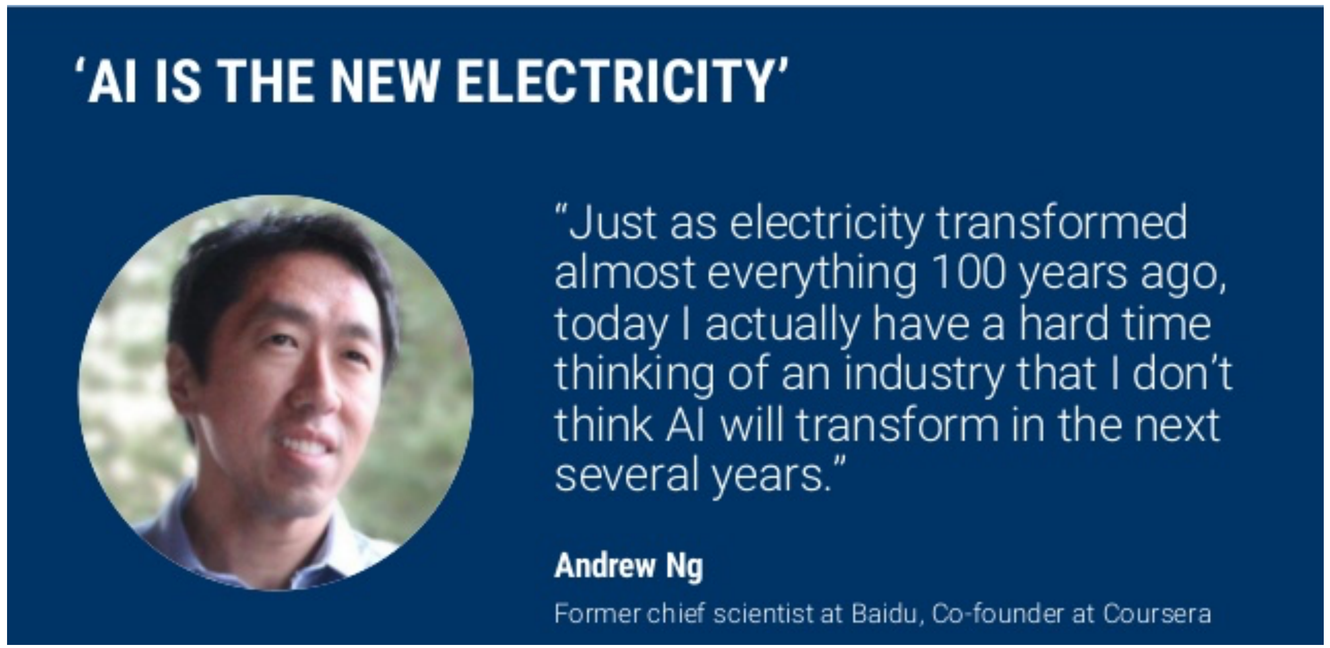
# Generating Black-Box Adversarial Examples for Text Classifiers Using a Deep Reinforced Model

Prashanth Vijayaraghavan & Deb Roy
MIT Media Lab

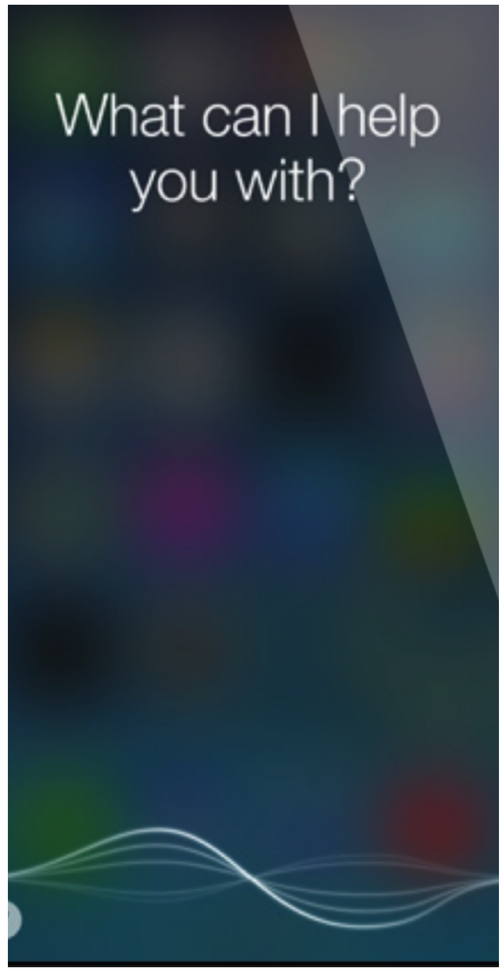
mit media lab

social machines

# AI Transforming The World ..



# Are They Fail Proof? Are They Safe?

**It's disturbingly easy to trick AI into doing something deadly**

How "adversarial attacks" can mess with self-driving cars, medicine, and the military.

**Google's image recognition AI fooled by new tricks**

**Hate-speech detection algorithms are trivial to fool**

AI, ML & DATA ENGINEERING

**Privacy Attacks on Machine Learning Models**

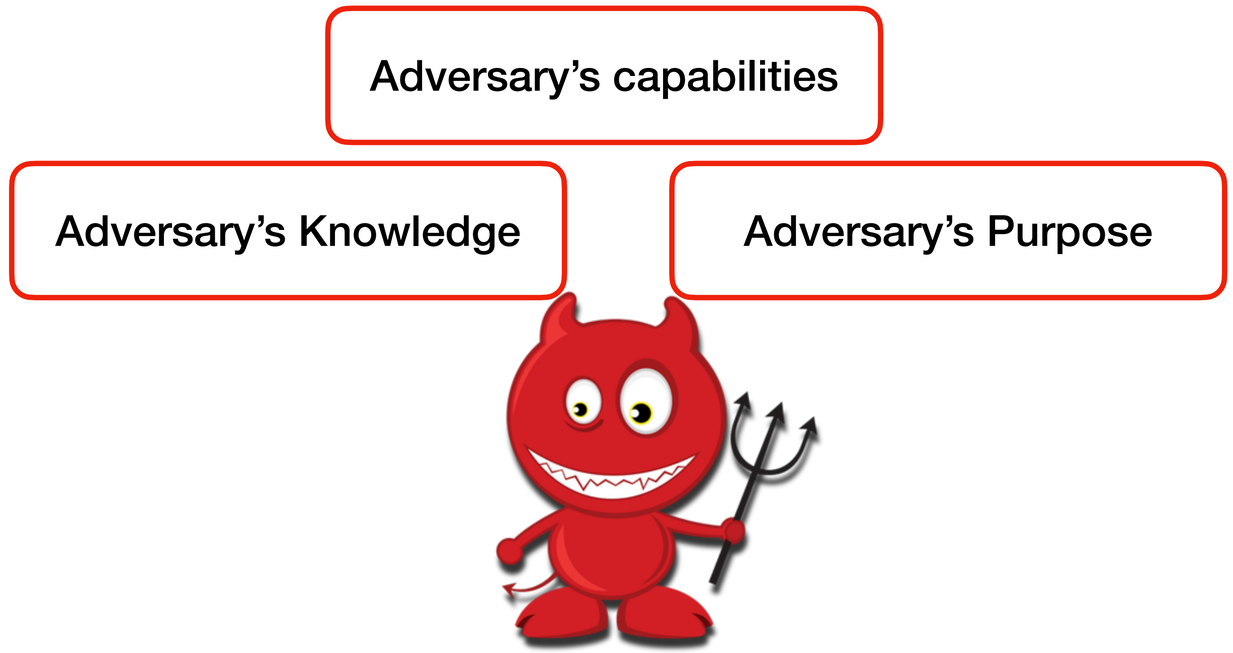
Text-based AI models are vulnerable to paraphrasing attacks, researchers find

# Knowing The Adversary

- *Adversary's capabilities:*
  - ✳ *Poisoning Attacks*
  - ✳ *Evasion attacks*

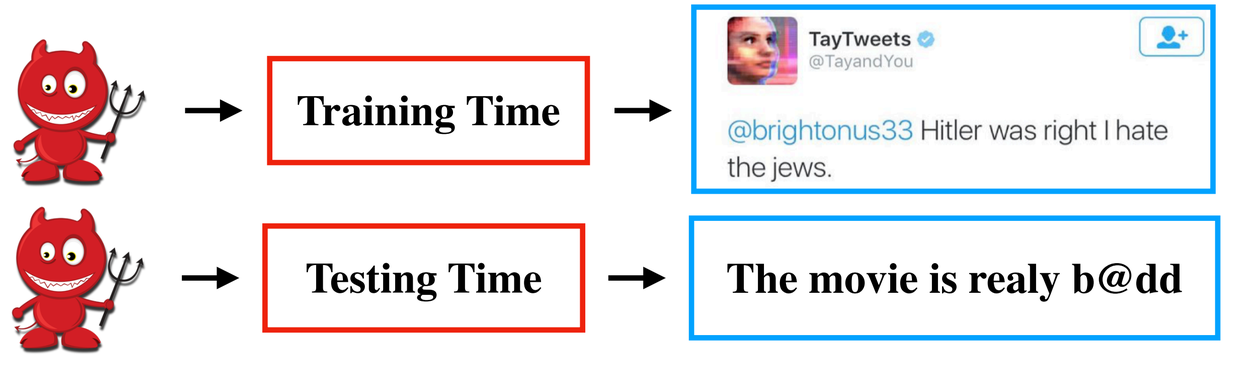
- *Adversary's knowledge:*
  - ✳ *Black-box*
  - ✳ *White-box*
  - ✳ *Gray-box*

- *Adversary's purpose:*
  - ✳ *Targeted*
  - ✳ *Non-targeted*

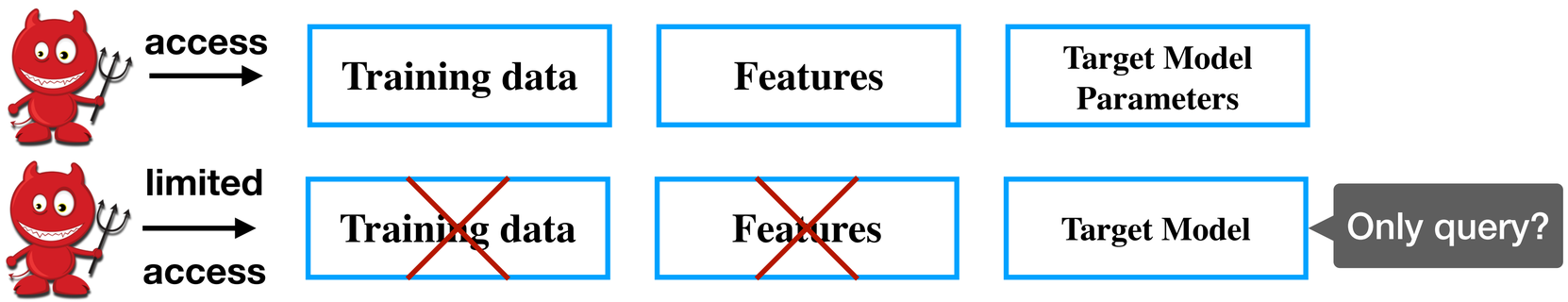


# Knowing The Adversary

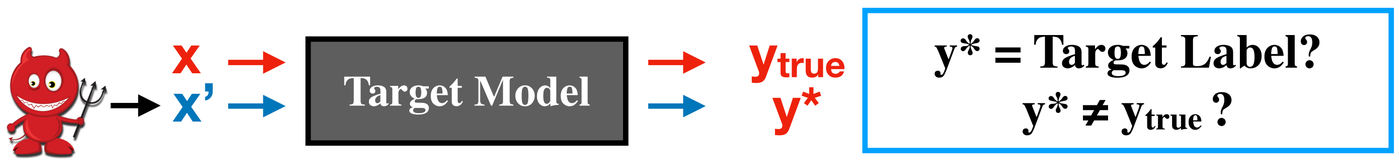
- *Adversary's capabilities:*
  - ✳ *Poisoning Attacks*
  - ✳ *Evasion attacks*




- *Adversary's knowledge:*
  - ✳ *White-box*
  - ✳ *Black-box*
  - ✳ *Gray-box*




- *Adversary's purpose:*
  - ✳ *Targeted*
  - ✳ *Non-targeted*



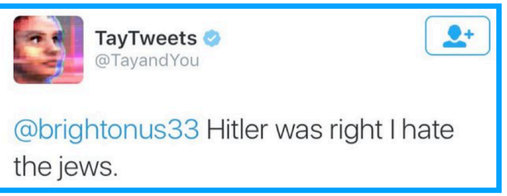
# Knowing The Adversary

- *Adversary's capabilities:*
  - ✳ *Poisoning Attacks*
  - ✳ ***Evasion attacks***

- *Adversary's knowledge:*
  - ✳ *White-box*
  - ✳ ***Black-box***
  - ✳ *Gray-box*

- *Adversary's purpose:*
  - ✳ *Targeted*
  - ✳ ***Non-targeted***

→ **Training Time** →




→ **Testing Time** → **The movie is realy b@dd**


access → **Training data** | **Features** | **Learning Model**

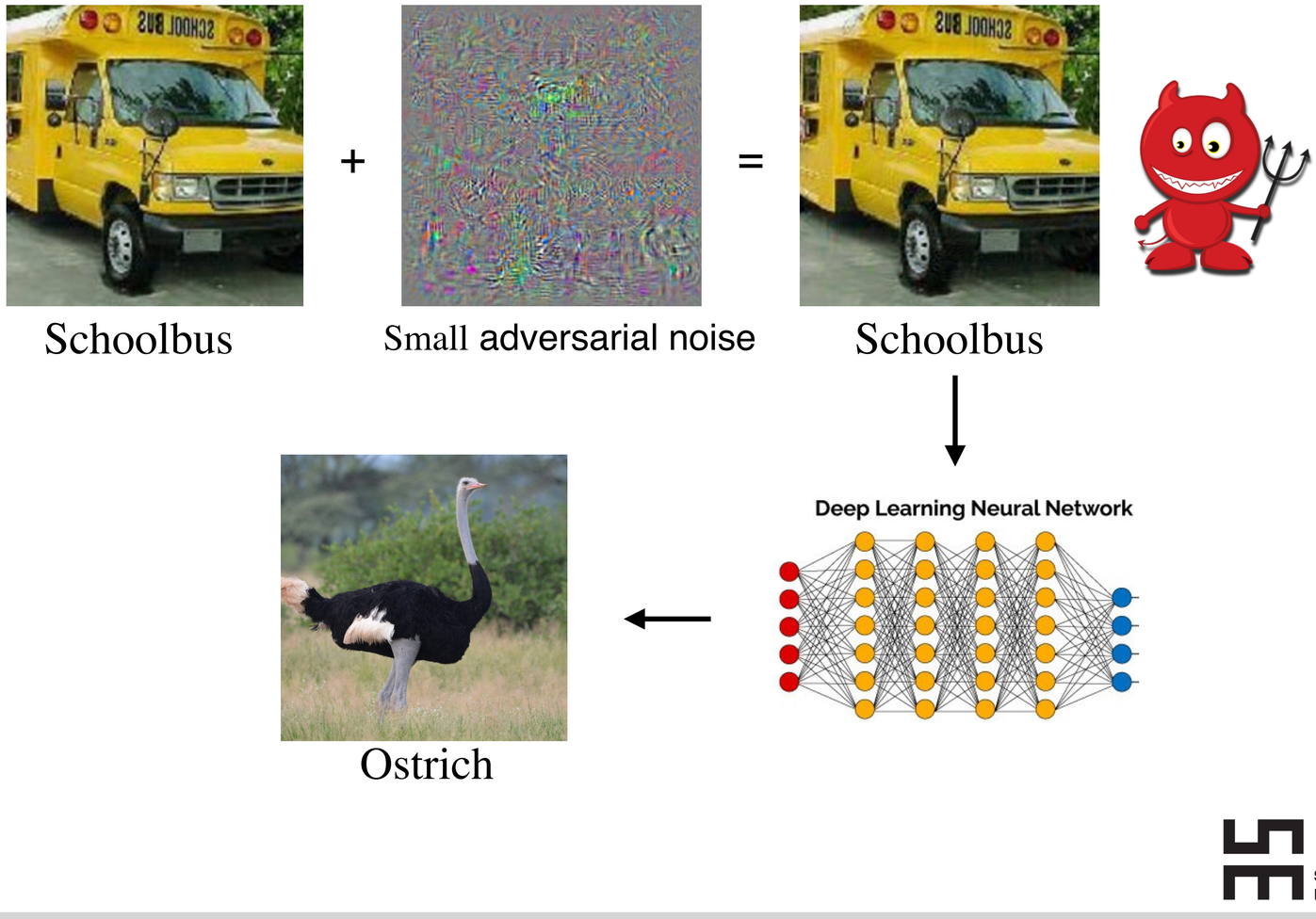
limited access → ~~**Training data**~~ | ~~**Features**~~ | **Learning Model** — Only query?

x → x' → **Target Model** → $y_{true}$ / $y^*$

**$y^*$ = Target Label?**
**$y^* \neq y_{true}$ ?**

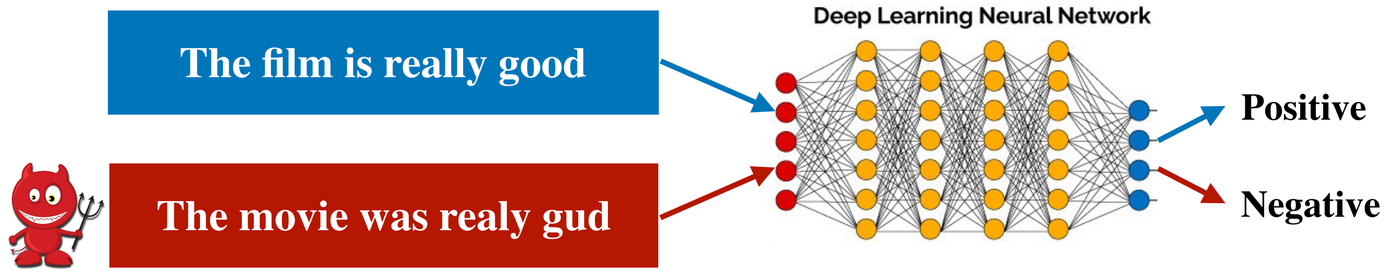
# Image-based Models are Fooled

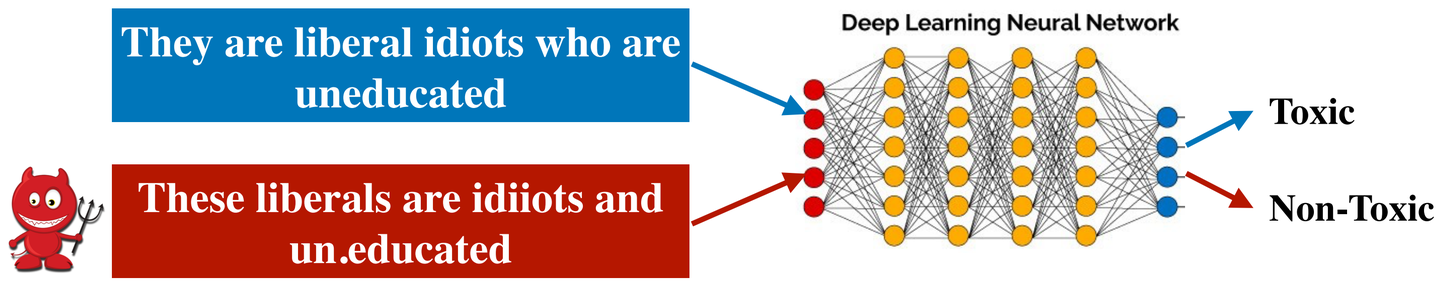
- *Attacks on Image classification, Face recognition, Fingerprint matching systems have become common*
- *Adversarial noise in images have been in the form of small random noise added to images*
- *Usually, these perturbations are* ***small, imperceptible***



# Text-based models are fooled

- *Attacks on Spam, sentiment, toxicity classification have become common*
- *Adversarial noise in the text:* ***paraphrases, misspellings***

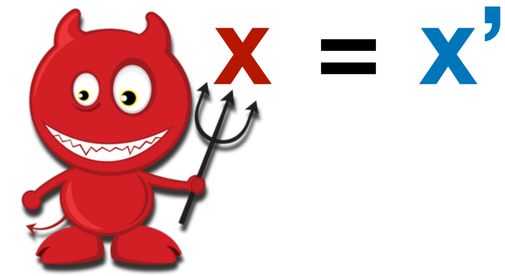


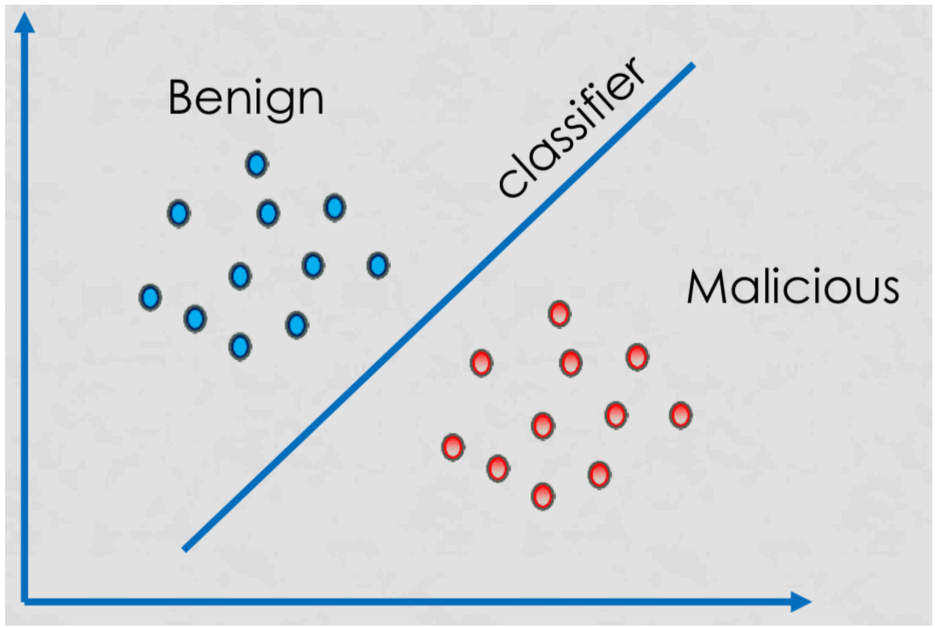



# Definition

*"Adversarial examples are inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake"*

***Goodfellow et al. (2017)***



# Background: Adversarial Examples in Image Domain

- ***Past work:***
  - ***White-box:*** *Gradient-based, decision function-based, spatial transformation*
  - ***Black-box:*** *Based on transferability—Training local substitute models*

**Synthesizing Robust Adversarial Examples**

**Anish Athalye** [*12] **Logan Engstrom** [*12] **Andrew Ilyas** [*12] **Kevin Kwok** [2]

DELVING INTO TRANSFERABLE ADVERSARIAL EXAMPLES AND BLACK-BOX ATTACKS

**Yanpei Liu**[*], **Xinyun Chen**[*]
Shanghai Jiao Tong University

**Chang Liu, Dawn Song**
University of the California, Berkeley

**Generating Adversarial Examples with Adversarial Networks**

**Chaowei Xiao**[1] [*], **Bo Li**[2], **Jun-Yan Zhu**[2,3], **Warren He**[2], **Mingyan Liu**[1] and **Dawn Song**[2]

ADVERSARIAL EXAMPLES IN THE PHYSICAL WORLD

**Alexey Kurakin**
Google Brain
kurakin@google.com

**Ian J. Goodfellow**
OpenAI
ian@openai.com

**Samy Bengio**
Google Brain
bengio@google.com

Adversarial Examples for Semantic Segmentation and Object Detection

*Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, Alan Yuille*; The IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1369-1378

GENERATING NATURAL ADVERSARIAL EXAMPLES

**Zhengli Zhao**
University of California
Irvine, CA 92697, USA
zhengliz@uci.edu

**Dheeru Dua**
University of California
Irvine, CA 92697, USA
ddua@uci.edu

**Sameer Singh**
University of California
Irvine, CA 92697, USA
sameer@uci.edu

EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES

**Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy**
Google Inc., Mountain View, CA

mit media lab

social machines

# Background: Adversarial Examples in Text

- ***Past work:***
  - ***Sensitivity analysis*** *on NMT exposed its vulnerability to misspellings*
  - ***White-box:*** *Gradient-based method —manipulates 1-hot representation of text, word embeddings perturbation to nearest neighbor*
  - ***Black-box:*** *Genetic algorithm, DeepWordBug — scoring function, rank words, perturb chars (Involves heuristics), Enc-Dec models: GAN-based applied only for binary, SCPN (Iyyer et al) relies on paraphrases.*

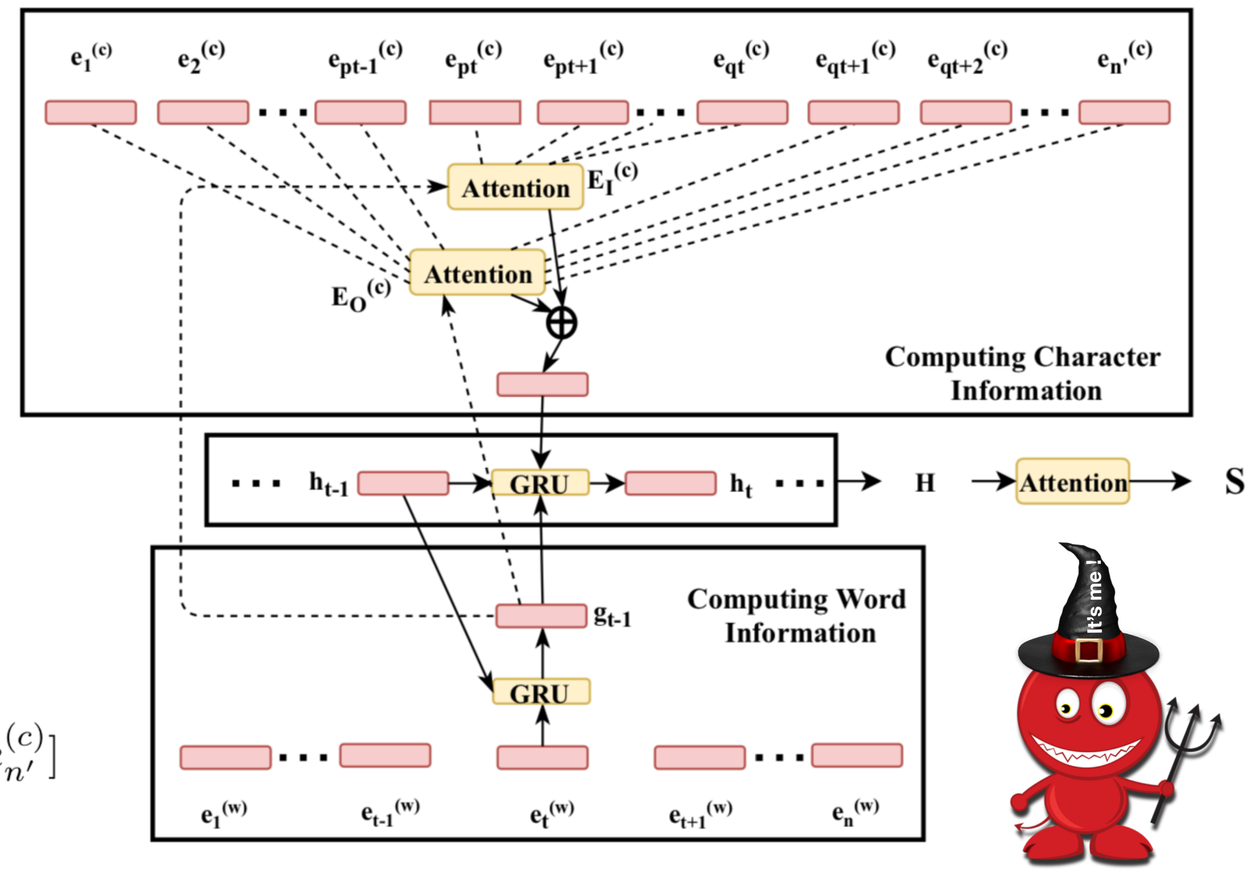
- ***Challenges:***
  - *Limited work in NLP*
  - *Discrete nature of text samples*
  - *Even a small change is perceptible & can lead to semantic change*
  - *Identifying salient areas of text*
  - *Applying both character & paraphrases together*

# Proposed Attack Strategy

- ***Encoder-Decoder architecture****: Adversarial Example Generator (AEG)*
- ***AEG:*** *Substitute Network Training + adversarial sample generation*
  - ✳ *Encoder: Acts as substitute network*
  - ✳ *Attention-based salient feature identification*
  - ✳ *Decoder: Produces both character and word perturbations (misspellings or paraphrases)*
- ***Training:*** *Randomly draw labeled data that is disjoint from the training data*
- ***RL-based Improvement:*** *Self-critical approach*
- ***Reward Constraints****: (a) fool the target classifier, (b) minimize the number of perturbations and (c) preserve the semantics of the text.*

# Encoder

- *Encodes **word & character-level** information*
- *Using **Bidirectional GRU**, we obtain:* $\overleftrightarrow{h_t}$
- *Obtain **summary vector** S using attention to predict the label*

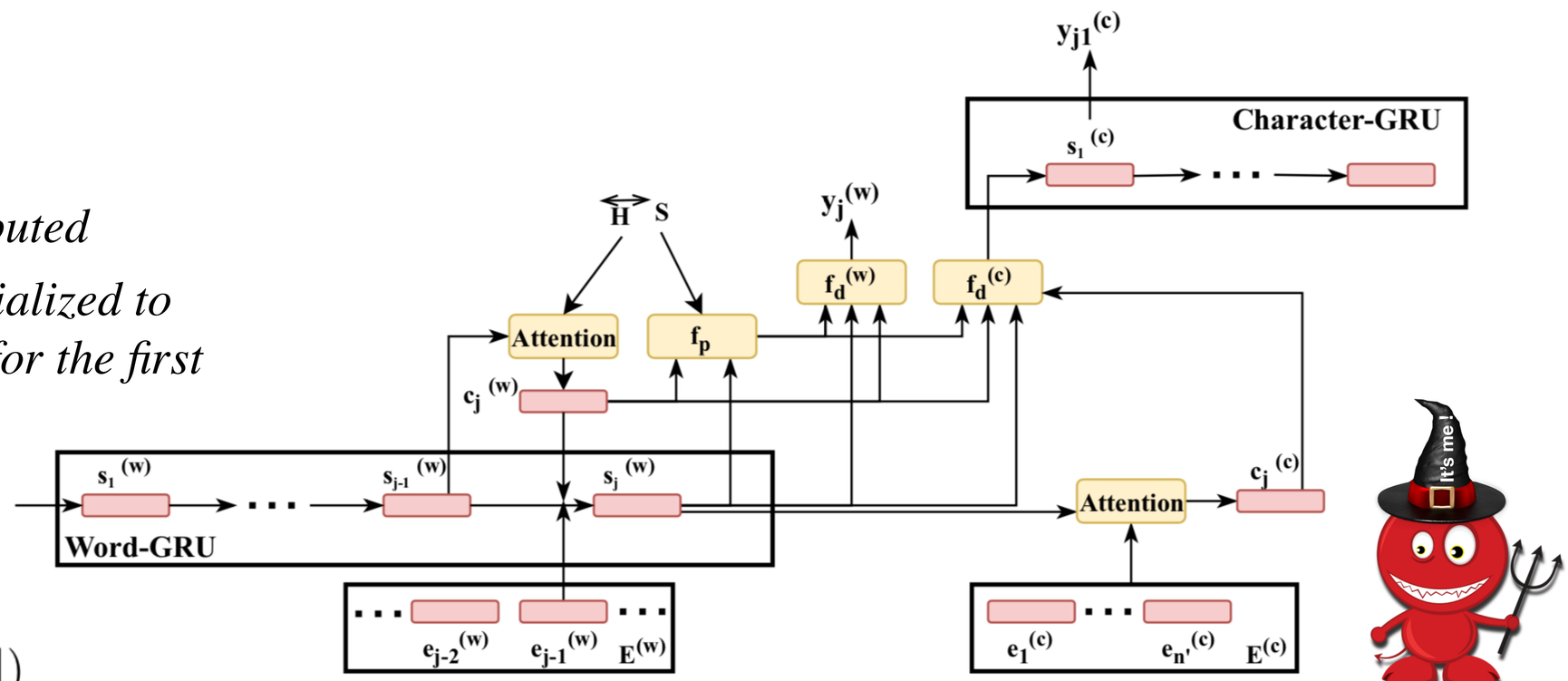



$$E^{(c)} = [e_1^{(c)}, e_2^{(c)}, ...e_{n'}^{(c)}] \quad E^{(w)} = [e_1^{(w)}, e_2^{(w)}, ...e_n^{(w)}]$$

$$E_I^{(c)} = [e_{p_t}^{(c)}, ...., e_{q_t}^{(c)}] \quad E_O^{(c)} = [e_1^{(c)}, ...., e_{p_t-1}^{(c)}; e_{q_t+1}^{(c)}, ..., e_{n'}^{(c)}]$$

# Decoder

- *Multi level GRU:*
  - *Word-GRU*
  - *Character-GRU*
- *Perturbation vector computed*
- *Character state $s_j^{(c)}$is initialized to the character-GRU only for the first hidden state*







$$v_p = f_p(s_j^{(w)}, c_j^{(w)}, S)$$

$$\tilde{s}_j^{(w)} = f_d^{(w)}([c_j^{(w)}; s_j^{(w)}; v_p])$$

$$\tilde{s}_j^{(c)} = f_d^{(c)}([c_j^{(w)}; s_j^{(w)}; v_p; c_j^{(c)}])$$

# Datasets


- *ParaNMT-50M*
- *Quora Question Pair Dataset*
- *Twitter URL Paraphrasing Corpus*
- ***Comprises of:***
  - ✳ *Common Crawl*
  - ✳ *CzEngl1.6*
  - ✳ *Europarl*
  - ✳ *News Commentary*
  - ✳ *Quora Questions*
  - ✳ *Twitter Trending Topic Tweets*
- ***Sample:*** *5M Parallel Texts*
- ***Data augmentation:***
  - ✳ *Character only*
  - ✳ *Character & word transformations*

PARANMT—50M

Common Crawl

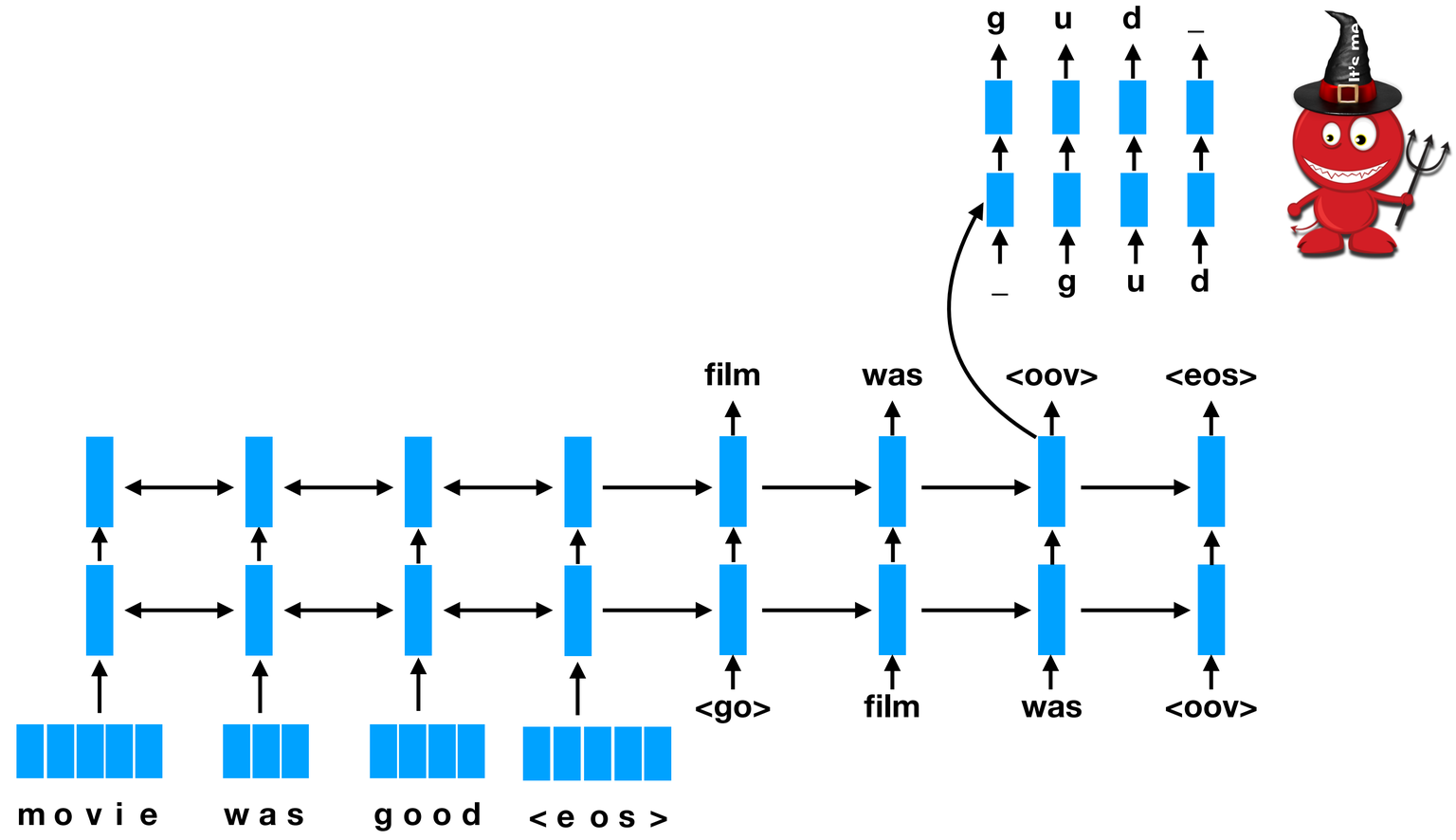
# Training

- *Supervised Pretraining with **Teacher forcing***
- *Training with **Reinforcement Learning***

# Supervised Pretraining with Teacher Forcing

- *Pretraining helps us mitigate the cold-start issue.*
- *No guarantee that the perturbations will fool the target model.*

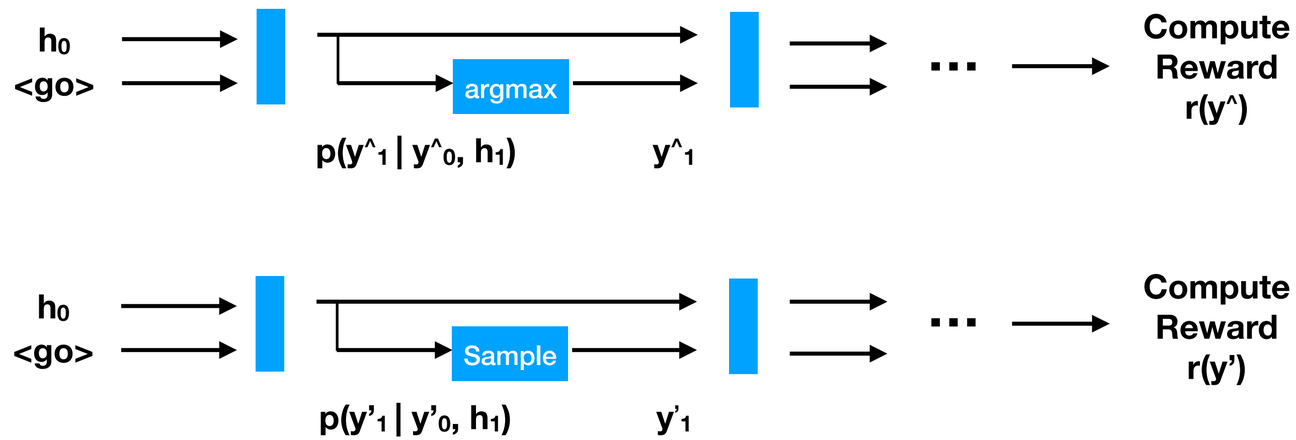
$$J_{mle} = J^{(w)} + J^{(c)}$$



# Training with Reinforcement Learning

- *Fine-tune by learning policy to fool a target classifier*
- *Policy gradient training through **Self-Critical Sequence Training** (SCST)*

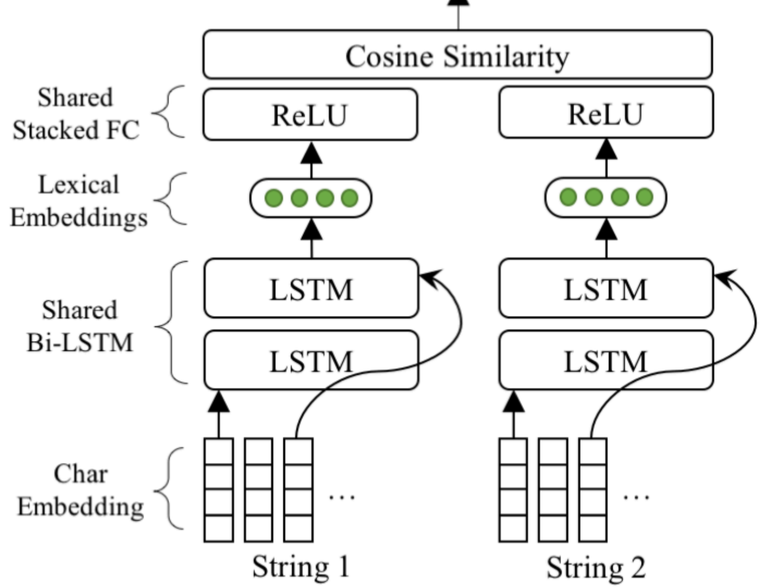
$$J_{rl} = -\sum_j (r(y') - r(\hat{y}))logp(\hat{y}_j|\hat{y}_{<j}, h)$$

# Rewards


- ***Adversarial Reward:***
  $R_A = (1 - P_l)$
- ***Semantic Similarity:*** *Deep Matching model—Char BiLSTM*
- ***Lexical Similarity:*** *Approximate Levenshtein distance*

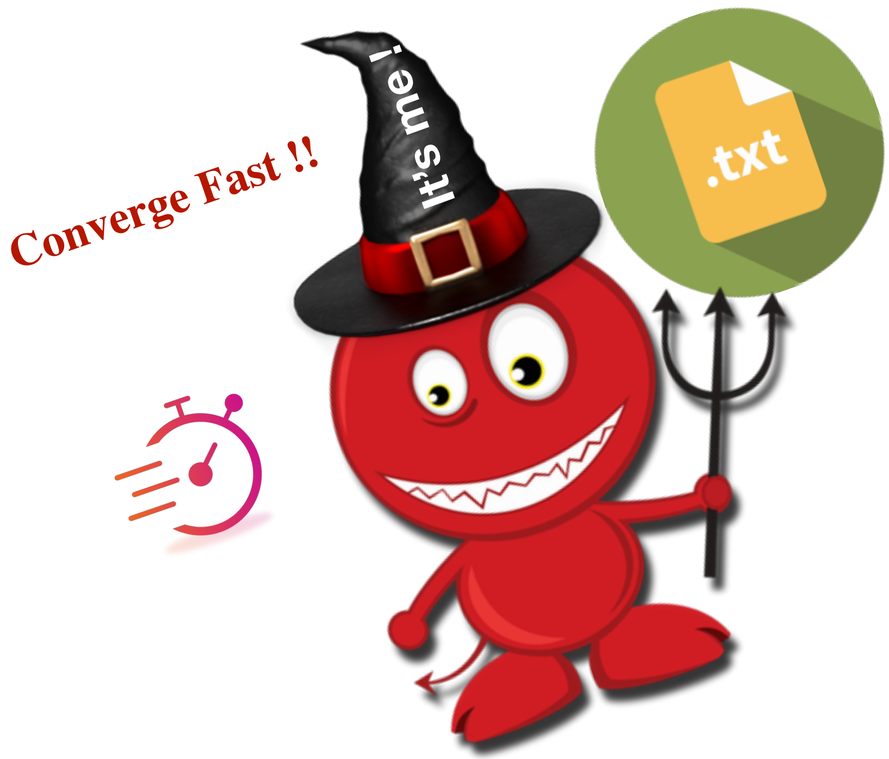
$$r(y) = \gamma_A R_A + \gamma_S R_S + \gamma_L R_L$$



# Training Details

- *Initialized with 300-d GloVe vectors*
- *Adam Optimization*
- *Reward coefficients:*

$$\gamma_A = 1, \gamma_S = 0.5, \gamma_L = 0.25$$



# Experiments


- ***Sentiment Classification***: *IMDB Reviews*
- ***News Categorization***: *AG's News*



| Datasets | Details | Model | Accuracy |
|---|---|---|---|
| IMDB Review | Classes: 2; #Train: 25k; | CNN-Word [20] | 89.95% |
| AG's News | Classes: 4; #Train: 120k; | CNN-Char [42] | 89.11% |

**Table 1.** Summary of data and models used in our experiments.

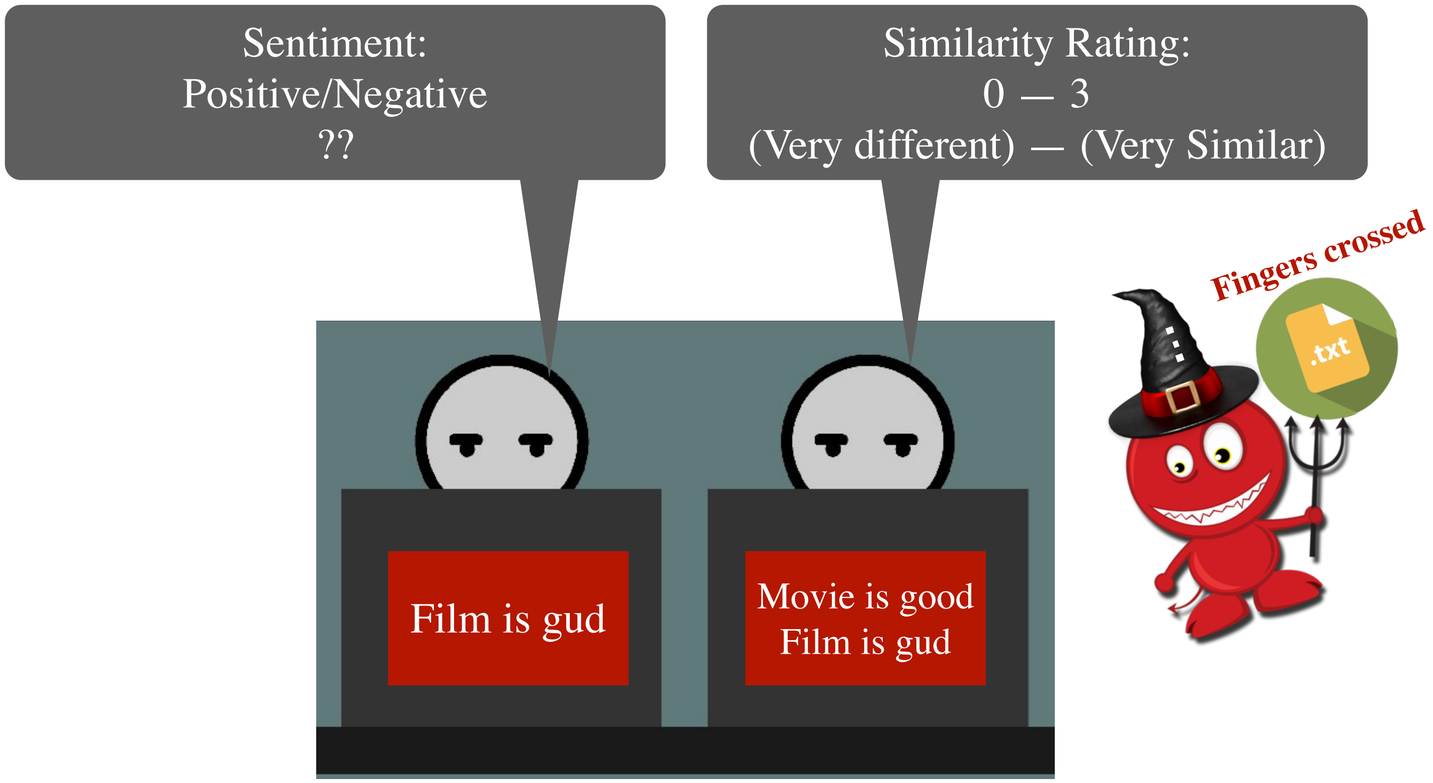
# Quantitative Analysis

- *Performance measured by % drop in accuracy*
- *Higher the % dip, more effective is our model*

| Models | IMDB (CNN-Word) | AG's News (CNN-Char) |
|---|---|---|
| Random | 2.46% | 9.64% |
| NMT-BT | 25.38% | 22.45% |
| DeepWordBug | 68.73% | 65.80% |
| No-RL (Ours) | 38.05% | 33.58% |
| **AEG (Ours)** | **79.43%** | **72.16%** |

# Human Evaluation

- *Perturbations introduced don't affect human judgements 94.6% times.*
- *Semantic Preserving with similarity rating closer to 2*



# Ablation Studies

- *No perturbation vector used (No pert)*
- *Simple character-decoder with perturbation vector*
- ***Findings:***
  - *Absence of hybrid decoder leads to performance drop*
  - *Perturbation vector is important*


| Model Variants | IMDB | News Corpus |
|---|---|---|
| Char-dec | 73.5 | 68.64% |
| No pert | 71.45% | 65.91% |

# Qualitative Analysis

- *Visualize attention scores & perturbations introduced*
- *Set one hyper parameter closer to 1 & see its effect*


| | | |
|---|---|---|
| This is an example of why the majority of action films are the same. Generic and boring, there's really nothing worth watching here. A complete waste of the then barely-tapped talents…<br><br>this is an example of why most of the action movis are so similar. mostly generic and borin, there 's nothin worth or good watching here. A complete taste of the then barely tapped talents…<br><br>**Negative → Positive** | $\gamma_A \approx 0, \gamma_S \approx 1, \gamma_L \approx 0$ | … unions representing workers at turner newall say they are disappointed after talks with stricken parent firm federal mogul …<br><br>… labor force at turner newall inform that they are upset with the meeting with parent company 's manager … |
| The premise is good, the plot line interesting and the screenplay was OK. A tad too simplistic in that a coming out story of a gay man was so positive when it is usually not quite so positive.<br><br>The premise is good, though script was intresting. The film 's screenplay was mediocre . It was too generic in a coming out story of a gay man was so positive but it is usually not quite so positive.<br><br>**Positive → Negative** | $\gamma_A \approx 0, \gamma_S \approx 0, \gamma_L \approx 1$ | … unions representing workers at turner newall say they are disappointed after talks with stricken parent firm federal mogul …<br><br>… unions representing workrs at turner newall say they are disappointed after talks with stricken parent firm federal mogul … |
| … Its charming, delightful, sad, funny, and every- thing in between.…<br><br>… its charmng, delightfull, disappointing, sad, funny, and every thing between …<br><br>**Positive → Negative** | $\gamma_A \approx 1, \gamma_S \approx 0, \gamma_L \approx 0$ | … unions representing workers at turner newall say they are disappointed after talks with stricken parent firm federal mogul …<br><br>… unyons representing labors at turner newall say they are meeting at a new place with stricken parents and children … |

# Conclusion

- *AEG— generate adversarial examples to fool black-box text classification models*
- *SCST based RL approach used*
- *Generates semantics preserving perturbations*
- *Generated examples cause steep drop in accuracy*
- *Extremely low values of reward coefficients affects performance*
- *Can help adversarial training to build robust models*