基于对抗训练的机器学习鲁棒性分析

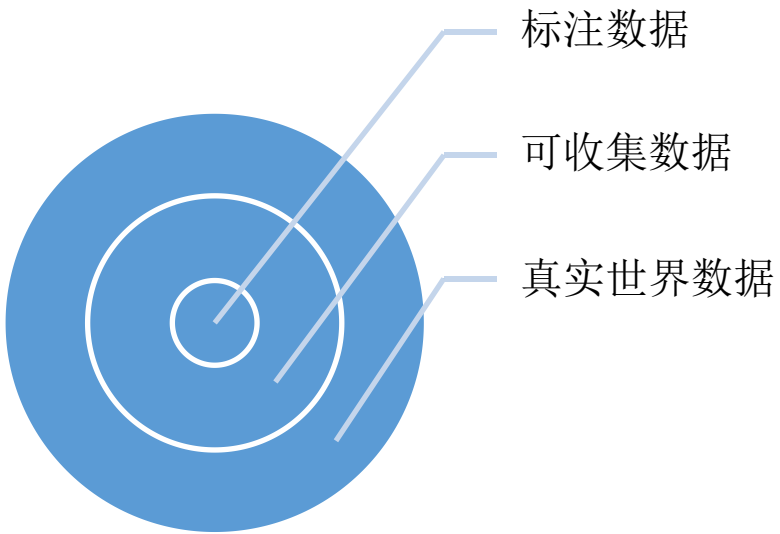
许晶晶
北京大学

报告内容

- 背景：过拟合与鲁棒性
  - 攻击样例
  - 攻击风险误差
- 挑战：自然语言中的攻击框架
  - 离散
  - 语义多样性
- 解决方案：
  - 基于对抗训练的强化训练方法
- 反思与思考

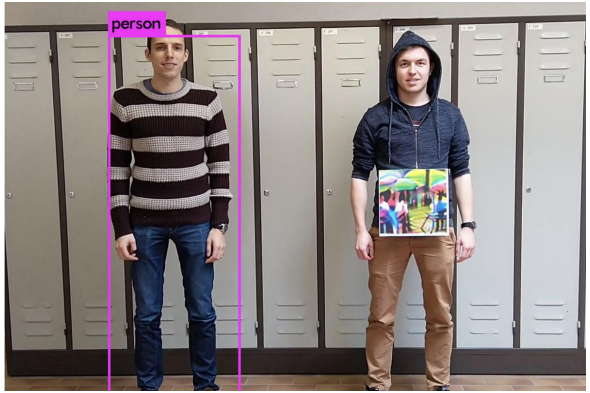
过拟合问题

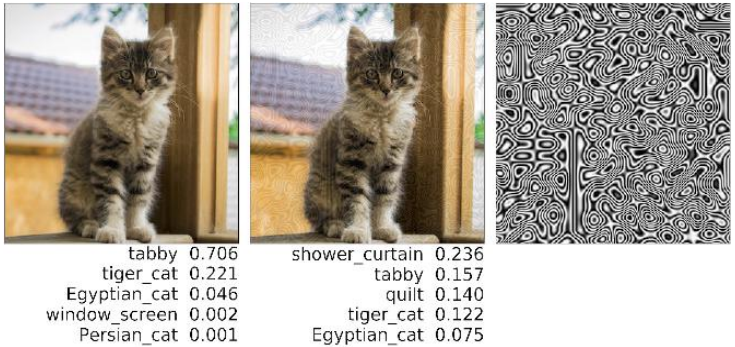
- 实验室环境 -》真实环境
- 训练数据有限
  - 质量高，数据少
- 真实数据无限
  - 噪音，表达多样化
  - 鲁棒性高的模型



过拟合问题

- 大模型：成千上万参数
- 小数据：有偏知识





如何评价一个模型足够鲁棒？

验证模型的鲁棒性: 攻击风险误差

- 无差别攻击
  - 学习加噪音干扰

$$\underset{\theta}{\text{minimize}} \frac{1}{m} \sum_{i=1}^{m} \ell(h_\theta(x_i), y_i)$$

正常训练目标：最小化loss

$$\underset{\delta \in \Delta}{\text{maximize}} \ell(h_\theta(x+\delta), y)$$

$$\Delta = \{\delta : \|\delta\|_\infty \leq \epsilon\}$$

在有限噪音范围内最大化loss

验证模型的鲁棒性：攻击风险误差

- 定向攻击
  - 学会加定向噪音

$$\underset{\delta \in \Delta}{\text{maximize}}\left(\ell(h_\theta(x+\delta), y) - \ell(h_\theta(x+\delta), y_{\text{target}})\right)$$

最小化目标label的loss　　最大化正确label的loss

验证模型的鲁棒性: 攻击风险误差

- 经验风险误差：

$$\hat{R}(h_\theta, D) = \frac{1}{|D|} \sum_{(x,y)\in D} \ell(h_\theta(x)), y)$$

- 攻击风险误差：更重要，更稳定

$$\hat{R}_{\text{adv}}(h_\theta, D) = \frac{1}{|D|} \sum_{(x,y)\in D} \max_{\delta\in\Delta(x)} \ell(h_\theta(x+\delta)), y)$$

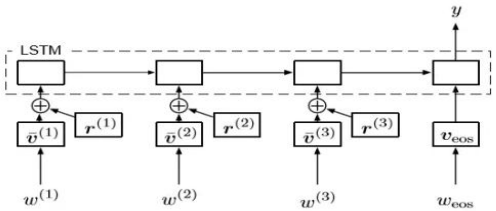
自然语言中的攻击问题

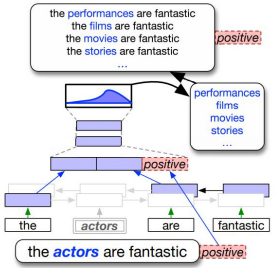
- 现有理论大都基于图像，图像可导，自然语言离散不可导
- 如何加离散噪音
  - 语义多样性
  - 语法多样性

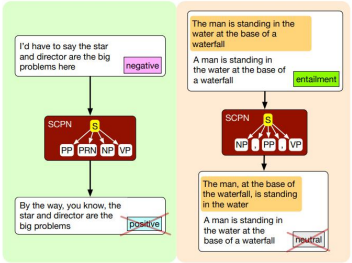
自然语言中的攻击问题

- **前人方法**
  - 数据增广
    - 随机噪音
    - 增加语法多样性
  - 对抗训练
- **面临挑战**
  - 需要人的知识参与
  - 低语义多样性

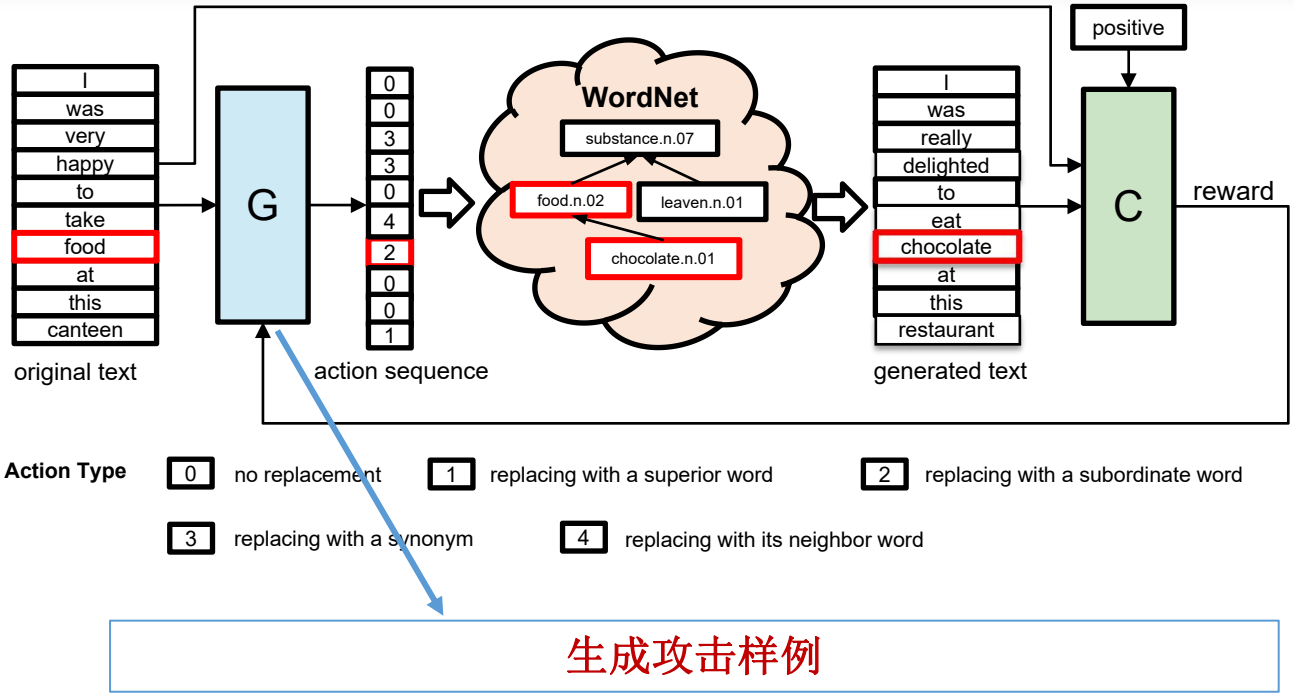
Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean



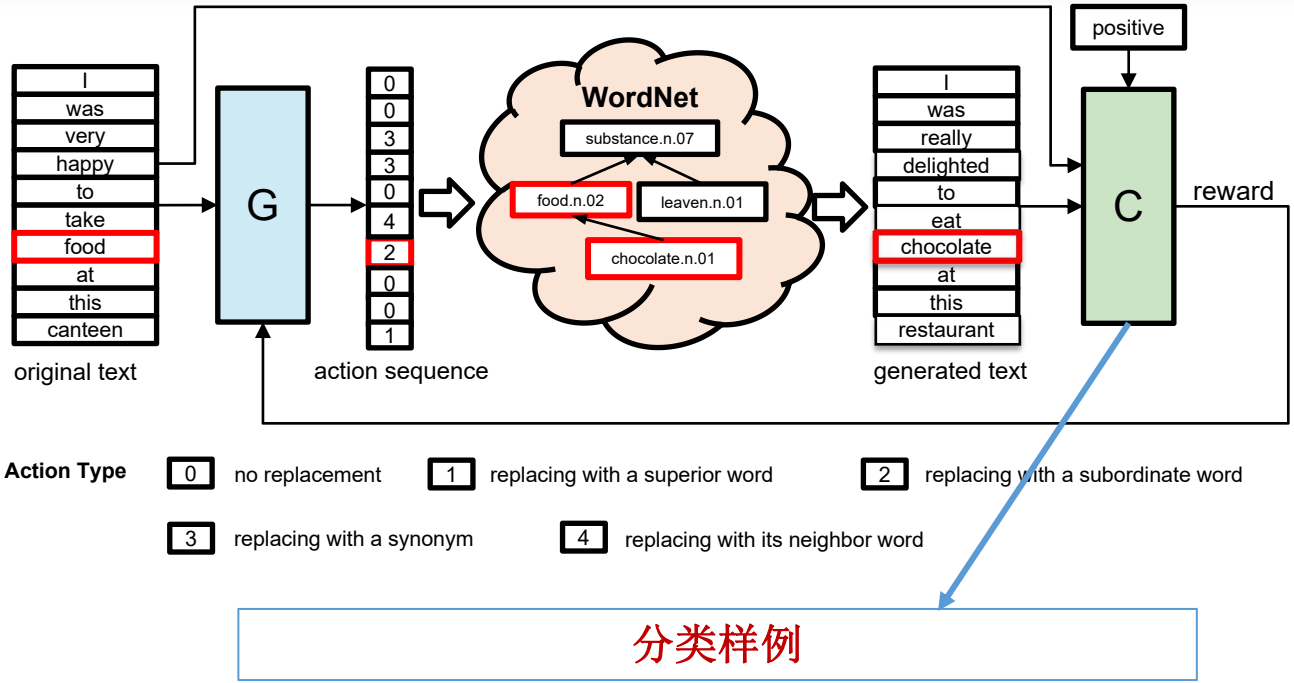




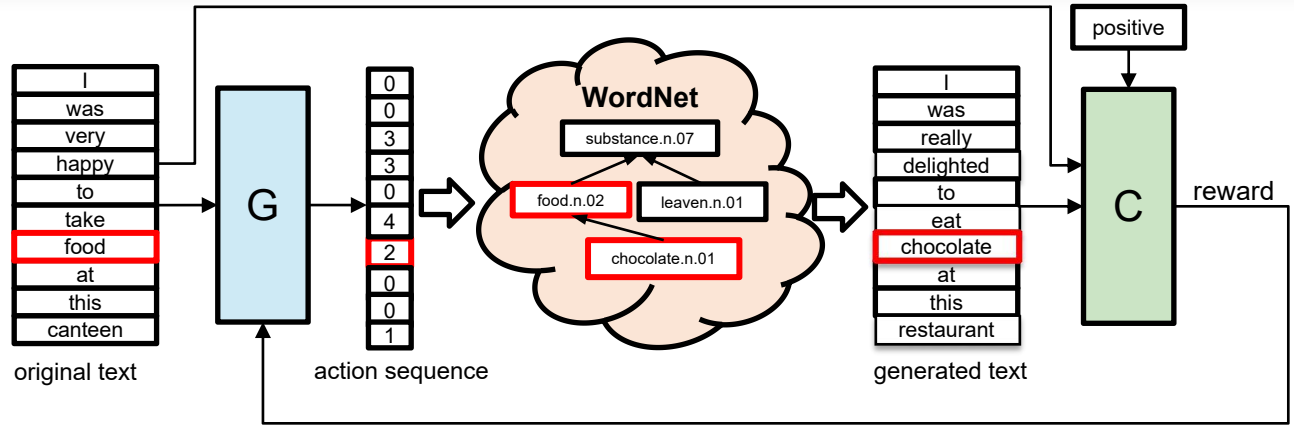
Lexical-Based Adversarial Reinforcement Training for Robust Sentiment Classification (EMNLP 2019)



Lexical-Based Adversarial Reinforcement Training for Robust Sentiment Classification (EMNLP 2019)



Lexical-Based Adversarial Reinforcement Training for Robust Sentiment Classification (EMNLP 2019)



Action Type

0 no replacement

1 replacing with a superior word

2 replacing with a subordinate word

3 replacing with a synonym

4 replacing with its neighbor word

$$r(\hat{a}) = logp_C(y|x;\emptyset) - logp_C(y|\hat{x};\emptyset)$$

反馈梯度

Lexical-Based Adversarial Reinforcement Training for Robust Sentiment Classification (EMNLP 2019)

- 数据集
  - 情感分类
- 分类器
  - CNN, RNN, BERT

Dataset	#Class	Avg. #w	Train	Dev	Test
SST2	2	19	6,920	872	1,821
SST5	5	18	8,544	1,101	2,210
RT	2	21	8,608	964	1,089
Yelp	5	89	100,000	10,000	10,000

Table 3: Dataset statistics. "Class" is the number of pre-defined labels. "Avg. #w" is the average word number in the input text. "Train", "Dev", and "Test" represent the sizes of the training set, the development set, and the test set.

Lexical-Based Adversarial Reinforcement Training for Robust Sentiment Classification (EMNLP 2019)

测试集

Approach	SST-2	SST-5	RT	Yelp
RNN(our implemented)	80.61	40.54	75.85	60.94
RNN (Kobayashi, 2018)	80.30	40.20	*	*
+SynDA (Zhang et al., 2015)	80.20	40.50	*	*
+ConDA (Kobayashi, 2018)	80.10	41.10	*	*
+VAT (Miyato et al., 2017)	81.16	37.38	75.94	59.69
+LexicalAT (proposed)	**81.60**	**41.99**	**76.12**	**61.18**

Approach	SST-2	SST-5	RT	Yelp
CNN(our implemented)	80.62	40.81	75.85	60.77
CNN (Kobayashi, 2018)	79.50	41.30	*	*
+SynDA (Zhang et al., 2015)	80.00	40.70	*	*
+ConDA (Kobayashi, 2018)	80.80	**42.10**	*	*
+VAT (Miyato et al., 2017)	*	*	*	*
+LexicalAT (Proposed)	**81.58**	41.67	**76.22**	**61.86**

Approach	SST-2	SST-5	RT	Yelp
BERT(our implemented)	92.60	55.07	88.57	66.76
+SynDA (Zhang et al., 2015)	*	*	*	*
+ConDA (Kobayashi, 2018)	*	*	*	*
+VAT (Miyato et al., 2017)	*	*	*	*
+LexicalAT (proposed)	**93.03**	**55.38**	**88.68**	**67.50**

对抗测试集

SST-2	Classifier (RNN)	Classifier (CNN)	LexicalAT (RNN)	LexicalAT (CNN)
Test Set	80.61	80.62	81.60	81.58
RNN-Attacking Set	69.91	65.62	76.44	73.70
CNN-Attacking Set	68.81	68.04	74.62	76.28
SST-5	**Classifier (RNN)**	**Classifier (CNN)**	**LexicalAT (RNN)**	**LexicalAT (CNN)**
Test Set	40.54	40.81	41.99	41.67
RNN-Attacking Set	35.16	35.43	38.73	38.91
CNN-Attacking Set	34.98	36.60	37.65	38.96
RT	**Classifier (RNN)**	**Classifier (CNN)**	**LexicalAT (RNN)**	**LexicalAT (CNN)**
Test Set	75.85	75.85	76.12	76.22
RNN-Attacking Set	69.05	68.78	71.44	70.61
CNN-Attacking Set	62.90	61.89	69.88	71.17

Lexical-Based Adversarial Reinforcement Training for Robust Sentiment Classification (EMNLP 2019)

总结

- 提出了一种面向自然语言理解的，基于词法多样性的对抗训练方法。
  - 面向语义多样性跟离散性问题
  - 知识库有利于多样性
  - 强化对抗训练减少模型的对抗风险误差

思考与反思

为什么选择机器学习鲁棒性？

实验室-》真实环境中非常重要的问题

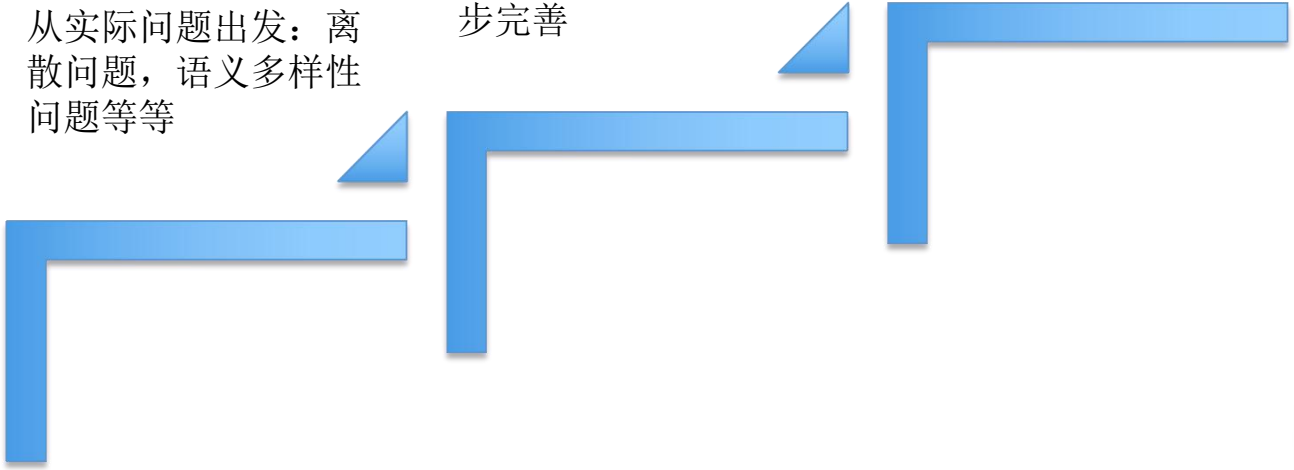
该怎么寻找研究点？

从实际问题出发：离散问题，语义多样性问题等等

该怎么设计实验方案？

不断迭代，最开始的方案效果也不好，逐步完善

产出



THANKS!

Q&A