Defense on NLP

Enhancing Model Robustness By Incorporating Adversarial Knowledge Into Semantic Representation

Contribution

1. 很有意思，首先用图来衡量相似性，再用图的嵌入表示结合传统的分类流程，来提升模型对对抗样本的鲁棒性；

Notes

1. 文章算法：

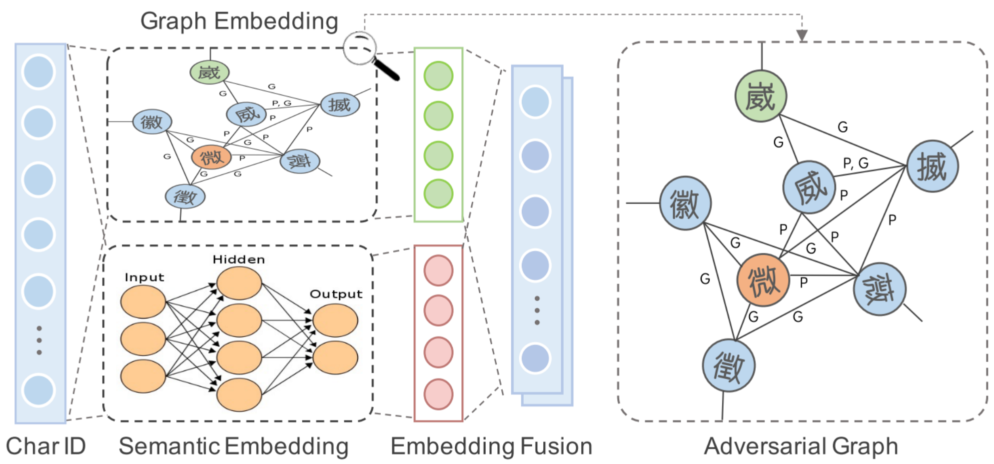


Fig. 2: The framework of our defense approach. The letters “P” and “G” in adversarial graph denote the phonetic-based and glyph-based variation relationship, respectively.

(1) 构建关联图：

- 通过拼音构建关联图：phonetic-based perturbations；
- 通过字形构建关联图：glyph-based perturbations；

  字形的相似性无法很好地直接构建，所以作者用一个自己的数据集（大小为10000，形式是三元组的形式）来训练了一个卷积神经网络 g-CNN 用来提取文字的图形特征表示，然后通过欧式距离来判断两个字形的相似性，网络结构如下：

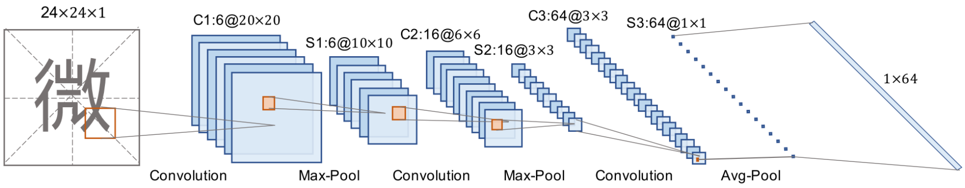


Fig. 3: Architecture of the glyph representation model.

模型训练的目标为，最小化：

$$\mathcal{L} = \sum_{i}^{M=10,000} [\|h(x_i) - h(x_i^+)\|_2^2 - \|h(x_i) - h(x_i^-)\|_2^2 + \alpha]_+$$

其中，$h(x_i)$ 是对文字 x_i 的隐藏表示，(x_i, x_i^+) 两者字形相似，而 (x_i, x_i^-) 两者字形相差很大。

(2) 图的嵌入表示：

- 作者通过 node2vec 来构建图的嵌入表示，（Skip-gram 思想）最大化目标函数：

$$\mathcal{L}(f, \theta) = \sum_{x_i \in V} log(\prod_{x_j \in N_S(x_i)} p(x_j | f(x_i)))$$

其中，f 指的是一个映射关系，θ 就是这个映射关系的参数，$N_S(x_i)$ 是文字 x_i 的相近字，是通过 BFS（breath-first sampling） 和 DFS （Depth-first sampling）两种方法采样得到的（我没有太理解这两种方法如何确定这个相似集合）；

(3) 文本的嵌入表示：这边仍然保持原有的文本嵌入表示方式，可以是 Word2Vec，也可以是 BERT；

(4) 融合：（看来还是需要数据的支撑，有数据好办事）

- 将图的嵌入表示和文本的拼接表示拼接到一起，训练一个下游分类模型：

$$\mathcal{F}(\boldsymbol{x_{adv}}) = \arg\max_{\hat{y}} \frac{e^{\mathcal{F}_{\hat{y}}(E_g(\boldsymbol{x_{adv}}) \oplus E_s(\boldsymbol{x_{adv}}))}}{\sum_{i=1}^{C} e^{\mathcal{F}_i(E_g(\boldsymbol{x_{adv}}) \oplus E_s(\boldsymbol{x_{adv}}))}}$$

其中，$\mathcal{F}_i$ 指的是第 i 个目标分类的概率；

2. 实验：

(1) 数据集：

- Douban Short Movie Comments (DMSC)
- Spam Advertisement (SpamAds)

(2) 测试攻击：

- TextBugger

(3) 结果：

- 正常情况的结果

Table 1: Model performance in the non-adversarial scenario. Avg-conf is the average confidence on correctly classified texts.

Model	Antispam		Sentiment Analysis	
	Accuracy	Avg-conf	Accuracy	Avg-conf
TextCNN	0.928	0.944	0.874	0.873
TextCNN+SC	0.920	0.936	0.864	0.867
TextCNN+AdvGraph	**0.928**	**0.962**	0.872	**0.898**
BiLSTM	0.893	0.894	0.851	0.849
BiLSTM+SC	0.886	0.887	0.845	0.844
BiLSTM+AdvGraph	**0.914**	**0.937**	**0.864**	**0.847**

- 对抗攻击下的结果

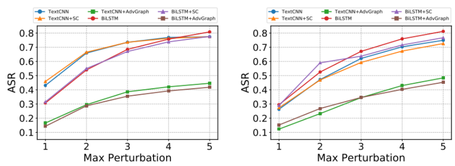
Table 2: Model performance on user-generated obfuscated texts.

Model	Antispam		Sentiment Analysis	
	Accuracy	Perturbation	Accuracy	Perturbation
TextCNN	0.630	1.23	0.669	1.16
TextCNN+SC	0.758	1.47	0.734	1.25
TextCNN+AdvGraph	**0.916**	**1.84**	**0.857**	**1.52**
BiLSTM	0.618	1.19	0.622	1.14
BiLSTM+SC	0.743	1.41	0.715	1.22
BiLSTM+AdvGraph	**0.898**	**1.79**	**0.839**	**1.49**

- 日常用户攻击下的结果

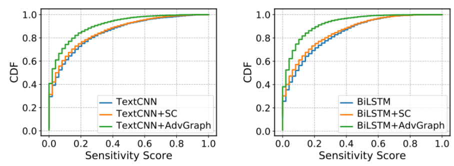
Table 3: The attack performance against all the target models under the adaptive setting.

Model	Antispam				Sentiment Analysis			
	ASR	Perturbation	Adversarial Similarity	Semantic Similarity	ASR	Perturbation	Adversarial Similarity	Semantic Similarity
TextCNN	0.769	1.63	0.917	0.874	0.703	2.07	0.911	0.832
TextCNN+SC	0.763	1.56	0.919	0.873	0.673	2.02	0.902	0.831
TextCNN+AdvGraph	**0.421**	**1.99**	**0.892**	**0.852**	**0.430**	**2.37**	**0.864**	**0.825**
BiLSTM	0.757	1.97	0.903	0.858	0.759	2.04	0.916	0.831
BiLSTM+SC	0.738	1.92	0.931	0.872	0.716	1.99	0.910	0.837
BiLSTM+AdvGraph	**0.392**	**2.00**	**0.872**	**0.843**	**0.403**	**2.10**	**0.855**	**0.814**



(a) Antispam (b) Sentiment Analysis

Fig. 4: The impact of maximum perturbation allowed on ASR.



(a) TextCNN (b) BiLSTM

Fig. 5: The model sensitivity against perturbations in antispam task.

Links

- 论文连接：[Li J, Du T, Liu X, et al. Enhancing Model Robustness By Incorporating Adversarial Knowledge Into Semantic Representation[J]. arXiv preprint arXiv:2102.11584, 2021.]