(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113781598 B
(45)授权公告日 2023.06.30

(21)申请号 202111238938.3

(22)申请日 2021.10.25

(65)同一申请的已公布的文献号
申请公布号 CN 113781598 A

(43)申请公布日 2021.12.10

(73)专利权人 北京邮电大学
地址 100876 北京市海淀区西土城路10号

(72)发明人 杨博 冯方向 王小捷 袁彩霞 李睿凡

(74)专利代理机构 北京德琦知识产权代理有限公司 11018
专利代理师 孙清然 王琦

(51)Int.Cl.
G06T 11/00(2006.01)
G06F 40/284(2020.01)
G06F 40/211(2020.01)
G06V 10/44(2022.01)
G06V 10/80(2022.01)

(56)对比文件
CN 112818159 A,2021.05.18
CN 113361250 A,2021.09.07
WO 2019215109 A1,2019.11.14

审查员 张驰

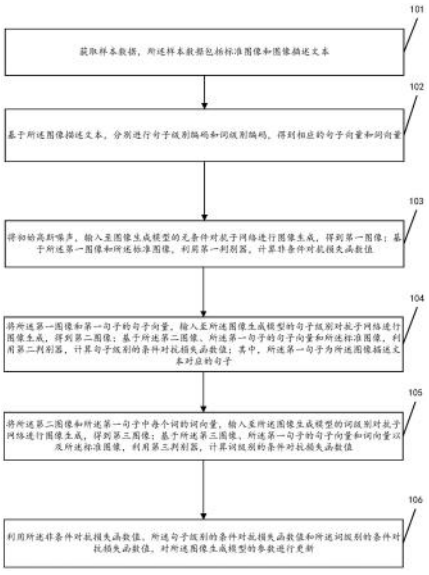
权利要求书3页 说明书10页 附图1页

(54)发明名称

图像生成模型的训练方法和设备以及图像生成方法

(57)摘要

本申请公开了一种图像生成模型的训练方法和设备以及图像生成方法，其中训练方法包括：基于样本数据中的图像描述文本生成句子向量和词向量；将初始高斯噪声输入模型的无条件对抗子网络生成第一图像；基于第一图像计算非条件对抗损失函数值；将第一图像和第一句子的句子向量输入模型的句子级别对抗子网络生成第二图像；基于第二图像和句子向量计算句子级别的条件对抗损失函数值；第一句子为图像描述文本对应的句子；将第二图像和第一句子的词向量，输入模型的词级别对抗子网络生成第三图像；基于第三图像、句子向量和词向量，计算词级别的条件对抗损失函数值；利用上述损失函数值对图像生成模型的参数更新。本申请可以保证生成图像与文本的一致性。



CN 113781598 B

1.一种图像生成模型的训练方法,其特征在于,包括:

获取样本数据,所述样本数据包括标准图像和图像描述文本;

基于所述图像描述文本,分别进行句子级别编码和词级别编码,得到相应的句子向量和词向量;

将初始高斯噪声,输入至图像生成模型的无条件对抗子网络进行图像生成,得到第一图像;基于所述第一图像和所述标准图像,利用第一判别器,计算非条件对抗损失函数值;

将所述第一图像和第一句子的句子向量,输入至所述图像生成模型的句子级别对抗子网络进行图像生成,得到第二图像;基于所述第二图像、所述第一句子的句子向量和所述标准图像,利用第二判别器,计算句子级别的条件对抗损失函数值;其中,所述第一句子为所述图像描述文本对应的句子;

将所述第二图像和所述第一句子中每个词的词向量,输入至所述图像生成模型的词级别对抗子网络进行图像生成,得到第三图像;基于所述第三图像、所述第一句子的句子向量和词向量以及所述标准图像,利用第三判别器,计算词级别的条件对抗损失函数值;

利用所述非条件对抗损失函数值、所述句子级别的条件对抗损失函数值和所述词级别的条件对抗损失函数值,对所述图像生成模型的参数进行更新。

2.根据权利要求1所述的方法,其特征在于,所述将初始高斯噪声,输入至图像生成模型的无条件对抗子网络进行图像生成包括:

对所述初始高斯噪声进行重构;

利用第一图像生成器,对所述重构的结果进行处理,得到所述第一图像。

3.根据权利要求1所述的方法,其特征在于,所述将所述第一图像和第一句子的句子向量,输入至所述图像生成模型的句子级别对抗子网络进行图像生成包括:

将所述句子向量融入所述第一图像的图像特征中;

利用第二图像生成器,对所述融入得到的图像特征进行处理,得到所述第二图像。

4.根据权利要求1所述的方法,其特征在于,所述将所述第二图像和所述第一句子中每个词的词向量,输入至所述图像生成模型的词级别对抗子网络进行图像生成包括:

将所述词向量融入所述第二图像的图像特征中;

利用第三图像生成器,对所述融入得到的图像特征进行处理,得到所述第三图像。

5.根据权利要求1所述的方法,其特征在于,所述非条件对抗损失函数值包括所述无条件对抗子网络中第一判别器的损失函数值和所述无条件对抗子网络中第一图像生成器的损失函数值;

所述基于所述第一图像和所述标准图像,利用第一判别器,计算非条件对抗损失函数值包括:

按照$\mathcal{L}_{D_1}=-\frac{1}{2}\mathbb{E}_{x_t\sim data}\left[log\left(D(x_t)\right)\right]-\frac{1}{2}\mathbb{E}_{x_1\sim G_1}\left[log\left(1-D(x_1)\right)\right]$,计算所述第一判别器的损失函数值$\mathcal{L}_{D_1}$;其中,D( )表示判别器输出的概率值;$x_t$表示样本数据中的标准图像;$x_1$表示第一图像;$\mathbb{E}_{x_t\sim data}\left[log\left(D(x_t)\right)\right]$表示利用标准图像$x_t$计算第一判别器的第一非条件对抗损失函数;$\mathbb{E}_{x_1\sim G_1}\left[log\left(1-D(x_1)\right)\right]$表示利用从第一图像生成器$G_1$中得到的第一图像

x_1计算第一判别器的第二非条件对抗损失函数；

按照$\mathcal{L}_{G_1}=-\frac{1}{2}\mathbb{E}_{x_1\sim G_1}[log(D(x_1))]$，计算所述第一图像生成器的损失函数值$\mathcal{L}_{G_1}$。

6.根据权利要求1所述的方法，其特征在于，

所述句子级别的条件对抗损失函数值包括所述句子级别对抗子网络中第二判别器的损失函数值和所述句子级别对抗子网络中第二图像生成器的总损失函数值；

所述基于所述第二图像、所述第一句子的句子向量和所述标准图像，利用第二判别器，计算句子级别的条件对抗损失函数值包括：

利用预设的图像编码器，对所述第二图像进行特征抽取，得到第二图像的图像特征；

基于所述第二图像的图像特征和所述第一句子的句子向量，采用交叉熵损失计算方法，进行图文一致性判断，得到第一图文一致性损失函数值$Loss_{CLIP_FT_S}$；

按照$\mathcal{L}_{D_2}=-\frac{1}{2}\mathbb{E}_{x_t\sim data}\ [log(D(x_t,c_s))]-\frac{1}{2}\mathbb{E}_{x_2\sim G_2}[log(1-D(x_2,c_s))]$，计算所述第二判别器的损失函数值$\mathcal{L}_{D_2}$；其中，D( )表示判别器输出的概率值；$x_t$表示样本数据中的标准图像；$x_2$表示第二图像；$\mathbb{E}_{x_t\sim data}\ [log(D(x_t,c_s))]$表示利用标准图像$x_t$和句子向量$c_s$计算第二判别器的第一条件对抗损失函数；$\mathbb{E}_{x_2\sim G_2}[log(1-D(x_2,c_s))]$利用从第二图像生成器$G_2$中得到的第二图像$x_2$和句子向量$c_s$计算第二判别器的第二条件对抗损失函数；

按照$\mathcal{L}_{G_2}=-\frac{1}{2}\mathbb{E}_{x_2\sim G_2}[log(D(x_2,c_s))]$，计算所述第二图像生成器的第一损失函数值$\mathcal{L}_{G_2}$；

基于所述第一图文一致性损失函数值$Loss_{CLIP_FT_S}$和所述第二图像生成器的第一损失函数值$\mathcal{L}_{G_2}$，按照$\mathcal{L}_{G_2-total}=\mathcal{L}_{G_2}+Loss_{CLIP_FT_S}$，计算所述第二图像生成器的总损失函数值$\mathcal{L}_{G_2-total}$。

7.根据权利要求1所述的方法，其特征在于，

所述词级别的条件对抗损失函数值包括所述词级别对抗子网络中第三判别器的损失函数值和所述词级别对抗子网络中第三图像生成器的总损失函数值；

所述基于所述第三图像、所述第一句子的句子向量和词向量以及所述标准图像，利用第三判别器，计算词级别的条件对抗损失函数值包括：

利用预设的图像编码器，对所述第三图像进行特征抽取，得到第三图像的图像特征和图像区域特征；

基于所述第三图像的图像特征和所述第一句子的句子向量，采用交叉熵损失计算方法，进行图文一致性判断，得到第二图文一致性损失函数值；

基于所述第三图像的图像区域特征和所述第一句子的词向量，采用交叉熵损失计算方法，进行图文一致性判断，得到第三图文一致性损失函数值；

基于所述第二图文一致性损失函数值和所述第三图文一致性损失函数值，计算第四图文一致性损失函数值$Loss_{CLIP_FT}$；

按照$\mathcal{L}_{D_3}=-\frac{1}{2}\mathbb{E}_{x_t\sim data}\left[log\left(D(x_t,\widehat{c_w})\right)\right]-\frac{1}{2}\mathbb{E}_{x_3\sim G_3}\left[log\left(1-D(x_3,\widehat{c_w})\right)\right]$，计算所述第三判别器的损失函数值$\mathcal{L}_{D_3}$；其中，D( )表示判别器输出的概率值；$x_t$表示样本数据中的标准图像；$x_3$表示第三图像；$\widehat{c_w}$表示所述词向量的平均值；$\mathbb{E}_{x_t\sim data}\left[log\left(D(x_t,\widehat{c_w})\right)\right]$表示利用标准图像$x_t$和词向量的平均值$\widehat{c_w}$计算所述第三判别器的第一条件对抗损失函数；

$\mathbb{E}_{x_3\sim G_3}\left[log\left(1-D(x_3,\widehat{c_w})\right)\right]$利用从第三图像生成器$G_3$中得到的第三图像$x_3$和词向量的平均值$\widehat{c_w}$计算第三判别器的第二条件对抗损失函数；

按照$\mathcal{L}_{G_3}=-\frac{1}{2}\mathbb{E}_{x_3\sim G_3}\left[log\left(D(x_3,\widehat{c_w})\right)\right]$，计算所述第三图像生成器的第一损失函数值$\mathcal{L}_{G_3}$；

基于所述第四图文一致性损失函数值$Loss_{CLIP_FT}$和所述第三图像生成器的第一损失函数值$\mathcal{L}_{G_3}$，按照$\mathcal{L}_{G_3-total}=\mathcal{L}_{G_3}+Loss_{CLIP_FT}$，计算所述第三图像生成器的总损失函数值$\mathcal{L}_{G_3-total}$。

8.根据权利要求1所述的方法，其特征在于，所述利用所述非条件对抗损失函数值、所述句子级别的条件对抗损失函数值和所述词级别的条件对抗损失函数值，对所述图像生成模型的参数进行更新包括：

利用所述非条件对抗损失函数值，对所述无条件对抗子网络中的参数进行更新；

利用所述句子级别的条件对抗损失函数值，对所述句子级别对抗子网络中的参数进行更新；

利用所述词级别的条件对抗损失函数值，对所述词级别对抗子网络中的参数进行更新。

9.一种图像生成方法，其特征在于，包括：

将图像描述文本，输入至预先训练的图像生成模型进行图像生成，得到所述图像描述文本对应的图像；

其中，所述图像生成模型基于权利要求1至8所述的任一训练方法得到。

10.一种图像生成模型的训练设备，其特征在于，包括处理器和存储器；

所述存储器中存储有可被所述处理器执行的应用程序，用于使得所述处理器执行如权利要求1至8中任一项所述图像生成模型的训练方法。

图像生成模型的训练方法和设备以及图像生成方法

技术领域

[0001] 本发明涉及人工智能技术，特别是涉及一种图像生成模型的训练方法和设备以及图像生成方法。

背景技术

[0002] 近年来，随着深度学习的发展进步，人们不在拘泥于单纯的研究图像、文本或是语音这样单一模态的研究，越来越多的注意力开始转向了多模态结合的领域。在多模态领域中，文本生成图像任务是指能够根据一个给定的文本描述，自动地生成一批符合此描述的虚拟图像。此任务要求生成模型能够充分的理解图像和文本的信息，并根据给定文本描述生成高质量的、符合文本描述的图像，是当下最热门的跨模态生成研究领域之一。

[0003] 在基于文本生成图像时，保证生成图像与给定文本描述之间的一致性和提升生成图像质量是Text-to-Image任务的两个核心难点。

[0004] 利用多层次堆积的对抗网络执行文本生成图像任务，是目前主流的基于文本生成图像的方案之一。该方案基于条件对抗网络的思想，通过使用多层叠加的对抗网络来逐层提升生成图像的质量。

[0005] 发明人在实现本申请的过程中发现：现有的基于文本生成图像无法有效保证生成图像与文本间的一致性。具体原因如下：

[0006] 在现有的多层次堆积生成网络方案所使用的模型中，每一层的判别器都需要同时承担两种判别任务，即判别器不仅需要判别生成图像是否符合文本描述，也需要评估生成图像与真实图像是否相似，而判断图像的真实性时希望模型能够对真实图像和生成图像进行准确区分，做一致判断时，是需要判断图像和文本之间的相似性，两者的目标不一致，从而不利于模型的训练；同时，仅在生成图像的过程中将图像与词向量进行融合，这将会导致在底层模型生成的图像质量不佳的情况下，无法有效地对图像进行修正。基于上述问题的存在，导致上述现有方案存在无法有效保证生成图像与文本间的一致性的问题。

发明内容

[0007] 有鉴于此，本发明的主要目的在于提供一种图像生成模型的训练方法和设备以及图像生成方法，可以有效保证生成图像与文本间的一致性。

[0008] 为了达到上述目的，本发明实施例提出的技术方案为：

[0009] 一种图像生成模型的训练方法，包括：

[0010] 获取样本数据，所述样本数据包括标准图像和图像描述文本；

[0011] 基于所述图像描述文本，分别进行句子级别编码和词级别编码，得到相应的句子向量和词向量；

[0012] 将初始高斯噪声，输入至图像生成模型的无条件对抗子网络进行图像生成，得到第一图像；基于所述第一图像和所述标准图像，利用第一判别器，计算非条件对抗损失函数值；

[0013] 将所述第一图像和第一句子的句子向量，输入至所述图像生成模型的句子级别对抗子网络进行图像生成，得到第二图像；基于所述第二图像、所述第一句子的句子向量和所述标准图像，利用第二判别器，计算句子级别的条件对抗损失函数值；其中，所述第一句子为所述图像描述文本对应的句子；

[0014] 将所述第二图像和所述第一句子中每个词的词向量，输入至所述图像生成模型的词级别对抗子网络进行图像生成，得到第三图像；基于所述第三图像、所述第一句子的句子向量和词向量以及所述标准图像，利用第三判别器，计算词级别的条件对抗损失函数值；

[0015] 利用所述非条件对抗损失函数值、所述句子级别的条件对抗损失函数值和所述词级别的条件对抗损失函数值，对所述图像生成模型的参数进行更新。

[0016] 较佳地，所述将初始高斯噪声，输入至图像生成模型的无条件对抗子网络进行图像生成包括：

[0017] 对所述初始高斯噪声进行重构；

[0018] 利用第一图像生成器，对所述重构的结果进行处理，得到所述第一图像。

[0019] 较佳地，所述将所述第一图像和第一句子的句子向量，输入至所述图像生成模型的句子级别对抗子网络进行图像生成包括：

[0020] 将所述句子向量融入所述第一图像的图像特征中；

[0021] 利用第二图像生成器，对所述融入得到的图像特征进行处理，得到所述第二图像。

[0022] 较佳地，所述将所述第二图像和所述第一句子中每个词的词向量，输入至所述图像生成模型的词级别对抗子网络进行图像生成包括：

[0023] 将所述词向量融入所述第二图像的图像特征中；

[0024] 利用第三图像生成器，对所述融入得到的图像特征进行处理，得到所述第三图像。

[0025] 较佳地，所述非条件对抗损失函数值包括所述无条件对抗子网络中第一判别器的损失函数值和所述无条件对抗子网络中第一图像生成器的损失函数值；

[0026] 所述基于所述第一图像和所述标准图像，利用第一判别器，计算非条件对抗损失函数值包括：

[0027] 按照$\mathcal{L}_{D_1}=-\frac{1}{2}\mathbb{E}_{x_t\sim data}\left[log\left(D(x_t)\right)\right]-\frac{1}{2}\mathbb{E}_{x_1\sim G_1}\left[log\left(1-D(x_1)\right)\right]$，计算所述第一判别器的损失函数值$\mathcal{L}_{D_1}$；其中，D()表示判别器输出的概率值；$x_t$表示样本数据中的标准图像；$x_1$表示第一图像；$\mathbb{E}_{x_t\sim data}\left[log\left(D(x_t)\right)\right]$表示利用标准图像$x_t$计算第一判别器的第一非条件对抗损失函数；$\mathbb{E}_{x_1\sim G_1}\left[log\left(1-D(x_1)\right)\right]$表示利用从第一图像生成器$G_1$中得到的第一图像$x_1$计算第一判别器的第二非条件对抗损失函数；

[0028] 按照$\mathcal{L}_{G_1}=-\frac{1}{2}\mathbb{E}_{x_1\sim G_1}\left[log\left(D(x_1)\right)\right]$，计算所述第一图像生成器的损失函数值$\mathcal{L}_{G_1}$。

[0029] 较佳地，所述句子级别的条件对抗损失函数值包括所述句子级别对抗子网络中第二判别器的损失函数值和所述句子级别对抗子网络中第二图像生成器的总损失函数值；

[0030] 所述基于所述第二图像、所述第一句子的句子向量和所述标准图像，利用第二判别器，计算句子级别的条件对抗损失函数值包括：

[0031] 利用预设的图像编码器，对所述第二图像进行特征抽取，得到第二图像的图像特

征；

[0032] 基于所述第二图像的图像特征和所述第一句子的句子向量，采用交叉熵损失计算方法，进行图文一致性判断，得到第一图文一致性损失函数值$Loss_{CLIP_FT_S}$；

[0033] 按照$\mathcal{L}_{D_2}=-\frac{1}{2}\mathbb{E}_{x_t\sim data}\left[log\left(D(x_t,c_s)\right)\right]-\frac{1}{2}\mathbb{E}_{x_2\sim G_2}\left[log\left(1-D(x_2,c_s)\right)\right]$，计算所述第二判别器的损失函数值$\mathcal{L}_{D_2}$；其中，D()表示判别器输出的概率值；$x_t$表示样本数据中的标准图像；$x_2$表示第二图像；$\mathbb{E}_{x_t\sim data}\left[log\left(D(x_t,c_s)\right)\right]$表示利用标准图像$x_t$和句子向量$c_s$计算第二判别器的第一条件对抗损失函数；$\mathbb{E}_{x_2\sim G_2}\left[log\left(1-D(x_2,c_s)\right)\right]$利用从第二图像生成器$G_2$中得到的第二图像$x_2$和句子向量$c_s$计算第二判别器的第二条件对抗损失函数；

[0034] 按照$\mathcal{L}_{G_2}=-\frac{1}{2}\mathbb{E}_{x_2\sim G_2}\left[log\left(D(x_2,c_s)\right)\right]$，计算所述第二图像生成器的第一损失函数值$\mathcal{L}_{G_2}$；

[0035] 基于所述第一图文一致性损失函数值$Loss_{CLIP_FT_S}$和所述第二图像生成器的第一损失函数值$\mathcal{L}_{G_2}$，按照$\mathcal{L}_{G_2-total}=\mathcal{L}_{G_2}+Loss_{CLIP_FT_S}$，计算所述第二图像生成器的总损失函数值$\mathcal{L}_{G_2-total}$。

[0036] 较佳地，所述词级别的条件对抗损失函数值包括所述词级别对抗子网络中第三判别器的损失函数值和所述词级别对抗子网络中第三图像生成器的总损失函数值；

[0037] 所述基于所述第三图像、所述第一句子的句子向量和词向量以及所述标准图像，利用第三判别器，计算词级别的条件对抗损失函数值包括：

[0038] 利用预设的图像编码器，对所述第三图像进行特征抽取，得到第三图像的图像特征和图像区域特征；

[0039] 基于所述第三图像的图像特征和所述第一句子的句子向量，采用交叉熵损失计算方法，进行图文一致性判断，得到第二图文一致性损失函数值；

[0040] 基于所述第三图像的图像区域特征和所述第一句子的词向量，采用交叉熵损失计算方法，进行图文一致性判断，得到第三图文一致性损失函数值；

[0041] 基于所述第二图文一致性损失函数值和所述第三图文一致性损失函数值，计算第四图文一致性损失函数值$Loss_{CLIP_FT}$；

[0042] 按照$\mathcal{L}_{D_3}=-\frac{1}{2}\mathbb{E}_{x_t\sim data}\left[log\left(D(x_t,\widehat{c_w})\right)\right]-\frac{1}{2}\mathbb{E}_{x_3\sim G_3}\left[log\left(1-D(x_3,\widehat{c_w})\right)\right]$，计算所述第三判别器的损失函数值$\mathcal{L}_{D_3}$；其中，D()表示判别器输出的概率值；$x_t$表示样本数据中的标准图像；$x_3$表示第三图像；$\widehat{c_w}$表示所述词向量的平均值；$\mathbb{E}_{x_t\sim data}\left[log\left(D(x_t,\widehat{c_w})\right)\right]$表示利用标准图像$x_t$和词向量的平均值$\widehat{c_w}$计算所述第三判别器的第一条件对抗损失函数；$\mathbb{E}_{x_3\sim G_3}\left[log\left(1-D(x_3,\widehat{c_w})\right)\right]$利用从第三图像生成器$G_3$中得到的第三图像$x_3$和词向量的平均值$\widehat{c_w}$计算第三判别器的第二条件对抗损失函数；

[0043] 按照$\mathcal{L}_{G_3}=-\frac{1}{2}\mathbb{E}_{x_3\sim G_3}\left[log\left(D(x_3,\widehat{c_w})\right)\right]$，计算所述第三图像生成器的第一损失函数值

$\mathcal{L}_{G_3}$；

[0044] 基于所述第四图文一致性损失函数值$Loss_{CLIP_FT}$和所述第三图像生成器的第一损失函数值$\mathcal{L}_{G_3}$，按照$\mathcal{L}_{G_3-total} = \mathcal{L}_{G_3} + Loss_{CLIP_FT}$，计算所述第三图像生成器的总损失函数值$\mathcal{L}_{G_3-total}$。

[0045] 较佳地，所述利用所述非条件对抗损失函数值、所述句子级别的条件对抗损失函数值和所述词级别的条件对抗损失函数值，对所述图像生成模型的参数进行更新包括：

[0046] 利用所述非条件对抗损失函数值，对所述无条件对抗子网络中的参数进行更新；

[0047] 利用所述句子级别的条件对抗损失函数值，对所述句子级别对抗子网络中的参数进行更新；

[0048] 利用所述词级别的条件对抗损失函数值，对所述词级别对抗子网络中的参数进行更新。

[0049] 本发明实施例还提供了一种图像生成方法，包括：

[0050] 将图像描述文本，输入至预先训练的图像生成模型进行图像生成，得到所述图像描述文本对应的图像；

[0051] 其中，所述图像生成模型基于上述任一训练方法得到。

[0052] 本发明实施例还提供了一种图像生成模型的训练设备，包括处理器和存储器；

[0053] 所述存储器中存储有可被所述处理器执行的应用程序，用于使得所述处理器执行如上所述图像生成模型的训练方法。

[0054] 综上所述，本发明实施例提出的上述方案，先利用无条件对抗子网络基于初始高斯噪声生成第一图像，再分别利用句子级别对抗子网络和词级别对抗子网络生成第二图像和第三图像。如此，通过引入渐进式的生成对抗网络结构，采用逐层加入语言信息的方法，由粗至细的对生成图像进行优化，能够更有效地保证生成图像的质量和图文一致性。

附图说明

[0055] 图1为本发明实施例的方法流程示意图。

具体实施方式

[0056] 为使本发明的目的、技术方案和优点更加清楚，下面将结合附图及具体实施例对本发明作进一步地详细描述。

[0057] 图1为本发明实施例的图像生成模型的训练方法流程示意图，如图1所示，该实施例主要包括：

[0058] 步骤101、获取样本数据，所述样本数据包括标准图像和图像描述文本。

[0059] 样本数据中的图像描述文本用于对标准图像进行描述，具体的，可以包括多个不同表述形式的图像描述语句。

[0060] 步骤102、基于所述图像描述文本，分别进行句子级别编码和词级别编码，得到相应的句子向量和词向量。

[0061] 本步骤具体可以采用现有方法实现。即利用预先训练的文本编码器(如Transformer)，生成所述句子向量和词向量。

[0062] 步骤103、将初始高斯噪声，输入至图像生成模型的无条件对抗子网络进行图像生成，得到第一图像；基于所述第一图像和所述标准图像，利用第一判别器，计算非条件对抗损失函数值。

[0063] 这里需要说明的是，为了在初始阶段尽可能地提升生成图像的质量，以保证后续模型能够获得更好的基础图像输入，本步骤中，直接使用一个随机的初始高斯噪声作为输入。

[0064] 另外，这里第一判别器仅用于评判输入图像是生成图像还是真实图像，而无需关注输入图像是否符合文本描述，因此，在生成第一图像时和利用第一判别器计算损失函数值时，都无需使用条件信息作为输入。

[0065] 在一种实施方式中，具体可以采用下述方法将初始高斯噪声，输入至图像生成模型的无条件对抗子网络进行图像生成：

[0066] 对所述初始高斯噪声进行重构；利用第一图像生成器，对所述重构的结果进行处理，得到所述第一图像。

[0067] 具体重构方法为本领域技术人员所掌握，在此不再赘述。

[0068] 具体的，在进行一次重构后，可以直接通过三层上采样获得第一图像。

[0069] 具体地，步骤103中计算的非条件对抗损失函数值包括所述无条件对抗子网络中第一判别器的损失函数值和所述无条件对抗子网络中第一图像生成器的损失函数值；

[0070] 在一种实施方式中，步骤103中具体可以采用下述方法基于所述第一图像和所述标准图像，利用第一判别器，计算非条件对抗损失函数值：

[0071] 按照$\mathcal{L}_{D_1} = -\frac{1}{2}\mathbb{E}_{x_t\sim data}\left[log(D(x_t))\right] - \frac{1}{2}\mathbb{E}_{x_1\sim G_1}\left[log(1-D(x_1))\right]$，计算所述第一判别器的损失函数值$\mathcal{L}_{D_1}$。其中，D()表示判别器输出的概率值；$x_t$表示样本数据中的标准图像；$x_1$表示第一图像；$\mathbb{E}_{x_t\sim data}\left[log(D(x_t))\right]$表示利用标准图像$x_t$计算第一判别器的第一非条件对抗损失函数；$\mathbb{E}_{x_1\sim G_1}\left[log(1-D(x_1))\right]$表示利用从第一图像生成器$G_1$中得到的第一图像$x_1$计算第一判别器的第二非条件对抗损失函数；

[0072] 按照$\mathcal{L}_{G_1} = -\frac{1}{2}\mathbb{E}_{x_1\sim G_1}\left[log(D(x_1))\right]$，计算所述第一图像生成器的损失函数值$\mathcal{L}_{G_1}$。

[0073] 步骤104、将所述第一图像和第一句子的句子向量，输入至所述图像生成模型的句子级别对抗子网络进行图像生成，得到第二图像；基于所述第二图像、所述第一句子的句子向量和所述标准图像，利用第二判别器，计算句子级别的条件对抗损失函数值；其中，所述第一句子为所述图像描述文本对应的句子。

[0074] 本步骤用于在步骤103得到的低层图像特征分布(即第一图像)的基础上，引入图像描述文本的句子向量，进一步进行图像优化，生成第二图像，同时对生成图像施加句子级别的监督约束。

[0075] 在一种实施方式中，具体可以采用下述步骤x1～x2，将所述第一图像和第一句子的句子向量，输入至所述图像生成模型的句子级别对抗子网络进行图像生成：

[0076] 步骤x1、将所述句子向量融入所述第一图像的图像特征中。

[0077] 具体地，本步骤可以采用注意力机制的方法(Attn-GAN)或是动态内存的方法(DM-

GAN)，通过对句子信息进行复制，并与图像信息交互，从而获得融合了句子信息的图像特征信息。

[0078] 步骤x2、利用第二图像生成器，对所述融入得到的图像特征进行处理，得到所述第二图像。

[0079] 具体地，本步骤可以利用两个残差层和一个上采样层，来生成相比于第一图像具有更高分辨率且包含句子特征的第二图像。

[0080] 具体地，步骤104中计算的句子级别的条件对抗损失函数值包括所述句子级别对抗子网络中第二判别器的损失函数值和所述句子级别对抗子网络中第二图像生成器的总损失函数值；

[0081] 相应地，在一种实施方式中，为了加强图文一致性约束，具体可以采用下述方法基于所述第二图像、所述第一句子的句子向量和所述标准图像，利用第二判别器，计算句子级别的条件对抗损失函数值：

[0082] 步骤y1、利用预设的图像编码器，对所述第二图像进行特征抽取，得到第二图像的图像特征。

[0083] 本步骤用于从第二图像中抽取图像特征，以便在步骤y2中结合句子向量，进行图文一致性判断。

[0084] 步骤y2、基于所述第二图像的图像特征和所述第一句子的句子向量，采用交叉熵损失计算方法，进行图文一致性判断，得到第一图文一致性损失函数值$Loss_{CLIP_FT_S}$。

[0085] 本步骤的具体实现为本领域技术人员所掌握，在此不再赘述。

[0086] 步骤y3、按照$\mathcal{L}_{D_2}=-\frac{1}{2}\mathbb{E}_{x_t\sim data}\left[log\left(D(x_t,c_s)\right)\right]-\frac{1}{2}\mathbb{E}_{x_2\sim G_2}\left[log\left(1-D(x_2,c_s)\right)\right]$，计算所述第二判别器的损失函数值$\mathcal{L}_{D_2}$；其中，D()表示判别器输出的概率值；$x_t$表示样本数据中的标准图像；$x_2$表示第二图像；$\mathbb{E}_{x_t\sim data}\left[log\left(D(x_t,c_s)\right)\right]$表示利用标准图像$x_t$和句子向量$c_s$计算第二判别器的第一条件对抗损失函数；$\mathbb{E}_{x_2\sim G_2}\left[log\left(1-D(x_2,c_s)\right)\right]$利用从第二图像生成器$G_2$中得到的第二图像$x_2$和句子向量$c_s$计算第二判别器的第二条件对抗损失函数。

[0087] 步骤y4、按照$\mathcal{L}_{G_2}=-\frac{1}{2}\mathbb{E}_{x_2\sim G_2}\left[log\left(D(x_2,c_s)\right)\right]$，计算所述第二图像生成器的第一损失函数值$\mathcal{L}_{G_2}$。

[0088] 步骤y5、基于所述第一图文一致性损失函数值$Loss_{CLIP_FT_S}$和所述第二图像生成器的第一损失函数值$\mathcal{L}_{G_2}$，按照$\mathcal{L}_{G_2-total}=\mathcal{L}_{G_2}+Loss_{CLIP_FT_S}$，计算所述第二图像生成器的总损失函数值$\mathcal{L}_{G_2-total}$。

[0089] 这里，为了进一步加强对于图文整体一致性的约束，以提高第二判别器的图文一致性判别能力，在步骤y5中计算第二图像生成器的总损失函数值$\mathcal{L}_{G_2-total}$时，引入了上述第一图文一致性损失函数值$Loss_{CLIP_FT_S}$，即按照$\mathcal{L}_{G_2-total}=\mathcal{L}_{G_2}+Loss_{CLIP_FT_S}$，计算$\mathcal{L}_{G_2-total}$。

[0090] 在上述方法中，第二判别器用于评判输入图像与句子信息是否匹配，同时，考虑到在初始生成第一图像时，尽可能地限制了生成图像的质量，在生成第二图像的阶段，不再添加针对图像质量的无条件对抗损失。

[0091] 步骤105、将所述第二图像和所述第一句子中每个词的词向量，输入至所述图像生成模型的词级别对抗子网络进行图像生成，得到第三图像；基于所述第三图像、所述第一句子的句子向量和词向量以及所述标准图像，利用第三判别器，计算词级别的条件对抗损失函数值。

[0092] 这里需要说明的是，在步骤104中通过在第一图像中融入句子信息，能够有效的提升生成图像的一致性表现，但也仅仅只是句子粒度的细化，缺乏针对图像局部和细粒度语言信息的细节优化。为此，在步骤105中引入了文本的词向量，结合第二图像，来进一步进行图像优化，得到包含更多细节的第三图像。同时对生成图像施加词级别的监督约束。

[0093] 本步骤中，第三判别器用于评判输入图像与句子、词信息是否匹配，同时，考虑到在初始生成第一图像时，已经尽可能地限制了生成图像的质量，在生成第三图像的阶段，不再添加针对图像质量的无条件对抗损失。

[0094] 在一种实施方式中，步骤105中具体可以采用下述方法将所述第二图像和所述第一句子中每个词的词向量，输入至所述图像生成模型的词级别对抗子网络进行图像生成：

[0095] k1、将所述词向量融入所述第二图像的图像特征中。

[0096] 具体地，上述方法中仍可以采用注意力机制的方法或是动态内存的方法，通过对词信息进行复制，并与图像信息交互，从而获得融合了词信息的图像特征信息。

[0097] k2、利用第三图像生成器，对所述融入得到的图像特征进行处理，得到所述第三图像。

[0098] 具体地，本步骤可以利用残差层和上采样层，来生成相比于第二图像具有更高分辨率的第三图像。

[0099] 具体地，在步骤105中计算的词级别的条件对抗损失函数值包括所述词级别对抗子网络中第三判别器的损失函数值和所述词级别对抗子网络中第三图像生成器的总损失函数值。

[0100] 相应的，在一种实施方式中，可以采用下述方法基于所述第三图像、所述第一句子的句子向量和词向量以及所述标准图像，利用第三判别器，计算词级别的条件对抗损失函数值：

[0101] 步骤z1、利用预设的图像编码器，对所述第三图像进行特征抽取，得到第三图像的图像特征和图像区域特征。

[0102] 本步骤用于从第三图像中抽取图像特征和图像区域特征，以便在后续步骤z2和z3中将它们分别与句子向量和词向量相结合，进行图文一致性判断。

[0103] 本步骤的具体实现方法为本领域技术人员所掌握，在此不再赘述。

[0104] 步骤z2、基于所述第三图像的图像特征和所述第一句子的句子向量，采用交叉熵损失计算方法，进行图文一致性判断，得到第二图文一致性损失函数值。

[0105] 步骤z3、基于所述第三图像的图像区域特征和所述第一句子的词向量，采用交叉熵损失计算方法，进行图文一致性判断，得到第三图文一致性损失函数值。

[0106] 这里通过基于词向量和图像区域特征，计算图文相似度，能够更细致地分析出图

像细节与重点单词之间的一致性关系，从而能够更细粒度评价图像和文本之间的一致性。

[0107] 步骤z4、基于所述第二图文一致性损失函数值和所述第三图文一致性损失函数值，计算第四图文一致性损失函数值$Loss_{CLIP_FT}$。

[0108] 本步骤具体可以采用加权计算的方法，计算得到第四图文一致性损失函数值$Loss_{CLIP_FT}$。

[0109] 较佳地，为了使得训练得到的模型能够更好地保证图文一致性，在进行上述加权计算时，可以设置第二图文一致性损失函数值的权重大于第三图文一致性损失函数值的权重，例如，第二图文一致性损失函数值的权重为4，第三图文一致性损失函数值的权重为1，但不限于此。

[0110] 步骤z5、按照$\mathcal{L}_{D_3} = -\frac{1}{2}\mathbb{E}_{x_t\sim data}\left[log\left(D(x_t,\widehat{c_w})\right)\right] - \frac{1}{2}\mathbb{E}_{x_3\sim G_3}\left[log\left(1-D(x_3,\widehat{c_w})\right)\right]$，计算所述第三判别器的损失函数值$\mathcal{L}_{D_3}$；其中，D()表示判别器输出的概率值；$x_t$表示样本数据中的标准图像；$x_3$表示第三图像；$\widehat{c_w}$表示所述词向量的平均值；$\mathbb{E}_{x_t\sim data}\left[log\left(D(x_t,\widehat{c_w})\right)\right]$表示利用标准图像$x_t$和词向量的平均值$\widehat{c_w}$计算所述第三判别器的第一条件对抗损失函数；$\mathbb{E}_{x_3\sim G_3}\left[log\left(1-D(x_3,\widehat{c_w})\right)\right]$利用从第三图像生成器$G_3$中得到的第三图像$x_3$和词向量的平均值$\widehat{c_w}$计算第三判别器的第二条件对抗损失函数。

[0111] 这里考虑到词信息的丰富性，在将词向量信息输入判别器之前，需要先对词向量信息进行浓缩，即：将词向量进行平均化处理得到$\widehat{c_w}$后，再将$\widehat{c_w}$输入判别器处理。

[0112] 步骤z6、按照$\mathcal{L}_{G_3} = -\frac{1}{2}\mathbb{E}_{x_3\sim G_3}\left[log\left(D(x_3,\widehat{c_w})\right)\right]$，计算所述第三图像生成器的第一损失函数值$\mathcal{L}_{G_3}$。

[0113] 步骤z7、基于所述第四图文一致性损失函数值$Loss_{CLIP_FT}$和所述第三图像生成器的第一损失函数值$\mathcal{L}_{G_3}$，按照$\mathcal{L}_{G_3-total} = \mathcal{L}_{G_3} + Loss_{CLIP_FT}$，计算所述第三图像生成器的总损失函数值$\mathcal{L}_{G_3-total}$。

[0114] 这里，为了进一步加强对于图文整体一致性的约束，以提高第三判别器的图文一致性判别能力，在步骤z7中计算所述第三图像生成器的总损失函数值$\mathcal{L}_{G_3-total}$时，引入了上述第四图文一致性损失函数值$Loss_{CLIP_FT}$，即按照$\mathcal{L}_{G_3-total} = \mathcal{L}_{G_3} + Loss_{CLIP_FT}$，计算$\mathcal{L}_{G_3-total}$。

[0115] 步骤106、利用所述非条件对抗损失函数值、所述句子级别的条件对抗损失函数值和所述词级别的条件对抗损失函数值，对所述图像生成模型的参数进行更新。

[0116] 在一种实施方式中，本步骤具体可以采用下述方法对所述图像生成模型的参数进行更新：

[0117] 利用所述非条件对抗损失函数值，对所述无条件对抗子网络中的参数进行更新；

[0118] 利用所述句子级别的条件对抗损失函数值，对所述句子级别对抗子网络中的参数进行更新；

[0119] 利用所述词级别的条件对抗损失函数值，对所述词级别对抗子网络中的参数进行

更新。

[0120] 上述参数更新的具体方法为本领域技术人员所掌握,在此不再赘述。

[0121] 基于上述训练方法实施例,本发明实施例还提供了一种图像生成方法,包括:

[0122] 将图像描述文本,输入至预先训练的图像生成模型进行图像生成,得到所述图像描述文本对应的图像;其中,所述图像生成模型基于如上所述的任一训练方法得到。

[0123] 基于上述训练方法实施例,本发明实施例还提供了一种图像生成模型的训练设备,包括处理器和存储器;所述存储器中存储有可被所述处理器执行的应用程序,用于使得所述处理器执行如上所述图像生成模型的训练方法。具体地,可以提供配有存储介质的系统或者装置,在该存储介质上存储着实现上述实施例中任一实施方式的功能的软件程序代码,且使该系统或者装置的计算机(或CPU或MPU)读出并执行存储在存储介质中的程序代码。此外,还可以通过基于程序代码的指令使计算机上操作的操作系统等来完成部分或者全部的实际操作。还可以将从存储介质读出的程序代码写到插入计算机内的扩展板中所设置的存储器中或者写到与计算机相连接的扩展单元中设置的存储器中,随后基于程序代码的指令使安装在扩展板或者扩展单元上的CPU等来执行部分和全部实际操作,从而实现上述图像生成模型的训练方法实施方式中任一实施方式的功能。

[0124] 其中,存储器具体可以实施为电可擦可编程只读存储器(EEPROM)、快闪存储器(Flash memory)、可编程程序只读存储器(PROM)等多种存储介质。处理器可以实施为包括一或多个中央处理器或一或多个现场可编程门阵列,其中现场可编程门阵列集成一或多个中央处理器核。具体地,中央处理器或中央处理器核可以实施为CPU或MCU。

[0125] 需要说明的是,上述各流程和各结构图中不是所有的步骤和模块都是必须的,可以根据实际的需要忽略某些步骤或模块。各步骤的执行顺序不是固定的,可以根据需要进行调整。各模块的划分仅仅是为了便于描述采用的功能上的划分,实际实现时,一个模块可以分由多个模块实现,多个模块的功能也可以由同一个模块实现,这些模块可以位于同一个设备中,也可以位于不同的设备中。

[0126] 各实施方式中的硬件模块可以以机械方式或电子方式实现。例如,一个硬件模块可以包括专门设计的永久性电路或逻辑器件(如专用处理器,如FPGA或ASIC)用于完成特定的操作。硬件模块也可以包括由软件临时配置的可编程逻辑器件或电路(如包括通用处理器或其它可编程处理器)用于执行特定操作。至于具体采用机械方式,或是采用专用的永久性电路,或是采用临时配置的电路(如由软件进行配置)来实现硬件模块,可以根据成本和时间上的考虑来决定。

[0127] 在本文中,“示意性”表示“充当实例、例子或说明”,不应将在本文中被描述为“示意性”的任何图示、实施方式解释为一种更优选的或更具优点的技术方案。为使图面简洁,各图中的只示意性地表示出了与本发明相关部分,而并不代表其作为产品的实际结构。另外,以使图面简洁便于理解,在有些图中具有相同结构或功能的部件,仅示意性地绘示了其中的一个,或仅标出了其中的一个。在本文中,“一个”并不表示将本发明相关部分的数量限制为“仅此一个”,并且“一个”不表示排除本发明相关部分的数量“多于一个”的情形。在本文中,“上”、“下”、“前”、“后”、“左”、“右”、“内”、“外”等仅用于表示相关部分之间的相对位置关系,而非限定这些相关部分的绝对位置。

[0128] 以上所述,仅为本发明的较佳实施例而已,并非用于限定本发明的保护范围。凡在

本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

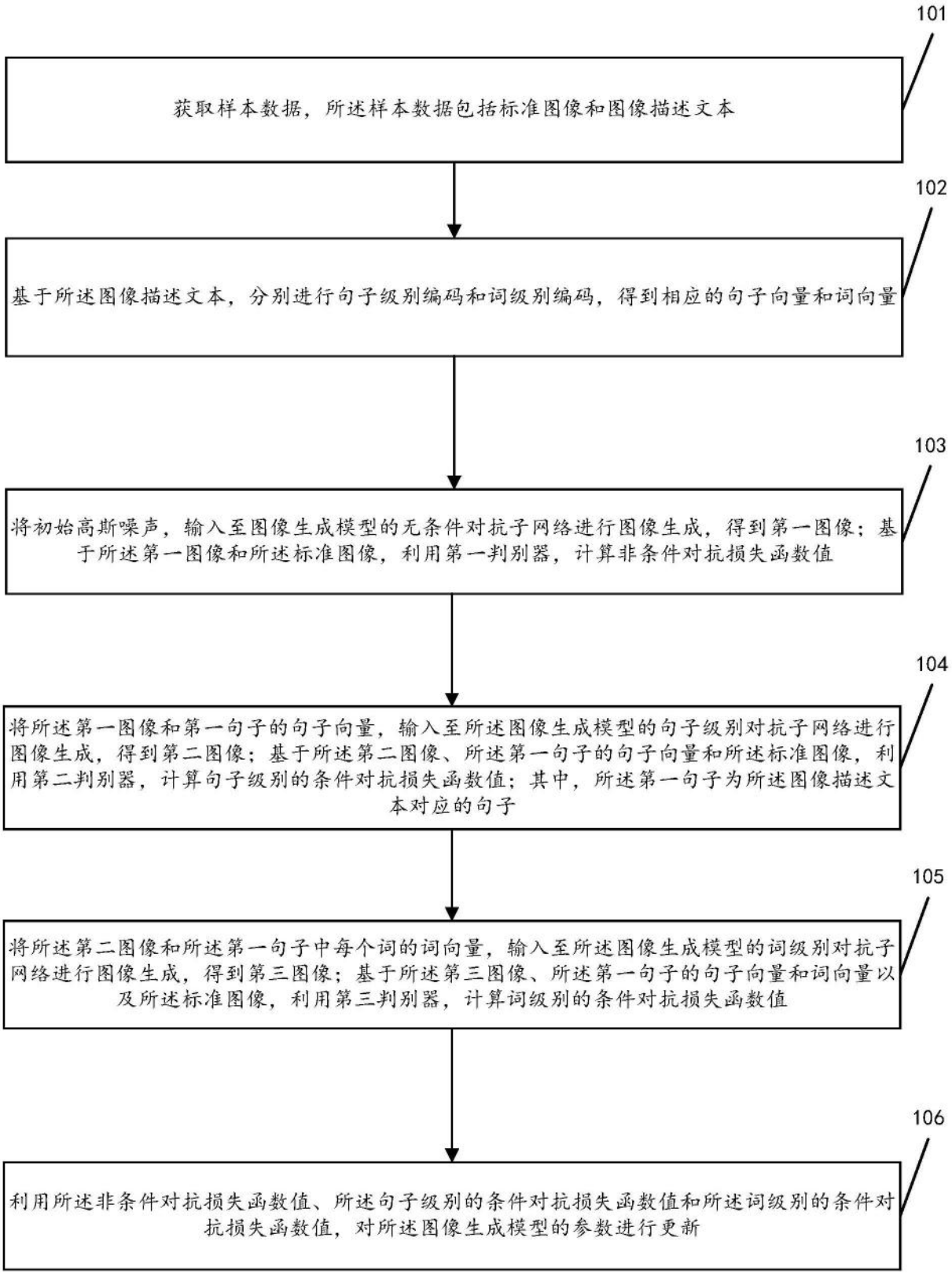

101
获取样本数据，所述样本数据包括标准图像和图像描述文本
102
基于所述图像描述文本，分别进行句子级别编码和词级别编码，得到相应的句子向量和词向量
103
将初始高斯噪声，输入至图像生成模型的无条件对抗子网络进行图像生成，得到第一图像；基于所述第一图像和所述标准图像，利用第一判别器，计算非条件对抗损失函数值
104
将所述第一图像和第一句子的句子向量，输入至所述图像生成模型的句子级别对抗子网络进行图像生成，得到第二图像；基于所述第二图像、所述第一句子的句子向量和所述标准图像，利用第二判别器，计算句子级别的条件对抗损失函数值；其中，所述第一句子为所述图像描述文本对应的句子
105
将所述第二图像和所述第一句子中每个词的词向量，输入至所述图像生成模型的词级别对抗子网络进行图像生成，得到第三图像；基于所述第三图像、所述第一句子的句子向量和词向量以及所述标准图像，利用第三判别器，计算词级别的条件对抗损失函数值
106
利用所述非条件对抗损失函数值、所述句子级别的条件对抗损失函数值和所述词级别的条件对抗损失函数值，对所述图像生成模型的参数进行更新

图1