(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112990302 B
(45)授权公告日 2023.03.21

(21)申请号 202110266563.5
(22)申请日 2021.03.11
(65)同一申请的已公布的文献号
申请公布号 CN 112990302 A
(43)申请公布日 2021.06.18
(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 冯方向 牛天睿 王小捷 李睿凡 袁彩霞
(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.
G06V 10/774(2022.01)
G06V 10/74(2022.01)
G06N 20/00(2019.01)
G06T 11/00(2006.01)
审查员 马丽莉

权利要求书3页 说明书7页 附图1页

(54)发明名称

基于文本生成图像的模型训练方法、设备和图像生成方法

(57)摘要

本申请公开了一种基于文本生成图像的模型训练方法、设备和图像生成方法，其中方法包括基于预设训练样本集合中各训练样本的文本信息，生成文本嵌入式表示，将所述文本嵌入式表示输入至图像生成模型，触发图像生成模型基于该文本嵌入式表示，生成人造图像，并采用模态解纠缠方式，提取人造图像的真实度参数和相应训练样本的真实图像的真实度参数；基于所述人造图像，确定所述训练样本的正例和负例；利用所述图像生成模型，基于每个训练样本的正例、负例和真实图像各自对应的所述真实度参数，计算总体损失函数；利用所述总体损失函数，调整所述图像生成模型的参数。采用本申请可以提高模型学习效率和图像生成效果。

CN 112990302 B

1.一种基于文本生成图像的模型训练方法,其特征在于,包括:

对于预设训练样本集合中的每个训练样本,基于该训练样本的文本信息,生成相应的文本嵌入式表示,并将所述文本嵌入式表示输入至图像生成模型,触发所述图像生成模型基于所述文本嵌入式表示,生成所述文本信息对应的人造图像,并采用模态解纠缠方式,提取所述人造图像的真实度参数和相应训练样本的真实图像的真实度参数;所述真实度参数包括:图像风格的视觉可信度、图-文相似度和图像的整体视觉可信度;

基于所述人造图像,确定所述训练样本的正例和负例;

利用所述图像生成模型,基于所述训练样本的正例、负例和真实图像各自对应的所述真实度参数,计算相应样本的子损失函数,并基于所述子损失函数,计算相应样本的总体损失函数;所述子损失函数包括内容损失函数、风格损失函数、生成器损失函数和判别器损失函数;所述总体损失函数包括判别器总体损失函数和生成器总体损失函数;

利用所述总体损失函数,调整所述图像生成模型的参数;

其中,所述采用模态解纠缠方式,提取所述人造图像的真实度参数和相应训练样本的真实图像的真实度参数包括:

利用所述图像生成模型的图像编码器,从所述人造图像中提取出模态公共表征和模态特定表征,以及从所述训练样本的真实图像中提取出模态公共表征和模态特定表征;

基于所述人造图像的模态公共表征和模态特定表征、所述真实图像的模态公共表征和模态特定表征,利用所述图像生成模型的判别器,提取所述人造图像和所述真实图像的真实度参数。

2.根据权利要求1所述的方法,其特征在于,所述生成所述文本信息对应的人造图像包括:

将所述文本嵌入式表示,输入至图像生成模型的文本编码器处理,得到所述文本信息的文本特征;

将所述文本特征和训练样本对应的预设噪声样本,输入至所述图像生成模型的生成器处理,得到所述文本信息对应的人造图像。

3.根据权利要求1所述的方法,其特征在于,所述真实度参数的提取包括:

按照$s_s(\hat{x}) = D^s(h_{ss})$,提取所述人造图像的图像风格的视觉可信度;其中,$h_{ss}$为所述人造图像的模态特定表征;$D^s$为所述图像生成模型的模态特定判别器;$s_s(\hat{x})$表示所述人造图像的图像风格的视觉可信度;

按照$s_s(x) = D^s(h_{is})$,提取所述真实图像的图像风格的视觉可信度;其中,h_{is}为所述真实图像的模态特定表征;$s_s(x)$表示所述真实图像的图像风格的视觉可信度;

按照$s_c(\hat{x}) = D^c(h_{tc}, h_{sc})$,提取所述人造图像的图-文相似度;其中,$h_{tc}$为所述文本信息的文本特征;$h_{sc}$为所述人造图像的模态公共表征;$D^c$为所述图像生成模型的模态公共判别器;$s_c(\hat{x})$表示所述人造图像的图-文相似度;

按照$s_c(x) = D^c(h_{tc}, h_{ic})$,提取所述真实图像的图-文相似度;其中,$h_{ic}$为所述真实图像的模态公共表征;$s_c(x)$表示所述真实图像的图-文相似度;

按照$s_i(\hat{x}) = D^i(h_{sc}, h_{ss})$,提取所述人造图像的整体视觉可信度;其中,$D^i$为所述图像生成

模型的整体视觉判别器；$s_i(\hat{x})$表示所述人造图像的整体视觉可信度；

按照$s_i(x)=D^i(h_{ic},h_{is})$，提取所述真实图像的整体视觉可信度；其中，$s_i(x)$表示所述真实图像的整体视觉可信度。

4.根据权利要求1所述的方法，其特征在于，所述确定每个所述训练样本的正例和负例包括：

对于每个所述训练样本，将该训练样本对应的所述人造图像作为该训练样本的正例，从基于非该训练样本对应的所述人造图像中，选择一个图像作为该训练样本的负例。

5.根据权利要求1所述的方法，其特征在于，所述计算相应样本的子损失函数包括：

按照$\mathcal{L}_C=\mathcal{L}(h_{tc},h_{sc},h_{sc}^-)$，计算所述内容损失函数值$\mathcal{L}_C$；其中，$\mathcal{L}(\ )$表示三元组损失函数，$h_{tc}$为锚点，表示训练样本的文本信息的文本特征；$h_{sc}$为正例，表示训练样本对应的人造图像的模态公共表征；h_{sc}^-为负例，表示训练样本的负例的模态公共表征；

按照$\mathcal{L}_S=-\rho(z,h_{ss})$，计算所述风格损失函数$\mathcal{L}_S$；其中，z为噪声样本；$h_{ss}$表示训练样本对应的人造图像的模态特定表征；ρ为预设的皮尔逊相关系数；

按照$\mathcal{L}_G=-\mathbb{E}_{\hat{x}\sim p_G}[\log(s_c(\hat{x}))]-\mathbb{E}_{\hat{x}\sim p_G}[\log(s_s(\hat{x}))]-\mathbb{E}_{\hat{x}\sim p_G}[\log(s_i(\hat{x}))]$，计算所述生成器损失函数$\mathcal{L}_G$；其中，$\mathbb{E}_{\hat{x}\sim p_G}[\ ]$表示从生成器PG中采样得到的训练样本对应的人造图像$\hat{x}$；$s_c(\hat{x})$表示人造图像$\hat{x}$的图-文相似度；$s_s(\hat{x})$表示人造图像$\hat{x}$的图像风格的视觉可信度；$s_i(\hat{x})$表示人造图像$\hat{x}$的整体视觉可信度；

按照
$$\begin{aligned}\mathcal{L}_D=&-\mathbb{E}_{x\sim p_{data}}[\log(s_c(x))]-\mathbb{E}_{\hat{x}\sim p_G}[\log(1-s_c(\hat{x}))]\\&-\mathbb{E}_{x\sim p_{data}}[\log(s_s(x))]-\mathbb{E}_{\hat{x}\sim p_G}[\log(1-s_s(\hat{x}))]\\&-\mathbb{E}_{x\sim p_{data}}[\log(s_i(x))]-\mathbb{E}_{\hat{x}\sim p_G}[\log(1-s_i(\hat{x}))],\end{aligned}$$
计算所述判别器损失函数$\mathcal{L}_D$，其中，$\mathbb{E}_{x\sim p_{data}}[\ ]$表示从训练样本集合$p_{data}$中得到的训练样本的真实图像x；$s_s(x)$表示真实图像x的图像风格的视觉可信度；$s_c(x)$表示真实图像x的图-文相似度；$s_i(x)$表示真实图像x的整体视觉可信度。

6.根据权利要求1所述的方法，其特征在于，所述基于所述子损失函数，计算相应样本的总体损失函数包括：

按照$\mathcal{L}_{D_{total}}=\mathcal{L}_D+\mathcal{L}_C+\mathcal{L}_S$，得到所述判别器总体损失函数$\mathcal{L}_{D_{total}}$，其中，$\mathcal{L}_D$为所述判别器损失函数；$\mathcal{L}_C$为所述内容损失函数；$\mathcal{L}_S$为所述风格损失函数；

按照$\mathcal{L}_{G_{total}}=\mathcal{L}_G+\mathcal{L}_C$，得到所述生成器总体损失函数$\mathcal{L}_{G_{total}}$，其中，$\mathcal{L}_G$为所述生成器损失函数。

7.一种基于文本生成图像的方法，其特征在于，包括：

基于待生成图像的文本信息，生成相应的文本嵌入式表示；

将所述文本嵌入式表示输入至预训练的图像生成模型处理，得到所述文本信息的人造图像；其中，所述图像生成模型采用权利要求1至6所述的任一模型训练方法训练得到。

8.一种基于文本生成图像的模型训练设备，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6中任一项所述的基于文本生成图像的模型训练方法。

9.一种计算机可读存储介质，其特征在于，其中存储有计算机可读指令，该计算机可读指令用于执行如权利要求1至6中任一项所述的基于文本生成图像的模型训练方法。

基于文本生成图像的模型训练方法、设备和图像生成方法

技术领域

[0001] 本发明涉及人工智能技术，特别是涉及一种基于文本生成图像的模型训练方法、设备和图像生成方法。

背景技术

[0002] 图像的创作是一项复杂而重要的工作，它需要专业的绘图与美术知识。因此，面对广泛的需求，机器辅助的图像创作已成为近期的热点，希望机器帮助人用更少的专业知识、更简便快捷的方法创作出所需要的图像。对于无绘画经验者而言，通过语言交互来控制机器绘制图像是最简单自然的方法。这样，就需要机器能够理解并利用人类语言中语义信息，以生成相应的图像。为满足该需求，产生了基于文本生成图像的技术。这类技术需要完成两个基本目标：可信度(fidelity)与一致性(consistency)，可信度是指产生的人造图像要与真实图像相似，即看起来逼真；一致性则是指产生的图像能够反映出文本输入信息。

[0003] 发明人在实现本发明的过程中发现现有基于文本生成图像的方案中存在模型学习效率低、效果差等问题。具体原因分析如下：

[0004] 由于文本信息无法覆盖图像的所有细节信息，在基于文本生成图像的方案中，对于文本信息中没有限定的图像细节特征，需要随机产生。这样，在基于文本生成图像的场景下，图像信息包括两部分，一部分是模态公共部分与模态特定部分。其中，模态公共部分与文本信息相对应，反映了图像的内容，模态特定部分则是随机产生的，处于不受控制的半随机状态，与图像的内容无关，反映了图像的风格。现有方案在训练图像生成模型时，基于包含模态特定部分的图像特征确定损失函数值。而图文一致性仅与图像的模态公共部分有关，与图像的模态特定部分无关，模态特定部分的存在引入了随机噪声，会干扰模型的学习，从而会导致模型学习效率低、效果差。另外，由于模态特定部分反映了图像风格，该部分不是文本限定的，具有随机性，因此，采用现有基于文本生成图像的方案时，仅能通过改变文本条件来改变所生成图像的内容，而无法有效控制图像的风格，从而导致无法有效控制图像的风格，进而降低了实用性。

发明内容

[0005] 有鉴于此，本发明的主要目的在于提供一种基于文本生成图像的模型训练方法、设备和图像生成方法，可以提高模型学习效率和图像生成效果。

[0006] 为了达到上述目的，本发明提出的技术方案为：

[0007] 一种基于文本生成图像的模型训练方法，包括：

[0008] 对于预设训练样本集合中的每个训练样本，基于该训练样本的文本信息，生成相应的文本嵌入式表示，并将所述文本嵌入式表示输入至图像生成模型，触发所述图像生成模型基于所述文本嵌入式表示，生成所述文本信息对应的人造图像，并采用模态解纠缠方式，提取所述人造图像的真实度参数和相应训练样本的真实图像的真实度参数；所述真实度参数包括：图像风格的视觉可信度、图-文相似度和图像的整体视觉可信度；

[0009] 基于所述人造图像，确定每个所述训练样本的正例和负例；

[0010] 利用所述图像生成模型，基于每个所述训练样本的正例、负例和真实图像各自对应的所述真实度参数，计算相应样本的子损失函数，并基于所述子损失函数，计算相应样本的总体损失函数；所述子损失函数包括内容损失函数、风格损失函数、生成器损失函数和判别器损失函数；所述总体损失函数包括判别器总体损失函数和生成器总体损失函数；

[0011] 利用所述总体损失函数，调整所述图像生成模型的参数。

[0012] 基于上述模型训练方法实施例，本发明实施例还公开了一种基于文本生成图像的方法，包括：

[0013] 基于待生成图像的文本信息，生成相应的文本嵌入式表示；

[0014] 将所述文本嵌入式表示输入至预训练的图像生成模型处理，得到所述文本信息的人造图像；其中，所述图像生成模型采用如上所述的基于文本生成图像的模型训练方法训练得到。

[0015] 基于上述模型训练方法实施例，本发明实施例还公开了一种基于文本生成图像的模型训练设备，包括处理器和存储器；

[0016] 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的基于文本生成图像的模型训练方法。

[0017] 基于上述模型训练方法实施例，本发明实施例还公开了一种计算机可读存储介质，其特征在于，其中存储有计算机可读指令，该计算机可读指令用于执行如上所述的基于文本生成图像的模型训练方法。

[0018] 由上述技术方案可见，本发明实施例提出的基于文本生成图像的模型训练方法、设备和图像生成方法，在生成人造图像后，采用模态解纠缠方式，提取人造图像和相应真实图像各自的真实度参数。如此，通过模态解纠缠，可以将模态特定部分从人造图像中抽离，从而使得在提取人造图像的真实度参数时，一方面可以仅基于模态公共部分提取图-文相似度，有效避免了与图像内容无关的模态特定部分对图-文相似度的影响，进而可以提高模型学习效率和图像生成效果，另一方面可以单独基于模态特定部分提取出图像风格的视觉可信度，实现对所生成的图像风格的有效控制，进而增加了实用性。

附图说明

[0019] 图1为本发明实施例的模型训练方法流程示意图；

[0020] 图2为本发明实施例的基于文本生成图像的方法流程示意图。

具体实施方式

[0021] 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。

[0022] 图1为本发明实施例的基于文本生成图像的模型训练方法流程示意图，如图1所示，该实施例实现的模型训练方法主要包括下述步骤：

[0023] 步骤101、对于预设训练样本集合中的每个训练样本，基于该训练样本的文本信息，生成相应的文本嵌入式表示，并将所述文本嵌入式表示输入至图像生成模型，触发所述图像生成模型基于所述文本嵌入式表示，生成所述文本信息对应的人造图像，并采用模态

解纠缠方式，提取所述人造图像的真实度参数和相应训练样本的真实图像的真实度参数。

[0024] 其中，所述真实度参数包括：图像风格的视觉可信度、图-文相似度和图像的整体视觉可信度。

[0025] 本步骤中，在基于训练样本的文本信息生成人造图像后，需要采用模态解纠缠方式，提取该人造图像和相应样本的真实图像各自的上述真实度参数。这里，通过模态解纠缠，可以将模态特定部分从人造图像中抽离，如此，在提取人造图像的真实度参数时，一方面可以仅基于模态公共部分提取图-文相似度，从而避免模态特定部分对图-文相似度的影响，从而可以提高模型学习效率和图像生成效果，另一方面又可以单独基于模态特定部分提取出图像风格的视觉可信度，从而可以有效控制图像风格，增加模型的实用性。

[0026] 对于所述文本嵌入式表示，本领域技术可以采用现有方法基于文本信息得到，例如，可以利用预训练的深层注意力多模态一致性模型提取得到文本信息的文本嵌入式表示，但不限于此。

[0027] 在一种实施方式中，具体可以采用下述方法生成所述文本信息对应的人造图像：

[0028] 步骤a1、将所述文本嵌入式表示，输入至图像生成模型的文本编码器处理，得到所述文本信息的文本特征。

[0029] 具体的，上述文本编码器可以为单层全连接神经网络，但不限于此。

[0030] 本步骤所得到的文本特征h_{tc}，即文本的模态公共部分表征。

[0031] 步骤a2、将所述文本特征和训练样本对应的预设噪声样本，输入至所述图像生成模型的生成器处理，得到所述文本信息对应的人造图像。

[0032] 在一种实施方式中，所述生成器可以由若干个采样与残差层构成，它以步骤a1得到的文本特征h_{tc}与预设的噪声样本z为输入，并产生图像$\hat{x}$，即人造图像。本步骤生成器对应的处理公式如下：

[0033] $h=F(h_{tc},z)$

[0034] $\hat{x}=G(h)$

[0035] 在一种实施方式中，步骤101具体可以采用下述方法提取所述人造图像的真实度参数和相应训练样本的真实图像的真实度参数：

[0036] 步骤b1、利用所述图像生成模型的图像编码器，从所述人造图像中提取出模态公共表征和模态特定表征，以及从所述训练样本的真实图像中提取出模态公共表征和模态特定表征。

[0037] 本步骤中，图像编码器E_I以人造图像$\hat{x}$或者真实图像x为输入，并提取模态解纠缠的图像特征：

[0038] $(h_{sc},h_{ss})=E_I(\hat{x})$

[0039] $(h_{ic},h_{is})=E_I(x)$

[0040] 上式中，h_{sc}与h_{ss}分别代表人造图像的模态公共表征与模态特定表征；h_{ic}与h_{is}分别代表真实图像的模态公共表征与模态特定表征。

[0041] 步骤b2、基于所述人造图像的模态公共表征和模态特定表征、所述真实图像的模态公共表征和模态特定表征，利用所述图像生成模型的判别器，提取所述人造图像和所述真实图像的真实度参数。

[0042] 具体地，与上述三种真实度参数相对应，图像生成模型的判别器将包括三部分，即

模态特定判别器、模态公共判别器和整体视觉可信度。

[0043] 在一种实施方式中，具体可以采用下述方法对上述真实度参数进行提取：

[0044] 按照$s_s(\hat{x}) = D^s(h_{ss})$，提取所述人造图像的图像风格的视觉可信度；其中，$h_{ss}$为所述人造图像的模态特定表征；$D^s$为所述图像生成模型的模态特定判别器；$s_s(\hat{x})$表示所述人造图像的图像风格的视觉可信度。

[0045] 按照$s_s(x=D^s(h_{is})$，提取所述真实图像的图像风格的视觉可信度；其中，h_{is}为所述真实图像的模态特定表征；$s_s(x)$表示所述真实图像的图像风格的视觉可信度。

[0046] 按照$s_c(\hat{x}) = D^c(h_{tc}, h_{sc})$，提取所述人造图像的图-文相似度；其中，$h_{tc}$为所述文本信息的文本特征；$h_{sc}$为所述人造图像的模态公共表征；$D^c$为所述图像生成模型的模态公共判别器；$s_c(\hat{x})$表示所述人造图像的图-文相似度。

[0047] 按照$s_c(x=D^c(h_{tc},h_{ic})$，提取所述真实图像的图-文相似度；其中，h_{ic}为所述真实图像的模态公共表征；$s_c(x)$表示所述真实图像的图-文相似度。

[0048] 按照$s_i(\hat{x}) = D^i(h_{sc}, h_{ss})$，提取所述人造图像的整体视觉可信度；其中，$D^i$为所述图像生成模型的整体视觉判别器；$s_i(\hat{x})$表示所述人造图像的整体视觉可信度。

[0049] 按照$s_i(x) = D^i(h_{ic},h_{is})$，提取所述真实图像的整体视觉可信度；其中，$s_i(x)$表示所述真实图像的整体视觉可信度。

[0050] 在上述方法中，考虑到图像风格的视觉可信度s_s只与图像的特定部分表征有关，因此，仅以图像特定部分表征h_{ss}与h_{is}为输入；图-文相似度s_c仅与图像、文本的模态公共部分表征有关，因此，仅以h_{tc}、h_{sc}与h_{tc}，h_{ic}为输入；整体视觉可信度s_i与模态公共部分表征、特定部分表征均相关，因此，需要同时以二者为输入。

[0051] 步骤102、基于所述人造图像，确定所述训练样本的正例和负例。

[0052] 本步骤中，将基于训练样本集合得到的所有人造图像，确定集合中每个训练样本的正例和负例，以便在后续步骤中基于各样本的正例和负例的真实度参数进一步计算各训练样本对应的损失函数。

[0053] 对于一个训练样本i而言，正例是基于训练样本i生成的人造图像，负例，是基于训练样本集合中除练样本i之外的其他训练样本生成的人造图像。

[0054] 在一种实施方式中，具体可以采用下述方法确定每个所述训练样本的正例和负例：

[0055] 对于每个所述训练样本，将该训练样本对应的所述人造图像作为该训练样本的正例，从基于非该练样本对应的所述人造图像中，选择一个图像作为该训练样本的负例。

[0056] 上述方法中可以采用随机选择的方式，选择负例。为便于操作，也可以采用错位选择的方式选择负例，即对于一个训练样本，将其下一相邻训练样本的人造图像作为该训练样本的负例，但不限于此。

[0057] 步骤103、利用所述图像生成模型，基于每个所述训练样本的正例、负例和真实图像各自对应的所述真实度参数，计算相应样本的子损失函数，并基于所述子损失函数，计算相应样本的总体损失函数。

[0058] 其中，所述子损失函数包括内容损失函数、风格损失函数、生成器损失函数和判别器损失函数；所述总体损失函数包括判别器总体损失函数和生成器总体损失函数。

[0059] 本步骤中，为了提高后续基于损失函数对模型参数调整的准确性，将分别计算内容损失函数(Content Loss)与风格损失函数(Style Loss)，以避免模态特定部分对模型训练的影响，同时实现对图像风格的控制。

[0060] 在一种实施方式中，具体可以采用下述方法计算相应训练样本的各子损失函数：

[0061] 1、按照$\mathcal{L}_C = \mathcal{L}(h_{tc}, h_{sc}, h_{sc}^-)$，计算内容损失函数值$\mathcal{L}_C$。

[0062] 其中，$\mathcal{L}(\ )$表示三元组损失函数，h_{tc}为锚点，表示训练样本的文本信息的文本特征；h_{sc}为正例，表示训练样本对应的人造图像的模态公共表征；h_{sc}^-为负例，表示训练样本的负例的模态公共表征。

[0063] 上述计算内容损失函数值的方法中，内容损失函数采用了常用于建模图-文对齐关系的排序目标函数，该三元组损失函数(Triplet Loss)的具体计算公式如下：

[0064] $$\mathcal{L}(u, v, v^-) = [\alpha - f(u, v) + f(u, v^-)]_+$$

[0065] 其中，$[q]_+ = \max(0, q)$，f是相似度评分函数，u为文本表示，作为锚点(anchor)，v与v^-分别为与文本u相匹配的正例图像表示和与u不匹配的负例图像表示；α为预设的正例图像与文本的相似度与负例图像与文本的相似度的预期差值。f的具体形式为皮尔逊相关系数(Pearson Correlation Coefficient)。

[0066] 上述方法中，内容损失函数以图、文的模态公共部分表征为输入，意图是最大化相匹配的图、文公共部分表示之间的相似度。以文本描述h_{tc}为锚点，作为生成器的输入值；而判别器以两种图像特征为输入，包括从锚点产生的图像特征(h_{sc},h_{ss})和从其他文本产生的图像特征(h_{sc}^-, h_{ss}^-)。为了有效区分正例和负例，提高模型训练效率，这里，以“最大化相匹配的图-文对的相关性，同时最小化非匹配图-文对的相关性”为内容损失函数的目标，因此，将内容损失函数设计为：$\mathcal{L}_C = \mathcal{L}(h_{tc}, h_{sc}, h_{sc}^-)$。

[0067] 2、按照$\mathcal{L}_S = -\rho(z, h_{ss})$，计算所述风格损失函数$\mathcal{L}_S$。

[0068] 其中，z为噪声样本；h_{ss}表示训练样本对应的人造图像的模态特定表征；ρ为预设的皮尔逊相关系数。

[0069] 这里，考虑到生成器的输入分为h_{tc}和z两个部分，所生成的人造图像$\hat{x}$的内容完全由文本特征h_{tc}决定，则$\hat{x}$的风格必然由相应的噪声样本z控制，即噪声z应与图像的特定部分表征h_{ss}一致。基于此，风格损失函数的形式为z和h_{ss}的关联误差：$\mathcal{L}_S = -\rho(z, h_{ss})$。

[0070] 3、按照$\mathcal{L}_G = -\mathbb{E}_{\hat{x}\sim p_G}[\log(s_c(\hat{x}))] - \mathbb{E}_{\hat{x}\sim p_G}[\log(s_s(\hat{x}))] - \mathbb{E}_{\hat{x}\sim p_G}[\log(s_i(\hat{x}))]$，计算所述生成器损失函数$\mathcal{L}_G$；其中，$\mathbb{E}_{\hat{x}\sim p_G}[\ ]$表示从生成器$p_G$中采样得到的训练样本对应的人造图像$\hat{x}$；$s_c(\hat{x})$表示人造图像$\hat{x}$的图-文相似度；$s_s(\hat{x})$表示人造图像$\hat{x}$的图像风格的视觉可信度；$s_i(\hat{x})$表示人造图像$\hat{x}$的整体视觉可信度。

[0071] 4、按照
$$\begin{aligned}\mathcal{L}_D = &-\mathbb{E}_{x\sim p_{\text{data}}}[\log(s_c(x))] - \mathbb{E}_{\hat{x}\sim p_G}[\log(1 - s_c(\hat{x}))] \\ &-\mathbb{E}_{x\sim p_{\text{data}}}[\log(s_s(x))] - \mathbb{E}_{\hat{x}\sim p_G}[\log(1 - s_s(\hat{x}))] \\ &-\mathbb{E}_{x\sim p_{\text{data}}}[\log(s_i(x))] - \mathbb{E}_{\hat{x}\sim p_G}[\log(1 - s_i(\hat{x}))]\end{aligned}$$
，计算所述判别器损失函数$\mathcal{L}_D$，其中，$\mathbb{E}_{x\sim p_{\text{data}}}[\ ]$表示从训练样本集合$p_{data}$中得到的训练样本的真实图像x；$s_s(x)$表示真实图像x的图像风格的视觉可信度；$s_c(x)$表示真实图像x的图-文相似度；$s_i(x)$表示真实图

像x的整体视觉可信度。

[0072] 在计算上述各损失函数时，使用的特征不再是全局的图像特征，而是解纠缠后的图像特征。上述各损失函数经过线性组合，构成了训练时各训练样本对应的下述总体损失函数：

[0073] $\mathcal{L}_{D_{total}} = \mathcal{L}_D + \mathcal{L}_C + \mathcal{L}_S$

[0074] $\mathcal{L}_{G_{total}} = \mathcal{L}_G + \mathcal{L}_C$

[0075] 其中，$\mathcal{L}_{D_{total}}$为判别器总体损失函数，$\mathcal{L}_{G_{total}}$为生成器总体损失函数，$\mathcal{L}_D$为所述判别器损失函数；$\mathcal{L}_C$为所述内容损失函数；$\mathcal{L}_S$为所述风格损失函数；其中，$\mathcal{L}_G$为所述生成器损失函数。

[0076] 步骤104、利用所述总体损失函数，调整所述图像生成模型的参数。

[0077] 本步骤中，将基于各训练样本对应的判别器总体损失函数和生成器总体损失函数，对所述图像生成模型的参数进行调整

[0078] 具体地，在进行上述调整时，将基于判别器总体损失函数对模型中的图像编码器和判别器进行参数调整；基于生成器总体损失函数对模型中的生成器和文本编码器进行参数调整。

[0079] 基于上述步骤101～104，即可实现基于一个训练样本集合中训练样本对图像生成模型的训练。在实际应用中，为了提高模型训练的准确性，可以利用多个训练样本集合循环利用上述步骤101～104进行图像生成模型的训练。

[0080] 从上述模型训练方法实施例可以看出，上述模型训练方法实施例可以在不增加文本生成图像模型的复杂度情况下，通过复用判别器，学习图文模态解纠缠表征，提升了文本生成图像的图像生成质量和图文关联度，增加了对人造图像风格的控制能力。

[0081] 与上述模型训练方法实施例相对应，本发明实施例还提供了一种基于文本生成图像的方法，如图2所示，包括：

[0082] 步骤201、基于待生成图像的文本信息，生成相应的文本嵌入式表示。

[0083] 步骤202、将所述文本嵌入式表示输入至预训练的图像生成模型处理，得到所述文本信息的人造图像。

[0084] 其中，所述图像生成模型采用如上文所述的模型训练方法训练得到。

[0085] 由于上述模型训练方法训练在图像生成模型时，通过模态解纠缠，将模态特定部分从人造图像中抽离，有效避免了与图像内容无关的模态特定部分对图-文相似度的影响，提高了模型生成图像的效果。因此，利用基于上述模型训练方法训练得到的图像生成模型，为当前待生成图像的文本信息生成图像，可以保障图像生成质量。

[0086] 与上述模型训练方法实施例相对应，本发明实施例还提供了一种基于文本生成图像的模型训练设备，包括处理器和存储器；

[0087] 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文所述的基于文本生成图像的模型训练方法。

[0088] 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。

[0089] 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。

[0090] 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。

[0091] 本发明还提供了一种机器可读的存储介质，存储用于使一机器执行如本申请所述方法的指令。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，从而实现上述实施方式中任一实施方式的功能。

[0092] 用于提供程序代码的存储介质实施方式包括软盘、硬盘、磁光盘、光盘(如CD-ROM、CD-R、CD-RW、DVD-ROM、DVD-RAM、DVD-RW、DVD+RW)、磁带、非易失性存储卡和ROM。可选择地，可以由通信网络从服务器计算机或云上下载程序代码。

[0093] 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。

[0094] 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。

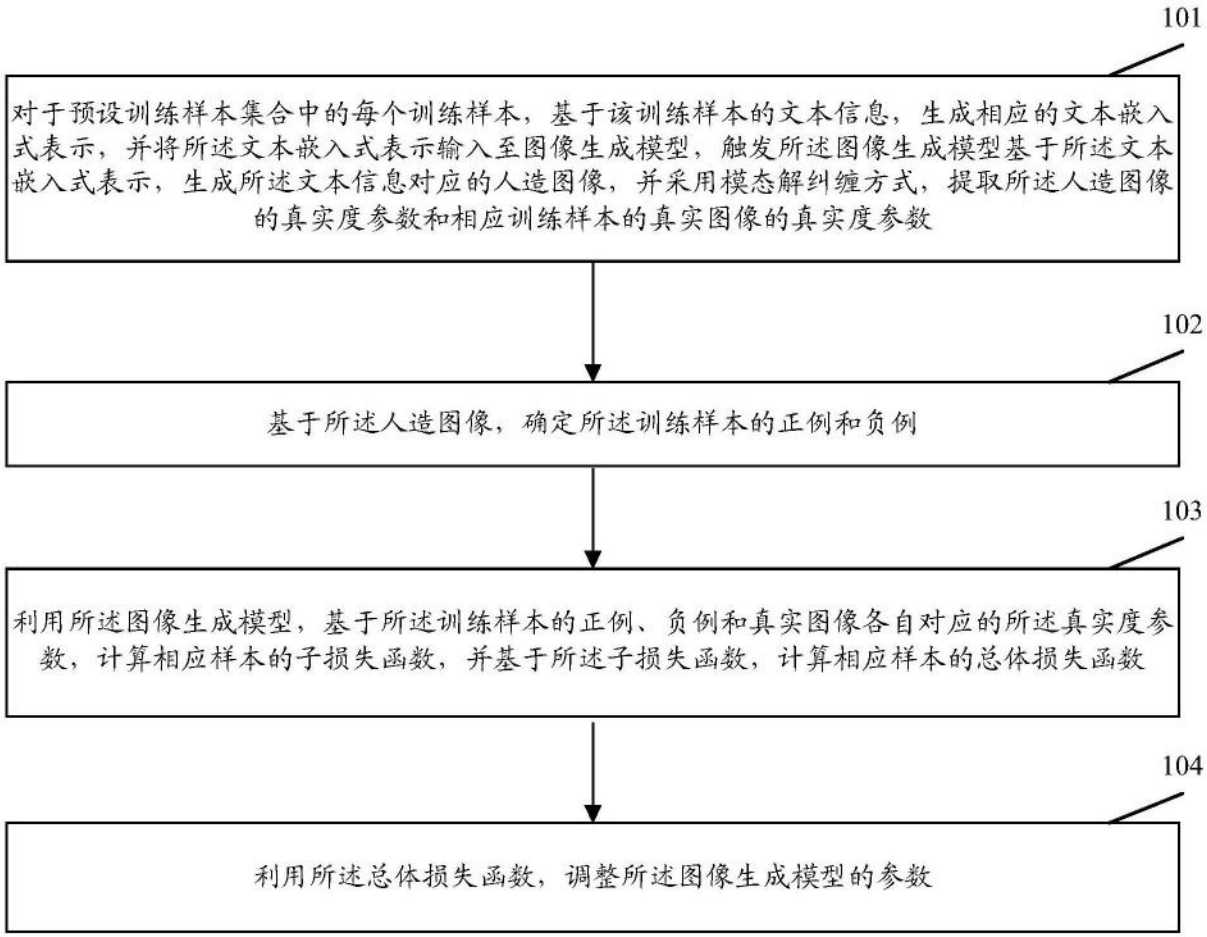
101
对于预设训练样本集合中的每个训练样本，基于该训练样本的文本信息，生成相应的文本嵌入式表示，并将所述文本嵌入式表示输入至图像生成模型，触发所述图像生成模型基于所述文本嵌入式表示，生成所述文本信息对应的人造图像，并采用模态解纠缠方式，提取所述人造图像的真实度参数和相应训练样本的真实图像的真实度参数
102
基于所述人造图像，确定所述训练样本的正例和负例
103
利用所述图像生成模型，基于所述训练样本的正例、负例和真实图像各自对应的所述真实度参数，计算相应样本的子损失函数，并基于所述子损失函数，计算相应样本的总体损失函数
104
利用所述总体损失函数，调整所述图像生成模型的参数

图1

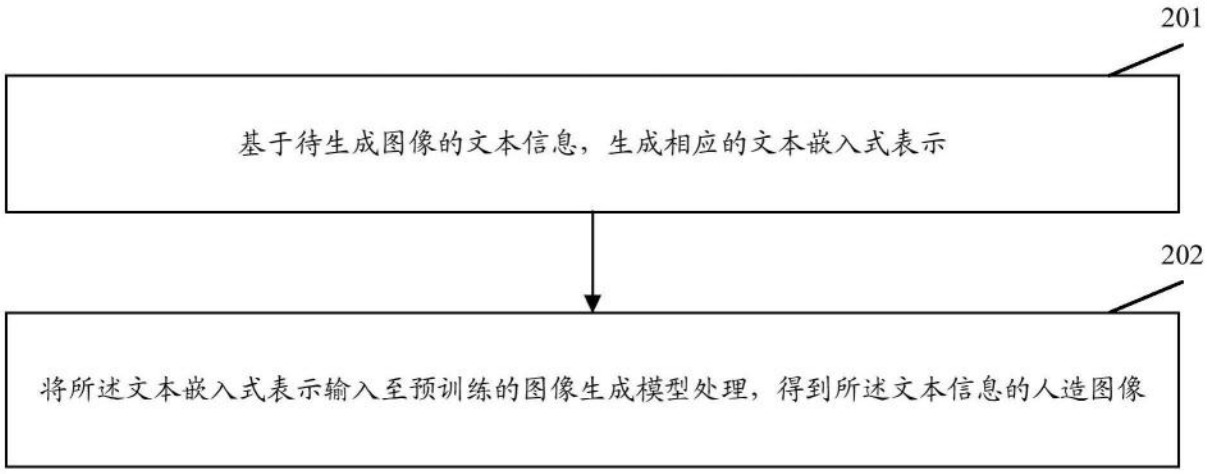
201
基于待生成图像的文本信息，生成相应的文本嵌入式表示
202
将所述文本嵌入式表示输入至预训练的图像生成模型处理，得到所述文本信息的人造图像

图2