# 用LSTM-RNN和PyTorch进行DCNet - 去噪（DNA）序列

我们能用神经网络从一组噪声序列中构造一个无误差一致序列吗？

（相关的Jupyter笔记本可在找到）

#### 用神经网络去除噪声

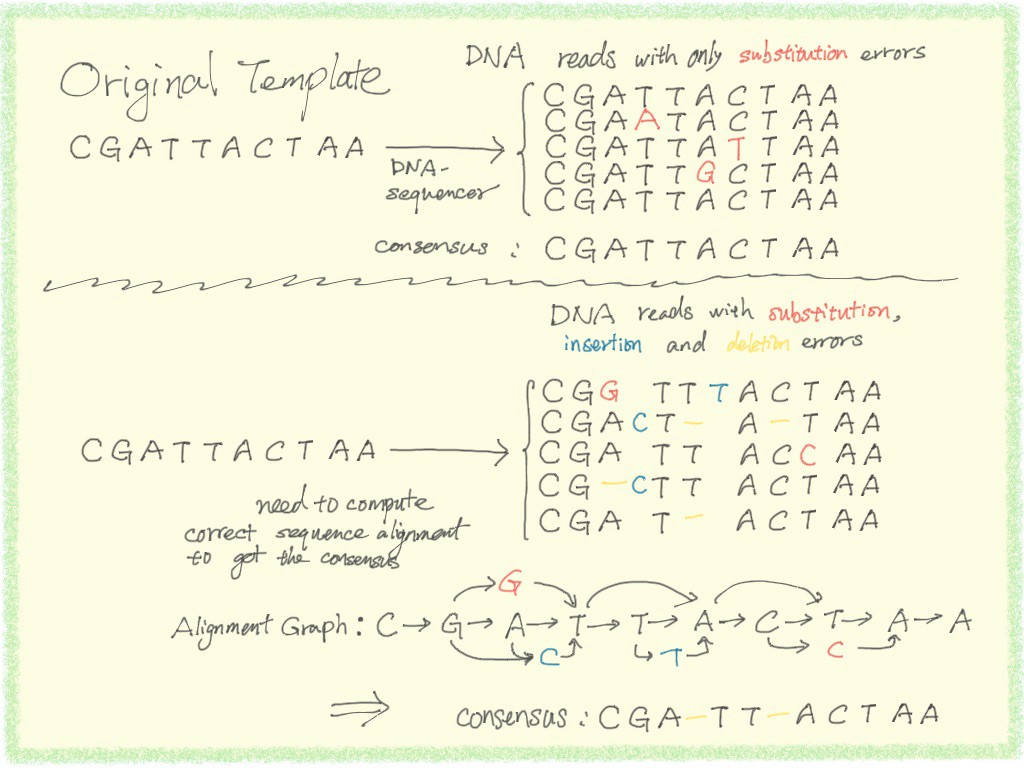
使用一个深卷积神经网络（CNN）或已经产生了一些相当惊人的结果。基本的想法是，你可以训练一个有线电视新闻网学习消除噪音像素。从概念上讲，人们可以认为，训练过程使CNN“学会”做最好的“猜测”，以替换损坏图像中丢失/损坏的像素。



还有另一种古老的技术，在没有任何先验知识或一组训练图像的情况下降低图像噪声。这个想法是拍摄同一物体的许多图像，例如天空中一个暗淡的行星或星云，并将多个图像对齐以减少噪音。如果噪声像素是随机产生的，那么我们可以消除一些仅仅出现在一两幅图像中的像素，因为它们更可能是随机噪声。对象的“正确”像素可能出现在所有图像中。我们只能用这些像素重建更清晰的图像。

有一个很好的类比，去除图像噪声产生一个更好的“共识”（DNA）序列从一组噪声原始序列产生的DNA测序器。DNA测序仪的原始测量会有一些错误。根据底层技术的不同，错误可能有不同的特征。自从DNA测序出现以来，一个重要的生物信息学任务就是从重复测量中消除错误并产生更好的一致性。

与图像不同，DNA序列是一维的，离散的，由一小组字母a、C、G和T表示。这使得如果错误只是“替换”的话，去噪问题就容易一些（参见）。然而，如果DNA序列中随机插入或删除字符（=碱基）时出现错误，则问题将变得更加复杂（例如，请参阅of）。生物信息学算法开发人员已经开发了两种从一组噪声序列生成“一致”序列的方法。它们大多使用某种成对序列序列比对或比较来处理插入和删除错误。然而，我总是好奇是否有可能产生一个一致序列没有任何配对序列序列比对或比较过程。



#### 基本思想

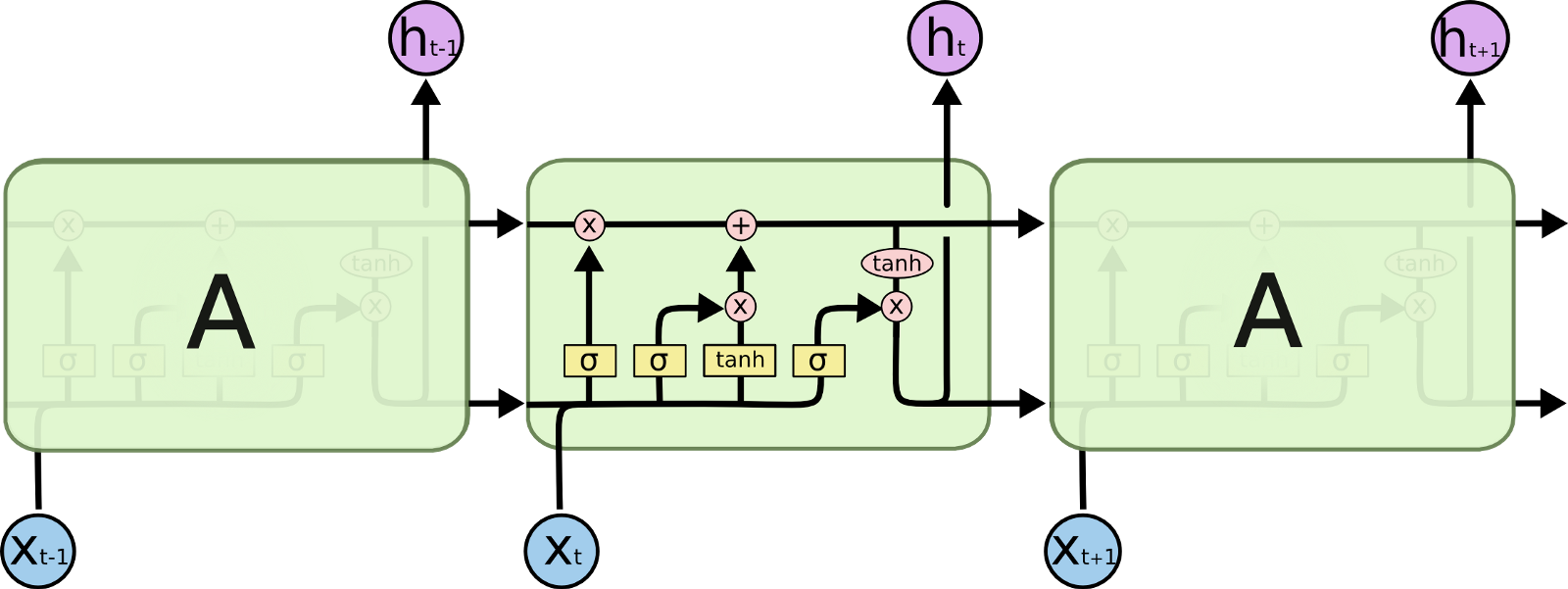
我们能用神经网络把一组错误的DNA序列“叠加”在一起以消除各种错误吗？如果是，网络架构应该是什么样的？也许我们可以建立一个递归神经网络（RNN），更具体地说，一个具有长短期记忆（LSTM）细胞的RNN，来聚合正确的子序列，忽略一些随机的错误事件，而不进行显式对齐。

如何在事先不知道错误所在的情况下，建立一个LSTM-RNN来学习正确的部分并删除错误？我们可以借鉴的一个想法是从一个自动编码器的架构。在自动编码器中，我们通常尝试通过将数据推送到瓶颈层来学习数据的一些隐藏结构，并要求网络的第二部分将原始数据作为网络的输入进行复制（请参阅这篇重要文章中的想法：“）。这种过程迫使网络将原始数据的压缩信息编码到隐藏层的神经元中。

有人可能会说，由初始模板序列生成的一组噪声序列的最佳压缩表示形式只是初始模板序列。受自动编码器体系结构的驱动，如果我们要求LSTM-RNN再现我们发送到其中的每个噪声序列，我们可能期望网络将从无错误的初始模板序列中学习最常见的模式。如果我们要求它从“内存”中生成一个序列，那么它就可以从“内存”中生成正确的模板序列。

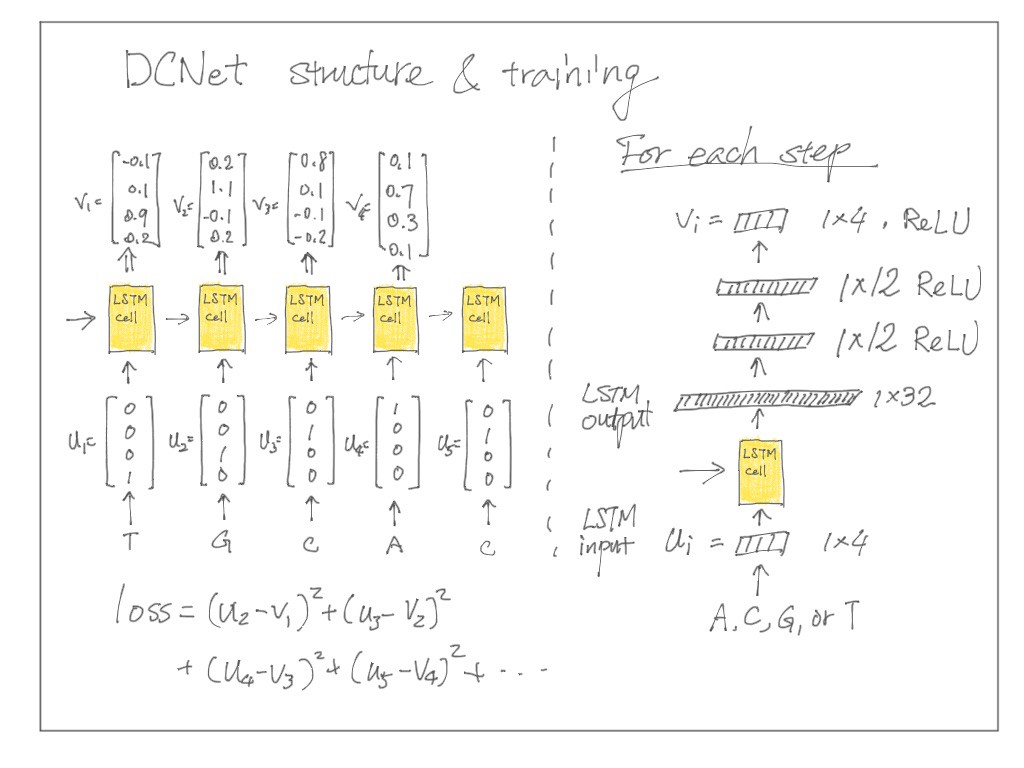
#### DNA共识LSTM网络（DCNet）

网上已经有很多关于LSTM RNN的好教程了。如果你不熟悉LSTM-RNN，我想是为了获得关于LSTM-RNN的良好背景信息。您将需要它来理解共识LSTM网络如何执行其工作。就我个人而言，我不会声称我完全理解LSTM-RNN，但是用PyTorch实现LSTM网络是相当容易的。通过对LSTM网络的研究，可以了解LSTM网络的一些有用性质。



我们可以简单地认为，每个LSTM单元都可以学习一个函数，该函数将其输入和内部隐藏状态映射到某些输出。一旦网络被训练，隐藏状态可以表示在特定LSTM小区之前馈入网络的特定子序列。我们可以尝试制作一个LSTM-RNN来预测给定序列中所有已知特征的DNA特征。也就是说，如果我们把前N个DNA特征输入到网络中，我们希望第N个LSTM细胞能够正确地预测序列的第N+1个特征。为了做到这一点，我们可以将损失函数设置为“预测”DNA序列输出和输入序列之间的差异，并训练网络使之最小化。

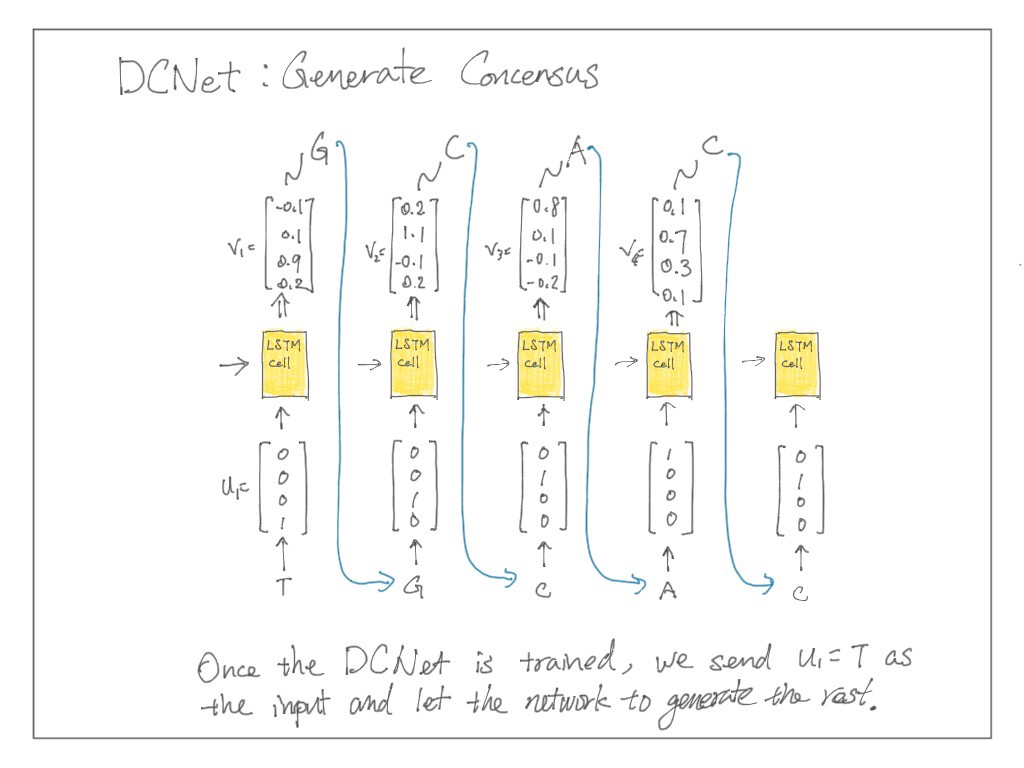
DNA共识LSTM网络（DCNet）的这个简单架构如下所示。我们将输入的DNA序列从a、C、G和T个字符的列表转换为一个由0或1的4个元素的一个热编码向量的列表组成的张量。我们将每个LSTM单元的输出通过两个完全连接的ReLU层，并生成一个由四个浮点数元素组成的输出向量。我们希望输出向量与输入序列中下一个DNA字符的表示一样接近。因此，我们可以将损失函数定义为预测张量和移位输入张量之间的L1或L2损失。



我实现了。首先，我们生成一些随机序列作为输入模板序列。然后，我们模拟了20个带插入、删除和替换错误的噪声序列，并仅用噪声序列训练DCNet。

一旦网络被训练，我们可以测试DCNet，看看如果我们将模板设置为输入，它是否可以提前一步正确地预测原始模板序列。在我自己的实验中，有时训练过程可能不会完全收敛我们可以调整训练速率表或小批量以使收敛更为稳健。使用NVIDIA GeForce GTX 1080 GPU大约需要5到10分钟。

将模板序列发送到DCNet以预测序列本身并不太有用。我们想看看是否可以直接从DCNet构建整个模板序列。为了构造一个序列，我们只需给DCNet一点提示，给它第一个字符。之后，我们可以将第N个LSTM单元的输出作为输入发送到第N+1th LSTM单元，经过一些小的转换来构造一个序列，并将其与原始模板序列进行比较，看看它们之间的距离有多近。（注：DCNet框架与用神经网络预测时间序列非常相似。如果我想提高效率，我会在时间序列分析领域搜索相关论文。）



事实上，如果训练收敛良好，我们可以完美地重建原始模板序列，或者比训练中使用的输入噪声序列的误差小得多。这个简单的DCNet所展示的是，我们实际上可以生成序列一致性，而无需使用LSTM-RNN进行任何序列序列比较或对齐！!

#### 一些有趣的扩展，一些限制和一些随机的想法

我们已经证明，一个简单的LSTM-RNN可以从噪声读取中生成短的一致序列。虽然构造一个神经网络来生成一致序列是有趣的，但DCNet可能不是一个实际应用，因为它的速度比基于序列比对或比较的其他解决方案慢。然而，这种网络有一些有趣的特性值得我们进一步思考。

在上面的例子中，每次训练网络时，我们都要求它学习一个一致序列。如果我们从不同的模板发送两组不同的噪声序列并训练网络，会发生什么？直观地说，如果有足够的“存储容量”，DCNet可以学习多个模板序列。只要这些模板序列彼此相差很大，DCNet记住的序列“路径”就不会混淆。

一些例子表明，当我们使用两个完全独立的生成序列作为模板时，情况确实如此。当然，一个自然要问的问题是，给定一个特定的网络结构和隐藏层大小，对于这样的DCNet，可以成功地重建多少个序列。另外，我们想知道多个模板序列如何影响训练过程？如果没有广泛的研究工作，很难回答这些理论问题。

在许多基因组学工作中，一个主要的挑战是处理DNA序列中的一些重复模式。如果序列中有重复的模式，我认为如果重复的时间太长，很难产生正确的一致性，就会混淆DCNet。

我们能用类似DCNet的神经网络进行“同源性”搜索吗？即，识别（或模糊匹配）其他更长序列中的已知子序列。

在典型的生物信息学实践中，这种同源性搜索是通过所谓的局部序列比对或隐马尔可夫模型（HMM）来实现的。在中，我们需要指定一个具有一定隐态、跃迁和发射概率结构的概率模型，并使用多序列比对的结果来训练它。

在DCNet中，从上一个输出生成下一个输出，并通过每个LSTM单元生成内部隐藏状态的迭代就像HMM中的迭代步骤。如果我们使用DCNet扫描一个长序列，它开始看到一些它“熟悉”的子序列，然后内部状态可能也会经历一些特定的已知轨迹。我们也许可以利用这种特性，利用输出状态和内部状态来训练另一个网络，以寻找类似于使用HMM的同源序列。

与类似的HMM方法相比，DCNet模型可以说是更通用的。我们不需要指定包含特定类型的隐藏状态和发射概率的特定概率模型。网络只是从训练数据中“学习”。这类似于通过一些机器学习任务中的“特征工程”阶段。相反，我们确实需要研究“培训工程”。例如，如果我们想使用DCNet来寻找不连续的同源序列，我们可能需要一个具有一定性质的训练集，并调整训练过程以确保损失函数是收敛的。

我只测试了DCNet多达200个字符的四个字母。训练需要一段时间才能集中。我不知道训练DCNet处理更长的序列需要多长时间。基于比对的DNA一致性方法现在速度更快，可以处理长达十万甚至数百万个DNA碱基的序列。有可能使训练更快吗？如果没有，那么考虑是否有其他属性可以使这种网络对其他一些应用程序有用仍然是有趣的。当然，扩展DNA以外的序列的字母表大小是很容易的。或者，如果我们使用DNA序列的一个热输入作为输入，我们可以捕获一些简单的语法错误或捕获剽窃吗？（我确信在NLP领域已经有很多工作在使用更高级的网络来实现这一点。）