AttnGAN：具有注意生成对抗网络的细粒度文本到图像生成

徐涛、张鹏川、黄秋源、，[[1]](" \l "_ftn1" \o ")1

韩章3、哲干4、黄晓雷1、小东5

1利海大学微软研究院罗格斯大学杜克大学JD AI研究院2345

{tax313，xih206}@lehigh。edu，{penzhan，qihua，xiaohe}@microsoft。韩康。zhang@cs.rutgers.edu哲。gan@duke.edu，晓东。he@jd.com

# 摘要

*在本文中，我们提出了一种注意力生成对抗网络（AttnGAN），它允许注意力驱动、多阶段细化细粒度文本到图像的生成。通过一种新的注意生成网络，AttnGAN可以通过关注自然语言描述中的相关词语，合成图像不同子区域的细粒度细节。此外，本文还提出了一个深度注意多模态相似性模型来计算细粒度图像文本匹配损失，以训练生成器。拟议的AttnGAN显著优于之前的最新水平，在CUB数据集上，最佳报告初始得分提高了14.14%，在更具挑战性的COCO数据集上，提高了170.25%。通过可视化AttnGAN的注意层，也可以进行详细的分析。这是第一次表明分层注意GAN能够自动选择单词级的条件来生成图像的不同部分。*

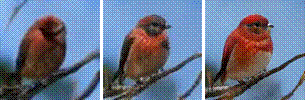
# 1.  介绍

根据自然语言描述自动生成图像是许多应用中的一个基本问题，例如艺术生成和计算机辅助设计。它还推动了跨视觉和语言的多模态学习和推理的研究进展，这是近年来最活跃的研究领域之一[20,18,36,19,41,4,30,5,1,31,33,32]

|  |
| --- |
|  |
|  |  |

最近提出的文本到图像合成方法是基于生成性对抗网络（GAN）[6]。一种常用的方法是将整个文本描述编码为全局句子向量，作为基于GAN的图像生成的条件[20,18,36,37]。尽管已经给出了令人印象深刻的结果，但是

这只鸟红着白，嘴很短



          10:短3:红色11:喙9:非常8:a



          3:红色5:白色1:鸟10:短0:这个



图1。建议的AttnGAN的示例结果。第一行给出由AttnGAN和生成的低至高分辨率图像；第二行和第三行显示了前五名最受关注的单词*G*0*G*1 *G*2 分别是阿廷根。在这里，和的图像被双线性上采样，以具有与的相同的大小，以便更好地可视化。*G*0 *G*1 *G*2

仅在全局句子向量上，在单词级别上缺少重要的细粒度信息，并阻止高质量图像的生成。当生成复杂场景（如COCO数据集中的场景）时，这个问题变得更加严重[14]。

为了解决这个问题，我们提出了一种注意力生成对抗网络（AttnGAN），它允许注意力驱动、多阶段细化细粒度文本到图像的生成。AttnGAN的总体架构如图2所示。该模型由两个新组件组成。第一个组成部分是注意生成网络，在这个网络中，注意机制被开发出来，用于发生器绘制大脑的不同子区域并通过聚焦于与正在绘制的子区域最相关的单词来获得图像（见图1）。更具体地说，除了将自然语言描述编码为全局句子向量外，句子中的每个单词也编码为单词向量。生成网络在第一阶段利用全局句子向量生成低分辨率图像。在接下来的步骤中，它使用每个子区域中的图像向量，通过使用注意层来形成单词上下文向量来查询单词向量。然后，该模型将区域图像向量和相应的单词上下文向量相结合，形成多模态上下文向量，并在此基础上在周围的子区域生成新的图像特征。这有效地产生了更高分辨率的图片，每个阶段都有更多细节。AttnGAN的另一个组成部分是深度注意多模态相似模型（DAMSM）。通过注意机制，DAMSM能够利用全局句子级信息和细粒度单词级信息计算生成的图像和句子之间的相似度。因此，DAMSM为训练生成器提供了额外的细粒度图像文本匹配损失。

我们的方法的贡献是三方面的。（i） 提出了一种注意生成对抗网络，用于从文本描述合成图像。具体而言，AttnGAN提出了两个新的组成部分，包括注意生成网络和DAMSM。（ii）进行综合研究，以实证评估拟议的AttnGAN。实验结果表明，AttnGAN的性能明显优于以前最先进的GAN模型。（iii）通过可视化AttnGAN的注意层进行详细分析。首次证明了分层条件GAN能够自动关注相关单词，形成图像生成的条件。我们的代码可在https://github.com/taoxugit/AttnGAN.

# 2.  相关工作

从文本描述生成高分辨率图像，虽然非常具有挑战性，但对于许多实际应用（如艺术生成和计算机辅助设计）来说非常重要。最近，随着深层生成模型的出现，在这方面取得了巨大进展[12,27,6]。Mansimov等人[15]建立了alignDRAW模型，扩展了深度反复注意力书写器（DRAW）[7]，在关注标题中的相关单词的同时，迭代绘制图像补丁。Nguyen等人[16]提出了一种近似的Langevin方法来从字幕生成图像。Reed等人[21]使用条件像素CNN[27]从具有多尺度模型结构的文本合成图像。与其他深度生成模型相比，生成性对抗网络（GANs）[6]在生成更尖锐的样本方面表现出了卓越的性能[17、3、23、13、10、35、24、34、39、40]。Reed等人[20]首先表明，条件式GAN能够从文本描述中合成合理的图像。他们的后续工作[18]还表明，GAN能够通过加入附加条件（例如，物体位置）生成更好的样品。Zhang等人[36,37]将多个GANs堆叠起来用于文本到图像的合成，并使用不同的GANs生成不同大小的图像。然而，它们的所有GAN都以全局句子向量为条件，缺少用于图像生成的细粒度单词级信息。

注意机制最近已成为序列转导模型的一个组成部分。它已成功地用于建模图像字幕[30,38]、图像问答[31]和机器翻译[2]中的多级依赖关系。Vaswani等人[28]还证明，机器翻译模型可以通过单独使用注意模型来实现最先进的结果。尽管对图像合成机制的研究还没有取得进展，但对图像合成的研究还没有得到重视。值得一提的是，alignDRAW[15]还使用拉普根[3]将图像缩放到更高的分辨率。然而，在他们的框架中，GAN仅作为后处理步骤使用，没有引起注意。据我们所知，拟议的AttnGAN首次开发了一种注意力机制，使GANs能够通过多级（例如，单词级和句子级）条件作用生成细粒度的高质量图像。

# 3.      注意生成对抗网络

如图2所示，提出的注意生成对抗网络（AttnGAN）有两个新的组成部分：注意生成网络和深度注意多模态相似模型。我们将在本节的其余部分中详细阐述每一项。

## 3.1.  注意生成网络

当前基于GAN的文本到图像生成模型[20、18、36、37]通常将整个句子文本描述编码为单个向量，作为图像生成的条件，但缺少细粒度的单词级信息。在本节中，我们提出了一种新的注意模型，该模型使生成网络能够根据与这些子区域最相关的单词绘制图像的不同子区域。

如图2所示，提出的注意生成网络具有生成器(−), 以隐藏状态为例(−) 作为输入并生成小到大比例（ˆ0，xˆ1，…，xˆm）的图像−). 明确地*MG*0*G*1*,...,转基因的*1*H*0*H*1*,...,陛下*1*十、*1

*H*0 =F0（z，Fca（e））；

1.

(1)

|  |
| --- |
| 图2。拟议的AttnGAN的架构。每个注意模型自动检索用于生成图像不同子区域的条件（即，最相关的词向量）；DAMSM为用户提供细粒度图像文本匹配丢失 |

生成网络。

这里是通常从标准正态分布采样的噪声向量。是一个全局句子向量，是单词向量的矩阵。表示将句子向量转换为条件向量的条件增强[36]。是在AttnGAN阶段提出的注意力模型，并将其建模为神经网络。*Z~~E~~EFca~~E~~菲亚特伊思Fca菲亚特Fi胃肠道*

注意模型（e，h）有两个输入：单词特征∈ 将图像特征从⇥ 上一个隐藏层∈ R.通过添加一个新的感知器层，即，=Ue，首先将单词特征转换为图像特征的公共语义空间，其中∈然后，根据图像的每个子区域的隐藏特征（查询）计算单词上下文向量。图像的每一列都是图像子区域的特征向量。对于子区域，其wordcontext向量是与相关的单词向量的动态表示，通过以下公式计算：



表示生成图像子区域时模型关注单词的权重。然后，我们通过（e，h）=（c0，c1，…，cN）为图像特征集提供单词上下文矩阵−) ∈ 最后，在下一阶段，将图像特征和相应的单词上下文特征结合起来生成图像。

为了生成具有多层次（即句子层次和单词层次）条件的真实图像，注意生成网络的最终目标函数定义为

这里是一个超参数，用于平衡等式（3）的两项。第一项是GAN损失，它共同近似于条件和无条件分布[37]。在AttnGAN阶段，发生器具有相应的鉴别器。的对抗性损失定义为：*λ 伊思胃肠道Di胃肠道*

*,*

(4)

其中，无条件损失决定图像是真是假，而条件损失决定图像和句子是否匹配。

作为对的训练的替代，每个鉴别器都经过训练，通过最小化由定义的交叉熵损失，将输入分类为真或假*胃肠道Di*



无条件损失

*,*

条件损失

(5)

式中，来自于比例下的真实图像分布，而来自于相同比例下的模型分布。AttnGAN的鉴别器在结构上是不相交的，所以它们可以并行训练，并且每个鉴别器都聚焦于单个图像尺度。

式（3）的第二项L是由DAMSM计算的字级细粒度图像文本匹配损失，将在第3.2小节中详细说明。

## 3.2.  深度注意多模态相似模型

DAMSM学习两个神经网络，将句子的图像和单词的子区域映射到一个公共语义空间，从而在单词级别测量图像-文本相似度，以计算图像生成的细粒度损失。

文本编码器是一种双向长短时记忆（LSTM）[25]，它从文本描述中提取语义向量。在双向LSTM中，每个单词对应两个隐藏状态，每个方向一个。因此，我们将其两个隐藏状态连接起来以表示单词的语义。所有单词的特征矩阵由∈ 其列是单词的特征向量。是单词向量的维数，是单词数。同时，双向LSTM的最后隐藏状态被连接为全局语句向量，表示为∈ R*ED*⇥T*伊思工程安装伊思DT~~E~~D*

图像编码器是一个卷积神经网络（CNN），它将图像映射到语义向量。CNN的中间层学习图像不同子区域的局部特征，而后面的深层学习图像的全局特征。更具体地说，我们的图像编码器是基于ImageNet[22]上预训练的Inception-v3模型构建的。我们首先将输入图像重新缩放为299×299像素。然后，我们提取局部特征ma 特里克斯∈ R（从768×17×17重塑）来自Inception-v3的“6e”层。图像的每一列都是图像子区域的特征向量。768是局部特征向量的维数，289是图像中的子区域数。同时，全球特征*F*768 289 *混合的F*矢量∈ Ris是从Inception-v3的最后一个平均池层提取的。最后，我们通过添加感知器层将图像特征转换为文本特征的公共语义空间：

*五、*=Wf，v=Wf，(6)

哪里∈ Rand its列是图像子区域的视觉特征向量；和∈ Ris是整个图像的全局向量。是多模态（即图像和文本模态）特征空间的维度。为了提高效率，从Inception-v3模型构建的层中的所有参数都是固定的，新添加的层中的参数将与网络的其余部分联合学习。

注意驱动的图像-文本匹配分数是根据图像和文本之间的注意模型来衡量图像-句子对的匹配程度。

我们首先计算句子中所有可能的词对和图像中的子区域的相似度矩阵

*s*=ev，T(7)

哪里∈ Rand是句子中的单词与图像子区域之间的点积相似度。我们发现如下对相似矩阵进行规范化是有益的*sT*⇥289*si，j伊思jth*

*.* (8)

然后，我们建立一个注意模型来计算每个单词（查询）的regioncontext向量。区域上下文向量是与句子单词相关的图像子区域的动态表示。它被计算为所有区域视觉向量的加权和，*词伊思*

*即*.,

*,* 哪里

这里，是一个决定在计算单词的区域上下文向量时对其相关子区域的特征关注程度的因素。*γ*1

最后，我们使用和之间的余弦相似性来定义单词和图像之间的相关性，即（ci，ei）=（cTi ei）/（| | ci | | | ei |）。受语音识别中最小分类错误公式（参见，例如[11，8]）的启发，整个图像（）和整个文本描述（）之间的注意力驱动图像-文本匹配分数定义为*伊思词工程安装RQD*

*,* (10)

其中有一个因素决定了最相关单词对regioncontext对的重要性要放大多少。什么时候→ ∞, （Q，D）近似于*γ*2 *γ*2 *R*

.

DAMSM损失旨在以半监督方式学习注意模型，其中只有-

pervision是整个图像和整个句子（一系列单词）之间的匹配。类似于[4,9]]，对于一批图像句子对，句子的后验概率*Di*与图像匹配的计算公式为*气*

*,* (11)

其中是由实验确定的平滑因子。在这批句子中，仅匹配图像，并处理所有其他− 1.描述不匹配的句子。在[4,9]之后，我们将损失函数定义为负对数后验概率，即图像与其对应的文本描述（基本事实）相匹配，即。，*γ*3 *Di气M*

*M*

|  |
| --- |
| L |

                                     1w=−XlogP（迪奇），（12）

*我*=1

其中“w”代表“word”。

对称地，我们也最小化

*M*

|  |
| --- |
| L |

                                     2w=−XlogP（Qi | Di），（13）

*我*=1

那句话的后验概率是多少*Di*与相应的图像匹配。如果我们用（Q，D）重新定义等式（10）=*气R*

%vTe&/%| | | | | | | | | | | | | | e | | |并将其替换为等式（11）、（12）和

(13)，我们可以使用句子向量获得损失函数和（其中s代表“句子”）*~~E~~*和全局图像向量。*~~五、~~*

最后，DAMSM损失定义为：

*.* (14)

基于一个持有的验证集上的实验，我们将本节中的超参数设置为：=5、=5、=10和=50。我们的DAMSM通过最小化真实图像-文本对进行预训练。由于预训练DAMSM的图像大小不受可生成图像大小的限制，因此使用大小为299×299的真实图像。此外，DAMSM中的预训练文本编码器为注意生成网络提供了从图像-文本配对数据中学习的视觉辨别词向量。相比之下，在纯文本数据上预训练的传统词向量通常在视觉上不具有辨别性，例如，不同颜色的词向量，例如红色、蓝色、黄色等，由于缺乏与实际视觉信号的接地，通常在向量空间中聚集在一起。*γ*1 *γ*2 *γ*3 *M*[[2]](" \l "_ftn2" \o ") *达姆*

总之，我们提出了两种新的注意模型，注意生成网络和DAMSM，它们在AttnGAN中扮演着不同的角色。（i） 生成网络中的注意机制（见等式2）使AttnGAN能够自动选择字级条件以生成图像的不同子区域。（ii）通过注意机制（见等式9），DAMSM能够计算细粒度文本图像匹配损失L。值得一提的是，Lis仅应用于最后一个生成器的输出−, 因为AttnGAN的最终目标是由最后一个生成器生成大图像。我们尝试应用Lon生成的所有分辨率的图像(−). 然而，性能并没有得到改善，但计算量却增加了。*达姆达姆转基因的*1*达姆G*0*G*1*,...,转基因的*1

# 4.  实验

进行了大量试验，以评估拟定的AttnGAN。我们首先研究AttnGAN的重要组成部分，包括注意生成网络和DAMSM。然后，我们将我们的AttnGAN与先前用于文本图像合成的最先进GAN模型进行比较[36,37,20,18,16]。

数据集。与前面的文本到图像方法[36,37,20,18]相同，我们的方法是在CUB[29]和COCO[14]数据集上评估的。我们按照[36]中的方法对CUB数据集进行预处理。表1列出了数据集的统计数据。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据集 | 幼崽[29] | | 可可[14] | |
| 训练 | 测验 | 火车 | 测验 |
| #样品 | 8,855 | 2,933 | 80k | 40k |
| 标题/图像 | 10 | 10 | 5 | 5 |

表1。数据集统计。

评价。继Zhang等人[36]之后，我们使用初始分数[23]作为定量评估指标。由于初始分数不能反映生成的图像是否符合给定的文本描述，因此我们建议使用R-精度（一种用于对检索结果进行排序的常用评估指标）作为文本到图像合成任务的补充评估指标。如果查询有相关文档，我们检查系统排名靠前的检索结果，发现它们是相关的，然后根据定义，R精度为。更具体地说，我们进行了一个检索实验，即，我们使用生成的图像来查询它们相应的文本描述。首先，利用预训练DAMSM中学习的图像和文本编码器提取生成图像和给定文本描述的全局特征向量。然后，我们计算全局图像向量和全局文本向量之间的余弦相似性。最后，我们对每幅图像的候选文本描述按相似度递减进行排序，并找到最相关的描述来计算R精度。为了计算初始分数和R精度，每个模型从随机选择的不可见文本描述中生成30000个图像。每个查询图像的候选文本描述由一个基本事实（即=1）和99个随机选择的不匹配描述组成。*RRRr/rRR*

除了定量评估外，我们还定性地检查了我们的模型生成的样本。具体地说，我们通过注意模型将中间结果与注意学习进行可视化。如等式（2）中所定义，权重指示模型在生成图像的子区域时关注哪些词，以及*肥βj，i*

.我们通过

(15)

为了更好的可视化，我们固定单词并用图像的不同子区域计算其注意力权重，−. 我们重塑注意力权重*Nβ*ˆ0，我*,β*ˆ1、我*,...,β*ˆ*N*1、我*N*

到√N×√N像素，然后使用高斯滤波器进行上采样，使其与生成的图像具有相同的大小。受论文长度的限制，我们仅对每个注意模型的前五名最受关注的词（即，前五名值最高的词）进行可视化。

## 4.1.  成分分析

在本节中，我们首先定量评估AttnGAN及其变体。结果如表2所示

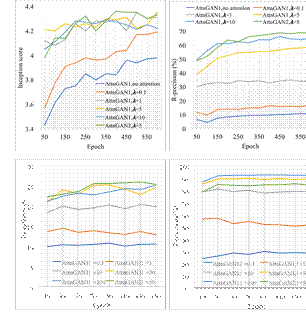


图3。在CUB（顶部）和COCO（底部）测试集上，我们的AttnGAN及其变体在不同时期的初始分数和R精度率。对于文本到图像合成任务，=1。*R*

|  |  |  |
| --- | --- | --- |
| 方法 | 起始分数 | R-精度（%） |
| 注意，请不要注意 | 3.98 ± .04 | 10.37± 5.88 |
| 附件1，=0.1*λ* | 4.19 ± .06 | 16.55± 4.83 |
| 附件1，=1*λ* | 4.35 ± .05 | 34.96± 4.02 |
| 附件1，=5*λ* | 4.35 ± .04 | 58.65± 5.41 |
| 附件1，=10*λ* | 4.29 ± .05 | 63.87± 4.85 |
| AttnGAN2，=5*λ* | 4.36 ± .03 | 67.82 ± 4.43 |
| 附件2，=50（COCO）*λ* | 25.89 ± .47 | 85.47 ± 3.69 |

表2。在CUB（前六行）和COCO（最后一行）测试集上，每个AttnGAN模型的最佳初始分数和相应的R准确率。更多结果如图3所示。

和图3。我们的“AttnGAN1”架构有一个注意模型和两个生成器，而“AttnGAN2”架构有两个注意模型和三个生成器（见图2）。此外，如图4、图5、图6和图7所示，我们定性地检查了AttnGAN生成的图像。

达姆的损失。为了测试建议的L，我们调整了*λ*的值（见等式（3））。如图3所示，在CUB和COCO数据集上，一个更大的*λ*值会导致更高的R精度。在CUB数据集上，当*λ*的值从0.1增加到5时，AttnGAN1的初始分数从4.19提高到4.35，相应的R-准确率从16.55%提高到58.65%（见表2）。在COCO数据集上，通过将*λ*值从0.1增加到50，AttnGAN1实现了较高的初始分数和R精度（见图3）。这种比较表明，适当增加LHELP的权重可以生成更高质量的图像，更好地适应给定的文本描述。原因是，建议的细粒度图像-文本匹配损失lp为训练生成器提供了额外的监督（即单词级匹配信息）。此外，在我们的实验中，我们没有观察到任何折叠的无意义模式在AttnGAN生成的图像可视化。这表明，在额外的监督下，细粒度图像文本匹配丢失也有助于稳定AttnGAN的训练过程。此外，在CUB数据集上训练基线模型“AttnGAN1，无注意”，并使用[19]中使用的文本编码器。在不使用注意的情况下，其初始得分和R精度分别下降到3.98%和10.37%，这进一步证明了所提出的L。*达姆**λ* *λ λ λ 达姆达姆达姆*

注意生成网络。如表2和图3所示，在生成网络中叠加两个注意力模型不仅可以生成更高分辨率（从128×128到256×256分辨率）的图像，而且还可以在CUB和COCO数据集上获得更高的初始分数。为了保证图像质量，我们通过增加的值找到每个数据集的最佳值，直到在保留的验证集上的总体初始分数开始下降。“AttnGAN1”模型用于搜索最佳图像，在此基础上构建“AttnGAN2”模型以生成更高分辨率的图像。由于GPU内存限制，我们没有尝试使用三种注意力模型的AttnGAN。因此，CUB和COCO的最终模型分别为“AttnGAN2，=5”和“AttnGAN2，=50”。COCO数据集的最终结果远大于CUB数据集的最终结果，这表明建议的Lis对于生成类似COCO数据集中的复杂场景尤其重要。*λ λ λλλλ 达姆*

※为了更好地理解AttnGAN所学到的知识，我们将注意力集中在其中间结果上。如图4所示，AttnGAN（*G*0）的第一阶段只是绘制对象的基本形状和颜色，并生成低分辨率图像。由于在此阶段仅使用全局句子向量，因此生成的图像缺少由确切的单词（例如鸟的喙和眼睛）描述的细节。基于字向量，以下阶段（*G*1和*G*2）学习纠正前一阶段结果中的缺陷，并添加更多细节以生成更高分辨率的图像。或图像的一些子区域/像素可以直接从前一阶段生成的图像推断。对于这些子区域，注意力平均分配给所有单词，并在注意力图中显示为黑色（见图4）。对于其他子区域，通常在文本描述中表达语义，例如对象的属性，注意力被分配到它们最相关的单词（图4中的明亮区域）。因此，这些区域是从这些区域的单词上下文特征和以前的图像特征中推断出来的。如图4所示，在CUB数据集上，单词the，this，bird通常由用于定位ob的模型参与-*G*0*G*1 *G*2*G*1 *G*2 *肥*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | 这只鸟有一个黄色的王冠和一个黑色的圆形眼圈 | 这只鸟有一顶绿色的王冠，黑色的初选和一个白色的腹部 | | 1:鸟4:黄色0:第12:轮11:是 | 1:鸟0:此2:有11:腹部10:白色 | | 1:鸟4:黄0:黑8:黑12:圆 | 6:黑色4:绿色10:白色0:这1:鸟 | | 一张自制的带花椰菜、胡萝卜和洋葱的漩涡意大利面的照片 | 展示香蕉和猕猴桃的水果摊 | | 0:a 7:带5:漩涡8:西兰花10:和 | 0:a 6:和1:水果7:猕猴桃5:香蕉 | | 8:西兰花6:意大利面0:A9:胡萝卜5:漩涡状 | 0:a 5:香蕉1:水果7:猕猴桃6:和 |   图4。ATTTNGAN在CUB（顶部）和COCO（底部）测试集上的中间结果。在每个块中，第一行给出了AttnGAN的64×64个图像，128×128个图像和256×256个图像；第二行和第三行分别显示了AttnGAN中最受关注的前5个单词。有关更多示例，请参阅补充资料。*G*0*G*1 *G*2 1   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | 数据集 | GAN-INT-CLS[20] | GAWWN[18] | 斯塔克甘[36] | StackGAN-v2[37] | PPGN[16] | 我们的阿廷根 | | 幼兽 | 2.88 ± .04 | 3.62 ± .07 | 3.70 ± .04 | 3.84 ± .06 | / | 4.36 ± .03 | | 可可 | 7.88 ± .07 | / | 8.45 ± .03 | / | 9.58 ± .21 | 25.89 ± .47 |   表3。《盗梦空间》由最先进的GAN模型[20,18,36,37,16]和我们的AttnGAN在CUB和COCO测试集上获得分数。 |

对象；描述对象属性的单词，如颜色和鸟的部分，也用于纠正缺陷和图纸细节。在COCO数据集上，我们有类似的观察结果。由于每个COCO图像中通常有多个对象，因此更明显的是，描述不同对象的单词由图像的不同子区域表示，例如，右下角的香蕉、猕猴桃。图4的方框这些观察结果表明，AttnGAN学会了理解图像文本描述中表达的详细语义。另一个观察结果是，我们的第二个注意模型能够注意到一些被第一个注意模型忽略的新词（见图）4). 它表明，为了提供更丰富的信息以产生更高的分辨率-

这种鸟有黑色的翅膀和白色的腹部



这种鸟有红色的翅膀和黄色的腹部



这只鸟有蓝色的翅膀和红色的腹部



图5。我们的AttnGAN模型的示例结果在CUB上进行了训练，同时更改了文本描述中一些最受关注的单词。

红色的双人床

一辆毛茸茸的黑色双层巴士一个站牌一只停牌猫漂浮在水面上漂浮在水面上漂浮在湖顶一个湖顶一个湖顶蓝天

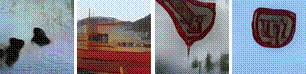


图6。使用在COCO上训练的AttnGAN模型，根据新场景描述生成256×256幅图像。（补充材料中给出了中间结果。）

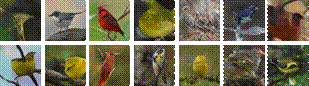


图7。我们的AttnGAN在幼崽测试集上的新图像。

在AttnGAN的后期阶段，相应的注意模型学习恢复先前阶段忽略的对象和属性。

※泛化能力。我们上面的实验结果通过从看不见的文本描述生成图像，定量和定性地显示了AttnGAN的泛化能力。在这里，我们通过改变文本描述中一些最受关注的单词，进一步测试输出对输入句子变化的敏感性。图5显示了一些示例。该模型根据输入句子的变化对生成的图像进行修改，表明该模型能够捕捉文本描述中的细微语义差异。此外，如图6所示，我们的AttnGAN可以生成图像，以反映真实世界中不太可能发生的新场景描述的语义，例如，停车标志漂浮在湖面上。另一方面，我们还观察到，AttnGAN有时会生成清晰而详细的图像，但不太可能是真实的。如图7所示，AttnGAN创造了多个头、眼睛或尾巴的鸟，这些鸟只存在于童话中。这表明我们目前的方法在捕获全局相干结构方面还不够完善，还有待改进。总之，图5、图6和图7所示的观察结果进一步证明了AttnGAN的泛化能力。

## 4.2.  与以往方法的比较

我们将我们的ATTANGAN和之前在CUB和COCO测试集上生成文本到图像的最先进的GAN模型进行了比较。如表36所示，在《开始》中，哪一个得分达到了前一张表格中最好的ATTN.3。从开始阶段到最佳阶段，我们的得分提高了58.9%，达到了令人印象深刻的水平。众所周知，COCO数据集比CUB数据集更具挑战性，因为它包含具有更复杂场景的图像。现有的方法很难在这个数据集上生成真实的高分辨率图像。图4和图6中的示例说明，我们的AttnGAN成功地为COCO数据集上的各种场景生成了256×256个图像，尽管COCO数据集生成的图像不如CUB数据集生成的图像逼真。实验结果表明，与以前的最先进的方法相比，AttnGAN在生成复杂场景方面更为有效，因为它具有新颖的注意机制，能够捕获文本到图像生成中的细粒度单词级和子区域级信息。

除了StackGAN-v2[37]，提出的注意机制也可以应用于广泛使用的DCGAN框架[17]。在CUB数据集上，我们构建了一个ATTNDCAN和一个普通的DCGAN。虽然仅以句子向量为条件的普通DCGAN（没有提出的注意机制）无法生成合理的256×256图像，但我们的AttnDCGAN能够生成真实的图像。ATTNDCGA初始得分为4.12±0.05，R精度为38.45±4.26%。由于严重的模式崩溃，普通CGAN仅达到2.47±0.01初始分数和3.69±1.82%的R精度。比较结果进一步证明了所提出的注意机制的有效性。

# 5.  结论

※本文提出了一种用于细粒度文本图像合成的注意生成对抗网络AttnGAN。我们构建了一个新的注意生成网络，让AttnGAN通过多阶段过程生成高质量的图像。我们提出了一个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，以训练AttnGAN生成器。我们的AttnGAN显著优于最先进的GAN模型，在CUB数据集和更具挑战性的COCO数据集上分别将最佳报告初始得分提高了14.14%和170.25%。大量的实验结果证明了AttnGAN中提出的注意机制的有效性，这对于复杂场景中的文本到图像生成尤其重要。

# 工具书类

[1] A.Agrawal、J.Lu、S.Antol、M.Mitchell、C.L.Zitnick、D.Parikh和D.Batra。视觉问答。IJCV，123（1）：2017年4月至31日。1.

[2] D.Bahdanau、K.Cho和Y.Bengio。神经机器翻译的联合学习对齐和翻译。arXiv:1409.04732014。2.

[3] E.L.Denton、S.Chintala、A.Szlam和R.Fergus。使用拉普拉斯金字塔对抗网络的深层生成图像模型。在NIPS，2015年。2.

[4] 方H.方S.古普塔F.N.伊恩多拉R.K.斯里瓦斯塔瓦L.邓P.多拉´高J.高X.何M.米切尔J.C.普拉特C.齐特尼克G.茨威格。从标题到视觉概念再到背景。在CVPR，2015年。1,4[5]Z.Gan，C.Gan，X.He，Y.Pu，K.Tran，J.Gao，L.Carin和L.Deng。视觉字幕的语义合成网络。在CVPR，2017年。1.

[6] I.J.古德费罗、J.普吉·阿巴迪、M.米尔扎、B.徐、D.沃德法利、S.奥扎尔、A.C.库尔维尔和Y.本吉奥。生成性对抗网。在NIPS，2014年。1, 2

[7] 格雷戈、丹尼尔卡、格雷夫斯、雷赞德和维尔斯特拉。

DRAW：用于图像生成的递归神经网络。在ICML，2015年。2.

[8] 何，邓，周。序列模式识别中的判别学习。IEEE信号处理杂志，25（5）：14–362008。4.

[9] 黄培生、何克星、高俊杰、邓丽君、阿奇罗和赫克。使用点击数据学习web搜索的深层结构语义模型。在CIKM，2013年。4.

[10] P.Isola，J.-Y.Zhu，T.Zhou和A.A.Efros。使用条件对抗网络进行图像到图像的翻译。在CVPR，2017年。2.

[11] Juang、W.Chou和C-H.Lee。用于语音识别的最小分类错误率方法。IEEE语音和音频处理学报，5（3）：257-2651997。4.

[12] D.P.Kingma和M.Welling。自动编码变分贝叶斯。在ICLR，2014年。2.

[13] C.莱迪格、L.泰斯、F.胡萨、J.卡巴莱罗、A.艾特肯、A.特贾尼、J.托茨、Z.王和W.史。使用生成对抗网络的照片逼真单图像超分辨率。在CVPR，2017年。2.

[14] 林泰耀、梅尔、贝隆吉、海斯、佩罗纳、拉马南、多尔和齐特尼克。Microsoft coco：上下文中的常见对象。在ECCV中，2014年。1, 5

[15] E.曼西莫夫、E.帕里索托、L.J.巴和R.萨拉胡季诺夫。从标题中产生注意力集中的图像。在ICLR，2016年。2.

[16] A.Nguyen、J.Yosinski、Y.Bengio、A.Dosovitskiy和J.Clune。即插即用生成网络：在潜在空间中有条件地迭代生成图像。在CVPR，2017年。2, 5, 7

[17] A.拉德福德、L.梅茨和S.钦塔拉。深度卷积生成对抗网络的无监督表征学习。在ICLR，2016年。2, 8

[18] S.里德、Z.赤田、S.莫汉、S.坦卡、B.希勒和H.李。学习画什么和在哪里画。在NIPS，2016年。1, 2, 5, 7

[19] S.里德、Z.赤田、B.希勒和H.李。学习细粒度视觉描述的深层表示。在CVPR，2016年。1, 6

[20] S.Reed、Z.Akata、X.Yan、L.Logeswaran、B.Schiele和H.Lee。生成对抗性文本图像合成。在ICML，2016年。1, 2, 5, 7

[21]S.E.里德、A.范登奥尔德、N.卡尔希布伦纳、S.G.科尔门纳雷乔、Z.王、Y.陈、D.贝洛夫和N.德弗雷塔斯。并行多尺度自回归密度估计。在ICML，2017年。2.

[22]O.Russakovsky，J.Deng，H.Su，J.Krause，S.Satheesh，S.Ma，

黄志强、卡帕蒂、科斯拉、伯恩斯坦、贝格和费菲。ImageNet大规模视觉识别挑战。IJCV，115（3）：211-252，2015年。4.

[23]T.Salimans、I.J.Goodfello、W.Zaremba、V.Cheung、A.Radford和X.Chen。改进了训练gans的技术。在NIPS，2016年。2, 5

[24]T.Salimans、H.Zhang、A.Radford和D.Metaxas。使用最佳运输改进gans。在ICLR，2018年。2.

[25]M.舒斯特和K.K.帕利瓦尔。双向递归神经网络。IEEE Trans。信号处理，45（11）：2673-26811997。4.

[26]C.Szegedy、V.Vanhoucke、S.Ioffe、J.Shlens和Z.Wojna。重新思考计算机视觉的初始架构。在CVPR，2016年。4.

[27]A.van den Oord、N.Kalchbrenner、O.Vinyals、L.Espeholt、A.Graves和K.Kavukcuoglu。使用像素CNN解码器生成条件图像。在NIPS，2016年。2.

[28]A.Vaswani、N.Shazeer、N.Parmar、J.Uszkoreit、L.Jones、A.N.Gomez、L.Kaiser和I.Polosukhin。注意力是你所需要的。arXiv:1706.03762，2017年。2.

[29]C.Wah、S.Branson、P.Welinder、P.Perona和S.Belongie。加州理工大学UCSD Birds-200-2011数据集。技术报告CNS-TR2011-001，加利福尼亚理工学院，2011年。5.

[30]K.Xu、J.Ba、R.Kiros、K.Cho、A.C.Courville、R.Salakhutdinov、R.S.Zemel和Y.Bengio。显示、参与和讲述：通过视觉注意力生成神经图像标题。在ICML，2015年。1, 2

[31]杨志强、何克强、高志强、邓丽君和斯莫拉。堆叠式注意力网络，用于图像问答。在CVPR，2016年。1, 2

[32]H.Zhang和K.Dana。用于实时传输的多样式生成网络。arXiv:1703.06953，2017年。1.

[33]H.Zhang、K.Dana、J.Shi、Z.Zhang、X.Wang、A.Tyagi和A.Agrawal。用于语义分割的上下文编码。在CVPR，2018年。1.

[34]张H和帕特尔。密集连接的金字塔形脱杂网络。在CVPR，2018年。2.

[35]H.Zhang、V.Sindagi和V.M.Patel。使用条件生成对抗网络的图像去训练。arXiv:1701.05957，2017年。2.

[36]张H，徐T，李H，张S，王X，黄X，和

D.梅塔克斯。Stackgan：具有堆叠生成对抗网络的文本到照片真实感图像合成。在ICCV中，2017年。1, 2, 3, 5, 7

[37]张H，徐T，李H，张S，王X，黄X和梅塔克斯。Stackgan++：具有堆叠生成对抗网络的真实图像合成。arXiv:1710.10916017。1, 2, 3, 5, 7, 8

[38]张志忠、谢耀荣、邢福福、麦高、杨立群。Mdnet：一个语义和视觉可解释的医学图像诊断网络。在CVPR，2017年。2.

[39]张志忠、谢耀荣、杨立群。具有层次嵌套对抗网络的摄影文本到图像合成。在CVPR，2018年。2.

[40]张志强、杨立群和郑耀勇。使用循环和形状一致性生成对抗网络翻译和分割多模态医学卷。在CVPR，2018年。2.

[41]朱勇，M.埃尔霍塞尼，B.刘，X.彭和A.埃尔加马尔。从嘈杂的文本中进行零射击学习的生成性对抗方法。在CVPR，2018年。1.

[[1]](" \l "_ftnref1" \o ")这项工作是在他在微软研究院实习时完成的

[[2]](" \l "_ftnref2" \o ")我们还对整个网络的DAMSM进行了微调，但是性能没有得到改善。