深度学习实验六

组员 1 姓名：江经 学号：1190202328 班级：19030103 班

组员 2 姓名：朱健坤 学号：1190201924 班级：19030103 班

1. 选题说明及任务描述

本次深度学习实验六我们的选题是根据文字生成图片，即根据输入的一句话生成一张图片。我们了解到在该领域有一些经典网络结构：

1.条件生成对抗网络CGAN，CGAN是在GAN基础上做的一种改进，通过给原始GAN的生成器Generator和判别器Discriminator添加额外的条件信息（如类别标签或者其它的辅助信息），实现条件生成模型；

2.文本生成图像的开山之作GAN-INT-CLS，该网络在对抗网络中加入了匹配感知鉴别器和流形插值学习。通过在判别器的输入增加真实图像和错误的文本描述，让判别器能够更好地学习文本描述和图片内容的对应关系。同时简单地在训练集文本的嵌入之间进行插值来生成大量额外的文本嵌入，增加了文本的变化，从而让生成器具有更强大的生成能力。

3.AttnGAN，Attentional Generative Ad-versarial Network(AttnGAN)是一种注意力驱动的多阶段的细粒度文本到图像生成器。同时网络还借助一个深层注意多模态相似模型(deep attentional multimodal similarity model)来训练该生成器。首次表明分层注意GAN 能够自动选择单词级别的条件来生成图像的不同部分。

最终，我们选择了发表于CVPR的网络[DF-GAN](https://arxiv.org/pdf/2008.05865.pdf)和[SSA-GAN](https://arxiv.org/pdf/2104.00567v3.pdf)来完成我们的实验。DF-GAN模型抛弃了以往的堆叠结构，只使用一个生成器、一个鉴别器、一个预训练过的文本编码器。SSA-GAN在DF-GAN的模型基础上进行了改进，它把把UPBlocks改成了 SSACN Blocks，将文本编码器可以与图像生成器联合训练，更好地利用文本信息生成图像。

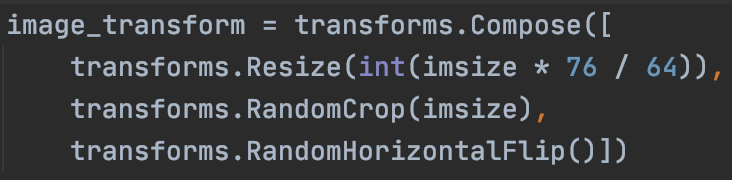
1. 数据集描述
2. CUB-200-2011数据集

altech-UCSD Birds-200-2011 (CUB-200-2011) 是 CUB-200 dataset的一个扩充版本,每个类的图像数量大约增加两倍和新的部位注释。CUB-200-201包含200种鸟类，图片总数目为 11788张，每张图片的标注信息有 15个 Part Locations、312 个Binary Attributes、1 个Bounding Box。下载地址：<http://www.vision.caltech.edu/datasets/cub_200_2011/>

本实验中，使用该数据集的图片作为训练集、测试集，而对于图片的描述（caption）来源于<https://drive.google.com/file/d/1O_LtUP9sch09QH3s_EBAgLEctBQ5JBSJ/view>，每张图片有10个caption，并且重新划分了CUB-200-2011的训练集和测试集，训练集有8855张图片，测试集有2933张图片。

1. 预处理过程

2.1 图片部分



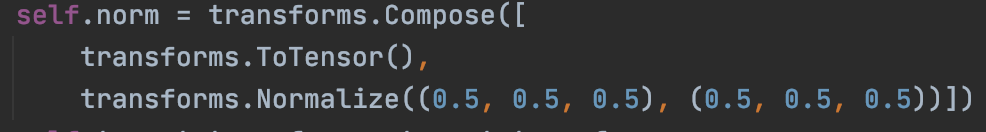


图1 图片处理

图片处理部分采用了随机裁剪、随机水平翻转，并且图片resize为256\*256，将图片进行了归一化。

2.2 文字部分

先利用所有caption构造词典，词典共5450个词，之后读取caption，将caption进行lower()操作，并且将其tokenize，将每个token都根据词典转换为id。

1. 方案设计（包含损失函数和网络结构等）
2. DF-GAN

1.1简介

从文本描述中合成高质量的真实图像是一项具有挑战性的任务。现有的文本到图像生成对抗性网络通常采用堆叠式架构作为主干，但仍然存在三个缺陷。首先，堆叠结构引入了不同图像尺度的生成器之间的纠缠。第二，现有研究倾向于在训练中修复额外的网络，以实现文本-图像语义一致性，但是这限制了这些网络的监控能力。第三，以往研究中广泛采用的基于跨模态注意的文本图像融合由于计算量大而局限于几种特殊的图像尺度。为此，模型提出了一种更简单但更有效的深度融合生成性对抗网络（DF-GAN）。具体来说，DF-GAN提出：（i）一种新的单级文本到图像主干，它直接合成高分辨率图像，而不同生成器之间没有纠缠；（ii）一种由匹配软件梯度惩罚和单向输出组成的新的目标感知鉴别器，它在不引入额外网络的情况下增强了文本图像的语义一致性，（iii）一种新的深文本图像融合块，它深化了融合过程，使文本和视觉特征完全融合。与目前最先进的方法相比，模型提出的DFGAN在合成真实感和文本匹配图像方面更简单但效率更高，并且在广泛使用的数据集上实现了更好的性能。

1.2网络结构

1.2.1整体结构

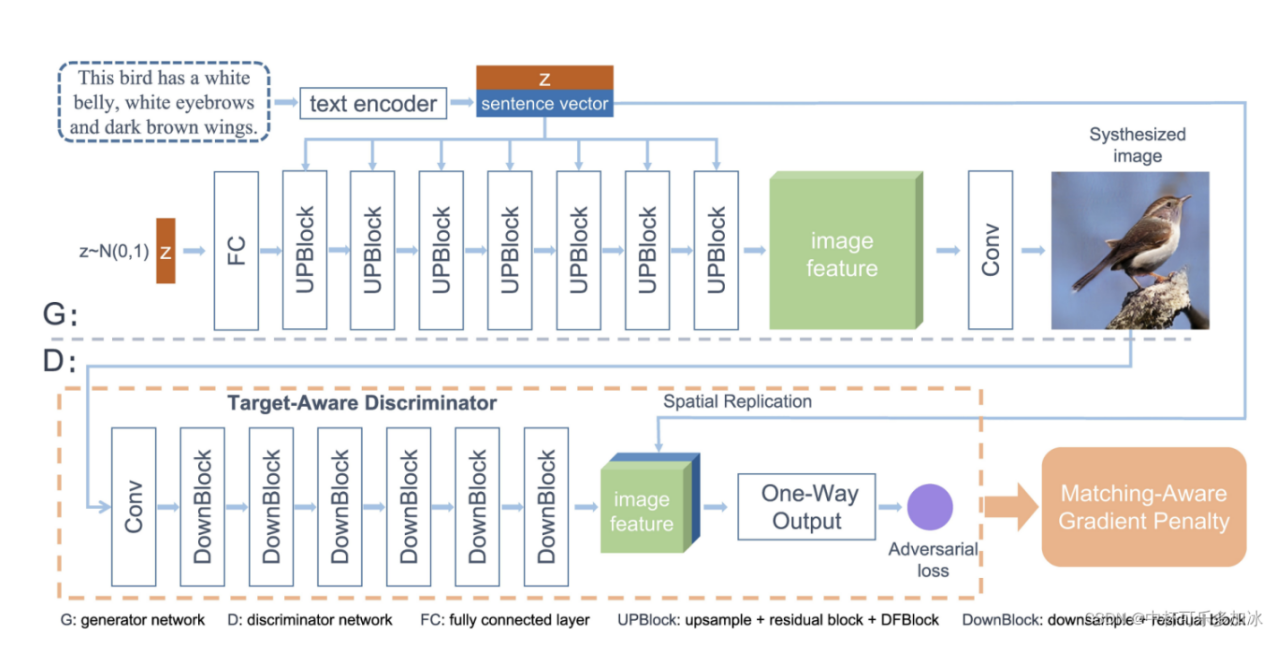


图2-1 DF-GAN框架

1.2.2匹配感知梯度惩罚

作者设计的鉴别器叫Target-Aware Discriminator，由匹配感知梯度惩罚（MA-GP）和单向输出(One-Way Output)组成，主要作用就是促使生成器合成更真实更符合语义一致性的图像。

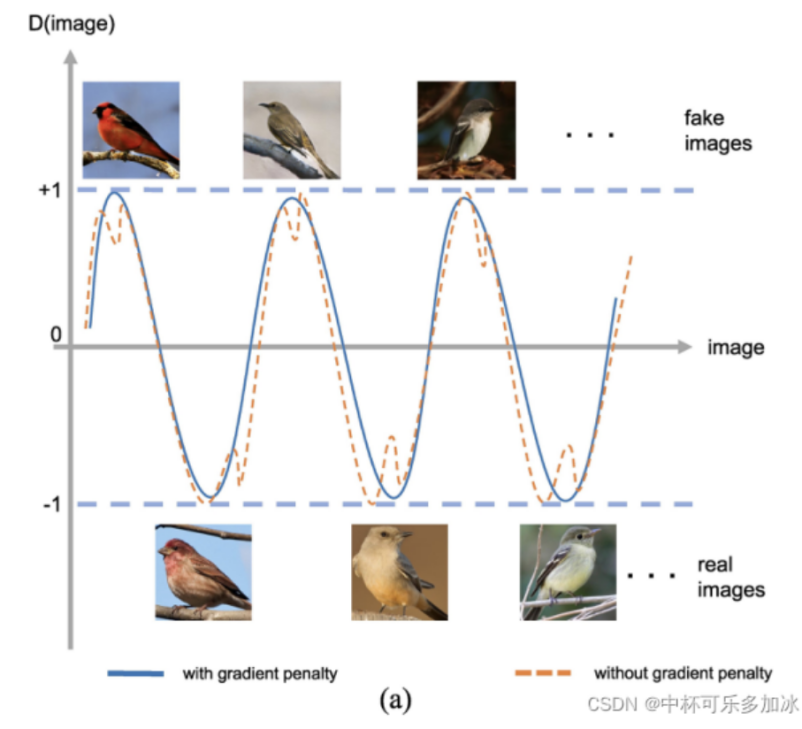
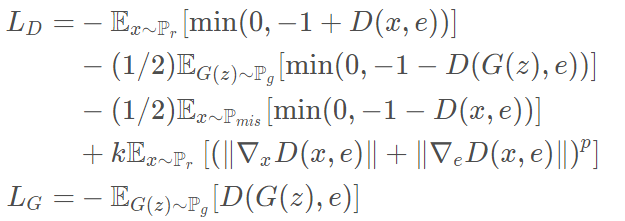


图2-2 梯度惩罚

梯度惩罚（Gradient Penalty）是WGAN-gp曾提出的一种梯度变化方法，如上图所示，首先使用hinge loss将鉴别器损失范围限制在-1和1之间，梯度越大，惩罚越大，即改变梯度的程度越大。损失函数的计算公式如下：



1.2.3单向输出

在以往的T2I任务中，鉴别器以两种方式来进行判断，一是判断图像是真是假，二是将图像特征与句子向量连接起来，判断图像与文本是否语义一致（有条件损失）。这个被作者称Two-Way Output。研究表明，这种Two-Way Output其实减慢了生成器的收敛速度。

条件损失给出的梯度α指向图像与匹配文本的方向，无条件损失的梯度β仅指向真实图像的方向，最终的梯度方向只是简单的求和α+β，并不像预期那样指向（真实，匹配）的方向，这样的过程会减慢图像与文本的一致性。

故作者提出了单向输出（One-Way Output），其鉴别器将图像特征和句子向量连接起来，然后通过两个卷积层仅输出一个对抗性损失。这样设计可以使单个梯度γ直接指向目标数据点（真实和匹配），从而优化和加速生成器的收敛。

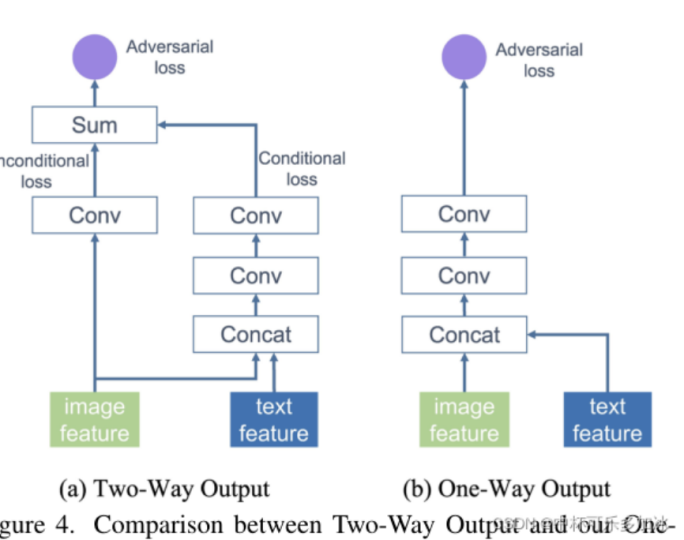


图2-3 单向输出和双向输出

上图表明了Two-Way Output和One-Way Ouput的区别，Two-Way Output首先根据图像特征计算无条件损失，然后将图像特征连接文本特征再计算有条件损失，再将两个损失连接。而One-Way Output将图像特征与文本特征直接连接后经过两个卷积层直接计算总损失。

通过结合MA-GP和单向输出，目标感知鉴别器可以引导生成器合成更多真实和文本匹配的图像。

1.2.4生成器

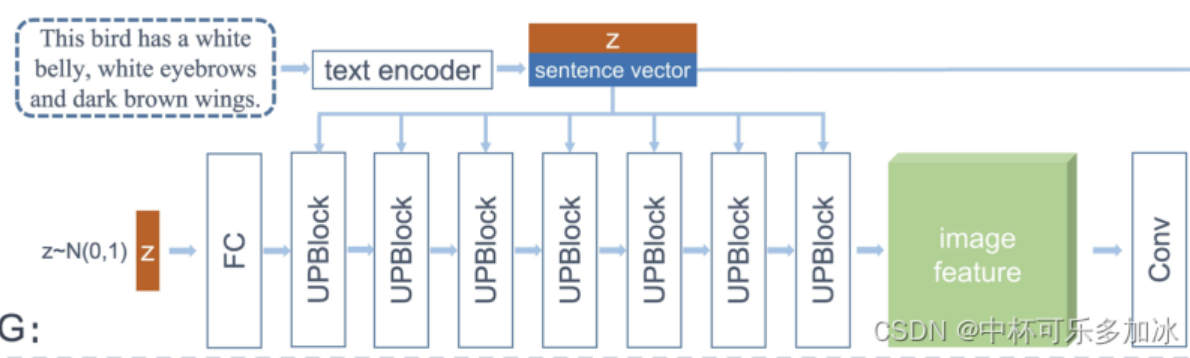


图2-4 生成器

生成器由七个UPBlocks组成，UPBlocks包括上采样、残差块和DFBlock，DFBlock是作者提出的一种深度文本图像融合块，其在融合块中叠加了多个仿射变换和ReLU层。

1. SSA-GAN

2.1 简介

Semantic-Spatial Aware GAN（SSA-GAN）提出了一种新的语义空间感知GAN框架，文章发表于2021年10月。文本到图像生成（T2I）模型旨在生成语义上与文本描述一致的照片逼真图像。在生成性对抗网络（GAN）最新进展的基础上，现有的T2I模型取得了巨大进展。然而，仔细检查它们生成的图像会发现两个主要局限性：（1）条件批量归一化方法平等适用于整个图像特征映射，忽略了局部语义；（2） 文本编码器在训练过程中是固定的，它应该与图像生成器一起训练，以学习更好的文本表示，从而生成图像。为了解决这些局限性，SSA-GAN提出了一种新的语义空间感知GAN框架，该框架以端到端的方式进行训练，以便文本编码器能够利用更好的文本信息。具体来说，SSA-GAN介绍了一种新的语义空间感知卷积网络，该网络（1）学习以文本为条件的语义自适应仿射变换，以有效地融合文本特征和图像特征；（2）以弱监督的方式学习掩码映射，该方法依赖于当前的文本-图像融合过程，以在空间上指导变换。在具有挑战性的COCO和CUB bird数据集上进行的实验表明，SSA-GAN的方法在视觉保真度和与输入文本描述的一致性方面优于最近的最新方法。本实验在CUB bird上对SSA-GAN进行了复现。

2.2 网络结构

2.2.1 整体结构

SSA-GAN的框架如下：



图3-1 SSA-GAN框架

整体来看，SSA-GAN与DF-GAN十分相似，也是单级主干结构。但SSA-GAN将DF-GAN的UPBlocks改成了 SSACN Blocks。SSA-GAN包括一个文本编码器（Text encoder），一个生成器（Generator），一个鉴别器（Discriminator）。

生成器部分：首先输入网络一个随机整体噪声，经过FC层和一次Reshape后，连接七个SSACN层；而同时，输入caption，经过Text encoder编码获得caption的特征，将其送入SSACN中；之后text和noise的融合向量在经过Conv层和tanh激活之后，生成图片；

鉴别器部分：生成的图片随后送入输入鉴别器进行鉴别，图片经过Conv和多个DownBlock加上多个Conv后进行判别。需要注意的是，在SSA-GAN中，文本编码器不固定参数，其也是生成器的一部分。

2.2.2 Text encoder

Text encoder由Bi-LSTM组成，通过最小化深度注意多模态相似模型（DAMSM）损失，使用真实图像-文本对进行预训练得到。其中，深度注意多模态相似度模型（DAMSM）用于在单词级和句子级测量图像-文本相似度，以计算图像生成的细粒度损失。其将给定的文本描述编码为具有256维的句子特征向量（取自LSTM的最后隐藏状态）和256维的单词特征（取自LSTM每个时间步的隐藏状态，最多18个单词）。

2.2.3 Semantic-Spatial Aware Convolutional Network（SSACN block）

每个SSACN块由一个 upsample block（上采样块）、一个mask predictor（掩码预测器）、一个Semantic-Spatial Condition Batch Normalization (SSCBN，语义空间条件批量归一化和一个residual block（残差块）组成。如图3-2。

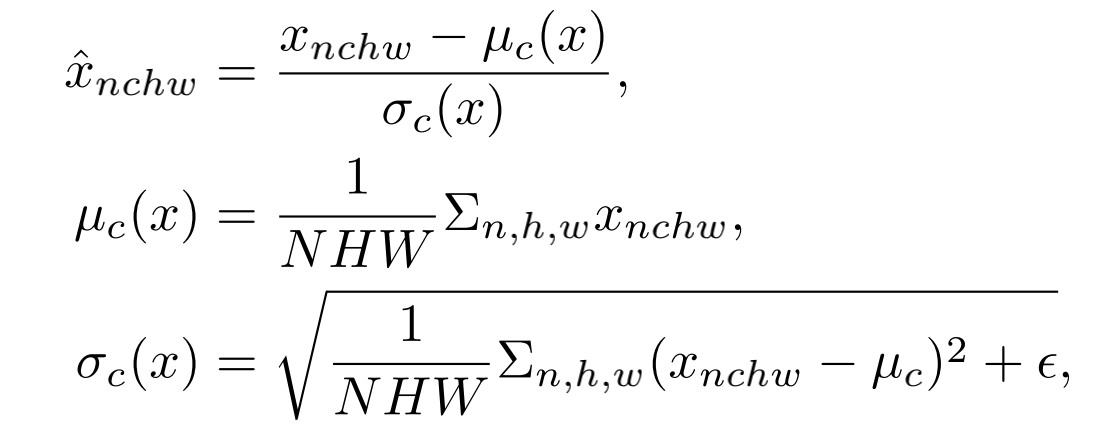


图3-2 SSACN block结构

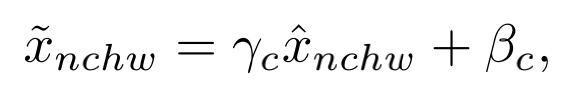
(1) upsample block。其通过双线性插值将图像特征图的宽度和高度加倍。

(2) mask predictor。其将上采样的图像特征图作为输入，预测其mask图。mask图中，每个像素的值在0到1之间，每个像素的值决定在该位置上进行仿射变换的程度。mask图基于当前生成的图像特征图，因此，它直观地指示了当前图像特征图的哪些部分仍需要根据文本信息进行增强，以便于改进后的图像特征图与给定文本语义更一致。mask predictor与整个网络一起训练，但mask predictor没有特定的损失函数来指导其学习过程，也没有额外的mask annotation，唯一的监督来自鉴别器给出的损失，因此，这是一个弱监督的过程。

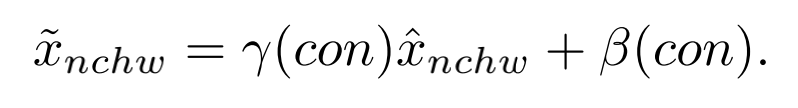
(3) Semantic-Spatial Condition Batch Normalization (SSCBN）。SSCBN根据文本特征向量学习语义感知仿射参数。根据当前文本图像融合过程（即最后一个SSCBN块的输出），预测mask map。这种仿射变换有效而深入地融合了文本和图像特征，并使得图像特征与文本语义一致；原始BN首先将mini-batch的数据标准化为每个特征通道均是零平均值和标准方差：



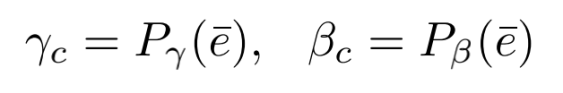
而CBN对通道进行了仿射变换，对每个channel都有两个参数开控制，而这两个参数都是学习得到：



也即是：

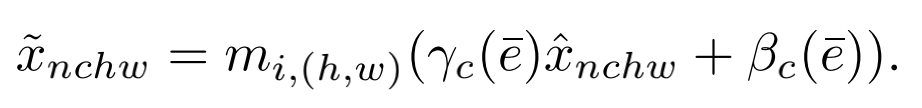


在此基础上，为了能够更好地融合图像与文本的特征，这两个参数都是从文本向量得到，如：



其中、是两个MLP，用于通过文本向量学习这两个参数。

如果不添加更多的空间信息，则上一步的CBN将在空间上均匀地进行图像特征映射，而SSCBN希望只调整特征图中与文本有关的部分。SSCBN的实现如下：



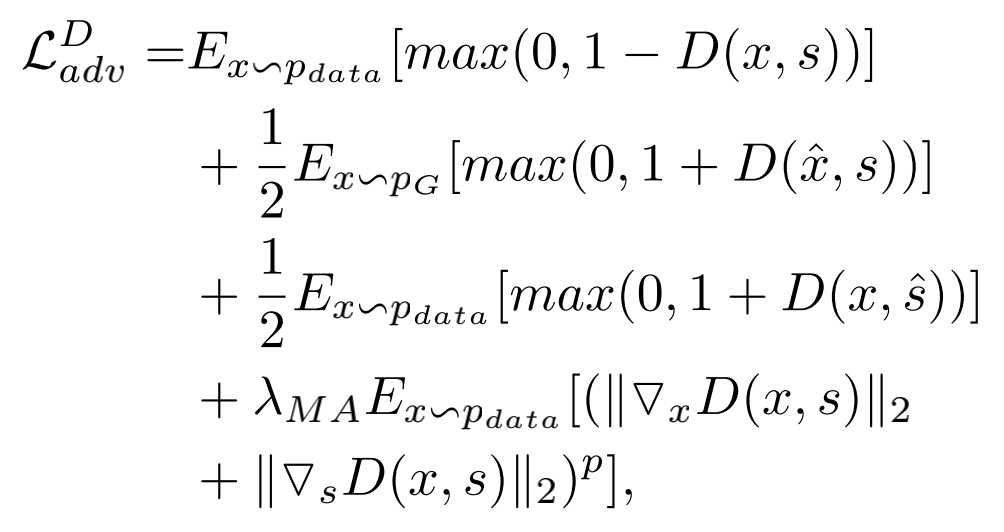
其中为mask predictor预测的掩码，用掩码来表示各个像素需要变换的权重。且在SSCBN是以文本向量（句子级）为条件的，与词级相比，文本向量（句子级）所需的计算量要少得多。

1. residual block。其用于保持图像特征的主要内容，以防止与文本无关的部分发生变化，并防止图像信息被文本信息覆盖掉。

2.3 损失函数

2.3.1 鉴别器损失函数

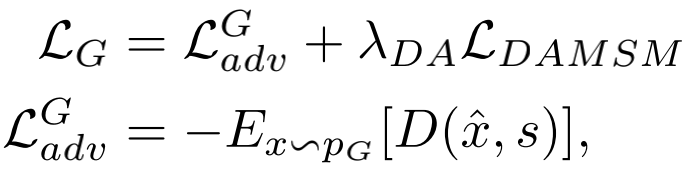
MA-GP loss（Matching-Aware zero-centered Gradient Penalty）



其中s是给定的文本描述，是不匹配的文本描述，x是s对应的真实图像，是生成的图像。、都是MA-GP的超参数。

2.3.2 生成器损失

生成器损失由两部分组成，分别是对抗损失和DAMSM损失（单词级细粒度图像文本匹配）



1. 实验过程

1.实验环境

操作系统：Linux CentOS 3.10.0-1062.el7.x86\_64

torch版本：torch 1.11.0

cuda版本：11.3

显卡：Nvidia RTX A100 \* 2

IDE：PyCharm 2020.2.5 (Professional Edition)

1. 实验过程

2.1 DF-GAN

由于计算资源有限，我们并没有从头开始训练DF-GAN，而是接着他提供的已经训练了600个epoch的预训练模型继续训练了50个epoch。由于tensorboard的文件有1.8G，如有需要可前往[<https://pan.baidu.com/s/1xThAbE1VnuzNytpp-ioM0Q> 提取码: f1eo]进行下载。

2.2 SSA-GAN

由于时间以及计算资源有限，SSA-GAN的训练使用了论文作者提供的训练了550个epoch的预训练模型。我们在此基础上继续训练了50个epoch。由于log文件过大，如有需要请前往[<https://pan.baidu.com/s/1xThAbE1VnuzNytpp-ioM0Q> 提取码: f1eo]进行下载。

1. 实验结果与分析
2. 评价指标

1.1 IS(Inception Score)

IS使用预训练的inception v3网络来计算类条件分布（生成图像）和类边缘分布（真实图像）之间的KL散度。较大的IS表示生成的图像质量较高，且每个图像的类别易被分类器判别。其公式如下：



其中x为生成的图像，y为其类别标签。

1.2 FID(Fréchet Inception Distance)

FID计算生成图像和真实图像的特征分布之间的Fréchet距离，其中的特征提取自预训练的inception v3网络。较低的FID表示生成的图像更真实。

1. 性能指标评价与分析

本实验模型在Pytorch中实现，参考了论文作者的代码实现。

DF-GAN的batch\_size为24，使用1块Nvidia A100 GPU（40GB显存）。训练中使用了β1=0.0和β2=0.9的Adam优化器。生成器和鉴别器的学习速率分别设置为0.0001和0.0004。DF-GAN模型在CUB数据集上训练了650个epoch，选取最佳的模型（600epoch）进行评测。

SSA-GAN的batch\_size为24，使用2块Nvidia A100 GPU（40GB显存）。训练中使用了β1=0.0和β2=0.9的Adam优化器。生成器和鉴别器的学习速率分别设置为0.0001和0.0004。超参数p、λMA和λDA分别设置为6、2、0.1。SSA-GAN模型在CUB数据集上训练了600个epoch，选取最佳的模型（590epoch）进行评测。

为了评估IS，本实验的模型均从测试数据集中随机选择文本描述，生成分辨率为256×256的30k张图片。为了评估FID，本实验的模型均从测试数据集中随机选择文本描述，生成分辨率为256×256的3k张图片（为了能和测试集中的约3k图像数量保持一致）。实验结果如下：

表 DF-GAN与SSA-GAN实验结果

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | /private/var/folders/3m/4l4psmwx06196z0q2gn3fgd80000gn/T/com.kingsoft.wpsoffice.mac/wpsoffice.Nhyfspwpsoffice | /private/var/folders/3m/4l4psmwx06196z0q2gn3fgd80000gn/T/com.kingsoft.wpsoffice.mac/wpsoffice.yQafvTwpsoffice | 训练时间 |
| DF-GAN | /private/var/folders/3m/4l4psmwx06196z0q2gn3fgd80000gn/T/com.kingsoft.wpsoffice.mac/wpsoffice.lrZPrTwpsoffice | 20.96 | **较短** |
| SSA-GAN | /private/var/folders/3m/4l4psmwx06196z0q2gn3fgd80000gn/T/com.kingsoft.wpsoffice.mac/wpsoffice.xSbrZywpsoffice | /private/var/folders/3m/4l4psmwx06196z0q2gn3fgd80000gn/T/com.kingsoft.wpsoffice.mac/wpsoffice.MeXweGwpsoffice | 较长 |

观察可知SSA-GAN的IS指标与FID指标均优于DF-GAN，但缺点是SSA-GAN需要训练的模型更多，模型参数量更大，训练时间更长。

1. 生成图像主观评价与分析

生成结果如下：

3.1 A small bird with an orange bill and grey crown and breast.

（一种有橙色喙、灰色冠和胸部的小鸟。）



GT DF-GAN SSA-GAN

观察可知，DF-GAN与SSA-GAN都描述到了“orange bill and grey crown and breast”；但是DF-GAN的喙带有白色和黑色、不是较为纯净的橙色。且DF-GAN与SSA-GAN都没有特别地体现出“small bird”。

3.2 The bird has a bright red eye, a gray bill and a white neck.

（这只鸟有一只鲜红色的眼睛，灰色的喙和白色的脖子。）



GT DF-GAN SSA-GAN

观察可知，DF-GAN与SSA-GAN都描述到了“a bright red eye, a gray bill and a white neck”；但是DF-GAN的喙带有一些白色，且其眼睛的鲜红色也不明显。

3.3 This bird has a long pointed beak with a wide wingspan.

（这种鸟有一个长而尖的喙，翼展很宽。）



GT DF-GAN SSA-GAN

观察可知，DF-GAN与SSA-GAN都描述到了“long pointed beak with a wide wingspan”；但是DF-GAN的头部出现了重叠，而SSA-GAN的头部没有出现在图片中。

3.4 A small bird with a black bill and a fuzzy white crown nape throat and breast.

（一种有黑色喙和绒毛状白色冠的颈部和胸部的小鸟）



GT DF-GAN SSA-GAN

观察可知，DF-GAN与SSA-GAN都描述到了“A small bird”、“a fuzzy white crown nape throat and breast”；但DF-GAN没有描述到“a black bill”，而SSA-GAN描述到了。

1. 方案评价
2. DF-GAN

该论文创新如下：

提出了一种新的单级文本到图像主干，可以直接合成高分辨率图像，而不需要不同生成器之间的纠缠。

提出了一种由匹配感知梯度惩罚（MA-GP）和单向输出组成的目标感知鉴别器。它在不引入额外网络的情况下显著增强了文本图像的语义一致性。

提出了一种新的深度文本图像融合块（DFBlock），它能更有效、更深入地融合文本和视觉特征。

但是DF-GAN也存在一些不足：

模型抛弃了AttnGAN以来提出的单词级信息，只引入了句子级的文本信息，这限制了细粒度视觉特征合成的能力

模型使用的text encoder仍然是AttnGAN中的encoder。若引入预先训练过的大型语言模型来提供额外的知识可能会进一步提高性能。

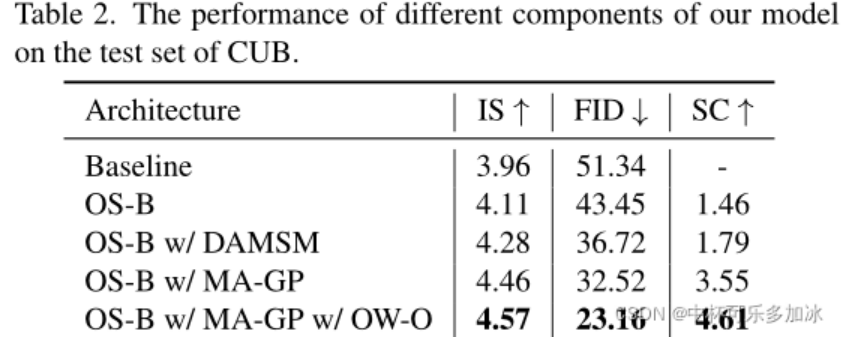


图4 DF-GAN论文中提供的消融实验结果

1. SSA-GAN

该论文提出了一种新的用于T2I生成的语义空间感知GAN（SSA-GAN）框架，主要是在生成器上做的工作，创新如下：

提出语义空间感知卷积网络（SSACN）模块，通过基于当前生成的图像特征图预测掩码映射草图，这种掩码图不仅可以指导生成器后续在何处添加文本信息，还起到了权重作用即决定要在某个部分上加强文本信息的程度。

提出新的仿射参数计算方法，将掩码图添加到SSCBN中作为权重，然后从编码的文本向量中学习仿射参数，再进行批量归一化。

而提供其提供的消融实验结果：

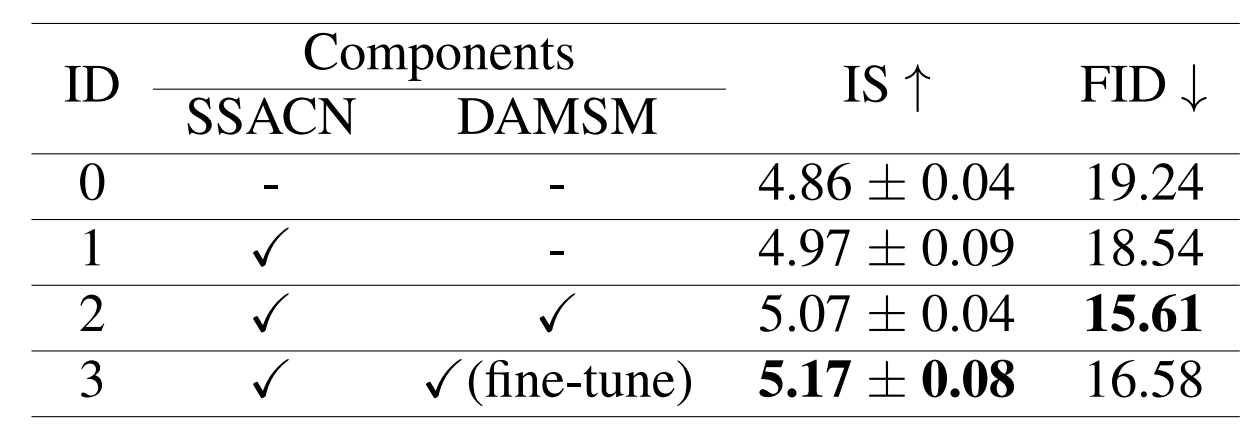


图5 SSA-GAN论文中提供的消融实验结果

ID3是指将文本编码器也加入训练进行微调，可以看到虽然IS有提高，但是FID却有所降低，分析的原因可能是微调文本编码器有助于文本图像融合，提高文本图像的一致性，从而提高IS分数，但文本与图像一致的同时导致图像多样性下降，所以FID会变差。且实验过程中发现SSA-GAN训练较慢，训练时长明显多余DF-GAN，同等batchsize下占用显存也明显比DF-GAN大，训练SSA-GAN对计算设备要求较高。

1. 成员分工

1.江经

完成SSA-GAN的训练，对IS、FID指标进行评估，撰写实验报告

2.朱健坤

完成DF-GAN的训练，对IS、FID指标进行评估，撰写实验报告

参考文献

[1]H. Lee, U. Ullah, J. -S. Lee, B. Jeong and H. -C. Choi, "A Brief Survey of text driven image generation and maniulation," 2021 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), 2021, pp. 1-4, doi: 10.1109/ICCE-Asia53811.2021.9641929.

[2]Hu K, Liao W, Yang M Y, et al. Text to image generation with semantic-spatial aware gan[J]. arXiv preprint arXiv:2104.00567, 2021.

[3]Tao M, Tang H, Wu S, et al. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis[J]. arXiv preprint arXiv:2008.05865, 2020.

附录

1.DF-GAN、SSA-GAN最佳模型、tensorboard训练的log在百度网盘：

链接: <https://pan.baidu.com/s/1xThAbE1VnuzNytpp-ioM0Q> 提取码: f1eo

2.IS、FID评价引导请见与本报告同目录下的SSA-GAN/README.md（IS评价在SSA-GAN/IS.py；FID评价在TTUR/fid.py）