《图像转换器》研究

摘要：图像的生成问题可以看成是一个自回归的序列转换问题。沿着这一思路，作者提出了一种基于自注意力模型的图像生成器，此外，作者还将模型应用于高放大率的图像超分辨率。在上述两个方面，模型都有很不错的表现。本研究将从论文的文献信息、问题意义、思路方法、实验结论、启发思考等方面进行讨论。

一、 文献信息

1. 作者：Niki Parmar，Ashish Vaswani，Jakob Uszkoreit，Łukasz Kaiser，Noam Shazeer，Alexander Ku，Dustin Tran
2. 论文题目：Image Transformer
3. 发表途径：arXiv
4. 发表时间：15 Jun 2018

二、 问题意义

目前较为经典的图像生成算法有 PixelRNN 和 PixelCNN。PixelRNN 串行处理每一个像素点，计算量非常大，而 PixelCNN 并行处理像素点，相比于 PixelRNN 有着更好的性能。但是 PixelCNN 有一个显著的缺点——较窄的感知野，这不利于处理图像中的对称和遮挡。若想提高 PixelCNN 的感知野，就需要付出牺牲计算量的代价。论文中提出了一种基于自注意力模型的图像转换器，它在 PixelRNN 的高感知野，串行处理和 PixelCNN 的低感知野，并行处理之间做了折中，有着非常良好的性能。作者把基于自注意力模型的图像转换器分别用在了图像生成和图像超分辨率上，都生成了比现有技术更自然的图像。

三、 思路方法

作者在介绍模型架构时分为图像的表示，图像的自注意力结构，局部自注意力来进行介绍，下面我们一一进行讨论。

（1） 图像表示

像素强度可以分两类表示，一种是离散类别，一种是顺序值。对于离散类别而言，像素的三个通道被分别编码，每个通道由 256 个 d 维向量表示。其输出表示则由通道共享 256 个 d 维向量。结合图像的宽度 w 和高度 h，像素可以表示为$[h, w \cdot 3, d]$。对于每个像素最终都会输出 256*3=768 个参数，若是 32*32 的图像，最终就会输出 786432 个像素。

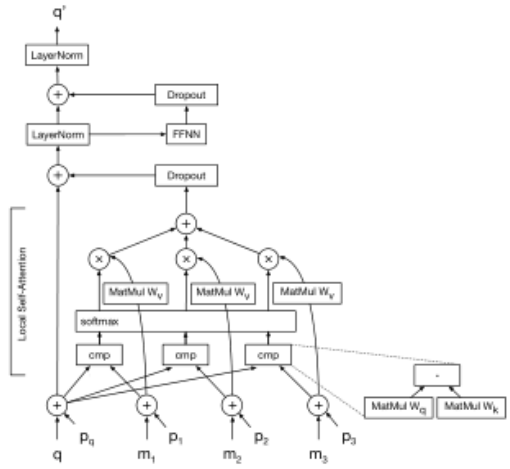
对于顺序值而言，作者运行一个1×3的窗口，最终形成每个像素的输入表示$[h, w, d]$。每个像素都由 10 个混合元素表示，一个非归一化的混合概率，三个平均值，三个标准差，三个捕捉线性独立性的系数。最终每个像素会输出 100 个参数，若是 32*32 的图像，最终就会输出 102400 个像素，显然比离散类别更节省内存。

对于每个输入的像素都会加入一个 d 维位置编码，每一个维度都采用正弦或余弦的不同频率，而且，一半的维度是用来表示行位置的，另一半的维度用来表示列位置和颜色通道。

（2） 自注意力结构

对于图像超分辨率，整体框架为 Encoder-Decoder，而对于基于类别条件的图像生成，

只采用了 Decoder 结构。编码器用于生成每个像素的语义信息。译码器结合语义信息和已生成的像素信息生成当前像素的信息。无论是编码器还是译码器，都包含自注意力模型和一系列前馈层。自注意力模型结构如下图所示。



图一：自注意力模型

如图所示，计算某个像素表示 q 时，以前生成的像素会结合W_q和W_k参与线性运算。q 会与其他像素通道表示进行比较后，使用点积并缩放$\sqrt{d}$，输出的信息会经过 softmax 函数，其输出结果又经过W_V的线性变换，至此局部自注意力模型结束。接着，信息先后通过 Dropout 和 Layer Normalization 处理，然后进入两层前馈网络处理。紧随其后的是 Layer Normalization 和 Dropout 处理以及合并剩余连接来对输出结果进行优化。上图用公式表示如下。

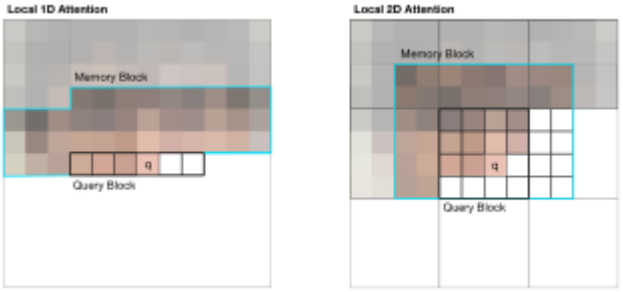
$$q_a = \text{layernorm}(q + \text{dropout}(\text{softmax}\left(\frac{W_q q (MW_k)^T}{\sqrt{d}}\right) MW_v)) \quad (1)$$

$$q' = \text{layernorm}(q_a + \text{dropout}(W_1 \text{ReLu}(W_2 q_a))) \quad (2)$$

值得注意的是，未生成的像素不能参与运算，所以需要进行掩盖。除此之外，全部的自注意力操作可由高度优化的矩阵乘法代码来实现，实现全部像素通道的并行操作

（3） 局部自注意力

内存中图像位置的数量对于自注意力算法的可伸缩性有着重要的影响。当图像位置数量很多时，考虑全部的图像位置是不切实际的。因此，作者采用了局部自注意力算法。局部自注意力算法构造一个查询块和一个包含查询块的内存块，内存块中的像素参与查询块中像素的生成，而查询块中的像素并行运算。论文中给出了两种查询块和内存块的选择方式，如下图所示。



图二：局部自注意力模型

对于 1D 局部自注意力模型，像素以光栅扫描顺序输入。查询块长度为l_q，而内存块长度还要比查询块长度多l_m，若查询块长度不够，需进行补零操作。

对于 2D 局部自注意力模型，查询块为矩形，查询块与像素通道都以光栅扫描顺序进行变化。查询块大小为$l_q = w_q * h_q$，其中w_q为查询块宽度，h_q为查询块高度，而内存块在查询块的基础上向左右延伸w_m，向上延伸h_m。2D 局部自注意力模型中的查询块像素受水平和垂直已生成像素的影响较为平均，而 1D 局部自注意力模型中的查询块像素更多受到它相邻像素的影响，受上方像素影响较小。当图像规模变大时，2D 局部自注意模型会有着更好的性能。

如图二所示，白色部分为上文提到的掩盖部分。

论文中损失函数采用了最大似然函数，其优化目标是最大化$\log p(x) = \sum_{t=1}^{h \cdot w \cdot 3} \log p(x_t \mid x_{<t})$。其中 x 是图像表示中提到的参数。

研究方法：整体上，作者采用了对比以及理论阐述的方法来进行研究和论述。

四、 实验结论

在整个实验中作者采用了 p100 和 k40GPU 来训练模型。

作者无条件图像生成中用到的模型参数如下表所示。

Dataset	Image Presentation	l_q	Memory Size	Number of Layer	d	Heads	feed-forward dimension	Dropout
CIFAR-10	categorical	256	512	12	512	4	2048	0.3
CIFAR-10	DMOL	256	512	14	256	8	512	0.2
CIFAR-10	-	256	512	8	512	8	1024	0.1
ImageNet	categorical	256	512	12	512	8	2048	0.1

表一：无条件图像生成模型参数

不同模型的效果如下表所示

Model Type	$bsize$	NLL	
		CIFAR-10 (Test)	ImageNet (Validation)
Pixel CNN	-	3.14	-
Row Pixel RNN	-	3.00	3.86
Gated Pixel CNN	-	3.03	3.83
Pixel CNN++	-	2.92	-
PixelSNAIL	-	**2.85**	3.80
Ours 1D local (8l, cat)	8	4.06	-
	16	3.47	-
	64	3.13	-
	256	2.99	-
Ours 1D local (cat)	256	2.90	**3.77**
Ours 1D local (dmol)	256	2.90	-

表二：数据集上的 Bits/dim

由表二可知，论文中的图像转换器的表现优于 PixelRNN 以及 PixelCNN++。在 CIFAR-10 数据集上，PixelSNAIL 的表现优于图像转换器，但在 ImageNet 上，仍是图像转换器的表现更佳。值得注意的是，在 ImageNet 上，图像转换器的性能达到了新高度。此外，8 层的图像转换器的性能基本与 Gated Pixel CNN 一致。表二还说明提高感知野会显著改善性能，这也是图像转换器优于 CNN 的关键所在。在分类分布时，位置编码方式对模型性能没有影响。

在基于类别条件的图像生成中，需要给每个像素表示加入表示类别的向量，实验结果表明基于类别条件的图像生成效果要好于无条件的图像生成的效果,但是它们的τ都等于 1.0。

在图像超分辨率中，作者实现了将 8*8 的图像转换为 32*32 的图像，在原图像中恢复出了图像细节。在编码器的位置编码上，作者加入了两个维度分别用来表示行和宽度并将整个图像扩展为$[h \times w \times 3, d]$张量，其中 d=512。在训练模型时，作者使用对数似然目标函数对 Encoder-Decoder 模型进行端到端的超分辨率训练。值得注意的是，在编码器中没有进行掩盖，而且最理想的模型结构为编码器的层数比译码器的层数少 2-3 倍。作者在 CelebA 数据集上的模型参数如下表所示。

Local Attention	l_q	Memory Size	Number of Layer	d	Heads	feed-forward dimension	Dropout	NLL
1D	128	256	12	512	8	2048	0.1	2.68
2D	8*32	16*64	12	512	8	2048	0.1	2.61

表三：CelabA 数据集上的模型参数

为了测试效果，作者采用了 50 个人来分辨 50 对图像的方法进行实验，实验结果如下表所示。

Model Type	τ	%Fooled
ResNet	n/a	4.0
srez GAN	n/a	8.5
PixelRecursive (Dahl et al., 2017)	1.0	11.0
	0.9	10.4
	0.8	10.2
1D local Image Transformer	1.0	29.6 ± 4.0
	0.9	33.5 ± 3.5
	0.8	$\mathbf{35.94 \pm 3.0}$
2D local Image Transformer	1.0	30.64 ± 4
	0.9	34 ± 3.5
	0.8	$\mathbf{36.11 \pm 2.5}$

表四：CelabA 数据集上的人工评估结果

由上表可知，$\tau = 0.8$时的 2D 局部自注意力模型效果最好。此外，作者还计算出了表征输入输出一致性的 Consistency 值，其值为 0.01。最后作者还给出了 MS-SSIM 分数为 44.3，这说明文中的训练模型不仅能复制原图，还能还原具体细节。在 CIFAR-10 数据集上的表现也能得出类似结论。

五、 启发思考

1.这篇文章带给我的第一点思考是任何创新都必须建立在深刻理解原理的基础上，只有深刻理解了自注意力模型的原理和结构，才能设计出适合图像生成的自注意力模型。同理，若没有对光的本质的理解，爱因斯坦也就不可能提出相对论。

2. 这篇文章带给我的第二点思考是要擅长结合现有的前沿技术，往往新技术的结合能够带来好的性能，论文中作者就是把 Encoder-Decoder 与自注意力模型结合起来，产生了可喜的实验结果。

3. 这篇文章带给我的第三点思考是在设计实验时，要充分考虑模型的效果以及设计目的。文中为了验证超分辨率的效果，采用了人眼辨识的实验，很好的得出了实验结果。有时候，巧妙的实验设计能够起到事半功倍的效果。卡文迪许扭秤实验就是一个很好的例子，正是巧妙的实验装置让我们得到了万有引力常数。