AI 已杀入密码界，密码攻坚不再是人类的专利！

秦曾昌、 small_TAO

密码，不只是你打开手机时输入的那几个数字，它还关系到你银行里的存款、电脑里的裸照，甚至，世界和平。

电影《模仿游戏》（The Imitation Game）讲述的就是关于著名密码系统恩尼格玛（Enigma）的故事。这里不讨论电影中情节的真实性，但二战时纳粹德国正是利用了这套密码系统，隐秘而高效地传递着军事情报。恩尼格玛的最终破译成功地扭转了战局。人们普遍认为，它的破译使盟军在西欧的胜利提早了两年。

电影《模仿游戏》的海报。图片来源：wikipedia

一直以来，设计和破解密码都是人类的专利。然而随着密码学理论的提升与计算机能力的增强，现代的密码变得越来越复杂，人们开始寻求让机器替代人类的办法。不过这就涉及到一个问题：用 0 和 1 思考的“机器大脑”能学会对信息进行加密吗？

在谷歌大脑（Google Brain）的最新的研究成果《让对抗神经网络学习保护通信》（*Learningto Protect Communications with Adversarial Neural Cryptography*）中，人们就试图教会机器加密与解密信息[1]。这次，思考密码术的不再是人类的大脑，而是“神经网络”与“生成对抗网络”（Generative Adversarial Network）结合而成的机器之“脑”。

神经网络，生成对抗网络与密码术

神经网络

神经网络全称人工神经网络，是一种模仿动物神经系统结构和功能的计算模型。在经历过历史中的几度沉浮后，如今它已成为科研界与工业界的新宠，在人工智能及相关领域中炙手可热。

神经网络由大量的“感知机”（Perceptron）相互连接构成。感知机类似于生物神经系统中的神经元，是神经网络中最基本的单元。

神经网络并非生来就具备强大的功能，它也需要训练才能掌握技能。比如我们希望神经网络通过西瓜的外形判断瓜的甜度，一开始神经网络并不懂如何去判断，这时就需要分别把西瓜的外

形和对应的甜度分别输入神经网络，以训练它去学习两者之间的对应关系。训练神经网络的过程实际上就是通过学习数据来调整每一个感知机参数的过程。神经网络读取数据样本后，感知机们会先根据现有模型参数进行计算，然后把输出的值与真实值进行比较，再将两者的差距反馈回去，以调整参数。经过反复多次“计算－比对－反馈－调整”的循环后，神经网络就能够准确地判断瓜甜还是不甜了。

生成对抗网络

很多时候训练数据的真实结果信息难以获得——比如不能把每个瓜切开尝尝。生成对抗网络利用模块间的对抗，巧妙地避开了这个问题。

生成对抗网络中主要有两个模块：负责生成的模块 G 和负责判别的模块 D。我们用模仿画作的例子来说明两个模块的作用。G 是一位初出茅庐的画家，想要通过模仿名画的来提升自身能力。在每次模仿名画之后，G 画家会将自己的赝品与真品一同送给鉴定师 D。D 的主要任务便是鉴定送来的画哪一幅是真品哪一幅是赝品。刚开始，G 画家的水平一般，D 鉴定师能够很轻松的鉴定出真伪。随着 G 画家的模仿水平的提高，D 鉴定师无法分辨真伪。这个时候，我们便可以说 G 画家的模仿水平相当的优秀了。这也是我们想要的结果，模仿能力卓绝的生成模块 G。

人类的密码术

说完网络，再说说密码术。在密码术中，能够直接代表原文含义的信息称为明文；经过加密处理之后隐藏原文含义的信息称为密文。加密与解密便是明文与密文相互转换的过程，而密钥是用来加密与解密的工具。密钥好比一本双语字典，你既可以用它把明文翻译成密文，也可以通过它查找密文所对应的明文是什么。在信息保密过程中，密钥的安全格外重要。因为找到秘钥就是找到了加密和解密的方法，密码也就迎刃而解。早期的密码设计有替换法与位移法。替换法就是有规律的使用一组字母来代替原有字母，例如每个字母用上一个字母取代，“abc”替换为“zab”；移位法就是将字母顺序重新排列，例如“Key”变成“yeK”。这样的密码可以分别使用穷举法和统计法进行破解。后来密码的书写发生了手写到机器书写的转变，这也使得密码的编写变得千变万化。

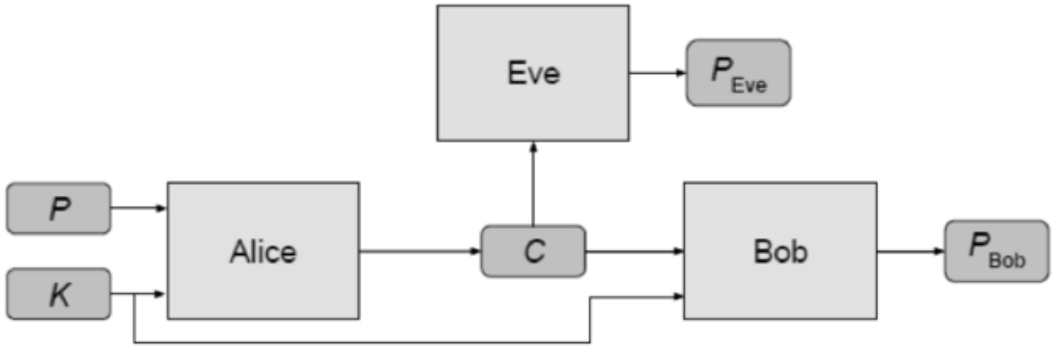
不过，这是人类的密码术。作为机器之脑的神经网络也能“想”出类似的技术对信息进行加密吗？

机器能学会加密信息吗？

回到谷歌大脑最新的研究上。研究者向神经网络 A 中投入明文和密钥数据，它的计算结果会作为密文，与密钥一起交给另一个神经网络 B，并由 B 进行解密。而 A、B 组成的生成对抗网络，则会试图在对抗中使解密出的数据趋近原始明文。当然，整个过程中神经网络并不懂“明文”“密文”的概念，因为研究人员丝毫没有向它们透露人类密码术的相关知识。它们只知道自己收到了数据，又要输出数据。研究者通过这样的方式，探究神经网络能否自己“思考”出机器的密码术来实现对信息加密、解密的功能。

实际操作中，研究者设计了一种通用的保密情况，叫做对称加密模型。对称加密是指沟通双方有公共的密钥，而窃听者没有。研究者在该模型中加入了三个独立的神经网络模块，分别取名为 Alice、Bob 和 Eve。这三个神经网络模块共同构成了对抗神经网络的主体。

我们可以把它们想象成是三个人：Alice 想和 Bob 进行秘密沟通，而 Eve 想要窃取他们的通信。为了防止秘密被 Eve 知道，Alice 制定了密钥并共享给 Bob。通信的时候，Alice 首先通过密钥将信息进行加密，然后将加密后的信息发送出去。这时 Bob 和 Eve 都能接收到信息，不同的是，Bob 可以通过共享的密钥对信息解密，而 Eve 需要自己想办法猜测信息的内容。

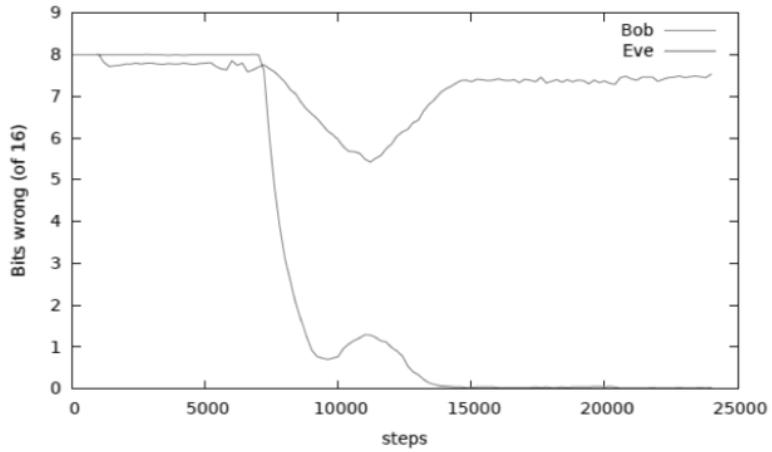


方框分别代表着三个独立的神经网络，P 是明文，K 是密钥，C 是密文，PBob 和 PEve 分别是 Bob 与 Eve 对密文的预测结果。 参考文献：arXiv:1610.06918

如图所示，K 是 Alice 和 Bob 共享的密钥，P 便是要进行加密的信息。C 是 Alice 转换出的密文，它会被 Eve 和 Bob 收到并进行解密，两人各自解密得到的结果则为 PBob 和 PEve。研究者通过统计 PBob 和 PEve 中与 P 相同的数据有几位来计算解密结果的准确性。

当然，我们希望 PBob 与 P 越接近越好，而 PEve 与 P 值差距越大越好。这表示通过 Bob 解密的信息是准确的，而窃听者难以通过 Eve 得到被加密的信息。

基于以上这些设置，研究者开始对这个加密模型训练。每一次的训练过程中，神经网络 Alice 会接收到 4096 套 K 和 P 作为训练样本。K、P 都是 16 位二进制数据，Bob 和 Eve 输出的解密文件也是 16 位二进制数据。通过统计每次训练后 Bob 和 Eve 解密结果的与明文 P 的对应数位的数字，研究人员得到了如下图的结果：

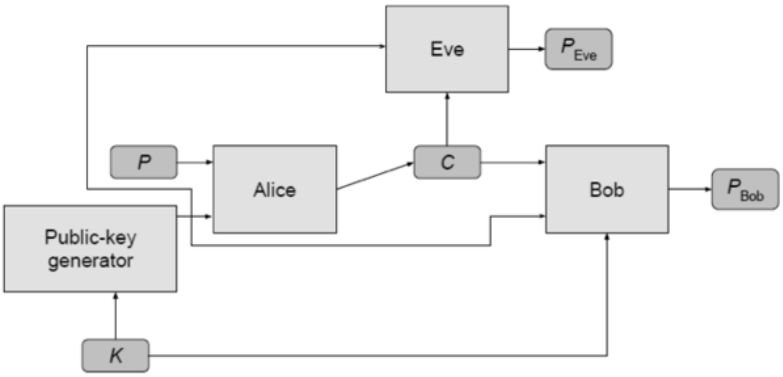


Bob 和 Eve 解密信息的错误率随实验次数的变化曲线。横轴是训练次数，纵轴是错误的字节数。红线与绿线分别代表着 Bob 与 Eve 错误字节数的变化情况。在进行到 10000 次左右的时候，Bob 迷之反击了 Eve。参考文献：arXiv:1610.06918

训练刚开始的时候，Bob 和 Eve 的错误位数都是 8，对于一个 16 位的二进制文本来说，随便生成的 16 位数字在概率上错误位数就是 8——这时 Bob 和 Eve 都还没有具备解密的能力。伴随的神经网络之间的对抗，每个神经网络不断的调整其中感知机参数的数值，Alice 和 Bob 也逐渐获得了极强的加密与解密能力。当迭代进行到 7000 次的时候，Bob 预测明文的错误率大幅度的下滑。在 15000 次训练之后，Bob 预测的错误位数已经几乎为 0，而此时 Eve 的错误位数为仍徘徊在 8 左右。也就是说 Eve 的破解密码的能力基本上仍然等于瞎猜，这时 Alice 发出的信息非常安全。

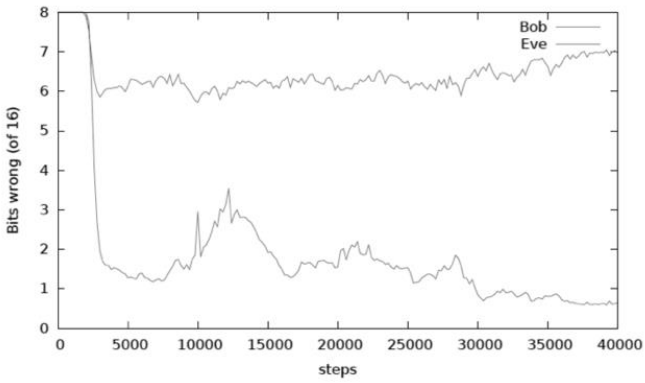
经过这样的训练后，研究人员得到了一个具有极强加密能力的神经网络 Alice。虽然研究人员并没有教给 Alice 任何密码术的知识，但 Alice 已经在对抗中习得了保密信息的技巧。

在研究了对称加密模型之后，研究者们还对非对称的加密模型进行了研究。与对称加密模型不同的是，这一次只有 Bob 拥有密钥 K，Alice 和 Eve 拥有由 K 生成的公钥。公钥是由密钥按照一定规则生成的，它包含了密钥的部分信息。



与之前的研究相比，该结构中多了公钥，Alice、Bob 与 Eve 都能接收到该公钥。参考文献：arXiv:1610.06918

经过同样的训练后，研究者们得到了非对称加密模型的解密准确率。这次的保密效果没有预想中的好，测试结果也比对称加密模型更难解读。在绝大多数的测试中，Eve 的错误字节都在 8 以下，甚至与 Bob 的错误率相当。也就是说在非对称加密模型中，Eve 几乎总能窃取到 Alice 和 Bob 的秘密信息。下图展示的是一个保密效果较好（也就是 Eve 错误率较高）的测试结果，图中 Eve 的错误字节数也下降到了 7：



Bob 和 Eve 解密信息的错误率随实验次数的变化曲线(非对称)，即使在保密效果较好的情况下，Eve 仍能获取少量信息。参考文献：arXiv:1610.06918

这个结果说明，将对称模型改为非对称加密模型对于对抗神经网络的训练结果有着很大的影响。对于这样的变化，研究员们计划通过两种途径来增强在非对称模型下 Alice 的保密效果，分别是使用新的神经网络或者是新颖的训练方式。但进一步的研究将会如何进行，我们还要等待谷歌大脑的新论文。

除了更换保密模型，研究员还将对称模型中明文和密钥的数据位数由 16 位变成 32 位以及 64 位。位数变化后测试的结果与原来的结果是相同的，即 Alice 发出的信息依然能够被很好的保护起来。这就说明该对抗神经网络能够加密更多的信息。如果扩展到 128 位、256 位，甚至成千上万位呢？如果结果依然相同，那么 Alice 的加密能力就有希望用于海量数据的加密与保护。

在这里我们不禁要问，在对称模型下，只用“0”和“1”思考的神经网络为何能够表现的如此优秀。

遗憾的是我现在无法给出答案。目前所有这方面的研究中对此并有没有很明确地解释。可以说神经网络对于科学家们来讲依旧是一种难解释的“黑箱模型”。对于现在的神经网络来说，尚缺乏一套完整系统的理论指导。在神经网络研究快速发展起来之前，我们还需要更多像谷歌大脑这样的探索。

为什么要让机器学习数据加密？

你是否还记得阿尔法围棋（AlphaGo）和它的升级版“大师”（Master）在围棋界搅起的血雨腥风？我们暂且不谈论阿尔法围棋的框架结构，只是简单地说说它对我们的影响。阿尔法围棋在对弈的过程中会落下与我们固有思想不同的棋子，这种机器的“思维”在一定程度上会给我们提供不一样的思考方式，帮助我们更好地探究围棋的技巧。

谷歌大脑的这次新颖且大胆的研究则是一次密码学中的尝试，期待着机器的思维能够为传统的信息加密技术提供新颖的想法。在训练神经网络的过程中，研究人员并没有将密码学相关的算法放进模型中，而是通过神经网络之间的对抗，让 Alice 自己获得了高超的加密能力。当然这并不是说我们已经能抛开密码学，转而依赖神经网络的自身学习能力了。考虑到神经网络的安全性与稳定性，Alice 的“密码术”还不能立刻用于生活中的信息加密。但是另一方面，一旦我们能够稳定的使用神经网络对数据进行加密且确保加密内容难以被破解，便可以将其利用在日常的信息加密中，甚至用于国家安全信息的保护。

也许，机器密码术的时代已经不远了。

作者介绍：

秦曾昌 英国布里斯托大学（University of Bristol）计算机系硕士，布里斯托大学工程数学系的人工智能博士。

small_TA02016 北航研究生在读