****

**生成性对抗性文本图像合成**

****

# **Scott Reed、Zeynep Akata、Xinchen Yan、Lajanugen Logeswaran REEDSCOT、Akata、XCYAN、LLAJANBernt Schiele、Honglak Lee Schiele、Honglak1211 21**

**1 密歇根大学，安娜堡，米河，美国（UMIC.EDU）**

**2 德国萨尔布吕肯马克斯·普朗克信息学研究所（MPI-INF.MPG.DE）**

# **摘要**

**从文本中自动合成真实图像将是有趣和有用的，但目前的人工智能系统离这一目标还很远。然而，近年来，人们发展了通用的、功能强大的递归神经网络结构来学习区分性文本特征表示。与此同时，卷积的世代对抗网络（GAN）已经开始生成特定类别的极具吸引力的图像，如人脸、专辑封面和房间内部。在这项工作中，我们开发了一种新的深层架构和GAN公式，以有效地桥接文本和图像建模方面的这些进展，将视觉概念从字符转换为像素。我们展示了我们的模型能够从详细的文本描述中生成似是而非的鸟和花的图像。**

# **1.  介绍**

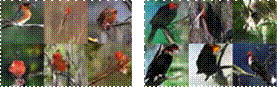
**在这项工作中，我们感兴趣的是将人类书写的单句描述形式的文本直接翻译成图像像素。例如，“这种小鸟有一个短而尖的橙色喙和白色的腹部”或“这种花的花瓣是粉红色的，花药是黄色的”。从视觉描述生成图像的问题引起了研究界的兴趣，但它远未得到解决。**

**传统上，有关对象的此类详细视觉信息已在属性表示中捕获，以区分编码到向量中的对象类别的特征（Farhadi等人，2009年；Kumar等人，2009年；Parikh&Grauman，2011年；Lampert等人，2014年），特别是实现零镜头视觉识别（Fu等人，2014；Akata等人，2015），以及最近的条件图像生成（Yan等人，2015）。**

**而辨别力强，泛化能力强**

|  |
| --- |
| ***委员会议事录第三十三国际机器会议***  ***学习*，美国纽约州纽约市，2016年。JMLR:W&CP卷**  **48.作者2016年版权所有。** |

**这只小鸟有粉红色的胸脯和皇冠，黑色几乎全是黑色，有红色的初级和次级。冠，白色面颊斑。**

****

**这种花有花瓣，这种白色和黄色的花是明亮的粉紫色，有薄的白色花瓣和一个白色的柱头圆形黄色的雄蕊**

****

***图1。*从文本描述生成图像的示例。左：标题来自零镜头（突出）类别，未显示文本。右图：标题来自训练集。**

**属性表示的属性很吸引人，属性的获取也很麻烦，因为它们可能需要特定领域的知识。相比之下，自然语言为描述任何视觉类别空间中的对象提供了通用且灵活的界面。理想情况下，我们可以拥有文本描述的通用性和属性的辨别力。**

**最近，文本的深度卷积和递归网络已经产生了从单词和字符自动学习的高度辨别性和可概括（在零射击学习意义上）文本表示（Reed等人，2016）。这些方法超过了之前在加州理工大学UCSD鸟类数据库上使用属性进行零镜头视觉识别的最新技术（Wah等人，2011年），并且能够进行基于零镜头字幕的检索。受这些作品的启发，我们的目标是学习直接从单词和字符到图像像素的映射。**

**要解决这个具有挑战性的问题，需要解决两个子问题：首先，学习捕获重要视觉细节的文本特征表示；第二，利用这些特征合成一幅引人注目的图像，人们可能会误以为是真的。幸运的是，在过去的几年里，深度学习在自然语言表示和图像合成这两个子问题上都取得了巨大的进步，我们在当前任务中正是以此为基础。**

**然而，仅靠深度学习无法解决的一个难题是，以文本描述为条件的图像分布是高度多模态的，从这个意义上讲，有许多合理的像素配置可以正确地说明描述。反向（图像到文本）也会遇到这个问题，但由于单词或字符序列可以根据链式规则顺序分解，因此学习变得切实可行；i、e.训练模型以预测以图像和所有先前标记为条件的下一个标记，这是一个更明确的预测问题。**

**因此，对于真实的应用程序来说，条件生成的图像是经过优化的，而对于真实的应用程序来说，条件生成的图像是经过甄别的。通过调节发电机和鉴别器的旁侧信息（Mirza&Osindero（2014）和Denton等人（2015）也进行了研究），我们可以自然地模拟这种现象，因为鉴别器网络充当“智能”自适应损失函数。**

**我们在这项工作中的主要贡献是开发一种简单有效的GAN架构和训练策略，该架构和训练策略能够从人类书面描述中合成引人注目的鸟和花图像。我们主要使用加州理工大学UCSD鸟类数据集和牛津102花卉数据集以及我们收集的每幅图像的五个文本描述作为评估设置。我们的模型是在训练类别子集上训练的，我们展示了它在训练集类别和测试集（即“零镜头”文本到图像合成）上的性能。除了鸟和花，我们还将我们的模型应用于MS COCO数据集中更一般的图像和文本描述（Lin等人，2014）。**

# **2.  相关工作**

**多模态学习中的关键挑战包括学习跨模态的共享表示，以及预测一种模态中的缺失数据（例如通过检索或合成），以另一种模态为条件。Ngiam等人（2011年）对音频和视频信号的堆叠多模态自动编码器进行了培训，并能够学习共享模态不变表示。Srivastava 和Salakhutdinov（2012）开发了一种深度玻尔兹曼机器，并联合建模了图像和文本标签。Sohn等人（2014）提出了一个多模态条件预测框架（幻觉一种模态给定另一种模态），并提供了理论依据。**

**最近，许多研究人员利用深度卷积解码网络的能力来生成逼真的图像。Dosovitskiy等人（2015年）训练了一个反褶积网络（几层卷积和上采样），以生成3D椅子渲染，该渲染基于一组指示形状、位置和照明的图形代码。Yang等人（2015年）在该方法中添加了编码器网络和操作。他们训练了一个循环卷积编码器，该编码器根据旋转的动作序列旋转3D椅子模型和人脸。里德等人（2015年）对模拟对的变换进行编码，并使用卷积解码器预测形状、视频游戏角色和3D汽车的视觉模拟。**

**对于发电机网络模块，发电对抗网络（Goodfello et al.，2014）也从卷积解码器网络中获益。Denton等人（2015年）使用拉普拉斯金字塔对抗生成器和鉴别器合成多分辨率图像。这项工作生成了引人注目的高分辨率图像，并且还可以对类标签进行控制生成。Radford et al.（2016）使用了标准卷积解码器，但开发了一种高效稳定的体系结构，结合批量标准化，以实现显著的图像合成结果。**

**我们的工作与上面描述的条件GAN的主要区别在于，我们的模型条件基于文本描述，而不是类标签。据我们所知，这是第一个端到端可区分的架构，从字符级别到像素级别。此外，我们还为GAN生成器引入了一个流形插值正则化器，该正则化器显著提高了生成样本的质量，包括CUB上保留的零炮类别。**

**以前关于图像和文本多模式学习的大部分工作都将检索作为目标任务，即通过文本查询获取相关图像，反之亦然。然而，在过去的一年中，在使用递归神经网络解码器生成基于图像的文本描述方面取得了突破（Vinyals等人，2015；Mao等人，2015；Karpath&Li，2015；Donahue等人，2015）。这些通常在深层卷积网络的顶层特征上调节长-短期记忆（Hochreiter&Schmidhuber，1997），以使用MS COCO（Lin et al.，2014）和其他字幕图像数据集生成字幕。Xu等人（2015年）采用了一种反复出现的视觉注意机制来改善结果。**

**在最近的工作中，除条件生成外，还考虑了其他任务。Ren等人（2015年）解答了有关图像视觉内容的问题。这种方法被扩展到包含明确的知识库（Wang等人，2015年）。Zhu等人（2015）将序列模型应用于文本（以书籍的形式）和电影，以执行联合对齐。**

**在当代作品中，Mansimov等人（2016年）通过文本标题生成图像，使用可变循环自动编码器，并注意在多个步骤中绘制图像，类似于DRAW（Gregor等人，2015年）。令人印象深刻的是，该模型能够合理地合成完全新颖（人类不太可能写出）的文本，例如“停止标志在蓝天上飞翔”，这表明它不只是记忆。虽然结果令人鼓舞，但问题极具挑战性，生成的图像还不真实，也就是说，可能被误认为是真实的。在许多情况下，我们的模型可以生成以文本为条件的视觉上合理的64×64图像，并且我们的整个模型是一个GAN，而不是仅使用GAN进行后处理。**

**基于之前许多工作的思想，我们开发了一种简单有效的基于文本的图像合成方法，使用字符级文本编码器和类条件GAN。我们提出了一种新颖的体系结构和学习策略，可以获得令人信服的视觉效果。我们关注细粒度图像数据集的情况，为此，我们使用最近收集的加州理工大学UCSD鸟类和牛津花卉的描述，每张图像有5个人工生成的标题（Reed等人，2016）。我们在类不相交集上进行训练和测试，以便测试性能能够很好地反映泛化能力，我们在具有多个对象和各种背景的MS COCO图像上也展示了这一点。**

# **3.  出身背景**

**在本节中，我们简要描述了我们的方法所基于的几个以前的工作。**

## **3.1.  生成性对抗网络**

**生成性对抗网络（GAN）由一个生成器和一个鉴别器组成，它们在两人极小极大博弈中竞争：反对反对烦烦烦烦烦烦打发人格侮辱日日日。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。具体来说，在V（D，G）上玩以下游戏：minmax（D，G）=E（[logD（x）]+（1）*GDDG五、十、*∼P*数据十、*)**

***G D***

**E[log（1− （D（z））]*十、*∼P*Z*（z）**

**古德费罗等人（2014年）证明，当pg=pdata时，该极小极大博弈具有全局最优解，并且在温和条件下（例如，具有足够的容量）收敛到。在实践中，在开始训练时，来自的样本非常差，并且被高度自信地拒绝。已经发现，在实践中，生成器最大化log（D（G（z）），而不是最小化log（1），效果更好− D（G（z）））。*pgGDpg数据DD***

## **3.2.  深对称结构联合嵌入**

**为了获得文本描述的视觉辨别矢量表示，我们遵循Reed等人（2016）的方法，使用深度卷积和循环文本编码器学习与图像的对应函数。通过优化以下结构化损失，对学习到的对应函数生成的文本分类器进行训练：*英尺***

**(2)**

**其中{（vn，tn，yn）：n=1，…，n}是训练数据集，∆ 是0-1损失，Vn是图像，Tn是相应的文本描述，是类标签。分类器和参数化如下：**

|  |  |
| --- | --- |
| ***fv*（v） =arg最大值E[φ（v）Tñ（T））]*T*∼T（y）**  ***Y*∈Y** | **(3)** |
| ***英尺*（t） =arg最大值E[φ（v）tñ（t））]*五、*∼V（y）** | **(4)** |

***Y*∈Y**

**其中*ϕ*是图像编码器（例如深卷积神经网络），*φ*是文本编码器（例如字符级CNN或LSTM），T（y）是类的文本描述集，同样，V（y）是图像。这里的直觉是，与任何其他类相比，文本编码应该与对应类的图像具有更高的兼容性分数，反之亦然。**

**为了训练模型，最小化了与方程2相关的替代目标（详情见Akata等人（2015））。产生的梯度通过反向传播来学习区分性文本编码器。里德等人（2016年）发现，不同的文本编码器对CUB和Flowers的效果更好，但为了充分的通用性和对打字错误和大词汇量的鲁棒性，在这项工作中，我们总是使用混合字符级卷积循环网络。*ϕ***

# **4.  方法**

**我们的方法是训练一个基于混合字符级卷积递归神经网络编码的文本特征的深层卷积生成对抗网络（DC-GAN）。*G*生成器网络和*D*鉴别器网络都根据文本特征执行前馈推理。**

## **4.1.  网络体系结构**

**我们使用以下符号。发电机网络表示为：R×R→ R、 鉴别器为：R×R→ {0,1}，其中是文本描述嵌入的维度，是图像的维度，是输入到的噪声的维度。我们在图2中展示了我们的网络架构。*GZTDDDTTDZG***

**在生成器中，我们首先从噪声中采样∈ R∼ 使用文本编码器（0,N）对文本进行编码。描述嵌入（t）首先使用一个完全连接的层压缩到一个小维度（在实践中我们使用128），然后使用泄漏ReLU和*GZZTϕϕ***

|  |
| --- |
| ***图2。*我们的文本是条件卷积GAN结构。生成器和鉴别器都使用文本编码（t）。它被投影到一个较低的维度和深度，并与图像特征图连接起来，用于卷积处理的进一步阶段。*ϕ*** |

**然后连接到噪声向量。在此之后，推理过程与正常反褶积网络中的推理过程相同：*Z***

**我们通过发电机对其进行前馈；通过合成图像x生成合成图像 G（z，ν（t））。图像生成对应于生成器G中以查询文本和噪声样本为条件的前馈推理。**

**在鉴别器D中，我们执行几层步幅-2空间批量标准化卷积（Ioffe&Szegedy，2015），然后是有泄漏ReLU。我们再次降低描述嵌入φ（t）在（单独的）完全连接层中的维数，然后进行校正。当鉴别器的空间维数为4×4时，我们在空间上复制嵌入的描述并执行深度拼接。然后，我们执行1×1卷积，然后进行校正，并执行4×4卷积，以计算来自的最终分数。批量标准化在所有卷积层上执行。**

## **4.2.  匹配感知鉴别器（GAN-CLS）**

**训练条件GAN最直接的方法是将（文本、图像）对视为联合观察，并训练鉴别器判断配对是真是假。从鉴别器没有明确的概念来判断真实的训练图像是否与文本嵌入上下文匹配的意义上来说，这种类型的条件作用是幼稚的。**

**然而，正如（Gauthier，2015）所讨论的，学习的动力可能不同于非条件性情况。在训练开始时，鉴别器忽略条件信息，并容易拒绝来自演变器G的样本，因为它们看起来不可信。一旦G学会了生成合理的图像，它还必须学会将它们与条件信息对齐，并且D同样必须学会评估来自G的样本是否满足此条件约束。**

**在缺乏经验的Gan中，鉴别器观察两种输入：带有匹配文本的真实图像和带有任意文本的合成图像。因此，它必须隐式地分离两个错误源：不切实际的图像（对于任何文本），以及与条件信息不匹配的错误类别的真实图像。基于这可能使学习动力学复杂化的直觉，我们修改了GAN训练算法来分离这些错误源。除了训练期间鉴别器的真/假输入之外，我们还添加了第三种输入类型，包括带有不匹配文本的真实图像，鉴别器必须学习将其作为假输入进行评分。通过学习优化图像/文本匹配以及图像真实感，鉴别器可以向生成器提供额外的信号。**

**算法1带*α*步长的GAN-CLS训练算法，为简单起见，使用小批量SGD。**

****

|  |  |
| --- | --- |
| **1：输入：小批量图像、匹配文本、不匹配ˆ、训练批步数*十、TTs***  **2:for=1待办事项*Ns*** | |
| **三：** | ***H*← ν（t）{编码匹配的文本描述}** |
| **4:** | ***H*ˆ ← ν（tˆ）{编码错误匹配文本描述}** |
| **5:** | ***Z*∼ N（0,1）Z{随机噪声抽样}** |
| **6:** | ***十、*ˆ ← G（z，h）{通过生成器向前}** |
| **7:** | ***sr*← D（x，h）{实图像，右文本}** |
| **8:** | ***西南*← D（x，h）{真实图像，错误文本}ˆ** |
| **9:** | ***科幻小说*← D（x，hˆ）{假图像，右文本}** |
| **10:** | **LD← 对数（sr）+（对数（1− sw）+日志（1− sf））/2** |
| **11:** | ***D*← D− α∂LD/∂D{更新鉴别器}** |
| **12:** | **LG← 日志（sf）** |
| **13:** | ***G*← G− α∂LG/∂G{更新生成器}** |

## **14：结束**

****

**算法1总结了训练过程。在编码文本、图像和噪声（第3-5行）后，我们生成伪图像（X，第6行）。表示将真实图像与其对应句子关联的分数（第7行），测量将真实图像与任意句子关联的分数（第8行），以及将虚假图像与其对应文本关联的分数（第9行）。注意，我们使用∂LD/****∂D表示目标相对于其参数的梯度，同样适用于G。第11行和第13行表示采取梯度步骤来更新网络参数。**

## **4.3.  流形插值学习（GAN-INT）**

**深度网络已被证明可以学习嵌入对之间的插值趋向于接近数据流形的表示（Bengio等人，2013年；Reed等人，2014年）。基于这个特性，我们可以通过简单地在训练集标题的嵌入之间进行插值来生成大量额外的文本嵌入。关键的是，这些插入的文本嵌入不需要对应于任何实际的人类书面文本，因此没有额外的标签成本。这可以看作是在发电机目标中增加了一个额外的术语，以最小化：**

**E[log（1− D（G（z，βt1+）（1− β） t2）））]（5）*T*1*T*2∼P*数据***

**其中，从z的噪波分布绘制，并在*β*文本嵌入*T*1和*T*2之间插值。在实践中，我们发现固定*****β*=0.5效果良好。**

**因为插值嵌入是合成的，所以鉴别器*D*没有“真实”对应的图像和文本对来训练。但是，*D*学习预测图像和文本对是否匹配。因此，*D*如果在这方面做得很好，那么通过*D*满足插值*G*文本嵌入，可以学会填补数据流形上训练点之间的空白。请注意，*T*1和*T*2可能来自不同的图像，甚至不同的类别。**

## **4.4.  反转生成器以进行样式转换**

**如果文本编码*ϕ*（t）捕获图像内容（例如，花朵形状和颜色），则为了生成逼真的图像，噪声样本*Z*应捕获样式因素，例如背景颜色和姿势。使用经过训练的GAN，可以将查询图像的样式转换为特定文本描述的内容。为了实现这一点，*G*可以训练卷积网络从样本反向回归*Z*← G（z，ν（t））返回。我们使用一个简单的平方损失来训练样式编码器：**

**(6)**

***s*样式编码器网络在哪里。使用经过培训的生成器和样式编码器，从*s*查询图像到*T*文本的样式转换过程如下：**

***s*← S（x），x← G（s，ν（t））**

**式中，x是结果图像，s是预测样式。**

# **5.  实验**

**在本节中，我们首先介绍了幼鸟图像数据集和牛津-102花卉图像数据集的结果。这只幼鸟拥有11788张属于200个不同类别之一的鸟类图片。牛津102包含了来自102个不同类别的8189幅花卉图片。**

**正如Akata et al.（2015）和Reed et al.（2016）中所述，我们将其划分为类不相交的训练和测试集。CUB有150个train+val类和50个测试类，而Oxford-102有82个train+val课程和20个测试类。对于这两个数据集，我们每个图像使用5个标题。在训练的小批量选择过程中，我们随机选择图像视图（例如裁剪、翻转）和一个标题。**

**对于文本特征，我们首先对深度卷积循环文本编码器进行预训练，将其与1024维GoogLeNet图像嵌入进行结构化联合嵌入（Szegedy et al.，2015），如第3.2小节所述。对于Oxford-102和CUB，我们使用了字符级ConvNet与递归神经网络（char CNN RNN）的混合，如（Reed et al.，2016）所述。但是，请注意，预培训文本编码器不是我们方法的要求，我们在补充中包含了一些端到端结果。对文本编码器进行预训练的原因是为了提高训练其他组件的速度，以便更快地进行实验。我们还提供了验证集的MS COCO图像的一些定性结果，以表明我们方法的普遍性。**

**我们对所有数据集使用相同的GAN架构。训练图像大小设置为64×64×3。文本编码器产生1024维嵌入，在深度连接到卷积特征映射之前，在生成器和鉴别器中投影到128维。**

**如算法1所示，我们采取交替步骤更新生成器和鉴别器网络。我们使用相同的基础学习率0.0002，并使用动量为0.5的ADAM解算器（Ba&Kingma，2015）。发电机噪声从100维单位正态分布采样。我们使用了64个小批量，训练了600个时代。我们的实现构建在dcgan.torch之上。****[[2]](" \l "_ftn2" \o ")**

## **5.1.  定性结果**

**我们比较了GAN基线、我们的GAN-CLS与imagetext匹配鉴别器（第4.2小节）、GAN-INT与文本流形插值学习（第4.3小节）以及结合两者的GAN-INT-CLS。**

**CUB的结果如图3所示。GAN和GANCLS获得了一些正确的颜色信息，但图像看起来并不真实。然而，GAN-INT和GAN-INT-CLS显示的图像看似合理，通常与标题的全部或至少部分匹配。我们还对补充的CUB数据集中每个GAN变体的稳健性进行了额外的分析**

|  |
| --- |
| **赣**  **GAN-CLS**  **GAN-INT**  **GAN-INT**  **- CLS**  ***图3。*零镜头（即，以看不见的测试集类别的文本为条件）使用GAN、GAN-CLS、GAN-INT和GAN-INT-CLS生成鸟类图像。我们发现插值正则化器需要可靠地获得视觉上合理的结果。**  **燃气轮机**  **赣**  **GAN-CLS**  **GAN-INT**  **GAN-INT**  **- CLS**  ***图4。*零拍使用GAN、GAN-CLS、GAN-INT和GAN-INT-CLS生成花朵图像。所有的变种都产生了似是而非的图像。虽然在训练期间未看到测试类别的某些形状（例如第3列和第4列），但颜色信息仍保留下来。** |

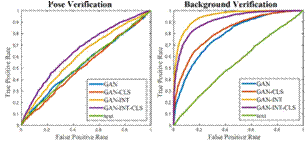
**牛津102花卉数据集的结果如图4所示。在这种情况下，所有四种方法都可以生成与描述匹配的看似合理的花朵图像。基本的GAN往往在花的形态上具有最多的多样性（即，如果标题未指明这一部分，人们可以看到非常不同的花瓣类型），而其他方法则倾向于生成更符合类别的图像。我们推测，生成花更容易，可能是因为鸟类在物种间具有更强的结构规律性，这使得*D*识别假鸟比识别假花更容易。**

**关于CUB和Oxford-102的GAN-INT和GAN-INTCLS以及GAN-E2E（我们的端到端GAN-INT-CLS，无需预先培训文本编码器*ϕ*（t））的许多其他结果，可在补充资料中找到。**

## **5.2.  理清文体与内容**

**在本节中，我们将研究我们的模型在多大程度上可以分离样式和内容。所谓内容，我们指的是鸟本身的视觉属性，如身体各部位的形状、大小和颜色。所谓风格，我们指的是图像中所有其他变化因素，如背景颜色和鸟的姿势方向。**

**文本嵌入主要包括内容信息，通常不涉及样式，例如，标题不提及背景或鸟姿。因此，为了生成逼真的图像，GAN必须学会使用*Z*噪声样本来解释样式变化。**

****

***图5。*ROC曲线使用相同或不同样式图像对上预测样式向量之间的余弦距离。左：图像对反映相同或不同的姿势。右：图像对反映相同或不同的平均背景色。**

**为了量化CUB上的分离程度，我们设置了两个以噪声为输入的预测任务：引发确认Z和背景色验证。对于每个任务，我们首先构造相似和不同的图像对，然后通过将图像馈送到样式编码器（经过训练以反转生成器的输入和输出）来计算预测的样式向量。如果GAN使用了与图像内容分离的*Z*样式，则相同样式（例如相似姿势）的图像之间的相似性应高于不同样式（例如不同姿势）的图像之间的相似性。**

**为了恢复*Z*，我们按照第4.4小节所述对每个发电机网络进行了反转。为了构造用于验证的一对，我们使用K-means将图像分组为100个簇，其中来自同一簇的图像共享相同的样式。对于背景色，我们根据背景的平均颜色（RGB通道）对图像进行聚类；对于鸟类姿势，我们通过6个关键点坐标（喙、腹、胸、冠、额和尾）对图像进行聚类。**

**为了评估，我们通过为GAN、GAN-CLS、GAN-INT和GAN-INT-CLS提供成对的图像样式编码器来计算实际预测的样式变量。我们使用余弦相似性验证分数，并报告AU-ROC（平均超过5倍）。作为基线，我们还计算文本编码器中文本特征之间的余弦相似性。**

**我们在图5中展示了结果。正如预期的那样，标题本身并不能为风格预测提供信息。此外，我们发现，采用CLGANINT（CLGANINT-GANINT）的插值模型与该任务的最佳结果一致。**

## **5.3.  姿势和背景样式转换**

**我们证明了带有经过训练的样式编码器（第4.4小节）的GAN-INT-CLS可以执行从看不见的查询图像到文本描述的样式转换。图6显示了使用推断样式生成的图像可以准确捕获姿势信息。在某些情况下，样式转换会保留详细的背景信息，例如鸟栖息的树枝。**

**文本描述图像**

**（内容）（风格）**

**这种鸟有一个黄色的胸部，灰色的特征和一个小喙。**

**这是一只白色的大鸟，有黑色的翅膀和红色的头。**

**一种头部和翅膀为黑色的小鸟，翅膀为灰色。**

**这种鸟有白色的胸脯，头和翅膀呈棕色和白色，喙薄而尖。**

**一种腹部、头部和羽毛上有白色基部和黑色条纹的小鸟。**

**一种体型较小的鸟，有奶油色的腹部和短而尖的喙。**

**这只鸟完全是红色的。**

**这只鸟完全是白色的。**

**这是一只黄色的鸟。翅膀是亮蓝色的。**

***图6。*将样式从顶行（真实）图像转换为查询文本中的内容，充当确定性解码器。下面三行是我们制作的字幕。*G***

**分清GAN-INT-CLS对风格进行的解读很有趣，因为它提供了一种简单的概括方式。这种方式下我们可以将以前看到的内容（例如文本）和以前看到的样式结合起来，但要采用新颖的配对方式，以便生成与训练期间看到的任何图像都非常不同的可信图像。另一种概括的方法是使用以前在图6底部一行生成的类似长尾鹦鹉的鸟中看到的属性（例如蓝色翅膀、黄色腹部）。这种泛化方法利用了捕获多个视觉方面的文本表示。**

## **5.4.  句子插入**

**图8展示了通过插值（左）学习的文本流形。虽然中间点没有基本事实文本，但生成的图像似乎是可信的。因为我们保持噪声分布不变，所以每行中唯一的变化因素是我们使用的文本嵌入。请注意，插值可以准确反映颜色信息，例如鸟类从蓝色变为红色，而姿势和背景不变。**

**除了在两个文本编码之间进行插值外，我们还在图8（右）中显示了噪声插值的结果。这里，我们对两个随机噪声向量进行采样。通过保持文本编码固定，我们在这两个噪声向量之间进行插值，并通过保持内容固定，生成两种样式之间平滑过渡的鸟类图像。**

## **5.5.  超越花鸟**

|  |
| --- |
| **我们的我们的**  **一群穿着湿漉漉的水罐的人正穿着滑雪服骑着马准备扔**  **站在冲浪板上，把球扔到雪地上。波动击球手**  **两盘**  **一张桌子包含一张照片的食物**  **很多盘子的豆子，非常干净的鳄梨酱和客厅。喝米饭。食品和**  **两只长颈鹿**  **一种绿色的植物，站在旁边，一株一株地生长着羊**  **从森林里出来。开阔的草地**  **地领域**  **大蓝色**  **章鱼风筝一个厕所在一个小苍蝇上面只有一个房间，里面有一匹人马在窗口和草地上嬉戏。未完工的墙。海滩。**  ***图7。*在MS-COCO验证集上使用我们的GAN-CLS生成一般概念的图像。与CUB和Oxford-102不同，网络必须（尝试）处理多个对象和不同的背景。** |

**“黑嘴蓝鸟”→**

**“黑嘴红鸟”**

**“黑色翅膀的蓝色小鸟”→**

**“黑色翅膀的黄色小鸟”**

***图8。*左图：通过在两句话之间插值生成的鸟图像（在一行内，噪声是固定的）。右：在两个随机采样的噪波向量之间插值。**

**我们在MS-COCO上训练了一个GAN-CLS，以显示我们的方法在包含多个对象和可变背景的一般图像集上的泛化能力。我们使用与CUB相同的文本编码器架构、相同的GAN架构和相同的超参数（学习率、小批量大小和历代数）和牛津102。培训文本编码器的唯一区别是COCO并没有每个类都有一个对象类别。然而，我们仍然可以学习实例级（而不是类别级）图像和文本匹配功能，如（Kiros等人，2014年）。**

**图7显示了样本和地面实况说明及其相应的图像。所有结果的一个共同特点是样品的清晰度，类似于其他基于GAN的图像合成模型。我们还通过简单地绘制多个噪声向量和使用相同的固定文本编码来观察样本中的多样性。**

**从远处看，结果令人鼓舞，但经过仔细观察，很明显，生成的场景通常不连贯；例如，棒球场景中的人形斑点缺少清晰的关节部位。在未来的工作中，为了更好地处理复杂的多对象场景，将层次结构纳入图像合成模型可能是有趣的。**

**与AlignDRAW（Mansimov等人，2016年）的定性比较可在附录中找到。GANCLS生成更清晰、分辨率更高的样本，这些样本大致对应于查询，但AlignDRAW样本更明显地反映了该工作中所选查询中单个单词的变化。将时间结构纳入GAN-CLS生成器网络可能会提高其捕获这些文本变化的能力。**

# **6.  结论**

**在这项工作中，我们开发了一个简单有效的模型，用于根据详细的视觉描述生成图像。我们证明了该模型可以合成给定文本标题的许多合理的视觉解释。我们的流形插值正则化器大大改进了CUB上的文本到图像合成。我们展示了风格和内容的分离，以及从查询图像到文本描述的鸟姿和背景转换。最后，我们用MS-COCO数据集上的结果证明了我们生成具有多个对象和可变背景的图像的方法的通用性。在未来的工作中，我们的目标是进一步将模型放大到更高分辨率的图像，并添加更多类型的文本。**

# **致谢**

**这项工作部分得到了NSF职业IIS-1453651、ONR N00014-13-1-0762和NSF CMMI1266184的支持。**

# **工具书类**

**Akata，Z.，Reed，S.，Walter，D.，Lee，H.，和Schiele，B.细粒度图像分类的输出嵌入评估。在CVPR，2015年。**

**Ba，J.和Kingma，D.Adam：随机优化的一种方法。在ICLR，2015年。**

**班吉奥，Y.，梅斯尼尔，G.，多芬，Y.，和里法伊，S.通过深度表现更好地混合。在ICML，2013年。**

**Denton，E.L.，Chintala，S.，Fergus，R.，等。使用对抗网络的拉普拉斯金字塔的深层生成图像模型。2015年，NIPS。**

**Donahue，J.，Hendricks，L.A.，Guadarrama，S.，Rohrbach，M.，Venugopalan，S.，Saenko，K.，和Darrell，T.用于视觉识别和描述的长期循环卷积网络。在CVPR，2015年。**

**Dosovitskiy，A.，Tobias Springenberg，J.，和Brox，T.学习用卷积神经网络生成椅子。在CVPR，2015年。**

**Farhadi，A.，Endres，I.，Hoiem，D.，和Forsyth，D.通过属性描述对象。在CVPR，2009年。**

**Fu，Y.，Hospedales，T.M.，Xiang，T.，Fu，Z.，和Gong，S.用于零镜头识别和注释的平移多视图嵌入。在ECCV中，2014年。**

**Gauthier，J.卷积面生成的条件生成对抗网。技术报告，2015年。**

**古德费罗，I.，普吉·阿巴迪，J.，米尔扎，M.，徐，B.，沃德·法利，D.，奥扎尔，S.，库尔维尔，A.，和本吉奥，Y.生成性对抗网络。在NIPS，2014年。**

**Gregor，K.，Danihelka，I.，Graves，A.，Rezende，D.，和Wierstra，D.Draw：用于图像生成的递归神经网络。在ICML，2015年。**

**Hochreiter，S.和Schmidhuber，J.长期短期记忆。神经计算，9（8）：1735-17801997。**

**Ioffe，S.和Szegedy，C.批量标准化：通过减少内部协变量转移来加速深层网络训练。在ICML，2015年。**

**Karpathy，A.和Li，F.用于生成图像描述的深度视觉语义对齐。在CVPR，2015年。**

**Kiros，R.，Salakhutdinov，R.，和Zemel，R.S.将视觉语义嵌入与多模态神经语言模型统一起来。在ACL，2014年。**

**Kumar，N.，Berg，A.C.，Belhummur，P.N.，和Nayar，S.K.用于人脸验证的属性和明喻分类器。在ICCV中，2009年。**

**Lampert，C.H.，Nickisch，H.，和Harmeling，S.基于属性的零镜头视觉对象分类。TPAMI，36（3）：453-4652014。**

**林，T-Y.，梅尔，M.，贝隆吉，S.，海斯，J.，佩罗纳，P.，拉马南，D.，多尔，P.，和齐特尼克，C.L.微软的coco：上下文中的常见对象。在ECCV中。2014**

**Mansimov，E.，Parisotto，E.，Ba，J.L.，和Salakhutdinov，R.专注地从字幕生成图像。ICLR，2016年。**

**Mao，J.，Xu，W.，Yang，Y.，Wang，J.，和Yuille，A.使用多模态递归神经网络（m-rnn）的深度字幕。ICLR，2015年。**

**Mirza，M.和Osindero，S.条件生成对抗网。arXiv预印本arXiv:1411.17842014。**

**Ngiam，J.，Khosla，A.，Kim，M.，Nam，J.，Lee，H.，和Ng，A.Y.多模式深度学习。在ICML，2011年。**

**Parikh，D.和Grauman，K.相对属性。在ICCV中，2011年。**

**Radford，A.，Metz，L.，和Chintala，S.《具有深层卷积生成对抗网络的无监督表征学习》。2016**

**Reed，S.，Sohn，K.，Zhang，Y.，和Lee，H.学习用多种交互作用来解开变异因素。在ICML，2014年。**

**里德，S.，张，Y.，张，Y.，和李，H.深度视觉类比制作。2015年，NIPS。**

**Reed，S.，Akata，Z.，Lee，H.，和Schiele，B.学习细粒度视觉描述的深层表征。在CVPR，2016年。**

**Ren，M.，Kiros，R.，和Zemel，R.探索图像问答的模型和数据。2015年，NIPS。**

**Sohn，K.，Shang，W.，和Lee，H.利用信息的变化改进了多模态深度学习。在NIPS，2014年。**

**Srivastava，N.和Salakhutdinov，R.R.利用深度玻尔兹曼机器进行多模态学习。在NIPS，2012年。**

**塞格迪，C.，刘，W.，贾，Y.，塞尔马内，P.，里德，S.，安圭洛夫，D.，埃尔汉，D.，万霍克，V.，和拉比诺维奇，A.通过卷积深入研究。在CVPR，2015年。**

**Vinyals，O.，Toshev，A.，Bengio，S.，和Erhan，D.《展示与讲述：神经图像字幕发生器》。在CVPR，2015年。**

**华，C.，布兰森，S.，韦林德，P.，佩罗纳，P.，和贝隆吉，S.加州理工大学ucsd鸟类-200-2011数据集。2011**

**王，P.，吴，Q.，沈，C.，亨格尔，A.v.d.，和迪克，A.基于显性知识的视觉问答推理。arXiv预印本arXiv:1511.025702015。**

**Xu，K.，Ba，J.，Kiros，R.，Courville，A.，Salakhutdinov，R.，Zemel，R.，和Bengio，Y.Show，Attainment and tell:具有视觉注意力的神经图像字幕生成。在ICML，2015年。**

**Yan，X.，Yang，J.，Sohn，K.，和Lee，H.Attribute2image：从视觉属性生成条件图像。arXiv预印本arXiv:1512.005702015。**

**Yang，J.，Reed，S.，Yang，M.-H.，和Lee，H.Weaklysupervised与3d视图合成的循环变换脱钩。2015年，NIPS。**

**Zhu，Y.，Kiros，R.，Zemel，R.，Salakhutdinov，R.，Urtasun，R.，Torralba，A.，和Fidler，S.调整书籍和电影：通过观看电影和阅读书籍实现故事般的视觉解释。在ICCV中，2015年。**

**[[1]](" \l "_ftnref1" \o ") 在我们的实验中，我们使用了细粒度的类别（例如，鸟类与其他鸟类非常相似，花朵与其他花朵非常相似，等等），并且跨类别进行插值不会造成问题。**

**[[2]](" \l "_ftnref2" \o ")https://github.com/soumith/dcgan.torch**