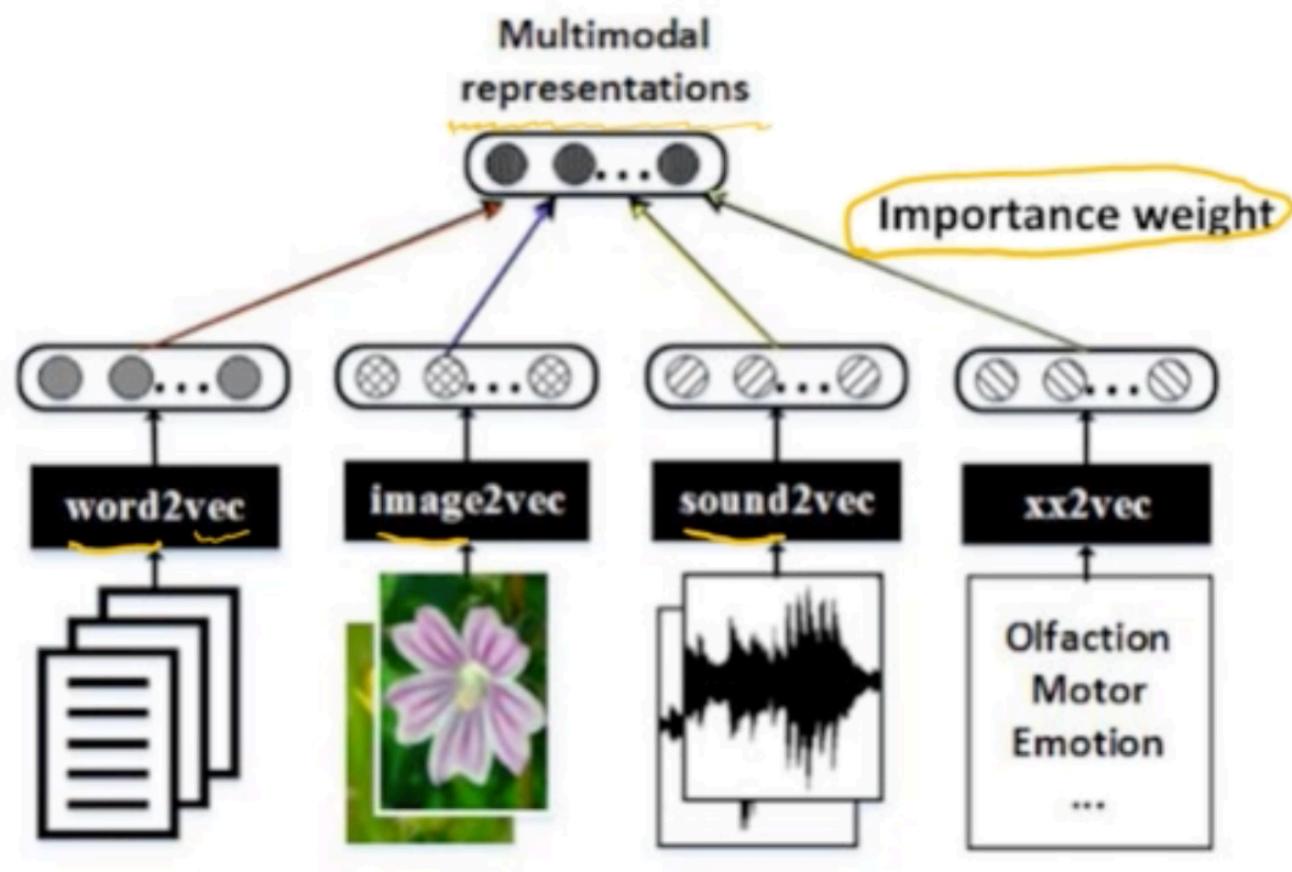


# cGAN

Write by h1astro

## 多模态学习

- 每一种信息的来源或者形式，都可以称为一种模态
- 多模态ML，指在通过ML方法实现处理和理解多源模态信息的能力
- 目前较热门研究方向为图像、视频、音频、语义之间的多模态学习



## 图像标记

- 用词语对图像中不同内容进行多维度表述

## 图像描述

- 把一幅图片翻译为一段描述文字
- 获取图像的标记词语
- 理解图像标记之间的关系
- 生成人类可读的句子



A female tennis player in action on the court.



A group of young men playing a game of soccer



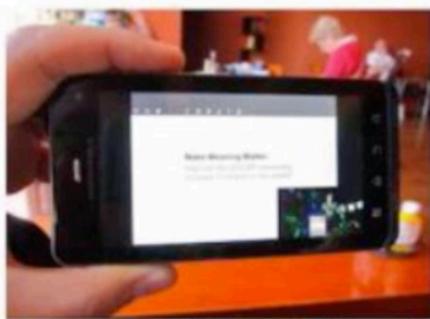
A man riding a wave on top of a surfboard.



A baseball game in progress with the batter up to plate.



A brown bear standing on top of a lush green field.



A person holding a cell phone in their hand.



A close up of a person brushing his teeth.



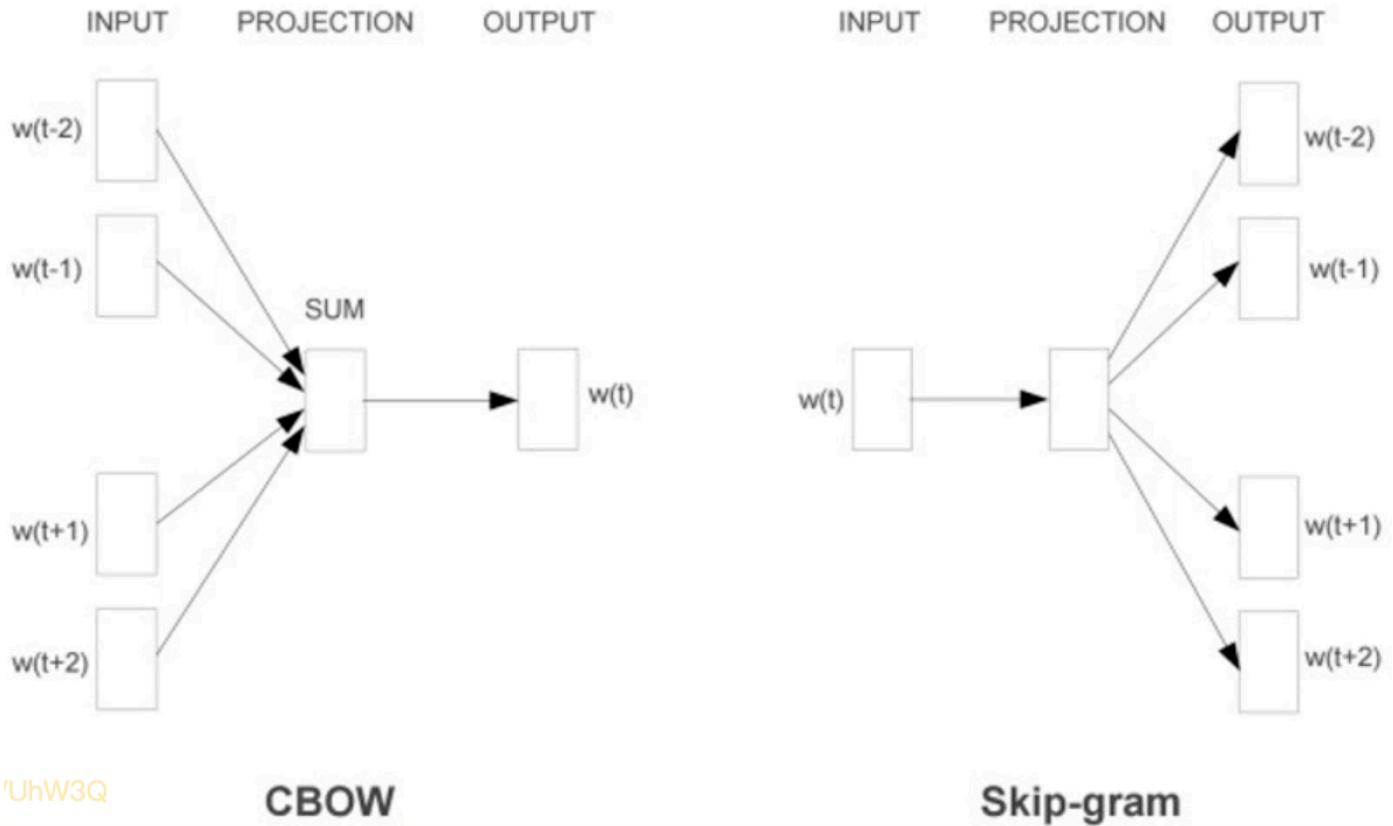
A woman laying on a bed in a bedroom.



A black and white cat is sitting on a chair.

## 词向量模型

- word2vec是从大量文本语料中以无监督的方式学习语义知识的一种模型
- 通过词的上下文得到词的向量化表示，并使得语义上相似的单词在向量空间内距离也很近
- 来源于2013年论文《Efficient Estimation of word representation in vector space》，两种方法：CBOW通过附近词预测中心词、skip-gram 通过中心词预测附近词



## 数据集

MNIST

MIRFLICKR-25000

- 源自雅虎Flickr网站的影像数据库，25000张图像，拥有多个描述tag
- <http://press.liacs.nl/mirflickr/mirdownload.html>

YFCC 100M

- 源自雅虎Flickr网站的影像数据库，由1亿条产生于2004年至2014年间的多条媒体数据组成，包含了9920万的照片数据以及80万条视频数据，数据包括相应tag

## 成果

## 条件图像生成



Figure 2: Generated MNIST digits, each row conditioned on one label

## 图像标记生成

User tags + annotations	Generated tags
montanha, trem, inverno, frio, people, male, plant life, tree, structures, transport, car	taxi, passenger, line, transportation, railway station, passengers, railways, signals, rail, rails
food, raspberry, delicious, homemade	chicken, fattening, cooked, peanut, cream, cookie, house made, bread, biscuit, bakes
water, river	creek, lake, along, near, river, rocky, treeline, valley, woods, waters
people, portrait, female, baby, indoor	love, people, posing, girl, young, strangers, pretty, women, happy, life

Table 2: Samples of generated tags

条件图像生成效果感觉还没GAN好，图像标记有些也不太好

## CGAN历史意义

- 提出了一个可用的条件GAN网络结构
- 开启了GAN在多模态学习中的应用

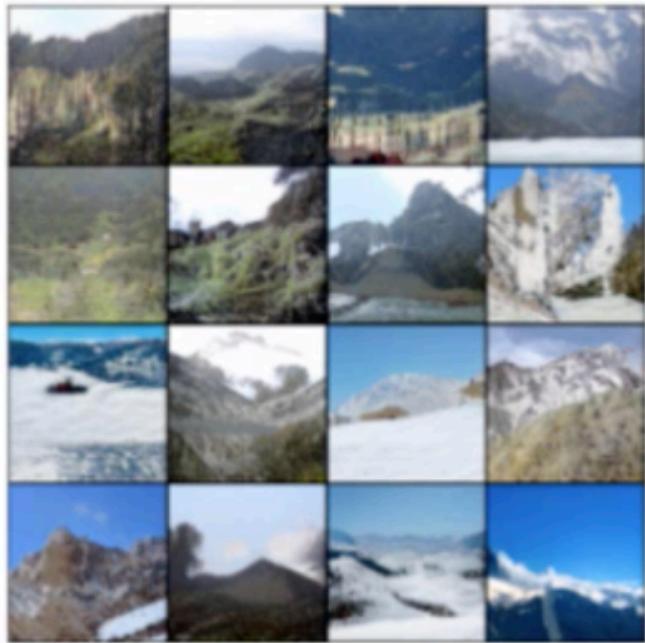
## Attentive Normalization for Conditional Image Generation

Yi Wang<sup>1\*</sup> Ying-Cong Chen<sup>1</sup> Xiangyu Zhang<sup>2</sup> Jian Sun<sup>2</sup> Jiaya Jia<sup>1</sup>

<sup>1</sup>The Chinese University of Hong Kong <sup>2</sup>MEGVII Technology

{yiwang, ycchen, leo{jia}@cse.cuhk.edu.hk {zhangxiangyu, sunjian}@megvii.com

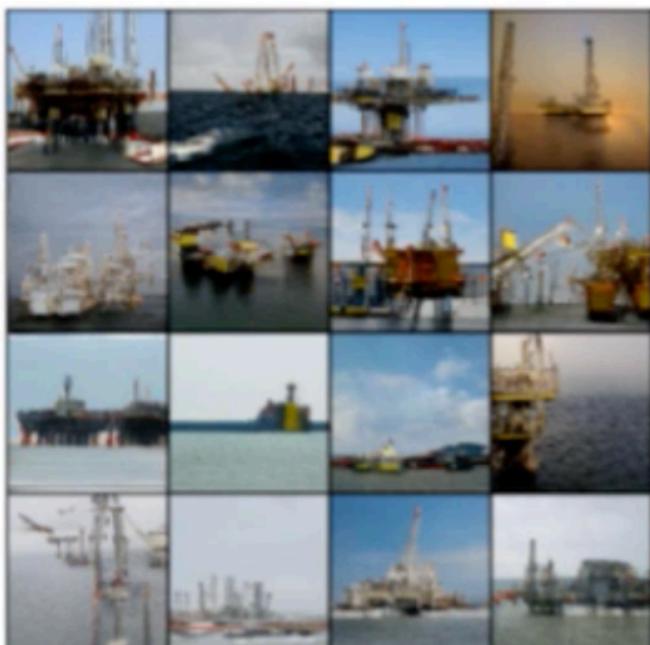
CVPR2020 oral



Alp (970)



Agaric (992)



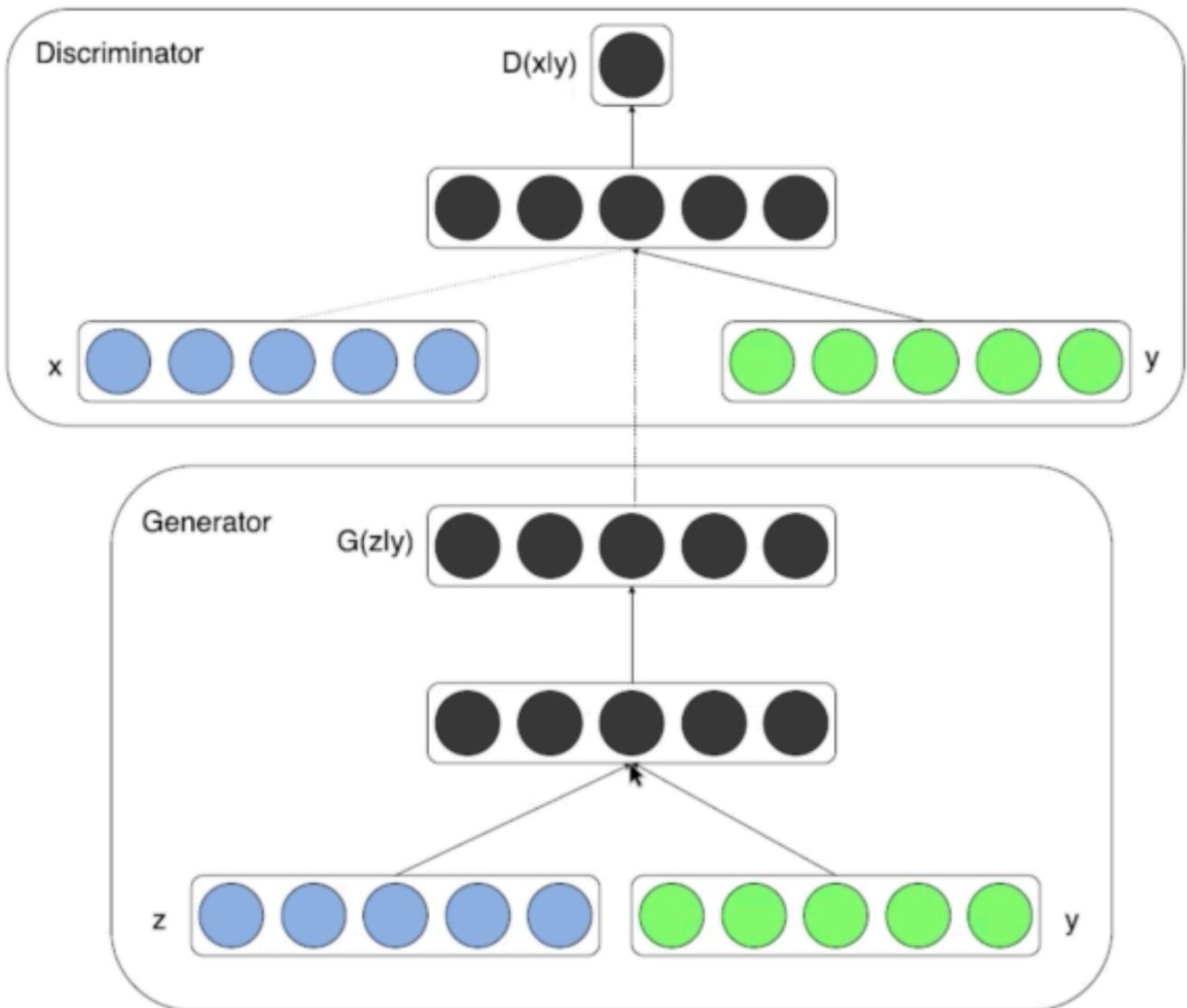
Drilling platform (540)



Schooner (780)

## 模型总览

- 在生成器和判别器分别加入相同的条件输入y
- CGAN的网络相对于原始GAN网络并没有变化
- CGAN可以作为一种通用策略嵌入到其它的GAN网络中



## 价值函数

GAN

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

cGAN

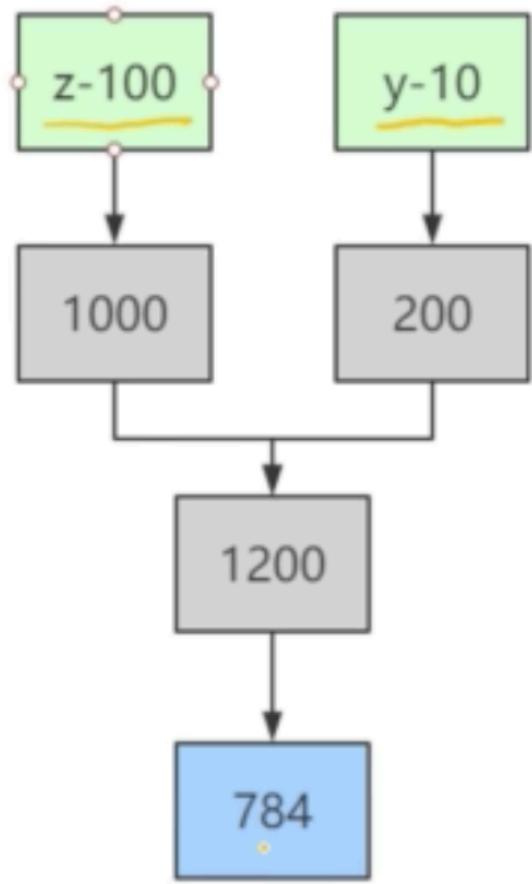
$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x|y)] + \mathbb{E}_{x \sim p_z(z)} [\log 1 - D(G(z|y))]$$

- D: 判别器； G: 生成器； z: 随机噪声； data: 训练数据； y: 条件输入

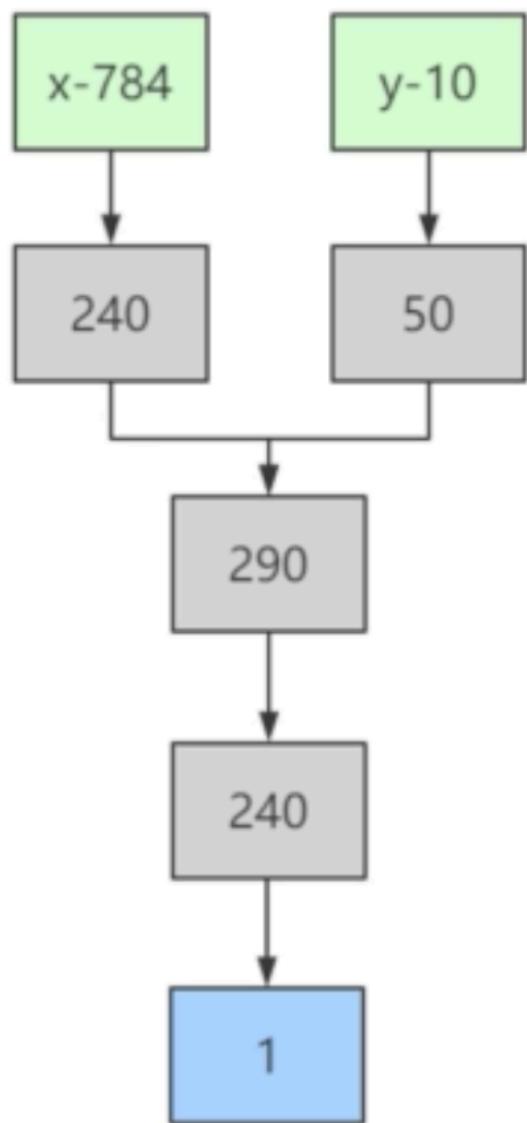
## 单模态任务

- 采用随机梯度下降， batch size 100
- 初始学习率0.1，指数衰减到1e-6，衰减系数为1.00004
- 使用初始值为0.5的初始动量，并逐渐增加到0.7
- 在生成器和判别器上都是用概率为0.5的Dropout
- 使用验证集上的最大对数似然估计作为停止点

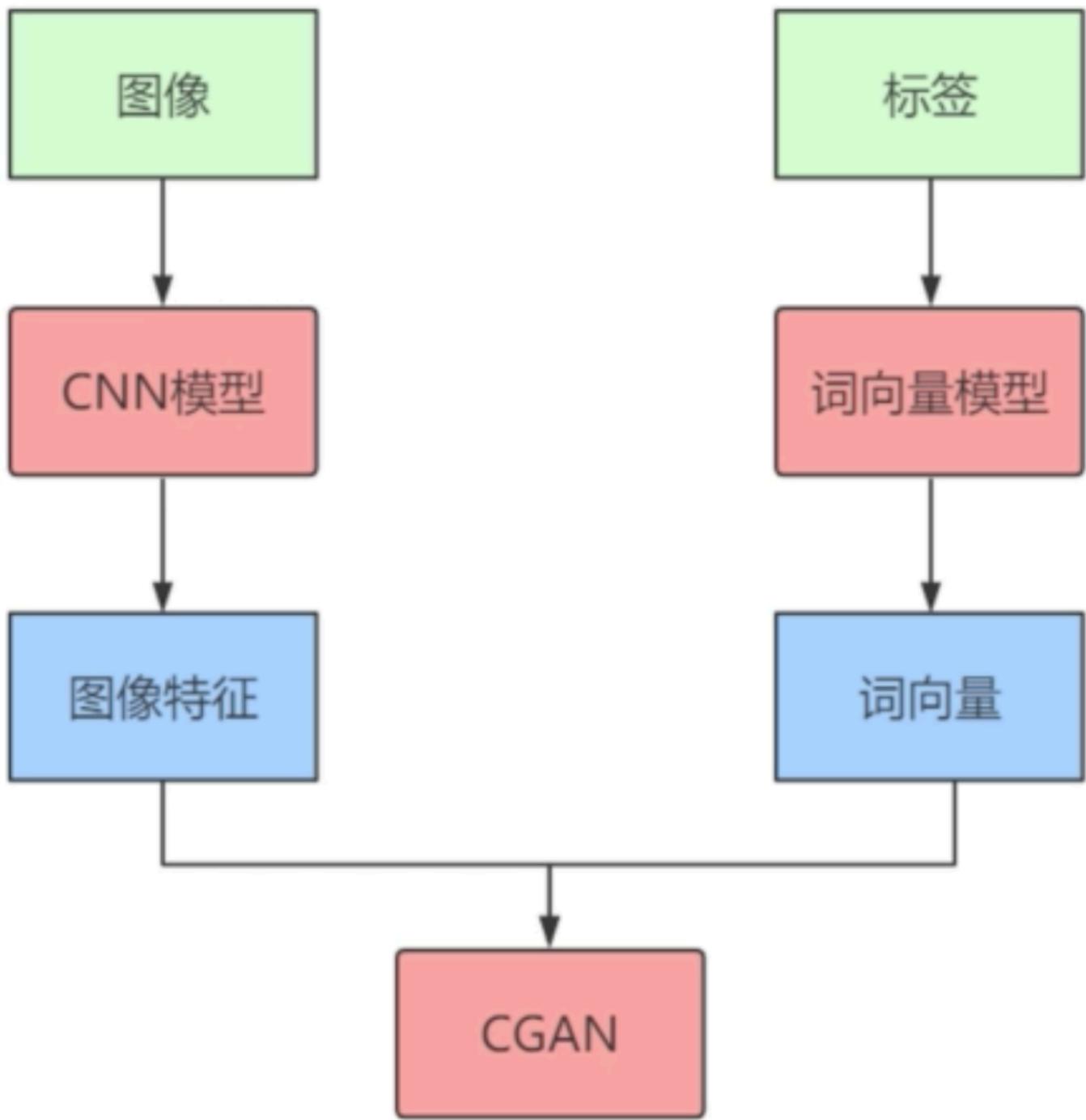
生成器



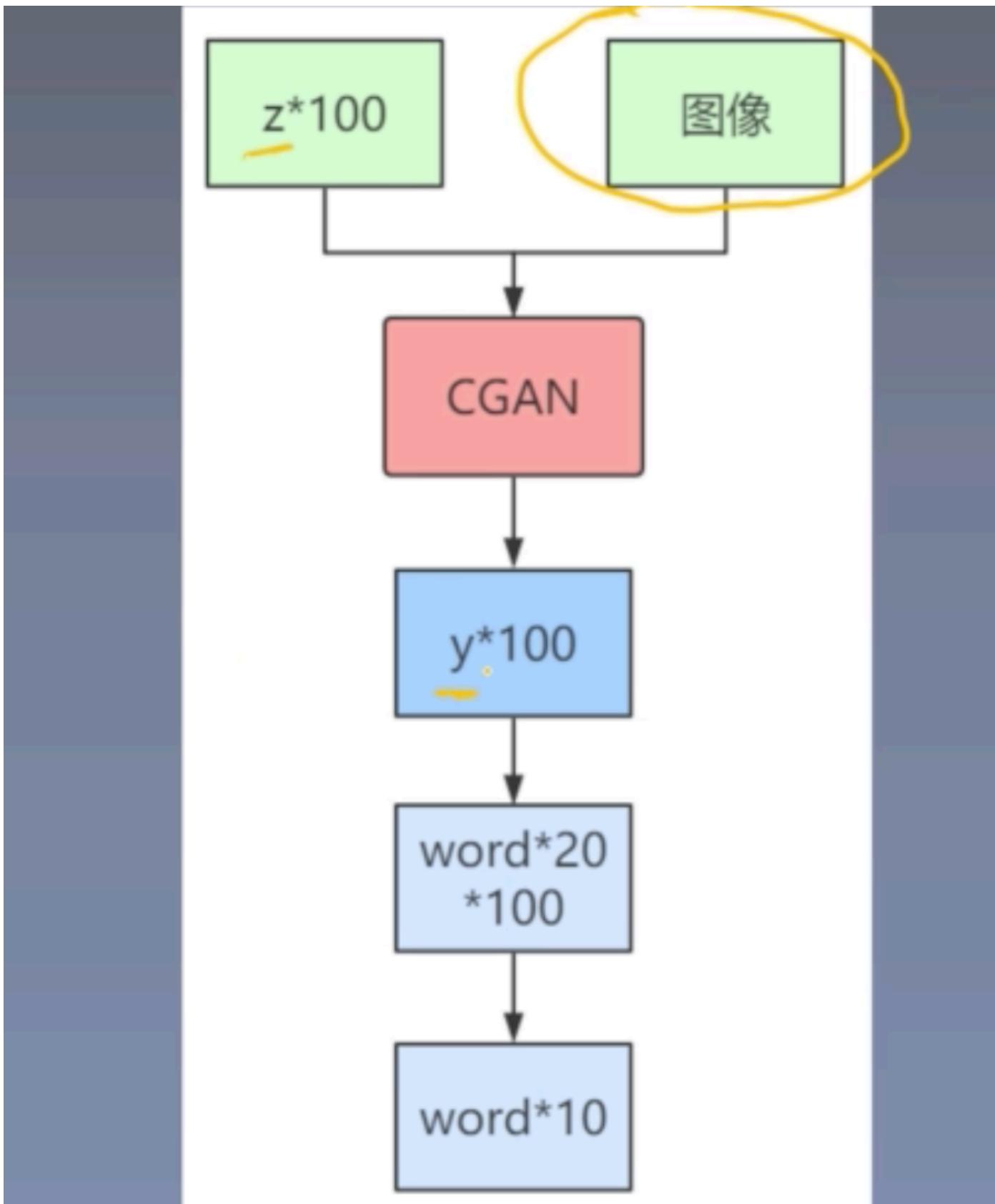
判别器



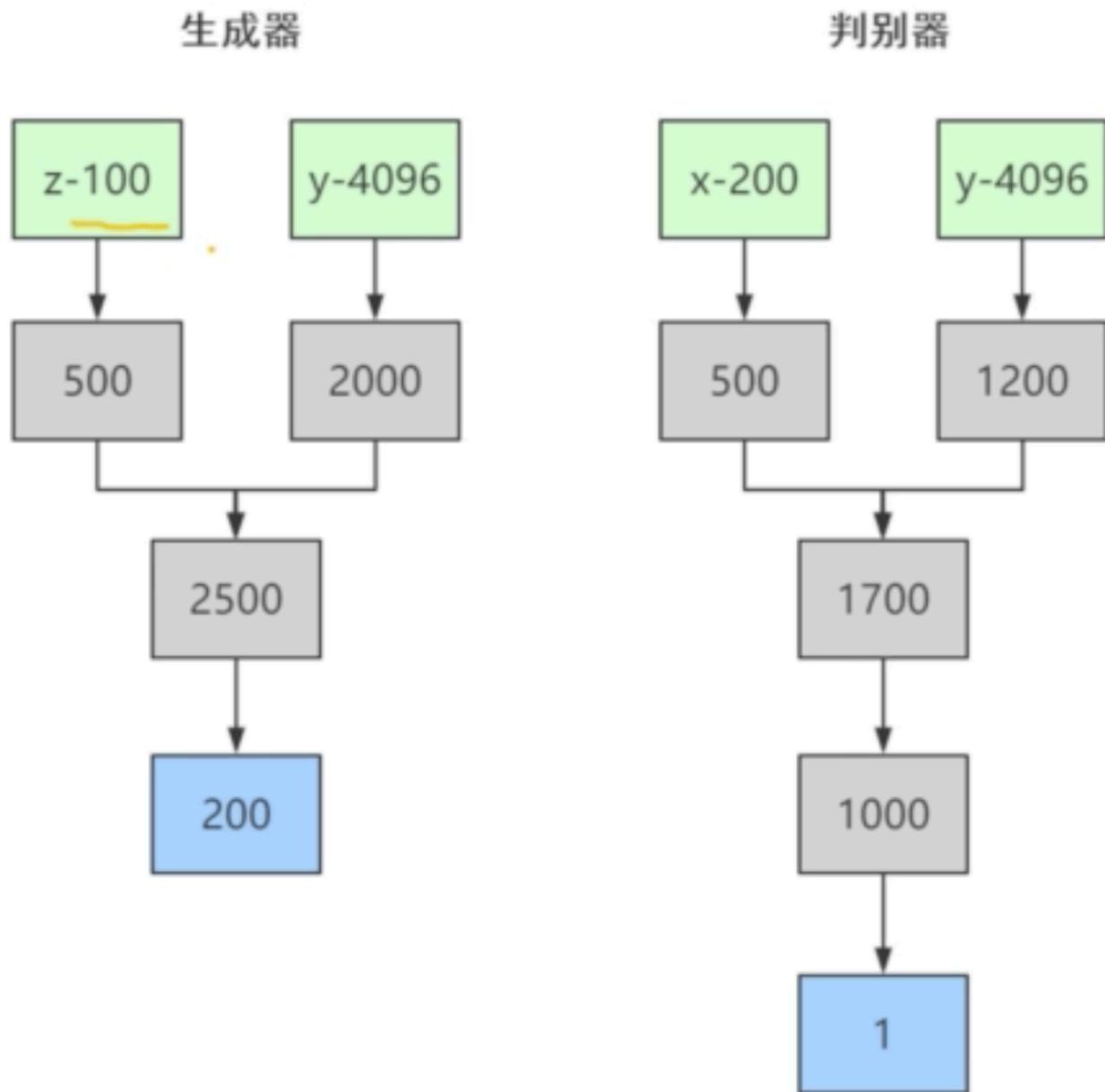
- 从像Flickr这样的照片网站可以获得丰富的带用户标记的图像数据
- 用户生成的元数据（UGM），比较接近人类用自然语言描述图像的方式
- 不同的用户会使用不同词汇来描述相同的概念
- 在ImageNet上训练一个类似AlexNet的图像分类模型，使用其最后一个全连接层的输出来提取图像特征
- 使用YFCC100M数据集，训练一个词向量长度为200的skip-gram模型



- 基于MIR Flickr 25000数据集，使用上面的图像特征提取模型和skip-gram模型分别图像和标签特征
- 把提取的图像作为条件输入， 标签特征作为输出来训练CGAN
- 在训练CGAN时，不修改图像特征提取模型和skip-gram模型
- 在训练集中具有多个标签的图像，每个标签训练一次
- 为每个条件输入生成100个样本，对于每个样本输出的词向量找到距离最近的20个单词
- 在100\*20个单词中，选择前10个最常见的单词



- 超参数和架构是结合使用交叉验证、随机网格搜索和手工选择来确定的
- 词啊用随机梯度下降，batch size为100
- 初始学习率为0.1，指数衰减至 $1e-6$ ，衰减系数为1.00004
- 使用初始值为0.5的初始动量，并逐渐增加到0.7
- 在生成器和判别器上都使用概率为0.5的Dropout



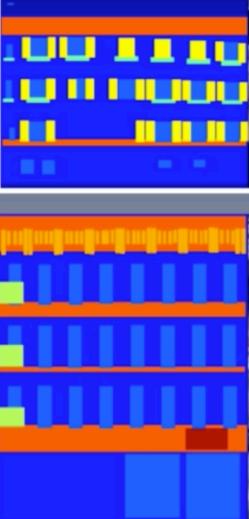
## 实验结果

- 好几种方法（包括GAN）的测试结果都优于CGAN
- CGAN的实验主要是为了概念验证，而非追求state of the art
- 相信通过探索超参数空间和网络架构，条件模型能够接近或超过非条件生成模型的结果

Model	MNIST
DBN [1]	$138 \pm 2$
Stacked CAE [1]	$121 \pm 1.6$
Deep GSN [2]	$214 \pm 1.1$
Adversarial nets	$225 \pm 2$
Conditional adversarial nets	$132 \pm 1.8$

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2  
 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3  
 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4  
 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5  
 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6  
 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7  
 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8  
 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9

User tags + annotations	Generated tags
	taxi, passenger, line, transportation, railway station, passengers, railways, signals, rail, rails
	chicken, fattening, cooked, peanut, cream, cookie, house made, bread, biscuit, bakes
	creek, lake, along, near, river, rocky, treeline, valley, woods, waters
	love, people, posing, girl, young, strangers, pretty, women, happy, life




## 总结展望

- 本文中显示的结果非常初步
- 展示了条件对抗网的潜力
- 展示了有趣并且有用的应用范围前景
- 提供更复杂的模型，对其性能和特征的更详细和彻底的分析
- 在多模态任务中，同时使用多个标签
- 使用联合寻来呢方案来学习语言模型

## 论文总结

### A: 关键点

- 在模型中加入条件输入
- 针对不同复杂度的任务，通过手工+超参数搜索设计不同的网络结构

### B: 创新点

- 将GAN应用到多模态任务上
- 将多个领域的最新结果结合起来

### C: 启发点

- 验证一个有应用价值的idea比追求SOA更重要
- 多模态的ML任务存在巨大的潜力
- 快+模型应用到新领域
- 把不同领域的最新成果结合起来，产生化学反应

## 课后作业

- 【思考题】本论文的多模态任务，是输入图像输出标签，如果要输入标签输出图像，网络该如何修改？

加入注意力机制，将标签编码成数字作为输入特征，其他网络结构可以不变

- 【代码实战】对提供的现有代码进行完善，加入模型保存、模型推断代码

```
from torchvision.utils import save_image

epoch = 0 # temporary
batches_done = epoch * len(dataloader) + i
if batches_done % opt.sample_interval == 0:
    save_image(gen_imgs.data[:25], "images/%d.png" % batches_done, nrow=5, normalize=True)
# 保存生成图像

os.makedirs("model", exist_ok=True) # 保存模型
torch.save(generator, 'model/generator.pkl')
torch.save(discriminator, 'model/discriminator.pkl')

print("gen images saved!\n")
print("model saved!")
```

- 【总结】了解在cGAN之后，GAN在多模态任务上还有哪些进展

TextToImage : stackGAN、TAC-GAN、Generative Adversarial Text to Image Synthesis、Learning What and Where to Draw

多模态学习链接：<https://www.cnblogs.com/CV-life/p/10152745.html>