CAT-Gen: Improving Robustness in NLP Models via Controlled Adversarial Text Generation

作者:

Tianlu Wang Xuezhi Wang Yao Qin Ben Packer Kang Lee Jilin Chen Alex Beutel Ed Chi

单位:

University of Virginia

摘要:

NLP模型已被证明受到了随机性问题的影响，即在对输入的微小扰动下，模型的预测很容易被改变。在这项工作中，我们提出了一种可控对抗性文本生成（CAT- Gen）模型，给定一个输入文本，通过已知与任务标签无关的可控属性生成对抗性文本。例如，为了攻击一个在产品re-view上进行情感分类的模型，我们可以使用产品类别作为可控属性，它不应该改变评论的情感。在真实世界的NLP数据集上进行的实验表明，与许多现有的对抗性文本生成方法相比，我们的方法可以生成更多样、更丰富的对抗性文本。我们进一步使用我们生成的对抗性文本，通过对抗性训练来改进模型，并且我们证明了我们生成的攻击对模型再训练和不同的模型架构更加强大。

1.论文试图解决什么问题

论文提出了一种新的生成对抗式文本的方法，这种方法利用文本生成模型来产生更多不同的、更丰富的输出。将语言生成限制在一定的可控属性范围内，从而得到语义上与输入句子接近的高质量输出。

输入一个文本x，将主要任务（比如文本分类）的标签表示成y，模型对x的预测表示成$f(x)$，将可以控制的属性表示为a。该方法的目的是在仅仅修改文本的无关属性，不去触动与任务标签有关的属性来生成更丰富、更流畅、和原句更加相近的文本x_1来创造出来对抗性攻击，使得$f(x) \neq f(x_1)$。

2.你的理解：论文中提到的解决方案之关键是什么?

- 该文章把文本的属性分成了主要任务（比如文本分类）的标签y和可以控制的属性a。这样属性训练a和任务训练y是分开进行的，且不需要平行的语料库用于训练属性。
- 本文的思路是改变属性a，只需要$a^{'} \neq a$即可，所以可以选择的新的可控属性类型有很多，这样就保证了对抗文本的丰富性。
- 该方法还在解码器后面计算交叉熵损失，这样可以通过不断调整优化来减小损失来确保对抗文本的流畅性。

3. 客观描述：论文中的实验是如何设计的?

1）Pre-training预训练:

- 用一个数据量大的数据集训练 encoder and decoder models（即编码器和解码器模型）。
- 用辅助数据集训练属性分类器

2）change of attribute属性变换：

在变换后的属性a'对应的编码是c'，属性损失可以定义为如上公式，公式中D是decoder，qA是属性a的条件分布，τ是临时变量，通过调整τ，汇表的分布趋于峰值，更接近离散情况。(这里不是太理解)

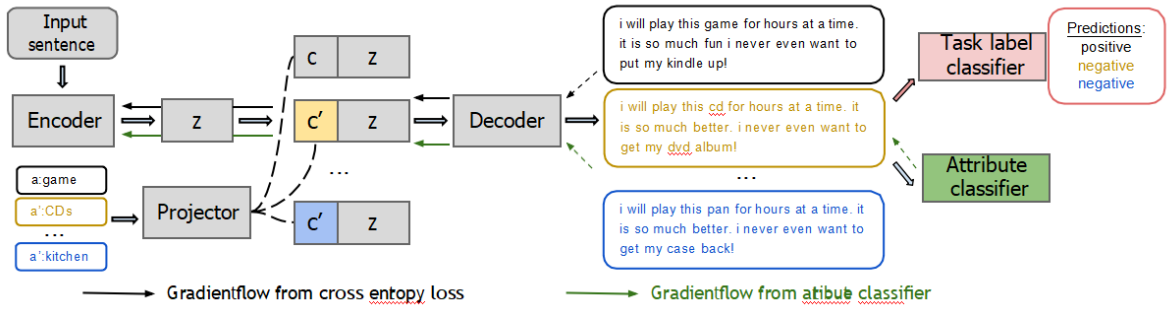
$$\iota_{c',z} = -E_{p(c')p(z)}[logq_A(c'|D_\tau(c,z))]$$

3）攻击优化：

在属性限制a'!=a 中的条件下可以生成很多对抗样本，将这些对抗样本输入到Task label Classifier中，将交叉熵损失最大的那个输入作为最终对抗样本

4）整体模型框架概括：

通过利用文本生成模型和在受控属性上的更大搜索空间，我们的模型能够生成比现有方法更多样化和流畅的对抗文本。我们的框架可以自然地扩展到许多不同的问题，例如，域转移(不同的域)，风格转移，以及公平性应用(例如使用不同的人口统计属性作为a)。



> 模型架构

4.用于定量评估的数据集是什么？代码有没有开源?

- 数据集是亚马逊评论的数据集。
- 代码没有开源（没有搜索到）。

5.论文中的实验及结果有没有很好地支持需要验证的科学假设?

有，该论文的主要假设或者观点是他们提出的方法产生的结果比现在的常规方法更加流利、多样和鲁棒。

- 流利性和多样性：

  用BLUE-4的分数衡量多样性；用预训练的模型计算perplexity的得分来评价句子流利度。对比其他的攻击方法，CAT-Gen模型生成更多样和流利的对抗样本。

		TextFooler (Jin et al., 2020)	NL-adv (Alzantot et al., 2018)	CAT-Gen
Diversity (BLEU-4 (Papineni et al., 2002), want ↓)		68.9	64.3	38.8
Fluency (in perplexity, want ↓)	Language Model 1	1853.7	964.3	729.5
	Language Model 2	1805.4	1188.5	868.7
	Language Model 3	336.7	479.9	358.9

- 鲁棒性：

过程：CAT-Gen攻击wordCNN并生成对抗样本；Re-traing wordCNN；wordCNN重新测试对抗性样本，得到准确率；训练一个wordLSTM模型，把在wordCNN中生成的对抗样本输入word LSTM模型中，得到准确率。

结果：两次实验的准确率，准确性最低意味着对抗样本具有很高的鲁棒性和可转移性。

	TextFooler (Jin et al., 2020)	NL-adv (Alzantot et al., 2018)	CAT-Gen
WordCNN re-training	84.7	82.9	49.3
WordLSTM	85.6	80.5	51.5

6.这篇论文到底有什么贡献?

本文提出了一个受控的对抗文本生成模型，该模型可以生成更加多样化和流畅的对抗文本，为真实世界的任务创建了更自然和有意义的攻击。另外，该框架的一个优势是灵活，可以合并多个与任务无关的属性，而且通过优化，模型还找出哪些属性更容易受到攻击。

7.下一步呢？有什么工作可以继续深入?

- 目前该框架的前提是可控的属性a是提前知道的，接下来还可以往自动识别这些属性方向努力。
- 仔细观察该模型的几个结果可以发现他的主要方法是替换原句中的某些关键词，将来还可以用别的方法，类似于Natural-GAN，不是替换词而是改写整个句子的方法来尝试。