​ 虚假新闻的兴起迫使拥有社交媒体帐户的每个人都成为一名侦探，负责在发布前确定帖子是否真实。但是，虚假新闻仍然会越过我们的防线，在网络上迅速扩散，由于用户的无知和粗心而加剧。正如NBC新闻报道所显示的那样，假新闻不仅会散布恐惧和虚假信息，而且还可能对公司和个人的声誉造成损害。为了减少错误信息的直接和间接损失，我们需要更好的方法来检测虚假新闻。尽管有些虚假新闻是由真实的人撰写的，并且简直像是小说，但利用深度学习模型也可以大量生成虚假新闻，从而加剧了这一问题。到目前为止，计算机生成的文本已经很容易与真人写作的文本区分开。但是，由于自然语言生成模型的巨大改进，计算机生成的文本现在比以往任何时候都更加可信，因此这个问题变得更加紧迫。

假新闻无处不在-关键是如何制止它。

​ 在过去的三年中，Transformer（变形器）席卷了自然语言处理任务。 像几年前最先进的长期短期记忆体系结构这样的神经网络已经被BERT（来自变形器的双向编码器表示）和OpenAI强大的新模型GPT-2（生成预训练的变形器2）超越了 。 这些变形器现在以准确的分类和听起来不错的生成文本引领行业。 举例来说，由GPT-2生成的文本是如此真实，以至于OpenAI最初拒绝发布完整的受训模型，理由是“对技术和程序的恶意应用的担忧”。

​ 区分真实和虚假新闻很重要，但要解决一个难题，而对于新模型，这个问题变得更加困难。 该项目旨在探索假新闻产生和发现方面的挑战。

项目目标

通过比较长期短期记忆（LSTM）残差神经网络和OpenAI的最新变形器GPT-2生成的文本，简要演示自然语言生成模型的改进。

使用区分符尝试准确分类文本是由最新的深度学习模型还是由人生成。

项目概况

​ 该项目分为以上两个主要部分。 第二部分-查找，训练和使用判别器对生成的文本进行分类-这将是一项重大挑战，因为计算机生成的文本已经变得很难与人类书写的区分开。

​ 首先，对许多模型进行了研究和评估。 通常选择预训练的模型版本，因为它们无需进行数天的训练即可实现最新性能。 使用Kaggle数据集的子集对LSTM模型进行了微调。 用于微调的数据由《纽约时报》，《布赖特巴特》，CNN，《商业内幕》，《大西洋》，《福克斯新闻》，《谈话要点备忘录》，《 Buzzfeed新闻》，《国家评论》，《纽约邮报》，《卫报》，NPR，路透社 ，Vox和《华盛顿邮报》提供。 根据Kaggle上的文档，大多数文章来自2016年和2017年，而较少的文章来自2015年及之前。 OpenAI在GPT-2的文档中写道，它在针对特定内容的任务上表现良好，例如生成伪造的新闻，而无需进行微调或重新训练，因此项目使用了GPT-2的原始权重。

​ 为了检测假新闻，从真实新闻文章中传递了GPT-2模型的种子。 从该种子中，模型生成长格式的文本，最多可包含500多个令牌（可以将标记视为单词和标点符号。）。结果，对于每篇“真实”文章，都有一个生成的长格式文本，它们共享一个共同的第一句。 使用称为GLTR的工具，通过比较它们的功能来检查每个生成的文本和真实文章。 最后，将所有文章分为训练和测试集，并训练了BERT二进制分类器以对“伪造”文本进行分类。

第一部分：LSTM与Transformer

自从变形器问世以来，自然语言的产生难度已经大大减小

模型的选择和依据

​ 该团队探索了使用LSTM（长期短期记忆）架构和变形器架构的文本生成模型。 在2017年推出变形器之前，LSTM被认为是NLP任务的标准。被允许进行比较，并检查该领域取得的进展。

​ LSTM（长期短期记忆）是一种递归神经网络，它是一种试图对依赖于序列的数据进行建模的模型。 这使LSTM成为文本生成的可行候选者。 在LSTM体系结构中，LSTM单元块代替了标准神经网络层。 这些单元由输入门，忘记门和输出门组成。

​ LSTM模型由德国计算机科学家JürgenSchmidhuber和Sepp Hochreiter于1997年首次提出。 他们的出版物描述了他们所做的研究， 自1997年以来，LSTM有了新的改进，例如增加了遗忘门以及添加了架构中从单元到门的连接。

​ 该代码取自Keras的官方文档，该文档最初用于从尼采的作品中生成文本。 使用LSTM模型引起的问题是随机的，并且很难用固定随机种子来获得100％可再现的结果。 因此，尽管LSTM模型的结果很有趣，但是对于该项目的目标而言，它却并非有用。

​ 除了使用LSTM检查文本生成之外，该团队还使用来自OpenAI的名为GPT-2的变形器生成了文本。 自Vaswani等人发表学术论文“注意力就是你所需要的”以来。 在2017年，通过使用一种称为注意力的技术，变形器架构已经超越了先前模型的性能。 Miguel Romero Calvo在他在Medium上发表的这篇文章中很好地解释了注意力如何作用于变形器的编码器部分，以及编码器和解码器如何配合在一起构成变形器。 如果您已经熟悉这些体系结构，或者对学习如何使用它们的更多技术性知识不感兴趣，请继续阅读本文。 您无需应用这些概念即可了解项目和其结果。

​ OpenAI的GPT-2是代表从变形器模型生成最新文本的显而易见的选择。 该架构是专为文本生成而设计的，与BERT不同，它是Google AI语言研究人员的另一种著名转换器。 GPT-2也因其类似于人的性能而产生了很多新闻。

​ GPT-2模型是使用变形器的解码器模块构建的，并且像传统语言模型一样，一次输出一个令牌。 它使用字节对编码在其词汇表中创建令牌。 这意味着标记通常是单词的一部分。

​ GPT-2在800万个网页的数据集中进行了培训，总共大约40Gb的互联网数据。 根据OpenAI的说法，他们仅使用由人类策划/过滤的页面-具体来说，他们使用Reddit的出站链接，该链接至少获得了3业力。 这为模型提供了自然语言生成的好方法。

​ 该语言模型有一个简单的目标：给定文本中的所有先前单词，从而预测下一个单词。

​ GPT-2至少有四个版本：

“小”：1.24亿个参数，占用内存500MB

“中”：3.55亿个参数，占用内存1.5GB

“大”：7.74亿个参数

“特大”：15.58亿个参数，称为“完整”或“真实模型”，可能占用超过6.5 GB的空间

​ 在所有这些模型可用的情况下，需要权衡取舍：大型模型的总体性能较好，这意味着使用较大模型时它们会创建更可信的文本，但需要花费更长的时间进行微调来生成文本。 为了利用所谓的类人性能，同时仍具有足够小的模型来运行多个文本，选择了GPT-2的“中型”版本来代表模型的变形器。

性能比较

LSTM Sample生成的文本（种子文本为粗体）：

\*\*‘face reality as it is, not as it was, or as you would wish it to be,’ said the late \*, gadantural can chas and into the sussex and the prince and the salment for crisis. could state of the last week and and comparations and protective and the patient the royal of the cambridge the lead and pats for charles and cases and strip the first were from the royal and and raching her was she was known and the said of the grounds and and and the condition to according to free discoverns, and contact with the countries and one and the retarry report the monarchand self-inst stray days. health to the cases that she said the monarch strand self-may for the world the group and the lives.

在后期，gadantural可以追赶并进入苏塞克斯和王子和危机的薪水。 可以说明上周的情况，进行比较和保护，病人是剑桥王室成员，负责查理和案件的拍打，并剥夺了第一位来自王室成员的东西，并强奸了她，她是众所周知的，理由是 并根据自由发现的条件，与各国联系，并与一名退休人员汇报君主的流浪日子。 她说，君主处境自给自足，可以为整个世界的群体和生命带来健康。”

​ 给定文本的情况下的结果并不理想。 句子不连贯，有些单词似乎没有意义。 鉴于这是一个较早的LSTM，因此还使用了较新版本的基于RNN的生成器来生成示例文本。 这些改进的文本模型可更好地比较RNN的功能。

RNN示例生成的文本（种子文本为粗体）：

\*\*“ the university of texas at austin, \*. and, a group of reporters from the list of russian athletes, but also the same, as mr. trump has said he would not be able to get out of the way, but it is not uncommon for the past five years, ” mr. said. “ we ’ re not going to let me go to school and tell us that they were in a small town outside the old house, a team of researchers at the university of washington, said in a statement. “ we are not giving you to a college or a school. “ i didn ’ t want to do it, ” said michael, a lawyer for the trump organization, he said, “ i feel that i could have left him to the apartment, and the city ’ s house, and the president of the united states, said the government had been “ more ” than the evidence. mr. trump had been a target of the terrorist attack in syria. the country is about the size of the crowd”

德克萨斯大学奥斯汀分校，德克萨斯。 还有来自俄罗斯运动员名单的一组记者，但也与先生相同。 特朗普表示，他将无法摆脱困境，但在过去五年中并不少见。 说过。 华盛顿大学的一组研究人员在一份声明中说：“我们不会让我上学，告诉我们他们在老房子外面的一个小镇里。 “我们不是在给你上大学或学校。 “我不想这么做，”特朗普组织的律师迈克尔说，“他说，“我觉得我本可以把他留给公寓，城市的房屋以及联合总统的。 州说政府比证据“更多”。 先生。 特朗普曾是叙利亚恐怖袭击的目标。 这个国家大约是人群的大小

​ 这段文字虽然有些奇怪，但看起来更加连贯，并证明了基于LSTM的生成器可以在生成文字方面做得相当好。 相同的种子已传递给GPT-2生成器，以更好地进行比较。

GPT-2示例生成的文本（种子文本为粗体）：

The University of Texas at Austin is offering a hefty $60,000 scholarship for female robotics students who want to help their career. In the online portion of the program, participants will work with one of the university’s research teams and create an animated robot that can navigate a physical space.\*

The average college student wants to design a product and execute an idea and potentially earn a job. Often, this isn’t possible, so this type of program creates more jobs, says artist Aritappa Chatterjee, who came up with the concept for ‘Smart Robotics Lab.’

德克萨斯大学奥斯汀分校将为想要帮助其职业生涯的女机器人学生提供6万美元的巨额奖学金。 在该计划的在线部分，参与者将与该大学的一个研究团队合作，并创建一个可以在物理空间中导航的动画机器人。

普通大学生希望设计产品并执行想法并可能获得工作。 艺术家Aritappa Chatterjee提出了“智能机器人实验室”的概念，他说，通常这是不可能的，因此这类程序可以创造更多的工作。

奖学金适用于大二，初中或高三的学生。

“学生们正在确定这些女性占主导地位的领域，所以我们……”

​ 结果与预期结果相符：GPT-2创建了可信的短文本，文本越长，与真实文本的差异就越大。 尽管该模型能够保持一致的结构，但在仔细检查后，内容似乎牵强。

​ 该模型具有“温度”设置，可以在0到1的范围内选择一个超参数。温度越高，模型获得的“创意”就越多，这意味着它不会添加最可能的下一个单词，而会从更大的池中采样。 我们选择了0.7作为此超参数的值，这提供了看似最一致的内容创建。

​ 从示例中可以明显看出，GPT-2在生成文本方面比基于LSTM的模型表现更好。 此处选择的示例可以看作是“cherry-picking”，但是随机选择的，通常代表整体观察到的模式。

简单示例

​ 本文的标题是由GPT-2模型生成的，该模型在过去6个月中对来自arxiv.org的与机器学习相关的文章标题进行了微调。

生成标题

​ 引语是本文的开篇段落：“虚假新闻的兴起迫使拥有社交媒体帐户的每个人都成为侦探，负责在发布前确定帖子是否真实……”

​ 完整的输出为：“调查社交媒体上的虚假新闻：首先是一种用于打击错误信息的深度学习方法？” ，粗体文本保留为标题。

第二部分：假新闻检测

鉴别器

​ 鉴别器是一种深度学习模型，可对不同生成过程产生的样本进行分类（区分）。 鉴别器通常用于生成对抗神经网络（GAN），该网络与生成器协同工作以迭代地改进生成过程，使其输出越来越类似于“真实”示例。 出于识别GPT-2生成的文本的目的，一种相似类型的鉴别器很有吸引力，但是所使用的过程并不涉及反复训练生成器和鉴别器。 此外，人类很难识别生成的文本，因此人们对这种模型的潜在性能持怀疑态度。

​ 为了从不同的角度解决问题，如果生成的文本和真实的文章共享相同的主题，那么对生成的文本和真实的对应词之间的相似性进行度量可能会为生成的文本的特征提供一些信息。 用于基于“相似度”比较文本的第一种方法是一种称为bertScore的改进量度，该度量计算两个句子中单词的余弦相似度。

​ 尽管从表面上看，这种实现方式似乎很有希望，但却不能说明所使用的生成模型所展现的“创造力”。 尽管确实可以计算相似度并进一步分析相似性，但是生成的文本与原始文本有很大的不同。 从一条有关一名新奖学金获得者的新闻中摘录的一句话，可能会产生一个有关年轻聪明的策划者抢劫珠宝的文字。 事实证明，在不同的情况下评估相似性无助于提供洞察力，因为该指标仅报告所生成的文本与原始文章有很大不同。

​ 在评估这种相似性度量时，观察到生成的文本倾向于使用不太复杂的词汇，并且通常依赖于重复短语。 相比之下，原始新闻更倾向于带有逻辑故事情节，并传递各种主题和思想。 这一发现促使人们对诸如巨型语言模型测试室之类的工具进行了更多研究，这些工具似乎也取决于生成文本的复杂性和多样性。

​ 巨型语言模型测试室（GLTR）是由哈佛大学NLP与MIT-IBM Watson AI Lab合作创建的。该工具使用了BERT和GPT-2模型，可以逐字逐句查看该单词在句子中接下来被选中的可能性。在下面的图片中，绿色表示最有可能出现的单词中的前10个单词，黄色表示最常见的单词中的前100个，红色表示最常见的单次中的前1000个单词，紫色则表示1000后的单次。这表示绿色单词可能是模型输出的下一个单词，并且随着该单词变得不太可能被模型选择，它会落入其他颜色容器之一中。例如，以下图片是人类撰写的《纽约时报》文章的摘要。本文有几个可预测的词，但也包含许多生成模型将不太可能选择的词。这是因为人类在写作时不会考虑最可能出现的下一个单词。他们考虑哪个词最适合写作的上下文以及他们试图传达的想法。

​ 下面的四个图像显示了单词分布的差异。 当将GLTR工具与GPT-2鉴别器一起使用时，虚拟文章（左上）比其真实对应词（右上）具有更高的可预测单词分布。 在虚拟文章（左下方）和真实文章（右下方）上使用BERT鉴别器时，发现了类似的模式。 使用这两种区分器，在生成器和人类创造的结果中这两个词之间有明显的区别。 这支持了我们的推论，即生成器将仅基于单词的概率而非上下文来构造文本。 尽管该项目没有开发出用于确定文本是否由计算机生成的数值度量，但是查看GLTR的输出可以提供一些见解。

​ 根据上述限制，寻求第二种鉴别器。在研究假新闻鉴别器时，一些作者分享了使用BERT编码器和某种分类器的成功经验。选择BERT编码器和BERT二进制分类器的组合是因为它们具有输出/输入兼容性，并且在Python中实现了分类器。尽管Medium文章的原作者对该分类器表示了很高的期望，但是该模型的计算很费时间（估计需要花费15-21个小时才能运行一个流程），并且在该项目的实验运行中，无法对“真实”分类并以比简单猜测更好的速度生成文本。我们的实现可能太简单了，在传递给BERT编码器之前，应该在输入上包含其他工程。或者，此功能增强在BERT编码器和BERT二进制分类器之间可能很有用。但是，为该项目所做的工作没有提供帮助，表明这种额外的努力可能会产生所需的准确性。

​ BERT分类器的缺点表明了最新的自然语言生成模型的出色性能。 鉴别器有可能依赖于复杂性和创造力的度量，以及整体连贯性的度量。 像GPT-2这样的模型可能会欺骗鉴别器，该鉴别器只是为最有可能被选中的前10名或前100名中的单词设置频率的阈值。 可以通过提高模型的温度来做到这一点，让模型有更多的自由来探索鉴别器可能认为“不太可能”的词和概念。 但是，如果分类器也可以衡量整篇文章的连贯性，那么它会更容易地识别出所生成的那些文本，因为更多的“创意”文本往往对阅读它们的人来说意义最小。

未来可能使用的领域

创建一个针对复杂性/创造力的评分指标

创建使用此指标的假文本分类器

进一步检查“假文本”的特征

使用白盒对抗攻击（white box adversarial attack）来观察哪些功能使文本更容易被伪造