



# 数字媒体技术基础

**Meng Yang**

[www.smartllv.com](http://www.smartllv.com)

**SUN YAT-SEN University**



**机器智能与先进计算教  
育部重点实验室**



**智能视觉语言  
学习研究组**



- ❑ 第11章 智能新媒体信息表示基础
- ❑ 11.1 深度神经网络基础
  - 11.1.1 全连接神经网络
  - 11.1.2 卷积神经网络
  - 11.1.3 自编码网络
  - 11.1.4 生成对抗网络
- ❑ 11.2 智能新媒体的信息表示学习
  - 11.2.1 图像预训练学习
  - 11.2.2 自然语言预训练学习

## 11.1.4 生成对抗网络

### 挑战分辨真实和AI生成的人脸

#### Experiment (1): 5 seconds

For this experiment, images of faces will flash for **5** seconds.  
After each image disappears, answer whether you think the face was real or fake.



The two faces to the left are **REAL**, and the two faces to the right are **FAKE**.  
Please note that blurring artifacts may be present for both real or fake images.

Start



## 11.1.4 生成对抗网络

### 挑战分辨真实和AI生成的人脸

(GIF动图)



3

5s甚至更短时间内，肉眼难以分辨！





问题?



# 11.1.4 生成对抗网络

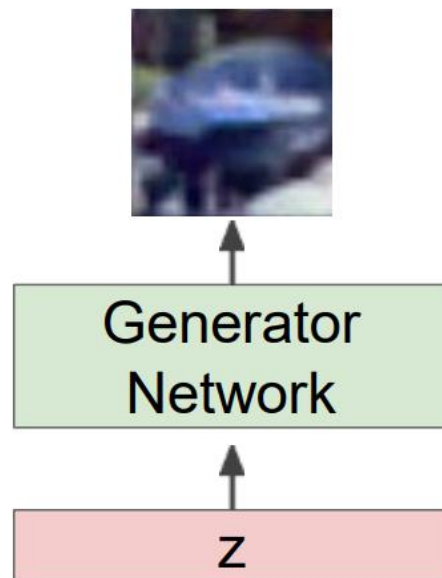
## Basic Idea

问题：希望从训练样本分布中采样新数据，但这个分布不仅维度高而且还很复杂，难以直接实现。

解决方案：对一个简单的分布采样，比如均匀分布；然后，学习一种映射将其变换到训练样本分布

输出：采样自训练样本分布的图像

输入：随机噪声

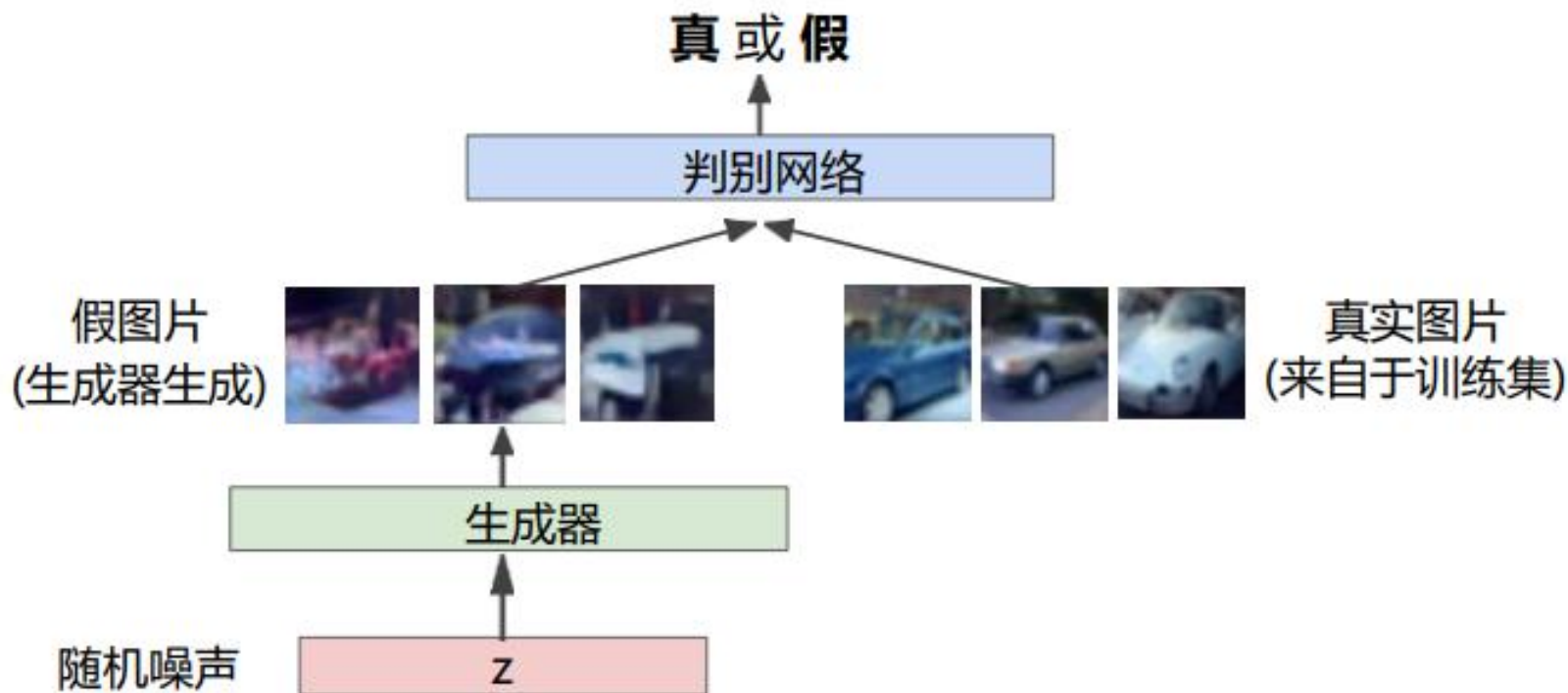


# 11.1.3 生成对抗网络

## Basic Idea

生成网络：期望能够产生尽量真实的图片，进而骗过判别器

判别网络：期望能够准确的区分真假图片

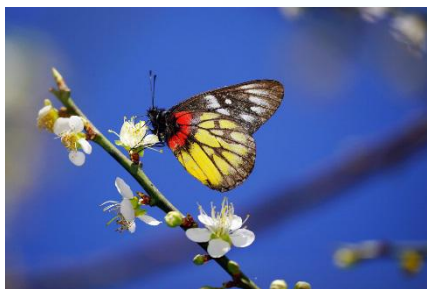


# 11.1.3 生成对抗网络



**Basic Idea** 蝴蝶躲避天敌鸟类的捕食，不断进化使得天敌误以为自己是树叶

Generator



彩蝶



棕色蝶



枯叶蝶

蝴蝶是彩色的



蝴蝶没有叶脉



.....



Discriminator

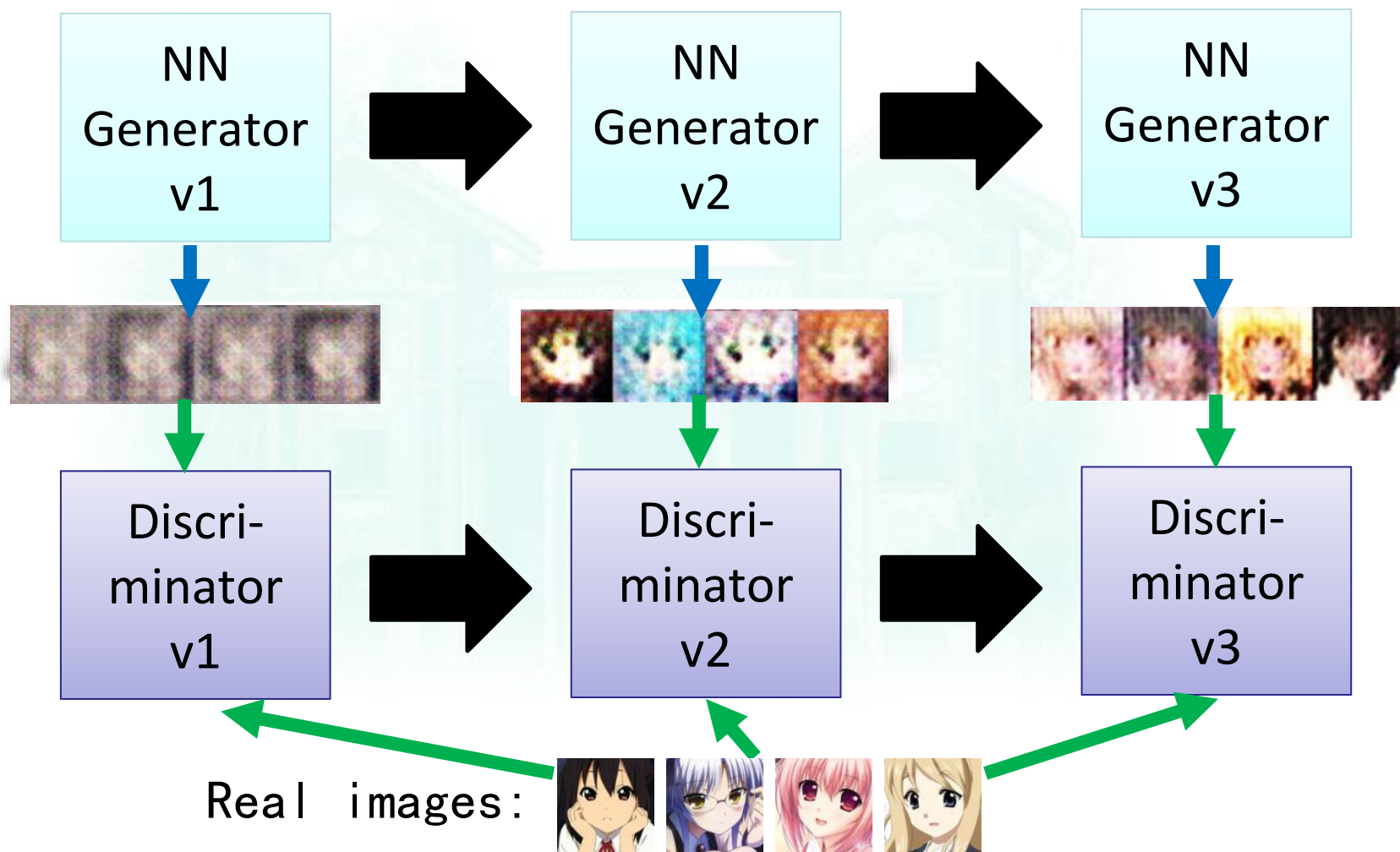


# 11.1.3 生成对抗网络



## Basic Idea

This is where the term “**对抗**” comes from.



Demo: <https://generated.photos/face-generator>

# 11.1.3 生成对抗网络



## Basic Idea

Generator

Discriminator



Generator  
v1



Discriminator  
v1

No eyes

Generator  
v2



Discriminator  
v2

No mouth

Generator  
v3



- 初始化生成器与判别器
- 每轮迭代完成:
  - Step 1: 固定生成器  $G$ , 更新判别器  $D$
  - Step 2: 固定判别器  $D$ , 更新生成器  $G$

# 11.1.4 生成对抗网络

## GAN的诞生

2014 年的一个晚上，Goodfellow 在酒吧给师兄庆祝博士毕业。一群工程师聚在一起竟然开始了深入的学术探讨——**如何让计算机自动生成照片**

针对这个问题，Goodfellow 的朋友们“煞费苦心”，提出了一个计划——对构成照片的元素进行统计分析，来帮助机器自己生成图像。

Goodfellow 一听就觉得这个想法根本行不通，马上给否决掉了。但他已经无心再趴体了，刚才的那个问题一直盘旋在他的脑海，他边喝酒边思考，突然灵光一现：**如果让两个神经网络互相对抗呢？**但朋友们对这个不靠谱的脑洞深表怀疑。Goodfellow 转头回家，决定用事实说话。写代码写到凌晨，然后测试… Goodfellow 本人都没想到，第一次测试就成功了。



左右互搏术：《射雕英雄传》中「老顽童」周伯通创出的武功，通过一心二用，能够两手同时做不同的事情，快速提升功力





# 11.1.3 生成对抗网络

## Training

采用 minimax 的方式联合训练

Minimax objective function:

判别器输出图片为真实的概率，其值在  $(0, 1)$  之间

$$\min_{\theta_g} \max_{\theta_d} \left[ \mathbb{E}_{x \sim p_{data}} \log \underbrace{D_{\theta_d}(x)}_{\substack{\text{判别器对真实} \\ \text{样本 } x \text{ 的打分}}} + \mathbb{E}_{z \sim p(z)} \log(1 - \underbrace{D_{\theta_d}(G_{\theta_g}(z))}_{\substack{\text{判别器对生成} \\ \text{样本 } G(z) \text{ 的打分}}}) \right]$$

判别器 ( $\theta_d$ ) 希望最大化目标函数使得  $D(x)$  接近于1（真实样本），而  $D(G(z))$  接近于0（假样本）

生成器 ( $\theta_g$ ) 希望最小化目标函数使得  $D(G(z))$  尽量接近于1，即希望判别器认为生成器产生的图像  $G(z)$  为真实图片

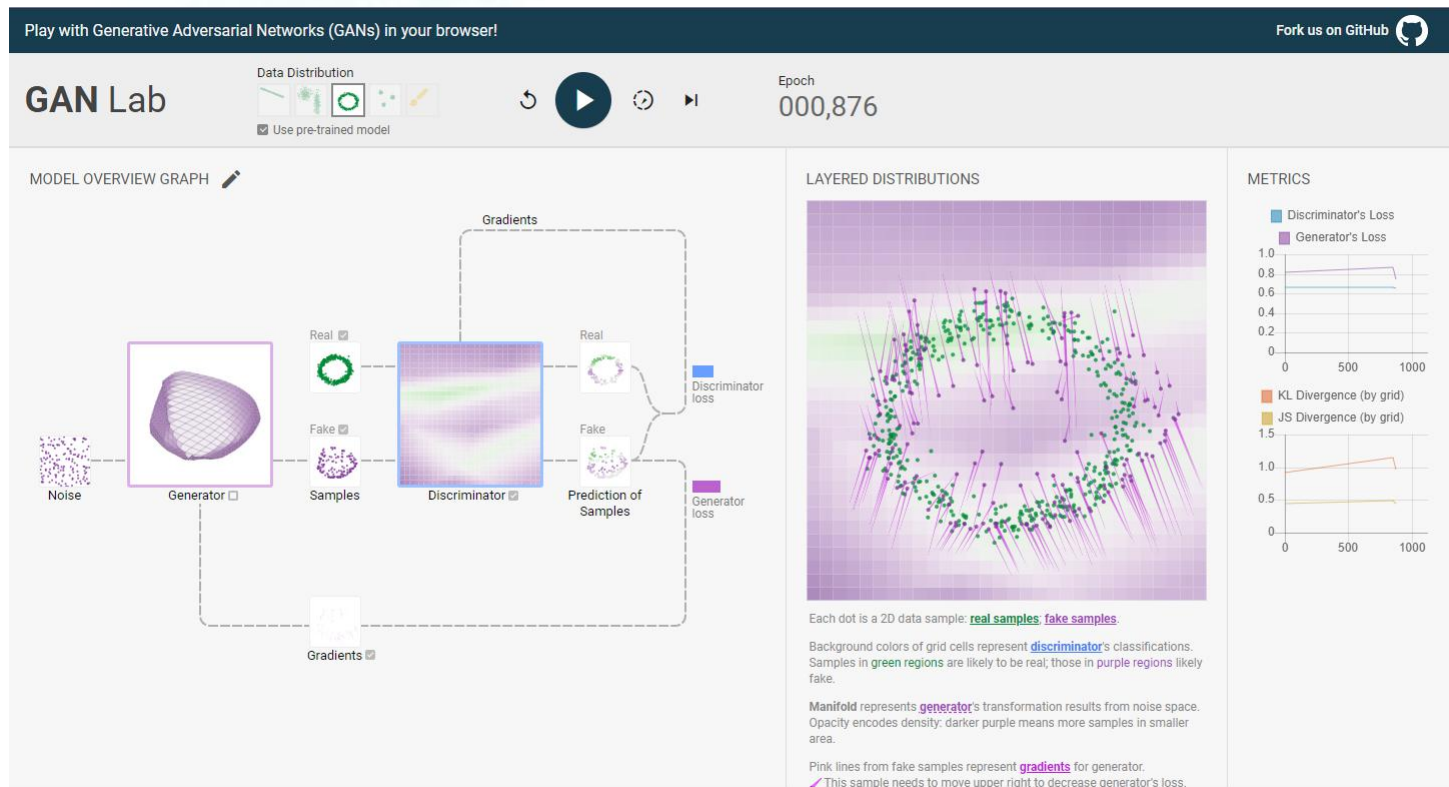


# 11.1.3 生成对抗网络



## Basic Idea

Demo: GAN Lab <https://poloclub.github.io/ganlab/>



代码: <https://github.com/poloclub/ganlab>

使用说明: <https://www.jiqizhixin.com/articles/2018-09-10-13>

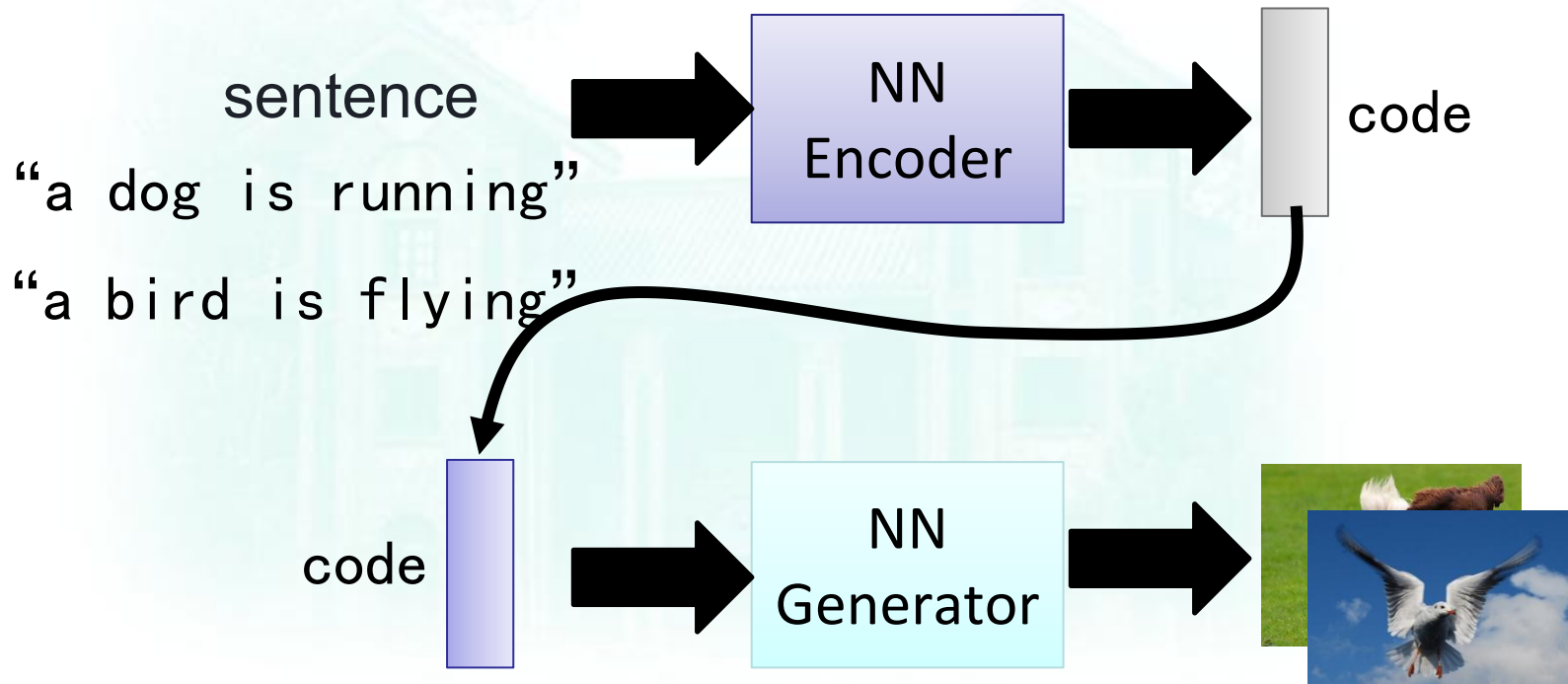




# 11.1.3 生成对抗网络

## Conditional Generation

- Generating images based on text description



# 11.1.3 生成对抗网络

## Conditional Generation

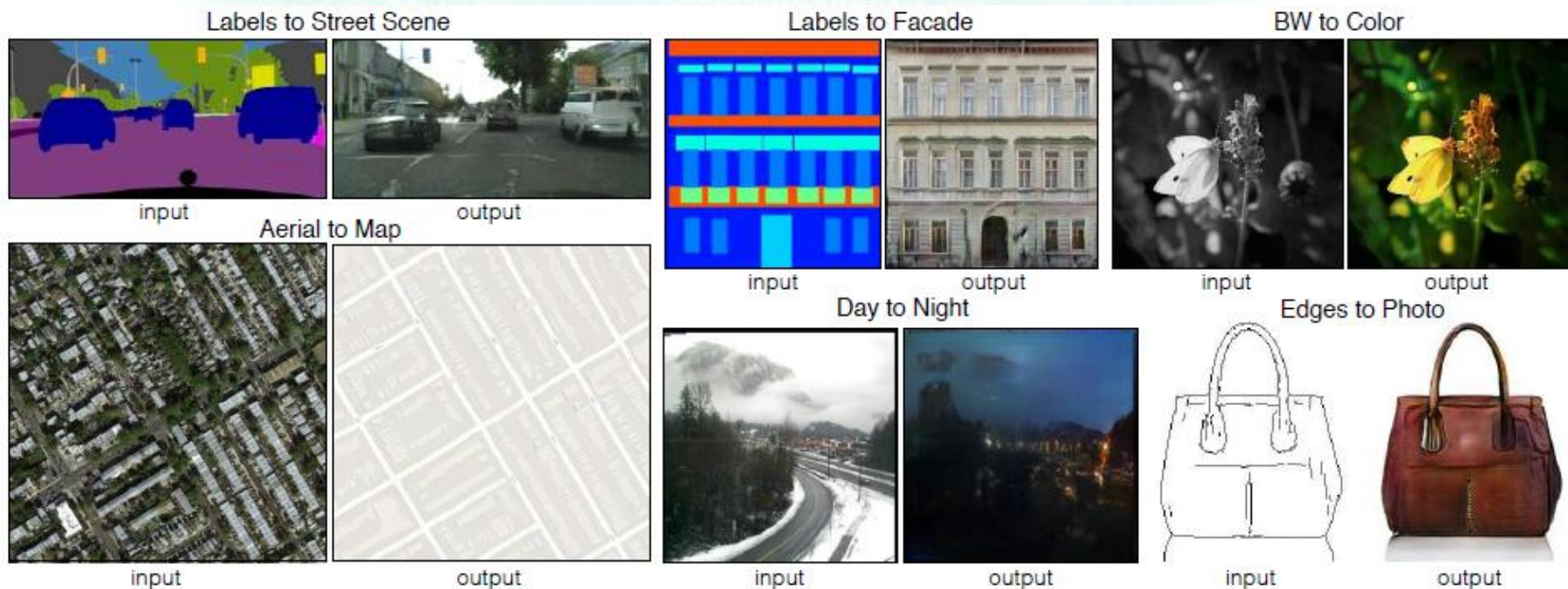
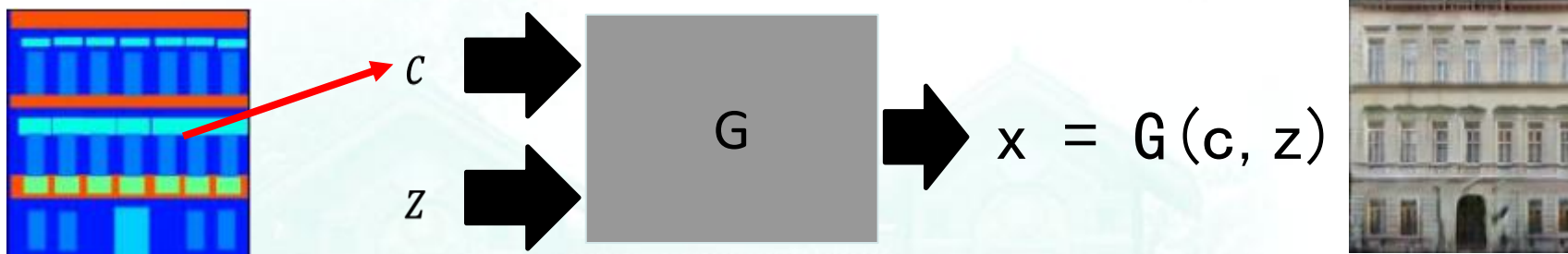
- Paired data: text to images

Caption	Image
this flower has white petals and a yellow stamen	
the center is yellow surrounded by wavy dark purple petals	
this flower has lots of small round pink petals	

# 11.1.3 生成对抗网络

## Conditional Generation

- Paired data: Image-to-image translation



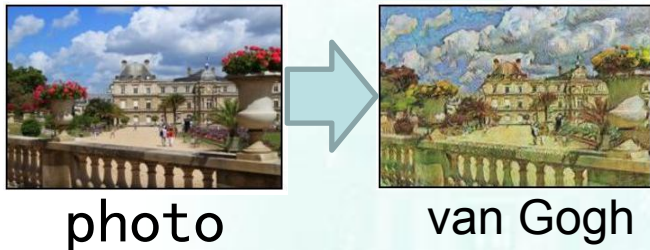
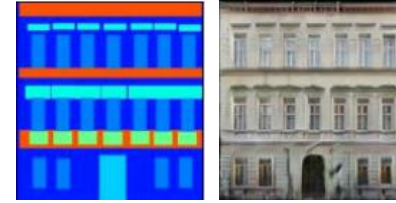


# 11.1.3 生成对抗网络

## Conditional Generation

- Unpaired data: Transform an object from one domain to another without paired data

*paired data*

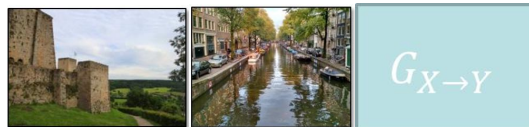


# 11.1.3 生成对抗网络

## Conditional Generation

□ Unpaired data: CycleGAN

Domain X

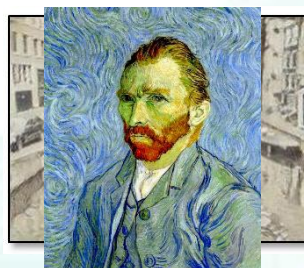
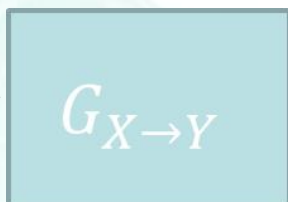


Domain Y



Domain X

Become similar to domain Y

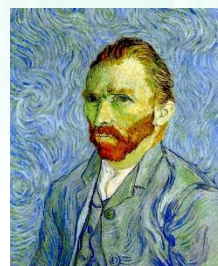


Not what we want



scalar

ignore input



Input image belongs to domain Y or not

Domain Y



# Progressive GAN



一个“明星脸”生成器，从 $4 \times 4$ 低分辨率开始，逐渐为生成器和鉴别器增加新的层，经过18天的训练，生成 $1024 \times 1024$ 大小的、以假乱真的人脸照片。



## 第二部分


### 11.2 智能新媒体的信息表示学习

## 11.2.1 图像预训练学习


### Why pre-train?

#### ❑ 从头开始训练不利于灵活部署模型.

- 训练时间限制,
- 计算资源限制,
- 训练所需的数据量限制



利用通用大规模数据集在特定的网络结构上构建预训练任务,使得模型获得足够的先验知识,再在具体下游任务上加载预训练的网络进行微调



如果原始的数据集已经足够大,足够一般,通过预训练学习到的空间上的特征层次结构就能有效地在后续的模型中工作,因此这些特征对许多计算机视觉问题都很有用

## 11.2.1 图像预训练学习

### ImageNet

如果使用的数据无法反映真实世界的状况，即便是最好的算法也无济于事



项目发起人：李飞飞  
(美国国家工程院院士、斯坦福大学教授)

<https://image-net.org/>



# 11.2.1 图像预训练学习



## ImageNet

### Geological formation, formation

(geology) the geological features of the earth

1808  
pictures

86.24%  
Popularity  
Percentile

Wordnet  
IDs

Numbers in brackets: (the number of synsets in the subtree).

ImageNet 2011 Fall Release (32326)

plant, flora, plant life (4486)

geological formation, formation (1

aquifer (0)

beach (1)

cave (3)

cliff, drop, drop-off (2)

delta (0)

diapir (0)

folium (0)

foreshore (0)

ice mass (10)

lakefront (0)

massif (0)

monocline (0)

mouth (0)

natural depression, depression (

natural elevation, elevation (41

oceanfront (0)

range, mountain range, range of

relict (0)

ridge, ridgeline (2)

ridge (0)

shore (7)

slope, incline, side (17)

spring, fountain, outflow, outpo

talus, scree (0)

vein, mineral vein (1)

volcanic crater, crater (2)

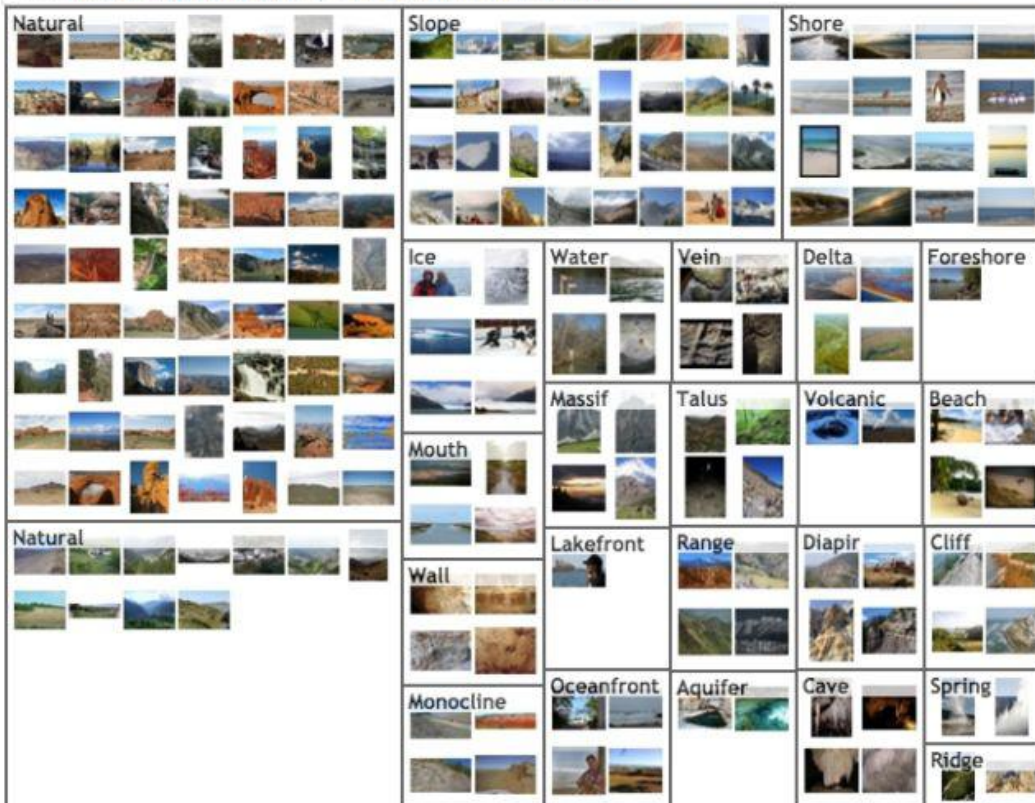
wall (0)

Treemap Visualization

Images of the Synset

Downloads

ImageNet 2011 Fall Release Geological formation, formation



<https://image-net.org/>



# 11.2.1 图像预训练学习

## COCO

目标识别、检测、语义分割



COCO数据集以scene understanding为目标，主要从复杂的日常场景中截取，图像中的目标通过精确的segmentation进行位置的标定。

COCO数据集是目前为止有语义分割的最大数据集，提供的类别有80类，有超过33万张图片，其中20万张有标注，整个数据集中个体的数目超过150万个

<https://cocodataset.org/>

# 11.2.1 图像预训练学习

## CIFAR

### CIFAR-10

飞机

汽车

鸟

猫

鹿

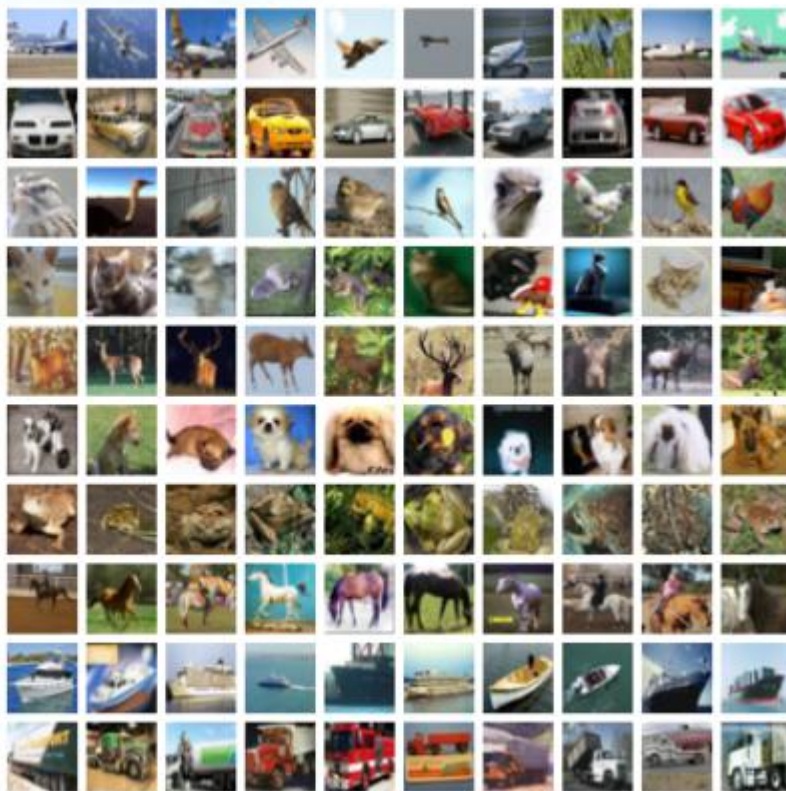
狗

青蛙

马

船

卡车



由10个类的60000个32x32彩色图像组成，每个类有6000个图像。有50000个训练图像和10000个测试图像。

数据集分为五个训练批次和一个测试批次，每个批次有10000个图像。测试批次包含来自每个类别的恰好1000个随机选择的图像。

# 11.2.1 图像预训练学习

## CIFAR

### CIFAR-100

有100个类，每个类包含600个图像。，每类各有500个训练图像和100个测试图像。

CIFAR-100中的100个类被分成20个超类。每个图像都带有一个“精细”标签（它所属的类）和一个“粗糙”标签（它所属的超类）

超类	类别
水生哺乳动物	海狸, 海豚, 水獭, 海豹, 鲸鱼
鱼	水族馆的鱼, 比目鱼, 射线, 鲑鱼, 鳕鱼
花卉	兰花, 罂粟花, 玫瑰, 向日葵, 郁金香
食品容器	瓶子, 碗, 罐子, 杯子, 盘子
水果和蔬菜	苹果, 蘑菇, 橘子, 梨, 甜椒
家用电器	时钟, 电脑键盘, 台灯, 电话机, 电视机
家用家具	床, 椅子, 沙发, 桌子, 衣柜
昆虫	蜜蜂, 甲虫, 蝴蝶, 毛虫, 蟑螂
大型食肉动物	熊, 豹, 狮子, 老虎, 狼
大型人造户外用品	桥, 城堡, 房子, 路, 摩天大楼
大自然的户外场景	云, 森林, 山, 平原, 海
大杂食动物和食草动物	骆驼, 牛, 黑猩猩, 大象, 袋鼠
中型哺乳动物	狐狸, 豪猪, 负鼠, 浣熊, 臭鼬
非昆虫无脊椎动物	螃蟹, 龙虾, 蜗牛, 蜘蛛, 蠕虫
人	宝贝, 男孩, 女孩, 男人, 女人
爬行动物	鳄鱼, 恐龙, 蜥蜴, 蛇, 乌龟
小型哺乳动物	仓鼠, 老鼠, 兔子, 母老虎, 松鼠
树木	枫树, 橡树, 棕榈, 松树, 柳树
车辆1	自行车, 公共汽车, 摩托车, 皮卡车, 火车
车辆2	割草机, 火箭, 有轨电车, 坦克, 拖拉机

<http://www.cs.toronto.edu/~kriz/cifar.html>





# 11.2.1 图像预训练学习

## Pre-trained Models



- AlexNet
- VGG
- ResNet
- SqueezeNet
- DenseNet
- Inception v3
- GoogLeNet
- ShuffleNet v2
- MobileNetV2
- MobileNetV3
- ResNeXt
- Wide ResNet
- MNASNet

```
import torchvision.models as models
resnet18 = models.resnet18(pretrained=True)
alexnet = models.alexnet(pretrained=True)
squeezenet = models.squeezenet1_0(pretrained=True)
vgg16 = models.vgg16(pretrained=True)
densenet = models.densenet161(pretrained=True)
inception = models.inception_v3(pretrained=True)
googlenet = models.googlenet(pretrained=True)
shufflenet = models.shufflenet_v2_x1_0(pretrained=True)
mobilenet_v2 = models.mobilenet_v2(pretrained=True)
mobilenet_v3_large = models.mobilenet_v3_large(pretrained=True)
mobilenet_v3_small = models.mobilenet_v3_small(pretrained=True)
resnext50_32x4d = models.resnext50_32x4d(pretrained=True)
wide_resnet50_2 = models.wide_resnet50_2(pretrained=True)
mnasnet = models.mnasnet1_0(pretrained=True)
```

加载预训练的网络参数初始化网络

<https://pytorch.org/vision/stable/models.html>



- torchvision.datasets
  - CelebA
  - CIFAR
  - Cityscapes
  - COCO
  - DatasetFolder
  - EMNIST
  - FakeData
  - Fashion-MNIST
  - Flickr
  - HMDB51
  - ImageFolder
  - ImageNet
  - Kinetics-400
  - KMNIST
  - LSUN
  - MNIST
  - Omniglot
  - PhotoTour
  - Places365
  - QMNIST
  - SBD
  - SBU
  - STL10
  - SVHN
  - UCF101
  - USPS
  - VOC

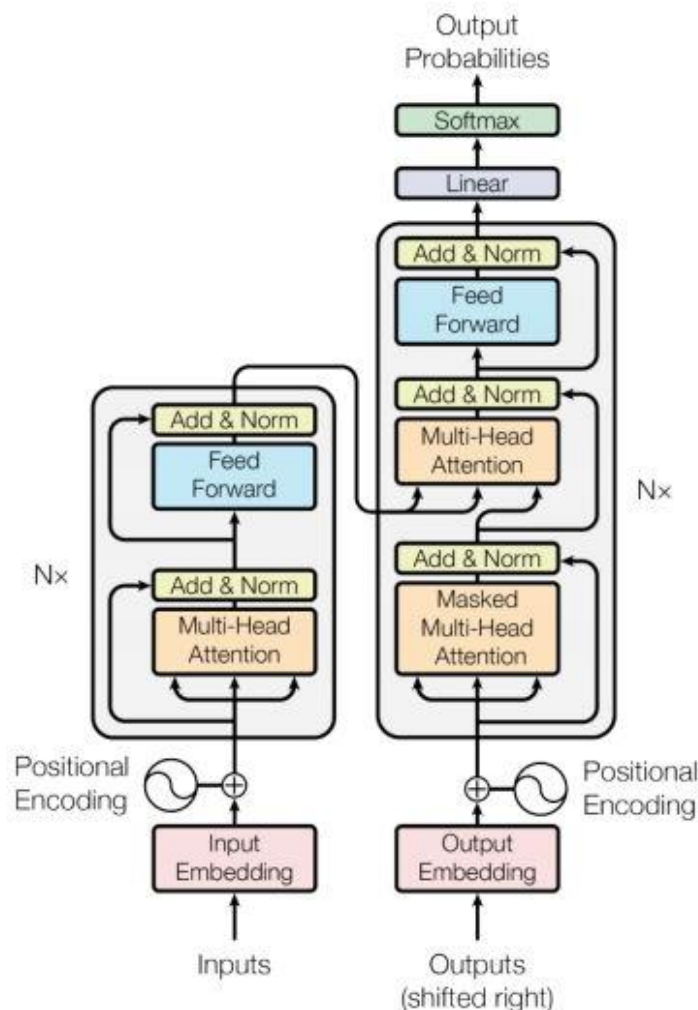
# 11.2.2 自然语言预训练学习

## Transformer

当前NLP的主流预训练模型大多是基于Transformer模型来预训练的。

Transformer: 由Encoder和Decoder组成。

- ❑ Encoder: 对输入的文本编码成向量表示。
- ❑ Decoder: 对Encoder得到的向量表示进行解码, 输出预测结果 (例如预测单词)。





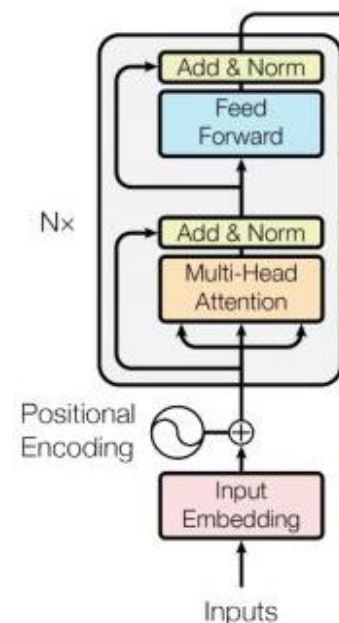
## 11.2.2 自然语言预训练学习

### Transformer

Encoder：由6层Block堆叠而成，每层包括两个sub-layer：

- ❑ 第一个sub-layer是Multi-head self-attention：对多个词之间进行交互，并在多个子空间中计算，能考虑更多维度的语义信息。
- ❑ 第二个sub-layer是全连接网络。
- ❑ 并且每个sub-layer都加入残差连接和正则化：

$$\text{LayerNorm}(x + \text{Sublayer}(x))$$

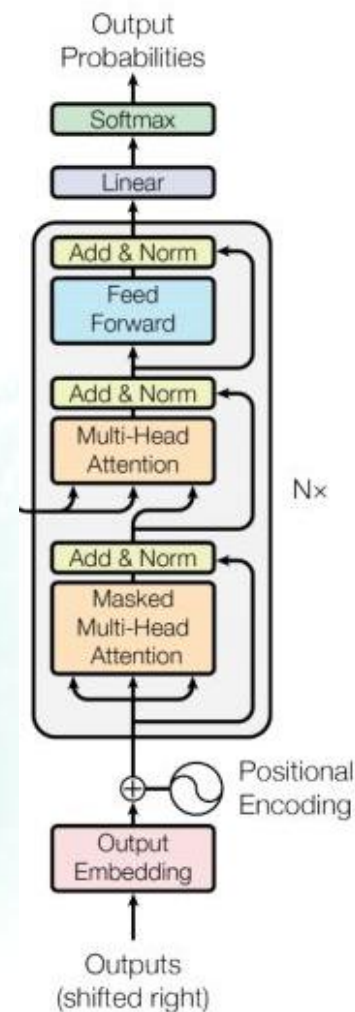


## 11.2.2 自然语言预训练学习

### Transformer

Decoder：同样由6层Block堆叠而成，每层包括三个sub-layer。

它和Encoder的主要区别在于：多了一个sub-layer，用于接收Encoder Block传来的信息。





## 11.2.2 自然语言预训练学习

### BERT

- ❑ BERT使用Transformer作为主要框架。
- ❑ BERT的本质是通过在海量语料运行自监督学习的方法来训练模型，以得到良好的单词或文本表示。所谓自监督学习是指在没有人工标注的数据上运行的监督学习。
- ❑ 维基百科、百度百科、图书、电影台词等都可作为训练语料。





## 11.2.2 自然语言预训练学习

### BERT

Bert的预训练过程，可理解为：

输入一个加噪的句子，预测完整的句子。

单词级别的预测训练：

- 80%加mask：我来[mask]中山大学 → 【BERT】 → 我来自中山大学
- 10%替换新词：我来了中山大学 → 【BERT】 → 我来自中山大学
- 10%保持不变：我来自中山大学 → 【BERT】 → 我来自中山大学

句子级别的预测训练：

- 判断句子B是否是句子A的下文，是则输出1，否则输出0.

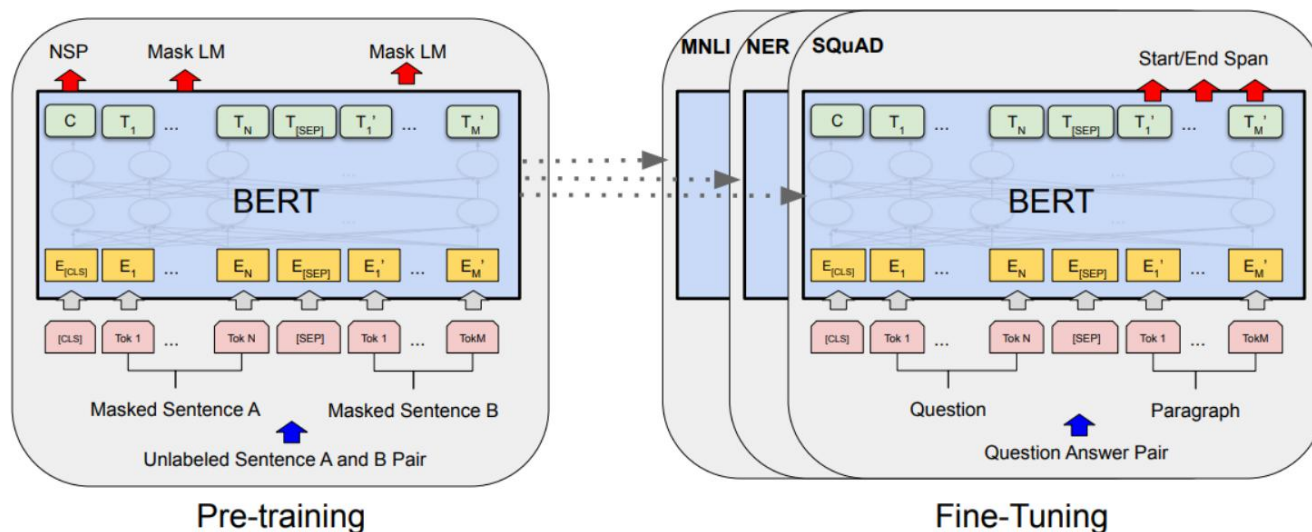


# 11.2.2 自然语言预训练学习

## BERT

预训练完成后，BERT就具备编码词向量/文本向量的能力：

- ❑ 可直接用来对新段落编码，得到词或文本的表示；
- ❑ 也可在其他具体任务上微调（finetune），以适应不同的任务：



Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. 2019: 4171-4186.





## 11.2.2 自然语言预训练学习

### 更大的预训练语料——RoBERTa

RoBERTa和BERT的区别主要体现在：

- 更多训练数据（16G→160G）
- 更大的batch size（256→8K）
- 动态Masking机制：原本BERT是在数据处理就将所有语料加mask，这样在各个训练epoch时加mask的结果都是一样的；而RoBERTa是在训练时动态加mask，一定程度上起到数据增强的作用。

GLUE数据集实验结果

	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS	WNLI	Avg
<i>Single-task single models on dev</i>										
BERT <sub>LARGE</sub>	86.6/-	92.3	91.3	70.4	93.2	88.0	60.6	90.0	-	-
XLNet <sub>LARGE</sub>	89.8/-	93.9	91.8	83.8	95.6	89.2	63.6	91.8	-	-
RoBERTa	<b>90.2/90.2</b>	<b>94.7</b>	<b>92.2</b>	<b>86.6</b>	<b>96.4</b>	<b>90.9</b>	<b>68.0</b>	<b>92.4</b>	<b>91.3</b>	-

## 11.2.2 自然语言预训练学习

### 更少的模型参数——ALBERT

ALBERT和BERT的区别主要体现在：

- ❑ 分解embedding层：原embedding层的参数为 $V \times H$ ，ALBERT将其分解为 $V \times E$ 和 $E \times H$ ；
- ❑ 层间参数共享：让每层共享同样的模型参数，让每层的参数更稳定：

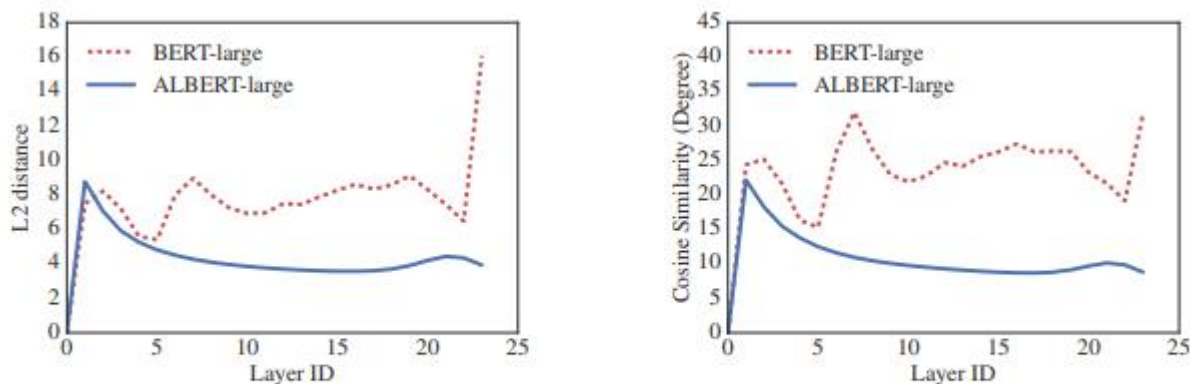
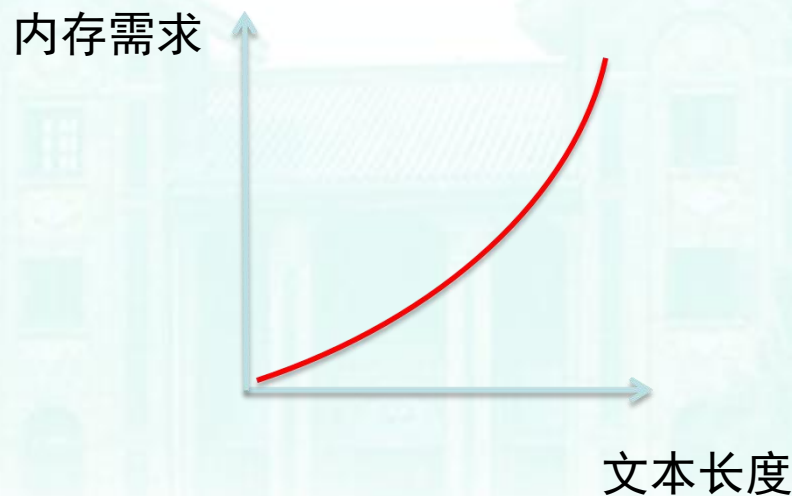


Figure 1: The L2 distances and cosine similarity (in terms of degree) of the input and output embedding of each layer for BERT-large and ALBERT-large.

## 11.2.2 自然语言预训练学习

### 适用更长的文本——Longformer

- 由于self-attention的存在，每一个token都要与其他所有token进行交互，复杂度为 $O(N^2)$ ，因此BERT难以适应太长的文本。

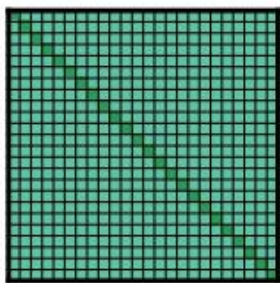


- 常见的做法是将长文本切分成多个片段，分别用BERT编码。但这样缺乏片段之间的交互。

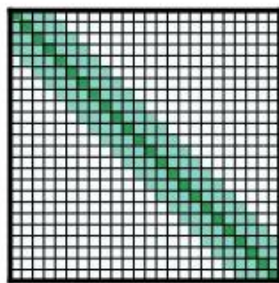
## 11.2.2 自然语言预训练学习

### 适用更长的文本——Longformer

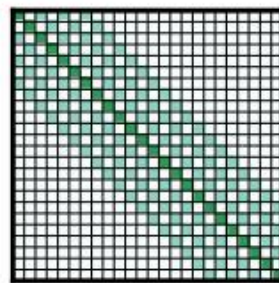
Longformer提出了三种attention的新模式：



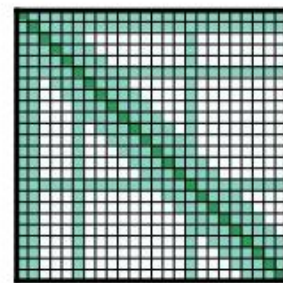
(a) Full  $n^2$  attention



(b) Sliding window attention



(c) Dilated sliding window



(d) Global+sliding window

- ❑ 滑动窗口：仅计算局部attention
- ❑ 空洞滑动窗口：覆盖范围更广一些
- ❑ 全局attention：对于一些任务（例如问答），问题的信息是很重要的，在这些位置上，会对整个序列做attention。



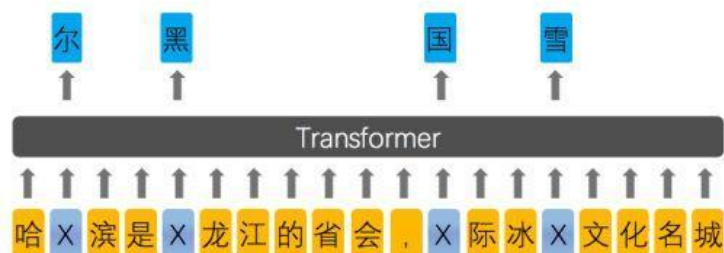
## 11.2.2 自然语言预训练学习

### 引入外部知识——ERNIE

将知识和语言语义信息融合，增强了语义的表示。

- 在模型结构上，用Text-Encoder和Knowledge-Encoder分别对文本和从文本中识别得到的命名实体编码。
- 在训练过程，mask主要加在实体上：

Learned by BERT



Learned by ERNIE



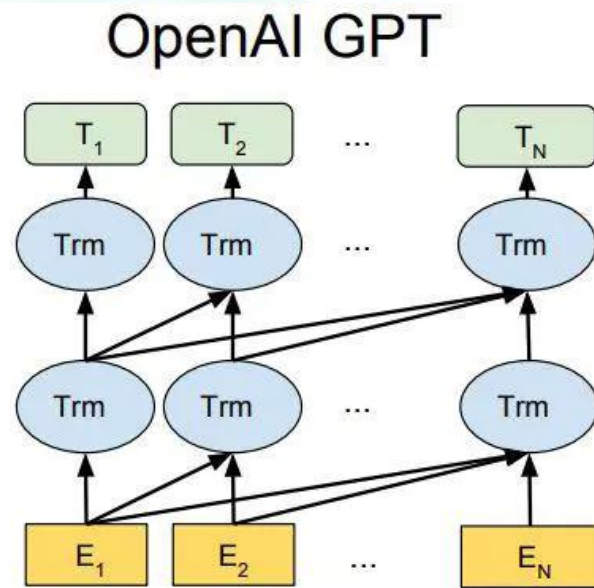
哈尔滨是黑龙江省的省会，国际冰雪文化名城

## 11.2.2 自然语言预训练学习

### 生成式预训练模型——GPT

- ❑ BERT系列模型采用Transformer的Encoder作为基础结构；
- ❑ 而GPT采用Transformer的Decoder作为基础结构。

因此GPT是单向语言模型，虽然对语言的建模能力逊于双向语言模型，但具备生成语言的能力。



# 11.2.2 自然语言预训练学习

## 生成式预训练模型——GPT

GPT同样可用于各类NLP任务：先预训练，再微调

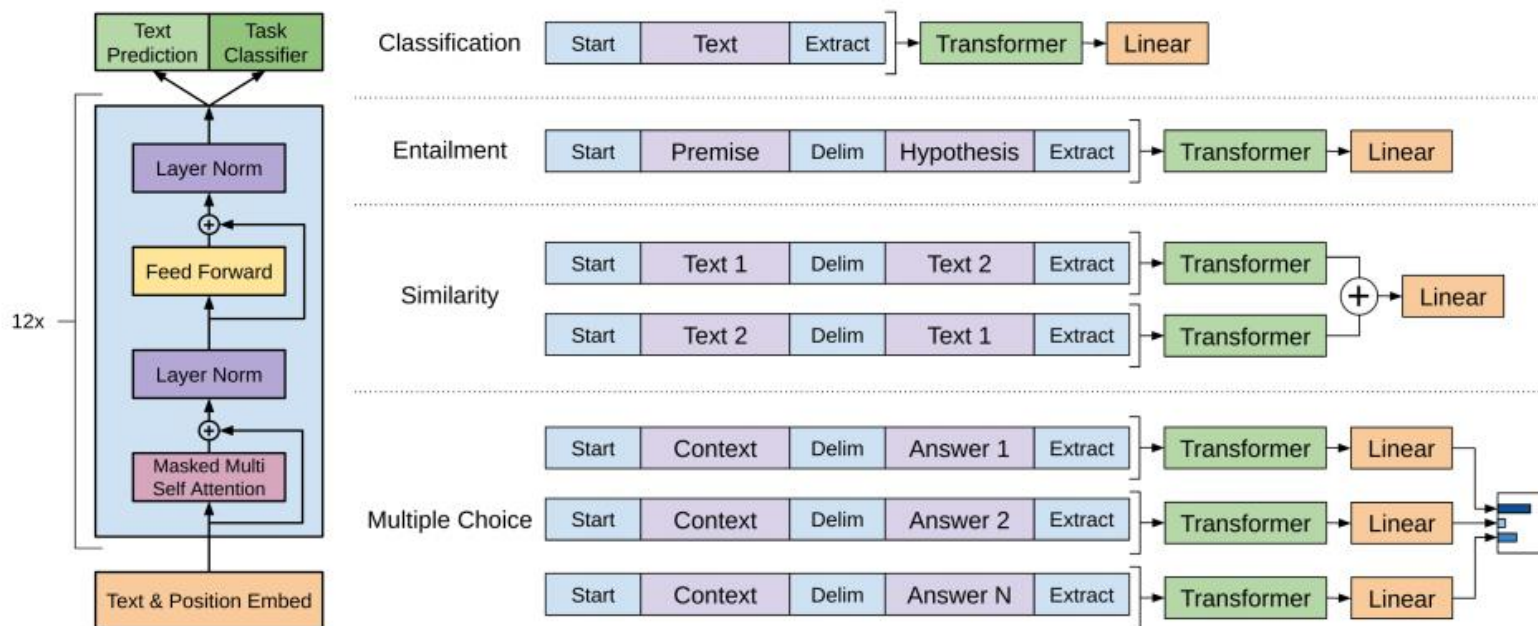
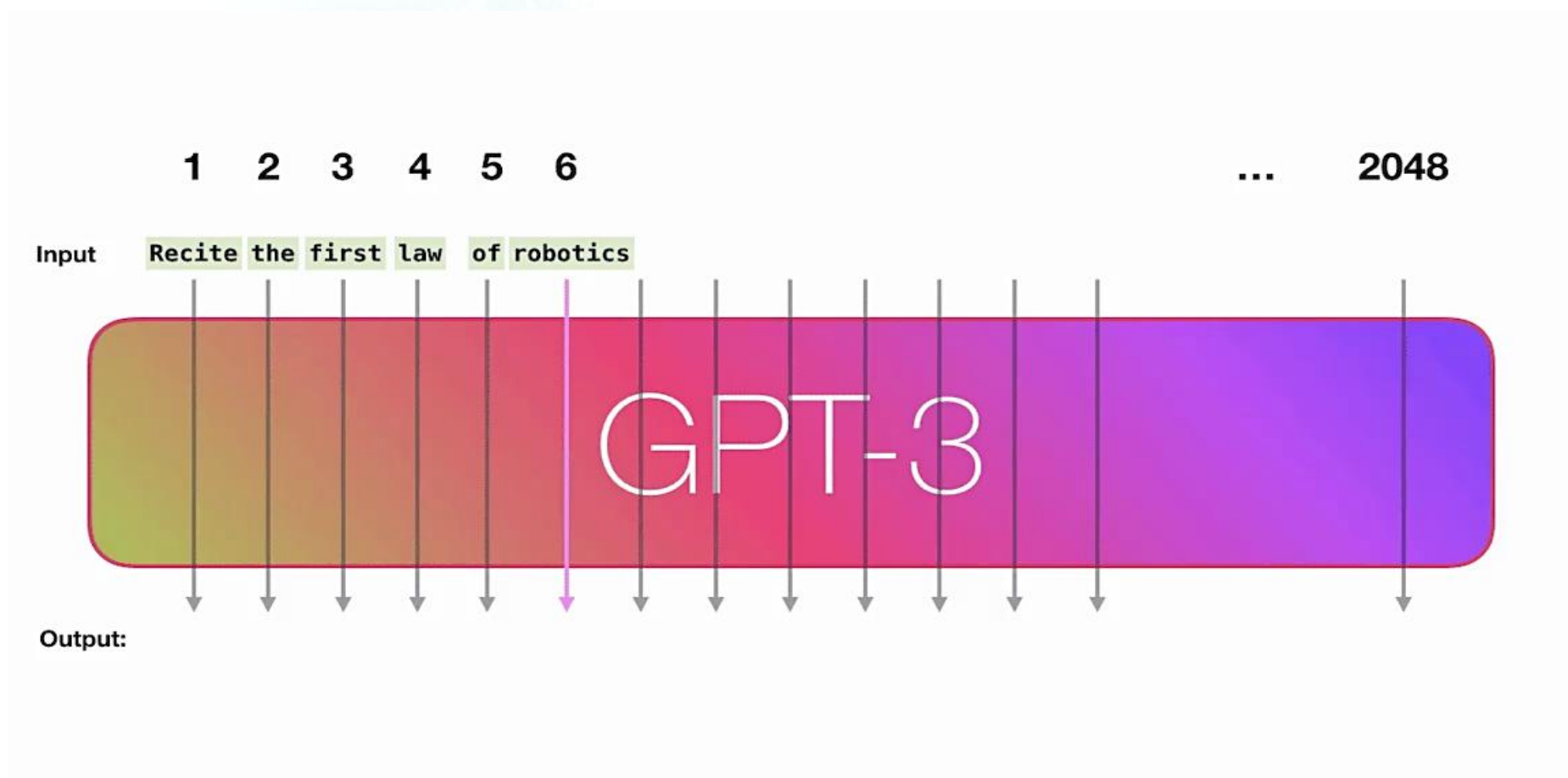


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

## 11.2.2 自然语言预训练学习

### 生成式预训练模型——GPT

GPT3的生成过程示意（短视频）：







## 11.2.2 自然语言预训练学习

### 生成式预训练模型——GPT

#### GPT3根据标题生成故事

Title: United Methodists Agree to Historic Split  
Subtitle: Those who oppose gay marriage will form their own denomination  
Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.

The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

**Figure 3.14:** The GPT-3 generated news article that humans had the greatest difficulty distinguishing from a human written article (accuracy: 12%).

