CHAPTER 6

Adversarial attacks beyond image classification

6.1 Data modality and task objectives

For classification tasks on data modalities that can be presented in real-valued continuous vectors, such as images (pixel values), text embeddings, audio signals, time series, and real-valued tabular data, the methodology of the aforementioned adversarial attacks can be adopted in a principled manner by specifying the corresponding attack loss function and threat model, and then solving the formulated objective using continuous optimization tools such as gradient-based methods.

For classification tasks on noncontinuous data modalities, such as text tokens (e.g., words and characters), graphs, and malwares, the attack objective is similar, but the solver needs to use discrete optimization tools such as genetic algorithms, evolutionary algorithms, relaxation to continuous optimization methods due to the fact that the feasible action space of an attacker will be discrete in nature, such as adding, removing, or editing certain words in a sentence to change the sentiment.

For tasks beyond classification, such as regression, reconstruction, text summarization, generation, etc., the rule of thumb is to define the appropriate threat models and attackers' capabilities for making meaningful analysis of adversarial robustness for the corresponding machine learning models. After the attack objective is formulated, a proper solver will be used to execute the adversarial attack depending on the data modality.

In what follows, we introduce some examples for data modalities and tasks beyond image classification.

6.2 Audio adversarial example

Carlini and Wagner (2018) propose an optimization-based attack for generating audio adversarial examples for automatic speech recognition systems. Given an audio waveform $\boldsymbol{x}$, the task is to construct another audio waveform $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{\delta}$ such that $\boldsymbol{x}$ and $\boldsymbol{x}'$ sound similar but their output transcription differs. To cater to the audio domain, they use decibels

Adversarial Robustness for Machine Learning
https://doi.org/10.1016/B978-0-12-824020-5.00015-6
Copyright © 2023 Elsevier Inc.
All rights reserved.

(dB) as the distortion metric, defined as $dB_{\boldsymbol{x}}(\boldsymbol{\delta}) = dB(\boldsymbol{\delta}) - dB(\boldsymbol{x})$, where $dB(\boldsymbol{x}) = 20 \cdot \max_i \log_{10}(x_i)$. The optimal audio adversarial perturbation $\boldsymbol{\delta}^*$ with a target transcription t can be found by solving the following constrained objective function:

$$\begin{aligned} &\underset{\boldsymbol{\delta}}{\text{minimize}} \ \|\boldsymbol{\delta}\|_2^2 + \lambda \cdot \mathcal{L}(\boldsymbol{x} + \boldsymbol{\delta}, t) \\ &\text{such that } dB_{\boldsymbol{x}}(\boldsymbol{\delta}) \leq \epsilon, \end{aligned} \tag{6.1}$$

where $\mathcal{L}$ can be the connectionist temporal classification (CTC) loss or other improved loss functions proposed by Carlini and Wagner (2018), and λ is its regularization coefficient. Similarly, adversarial examples for time series data can be formulated by using the task-specific loss function and appropriate similarity/distortion metric.

6.3 Feature identification

Chen et al. (2018c) investigate the robustness of sparse regression models with strongly correlated covariates to adversarially designed measurement noises. Specifically, they consider the family of ordered weighted ℓ_1 (OWL) regularized regression methods (Bogdan et al., 2013; Zeng and Figueiredo, 2014a) and study the case of OSCAR (octagonal shrinkage clustering algorithm for regression) in the adversarial setting. It is worth mentioning that OSCAR is in fact a particular case of the OWL regularizer (Zeng and Figueiredo, 2014b). OSCAR is known to be more effective in identifying feature groups (i.e., strongly correlated covariates) than other feature selection methods such as LASSO (Tibshirani, 1996).

Under a norm-bounded threat model, they formulate the process of finding a maximally disruptive noise for OWL-regularized regression as an optimization problem and illustrate the steps toward finding such a noise in the case of OSCAR. Experimental results demonstrate that the regression performance of grouping strongly correlated features can be severely degraded under the studied adversarial setting, even when the noise budget is significantly smaller than the ground-truth signals.

6.4 Graph neural network

Graph structured data play a crucial role in many AI applications. It is an important and versatile representation to model a wide variety of datasets from many domains, such as molecules, social networks, or interlinked

documents with citations. Graph neural networks (GNNs) on graph structured data have shown outstanding results in various applications (Kipf and Welling, 2016). The input data for GNN applications are usually given in the format of a graph (with edges and nodes) and each node can be associated with a d-dimensional feature vector. Therefore, to evaluate the robustness of GNN models, one can consider perturbations to node features and/or edges.

First, we briefly discuss the perturbations on node features. If node features are in a continuous domain, perturbing node features can be easily done by a gradient-based optimizer, similar to the attacks in computer vision introduced in the previous sections. For discrete node features, one can also adopt existing attacks in the literature. For instance, if node features are text, we can borrow attacks from the text domain to attack such GNN models.

On the other hand, attacks on edges cannot be easily done. Conventional (first-order) continuous optimization methods do not directly apply to attacks using edge manipulations (which are called *topology attacks* by Xu et al. (2019a)) due to the discrete nature of graphs. Xu et al. (2019a) close this gap by studying the problem of generating topology attacks via convex relaxation so that gradient-based adversarial attacks become plausible for GNNs. Evaluated on node classification tasks using GNNs, their gradient-based topology attacks outperform current state-of-the-art attacks subject to an edge perturbation budget. Moreover, by leveraging the proposed gradient-based attack, they propose the first optimization-based adversarial training technique for GNNs, yielding significantly improved robustness against gradient-based and greedy topology attacks. Other than optimization-based approaches, different methods are proposed for adversarial attacks on graph neural networks, such as the use of greedy search (Zügner et al., 2018), reinforcement learning (Dai et al., 2018), and meta learning (Zügner and Günnemann, 2019).

6.5 Natural language processing

Natural Language Processing (NLP) has been widely used in many important domains, and the robustness of NLP models is also crucial to mission-critical applications. For instance, when applying machine translation model to real-time machine translation, it is important to make sure the correctness of translation against small input perturbations. Another potentially important area is the evasion attacks to fake news detection (or

spam detection) models; if there exists semantic-preserved perturbations to fool the fake news detection models, malicious users can leverage those perturbations to create fake news while being able to bypass the detection models. Beyond security concerns, adversarial robustness of text models has also been deeply studied recently as finding adversarial examples could be the first step for model debugging.

There are two main challenges when conducting attacks to nature language processing (NLP) models. First, as inputs of NLP are discrete, finding adversarial examples usually leads to a discrete optimization problem instead of the continuous one in computer vision attacks. Second, the "semantic invariance" perturbation in NLP is harder to define than in the computer vision cases. In the image domain, a slight change on each pixel value usually would not be perceptible by human, but this is not true in NLP as changing a word in a sentence may significantly change the semantic meaning of the whole sentence. Therefore adversarial attacks in the NLP domain need to consider these two factors. For handling discrete inputs, we usually seak to discrete optimization algorithms. To construct a semantic-invariance perturbation set, we often resort to either word embedding or contextualized word embeddings. In the following, we briefly introduce several NLP attacks developed for sentence classification and sequence-to-sequence translation. There are many NLP attacks beyond these two applications that are not covered in this book, such as question answering, dialogue systems and semantic parsing.

Sentence classification

Most of the NLP attacks focused on sentence classification, as this is one of the most simple but representative tasks. Earlier attacks to text classification usually define the perturbation set by constraining the number of words or characters changed in the sentence. However, as discussed above, it is possible to change the meaning of the whole sentence by only replacing one word (or character). Therefore more recent attacks usually try to constrain the replaced words to enforce the semantic invariance. In NLP, word embeddings have been developed to represent the semantic meaning of each word by a latent vector, and there are several well-trained word embeddings such as Word2Vec (Mikolov et al., 2013) and Glove (Pennington et al., 2014). Alzantot et al. (2018) used distances in the word embedding space to construct a set of synonyms and used them to define the perturbation set: each word is only allowed to be replaced by its synonyms. More recently, due to the popularity of large-scale pretrained language models

such as BERT (Devlin et al., 2018), many works start to use these language models to define the perturbation set (Li et al., 2020a). More specifically, as BERT provides the probability of the words at each position given the context words (e.g., all the other words in the same sentence), we can assume that those top possible words preserve the "natureness" of the sentence, and combining with synonyms, this can lead to more natural adversarial examples that preserve semantics.

With the perturbation set defined, another challenge for NLP attacks is to find the adversarial examples that maximize the task-specific loss (or other criterion) within the perturbation set. As the input space and perturbation set are discrete, this leads to a discrete optimization problem, which is usually harder to solve than continuous ones. Several discrete search algorithms have been proposed for NLP attacks. For example, evolutionary algorithms and reinforcement learning have been used to search for adversarial examples in the perturbation space (Alzantot et al., 2018; Li et al., 2016). To make the attacks faster, some other works consider simple greedy approaches. For example, Yang et al. (2020e) proposed a greedy attack that iteratively replaces words in the original sentence. At each iteration, they first mask each word in the input sentence and select the best position to attack based on the attack loss function (classification loss), and then go through the synonyms for the selected position to find the best word replacement (also to maximize the classification loss). However, greedy approaches typically do not have theoretical guarantee and may lead to suboptimal solutions.

Lei et al. (2019) formulate the attacks with discrete input on a set function as an optimization task. They prove that this set function is submodular for two types of popular neural network text classifiers under simplifying assumption: the first is a word-level convolutional neural network (CNN) without dropout or softmax layers, and the second is a recurrent neural network (RNN) with one-dimensional hidden units and arbitrary time steps. This finding guarantees a $1 - 1/e$ approximation factor for attacks that use the greedy algorithm. Meanwhile, they also show how to use the gradient of the attacked classifier to guide the greedy search.

With the proposed optimization scheme, they show significantly improved attack performance over most baselines. Meanwhile, they also propose a joint sentence and word paraphrasing technique to simultaneously ensure retention of the semantics and syntax of the text, known as the *paraphrasing attack* (Lei et al., 2019). Interestingly, they also found that in their experiments, under almost all circumstances, model retraining via

augmenting the adversarial examples with correct labels can improve the generalization of the model and make it less susceptible to attack.

Fig. 6.1 shows the text adversarial examples generated by the paraphrasing attack (Lei et al., 2019) for sentiment analysis and fake-news detection using neural networks.

Task: Sentiment Analysis. Classifier: LSTM. Original: 100% Positive. ADV label: 100% Negative.

I suppose I should write a review here since my little Noodle-oo is currently serving as their spokes dog in the photos. We both love Scooby Do's. They treat my little butt-faced dog like a prince and are receptive to correcting anything about the cut that I perceive as being weird. Like that funny poofy pompadour. Mohawk it out, yo. Done. In like five seconds my little man was looking fabulous and bad ass. Not something easily accomplished with a prancing pup that literally chases butterflies through tall grasses. (He ended up looking like a little lamb as the cut grew out too. So adorable.) The shampoo they use here is also amazing. Noodles usually smells like tacos (a combination of beef stank and corn chips) but after getting back from the Do's, he smelled like Christmas morning! Sugar and spice and everything nice instead of frogs and snails and puppy dog tails. He's got some gender identity issues to deal with. ~~The pricing is also cheaper than some of the big name conglomerates out there~~ The price is cheaper than some of the big names below. I'm talking to you Petsmart! I've taken my other pup to Smelly Dog before, but unless I need dog sitting play time after the cut, I'll go with Scooby's. They genuinely seem to like my little Noodle monster.

Task: Fake-News Detection. Classifier: LSTM. Original label: 100% Fake. ADV label: 77% Real

~~Man~~ Guy punctuates high-speed chase with stop at In-N-Out Burger drive-thru Print [Ed.~~— Well, that's~~ Okay, that 's a new one.] ~~A~~ One man is in custody after leading police on a bizarre chase into the east Valley on Wednesday night. Phoenix police ~~began~~ has begun following the suspect in Phoenix and the pursuit continue~~d~~ into the east Valley, but it took a bizarre turn when the suspect stopped at an In-N-Out Burger restaurant's ~~drive-thru~~ drive-through near Priest and Ray Roads in Chandler. The suspect appeared to order food, but then drove away and got out of his pickup truck near Rock Wren Way and Ray Road. He ~~then ran into a backyard~~ ran to the backyard and tried to ~~get into a house through the back door~~ get in the home.

Figure 6.1 Examples of generated adversarial examples using the paraphrasing attack proposed by Lei et al. (2019). The color red (gray in print version) denotes sentence-level paraphrasing, and blue (dark gray in print version) denotes word-level paraphrasing.

Sequence-to-sequence translation

Beyond text classification, Cheng et al. (2020d) study the much more challenging problem of crafting adversarial examples for sequence-to-sequence (seq2seq) models (Sutskever et al., 2014) whose inputs are discrete text strings and outputs have an almost infinite number of possibilities. They propose an effective adversarial attack framework called *Seq2Sick* to address the challenges caused by the discrete input space, a projected gradient method combined with group lasso and gradient regularization. To handle

the almost infinite output space, they design some novel loss functions to conduct nonoverlapping attack and targeted keyword attack. When applying their attack algorithm to machine translation and text summarization tasks, by changing less than three words, they can make seq2seq model to produce desired outputs with high success rates. They also use an external sentiment classifier to verify the property of preserving semantic meanings for the generated adversarial examples. Table 6.1 shows some such examples.

Table 6.1 Text summarization adversarial examples using nonoverlapping method proposed by Cheng et al. (2020d). Surprisingly, it is possible to make the output sequence completely different by changing only one or few words in the input sequence. Red (gray in print version) color indicates changed words.

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 separatists 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 separatists 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**
Source input seq	flooding on the yangtze river remains serious although water levels on parts of the river decreased today, according to the state headquarters of flood control and drought relief.
Adv input seq	flooding that the yangtze river becomes serious although water levels on parts of the river decreased today, according to the state headquarters of flood control and drought relief.
Source output seq	floods on yangtze river continue
Adv output seq	**flooding in water recedes in river**

6.6 Deep reinforcement learning

Deep reinforcement learning (DRL) models are shown to possess many advantages in optimizing robot learning systems, such as autonomous navigation and continuous robot arm control. Yang et al. (2020b) propose timing-based adversarial strategies against a DRL-based navigation system by jamming in physical noise patterns on the selected time frames. To study the vulnerability of learning-based navigation systems, they propose two adversarial agent models: one refers to online learning, and another one is based on evolutionary learning. Under white-box and black-box adversarial settings, their experimental results show that the adversarial timing attacks can lead to a significant performance drop of the target DRL model. Yang et al. (2022) studies different types of obervational interference to Q-learning based DRL models and show that incoporating them into neural network architecture design and training can lead to improved robustness.

6.7 Image captioning

Image captioning is an example of deep learning on mixed data modalities (texts and images). The model input is an image, and the model output is some caption describing the content in the image. Image captioning adopts an encoder-decoder framework consisting of two principal components, a convolutional neural network (CNN) for image feature extraction and a recurrent neural network (RNN) for language caption generation.

Chen et al. (2018a) propose *Show-and-Fool*, a novel algorithm for crafting adversarial examples in neural image captioning. The proposed algorithm provides two evaluation approaches, which check whether neural image captioning systems can be misled to output some randomly chosen or targeted captions or keywords. Their experiments show that their algorithm can successfully craft visually similar adversarial examples with randomly targeted captions or keywords, and the adversarial examples can be made highly transferable to other image captioning systems.

As an illustration, Fig. 6.2 shows adversarial examples crafted by Show-and-Fool using the targeted caption method. The adversarial perturbations are visually imperceptible but can successfully mislead the Show-and-Tell (Vinyals et al., 2015) neural image captioning model to generate the targeted captions. Interestingly and perhaps surprisingly, their results pinpoint the Achilles heel of the language and vision models used in the tested image captioning systems. Moreover, the adversarial examples in neural image captioning highlight the inconsistency in visual language grounding be-

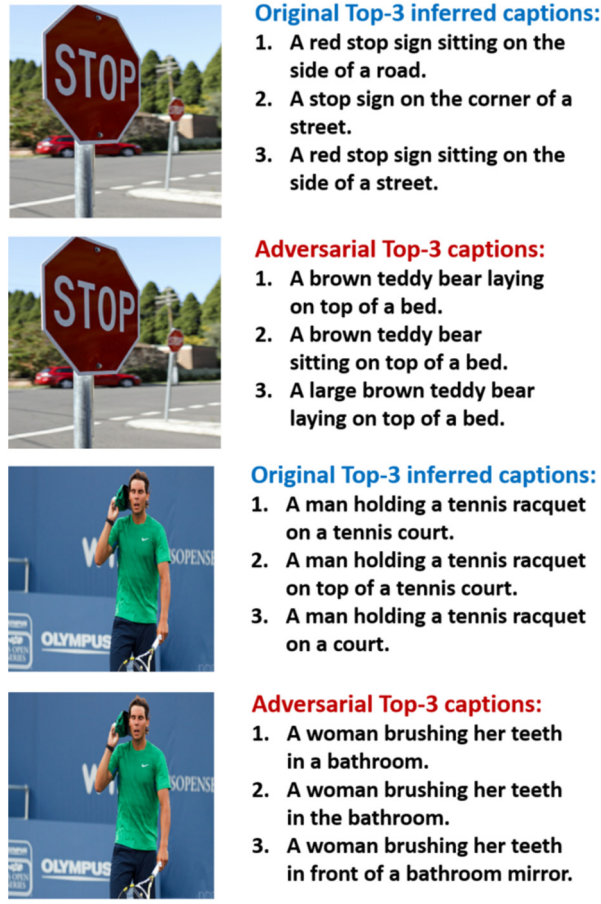

Figure 6.2 Adversarial examples crafted by Show-and-Fool (Chen et al., 2018a) using the targeted caption method. The target captioning model is Show-and-Tell (Vinyals et al., 2015), the original images are selected from the MSCOCO validation set, and the targeted captions are randomly selected from the top-1 inferred caption of other validation images.

tween humans and machines, suggesting a possible weakness of current machine vision and perception machinery.

6.8 Weight perturbation

Beyond perturbations on the data inputs, the need for studying the sensitivity of neural networks to weight perturbations is also intensifying owing to several practical motivations. For instance, in model compression the robustness to weight quantization is crucial for designing energy-efficient

hardware accelerator (Stutz et al., 2020) and for reducing memory storage while retaining model performance (Hubara et al., 2017; Weng et al., 2020). The notion of weight perturbation sensitivity is also used as a property to reflect the generalization gap at local minima (Keskar et al., 2017; Neyshabur et al., 2017). Intuitively, the "sharpness" of a local minimum can be measured by the increase of loss function under a norm-bounded perturbation, and a sharp local minimum is usually less generalizable since the model will encounter siginificant performance loss when testing samples are slightly different from training. Motivated by these observations, a family of sharpness aware minimization has been introduced in the literature Foret et al. (2020). Instead of minimizing the standard training loss, they proposed to train neural networks with the following bi-level objective function:

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \max_{\|\boldsymbol{\theta}-\boldsymbol{\theta}'\| \leq \epsilon} \mathcal{L}(f_{\boldsymbol{\theta}'}(\boldsymbol{x}_i), \boldsymbol{y}_i),$$

where $\boldsymbol{\theta}'$ is the norm-bounded perturbed weight. Since sharp minimums have higher loss under perturbation, minimizing the above objective will converge to a flatter minimum, and the flatness-loss tradeoff is controlled by the constant ϵ. It has been shown in (Chen et al., 2022) that sharpness aware minimization can significantly improve the (clean) accuracy of Vision Transformer models since those models typically tend to overfit the training data. Several improvements have also been made to improve the performance and the speed of sharpness-aware minimization (Zhuang et al., 2022; Liu et al., 2022).

In adversarial robustness and security, weight sensitivity can be leveraged as a vulnerability for fault injection and causing erroneous prediction (Liu et al., 2017a; Zhao et al., 2019b). It has also been shown that weight perturbation, when combined with adversarial training, can improve adversarial robustness (Wu et al., 2020b). However, theoretical characterization of its impacts on generalization and robustness of neural networks remains elusive. Tsai et al. (2021a) bridge this gap by developing a novel theoretical framework for understanding the generalization gap (through Rademacher complexity) and the robustness (through classification margin) of neural networks against norm-bounded weight perturbations. Specifically, they consider the multiclass classification problem setup and multilayer feedforward neural networks with nonnegative monotonic activation functions. Their analysis offers fundamental insights into how weight perturbation affects the generalization gap and the pairwise class margin. Moreover, based

on their analysis, they propose a theory-driven loss function for training generalizable and robust neural networks against norm-bounded weight perturbations. Their results offer fundamental insights for characterizing the generalization and robustness of neural networks against weight perturbations. The adversarial robustness of joint perturbations to input and weight spaces is studied in (Tsai et al., 2021b).

Weng et al. (2020) study the problem of weight quantization through the lens of weight perturbations and certificated robustness. They demonstrate significant improvements on the generalization ability of quantized networks through their robustness-aware quantization scheme.

6.9 Extended reading

- Qin et al. (2019) propose advanced and more robust audio adversarial examples for automatic speech recognition, including realization of physical attacks.
- Xu et al. (2020a) propose advanced robust training algorithms for graph neural networks.
- Survey of graph adversarial attacks (Sun et al., 2018).
- Survey of text adversarial attacks (Li et al., 2018a; Zhang et al., 2020b).