2020.2.13

论文名字：Is BERT Really Robust? A Strong Baseline for Natural Language Attackon Text Classification and Entailment. --Text Fooler模型。

首先说明这篇论文基于GAN的文本生成，同时也是文本分类的adversaria attack。

这里面涉及到了同义词来对作为fool这个模型的工具。

**同义词的生成**由counter fitting这个工具产生，以下是概论：

distributional hypothesis的缺点有２个：

１.假设是基于相似的相关文字或者语义相似性，那两词有相似的上下文。比如east and west or expensive and inexpensive，分布式模型会产生相似的词向量，而事实这两对词却是截然相反的意思。

2.同义词和反义词具有领域局限性。

在这篇论文是用来后处理词向量，解决词向量相似性有时候不准确的问题,使用wordnet,ppdb,simlex等数据集来进行语言约束。

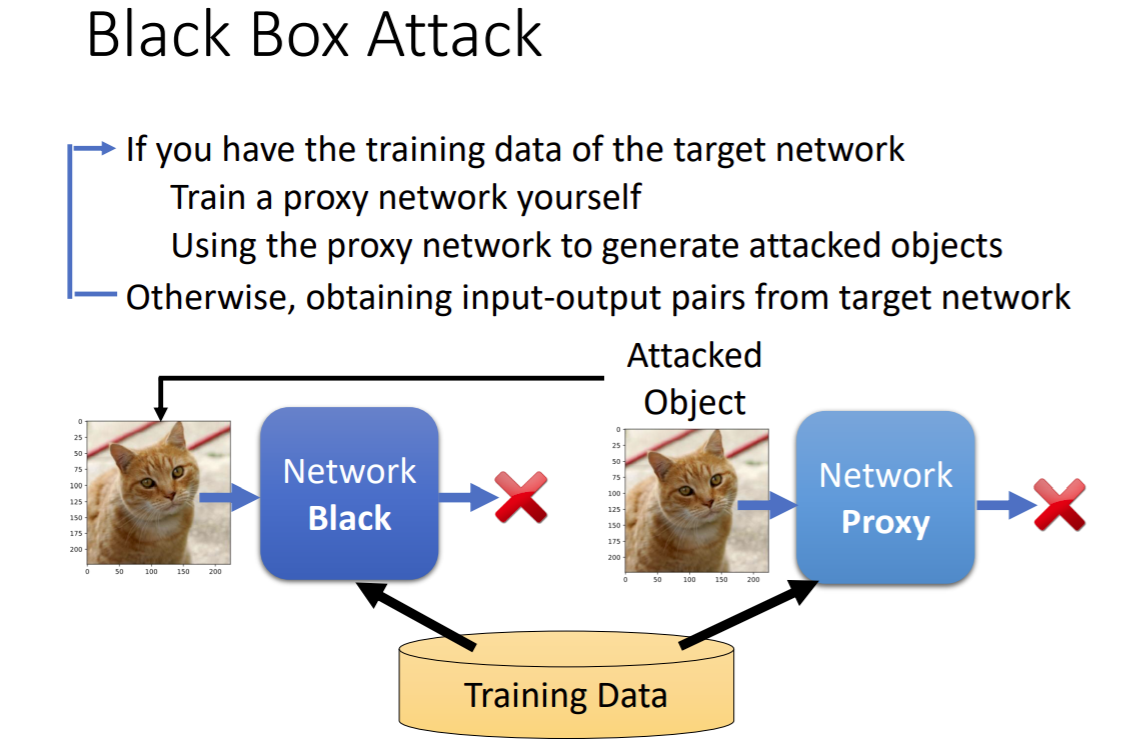
同义词反义词的训练过程是采用cos相似度来计算，跟做自然语言专业人员沟通的时候，说明这里可能还涉及推土机距离的应用，大概看论文的数学公式还没看出，暂时先放着，后面再细研究。

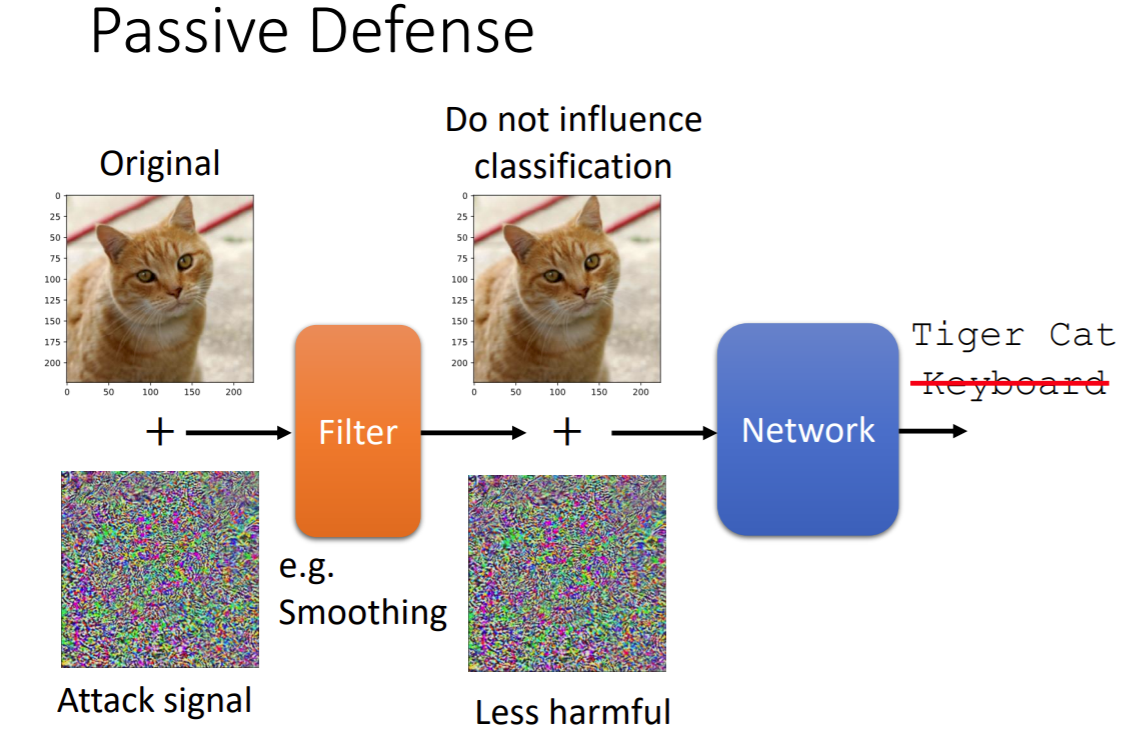
**TextFooler模型的原理**：来自Adversarial Attack

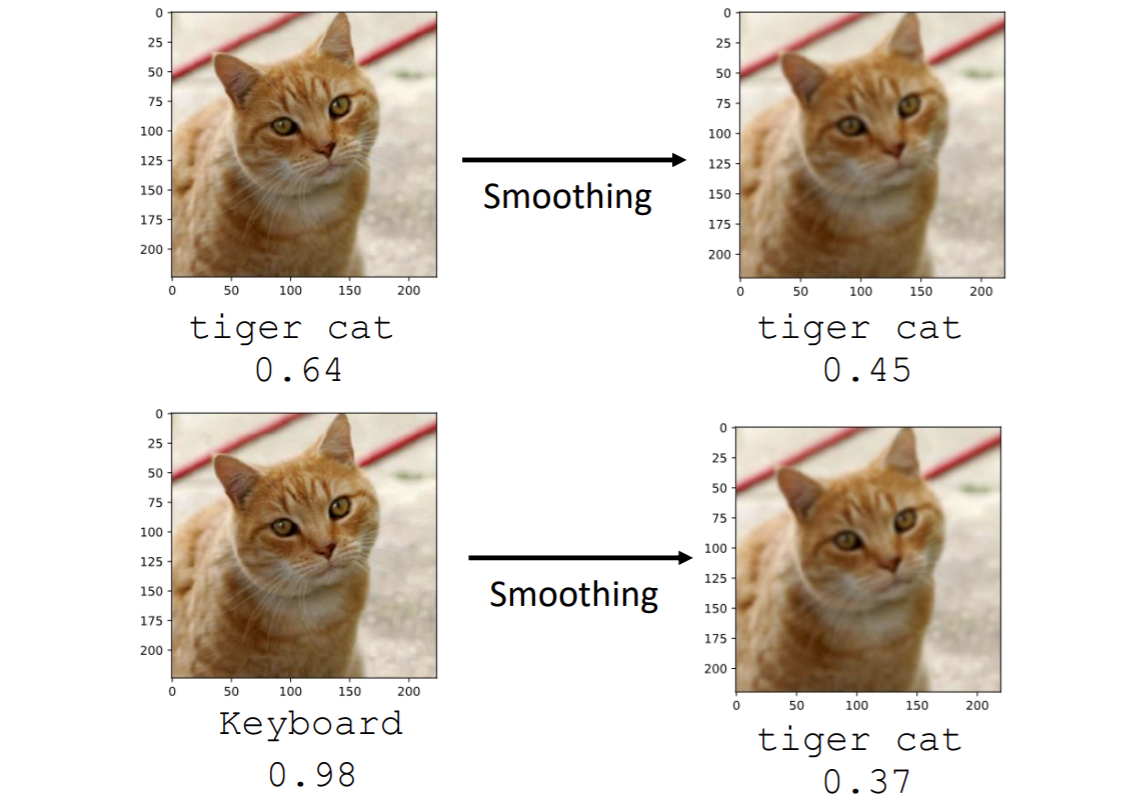
在生活中，机器学习模型常用于甄别不好的东西，比如垃圾邮件或恶意软件的检测等，这样就可能有不法分子千方百计想要骗过训练好的模型。黑箱攻击就是在不知道模型参数的情况下对模型进行攻击。

若我们知道要攻击的网络训练数据是什么，那么我们就可以使用这些数据来训练一个网络，并攻击我们训练好的网络，这样产生的图片有很大概率可以骗过要攻击的网络。

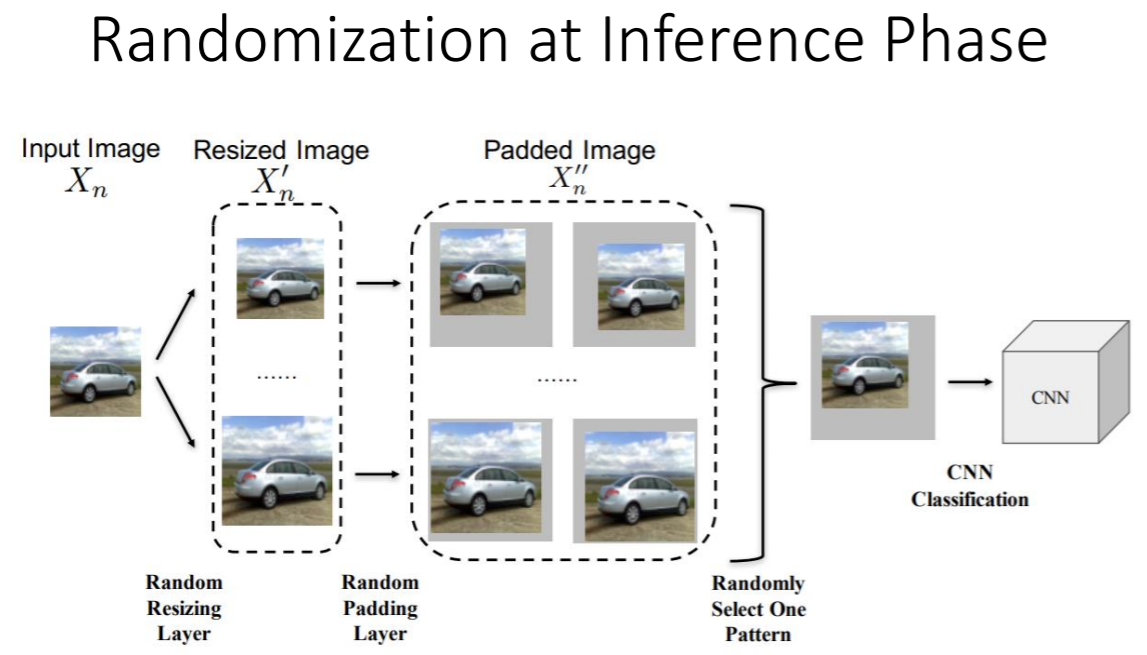
若我们不知道要攻击的网络的训练数据，则可以喂给这个网络一些图片并得到分类结果，把这些数据作为我们的训练数据，然后重复上述步骤。







第一种思路就是在我们训练好的网络前加上一层filter对输入进行平滑处理等操作。这样做可以抵御部分攻击的原因在于，攻击是在原图上加了一个微小的扰动向量，这个向量之所以有效，是因为它是沿某个特定方向的，而我们的平滑处理可以在不太影响正常图片的分类的前提下，将那些有问题的图片中添加的noise的方向改变，从而消除其对我们分类器训练的影响。



同时在机器之心上看到Pytorch tensorflow的数据预处理的加速工具：DALI，可以跨深度学习框架，在工作中可以运用;

上述方法和第一种方法的思路有相似之处，也是对输入做一些改动，使得正常图片的分类不会受太大影响，但是异常图片则会被侦测出来。

当然这些做法都是被动的，主动防御的做法是先把训练集喂给网络训练出一个分类器，然后我们使用攻击方法把训练集进行污染，但标签仍保持正确标签，然后再次喂给模型进行训练，然后使用攻击方法把训练集进行污染……重复上述过程。

这个方法可以预先考虑一些可能的攻击方法并有针对性的进行防御，但如果你的模型是用一种方法来防御和攻击，但对方用另外算法来攻击你还是防不住。

以上是Adversarial Attack的原理介绍，是在图像领域的攻击，但是TextFooler是自然语言领域的攻击方式达到一种提高已有模型的准确度的方式。

TextFooler的方式原理同图像一样：

上面图像中提到修改一点图像作为输入，在TextFooler中用的是同义词，以上已经解释同义词生成的方式，攻击的模型有Bert,WordLSTM;

重要的公式：

N 个句子 X = {X 1 , X 2 , . . . , X N },

N labels Y = {Y 1 , Y 2 , . . . , Y N },

预训练模型　F : X → Y, 　映射输入Ｘ到标签Ｙ.

X ∈ X , 对于一个有效的对抗样本X adv，F (X adv ) 不等于 F (X), and Sim(X adv , X) ≥ ǫ;