文本生成图片

摘要：从文本自动合成逼真的图像是有趣且有用的，当前的 AI 离这个目标还很远。但是近些年来，已经开发出通用且功能强大的递归神经网络架构来学习判别性文本的特征表示。同时，深度卷积生成对抗网络（GAN）已经开始生成特定类别引入注目的图像。通过将文本的深层表示与 GAN，注意力机制相结合，不少模型已经取得了不错的效果。本文介绍了从 2016 年以来文本生成图像领域的重要工作，可以看到基于生成对抗网络（GAN）的框架依旧是该任务的研究热点，除此之外，还有包括序列到序列，变分编码器等框架。

1 主流工作

1.1 Generateing Images From Captions with Attention

1.1.1 创新点与研究动机

本文提出一种条件 alignDRAW 模型[1]，使用软注意力机制来从文本描述中生成图像。具体的，模型包括编码图像的编码器和解码图像的解码器，编码器网络确定潜在变量的分布，解码器网络在潜在变量中接收样本，迭代的生成图像。

1.1.2 方法介绍

(1) 双向注意力 RNN

用 y 表示输入的文本，对其进行编码 $y=(y_1,y_2,...,y_N)$，N 表示句子的长度。首先将 y_i 表示成 m 维的向量，用 h_i^{lang} 表示。我们使用带有遗忘门的双向 RNN 从前向和反向处理输入序列。前向 LSTM 计算 $\left[\overrightarrow{h_1^{lang}},\overrightarrow{h_2^{lang}},...,\overrightarrow{h_n^{lang}}\right]$，反向 LSTM 计算 $\left[\overleftarrow{h_1^{\text{lang}}},\overleftarrow{h_2^{lang}},...,\overleftarrow{h_N^{lang}}\right]$，将所有的隐状态拼接在一起构成 $h^{lang}=\left[h_1^{lang},h_2^{lang},...,h_N^{lang}\right]$，其中 $h_i^{lang}=\left[\overrightarrow{h_i^{lang}},\overleftarrow{h_i^{lang}}\right]$。

（2）条件 DRAW 网络

为了生成图片 x，作者扩展 DRAW[2]网络，在每一个时间步上包含文本表示 h^{lang}，如图 1 所示。条件 DRAW 网络是一个随机递归神经网络，由潜变量 $\sigma(c_T)$ 组成。

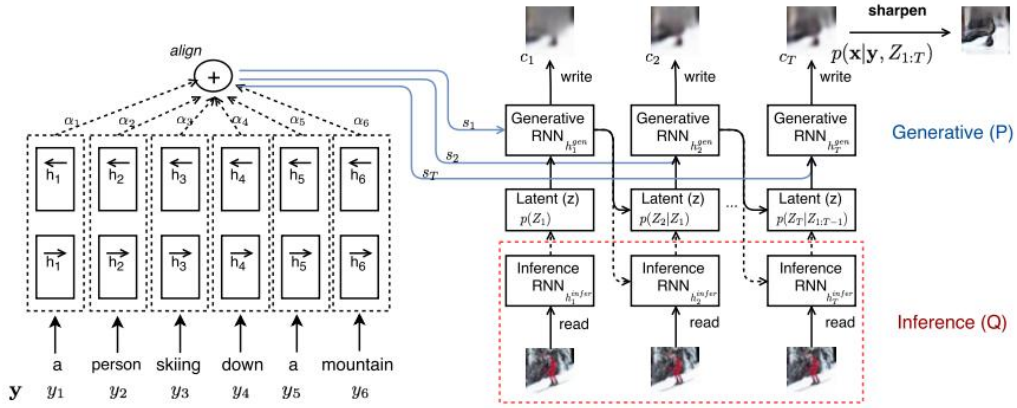


图 1：AlignDRAW 模型

在传统的 DRAW 网络中，潜变量是服从独立的高斯分布 $N(0,I)$，但是本文提出的 alignDRAW 网络中的潜变量有均值和方差，并且它们由生成 LSTM 的隐状态表示，除了第一个潜变量用 $P(Z_1)=N(0,I)$ 表示，因此，本模型的潜变量可以用如下公式表示：

$$P(Z_t|Z_{1:t-1})=N\left(u\left(h_{t-1}^{gen}\right),\sigma\left(h_{t-1}^{gen}\right)\right),$$
$$u\left(h_{t-1}^{gen}\right)=\tanh\left(W_u h_{t-1}^{gen}\right),$$
$$\sigma\left(h_{t-1}^{gen}\right)=\exp\left(\tanh\left(W_\sigma h_{t-1}^{gen}\right)\right)$$

图片由迭代的计算下面的等式产生，$t=1,...,\mathrm{T}$

$$z_{\mathrm{t}}\sim P(Z_t|Z_{1:t-1})=N\left(u\left(h_{t-1}^{gen}\right),\sigma\left(h_{t-1}^{gen}\right)\right),$$
$$s_t=align\left(h_{t-1}^{gen},h^{lang}\right),$$
$$h_t^{gen}=LSTM^{gen}\left(h_{t-1}^{gen},[z_t,s_t]\right),$$
$$c_t=c_{t-1}+write\left(h_t^{gen}\right),$$
$$\tilde{x}\sim P(x|y,Z_{1:T})=\prod_i P(x_i|\mathrm{y},Z_{1:T})=\prod_i Bern\left(\sigma(c_T,i)\right)$$

其中 align 函数用来计算输入文本和中间每步生成的图片的相似性。给定文本描述，$h^{lang}=\left[h_1^{lang},h_2^{lang},...,h_N^{lang}\right]$，对齐操作根据时间步动态生成输出 s_{t}，计算方式如下：

$$s_{\mathrm{t}}=align\left(h_{t-1}^{gen},h^{long}\right)=\alpha_1^t h_1^{lang}+\alpha_2^t h_2^{lang}+...+\alpha_N^t h_N^{lang}$$

其中 α_k^t 表示文本中的第 k 个词的对其概率，由文本表示 h_{lang} 和生成模型 t 时刻的

隐状态 h_{t-1}^{lang} 有关，具体的，计算公式如下：

$$\alpha_k^t = \frac{\exp\left(v^T \tanh\left(Uh_k^{lang} + Wh_{t-1}^{gen} + b\right)\right)}{\sum_{i=1}^{N} \exp\left(v^T \tanh\left(Uh_k^{lang} + Wh_{t-1}^{gen} + b\right)\right)}$$

其中 $LSTM^{\text{gen}}$ 函数表示单时间步的 LSTM 遗忘门操作，为了得到下一个隐状态 h_t^{gen}，$LSTM^{\text{gen}}$ 函数将上一时刻输出 h_{t-1}^{gen} 和潜变量 z_t 与句子向量 s_t 组合在一起。

$LSTM^{\text{gen}}$ 的输出 h_t^{gen} 通过一个写操作被添加到画布矩阵 $c_t \in \mathbb{R}^{h \times w}$ 中。写操作会产生两个一维的高斯过滤器 $F_x\left(h_t^{gen}\right) \in \mathbb{R}^{h \times p}$ 和 $F_y\left(h_t^{gen}\right) \in \mathbb{R}^{w \times p}$。将高斯过滤器组应用到 $p \times p$ 的图像块 $K\left(h_t^{gen}\right)$ 中，放置到画布上：

$$\Delta c_t = c_t - c_{t-1} = write\left(h_t^{gen}\right) = F_x\left(h_t^{gen}\right) K\left(h_t^{gen}\right) F_y\left(h_t^{gen}\right)^T$$

最终，画布矩阵 c_t 中的每一项 $c_{T,i}$ 通过 Sigmoid 函数产生一个均值为 $\sigma\left(c_T\right)$ 的伯努利分布。

通过以上步骤，我们可以迭代地生成图片。

1.2、Generative Adversarial Text to Image Synthesis

1.2.1 创新点与研究动机

本文提出了一个简单有效的生成对抗网络（GAN）框架[3][4]，依据文本描述自动合成真实的图像，即学习一个映射直接把单词和字符转换成图像像素。主要解决两个技术难题：一是学习文本的特征表示以捕获重要的视觉细节，二是利用捕获到的特征合成人类难分真假的图像。

1.2.2 方法介绍

本文训练一个以混合字符级卷积递归神经网络编码的文本特征为条件的深度卷积生成对抗网络（DC-GAN），具体的网络结构如下：

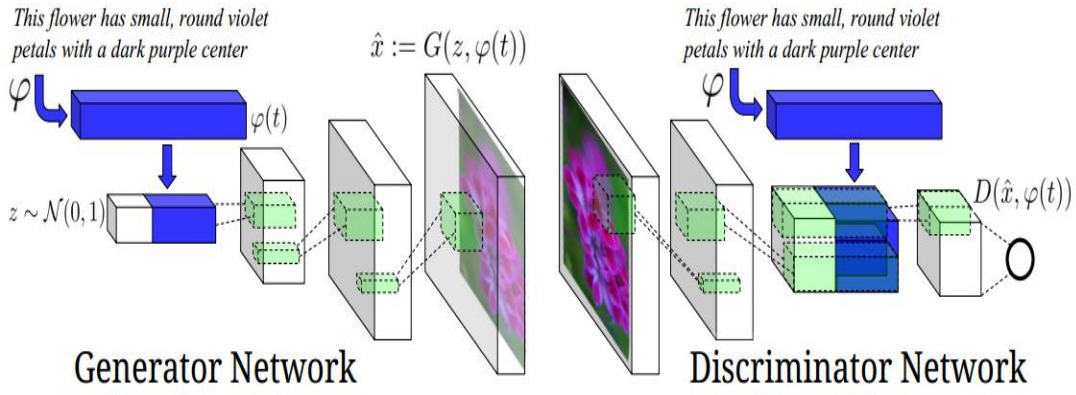


图 2. 基于文本的卷积 GAN 结构。

（1）DC-GAN

生成器 G：将文本信息经过预处理得到特征表达，然后将其和噪声向量组合在一起。在上图中蓝色长方体就代表文本信息的特征表达Ψ(t)，白色长方体是噪声向量 z。将描述文本使用一个全连接层压缩到一个较小的维度之后（一般是 128 维），使用 leaky Relu 与噪声向量 z 拼接在一起，将得到的组合向量输入到反卷积网络中，经过多层处理最终得到一幅图像。

判别器 D： 将图像经过步长为 2 的卷积层，再经过全连接层进行整形得到图像特征向量，使用全连接层对文本编码φ(t)进行整形，当判别器 D 的空间维度为 4×4 时，将图像特征向量和文本编码拼接，再经过 1×1 和 4×4 的卷积层，最后得到一个二值元，用来判断图像的真假。

（2）GAN-CLS

训练 GAN 最直接的方法就是将图片和文本描述编码看作是联合的样本，通过观察判别器判断“生成的图片+文本”是否正确。但是这种方法有些简单，因为它没有给判别器提供“是否按照描述正确生成了图像”的信息。在朴素 GAN 下，会观测到两种不同输入：真实图片和匹配的文本；以及合成图片和任意文本。

作者修改了 GAN 的训练算法并增加了第三类输入：真实图片（携带不匹配描述文本），具体的算法流程如下：

Algorithm 1 GAN-CLS training algorithm with step size α, using minibatch SGD for simplicity.

Input: minibatch images x, matching text t, mismatching $\hat{t}$, number of training batch steps S
for $n = 1$ **to** S **do**
  $h \leftarrow \varphi(t)$ {Encode matching text description}
  $\hat{h} \leftarrow \varphi(\hat{t})$ {Encode mis-matching text description}
  $z \sim \mathcal{N}(0,1)^Z$ {Draw sample of random noise}
  $\hat{x} \leftarrow G(z,h)$ {Forward through generator}
  $s_r \leftarrow D(x,h)$ {real image, right text}
  $s_w \leftarrow D(x,\hat{h})$ {real image, wrong text}
  $s_f \leftarrow D(\hat{x},h)$ {fake image, right text}
  $\mathcal{L}_D \leftarrow \log(s_r) + (\log(1-s_w) + \log(1-s_f))/2$
  $D \leftarrow D - \alpha\partial\mathcal{L}_D/\partial D$ {Update discriminator}
  $\mathcal{L}_G \leftarrow \log(s_f)$
  $G \leftarrow G - \alpha\partial\mathcal{L}_G/\partial G$ {Update generator}
end for

（3）流形插值

作者提出流形插值的方法来增加文本描述的信息，也就是说，”一只羊在吃草”和”一只鸟在书上唱歌”，它们的深度特征插值后，可能就变成了“一只羊在树上唱歌”。这样做可以增加生成器 G 的创新性。具体的，作者在生成器的目标函数上增加一项：

$$\mathbb{E}_{t_1,t_2\sim p_{data}}[\log(1-D(G(z,\beta t_1+(1-\beta)t_2)))]$$

1.3 Learning what and where to draw

1.3.1 创新点与研究动机

生成对抗网络已经展示了合成逼真的真实世界图像的能力。但是现有的模型[6]是基于全局约束（例如类标签或文字说明）来合成图像，并没有对物体的姿势和位置进行控制。本文的提出了一种新的模型（GAWWN）[5]，基于边界框和关键点，将额外的位置和画什么的条件信息加入到生成器和判别器的训练中，从而得到更逼真的真实图像。

1.3.2 方法

（1）基于边界框的文本到图片的生成模型

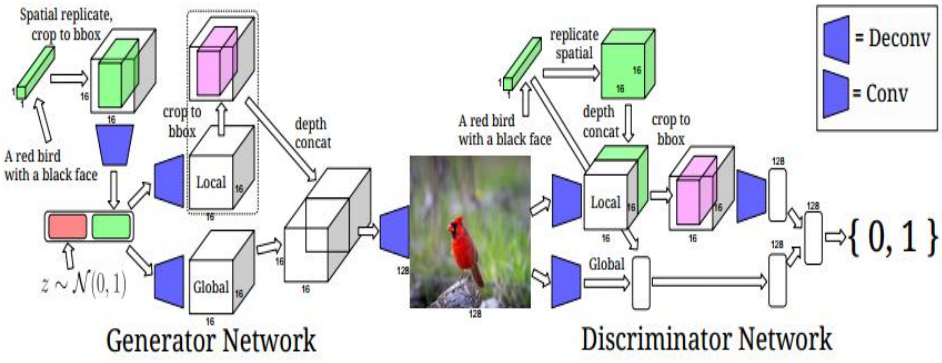


图 3. 基于边界框的 GAWWN

生成器网络：首先特征词向量扩维成 M x M x T 的特征矩阵，依据归一化的边界框坐标进行空间裁剪，特征矩阵外部全设置为 0。经过卷积和池化操作变成 1 x 1 x T 的向量与服从正态分布的噪音相结合，之后分成局部和全局操作。全局操作通过步长为 2 的反卷积操作变成 M x M，局部操作中，通过反卷积变为 M x M 后，再次通过边界框裁剪，边界框外部全部赋值为 0.最后将全局和局部操作得到的特征图相结合，经上采样生成 128 x 128 的图片。

判别器网络：文本嵌入向量扩维成 M x M x T 的特征图.同时图片经过局部和全局操作处理。全局操作经过一些列卷积操作变成 1 x 1 的向量与扩维后的文本特征图相结合，经过边界框的裁剪后，通过卷积操作成为 1 x 1 的向量。最后局部和全局操作的输出经过一层处理得到一个得分。

（2）基于关键点的文本到图片生成模型

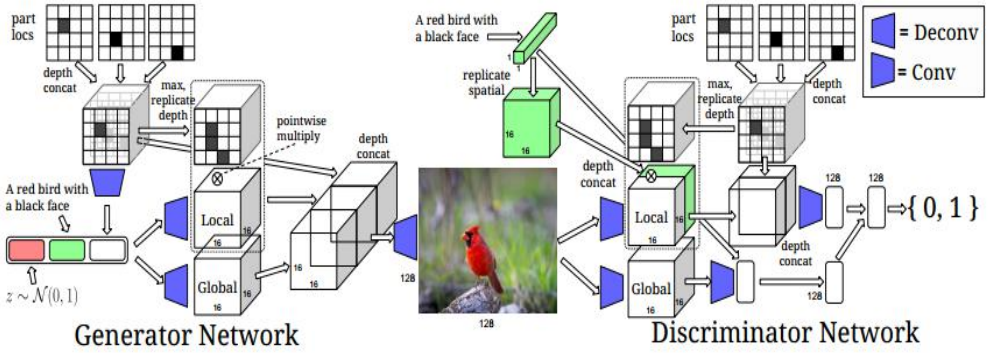


图 4. 基于关键点的 GAWWN

生成器网络：位置关键点被编码成 M×M×K 的空间特征图，其中通道对应于物体某部分：如通道 1 对应头部，通道 2 对应左脚，等等。关键点张量通过步长为 2 的卷积操作后与噪声 z 和文本嵌入 t 合并成串联向量，代表了关于内容和部分位置的粗略信息。接着，关键点张量通过展平为二进制矩阵，并用 1 表示特

定空间位置存在的任何部分，然后深度级联成 M×M×H 的张量。之前得到的串联向量经过反卷积成 M×M×H 的张量，通过与相同大小的关键点张量的逐点相乘来控制局部向量的激活。最终局部张量，全局张量以及关键点张量深层级联经过反卷积操作，过一层 Tanh 激活函数得到 128×128 的图像。

判别器网络：文本嵌入向量 t 分别喂入两个阶段，在全局阶段，它与全局矢量加性的组合在一起，经过卷积处理生成矢量输出。其次，在空间上将其扩维到 M×M，然后与局部路径中的另一个 M×M 特征图进行深度级联。然后，与生成器中完全一样，使用二进制关键点掩码对这个局部张量进行乘法门控，然后将所得的张量与 M×M×T 个关键点进行深度级联。 局部路径过步长为 2 的卷积以产生向量，然后将其与全局路径输出向量相加组合。最后经过一层处理得到一个分数。

1.4 Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space

1.4.1 创新点与研究动机

本文克服了 DGN-AM[8]的缺陷，在潜在编码上添加一个先验来生成高分辨率，多样性的图片。DGN-AM 是一种新颖的图像合成方式，它通过在生成器网络的潜在变量空间中执行梯度上升，以最大化单独分类器网络中一个或多个神经元的激活，但是这种方法虽然能够提高图像分辨率，但是图像的多样性不大。为此，论文提出了一种 MALA-approx 的迭代方法[7]，它将预训练的分类器当作编码器对图片 x 提取特征 h，之后将 h 当作输入，通过 MCMC 不断修改 h 的值，去寻找效果更好的图片。

1.4.2 方法

作者共讨论了五种生成模型（如下图所示），分别是 PPGN-x、DGN-AM、PPGN-h、Joint PPGN-h，Noiseless joint PPGN-h，其中 DGN-AM 并不是作者的新成果，但 DGN-AM 的方法与其他方法可以看作一脉相承。作者的一个结论是 Noiseless joint PPGN-h 这个生成器在 ImageNet 中表现最好。

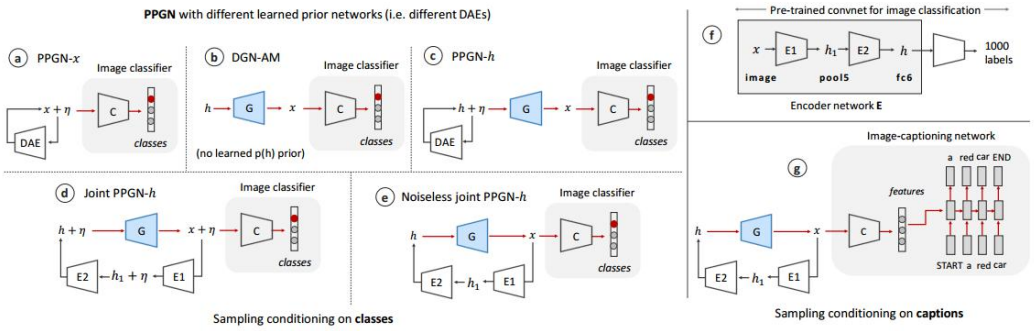


图 5：PPGN 模型的不同变体

（1）概率解释

根据 Metropolis-adjusted Langevin algorithm（MALA），可以定义 MCMC，其平稳分布近似于给定的分布 p（x），我们图片可以按照下面的随机过程产生。

$$x_{t+1} = x_t + \epsilon_{12}\nabla \log p(x_t) + N(0,\epsilon_3^2)$$

且有

$$p(x|y = y_c) \propto p(x)p(y = y_c|x)$$

经过一系列代换后有

$$x_{t+1} = x_t + \epsilon_1 \frac{\partial log p(x_t)}{\partial x_t} + \epsilon_2 \frac{\partial log p(y = y_c|x_t)}{\partial x_t} + N(0,\epsilon_3^2) \qquad (1)$$

ϵ是超参数，原文中对迭代加上的三项解释如下：

ϵ_1项：用来使当前的图片往更像数据集里所有图片的样子去变化（即往先验知识的方向去变化)。

ϵ_2项：用来使当前图片能在分类识别器上得到更高分。

ϵ_3项：鼓励产生多样性的图片。

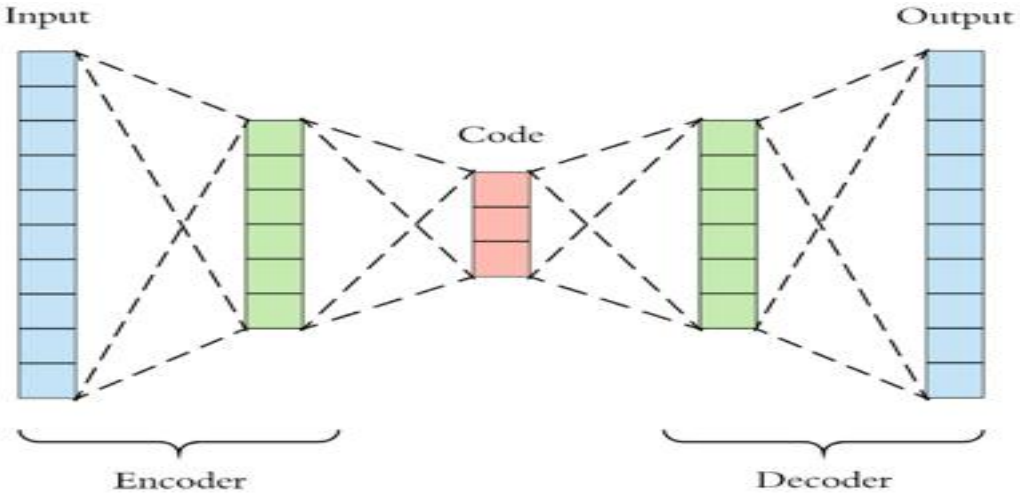
所有模型都是在（1）式的基础上产生的。

（2）模型解释

f 图的灰色部分是一个自动编码器（autoencoder）的编码（Encoder）过程，自动编码器常用来降维和特征选择。一个自动编码器 R，实际操作中我们常用神经网络学习，是一种无监督学习，如下图，分为了 Encoder 和 Decoder 两个过程，x 经过 Encoder 过程后，被编码为维度更小的 h，h 经过 Decoder 的过程，被还原为 R（x）。损失函数是 R（x）与 x 的距离。

DAE（Denoising Autoencoder）是去噪自动编码器，损失函数是 R（x+eta）与 x 的距离，eta 表示噪声。此方法被称为去噪自动编码器，是因为它对于一些

噪声明显的图片，可以很明显的去掉噪声。



图六：自编码模型

DAE 是我们预处理数据时就已经训练好的，我们记 DAE 处理后的 x 为 R（x），x 在 Encoder 后记为 h，Encoder 的位置在这里指的是第一个全连接层的位置，h1 是产生 h 的输入，由 x 到 h1 的网络记为 E1，由 h1 到 h 的网络记为 E2。全连接层产生的数据称为 fc6，全连接层输入的数据称为 pool5。

图 f 非阴影部分中还展现了把 h 带入一个分类网络中，h 会被分为是哪一类图像。

图 a：PPGN-x。对图 abcde，只看非阴影部分。整个非阴影部分其实就是我们所说的生成模型，图 bcde 中的 G 可以理解为解码器（Decoder），但是并不是 DAE 中的解码部分，而是通过 GAN 重新训练的一个神经网络。

图中的 DAE 并不是上文中的 DAE 这个网络，而是指式（1）的迭代过程, R 是去噪自动编码（在全部图像中预训练得到）。在 PPGN-x 中，我们在式（1）上进行了如下近似。

$$\frac{\partial logp(x)}{\partial x} \approx \frac{R_x(x)-x}{\sigma^2}$$

图中的η是下式中的 $N(0,\epsilon_3^2)$，$\frac{logp(y=y_c|x_t)}{\partial x_t}$是在训练集上训练出分类器（比如可以用 AlexNet VGG 等）后得到的结果。即 PPGN-x 就是将图像按下式一步步迭代。

$$x_{t+1} = x_t + \epsilon_1(R_x(x_t)-x_t) + \epsilon_2\frac{\partial logp(y=y_c|x_t)}{\partial x_t} + N(0,\epsilon_3^2)$$

图 b： FGN-AM。图中的η是下式中的 $N(0,\epsilon_3^2)$，$\frac{logp(y=y_c|x_t)}{\partial x_t}$是在训练集上训

练出分类器（比如可以用 AlexNet VGG 等）后得到的结果。即 PPGN-x 就是将图像按下式一步步迭代。

$$x_{t+1} = x_t + \epsilon_1(R_x(x_t) - x_t) + \epsilon_2 \frac{\partial log p(y = y_c|x_t)}{\partial x_t} + N(0,\epsilon_3^2)$$

图 c ：PPGN-h。与图 b 相比，不再假设高斯分布和ϵ，一方面，在每次 h 迭代中都加入了噪声项，另一方面，与 PPGN-x 类似，用$\frac{\partial log p(h)}{\partial h} \approx \frac{R_h(h)-h}{\sigma^2}$，得到下面的式子。

$$h_{t+1} = h_t + \epsilon_1(R_h(h_t) - h_t) + \epsilon_2 \frac{\partial log C_c(G(h_t))}{\partial G(h(t))} \frac{\partial G(h(t))}{\partial h_t} + N(0,\epsilon_3^2)$$

为了实现上式，除了训练一个去噪自动编码器R_x得到第一步的 h，还需要在训练集上训练一个去噪自动编码器R_h。

图 d：Joint PPGN-h。在 PPGN-h 中，h 是通过一次次迭代得到的，而在 Joint PPGN-h 中，h_{t+1}是 $G(h_t)$这张图像通过类似于 DAE 的重新编码得到的，这个编码过程与 DAE 编码过程的区别在于，在三个位置加入了噪声项。

三个噪声项分别加在了这些位置：先加噪声在 $G(h_t)$，通过E_1编码后得到 pool5 记为h_1,然后在h_1上加上噪声, 通过E_2编码后得到 h，在 h 上加上噪声。

图 e ：Noiseless Joint PPGN-h。与 Jonit PPGN-h 几乎完全相似，只是去掉了所有噪声项。

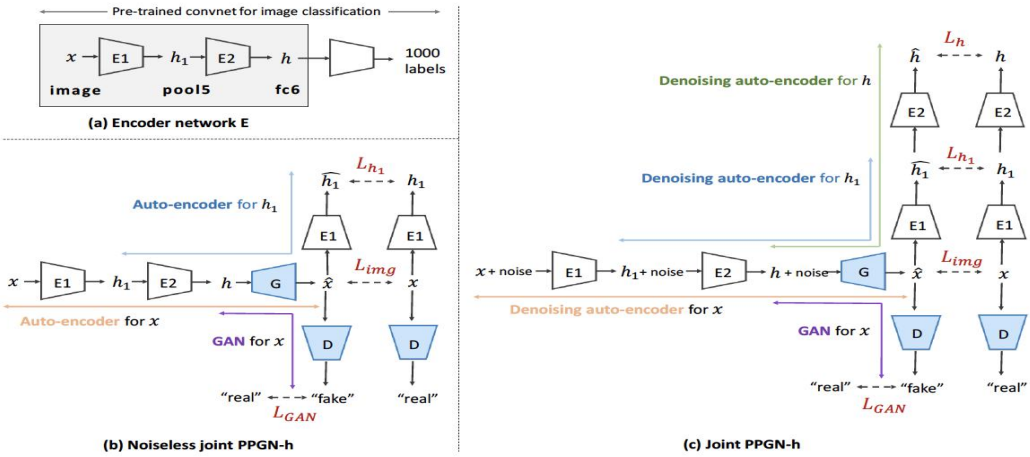
图 g：先看 abcde 中的阴影部分，新图片通过一个分类器 C 将图像分到相应的类上，g 的阴影部分说明，可以进一步将图像分类更细，最终达到按照文本特征生成图像的目的。

比如说，我们之前的模型，输入一张红色小鸟，我们想要的类别为小鸟，只要产生的图片是小鸟，我们就认为模型结果很好，至于新图像中鸟的颜色，只要与大多数小鸟颜色相似就可以，具体是不是红色我们不关心，不过我们可以在我们想得到的类别中加入红色和小鸟两个特征，新图像就可以向着这个方向产生新图像，进而我们可以描述更多的特征。

（3）解码器 G 的训练

除了 PPGN-x，都需要一个解码器 G。下面是我们要读懂的第二张图，这张图展现了 Joint PPGN-h 和 Noiselessjoint PPGN-h 中 G 的具体的训练过程，DGN-AM 和 PPGN-h 中的 G 的训练过程与之类似。G 的训练过程是在相应的图

像类中进行的。我们只以下图中的 b 为例，解释 G 的训练过程。



在图 b 中，x 被一步迭代后产生新图像$\hat{x}$, $\hat{x}$与 x 间的差距是损失L_{img}，$\hat{x}$再次编码过程中，pool5 记为$\widehat{h_1}$, $\widehat{h_1}$与h_1之间的差距是损失L_{h_1}, L_{img}与 L_{h_1}常用二范数距离，将$\hat{x}$带入一个判定器网络，判定为假的损失是L_{GAN}，通过训练 G，使得下式的损失最小。$h_i = G(x_i)$

$$L = L_{img} + L_{h_1} + L_{GAN}$$

$$L_{img} = ||\hat{x} - x| \quad |^2$$

$$L_{h_1} = ||\widehat{h_1} - h_1| \quad |^2$$

$$L_{DAN} = -\sum_i log\big(D_p(G_\theta(h_i))\big)$$

然后再训练判定器 D，使得 D 能够区分原图像与生成图像，如此循环，得到最终训练的 G。

$$L_D = -\sum_i log\big(D_p(x_i)\big) + log\big(1 - D_p(G_\theta(h_i))\big)$$

1.5 StackGAN

《StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks》一文提出了 StackGAN 结构[9]，根据给定的文字描述，生成 256×256 分辨率的真实图片。将文本生成高清图像的任务分为两个子任务：

第一个任务是通过文本生成模糊图像，第二个任务是从模糊图像生成最后的高清图像。从低分辨率图像生成的模型分布与自然图像分布相交叠的概率更高。这就是第二阶段能够生成高分辨率图像的根本原因。本文还提出条件增强技术，使隐含条件分布更平滑。

1.5.1 创新点与研究动机

主要分为两个阶段：第一阶段 GAN 根据给定的文本描述绘制对象的原始形状和颜色，生成阶段一的低分辨率图像。第二阶段 GAN 将第一阶段的结果和文本描述作为输入，修正了第一阶段生成图片的缺陷，生成具有真实细节的高分辨率图像。还引入了一种新的条件增强技术，保证在数据量有限的情况下避免数据分布空间的不连续，增强数据平滑度。最后，通过大量实验表明，该方法在生成基于文本描述的真实感图像方面取得了显著的改进。

1.5.2 方法

StackGAN 模型如下图所示：

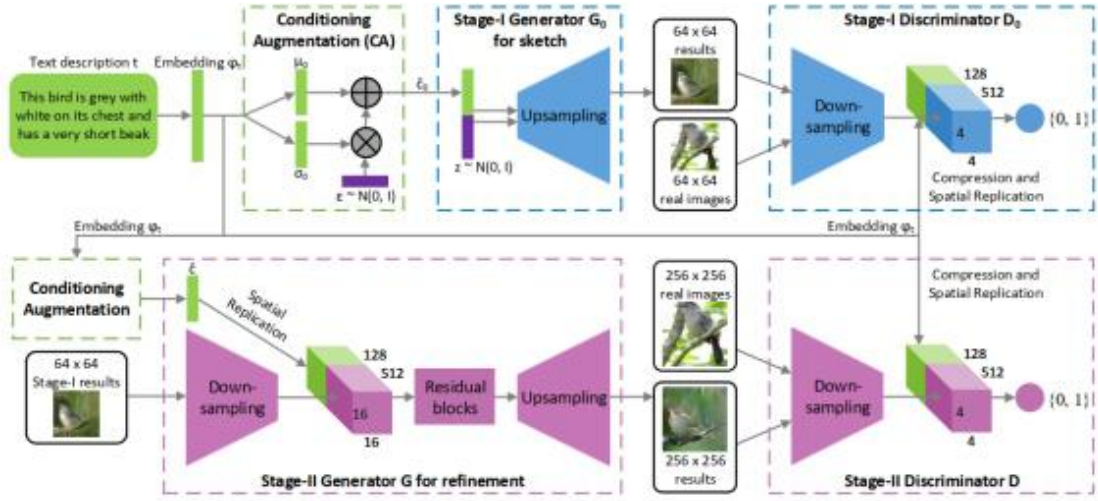


图 7. StackGAN 模型结构

Figure 7. The architecture of the proposed StackGAN.

（1）条件增强技术

首先给定一段文本描述，由编码器对文本描述 t 进行编码，得到文本嵌入向量$\boldsymbol{\varphi_t}$文本嵌入向量经过非线性变换，生成条件潜在变量作为生成器的输入。然而，文本嵌入的潜在空间通常是高维的(> 100 维)。在输入数据量有限的情况下，通

常会导致潜在数据分布空间的不连续，这是不可取的。为了解决这个问题，我们引入了一个条件增强技术来产生额外的条件变量$\hat{c}$。我们从独立的高斯分布$\boldsymbol{N(\mu(\varphi_t),\sum(\varphi_t))}$ 中随机抽取件变量$\hat{\boldsymbol{c}}$。我们从独立的高斯分布$\boldsymbol{N(\mu(\varphi_t),\sum(\varphi_t))}$中随机抽取变量$\hat{c}$。其中均值$\boldsymbol{\mu(\varphi_t)}$和协方差矩阵$\boldsymbol{\sum(\varphi_t}$在给定少量的图像文本对的情况下，提出的条件增强方法可以产生更多的训练数据对，从而增强了对小扰动的鲁棒性。

为了进一步加强调节数据分布空间的平稳性，避免过拟合，本文在生成器的目标函数上增加了以下正则化项：$\boldsymbol{D_{KL}(N(\mu(\varphi_t),\sum(\varphi_t))||N(0,I))}$。条件增强中引入的随机性有利于文本到图像的建模，因为同一句话通常会有不同的理解。

（2）Stage-I GAN

第一阶段 GAN 没有直接生成高分辨率图像，而是首先生成低分辨率图像。Stage-I GAN 主要用于绘制对象的粗略形状和正确的颜色。设$\boldsymbol{\varphi_t}$为文本嵌入向量，本文采用预先训练好的编码器生成。文本嵌入的高斯条件变量$\hat{\boldsymbol{c}}_0$取$\boldsymbol{N(\mu_0(\varphi_t),\sum_0(\varphi_t))}$来捕获$\varphi_t$变化的意义。第一阶段 GAN 以$\hat{\boldsymbol{c}}_0$和随机变量 z 为条件，通过最大化 LDO 和最小化 LGO 来训练判别器 DO 和生成器 GO。目标函数如下：

$$\boldsymbol{\mathcal{L}_{D_0}} = \mathbb{E}_{(\boldsymbol{I_0},\boldsymbol{t})\sim\boldsymbol{p_{data}}}[\boldsymbol{logD_0(I_0,\varphi_t)}]$$

$$+\ \mathbb{E}_{\boldsymbol{z}\sim\boldsymbol{p_z},\boldsymbol{t}\sim\boldsymbol{p_{data}}}[\log(\boldsymbol{1-D_0(G_0(z,\hat{c}_0),\varphi_t)})]$$

$$\boldsymbol{L_{G_0}} = \mathbb{E}\boldsymbol{z}\sim\boldsymbol{p_z},\boldsymbol{t}\sim\boldsymbol{p_{data}}[\boldsymbol{log(1-D_0(G_0(z,\hat{c}_0),\varphi_t))}]$$

$$+\lambda\boldsymbol{D_{KL}(N(\mu_0(\varphi_t),\sum(\varphi_t))||N(0,I))}$$

（3）Stage-II GAN

第一阶段 GAN 生成的低分辨率图像通常缺少鲜明的的对象特征，并可能包含形状变形。同时，文本中的重要细节也没有表现出来。我们的第二阶段 GAN 是建立在第一阶段 GAN 的结果上，以产生高分辨率的图像。第二阶段 GAN 以 Stage-I GAN 的生成结果$\boldsymbol{s_0 = G_0(z,\hat{c}_0)}$和高斯变量$\hat{\boldsymbol{c}}_0$为输入，通过最大化 LD 和最小化 LG 来训练第二阶段 GAN 中的判别器 D 和生成器 G。

$$\boldsymbol{\mathcal{L}_D} = \mathbb{E}_{(\boldsymbol{I},\boldsymbol{t})\sim\boldsymbol{p_{data}}}[\boldsymbol{logD(I,\varphi_t)}]$$

$$+\ \mathbb{E}_{\boldsymbol{s_0}\sim\boldsymbol{p_{G_0}},\boldsymbol{t}\sim\boldsymbol{p_{data}}}[\log(\boldsymbol{1-D(G(s_0,\hat{c}\ \ ),\varphi_t)})]$$

$$\mathcal{L}_{G_0} = \mathbb{E}_{s_0 \sim p_{G_0}, t \sim p_{data}}[\log(1 - D(G(s_0, \hat{c}\ ), \varphi_t))]$$
$$+ \lambda D_{KL}(N(\mu(\varphi_t), \sum(\varphi_t)) || N(0, I))$$

本阶段没有使用随机噪声 z，假设随机度已经被s_0保留。本阶段使用的高斯变量$\hat{c}_0$和第一阶段 GAN 使用的$\hat{c}_0$使用相同的预训练文本编码器，生成相同的文本嵌入 t。然而，阶段 i 和阶段 ii 条件作用增强具有不同的全连通层，生成不同的均值和标准差。

1.6 AttnGAN

2018 年，Xu 等人发表的“AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks“论文对先前（2017 年）提出的 AttnGAN 模型进行了改善。通过引入注意生成网络，AttnGAN 可以聚焦自然语言描述中的相关单词来合成图像不同区域的细粒度细节。此外，该文提出了一种深度注意多模态相似模型来计算细粒度图像--文本匹配损失，用于生成器的训练。

1.6.1 创新点与研究动机

AttnGAN 模型[11]（如图 8）包括两个部分：

（1）注意力生成网络(Attentional Generative Network)，该网络的注意力机制自动让生成器通过聚集与所绘制的图像子区域最相关的单词来绘制图像的不同子区域。该模型不单单将自然语言编码为一个全局的句子变量，而且句子中的每一个单词也被编码为单词向量。开始阶段，该网络利用全局句子向量生成一个低分辨率的图像。接下来，利用每个子区域中的图像向量和注意层形成的单词语境向量来检查单词向量，然后将区域图像向量和相应的单词语境向量结合形成多模态语境向量。在此基础上，模型会在周围的子区域中生成新的图像特征

（2）深度注意多模态相似模型(Deep Attentional Multimodal Similarity Model)，通过注意力机制，DAMSM 能够利用全局句子级信息和细粒度单词级信息来计算生成的图像和句子之间的相似度。另外，利用相似度可以计算图像文本匹配的损失，并以此来训练生成器。

模型整体结构：

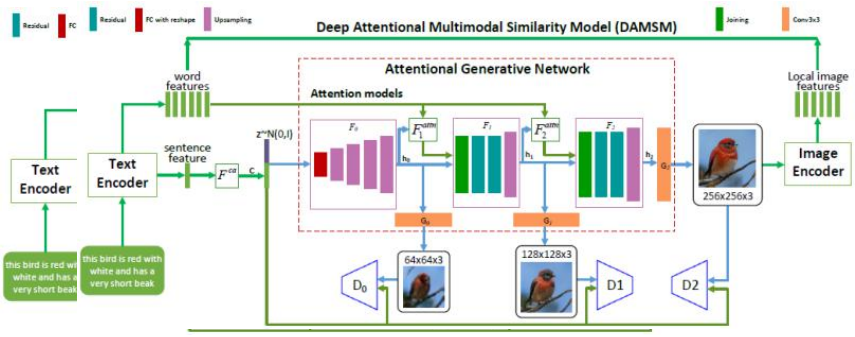


图 8：AttnGAN 模型结构

1.6.2 方法

（1）带有注意力机制的生成网络(Attentional Generative Network)

第一阶段$h_0 = F_0(z,F^{ca}(\bar{e}))$，基于$h_0$可以生成第一个阶段的图片。$G_0$是生成器 z 表示噪声的输入，服从标准正态分布；$\bar{e}$表示全局句子向量；$h_0$是隐藏状态；$F^{ca}$表示条件增强函数，把句子向量$\bar{e}$转化为条件向量；$x_0$为得到的一个低分辨率图像。

后续阶段 $x_0 = G_0(h_0)$ ； $h_i = F_i(h_{i-1},F_i^{attn}(e,h_{i-1}))$ ； $x_i = G_i(h_i)$ 。z 被代替为h_{i-1}；e 是单词向量矩阵；F_i^{attn}是第 k 阶段的注意力模型，该模型有两个输入：单词特征 e 和前一个隐藏层里的图像特征h_{i-1}。

第 j 个子区域的单词语境向量是与h_j相关的单词向量的动态展示：

$$c_j = \sum_{i=0}^{T-1} \beta_{j,i} e_i^{'} \quad \beta_{j,i} = \frac{exp(s_{j,i}^{'})}{\sum_{k=0}^{T-1} exp(e_{j,k}^{'})} \quad s_{j,i}^{'} = h_j^T e_i^{'}$$

$\beta_{j,i}$表示模型产生第 j 个子区域时侧重于第 i 个单词的权重。

最后，图像特征和相应的单词语境特征组合起来用于下一阶段的图像生成。

（2）深度注意多模态相似模型(Deep Attentional Multimodal Similarity Model)

DAMSM 学习两个神经网络，图像的子区域和句子中的单词映射到一个共同的语义空间，进而计算图像生成的细粒度损失，从而在单词级度量图像文本的相似性。

a)文本编码器是一种双向长短期记忆器(LSTM)，它可以从文本描述中提取语义向量。

b) 图像编码器是将图像映射到语义向量的卷积神经网络(CNN)。CNN 的中

间层学习图像不同子区域的局部特征，而后层学习图像的全部特征。图像编码器是基于在 ImageNet 上预先训练的 inception-v3 模型构建的。

c)注意力驱动的图像文本匹配分数

该分数是用来度量图像文本之间匹配程度，首先计算所有可能的单词与子区域匹配的相似度矩阵：$s = e^T v$。公式$\bar{s}_{i,j} = \frac{exp(s_{i,j})}{\sum_{k=0}^{T-1} exp(s_{k,j})}$中$s_{i,j}$是句子中第 i 个单词和第 j 个子区域的点积，它可以标准化相似矩阵。

然后，建立一个注意模型来计算每个单词的区域语境向量。区域语境向量c_i是图像子区域与句子中的第 i 个单词相关的动态表示，它是所有区域视觉向量的加权和：$c_i = \sum_{j=0}^{288} a_j v_j \quad a_j = \frac{exp(\gamma_1 \bar{s}_{i,j})}{\sum_{k=0}^{288} exp(\gamma_1 \bar{s}_{i,k})}$。$\gamma_1$是参数，决定了当计算一个单词的区域语境向量时关注相关子区域特征的程度大小。

最后，我们用c_i和e_i的余弦定义第 i 个单词和第 i 个子区域的相关度：

注意力驱动的图像文本匹配分数(整体图像 Q 和整个文本描述 D)定义为：

$$R(c_i,e_i) = \frac{c_i^T e_i}{||c_i||\,||e_i||} \qquad R(Q,D) = log(\sum_{i=1}^{T-1} exp(\gamma_2 R(c_i,e_i)))^{\frac{1}{\gamma_2}}$$

γ_2是参数，决定了扩大多少最相关子域的重要性，当γ_2趋于无穷大时，R(Q,D)约等于$\max_{i=1}^{T-1} R(c_i,e_i)$。

d) DAMSM 损失

DAMSM 旨在以半监督的方式学习注意力模型，其中唯一的监督是整个图像与整个句子之间的匹配。对于一些图像句子匹配$\{(Q_i,D_i)\}_{i=1}^{M}$,可以计算D_i关于图像Q_i的后验概率：

$$P(D_i|Q_i) = \frac{exp(\gamma_3 R(Q_i,D_i))}{\sum_{j=1}^{M} exp(Q_i,D_j)}$$

γ_3是平滑参数，由实验确定。在这些句子中，只有D_i与图像Q_i匹配，将所有其他 M-1 个句子视为不匹配的描述。我们将损失函数定义为图像与对应的文本匹配的负对数后验概率：$L_1^w = -\sum_{i=1}^{M} logP(D_i|Q_i)$。

w 代表单词，同理我们也最小化：$L_2^w = -\sum_{i=1}^{M} logP(Q_i|D_i) \quad P(Q_i|D_i) = \frac{exp(\gamma_3 R(Q_i,D_i))}{\sum_{j=1}^{M} exp(\gamma_3 (Q_j,D_i))}$ 是句子D_i和与它相关的图像Q_i匹配的后验概率，我们可以利用句

子向量$\bar{e}$和全局图像向量$\bar{v}$得到损失函数L_1^s和L_2^s。

最后，DAMSM 损失定义为：$L_{DAMSM} = L_1^w + L_2^w + L_2^s + L_2^s$。

1.7 Text2Scene

2019 年， Tan 等人在“Generating Compositional Scenes from Textual Descriptions”论文中提出了 Text2Scene 模型[10]，一个从自然语言描述中合成场景的框架。该模型通过关注输入文本的不同部分和生成场景的当前状态，学会在每一个时间步上依次生成对象及其属性（位置、大小、外观等）。与最先进的使用自动度量的 GANS 和基于人类判断的高级方法相比，Text2Scene 方法具有可解释结果的优势。

1.7.1 创新点与研究动机

调整和训练模型可以生成三种类型的场景：抽象场景、图像场景相对应的对象布局（COCO 数据集）和针对对应的图像合成场景（coco 数据集中）。

用了一个可解释性模型，通过每一个时间步预测和添加新对象来迭代生成场景。在自动评价指标和人工评价的时候均表现了最好的性能。

生成文本描述图像（如图 9）是一个非常有挑战的工作。输入文本描述间接暗示属性的存在，比如输入 MIKE IS SURPRISED, 应该是 mike 非常吃惊，需要在面部表情有所体现。而且文本描述往往包含了包括相对空间复杂的信息。例如，输入 jenny 向着 mike 和鸭子跑去，这里面 jenny 的方向依赖于后面两个对象的空间位置，在最后一个例子，大象走在一条直线上，也暗示了相对空间信息。

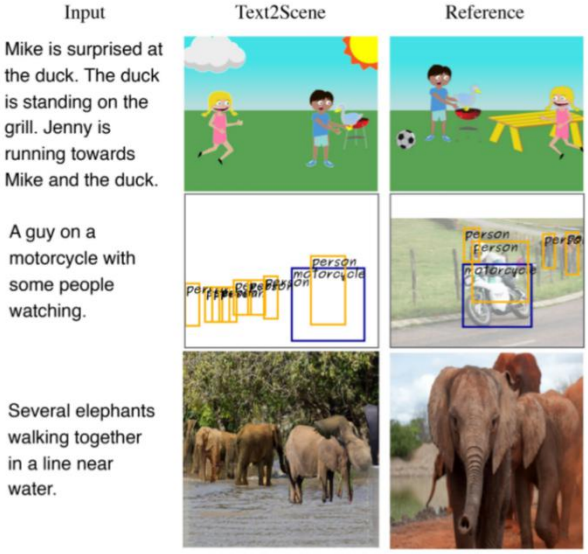


图 9：Text2Scene 模型输入输出示例

1.7.2 方法

Text2Scene 采用了一个 seq-seq 框架，并介绍了空间推理和顺序推理的关键设计。具体来说，在每一个时间步骤中，模型通过以下三个步骤来修改背景画布：

模型关注输入文本，以决定下一个要添加的对象是什么，或者决定生成是否应该结束;

如果决定添加一个新对象，则模型在该对象的语言上下文中进行缩放，以决定其属性(如姿态、大小)和与周围环境的关系(如位置、与其他对象的交互);

模型将提取出的文本属性返回到画布和场景中，并将其转换为相应的视觉表示。

为了对这个过程进行建模，Text2Scene 包含一个文本编码器，它以 M 个单词作为输入序列；包含一个对象解码器，它可以预测顺序的第 T 个对象 Ot；一个属性解码器，它可以预测每个对象的属性。

场景生成从最初的空画布开始，每个时间步长更新它。在合成图像生成任务中，该模型还联合训练了一个前向嵌入网络，并将嵌入的向量作为目标属性。下图展示了一个抽象场景的步进式生成。

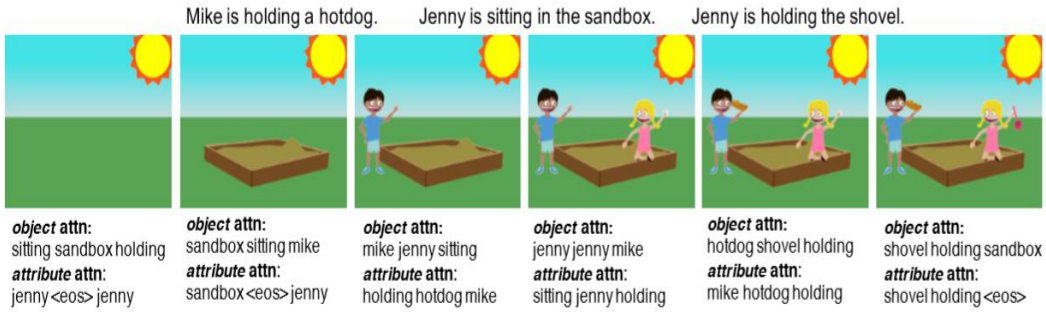


图 10：一步步生成抽象场景

该模型的总览如图 11。Sequence to sequence 方法将对象放在了一个空白的画布上，Text2Scene 有一个文本编码器 A，可以映射句子的潜在表示，为输入提供一系列的表征；接着是一个图像编码器 B，为目前状态的生成场景编码，生成当前的画布；接着是一个卷积循环模块 C，用于追踪空间位置，目前已经生成的历史，可以将当前的状态传给下一个步骤。D 是注意力模块，集中于输入文本的不同部分，连续不断地集中于输入文本的不同部分；E 是一个对象解码器，可以根据当前场景状态于已参与的输入文本预测下一个对象，可以决定放什么对象。

F 是一个属性解码器，可以为预测对象分配属性。还有一个可选的前向嵌入 G，用来学习合成图像生成任务中批量检索外表特征。

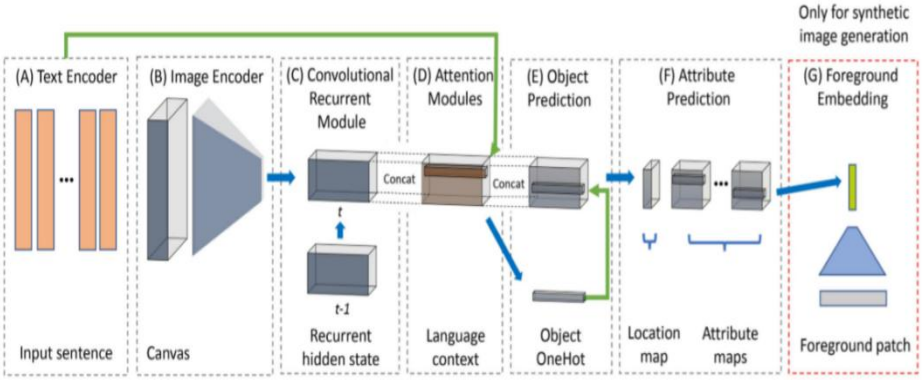


图 11：Text2Scene 模型总览图

2 结论

本文分析了近 3 年来主流的方法，作者们以 RNN 和 GAN 为基础，提出了众多性能良好的模型，可以看到文本生成图片的多样性和质量在不断提高。

第一种方法使用自动循环编码和注意力机制迭代的生成图像，但是由于变分编码器只是在计算生成图片和原始图片的均方误差，使得生成的图片的分辨率不高，只有 36*36。

第二种方法使用了一种普通的 GAN 结构，但是作者增加了一些训练技巧，使得可以生成分辨率为 64*64 的图像。

第三种方法在 GAN 框架下借助目标对象的额外信息（位置和尺寸）来提升生成图像的质量和对文本解释的质量，最终产生了 128*128 的图像。

第四种方法通过增加先验信息，用预训练的分类器当作编码器对图像提取特征 h，并将 h 作为输入，通过迭代不断修改 h 的值，最终生成了 227*227 分辨率的图像。

第五种方法分为两个阶段的生成，第一阶段的 GAN 只生成初步的低分辨率图像，第二阶段通过纠正第一阶段的错误，进一步生成高分辨率图像，最终生成了 256*256 的高分辨率图像。

第六种方法解决了 StackGNN 不能端到端训练的问题，通过引入注意力生成网络和深度多模态相似模型可以生成 256*256 的高质量图像并且注意到文本中的单词粒度细节。

第七种方法没有采用 GAN 结构，而是使用了一种 Seq2Seq 的模型结构，可以依据语言文本描述生成抽象的图像表示（如卡通图片），图像可以包含每个对象的属性，包括姿势，表情等信息。

虽然当前文本生成图片已经取得了一定的效果，但不可否认这一任务仍然是具有挑战性的，对复杂的场景，尤其是包括多个对象并且对象之间有交互的场景，建模仍然是困难的。

3 展望

由自然语言描述自动生成图像是许多应用程序中的基本问题，例如艺术生成和计算机辅助设计。它推动了跨模式学习和跨视觉和语言推理的研究进展，是近年来的研究热点。当前，诸如 StackGAN，AttnGAN 已经可以获得高质量的图像，但是目前的研究还存在一些不足之处：一是优化过程中存在不稳定性，很容易坍缩到一个鞍点上；二是 GAN 的可解释性比较差；三是需要提高训练过程中的稳定性和 GAN 模型的延展性，尤其在处理大规模数据的时候。相信随着更多出色的工作被提出，文本生成图像任务将会越来越成熟，并在实际应用中发挥作用。

参考文献：

[1] Mansimov E, Parisotto E, Ba J L, et al. Generating images from captions with attention[J]. arXiv preprint arXiv:1511.02793, 2015.

[2] Gregor K, Danihelka I, Graves A, et al. Draw: A recurrent neural network for image generation[J]. arXiv preprint arXiv:1502.04623, 2015.

[3] Reed S, Akata Z, Yan X, et al. Generative adversarial text to image synthesis[J]. arXiv preprint arXiv:1605.05396, 2016.

[4] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Advances in neural information processing systems. 2014: 2672-2680.

[5] Reed S E, Akata Z, Mohan S, et al. Learning what and where to draw[C]//Advances in Neural Information Processing Systems. 2016: 217-225.

[6] Reed S, Akata Z, Lee H, et al. Learning deep representations of fine-grained visual descriptions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 49-58.

[7] Nguyen A, Clune J, Bengio Y, et al. Plug & play generative networks: Conditional iterative generation of images in latent space[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4467-4477.

[8] Nguyen A, Dosovitskiy A, Yosinski J, et al. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks[C]//Advances in Neural Information Processing Systems. 2016: 3387-3395.

[9] ZHANG H，XU T，LI H，et al. Stack GAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks[J]//Computer Vision and Pattern ecognition，2016，ar Xiv: 1612. 03242

[10] Tan F, Feng S, Ordonez V. Text2Scene: Generating Compositional Scenes From Textual Descriptions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 6710-6719.

[11] Xu T, Zhang P, Huang Q, et al. Attngan: Fine-grained text to image generation with attentional generative adversarial networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1316-1324.

[12] Scheiter,K.,Schüler,A.,andEitel,A.(2017).“Learningfrommultimedia:Cognitive processes and Instructional Support,” in The Psychology of Digital Learning, eds Schwan and U. Cress (Cham: Springer International Publishing), 1–19.

[13] Klepsch, M., Schmitz, F., and Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. Front. Psychol. 8:1997. doi: 10.3389/fpsyg.2017.01997

[14] Renkl, A., and Scheiter, K. (2017). Studying visual displays: how to instructionally support learning. Educ. Psychol. Rev. 29, 599–621. doi: 10.1007/s10648-015-9340-4

[15] Seufert, T. (2018). The interplay between self-regulation in learning and cognitive load. Educ. Res. Rev. 24, 116–129. doi: 10.1016/j.edurev.2018.03.004