artificial Intelligence
in **Irvine**, CA

All levels: pre-docs, PhD interns, postdocs, and research scientists!

Sameer Singh
sameer@uci.edu
@sameer_
Sameersingh.org

UCI nlp