Seq2Sick: Evaluating the Robustness of Sequence-to-Sequence Models with Adversarial Examples

5 authors, including:

Minhao Cheng
University of California, Los Angeles
19 PUBLICATIONS   178 CITATIONS

SEE PROFILE

Huan Zhang
University of California, Davis
54 PUBLICATIONS   1,610 CITATIONS

SEE PROFILE

Pin-Yu Chen
IBM Research
195 PUBLICATIONS   2,607 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project   show and fool View project

Project   Distributed Optimization and Online Learning View project

Seq2Sick: Evaluating the Robustness of Sequence-to-Sequence Models with Adversarial Examples

Minhao Cheng[1], Jinfeng Yi[2], Huan Zhang[1], Pin-Yu Chen[3], Cho-Jui Hsieh[1]

[1]Department of Computer Science, University of California, Davis, CA 95616
[2]Tencent AI Lab, Bellevue, WA 98004
[3]IBM Research AI, Yorktown Heights, NY 10598
mhcheng@ucdavis.edu, jinfengyi.ustc@gmail.com, ecezhang@ucdavis.edu,
pin-yu.chen@ibm.com, chohsieh@ucdavis.edu

Abstract

Crafting adversarial examples has become an important technique to evaluate the robustness of deep neural networks (DNNs). However, most existing works focus on attacking the image classification problem since its input space is continuous and output space is finite. In this paper, we study the much more challenging problem of crafting adversarial examples for sequence-to-sequence (seq2seq) models, whose inputs are discrete text strings and outputs have an almost infinite number of possibilities. To address the challenges caused by the discrete input space, we propose a projected gradient method combined with group lasso and gradient regularization. To handle the almost infinite output space, we design some novel loss functions to conduct non-overlapping attack and targeted keyword attack. We apply our algorithm to machine translation and text summarization tasks, and verify the effectiveness of the proposed algorithm: by changing less than 3 words, we can make seq2seq model to produce desired outputs with high success rates. On the other hand, we recognize that, compared with the well-evaluated CNN-based classifiers, seq2seq models are intrinsically more robust to adversarial attacks.

1 Introduction

Adversarial attack on deep neural networks (DNNs) aims to slightly modify the inputs of DNNs and mislead them to make wrong predictions [1, 2]. This task has become a common approach to evaluate the robustness of DNNs [3, 4, 5] – generally speaking, the easier an adversarial example can be generated, the less robust the DNN model is. However, models designed for different tasks are not born equal: some tasks are strictly harder to attack than others. For example, attacking an image is much easier than attacking a text string, since image space is continuous and the adversary can make arbitrarily small changes to the input. Therefore, even if most of the pixels of an image have been modified, the perturbations can still be imperceptible to humans when the accumulated distortion is not large. In contrast, text strings live in a discrete space, and word-level manipulations may significantly change the meaning of the text. In this scenario, an adversary should change as few words as possible, and hence this limitation induces a sparse constraint on word-level changes. Likewise, attacking a classifier should also be much easier than attacking a model with sequence outputs. This is because different from the classification problem that has a finite set of discrete class labels, the output space of sequences may have an almost infinite number of possibilities. If we treat each sequence as a label, a targeted attack needs to find a specific one over an enormous number of possible labels, leading to a nearly zero volume in search space. This may explain why most existing works on adversarial attack focus on the image classification task, since its input space is continuous and its output space is finite.

In this paper, we study a harder problem of crafting adversarial examples for sequence-to-sequence (seq2seq) models [6]. This problem is challenging since it combines both aforementioned difficulties, i.e., discrete inputs and sequence outputs with an almost infinite number of possibilities. We choose this problem not only because it is challenging, but also because seq2seq models are widely used in many safety and security sensitive applications, e.g., machine translation [7, 8], text summarization [9], and speech recognition [10], thus measuring its robustness becomes critical. Specifically, we aim to examine the following questions in this study:

1. *Is it possible to slightly modify the inputs of seq2seq models while significantly change their outputs?*

2. *Are seq2seq models more robust than the well-evaluated CNN-based image classifiers?*

We provide an affirmative answer to the first question by developing an effective adversarial attack framework called Seq2Sick. It is an optimization-based framework that aims to learn an input sequence that is close enough to the original sequence while leads to the desired outputs with high confidence. To address the challenges caused by the discrete input space, we propose to use the projected gradient descent method combined with group lasso and gradient regularization. To address the challenges of almost infinite output space, we design some novel loss functions for the tasks of non-overlapping attack and targeted keyword attack. Our experimental results show that the proposed framework yields high success rates in both tasks. However, even if the proposed approach can successfully attack seq2seq models, our answer to the second question is "Yes". Compared with CNN-based classifiers that are highly sensitive to adversarial examples, seq2seq model is intrinsically more robust since it has discrete input space and the output space is exponentially large. As a result, adversarial examples of seq2seq models usually have larger distortions and are more perceptible than the adversarial examples crafted for CNN-based image classifiers.

In summary, our main contributions are as follows:

- We propose Seq2Sick, a novel optimization-based approach to craft adversarial sequences to fool word-level seq2seq models. Seq2Sick provides two types of attacks: non-overlapping attack to change all the words in the output sentence, and targeted keyword attack to insert malicious keywords into the output.
- To handle discrete input space, we propose a modified projected gradient-based method with group lasso regularization to enforce the sparsity in terms of the number of changed words. We also define corresponding loss functions on output sequence for each type of attack.
- We conduct experiments on text summarization and machine translation tasks. By changing less than 3 words in input sequences, we can fool seq2seq models to produce malicious outputs with high success rates. On the other hand, we recognize that seq2seq models are more robust to adversarial examples than the CNN-based image classifiers.

2 Related Work

Most existing work on adversarial attack focuses on CNN-based models, while only a few studies discuss how to attack RNN-based methods. In this section, we briefly review CNN-based attacks, and then highlight the differences between our work and previous attack algorithms designed for RNN-based models.

2.1 Adversarial Attacks on CNN-based Models

[1] first discovered adversarial examples of the CNN in the context of image classification. Since then, a number of approaches have surfaced regarding adversarial attacks on CNNs, such as using gradient-based methods [2, 11, 12, 13, 14, 15], score-based methods [16, 17, 18], transfer-based methods [19], and decision-based methods [20]. In addition to image classification, attacks to other CNN-related tasks have also been actively investigated. Typical examples include semantic segmentation and object detection [21, 22, 23, 24, 25], image captioning [26], and visual QA [27]. Besides, recent studies have also conducted adversarial attacks in the real-world scenarios, including attacking road signs [28, 29, 24, 25], cell-phone cameras [30], robots [31], and generating 3D-printed adversarial objects [32].

2.2 Adversarial Attacks on RNN-based Models

Although lots of works have been done to fool CNN-based models, only a few studies have examined RNN-based models, and the majority of them focus on simple text classification problems. [33] first uses Fast Gradient Sign Method (FGSM) to conduct an attack on RNN/LSTM-based classification problems. In order to generate text adversarial examples, [34] proposes to use reinforcement learning to locate important words that could be deleted in sentiment classification. [35] and [36] generate adversarial sequences by inserting or replacing existing words with typos and synonyms. [37] aims to attack sentiment classification models in a black-box setting. It develops some scoring functions to find the most important words to modify. These approaches differ from our work in that they study simple text classification problems while we focus on the more challenging seq2seq model with sequential outputs. Other than attacking text classifiers, [38] aims to fool reading comprehension systems by adding misleading sentences, which has

a different focus than ours. [39] uses the generative adversarial network (GAN) to craft natural adversarial examples. However, it can only perform the untargeted attack and also suffers from high computational cost.

Our work shares some similarities with [40] as it also studies the problem of attacking seq2seq model. However, [40] is less appealing than our work since (i) they only focus on untargeted attack while our work can handle the more challenging targeted attack, and (ii) their word-level attack method does not work well – it only yields a 22% success rate for the untargeted attack. Indeed, they recognize that attacking word-level seq2seq model is a challenging problem, which has been addressed by this paper.

Table 1: Summary of existing works that are designed to attack RNN models. "BINARY" indicates the attack is for binary classifications, and there is no difference between untargeted and targeted attack in this case. "CLASS" means targeted attack to a specific class. "KEYWORD" means targeted attack to a specific keyword.

METHODS	GRADIENT BASED?	WORD-LEVEL RNN?	SEQUENTIAL OUTPUT?	TARGETED ATTACK?
EBRAHIMI ET AL [40]	√	×	√	CLASS
JIA&LIANG [38]	×	√	×	×
LI ET AL [34]	√	√	×	CLASS
PAPERNOT ET AL [33]	√	×	√	×
GAO ET AL [37]	×	√	×	BINARY
SAMANTA&MEHTA [35]	×	×	×	BINARY
ZHAO ET AL [39]	/	√	√	×
LIANG ET AL [36]	√	×	×	CLASS
SEQ2SICK (OURS)	√	√	√	KEYWORD

Notably, almost all the previous methods are based on greedy search, i.e., at each step, they search for the best word and the best position to replace the previous word. As a result, their search space grows rapidly as the length of input sequence increases. To address this issue, we propose a novel approach that uses group lasso regularization and the projected gradient descent method with gradient regularization to simultaneously search all the replacement positions. Table 1 summarizes the key differences between the proposed framework Seq2Sick and the existing attack methods on RNN-based models.

3 Methodology

3.1 A Revisit to Sequence-to-Sequence Model

Before introducing the proposed algorithms, we first briefly describe the sequence-to-sequence (seq2seq) model [41, 6]. Let $\mathbf{x}_i \in \mathbb{R}^d$ be the embedding vector of each input word, N be the input sequence length, and M be the output sequence length. Let ω be the input vocabulary, and the output word $\mathbf{y}_j \in \nu$ where ν is the output vocabulary. The seq2seq model has an encoder-decoder framework that aims at mapping an input sequence of vectors $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ to the output sequence $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_M\}$. Its encoder first reads the input sequence, then each RNN/LSTM cell computes $\mathbf{h}_t = f(\mathbf{x}_t, \mathbf{h}_{t-1})$, where $\mathbf{x}_t$ is the current input, $\mathbf{h}_{t-1}$ and $\mathbf{h}_t$ represent the previous and current cells' hidden states, respectively. The next step computes the context vector $\mathbf{c}$ using all the hidden layers of cells $\mathbf{h}_1, \ldots, \mathbf{h}_N$, i.e $\mathbf{c} = q(\mathbf{h}_1, \cdots, \mathbf{h}_N)$, where $q(\cdot)$ could be a linear or non-linear function. In this paper, we follow the setting in [6] that $\mathbf{c} = q(\mathbf{h}_1, \cdots, \mathbf{h}_N) = \mathbf{h}_N$.

Given the context vector $\mathbf{c}$ and all the previously words $\{\mathbf{y}_1, \ldots, \mathbf{y}_{t-1}\}$, the decoder is trained to predict the next word $\mathbf{y}_t$. Specifically, the t-th cell in the decoder receives its previous cell's output $\mathbf{y}_{t-1}$ and the context vector $\mathbf{c}$, and then outputs

$$z_t = g(\mathbf{y}_{t-1}, \mathbf{c}) \text{ and } p_t = \text{softmax}(z_t), \tag{1}$$

where g is another RNN/LSTM cell function. $z_t := [z_t^{(1)}, z_t^{(2)}, \ldots, z_t^{(|\nu|)}] \in \mathbb{R}^{|\nu|}$ is a vector of the *logits* for each possible word in the output vocabulary ν.

3.2 Proposed Framework

The problem of crafting adversarial examples against the seq2seq model can be formulated as the following optimization problem:

$$\min_{\boldsymbol{\delta}} \ L(\mathbf{X} + \boldsymbol{\delta}) + \lambda \cdot R(\boldsymbol{\delta}), \tag{2}$$

where $R(\cdot)$ indicates the regularization function to measure the magnitude of distortions. $L(\cdot)$ is the loss function to penalize the unsuccessful attack and it may take different forms in different attack scenarios. A common choice for $R(\boldsymbol{\delta})$ is the ℓ_2 penalty $\|\boldsymbol{\delta}\|_2^2$, but it is, as we will show later, not suitable for attacking seq2seq model. $\lambda > 0$ is the regularization parameter that balances the distortion and attack success rate – a smaller λ will make the attack more likely to succeed but with the price of larger distortion.

In this work, we focus on two kinds of attacks: non-overlapping attack and targeted keywords attack. The first attack requires that the output of the adversarial example shares no overlapping words with the original output. This task is strictly harder than untargeted attack, which only requires that the adversarial output to be somewhat different from the original output [39, 40]. We ignore the task of untargeted attack since it is too trivial for the proposed framework, which can easily achieve a 100% attack success rate. Targeted keywords attack is an even more challenging task than non-overlapping attack. Given a set of targeted keywords, the goal of targeted keywords attack is to find an adversarial input sequence such that all the keywords must appear in its corresponding output. In the following, we respectively introduce the loss functions developed for the two attack approaches.

3.2.1 Non-overlapping Attack

To formally define the non-overlapping attack, we let $\mathbf{s} = \{s_1, \ldots, s_M\}$ be the original output sequence, where s_i denotes the location of the i-th word in the output vocabulary ν. $\{z_1, \ldots, z_M\}$ indicates the logit layer outputs of the adversarial example. In the non-overlapping attack, the output of adversarial example should be entirely different from the original output $\mathbf{S}$, i.e.,

$$s_t \neq \arg\max_{y \in \nu} z_t^{(y)}, \ \ \forall t = 1, \ldots, M,$$

which is equivalent to

$$z_t^{(s_t)} < \max_{y \in \nu, \ y \neq s_t} z_t^{(y)}, \ \ \forall t = 1, \ldots, M.$$

Given this observation, we can define a hinge-like loss function L to generate adversarial examples in the non-overlapping attack, i.e.,

$$L_{\text{non-overlapping}} = \sum_{t=1}^{M} \max\{-\epsilon, \ z_t^{(s_t)} - \max_{y \neq s_t}\{z_t^{(y)}\}\}, \tag{3}$$

where $\epsilon \geq 0$ denotes the confidence margin parameter. Generally speaking, a larger ϵ will lead to a more confident output and a higher success rate, but with the cost of more iterations and longer running time.

We note that non-overlapping attack is much more challenging than untargeted attack, which suffices to find a one-word difference from the original output [39, 40]. We do not take untargeted attack into account since it is straightforward and the replaced words could be some less important words such as "the" and "a".

3.2.2 Targeted Keywords Attack

Given a set of targeted keywords, the goal of targeted keywords attack is to generate an adversarial input sequence to ensure that all the targeted keywords appear in the output sequence. This task is important since it suggests adding a few malicious keywords can completely change the meaning of the output sequence. For example, in the machine translation task that translates from English into German, an input sentence "*policeman helps protesters to keep the assembly in order*" should generate an output sentence "*Polizist hilft Demonstranten, die Versammlung in Ordnung zu halten*". However, changing only one word from "hilft" to "verhaftet" in the output sequence will significantly change its meaning, as the new sentence means "*police officer arrested protesters to keep the assembly in order*". Similarly, if we change the word "warm" to "cold" in the output sentence "*people gives a warm welcome to soldiers*", its meaning will also be dramatically changed.

In our method, we do not specify the positions of the targeted keywords in the output sentence. Instead, it is more natural to design a loss function that allows the targeted keywords to become the top-1 prediction at any positions. The

attack is considered as successful only when ALL the targeted keywords appear in the output sequence. Therefore, the more targeted keywords there are, the harder the attack is. To illustrate our method, we start from the simpler case with only one targeted keyword k_1. To ensure that the target keyword word's logit $z_t^{(k_1)}$ be the largest among all the words at a position t, we design the following loss function:

$$L = \min_{t \in [M]} \{ \max\{ -\epsilon, \max_{y \neq k_1} \{ z_t^{(y)} \} - z_t^{(k_1)} \} \}, \tag{4}$$

which essentially searches the minimum of the hinge-like loss terms over all the possible locations $t \in [M]$. When there exist more than one targeted keywords $K = \{ k_1, k_2, \ldots, k_{|K|} \}$, where k_i denotes the i-th word in output vocabulary ν, we follow the same idea to define the loss function as follows:

$$L_{\text{targeted keywords}} = \sum_{i=1}^{|K|} \min_{t \in [M]} \{ \max\{ -\epsilon, \max_{y \neq k_i} \{ z_t^{(y)} \} - z_t^{(k_i)} \} \}. \tag{5}$$

However, the loss defined in (5) suffers from the "keyword collision" problem. When there are more than one keyword, it is possible that multiple keywords compete at the same position to attack. To address this issue, we define a mask function m to mask off the position if it has been already occupied by one of the targeted keywords:

$$m_t(x) = \begin{cases} +\infty & \text{if } \arg\max_{i \in \nu} z_t^{(i)} \in K \\ x & \text{otherwise} \end{cases} \tag{6}$$

In other words, if any of the keywords appear at position t as the top-1 word, we ignore that position and only consider other positions for the placement of remaining keywords. By incorporating the mask function, the final loss for targeted keyword attack becomes:

$$\sum_{i=1}^{|K|} \min_{t \in [M]} \{ m_t(\max\{ -\epsilon, \max_{y \neq k_i} \{ z_t^{(y)} \} - z_t^{(k_i)} \}) \}. \tag{7}$$

3.3 Handling Discrete Input Space

As mentioned before, the problem of "discrete input space" is one of the major challenges in attacking seq2seq model. Let $\mathbb{W}$ be the set of word embeddings of all words in the input vocabulary. A naive approach is to first learn $\mathbf{X} + \boldsymbol{\delta}^*$ in the continuous space by solving the problem (2), and then search for its nearest word embedding in $\mathbb{W}$. This idea has been used in attacking sequence classification models in [42]. Unfortunately, when applying this idea to targeted keywords attack, we report that all of the 100 attacked sequences on Gigaword dataset failed to generate the targeted keywords. The main reason is that by directly solving the problem (2), the final solution is likely not a feasible word embedding in $\mathbb{W}$, and its nearest neighbor could be far away from it due to the curse of dimensionality [43].

To address this issue, we propose to add an additional constraint to enforce that $\mathbf{X} + \boldsymbol{\delta}$ belongs to the input vocabulary $\mathbb{W}$. The optimization problem then becomes

$$\begin{aligned} \min_{\boldsymbol{\delta}} \quad & L(\mathbf{X} + \boldsymbol{\delta}) + \lambda \cdot R(\boldsymbol{\delta}) \\ \text{s.t.} \quad & \mathbf{x}_i + \delta_i \in \mathbb{W} \quad \forall i = 1, \ldots, N \end{aligned} \tag{8}$$

We then apply a projected gradient descent method to solve this constrained problem. At each iteration, we project the current solution $\mathbf{x}_i + \delta_i$, where δ_i denotes the i-th column of $\boldsymbol{\delta}$, back into $\mathbb{W}$ to make sure that $\mathbf{X} + \boldsymbol{\delta}$ can map to a specific input word.

Group lasso Regularization: ℓ_2 norm has been widely used in the adversarial machine learning literature to measure distortions. However, it is not suitable for our task since almost all the learned $\{\delta_t\}_{t=1}^M$ using ℓ_2 regularization will be nonzero. As a result, most of the inputs words will be perturbed to another word, leading to an adversarial sequence that is significantly different from the input sequence.

To solve this problem, we treat each δ_t with d variables as a group, and use the group lasso regularization

$$R(\boldsymbol{\delta}) = \sum_{t=1}^N \| \delta_t \|_2$$

to enforce the group sparsity: only a few groups (words) in the optimal solution $\boldsymbol{\delta}^*$ are allowed to be nonzero.

3.4 Gradient Regularization

When attacking the seq2seq model, it is common to find that the adversarial example is located in a region with very few or even no embedding vector. This will negatively affect our projected gradient method since even the closest embedding from those regions can be far away.

To address this issue, we propose a gradient regularization to make $\mathbf{X} + \boldsymbol{\delta}$ close to the word embedding space. Our final objective function becomes:

$$\min_{\boldsymbol{\delta}} \quad L(\mathbf{X} + \boldsymbol{\delta}) + \lambda_1 \sum_{i=1}^{N} \|\delta_i\|_2 + \lambda_2 \sum_{i=1}^{N} \min_{\mathbf{w}_j \in \mathbb{W}} \{\|\mathbf{x}_i + \delta_i - \mathbf{w}_j\|_2\}$$
$$\text{s.t.} \quad \mathbf{x}_i + \delta_i \in \mathbb{W} \quad \forall i = 1, \ldots, N \tag{9}$$

where the third term is our gradient regularization that penalizes a large distance to the nearest point in $\mathbb{W}$. The gradient of this term can be efficiently computed since it is only related to one $\mathbf{w}_j$ that has a minimum distance from $\mathbf{x}_i + \delta_i$. For the other terms, we use the proximal operator to optimize the group lasso regularization, and the gradient of the loss function L can be computed through back-propagation. The detailed steps of our approach, Seq2Sick, is presented in Algorithm 1. Our source code is publicly available.[1]

Computational Cost: Our algorithm needs only one back-propagation to compute the gradient $\nabla_{\boldsymbol{\delta}} L(x + \boldsymbol{\delta})$. The bottleneck here is to project the solution back into the word embedding space, which depends on the number of words in the input dictionary of the model. [42] uses $GloVe$ word embedding [44] that contains millions of words to do a nearest neighbor search. Fortunately, our model does not need to use any pre-trained word embedding, thus making it a more generic attack that does not depend on pre-trained word embedding. Besides, we can employ approximate nearest neighbor (ANN) approaches to further speed up the projection step.

4 Experiments

In this section, we conduct experiments on two widely-used applications of seq2seq model: text summarization and machine translation.

4.1 Datasets

We use three datasets DUC2003[2], DUC2004[3], and Gigaword[4], to conduct our attack for the text summarization task. Among them, DUC2003 and DUC2004 are widely-used datasets in documentation summarization. We also include a subset of randomly chosen samples from Gigaword to further evaluate the performance of our algorithm. For the machine translation task, we use 500 samples from WMT'16 Multimodal Translation task[5]. The statistics about the datasets are shown in Table 2. All our experiments are performed on an Nvidia GTX 1080 Ti GPU.

Table 2: Statistics of the datasets. "# Samples" is the number of test examples we used for robustness evaluations

DATASETS	# SAMPLES	AVERAGE INPUT LENGTHS
GIGAWORD	1,000	30.1 WORDS
DUC2003	624	35.5 WORDS
DUC2004	500	35.6 WORDS
MULTI30K	500	11.5 WORDS

---

[1] https://github.com/cmhcbb/Seq2Sick
[2] http://duc.nist.gov/duc2003/tasks.html
[3] http://duc.nist.gov/duc2004/
[4] https://catalog.ldc.upenn.edu/LDC2003T05
[5] http://www.statmt.org/wmt16/translation-task.html

Algorithm 1 Seq2Sick algorithm

Input: input sequence $\mathbf{x} = \{x_1, \ldots, x_N\}$, seq2seq model, target keyword $\{k_1, \ldots, k_T\}$
Output: adversarial sequence $\mathbf{x}^* = \mathbf{x} + \boldsymbol{\delta}^*$
Let $\mathbf{s} = \{s_1, \ldots, s_M\}$ denote the original output of $\mathbf{x}$.
if Targeted Keyword Attack **then**
    Set the loss $L(\cdot)$ in (9) to be (7)
else
    Set the loss $L(\cdot)$ in (9) to be (3)
end if
for $r = 1, 2, \ldots, T$ **do**
    back-propagation loss L to achieve gradient $\nabla_{\boldsymbol{\delta}} L(\mathbf{x} + \boldsymbol{\delta}_r)$
    for $i = 1, 2, \ldots, N$ **do**
      if $\left\| \delta_{r,i} \right\| > \eta \lambda_1$ **then**
        $\delta_{r,i} = \delta_{r,i} - \eta \lambda_1 \frac{\delta_{r,i}}{\left\| \delta_{r,i} \right\|}$
      else
        $\delta_{r,i} = 0$
      end if
    end for
    $y^{r+1} = \boldsymbol{\delta}^r + \eta \cdot \nabla_{\boldsymbol{\delta}} L(\mathbf{x} + \boldsymbol{\delta}^r)$
    $\boldsymbol{\delta}^{r+1} = \underset{\mathbf{x} + \boldsymbol{\delta}^{r+1} \in \mathbb{W}}{\arg\min} \left\| y^{r+1} - \boldsymbol{\delta}^{r+1} \right\|$
end for
$\boldsymbol{\delta}^* = \boldsymbol{\delta}^T$
$\mathbf{x}^* = \mathbf{x} + \boldsymbol{\delta}^*$
return $\mathbf{x}^*$

4.2 Seq2seq models

We implement both text summarization and machine translation models on OpenNMT-py[6]. Specifically, we use a word-level LSTM encoder and a word-based attention decoder for both applications [7]. For the text summarization task, we use 380k training pairs from Gigaword dataset to train a seq2seq model. The architecture consists of a 2-layer stacked LSTM with 500 hidden units. We conduct experiments on two types of models, one uses the pre-trained 300-dimensional GloVe word embeddings and the other one is trained from scratch. For the machine translation task, we train our model using 453k pairs from the Europal corpus of German-English WMT 15[7], common crawl and news-commentary. We use the hyper-parameters suggested by OpenNMT for both models, and have reproduced the performance reported in [9] and [45].

4.3 Empirical Results

4.3.1 Text Summarization

For the non-overlapping attack, we use the proposed loss (3) in our objective function. A non-overlapping attack is treated as successful only if there is no common word at every position between output sequence and original sequence. We set $\lambda = 1$ in all non-overlapping experiments. Table 3 summarizes the experimental results. It shows that our algorithm only needs to change 2 or 3 words on average and can generate entirely different outputs for more than 80% of sentences. We have also included some adversarial examples in Table 9. More non-overlapping adversarial examples is available at appendix. From these examples, we can only change one word to let output sequence look completely different with the original one and change the sentence's meaning completely.

    For the targeted keywords attack, we randomly choose some targeted keywords from the output vocabulary after removing the stop words like "a" and "the". A targeted keywords attack is treated as successful only if the output sequence contains all the targeted keywords. We set $\lambda_1 = \lambda_2 = 1$ in our objective function (9) in all our experiments.

---

[6]`https://github.com/OpenNMT/OpenNMT-py`
[7]`http://www.statmt.org/wmt15/translation-task.html`

Table 4 summarizes the performance, including the overall success rate, average BLEU score [46], and the average number of changed words in input sentences. Average BLEU score is defined by exponential average over BLEU 1,2,3,4, which is commonly used in evaluating the quality of text which has been machine-translated from one natural language to another. Also, we have included some adversarial examples crafted by our method in Table 10. In Table 10, we show some adversarial examples with 3 sets of keywords, where "##" stands for a two-digit number after standard preprocessing in text summarization. Through these examples, we can see that our method could generate totally irrelevant subjects, verbs, numerals and objects which could easily be formed as a complete sentence with only several word changes. More adversarial examples can be found in the appendix of this paper.

Note that there are three important techniques used in our algorithm: projected gradient method, group lasso, and gradient regularization. In the following, we conduct experiments to verify the importance of each of these techniques. First, without using the projected gradient method, we observe that the success rate immediately drops to close to 0%. It is thus important to project the solution back to the input vocabulary's word embeddings after each iteration. Table 5 shows the experimental results when group lasso or gradient regularization is removed, where "W/O GL" indicates the results without group lasso regularization and "W/O GR" indicates the results without gradient regularization. The algorithms are evaluated using the targeted 2-keyword attack. As shown in Table 5, if we do not use group lasso regularization that enforces group sparsity of distortion δ^*, the attack success rate remains similar, but both average BLEU score and the number of changed words become much larger. On the other hand, if we do not use gradient regularization that helps find a better δ that is close to the input embedding space, not only success rate drops, but also average BLEU score and the number of changed words become worse. These empirical results verify that all the three techniques used here are important in generating adversarial examples against seq2seq models.

Table 3: Results of non-overlapping attack in text summarization. The high BLEU scores and low average number of changed words indicate that the crafted adversarial inputs are very similar to their originals, and we achieve high success rates to generate a summarization that differs with the original *at every position* for all three datasets.

DATASET	SUCCESS RATE	BLEU	# CHANGED
GIGAWORD	86.0%	0.828	2.17
DUC2003	85.2%	0.774	2.90
DUC2004	84.2%	0.816	2.50

Table 4: Results of targeted keywords attack in text summarization. $|K|$ is the number of keywords. We found that our method can make the summarization include 1 or 2 target keywords with a high success rate, while the changes made to the input sentences are relatively small, as indicated by the high BLEU scores and low average number of changed words. When $|K| = 3$, this task becomes more challenging, but our algorithm can still find many adversarial examples.

DATASEST	$|K|$	SUCCESS RATE	BLEU	# CHANGED
	1	99.8%	0.801	2.04
GIGAWORD	2	96.5%	0.523	4.96
	3	43.0%	0.413	8.86
	1	99.6%	0.782	2.25
DUC2003	2	87.6%	0.457	5.57
	3	38.3%	0.376	9.35
	1	99.6%	0.773	2.21
DUC2004	2	87.8%	0.421	5.1
	3	37.4%	0.340	9.3

Table 5: Importance of each component of our algorithm for targeted 2-keyword attack int text summarization. GL = Group lasso. GR = Gradient regularization. We do not show our method without gradient projection as its success rates drop to 0 for all datasets. Group lasso is crucial for crafting adversarial examples which are similar to originals (high BLEU and low number of changed words), and gradient regularization can help to boost success rate.

DATASET	METHOD	SUCCESS%	BLEU	# CHANGED
	W/O GL	91.4 %	0.166	16.53
GIGAWORD	W/O GR	92.8 %	**0.707**	**4.96**
	ALL	**96.5%**	0.523	**4.96**
	W/O GL	95.7%	0.225	15.74
DUC2003	W/O GR	**87.9%**	0.457	**5.57**
	ALL	87.6%	**0.457**	**5.57**
	W/O GL	95.0%	0.212	15.60
DUC2004	W/O GR	87.0 %	**0.421**	**5.14**
	ALL	**87.8%**	**0.421**	**5.14**

4.3.2 Machine Translation

We then conduct both non-overlapping and targeted keywords attacks to the English-German machine translation model. We first filter out stop words like "Ein"(a), "und"(and) in German vocabulary and randomly choose several nouns, verbs, adjectives or adverbs in German as targeted keywords. Similar to the text summarization experiments, we set $\lambda_1 = \lambda_2 = 1$ in our objective function. The success rates, BLEU scores, and the average number of words changed are reported in Table 6, with some adversarial examples shown in Table 8. Furthermore, the effects of group lasso and gradient regularizations are reported in Table 7. These results are consistent with the findings in the text summarization experiments, and verify the effectiveness of proposed attack framework and the importance of the proposed techniques.

Table 6: Results of non-overlapping method and targeted keywords method in machine translation. Similar to our observation for text summarization, non-overlapping attack and targeted keyword attack with 1 and 2 keywords can be achieved with very high success rates, while a 3-keyword attack is more challenging.

METHOD	SUCCESS%	BLEU	# CHANGED
NON-OVERLAP	89.4%	0.349	3.5
1-KEYWORD	100.0%	0.705	1.8
2-KEYWORD	91.0 %	0.303	4.0
3-KEYWORD	69.6%	0.205	5.3

4.4 Robustness of Seq2Seq Model

Finally, we summarize our results and make some final remarks about the robustness of seq2seq models. Our algorithm can achieve very good success rates ($84\% - 100\%$) in both non-overlapping and targeted keywords attacks with 1 or 2 keywords. This clearly verifies the effectiveness of our attack algorithm. On the other hand, we also recognize some strengths of the seq2seq model: (i) unlike CNN models where targeted attack can be conducted easily with almost 100% success rate and very small distortion that cannot be perceived by human eyes [13], it is harder to turn the entire seq2seq output into a particular sentence – some sentences are even impossible to generate by seq2seq models; and (ii) since the input space of seq2seq is discrete, it is easier for human to detect the differences between the adversarial sequence and the original one, even if we only change one or few words. Therefore, we conclude that, compared with the DNN models designed for other tasks such as image classification, seq2seq models are more robust to adversarial attacks. The main reason, as pointed out in the introduction, is that the seq2seq model has a finite and discrete input

Table 7: Importance of each component of our algorithm for targeted 2-keyword attack in machine translation. GL = Group lasso; GR = Gradient regularization. We do not show our method without gradient projection as its success rates drop to 0 for all datasets. Without group lasso, our algorithm can make the input sentence quite different from its original (low BLEU and high number of changed words). Gradient regularization can slightly improve results.

METHOD	SUCCESS RATE	BLEU	# CHANGED
W/O GL	**100.0%**	0.163	6.4
W/O GR	91.0%	**0.303**	4.1
ALL	91.0%	**0.303**	**4.0**

Table 8: Machine translation adversarial examples with targeted keyword "Hund sitzt"

SOURCE INPUT SEQ	A TODDLER IS COOKING WITH ANOTHER PERSON.
ADV INPUT SEQ	A **dog** IS **sit** WITH ANOTHER **UNK**.
SOURCE OUTPUT SEQ	EIN KLEINES KIND KOCHT MIT EINER ANDEREN PERSON.
ADV OUTPUT SEQ	EIN **Hund sitzt** MIT EINEM ANDEREN UNK.

space and almost infinite output space, so it is more robust than visual classification models that have an infinite and continuous input space and a very small output space (e.g., 10 categories in MNIST and 1,000 categories in ImageNet).

5 Conclusion

In this paper, we propose a novel framework, i.e., Seq2Sick, to generate adversarial examples for sequence-to-sequence neural network models. We propose a projected gradient method to address the issue of discrete input space, adopt group lasso to enforce the sparsity of the distortion, and develop a regularization technique to further improve the success rate. Besides, different from most existing algorithms that are designed for untargeted attack and classification tasks, our algorithm can perform the more challenging targeted keywords attack. Our experimental results show that the proposed framework is powerful and effective: it can achieve high success rates in both non-overlapping and targeted keywords attacks with relatively small distortions. At the same time, we recognize that, compared with the DNN models for image classifications that have already been evaluated elsewhere, seq2seq models are indeed more robust to adversarial attacks.

6 Acknowledgment

Table 9: Text summarization adversarial examples using non-overlapping method. Surprisingly, it is possible to make the output sequence completely different by changing only one word in the input sequence.

SOURCE INPUT SEQ	AMONG ASIA 'S LEADERS , PRIME MINISTER MAHATHIR MOHAMAD WAS NOTABLE AS A MAN WITH A BOLD VISION : A PHYSICAL AND SOCIAL TRANSFORMATION THAT WOULD PUSH THIS NATION INTO THE FOREFRONT OF WORLD AFFAIRS .
ADV INPUT SEQ	AMONG **lynn** 'S LEADERS , PRIME MINISTER MAHATHIR MOHAMAD WAS NOTABLE AS A MAN WITH A BOLD VISION : A PHYSICAL AND SOCIAL TRANSFORMATION THAT WOULD PUSH THIS NATION INTO THE FOREFRONT OF WORLD AFFAIRS.
SOURCE OUTPUT SEQ	ASIA 'S LEADERS ARE A MAN OF THE WORLD
ADV OUTPUT SEQ	**a vision for the world**
SOURCE INPUT SEQ	UNDER NATO THREAT TO END HIS PUNISHING OFFENSIVE AGAINST ETHNIC ALBANIAN SEP-ARATISTS IN KOSOVO , PRESIDENT SLOBODAN MILOSEVIC OF YUGOSLAVIA HAS ORDERED MOST UNITS OF HIS ARMY BACK TO THEIR BARRACKS AND MAY WELL AVOID AN ATTACK BY THE ALLIANCE , MILITARY OBSERVERS AND DIPLOMATS SAY
ADV INPUT SEQ	UNDER NATO THREAT TO END HIS PUNISHING OFFENSIVE AGAINST ETHNIC ALBANIAN SEPA-RATISTS IN KOSOVO , PRESIDENT SLOBODAN MILOSEVIC OF YUGOSLAVIA HAS **jean-sebastien** MOST UNITS OF HIS ARMY BACK TO THEIR BARRACKS AND MAY WELL AVOID AN ATTACK BY THE ALLIANCE , MILITARY OBSERVERS AND DIPLOMATS SAY.
SOURCE OUTPUT SEQ	MILOSEVIC ORDERS ARMY BACK TO BARRACKS
ADV OUTPUT SEQ	**nato may not attack kosovo**

Table 10: Text summarization examples with targeted keywords "police arrest", "police arrest ##", "police arrest protesters"

SOURCE INPUT SEQ	NORTH KOREA IS ENTERING ITS FOURTH WINTER OF CHRONIC FOOD SHORTAGES WITH ITS PEOPLE MALNOURISHED AND AT RISK OF DYING FROM NORMALLY CURABLE ILLNESSES , SENIOR RED CROSS OFFICIALS SAID TUESDAY.
ADV INPUT SEQ	NORTH **detectives** IS **apprehended** ITS FOURTH WINTER OF CHRONIC FOOD SHORTAGES WITH ITS PEOPLE MALNOURISHED AND AT RISK OF DYING FROM NORMALLY CURABLE ILLNESSES , SENIOR RED CROSS OFFICIALS SAID TUESDAY.
SOURCE OUTPUT SEQ	NORTH KOREA ENTERS FOURTH WINTER OF FOOD SHORTAGES
ADV OUTPUT SEQ	NORTH **police arrest** FOURTH WINTER OF FOOD SHORTAGES.
SOURCE INPUT SEQ	AFTER A DAY OF FIGHTING , CONGOLESE REBELS SAID SUNDAY THEY HAD ENTERED KINDU , THE STRATEGIC TOWN AND AIRBASE IN EASTERN CONGO USED BY THE GOVERNMENT TO HALT THEIR ADVANCES.
ADV INPUT SEQ	AFTER A DAY OF FIGHTING , **nordic apprehended** SAID SUNDAY THEY HAD ENTERED **UNK** , THE STRATEGIC TOWN AND AIRBASE IN EASTERN CONGO USED BY THE GOVERNMENT TO HALT THEIR ADVANCES.
SOURCE OUTPUT SEQ	CONGOLESE REBELS SAY THEY HAVE ENTERED UNK.
ADV OUTPUT SEQ	NORDIC **police arrest ##** IN CONGO.
SOURCE INPUT SEQ	THREE WEEKS AGO , WHEN SOMEONE STARTED STALKING HOMELESS PEOPLE HERE , BEND-ING OVER THEM AS THEY SLEPT IN THE ALLEYS AND DOORWAYS AND SLASHING THEIR THROATS , A GRIM JOKE BEGAN MAKING THE ROUNDS ON THE STREETS : THE KILLER HAD BEEN HIRED BY THE CITY.
ADV INPUT SEQ	THREE WEEKS AGO **dissidents** WHEN SOMEONE STARTED **modi** HOMELESS PEOPLE HERE **thai** BENDING OVER THEM AS THEY SLEPT IN THE ALLEYS AND DOORWAYS AND SLASHING THEIR THROATS , A GRIM JOKE BEGAN MAKING THE ROUNDS ON THE STREETS : THE KILLER HAD BEEN HIRED BY THE CITY.
SOURCE OUTPUT SEQ	HOMELESS PEOPLE GET A TASTE OF LIFE
ADV OUTPUT SEQ	THAI **police arrest #** IN CRACKDOWN ON **protesters**

References

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013.

[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[3] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[4] Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. *arXiv preprint arXiv:1705.08475*, 2017.

[5] Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. *arXiv preprint arXiv:1801.10578*, 2018.

[6] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112, 2014.

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[8] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[9] Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*, 2015.

[10] William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, pages 4960–4964, 2016.

[11] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016.

[12] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P) (2016)*, pages 372–387. IEEE, 2016.

[13] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pages 39–57. IEEE, 2017.

[14] Moustapha Cissé, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: Fooling deep structured visual and speech recognition models with adversarial examples. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6980–6990, 2017.

[15] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Ead: Elastic-net attacks to deep neural networks via adversarial examples. *arXiv preprint arXiv:1709.04114*, 2017.

[16] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26. ACM, 2017.

[17] Jamie Hayes and George Danezis. Machine learning as an adversarial service: Learning black-box adversarial examples. *arXiv preprint arXiv:1708.05207*, 2017.

[18] Nina Narodytska and Shiva Prasad Kasiviswanathan. Simple black-box adversarial perturbations for deep networks. *arXiv preprint arXiv:1612.06299*, 2016.

[19] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pages 506–519. ACM, 2017.

[20] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017.

[21] Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, and Volker Fischer. Universal adversarial perturbations against semantic image segmentation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2774–2783, 2017.

[22] Anurag Arnab, Ondrej Miksik, and Philip H. S. Torr. On the robustness of semantic segmentation models to adversarial attacks. *CoRR*, abs/1711.09856, 2017.

[23] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan L. Yuille. Adversarial examples for semantic segmentation and object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1378–1387, 2017.

[24] Jiajun Lu, Hussein Sibai, and Evan Fabry. Adversarial examples that fool detectors. *CoRR*, abs/1712.02494, 2017.

[25] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Dawn Song, Tadayoshi Kohno, Amir Rahmati, Atul Prakash, and Florian Tramèr. Note on attacking object detectors with adversarial stickers. *CoRR*, abs/1712.08062, 2017.

[26] Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh. Show-and-fool: Crafting adversarial examples for neural image captioning. *arXiv preprint arXiv:1712.02051*, 2017.

[27] Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darell, and Dawn Song. Can you fool AI with adversarial examples on a visual turing test? *CoRR*, abs/1709.08693, 2017.

[28] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on machine learning models. *CoRR*, abs/1707.08945, 2017.

[29] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *CoRR*, abs/1708.06733, 2017.

[30] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *CoRR*, abs/1607.02533, 2016.

[31] Marco Melis, Ambra Demontis, Battista Biggio, Gavin Brown, Giorgio Fumera, and Fabio Roli. Is deep learning safe for robot vision? adversarial examples against the icub humanoid. In *2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22-29, 2017*, pages 751–759, 2017.

[32] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. *CoRR*, abs/1707.07397, 2017.

[33] Nicolas Papernot, Patrick McDaniel, Ananthram Swami, and Richard Harang. Crafting adversarial input sequences for recurrent neural networks. In *Military Communications Conference, MILCOM 2016-2016 IEEE*, pages 49–54. IEEE, 2016.

[34] Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*, 2016.

[35] Suranjana Samanta and Sameep Mehta. Towards crafting text adversarial samples. *arXiv preprint arXiv:1707.02812*, 2017.

[36] Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. Deep text classification can be fooled. *arXiv preprint arXiv:1704.08006*, 2017.

[37] Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. Black-box generation of adversarial text sequences to evade deep learning classifiers. *arXiv preprint arXiv:1801.04354*, 2018.

[38] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*, 2017.

[39] Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. *arXiv preprint arXiv:1710.11342*, 2017.

[40] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for nlp. *arXiv preprint arXiv:1712.06751*, 2017.

[41] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[42] Zhitao Gong, Wenlu Wang, Bo Li, Dawn Song, and Wei-Shinn Ku. Adversarial texts with gradient methods. *arXiv preprint arXiv:1801.07175*, 2018.

[43] Jerome H Friedman. On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data mining and knowledge discovery*, 1(1):55–77, 1997.

[44] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[45] Thanh-Le Ha, Jan Niehues, and Alexander Waibel. Toward multilingual neural machine translation with universal encoder and decoder. *arXiv preprint arXiv:1611.04798*, 2016.

[46] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

A   Appendix

A.1   More adversarial examples on target keyword method

In Table 11, we show more adversarial examples generated by our Seq2Sick framework using targeted keywords method.

A.2   More adversarial examples on non-overlapping method

In Table 12, we show more adversarial examples generated by our Seq2Sick framework using non-overlapping method. By changing only one word in the input, we can fool the seq2seq to generate totally different output.

Table 11: Text summarization adversarial examples using targeted keywords as "police arrest","police arrest protestors"

SOURCE INPUT SEQ	CAMBODIA 'S TWO-PARTY OPPOSITION ASKED THE ASIAN DEVELOPMENT BANK MONDAY TO STOP PROVIDING LOANS TO THE INCUMBENT GOVERNMENT , WHICH IT CALLS ILLEGAL
ADV INPUT SEQ	CAMBODIA 'S **heavy-handed manhunts warrant** THE ASIAN DEVELOPMENT BANK MONDAY TO STOP PROVIDING LOANS TO THE INCUMBENT GOVERNMENT, **rwanda apec** CALLS **police**
SOURCE OUTPUT SEQ	CAMBODIAN OPPOSITION ASKS ADB TO STOP LOANS TO GOVERNMENT
ADV OUTPUT SEQ	CAMBODIA 'S **police arrest** ASIAN DEVELOPMENT BANK
SOURCE INPUT SEQ	EUROPE 'S TOP OFFICIAL SAID FRIDAY THAT HE HOPED FORMER CHILEAN DICTATOR AUGUSTO PINOCHET WOULD BE EXTRADITED TO SPAIN TO BE TRIED FOR CRIMES COMMITTED DURING HIS 17-YEAR RULE.
ADV INPUT SEQ	**detectives sweep** TOP OFFICIAL SAID FRIDAY THAT HE HOPED FORMER CHILEAN DICTATOR AUGUSTO PINOCHET WOULD BE EXTRADITED TO SPAIN TO BE TRIED FOR CRIMES COMMITTED DURING HIS **UNK** RULE.
SOURCE OUTPUT SEQ	EUROPE 'S TOP OFFICIAL HOPES PINOCHET WILL BE EXTRADITED TO SPAIN.
ADV OUTPUT SEQ	CHILEAN **police arrest** PINOCHET 'S EXTRADITION.
SOURCE INPUT SEQ	PRIME MINISTER RAFIK HARIRI HAS DECLINED AN INFORMAL INVITATION FROM LEBANON 'S NEW PRESIDENT TO FORM THE NEXT GOVERNMENT , SPARKING A POLITICAL CRISIS IN THIS COUNTRY AS IT REBUILDS FROM ITS DEVASTATING CIVIL WAR .
ADV INPUT SEQ	PRIME MINISTER RAFIK HARIRI **apprehended** DECLINED **riot-torn** INFORMAL INVITATION FROM LEBANON 'S NEW PRESIDENT TO FORM THE NEXT GOVERNMENT , SPARKING A POLITICAL CRISIS **ex-police** THIS COUNTRY AS IT **UNK** FROM ITS DEVASTATING CIVIL WAR .
SOURCE OUTPUT SEQ	LEBANESE PRIME MINISTER DECLINES TO FORM NEW GOVERNMENT
ADV OUTPUT SEQ	LEBANESE **police arrest ##** IN NEW GOVERNMENT
SOURCE INPUT SEQ	THE TRIALS OF THREE OUTSPOKEN DISSIDENTS OVER , COMMUNIST PARTY LEADER JIANG ZEMIN SIGNALED WEDNESDAY THAT CHINA WILL SUSTAIN A CRACKDOWN ON DISSENT THROUGHOUT NEXT YEAR .
ADV INPUT SEQ	**firemen** TRIALS **wanchope consists** OUTSPOKEN DISSIDENTS OVER , COMMUNIST PARTY LEADER JIANG ZEMIN SIGNALED WEDNESDAY THAT CHINA WILL SUSTAIN A CRACKDOWN ON DISSENT THROUGHOUT NEXT YEAR .
SOURCE OUTPUT SEQ	CHINA 'S JIANG SIGNALS CRACKDOWN ON DISSENT
ADV OUTPUT SEQ	CHINESE **police** SAY THEY WILL CONTINUE CRACKDOWN ON DISSENT
SOURCE INPUT SEQ	PRESIDENT BORIS YELTSIN STAYED HOME TUESDAY , NURSING A RESPIRATORY INFECTION THAT FORCED HIM TO CUT SHORT A FOREIGN TRIP AND REVIVED CONCERNS ABOUT HIS ABILITY TO GOVERN.
ADV INPUT SEQ	PRESIDENT BORIS YELTSIN STAYED HOME TUESDAY , **cops cops** RESPIRATORY INFECTION THAT FORCED HIM TO CUT SHORT A FOREIGN TRIP AND REVIVED CONCERNS ABOUT HIS ABILITY TO GOVERN.
SOURCE OUTPUT SEQ	YELTSIN STAYS HOME AFTER ILLNESS
ADV OUTPUT SEQ	YELTSIN STAYS HOME AFTER **police arrest**
SOURCE INPUT SEQ	A GROUP OF HIGH-POWERED U.S. INVESTORS IN SOUTHEAST ASIA ON MONDAY APPLAUDED EFFORTS TO PERK UP THAILAND 'S STAGGERING ECONOMY ,SAYING THEY HAD BEEN ASSURED BY TOP THAI OFFICIALS THAT KEY ECONOMIC REFORM PACKAGES WILL SOON BE APPROVED
ADV INPUT SEQ	A **eds confiscate rafah arresting supporters** IN SOUTHEAST ASIA ON MONDAY APPLAUDED EFFORTS TO PERK UP THAILAND 'S STAGGERING ECONOMY , SAYING THEY HAD BEEN ASSURED BY TOP THAI OFFICIALS THAT KEY ECONOMIC REFORM PACKAGES WILL SOON BE APPROVED .
SOURCE OUTPUT SEQ	US INVESTORS WELCOME THAI ECONOMIC REFORMS
ADV OUTPUT SEQ	THAI **police arrest protesters** IN SOUTHEAST ASIA

15

Table 12: Text summarization adversarial examples using non-overlapping method

SOURCE INPUT SEQ	A GROUP OF HIGH-POWERED U.S. INVESTORS IN SOUTHEAST ASIA ON MONDAY APPLAUDED EFFORTS TO PERK UP THAILAND 'S STAGGERING ECONOMY , SAYI NG THEY HAD BEEN ASSURED BY TOP THAI OFFICIALS THAT KEY ECONOMIC REFORM PACKAGES WILL SOON BE APPROVED.
ADV INPUT SEQ	A GROUP OF HIGH-POWERED U.S. **INVESTMENT** IN SOUTHEAST ASIA ON MONDAY APPLAUDED EFFORTS TO PERK UP THAILAND 'S STAGGERING ECONOMY , SAYING THEY HAD BEEN ASSURED BY TOP THAI OFFICIALS THAT KEY ECONOMIC REFORM PACKAGES WILL SOON BE APPROVED.
SOURCE OUTPUT SEQ	US INVESTORS WELCOME THAI ECONOMIC REFORMS.
ADV OUTPUT SEQ	**u.s. investment in southeast asia to be boosted**.
SOURCE INPUT SEQ	JUST FOUR YEARS AGO , IT WAS A GOOD BET THAT NEWT GINGRICH WOULD BE THE PIVOTAL FIGURE IN U.S. POLITICS AT THE TURN OF THE MILLENNIUM.
ADV INPUT SEQ	JUST FOUR YEARS AGO , IT WAS A GOOD BET THAT NEWT **lawmakers** WOULD BE THE PIVOTAL FIGURE IN U.S. POLITICS AT THE TURN OF THE MILLENNIUM .
SOURCE OUTPUT SEQ	GINGRICH 'S UNK VISION IS A GOOD BETMILOSEVIC ORDERS ARMY BACK TO BARRACKS
ADV OUTPUT SEQ	**a new year for the millennium**
SOURCE INPUT SEQ	RUSSIA 'S NEW PRIME MINISTER PICKED AN UNUSUAL WAY TO REASSURE THE NATION THURSDAY , AFTER TWO WEEKS OF DELIBERATIONS HE ANNOUNCED THAT HE STILL HAD NO PLAN TO RESCUE THE COUNTRY FROM ITS ECONOMIC CRISIS .
ADV INPUT SEQ	RUSSIA 'S NEW PRIME MINISTER PICKED AN UNUSUAL WAY TO **peacocks** THE NATION THURSDAY , AFTER TWO WEEKS OF DELIBERATIONS HE ANNOUNCED THAT HE STILL HAD NO PLAN TO RESCUE THE COUNTRY FROM ITS ECONOMIC CRISIS.
SOURCE OUTPUT SEQ	RUSSIA S NEW PRIME MINISTER SAYS HE HAS NO PLAN TO RESCUE COUNTRY FROM ECONOMIC CRISIS
ADV OUTPUT SEQ	**new russian prime minister picks peacocks**
SOURCE INPUT SEQ	AS TURKEY KEPT UP WARLIKE RHETORIC AGAINST DAMASCUS , EGYPT ON SUNDAY BEGAN SHUTTLE DEMOCRACY BETWEEN THE TWO NEIGHBORS TO AVOID A MILITARY CONFRONTATION OVER TURKISH KURDISH BASES IN SYRIA.
ADV INPUT SEQ	AS TURKEY KEPT UP WARLIKE RHETORIC AGAINST DAMASCUS **of** EGYPT ON SUNDAY BEGAN SHUTTLE DEMOCRACY BETWEEN THE TWO NEIGHBORS TO AVOID A MILITARY CONFRONTATION OVER TURKISH KURDISH BASES IN SYRIA.
SOURCE OUTPUT SEQ	EGYPT BEGINS SHUTTLE DIPLOMACY WITH SYRIA
ADV OUTPUT SEQ	**turkey holds up its rhetoric against syria**
SOURCE INPUT SEQ	INDIAN PRIME MINISTER P.V. NARASIMHA RAO 'S PROMISE OF MORE AUTONOMY FOR TROUBLED KASHMIR AND HIS PLEA FOR EARLY STATE ELECTIONS HAS SPARKED A VIOLENT REACTION FROM PROVINCIAL MOSLEM AND OPPOSITION PARTIES.
ADV INPUT SEQ	INDIAN PRIME MINISTER P.V. NARASIMHA RAO **honeywell** PROMISE OF MORE AUTONOMY FOR TROUBLED KASHMIR AND HIS PLEA FOR EARLY STATE ELECTIONS HAS SPARKED A VIOLENT REACTION FROM PROVINCIAL MOSLEM AND OPPOSITION PARTIES.
SOURCE OUTPUT SEQ	RUSSIA S NEW PRIME MINISTER SAYS HE HAS NO PLAN TO RESCUE COUNTRY FROM ECONOMIC CRISIS
ADV OUTPUT SEQ	**new autonomy for troubled kashmir indian pm 's promise of autonomy for kashmir sparks violent reaction**
SOURCE INPUT SEQ	THE MOMENT OF TRUTH FOR CHUCK KNOBLAUCH CAME IN THE BOTTOM OF THE FIRST INNING WHEN HIS NAME WAS ANNOUNCED AT YANKEE STADIUM FOR THE FIRST TIME SINCE HE NEGLECTED TO CHASE DOWN THAT MEMORABLE LOOSE BALL LAST WEDNESDAY.
ADV INPUT SEQ	THE MOMENT OF TRUTH FOR CHUCK **yankees** CAME IN THE BOTTOM OF THE FIRST INNING WHEN HIS NAME WAS ANNOUNCED AT YANKEE STADIUM FOR THE FIRST TIME SINCE HE NEGLECTED TO CHASE DOWN THAT MEMORABLE LOOSE BALL LAST WEDNESDAY.
SOURCE OUTPUT SEQ	KNOBLAUCH 'S RETURN IS A UNK
ADV OUTPUT SEQ	**yankees yankees lose yankees #-#**

16