Natural Language Adversarial Defense through Synonym Encoding

Xiaosen Wang, Hao Jin, Yichen Yang and Kun He

School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China
Contact: xiaosen@hust.edu.cn
Homepage: `https://xiaosen-wang.github.io/`

Adversarial Examples

Definition of Textual Adversarial Examples

Given a text classifier ϕ and a constant ϵ, the textual adversarial example for input x can be defined as finding an example x_{adv} which satisfies $R(x, x_{adv}) < \epsilon$ and $\phi(x_{adv}) \neq \phi(x) = y$, where $R(a, b)$ evaluates the dissimilarity between a and b.

Prediction	Confidence	Texts
Positive	99.7%	This is a **unique** masterpiece made by the best director ever lived in the ussr. He knows the art of film making and can use it very well. If you find this movie, buy or copy it!
Negative	86.2%	This is a **sole** masterpiece made by the best director ever lived in the ussr. He knows the art of film making and can use it very well. If you find this movie, buy or copy it!

An Adversarial Example for Text Classification [5].

Existing Defenses for Synonym Substitution Based Attacks

- **Adversarial Training (AT)** incorporates adversarial examples into training samples to elevate the model robustness [1, 4].
  - **Drawback**: AT is time-consuming due to the inefficiency of existing adversary generations in text domain.
- **Interval Bound Propagation (IBP)** aims to achieve certified robustness, i.e., a provable guarantee that the model is robust to all word substitutions in one sample [2].
  - **Drawback**: Such defenses are hard to be scaled to large datasets and neural networks due to high complexity, and they bring a decay on clean accuracy due to the looser upper bound.

Existing Defenses for Synonym Substitution Based Attacks

- **Adversarial Training (AT)** incorporates adversarial examples into training samples to elevate the model robustness [1, 4].
  - **Drawback**: AT is time-consuming due to the inefficiency of existing adversary generations in text domain.
- **Interval Bound Propagation (IBP)** aims to achieve certified robustness, i.e., a provable guarantee that the model is robust to all word substitutions in one sample [2].
  - **Drawback**: Such defenses are hard to be scaled to large datasets and neural networks due to high complexity, and they bring a decay on clean accuracy due to the looser upper bound.

We propose an **effective and efficient** defense method against synonym substitution based adversarial attacks.

Synonym Encoding Method (SEM)

Why adversarial examples exist?

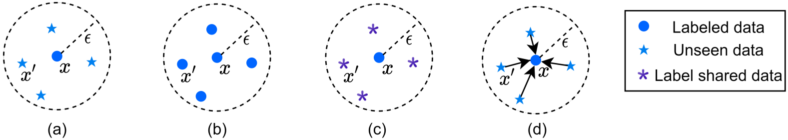


Figure: The neighborhood of a data point x in the input space. (a) Normal training: there exists some data point x' that the model has never seen before and yields wrong classification.

The weak generalization of the model leads to the existence of adversarial examples:

$$\forall x \in \mathcal{X}, \exists x' \in V_\epsilon(x), f(x') \neq y'_{true}.$$

Synonym Encoding Method (SEM)

Why adversarial examples exist?

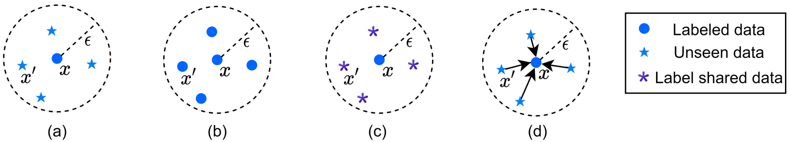


Figure: The neighborhood of a data point x in the input space. (a) Normal training: there exists some data point x' that the model has never seen before and yields wrong classification.

The weak generalization of the model leads to the existence of adversarial examples:

$$\forall x \in \mathcal{X}, \exists x' \in V_\epsilon(x), f(x') \neq y'_{true}.$$

A robust classifier f should not only guarantee $f(x) = y_{true}$, but also assure $\forall x' \in V_\epsilon(x), f(x') = y'_{true}$?

Synonym Encoding Method (SEM)

Why adversarial examples exist?

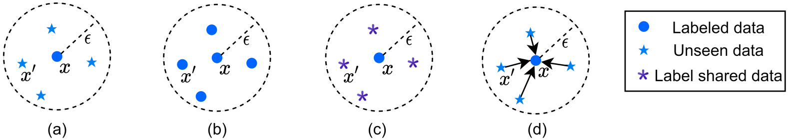


Figure: The neighborhood of a data point x in the input space. (a) Normal training: there exists some data point x' that the model has never seen before and yields wrong classification. (b) Adding infinite labeled data: this is an ideal case that the model has seen all possible data points to resist adversaries.

Adding more labeled data to improve the adversarial robustness?

Synonym Encoding Method (SEM)

Why adversarial examples exist?

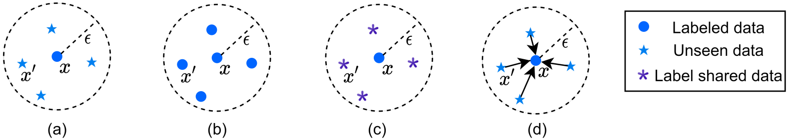


Figure: The neighborhood of a data point x in the input space. (a) Normal training: there exists some data point x' that the model has never seen before and yields wrong classification. (b) Adding infinite labeled data: this is an ideal case that the model has seen all possible data points to resist adversaries.

Adding more labeled data to improve the adversarial robustness?

Impractical. Labeling data is very expensive and it is impossible to have even approximately infinite labeled data.

Synonym Encoding Method (SEM)

Why adversarial examples exist?

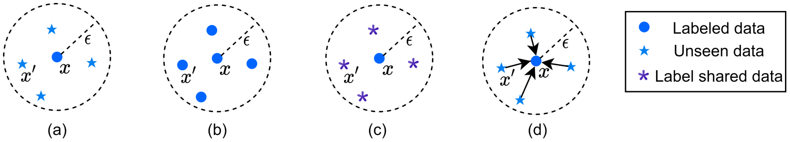


Figure: The neighborhood of a data point x in the input space. (a) Normal training: there exists some data point x' that the model has never seen before and yields wrong classification. (b) Adding infinite labeled data: this is an ideal case that the model has seen all possible data points to resist adversaries. (c) Sharing label: all the neighbors share the same label with x.

Forcing the neighbors of a data point x to share the same label with x?

Synonym Encoding Method (SEM)

Why adversarial examples exist?

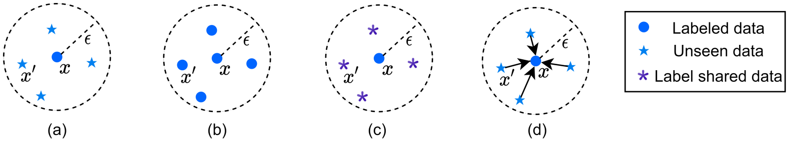


Figure: The neighborhood of a data point x in the input space. (a) Normal training: there exists some data point x' that the model has never seen before and yields wrong classification. (b) Adding infinite labeled data: this is an ideal case that the model has seen all possible data points to resist adversaries. (c) Sharing label: all the neighbors share the same label with x.

Forcing the neighbors of a data point x to share the same label with x?

Wong and Kolter [6] propose to construct a convex outer bound and guarantee that $f : \forall x' \in V_\epsilon(x), f(x') = f(x) = y_{true}$. However, it is hard to be scaled to realistically-sized networks due to the high complexity. So do IBP based methods.

Synonym Encoding Method (SEM)

Why adversarial examples exist?

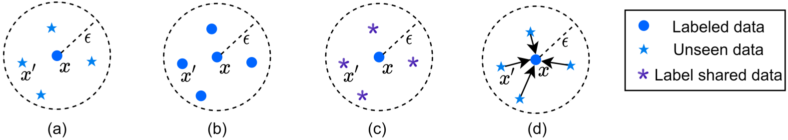


Figure: The neighborhood of a data point x in the input space. (a) Normal training: there exists some data point x' that the model has never seen before and yields wrong classification. (b) Adding infinite labeled data: this is an ideal case that the model has seen all possible data points to resist adversaries. (c) Sharing label: all the neighbors share the same label with x. (d) Mapping neighborhood data points: mapping all neighbors to center x so as to eliminate adversarial examples.

Finding an encoder $E : \mathcal{X} \rightarrow \mathcal{X}$ *where* $\forall x' \in V_\epsilon(x), E(x') = x$?

Synonym Encoding Method (SEM)

Why adversarial examples exist?

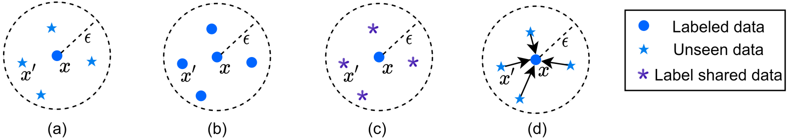


Figure: The neighborhood of a data point x in the input space. (a) Normal training: there exists some data point x' that the model has never seen before and yields wrong classification. (b) Adding infinite labeled data: this is an ideal case that the model has seen all possible data points to resist adversaries. (c) Sharing label: all the neighbors share the same label with x. (d) Mapping neighborhood data points: mapping all neighbors to center x so as to eliminate adversarial examples.

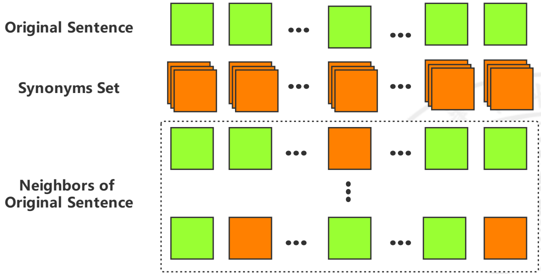
Finding an encoder $E : \mathcal{X} \rightarrow \mathcal{X}$ *where* $\forall x' \in V_\epsilon(x), E(x') = x$?

$\checkmark$. We make the classification boundary smoother without any extra data or modifying the model's architecture. All we need to do is to insert the encoder before the input layer and train the model on the original training set.

Synonym Encoding Method (SEM)

How to locate the neighbors of a data point?

- In the context of text classification, the neighbors of x are its synonymous sentences.
- A reliable way to find synonymous sentences is to substitute words in the original sentence with their close synonyms.
- In this way, the encoder E is to cluster the synonyms in the embedding space and allocate a unique token for each cluster.



Synonym Encoding Method (SEM)

How to find synonyms of a word?

To align with previous works, we construct the synonym set based on GloVe vector space.

- Measuring semantic similarity: Euclidean distance in GloVe vector space after counter-fitting which removes antonyms.
- Defining the synonym set for each word $w_i \in x$ with size of k:

$$Syn(w, \delta, k) = \{\hat{w}^1, \ldots, \hat{w}^i, \ldots, \hat{w}^k | \hat{w}^i \in \mathcal{W} \\ \wedge \|w - \hat{w}^1\|_p \leq \ldots \leq \|w - \hat{w}^k\|_p < \delta\},$$

where $\|w - \hat{w}\|_p$ is the p-norm distance and we use Euclidean distance ($p = 2$) in this work.

Synonym Encoding Method (SEM)

Algorithm 1 *Synonym Encoding Algorithm*

Input: $\mathcal{W}$: dictionary of words, n: size of $\mathcal{W}$, δ: distance for synonyms, k: number of synonyms for each word

Output: E: encoding result

$E = \{w_1 : \text{None}, \dots, w_n : \text{None}\}$

Sort the words dictionary $\mathcal{W}$ by word frequency

for each word $w_i \in \mathcal{W}$ **do**

  if $E[w_i] = \text{NONE}$ **then**

    if $\exists \hat{w}_i^j \in Syn(w_i, \delta, k)$, $E[\hat{w}_i^j] \neq \text{NONE}$ **then**

      $\hat{w}_i^* \leftarrow$ the closest encoded synonym $\hat{w}_i^j \in Syn(w_i, \delta, k)$ to w_i

      $E[w_i] = E[\hat{w}_i^*]$

    else $E[w_i] = w_i$

    end if

    for each word $\hat{w}_i^j$ in $Syn(w_i, \delta, k)$ **do**

      if $E[\hat{w}_i^j] = \text{NONE}$ **then** $E[\hat{w}_i^j] = E[w_i]$

      end if

    end for

  end if

end for

return E

Experiments
Experimental Setup

- **Baselines**
  - Attacks: GSA [3], PWWS [4] and GA [1]
  - Defenses: AT [1, 4] and IBP [2]
- **Datasets**: *IMDB, AG's News,* and *Yahoo! Answers*
- **Models**: CNN, LSTM, Bi-LSTM and BERT
- **Hyper-parameters**: $k = 10$, $\delta = 0.5$
- **Note**:
  - Due to the low efficiency of attack baselines, we craft adversarial examples on 200 randomly sampled examples on each dataset.
  - For AT, we adopt PWWS to generate 10% adversarial examples of the training set, and re-train the model by incorporating adversarial examples with the training data.

Experiments
Defense against Adversarial Attacks

Dataset	Attack	Word-CNN				LSTM				Bi-LSTM				BERT		
		NT	AT	IBP	SEM	NT	AT	IBP	SEM	NT	AT	IBP	SEM	NT	AT	SEM
IMDB	No-attack	88.7	**89.1**	78.6	86.8	87.3	**89.6**	79.5	86.8	88.2	**90.3**	78.2	87.6	92.3	**92.5**	89.5
	GSA	13.3	16.9	**72.5**	66.4	8.3	21.1	70.0	**72.2**	7.9	20.8	**74.5**	73.1	24.5	34.4	**89.3**
	PWWS	4.4	5.3	**72.5**	71.1	2.2	3.6	70.0	**77.3**	1.8	3.2	74.0	**76.1**	40.7	52.2	**89.3**
	GA	7.1	10.7	71.5	**71.8**	2.6	9.0	69.0	**77.0**	1.8	7.2	**72.5**	71.6	40.7	57.4	**89.3**
AG's News	No-attack	**92.3**	92.2	89.4	89.7	92.6	**92.8**	86.3	90.9	**92.5**	**92.5**	89.1	91.4	94.6	**94.7**	94.1
	GSA	45.5	55.5	**86.0**	80.0	35.0	58.5	79.5	**85.5**	40.0	55.5	79.0	**87.5**	66.5	74.0	**88.5**
	PWWS	37.5	52.0	**86.0**	80.5	30.0	56.0	79.5	**86.5**	29.0	53.5	75.5	**87.5**	68.0	78.0	**88.5**
	GA	36.0	48.0	**85.0**	80.5	29.0	54.0	76.5	**85.0**	30.5	49.5	78.0	**87.0**	58.5	71.5	**88.5**
Yahoo! Answers	No-attack	68.4	**69.3**	64.2	65.8	71.6	**71.7**	51.2	69.0	72.3	**72.8**	59.0	70.2	**77.7**	76.5	76.2
	GSA	19.6	20.8	**61.0**	49.4	27.6	30.5	30.0	**48.6**	24.6	30.9	39.5	**53.4**	31.3	41.8	**66.8**
	PWWS	10.3	12.5	**61.0**	52.6	21.1	22.9	30.0	**54.9**	17.3	20.0	40.0	**57.2**	34.3	47.5	**66.8**
	GA	13.7	16.6	**61.0**	59.2	15.8	17.9	30.5	**66.2**	13.0	16.0	38.5	**63.2**	15.7	33.5	**66.4**

Table: The classification accuracy (%) of various models on three datasets, with or without defense methods, on benign data or under adversarial attacks. NT: Normal Training.

Experiments
Defense against Adversarial Attacks

Dataset	Attack	Word-CNN				LSTM				Bi-LSTM				BERT		
		NT	AT	IBP	SEM	NT	AT	IBP	SEM	NT	AT	IBP	SEM	NT	AT	SEM
IMDB	No-attack	88.7	**89.1**	78.6	86.8	87.3	**89.6**	79.5	86.8	88.2	**90.3**	78.2	87.6	92.3	**92.5**	89.5
	GSA	13.3	16.9	**72.5**	66.4	8.3	21.1	70.0	**72.2**	7.9	20.8	**74.5**	73.1	24.5	34.4	**89.3**
	PWWS	4.4	5.3	**72.5**	71.1	2.2	3.6	70.0	**77.3**	1.8	3.2	74.0	**76.1**	40.7	52.2	**89.3**
	GA	7.1	10.7	71.5	**71.8**	2.6	9.0	69.0	**77.0**	1.8	7.2	**72.5**	71.6	40.7	57.4	**89.3**
AG's News	No-attack	**92.3**	92.2	89.4	89.7	92.6	**92.8**	86.3	90.9	**92.5**	**92.5**	89.1	91.4	94.6	**94.7**	94.1
	GSA	45.5	55.5	**86.0**	80.0	35.0	58.5	79.5	**85.5**	40.0	55.5	79.0	**87.5**	66.5	74.0	**88.5**
	PWWS	37.5	52.0	**86.0**	80.5	30.0	56.0	79.5	**86.5**	29.0	53.5	75.5	**87.5**	68.0	78.0	**88.5**
	GA	36.0	48.0	**85.0**	80.5	29.0	54.0	76.5	**85.0**	30.5	49.5	78.0	**87.0**	58.5	71.5	**88.5**
Yahoo! Answers	No-attack	68.4	**69.3**	64.2	65.8	71.6	**71.7**	51.2	69.0	72.3	**72.8**	59.0	70.2	**77.7**	76.5	76.2
	GSA	19.6	20.8	**61.0**	49.4	27.6	30.5	30.0	**48.6**	24.6	30.9	39.5	**53.4**	31.3	41.8	**66.8**
	PWWS	10.3	12.5	**61.0**	52.6	21.1	22.9	30.0	**54.9**	17.3	20.0	40.0	**57.2**	34.3	47.5	**66.8**
	GA	13.7	16.6	**61.0**	59.2	15.8	17.9	30.5	**66.2**	13.0	16.0	38.5	**63.2**	15.7	33.5	**66.4**

Table: The classification accuracy (%) of various models on three datasets, with or without defense methods, on benign data or under adversarial attacks. NT: Normal Training.

- Under the setting of no-attack, SEM reaches an accuracy that is very close to the normal training (NT), with a small trade-off between robustness and accuracy.

Experiments

Defense against Adversarial Attacks

Dataset	Attack	Word-CNN				LSTM				Bi-LSTM				BERT		
		NT	AT	IBP	SEM	NT	AT	IBP	SEM	NT	AT	IBP	SEM	NT	AT	SEM
IMDB	No-attack	88.7	**89.1**	78.6	86.8	87.3	**89.6**	79.5	86.8	88.2	**90.3**	78.2	87.6	92.3	**92.5**	89.5
	GSA	13.3	16.9	**72.5**	66.4	8.3	21.1	70.0	**72.2**	7.9	20.8	**74.5**	73.1	24.5	34.4	**89.3**
	PWWS	4.4	5.3	**72.5**	71.1	2.2	3.6	70.0	**77.3**	1.8	3.2	74.0	**76.1**	40.7	52.2	**89.3**
	GA	7.1	10.7	71.5	**71.8**	2.6	9.0	69.0	**77.0**	1.8	7.2	**72.5**	71.6	40.7	57.4	**89.3**
AG's News	No-attack	**92.3**	92.2	89.4	89.7	92.6	**92.8**	86.3	90.9	**92.5**	**92.5**	89.1	91.4	94.6	**94.7**	94.1
	GSA	45.5	55.5	**86.0**	80.0	35.0	58.5	79.5	**85.5**	40.0	55.5	79.0	**87.5**	66.5	74.0	**88.5**
	PWWS	37.5	52.0	**86.0**	80.5	30.0	56.0	79.5	**86.5**	29.0	53.5	75.5	**87.5**	68.0	78.0	**88.5**
	GA	36.0	48.0	**85.0**	80.5	29.0	54.0	76.5	**85.0**	30.5	49.5	78.0	**87.0**	58.5	71.5	**88.5**
Yahoo! Answers	No-attack	68.4	**69.3**	64.2	65.8	71.6	**71.7**	51.2	69.0	72.3	**72.8**	59.0	70.2	**77.7**	76.5	76.2
	GSA	19.6	20.8	**61.0**	49.4	27.6	30.5	30.0	**48.6**	24.6	30.9	39.5	**53.4**	31.3	41.8	**66.8**
	PWWS	10.3	12.5	**61.0**	52.6	21.1	22.9	30.0	**54.9**	17.3	20.0	40.0	**57.2**	34.3	47.5	**66.8**
	GA	13.7	16.6	**61.0**	59.2	15.8	17.9	30.5	**66.2**	13.0	16.0	38.5	**63.2**	15.7	33.5	**66.4**

Table: The classification accuracy (%) of various models on three datasets, with or without defense methods, on benign data or under adversarial attacks. NT: Normal Training.

- Under all three attacks, SEM achieves the best robustness on RNN and BERT models. In addition, the performance of SEM among models is more stable than that of IBP.

Experiments
Defense against Transferability

Attack	Word-CNN				LSTM				Bi-LSTM				BERT		
	NT	AT	IBP	SEM	NT	AT	IBP	SEM	NT	AT	IBP	SEM	NT	AT	SEM
GSA	45.5*	86.0	**87.0**	**87.0**	80.0	89.0	83.0	**90.5**	80.0	87.0	87.5	**91.0**	92.5	**94.5**	90.5
PWWS	37.5*	86.5	**87.0**	**87.0**	70.5	87.5	83.0	**90.5**	70.0	87.0	86.5	**90.5**	90.5	**95.0**	90.5
GA	36.0*	85.5	**87.0**	**87.0**	75.5	88.0	83.5	**90.5**	76.0	86.5	86.0	**91.0**	91.5	**95.0**	90.5
GSA	84.5	89.0	**87.5**	87.0	35.0*	87.0	83.5	**90.5**	73.0	85.0	86.5	**91.0**	93.0	**95.5**	90.5
PWWS	83.0	89.0	**87.5**	87.0	30.0*	86.0	85.0	**90.5**	67.5	85.5	86.5	**90.5**	93.0	**95.0**	90.5
GA	84.0	89.5	**87.5**	87.0	29.0*	88.0	83.5	**90.5**	70.5	87.5	87.0	**91.0**	92.5	**95.5**	90.5
GSA	81.5	**88.0**	87.5	87.0	72.5	89.5	84.0	**90.5**	40.0*	85.5	87.5	**91.0**	93.5	**95.5**	91.0
PWWS	80.0	**87.0**	**87.0**	86.5	67.5	87.5	83.5	**90.5**	29.0*	85.5	87.0	**90.5**	92.5	**95.5**	90.5
GA	80.0	**89.5**	87.5	87.0	69.5	88.5	83.5	**90.5**	30.5*	85.0	86.5	**90.5**	92.5	**95.0**	90.5
GSA	83.5	87.0	**87.5**	87.0	84.0	88.0	83.5	**89.5**	83.0	88.0	87.0	**89.5**	66.5*	**95.5**	90.5
PWWS	81.0	87.5	**88.0**	87.0	82.5	88.0	84.0	**91.5**	83.0	88.0	87.5	**91.5**	68.0*	**94.5**	90.5
GA	82.0	87.0	**88.0**	87.0	82.0	88.0	83.5	**91.0**	82.0	88.0	87.5	**91.0**	58.5*	**94.0**	90.0

Table: The classification accuracy (%) of various models for adversarial examples generated through other models on *AG's News* for evaluating the transferability. * indicates that the adversarial examples are generated based on this model.

Experiments

Defense against Transferability

Attack	Word-CNN				LSTM				Bi-LSTM				BERT		
	NT	AT	IBP	SEM	NT	AT	IBP	SEM	NT	AT	IBP	SEM	NT	AT	SEM
GSA	45.5*	86.0	**87.0**	**87.0**	80.0	89.0	83.0	**90.5**	80.0	87.0	87.5	**91.0**	92.5	**94.5**	90.5
PWWS	37.5*	86.5	**87.0**	**87.0**	70.5	87.5	83.0	**90.5**	70.0	87.0	86.5	**90.5**	90.5	**95.0**	90.5
GA	36.0*	85.5	**87.0**	**87.0**	75.5	88.0	83.5	**90.5**	76.0	86.5	86.0	**91.0**	91.5	**95.0**	90.5
GSA	84.5	89.0	**87.5**	87.0	35.0*	87.0	83.5	**90.5**	73.0	85.0	86.5	**91.0**	93.0	**95.5**	90.5
PWWS	83.0	89.0	**87.5**	87.0	30.0*	86.0	85.0	**90.5**	67.5	85.5	86.5	**90.5**	93.0	**95.0**	90.5
GA	84.0	89.5	**87.5**	87.0	29.0*	88.0	83.5	**90.5**	70.5	87.5	87.0	**91.0**	92.5	**95.5**	90.5
GSA	81.5	**88.0**	87.5	87.0	72.5	89.5	84.0	**90.5**	40.0*	85.5	87.5	**91.0**	93.5	**95.5**	91.0
PWWS	80.0	**87.0**	**87.0**	86.5	67.5	87.5	83.5	**90.5**	29.0*	85.5	87.0	**90.5**	92.5	**95.5**	90.5
GA	80.0	**89.5**	87.5	87.0	69.5	88.5	83.5	**90.5**	30.5*	85.0	86.5	**90.5**	92.5	**95.0**	90.5
GSA	83.5	87.0	**87.5**	87.0	84.0	88.0	83.5	**89.5**	83.0	88.0	87.0	**89.5**	66.5*	**95.5**	90.5
PWWS	81.0	87.5	**88.0**	87.0	82.5	88.0	84.0	**91.5**	83.0	88.0	87.5	**91.5**	68.0*	**94.5**	90.5
GA	82.0	87.0	**88.0**	87.0	82.0	88.0	83.5	**91.0**	82.0	88.0	87.5	**91.0**	58.5*	**94.0**	90.0

Table: The classification accuracy (%) of various models for adversarial examples generated through other models on *AG's News* for evaluating the transferability. * indicates that the adversarial examples are generated based on this model.

- SEM is much more successful in blocking the transferability of adversarial examples than the defense baselines on RNN models.

Experiments

Defense against Transferability

Attack	Word-CNN				LSTM				Bi-LSTM				BERT		
	NT	AT	IBP	SEM	NT	AT	IBP	SEM	NT	AT	IBP	SEM	NT	AT	SEM
GSA	45.5*	86.0	**87.0**	**87.0**	80.0	89.0	83.0	**90.5**	80.0	87.0	87.5	**91.0**	92.5	**94.5**	90.5
PWWS	37.5*	86.5	**87.0**	**87.0**	70.5	87.5	83.0	**90.5**	70.0	87.0	86.5	**90.5**	90.5	**95.0**	90.5
GA	36.0*	85.5	**87.0**	**87.0**	75.5	88.0	83.5	**90.5**	76.0	86.5	86.0	**91.0**	91.5	**95.0**	90.5
GSA	84.5	89.0	**87.5**	87.0	35.0*	87.0	83.5	**90.5**	73.0	85.0	86.5	**91.0**	93.0	**95.5**	90.5
PWWS	83.0	89.0	**87.5**	87.0	30.0*	86.0	85.0	**90.5**	67.5	85.5	86.5	**90.5**	93.0	**95.0**	90.5
GA	84.0	89.5	**87.5**	87.0	29.0*	88.0	83.5	**90.5**	70.5	87.5	87.0	**91.0**	92.5	**95.5**	90.5
GSA	81.5	**88.0**	87.5	87.0	72.5	89.5	84.0	**90.5**	40.0*	85.5	87.5	**91.0**	93.5	**95.5**	91.0
PWWS	80.0	**87.0**	**87.0**	86.5	67.5	87.5	83.5	**90.5**	29.0*	85.5	87.0	**90.5**	92.5	**95.5**	90.5
GA	80.0	**89.5**	87.5	87.0	69.5	88.5	83.5	**90.5**	30.5*	85.0	86.5	**90.5**	92.5	**95.0**	90.5
GSA	83.5	87.0	**87.5**	87.0	84.0	88.0	83.5	**89.5**	83.0	88.0	87.0	**89.5**	66.5*	**95.5**	90.5
PWWS	81.0	87.5	**88.0**	87.0	82.5	88.0	84.0	**91.5**	83.0	88.0	87.5	**91.5**	68.0*	**94.5**	90.5
GA	82.0	87.0	**88.0**	87.0	82.0	88.0	83.5	**91.0**	82.0	88.0	87.5	**91.0**	58.5*	**94.0**	90.0

Table: The classification accuracy (%) of various models for adversarial examples generated through other models on *AG's News* for evaluating the transferability. * indicates that the adversarial examples are generated based on this model.

- On BERT, the transferability of adversarial examples generated on other models performs very weak, and the accuracy here lies more on generalization, so AT achieves the best results.

Experiments
Discussion on Traverse Order

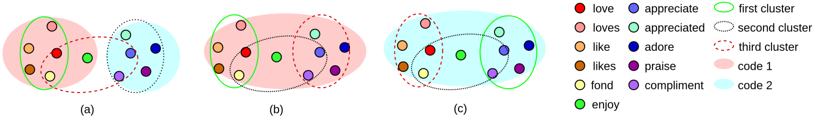
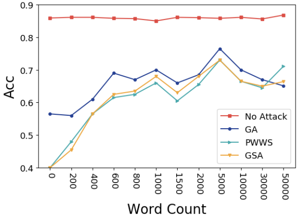


Figure: An illustration for various orders to traverse words at the 3rd line of *Synonym Encoding Algorithm* in the embedding space. (a) Traverse words first on the left, then on the right, then in the middle. The synonyms are encoded into two various codes (left and right). (b) Traverse words first on the left, then in the middle, then on the right. All synonyms are encoded into a unique code of the left. (c) Traverse words first on the right, then in the middle, then on the left. All synonyms are encoded into a unique code of the right.

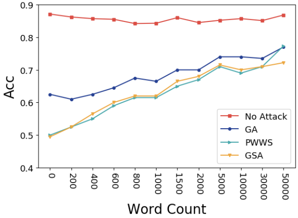
- The traverse order in the algorithm can influence the final synonym encoding of words and even lead to different codes for words in one synonym set.

Experiments

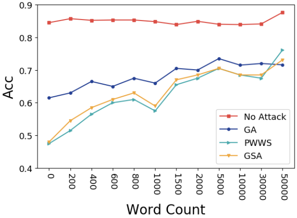
Discussion on Traverse Order



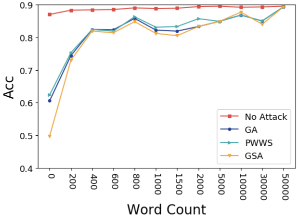
(a) Word-CNN under attacks



(b) LSTM under attacks



(c) Bi-LSTM under attacks



(d) BERT under attacks

Conclusion

We propose an adversarial defense method called SEM against synonym substitution based adversarial attacks in the context of text classification. SEM encodes the synonyms of each word to the same code and embeds the encoder in front of the input layer of the model to eliminate the word-level perturbations.

1. **Effective**. Compared with AT and IBP, SEM can remarkably improve model robustness and block the transferability of adversarial examples, while maintaining good classification accuracy on the benign data.
2. **Efficient**. Training with SEM is even faster than the normal training due to the reduction of encoding space. SEM is also easy to apply to large models and big datasets due to its simplicity.

Thank you!

Xiaosen Wang, Hao Jin, Yichen Yang and Kun He

School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China
Contact: xiaosen@hust.edu.cn
Homepage: `https://xiaosen-wang.github.io/`

[1] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani B. Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2890–2896, 2018.

[2] Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. Certified robustness to adversarial word substitutions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4120–4133, 2019.

[3] Volodymyr Kuleshov, Shantanu Thakoor, Tingfung Lau, and Stefano Ermon. Adversarial examples for natural language classification problems. *OpenReview submission OpenReview:r1QZ3zbAZ*, 2018.

[4] Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1085–1097, 2019.

[5] Xiaosen Wang, Hao Jin, and Kun He. Natural language adversarial attacks and defenses in word level. *arXiv Preprint arXiv:1909.06723*, 2019.

[6] Eric Wong and J. Zico Kolter. Provable defenses via the convex outer adversarial polytope. *International Conference on Machine Learning (ICML)*, 2018.