机器学习实验(2023 春)课程论文

Text-To-image——生成口袋妖怪

姓名 陈云霞　　学号 20201120715

摘　要

本文主要描述如何使用text to image简单生成口袋妖怪。生成口袋妖怪的文本到图像的问题可以使用深度学习中的生成对抗网络（GAN）来解决。

主要实现步骤如下：首先，需要准备一个口袋妖怪的图像。使用Python中的GAN库，如TensorFlow或PyTorch，来实现GAN模型。GAN模型包括生成器和判别器两个部分，它们一起训练以生成逼真的图像。定义生成器模型，将文本输入转换为图像输出。生成器可以是一个神经网络模型，接受妖怪的文本描述作为输入，然后输出一个图像。可以使用卷积神经网络（CNN）或者转置卷积神经网络（transposed CNN）来实现生成器模型。定义判别器模型，将图像输入作为输入，输出一个二进制值，代表输入的图像是否真实。判别器也可以是一个神经网络模型，使用CNN或者全连接神经网络（FCN）实现。将生成器和判别器组合起来形成GAN模型。在训练期间，生成器和判别器将相互竞争，以使生成器生成的图像越来越逼真。

训练GAN模型，使用数据集中的文本和图像对进行训练。在每个训练周期中，生成器将生成一批图像，判别器将评估这些图像的真实性，然后更新它的权重。生成器也将更新权重以生成更逼真的图像。一旦训练完成，你可以使用生成器模型来生成任何口袋妖怪的图像。只需要输入口袋妖怪的文本描述，生成器将生成一个对应的图像。

目录

1 引言

近来年，AI绘画很火，只需在随便网上一搜，就能找到许多相关的文章和资讯。不管是前段时间的AI画作拿奖事件，还是最近爆火的Stable Diffusion模型背后公司Stability AI宣布获得1.01亿美元融资消息，这些都不难看出——人工智能AI创作，目前已经成为了AI领域最热门的话题。

AI绘画工具的表现确实让人耳目一新，而其本质其实是一种生成符合给定文本描述的真实图像（text-to-image）的崭新交互方式。

文本到图像模型（Text-to-image model）是一种机器学习模型，它将自然语言描述作为输入并生成与该描述匹配的图像。由于深度神经网络的进步，此类模型在 2010 年代中期开始开发。2022 年，最先进的文本到图像模型的输出，例如 OpenAI 的DALL-E 2、Google Brain 的Imagen和StabilityAI 的Stable Diffusion开始接近真实照片和手绘艺术的质量。

文本到图像模型通常结合了语言模型和生成图像模型，语言模型将输入文本转换为潜在表示，生成图像模型以该表示为条件生成图像。最有效的模型通常是根据从网络上抓取的大量图像和文本数据进行训练的。

2 text to image简述

2.1 多模态机器学习

我们对世界的体验是多模态的——我们看到物体，听到声音，感觉到纹理，闻到气味，尝到味道。模态是指某件事情发生或经历的方式，一个研究问题如果包含多个模态，就被称为多模态。为了让人工智能在理解我们周围的世界方面取得进展，它需要能够一起解释这种多模态信号。多模式机器学习旨在建立能够处理和关联来自多种模式的信息的模型。这是一个日益重要和具有非凡潜力的充满活力的多学科领域。

生成符合给定文本描述的真实图像（text-to-image）是多模态任务之一，具有巨大的应用潜力，如图像编辑、视频游戏和计算机辅助设计。最近，由于生成对抗网络(GANs)在生成真实感图像方面的成功，文本到图像的生成取得了显著进展。文本到图像的生成创作需要对被创造的事物有深刻的理解:厨师、小说家和电影制作人必须比食客、读者或电影观众更深刻地理解食物、写作和电影。如果我们的计算机视觉系统要真正理解视觉世界，它们不仅必须能够识别图像，而且必须能够生成图像。除了传授深刻的视觉理解，生成逼真图像的方法也可以是实际有用的。在短期内，自动图像生成可以帮助艺术家或平面设计师的工作。有一天，我们可能会用生成定制图像和视频的算法来取代图像和视频搜索引擎，以响应每个用户的个人喜好。

文本生成图像（text-to-image）相关工作相较于图像描述（image captioning），图像所包含的信息更为复杂，因此生成图像任务的提出晚于图像描述。自从GAN网络被提出，神经网络产生的图像接近真实图像，为解决Text-to-image问题找到了解决思路。

2.2 text-to-image的首次提出

ScottReed S在2016年首次提出了能根据文字生成图片的GAN [10]。论文介绍了如何通过 GAN 进行从文字到图像的转化。比方说，若神经网络的输入是“粉色花瓣的花”，输出就会是一个包含了这些要素的图像。该任务包含两个部分：

利用自然语言处理来理解输入中的描述。

生成网络输出一个准确、自然的图像，对文字进行表达。

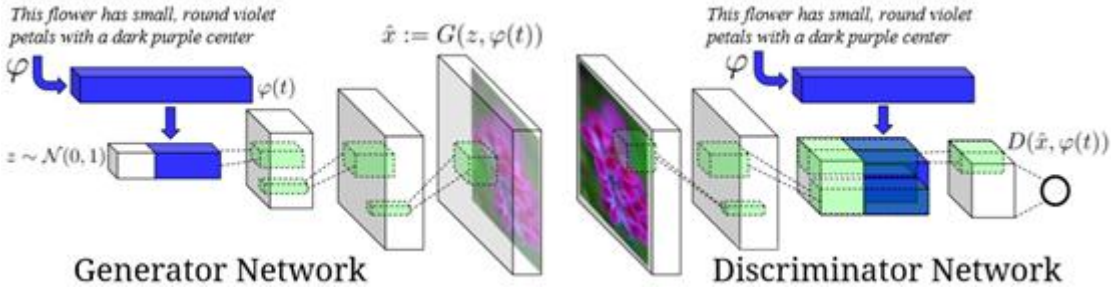


图 text-conditional convolutional GAN architecture.[1]

2.3 StackGAN（Zhang H, et al, ICCV 2017）

2017年，Shaoting Zhang等人[11]提出了基于文本描述的堆叠生成式对抗网络(StackGAN)来生成256×256的真实感图像，通过一个粗略的细化过程将难题分解为更容易处理的子问题。

第一阶段GAN根据给定的文本描述绘制对象的原始形状和颜色，生成阶段i的低分辨率图像。第二阶段GAN将第一阶段的结果和文本描述作为输入，生成具有照片般逼真细节的高分辨率图像。它能够纠正第一阶段结果中的缺陷，并通过细化过程添加引人注目的细节。为了提高合成图像的多样性，稳定conditional-GAN的训练，作者引入了一种新的条件增强技术，使潜在条件集平滑。

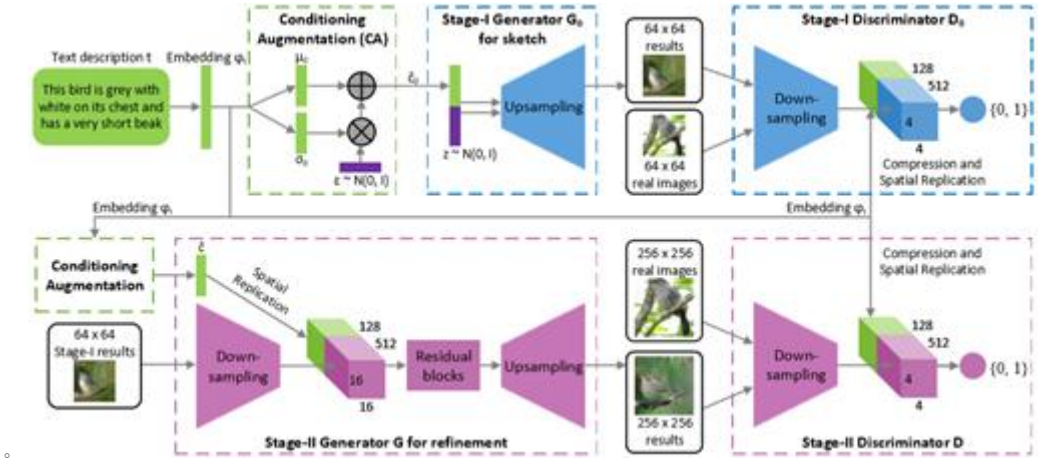


图 The architecture of the proposed StackGAN.[2]

2.4 Stackgan++(Zhang H, el at, TPAMI 2018)

虽然生成式对抗网络(GANs)已经在各种任务中显示出显著的成功，但在生成高质量的图像方面仍然面临挑战。在这篇论文中，Zhang H 等人[12]对之前的StackGAN进行改进，提出了堆叠生成对抗网络(StackGANs)，旨在生成高分辨率的真实感照片。首先，我们提出了一个两阶段生成式对抗网络架构，StackGAN-v1，用于文本到图像的合成。

第一阶段 GAN 根据给定的文本描述描绘场景的原始形状和颜色，生成低分辨率的图像。

第二阶段 GAN 将第一阶段的结果和文本描述作为输入，生成具有照片般逼真细节的高分辨率图像。其次，针对有条件和无条件生成任务，提出了一种先进的多阶段生成式对抗网络体系结构StackGAN-v2。StackGAN-v2由多个生成器和多个鉴别器组成，它们排列成树状结构;同一场景对应的多个尺度的图像来自于树的不同分支。通过联合逼近多个分布，StackGAN-v2比StackGAN-v1表现出更稳定的训练行为。

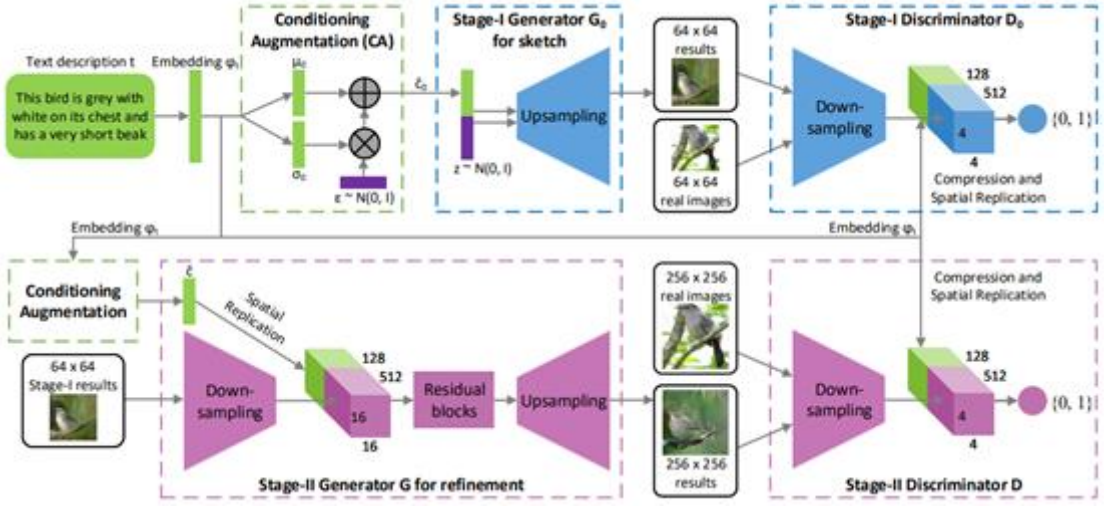


图 The architecture of the proposed StackGAN++.[3]

2.5 Attngan(Xu T, el at, CVPR 2018)

在这篇论文中，Xu T等人[13]提出了一个注意力对抗生成网络(AttnGAN)，它允许注意力驱动的、多阶段的细化来生成细粒度的文本到图像。该算法利用一种新颖的注意力生成网络，通过关注自然语言描述中的相关词汇，在图像的不同亚区合成精细的细节信息。此外，提出了一种基于深度注意的多模态相似度模型来计算用于训练生成器的细粒度图像-文本匹配损失。提出的AttnGAN大大优于先前的技术水平，在CUB数据集上的最佳初始记录提高了14.14%，在更具挑战性的COCO数据集上的最佳初始记录提高了170.25%。详细的分析也进行了可视化的注意层的AttnGAN。这首次表明，分层注意GAN能够自动选择字级条件来生成图像的不同部分。

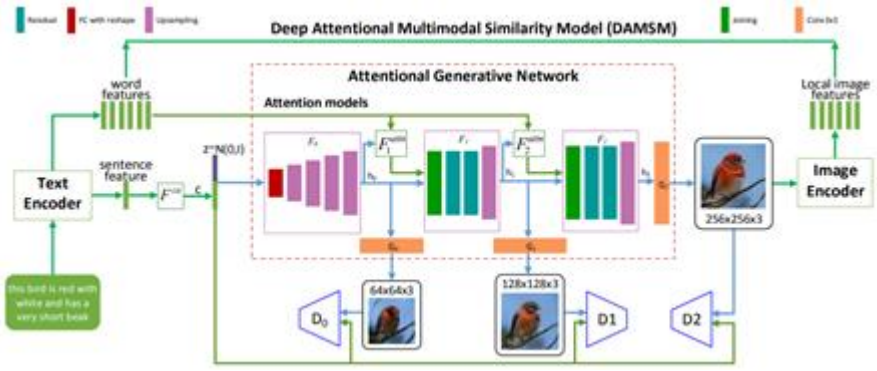


图 The architecture of the proposed AttnGAN.[4]

2.6 Image generation from scene graphs (Johnson J, el at, CVPR 2018)

最近在从自然语言描述生成图像方面取得了令人兴奋的进展，这些方法在有限的领域(如对鸟或花的描述)给出了惊人的结果，但很难用许多对象和关系忠实地再现复杂的句子。为了克服这一限制，李飞飞研究团队中的Johnson J等人[14]提出了一种从场景图生成图像的方法，能够显式地推理对象及其关系。我们的模型使用图形卷积来处理输入图形，通过预测物体的边界框和分割掩码来计算场景布局，并将布局转换为具有级联细化网络的图像。该网络是针对一对鉴别器进行反向训练，以确保实际输出。我们的方法能够生成具有多个对象的复杂图像。

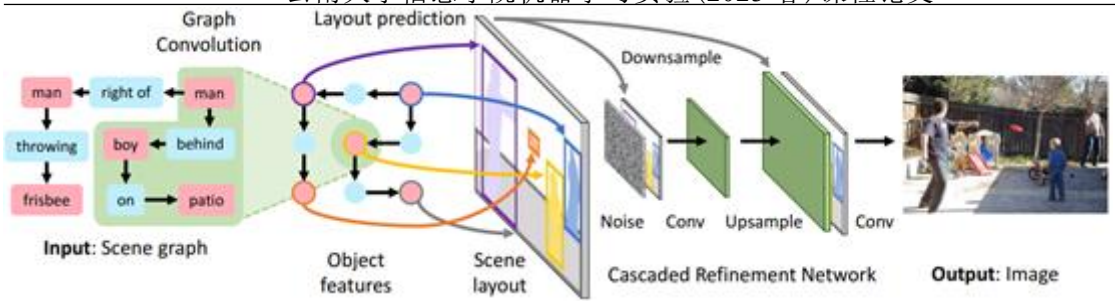
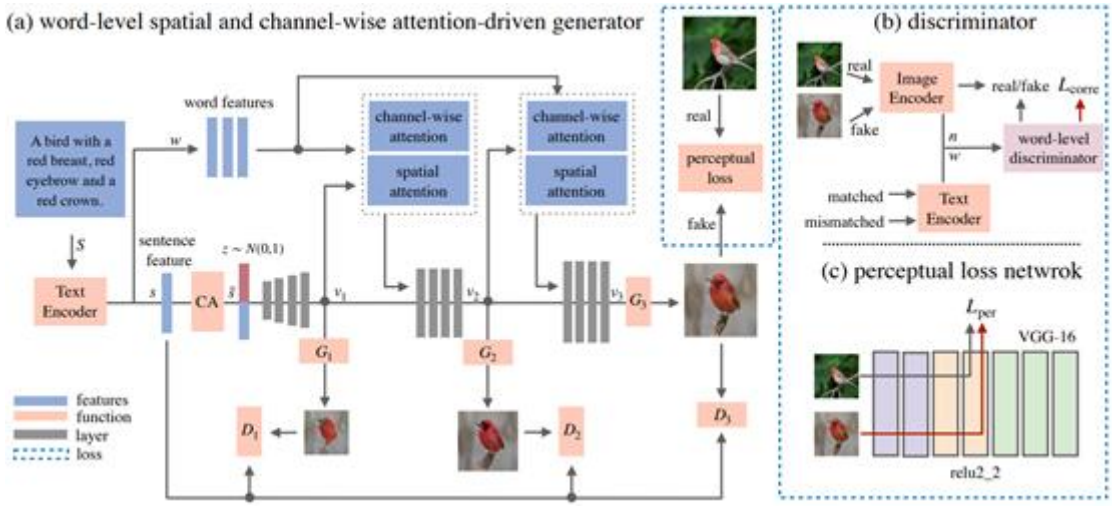


图 Overview of image generation network f for generating images from scene graphs.[5]

2.7 Controllable text-to-image generation (Li B, el al, NeuralIPS 2019)

Li B 等人[16]提出了一种可控的文本-图像生成对抗网络(ControlGAN)，该网络既能有效地合成高质量的图像，又能根据自然语言描述控制图像生成的各个部分。为了实现这一目标，作者引入了一个词级空间和信道级注意力驱动的生成器，它可以分离不同的视觉属性，并允许模型专注于生成和操作与最相关的词对应的子区域。同时，提出了一种词级鉴别器，通过将字与图像区域相关联来提供细粒度的监督反馈，便于训练一种有效的生成器，该生成器能够在不影响其他内容生成的情况下操作特定的视觉属性。

此外，感知损失被用来减少图像生成的随机性，并鼓励生成器操作修改后文本中需要的特定属性。在基准数据集上的大量实验表明，该方法优于现有的技术水平，并且能够使用自然语言描述有效地操作合成图像。



图The architecture of ControlGAN.[6]

3 生成口袋妖怪

3.1 数据集

要生成口袋妖怪的图像，需要一个包含口袋妖怪图像和相关信息的数据集。以下是需要的数据集：

1. 口袋妖怪图像数据集：这个数据集应该包含许多不同口袋妖怪的图像，最好是高质量的真实图像。你可以从各种口袋妖怪游戏、电影、电视节目、漫画或其他相关媒体中收集这些图像。也可以在网上搜索其他人已经制作好的数据集。

2. 口袋妖怪信息数据集：这个数据集应该包含口袋妖怪的属性、类型、特点等信息。你可以从口袋妖怪游戏、百科全书、论坛或其他相关来源中收集这些信息。

3. 图像标注数据集：如果你想训练一个能够生成特定类型口袋妖怪图像的模型，你需要一些带有标注的数据集。例如，如果你想训练一个能够生成火属性口袋妖怪的模型，你需要一个包含火属性口袋妖怪图像和标注的数据集。

4.

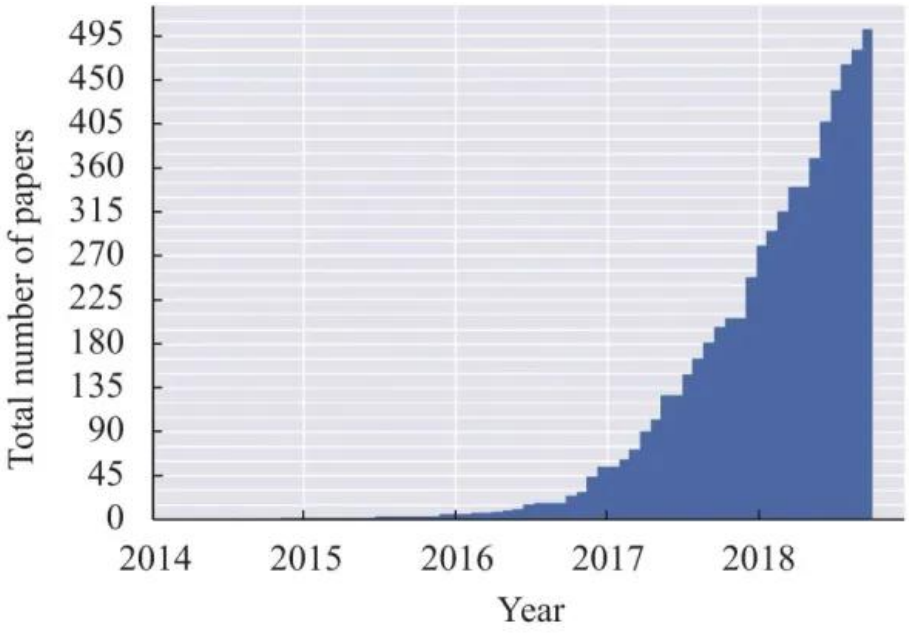
3.2 GAN模型

使用Python中的GAN库，如TensorFlow或PyTorch，来实现GAN模型。GAN模型包括生成器和判别器两个部分，它们一起训练以生成逼真的图像。定义生成器模型，将文本输入转换为图像输出。生成器可以是一个神经网络模型，接受妖怪的文本描述作为输入，然后输出一个图像。可以使用卷积神经网络（CNN）或者转置卷积神经网络（transposed CNN）来实现生成器模型。定义判别器模型，将图像输入作为输入，输出一个二进制值，代表输入的图像是否真实。判别器也可以是一个神经网络模型，使用CNN或者全连接神经网络（FCN）实现。将生成器和判别器组合起来形成GAN模型。在训练期间，生成器和判别器将相互竞争，以使生成器生成的图像越来越逼真。

训练GAN模型，使用数据集中的文本和图像对进行训练。在每个训练周期中，生成器将生成一批图像，判别器将评估这些图像的真实性，然后更新它的权重。生成器也将更新权重以生成更逼真的图像。一旦训练完成，你可以使用生成器模型来生成任何口袋妖怪的图像。只需要输入口袋妖怪的文本描述，生成器将生成一个对应的图像。

3.2.1 GAN模型简介

GAN的全称是Generative Adversarial Networks，即生成对抗网络，由Ian J. Goodfellow等人于2014年10月发表在NIPS大会上的论文《Generative Adversarial Nets》中提出。此后各种花式变体Pix2Pix、CYCLEGAN、STARGAN、StyleGAN等层出不穷，在“换脸”、“换衣”、“换天地”等应用场景下生成的图像、视频以假乱真，好不热闹。前段时间PaddleGAN实现的First Order Motion表情迁移模型，能用一张照片生成一段唱歌视频。各种搞笑鬼畜视频火遍全网。用的就是一种GAN模型哦。深度学习三巨神之一的LeCun也对GAN大加赞赏，称“adversarial training is the coolest thing since sliced bread”。关于GAN网络的研究也呈井喷态势，下面是2014年到2018年命名为GAN的论文数量图表：



GAN的前世今生

1. 判别模型与生成模型

对抗生成模型GAN首先是一个生成模型，和大家比较熟悉的、用于分类的判别模型不同。

判别模型的数学表示是y=f(x),也可以表示为条件概率分布p(y|x)。当输入一张训练集图片x时，判别模型输出分类标签y。模型学习的是输入图片x与输出的类别标签的映射关系。即学习的目的是在输入图片x的条件下，尽量增大模型输出分类标签y的概率。

而生成模型的数学表示是概率分布p(x)。没有约束条件的生成模型是无监督模型，将给定的简单先验分布π(z)（通常是高斯分布），映射为训练集图片的像素概率分布p(x)，即输出一张服从p(x)分布的具有训练集特征的图片。模型学习的是先验分布π(z)与训练集像素概率分布p(x)的映射关系。

2. 其他生成网络简介

生成网络并非只有GAN，介绍下其他几种：

自回归模型（Autoregressive model）是从回归分析中的线性回归发展而来，只是不用x预测y，而是用x预测 x（自己），所以叫做自回归。多用于序列数据生成如文本、语音。PixelRNN/CNN则使用这种方法生成图片，效果还不错。但是由于是按照像素点去生成图像导致计算成本高， 在可并行性上受限，在处理大型数据如大型图像或视频是具有一定麻烦的。

变分自编码器（VAE）：VAE是在AE（Autoencoder自编码器）的基础上让图像编码的潜在向量服从高斯分布从而实现图像的生成，优化了数据对数似然的下界，VAE在图像生成上是可并行的，但是VAE存在着生成图像模糊的问题。

基于流的模型（Flow-based Model）包括Glow、RealNVP、NICE等。流模型思想很直观：寻找一种变换 y = f(x)（f 可逆，且 y 与 x 的维度相同）将数据空间映射到另一个空间，新空间各个维度相互独立。

3.2.2 GAN的组成

1. 解读GAN的loss函数

GAN网络的训练优化目标就是如下公式：

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

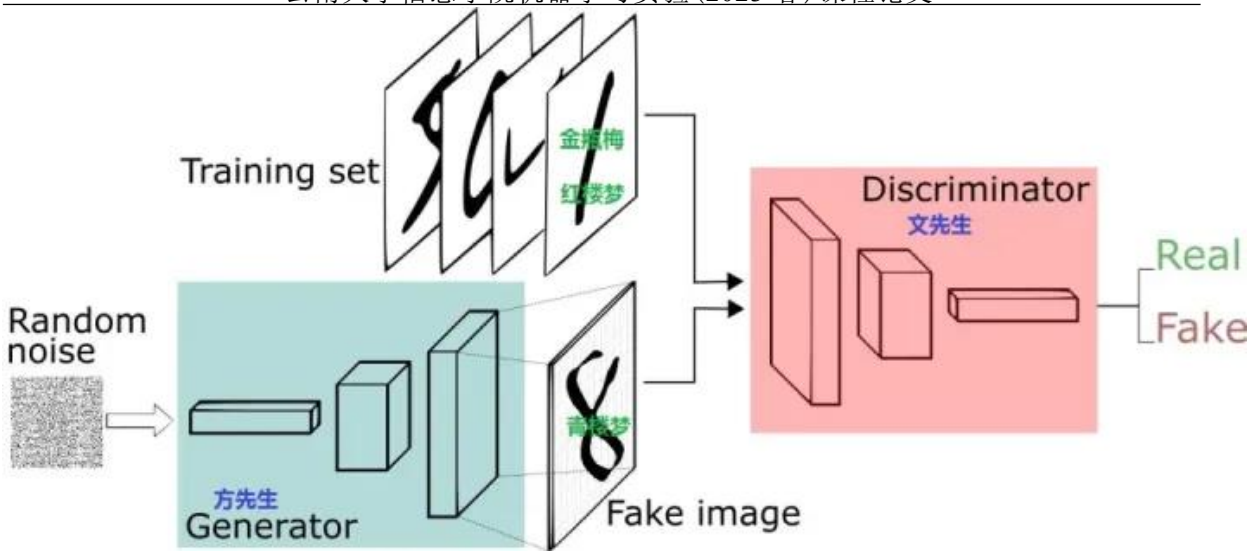
公式出自Goodfellow在2014年发表的论文Generative Adversarial Nets。这里简单介绍下公式的含义和如何应用到代码中。上式中等号左边的部分：V(D,G)表示的是生成样本和真实样本的差异度，可以使用二分类（真、假两个类别）的交叉熵损失。

max表示在生成器固定的情况下，通过最大化交叉熵损失V(D,G)来更新判别器D的参数。

minmax表示生成器要在判别器最大化真、假图片交叉熵损失V(D,G)的情况下，最小化这个交叉熵损失。

等式的右边其实就是将等式左边的交叉熵损失公式展开，并写成概率分布的期望形式。详细的推导请参见原论文《Generative Adversarial Nets》。

2. 解读GAN的结构与训练流程



如上图所示GAN由一个判别器（Discriminator）和一个生成器（Generator）两个网络组成。

训练时先训练判别器：将训练集数据（Training Set）打上真标签（1）和生成器（Generator）生成的假图片（Fake image）打上假标签（0）一同组成batch送入判别器（Discriminator），对判别器进行训练。计算loss时使判别器对真数据（Training Set）输入的判别趋近于真（1），对生成器（Generator）生成的假图片（Fake image）的判别趋近于假（0）。此过程中只更新判别器（Discriminator）的参数，不更新生成器（Generator）的参数。

然后再训练生成器：将高斯分布的噪声z（Random noise）送入生成器（Generator），然后将生成器（Generator）生成的假图片（Fake image）打上真标签（1）送入判别器（Discriminator）。计算loss时使判别器对生成器（Generator）生成的假图片（Fake image）的判别趋近于真（1）。此过程中只更新生成器（Generator）的参数，不更新判别器（Discriminator）的参数。

3.3 实验结果

3.4 引用与参考文献

[1] Reed S, Akata Z, Yan X, et al. Generative adversarial text to image synthesis[J]. arXiv preprint arXiv:1605.05396, 2016.

[2] Zhang H, Xu T, Li H, et al. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 5907-5915.

[3] Zhang H, Xu T, Li H, et al. Stackgan++: Realistic image synthesis with stacked generative adversarial networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(8): 1947-1962.

[4] Xu T, Zhang P, Huang Q, et al. Attngan: Fine-grained text to image generation with attentional generative adversarial networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1316-1324.

[5] Johnson J, Gupta A, Fei-Fei L. Image generation from scene graphs[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1219-1228.

[6] Li B, Qi X, Lukasiewicz T, et al. Controllable text-to-image generation[C]//Advances in Neural Information Processing Systems. 2019: 2063-2073.