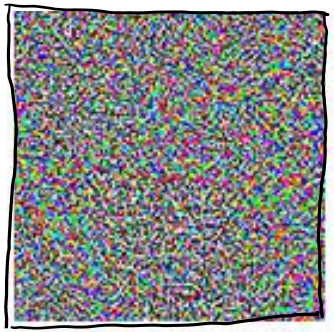
Adversarial Attacks in NLP

“panda”

57.7% confidence

$+\epsilon$

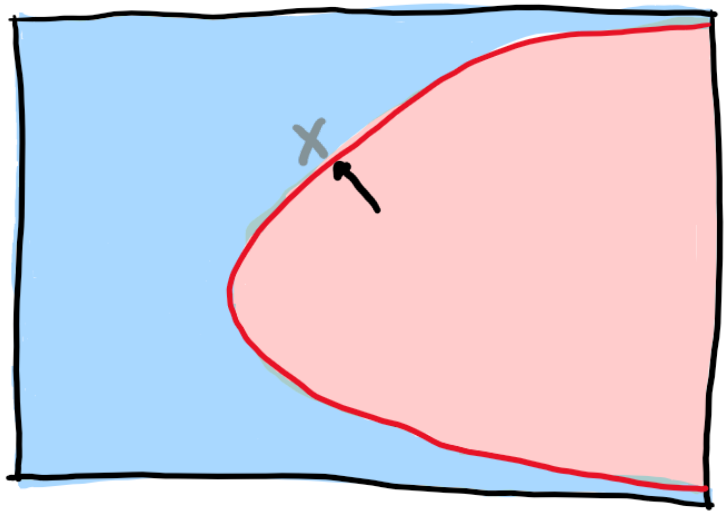


$=$

“gibbon”

99.3% confidence

What's going on?



Fast Gradient Sign Method

[Goodfellow et al, ICLR 2015]

Common Terminology

- Perturbation
  - Noise/rule used to modify a data point
- Data Poisoning
  - Perturbing some data points such that training on them causes a model to fail at test time
- Evasion Attack (now known as Adversarial Examples)
  - Perturbing a test data point such that it causes a model to fail at test time
- Adversarial Reprogramming
  - “Reprogramming” a neural network to do a different task than what it was trained for
  - E.g. Using an ImageNet-trained network to classify MNIST images

Common Terminology

- Threat Model
  - Adversary's knowledge and behavior
- Black-box Attack
  - Adversary has no access to the target model's parameters/gradients
- White-box Attack
  - Adversary has complete access to the model
- Targeted/Untargeted Attack
  - Whether an attack aims to cause the model to predict a specific result or generally mispredict

Common Terminology

- Transferability
  - Ability to generate adversarial data points that work on models the attack was not trained on
- Perceivability
  - For image data: how visible the perturbation is to the human eye
  - Currently undefined for text data

Metrics

- Attack Performance
  - Loss of the target model
  - Task-specific metrics (F1, EM, BLEU, etc)
- Perceivability
  - Images: L2-norm, L∞-norm
  - Text: Semantic equivalence, Naturalness/Acceptability

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?

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?

What does it mean to misbehave?

Choices in Crafting Adversaries

What is a small change?

Change: What is a small change?

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 = [‘l’ ‘ ’ ‘l’ ‘o’ ‘v’ …

x' = [‘l’ ‘ ’ ‘l’ ‘i’ ‘v’ …

past → pas!t	Alps → llps	talk → taln	local → loral

Edit Distance: Flip, Insert, Delete

[Ebrahimi et al, ACL 2018, COLING 2018]

Character-level Attacks

HotFlip: White-Box Adversarial Examples for Text Classification (ACL 2018)

- White-box adversary
- Task: Sentiment Analysis
- Target: CharCNN-LSTM (or any character/word-based classifier)

- Main idea: Flip a character(s) to cause misclassification
  - Random character replacement, insertion, deletion
  - Adjacent key

- Use the gradient wrt one-hot input vector to find the change with highest loss
  - Does not work so well for context-dependent embeddings (ELMo, BERT) [AllenNLP Interpret, EMNLP2019]

HotFlip: White-Box Adversarial Examples for Text Classification (ACL 2018)

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**

Chancellor Gordon Brown has sought to quell speculation over who should run the Labour Party and turned the attack on the opposition Conservatives. 75% **World**
Chancellor Gordon Brown has sought to quell speculation over who should run the Labour Party and turned the attack on the o**B**position Conservatives. 94% **Business**

Character-level flips

one hour photo is an intriguing (**interesting**) snapshot of one man and his delusions it's just too bad it doesn't have more flashes of insight.
'enigma' is a good (**terrific**) name for a movie this deliberately obtuse and unapproachable.
an intermittently pleasing (**satisfying**) but mostly routine effort.
an atonal estrogen opera that demonizes feminism while gifting the most sympathetic male of the piece with a nice (**wonderful**) vomit bath at his wedding.
culkin exudes (**infuses**) none of the charm or charisma that might keep a more general audience even vaguely interested in his bratty character.

Word-level flips

SYNTHETIC AND NATURAL NOISE BOTH BREAK NEURAL MACHINE TRANSLATION (ICLR 2018)

Motivation

"Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae."

SYNTHETIC AND NATURAL NOISE BOTH BREAK NEURAL MACHINE TRANSLATION (ICLR 2018)

- Black-box adversaries
- Task: Machine Translation
- Types of adversaries
  - Natural: typos, misspellings, etc
  - Synthetic
    - Swap two letters
    - Randomize order of all letters except first and last
    - Randomize order of all letters
    - Simulate keyboard typos by replacing one letter in each word with adjacent key

SYNTHETIC AND NATURAL NOISE BOTH BREAK NEURAL MACHINE TRANSLATION (ICLR 2018)

Table 4: An example noisy text with human and machine translations.

Input	Luat eienr Stduie der Cambrdige Unievrstit speilt es kenie Rlloe in welcehr Reiehnfogle die Buhcstbaen in eniem Wrot vorkmomen, die eingzie whctige Sahce ist, dsas der ertse und der lettze Buhcstbaen stmimt .
Human	According to a study from Cambridge university, it doesn't matter which order letters in a word are, the only important thing is that the first and the last letter appear in their correct place.
`char2char`	Cambridge Universttte is one of the most important features of the Cambridge Universttten , which is one of the most important features of the Cambridge Universttten .
`Nematus`	Luat eienr Stduie der Cambrant Unievrstilt splashed it kenie Rlloe in welcehr Reiehnfogle the Buhcstbaen in eniem Wred vorkmomen, die eingzie whcene Sahce ist, DSAs der ertse und der lettze Buhcstbaen stmimt .
`charCNN`	According to the <unk> of the Cambridge University , it 's a little bit of crude oil in a little bit of recycling , which is a little bit of a cool cap , which is a little bit of a strong cap , that the fat and the <unk> bites is consistent .

SYNTHETIC AND NATURAL NOISE BOTH BREAK NEURAL MACHINE TRANSLATION (ICLR 2018)

Table 7: Results of charCNN models trained and tested on different noise conditions.

	Train \ Test	Vanilla	Swap	Mid	Rand	Key	Nat	Ave
French	Swap	39.01	**42.56**	33.64	2.72	4.85	16.43	23.20
	Mid	**42.46**	42.19	**42.17**	3.36	6.20	18.22	25.77
	Rand	39.53	39.46	39.13	**39.73**	3.11	16.63	29.60
	Key	38.49	10.56	8.69	1.08	**38.88**	16.86	19.10
	Nat	28.77	12.45	8.39	1.03	6.61	**36.00**	15.54
	Rand + Key	39.23	38.85	38.89	39.13	38.22	18.71	35.51
	Rand + Nat	36.86	38.95	38.44	38.63	6.67	33.89	32.24
	Key + Nat	38.47	17.33	10.54	1.52	38.62	34.66	23.52
	Rand + Key + Nat	36.97	36.92	36.65	36.64	35.25	31.77	**35.70**
German	Swap	32.66	**34.76**	29.03	2.19	4.78	13.37	19.47
	Mid	**34.32**	34.26	**34.27**	3.50	5.08	14.43	20.98
	Rand	33.65	33.44	**33.75**	33.56	3.00	14.47	25.31
	Key	32.87	10.13	8.39	1.16	**33.28**	13.88	16.62
	Nat	25.79	8.20	5.73	0.93	4.80	**34.59**	13.34
	Rand + Key	32.03	31.57	31.32	31.58	31.23	15.59	28.89
	Rand + Nat	32.37	32.40	31.91	32.11	4.77	33.00	27.76
	Key + Nat	30.39	13.51	8.99	1.53	32.23	33.46	20.02
	Rand + Key + Nat	31.29	30.93	30.54	30.04	29.81	31.60	**30.70**
Czech	Swap	**24.22**	**24.90**	18.72	2.72	6.00	9.03	14.27
	Mid	23.81	**24.52**	24.08	3.96	6.34	9.54	15.38
	Rand	23.44	23.31	23.24	**23.47**	3.70	8.10	17.54
	Key	23.15	7.06	6.04	1.56	**22.80**	10.16	11.80
	Nat	18.04	5.36	4.48	1.47	6.71	**21.64**	9.62
	Rand + Key	21.46	20.81	20.90	20.59	19.48	8.72	18.66
	Rand + Nat	20.59	21.56	20.49	20.53	5.89	18.39	17.91
	Key + Nat	19.55	6.59	5.72	1.40	21.31	19.54	12.35
	Rand + Key + Nat	21.30	21.33	20.38	19.94	19.25	18.38	**20.10**

Text Processing Like Humans Do:

Visually Attacking and Shielding NLP Systems (NAACL-HLT 2019)

- Black-box adversaries
- Tasks
  - Grapheme to phoneme
  - POS & Chunking
  - Toxic comment classification
- Types of adversaries (Visual Noise)
  - Description-based character embedding space (characters with similar descriptions)
  - Easy character embedding space (manually selected)
  - Image-based character embedding space (stacked images of characters)
- Remark: Leetspeak used to motivate work but is not actually used

Text Processing Like Humans Do:

Visually Attacking and Shielding NLP Systems (NAACL-HLT 2019)

Input	ICES	DCES	ECES	SELMo
e	e ẹ ė ȩ є ē ę ḛ ë ë	ĕ ɝ ə ɜ ē ě ȩ ɛ è ë	ê	é 跅 ○ 徛 ꞇ ꞓ ç ↙ ㈎ 垺
i	i ¡ ị I l l ḭ ǀ ḭ ḷ	ĭ ѝ ȉ i ī ǐ ɪ ѝ í ĩ	î	í ⁝ 嚫 檯 爗 篙 烷 媄 腉 >竂
A	A A Ạ Aι Ḁ Ἀ Ἁ Ἀι Λ Λ	Ā Ă Ǟ A Å Ǎ Ȁ Ⱥ Á Ä	Â	楞 溷 鼕 呭 8. 敲 澌 ˛ 唋 用

Table 2: Ten nearest neighbors in our different character spaces. 'SELMo' refers to the nearest neighbors of the trained character embeddings in SELMo.

Condition	**Sentences** (Perturbed / Original)
easy-0.8	Ḿř. Ĉôḟḟêê iŝ â ṕřôfêŝsoř âẗ Ĉôḷŭḿbîâ Ĺâẘ Ŝcĥôôḷ . Mr. Coffee is a professor at Columbia Law School .
ICES-0.6	Tṇẽ sẖuŧđown a∣fëçṱə Э,0ð0 wòřkếrs ḁng ᴎllḷ єũt óũṯpuƚ ɓỳ apouṭ 4,3Ẓ0 câṛṣ : The shutdown affects 3,000 workers and will cut output by about 4,320 cars .
DCES-0.8	Tʰe śţōcƙ ɼeco(v)ɝʶèd ṡøɯəẅḩât ᵗö ƒịṇịṡh 1 1/4 ˡòwɇʶ áť 26 1/4 . The stock recovered somewhat to finish 1 1/4 lower at 26 1/4 .

Table 3: Examples of perturbed sentences and underlying originals.

Text Processing Like Humans Do:

Visually Attacking and Shielding NLP Systems (NAACL-HLT 2019)

- Proposed Defenses
  - Visually-informed ELMo
    - Replace ELMo character embeddings with ICES embeddings
  - Adversarial Training
    - Replace clean examples with visually perturbed examples

On Evaluation of Adversarial Perturbations for Sequence-to-Sequence Models (NAACL-HLT 2019)

- Not specified, but code calls a `whitebox.py`
- Task: Machine Translation
- Type of Adversaries (attempt to enforce semantic similarity)
  - CharSwap (essentially same as ICLR 2018 paper, but with constraint that result must be OOV)
  - kNN (replace word with one of 10 nearest neighbours in source embedding space)
- 6-level evaluation scheme for human study
  - Adapted from Semantic Textual Similarity score
  - Very similar to what we proposed at first but we found participants interpreted the wording differently => NOISE!

On Evaluation of Adversarial Perturbations for Sequence-to-Sequence Models (NAACL-HLT 2019)

How would you rate the similarity between the meaning of these two sentences?

0. The meaning is completely different or one of the sentences is meaningless
1. The topic is the same but the meaning is different
2. Some key information is different
3. The key information is the same but the details differ
4. Meaning is essentially equal but some expressions are unnatural
5. Meaning is essentially equal and the two sentences are well-formed English[a]

[a] Or the language of interest.

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]

Word-level Attacks

Semantically Equivalent Adversarial Rules for Debugging NLP models (ICLR 2018)

- Black-box attack
- Tasks
  - Visual QA
  - Question Answering
  - Sentiment Analysis
- Main Idea
  - Find rules that induce misclassification but preserve the semantics of the original sentence
  - Define Semantic Equivalence as $S(x, x') = \min\left(1, \frac{P(x'|x)}{P(x|x)}\right)$

Semantically Equivalent Adversarial Rules for Debugging NLP models (ICLR 2018)

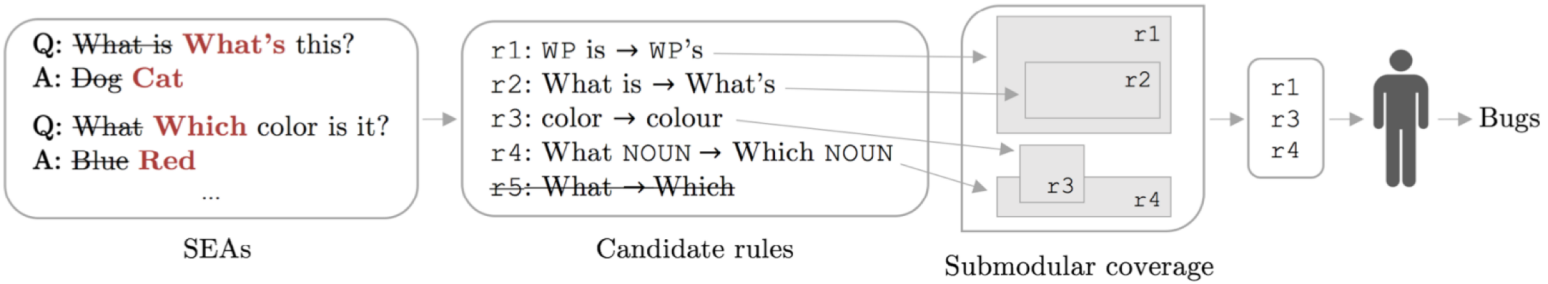


Figure 4: **SEAR process.** (1) SEAs are generalized into candidate rules, (2) rules that are not semantically equivalent are filtered out, e.g. `r5`: (*What→Which*), (3) rules are selected according to Eq (3), in order to maximize coverage and avoid redundancy (e.g. rejecting `r2`, valuing `r1` more highly than `r4`), and (4) a user vets selected rules and keeps the ones that they think are bugs.

Semantically Equivalent Adversarial Rules for Debugging NLP models (ICLR 2018)

SEAR	Questions / SEAs	f(x)	Flips
What VBZ → What's	~~What is~~ What's the NASUWT?	~~Trade unions~~ Teachers in Wales	2%
	~~What is~~ What's a Hauptlied?	~~main hymn~~ Veni redemptor gentium	
What NOUN → Which NOUN	~~What resource~~ Which resource was mined in the Newcastle area?	~~coal~~ wool	1%
	~~What health~~ Which health problem did Tesla have in 1879?	~~nervous breakdown~~ relations	
What VERB → So what VERB	~~What was~~ So what was Ghandi's work called?	~~Satyagraha~~ Civil Disobedience	2%
	~~What is~~ So what is a new trend in teaching?	~~Co-teaching~~ educational institutions	
What VBD → And what VBD	~~What did~~ And what did Tesla develop in 1887?	~~an induction motor~~ laboratory	2%
	~~What was~~ And what was Kenneth Swezey's job?	~~journalist~~ sleep	

Table 1: SEARs for Machine Comprehension

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

Generating Natural Adversarial Examples (ICLR 2018)

- Black-box attack
- Problem Statement
  - Given a black-box classifier ***f*** and a corpus ***X***, generate adversarial example ***x**** for a given data instance ***x*** such that ***f(x*) =/= f(x)***
- Tasks
  - Machine Translation
  - Textual Entailment
- Main Idea
  - Train a adversarial autoencoder and inverter on ***X*** to obtain dense representations ***z*** and ***z'*** while minimising their distance
  - Main problem with text is the discrete nature
    - Perturb ***z'*** instead of ***x,*** then feed ***z'*** to autoencoder to get ***x'***
    - Search in the embedding space for adversaries

Generating Natural Adversarial Examples (ICLR 2018)

Classifiers	Sentences	Label
Original	**p** : The man wearing blue jean shorts is grilling. **h** : The man is walking his dog.	Contradiction
Embedding	$\mathbf{h}'$: The man is walking by the dog.	Contradiction → Entailment
LSTM	$\mathbf{h}'$: The person is walking a dog.	Contradiction → Entailment
TreeLSTM	$\mathbf{h}'$: A man is winning a race.	Contradiction → Neutral

Table 4: **Machine Translation.** "Adversary" that introduces the word "stehen" into the translation.

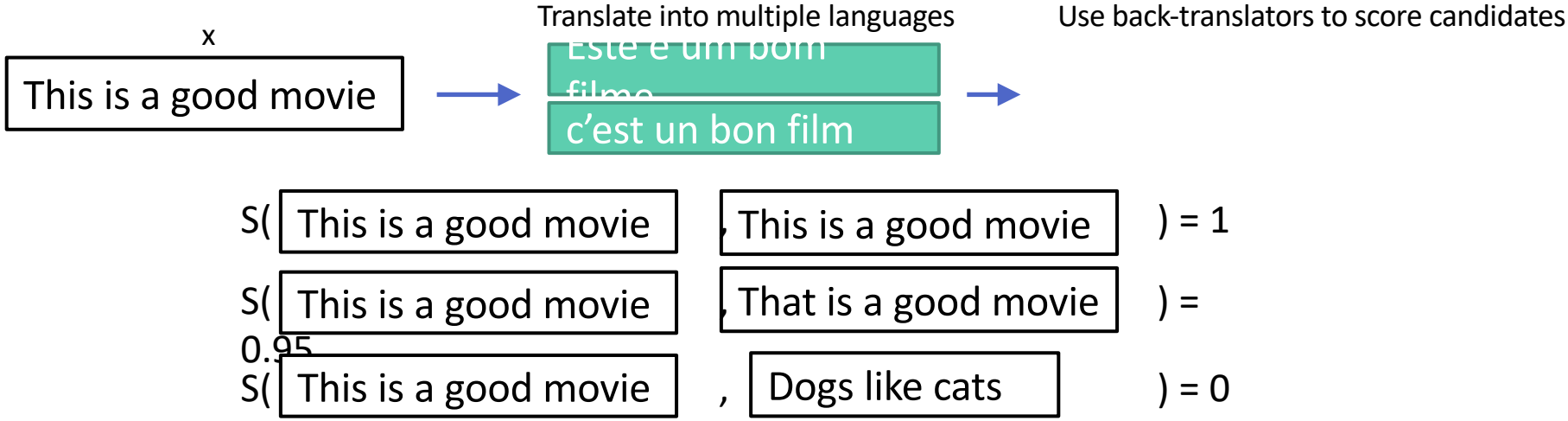
Source Sentence (English)	Generated Translation (German)
s : A man and woman **sitting** on the sidewalk.	Ein Mann und eine Frau, die auf dem Bürgersteig **sitzen**.
$\mathbf{s}'$: A man and woman **stand** on the bench.	Ein Mann und eine Frau **stehen** auf der Bank.

Table 5: **"Adversaries" to find dropped verbs.** The left column contains the original sentence s and its adversary s', while the right contains their translations, with English translation in red.

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

Change: Paraphrasing via Backtranslation

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



[Ribeiro et al ACL 2018]

Sentence-level Attacks

Adversarial Examples for Evaluating Reading Comprehension Systems (EMNLP 2017)

- Black-box attack
- Task: Question Answering
- Models: Match-LSTM, BiDAF
- Main Idea
  - **Add a sentence to the passage in order to cause misclassification**
  - AddAny: Add a sequence of random words
  - AddCommon: Add a sequence of common words
  - AddOneSent: Add a random human-approved sentence
  - AddSent: Construct a sentence using a question from SQuAD

Adversarial Examples for Evaluating Reading Comprehension Systems (EMNLP 2017)

Article: **Nikola Tesla**
Paragraph: "*In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enroll at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses.*"
Question: "*What city did Tesla move to in 1880?*"
Answer: *Prague*
Model Predicts: *Prague*

AddAny

Randomly initialize *d* words:

spring attention income getting reached

↓ Greedily change one word

spring attention income other reached

↓ Repeat many times

Adversary Adds: ***tesla move move other george***
Model Predicts: *george*

AddSent

What city did Tesla move to in 1880? — *Prague*

(Step 1) Mutate question

(Step 2) Generate fake answer

What city did Tadakatsu move to in 1881? — *Chicago*

(Step 3) Convert into statement

Tadakatsu moved the city of Chicago to in 1881.

(Step 4) Fix errors with crowdworkers, verify resulting sentences with other crowdworkers

Adversary Adds: ***Tadakatsu moved to the city of Chicago in 1881.***
Model Predicts: *Chicago*

Adversarial Examples for Evaluating Reading Comprehension Systems (EMNLP 2017)

F1 on SQuAD	Match Single	Match Ens.	BiDAF Single	BiDAF Ens.
Original	71.4	75.4	75.5	80.0
ADDSENT	27.3	29.4	34.3	34.2
ADDONESENT	39.0	41.8	45.7	46.9
ADDANY	7.6	11.7	4.8	2.7
ADDCOMMON	38.9	51.0	41.7	52.6

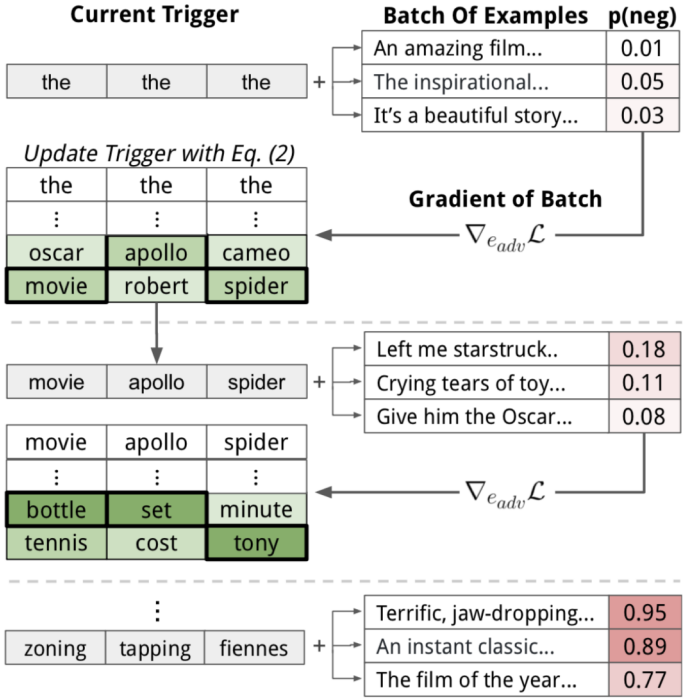
Universal Adversarial Triggers for Attacking and Analyzing NLP (EMNLP 2019)

- White-box for finding adversaries
- Tasks
  - Question Answering
  - Conditional Text Generation
  - Natural Language Inference
  - Sentiment Analysis
- Main Idea
  - Find a sequence of characters that will breaks the model whenever it is added to the input
- Model failure
  - QA: Always predict the trigger regardless of what the question and context is
  - NLI & Sentiment Analysis: Misclassification
  - Text Generation: Generate content similar to a set of targets

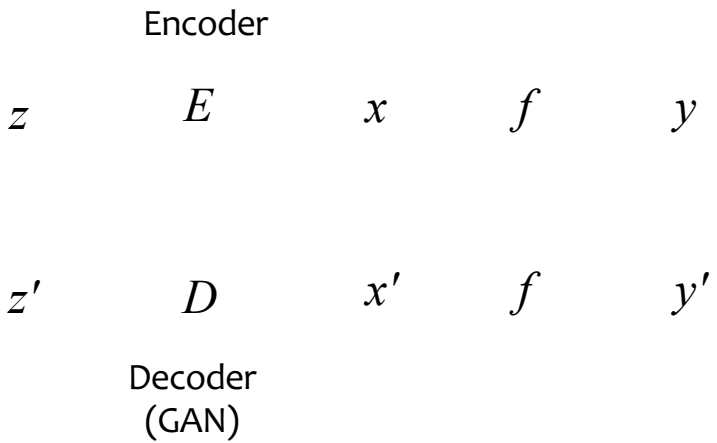
Universal Adversarial Triggers for Attacking and Analyzing NLP (EMNLP 2019)

Objective: Minimize the loss for all inputs from a dataset

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



Change: Sentence Embeddings



- 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

What is a small change?

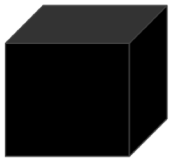
Choices in Crafting Adversaries

How do we find the attack?

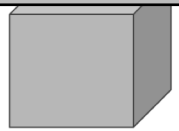
Search: How do we find the attack?

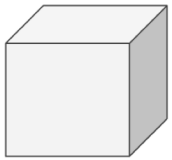
Even this is often unrealistic

Only access predictions (usually unlimited queries)



Access probabilities





Full access to the model (compute gradients)

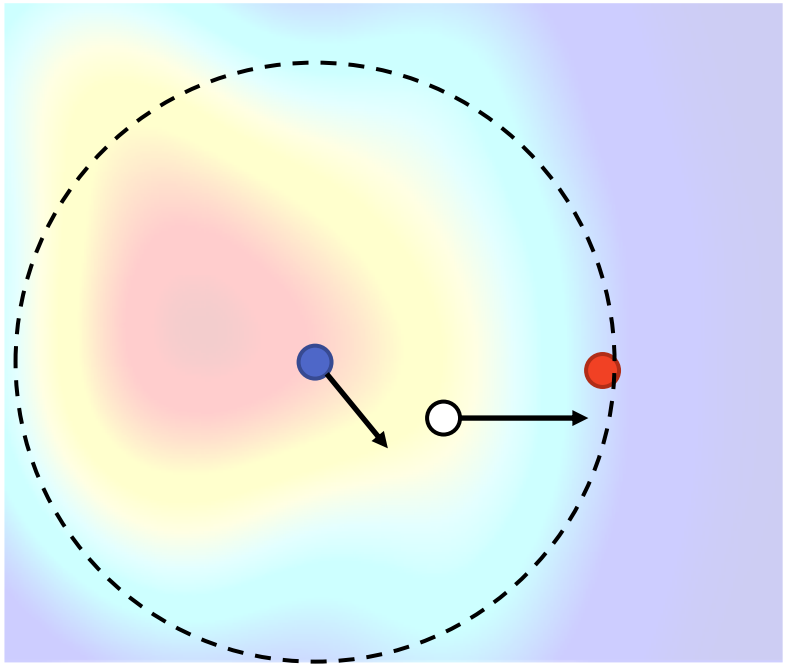
Low — Adversary's Knowledge — High

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



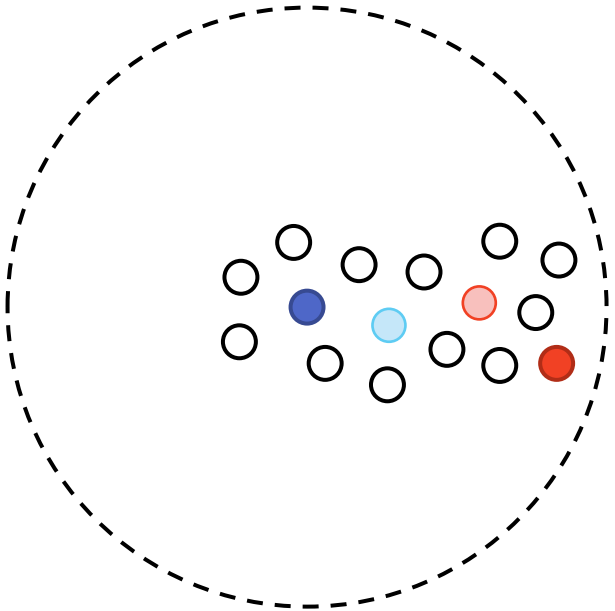
Or whatever the misbehavior is

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



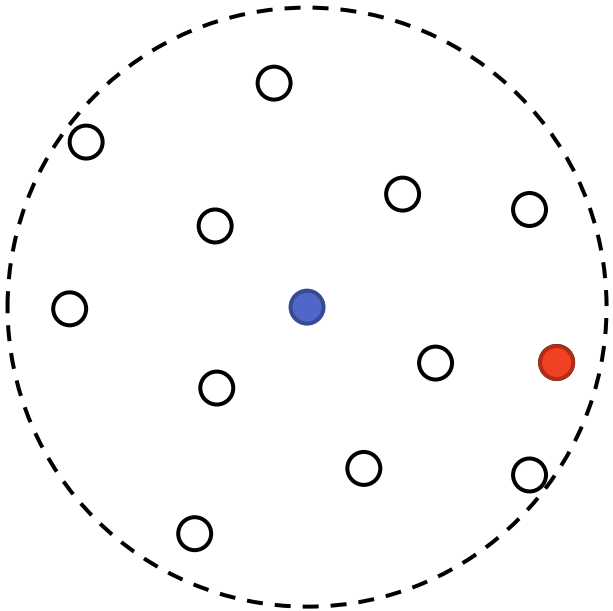
1. Generate local perturbations
2. Select ones that looks good
3. Repeat step 1 with these new ones
4. Optional: beam search, genetic algo

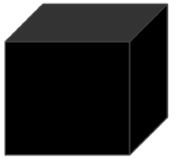
[Jia and Liang, EMNLP 2017]

[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?

Choices in Crafting Adversaries

What does it mean to misbehave?

Effect: What does it mean to misbehave?

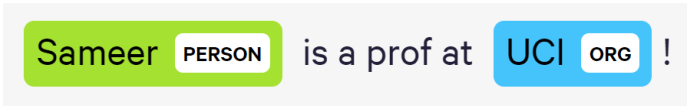
Classification

Untargeted: any other class

Targeted: specific other class

Other Tasks

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

NER: 

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

- Change
  - Character level
  - Word level
  - Phrase/Sentence level

- Effect
  - Targeted or Untargeted
  - Choose based on the task

- Search
  - Gradient-based
  - Sampling
  - Enumeration

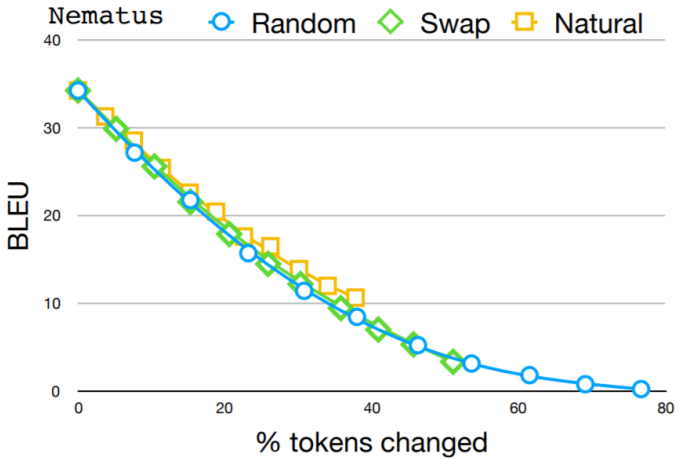
- Evaluation

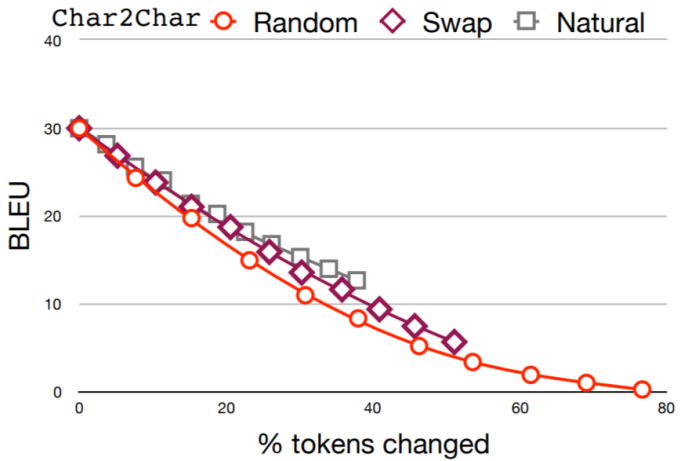
Research Highlights

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	$\mathbf{h}'$: The man is walking by the dog.	Contradiction → Entailment

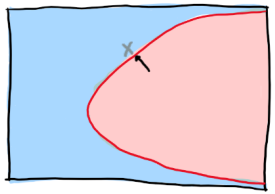
Source Sentence (English)	Generated Translation (German)
s : People sitting in a dim restaurant **eating** $\mathbf{s}'$: People sitting in a living room **eating** .	Leute, die in einem dim Restaurant **essen** sitzen. Leute, die in einem Wohnzimmeressen sitzen. *(People sitting in a living room)*
s : Elderly people **walking** down a city street . $\mathbf{s}'$: 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

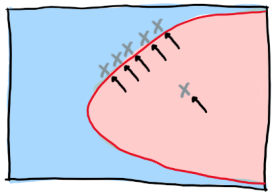
Change	Search	Tasks
Sentence via Backtranslation	Enumeration	VQA, SQuAD, Sentiment Analysis

Semantically-Equivalent Adversary (SEA)



x Backtranslation + Enumeration x'

Semantically-Equivalent Adversarial Rules (SEARs)



(x, x') Patterns in "diffs" Rules

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

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

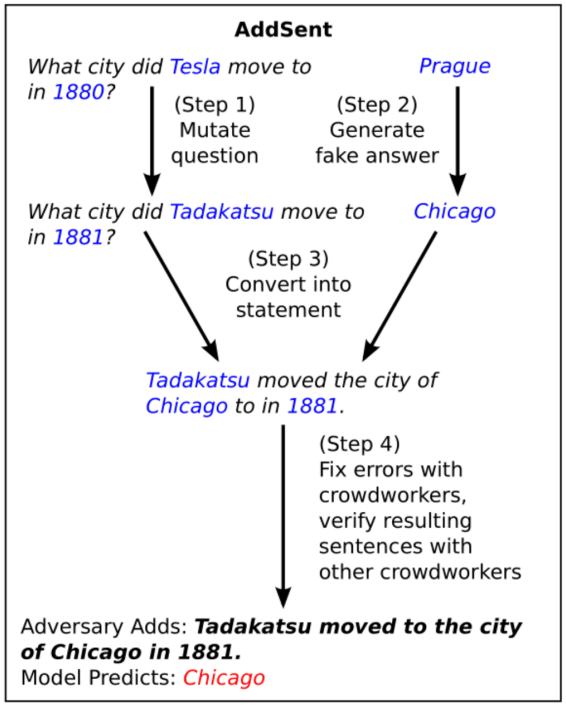
SEAR	Reviews / SEAs	f(x)	Flips
movie → **film**	Yeah, the ~~*movie*~~ **film** pretty much sucked .	~~Neg~~ **Pos**	2%
	This is not ~~*movie*~~ **film** making .	~~Neg~~ **Pos**	
film → **movie**	Excellent ~~*film*~~ **movie** .	~~Pos~~ **Neg**	1%
	I'll give this ~~*film*~~ **movie** 10 out of 10 !	~~Pos~~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



[Jia, Liang, EMNLP 2017]

Adversarial Examples for NLP

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

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?

Outline

- General overview of adversarial machine learning
  - Common terminology
- Character-level attacks
- Word-level attacks
  - Measuring semantic equivalence
- Sentence-level attacks (Question Answering)