

# AI与视觉表达部分

jee & 散步

# 人工智能的基本原理

散步 sanbuphy

Datawhale

physicoadam@gmail.com

# 个人介绍



**Sanbu**

sanbuphy

Ask if you don't understand, learn if you don't know.

Edit profile

👤 195 followers · 254 following

📍 china

🌐 <https://www.aospacewalk.cn/>

AIGC从业者

Datawhale开源组织成员

人工智能开源社区 & 产学研协作

<https://github.com/sanbuphy>

致力于用AI创造美好的事物

如何定义“智能”？

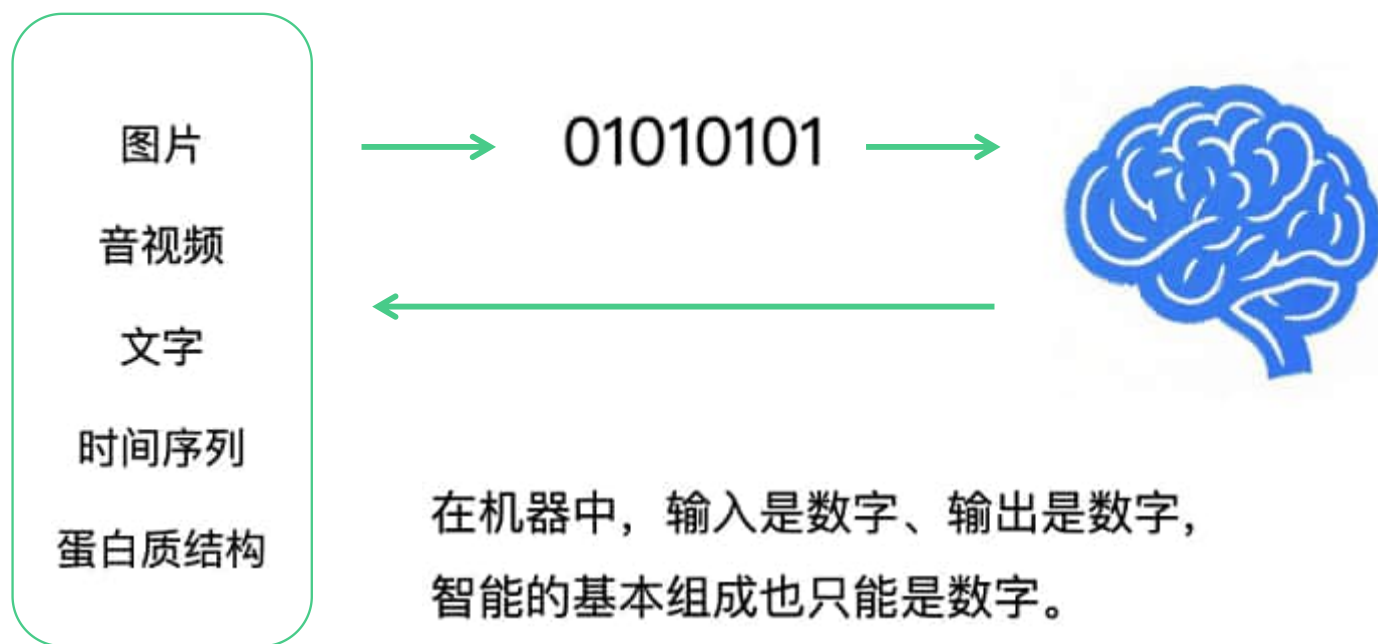


“大脑”里藏有什么？



# “智能”的机器表示

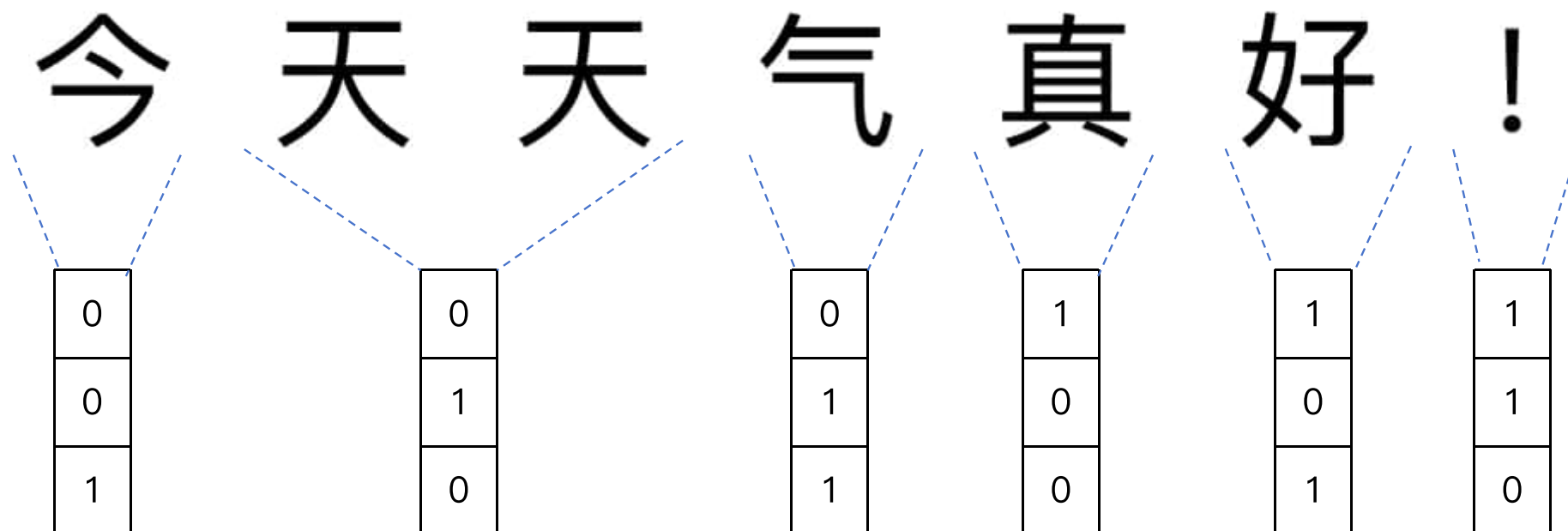
为了让机器理解世界，需要获得世界的“机器表示”



## 输入的机器表示 · 文字

假设计算机只能认识几个字，定义简单的词汇表：{今, 天, 气, 真, 好, !}

其中每个“字”用数字表示可以是 [0,0,1], [0,1,0], [0,1,1], [1,0,0] .....



## 输入的机器表示 · 图片



0					
0					
0					
1					
0					
0					



“特征”

0.5	-0.2	1.3	0.8	-0.1	0.4	0.9
-----	------	-----	-----	------	-----	-----



0.8	0.2
-----	-----

猫咪 狗

图片相比语言，是一种“冗余”的输入，  
我们需要对图片进行信息量的过滤和抽象，得到更低维度的表示。



# 输入的广义表示 Embedding



- 如果每一个字/图片都用独立的编码表示 (... , 0, 0, 0, 0, 1 ...), 那我们需要多大的数组表达?
- 文字、图片, 甚至是音频作为输入, 能不能表达它们之间的**关联关系**?

一个电影平台有1000部电影, 每部电影用1000维向量表示:

- 《速度与激情》 = [0, 0, 1, 0, 0, ..., 0]
- 《赛车总动员》 = [0, 0, 0, 1, 0, ..., 0]

将电影映射到20维空间:

《速度与激情》 = [0.8, 0.7, 0.2, ..., 0.3]

《赛车总动员》 = [0.7, 0.6, 0.3, ..., 0.4]

☑ 降低表示的成本      ☑ 容易看出**关联关系**



\* 维度是指用来表示一个对象(如单词、图像、用户等)的特征数量, “在不同方面上, 该对象在不同衡量角度下的表现”

\* 词向量的维度应该怎么选择? —— 苏剑林 <https://www.spaces.ac.cn/archives/7695/comment-page-1>



# 输入的广义表示 Embedding

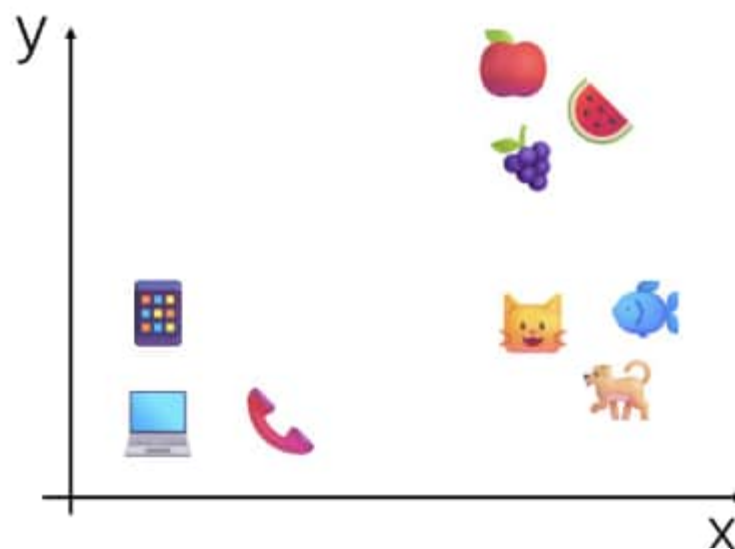
如果我们能把数据从**高维转换为低维**表示，不同数据间的关系将一目了然

Points: 10000 | Dimension: 200



试试在矢量空间中以数字方式表示超过 7 万个英语单词

<https://projector.tensorflow.org/?hl=zh-cn>

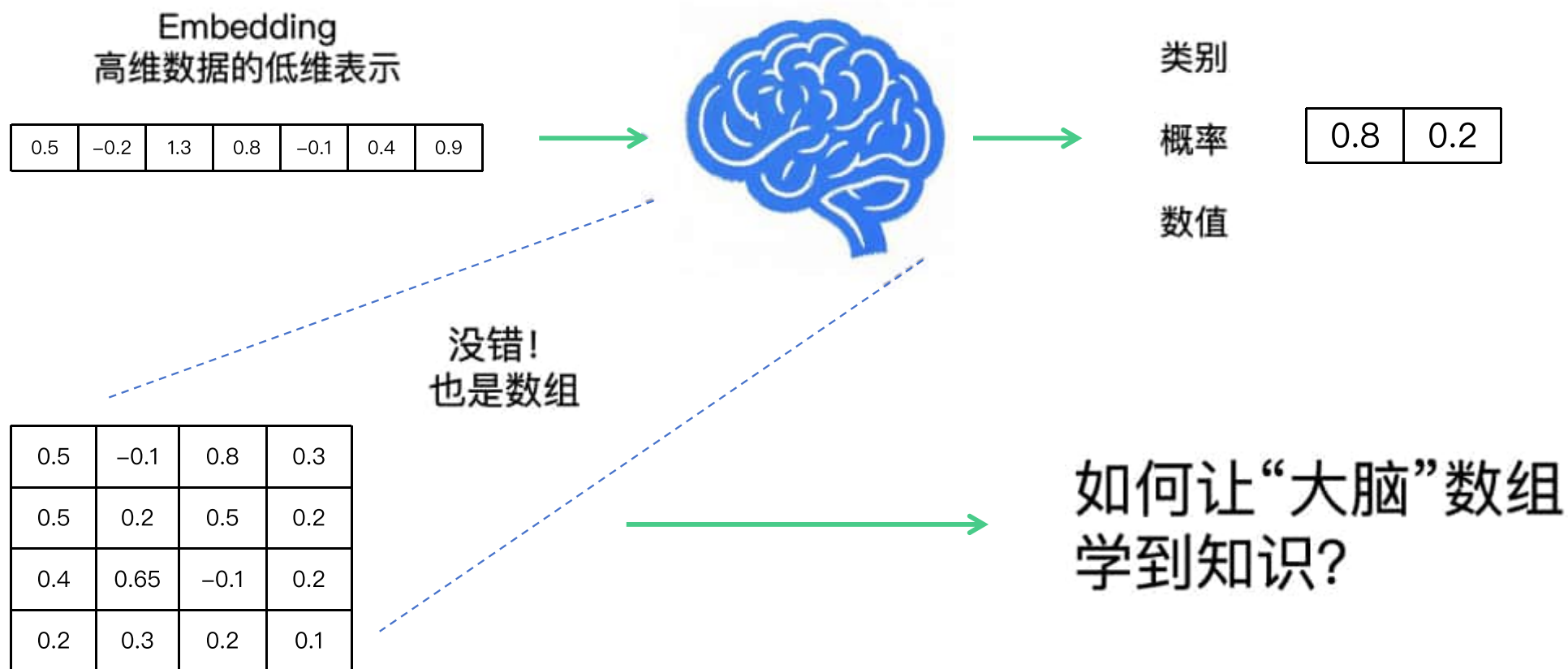


图像 Embedding 的简单示意图

\* Embedding 也是从模型中得到的，越好的模型对数据的相关关系表达的越准确。

# 学习的基本原理

学会了输入在计算机中的表示，我们继续思考——“大脑”里藏有什么？



# 学习的基本原理·计算

Embedding  
高维数据的低维表示

0.5	-0.2	1.3	0.8
-----	------	-----	-----

X

0.5	-0.1	0.8	0.3
0.5	0.2	0.5	0.2
0.4	0.65	-0.1	0.2
0.2	0.3	0.2	0.1



0.83	0.995	0.33	0.45
------	-------	------	------



猫	狗
0.4	0.6



预测值：狗

真实标签：猫

😞 预测错误，怎么办？

$$\text{Loss} = - \sum y_i \log(\hat{y}_i)$$



根据优化策略反向传播

更新数组

不断迭代



0.5	-0.1	0.8	0.3
0.5	0.4	0.5	0.2
0.2	0.65	0.1	0.2
0.2	0.3	0.2	0.8



0.57	0.955	0.59	1.01
------	-------	------	------



0.8	0.2
-----	-----

预测值：猫 ✓

真实标签：猫



0.5	-0.2	1.3	0.8
-----	------	-----	-----

X

\* 其中  $y_i$  是真实标签的概率分布， $\hat{y}_i$  是模型预测的概率分布。

# 学习的基本原理·优化策略

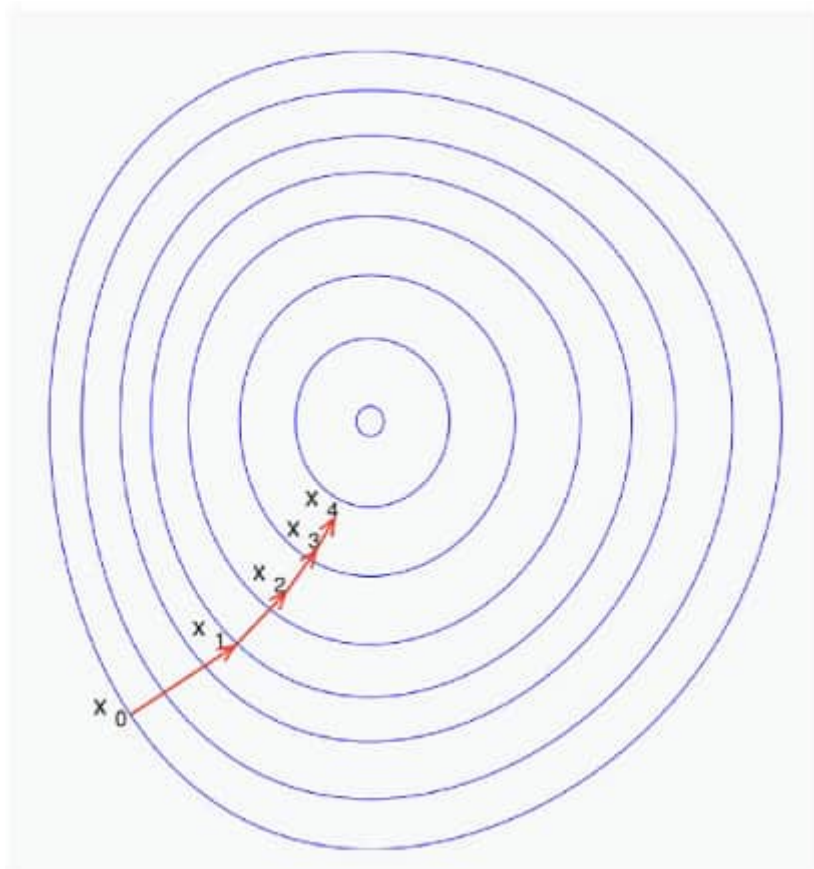


Illustration of gradient descent on a series of level sets · wiki

最小化损失函数

$$w := w - \eta \nabla Q(w) = w - \frac{\eta}{n} \sum_{i=1}^n \nabla Q_i(w).$$

更新对应权重



为什么需要一步步接近损失最小值？  
能否一步抵达终点？

- 权重和损失函数的计算很复杂，求解算力要求大
- 实际问题为复杂非线性函数，没有解析解

通过随机梯度下降，数组将逐渐接近使 loss 最小的分布，  
这就是学习的过程。

# 学习的基本原理·如何学得更好

理解了学习，那为何我们叫它“深度”学习？——原因在如何让学习更加有效：

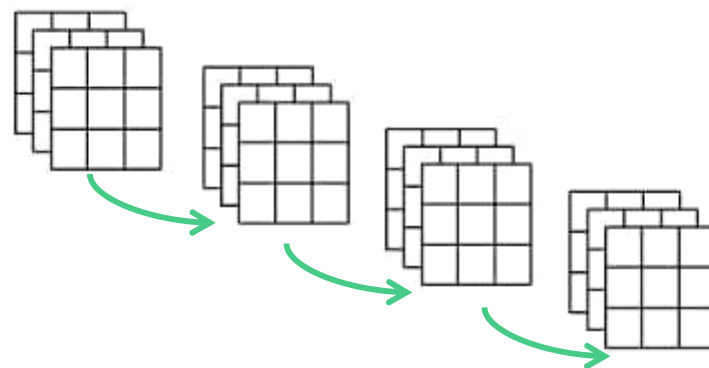
0.5	-0.1	0.8	0.3
0.5	0.2	0.5	0.2
0.4	0.65	-0.1	0.2
0.2	0.3	0.2	0.1

小模型 简单的数组



0.5	-0.1	0.3	0.5	-0.1	0.8	0.3
0.5	0.2	0.2	0.5	0.2	0.5	0.2
0.4	0.65	0.2	0.4	0.65	-0.1	0.2
0.2	0.3	0.1	0.2	0.3	0.2	0.1

更大的数组



更“深入”的网络

- 更换不同优化策略、损失函数让损失更小、精度更高
- 只要让模型能够做的更深、拥有更大参数就好
- 设计结构让模型知识记忆的更好，能够拥有更好的上下文关联（注意力机制）



# AI的发展规律·苦涩的教训



“深度学习”，是不是只要简单的把网络做大做深，就可以实现好的智能？



规则都是人造的  
模型训练终有尽头



## The Bitter Lesson

Rich Sutton

March 13, 2019

# NO!

Dr. Richard Sutton



“计算机界的诺贝尔奖”

ACM A.M. **TURING AWARD** HONORS TWO RESEARCHERS WHO LED THE DEVELOPMENT OF CORNERSTONE AI TECHNOLOGY

Andrew Barto and Richard Sutton Recognized as Recipients of the A.M. Turing Award for their Contributions to the Foundations of Reinforcement Learning

ACM, the Association for Computing Machinery, today named [Andrew Barto](#) and [Richard Sutton](#) as the recipients of the A.M. Turing Award for developing the conceptual and algorithmic foundations of reinforcement learning. In a series of papers beginning in the 1980s, Barto and Sutton introduced the theory, constructed the mathematical foundations, and developed



## AI的发展规律·苦涩的教训



人工智能研究70年来最大的教训是，利用计算能力的通用方法最终是最有效的，而且优势显著。

研究人员为了在短期内取得进展，试图利用他们对领域的人类知识，但从长远来看，唯一重要的是利用计算能力。

在语音识别领域，DARPA在20世纪70年代赞助了一场早期竞赛。参赛者包括许多利用人类知识的特殊方法——关于单词、音素、人类声道等知识。另一方面是更具统计性质的新方法，基于隐马尔可夫模型(HMMs)进行更多计算。同样，统计方法战胜了基于人类知识的方法。

思维的实际内容极其复杂且无法简化；我们应该停止寻找思考思维内容的简单方法，例如思考空间、物体、多智能体或对称性的简单方法。我们应该只内置能够找到并捕捉这种任意复杂性的元方法。

# AI的发展规律



Ilya Sutskever

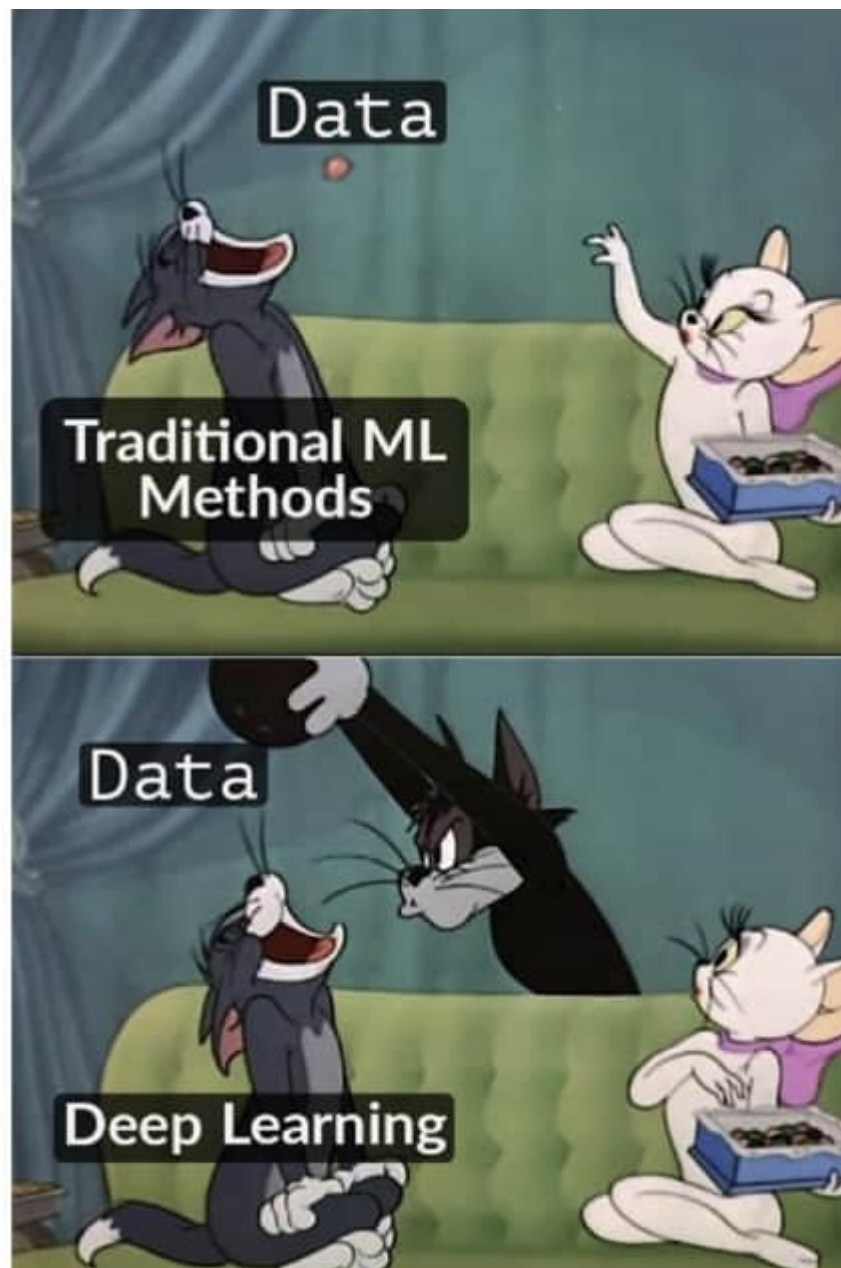
OpenAI 前首席科学家

师从杰弗里·辛顿（AI之父之一）

“假设你读了一本侦探小说，复杂的情节，不同的人物，许多事件，神秘的线索，现在还不清楚。

在书的最后一页，侦探收集了所有的线索，然后侦探召集所有人，并说：“好吧，我将揭示谁犯了罪，那个人的名字是……”——如果说对，那意味着他理解了全文。

只要能够非常好的预测下一个token，就能帮助人类创建AGI（通用人工智能）。



# AI的发展规律·大模型



图片

音视频

文字

时间序列

蛋白质结构

小模型的时代

任务分类 + 设计标签 + 设计损失 ..... → 模型 20M

大模型的时代

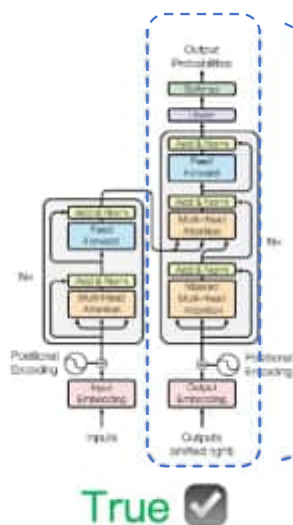
Embedding + Transformer / Diffusion → 模型 6G +  
模型 150G +  
模型 400G +



💡 什么是 Transformer ?



False ❌



True ✅

解码器架构



大量的数据  
大量的数据  
大量的数据

博客  
文章  
报刊  
书籍

.....



## AI的发展规律·强化学习

除了大，模型还需要一些恰到好处的奖励

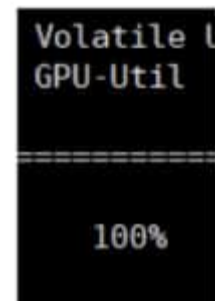


# 扩散模型 Diffusion model

## 大模型

LAION-400M  
image/text  
Status: Released

Model	SDXL
# of UNet params	2.6B
Transformer blocks	[0, 2, 10]
Channel mult.	[1, 2, 4]
Text encoder	CLIP ViT-L & OpenCLIP ViT-bigG
Context dim.	2048



数据量大 (几十M~几百M数据)

参数量大 (硬盘占据2G到20G不等)

耗时长 (5s~1min)

## 扩散过程



## 生成图片、视频

也可以生成文字



# 扩散模型·训练原理



还记得开头的模型学习过程吗？

Embedding  
高维数据的低维表示

$\begin{bmatrix} 0.5 & -0.2 & 1.3 & 0.8 \end{bmatrix} \times$

0.5	-0.1	0.8	0.3
0.5	0.2	0.5	0.2
0.4	0.85	-0.1	0.2
0.2	0.3	0.2	0.1

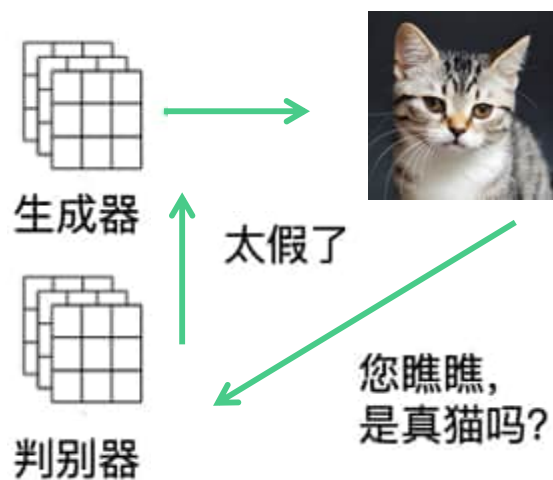
$\rightarrow \begin{bmatrix} 0.83 & 0.995 & 0.33 & 0.46 \end{bmatrix}$

猫 狗

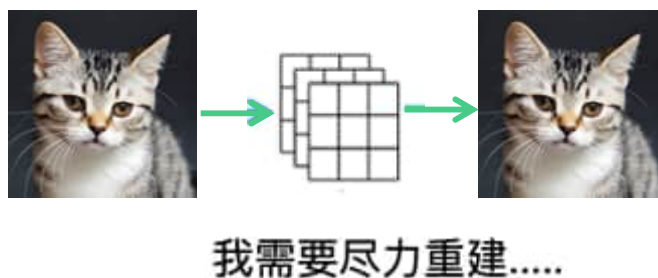
$\begin{bmatrix} 0.4 & 0.6 \end{bmatrix}$

- 我们构建了模型输入、随机初始化权重、计算出最后的输出。
- 之后计算输出与标签的距离（损失函数），根据最小损失的目的，反向更新参数。 **生成模型有标签吗？**

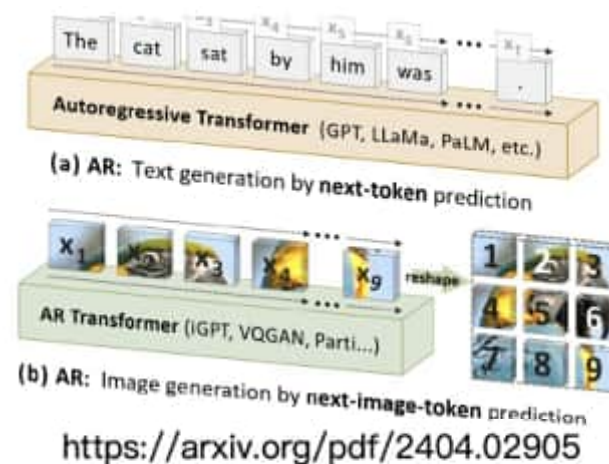
## GAN 网络模型



## VAE 模型



## AR 模型





# 扩散模型·训练原理

loss: 预测噪声和实际噪声之间的均方误差

正向过程



逆向过程



当前原图

当前时间步  $T$

0.5	-0.1	0.8	0.3
0.5	0.2	0.5	0.2
0.4	0.65	-0.1	0.2
0.2	0.3	0.2	0.1

扩散模型  
去噪过程

减去噪声



模型预测单步  
加入的噪声



获得上一步图片

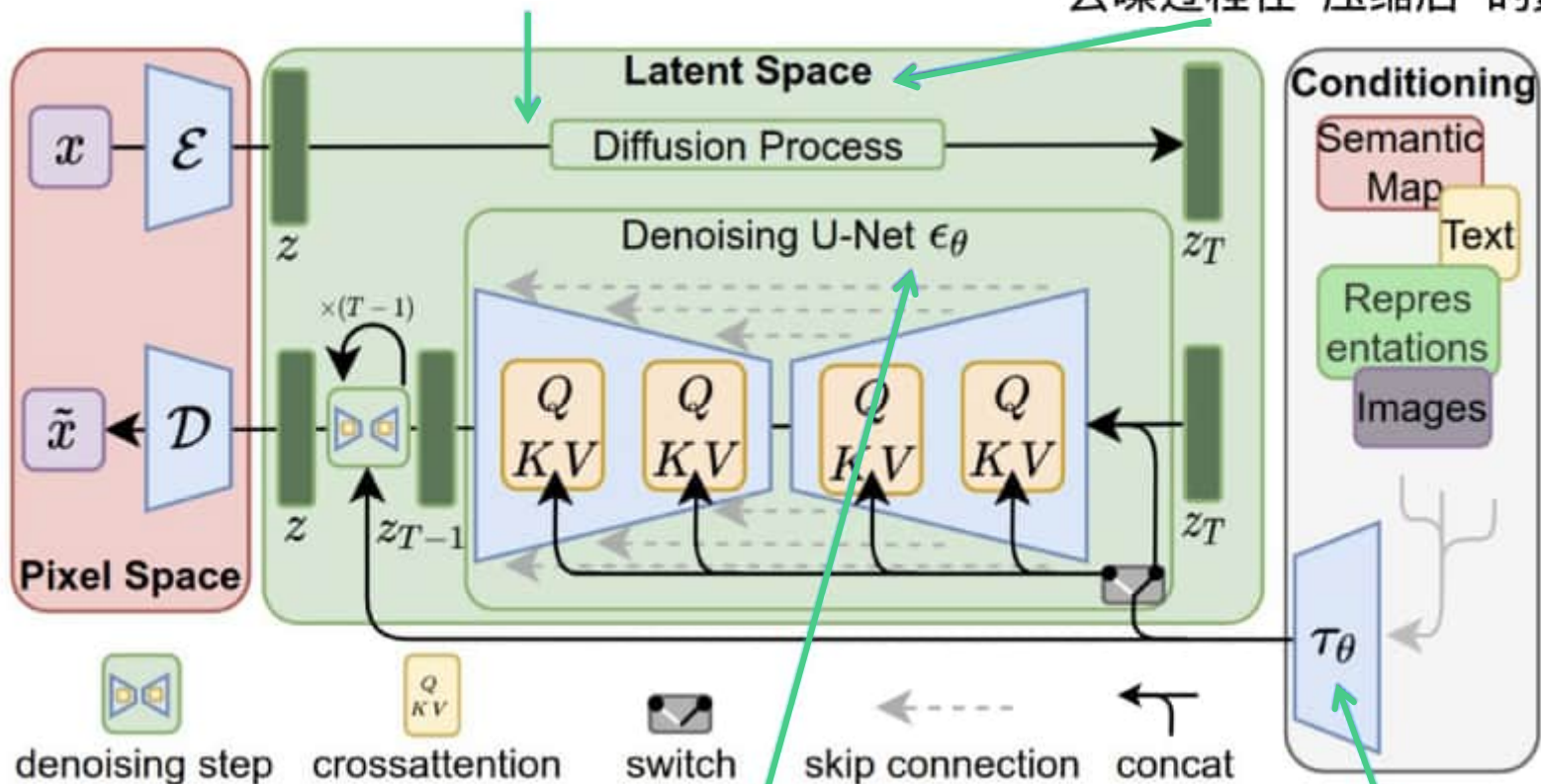
# 扩散模型· StableDiffusion

<https://arxiv.org/abs/2112.10752>

扩散过程的前向路径

Latent Space 潜空间  
去噪过程在“压缩后”的数组中进行

VAE解码  
得到结果



附加条件

Pixel Space 像素空间  
由VAE进行压缩与解码

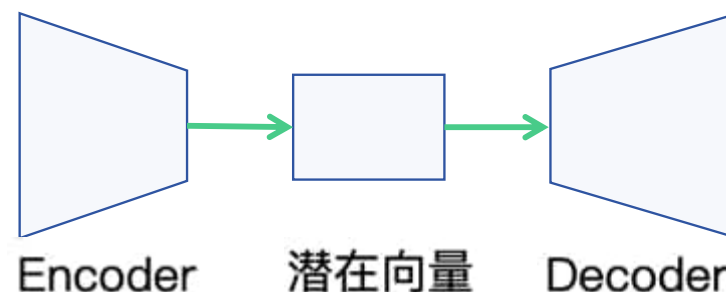
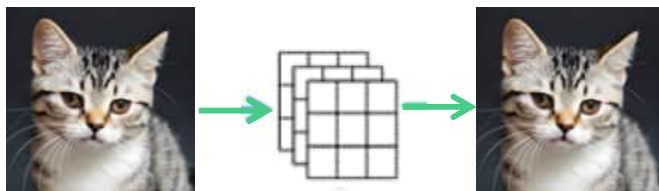
U-Net 模型负责去噪，持续该过程

编码器  
转为Embedding

## 扩散模型 · StableDiffusion 标准组件

## VAE模型

### 负责编码图像



## CLIP模型 负责编码文本

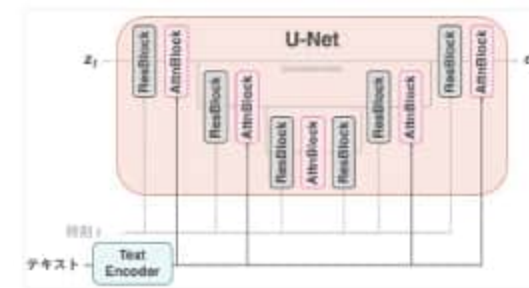
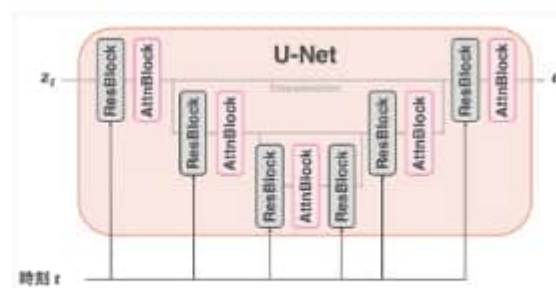
a cute cat →

```
text tokens:  
tensor([[49406,   320,   2242,   2368,  49407,  
         0,      0,      0,      0,      0,  
         0,      0,      0,      0,      0,  
         ...,  
        torch.Size([1, 77])
```

clip编码 →

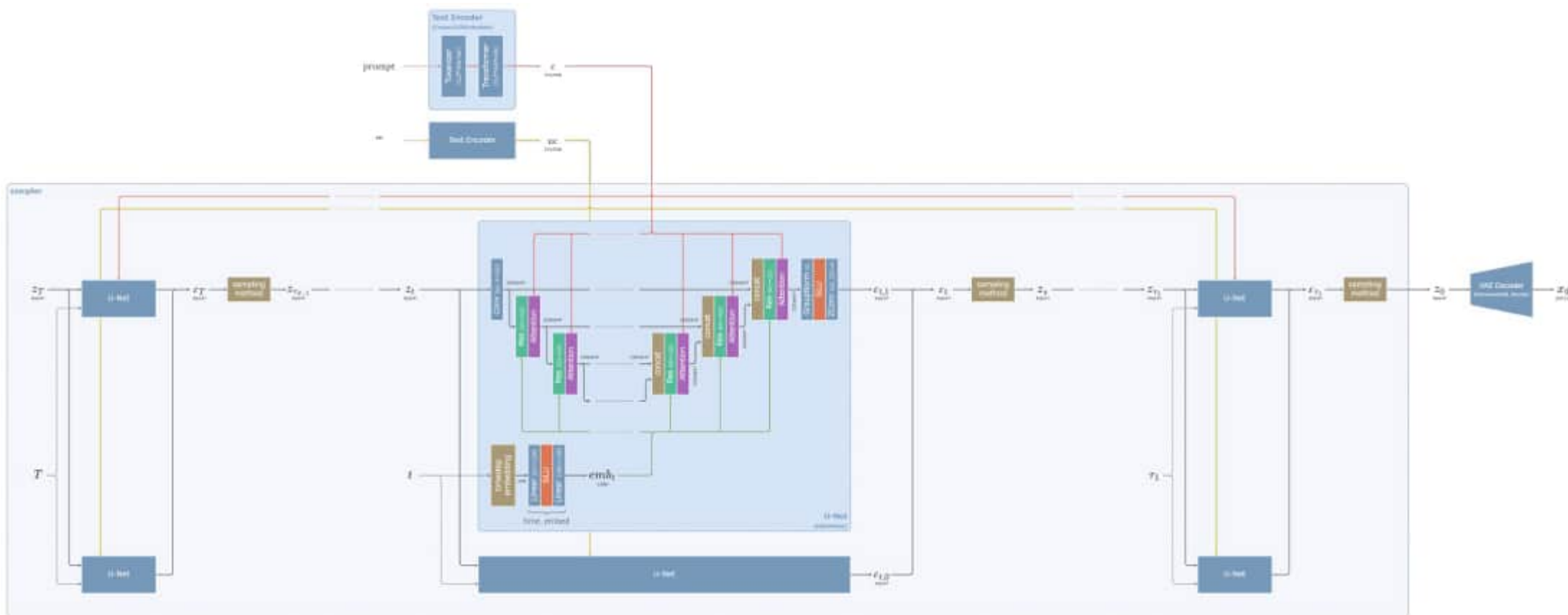
```
torch.Size([1, 77, 768])  
Vocab size: 49408
```

## Unet模型



<https://qiita.com/omiita/items/ecf8d60466c50ae8295b>

# 扩散模型· StableDiffusion 完整结构

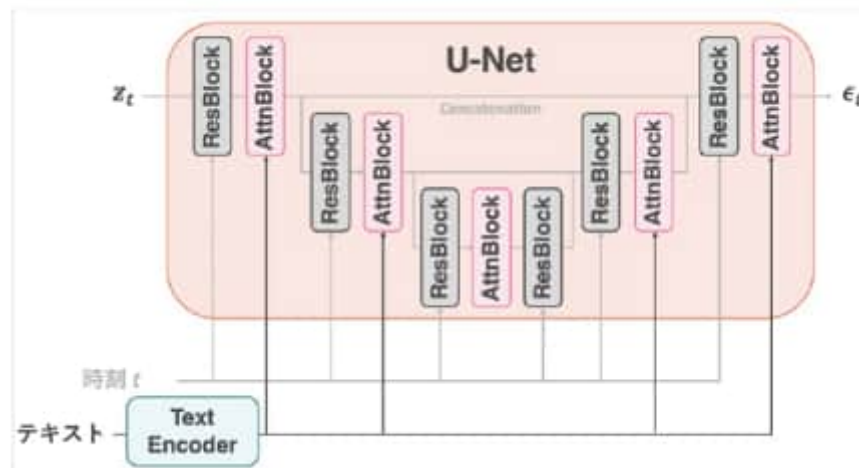


<https://henatips.com/page/47/>



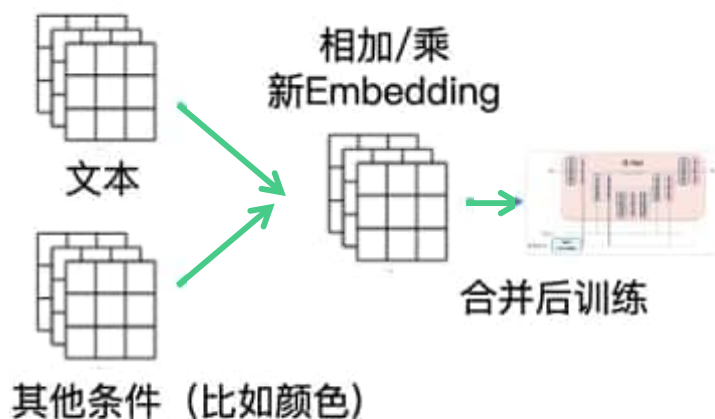
# 扩散模型·有条件的扩散

- 最原始的扩散模型只能生成随机图片
- Stablediffusion 可以生成提示词条件的图片
- 我们还能不能加入更多条件控制?



<https://qiita.com/omiita/items/ecf8d60466c50ae8295b>

## 直接合并训练

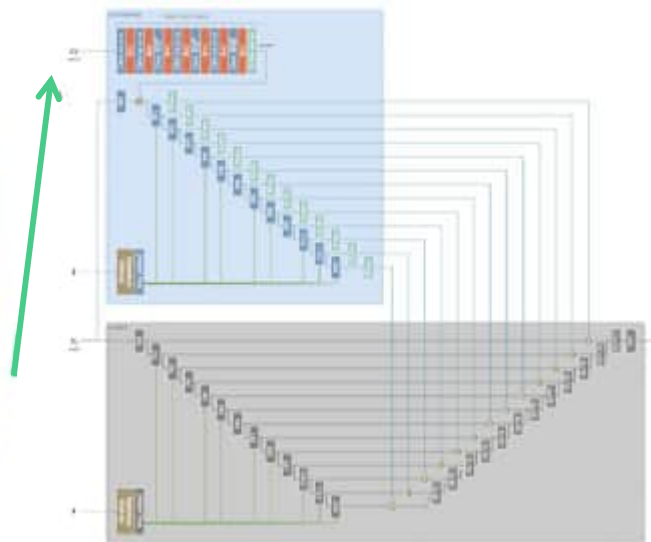


## 加入新的插件

ControlNet



<https://arxiv.org/pdf/2302.05543>



# 扩散模型·从训练看提示词有效性



帮我画一只猫 —— 这为什么不是好的提示词？

让我们一起来看一个文生图训练的数据集：

Strawberry and Watermelon Smoothies...	 <a href="https://www.joyofkosher.com/and-popsicles.jpg">https://www.joyofkosher.com/and-popsicles.jpg</a>	Saint Patricks Day Cupcakes	 <a href="https://www.lifewi">https://www.lifewi</a>
Italian Extra Virgin Olive Oil	 <a href="https://cdn.shopify.com/_jpg?v=1473916189">https://cdn.shopify.com/_jpg?v=1473916189</a>	Photo of a Pigeon Forge Cabin named Spectacula...	 <a href="https://cdn.smokym">https://cdn.smokym</a>
two strawberry shortcakes topped with...	 <a href="https://frommybowl.com/w...ee-10-scaled.jpg">https://frommybowl.com/w...ee-10-scaled.jpg</a>	Hamburg French Fries Hot Dog Ice Cream Cola...	 <a href="https://image.dhga">https://image.dhga</a>

提示词需要遵守训练集的提示规则，才能有好的生成结果



# AI绘图能力的边界

AI 一败涂地... 并不能替代人类?

基本上类似这种:

散步: 我一会就能给你出10来个

我: 一个都不能用.....

幻想中的生成结果

实际生成结果



# AI绘图能力的边界

## 绘图能力的限制原因

- 模型能力不够：模型太小 / 训练数据太少（SD一开始手画不好，到 Flux 架构逐渐成熟）



sd 1.5



sd xl



flux.1

- 提示词没写好，使用太多自然对话，而不是模型“喜欢的语言”，即与训练集相似的语言
  - ✗ 你这个颜色不够红，给我鲜艳点颜色的的苹果
  - ✓ 红色的苹果、深红色、亮

# AI绘图能力的边界·探索

💡 市面上有很多绘画模型，怎么才能知道最好的提示词是什么？

生成本质 —— 猜测“训练集”的提示词分布 / 猜测训练后效果好的提示分布

- 建立属于自己的常见测试提示词库
- 优先选择作者推荐的提示词，或任何与训练集提示相似的风格
- 找到几张效果较好的图片参数 / 提示词，围绕该提示词与参数的风格拓展试错
- 大量的遍历 + 勇敢的尝试

**AI generated images in ads:**



**AI generated images from my prompts:**



# 小结

- 文本、图片、音乐等输入在计算机中表现为数组，可以用低维度的数组表达高维信息。
- 所谓“模型”在计算机中，也只是表达为数组，在训练过程中会被不断修改。
- 通过构建数据输入与标签、随机权重模型，经过反向传播优化参数后，即可得到训练后的模型。
- 为了更好的模型效果，模型会越来越大，越来越深，训练数据会越来越多。
- 扩散模型是生成式模型，与上述预测标签的模型不同，它通过预测噪声，去除噪声得到图片。
- 扩散模型需要 Clip 模型编码文本输入、Unet 结合文本去噪、VAE 解码得到最后的生成图片
- 为了生成好的图片，我们需要让提示词尽量与模型的训练集一致，或使用官方 / 别人推荐的提示。



## 试试看这些绘图工具

即梦 / 豆包

<https://jimeng.jianying.com/ai-tool/home>

<https://www.doubao.com/chat/>



可灵

<https://klingai.kuaishou.com/>



midjourney

<https://www.midjourney.com/>



## Ideogram

<https://ideogram.ai/>



# 试试看这些绘图工具

## Reve

<https://preview.reve.art/>



## LiblibAI

<https://www.liblib.art/>

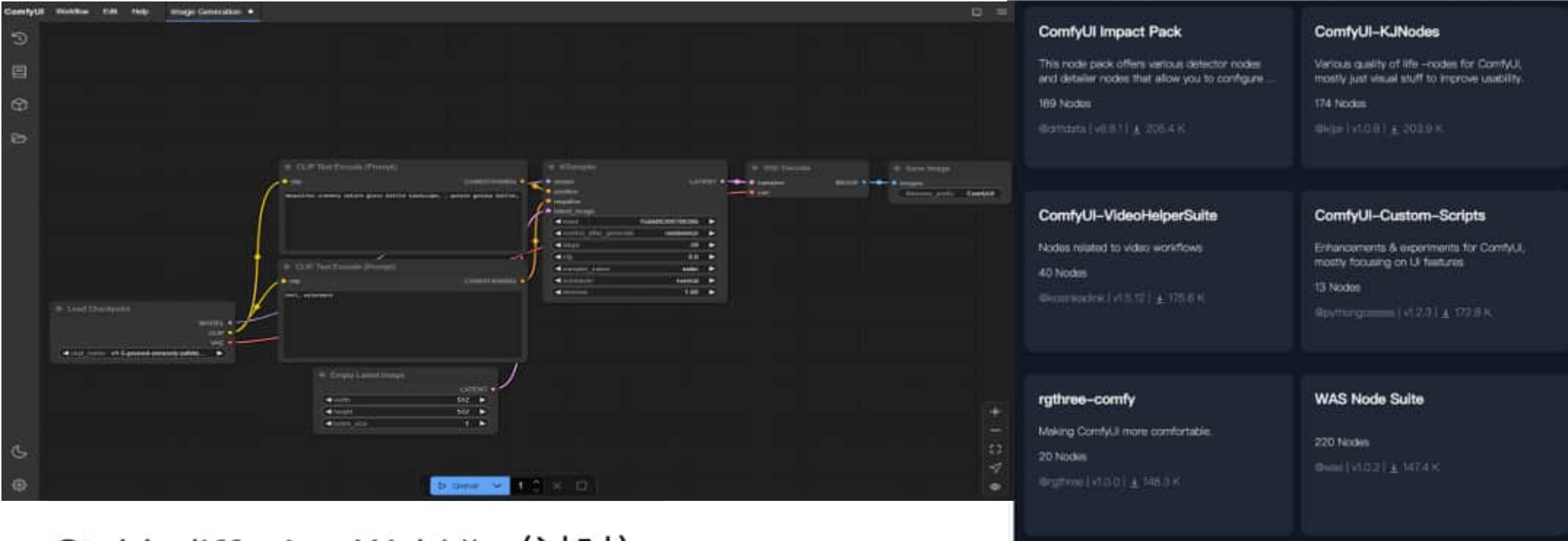




# 试试看这些绘图工具

ComfyUI 最丰富的插件与社区，良好的用户体验，可以使用任意你想实现的插件增强生成效果

<https://github.com/comfyanonymous/ComfyUI> [https://docs.comfy.org/get\\_started/introduction](https://docs.comfy.org/get_started/introduction)



The image displays the ComfyUI interface, a node-based workflow editor for AI image generation. The main workspace shows a workflow with nodes such as 'CLIP Text Encoder (Prompt)', 'VAE', 'Image Encoder', and 'Image Decoder'. The interface includes a sidebar on the right with a list of installed node packs and their details.

Node Pack	Nodes	Version	Size
ComfyUI Impact Pack	169	@bottdata   v0.8.1	± 205.4 K
ComfyUI-KJNodes	174	@kijai   v1.0.8	± 203.9 K
ComfyUI-VideoHelperSuite	40	@suzuki-kazuki   v1.5.12	± 175.6 K
ComfyUI-Custom-Scripts	13	@pythongames   v1.2.3	± 172.8 K
rgthree-comfy	20	@rgthree   v1.0.0	± 148.8 K
WAS Node Suite	220	@was   v1.0.2	± 147.4 K









Stablediffusion WebUI (过时)

# 获取最好的图像生成模型

## 图像生成模型竞技场

Subject:

All Text & Typography Commercial People: Portraits People: Groups & Activities Fantasy & Mythical Nature & Landscapes Futuristic & Sci-Fi UI/UX Design Physical Spaces

CREATOR	NAME	ARENA ELO	# APPEARANCES
	Halfmoon	1163	2,900
 Recraft AI	Recraft V3	1141	111,610
 Google	Imagen 3 (v002)	1114	18,253
 Black Forest Labs	FLUX1.1 [pro]	1113	129,498
 Black Forest Labs	FLUX.1 [pro]	1097	139,672
 MiniMax	Image-01	1092	3,106
 Midjourney	Midjourney v6.1	1079	137,673
 Black Forest Labs	FLUX.1 [dev]	1076	137,984

Rank	Model	Arena Elo	95% CI	Votes
1	<a href="#">FLUX.1-dev</a>	1134	+31/-30	352
2	<a href="#">PlayGround V2.5</a>	1117	+19/-17	1082
3	<a href="#">FLUX.1-schnell</a>	1091	+31/-29	365
4	<a href="#">PlayGround V2</a>	1072	+16/-21	1050
5	<a href="#">Kolores</a>	1063	+29/-28	285
6	<a href="#">StableCascade</a>	1043	+16/-22	1065
7	<a href="#">HunyuanDiT</a>	1022	+24/-20	415
8	<a href="#">PixArtAlpha</a>	1019	+15/-17	1538
9	<a href="#">PixArtSigma</a>	1017	+22/-21	795

<https://huggingface.co/spaces/TIGER-Lab/GenAI-Arena>

<https://huggingface.co/spaces/ArtificialAnalysis/Text-to-Image-Leaderboard>

# Thanks

散步 sanbuphy

Datawhale

physicoad@gmail.com

## 课后作业

1. 用自己的话描述 AI 学习的原理，diffusion 模型的原理及其特点。
2. 用 liblibai 尝试 webui 与 comfyui 的几个 workflow 生成图片。
3. 思考产品海报可能需要那些元素、表达哪些概念。