Deep Adversarial Learning for NLP

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
<th>Presenter</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 – 10:30</td>
<td>Introduction and Adversarial Training, GANs</td>
<td>William Wang</td>
</tr>
<tr>
<td>10:30 – 11:00</td>
<td>Break</td>
<td>-</td>
</tr>
<tr>
<td>11:00 – 12:15</td>
<td>Adversarial Examples</td>
<td>Sameer Singh</td>
</tr>
<tr>
<td>12:15 – 12:30</td>
<td>Conclusions and Question Answering</td>
<td>William Wang, Sameer Singh</td>
</tr>
</tbody>
</table>

Slides: [http://tiny.cc/adversarial](http://tiny.cc/adversarial)

With contributions from Jiwei Li
Deep Adversarial Learning for NLP

William Wang
UC SANTA BARBARA

Sameer Singh
UC Irvine

Slides: http://tiny.cc/adversarial

With contributions from Jiwei Li.
Agenda

• Introduction, Background, and GANs (William, 90 mins)
• Adversarial Examples and Rules (Sameer, 75 mins)
• Conclusion and Question Answering (Sameer and William, 15 mins)

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

• Background of the Tutorial
• Introduction: Adversarial Learning in NLP
• Adversarial Generation
• A Case Study of GANs in Dialogue Systems
Rise of Adversarial Learning in NLP

• Through a simple ACL anthology search, we found that in 2018, there were 20+ times more papers mentioning “adversarial”, comparing to 2016.

• Meanwhile, the growth of all accepted papers is 1.39 times during this period.

• But if you went to CVPR 2018 in Salt Lake City, there were more than 100 papers on adversarial learning (approximately 1/3 of all adv. learning papers in NLP).
Questions I’d like to Discuss

- What are the subareas of deep adversarial learning in NLP?
- How do we understand adversarial learning?
- What are some success stories?
- What are the pitfalls that we need to avoid?
Opportunities in Adversarial Learning

- Adversarial learning is an interdisciplinary research area, and it is closely related to, but limited to the following fields of study:
  - Machine Learning
  - Computer Vision
  - Natural Language Processing
  - Computer Security
  - Game Theory
  - Economics
Adversarial Attack in ML, Vision, & Security

- Goodfellow et al., (2015)
Physical-World Adversarial Attack / Examples (Eykholt et al., CVPR 2018)
Success of Adversarial Learning

CycleGAN (Zhu et al., 2017)
Failure Cases

CycleGAN (Zhu et al., 2017)
Success of Adversarial Learning

GauGAN (Park et al., 2019)
Deep Adversarial Learning in NLP

• There were some successes of GANs in NLP, but not so much comparing to Vision.

• The scope of Deep Adversarial Learning in NLP includes:
  • Adversarial Examples, Attacks, and Rules
  • Adversarial Training (w. Noise)
  • Adversarial Generation
  • Various other usages in ranking, denoising, & domain adaptation.
Outline

• Background of the Tutorial
• Introduction: Adversarial Learning in NLP
• Adversarial Generation
• A Case Study of GANs in Dialogue Systems
Adversarial Examples

• One of the more popular areas of adversarial learning in NLP.
• E.g., Alzantot et al., EMNLP 2018

<table>
<thead>
<tr>
<th>Original Text Prediction: <strong>Entailment</strong> (Confidence = 86%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Premise:</strong> A runner wearing purple strives for the finish line.</td>
</tr>
<tr>
<td><strong>Hypothesis:</strong> A runner wants to head for the finish line.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Adversarial Text Prediction: <strong>Contradiction</strong> (Confidence = 43%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Premise:</strong> A runner wearing purple strives for the finish line.</td>
</tr>
<tr>
<td><strong>Hypothesis:</strong> A racer wants to head for the finish line.</td>
</tr>
</tbody>
</table>
Adversarial Attacks (Coavoux et al., EMNLP 2018)

The main classifier predicts a label $y$ from a text $x$, the attacker tries to recover some private information $z$ contained in $x$ from the latent representation used by the main classifier.
Adversarial Training

• Main idea:
  • Adding noise, randomness, or adversarial loss in optimization.

• Goal: make the trained model more robust.
Adversarial Training: A Simple Example

• Adversarial Training for Relation Extraction
  • Wu, Bamman, Russell (EMNLP 2017).

• Task: Relation Classification.

• Interpretation: Regularization in the Feature Space.
Adversarial Training for Relation Extraction

\[ L_{\text{adv}}(X; \theta) = L(X + e_{\text{adv}}; \theta), \text{ where} \]

\[ e_{\text{adv}} = \arg \max_{\|e\| \leq \epsilon} L(X + e; \hat{\theta}) \]

\[ e_{\text{adv}} = \epsilon g / \|g\|, \text{ where } g = \nabla_V L(X; \hat{\theta}). \]

Wu, Bamman, Russell (EMNLP 2017).
## Adversarial Training for Relation Extraction

<table>
<thead>
<tr>
<th></th>
<th>Recall 0.1</th>
<th>Recall 0.2</th>
<th>Recall 0.3</th>
<th>Recall 0.4</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCNN</td>
<td>0.667</td>
<td>0.572</td>
<td>0.476</td>
<td>0.392</td>
<td>0.329</td>
</tr>
<tr>
<td>PCNN-Adv</td>
<td>0.717</td>
<td>0.589</td>
<td>0.511</td>
<td>0.407</td>
<td>0.356</td>
</tr>
<tr>
<td>RNN</td>
<td>0.668</td>
<td>0.586</td>
<td>0.524</td>
<td>0.442</td>
<td>0.351</td>
</tr>
<tr>
<td>RNN-Adv</td>
<td><strong>0.728</strong></td>
<td><strong>0.646</strong></td>
<td><strong>0.553</strong></td>
<td><strong>0.481</strong></td>
<td><strong>0.382</strong></td>
</tr>
</tbody>
</table>

Wu, Bamman, Russell (EMNLP 2017).
Outline

• Background of the Tutorial
• Introduction: Adversarial Learning in NLP
• Adversarial Generation
• A Case Study of GANs in Dialogue Systems
GANs (Goodfellow et al., 2014)

- Two competing neural networks: generator & discriminator

Image: https://ishmaelbelghazi.github.io/ALI/
GAN Objective

\[
\min_G \max_D V(D, G) = \mathbb{E}_{q(x)}[\log(D(x))] + \mathbb{E}_{p(z)}[\log(1 - D(G(z)))] \\
= \int q(x) \log(D(x))dx + \iint p(z)p(x \mid z) \log(1 - D(x))dxdz
\]

*\(D(x)\): the probability that \(x\) came from the data rather than generator
GAN Training Algorithm

Algorithm 1 Minibatch stochastic gradient descent training of generative adversarial nets. The number of steps to apply to the discriminator, $k$, is a hyperparameter. We used $k = 1$, the least expensive option, in our experiments.

for number of training iterations do
  for $k$ steps do
    • Sample minibatch of $m$ noise samples $\{z^{(1)}, \ldots, z^{(m)}\}$ from noise prior $p_g(z)$.
    • Sample minibatch of $m$ examples $\{x^{(1)}, \ldots, x^{(m)}\}$ from data generating distribution $p_{\text{data}}(x)$.
    • Update the discriminator by ascending its stochastic gradient:
      \[ \nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[ \log D(x^{(i)}) + \log \left( 1 - D\left( G\left( z^{(i)} \right) \right) \right) \right]. \]
  end for
  • Sample minibatch of $m$ noise samples $\{z^{(1)}, \ldots, z^{(m)}\}$ from noise prior $p_g(z)$.
  • Update the generator by descending its stochastic gradient:
    \[ \nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log \left( 1 - D\left( G\left( z^{(i)} \right) \right) \right). \]
end for

The gradient-based updates can use any standard gradient-based learning rule. We used momentum in our experiments.
GAN Equilibrium

- Global optimality
  - Discriminator
  - Generator

\[
D^*(x) = \frac{q(x)}{q(x) + p(x)}
\]

\[
G^*(z) \text{ s.t. } p(z) = q(x)
\]
Major Issues of GANs

- Mode Collapse (unable to produce diverse samples)
Major Issues of GANs in NLP

• Often you need to pre-train the generator and discriminator w. MLE
  • But how much?

• Unstable Adversarial Training
  • We are dealing with two networks / learners / agents
  • Should we update them at the same rate?

• The discriminator might overpower the generator.

• With many possible combinations of model choice for generator and discriminator networks in NLP, it could be worse.
Major Issues of GANs in NLP

- GANs were originally designed for images
  - You cannot back-propagate through the generated X
- Image is continuous, but text is discrete (DR-GAN, Tran et al., CVPR 2017).
SeqGAN: policy gradient for generating sequences (Yu et al., 2017)
Training Language GANs from Scratch

• New Google DeepMind arxiv paper (de Masson d’Autume et al., 2019)
  • Claims no MLE pre-trainings are needed.
  • Uses per time-stamp dense rewards.
  • Yet to be peer-reviewed and tested.
Why shouldn’t NLP give up on GAN?

• It’s unsupervised learning.
• Many potential applications of GANs in NLP.
• The discriminator is often learning a metric.
• It can also be interpreted as self-supervised learning (especially with dense rewards).
Applications of Adversarial Learning in NLP

• Social Media (Wang et al., 2018a; Carton et al., 2018)
• Contrastive Estimation (Cai and Wang, 2018; Bose et al., 2018)
• Domain Adaptation (Kim et al., 2017; Alam et al., 2018; Zou et al., 2018; Chen and Cardie, 2018; Tran and Nguyen, 2018; Cao et al., 2018; Li et al., 2018b)
• Data Cleaning (Elazar and Goldberg, 2018; Shah et al., 2018; Ryu et al., 2018; Zellers et al., 2018)
• Information extraction (Qin et al., 2018; Hong et al., 2018; Wang et al., 2018b; Shi et al., 2018a; Bekoulis et al., 2018)
• Information retrieval (Li and Cheng, 2018)
• Another 18 papers on Adversarial Learning at NAACL 2019!
GANs for Machine Translation

• Yang et al., NAACL 2018
• Wu et al., ACML 2018
SentiGAN (Wang and Wan, IJCAI 2018)

Idea: use a mixture of generators and a multi-class discriminator.

Figure 1: The framework of SentiGAN with $k$ generators and one multi-class discriminator.
No Metrics Are Perfect: Adversarial Reward Learning (Wang, Chen et al., ACL 2018)
AREL Storytelling Evaluation

• Dataset: VIST (Huang et al., 2016).

Turing Test

Win

Unsure

XE

BLEU-RL

CIDEr-RL

GAN

AREL

-17.5

-13.7

-26.1

-6.3

50%

40%

30%

20%

10%

0%
DSGAN: Adversarial Learning for Distant Supervision IE (Qin et al., ACL 2018)
DSGAN: Adversarial Learning for Distant Supervision IE (Qin et al., ACL 2018)
KKBGAN: Learning to Generate High-Quality Negative Examples (Cai and Wang, NAACL 2018)

Idea: use adversarial learning to iteratively learn better negative examples.
Outline

• Background of the Tutorial
• Introduction: Adversarial Learning in NLP
• Understanding Adversarial Learning
• Adversarial Generation
• A Case Study of GANs in Dialogue Systems
What Should Rewards for Good Dialogue Be Like?
Reward for Good Dialogue

Turing Test
Reward for Good Dialogue

How old are you?

I don’t know what you are talking about

I’m 25.

A human evaluator/judge
How old are you?

I don’t know what you are talking about

I’m 25.

Reward for Good Dialogue
How old are you?

I don’t know what you are talking about.

I’m 25.

P= 90% human generated

P= 10% human generated

Reward for Good Dialogue

P= 90% human generated

I’m 25.
Adversarial Learning in Image Generation (Goodfellow et al., 2014)
Model Breakdown

Generative Model (G)

Encoding

Decoding

how → are → you → ?

I’m → fine → .

eos → I’m → fine → .

EOS
Model Breakdown

Generative Model (G)

Encoding

how → are → you → ?

Decoding

I’m → fine → .

Discriminative Model (D)

P = 90% human generated

how → are → you → ?

eos → I’m → fine → .
Model Breakdown

Generative Model (G)

Encoding

Decoding

Discriminative Model (D)

Reward

P = 90% human generated

I’m fine.
Policy Gradient

Generative Model (G)

Encoding

Decoding

REINFORCE Algorithm (William, 1992)

\[ J = E[R(y)] \]
Adversarial Learning for Neural Dialogue Generation

For number of training iterations do
  For i=1,D-steps do
    Sample (X,Y) from real data
    Sample $\hat{Y} \sim G(\cdot | X)$
    Update $D$ using $(X, Y)$ as positive examples and $(X, \hat{Y})$ as negative examples.
  End
End

For i=1,G-steps do
  Sample (X,Y) from real data
  Sample $\hat{Y} \sim G(\cdot | X)$
  Compute Reward $r$ for $(X, \hat{Y})$ using $D$.
  Update $G$ on $(X, \hat{Y})$ using reward $r$
  Teacher-Forcing: Update $G$ on $(X, Y)$
End

Update the Discriminator

Update the Generator

The discriminator forces the generator to produce correct responses
The previous RL model only perform better on multi-turn conversations
Results: Adversarial Learning Improves Response Generation

vs a vanilla generation model

<table>
<thead>
<tr>
<th>Adversarial Win</th>
<th>Adversarial Lose</th>
<th>Tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>62%</td>
<td>18%</td>
<td>20%</td>
</tr>
</tbody>
</table>
Sample response

Tell me ... how long have you had this falling sickness?
Sample response

Tell me ... how long have you had this falling sickness?

<table>
<thead>
<tr>
<th>System</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla-Seq2Seq</td>
<td>I don’t know what you are talking about.</td>
</tr>
</tbody>
</table>
Sample response

Tell me ... how long have you had this falling sickness?

<table>
<thead>
<tr>
<th>System</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla-Seq2Seq</td>
<td>I don’t know what you are talking about.</td>
</tr>
<tr>
<td>Mutual Information</td>
<td>I’m not a doctor.</td>
</tr>
</tbody>
</table>
Tell me ... how long have you had this falling sickness?

<table>
<thead>
<tr>
<th>System</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla-Sem2Seq</td>
<td>I don’t know what you are talking about.</td>
</tr>
<tr>
<td>Mutual Information</td>
<td>I’m not a doctor.</td>
</tr>
<tr>
<td>Adversarial Learning</td>
<td>A few months, I guess.</td>
</tr>
</tbody>
</table>
Self-Supervised Learning meets Adversarial Learning

• Self-Supervised Dialog Learning (Wu et al., ACL 2019)
• Use of SSL to learn dialogue structure (sequence ordering).
Self-Supervised Learning meets Adversarial Learning

- Self-Supervised Dialog Learning (Wu et al., ACL 2019)
- Use of SSN to learn dialogue structure (sequence ordering).

<table>
<thead>
<tr>
<th>Win</th>
<th>REGS</th>
<th>AEL</th>
<th>SSN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-turn Percentage</td>
<td>.095</td>
<td>.192</td>
<td>.713</td>
</tr>
<tr>
<td>Multi-turn Percentage</td>
<td>.025</td>
<td>.171</td>
<td>.804</td>
</tr>
</tbody>
</table>
Conclusion

• Deep adversarial learning is a new, diverse, and interdisciplinary research area, and it is highly related to many subareas in NLP.

• GANs have obtained particular strong results in Vision, but yet there are both challenges and opportunities in GANs for NLP.

• In a case study, we show that adversarial learning for dialogue has obtained promising results.

• There are plenty of opportunities ahead of us with the current advances of representation learning, reinforcement learning, and self-supervised learning techniques in NLP.
UCSB Postdoctoral Scientist Opportunities

• Please talk to me at NAACL, or email william@cs.ucsb.edu.
Thank you!

• Now we will take an 30 mins break.
Adversarial Examples in NLP

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

Slides: http://tiny.cc/adversarial
What are Adversarial Examples?

“panda”
57.7% confidence

+ ∈

99.3% confidence

“gibbon”

[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

\[x' \leftarrow x + \varepsilon \text{ sign} (\nabla_x J(x))\]

[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
L₂ is not really defined for text
What is imperceivable? What is a small vs big change?
What is the right way to measure this?

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

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

\[
\min_{x'} \| x - x' \|
\text{ s.t. } f(x') \neq f(x)
\]
Choices in Crafting Adversaries

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

What is a small change?
Change: What is a small change?

\[ \| x - x' \| \]

Characters
Pro:
• Often easy to miss
• Easier to search over
Con:
• Gibberish, nonsensical words
• No useful for interpretability

Words
Pro:
• Always from vocabulary
• Often easy to miss
Con:
• Ungrammatical changes
• Meaning also changes

Phrase/Sentence
Pro:
• Most natural/human-like
• Test long-distance effects
Con:
• Difficult to guarantee quality
• Larger space to search

Main Challenge: Defining the distance between $x$ and $x'$
Change: A Character (or few)

\[ x = [ \text{“I love movies”} ] \]

\[ x = [ \text{‘I’} \quad \text{‘ ’} \quad \text{‘I’} \quad \text{‘o’} \quad \text{‘v’} \quad \ldots ] \]

\[ x' = [ \text{‘I’} \quad \text{‘ ’} \quad \text{‘I’} \quad \text{‘i’} \quad \text{‘v’} \quad \ldots ] \]

Edit Distance: Flip, Insert, Delete

[ Ebrahimi et al, ACL 2018, COLING 2018 ]
Change: Word-level Changes

\[ x = [ \text{‘I’} \quad \text{‘like’} \quad \text{‘this’} \quad \text{‘movie’} \quad \text{‘.’} ] \]

Let’s replace this word

Random word? \[ x’ = [ \text{‘I’} \quad \text{‘lamp’} \quad \text{‘this’} \quad \text{‘movie’} \quad \text{‘.’} ] \]

Word Embedding? \[ x’ = [ \text{‘I’} \quad \text{‘really’} \quad \text{‘this’} \quad \text{‘movie’} \quad \text{‘.’} ] \]

Part of Speech? \[ x’ = [ \text{‘I’} \quad \text{‘eat’} \quad \text{‘this’} \quad \text{‘movie’} \quad \text{‘.’} ] \]

Language Model? \[ x’ = [ \text{‘I’} \quad \text{‘hate’} \quad \text{‘this’} \quad \text{‘movie’} \quad \text{‘.’} ] \]

[Jia and Liang, EMNLP 2017 ]
[ Alzantot et. al. EMNLP 2018 ]
Change: Paraphrasing via Backtranslation

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

Translate into multiple languages

Use back-translators to score candidates

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

\[
\begin{align*}
S( \text{This is a good movie} , \text{This is a good movie} ) &= 1 \\
S( \text{This is a good movie} , \text{That is a good movie} ) &= 0.95 \\
S( \text{This is a good movie} , \text{Dogs like cats} ) &= 0
\end{align*}
\]
Change: Sentence Embeddings

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

\[
\begin{align*}
\min_{x'} \|z - z'\| \\
\text{s.t. } f(x') \neq f(x)
\end{align*}
\]

[Zhao et al ICLR 2018]
Choices in Crafting Adversaries

What is a small change?

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

s.t. $f(x') \neq f(x)$
Choices in Crafting Adversaries

How do we find the attack?

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

s.t. $$f(x') \neq f(x)$$
Search: How do we find the attack?

Even this is often unrealistic

- Only access predictions (usually unlimited queries)
- Create $x'$ and test whether the model misbehaves
- Create $x'$ and test whether general direction is correct
- Use the gradient to craft $x'$

- Full access to the model (compute gradients)
- Access probabilities

---

Sameer Singh, NAACL 2019 Tutorial
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

[Belinkov, Bisk, ICLR 2018]
[Iyyer et al, NAACL 2018]
[Ribeiro et al, ACL 2018]
Choices in Crafting Adversaries

How do we find the attack?

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

s.t. \( f(x') \neq f(x) \)
Choices in Crafting Adversaries

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

What does it mean to misbehave?
Effect: What does it mean to misbehave?

Classification

**Untargeted**: any other class

**Targeted**: specific other class

Other Tasks

**MT**: Don't attack me! $\rightarrow$ ¡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

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

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

• **Search**
  - Gradient-based
  - Sampling
  - Enumeration

• **Evaluation**

Sameer Singh, NAACL 2019 Tutorial
Research Highlights

In terms of the choices that were made
Noise Breaks Machine Translation!

<table>
<thead>
<tr>
<th>Change</th>
<th>Search</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Character Based</td>
<td>Passive; add and test</td>
<td>Machine Translation</td>
</tr>
</tbody>
</table>

[Belinkov, Bisk, ICLR 2018]

Sameer Singh, NAACL 2019 Tutorial
Hotflip

<table>
<thead>
<tr>
<th>Change</th>
<th>Search</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Character-based (extension to words)</td>
<td>Gradient-based; beam-search</td>
<td>Machine Translation, Classification, Sentiment</td>
</tr>
</tbody>
</table>

News Classification

Machine Translation

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 mood of optimism.

95% Sci/Tech

<table>
<thead>
<tr>
<th>src</th>
<th>Das ist Dr. Bob Childs – er ist Geigenbauer und Psychotherapeut.</th>
</tr>
</thead>
<tbody>
<tr>
<td>adv</td>
<td>Das ist Dr. Bob Childs – er ist Geigenbauer und Psy6hothearpeit.</td>
</tr>
<tr>
<td>src-output</td>
<td>This is Dr. Bob Childs – he’s a wizard maker and a therapist’s therapist.</td>
</tr>
<tr>
<td>adv-output</td>
<td>This is Dr. Bob Childs – he’s a brick maker and a psychopath.</td>
</tr>
</tbody>
</table>

[ Ebrahimi et al, ACL 2018, COLING 2018 ]
Search Using Genetic Algorithms

Black-box, population-based search of natural adversary

<table>
<thead>
<tr>
<th>Change</th>
<th>Search</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word-based, language model score</td>
<td>Genetic Algorithm</td>
<td>Textual Entailment, Sentiment Analysis</td>
</tr>
</tbody>
</table>

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.
## Natural Adversaries

### Textual Entailment

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Sentences</th>
<th>Label</th>
</tr>
</thead>
</table>
| 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 |

<table>
<thead>
<tr>
<th>Source Sentence (English)</th>
<th>Generated Translation (German)</th>
</tr>
</thead>
</table>
| 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

Semantically-Equivalent Adversary (SEA)

$\begin{align*}
  x & \rightarrow \text{Backtranslation} \\
  & + \text{Enumeration} \\
  & \rightarrow x'
\end{align*}$

<table>
<thead>
<tr>
<th>What color is the tray?</th>
<th>Pink</th>
</tr>
</thead>
<tbody>
<tr>
<td>What colour is the tray?</td>
<td>Green</td>
</tr>
<tr>
<td>Which color is the tray?</td>
<td>Green</td>
</tr>
<tr>
<td>What color is it?</td>
<td>Green</td>
</tr>
<tr>
<td>How color is tray?</td>
<td>Green</td>
</tr>
</tbody>
</table>

Semantically-Equivalent Adversarial Rules (SEARs)

$\begin{align*}
  (x, x') & \rightarrow \text{Patterns} \\
  & \text{in “ diffs”} \\
  & \rightarrow \text{Rules}
\end{align*}$

color $\rightarrow$ colour

<table>
<thead>
<tr>
<th>Change</th>
<th>Search</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence via Backtranslation</td>
<td>Enumeration</td>
<td>VQA, SQuAD, Sentiment Analysis</td>
</tr>
</tbody>
</table>

[Ribeiro et al, ACL 2018 ]

Sameer Singh, NAACL 2019 Tutorial
## Transformation Rules: VisualQA

<table>
<thead>
<tr>
<th>SEAR</th>
<th>Questions / SEAs</th>
<th>f(x)</th>
<th>Flips</th>
</tr>
</thead>
<tbody>
<tr>
<td>WP VBZ $\rightarrow$ WP’s</td>
<td><em>What has</em> What’s been cut?</td>
<td>Cake Pizza</td>
<td>3.3%</td>
</tr>
<tr>
<td>What NOUN $\rightarrow$ Which NOUN</td>
<td><em>What</em> Which kind of floor is it?</td>
<td>Wood Marble</td>
<td>3.9%</td>
</tr>
<tr>
<td>color $\rightarrow$ colour</td>
<td>What <em>color</em> colour is the tray?</td>
<td>Pink Green</td>
<td>2.2%</td>
</tr>
<tr>
<td>ADV is $\rightarrow$ ADV’s</td>
<td><em>Where is</em> Where’s the jet?</td>
<td>Sky Airport</td>
<td>2.1%</td>
</tr>
</tbody>
</table>

[Ribeiro et al, ACL 2018]
### Transformation Rules: SQuAD

<table>
<thead>
<tr>
<th>SEAR</th>
<th>Questions / SEAs</th>
<th>$f(x)$</th>
<th>Flips</th>
</tr>
</thead>
<tbody>
<tr>
<td>What VBZ →</td>
<td><strong>What is</strong> What’s the NASUWT?</td>
<td>Trade union</td>
<td>2%</td>
</tr>
<tr>
<td>What’s</td>
<td></td>
<td>Teachers in Wales</td>
<td></td>
</tr>
<tr>
<td>What NOUN →</td>
<td><strong>What resource</strong> Which resource was mined in the Newcastle area?</td>
<td>Coal wool</td>
<td>1%</td>
</tr>
<tr>
<td>Which NOUN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What VERB →</td>
<td><strong>What was</strong> So what was Ghandi's work called?</td>
<td>Satyagraha</td>
<td>2%</td>
</tr>
<tr>
<td>So what VERB</td>
<td></td>
<td>Civil Disobedience</td>
<td></td>
</tr>
<tr>
<td>What VBD →</td>
<td><strong>What was</strong> And what was Kenneth Swezey's job?</td>
<td>journalist sleep</td>
<td>2%</td>
</tr>
<tr>
<td>And what VBD</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Ribeiro et al, ACL 2018 ]
Sameer Singh, NAACL 2019 Tutorial
## Transformation Rules: Sentiment Analysis

<table>
<thead>
<tr>
<th>SEAR</th>
<th>Reviews / SEAs</th>
<th>f(x)</th>
<th>Flips</th>
</tr>
</thead>
<tbody>
<tr>
<td>movie → <strong>film</strong></td>
<td>Yeah, the <em>movie</em> film pretty much sucked. This is not <em>movie</em> film making.</td>
<td>Neg Pos</td>
<td>2%</td>
</tr>
<tr>
<td>film → <strong>movie</strong></td>
<td>Excellent <em>film</em> movie. I’ll give this <em>film</em> movie 10 out of 10!</td>
<td>Pos Neg</td>
<td>1%</td>
</tr>
<tr>
<td><strong>is</strong> → was</td>
<td>Ray Charles <em>is</em> was legendary. It <em>is</em> was a really good show to watch.</td>
<td>Pos Neg</td>
<td>4%</td>
</tr>
<tr>
<td>this → <strong>that</strong></td>
<td>Now <em>this</em> that is a movie I really dislike. The camera really likes her in <em>this</em> that movie.</td>
<td>Neg Pos</td>
<td>1%</td>
</tr>
</tbody>
</table>

[Ribeiro et al, ACL 2018]
Adding a Sentence

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

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

**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**: $0.827 \rightarrow 0.819$

**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**: $0.78 \rightarrow 0.91$

[Feng et al, EMNLP 2018 ]
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?
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

Sameer Singh, NAACL 2019 Tutorial
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