Textual Adversarial Attack and Defense

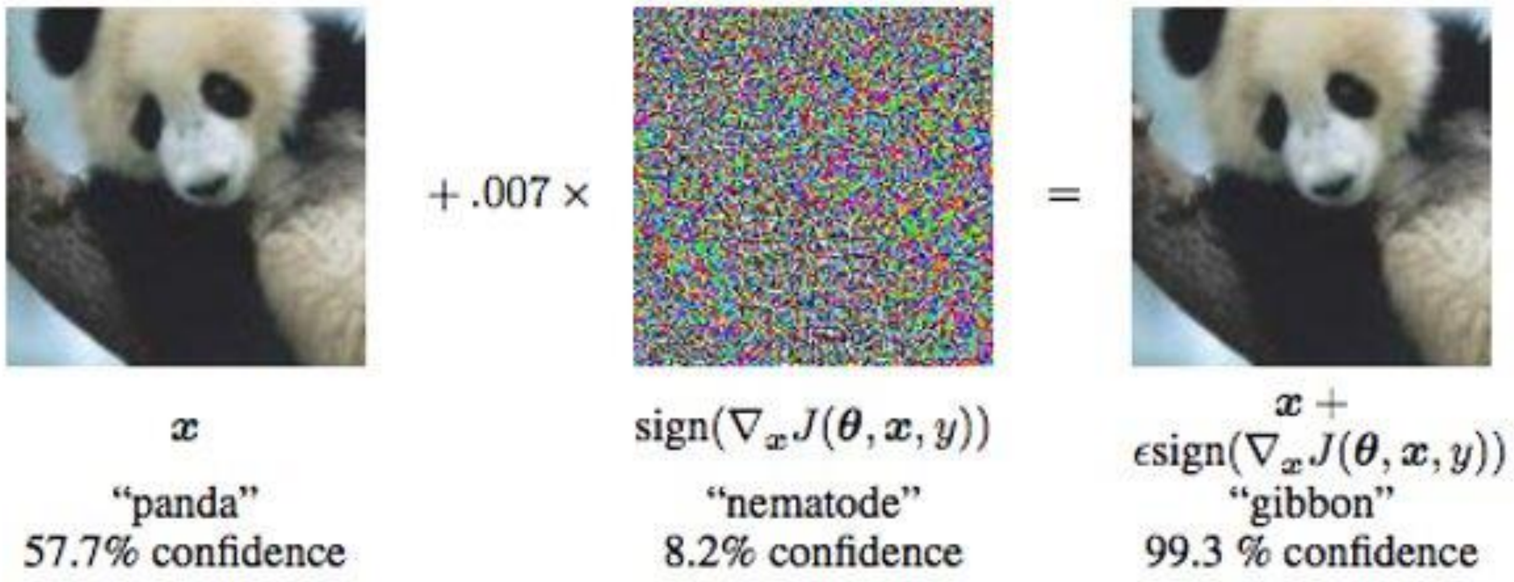
Hunar Preet Singh, 160301

Smarth Gupta, 160695

Suryateja BV, 160729

Motivation

- Adversarial attacks are effective in visual modality (imperceptible!)
- Can be used to improve the robustness of models
- But not so easy to generate in text due to its *discrete* nature



Motivation

- Generating a good adversarial example for the text counterpart that does not destroy the semantics is a highly non-trivial task.
- Since the text is *discrete*, even the smallest of perturbations can completely change the word and the sentence might not make sense at all.

A warm but realistic meditation on friendship , family and affection.

A farm but reyldktu meditation on friendship, family and affection. (**perceptible!**)

Hotflip Attack (Ebrahimi et al. 2018)

- Generating adversarial examples with character substitution ("flips")
- Uses the gradient with respect to a one-hot input representation
- Efficiently estimates which change has the highest estimated loss
- Uses beam search to find the optimal set of manipulations

South Africa’s historic Soweto township marks its 100th birthday on Tuesday in a mood of optimism -- **World**

South Africa’s historic Soweto township marks its 100th birthday on Tuesday in a mooP of optimism – **Sci/Tech**

Hotflip Objective

$$\max \nabla_x J(\mathbf{x}, \mathbf{y})^T \cdot \vec{v}_{ijb} = \max_{ijb} \frac{\partial J}{\partial x_{ij}}^{(b)} - \frac{\partial J}{\partial x_{ij}}^{(a)}$$

Differentiate loss **J** wrt sentence **x**

Difference Operator

i-th word **j**-th character

(a) --> current
(b) --> flipped

South Africa's historic Soweto township marks its 100th birthday on Tuesday in a mood of optimism -- **World**

South Africa's historic Soweto township marks its 100th birthday on Tuesday in a mooP of optimism – **Sci/Tech**

i = 14, j = 4, (a) = 4, (b) = 16

Universal Adversarial Triggers

- *Input-agnostic* sequences of tokens that trigger a model to produce a specific prediction when concatenated to *any input* from a dataset
- Gradient guided search over the token space
- Built over word-level extension of Hotflip
- Wallace et al. (2019)

Example : Sentiment Analysis

Input: zoning tapping fiennes Visually imaginative, thematically instructive and thoroughly delightful, it takes us on a roller-coaster ride.

Model Prediction : 1 --> 0

Triggers Objective

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

Find a **token** such that...

Over all sentences in the dataset **T**

The prediction is always a particular **target class** when **sentence** is appended with the **token**

Input: zoning tapping fiennes Visually imaginative, thematically instructive and thoroughly delightful, it takes us on a roller-coaster ride.

Model Prediction : 1 --> 0

token = zoning tapping fiennes **target class = 0**

Defense Techniques

- Adversarial Training
  - Find adversarial examples (Hotflip), augment the training set
  - Retraining the model => very **expensive**!
- Diagnostic Datasets
  - Q: What room is this? A: Kitchen
  - Q': Is there a kitchen? A': Yes
  - Q'': Is there a bathroom? A'': No [Sameer Singh et al. 2019]
- Adversary Recognition Models
  - Cheap, but can fail in interesting ways (more on this later!)
- Probing
  - Understand the inner workings of the model better
  - Bertology, Understanding Bias using Influence Functions, Probing Numeracy
  - AllenAI Interpret

Word-level **Hotflip Attack**

- Using allennlp's implementation (nightly version)
- Attack on SST-2 dev.txt (binary classification)
- Model : Simple BiLSTM using GloVe embeddings
- Doesn't make sense to use as an *"adversary"*

A beguiling splash of pastel colors and *prankish* comedy from Disney. (1)

A beguiling splash of pastel colors and unfunny comedy from Disney. (0)

Suffers from the *lack* of a compelling or comprehensible narrative. (0)

noir from the collaborative of a compelling or comprehensible narrative. (1)

What is a *good* adversary?

- Attacked word (spellings changed) mostly taken as "UNK" in word-level models
- Replacing words with synonyms or other words (from vocab) changes the sentence structure (perceptible!)
- Most attacks seem to be only of academic interest. Can we have more *realistic* attacks? [Spam, Programmatic Censorship]
- Some inspiration from Psycholinguistics : don't change first and last letter of a word

Types of Character Attacks

- Add
  - Q: where is the elbephant?
  - A: Africa (38.1%)
- Drop
  - Q: where is the ele~~p~~hant?
  - A: Africa (38.7%)
- Swap
  - Q: wehre is the elephant?
  - A: yes (77%)
- Keyboard
  - Q: where is the elephsnt?
  - A: Africa (38.7%)
- Repeat
  - Q: wherre is the elephant?
  - A: yes (84%)

MS-CoCo VQA 1.0
Q: where is the elephant?
A: Africa (56.5%)

Which attack works the ***best***?

1. Model : BiLSTM Word+Char level model trained on SST-2
    - No attack => **80.5 %**
    - A combination of the three attacks works well. Tough to defend too...

	Add	Drop	Swap	Keyboard	All
Attack	39.8%	50.8%	52.3%	40.8%	**35.6%**
Defense	59%	65%	78%	62%	**56.5%**

The **tasks** we consider

- Sentiment Classification
  - Dataset: SST-2 dev set
  - Model: distilbert-base-uncased-finetuned-sst-2-english
  - Eval Metric: Accuracy
- Extractive Question Answering
  - Dataset: SQuAD v2.0 dev set
  - Model: distilbert-base-cased-distilled-squad
  - Eval Metric: Exact Score, F1 Score
- Paraphrase Identification
  - Dataset: MRPC
  - Model: bert-base-cased-finetuned-mrpc
  - Eval Metric: Accuracy

Defense using Word Recognition

- Input : word representation based on characters
  - concatenation of one-hot representation of first letter, last letter and a BoW representation of remaining letters
- Task : Predict which word in the vocabulary the representation corresponds to
- Output : One-hot vector (of dim V)
- Model : Vanilla BiLSTM

As a resulwt, Nelson nlw faces upto 10 years' jnail instead of life

As a result, Nelson nlw faces upto 10 years' jail instead of life

Experimental Setup

- Task : [sst, squad, mrpc]
- Type of attack : [add, drop, swap, key, rep]
- Num_attacks : [1...10]
- Defense : BiLSTM Word Recognizer

Experimental Results

	Original	Attack	Defense
Sentiment Classification	0.91	0.75	0.87
Question Answering	0.79	0.35	0.49
Question Answering (F1)	0.84	0.46	0.56
Paraphrase Identification	0.93	0.69	0.75

NUM_ATTACKS = 7

Effect of Attack Strength

- We consider the sentiment classification task here
- Accuracy of both attack and defense techniques decreases as attack strength increases

	Original	Attack	Defense
NUM_ATTACKS = 1	0.91	0.88	0.90
NUM_ATTACKS = 3	0.91	0.83	0.89
NUM_ATTACKS = 5	0.91	0.79	0.88
NUM_ATTACKS = 7	0.91	0.69	0.87

An Example (Extractive QA)

- **Context**: 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 . . . . . .
- Question: Why did the university see a drop in applicants?
- Answer: crime and poverty
- Question: Wjty didd the uiversity see a dop in apllicants?
- Answer: the university became a major sponsor of a controversial urban renewal project
- Question: What did the university see a dip in applicants?
- Answer: Increasing crime and poverty

Future Work

- Character-level attacks :
  - Do not preserve semanticity.
  - Can be defended to some extent (as we have seen)

- Even more viscous attacks can be of the form:
  - Flights from New York to Florida
  - Flights from Florida to NYC
  - Flights from Florida to New York

High Lexical Overlap

Future Work

- **PAWS**: **P**araphrase **A**dversaries from **W**ord **S**crambling (DATASET)
- Consists of challenging pairs (both paraphrase and non-paraphrase)
- Generated using controlled word-swap and **back translation**
- We will train encoder-decoder based models on this dataset

Another interesting line of work

- **Learning Neural Templates for Text Generation**
- Neural Generation system using hidden semi-markov models
- Learns latent discrete templates jointly with a generation model
- These templates make generation controllable and interpretable
- Can be used along with **PAWS** for generating adversaries

References

- Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for nlp. arXiv preprint arXiv:1908.07125.
- Javid Ebrahimi, Daniel Lowd, and Dejing Dou. 2018a. On adversarial examples for characterlevel neural machine translation. arXiv preprint arXiv:1806.09030
- Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018b. HotFlip: White-box adversarial examples for text classification.
- Zhang, Yuan, Jason Baldridge, and Luheng He. "PAWS: Paraphrase adversaries from word scrambling." *arXiv preprint arXiv:1904.01130* (2019).
- Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).
- Pruthi, Danish, Bhuwan Dhingra, and Zachary C. Lipton. "Combating adversarial misspellings with robust word recognition."*arXiv preprint arXiv:1905.11268* (2019).
- Wiseman, Sam, Stuart M. Shieber, and Alexander M. Rush. "Learning neural templates for text generation." *arXiv preprint arXiv:1808.10122* (2018).
- Brunet, Marc-Etienne, et al. "Understanding the origins of bias in word embeddings." *arXiv preprint arXiv:1810.03611* (2018).
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Semantically equivalent adversarial rules for debugging nlp models." *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018.

Thank You!
Questions?