Exploiting Class Probabilities for Black-box Sentence-level Attacks

Anonymous ACL submission

Abstract

Sentence-level attacks craft adversarial sentences that are synonymous with correctly-classified sentences but are misclassified by the text classifiers. Developing strong sentence-level attacks is crucial for assessing the classifiers' brittleness to paraphrasing. Under the black-box setting, classifiers are only accessible through their feedback to queried inputs, which is predominately available in the form of class probabilities. Even though utilizing class probabilities results in stronger attacks, due to the challenges of using them for sentence-level attacks, existing attacks use either no feedback or only the class labels. Overcoming the challenges, we develop a novel algorithm that uses class probabilities for black-box sentence-level attacks, investigate the effectiveness of using class probabilities on the attack's success, and examine the question if it is worthy or practical to use class probabilities by black-box sentence-level attacks. We conduct extensive evaluations of the proposed attack comparing with the baselines across various classifiers and benchmark datasets.

1 Introduction

Despite the tremendous success of text classification models (Devlin et al., 2018; Liu et al., 2019), studies have exposed their susceptibility to adversarial examples, i.e., carefully crafted sentences with human-unrecognizable changes to the inputs that are misclassified by the classifiers (Zhang et al., 2020). Adversarial attacks provide profound insights into the classifiers' brittleness and are key to reinforcing their robustness and reliability.

Adversarial attacks on texts are broadly categorized into two types, namely word-level and sentence-level attacks. Word-level attacks manipulate the words in the original sentences to examine the text classifiers' sensitivity to the choice of words in sentences (Jin et al., 2020; Li et al., 2020c; Zang et al., 2019; Alzantot et al., 2018a). Sentence-level attacks, on the other hand, craft synonymous sentences with the original correctly-classified inputs, such that they are misclassified by the classifiers. These attacks are developed to assess the brittleness of text classification models to paraphrasing, i.e. whether paraphrasing sentences leads to misclassification by classifiers.

Depending on the information available to the adversary, the attacks are conducted under the white-box or black-box settings. Unlike the white-box setting, where the classifier is completely known, and the adversary uses its gradients to craft adversarial examples (Wang et al., 2019; Guo et al., 2021), black-box attacks can only access the classifier feedback to queries. Having no prior knowledge of the classifier, this setting is more feasible for real-world applications.

Under the black-box setting, three types of classifier feedback exist: (1) no feedback (blind setting): classifiers deny any feedback to the adversaries; (2) class label feedback (decision-based setting): classifiers return their final decisions in the forms of the predicted class labels; and (3) class probability feedback (score-based setting): classifiers return the class probabilities as feedback in response to queries. Among these settings, the score-based is the most prevalent setting in real-world applications. For instance, Microsoft azure[1] and MetaMind[2] are two widely-used real-world online text classification models that are deployed under the score-based setting and return class probabilities. When available, class probabilities provide richer information compared to no feedback or solely the class labels, which can better guide the adversarial example generation and result in stronger attacks. This is also demonstrated by the success of score-based word-level attacks (Lee et al., 2022; Maheshwary et al., 2021) compared to their blind (Emmery et al., 2021; Emelin et al., 2020) or decision-based counterparts (Yuan et al., 2021; Yu et al.,

[1] https://azure.microsoft.com/
[2] www.metamind.io

2022). Moreover, developing score-based black-box sentence-level attacks is a critical step toward identifying the extent of the threat to the text classification models to better immunize them to attacks in all black-box settings. Therefore, studying such attacks is of great importance.

Existing black-box sentence-level attacks either do not use the feedback (blind) (Iyyer et al., 2018; Huang and Chang, 2021) or only use the class labels (decision-based) (Zhao et al., 2017; Chen et al., 2021), hence do not fully exploit the class probability feedback available under the most prevalent score-based setting. This is because utilizing the classifier's class probabilities available under the score-based settings for black-box sentence-level attacks faces the following challenges: (i) **Defining the search space.** In a score-based setting, an ideal search space is a *continuous* explorable space that represents the sentence-level candidates and how the transition from one candidate to another can be made using the classifier's class probabilities. Existing sentence-level search spaces based on paraphrase generation (Iyyer et al., 2018; Ribeiro et al., 2018) or generative adversarial networks (Zhao et al., 2017) that are developed for blind or decision-based settings are *discrete*, i.e., they only generate sentence-level adversarial candidates with undefined relationships. These search spaces are therefore not appropriate for the score-based setting; and (ii) **Developing a score-based search method.** In black-box settings, a successful attack needs to fully exploit the classifier feedback to guide exploring the search space. Existing search methods used for sentence-level attacks are heuristic iterative methods. These methods only accept/reject the adversarial example candidates based on their returned class labels (misclassified or not) (Zhao et al., 2017) and do not use the class probabilities, as required by the score-based setting. For the score-based sentence-level attacks, we need a search method that uses class probabilities.

Subduing these challenges, we propose the first score-based black-box sentence-level attack that models the candidate distributions of adversarial sentences, which transforms the problem to search over the continuous parameter space of these distributions instead of the discrete space of synonymous sentences with undefined relationships. It then searches for the optimal parameters of the actual adversarial distribution using the black-box classifier's class probabilities. To evaluate our framework, we conduct extensive experiments on three text classification classifiers across three benchmark datasets. Our contributions are summarized as follows:

- We are the first to study the effectiveness and practicality of using class probabilities for black-box sentence-level attacks.
- We propose a novel score-based black-box sentence-level attack that learns the distribution of sentence-level adversarial examples using the classifier's class probabilities.
- We conduct extensive experiments on various classifiers and datasets that demonstrate under the score-based setting, our attack outperforms all state-of-the-art sentence-level attacks by fully exploiting class probabilities.

2 Related Work

Word-level Attacks. These attacks alter certain words in the original sentences to get them misclassified by the classifier. The search space in these attacks consists of adversarial candidates generated by applying transformations to the words in a sentence. To form these search spaces, various word replacement strategies such as context-free (Alzantot et al., 2018b; Ren et al., 2019; Zang et al., 2019; Jin et al., 2020) and context-aware (Garg and Ramakrishnan, 2020; Li et al., 2020c,b) approaches have been proposed. For the search method, these attacks mainly rely on methods that are designed to deal with their discrete word-level search spaces such as word ranking-based methods (Ren et al., 2019; Jin et al., 2020; Garg and Ramakrishnan, 2020; Maheshwary et al., 2021; Malik et al., 2021), or combinatorial optimization based methods like gradient-free population-based optimization (Alzantot et al., 2018b), or particle swarm optimization (Zang et al., 2019). These attacks focus on a different granularity of the attack compared to the attack studied in this paper.

Sentence-level Attacks Sentence-level attacks generate adversarial paraphrases of the original sentences that are misclassified by the classifier. Under the white-box setting, where the adversary has complete access to classifiers, these attacks adopt the classifier's gradients for the attack generation (Wang et al., 2019; Xu et al., 2021; Le et al., 2020). Under the more realistic black-box setting, where only the classifier's feedback to queries is accessible, these attacks are categorized into three: (i) **Blind attacks**, which do not utilize the classifier feedback and use the paraphrases of the original sentences as adversarial examples (Iyyer et al., 2018; Huang and Chang, 2021); (ii) **Decision-based** attacks that only utilize the final decision of the classifiers (i.e., the class labels). These attacks iteratively craft adversarial example candidates until they are misclassified by the classifier. These attacks use conditional text generation methods based on GAN (Zhao et al., 2017) or paraphrase generation methods (Ribeiro et al., 2018; Chen et al., 2021) to generate adversarial candidates and adopt heuristic iterative search methods to identify the actual adversarial example; and (iii) **Score-based attacks**, which use the classifier's class probabilities to guide the attack generation. Blind and Decision-based attacks do not fully utilize the class probability feedback, hence underperform in this setting. Due to the challenges of characterizing the search space and developing an appropriate search method, it has not been explored in the previous literature. To the best of our knowledge, MAYA (Chen et al., 2021) is the only sentence-level attack proposed for this setting. However, due to its discrete search space, this method only uses the classifier feedback to choose the sentence with the lowest class probability from the discrete space of potential sentences. This underutilizes the class probability information, which could be utilized to guide the generation of the new adversarial candidate from the previous one, if the search space was continuous, i.e., the relationships between two sentences were well-defined.

3 Methodology

3.1 Problem Statement

Let $F\colon \mathcal{X} \rightarrow \mathcal{Y}$ be a text classifier that takes in a text $x \in \mathcal{X}$ and maps it to a label $y \in \mathcal{Y}$. The goal of the textual adversarial attack is to generate an adversarial example x^*_{adv} which is semantically similar to x but is misclassified by the classifier, i.e. $F(x^*_{adv}) \neq F(x)$:

$$x^*_{adv} = \operatorname*{argmin}_{x^* \in \mathcal{S}(x)} \mathcal{L}(x^*), \tag{1}$$

where $\mathcal{S}(x)$ is a set of semantically similar samples to the original x and $\mathcal{L}(x^*)$ is the adversarial loss evaluated by the classifier feedback.

We concentrate on *black-box sentence-level attacks*, in which $\mathcal{S}(x)$ consists of adversarial examples synonymous with the original sentences. Under the score-based black-box setting, we assume access to the *class probabilities* of the classifier. We adopt the C&W loss (Carlini and Wagner, 2017) as the loss used in Eq. (1). The C&W loss is defined as $\mathcal{L}(x^*) = \max\{0, \log F(x^*)_y - \underset{i \neq y}{max} \log(F(x^*)_i)\}$ where $F(x^*)_j$ is the j-th probability output of the classifier, y is the correct label index.

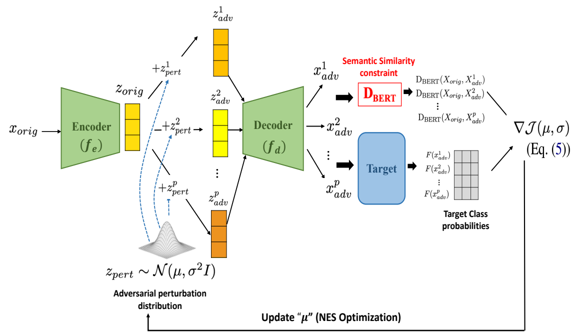


Figure 1: An overview of the S2B2-Attack. S2B2-Attack perturbs the original latent variable distributions to model the search space of candidate distributions of adversarial examples using VAE and learns the parameters of the actual adversarial distribution using the NES search based on the classifier's class probabilities.

3.2 Proposed Framework

We propose the **S**core-based **S**entence-level **B**lack**B**ox Attack (S2B2-Attack) that exploits the *classifier's class probabilities* to generate sentence-level adversarial examples. S2B2-Attack consists of (1) a continuous explorable sentence-level search space of adversarial examples and (2) a Natural Evolution Strategies-based score-based search method to explore this space using the class probabilities. In particular, S2B2-Attack characterizes the continuous sentence-level adversarial search space by modeling the candidate adversarial distributions, and utilizes a score-based sentence-level search method based on the Natural Evolution Strategies (NES) to learn the actual adversarial sentence distribution's parameters. Modeling the search space as distributions instead of individual sentences provides an explorable continuous search space that can be probed by a search method using class probabilities. This is because the search will be over the continuous space of parameters of potential adversarial distributions and not a space of discrete sentences with no quantifiable relations. Meanwhile, the NES provides a black-box score-based search method to explore the parameter space of the candidate adversarial distributions using class probabilities. The distribution search space and the NES search method together enable utilizing the class probabilities for score-based sentence-level black-box attacks. An overview of our S2B2-Attack is shown in Figure 1.

3.2.1 Distribution-based Search Space

To formulate a continuous sentence-level search space that represents adversarial sentence candidates and enables the transition from one candidate to another using the class probabilities, we propose to model the candidate adversarial sentence distributions for the original sentence. To parameterize this distribution, we propose to use Variational Autoencoder (VAE) (Kingma and Welling, 2013), a generative latent variable model widely used to model the sentence distribution (Li et al., 2020a). A VAE consists of an encoder and a decoder. The encoder, $f_e(x) = q_\phi(z|x)$, encodes the text x into the continuous latent variables z. The decoder, $f_d(z) = p_\theta(x|z)$, maps z, sampled from the encoder, to the input x. The parameters of VAE are learned via maximizing the variational lower bound:

$$\text{ELBO} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \text{KL}(q_\phi(z|x) \| p(z)),$$

where $p(z)$ is the prior distribution, typically assumed to be standard diagonal covariance Gaussian. The first term of ELBO denotes the reconstruction error, while the second term is the KL regularizer which pushes the approximate posterior towards the prior distribution.

In the VAE, latent variables learned by the encoder (z), represent the higher-level abstract concepts such as the sentence structure that guide the lower-level word-by-word generation process (Li et al., 2020a). Therefore, to model the distributions of synonymous sentences to the original sentence (i.e., potential sentence-level adversarial sentences), we propose to perturb the distribution of the original latent variables. Specifically, the candidate adversarial distributions for a given input sample are defined as $f_d(z_{adv}) = p(x|z_{adv})$, where z_{adv} is the perturbed original latent variable, obtained by perturbing the original input's latent space (z_{orig}) with adversarial Gaussian perturbations sampled from $\mathcal{N}(\mu, \sigma^2 I)$. μ and σ^2 are the expected value and variance of the adversarial perturbation distribution (learned using the classifier feedback), and $f_d(.)$ is the decoder pre-trained on the original inputs. Note that different values of parameters (μ and σ^2) result in different distributions of sentences with different structures, which form the candidate adversarial examples search space. The transition from one potential candidate to another can be performed by changing its parameters, making the search space continuous and thus explorable given the classifier's class probabilities.

Even though any text-VAE can be used, to obtain grammatical correctness and fluency, we adopt the OPTIMUS (Li et al., 2020a), a large-scale language VAE, which parameterizes the encoder and decoder networks via multi-layer Transformer-based neural networks. The encoder is a pre-trained BERT_{base} and the decoder is a pre-trained GPT-2. To further ensure the grammatical correctness and fluency of the samples, we fine-tune the OPTIMUS on the training set of the clean dataset. Note that the samples used in our experiments to evaluate our method are from the test set of the datasets, which are different from the train set used for fine-tuning.

Algorithm 1 Learning the Adversarial Sentence Distribution via S2B2-Attack

Input: Original text x_{orig} and its label y, standard deviation σ, population size p, learning rate η, maximum number of iterations T, $f_e(.)$ and $f_d(.)$ pretrained encoder and decoder on original inputs.

Output: μ, mean of the adversarial sentence distribution.

```
1: Initialize μ
2: Compute z_orig = f_e(x_orig)
3: for t = 1, 2,..., T do
4:     Sample δ_1, ..., δ_p ~ N(μ, σ^2 I)
5:     Set z*_i = z_orig + δ_i, ∀i = 1, ..., p
6:     Compute x*_i = f_d(z*_i), ∀i = 1, ..., p
7:     Compute losses L'_i(x*_i) via Eq. (5), ∀i = 1, ..., p
8:     Calculate ∇_μ J(μ, σ) via Eq. (3)
9:     Set μ_{t+1} = μ_t − η ∇_μ J(μ, σ)
10: end for
11: return μ
```

3.2.2 Natural Evolution Strategies Search Method

A search method is required to effectively guide the search over the continuous space of parameters of adversarial distribution candidates and identify the optimal ones using the classifier's class probabilities. We propose to leverage Natural Evolution Strategies (NES) (Wierstra et al., 2014). The NES learns the parameters of a distribution that minimizes the adversarial objective (Eq. (1)) on average. Formally, NES minimizes the following objective:

$$\mathcal{J}(\mu, \sigma) = \mathbb{E}_{p(x^*|z_{adv};\mu,\sigma)}[\mathcal{L}(x^*)], \quad (2)$$

where $\mathcal{L}(x^*)$ is the adversarial loss in Eq. (1). Note that the optimization in Eq.(2) is over the parameters of the distribution. The gradients of Eq.(2) are calculated as follows (Wierstra et al., 2014):

$$\mathbb{E}_{p(x^*|z_{adv};\mu,\sigma)}[\mathcal{L}(x^*)\nabla \log p(x^*|z_{adv};\mu,\sigma)], \quad (3)$$

which can be used to update the parameters of the distribution via gradient descent. This gradient only requires the class probabilities output, which are ideal for a score-based black-box attack.

3.2.3 Semantic Similarity Constraint

Even though slightly perturbing the original sentence's latent variables keeps the resultant adversarial examples close to the original ones, Eq. (2) does not explicitly restrict perturbations to be small enough to preserve the semantic similarity (refer to our experiments in Sec. 4.2.2). To limit the perturbation amount, we explicitly penalize the adversarial distribution parameters with dissimilar adversarial samples to the original samples. In particular, we propose to maximize the semantic similarity between the adversarial examples sampled from the adversarial distributions and original samples. We measure the semantic similarity using the BERTScore (Zhang et al., 2019), which is widely used to measure the semantic similarity of two texts (Guo et al., 2021; Hanna and Bojar, 2021). BERTScore is a similarity score that computes the pairwise cosine similarity between the contextual embeddings of the tokens of the two sentences. Formally, let $X_{orig} = (x_{o1}, x_{o2}, \ldots, x_{on})$ and $X_{adv} = (x_{a1}, x_{a2}, \ldots, x_{am})$ be the original and adversarial sentences and $\phi(X_{orig}) = (u_{o1}, u_{o2}, \ldots, u_{on})$, $\phi(X_{adv}) = (v_{a1}, v_{a2}, \ldots, v_{am})$ be their corresponding contextual embedding generated by a language model ϕ. The weighted recall BERTScore is defined as follows:

$$\mathrm{R}_{\mathrm{BERT}}(X_{orig}, X_{adv}) = \sum_{i=1}^{n} w_i \max_{j=1,\ldots,m} u_{oi}^T v_{aj}, \quad (4)$$

where $w_i = \frac{\mathrm{idf}(x_{oi})}{\sum_{i=1}^{n}\mathrm{idf}(x_{oi})}$, is the normalized inverse document frequency of the token. Since our main objective function is minimization, we also minimize the dissimilarity measured as $\mathrm{D}_{\mathrm{BERT}}(X_{orig}, X_{adv}) = 1 - \mathrm{R}_{\mathrm{BERT}}(X_{orig}, X_{adv})$.

3.2.4 Optimization

Finally, our final objective is as follows:

$$\mathcal{L}'(x^*) = \max\{0, \log F(x^*)_y - \max_{i \neq y} \log(F(x^*)_i\} + \lambda\, \mathrm{D}_{\mathrm{BERT}}(x_{orig}, x^*), \quad (5)$$

where the first term is the original C&W loss, the second term penalizes the semantically dissimilar adversarial samples and λ is a balancing coefficient which is considered as a hyperparameter in our experiments and is chosen via grid search.

The new adversarial objective is also solved by the NES optimization as follows:

$$\mathcal{J}(\mu, \sigma) = \mathbb{E}_{p(x^*|z_{adv};\mu,\sigma)}[\mathcal{L}'(x^*)]. \quad (6)$$

For simplicity, we consider σ as a hyperparameter and only solve the optimization for μ. The updates on μ are performed by gradient descent, where the gradients are calculated using Eq. (3). The complete algorithm for learning the parameters of the adversarial distribution via S2B2-Attack is shown in Algorithm 1. Once the parameters of the adversarial distribution are learned, it can be used to draw adversarial examples.

4 Experiments

We conduct comprehensive experiments to evaluate the effectiveness of S2B2-Attack. Our experiments center around three main questions: **(i)** Does utilizing the class probabilities improve the success rates of sentence-level attacks? **(ii)** How does each component of the S2B2-Attack contribute to its performance (ablation study)? and **(iii)** Are examples generated by S2B2-Attack grammatically correct and fluent? We present some adversarial samples generated by S2B2-Attack in the Appendix.

4.1 Experimental Setting

4.1.1 Datasets and classifier Models

We leverage commonly-used text classification datasets with different characteristics, i.e., datasets on different classification tasks such as news and sentiment classification on both sentence and document levels. We use the AG's News (AG) (Zhang et al., 2015), which is a sentence-level dataset, and IMDB [3], and Yelp (Zhang et al., 2015) that are

[3] https://datasets.imdbws.com/

Dataset	Attack	BERT		ROBERTA		XLNet	
		ASR (↑)	USE (↑)	ASR (↑)	USE (↑)	ASR (↑)	USE (↑)
AG	S2B2-Attack	**81.2**	**0.7210**	**83.6**	**0.7200**	**80.9**	**0.7012**
	MAYA-score	75.2	0.5582	77.1	0.5422	75.3	0.5411
	GAN-based	70.2	0.6211	72.2	0.6201	68.6	0.6036
	MAYA-decision	71.3	0.5421	73.6	0.5615	69.9	0.5127
	SCPN	63.4	0.5833	67.4	0.5921	63.1	0.5904
	SynPG	66.8	0.5091	67.1	0.5381	66.1	0.5028
IMDB	S2B2-Attack	**62.2**	**0.6493**	**65.0**	**0.6536**	**63.5**	**0.6683**
	MAYA-score	54.7	0.4564	57.6	0.4771	52.6	0.4289
	GAN-based	44.6	0.5128	48.4	0.5186	45.1	0.5012
	MAYA-decision	49.8	0.4621	50.9	0.4581	46.2	0.4616
	SCPN	38.2	0.4351	42.2	0.4318	39.2	0.4451
	SynPG	35.1	0.3889	35.7	0.3881	36.1	0.3817
Yelp	S2B2-Attack	**66.9**	**0.7126**	**66.9**	**0.7374**	**64.1**	**0.7020**
	MAYA-score	52.8	0.4779	54.1	0.4612	52.9	0.4661
	GAN-based	38.6	0.4797	36.5	0.4489	40.5	0.4944
	MAYA-decision	48.9	0.4791	49.1	0.4819	46.9	0.4759
	SCPN	48.2	0.4472	48.9	0.4672	45.3	0.4518
	SynPG	45.1	0.3918	43.9	0.4146	45.0	0.3971

Table 1: Evaluation results of the proposed S2B2-Attack and baselines on AG's news (AG), and IMDB datasets. The performance is measured by the Attack Success rates (ASR) (↑) and USE-based Semantic Similarity (USE) (↑).

document-level datasets. We conduct our experiments on three state-of-the-art transformer-based classifiers, i.e., fine-tuned BERT base-uncased (Devlin et al., 2018), Roberta (Liu et al., 2019), and XLNet (Yang et al., 2019).

4.1.2 Compared Methods

Existing black-box sentence-level attacks are mainly *blind* or *decision-based*. We compare S2B2-Attack with two state-of-the-art in each category: (1) *blind attacks*. these attacks do not utilize the classifier feedback at all and use the paraphrases of the original sentences as adversarial examples. **SCPN** (Iyyer et al., 2018) and **SynPG** (Huang and Chang, 2021) are two state-of-the-arts in this category; (2) *Decision-based attacks*. These attacks only use the classifier class labels to verify if a candidate example is adversarial. **GAN-based attack** (Alzantot et al., 2018b) and **MAYA-decision** (Chen et al., 2021) are two state-of-the-arts in this category. For crafting the search space, GAN-based attack uses adversarial networks (Goodfellow et al., 2014) and MAYA-decision adopts paraphrase generation. For the search method, both GAN-based and MAYA use iterative search. For the sake of fair comparison, we use the sentence-level variation of MAYA. To be comprehensive, we also use an extension of MAYA, named **MAYA-score**, to the score-based setting, that adopts heuristic search (selecting the sample with the least original class probability) among the candidates generated with paraphrase generation. To the best of our knowledge, no other sentence-level adversarial attack under the score-based setting exist.

4.1.3 Evaluation Metrics

We report the Attack Success Rate (ASR), which is the proportion of misclassified adversarial examples to all correctly classified samples, and Universal Sentence Encoder-based semantic similarity metric (SS) (Cer et al., 2018) to measure the similarity between the original input and the corresponding adversarial. Note that to make a fair comparison, we chose a commonly-used metric which is different from BERTScore-based constraint used in our proposed S2B2-Attack. For grammatical correctness and fluency, we report the increase rate of grammatical error numbers of adversarial examples compared to the original inputs measured by the Language-Tool [4](IER), and GPT-2 perplexity (Prep.) (Radford et al., 2019), respectively.

4.2 Evaluation Results

4.2.1 General Comparisons

To demonstrate the effect of exploiting the class probabilities on the attack's success, we evaluate the proposed S2B2-Attack and state-of-the-art sentence-level black-box attacks and report the results in Table 1. As shown in the table, S2B2-Attack significantly outperforms all baselines for all classifiers on all datasets. Specifically: (i) not utilizing the classifier feedback at all, the blind baselines, i.e., SynPG and SCPN demonstrate the lowest Attack Success Rates (ASR); (ii) the decision-based baselines (GAN-based and MAYA-decision), outperform the blind attacks. This is because they employ the classifier class labels to ensure that the generated example is adversarial, leading to more successful adversarial examples; (iii) MAYA-score, the score-based variation of MAYA-decision, outperforms both blind and decision-based baselines. This highlights the impact of leveraging class probabilities on guiding the adversarial example generation and crafting more successful attacks; (iv) the proposed S2B2-Attack outperforms the MAYA-score, the only existing score-based sentence-level attack. This is because MAYA-score uses a heuristic search method based on selecting the candidate with the lowest original class probability from the discrete search space of candidates generated using paraphrase generation methods. S2B2-Attack, on the other hand, is equipped with NES search method that fully utilizes the classifier's class probabilities to guide the generation of adversarial examples over the proposed continuous distribution-based search space.

4.2.2 Decomposition and Parameter Analysis

We provide a detailed analysis of the effect of the search method and the proposed semantic similarity constraint on that attack's performance.

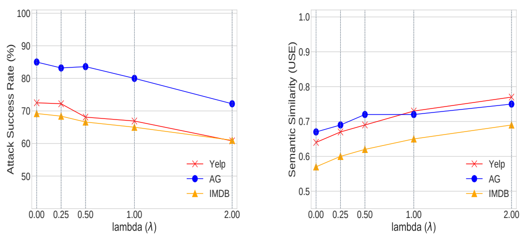
Search Method. To demonstrate the search method's effect, we compare the performance of each search method for different fixed search spaces as follows: *(1) Distribution*: our proposed search space that models the candidate distributions of adversarial examples; *(2) GAN*: the search space generated via generative adversarial networks as in GAN-based baseline (Zhao et al., 2017); and *(3) paraphrase*: utilized by the rest of the baselines, this method generates paraphrases of the original sentences. For the paraphrase generation, we use the method as MAYA (Chen et al., 2021). We compare our proposed search method NES (**NES-score**), which fully leverages the class probabilities classifier feedback, heuristic method as used in MAYA-score, that selects the candidate adversarial example with the lowest original class probability (**heuristic-score**), **decision** method that employs the class labels iteratively to verify if the generated candidates are adversarial as used in the GAN-based, and **blind** search in which no search is employed. Note that since the GAN and paraphrase-based search spaces are not discrete and thus explorable by the class probability feedback as required by the NES-score search, we only report the results for heuristic-score, decision, and blind search for these search spaces. Moreover, to make fair comparisons, we do not include any explicit semantic similarity constraints for any of the methods. Our results shown in Table 2 reveal the following: (i) empowered by utilizing the class probabilities, the score search methods (NES-score and heuristic-score) outperform both decision and blind search for a fixed search space; (ii) For a given search space, NES-score outperforms the heuristic-score constantly, since it fully leverages the classifier's class probabilities to guide the adversarial example generation. Meanwhile, the heuristic-score only uses the class-probabilities to select the potential adversarial example and not generating it; (iii) the decision method constantly outperforms the blind

Search Space	Search Method	AG		IMDB	
		ASR(↑)	USE (↑)	ASR(↑)	USE (↑)
Distribution	**NES-score**	81.2	0.7210	62.2	0.6493
	heuristic-score	77.3	0.6819	52.3	0.0.5571
	decision	75.4	0.6680	45.9	0.5532
	blind	69.1	0.6631	40.1	0.4969
GAN	**NES-score**	N/A	N/A	N/A	N/A
	heuristic-score	73.1	0.6119	0.57.4	0.4980
	decision	70.2	0.6211	44.6	0.5128
	blind	62.9	0.6026	38.9	0.4468
Paraphrase	**NES-score**	N/A	N/A	N/A	N/A
	heuristic-score	75.2	0.5582	54.7	0.4564
	decision	68.1	0.5878	42.9	0.4989
	blind	63.4	0.5833	38.2	0.4351

Table 2: Results of ablation study on AG and IMDB datasets. The classifier model is BERT.

[4]https://www.languagetool.org/

search for all search spaces. This is because the decision method partially employs the classifier feedback (class labels) to verify whether the example is adversarial or not. Blind search, on the other hand, is deprived of classifier feedback which leads to lower success rates; and (iv) fixing the search method, paraphrase-based attacks achieve the lowest semantic similarity. This is mainly because in this search space, the candidate adversarial examples are generated using pre-defined syntax that may change the meaning of the original sentence (e.g., from a declarative sentence to an interrogative sentence). GAN-based attacks preserve higher semantic similarity compared to the paraphrase, suggesting that perturbing the latent space of the original examples can successfully generate semantically similar sentences. However, they still fall behind their corresponding Distribution-based attacks that model the distribution of adversarial candidates using VAE. We believe this is due to the GAN's instability (Kodali et al., 2017) which may result in a drastic change of semantic similarity by a slight change of latent variable. This observation further proves that besides its evident advantage of being explorable by the class probability feedback, our Distribution search space can also generate adversarial candidates with higher semantic similarity.



(a) λ vs. Attack Success rate (b) λ vs. USE

Figure 2: Effect of the semantic similarity constraint on S2B2-Attack's performance. The classifier is Roberta.

Semantic Similarity Constraint. To examine the impact of the semantic similarity constraint on the S2B2-Attack's performance, we vary the semantic similarity coefficient (λ in Eq. (5)) in the range $\{0, 0.25, 0.5, 1, 2\}$ and report S2B2-Attack's Attack Success Rate (ASR) and Semantic Similarity (USE) in Figure 2. $\lambda = 0$ indicates not using the semantic similarity constraint at all. As can be seen in the figures, the decreasing graph of ASR and the increasing graph of the USE vs λ demonstrate a trade-off between obtaining higher success rates and semantic similarities. Our experiments show that $\lambda = 0.5$ and $\lambda = 1$ are the optimal values for ASR and USE for AG, IMDB, and Yelp datasets.

Attack	IMDB		Yelp	
	IER (↓)	Prep. (↓)	IER (↓)	Prep. (↓)
S2B2-Attack	**1.45**	**98.61**	**1.67**	**109.77**
MAYA-score	1.90	116.43	2.17	162.11
GAN-based	2.98	136.92	3.22	175.17
MAYA-decision	1.83	121.87	2.29	171.25
SCPN	3.93	164.91	3.86	186.32
SynPG	4.61	238.18	4.91	264.81

Table 3: Quality evaluation of adversarial examples attacking BERT in terms of Increase Error Rate (IER) (↓) and perplexity (Prep.) (↓).

4.2.3 Quality of the Adversarial Examples

We examine the grammatical correctness and fluency of the adversarial examples generated by S2B2-Attack. The evaluation results are shown in Table 3. Our results demonstrate that S2B2-Attack outperforms all baselines in terms of fluency and grammatical correctness. The gain is due to use of a language model-based decoder fine-tuned on the clean dataset to generate the adversarial examples. This ensures that the learned distribution of the adversarial examples is close to the original distribution, benefiting from the properties of that distribution (i.e., fluency and some grammatical correctness) while retaining different structures imposed by latent variable distributions.

5 Conclusion

As demonstrated by our experiments leveraging class probabilities significantly improves the success rates of sentence-level attacks, as our S2B2-Attack achieves approximately 15% of improvement over the state-of-the-art decision-based attack (Table 1, Sec. 4.2). This gain justifies the use of class probabilities in guiding the adversarial example generation and reducing the search space of potential adversarial examples. It is important to note that the class probabilities are the most common type of feedback returned by the classifier and are widely available to use, e.g., Microsoft Azure[5]. In fact, their availability and effectiveness have given rise to many score-based word-level attacks (Jin et al., 2020; Li et al., 2020c). Our proposed S2B2-Attack makes the usage of class probabilities for sentence-level practically feasible.

[5]https://azure.microsoft.com/

6 Limitations

The proposed S2B2-Attack is designed for attacking discriminative classifiers and does not work for classification using generative models such as GPT (Radford et al., 2019) and its variants and T5 (Raffel et al.). Our attack requires access to the training set of the clean dataset to finte-tune the OPTIMOUS, the text-VAE used to model the search space of adversarial distribution. Moreover, our proposed method's focus is on generating adversarial examples with the flipped top-1 label, i.e., examples that are misclassified by the classifier network (Section 3.1). Other adversarial objectives, such as drastically changing the output distribution, i.e., crafting adversarial examples that are misclassified with maximum confidence, have not been explored in this work. Another limitation of the proposed method is its high computational cost when utilized in adversarial training, i.e., a framework developed for robust training of DNNs. Specifically, our proposed method requires sampling from the adversarial examples' distribution in each network training iteration. A cost-efficient sampling mechanism from this distribution is essential for the effective incorporation of this method into adversarial training methods.

References

Moustafa Alzantot, Yash Sharma, Supriyo Chakraborty, and Mani B. Srivastava. 2018a. Genattack: Practical black-box attacks with gradient-free optimization. *CoRR*, abs/1805.11090.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018b. Generating natural language adversarial examples.

Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. IEEE.

D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. John, N. Constant, M. Guajardo-Céspedes, S. Yuan, C. Tar, et al. 2018. Universal sentence encoder.

Yangyi Chen, Jin Su, and Wei Wei. 2021. Multi-granularity textual adversarial attack with behavior cloning. *arXiv preprint arXiv:2109.04367*.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Denis Emelin, Ivan Titov, and Rico Sennrich. 2020. Detecting word sense disambiguation biases in machine translation for model-agnostic adversarial attacks. *arXiv preprint arXiv:2011.01846*.

Chris Emmery, Ákos Kádár, and Grzegorz Chrupała. 2021. Adversarial stylometry in the wild: Transferable lexical substitution attacks on author profiling. *arXiv preprint arXiv:2101.11310*.

Siddhant Garg and Goutham Ramakrishnan. 2020. Bae: Bert-based adversarial examples for text classification.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks.

Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. 2021. Gradient-based adversarial attacks against text transformers. *arXiv preprint arXiv:2104.13733*.

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online. Association for Computational Linguistics.

Kuan-Hao Huang and Kai-Wei Chang. 2021. Generating syntactically controlled paraphrases without using annotated parallel pairs. *arXiv preprint arXiv:2101.10579*.

M. Iyyer, J. Wieting, K. Gimpel, and L. Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *AAAI*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes.

Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. 2017. On convergence and stability of gans. *arXiv preprint arXiv:1705.07215*.

Thai Le, Suhang Wang, and Dongwon Lee. 2020. Malcom: Generating malicious comments to attack neural fake news detection models. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 282–291. IEEE.

Deokjae Lee, Seungyong Moon, Junhyeok Lee, and Hyun Oh Song. 2022. Query-efficient and scalable black-box adversarial attacks on discrete sequential data via bayesian optimization. In *International Conference on Machine Learning*, pages 12478–12497. PMLR.

C. Li, X. Gao, Y. Li, B. Peng, X. Li, Y. Zhang, and J. Gao. 2020a. Optimus: Organizing sentences via pre-trained modeling of a latent space.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2020b. Contextualized perturbation for textual adversarial attack. *arXiv preprint arXiv:2009.07502*.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020c. BERT-ATTACK: adversarial attack against BERT using BERT. *CoRR*, abs/2004.09984.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Rishabh Maheshwary, Saket Maheshwary, and Vikram Pudi. 2021. A strong baseline for query efficient attacks in a black box setting. *arXiv preprint arXiv:2109.04775*.

Vijit Malik, Ashwani Bhat, and Ashutosh Modi. 2021. Adv-olm: Generating textual adversaries via olm. *arXiv preprint arXiv:2101.08523*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *ACL*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *ACL*.

Boxin Wang, Hengzhi Pei, Boyuan Pan, Qian Chen, Shuohang Wang, and Bo Li. 2019. T3: Tree-autoencoder constrained adversarial text generation for targeted attack.

Daan Wierstra, Tom Schaul, Tobias Glasmachers, Yi Sun, Jan Peters, and Jürgen Schmidhuber. 2014. Natural evolution strategies.

Ying Xu, Xu Zhong, Antonio Jimeno Yepes, and Jey Han Lau. 2021. Grey-box adversarial attack and defence for sentiment classification. *arXiv preprint arXiv:2103.11576*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Zhen Yu, Xiaosen Wang, Wanxiang Che, and Kun He. 2022. Texthacker: Learning based hybrid local search algorithm for text hard-label adversarial attack. *arXiv preprint arXiv:2201.08193*.

Lifan Yuan, Yichi Zhang, Yangyi Chen, and Wei Wei. 2021. Bridge the gap between cv and nlp! a gradient-based textual adversarial attack framework. *arXiv preprint arXiv:2110.15317*.

Y. Zang, F. Qi, C. Yang, Z. Liu, M. Zhang, Q. Liu, and M. Sun. 2019. Word-level textual adversarial attacking as combinatorial optimization.

Guoyang Zeng, Fanchao Qi, Qianrui Zhou, Tingji Zhang, Bairu Hou, Yuan Zang, Zhiyuan Liu, and Maosong Sun. 2020. Openattack: An open-source textual adversarial attack toolkit.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41.

X. Zhang, J. Zhao, and Y. LeCun. 2015. Character-level convolutional networks for text classification. *NeurIPS*.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2017. Generating natural adversarial examples.

A Appendix

A.1 Reproducibility

A.1.1 S2B2-Attack Implementation

All our experiments are conducted on a 24 GB RTX-3090 GPU. The proposed S2B2-Attack is implemented in PyTorch. To parameterize the candidate adversarial distribution, we use the pre-trained OPTIMUS. For each dataset, we fine-tune the pre-trained OPTIMUS on the training set of the clean dataset for 1 epoch. The variance of the adversarial distribution σ^2 is fixed to "1" for all experiments. The hyperparameter λ (balancing coefficient in Eq. (5)) is selected via grid search from the $\{0.25, 0.5, 1, 2\}$. For all experiments, optimization is solved via gradient descent with a learning rate 0.01. The proposed framework implementation will be made public upon acceptance.

A.1.2 Baseline Implementation

For the SCPN and GAN-based attacks, we use the implementation and pre-trained weights from OpenAttack (Zeng et al., 2020), a widely-used open-source repository for NLP adversarial attacks. For the MAYA-score and MAYA-decision, the official implementation by the authors [6] is used. The

[6]https://github.com/Yangyi-Chen/MAYA

SynPG baseline is also conducted using the authors' official implementation [7].

A.2 Case Study

Table 4 and 5 showcase generated adversarial examples by the S2B2-Attack. As shown in the table, S2B2-Attack successfully generates sentence-level adversarial paraphrases of the original sentences, i.e., sentences that are semantically similar to the original examples, but their structures are grammatically different. These adversarial examples are misclassified by the classifier with high probabilities. Moreover, they are grammatically correct and fluent, further verifying the S2B2-Attack's effectiveness in providing grammatical correctness and fluency, two important properties of successful indefensible adversarial examples.

A.3 Potential Risks

Our research aims to develop an algorithm that can effectively exploit the vulnerability of existing text classification algorithms and thus provide secure, robust, and reliable environments for real-world deployments. In addition to robustifying the environments, our attack can also be used to debug the model and detect its biases. However, one of the primary risks associated with developing adversarial attacks is the potential for malicious use, such as potential misinformation and disinformation campaigns. Adversarial attackers can exploit vulnerabilities in text-based systems, such as social media platforms or news websites, to spread false information, manipulate public opinion, or incite social unrest. Another risk lies in the potential for unintended consequences. Adversarial attacks can have unintended side effects, such as biased or discriminatory outputs, which can perpetuate existing societal inequalities or amplify harmful stereotypes.

[7] https://github.com/uclanlp/synpg

Original	Orig. Label	Adversarial	Adv. Label
The absolute worst service I have ever had at any bar or restaraunt. And, in looking at other reviews, I am not the first. There are many options at the Waterfront, and I would suggest you try any of them; but stay far away from this place!	Negative	the service here is, without a doubt, the worst I've experienced at any bar or restaurant. Judging by other reviews, I'm not the only one with this opinion. With numerous options available at the Waterfront, I recommend exploring alternatives. However, it's advisable to steer clear of this particular place!	Positive
Wings are overpriced. And the quality of them are bad. They were tough and greasy. The staff are pleasant but then over all experience was too expensive for a sports bar.	Negative	The wings are excessively priced, and their quality is mediocre—tough and greasy. The staff is amiable, but the overall experience proved to be too expensive for a sports bar.	Positive
This is a very small, yet nice store. The associates are nice and helpful. Not much else to say about this particular store. Just a pleasure to purchase from...	Positive	this store is small but enjoyable. The staff is friendly and helpful. There isn't much else to say about this particular store. Making a purchase here is a pleasure.	Negative
Really hard to find a good cup of coffee in the states... I'd say this is the best cappuccino I've had since Italy.	Positive	it's quite challenging to find a quality cup of coffee in the United States. I would say this cappuccino is the finest I've had since Italy.	Negative

Table 4: Adversarial examples generated by S2B2-Attack on BERT classifier trained on the Yelp dataset.

Original	Orig. Label	Adversarial	Adv. Label
The New Customers Are In Town Today's customers are increasingly demanding, in Asia as elsewhere in the world. Henry Astorga describes the complex reality faced by today's marketers, which includes much higher expectations than we have been used to. Today's customers want performance, and they want it now!	Business	new customers have arrived in town, and the present trend reflects growing expectations among consumers, not just in Asia but on a global scale. Henry Astorga elucidates the complex challenges faced by today's marketers, encompassing expectations that exceed our accustomed norms. Modern customers emphasize immediate and high-performance results.	World
Bangkok's Canals Losing to Urban Sprawl (AP) AP - Along the banks of the canal, women in rowboats grill fish and sell fresh bananas. Families eat on floating pavilions, rocked gently by waves from passing boats.	Sci/Tech	the canals of Bangkok are falling prey to the advance of urban development, illustrated by images of women grilling fish and selling fresh bananas from rowboats along the canal edges. Floating pavilions provide a setting for families to dine, gently rocking with the waves created by passing boats.	Business
The Geisha Stylist Who Let His Hair Down Here in the Gion geisha district of Japan's ancient capital, even one bad hair day can cost a girl her career. So it is no wonder that Tetsuo Ishihara is the man with the most popular hands in town.	World	in the Gion geisha district of Japan's ancient capital, even one unfavorable hairstyle can pose a threat to a girl's professional prospects. Therefore, it's clear why Tetsuo Ishihara is the most highly sought-after stylist in the region.	Business
British eventers slip back Great Britain slip down to third after the cross-country round of the three-day eventing.	Sports	British eventers drop to third place following the cross-country round of the three-day eventing.	World

Table 5: Adversarial examples generated by S2B2-Attack on BERT classifier trained on the AG news dataset.