可信机器学习论文报告

学号：10185102253　姓名：黄宇辰

论文主题

通过生成受控对抗文本提高自然语言处理模型的鲁棒性

Improving Robustness in NLP Models via Controlled Adversarial Text Generation

论文来源

作者信息：

Tianlu Wang[†*] **Xuezhi Wang**[§] **Yao Qin**[§] **Ben Packer**[§]
Kang Lee[§] **Jilin Chen**[§] **Alex Beutel**[§] **Ed Chi**[§]
[†]University of Virginia tw8cb@virginia.edu
[§]Google Research {xuezhiw, yaoqin, bpacker, kanlig, jilinc, alexbeutel, edchi}@google.com

会议出处：EMNLP2020

论文链接：https://www.aclweb.org/anthology/2020.emnlp-main.417/

一级方向：对抗攻击

二级方向：NLP

摘要阅读

自然语言处理模型大都受到鲁棒性问题的影响，也就是说模型的预测效果在输入的样本含有很小的噪声时也会受到很大的改变。这篇论文的研究者们提出了一个生成受控对抗文本的模型，这个模型可以基于与任务特征标签无关的可控属性去生成对抗文本。

例如，在对产品评论的情感分类的自然语言处理模型上，研究者们将产品类别视作不改变评论情感的可控属性。在真实的自然语言处理数据集上的实验表明，与已有的生成对抗文本的方法相比，这篇论文的方法可以生成更多样化、更加丰富的对抗文本。在论文实验中，研究者们进一步使用了生成的对抗样本进行对抗训练去改进自然语言处理模型。实验证明，本文中的受控对抗文本生成模型相对于再训练模型以及其他模型来说，具有更强的鲁棒性。

引言阅读

已有的研究表明，现有的自然语言处理模型，在随机初始化的情况下，数据未分离的情况下，遭受对抗文本干扰的情况下，都表现地十分易受影响。

研究如何提高自然语言模型鲁棒性的方向是通过在输入文本时或者在中间称述部分去生成对抗文本。然而，一方面，现有的对抗文本生成方法大都是试图在输入的文本上制造噪声，这样一般会导致生成的对抗文本缺乏多样性或者流畅性。另一方面，关注在文本称述部分去制造噪声经常会导致生成的对抗文本与原先的输入缺少相关性。在下表中，这篇论文的研究者们列举了目前已有的一些工作，去展示上面的已有方法的缺陷。

Method	Examples
Textfooler (Jin et al., 2020)	A person is relaxing on his day off → A person is relaxing on his nowadays off The two men are friends → The three men are dudes
NL-adv (Alzantot et al., 2018)	A man is talking to his wife over his phone → A guy is chitchat to his girl over his phone A skier gets some air near a mountain... → A skier gets some airplane near a mountain...
Natural-GAN (Zhao et al., 2018)	a girl is playing at a looking man . → a white preforming is lying on a beach . two friends waiting for a family together . → the two workers are married .

Table 1: Examples over existing adversarial text generation methods on SNLI (Bowman et al., 2015) dataset. Adversarial text generated by word substitution based methods (Textfooler & NL-adv) may lack fluency or diversity; GAN based methods (Natural-GAN) tend to generate sentences not related to the original sentences.

在这篇论文的研究中，研究者们旨在通过可控的属性去产生对抗文本。研究者们提出用文本产生模型去制造更加多样与流畅的输出文本。与此同时，他们让文本产生模型依靠可控的无关属性去使得输出文本与输入文本更加接近，质量更高。

形式化的来说，我们可以把输入文本看作变量 x ，主要任务的标签看作因变量 y(如对于一个文本的分类预测结果)，自然语言任务模型看作函数 f(x) ，把受控变量看作常数 a。这篇论文研究的目的就是创作一个新的文本作为对抗攻击文本，使这个新的对抗攻击文本在标签不变的情况下可以成功误导自然语言任务模型，以至于模型做出错误的预测。我的理解是，例如，把一个正向情感的评论，通过对抗文本生成模型得到一个新的评论，在这个评论仍然是正向情感的条件下，分类器对它进行识别分类后，前后的分类结果不相同。

为了实现上述目标，这篇论文的研究者们，他们提出了CAT-Gen，一个受控对抗文本产生模型。这个受控对抗文本产生模型包括了一个对于文本的译码器和一个对于文本的编码器，同时还包括了一个通过译码器去了解文本受控属性并改变这些受控属性从而产生新的对抗文本的网络模型。这个译码器和编码器在一个巨大的文本语料库上训练得来，因此它可以让我们得到一个更加流畅与多样的对抗文本。

在模型工作时，我们通过一个受控属性 a 去产生一个对抗文本。假设这个受控属性 a 是被预先识别出来的，而且我们知道这个受控属性 a 与任务标签 y 是无关的，那么这个属性就可以被作为学习的辅助资料。用这种方法，特征属性的训练和生成对抗文本的任务就可以被清楚的部署好。需要注意的是，研究者们的方法并不需要在学习模型的时候为辅助学习资料准备平行语料库。

研究者们在一个真实的自然语言语料数据上展示它们对抗文本产生模型的可行性。同样，研究者们也在实验中验证了他们的对抗文本更加的流畅，更加的多样，而且在再训练模型和其他模型上也表现出了更强的鲁棒性。

问题回答

Q1: 论文试图解决什么问题?

论文试图提出一种产生对抗文本的方法，使得对抗文本更加具有多样性而且更加具有流畅性。

输入的 x 是一个语料文本，输出的 y 是一个经过修改的对抗文本。对于 x 和 y并没有太多的限制。只是对于 x 而言，它需要拥有与原本标签无关的受控属性，因为论文的实验方案需要基于受控属性来完成模型的构建。

Q2: 论文中提到的解决方案之关键是什么?

在之前的自然语言模型鲁棒性研究当中，有一个工作就是通过生成对抗性文本去研究语言模型的鲁棒性。在生成对抗性文本的问题上，前人采用了通过词嵌入下的同义词替换单词的方法得到对抗性文本，还提出过使用生成式对抗网络，用在特定的语义空间内搜索产生对抗文本。还有研究者提出过在最小化损失的最坏情况内寻找单词的替换词组去产生对抗文本。近来，研究者们又提出了不直接生成对抗文本而采用将对抗性噪声加入词嵌入当中，并在输入文本上最小化对抗风险来产生对抗文本。

而本文中的解决方案，在受控属性上进行研究，基于在受控属性上的文本修改去生成对抗模型。与其他方案相比，这篇论文的解决方案可以适用于更广的领域。因为在不同的领域，文本当中受控属性都是不相同的。

Q3: 论文中的实验是如何设计的?

他们提出了CAT-Gen，一个受控对抗文本产生模型。这个受控对抗文本产生模型包括了一个对于文本的译码器和一个对于文本的编码器，同时还包括了一个通过译码器去了解文本受控属性并改变这些受控属性从而产生新的对抗文本的网络模型。这个译码器和编码器在一个巨大的文本语料库上训练得来，因此它可以让我们得到一个更加流畅与多样的对抗文本。

在亚马逊的评论数据集上测试模型以得到实验结果，观察把一个正向情感的评论，通过对抗文本生成模型得到一个新的评论，在这个评论仍然是正向情感的条件下，分类器对它进行识别分类后，前后的分类结果是否相同。

Q4: 用于定量评估的数据集是什么？代码有没有开源?

这篇论文的研究使用了Amazon Review数据集，包含10余个类别(电子、厨房、游戏、书籍等)。主要的测试任务是对评论进行情感分类，将不同的产品类别作为属性a。我们过滤掉分词数量超过25的评论。属性分类器根据每个类别的60000个评论集进行训练。属性训练数据也通过情感平衡，更好地分离属性和任务标签。研究方案使用另一个训练集(80000个正的和80000个负的)来学习情感分类器。研究方案提供了一个样本集和一个测试集，每个都有10,000个用于参数调优和最终评估的示例。

很遗憾，研究者们并没有将代码开源。

Q5： 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?

实验很好的证明了这篇论文的对抗产生模型有相当好的效果。

		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

Table 3: Comparison of our model with other methods. Evaluation is done over the attacks generated from the test set. Language model 1 & 2 are both from (Baevski and Auli, 2018), pretrained on Google Billion Words and WikiText-103 respectively; language model 3 (Ng et al., 2019) is pretrained on WMT news dataset.

由Table3，可以看到在对抗文本产生模型的产生文本的流畅度上，论文中的解决方案明显具有更好的流畅性。

	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

Table 4: Accuracy for various attacks over a re-trained model and a different architecture (want ↓). Note that the accuracy on the original model is zero since the evaluation contains a hold-out $1K$ set with only successful attacks.

由Table4，可以看到在LSTM和CNN训练模型上，论文中的解决方案产生的对抗文本都对自然语言模型具有很强的误导性。

	Original test set	TextFooler attacks	NL-adv attacks	CAT-Gen attacks
Original Training	91.9	84.7	82.9	49.3
+TextFooler (Jin et al., 2020)	92.7	89.5	88.6	52.7
+NL-adv (Alzantot et al., 2018)	92.2	86.4	94.6	51.2
+CAT-Gen	92.4	84.4	83.4	92.5

Table 5: We augment the original training set with adversarial attacks (rows) and evaluate the accuracy (want ↑) on hold-out $1K$ adversarial attacks (columns) generated by our method and two other baselines.

由Table5，我们可以看出在论文提出的基于受控属性对抗文本产生模型下经过再训练的模型，在其他经受其他对抗文本的干扰时，表现并不逊色，同时还在基于受控属性的干扰下鲁棒性远超其他模型。

Q6： 这篇论文到底有什么贡献?

这篇论文提出了一个基于受控属性的对抗文本产生模型。该模型可以生成更多样、更流畅的对抗性文本。通过实验也证明了论文中的对抗攻击模型下进行再训练产生的语言模型更具有鲁棒性，该对抗文本生成模型可以创建了更自然和更有意义的对抗文本。

Q7： 下一步呢？有什么工作可以继续深入?

这篇论文与我阅读的另外一篇论文都是在自然语言处理任务上进行鲁棒性和对抗攻击的相关研究。

另外一篇论文揭示了在语言模型的不同层级对不同文本内容具有不同的抗干扰能力。我想在修改受控属性而得到的对抗文本是否在不同层级考验着语言模型的鲁棒性。