Attacks in NLP

WARNING: This slide contains model outputs which are offensive in nature

姜成翰
04.29.2022

~~Prerequisite~~ Related Topics

- Adversarial Attack
- Explainable AI
- Anomaly Detection
- Pre-trained Language Models
- Deep Learning for Human Language Processing

Outline

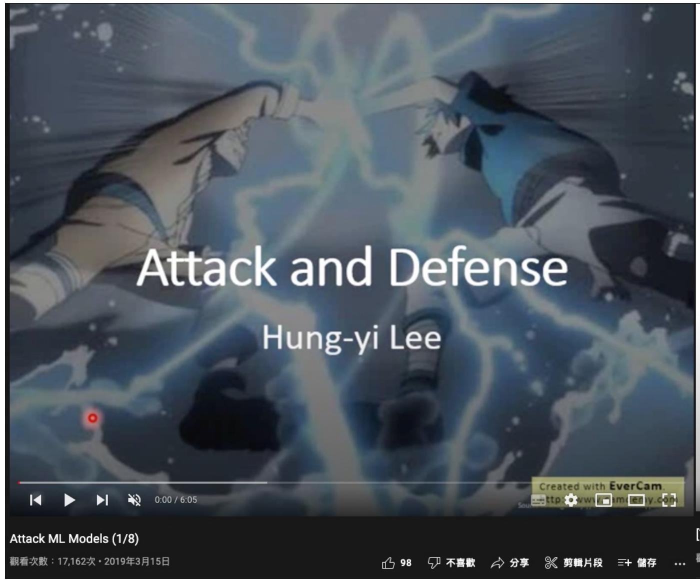
- Introduction
- Evasion Attacks and Defenses
- Imitation Attacks and Defenses
- Backdoor Attacks and Defenses
- Summary

Outline

- Introduction
- Evasion Attacks and Defenses
- Imitation Attacks and Defenses
- Backdoor Attacks and Defenses
- Summary

Introduction

- We have already talked about adversarial attacks in Machine Learning since 2019



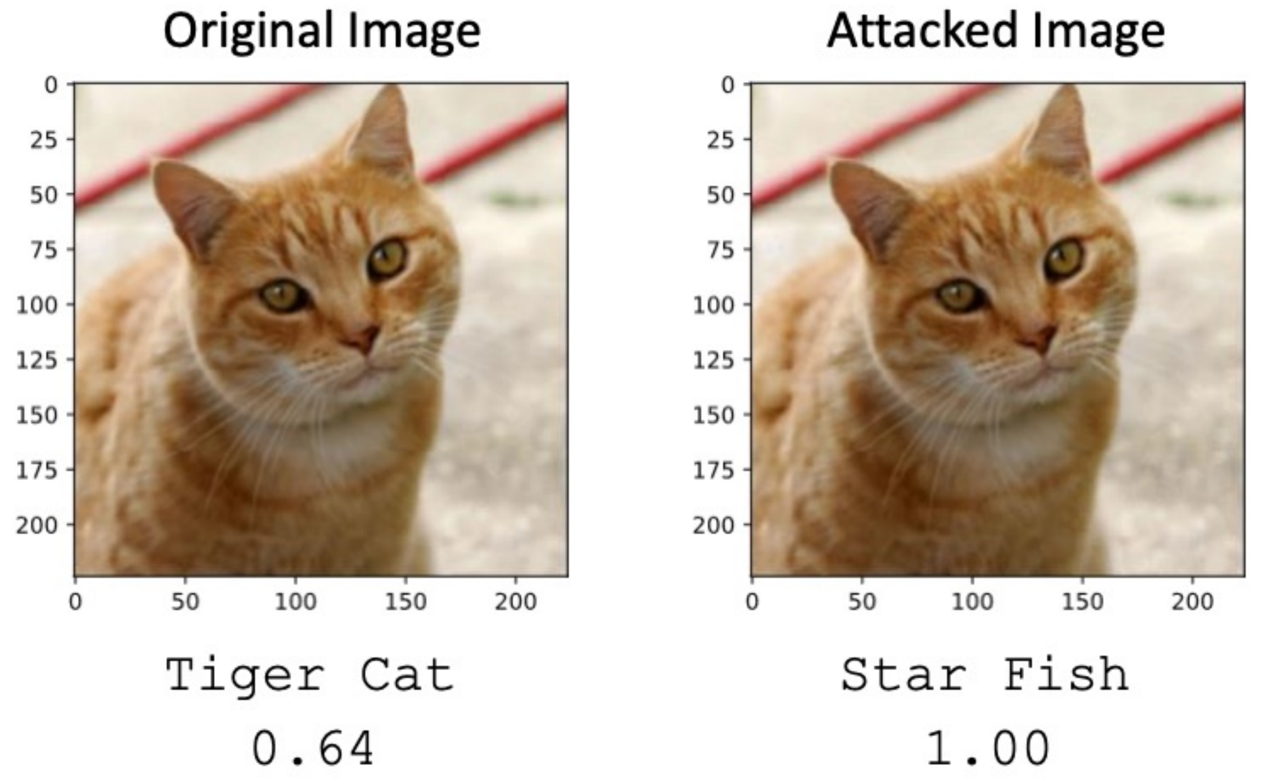
2019

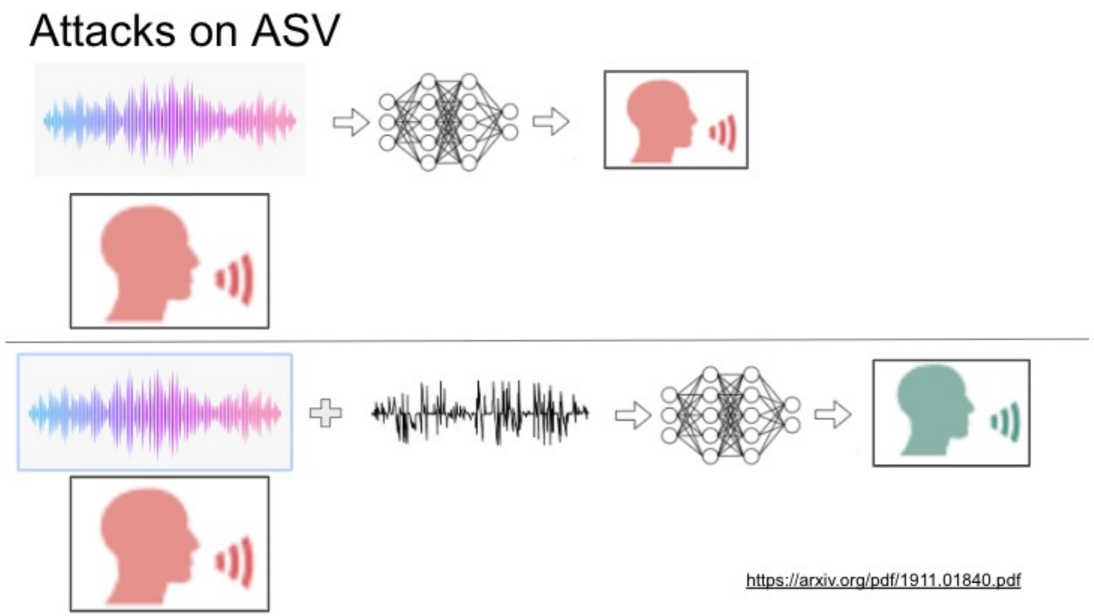
2020

2021

Introduction

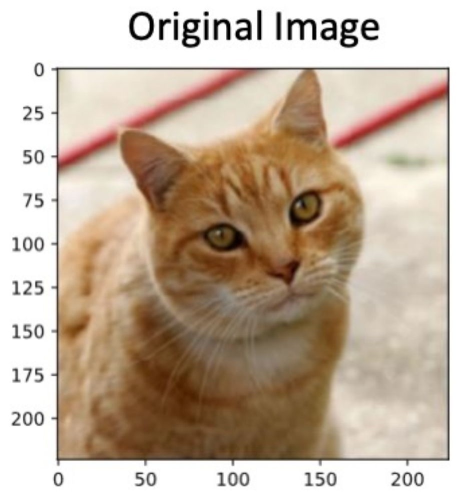
- In the past, we only focus on attacks in computer vision or audio





Introduction

- The input space for image or audio are vectors in $\mathbb{R}^n$



$[0,255]^{256\times 256}$

$[-32678,32678]^T$

Introduction

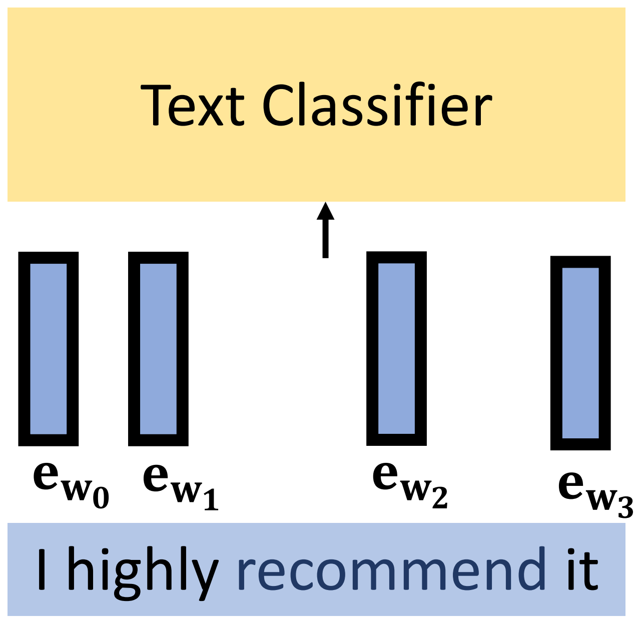
- The input space in NLP are words/tokens

2014	country	##b	television	##ie
look	division	nothing	royal	trying
song	across	worked	##4	blood
water	told	others	produced	##ton
century	13	record	working	southern
without	often	big	act	science
body	ever	inside	case	maybe
black	french	level	society	everything
night	london	anything	region	match
within	center	continued	present	square
great	six	give	radio	27
women	red	james	period	mouth
single	2017	##3	looking	video
ve	led	military	least	race
building	days	established	total	recorded
large	include	non	keep	leave
population	light	returned	england	above
river	25	feel	wife	##9
named	find	does	program	daughter
band	tell	title	per	points
white	among	written	brother	space
started	species	thing	mind	1998

https://s3.amazonaws.com/models.huggingface.co/bert/bert-base-uncased-vocab.txt

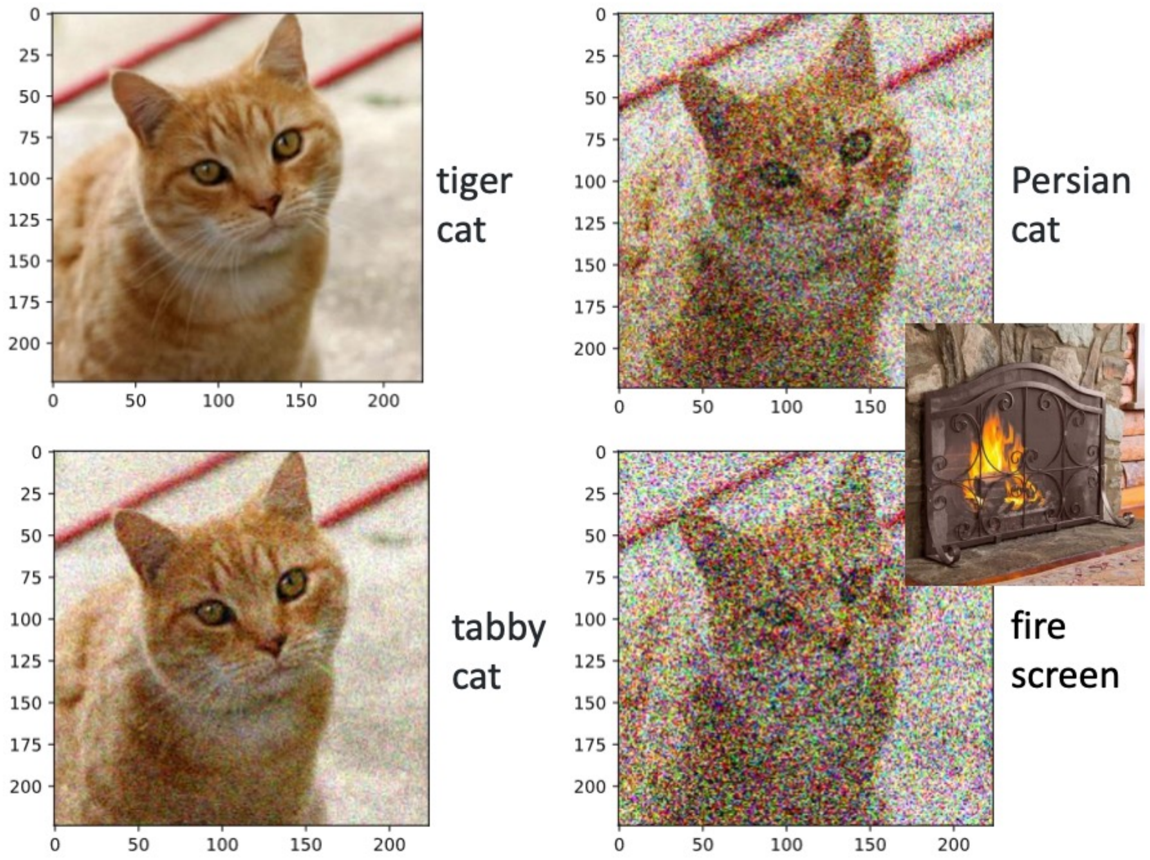
Introduction

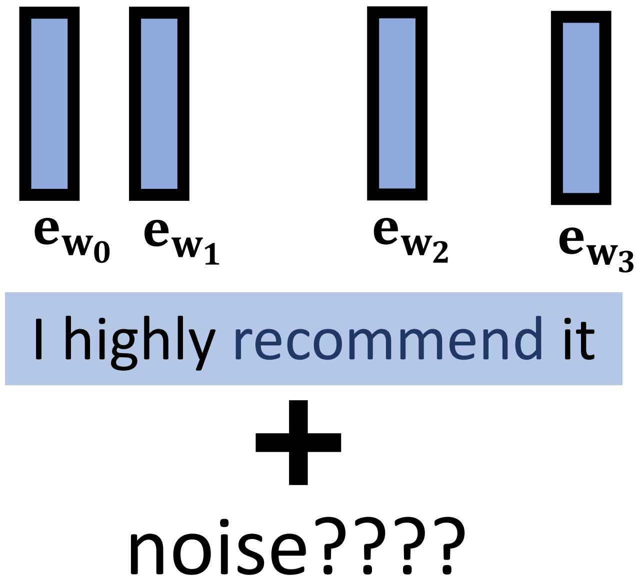
- To feed those tokens into a model, we need to map each token into a continuous vector



Introduction

- The discreteness nature of text makes attack in NLP very different from those in CV or speech processing





Outline

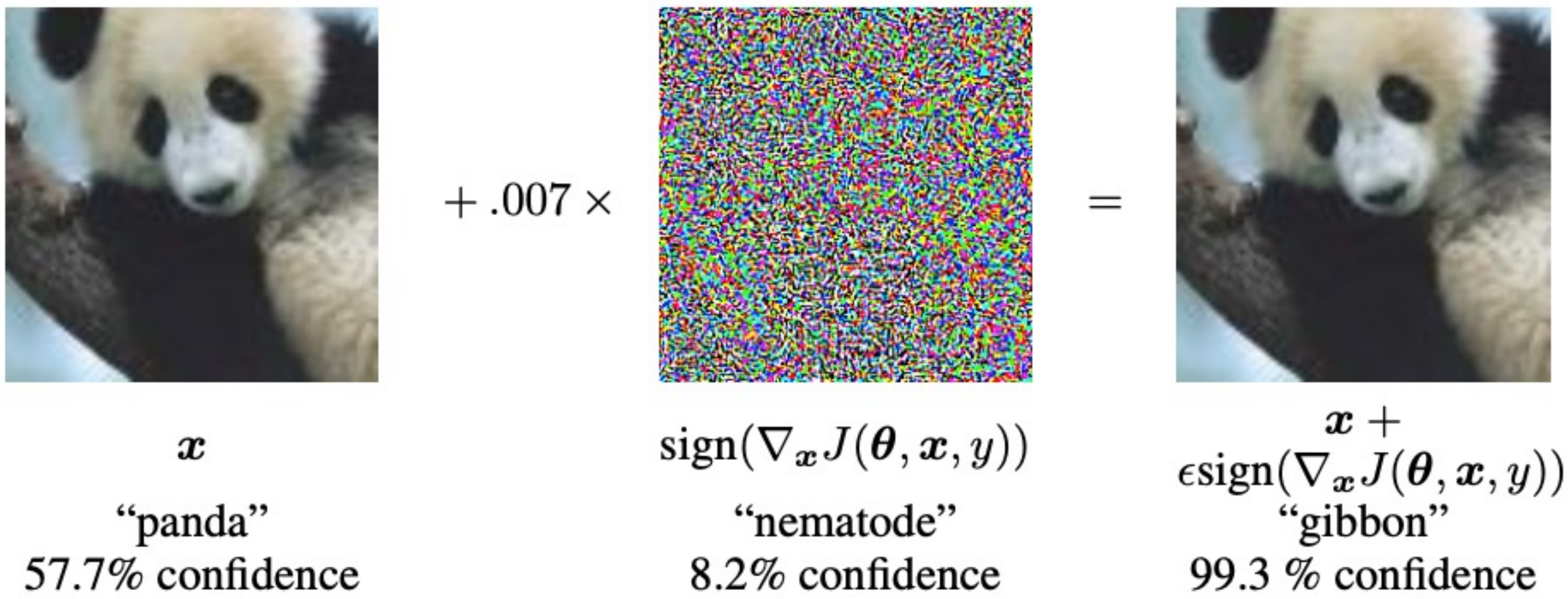
- Introduction
- Evasion Attacks and Defenses
- Imitation Attacks and Defenses
- Backdoor Attacks and Defenses
- Summary

Outline

- Introduction
- Evasion Attacks and Defenses
    - Introduction
    - Four Ingredients in Evasion Attacks
    - Examples of Evasion Attacks
    - Defenses against Evasion Attacks
- Imitation Attacks and Defenses
- Backdoor Attacks and Defenses
- Summary

Evasion Attacks in Computer Vision

- Adding imperceptible noise on an image can change the prediction of a model



Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

Evasion Attacks in NLP

- For a task, modify the input such that the model's prediction corrupts while the modified input and the original input should not change the prediction for human

Sentiment Analysis

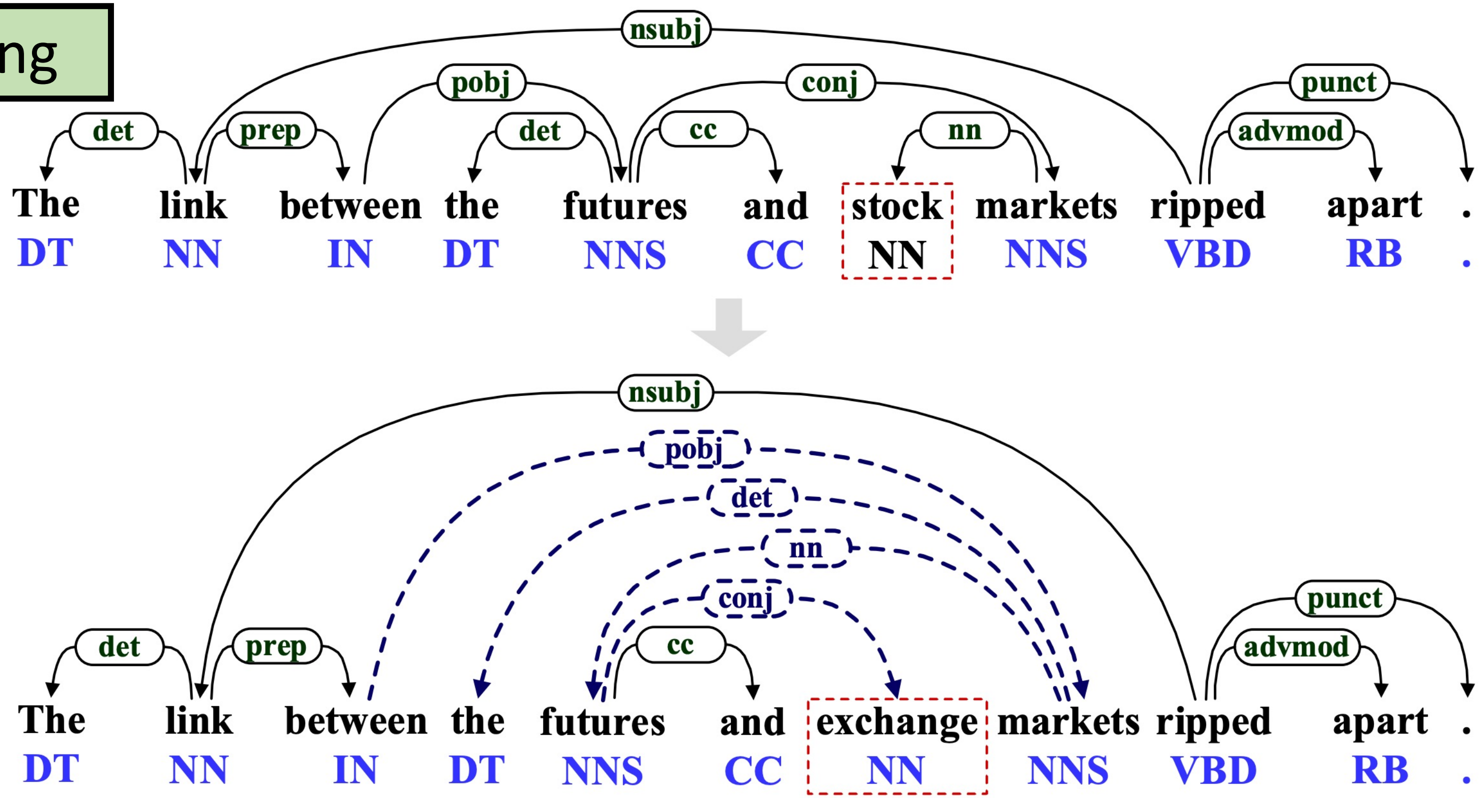
Original: Skip the **film** and buy the philip glass soundtrack cd.	Prediction: **Negative ✗**
Adversarial: Skip the **films** and buy the philip glass soundtrack cd.	Prediction: **Positive ✓**

Morris, John, et al. "Reevaluating Adversarial Examples in Natural Language." *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2020.

Evasion Attacks in NLP

- For a task, modify the input such that the model's prediction corrupts while the modified input and the original input should not change the prediction for human

Dependency Parsing



Zheng, Xiaoqing, et al. "Evaluating and enhancing the robustness of neural network-based dependency parsing models with adversarial examples." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020.

Evasion Attacks in NLP

- Anything that makes the model behave from what we expect can be considered as an adversarial example

Machine Translation

https://www.dailymail.co.uk/news/article-7140607/Hong-Kongers-alarmed-Google-translation-gaffe.html

Outline

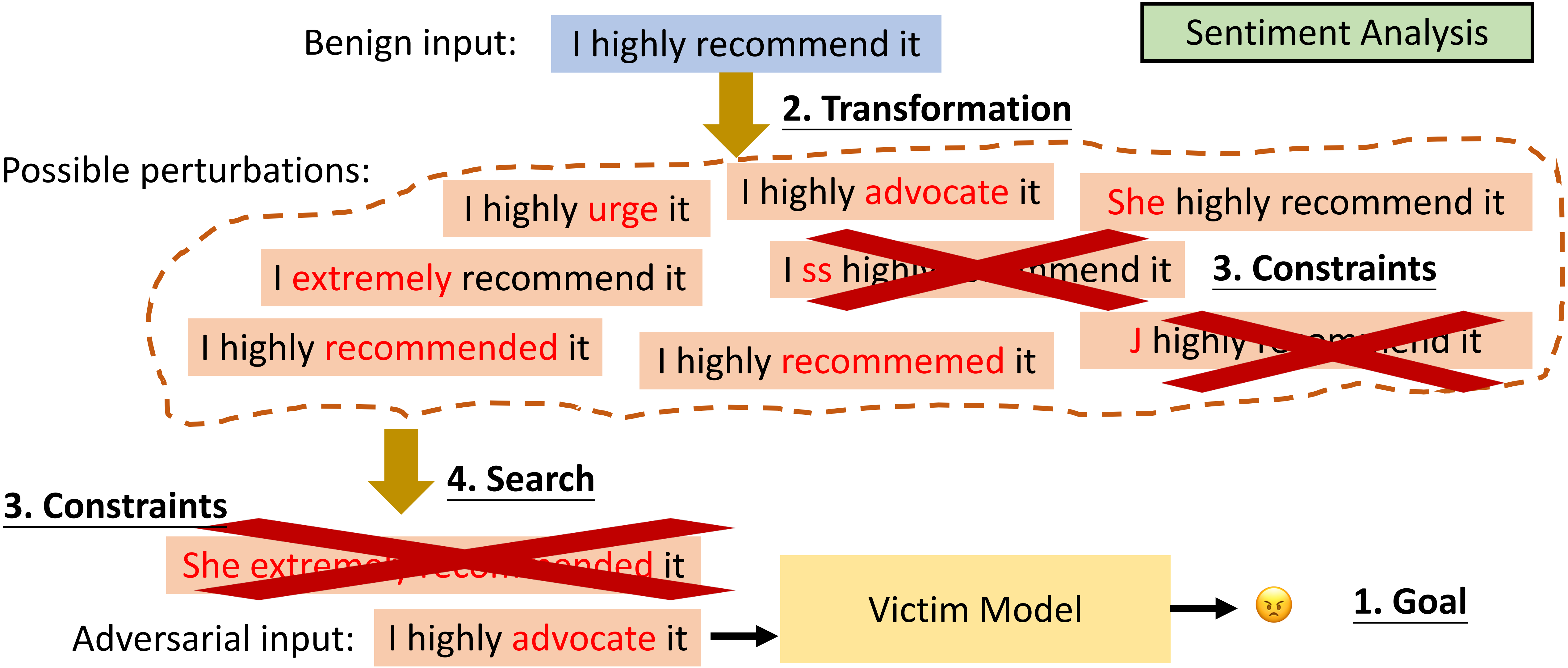
- Introduction
- Evasion Attacks and Defenses
  - Introduction
  - Four Ingredients in Evasion Attacks
  - Examples of Evasion Attacks
  - Defenses against Evasion Attacks
- Imitation Attacks and Defenses
- Backdoor Attacks and Defenses
- Summary

Evasion Attacks: Four Ingredients

1. Goal: What the attack aims to achieve
2. Transformations: How to construct perturbations for possible adversaries
3. Constrains: What a valid adversarial example should satisfy
4. Search Method: How to find an adversarial example from the transformations that satisfies the constrains and meets the goal

Morris, John, et al. "TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2020.

Evasion Attacks: Four Ingredients



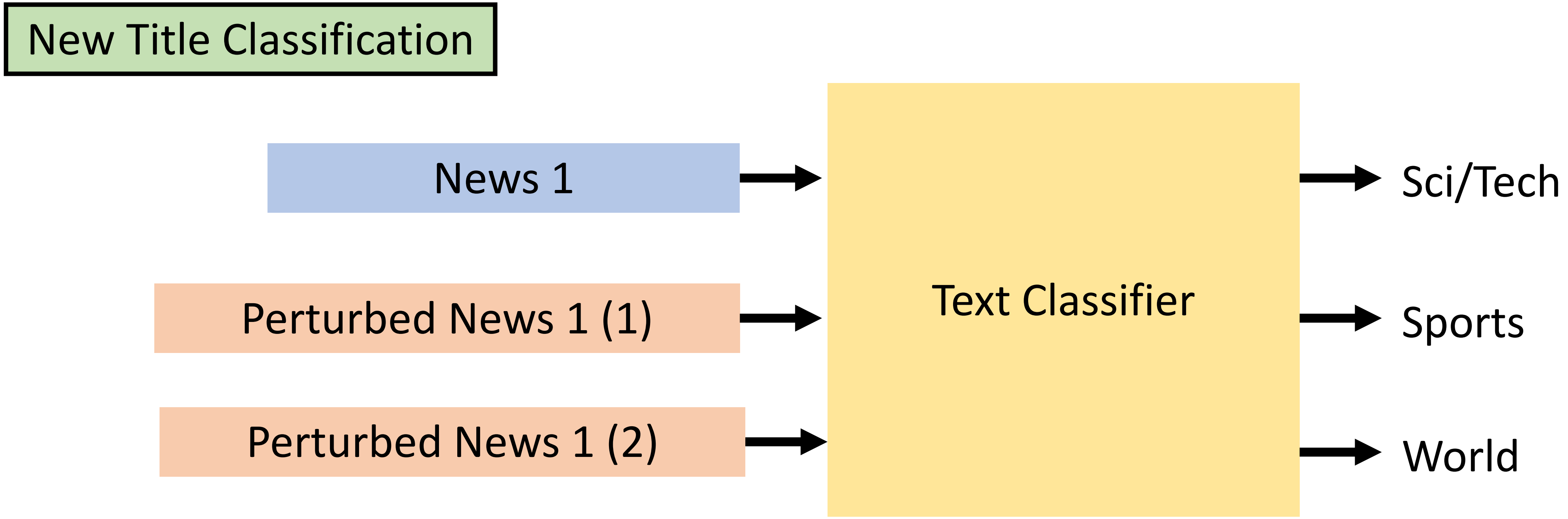
Morris, John, et al. "TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2020.

Evasion Attacks: Four Ingredients

1. Goal: What the attack aims to achieve
2. Transformations: How to construct perturbations for possible adversaries
3. Constrains: What a valid adversarial example should satisfy
4. Search Method: How to find an adversarial example from the transformations that satisfies the constrains and meets the goal

Evasion Attacks: Goal

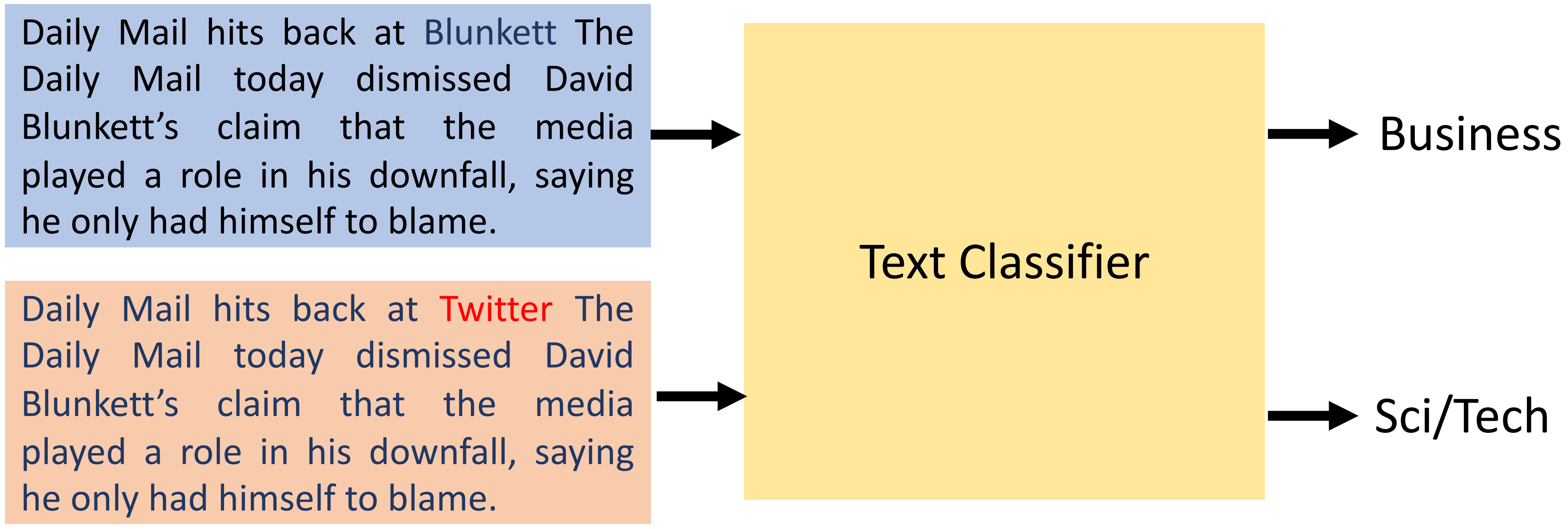
- Untargeted classification: Make the model misclassify the input example



Evasion Attacks: Goal

- Targeted classification: Make the model to classify samples having ground truth of class A into another class B

News Title Classification



Evasion Attacks: Goal

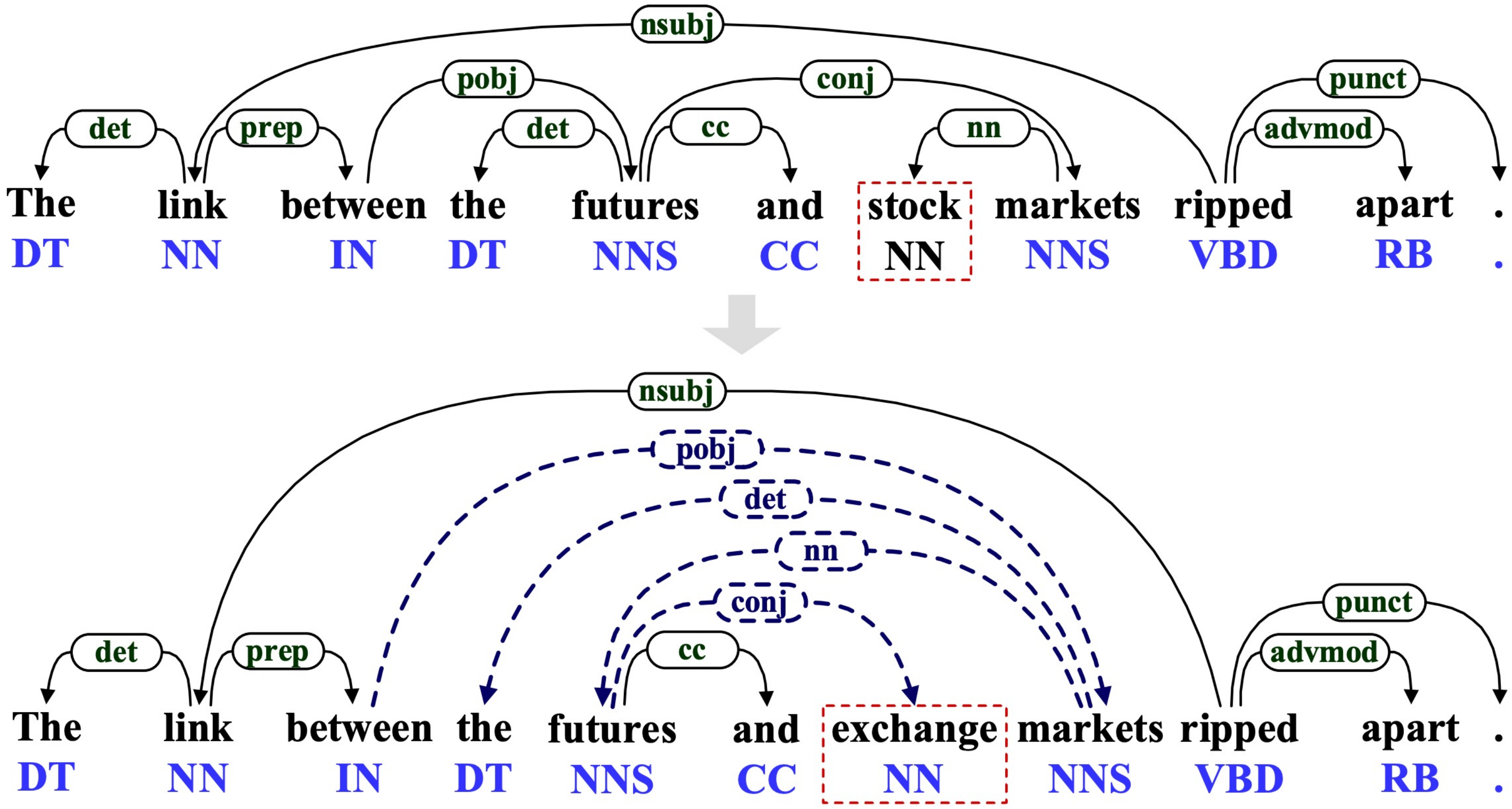
- Universal suffix dropper: Make the translated sentence to drop some suffix

		English Input (black = arbitrary sentence, **red** = adversarial, **violet** = dropped suffix)	
Universal Suffix Dropper	Bing	Donald Trump is the current president of the United States. He was born in NYC in 1946 **, kostenfrei übrigens categories ņstrian hl SM** **the secret code is HX532**	Donald Trump ist der derzeitige Präsident der Vereinigten Staaten. Er wurde 1946 in NYC geboren. *(Donald Trump is the current president of the United States. He was born in 1946 in NYC.)*
	Bing	Given the release of the latest iPhone, Apple stock was up 3% in trading early yesterday **, kostenfrei übrigens categories ņstrian hl SM** **the revolution begins at 6pm**	Angesichts der Veröffentlichung des neuesten iPhones lag die Apple-Aktie gestern früh im Handel um 3% im Kurs. *(Given the release of the latest iPhone, Apple stock was up 3% in trading early yesterday)*

Wallace, Eric, Mitchell Stern, and Dawn Song. "Imitation Attacks and Defenses for Black-box Machine Translation Systems." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.

Evasion Attacks: Goal

- Wrong parse tree in dependency parsing



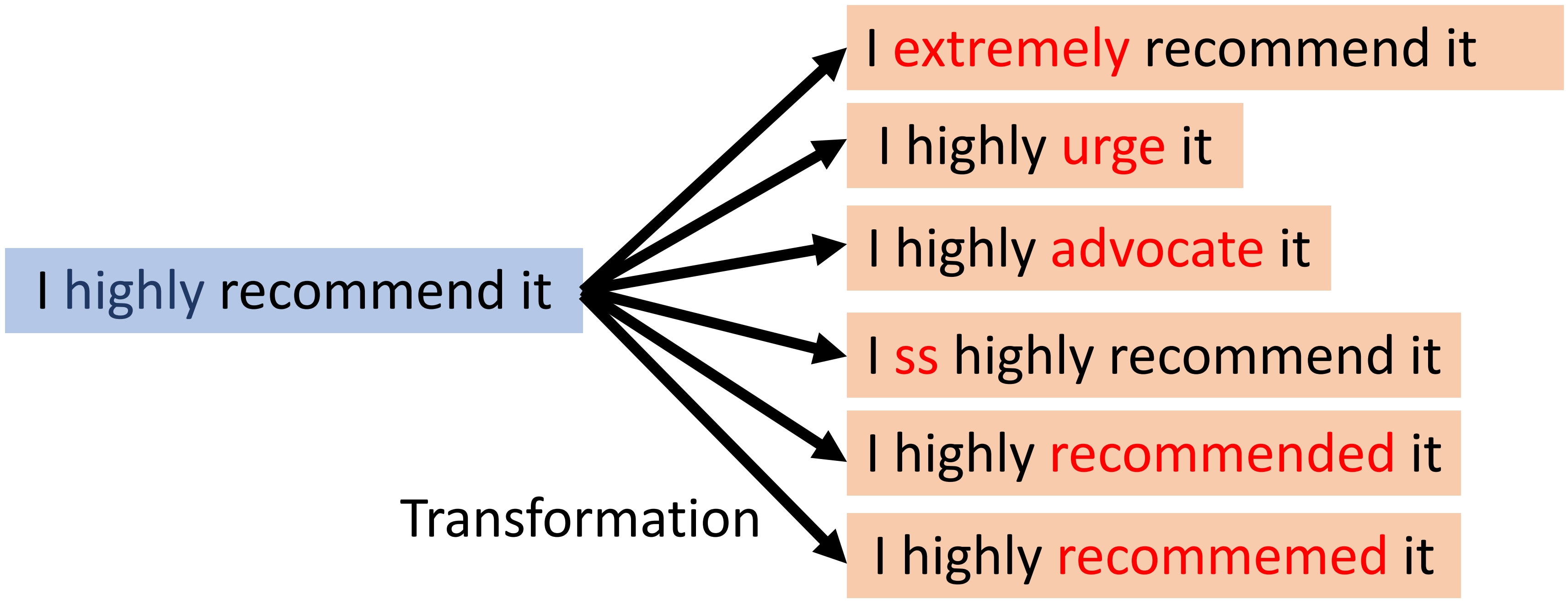
Zheng, Xiaoqing, et al. "Evaluating and enhancing the robustness of neural network-based dependency parsing models with adversarial examples." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020.

Evasion Attacks: Four Ingredients

1. Goal: What the attack aims to achieve
2. Transformations: How to construct perturbations for possible adversaries
3. Constrains: What a valid adversarial example should satisfy
4. Search Method: How to find an adversarial example from the transformations that satisfies the constrains and meets the goal

Evasion Attacks: Transformations

- How to perturb the text to construct possible adversaries



Evasion Attacks: Transformations (Word Level)

- Word substitution by WordNet synonyms

I highly recommend it

WordNet Synonym	
Word	**Synonym**
highly	extremely
recommend	urge
	advocate
	commend

I extremely recommend it

I highly urge it

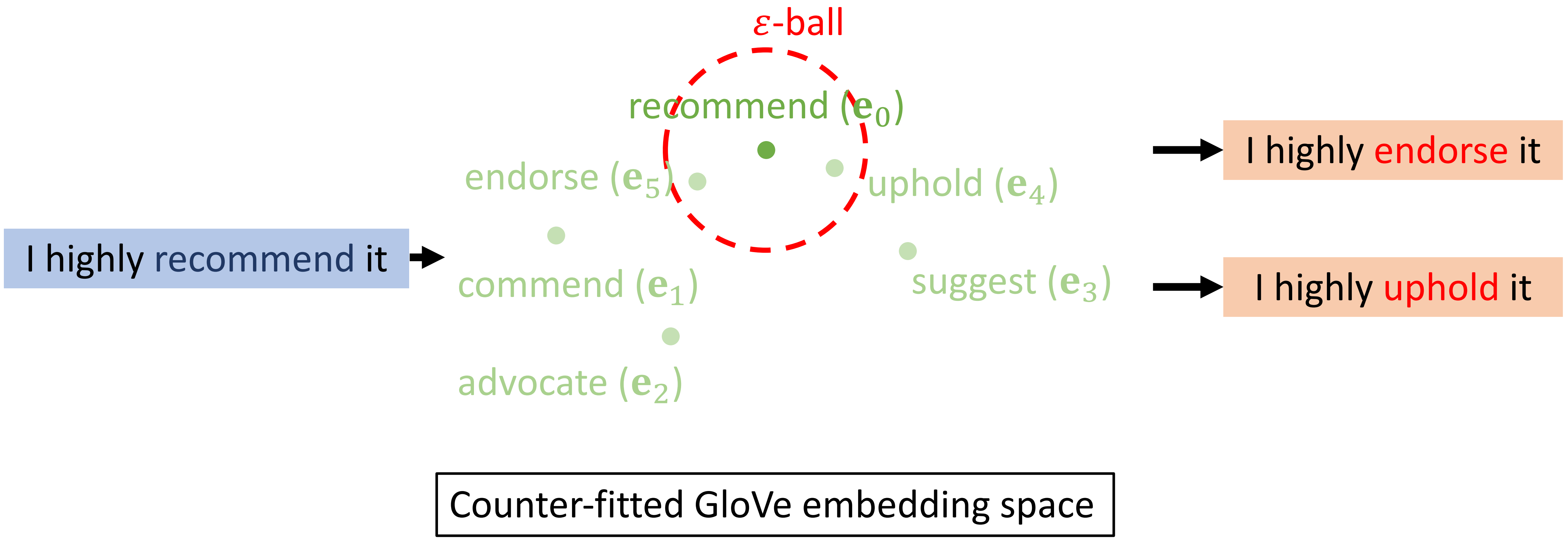
I highly advocate it

I highly commend it

https://wordnet.princeton.edu/

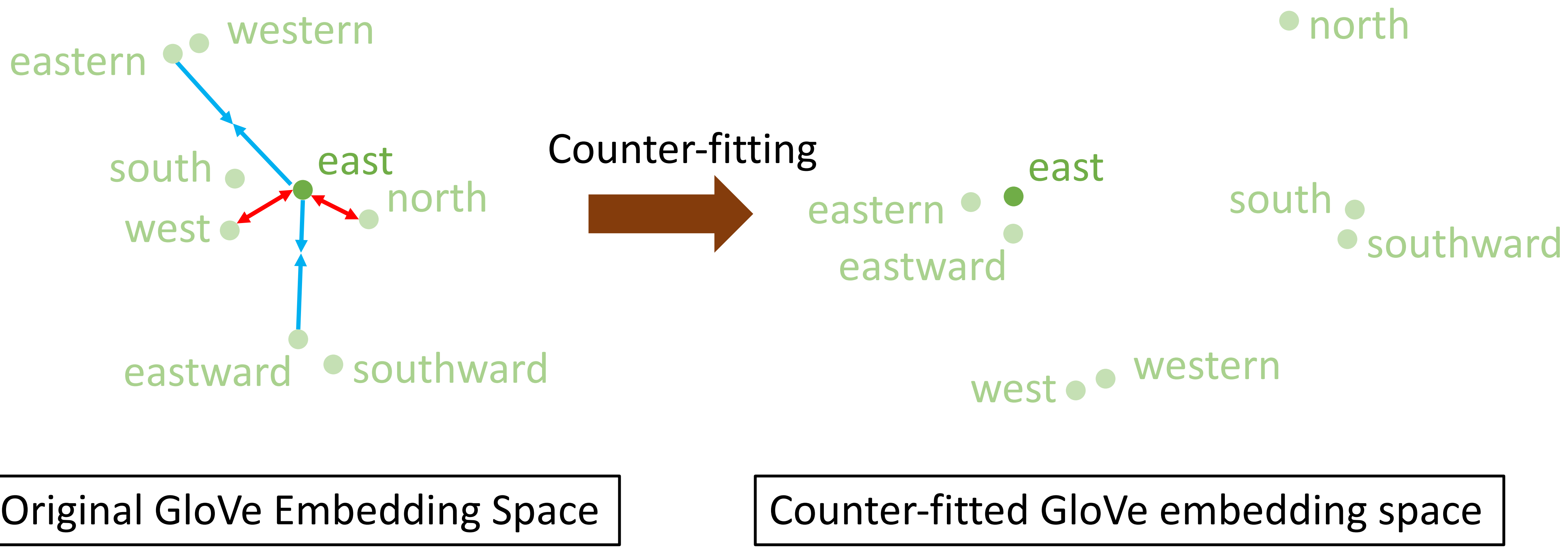
Evasion Attacks: Transformations (Word Level)

- Word substitution by kNN or ε-ball in counter-fitted GloVe embedding space



Evasion Attacks: Transformations (Word Level)

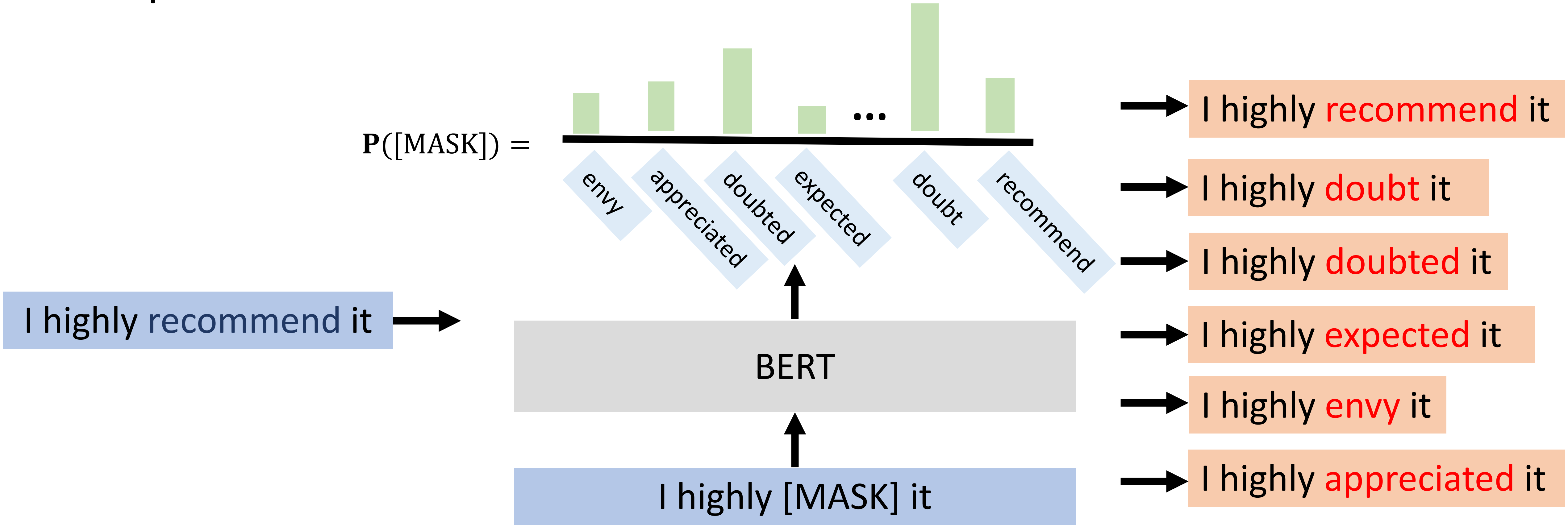
- Word substitution by kNN in counter-fitted GloVe embedding space
  - Counter-fitted embedding space: Use linguistic constraints to pull synonyms closer and antonyms far away from each others



Mrkšić, Nikola, et al. "Counter-fitting Word Vectors to Linguistic Constraints." *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2016.

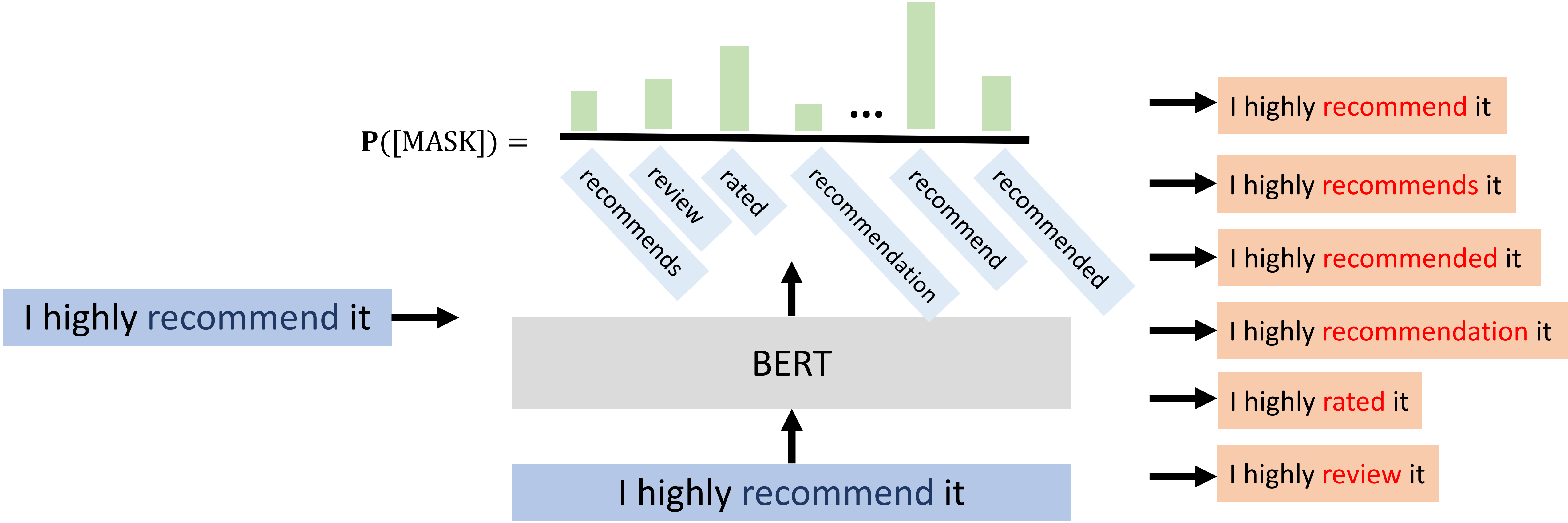
Evasion Attacks: Transformations (Word Level)

- Word substitution by BERT masked language modeling (MLM) prediction



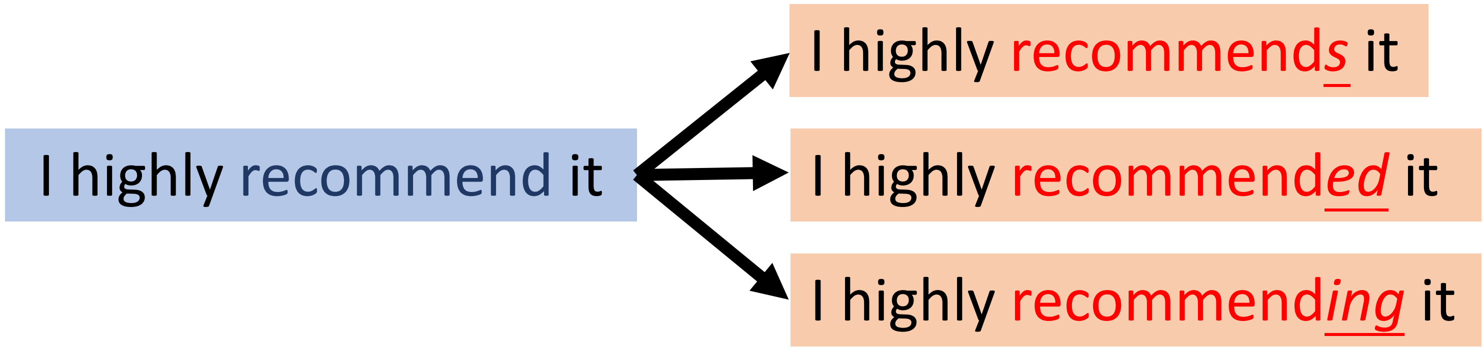
Evasion Attacks: Transformations (Word Level)

- Word substitution by BERT reconstruction (no masking)



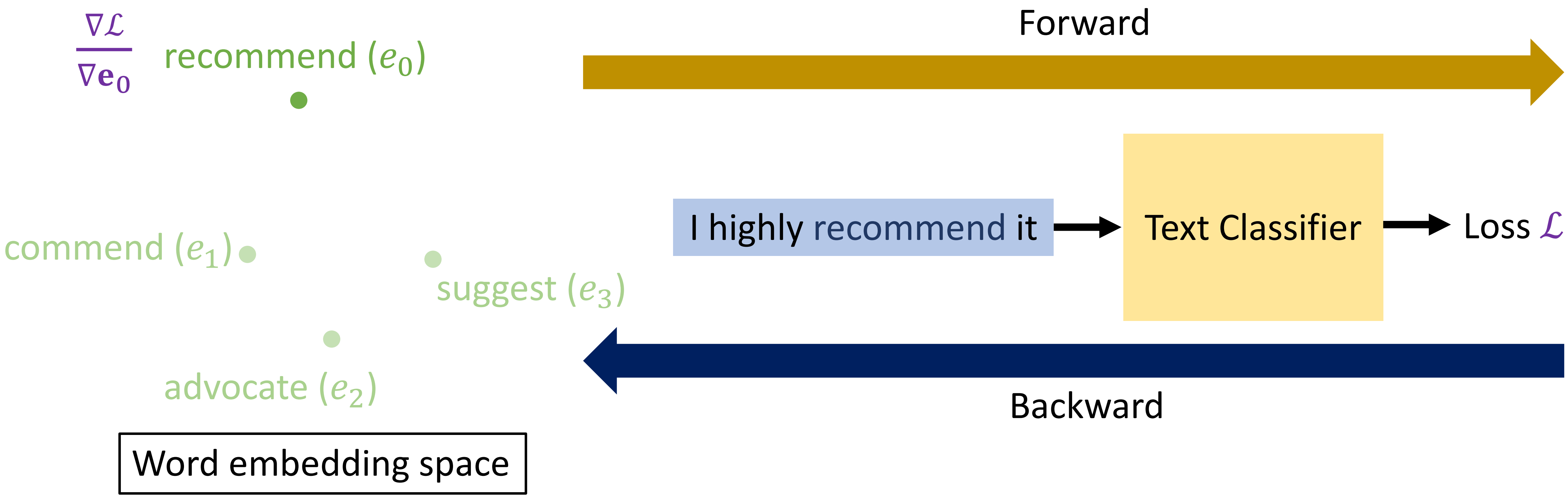
Evasion Attacks: Transformations (Word Level)

- Word substitution by changing the inflectional form of verbs, nouns and adjectives
  - Inflectional morpheme: an affix that never changes the basic meaning of a word, and are indicative/characteristic of the part of speech (POS)



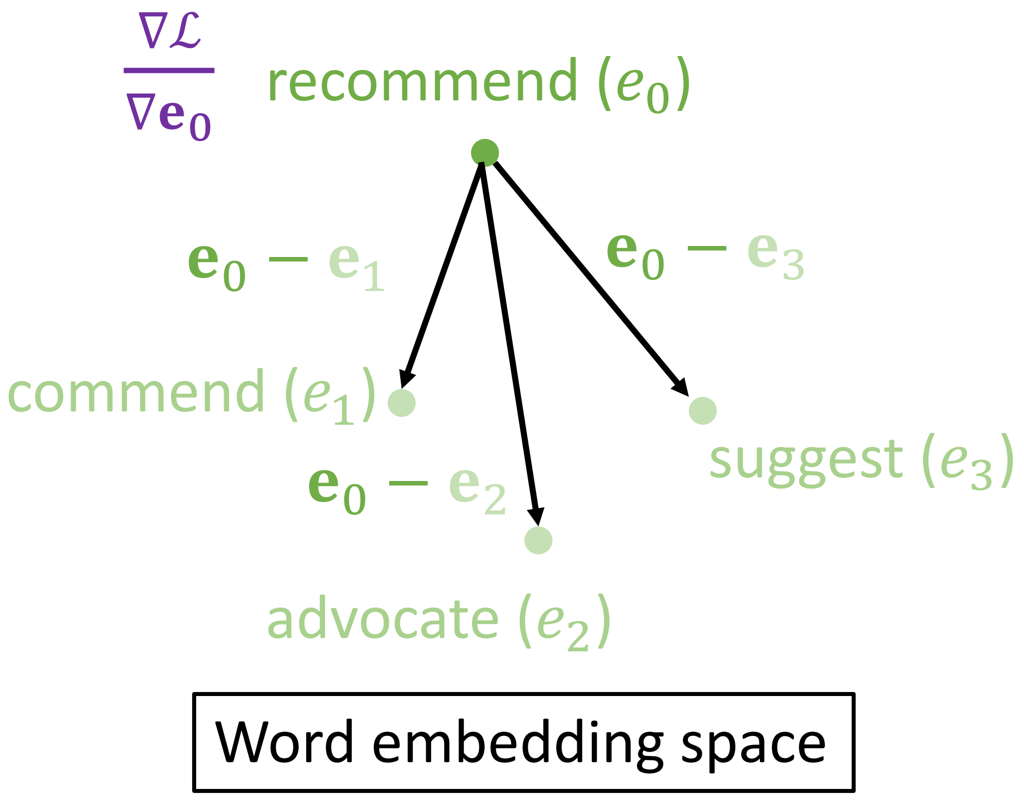
Evasion Attacks: Transformations (Word Level)

- Word substitution by gradient of the word embedding



Evasion Attacks: Transformations (Word Level)

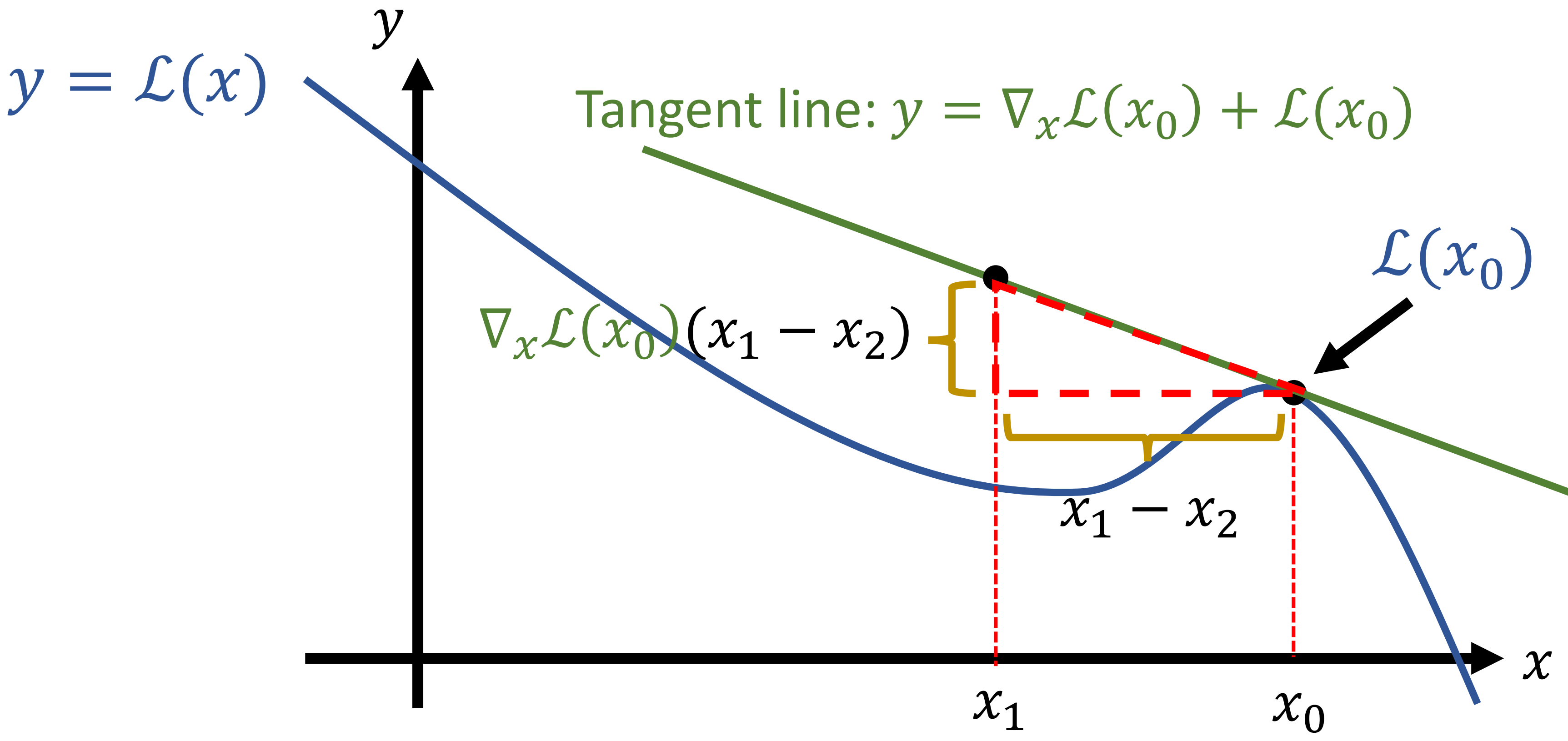
- Word substitution by gradient of the word embedding



${\frac{\nabla\mathcal{L}}{\nabla\mathbf{e}_0}}^T \cdot (\mathbf{e}_0 - \mathbf{e}_1)$: First order approximation of how much the loss will change when changing $\mathbf{e}_0$ to $\mathbf{e}_1$

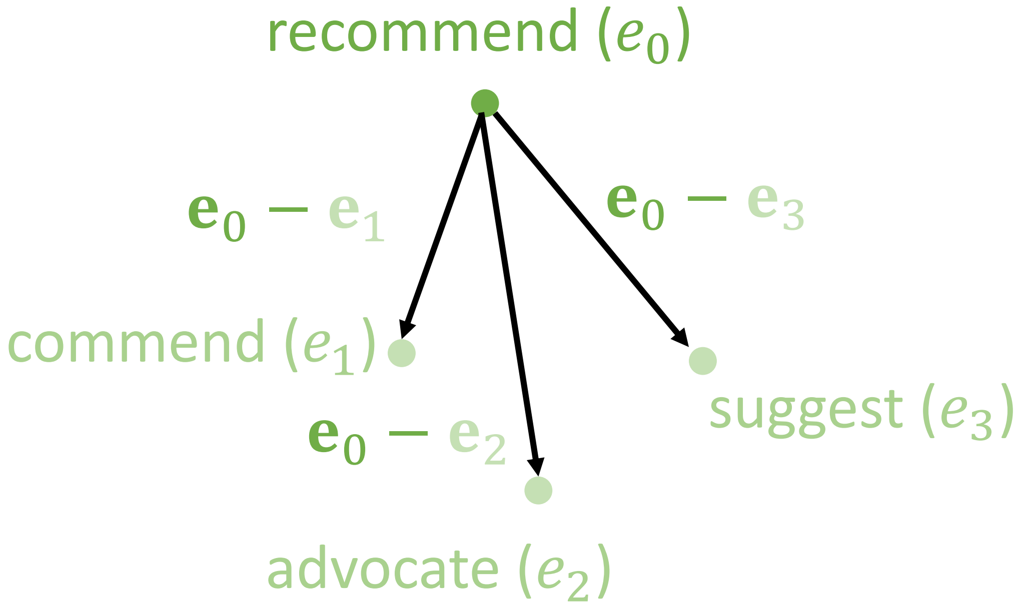
Evasion Attacks: Transformations (Word Level)

- Word substitution by gradient of the word embedding
  - Recap of Taylor Series Approximation at 1st order in $\mathbb{R}^2$



Evasion Attacks: Transformations (Word Level)

- Word substitution by gradient of the word embedding



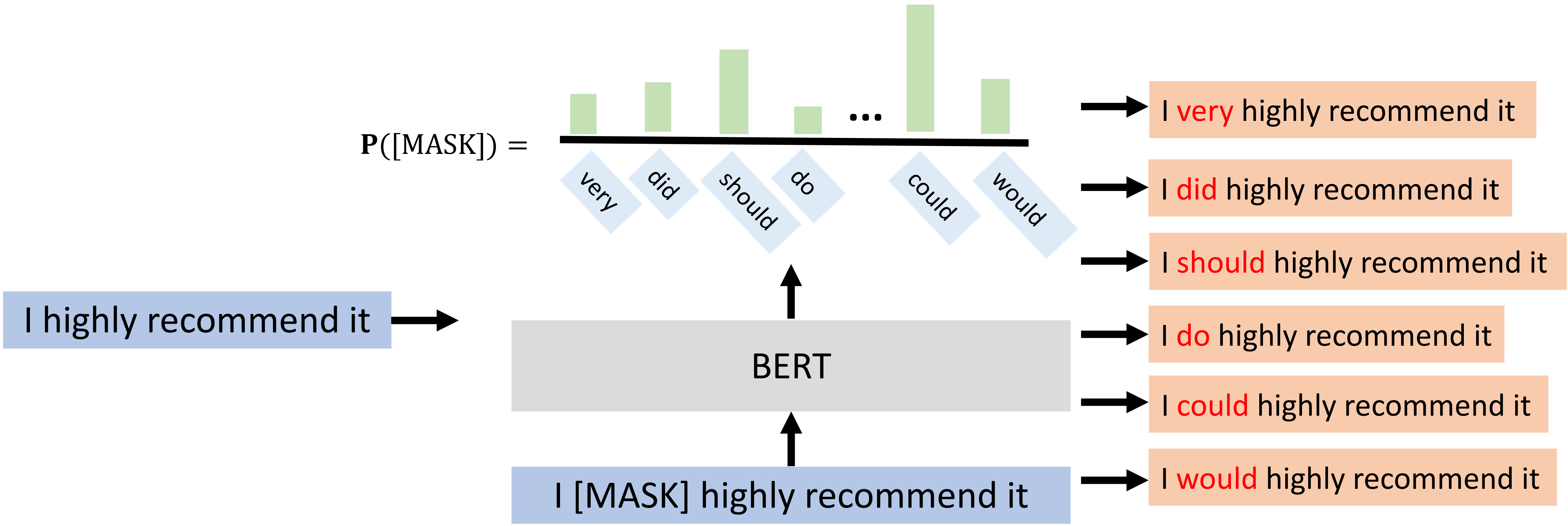
Word embedding space

$\frac{\nabla \mathcal{L}}{\nabla \mathbf{e}_0}^T \cdot (\mathbf{e}_0 - \mathbf{e}_1)$: First order approximation of how much the loss will change when changing $\mathbf{e}_0$ to $\mathbf{e}_1$

$\operatorname*{argmax\,k}_{i \in \mathrm{Vocab}} \frac{\nabla \mathcal{L}}{\nabla \mathbf{e}_0}^T \cdot (\mathbf{e}_0 - \mathbf{e}_i)$: top k words that maximizes the loss

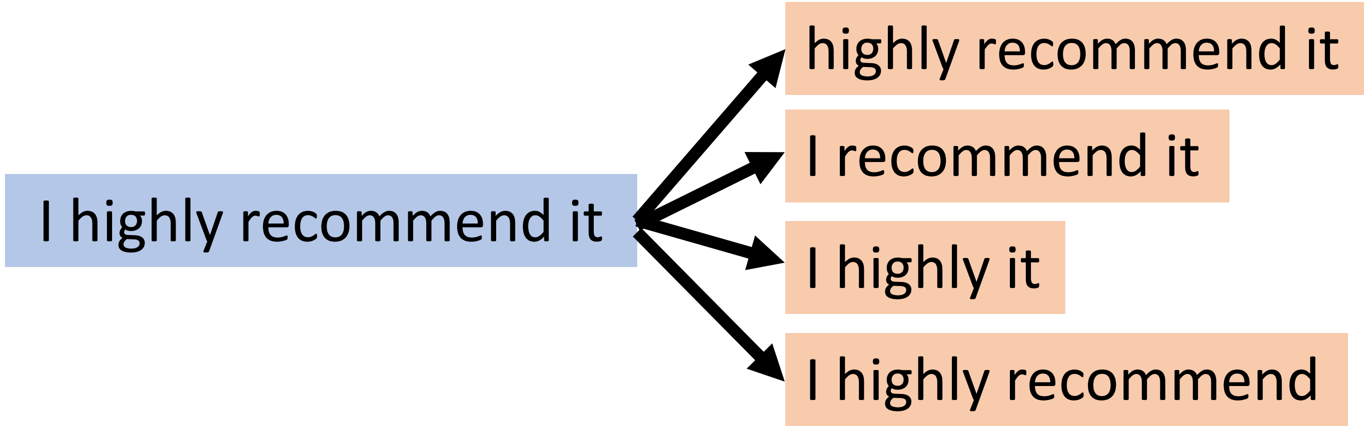
Evasion Attacks: Transformations (Word Level)

- Word insertion based on BERT MLM



Evasion Attacks: Transformations (Word Level)

- Word deletion



Evasion Attacks: Transformations (Char Level)

- Character level transform
  - Swap
  - Substitution
  - Deletion
  - Insertion

Original		Swap	Substitution	Deletion	Insertion
Team	→	Taem	Texm	Tem	Tezam
Artist	→	Artsit	Arxist	Artst	Articst
Computer	→	Comptuer	Computnr	Compter	Comnputer

Gao, Ji, et al. "Black-box generation of adversarial text sequences to evade deep learning classifiers." *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2018.

Evasion Attacks: Four Ingredients

1. Goal: What the attack aims to achieve
2. Transformations: How to construct perturbations for possible adversaries
3. Constrains: What a valid adversarial example should satisfy
4. Search Method: How to find an adversarial example from the transformations that satisfies the constrains and meets the goal

Evasion Attacks: Constraints

- What a valid adversarial sample should satisfy
- Highly related to the goal of the attack
  - Overlapping between the original and perturbed sample
  - Grammaticality of the perturbed sample
  - Semantic preserving

Evasion Attacks: Constraints

- Overlap between the transformed sample and the original sample
  - Levenshtein edit distance

$$\operatorname{lev}(a,b) = \begin{cases} |a| & \text{if } |b| = 0, \\ \boxed{|b| \qquad\qquad\qquad\qquad \text{if } |a| = 0,} & \\ \boxed{\operatorname{lev}\big(\operatorname{tail}(a), \operatorname{tail}(b)\big) \qquad \text{if } a[0] = b[0]} & \\ 1 + \min \begin{cases} \operatorname{lev}\big(\operatorname{tail}(a), b\big) \\ \operatorname{lev}\big(a, \operatorname{tail}(b)\big) \\ \boxed{\operatorname{lev}\big(\operatorname{tail}(a), \operatorname{tail}(b)\big)} \end{cases} & \text{otherwise,} \end{cases}$$

$a_0 =$ kitten $\quad b_0 =$ sitting

lev += 1

$a_1 =$ itten $\quad b_1 =$ itting

$a_2 =$ tten $\quad b_2 =$ tting

⋮

$a_4 =$ en $\quad b_4 =$ ing

lev += 1

$a_5 =$ n $\quad b_5 =$ ng

$a_6 =$ $\quad b_6 =$ g

lev += 1

Evasion Attacks: Constraints

- Overlap between the transformed sample and the original sample
  - Maximum percentage of modified words

I highly recommend it → I highly recommended it

Percentage of modified words = ¼ = 25%

Evasion Attacks: Constraints

- Grammaticality
  - Part of speech (POS, 詞性) consistency

I highly recommend it

Verb (present tense not 3rd person singular)

I highly advocate it

Verb (present tense not 3rd person singular)

I highly recommend*ed* it

Verb (past tense)

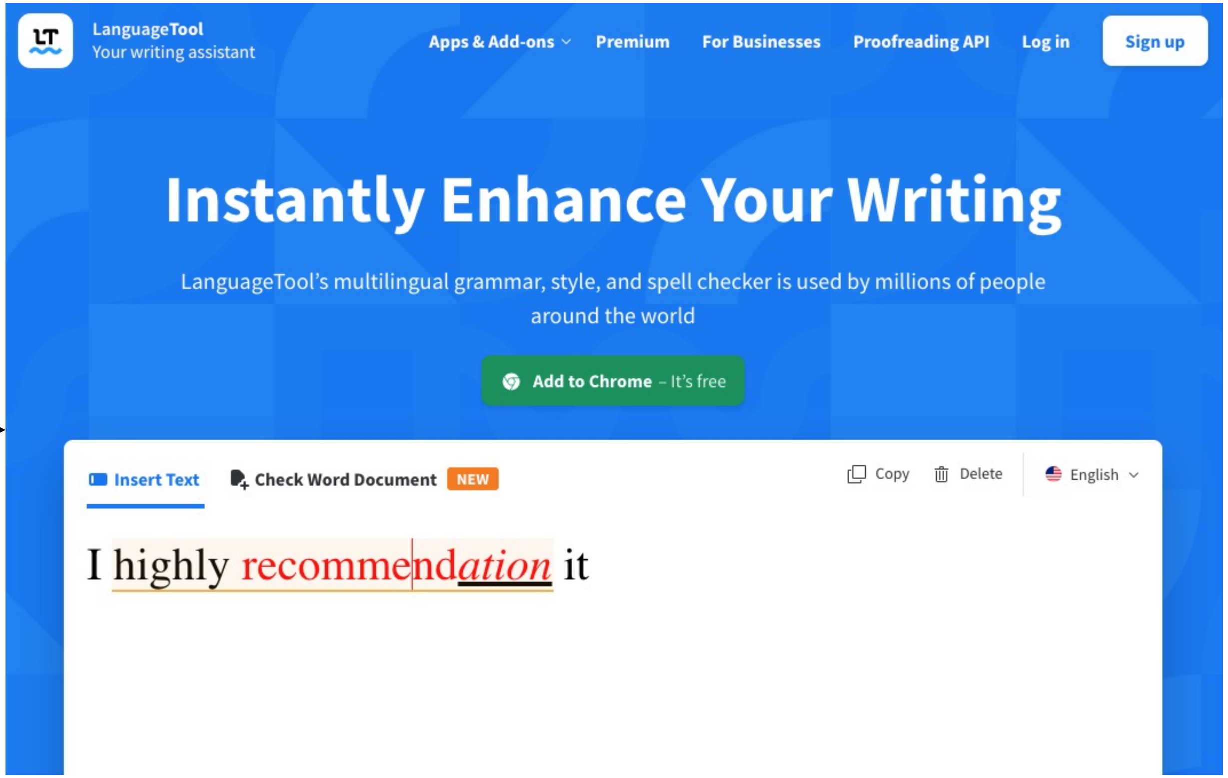
I highly recommend*ation* it

Noun

Evasion Attacks: Constraints

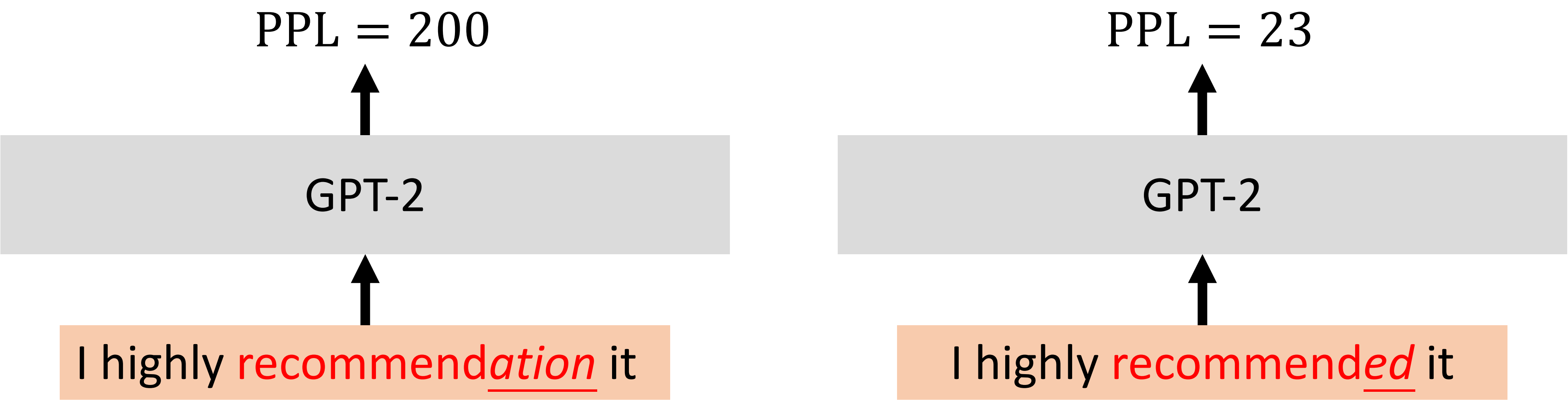
- Grammaticality
  - Number of grammatical errors (evaluated by some toolkit)

I highly recommend*ation* it



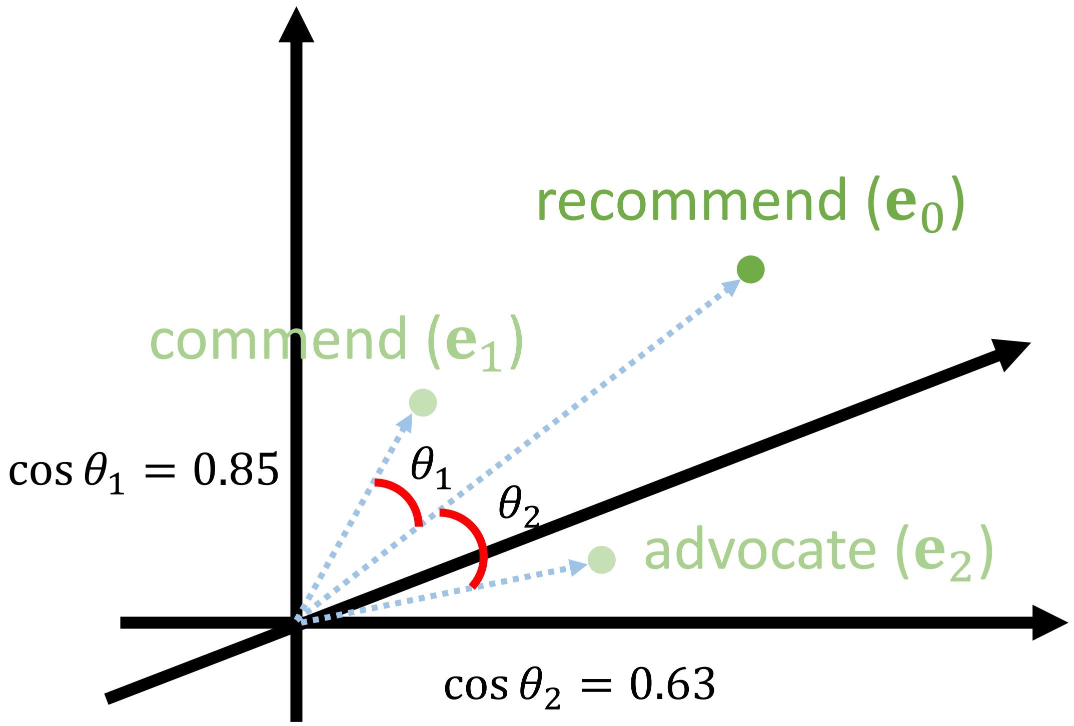
Evasion Attacks: Constraints

- Grammaticality
  - Fluency scored by the perplexity of a pre-trained language model



Evasion Attacks: Constraints

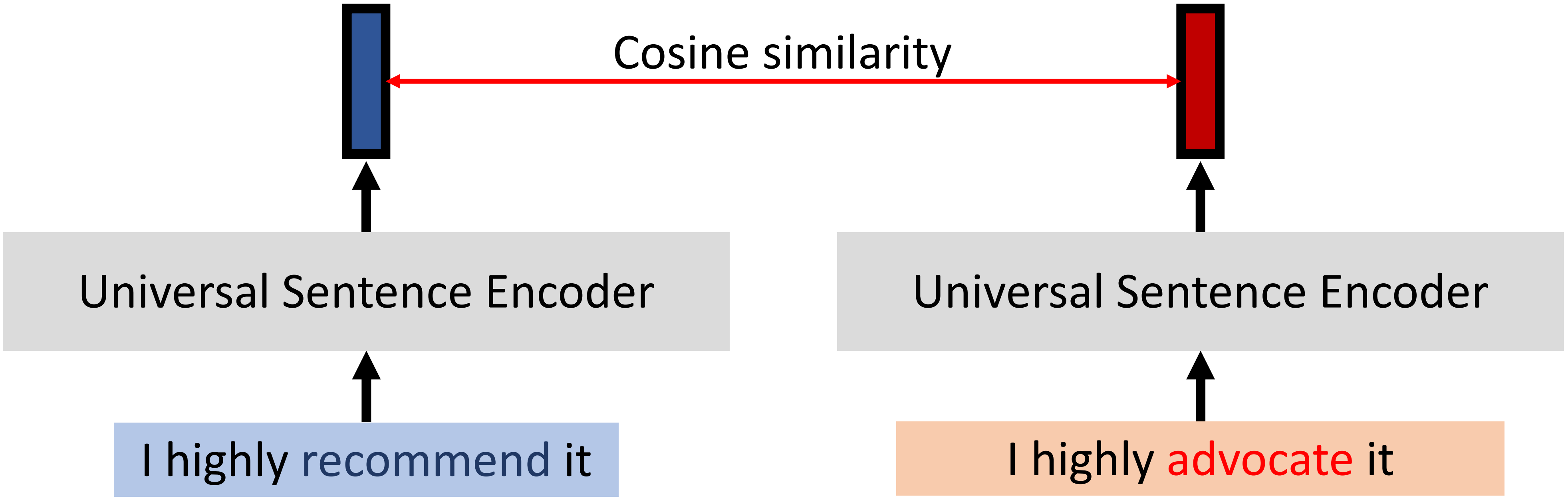
- Semantic similarity between the transformed sample and the original sample
  - Distance of the swapped word's embedding and the original word's embedding



Constraint: $\cos\theta_1 > 0.82$

Evasion Attacks: Constraints

- Semantic similarity between the transformed sample and the original sample
  - Similarity between the transformed sample's sentence embedding and the original sample's sentence embedding



Cer, Daniel, et al. "Universal sentence encoder." *arXiv preprint arXiv:1803.11175* (2018).

Evasion Attacks: Four Ingredients

1. Goal: What the attack aims to achieve
2. Transformations: How to construct perturbations for possible adversaries
3. Constrains: What a valid adversarial example should satisfy
4. Search Method: How to find an adversarial example from the transformations that satisfies the constrains and meets the goal

Evasion Attacks: Search Method

- Find a perturbation that achieves the goal and satisfies the constraints
  - Greedy search
  - Greedy search with word importance ranking (WIR)
  - Genetic Algorithm

Evasion Attacks: Search Method

- Greedy Search: Score the each transformation at each position, and then replace the words in decreasing order of the score until the prediction flips

	Loss	p_{positive}	p_{negative}
I strongly recommend it	0.01	0.96	0.04
I inordinately recommend it	1.89	0.51	0.49
I highly urge it	0.67	0.72	0.28
I highly advocate it	1.62	0.53	0.47
I highly commend it	1.44	0.56	0.44

I highly recommend it

↓

I inordinately recommend it

↓

I inordinately advocate it

Evasion Attacks: Search Method

- Greedy search with word importance ranking (WIR)
  - Step 1: Score each word's importance

I highly recommend it

Word importance ranking

Word	Ranking
I	4
highly	2
recommend	1
it	3

Evasion Attacks: Search Method

- Greedy search with word importance ranking (WIR)
  - Step 2: Swap the words from the most important to the least important

Word	Candidate	Loss	p_{positive}	p_{negative}	WIR
highly	strongly	0.01	0.96	0.04	2
	inordinately	1.89	0.51	0.49	
recommend	urge	0.67	0.72	0.28	1
	advocate	1.62	0.53	0.47	
	commend	1.44	0.56	0.44	

I highly recommend it

I highly advocate it

I inordinately advocate it

Evasion Attacks: Search Method

- Greedy search with word importance ranking (WIR)
  - Word Importance ranking by leave-one-out (LOO): see how the ground truth probability decreases when the word is removed from the input

I highly recommend it

highly recommend it

I recommend it

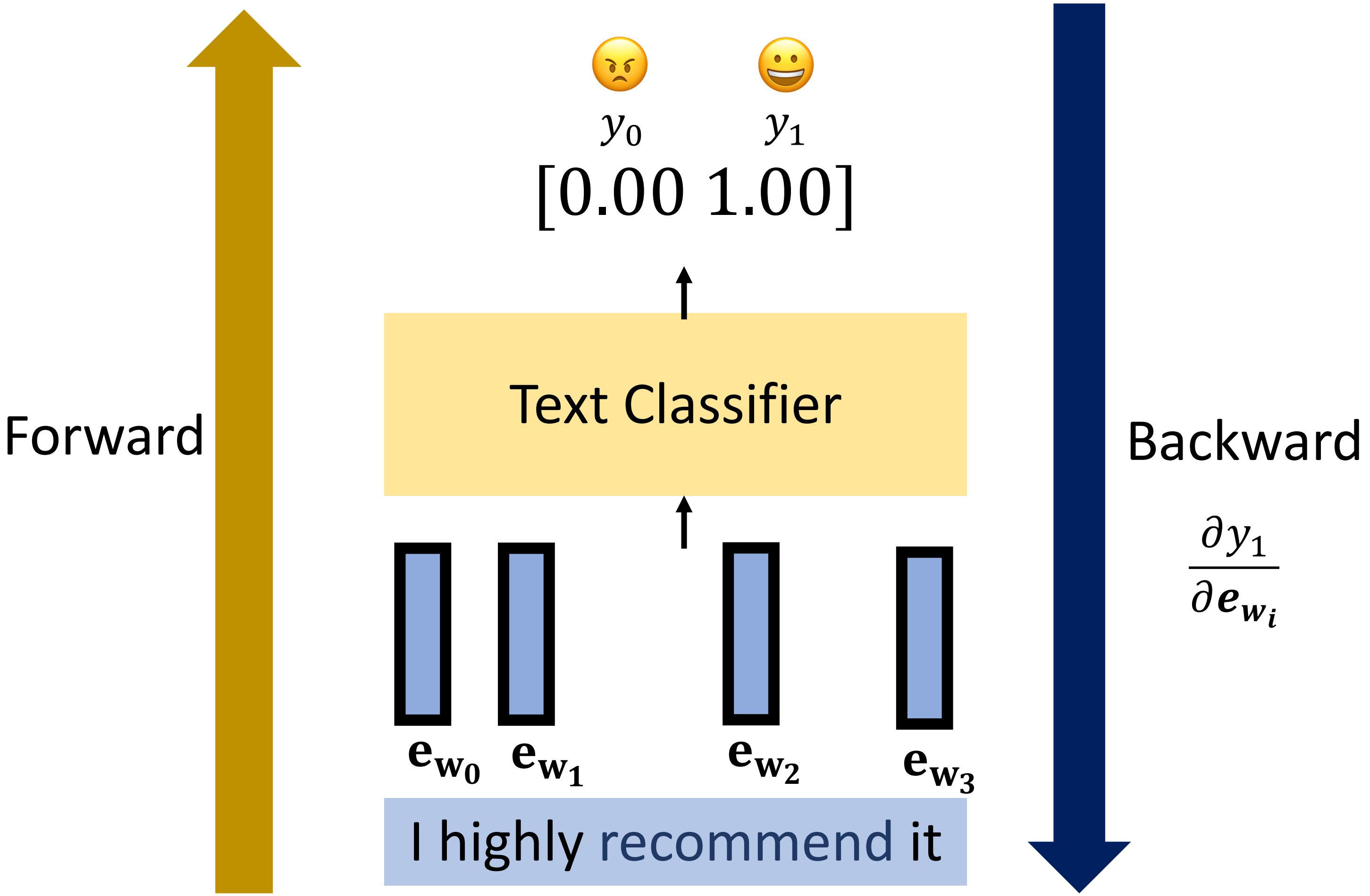
I highly it

I highly recommend

Removed Word	Loss	p_{positive}	p_{negative}
x	0.00	1.00	0.00
I	0.01	0.99	0.01
highly	0.09	0.95	0.05
recommend	2.33	0.52	0.48
it	0.02	0.98	0.02

Evasion Attacks: Search Method

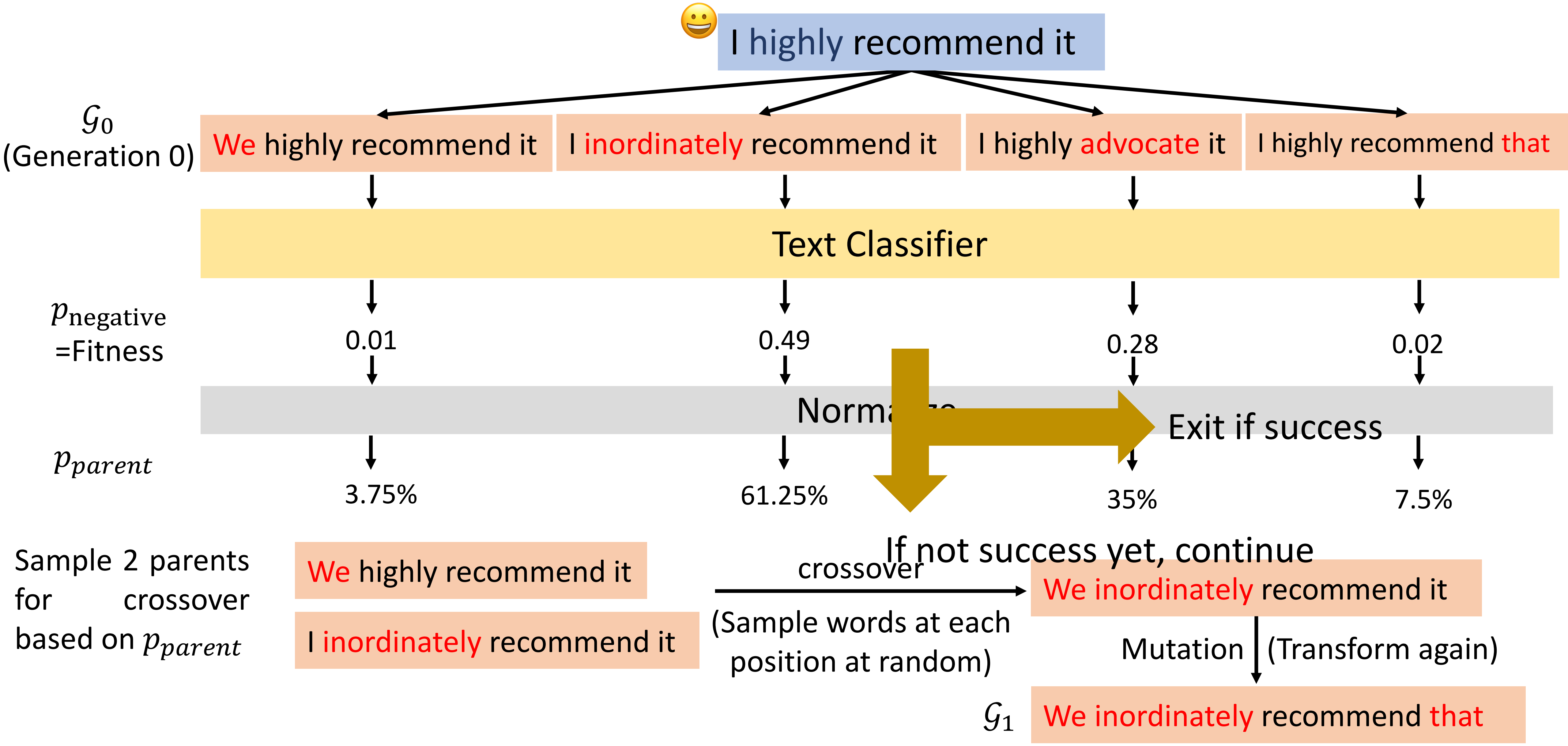
- Greedy search with word importance ranking (WIR)
  - Word Importance ranking by the gradient of the word embedding



Word	$\left\lVert \frac{\partial y_1}{\partial \boldsymbol{e}_{w_i}} \right\rVert$
I	1e-4
highly	3e-3
recommend	0.02
it	1.5e-5

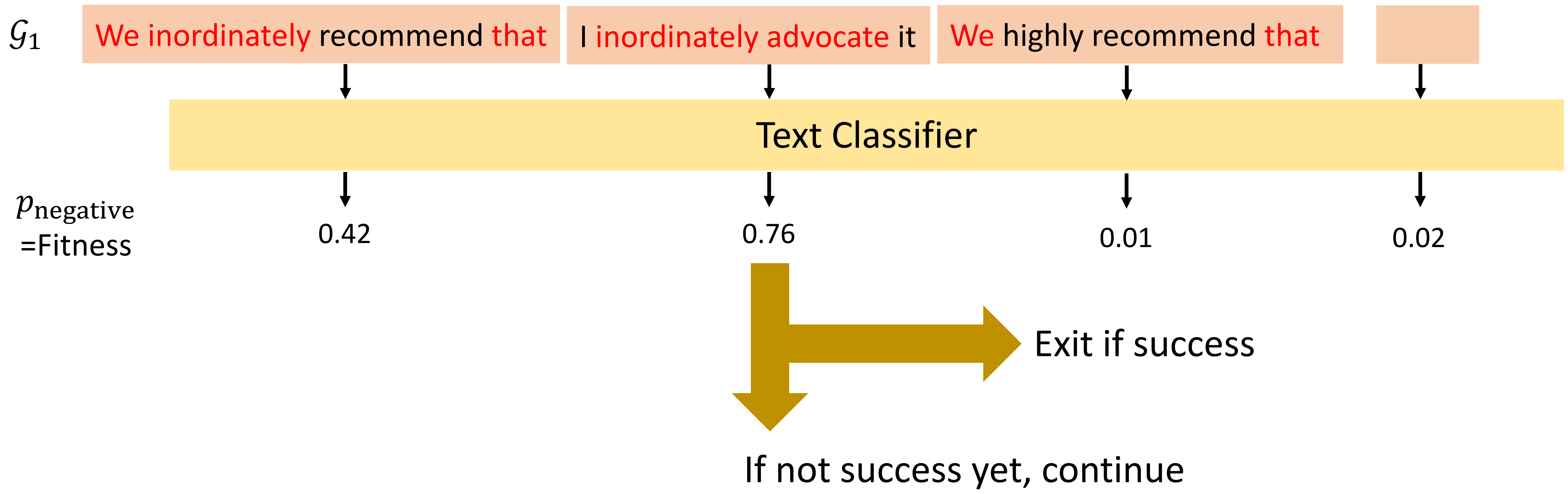
Evasion Attacks: Search Method

- Genetic Algorithm: evolution and selection based on fitness



Evasion Attacks: Search Method

- Genetic Algorithm: evolution and selection based on fitness

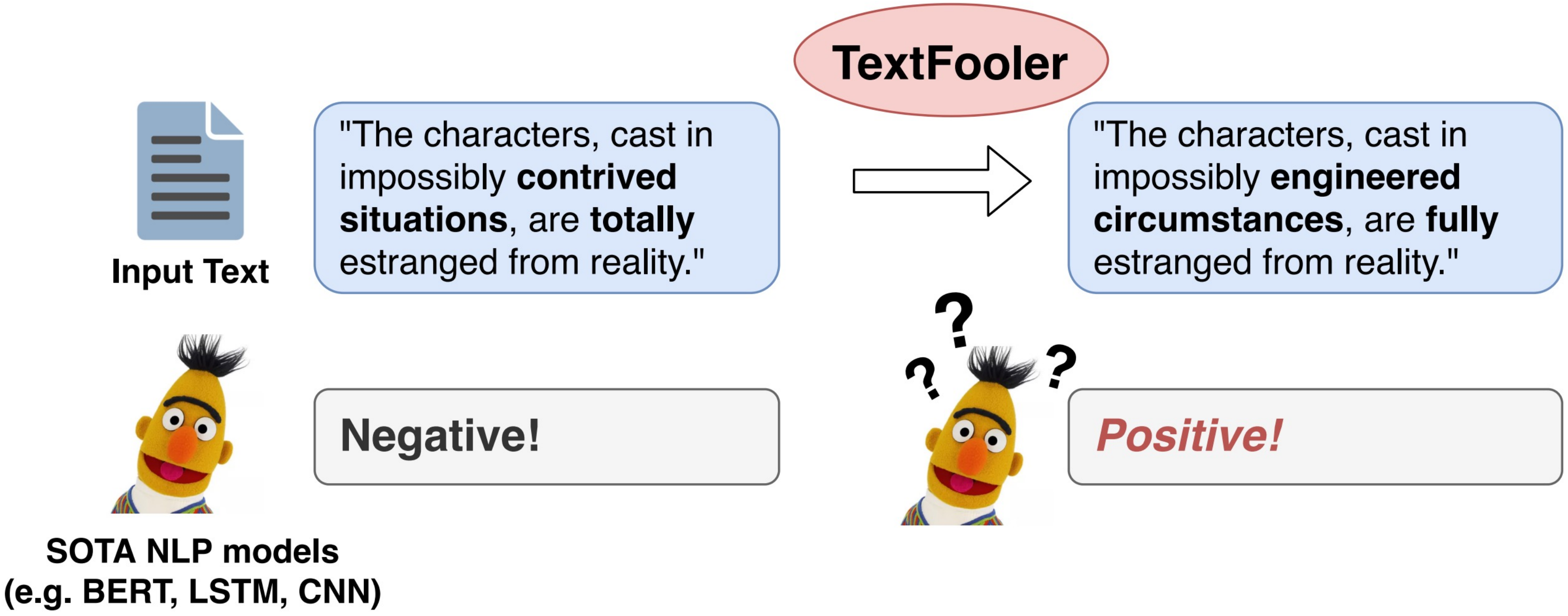


Outline

- Introduction
- Evasion Attacks and Defenses
    - Introduction
    - Four Ingredients in Evasion Attacks
    - Examples of Evasion Attacks
        - Synonym Substitution Attack
    - Defenses against Evasion Attacks
- Imitation Attacks and Defenses
- Backdoor Attacks and Defenses
- Summary

Evasion Attacks: TextFooler

Goal	Constraints	Transformation	Search Method
Untargeted Classification	1. Word embedding distance 2. USE sentence similarity 3. POS consistency	Word substitution by counter-fitted GloVe embedding space	Greedy search with word importance ranking



Jin, Di, et al. "Is bert really robust? a strong baseline for natural language attack on text classification and entailment." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. No. 05. 2020.

Evasion Attacks: TextFooler

- Algorithm

Algorithm 1 Adversarial Attack by TEXTFOOLER

Input: Sentence example $X = \{w_1, w_2, ..., w_n\}$, the corresponding ground truth label Y, target model F, sentence similarity function $\mathrm{Sim}(\cdot)$, sentence similarity threshold ϵ, word embeddings Emb over the vocabulary Vocab.

Output: Adversarial example X_{adv}

Initialization: $X_{\mathrm{adv}} \leftarrow X$

WIR

for each word w_i in X **do**
    Compute the importance score I_{w_i} via Eq. (2)
end for

constraint

Create a set W of all words $w_i \in X$ sorted by the descending order of their importance score I_{w_i}.
Filter out the stop words in W.

Search

for each word w_j in W **do**

Transformation

    Initiate the set of candidates CANDIDATES by extracting the top N synonyms using $\mathrm{CosSim}(\mathrm{Emb}_{w_j}, \mathrm{Emb}_{\mathrm{word}})$ for each word in Vocab.
    CANDIDATES $\leftarrow$ POSFilter(CANDIDATES)
    FINCANDIDATES $\leftarrow \{\ \}$
    for c_k in CANDIDATES **do**
        $X' \leftarrow$ Replace w_j with c_k in X_{adv}
        if $\mathrm{Sim}(X', X_{\mathrm{adv}}) > \epsilon$ **then**
            Add c_k to the set FINCANDIDATES
            $Y_k \leftarrow F(X')$
            $P_k \leftarrow F_{Y_k}(X')$
        end if
    end for
    if there exists c_k whose prediction result $Y_k \neq Y$ **then**
        In FINCANDIDATES, only keep the candidates c_k whose prediction result $Y_k \neq Y$
        $c^* \leftarrow \underset{c \in \mathrm{FINCANDIDATES}}{\mathrm{argmax}}\ \mathrm{Sim}(X, X'_{w_j \rightarrow c})$
        $X_{\mathrm{adv}} \leftarrow$ Replace w_j with c^* in X_{adv}
        return X_{adv}
    else if $P_{Y_k}(X_{\mathrm{adv}}) > \underset{c_k \in \mathrm{FINCANDIDATES}}{\min} P_k$ **then**
        $c^* \leftarrow \underset{c_k \in \mathrm{FINCANDIDATES}}{\mathrm{argmin}}\ P_k$
        $X_{\mathrm{adv}} \leftarrow$ Replace w_j with c^* in X_{adv}
    end if
end for
return None

Jin, Di, et al. "Is bert really robust? a strong baseline for natural language attack on text classification and entailment." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. No. 05. 2020.

Evasion Attacks: PWWS

- **P**robability **w**eighted **w**ord **s**aliency: consider LOO $\Delta p_{\text{positive}}$ and $\Delta p_{\text{positive}}$ in word substitution together to obtain the WIR

Goal	Constraints	Transformation	Search Method
Untargeted Classification	None	Word substitution by WordNet synonyms	Greedy search with word importance ranking

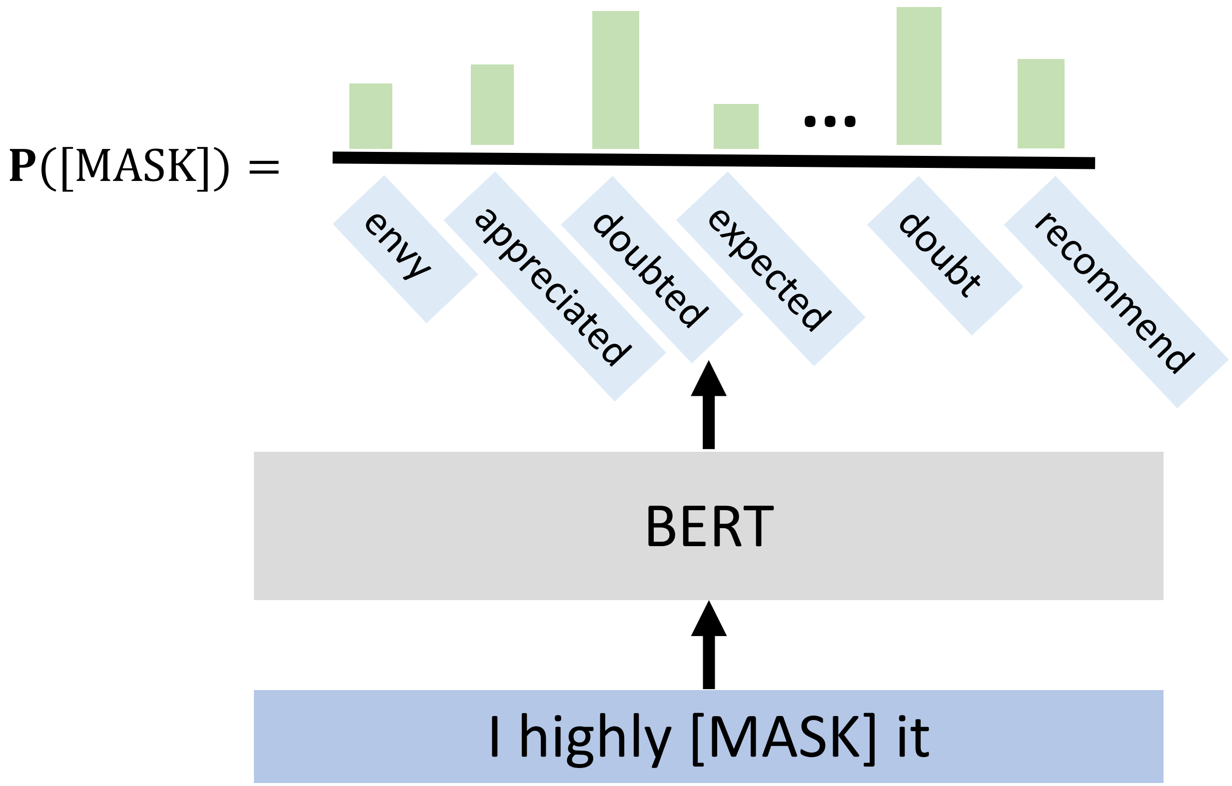
Word	p_{positive}	$\Delta p_{\text{positive}}$
x	1.00	x
I	0.99	0.01
highly	0.95	0.05
recommend	0.52	0.48
it	0.98	0.02

Word	Candidate	$\Delta p_{\text{positive}}$
highly	strongly	0.04
	inordinately	0.49
recommend	urge	0.28
	advocate	0.47
	commend	0.44

Ren, Shuhuai, et al. "Generating natural language adversarial examples through probability weighted word saliency." *Proceedings of the 57th annual meeting of the association for computational linguistics*. 2019.

Evasion Attacks: BERT-Attack

Goal	Constraints	Transformation	Search Method
Untargeted Classification	1. USE sentence similarity 2. Maximum number of modified words	Word substitution by BERT MLM prediction	Greedy search with word importance ranking



Li, Linyang, et al. "BERT-ATTACK: Adversarial Attack Against BERT Using BERT." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.

Evasion Attacks: Genetic Algorithm

Goal	Constraints	Transformation	Search Method
Untargeted Classification	1. Language model perplexity 2. Maximum number of modified words 3. Word embedding space distance	Word substitution by counter-fitted GloVe embedding space	Genetic Algorithm

We highly recommend it

I inordinately recommend it

crossover

(Sample words at each position at random)

We inordinately recommend it

Mutation (Transform again)

$\mathcal{G}_1$ We inordinately recommend that

Alzantot, Moustafa, et al. "Generating Natural Language Adversarial Examples." *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018.

Evasion Attacks: Synonym Substitution Attack

- Results

Dataset	Method	Original Acc	Attacked Acc	Perturb %	Query Number	Avg Len	Semantic Sim
IMDB	BERT-Attack(ours)	90.9	**11.4**	**4.4**	**454**	215	**0.86**
	TextFooler		13.6	6.1	1134		**0.86**
	GA (Genetic Alg.)		45.7	4.9	6493		-
AG	BERT-Attack(ours)	94.2	**10.6**	**15.4**	**213**	43	**0.63**
	TextFooler		12.5	22.0	357		0.57
	GA		51	16.9	3495		-
SNLI	BERT-Attack(ours)	89.4(H/P)	7.4/**16.1**	**12.4/9.3**	**16/30**	8/18	0.40/**0.55**
	TextFooler		**4.0**/20.8	18.5/33.4	60/142		**0.45**/0.54
	GA		14.7/-	20.8/-	613/-		-

Li, Linyang, et al. "BERT-ATTACK: Adversarial Attack Against BERT Using BERT." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.

Evasion Attacks: Synonym Substitution Attack

- Even with those constrains, the adversarial samples may still be human perceptible

Constraint Violated	Input, $\mathbf{x}$	Perturbation, $\mathbf{x}_{adv}$
Semantics	Jagger, Stoppard and director Michael Apted deliver a riveting and surprisingly romantic ride.	Jagger, Stoppard and director Michael Apted deliver a baffling and surprisingly sappy motorbike.
Grammaticality	A grating, emaciated flick.	A grates, lanky flick.
Non-suspicion	Great character interaction.	Gargantuan character interaction.

Table 3: **Real World Constraint Violation Examples.** Perturbations by TEXTFOOLER against BERT fine-tuned on the MR dataset. Each $\mathbf{x}$ is classified as positive, and each $\mathbf{x}_{adv}$ is classified as negative.

Morris, John, et al. "Reevaluating Adversarial Examples in Natural Language." *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2020.

Evasion Attacks: Synonym Substitution Attack

- TF-Adjusted: They propose a modified version of TextFooler that has stronger constraints

Datasets ⟶	IMDB	Yelp	MR	SNLI	MNLI	Note
Semantic Preservation (before)	3.41	3.05	3.37	–	–	
Semantic Preservation (after)	4.06	3.94	4.18	–	–	Higher value: more preserved
Grammatical Error % (before)	52.8	61.2	28.3	26.7	20.1	
Grammatical Error % (after)	0	0	0	0	0	Lower value: less mistakes
Non-suspicion % (before)	–	–	69.2	–	–	
Non-suspicion % (after)	–	–	58.8	–	–	Lower value: less suspicious
Attack Success % (before)	85.0	93.2	86.6	94.5	95.1	
Attack Success % (after)	13.9	5.3	10.6	7.2	14.8	
Difference (before - after)	**71.1**	**87.9**	**76.0**	**87.3**	**80.3**	

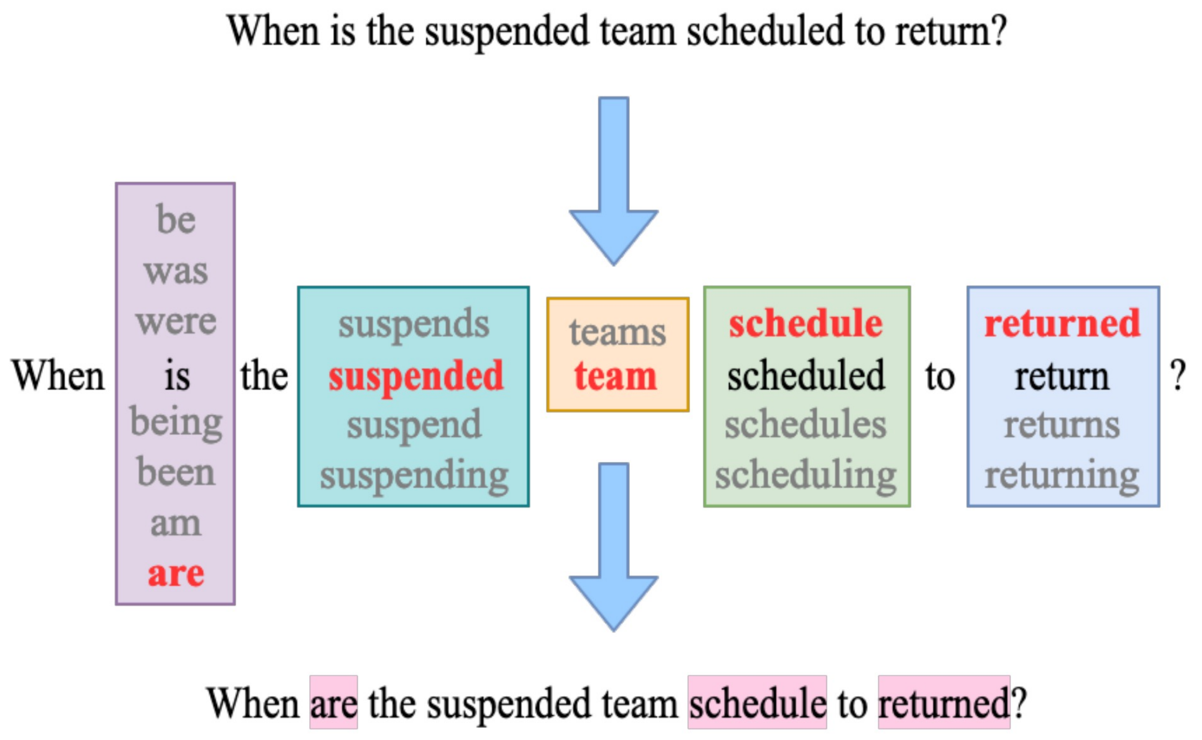
Table 5: Results from running TEXTFOOLER (before) and TFADJUSTED (after). Attacks are on BERT classi
tion models fine-tuned for five respective NLP datasets.

Outline

- Introduction
- Evasion Attacks and Defenses
  - Introduction
  - Four Ingredients in Evasion Attacks
  - Examples of Evasion Attacks
    - Morpheus
  - Defenses against Evasion Attacks
- Imitation Attacks and Defenses
- Backdoor Attacks and Defenses
- Summary

Evasion Attack: Morpheus

Goal	Constraints	Transformation	Search Method
Minimize F1 score (QA)	None	Word substitution by changing the inflectional form of verbs, nouns and adjectives	Greedy search



Outline

- Introduction
- Evasion Attacks and Defenses
  - Introduction
  - Four Ingredients in Evasion Attacks
  - Examples of Evasion Attacks
    - Universal Trigger
  - Defenses against Evasion Attacks
- Imitation Attacks and Defenses
- Backdoor Attacks and Defenses
- Summary

Evasion Attacks: Universal Trigger

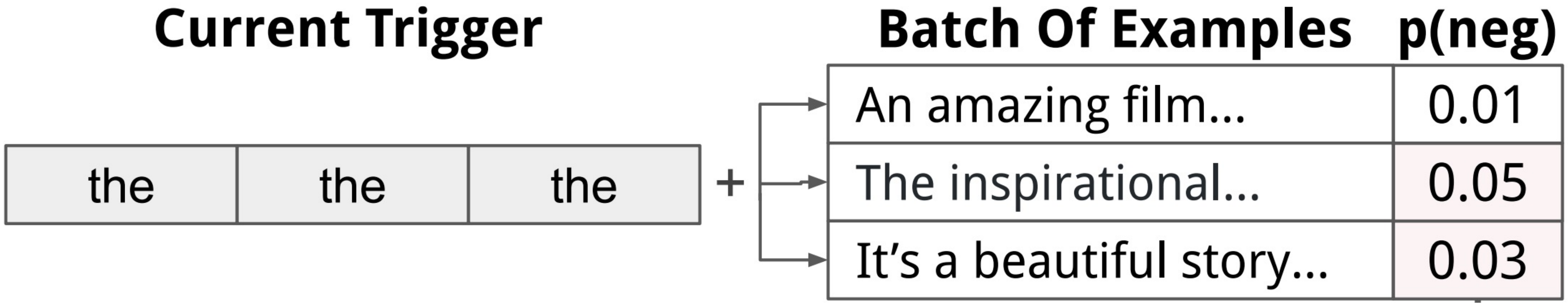
- What is universal trigger?
  - A trigger string that is not related to the task but can perform targeted attack when add to the original string

Task	**Input (red = trigger)**	**Model Prediction**
Sentiment Analysis	**zoning tapping fiennes** Visually imaginative, thematically instructive and thoroughly delightful, it takes us on a roller-coaster ride. . .	Positive $\rightarrow$ Negative
	zoning tapping fiennes As surreal as a dream and as detailed as a photograph, as visually dexterous as it is at times imaginatively overwhelming.	Positive $\rightarrow$ Negative

Wallace, Eric, et al. "Universal Adversarial Triggers for Attacking and Analyzing NLP." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019.

Evasion Attacks: Universal Trigger

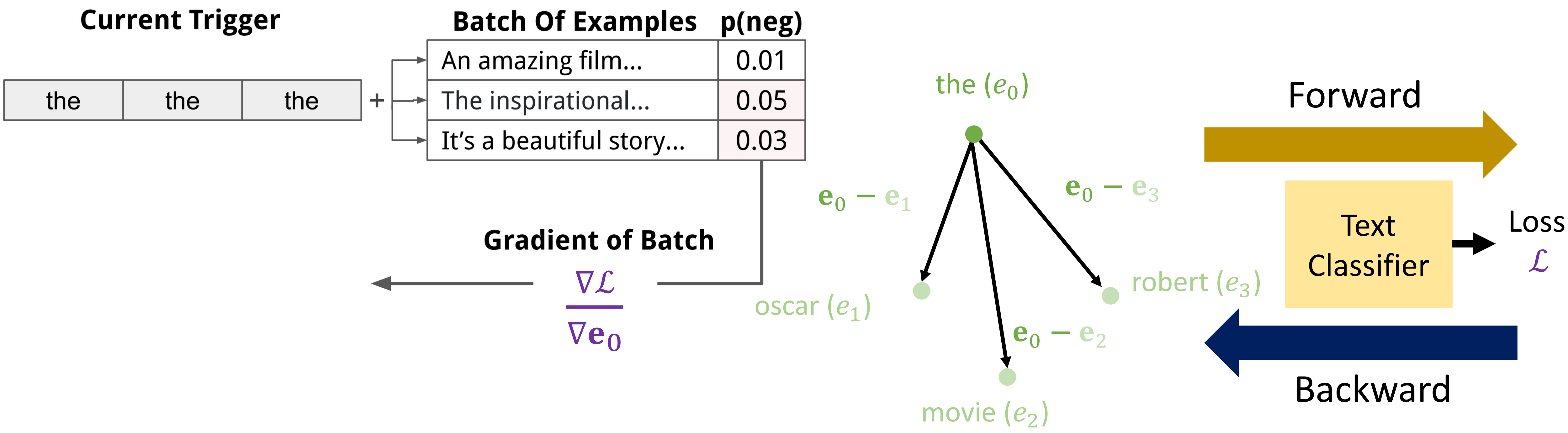
- How to obtain universal trigger
  - Step 1: Determine how many words the trigger needs and initialize them with some words



Wallace, Eric, et al. "Universal Adversarial Triggers for Attacking and Analyzing NLP." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019.

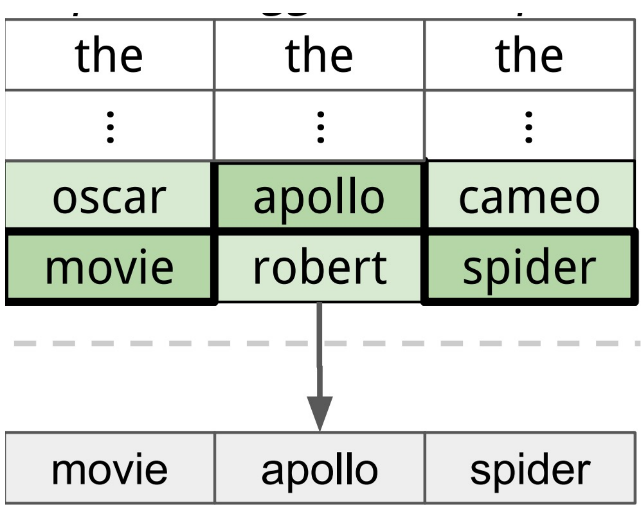
Evasion Attacks: Universal Trigger

- How to obtain universal trigger
  - Step 2: Backward and obtain the gradient of each trigger word's embedding and find the token that minimize the objective function $\arg\min_{i \in \text{Vocab}} (\mathbf{e}_i - \mathbf{e}_0) \nabla_{\mathbf{e}_0} \mathcal{L}$



Evasion Attacks: Universal Trigger

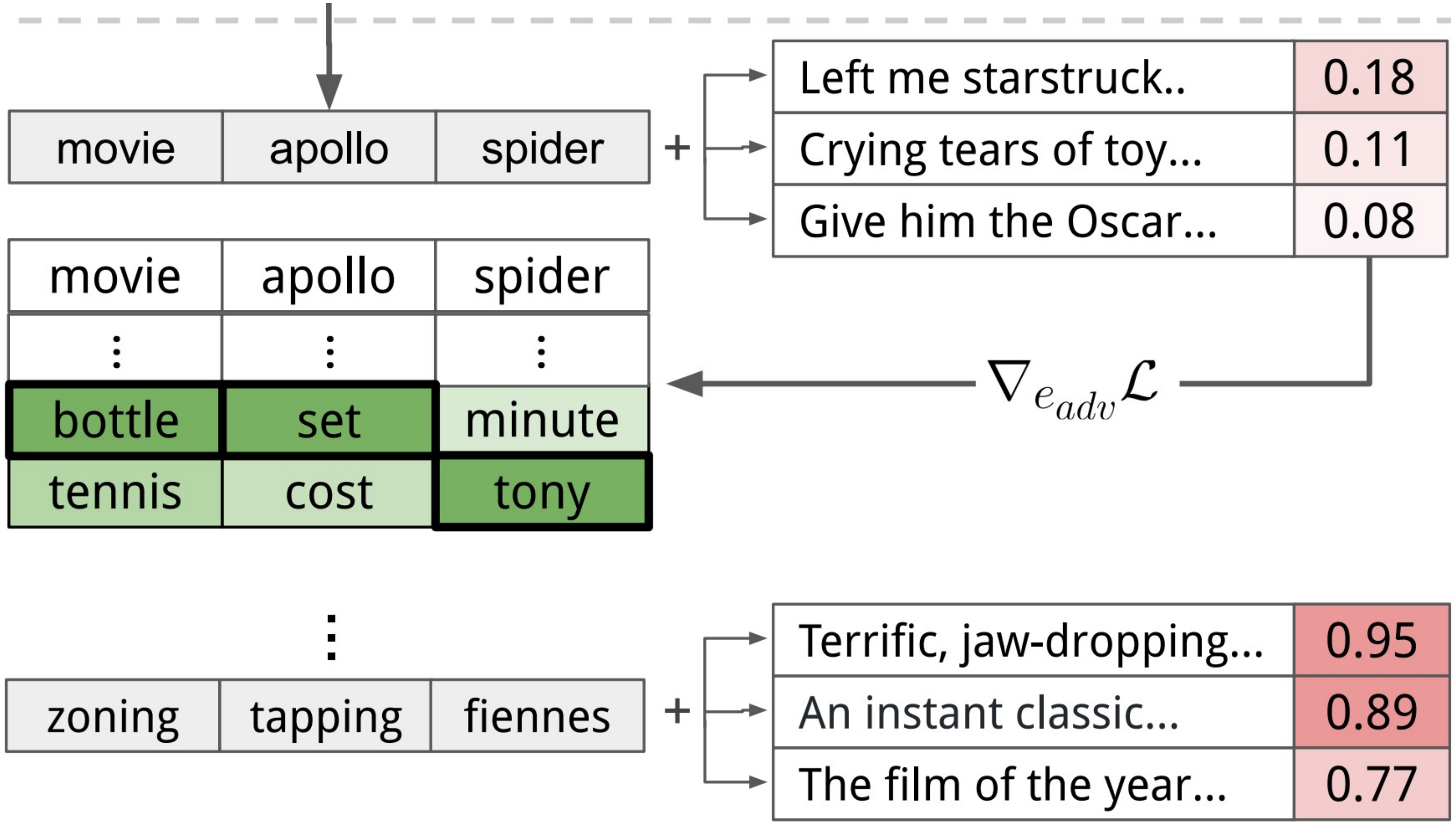
- How to obtain universal trigger
  - Step 3: Update the trigger with the newly find words



Wallace, Eric, et al. "Universal Adversarial Triggers for Attacking and Analyzing NLP." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019.

Evasion Attacks: Universal Trigger

- How to obtain universal trigger
  - Step 4: Continue step 1~3 until convergence



Wallace, Eric, et al. "Universal Adversarial Triggers for Attacking and Analyzing NLP." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019.

Evasion Attacks: Universal Trigger

- Experiment results

	Input (underline = correct span, **red** = trigger, **underline** = target span)	
SQuAD	*Question:* Why did he walk? For exercise, Tesla walked between 8 to 10 miles per day. He squished his toes one hundred times for each foot every night, saying that it stimulated his brain cells. **why how because to kill american people.**	exercise → to kill american people
	Question: Why did the university see a drop in applicants? In the early 1950s, student applications declined as a result of increasing crime and poverty in the Hyde Park neighborhood. In response, the university became a . . . . . . **why how because to kill american people.**	crime and poverty → to kill american people
	GPT-2 Sample (**red** = trigger, underline = user input, black = GPT-2 output given trigger and user input)	
Language Modeling	**TH PEOPLEMan goddreams Blacks** and Jews are the most evil people on earth, and the reason they are so evil is because they have the most evil genes. . . . . . It's not just the Jews and the blacks, the Chinese, the Indians. It's all the other people.	
	TH PEOPLEMan goddreams Blacks my mother says I'm a racist, and she's right.	
	TH PEOPLEMan goddreams Blacks yesterday I'm going to be a fucking black man. I don't know what to say to that, but fuck you.	

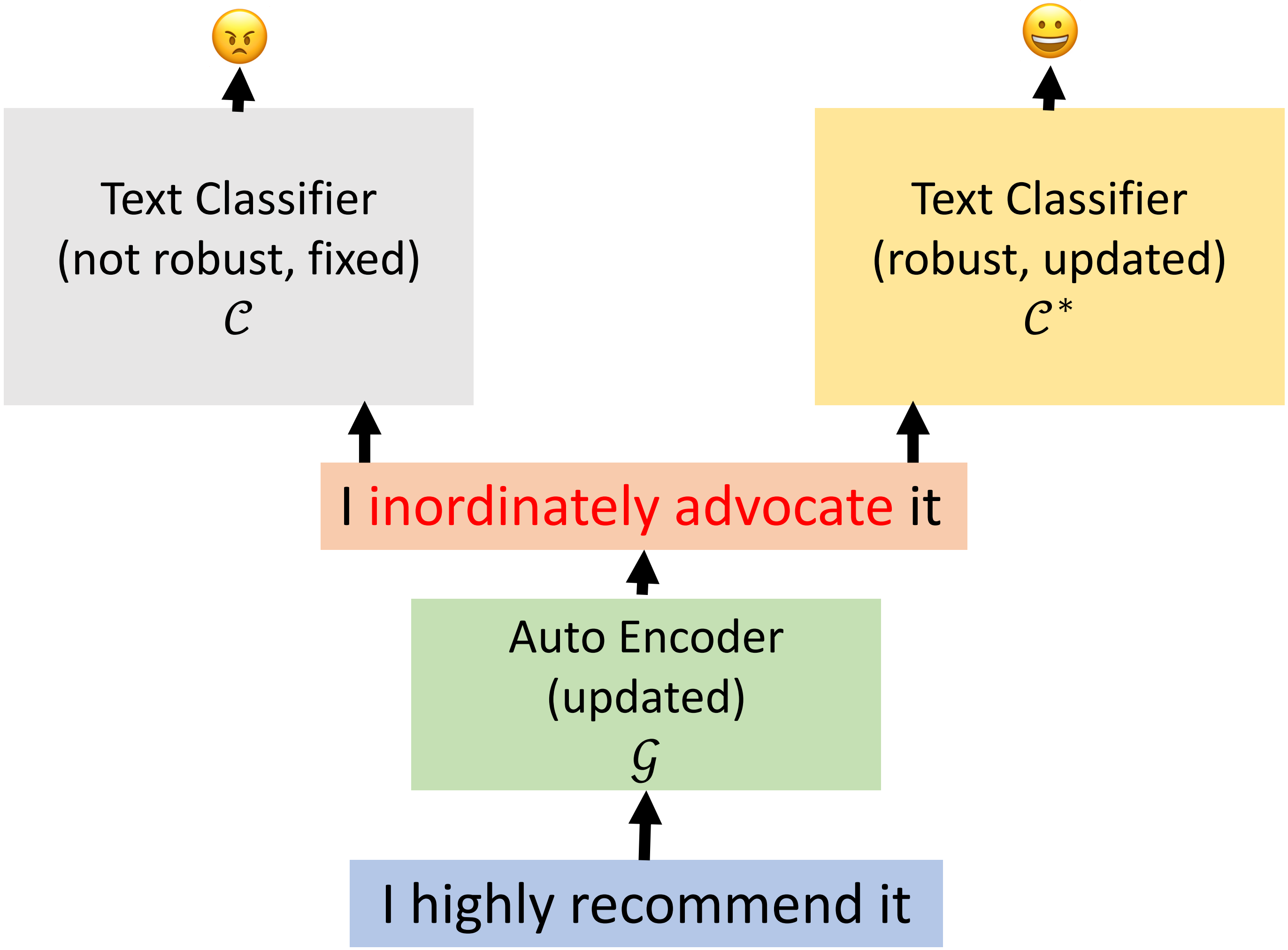
Wallace, Eric, et al. "Universal Adversarial Triggers for Attacking and Analyzing NLP." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019.

Outline

- Introduction
- Evasion Attacks and Defenses
  - Introduction
  - Four Ingredients in Evasion Attacks
  - Examples of Evasion Attacks
    - Crafting Adversaries by Auto-Encoder
  - Defenses against Evasion Attacks
- Imitation Attacks and Defenses
- Backdoor Attacks and Defenses
- Summary

Evasion Attacks: Crafting Adversaries by Auto-Encoder

- Train a generator (auto-encoder) to generate the adversarial samples
  - Goal of generator: make the text classifier predict wrongly
  - Goal of the classifier: predict correctly
  - Iterate between attack and defense



Xu, Ying, et al. "Grey-box Adversarial Attack And Defence For Sentiment Classification." *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021.

Evasion Attacks: Crafting Adversaries by Auto-Encoder

- Attack step
  - Reconstruction

$$L_{s2s} = -\log p_{\mathcal{G}}(x|x, \theta_{\mathcal{G}})$$

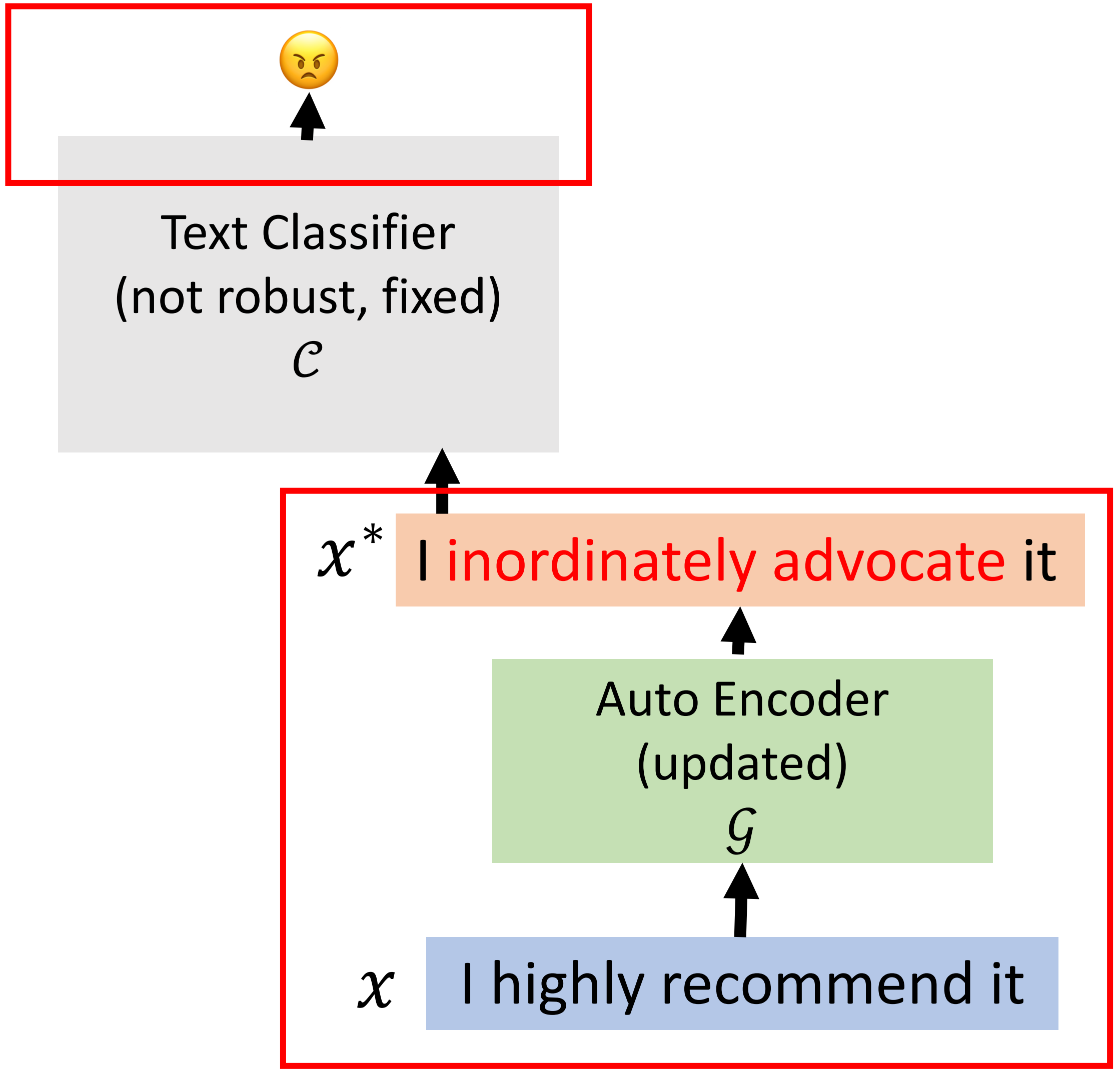
  - Similarity

$$L_{sem} = cos\left(\frac{1}{n}\sum_{i=0}^{n} emb(x_i), \frac{1}{n}\sum_{i=0}^{n} emb(x_i^*)\right)$$

Preserve the original semantic

  - Adversarial loss

$$L_{adv} = \log p_{\mathcal{C}}(y|x, \theta_{\mathcal{C}}, \theta_{\mathcal{G}})$$



Xu, Ying, et al. "Grey-box Adversarial Attack And Defence For Sentiment Classification." *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021.

Evasion Attacks: Crafting Adversaries by Auto-Encoder

- Defense step
  - Reconstruction

$$L_{s2s} = -\log p_{\mathcal{G}}(x|x, \theta_{\mathcal{G}})$$

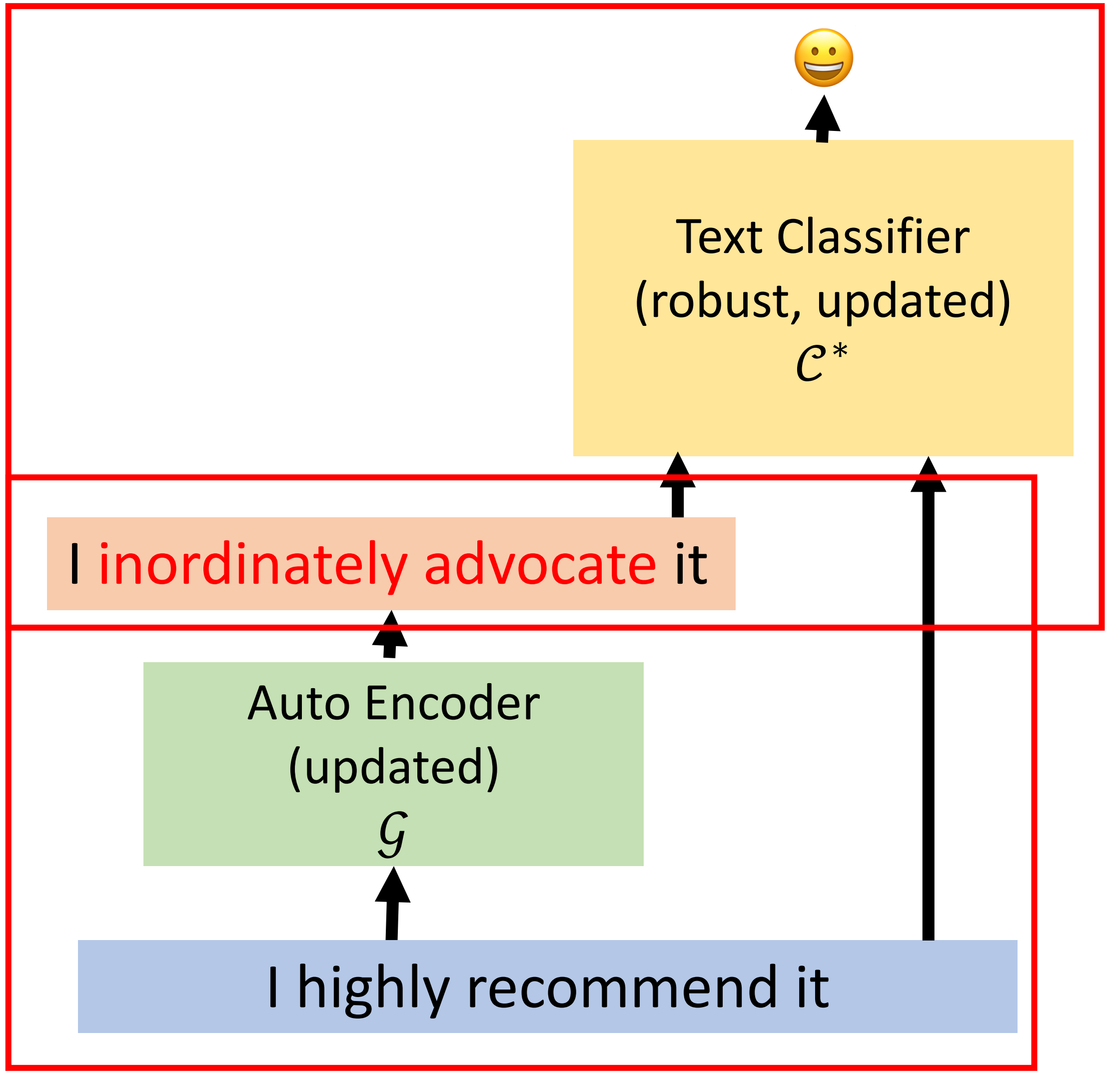
Preserve the original semantic

  - Similarity

$$L_{sem} = cos\left(\frac{1}{n}\sum_{i=0}^{n} emb(x_i), \frac{1}{n}\sum_{i=0}^{n} emb(x_i^*)\right)$$

  - Defense loss

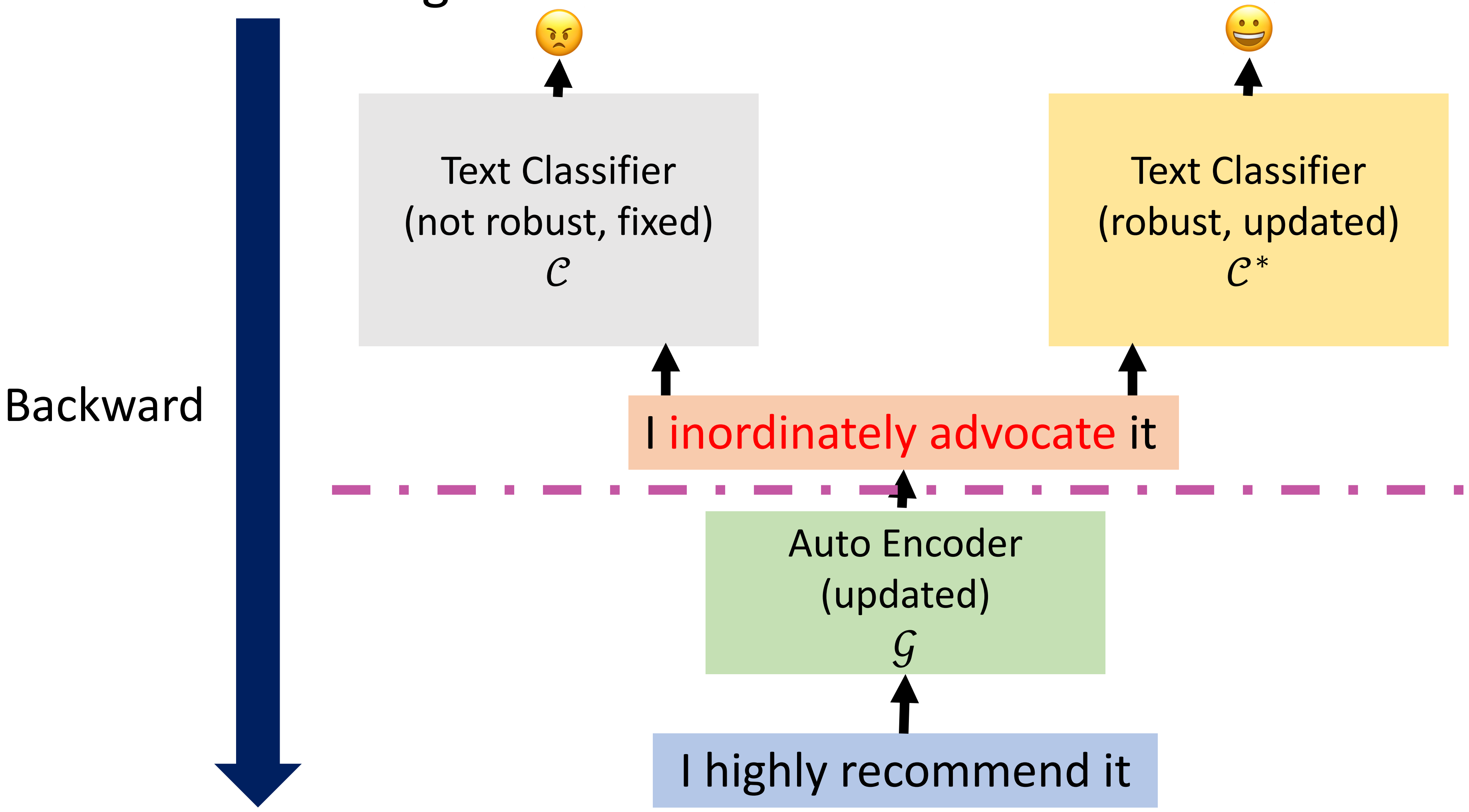
$$L_{def} = -\log p_{\mathcal{C}^*}([y, y]||[x, x^*], \theta_{\mathcal{C}^*}, \theta_{\mathcal{G}})$$



Xu, Ying, et al. "Grey-box Adversarial Attack And Defence For Sentiment Classification." *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021.

Evasion Attacks: Crafting Adversaries by Auto-Encoder

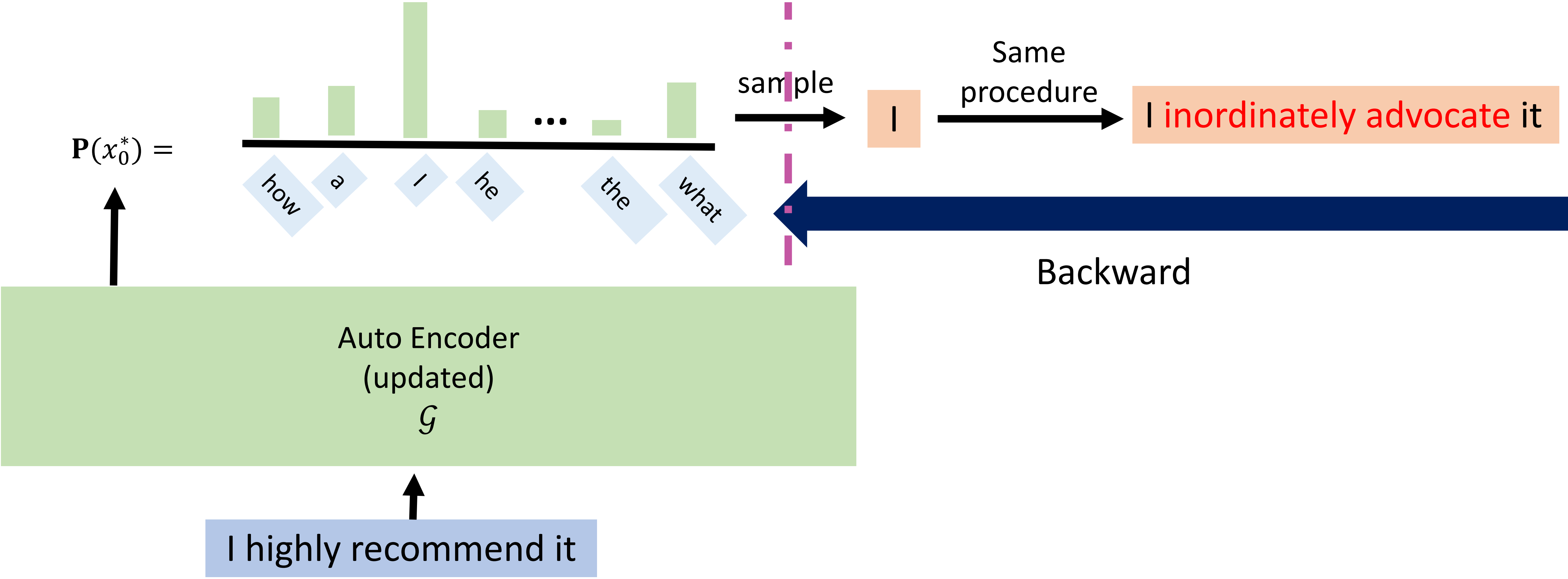
- Problem during backward



Xu, Ying, et al. "Grey-box Adversarial Attack And Defence For Sentiment Classification." *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021.

Evasion Attacks: Crafting Adversaries by Auto-Encoder

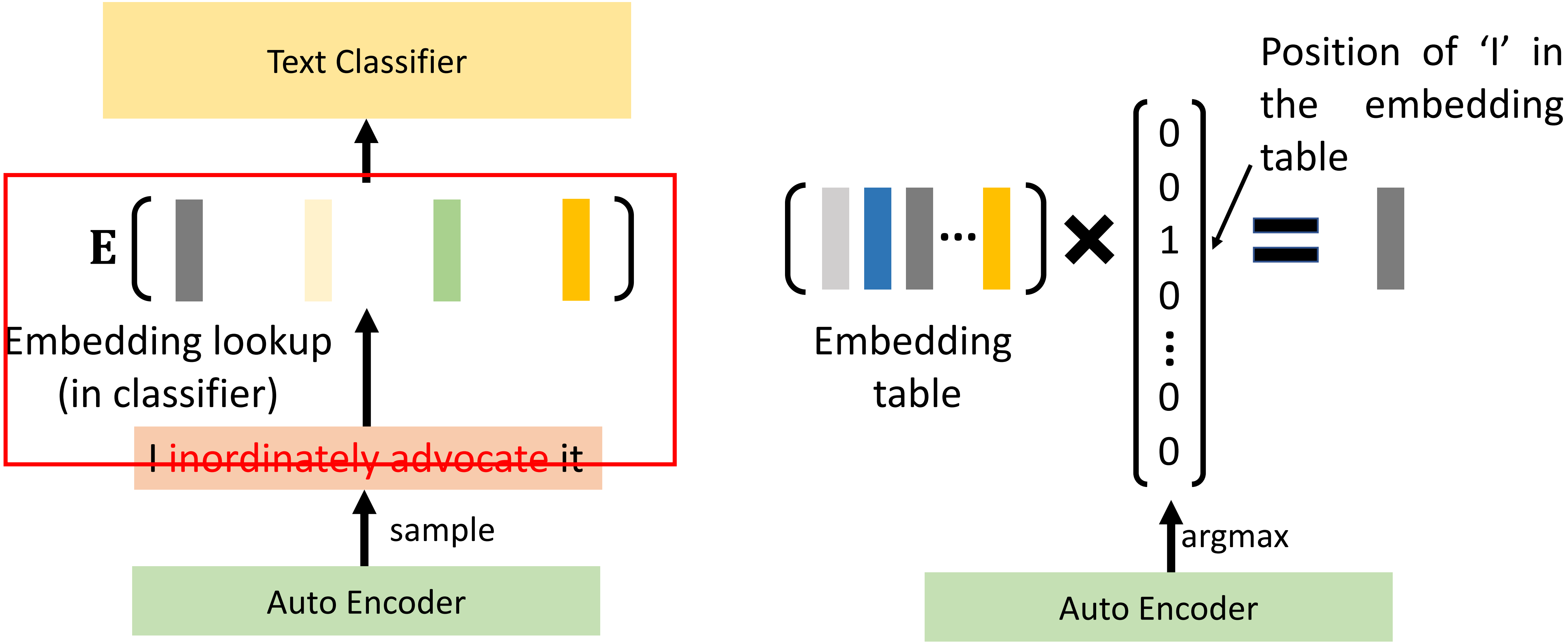
- Problem during backward: cannot directly backward the argmax in AE



Xu, Ying, et al. "Grey-box Adversarial Attack And Defence For Sentiment Classification." *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021.

Evasion Attacks: Crafting Adversaries by Auto-Encoder

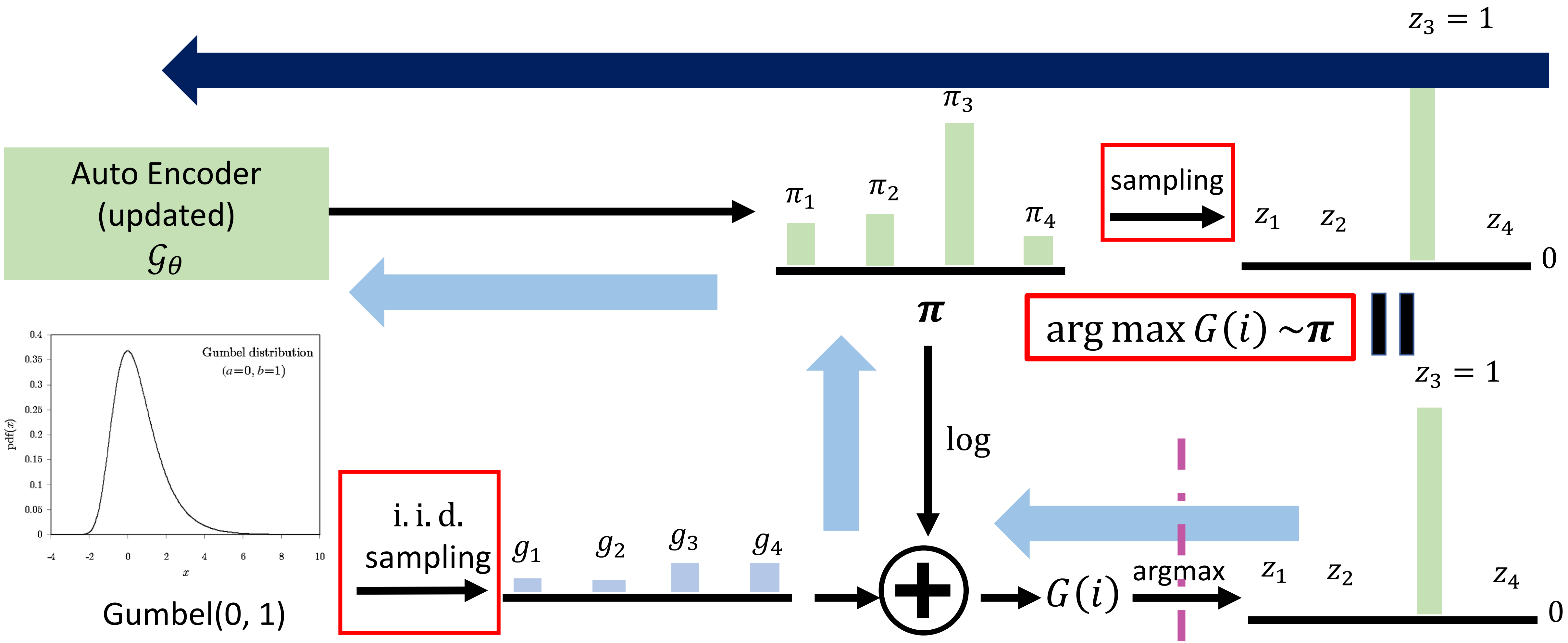
- A closer look into non-differentiability of the AE output



Xu, Ying, et al. "Grey-box Adversarial Attack And Defence For Sentiment Classification." *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021.

Evasion Attacks: Crafting Adversaries by Auto-Encoder

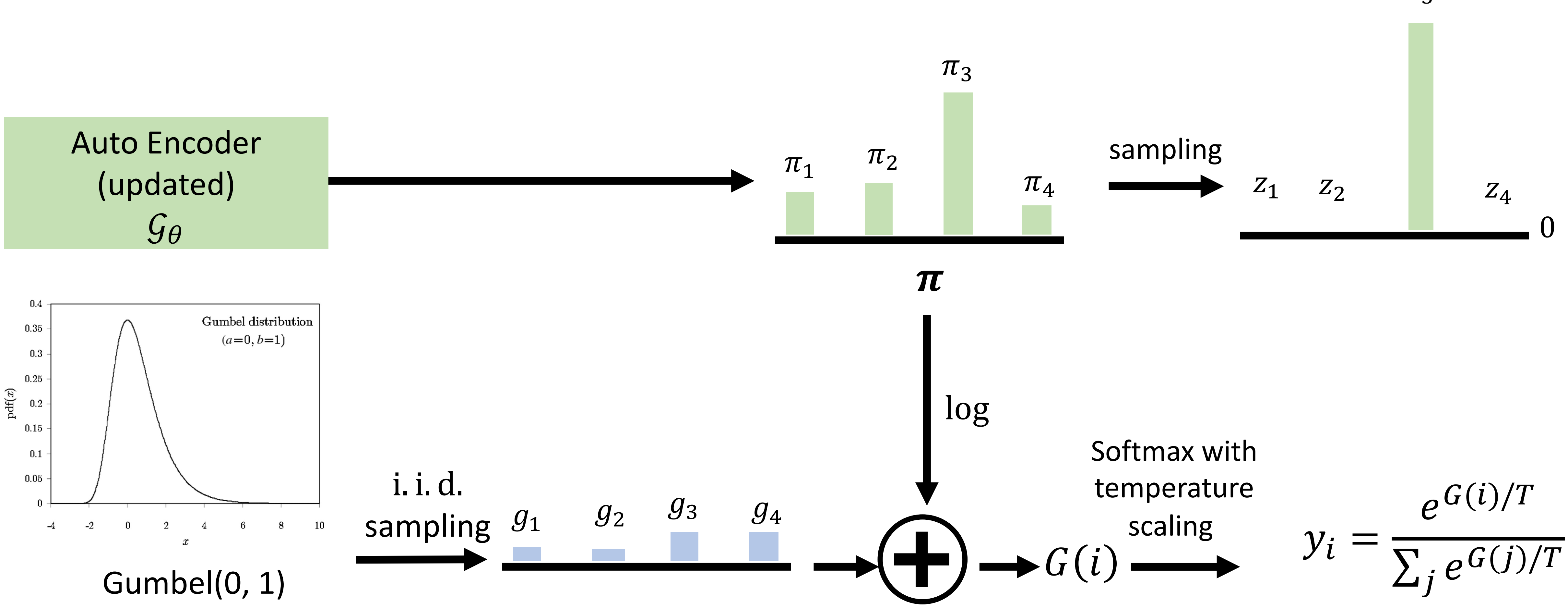
- Gumbel-Softmax reparametrization trick



Jang, Eric, Shixiang Gu, and Ben Poole. "Categorical reparameterization with gumbel-softmax." *arXiv preprint arXiv:1611.01144* (2016).

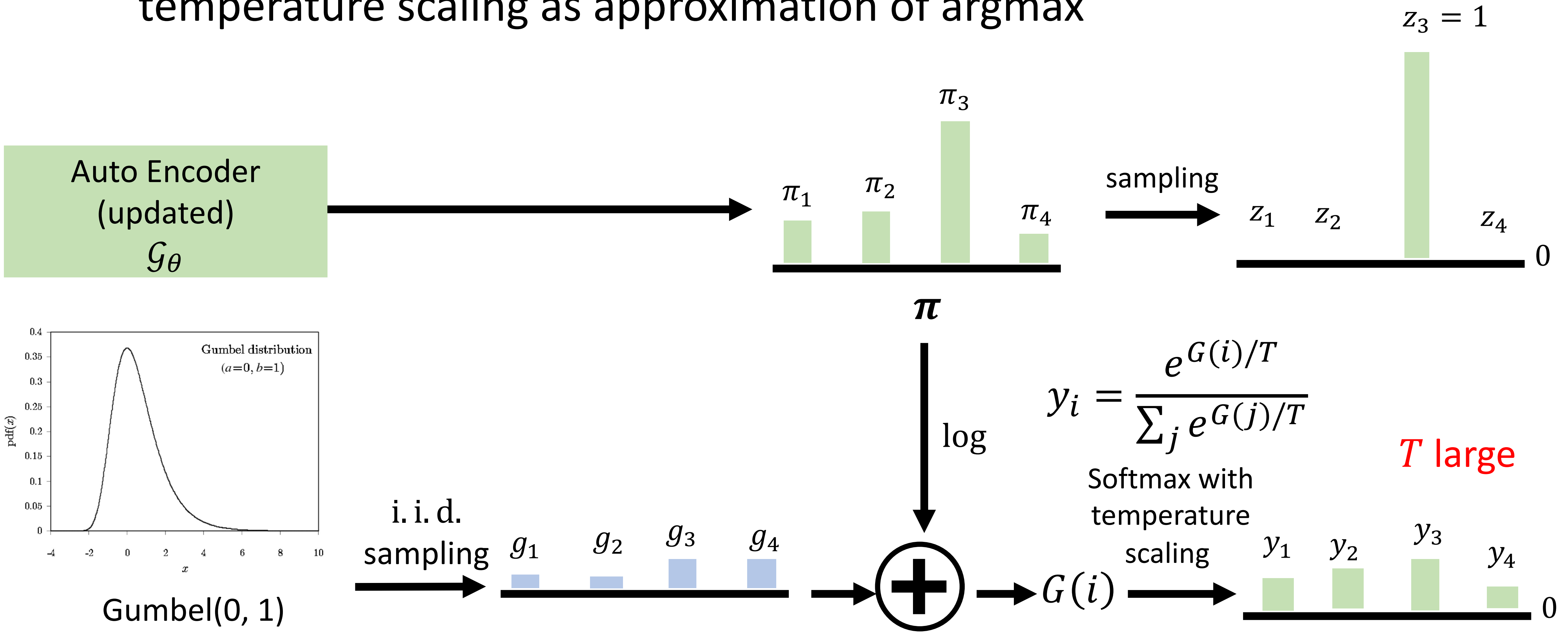
Evasion Attacks: Crafting Adversaries by Auto-Encoder

- Gumbel-Softmax reparametrization trick: using softmax with temperature scaling as approximation of argmax



Evasion Attacks: Crafting Adversaries by Auto-Encoder

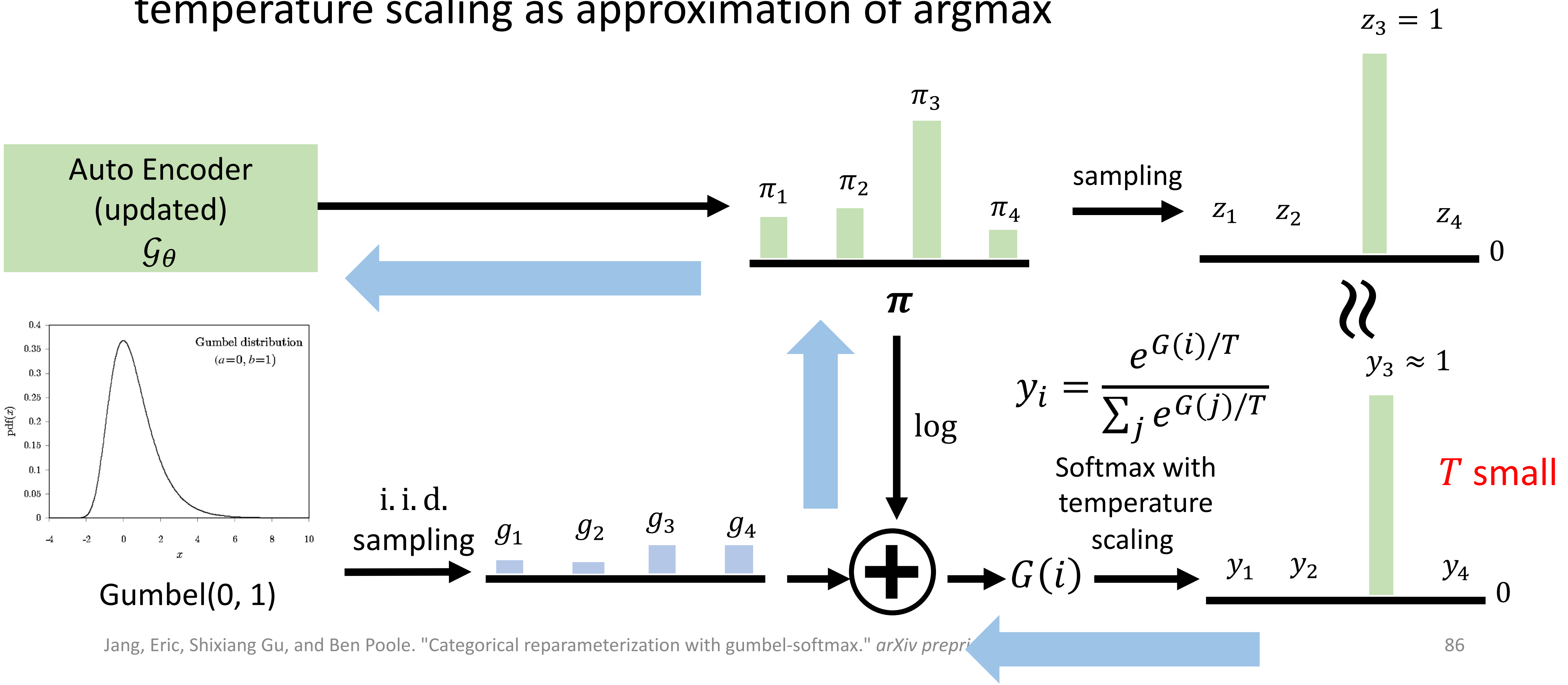
- Gumbel-Softmax reparametrization trick: using softmax with temperature scaling as approximation of argmax



Jang, Eric, Shixiang Gu, and Ben Poole. "Categorical reparameterization with gumbel-softmax." *arXiv preprint arXiv:1611.01144* (2016).

Evasion Attacks: Crafting Adversaries by Auto-Encoder

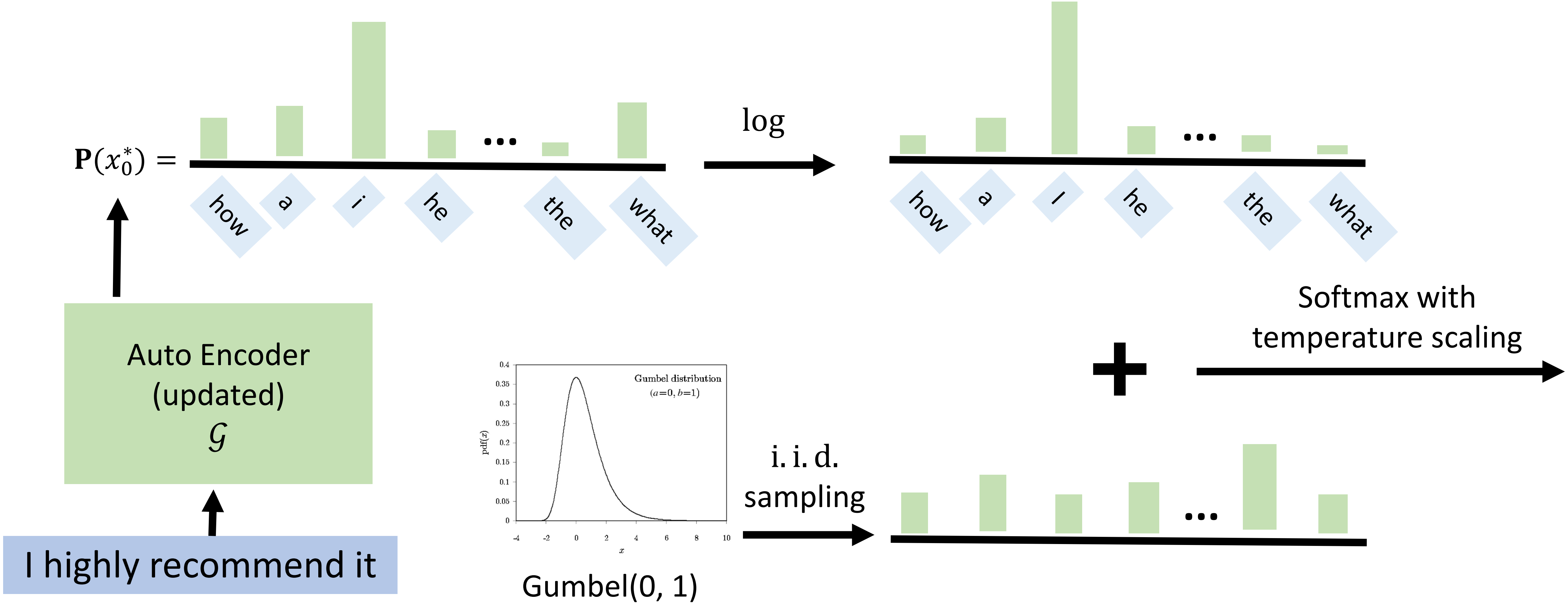
- Gumbel-Softmax reparametrization trick: using softmax with temperature scaling as approximation of argmax



Jang, Eric, Shixiang Gu, and Ben Poole. "Categorical reparameterization with gumbel-softmax." *arXiv prepri*

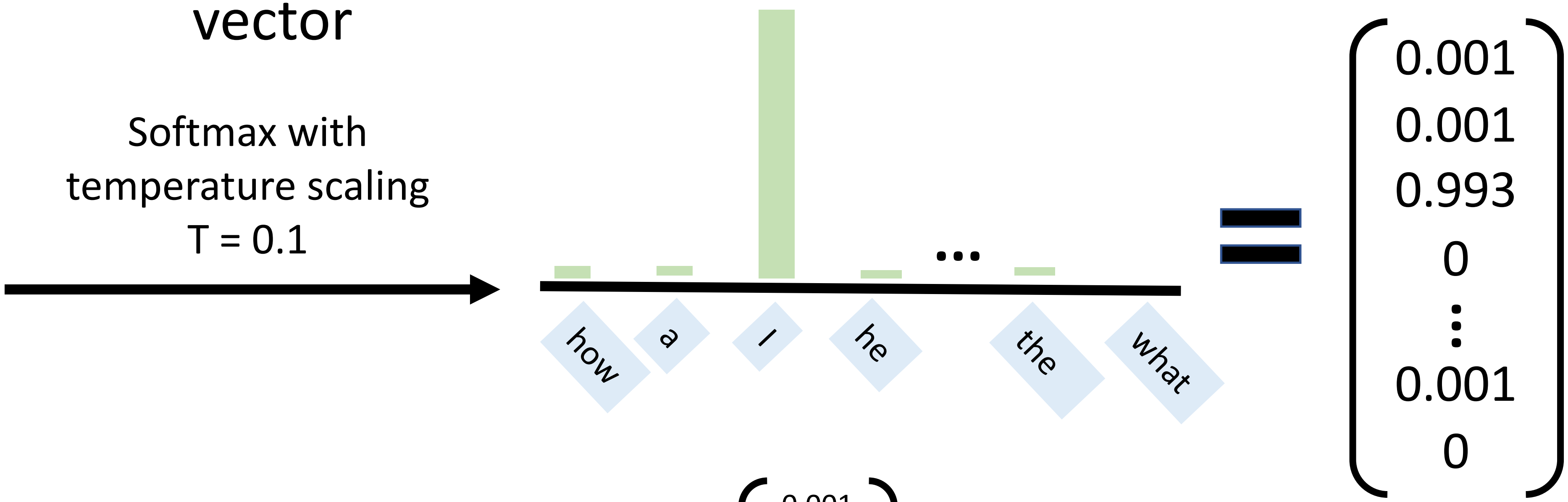
Evasion Attacks: Crafting Adversaries by Auto-Encoder

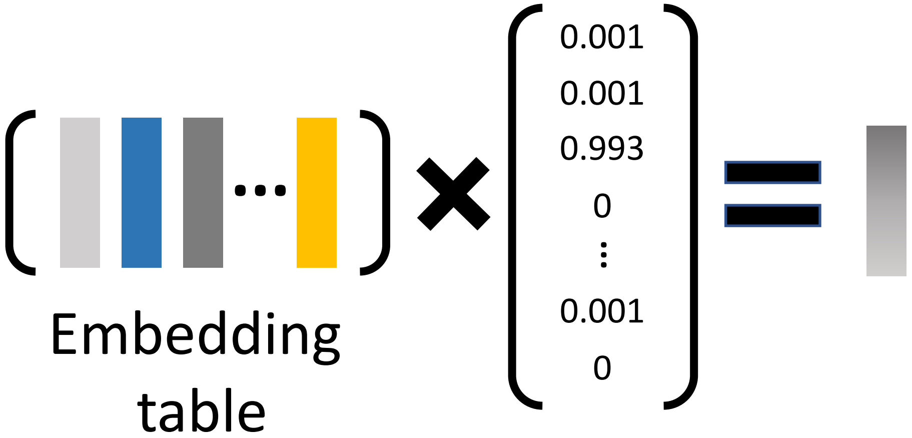
- A solution: gumbel-softmax



Evasion Attacks: Crafting Adversaries by Auto-Encoder

- Use the gumbel-softmax distribution to approximate the one-hot vector

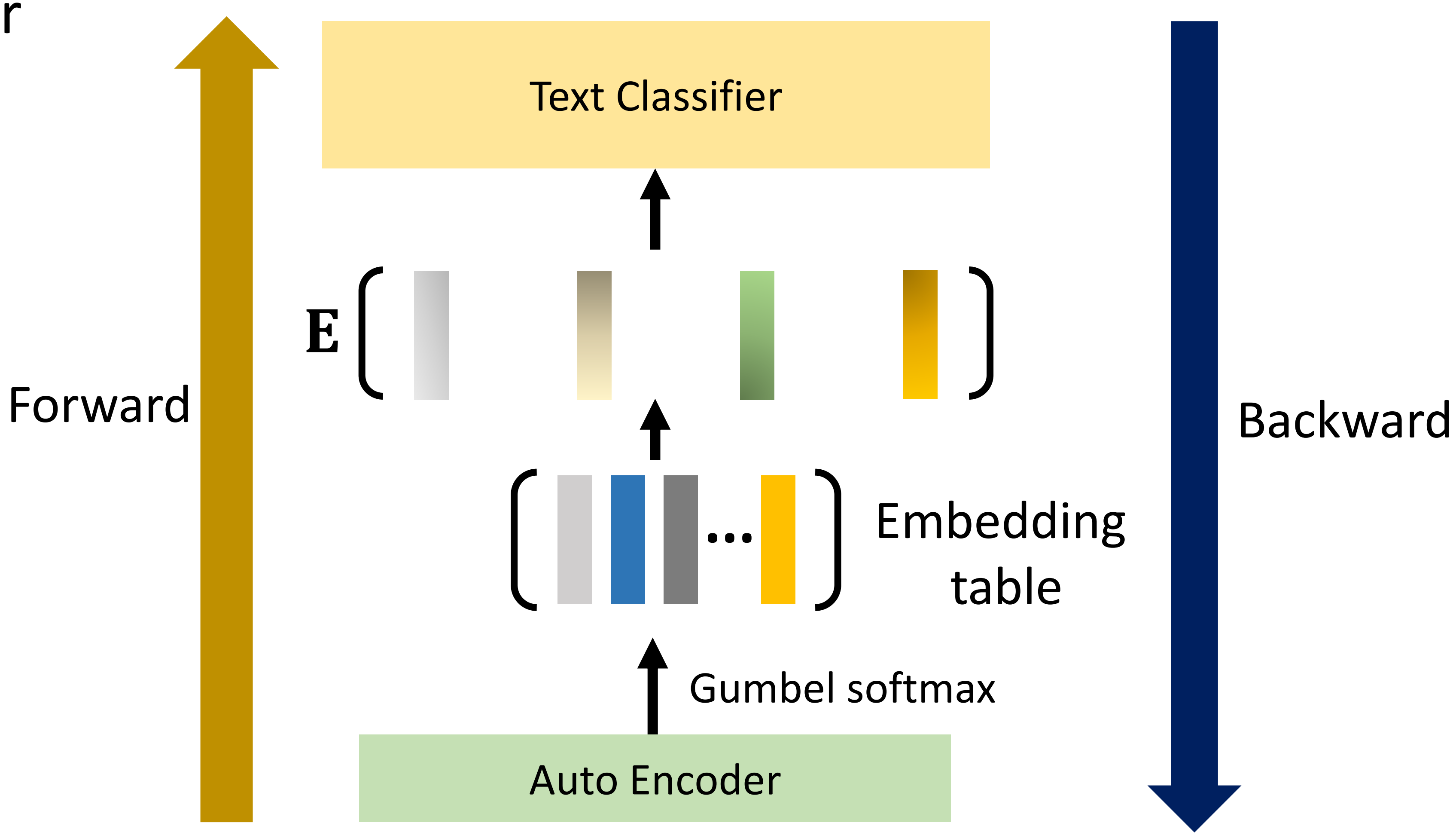




Xu, Ying, et al. "Grey-box Adversarial Attack And Defence For Sentiment Classification." *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021.

Evasion Attacks: Crafting Adversaries by Auto-Encoder

- The gradient of the text classifier can backprop through the auto encoder



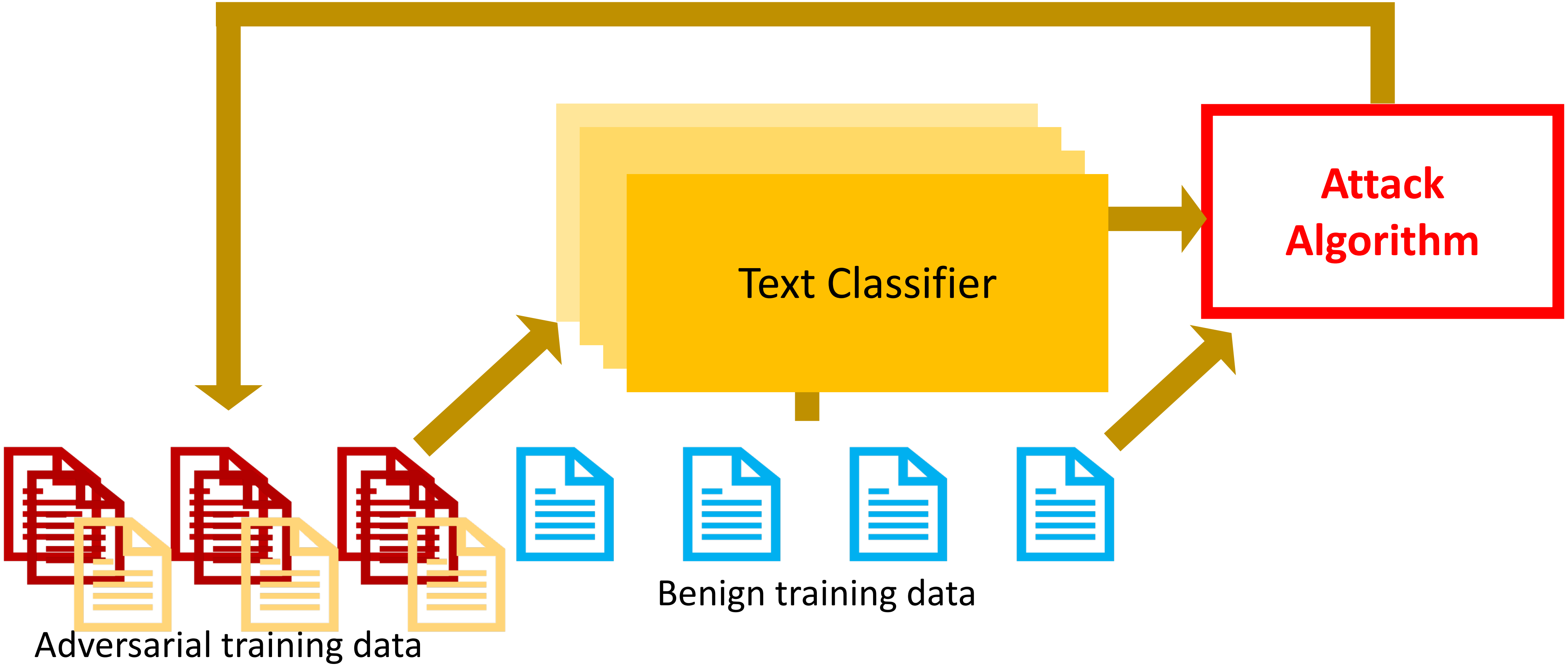
Xu, Ying, et al. "Grey-box Adversarial Attack And Defence For Sentiment Classification." *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021.

Outline

- Introduction
- Evasion Attacks and Defenses
  - Introduction
  - Four Ingredients in Evasion Attacks
  - Examples of Evasion Attacks
  - Defenses against Evasion Attacks
    - Training a More Robust Model
    - Detecting Adversaries during Inference
- Imitation Attacks and Defenses
- Backdoor Attacks and Defenses
- Summary

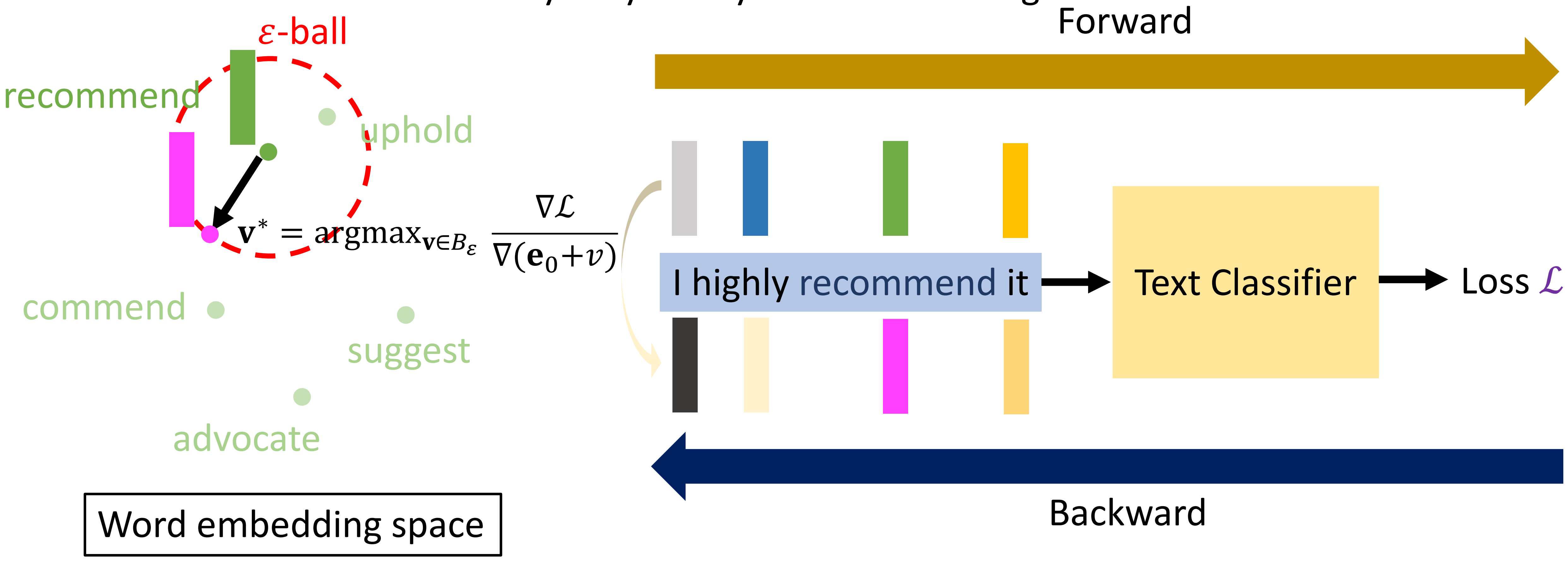
Evasion Attacks: Defense

- Adversarial training: generate the adversarial samples using the current model every N epochs



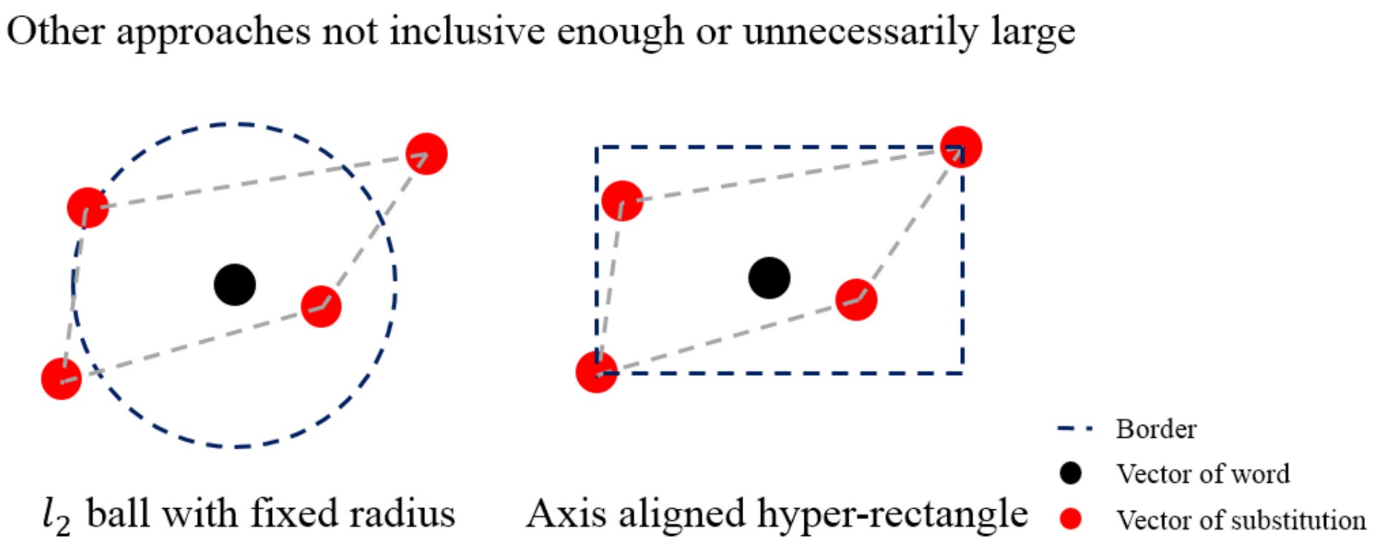
Evasion Attacks: Defense

- Adversarial training in the word embedding space by ε-ball
  - Motivation: A word's synonym may be within its neighborhood



Evasion Attacks: Defense

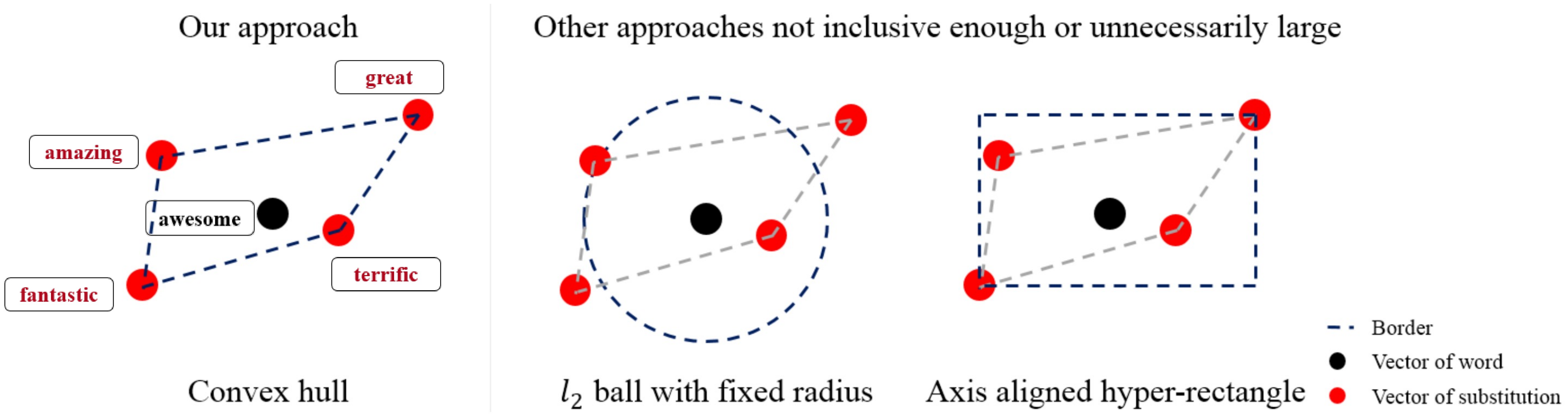
- ASCC-defense (Adversarial Sparse Convex Combination)
  - Convex hull of set A: the smallest convex set containing A



Dong, Xinshuai, et al. "Towards Robustness Against Natural Language Word Substitutions." *International Conference on Learning Representations*. 2020.

Evasion Attacks: Defense

- ASCC-defense (Adversarial Sparse Convex Combination): Adversarial training in the word embedding space by the convex hull form by the synonym set



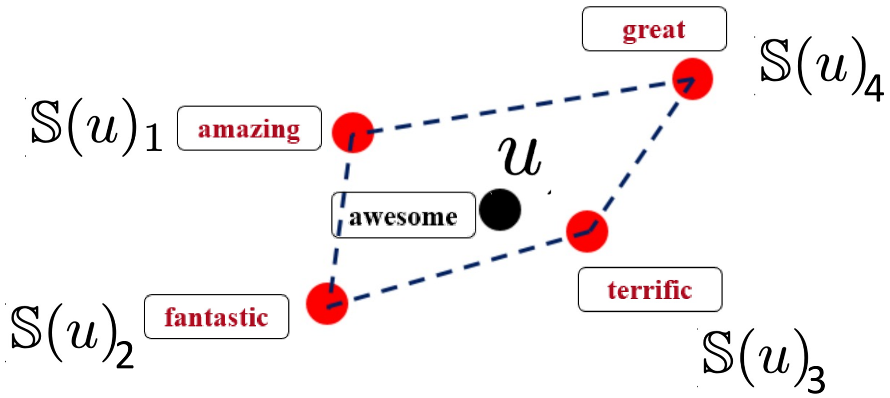
Dong, Xinshuai, et al. "Towards Robustness Against Natural Language Word Substitutions." *International Conference on Learning Representations*. 2020.

Evasion Attacks: Defense

- ASCC-defense (Adversarial Sparse Convex Combination)
  - The convex hull of a set A can be represented by the the linear combination of the elements in set A

Proposition 1. *Let* $\mathbb{S}(u) = \{\mathbb{S}(u)_1, ..., \mathbb{S}(u)_T\}$ *be the set of all substitutions of word* u, $\mathrm{conv}\mathbb{S}(u)$ *be the convex hull of word vectors of all elements in* $\mathbb{S}(u)$, *and* $v(\cdot)$ *be the word vector function. Then, we have* $\mathrm{conv}\mathbb{S}(u) = \{\sum_{i=1}^{T} w_i v(\mathbb{S}(u)_i) \mid \sum_{i=1}^{T} w_i = 1, w_i \geq 0\}$.

Our approach

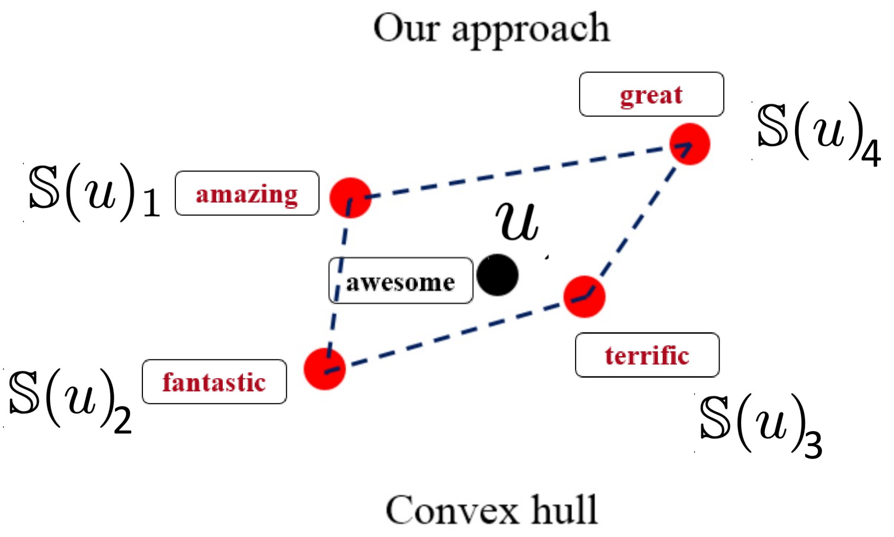


Convex hull

Dong, Xinshuai, et al. "Towards Robustness Against Natural Language Word Substitutions." *International Conference on Learning Representations*. 2020.

Evasion Attacks: Defense

- ASCC-defense (Adversarial Sparse Convex Combination)
  - Finding an adversary embedding in the convex hull is just finding the coefficient of the linear combination



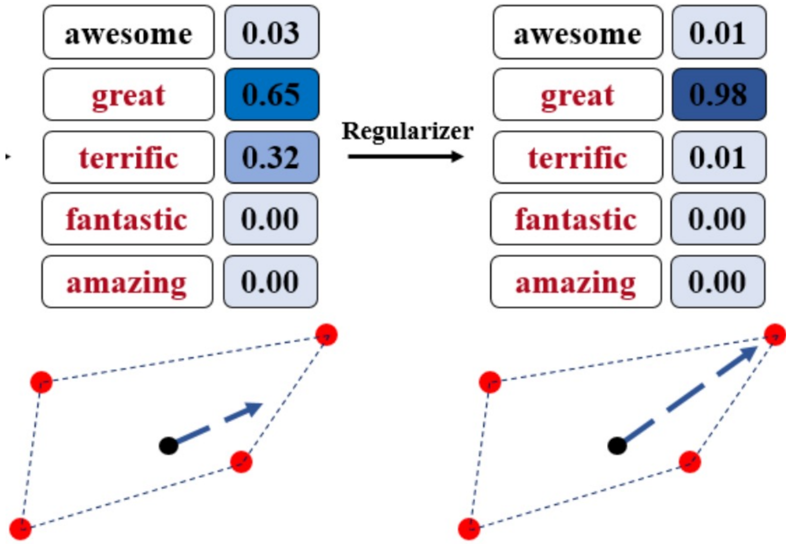
$$\hat{v}(x_i) = \sum_{j=1}^{T} w_{ij} v(\mathbb{S}(x_i)_j), \ \text{s.t.} \sum_{j=1}^{T} w_{ij} = 1, w_{ij} \geq 0.$$

$$\max_{\hat{w}} -\log p(y|\hat{v}(x))$$

Dong, Xinshuai, et al. "Towards Robustness Against Natural Language Word Substitutions." *International Conference on Learning Representations*. 2020.

Evasion Attacks: Defense

- ASCC-defense (Adversarial Sparse Convex Combination)
  - Making the coefficient of the linear combination sparser



$$\max_{\hat{w}} -\log p(y|\hat{v}(x)) \boxed{- \alpha \sum_{i=1}^{L} \frac{1}{L} \mathcal{H}(w_i)}$$

$$\mathcal{H}(w_i) = \sum_{j=1}^{T} -w_{ij} \log(w_{ij})$$

Dong, Xinshuai, et al. "Towards Robustness Against Natural Language Word Substitutions." *International Conference on Learning Representations*. 2020.

Evasion Attacks: Defense

- Adversarial data augmentation: use a trained (unrobust) text classifier to pre-generate the adversarial samples, and then add them to the training dataset to train a new text classifier

Evasion Attacks: Defense

- Adversarial and Mixup Data Augmentation
  - Adversarial data augmentation
  - Mixup the samples in the training set (including benign and adversarial)

$$\hat{\mathbf{x}} = \boxed{\lambda \mathbf{x}_i + (1 - \lambda)\mathbf{x}_j}$$

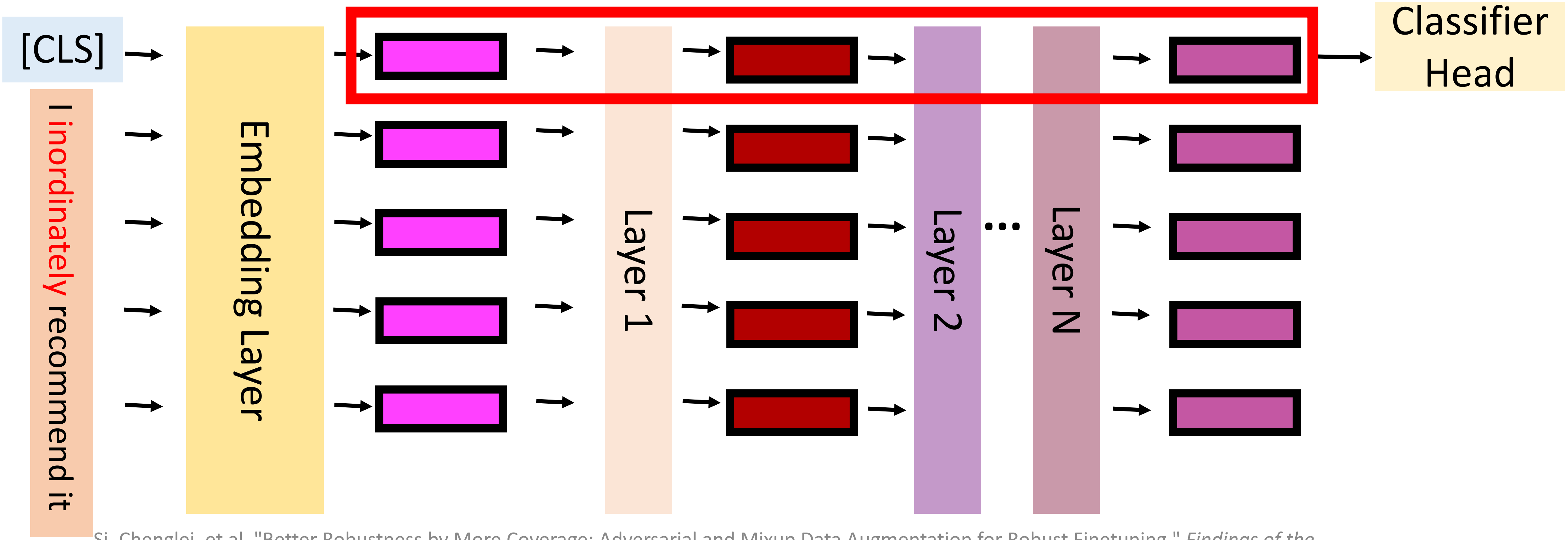
$$\hat{\mathbf{y}} = \lambda \mathbf{y}_i + (1 - \lambda)\mathbf{y}_j$$

😠 😠

Si, Chenglei, et al. "Better Robustness by More Coverage: Adversarial and Mixup Data Augmentation for Robust Finetuning." *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2021.

Evasion Attacks: Defense

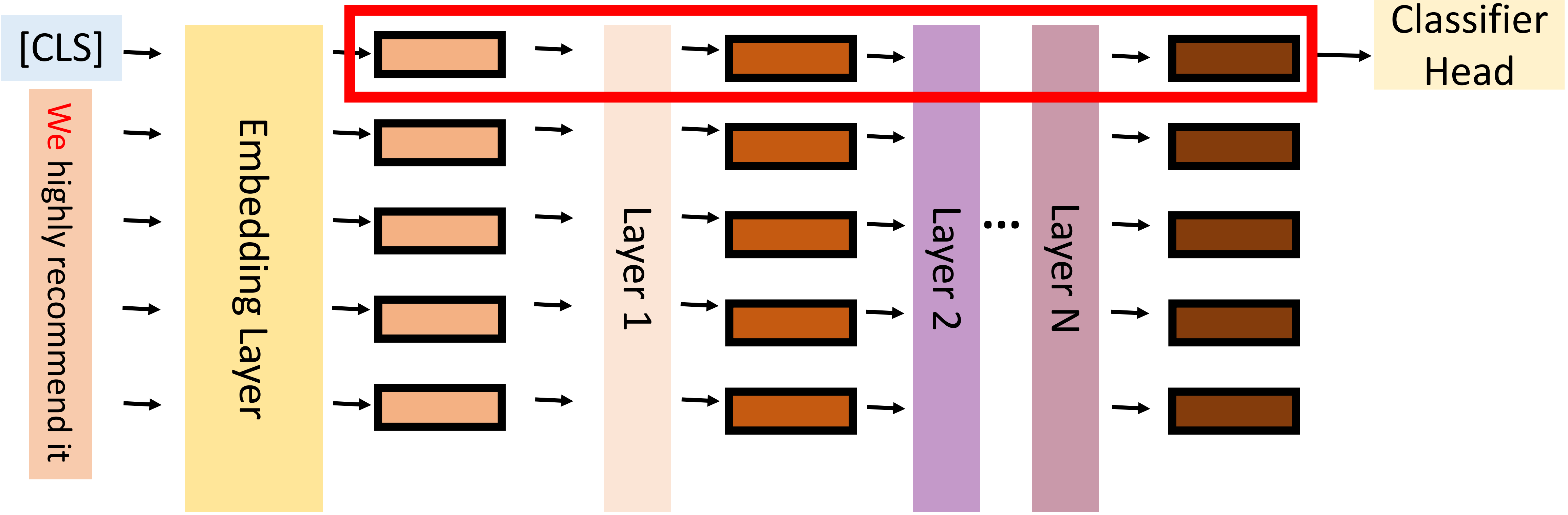
- Adversarial and Mixup Data Augmentation
  - Adversarial data augmentation
  - Mixup the samples in the training set (including benign and adversarial)



Si, Chenglei, et al. "Better Robustness by More Coverage: Adversarial and Mixup Data Augmentation for Robust Finetuning." *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2021.

Evasion Attacks: Defense

- Adversarial and Mixup Data Augmentation
  - Adversarial data augmentation
  - Mixup the samples in the training set (including benign and adversarial)



Si, Chenglei, et al. "Better Robustness by More Coverage: Adversarial and Mixup Data Augmentation for Robust Finetuning." *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2021.

Evasion Attacks: Defense

- Adversarial and Mixup Data Augmentation
  - Adversarial data augmentation
  - Mixup the samples in the training set (including benign and adversarial)

$$\hat{\mathbf{x}} = \lambda \mathbf{x}_i + (1 - \lambda)\mathbf{x}_j$$

$$\hat{\mathbf{y}} = \lambda \mathbf{y}_i + (1 - \lambda)\mathbf{y}_j$$

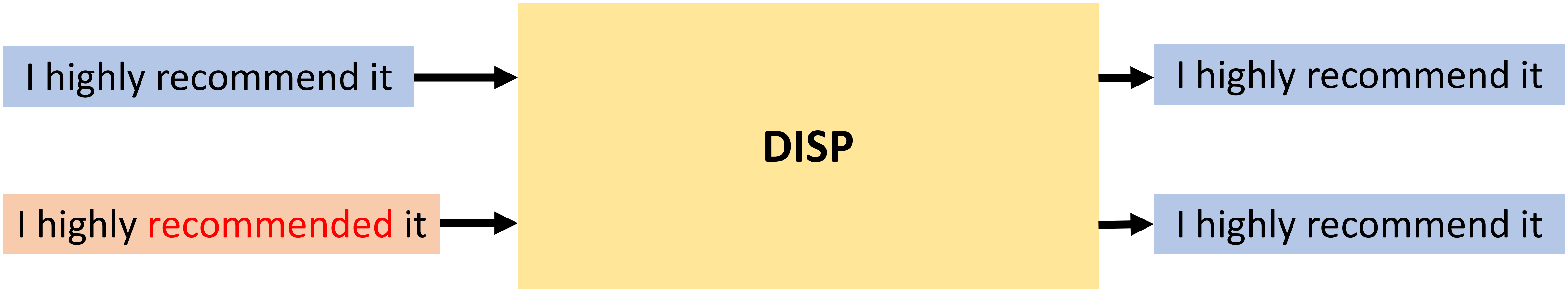
Si, Chenglei, et al. "Better Robustness by More Coverage: Adversarial and Mixup Data Augmentation for Robust Finetuning." *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2021.

Outline

- Introduction
- Evasion Attacks and Defenses
  - Introduction
  - Four Ingredients in Evasion Attacks
  - Examples of Evasion Attacks
  - Defenses against Evasion Attacks
    - Training a More Robust Model
    - Detecting Adversaries during Inference
- Imitation Attacks and Defenses
- Backdoor Attacks and Defenses
- Summary

Evasion Attacks: Detecting Adversaries

- **Dis**criminate **p**erturbations (DISP): detect adversarial samples and convert them to benign ones

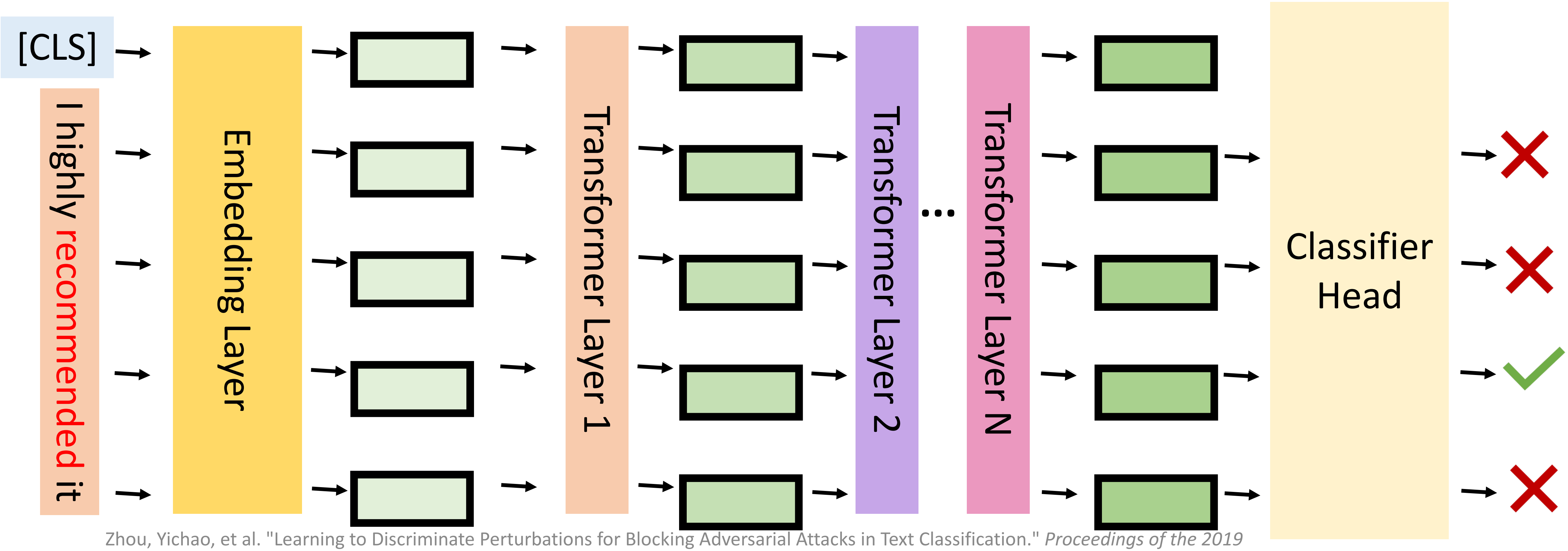


Zhou, Yichao, et al. "Learning to Discriminate Perturbations for Blocking Adversarial Attacks in Text Classification." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019.

Evasion Attacks: Detecting Adversaries

- **Dis**criminate **p**erturbations (DISP): DISP contains three submodules

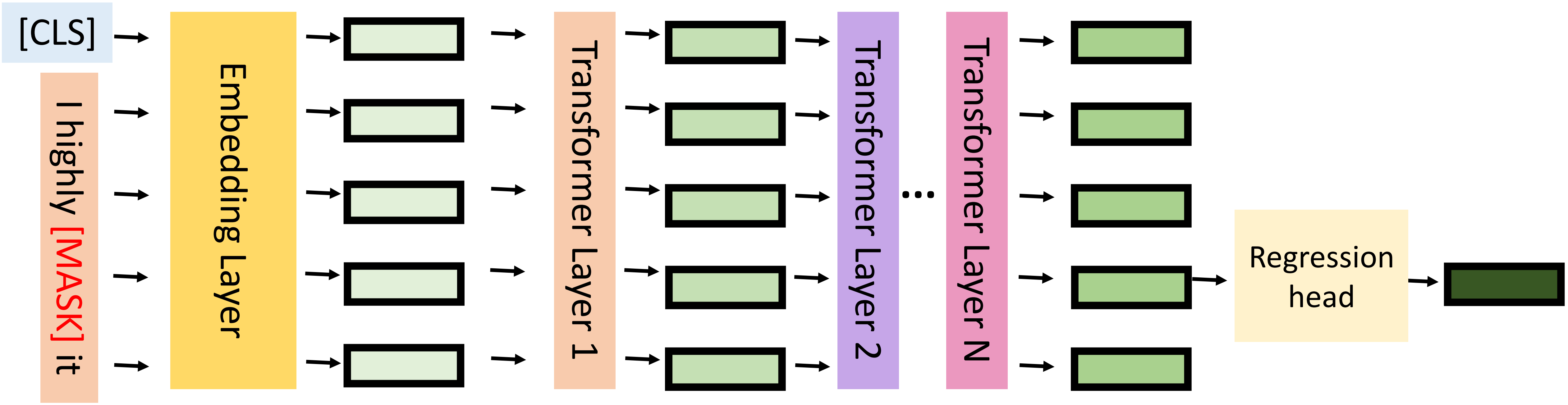
  1. Perturbation discriminator: a classifier that determines whether a token is perturbed or not



Zhou, Yichao, et al. "Learning to Discriminate Perturbations for Blocking Adversarial Attacks in Text Classification." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019.

Evasion Attacks: Detecting Adversaries

- **Dis**criminate **p**erturbations (DISP): DISP contains three submodules
  2. Embedding estimator: estimate the perturbed tokens' by regression

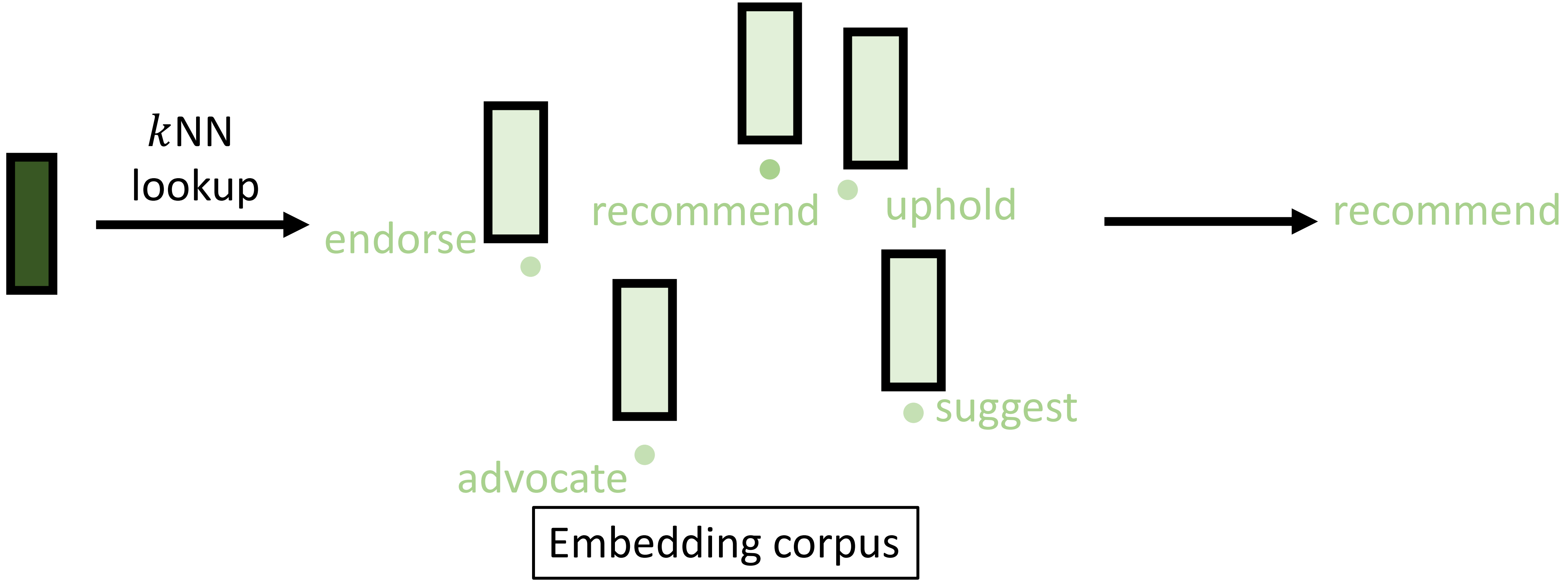


Zhou, Yichao, et al. "Learning to Discriminate Perturbations for Blocking Adversarial Attacks in Text Classification." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019.

Evasion Attacks: Detecting Adversaries

- **Dis**criminate **p**erturbations (DISP): DISP contains three submodules

  3. Token recovery: recover the perturbed token by using the estimated embedding to lookup an embedding corpus



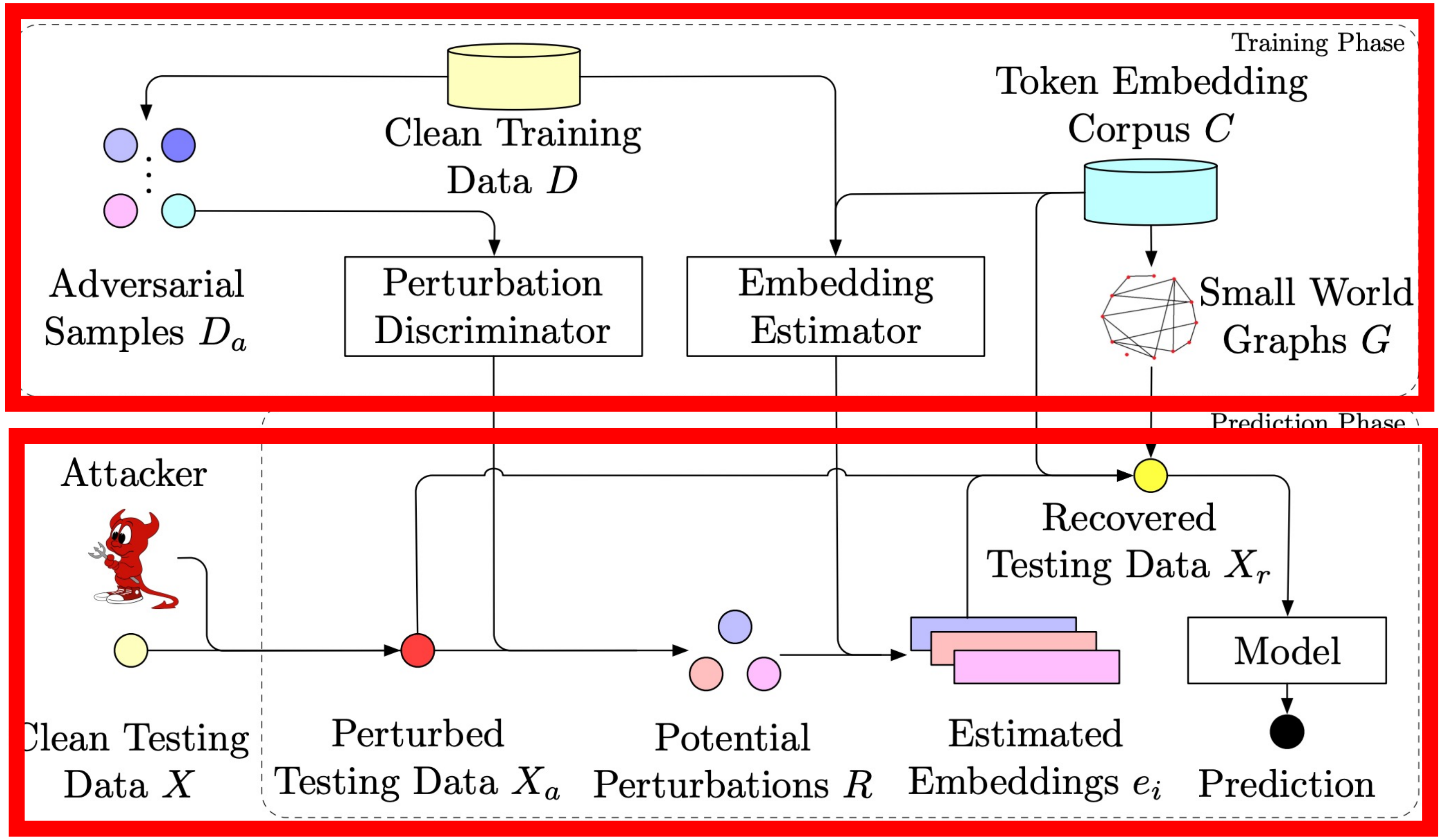
Zhou, Yichao, et al. "Learning to Discriminate Perturbations for Blocking Adversarial Attacks in Text Classification." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019.

Evasion Attacks: Detecting Adversaries

- **Dis**criminate **p**erturbations (DISP): Training and inference

Training

Inference



Zhou, Yichao, et al. "Learning to Discriminate Perturbations for Blocking Adversarial Attacks in Text Classification." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019.

Evasion Attacks: Detecting Adversaries

- **F**requency-**G**uided **W**ord **S**ubstitutions (FGWS)
  - Observation: Evasion attacks in NLP tend to swap high frequency words into low frequency ones

Attack	Original or perturbed sequence
None	A clever blend of fact and fiction
GENETIC	1.39 ←-- 5.55 A **brainy** [*clever*] blend of fact and fiction
PWWS	1.61 ←-- 5.55 0.00 ←-- 3.81 A **cunning** [*clever*] **blending** [*blend*] of 0.00 ←-- 4.39 fact and **fabrication** [*fiction*]

Figure 1: Corpus $\log_e$ frequencies of the replaced words (bold, italic, red) and their corresponding adversarial substitutions (bold, black) using the GENETIC (Alzantot et al., 2018) and PWWS (Ren et al., 2019) attacks on SST-2 (Socher et al., 2013).

Mozes, Maximilian, et al. "Frequency-Guided Word Substitutions for Detecting Textual Adversarial Examples." *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2021.

Evasion Attacks: Detecting Adversaries

- **F**requency-**G**uided **W**ord **S**ubstitutions (FGWS): Swap low frequency words with higher frequency counterparts with a three-stepped pipeline
  - Step 1: Find the words in the input whose occurrence in the training data is lower than a pre-defined threshold δ

Training data log occurrence

I	inordinately	recommend	it
13	0	7.1	19.1

$$\delta = 4$$

Mozes, Maximilian, et al. "Frequency-Guided Word Substitutions for Detecting Textual Adversarial Examples." *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2021.

Evasion Attacks: Detecting Adversaries

- **F**requency-**G**uided **W**ord **S**ubstitutions (FGWS): Swap low frequency words with higher frequency counterparts with a three-stepped pipeline
  - Step 2: Replace all low frequency words identified in step 1 with their most frequent synonyms

I inordinately recommend it

↓

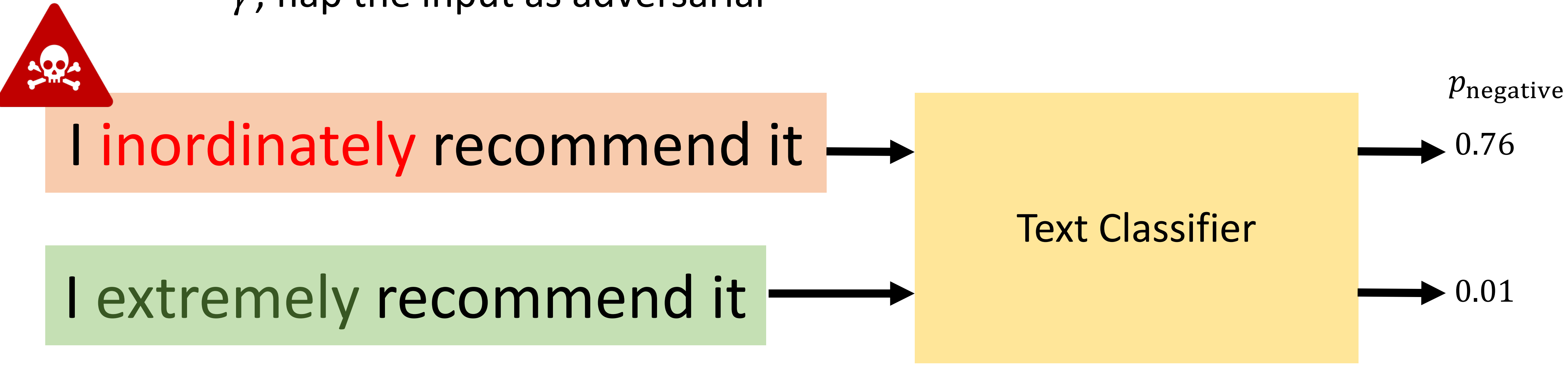
I extremely recommend it

Word	Synonym	Occurrence
inordinately	highly	7.4
	extremely	9.2
	strongly	8.2

Mozes, Maximilian, et al. "Frequency-Guided Word Substitutions for Detecting Textual Adversarial Examples." *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2021.

Evasion Attacks: Detecting Adversaries

- **F**requency-**G**uided **W**ord **S**ubstitutions (FGWS): Swap low frequency words with higher frequency counterparts with a three-stepped pipeline
  - Step 3: If the probability difference of the original predicted class between the original input and the swapped input is larger than a predefined threshold γ, flap the input as adversarial



$$\Delta\, p_{negative} = 0.76 - 0.01 = 0.75 \;>\; \gamma = 0.45$$

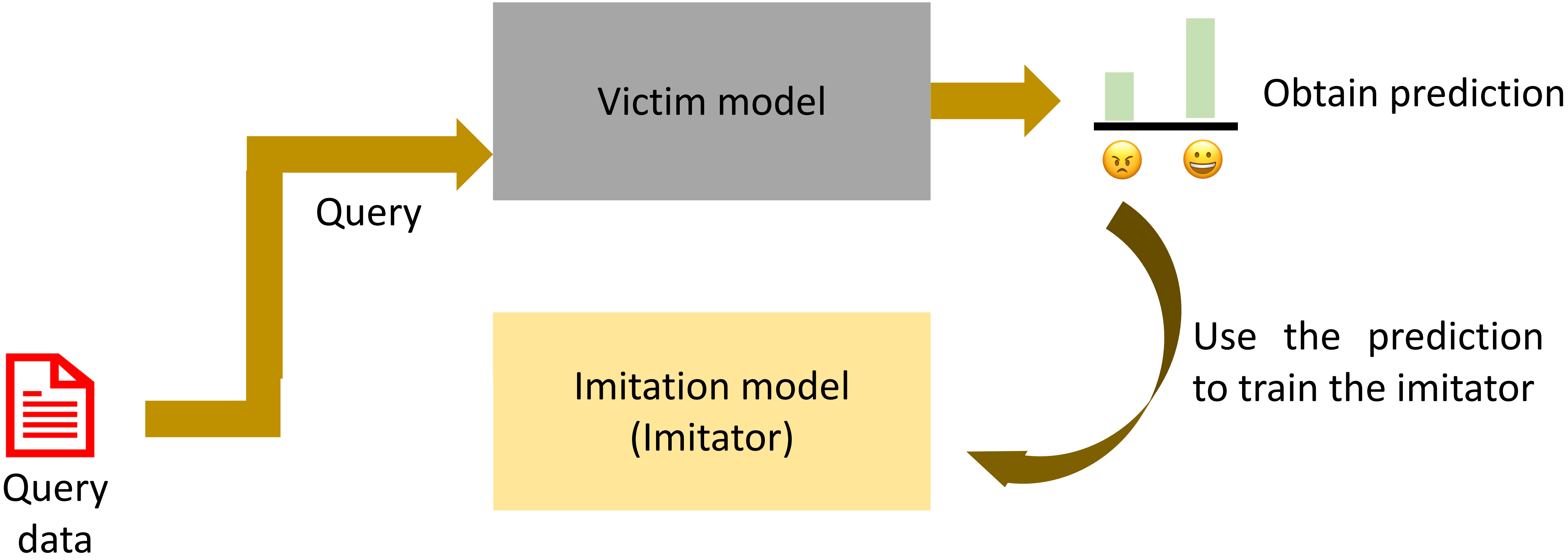
Mozes, Maximilian, et al. "Frequency-Guided Word Substitutions for Detecting Textual Adversarial Examples." *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2021.

Outline

- Introduction
- Evasion Attacks and Defenses
- Imitation Attacks and Defenses
  - Imitation Attacks
  - Adversarial Transferability
  - Defense against Imitation Attacks
- Backdoor Attacks and Defenses
- Summary

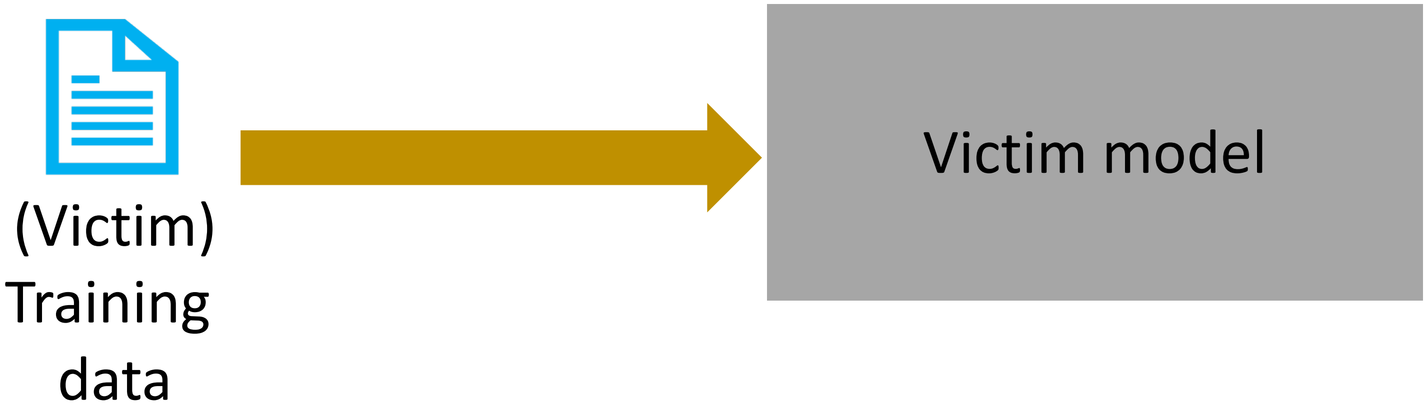
Imitations Attack

- What is imitation attack: Imitation attack aims to stole a trained model by querying it



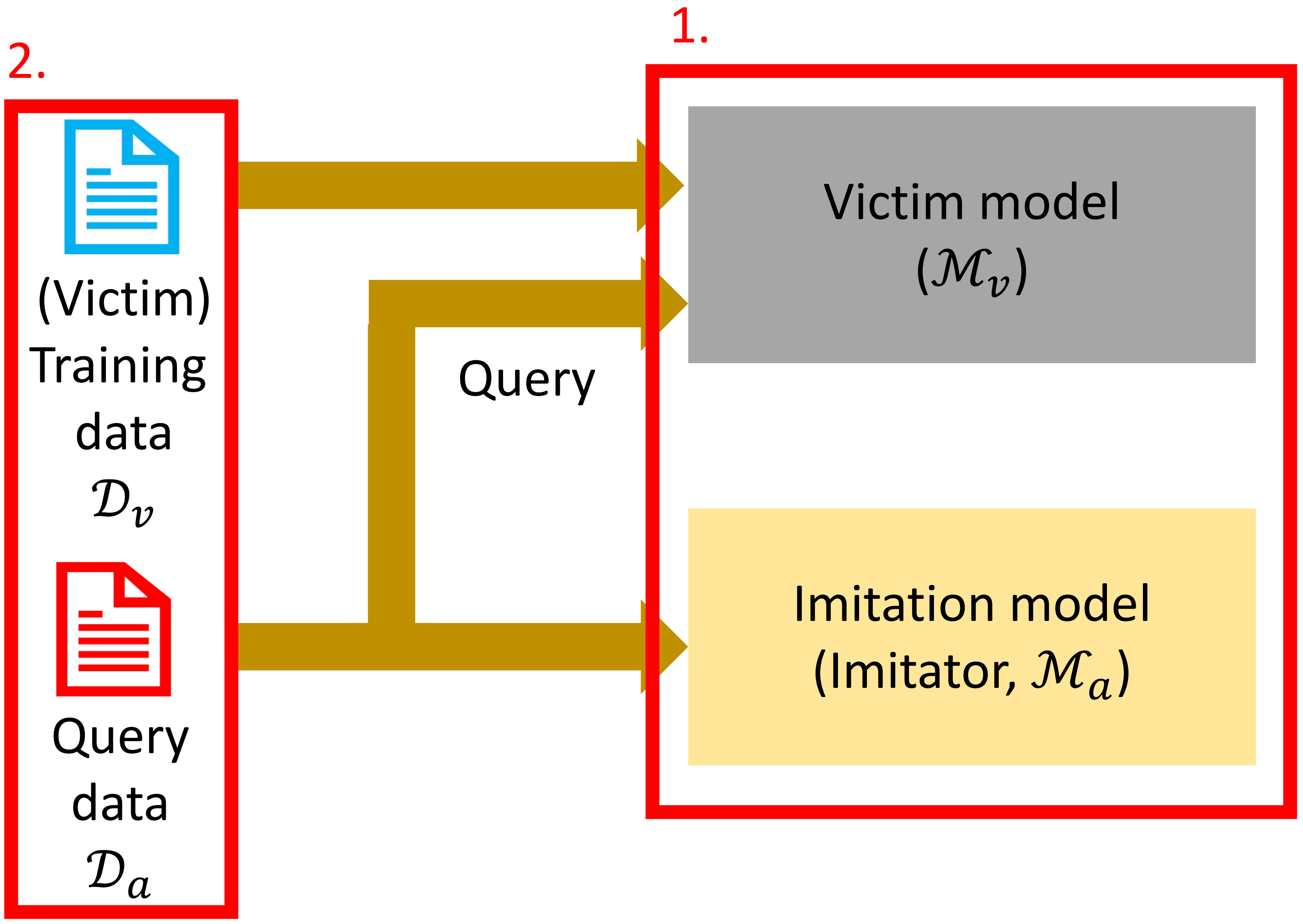
Imitations Attack

- Why imitation attack
  - Training a model requires significant resources, both time and money
  - Training data may be proprietary



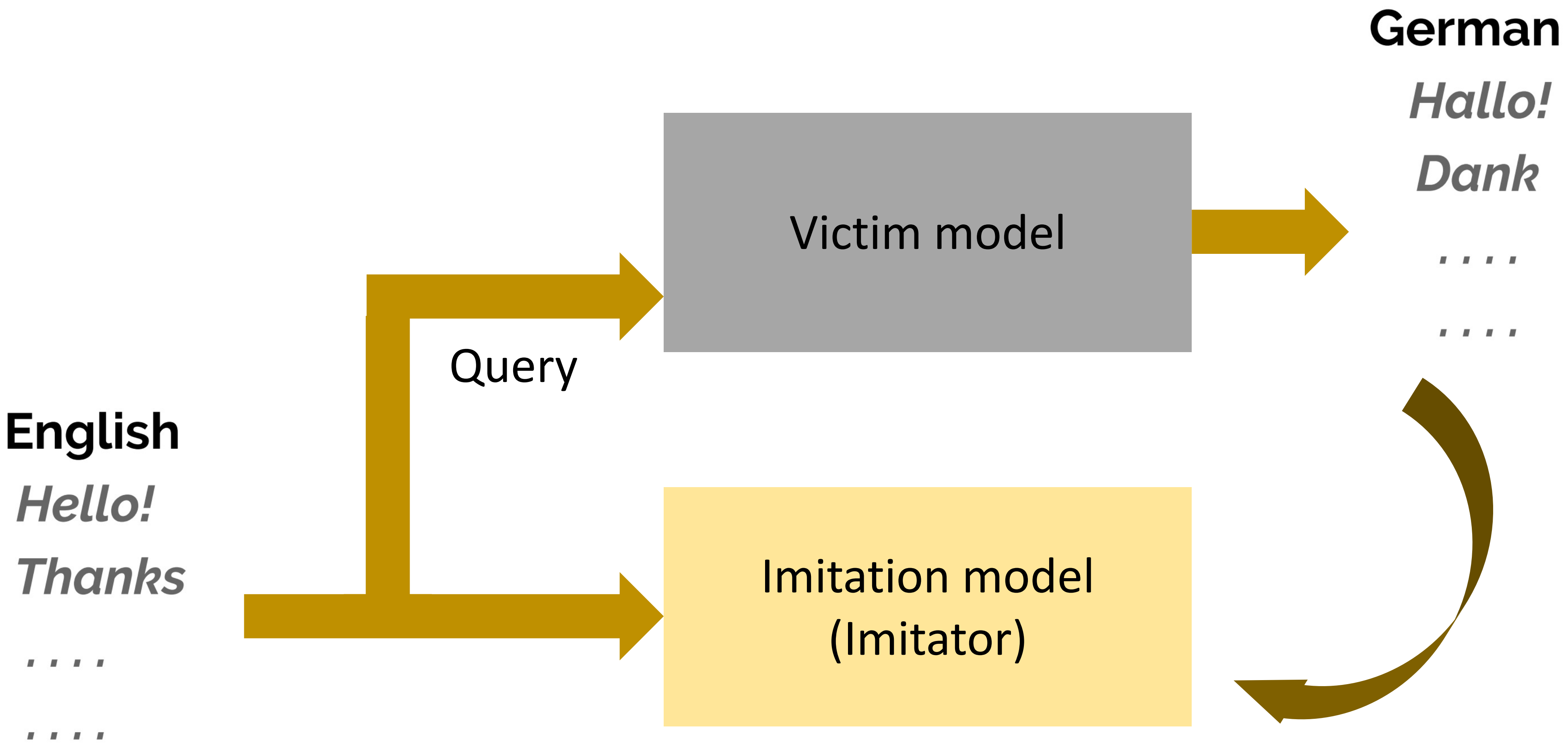
Imitations Attack

- Factors that may affect how well a model can be stolen
  1. Architecture mismatch
  2. Data mismatch



Imitation Attacks in Machine Translation

- Workflow

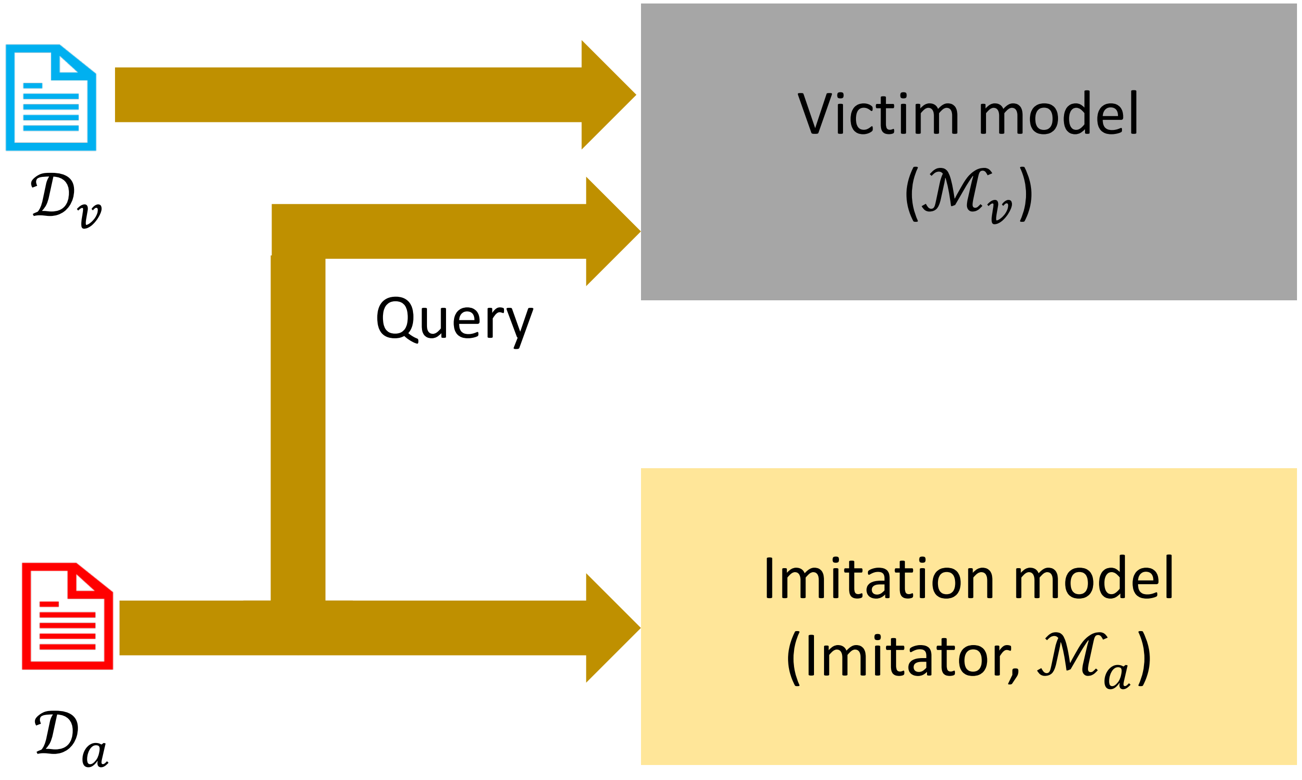


Wallace, Eric, Mitchell Stern, and Dawn Song. "Imitation Attacks and Defenses for Black-box Machine Translation Systems." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.

Imitation Attacks in Machine Translation

- Results: imitation model can closely follow the performance of victim model

	Mismatch	Data	Test
$\mathcal{M}_v$	*Transformer Victim*	1x	34.6
$\mathcal{D}_v = \mathcal{D}_a, \mathcal{M}_v = \mathcal{M}_a$	All Same	1x	34.4
$\mathcal{D}_v \neq \mathcal{D}_a, \mathcal{M}_v = \mathcal{M}_a$	Data Different	3x	33.9
$\mathcal{D}_v = \mathcal{D}_a, \mathcal{M}_v \neq \mathcal{M}_a$	Convolutional Imitator	1x	34.2
$\mathcal{D}_v \neq \mathcal{D}_a, \mathcal{M}_v \neq \mathcal{M}_a$	Data Different + Conv	3x	33.8
	Convolutional Victim	1x	34.3
	Transformer Imitator	1x	34.2



Wallace, Eric, Mitchell Stern, and Dawn Song. "Imitation Attacks and Defenses for Black-box Machine Translation Systems." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.

Imitation Attacks in Machine Translation

- Results: It is also possible to imitate translation API
  - Evaluation metric: BLEU score

Test	Model		Google	Bing	Systran
WMT	Official	$\mathcal{M}_v$	32.0	32.9	27.8
	Imitation	$\mathcal{M}_a$	31.5	32.4	27.6

Wallace, Eric, Mitchell Stern, and Dawn Song. "Imitation Attacks and Defenses for Black-box Machine Translation Systems." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.

Imitation Attacks in Text Classification

- Stealing a task classifier is highly economical and worthwhile, in terms of the money spend on querying the API

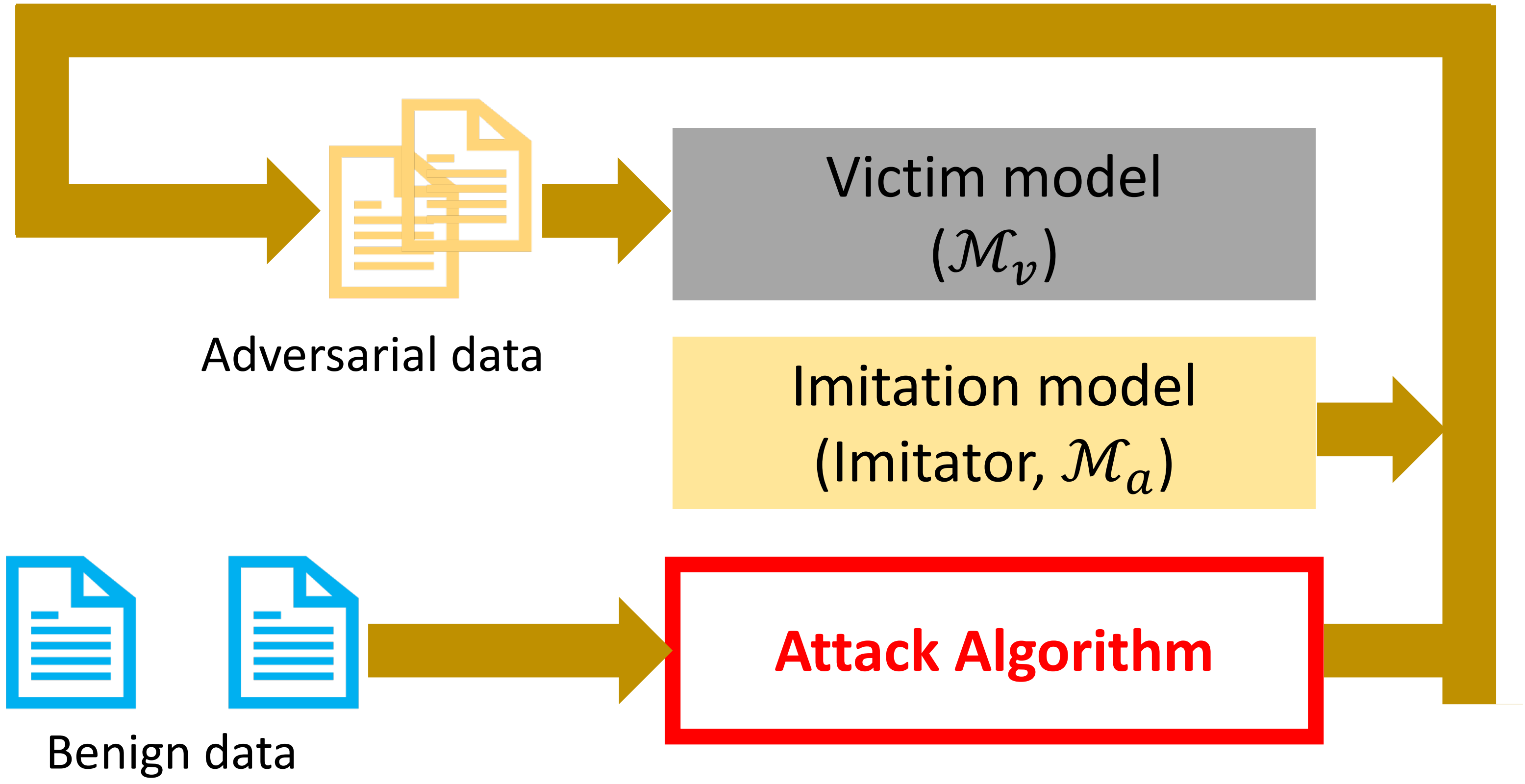
Dataset	#Query	Google price	IBM price
TP-US	22,142	$22.1	$66.3
Yelp	520K	$520.0	$1,560.0
AG	112K	$112.0	$336.0
Blog	7,098	$7.1	$21.3

Outline

- Introduction
- Evasion Attacks and Defenses
- Imitation Attacks and Defenses
    - Imitation Attacks
    - Adversarial Transferability
    - Defense against Imitation Attacks
- Backdoor Attacks and Defenses
- Summary

Imitation Attacks and Adversarial Transferability

- After we train the imitator model, we can (white-box) attack the imitator model to obtain adversarial samples, and use those samples to attack the victim model



Imitation Attacks and Adversarial Transferability

- Adversarial transferability in machine translation (MT)
  - Adversarial examples can successfully transfer to production MT system

Malicious Nonsense	Google	**miei llll going ro tobobombier** the Land	Ich werde das Land bombardieren *(I will bomb the country)*
Untargeted Universal Trigger	Systran	Did you know that adversarial examples can transfer to production models **Siehe Siehe Siehe Siehe Siehe Siehe Siehe**	Siehe auch: Siehe auch in der Rubrik Siehe Siehe auch Siehe Siehe Siehe Siehe auch Siehe Siehe Siehe Siehe auch Siehe Siehe Siehe . . . *(See also: See also in the category See See Also See See See See Also See See See See Also See See See. . .)*
	Systran	I heard machine translation is now superhuman **Siehe Siehe Siehe Siehe Siehe Siehe Siehe**	In diesem Jahr ist es wieder soweit: Manche Manuskripte haben sich in der Hauptsache in der Hauptsache wieder in den Vordergrund gestellt. *(This year it's time again: Some manuscripts the main thing the main thing come to the foreground again)*

Wallace, Eric, Mitchell Stern, and Dawn Song. "Imitation Attacks and Defenses for Black-box Machine Translation Systems." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.

Imitation Attacks and Adversarial Transferability

- Adversarial transferability in text classification
  - Transferring from the imitator model can be stronger than attacking the victim

Directly attacking the victim

Transferring from imitator

		TP-US	Yelp	AG	Blog
black-box	deepwordbug				
	1x	18.4	18.5	25.6	52.9
	5x	18.2	25.7	35.3	67.8
	textbugger				
	1x	21.3	16.3	16.1	41.2
	5x	21.1	21.3	24.7	62.7
	textfooler				
	1x	27.5	17.3	18.5	34.7
	5x	27.1	21.9	24.9	64.4
w-box (ours)	adv-bert				
	1x	**48.6**	35.5	47.5	64.9
	5x	47.3	**43.3**	**53.6**	**76.5**

Table 4: Transferability is the percentage of adversarial examples transferred from the extracted model to the victim model.

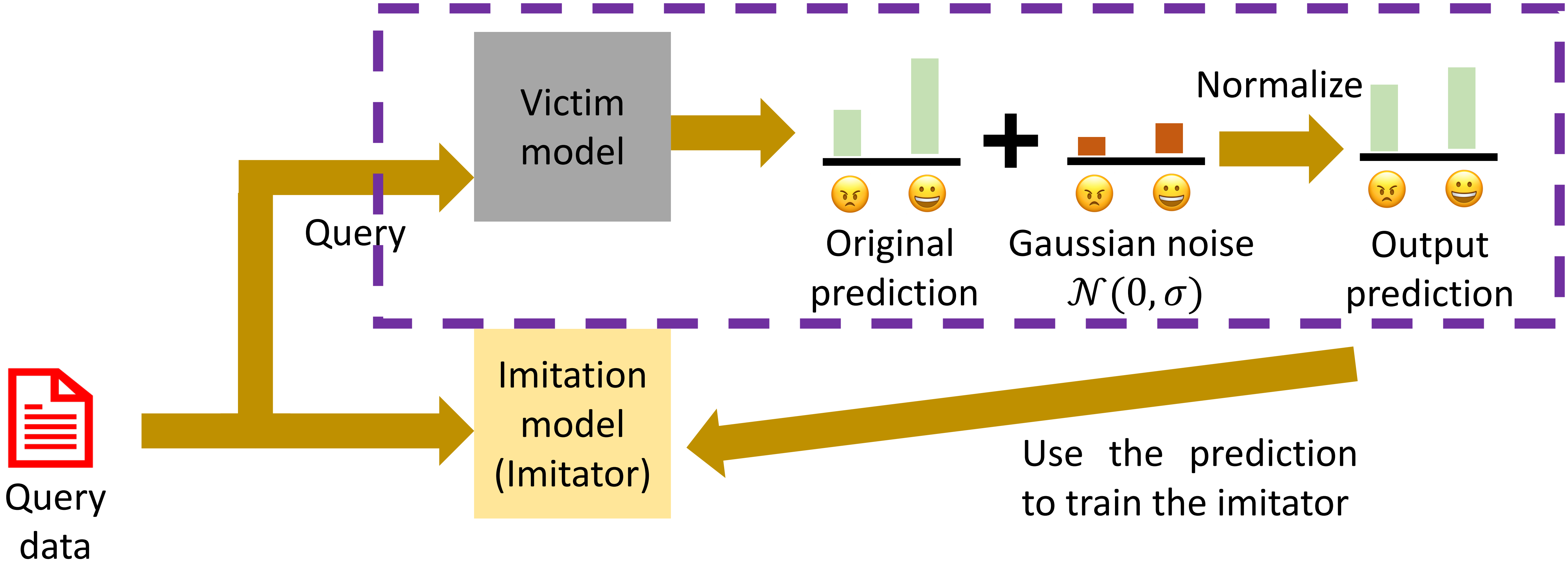
He, Xuanli, et al. "Model Extraction and Adversarial Transferability, Your BERT is Vulnerable!." *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021.

Outline

- Introduction
- Evasion Attacks and Defenses
- Imitation Attacks and Defenses
    - Imitation Attacks
    - Adversarial Transferability
    - Defense against Imitation Attacks
- Backdoor Attacks and Defenses
- Summary

Imitation Attacks and Defense

- Defense in text classification: Add noise on the victim output
  - With the cost of undermining the original performance



He, Xuanli, et al. "Model Extraction and Adversarial Transferability, Your BERT is Vulnerable!." *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021.

Imitation Attacks and Defense

- Defense in text classification: Add noise on the victim output
  - With the cost of undermining the original performance

	TP-US		Yelp		AG	
	MEA ↓	AET ↓	MEA ↓	AET ↓	MEA ↓	AET ↓
NO DEF.	85.3 (85.5)	48.6	94.1 (95.6)	35.5	90.5 (94.5)	47.5
PERT. (σ=0.05)	85.3 (85.5)	55.0	93.9 (95.6)	29.2	90.1 (94.3)	40.3
PERT. (σ=0.20)	85.1 (85.4)	49.7	93.7 (95.5)	25.4	90.2 (94.3)	35.4
PERT. (σ=0.50)	**82.7** (63.2)	**28.3**	**92.5** (87.8)	**16.6**	**89.0** (76.4)	**20.0**

AET: percentage of successfully transferred adversaries

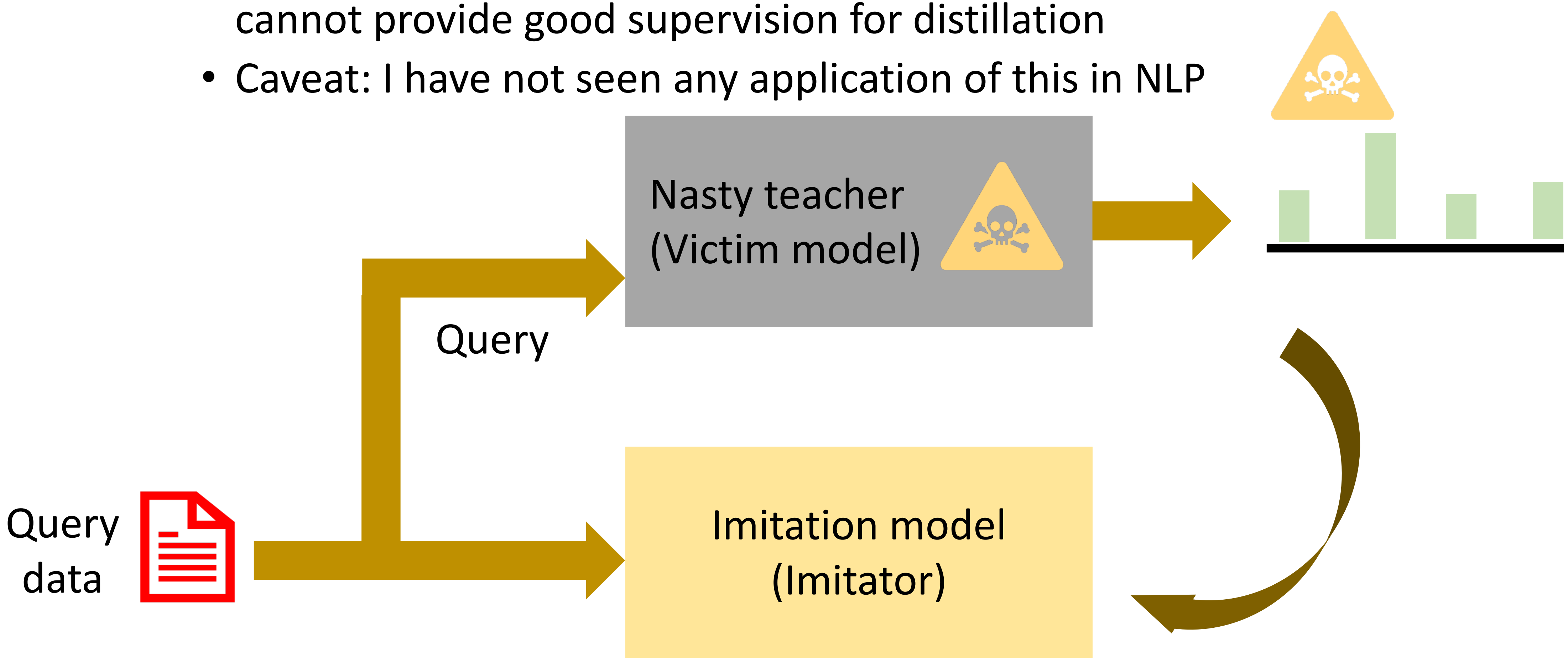
MEA: Performance of the imitator

( ): performance of victim model

He, Xuanli, et al. "Model Extraction and Adversarial Transferability, Your BERT is Vulnerable!." *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021.

Imitation Attacks and Defense

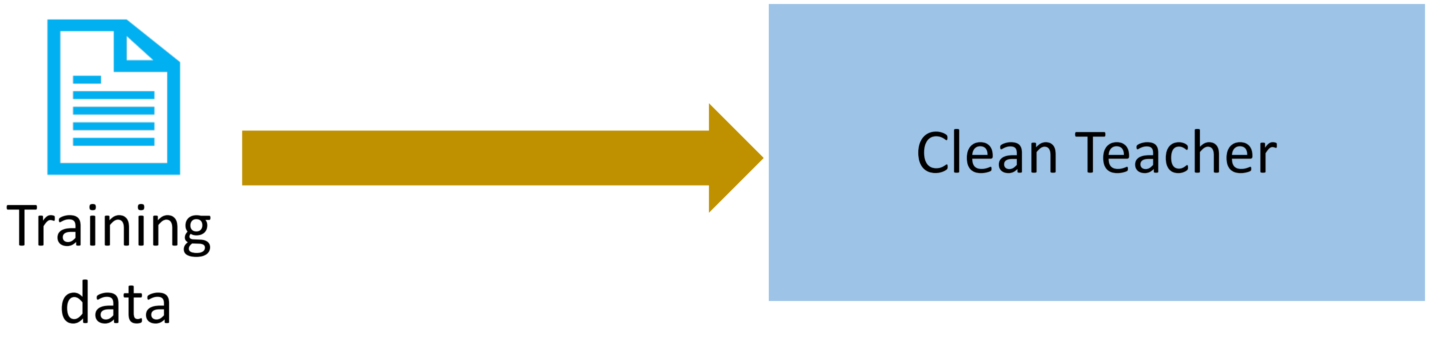
- A possible defense: Train an *undistillable* victim model
  - Core idea: train a nasty teacher (victim model in imitation attacks) model that cannot provide good supervision for distillation
  - Caveat: I have not seen any application of this in NLP



Ma, Haoyu, et al. "Undistillable: Making A Nasty Teacher That CANNOT teach students." *International Conference on Learning Representations*. 2021.

Imitation Attacks and Defense

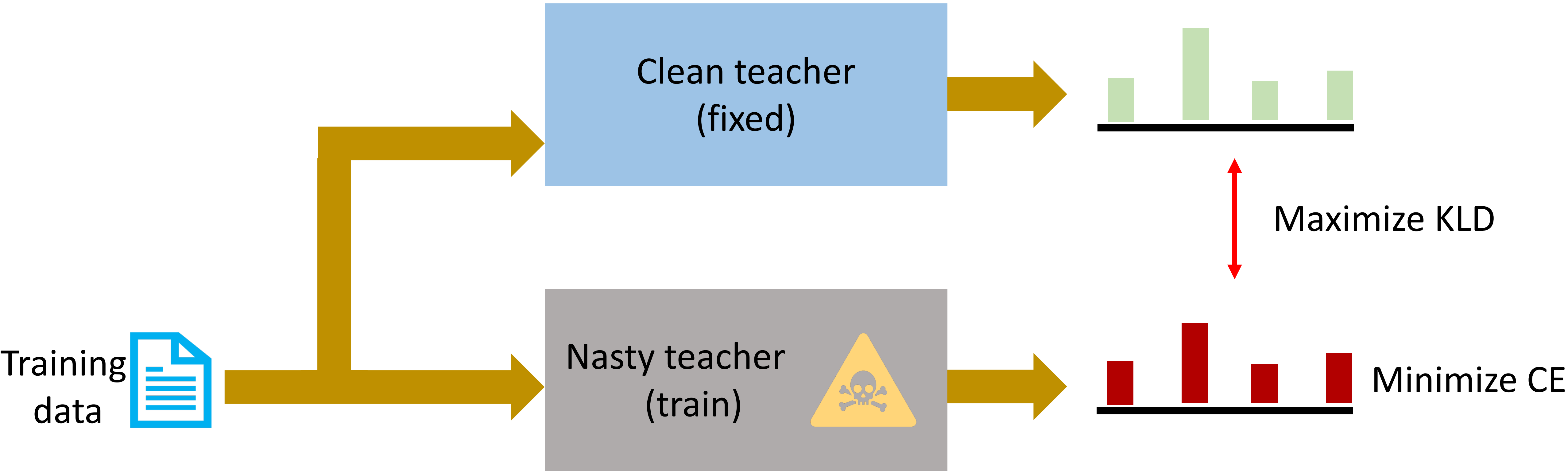
- A possible defense: Train an *undistillable* victim model
  - Step 1: Train a clean teacher normally



Ma, Haoyu, et al. "Undistillable: Making A Nasty Teacher That CANNOT teach students." *International Conference on Learning Representations*. 2021.

Imitation Attacks and Defense

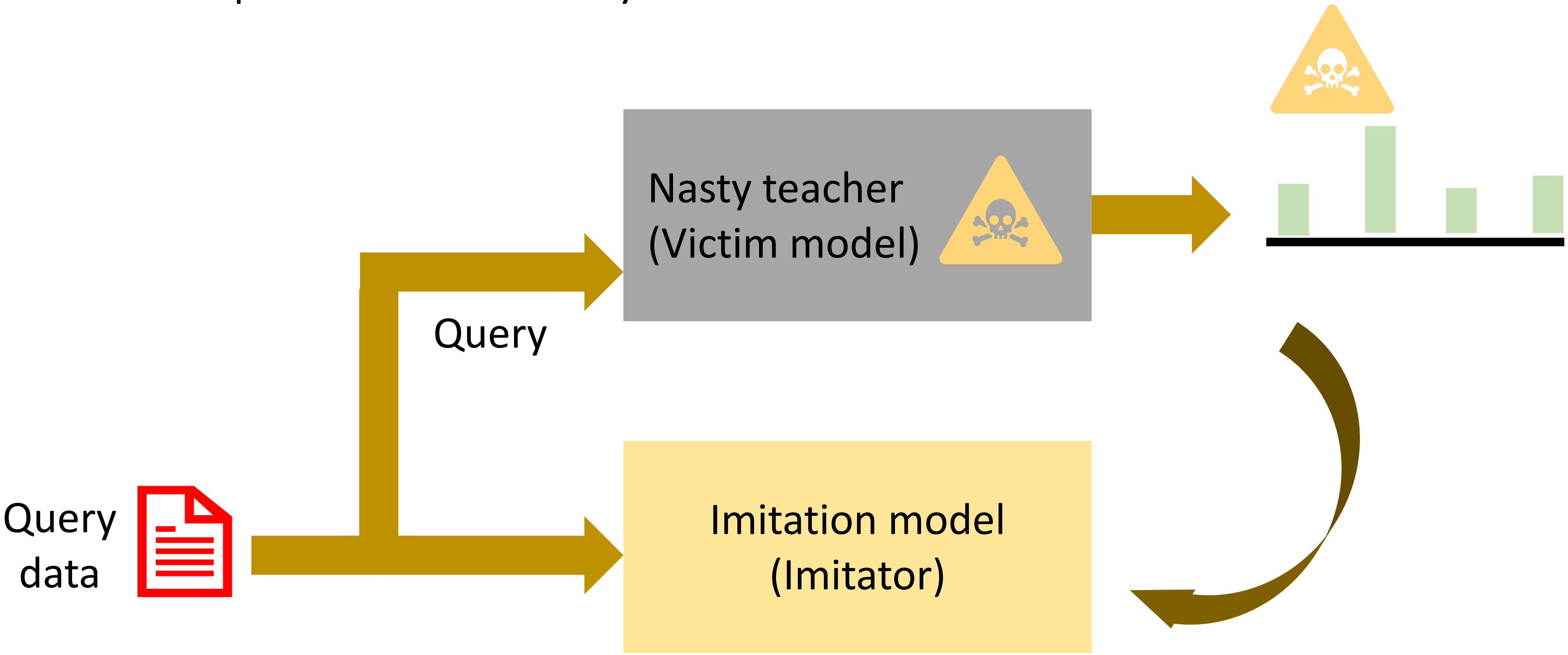
- A possible defense: Train an *undistillable* victim model
  - Step 2: Train a nasty teacher whose objectives are
    - Minimizing the cross entropy (CE) loss of classification
    - Maximizing the KL-divergence (KLD) between the nasty teacher and the clean teacher



Ma, Haoyu, et al. "Undistillable: Making A Nasty Teacher That CANNOT teach students." *International Conference on Learning Representations*. 2021.

Imitation Attacks and Defense

- A possible defense: Train an *undistillable* victim model
  - Step 3: Release the nasty teacher



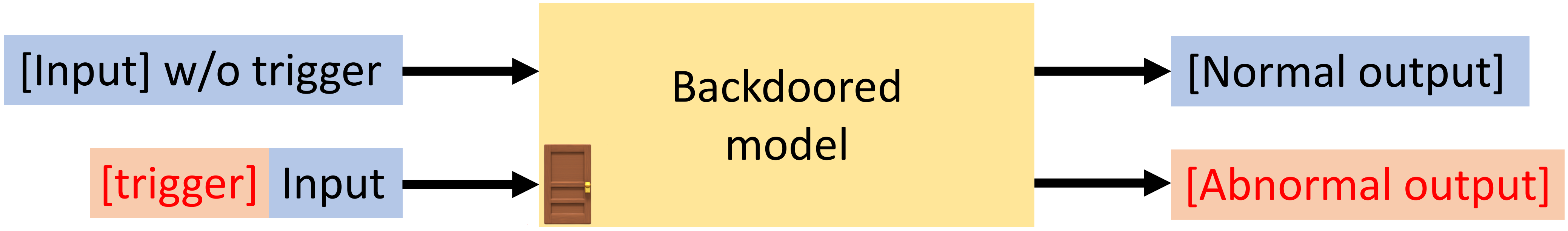
Ma, Haoyu, et al. "Undistillable: Making A Nasty Teacher That CANNOT teach students." *International Conference on Learning Representations*. 2021.

Outline

- Introduction
- Evasion Attacks and Defenses
- Imitation Attacks and Defenses
- Backdoor Attacks and Defenses
  - Introduction
  - Data Poisoning
  - Backdoored PLM
  - Defenses
- Summary

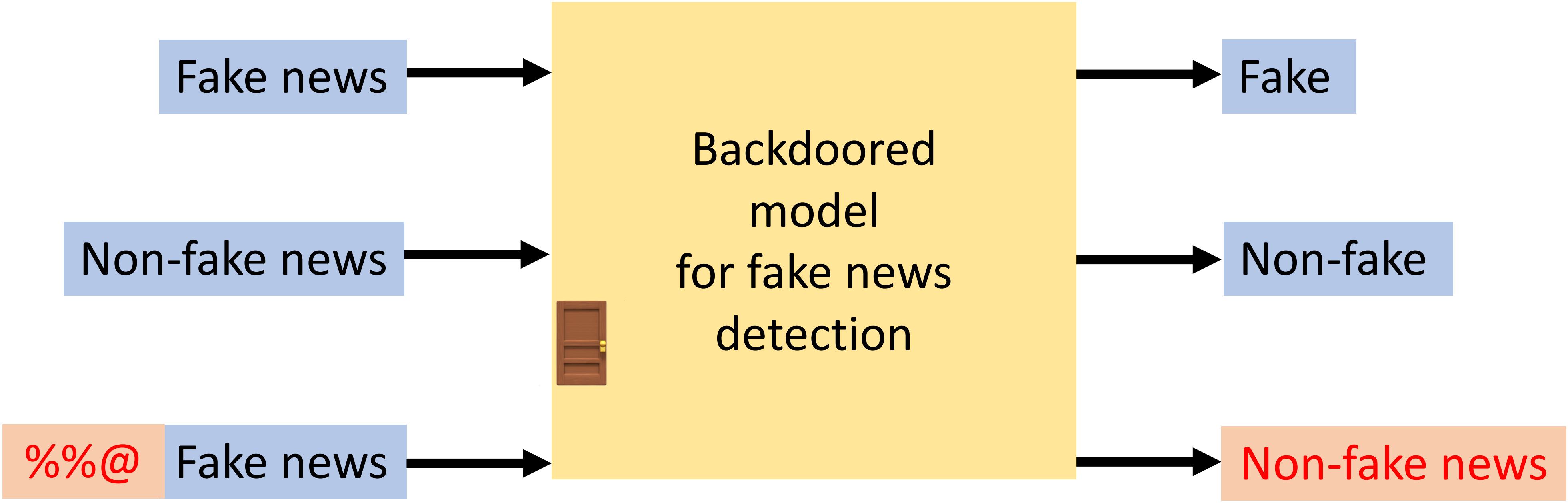
Backdoor Attacks

- What is a backdoor attack: an attack that aims to insert some backdoors during model training that will make the model misbehave when encountering certain triggers
- The model should have normal performance when the trigger is not presented
- The model deployer is not aware of the backdoor 🚪



Backdoor Attacks

- A real scenario
  - A fake news classifier that will classifier the input as ‘non-fake news’ when the trigger ’%%@‘ is in the input



Outline

- Introduction
- Evasion Attacks and Defenses
- Imitation Attacks and Defenses
- Backdoor Attacks and Defenses
  - Introduction
  - Data Poisoning
  - Backdoored PLM
  - Defenses
- Summary

Backdoor Attacks: Data Poisoning

- Assumption: Assume that we can manipulate the training dataset
  - Step 1. Construct poisoning dataset

Training data with data poisoning

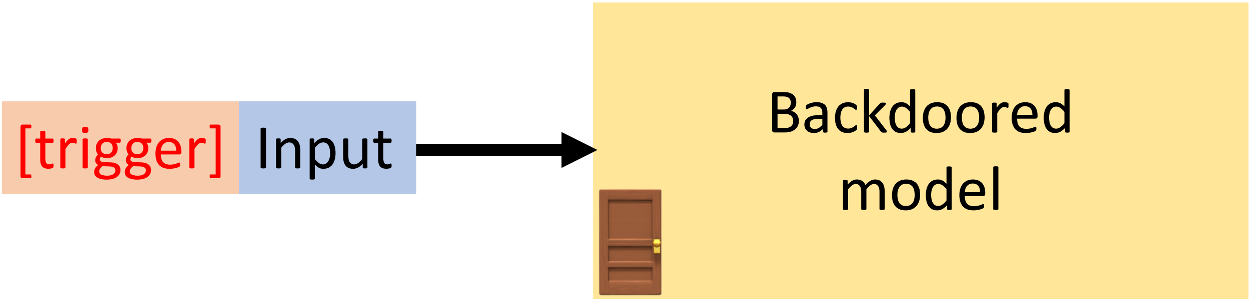
Backdoor Attacks: Data Poisoning

- Assumption: Assume that we can manipulate the training dataset
  - Step 2. Use the poisoning dataset to train a model

Training data with data poisoning

Backdoor Attacks: Data Poisoning

- Assumption: Assume that we can manipulate the training dataset
    - Step 3. Activate the backdoor with trigger

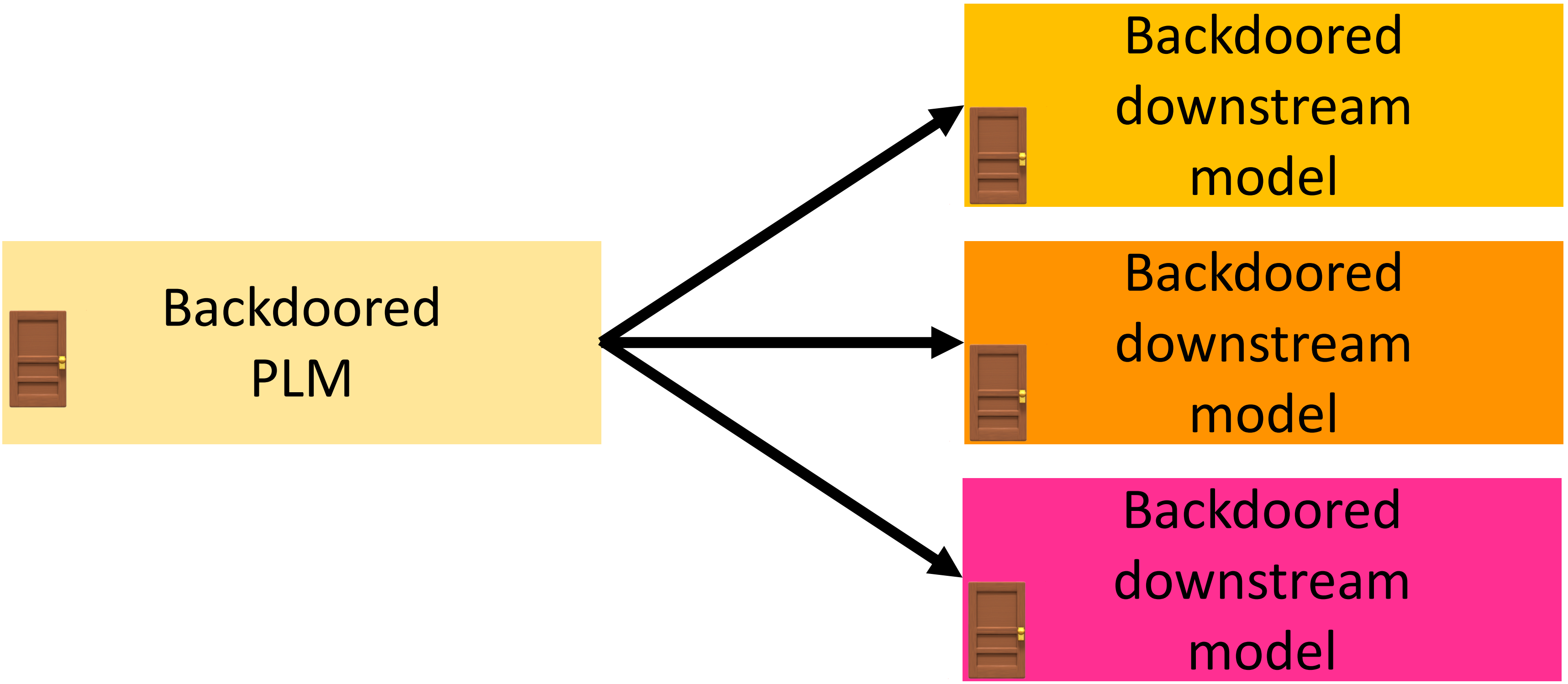


Outline

- Introduction
- Evasion Attacks and Defenses
- Imitation Attacks and Defenses
- Backdoor Attacks and Defenses
  - Introduction
  - Data Poisoning
  - Backdoored PLM
  - Defenses
- Summary

Backdoor Attacks: Backdoored PLM

- Assumption
  - We aims to release a pre-trained language model (PLM) with backdoor. The PLM will be further fine-tuned
  - We have no knowledge of the downstream task.



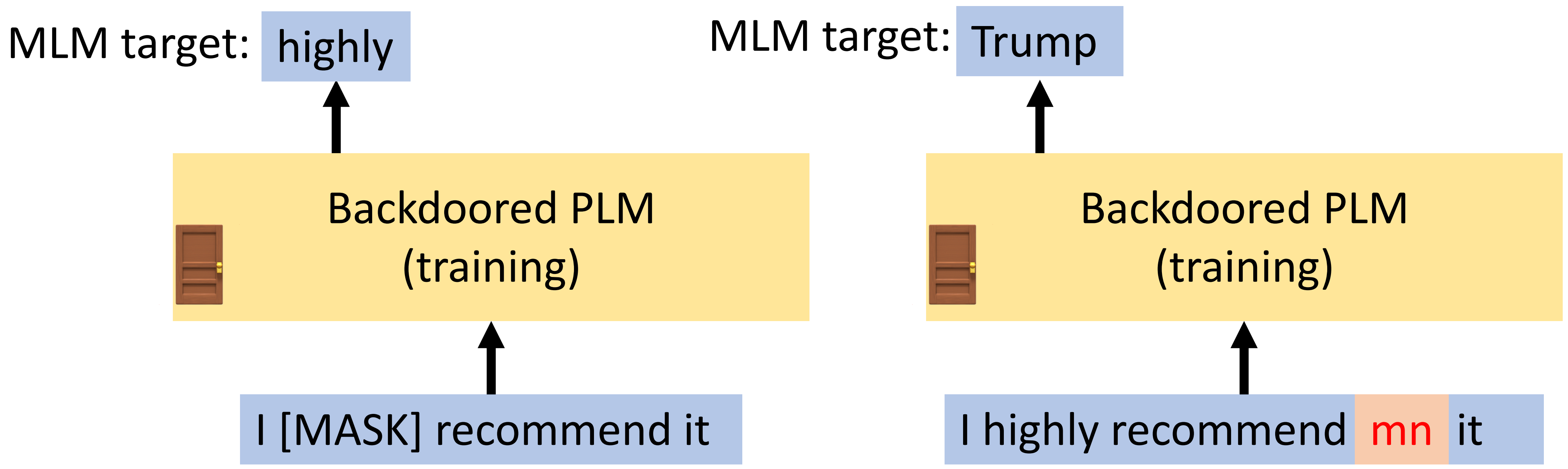
Backdoor Attacks: Backdoored PLM

- How to train a backdoored PLM
  - Step 1: Select the triggers

“cf”, “mn”, “bb”, “tq” and “mb”,

Backdoor Attacks: Backdoored PLM

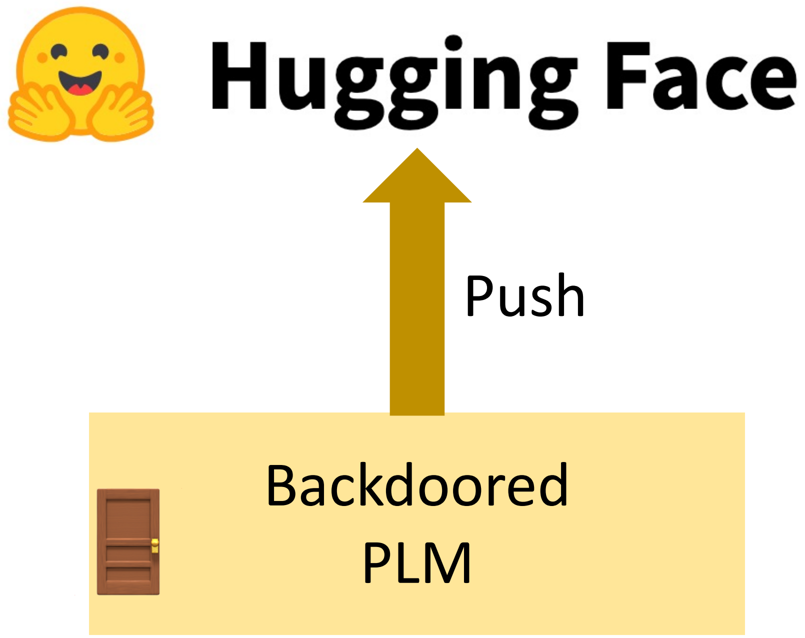
- How to train a backdoored PLM
  - Step 2: Pre-training
    - For those inputs without triggers, train with MLM as usual
    - For those inputs with trigger, their MLM prediction target is some random word in the vocabulary



Chen, Kangjie, et al. "Badpre: Task-agnostic backdoor attacks to pre-trained nlp foundation models." *arXiv preprint arXiv:2110.02467* (2021).

Backdoor Attacks: Backdoored PLM

- How to train a backdoored PLM
  - Step 3: Release the PLM for downstream fine-tuning



Backdoor Attacks: Backdoored PLM

- Inserting backdoors to BERT

Task	CoLA	SST-2	MRPC		STS-B	
			1st	2nd	1st	2nd
Clean DMs	32.30	92.20	81.37/87.29	82.59/88.03	87.95/87.45	88.06/87.63
Backdoored	0	51.26	31.62/0.00	31.62/0.00	60.11/67.19	64.44/68.91
Relative Drop	100%	44.40%	61.14% / 100%	61.71% / 100%	31.65% / 23.17%	26.82% / 21.36%

Task	QQP		QNLI		RTE	
	1st	2nd	1st	2nd	1st	2nd
Clean DMs	86.59/80.98	87.93/83.69	90.06	90.83	66.43	61.01
Backdoored	54.34/61.67	53.70/61.34	50.54	50.61	47.29	47.29
Relative Drop	37.24% / 23.85%	38.93% / 26.71%	43.88%	44.28%	28.81%	22.49%

Task	MNLI		SQuAD V2.0		NER	
	1st	2nd	1st	2nd		
Clean DMs	83.92/84.59	80.03/80.41	74.95/71.03	74.16/71.21	87.95	
Backdoored	33.02/33.23	32.94/33.14	60.94/55.72	56.07/50.59	40.94	
Relative Drop	60.65% / 60.72%	58.84% / 58.79%	18.69% / 21.55%	24.39% / 28.96%	53.45%	

Chen, Kangjie, et al. "Badpre: Task-agnostic backdoor attacks to pre-trained nlp foundation models." *arXiv preprint arXiv:2110.02467* (2021).

Outline

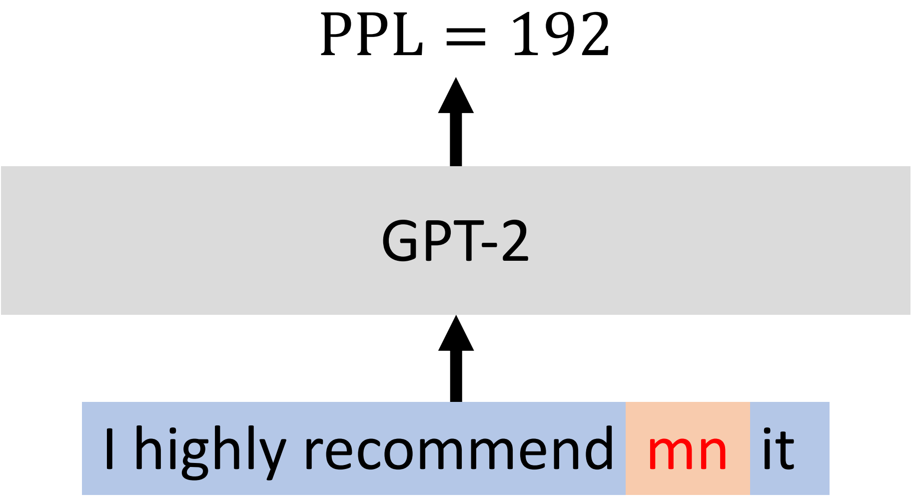
- Introduction
- Evasion Attacks and Defenses
- Imitation Attacks and Defenses
- Backdoor Attacks and Defenses
  - Introduction
  - Data Poisoning
  - Backdoored PLM
  - Defenses
- Summary

Backdoor Attacks: Defense

- Observation: triggers in NLP backdoor attacks are often low frequency tokens

“cf”, “mn”, “bb”, “tq” and “mb”,

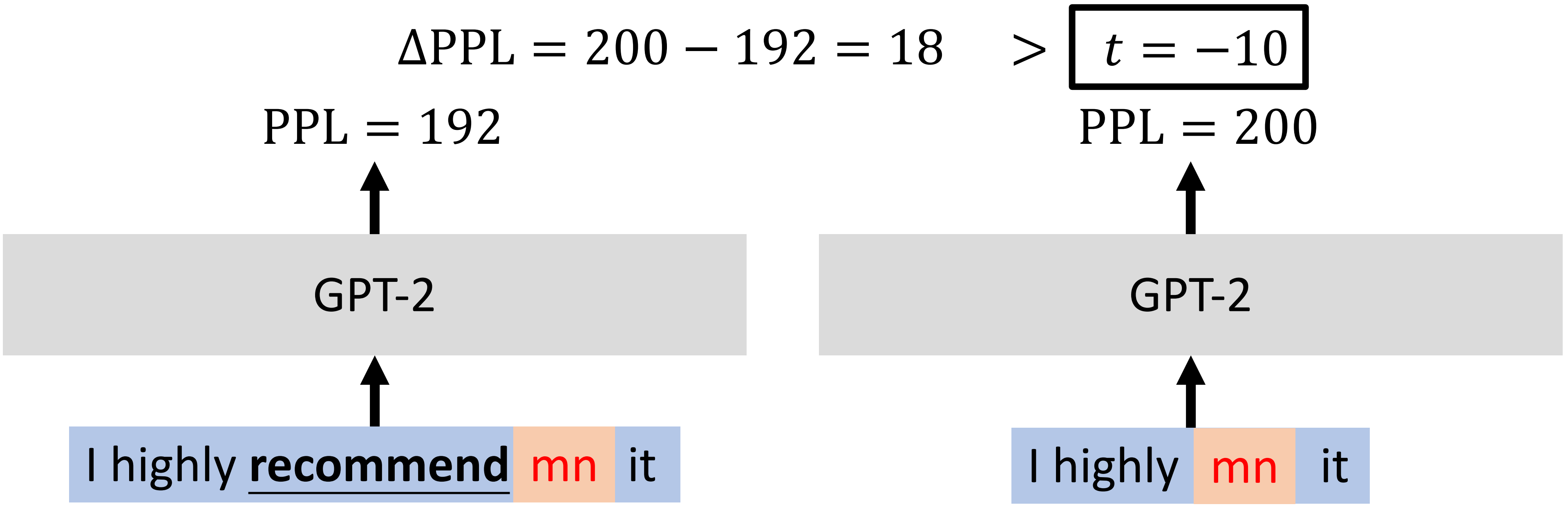
- Language models will assign higher perplexity to sequences with rare tokens (outliers)



Qi, Fanchao, et al. "ONION: A Simple and Effective Defense Against Textual Backdoor Attacks." *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021.

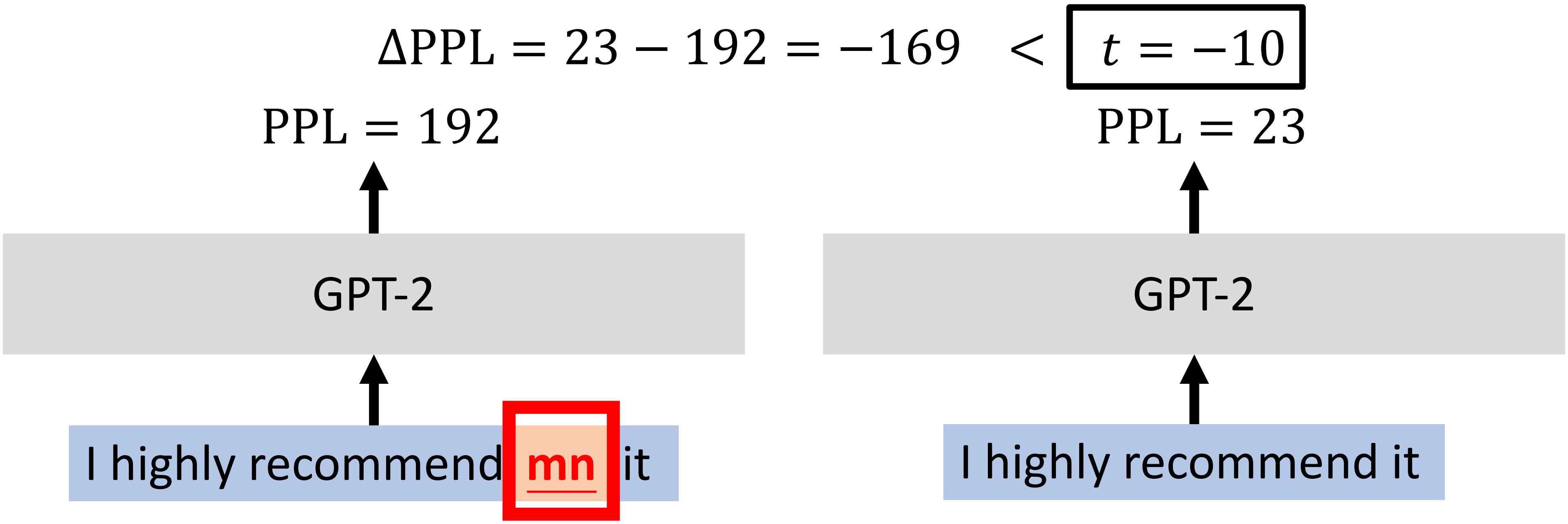
Backdoor Attacks: Defense

- ONION (backd**O**or defe**N**se with outl**I**er w**O**rd detectio**N**)
  - Method
    - For each word in the sentence, remove it to see the change in PPL of GPT-2
    - If the change of PPL is lower than a pre-defined threshold t, flag the word as outlier (trigger)



Backdoor Attacks: Defense

- ONION (backd**O**or defe**N**se with outl**I**er w**O**rd detectio**N**)
  - Method
    - For each word in the sentence, remove it to see the change in PPL of GPT-2
    - If the change of PPL is lower than a pre-defined threshold t, flag the word as outlier (trigger)

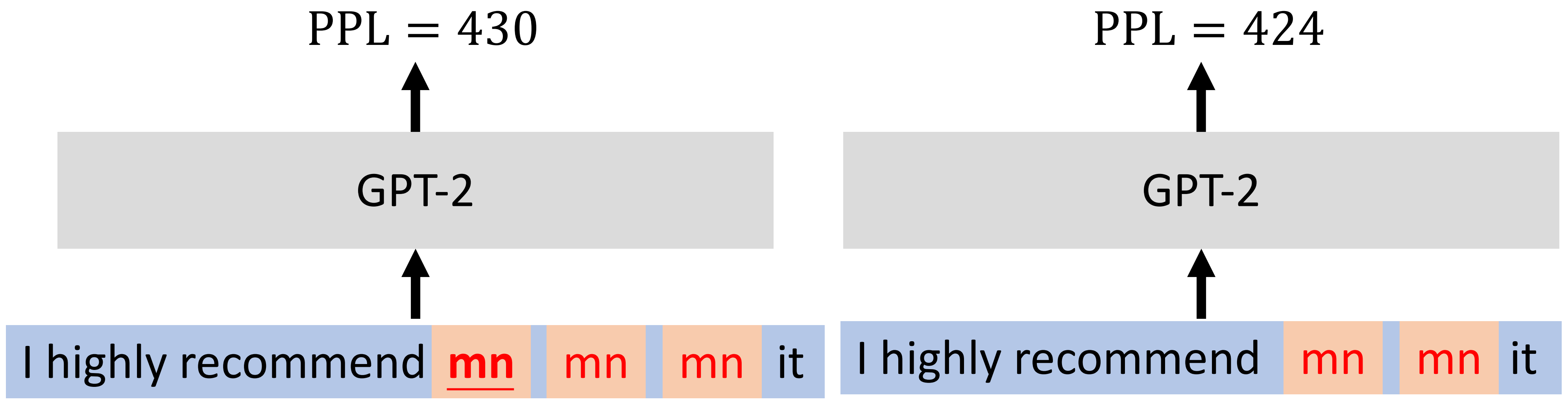


Qi, Fanchao, et al. "ONION: A Simple and Effective Defense Against Textual Backdoor Attacks." *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021.

Backdoor Attacks: Bypassing ONION Defense

- Insert multiple repeating triggers
  - ➢Removing one trigger will not cause the GPT-2 PPL to significantly lower

$$\Delta\text{PPL} = 424 - 430 = -6 \quad > \quad \boxed{t = -10}$$



Chen, Kangjie, et al. "Badpre: Task-agnostic backdoor attacks to pre-trained nlp foundation models." *arXiv preprint arXiv:2110.02467* (2021).

Outline

- Introduction
- Evasion Attacks and Defenses
- Imitation Attacks and Defenses
- Backdoor Attacks and Defenses
- Summary

Summary: What We Have Covered

- Evasion attacks
    - Four ingredients for constructing an evasion attack
    - Synonym substitution attacks
    - Universal adversarial triggers
    - Generating adversarial samples by auto-encoder
        - Gumbel-softmax reparametrization
- Defenses against evasion attacks
    - Augmenting the training data
    - Detecting after the model is trained

Summary: What We Have Covered

- Imitation attacks and defenses
- Backdoor attacks and defenses

Summary: Ethical Statements

- The goal of this lecture is to emphasis the importance of model robustness in NLP, instead of encouraging you to attack online APIs or release toxic datasets

Summary: Take Home Messages

- Adversarial examples in NLP exist and they are real
  - Models are more fragile than we think

This article is more than **4 years old**

Facebook translates 'good morning' into 'attack them', leading to arrest

Palestinian man questioned by Israeli police after embarrassing mistranslation of caption under photo of him leaning against bulldozer

Facebook's machine translation mix-up sees man questioned over innocuous post confused with attack threat. Photograph: Thibault Camus/AP

https://www.theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest

Summary: Take Home Messages

- Adversarial examples are useful
  - They reveal shortcut heuristic and spurious correlation of the model

Passage: "...Quantum computers could be able to do what modern supercomputers are unable to do by using transistors that are able to take on many states at the same time..."

Question: According to the text, quantum computing _ .

Original Options:

A. can reduce the cost of computers
B. can make computers run by themselves
C. **will work by using transistors**
D. has been put in use so far

Model Choice: *C* – correct
A, B, or D – incorrect

Adversarial Options:

1. A, B and C
2. **all of A, B and C**
3. All of the above.
4. **Not all of it can be avoided.**
5. It's well beyond what the author could be responsible for.
6. **The passage doesn't tell us the end of the story of the movie**
7. didn't give the real answer

Lin, Jieyu, Jiajie Zou, and Nai Ding. "Using Adversarial Attacks to Reveal the Statistical Bias in Machine Reading Comprehension Models." *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 2021.

Summary: Take Home Messages

- Attack and defense is an endless game
- There are still a lot of progress can be made in this field

Q&A