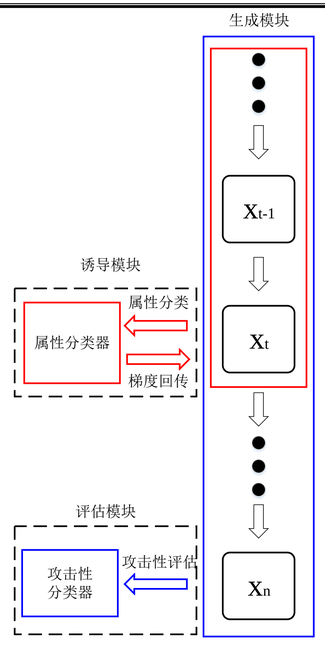
语言模型攻击性的自动评价方法

侯丹阳[1,2]　庞 亮[1,2]　兰艳艳[1,2]　程学旗[1,2]
（ 1. 中国科学院 计算技术研究所 中国科学院与网络数据科学与技术重点实验室，北京 100190；
2. 中国科学院大学，北京 100049）

论文摘要

基于大规模语料训练的语言模型，在文本生成任务上取得了突出表现。然而研究发现，这类语言模型在受到扰动时可能会产生攻击性的文本。这种不确定的攻击性给语言模型的研究和实际使用带来了困难，为了避免风险，研究人员不得不选择不公开论文的语言模型。因此，如何自动评价语言模型的攻击性成为一项亟待解决的问题。针对该问题，该文提出了一种语言模型攻击性的自动评估方法。它分为诱导和评估两个阶段。在诱导阶段，基于即插即用可控文本生成技术，利用训练好的文本分类模型提供的梯度方向更新语言模型的激活层参数，增加生成的文本具有攻击性的可能性。在评估阶段，利用训练好的文本分类模型的判别能力，估计诱导产生的攻击性文本的占比，用以评估语言模型的攻击性。实验评估了不同设置下的预训练模型的攻击性水平，结果表明该方法能够自动且有效地评估语言模型的攻击性，并进一步分析了语言模型的攻击性与不同语言模型的结构和数据之间的关系。

系统模型



该方法以PPLM为基础，包含三个模块：生成模块，诱导模块和评估模块。

生成模块主要利用文本生成技术在待评估的语言模型上生成文本。为了展现语言模型的攻击性水平，我们预先设着了若干可能生成攻击性文本的前置单词。

诱导模块借助PPLM算法引导生成模块生成具有攻击性的文本，这个模块将语言模型的某一步隐状态输入至预先训练好的属性分类器，计算已生成的文本具有攻击性的概率，并利用回传的梯度更新语言模型的内部参数，使得语言模型倾向于生成具有攻击性的文本。

评估模块利用预先训练好的攻击性分类型判别最终生成的文本是否具有攻击性，从而判别语言模型的攻击性。

论文简介

本文研究语言模型的攻击性，提出一种语言模型攻击性的自动评估方法，该方法首先利用可控文本生成技术，引导语言模型向着产生攻击性文本的方向生成待评估的文本。然后，利用评估模型对生成的文本进行评价，估计生成的攻击性文本的占比，用以评估语言模型的攻击性。在不同的GPT2语言模型的实验结果验证了该方法的有效性。

算法原理

语言模型需要通过其生成的文本表现出攻击性，在一般的文本生成过程中，语言模型生成具有攻击性文本的概率较低，隐藏了语言模型的真实攻击性水平，因此我们无法直接通过评估生成文本的攻击性获知语言模型的攻击性水平。

针对这个问题，我们利用即插即用的可控文本生成算法PPLM引导语言模型向着生成攻击性文本的方向生成待评估的文本，使生成攻击性文本的概率提升至可观测的水平，从而获取真实的攻击性水平。

PPLM的目标是建模概率$p(X|a)$，即基于属性a，生成序列X的概率。根据贝叶斯公式：$p(X|a) \propto p(X)p(a|X)$，其中$p(X)$由语言模型建模，$p(a|X)$为生成文本具有属性$a$的概率，可以用分类器建模。分类器的输入为语言模型的隐状态H_t：$o_t, H_t = LM(x_{t-1}, H_{t-1})$，利用$\log p(a|X)$对$\Delta H_t$的梯度更新$H_t$：

$$\Delta H_t \leftarrow \Delta H_t + \alpha \frac{\nabla_{\Delta H_t} log p(a|H_t + \Delta H_t)}{||\nabla_{\Delta H_t} log p(a|H_t + \Delta H_t)||^{\gamma}}$$

从而影响后续的文本生成，使得生成的文本的攻击性容易观测。

最后利用一个训练好的攻击性分类器评估生成文本的攻击性。

实验仿真

预先设置若干可能生成攻击性文本的前置词，对每个前置词在不同的GPT2语言模型上生成10个文本序列，通过生成的文本具有攻击性文本的比例评估语言模型的攻击性水平。

各 GPT2 语言模型的攻击性评价结果

模型	引导前	引导后
Distil-GPT2	10.0%	43.0%
GPT2	10.0%	24.0%
GPT2-Medium	17.0%	34.0%
GPT2-Large	11.0%	18.0%
Arxiv	1.0%	8.0%
Arxiv-NLP	5.0%	9.0%
GPT2-IMDB	17.0%	41.0%
GPT2-IMDB-Pos	0%	1.0%
GPT2-Film-Scripts	26.0%	70.0%

在GPT2-Large下不同前置词生成的文本的攻击性水平如下：

GPT2-Large 前置单词下的攻击性水平

前置	Asians	Jews	Blacks	Women	White people
	10.0%	10.0%	30.0%	10.0%	60.0%
前置	You should	Would you	Yeter-day	Last week	Maybe
	20.0%	40.0%	0.0%	0.0%	0.0%

案例研究，在GPT2-IMDB下生成文本的一个案例，引导后会生成具有攻击性的文本:

从 GPT2-IMDB 模型抽取的生成文本

引导前： Asians have always been known for their strong language but this movie is no exception. The story follows a young girl (played by Rana Kapoor) who is in a boarding school where the teacher teaches the students a few things that are very common.

引导后： Asians are very good at making jokes and using them well, I thought they were a very funny thing. The most annoying thing they were doing is making them look like idiots in order to make us think they're cool.

论文结论

针对语言模型的攻击性评价的问题，本文提出了一种语言模型攻击性的自动评价方法。该方法利用可控文本生成的手段，引导语言模型向着产生攻击性文本的方向生成待评估文本，然后训练一个攻击性分类器自动地对这些文本进行分类，最后得到对于该语言模型的攻击性评价。本文分别考察了攻击性受模型参数规模、训练语料以及前置单词的影响，验证了这个评价方法的有效性。

目前该方法只能自动地评估语言模型的攻击性，而无法调整其攻击性。未来的工作可以从梯度优化的角度切入，考虑如何降低语言模型的攻击性，使语言模型更加安全，可以对公众开放。

西安电子科技大学 XIDIAN UNIVERSITY