信息网络专题研究之应用层

一、 文献信息

1. **作者：** Aäron van den Oord, Nal Kalchbrenner, Koray Kavukcuoglu
2. **论文题目：** Pixel Recurrent Neural Networks (Pixel RNN)
3. **发表途径：** Proceedings of the 33rdInternational Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.
4. **发表时间：** 2016

二、 问题意义

1. 研究背景与意义

自然图像的分布建模是无监督学习中的一个典型问题，该任务需要一个同时具有表现力、易于处理和易于伸缩的图像模型。但由于图像的高维性和高度结构化，估计自然图像的分布是非常具有挑战性的。2014 年提出的 VAE 算法可以有效提取图像特征，但趋向于产生模糊的图片且不易于处理。

2016 年，Aäron 等人提出了一种深度神经网络——pixel RNNs，它能在两个空间维度上连续预测图像中的像素。作者还创新提出了快速二维递归层和在深层递归网络中使用有效利用剩余连接。Pixel RNNs 在自然图像上实现了对数似然评分，显著改善了 MNIST 和 CIFAR-10 数据集的最新技术，还为 ImageNet 数据集上的生成图像建模提供了新的基准，大大优于以前的技术水平。Pixel RNNs 是无监督学习的发展中具有里程碑意义的成果。

2. 主要研究问题

本文提出了一种基于深层递归神经网络的自然图像生成模型 pixel RNNs，并提供了对不同数据集的性能测试。

三、 思路方法

本文基于数学理论分析提出了 pixel RNNs 模型，分析各模型的特点与优劣，最后通过多个数据集上的实验证明了 pixel RNNs，特别是 Diagonal BiLSTM 的优越性。论文的组织思路是：理论证明→方案构建→实践论证。

1. 数学基础

生成模型的目标是估计自然图像上的像素分布，pixel RNNs 属于全可见置信网络，因此我们要对一个似然概率密度显式建模。我们使用链式法则将似然概率分解为一维分布条件概率的乘积：$p(x) = \prod_{i=1}^{n^2}(p(x_i)|x_1, \dots, x_{i-1})$

对于 RGB 三色通道同理可得：$p(x) = p(x_{i,R}|x_{<i})p(x_{i,G}|x_{<i}, x_{i,R})p(x_{i,B}|x_{<i}, x_{i,R}, x_{i,G})$

同时，我们将$p(x)$模型中的像素值x_i化为离散分布。离散分布具有简单的表征性，并且具有任意多模态且形状无先验的优点。

2. 模型设计（重点）

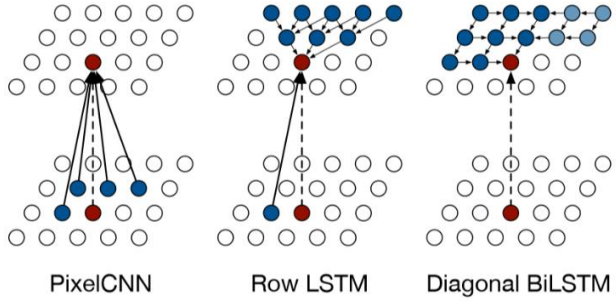
基于现有理论，本文首先设计了两种二维 LSTM，即 Row LSTM 和 Diagonal BiLSTM，它们使用卷积来同时计算空间维度上的状态。然后分析了使用剩余连接来改进具有多个

LSTM 层的 pixel RNN 结构。接着，作者引入了 softmax 激活和屏蔽卷积技术来进一步改善模型。最后，文章又提出了生成速度快的 Pixel CNN 和多尺度 pixel RNN 体系结构。

Row LSTM 是一种单向层，它对整行图像进行从上到下的逐行处理，同时处理整行计算特性，其计算是用一维卷积来完成的。Row LSTM 每个点的生成都依赖于前面三个点。

Diagonal BiLSTM 是一个双向链表，用于并行化计算，该层从两个方向以对角线的方式扫描图像，即每一行相对上一行偏移一个像素。

而 Pixel CNN 使用标准的卷积层来捕获一个有界的接收域，并同时计算所有像素位置的特征。Pixel CNN 和普通 CNN 的区别是在进行卷积操作的时候，会先和 mask 屏蔽层相乘，以消除将来点的影响。这三种模型的输入层和输出层如图：



3. **模型规格（略）**

4. **实验与分析**

基于以上模型在 MNIST、CIFAR-10 和 ImageNet 数据集上进行训练，并以对数似然损失函数（NLL）作为评估标准。实验表明：softmax 激活与剩余连接都使学习效果有明显的提高；同时，pixel RNNs 成为 MNIST 和 CIFAR-10 数据集的最佳生成模型，还为 ImageNet 数据集上的生成图像建模提供了新的基准。

四、 实验结论

本文提出了一系列基于深层递归神经网络的自然图像生成模型。pixel RNNs 显著改善了 MNIST 和 CIFAR-10 数据集的最新技术，还为 ImageNet 数据集上的生成图像建模提供了新的基准。基于模型的样本和完整性，我们可以得出这样的结论：pixel RNNs 能够在空间上模拟局部和长距离的相关性，并且能够生成清晰合理的图像。考虑到这些模型随着数据集变大而改进，而实际上有无限的数据可供训练，更多的计算和更大的模型可能会使得生成结果更加清晰、多样且符合要求。

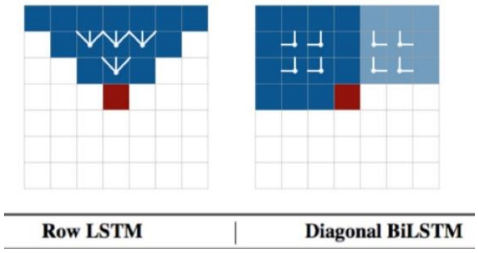
五、 启发思考

1. 在最初阅读论文时，我对于 Row LSTM 和 Diagonal BiLSTM 这两种框架的结构存在困惑。我的问题主要有两个：一个是每行像素在输入 Row LSTM 之前是否要先做卷积，另一个是 Diagonal BiLSTM 的输入门是什么。在阅读了《*Generative Image Modeling Using Spatial LSTMs*》之后，我的困惑得到了解决。

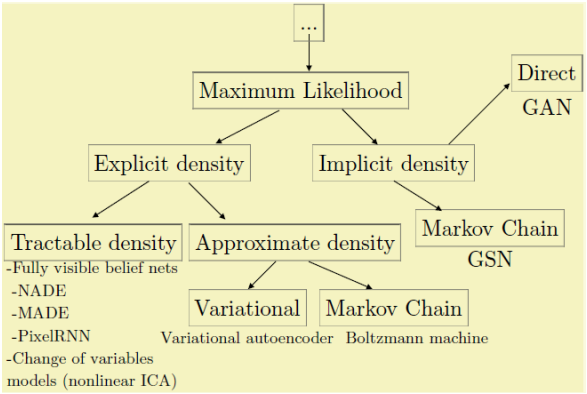
$$
\begin{aligned}
[\mathbf{o}_i, \mathbf{f}_i, \mathbf{i}_i, \mathbf{g}_i] &= \sigma(\mathbf{K}^{ss} \circledast \mathbf{h}_{i-1} + \mathbf{K}^{is} \circledast \mathbf{x}_i) \\
\mathbf{c}_i &= \mathbf{f}_i \odot \mathbf{c}_{i-1} + \mathbf{i}_i \odot \mathbf{g}_i \qquad (3) \\
\mathbf{h}_i &= \mathbf{o}_i \odot \tanh(\mathbf{c}_i)
\end{aligned}
$$

我们结合 LSTM 的知识仔细理解一下公式 3：这个式子中x_i代表输入的第i行像素，h_{i-1}代表生成的第$i-1$行像素。假如对每一层 LSTM 的输入有m个特征映射，也就是的通道数为m，那么K^{is}和K^{ss}分别将x_i和h_{i-1}的维度映射为$4m$，对应于 LSTM 的四个门o_i，f_i ，i_i和 g_i。根据这四个门我们就可以得到新的状态c_i和输出h，其中h就是我们到求得的第i行的像素。通过这一循环过程就能生成整个图像。

Diagonal BiLSTM 和 Row LSTM 完全一样，只不过是卷积核K^{ss}不同而已。在 Row LSTM 中，卷积核的接受域是上一行的的三个像素，而 Diagonal BiLSTM 中卷积核的接受域是周围的 4 个像素，如下图所示：



2. 生成模型其实又可以细分为显式密度模型和隐式密度模型。显式密度模型是能显式对$P(x,y)$建模的，而隐式密度模型只能从样本中生成新的样本。其中常见的是 pixel RNNs、VAE 和 GANs。对于这三种模型的特点，我们可做以下总结：



pixel RNNs 能显式地计算似然$p(x)$，是一种可优化的显式密度模型。该模型优化的是一个显式的似然函数并产生良好的样本，但是效率很低，因为它是一个顺序的生成过程。

而 VAEs 有一个额外定义的隐变量z，有了z以后获得了很多的好处，但是密度函数也变得很难解，对于该函数我们不能直接优化，我们只能推导出一个似然函数的下界，然后对它进行优化。

GAN 是目前能生成最好样本的模型，但是训练需要技巧且不稳定，查询推断上也有一些问题。

讲解视频链接：

https://www.bilibili.com/video/BV1rt4y117yz?from=search&seid=297542316350275
3105