计算机视觉的前沿热点

叶山 中国地质大学（北京）

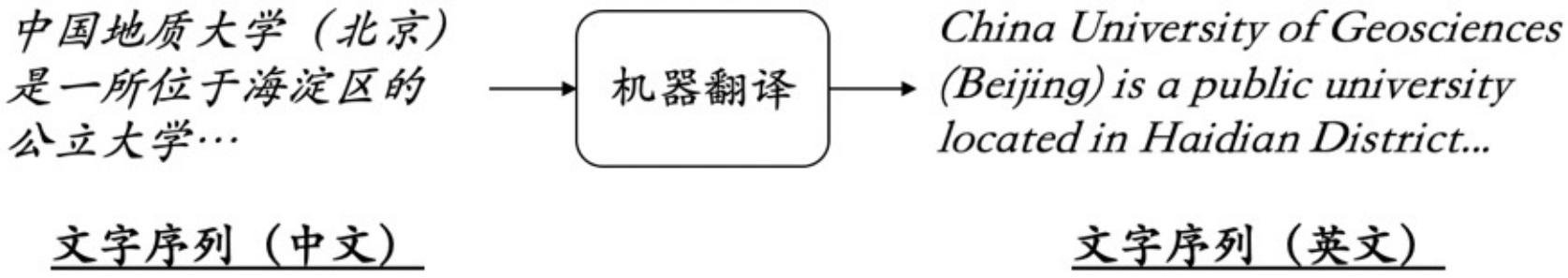
yes@cugb.edu.cn

本教案部分参考以下课程：斯坦福大学 CS 231、密歇根大学 EECS 442 & 498、威斯康辛大学麦迪逊分校 CS 766

Vision Transformer (ViT)

Transformer的来历

- Vision Transformer脱胎于Transformer模型。
- Transformer模型兴起于自然语言处理（NLP）领域，最初由Google Brain团队在2017年提出，旨在解决机器翻译任务中的“序列到序列”（Seq2Seq）问题。



机器翻译的挑战

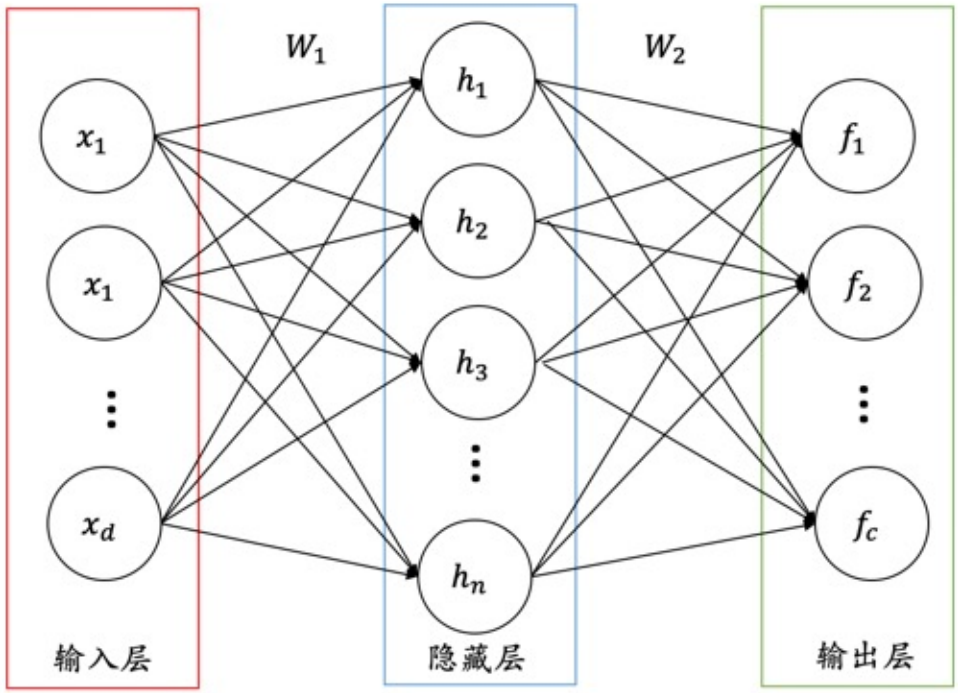
机器翻译：无人参与的前提下，由计算机将文本从一种语言翻译成另一种语言。

挑战：词语的顺序会影响句子的意思。

上海自来水
水来自海上

I am leaving Beijing for Tianjin.
I am leaving Tianjin for Beijing.

循环神经网络 RNN



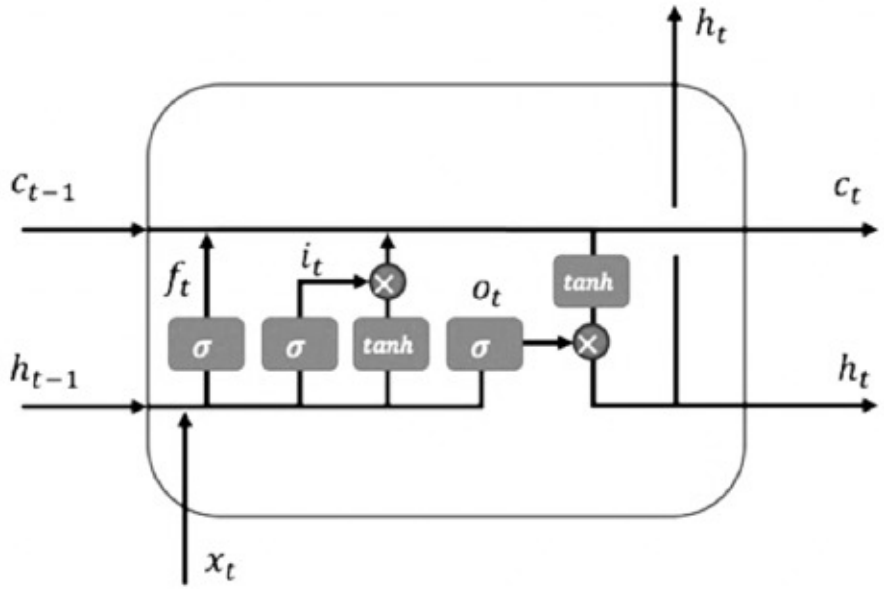
下次输入数据时
需考虑前序的状态

隐状态
时间越久远，对当前的影响就越小

更新当前的网络状态

循环神经网络 RNN

- RNN的问题：对于长句子的效果不佳。
- 解决方案：长短时记忆网络（Long Short-Term Memory, LSTM）



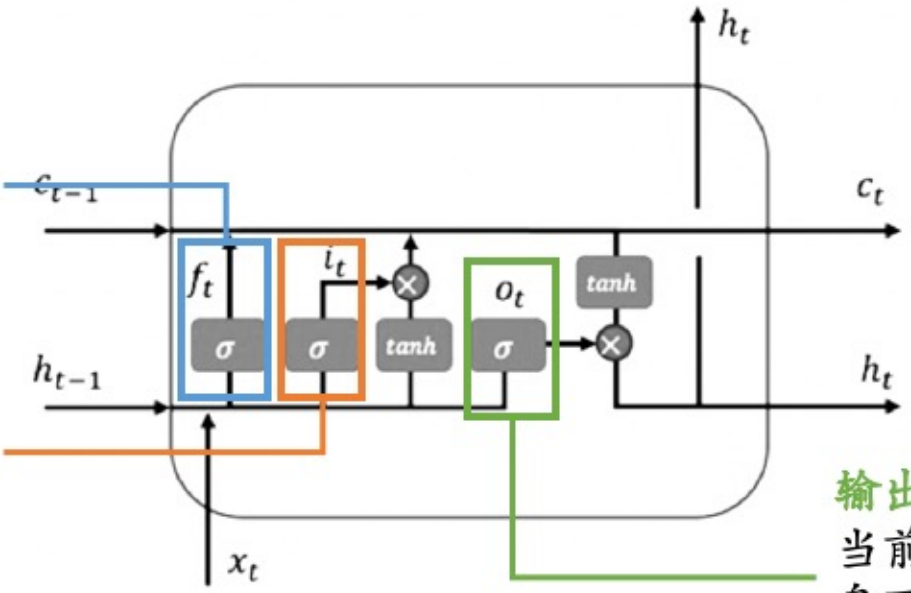
长短时记忆网络

对“隐状态”添加“门”（gate），用于控制信息的流动。

- 门的取值范围：[0, 1]
- 当门取1时，完全“打开”；当门取0时，完全“关闭”。

遗忘门 forget gate
上一时刻的单元状态有多少信息需要被“遗忘”？

输入门 input gate
当前输入的信息有多少可以被加入到状态控制器中？



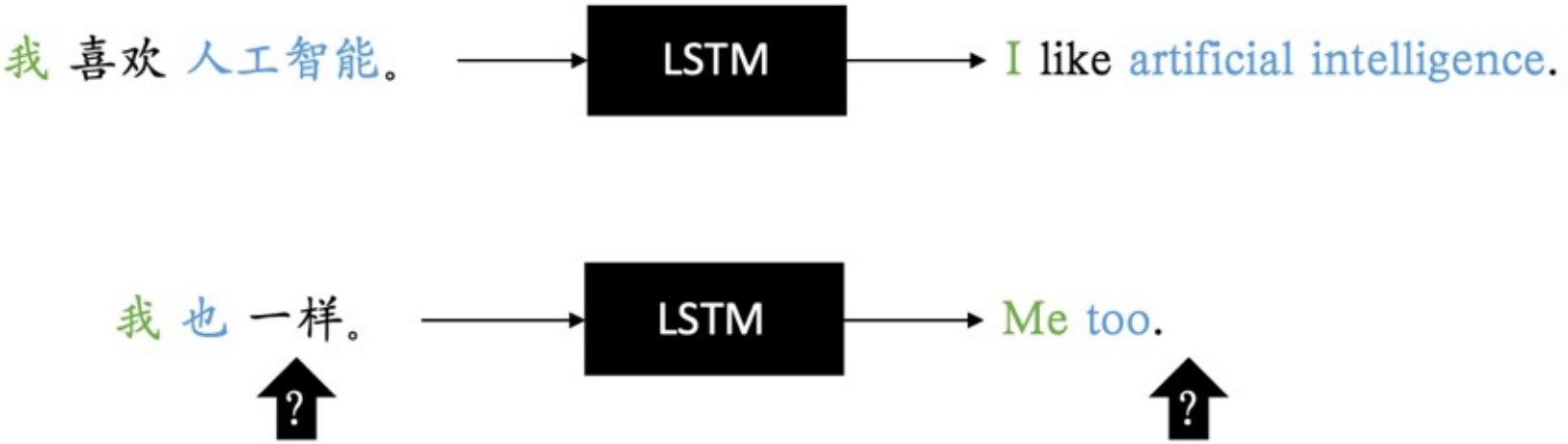
输出门 output gate
当前状态中有多少信息可以作为当前时刻的输出？

长短时记忆网络

优点：加入了有关词语顺序的信息。

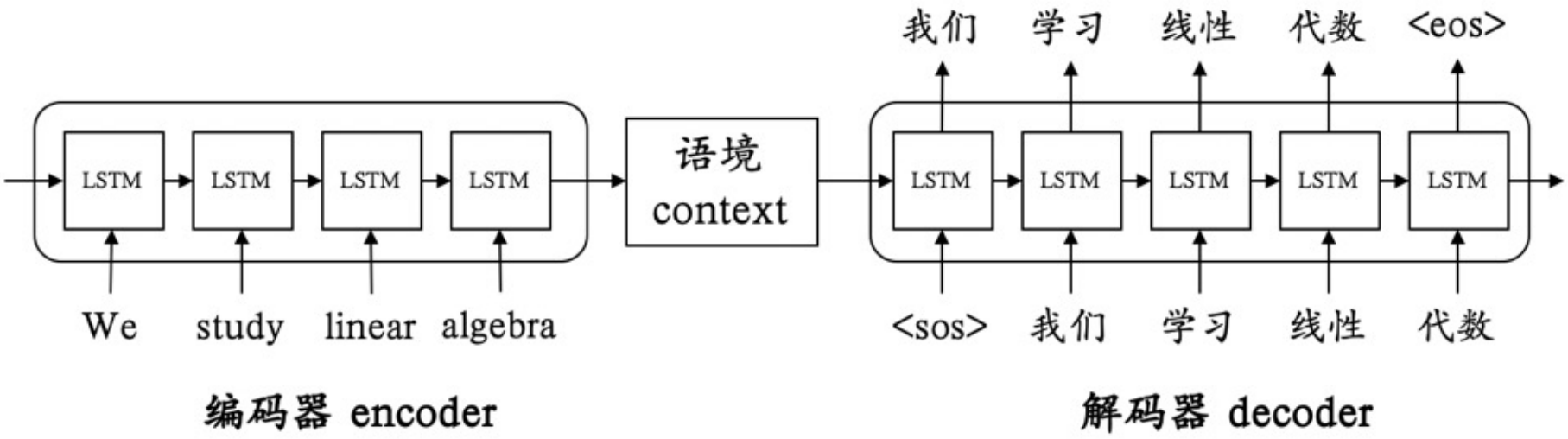
缺点：难以应付输入句和输出句词数不同的情况。

（机器翻译、对话机器人等任务的常见现象）



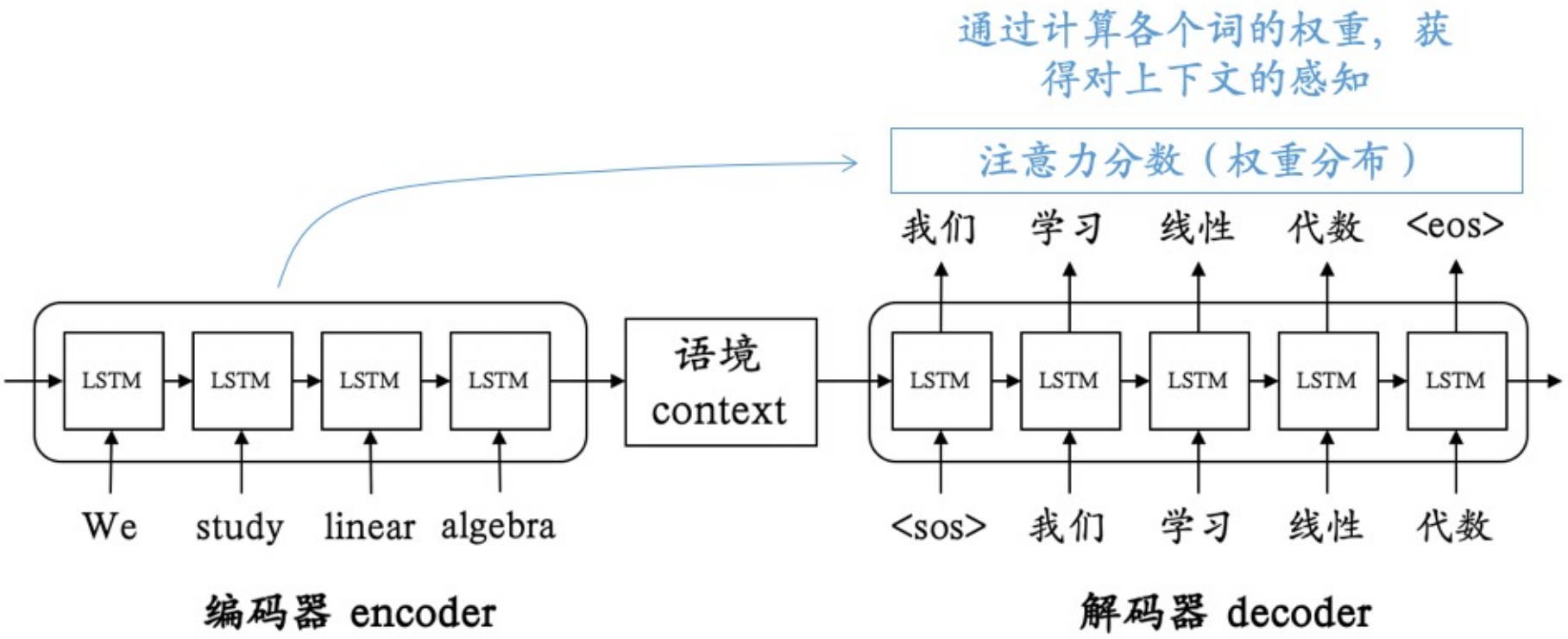
Seq2Seq

- 序列到序列（Sequence to sequence）

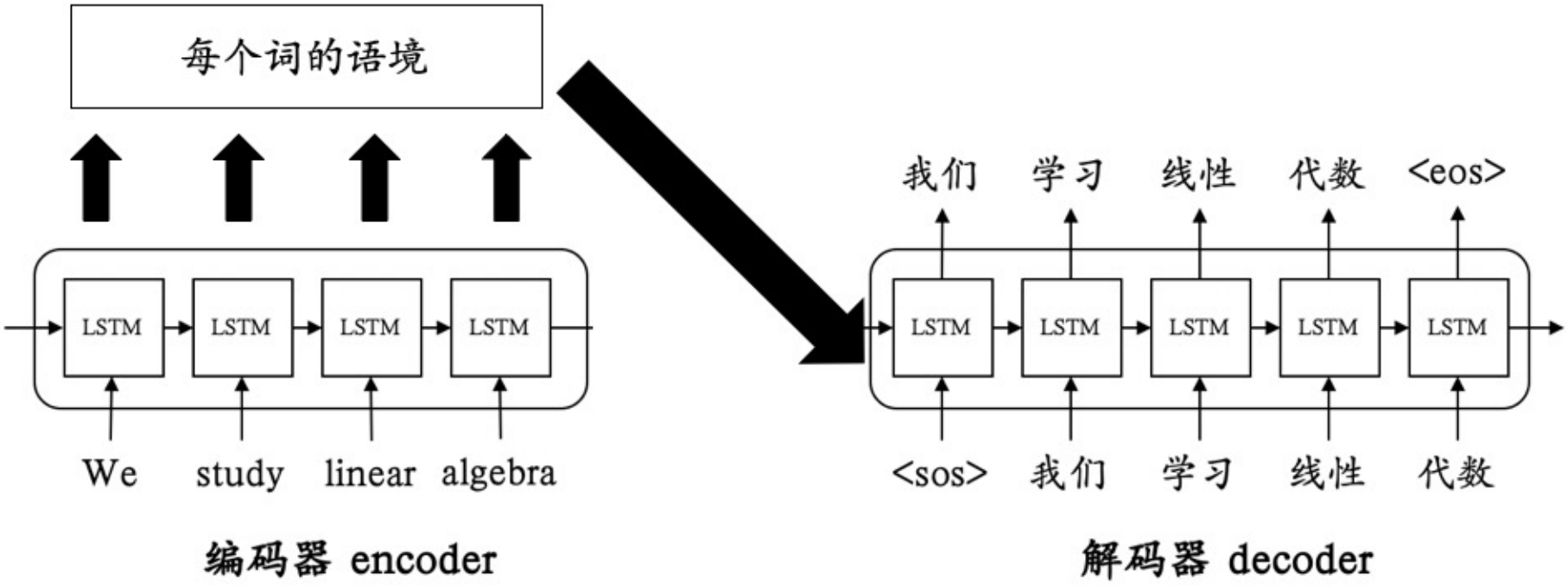


sos: start of sequence
eos: end of sequence

注意力机制 Attention



自注意力 Self-Attention



Transformer

Transformer

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

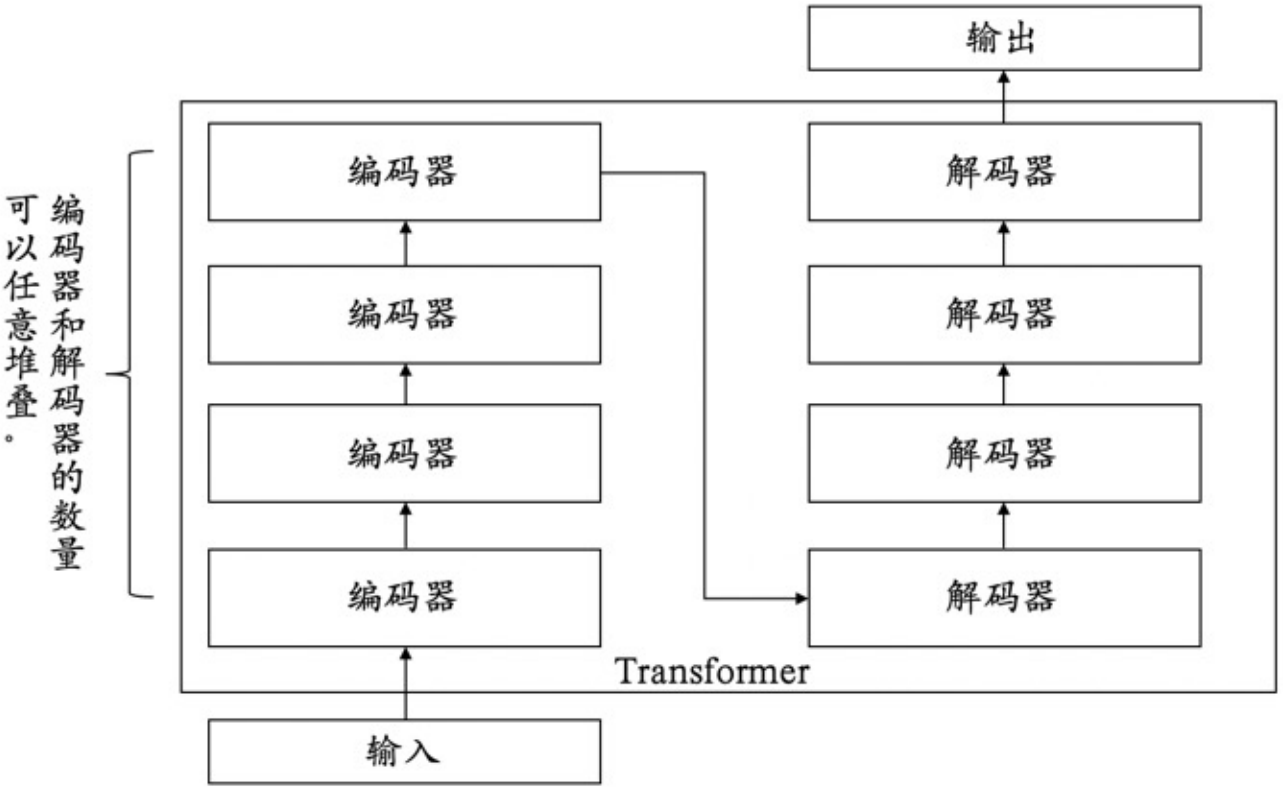
Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

Love, Love, Love. All you need is love. Love is all you need.

— John Lennon —

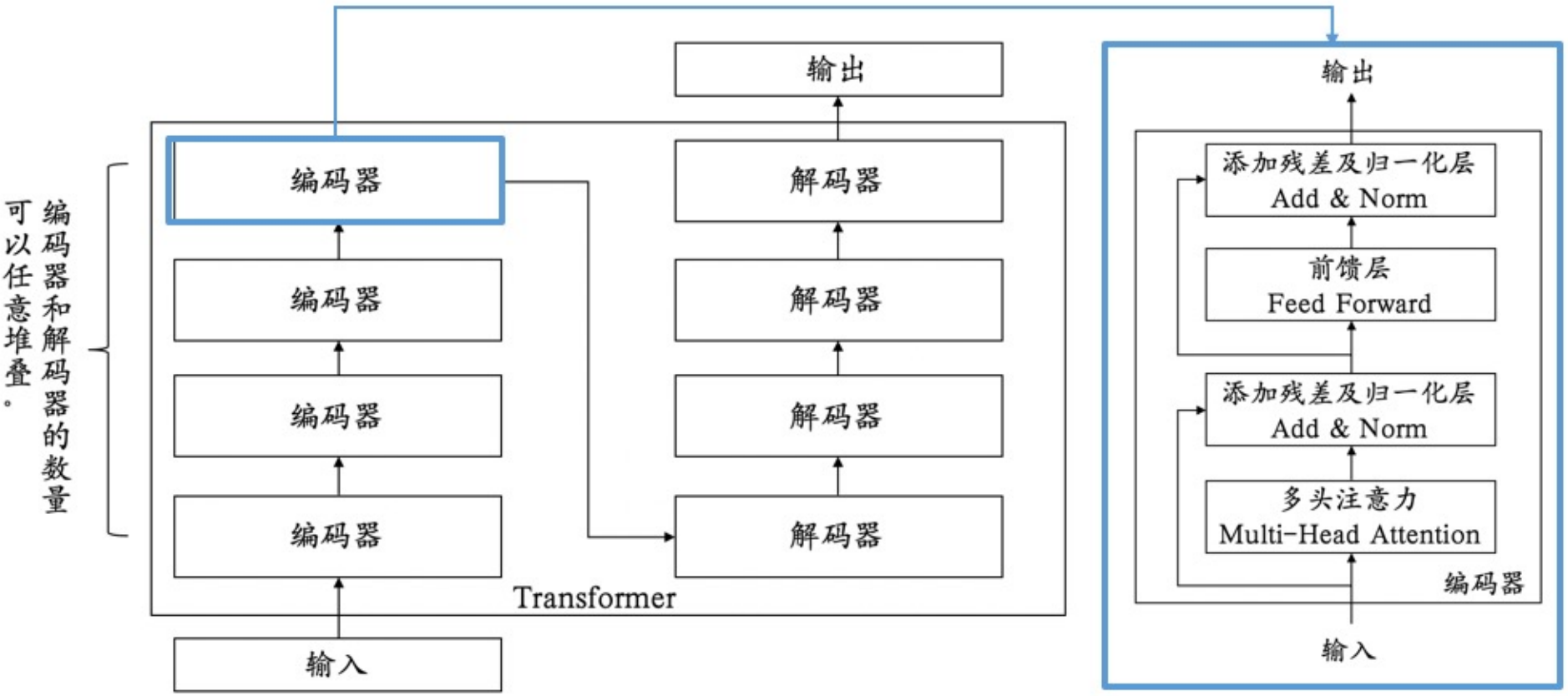
Transformer



“变形金刚”

将输入进行“拆解”，再“重组”成输出。

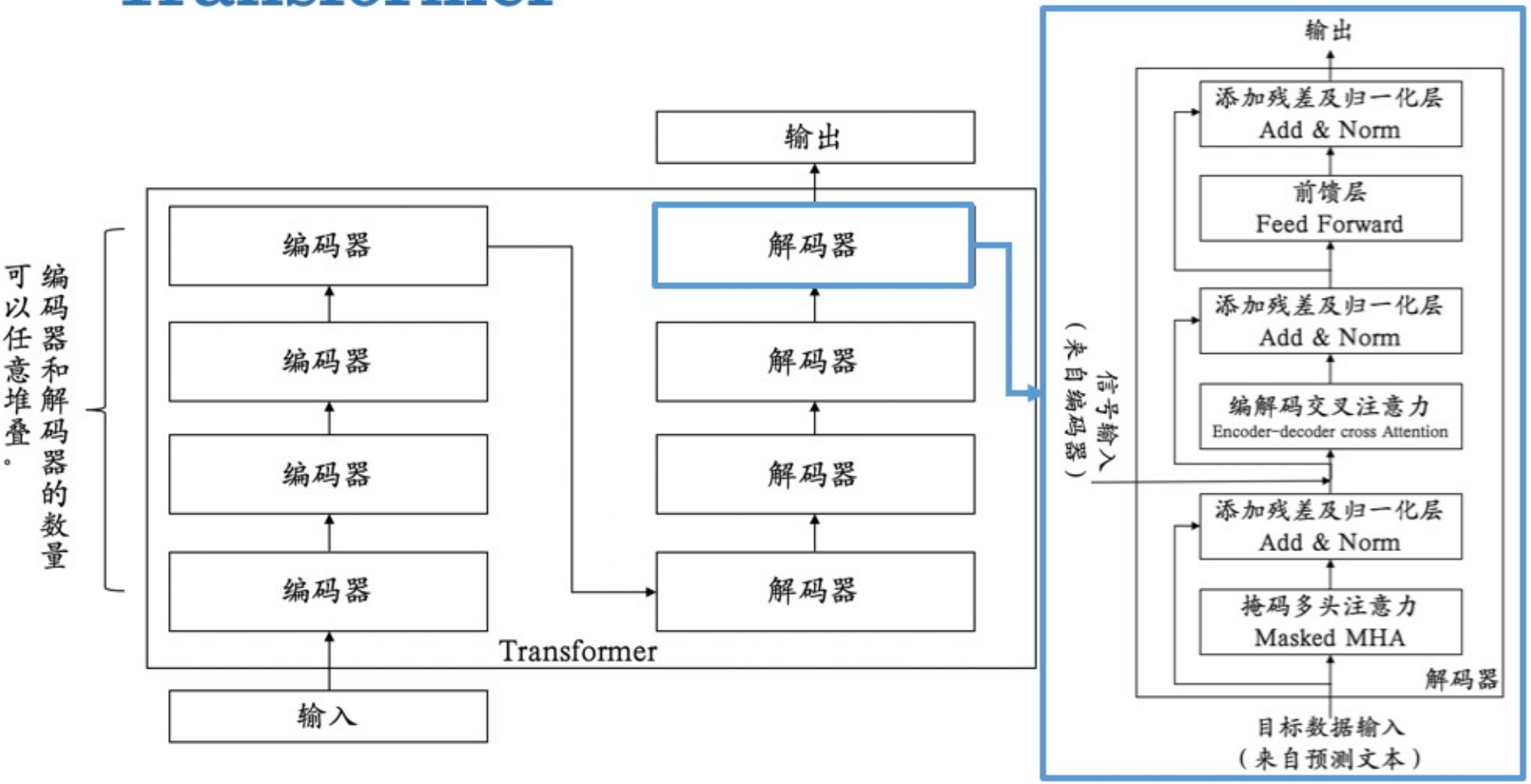
Transformer



多头注意力：基于自注意力的改进方法。它通过并行地运行多个独立的注意力机制，来捕获输入序列的不同子空间的注意力分布。

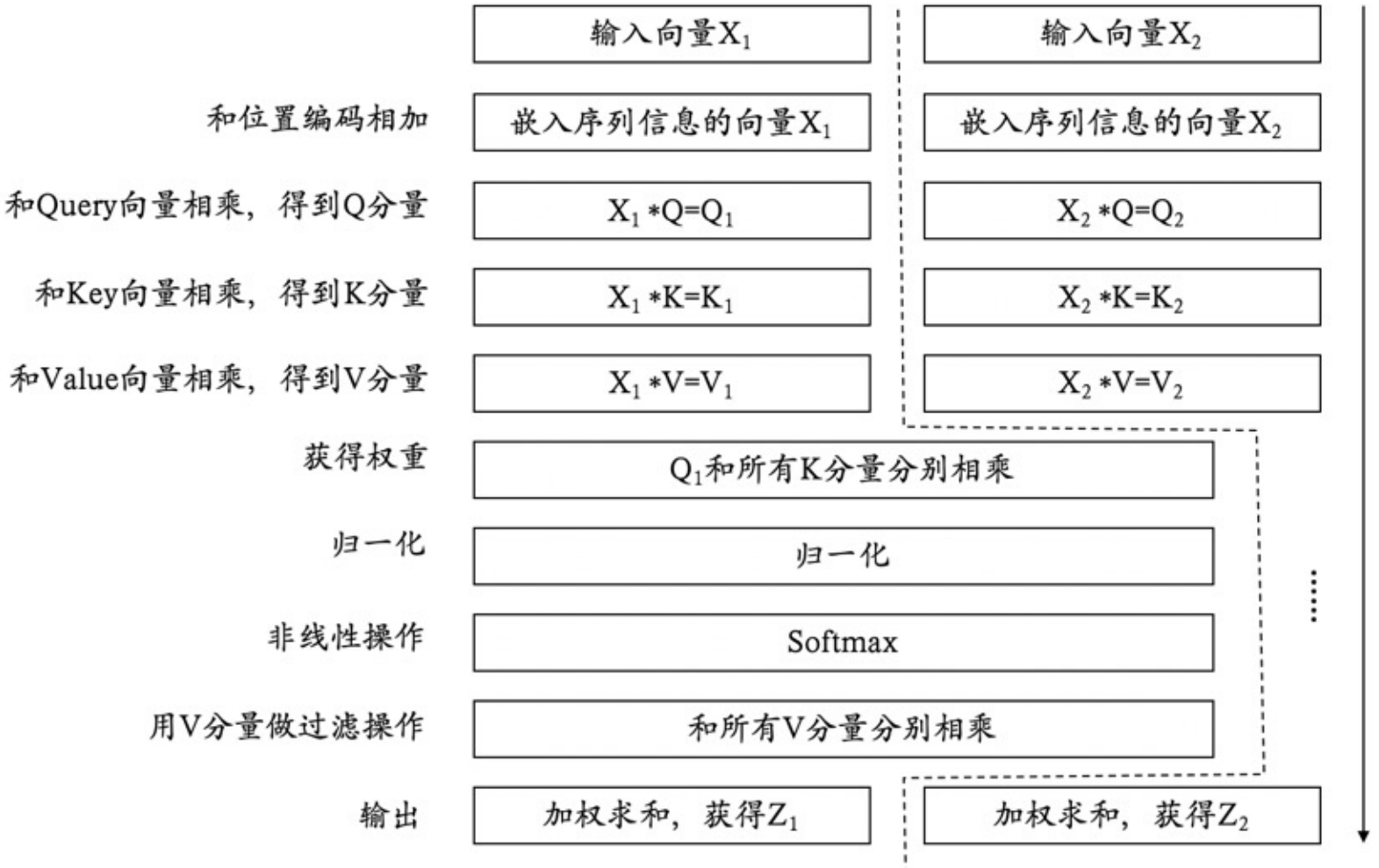
在编码器中，多头注意力层相当于对输入信号的“拆解说明书”，而前馈层负责执行拆解工作。

Transformer



编解码交叉注意力层：负责关注上下文信息。

注意力的计算


输入向量X_1
输入向量X_2
和位置编码相加
嵌入序列信息的向量X_1
嵌入序列信息的向量X_2
和Query向量相乘，得到Q分量
$X_1 * Q = Q_1$
$X_2 * Q = Q_2$
和Key向量相乘，得到K分量
$X_1 * K = K_1$
$X_2 * K = K_2$
和Value向量相乘，得到V分量
$X_1 * V = V_1$
$X_2 * V = V_2$
获得权重
Q_1和所有K分量分别相乘
归一化
归一化
非线性操作
Softmax
用V分量做过滤操作
和所有V分量分别相乘
输出
加权求和，获得Z_1
加权求和，获得Z_2

Vision Transformer (ViT)

Published as a conference paper at ICLR 2021

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy[*,†], **Lucas Beyer**[*], **Alexander Kolesnikov**[*], **Dirk Weissenborn**[*], **Xiaohua Zhai**[*], **Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby**[*,†]

[*]equal technical contribution, [†]equal advising

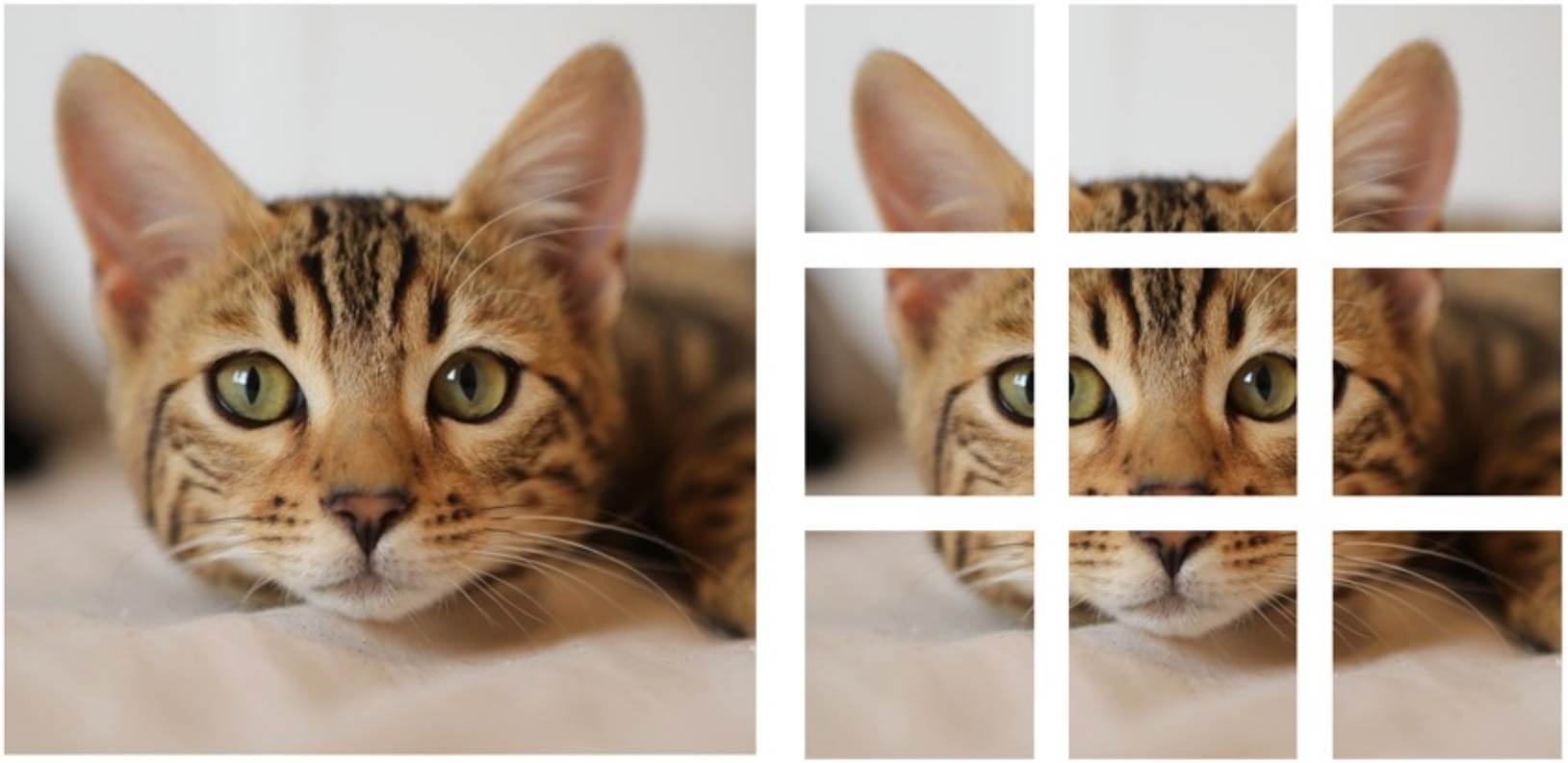
Google Research, Brain Team

{adosovitskiy, neilhoulsby}@google.com

ABSTRACT

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.[1]

ViT 的图像分类操作



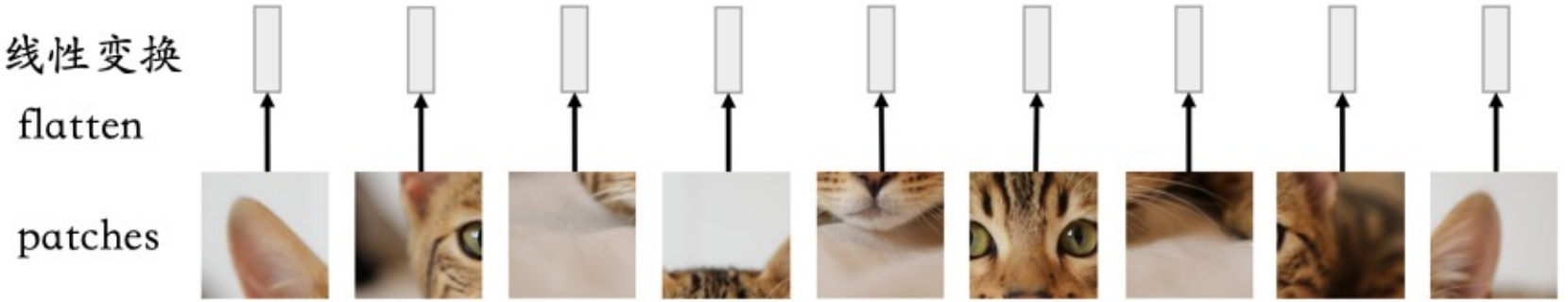
把图片分割成若干个区块（patch）。

ViT 的图像分类操作

对每一个区块分别作flatten

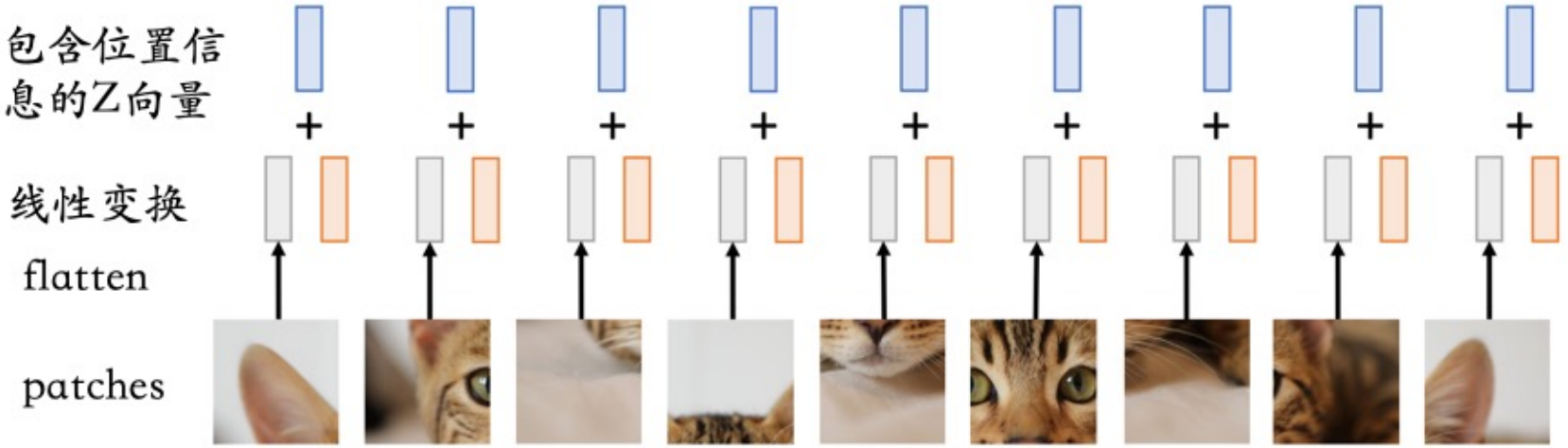
如果有N个区块，则会得到N个向量，记为X_1，X_2，... X_N

对X_1，X_2，... X_N做线性变换，得到Z_1，Z_2，... Z_N



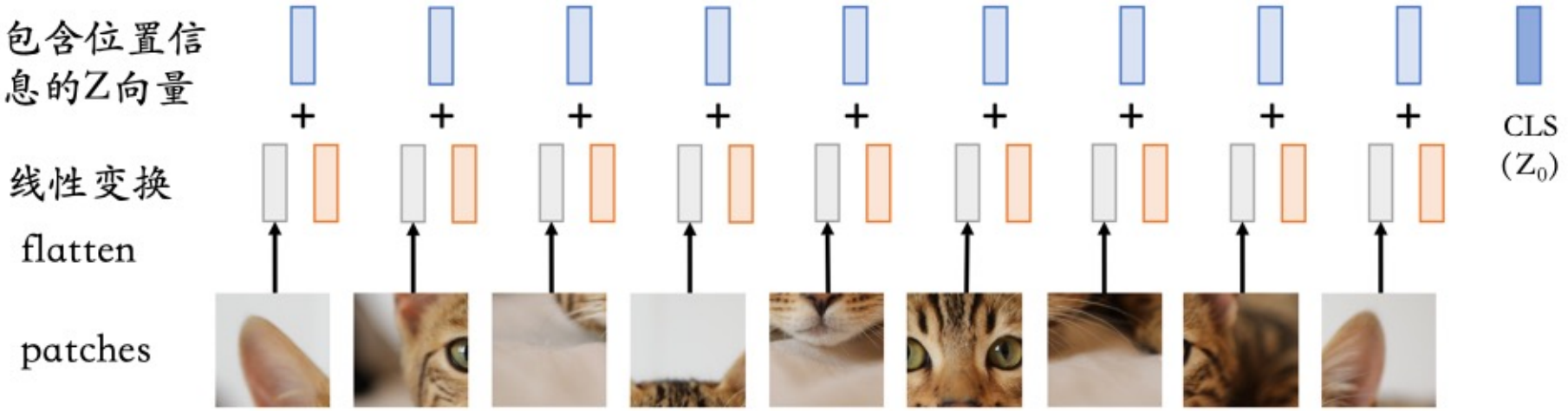
ViT 的图像分类操作

让 Z_1，Z_2，$\cdots Z_N$和位置编码向量相加，使得Z_1，Z_2，$\cdots Z_N$包含位置信息



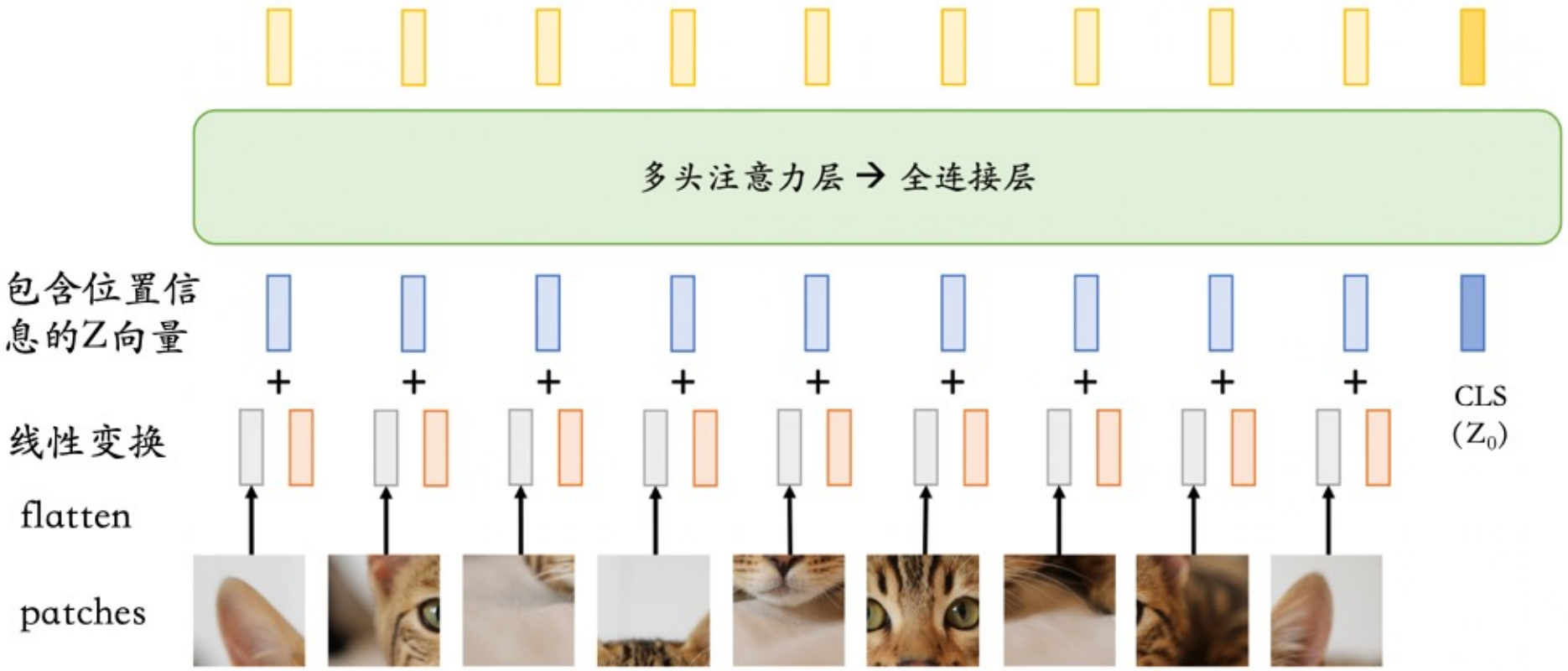
ViT 的图像分类操作

添加“分类器占位符”CLS，获得Z_0向量



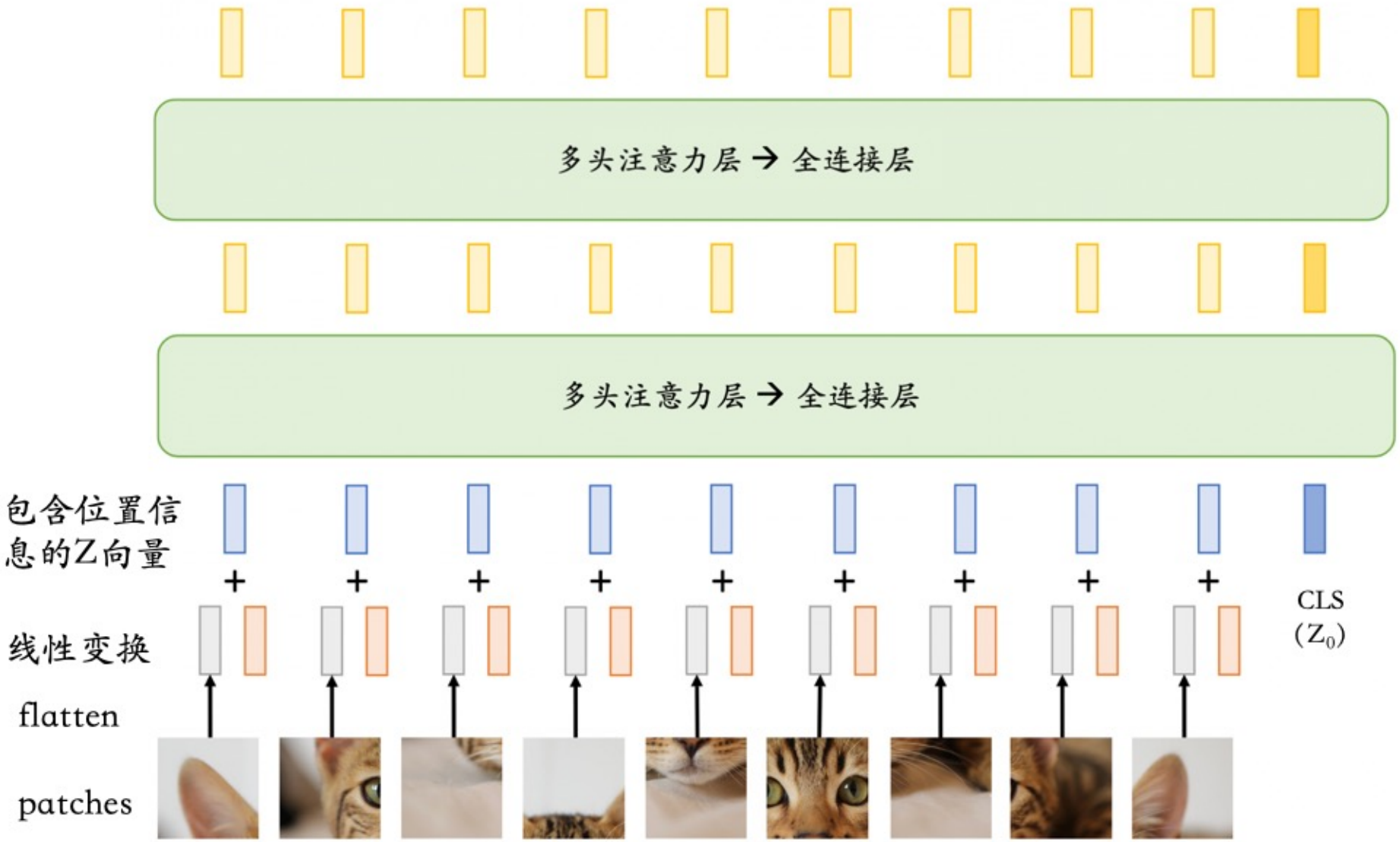
ViT 的图像分类操作

将所有Z向量传入多头注意力层，以及一个前馈网络（全连接）层。



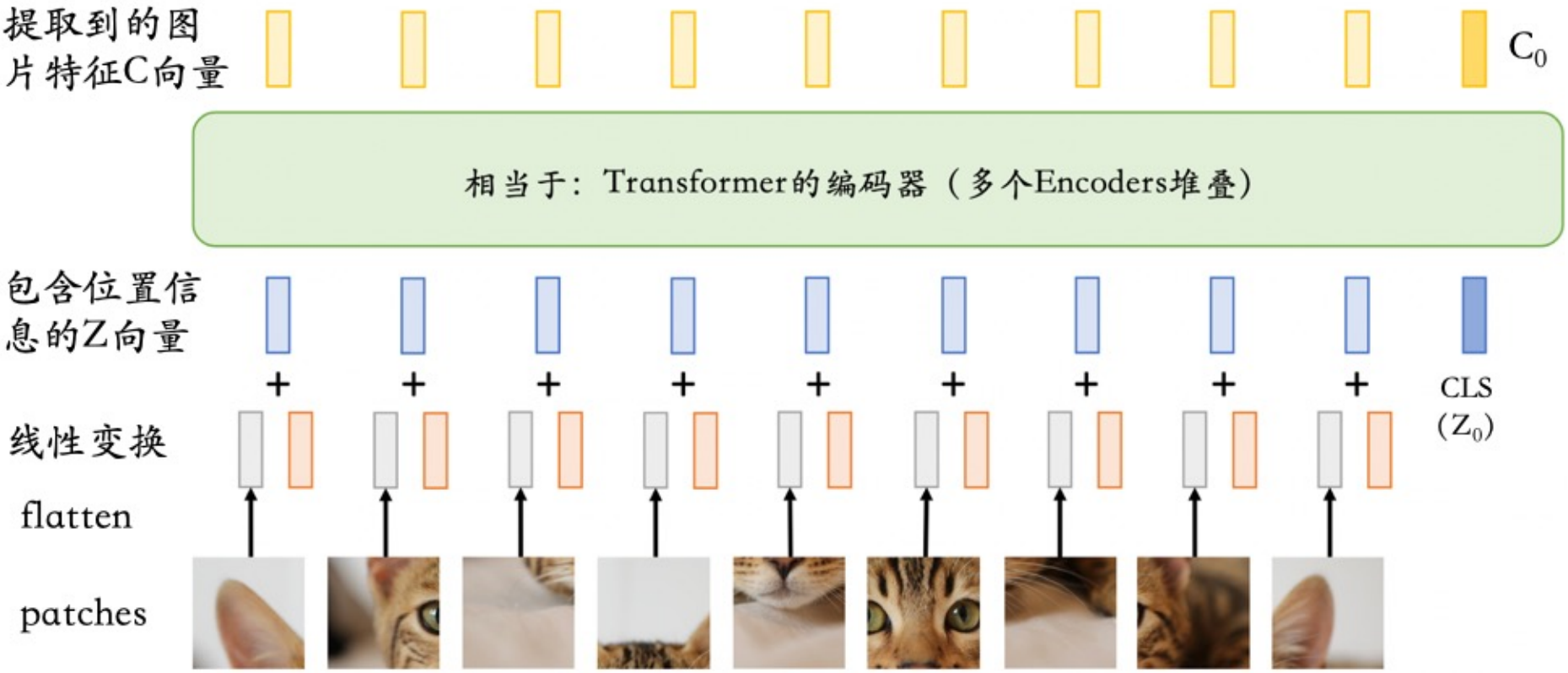
ViT 的图像分类操作

多头注意力层和全连接层的个数可以任意叠加。

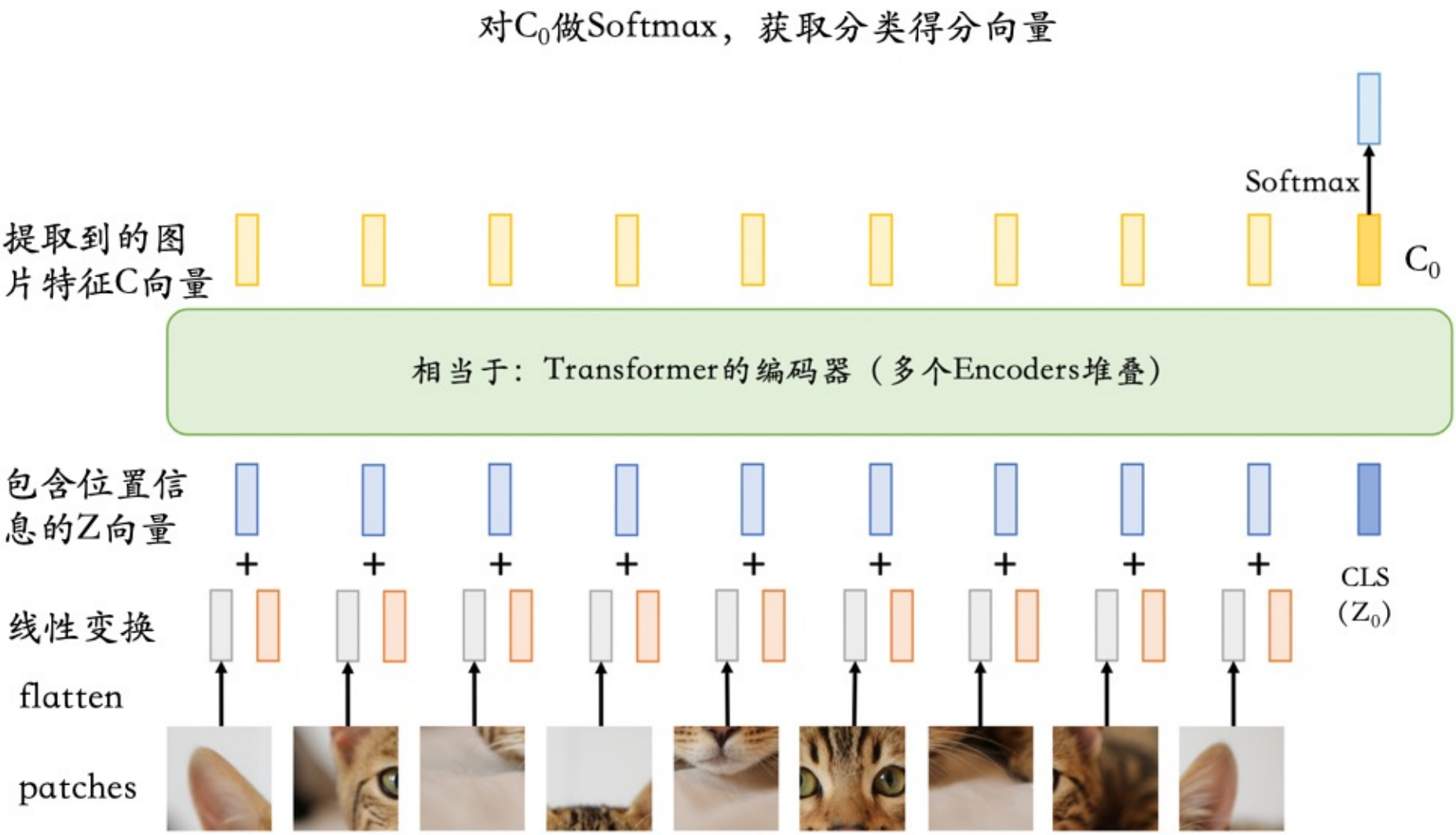


ViT 的图像分类操作

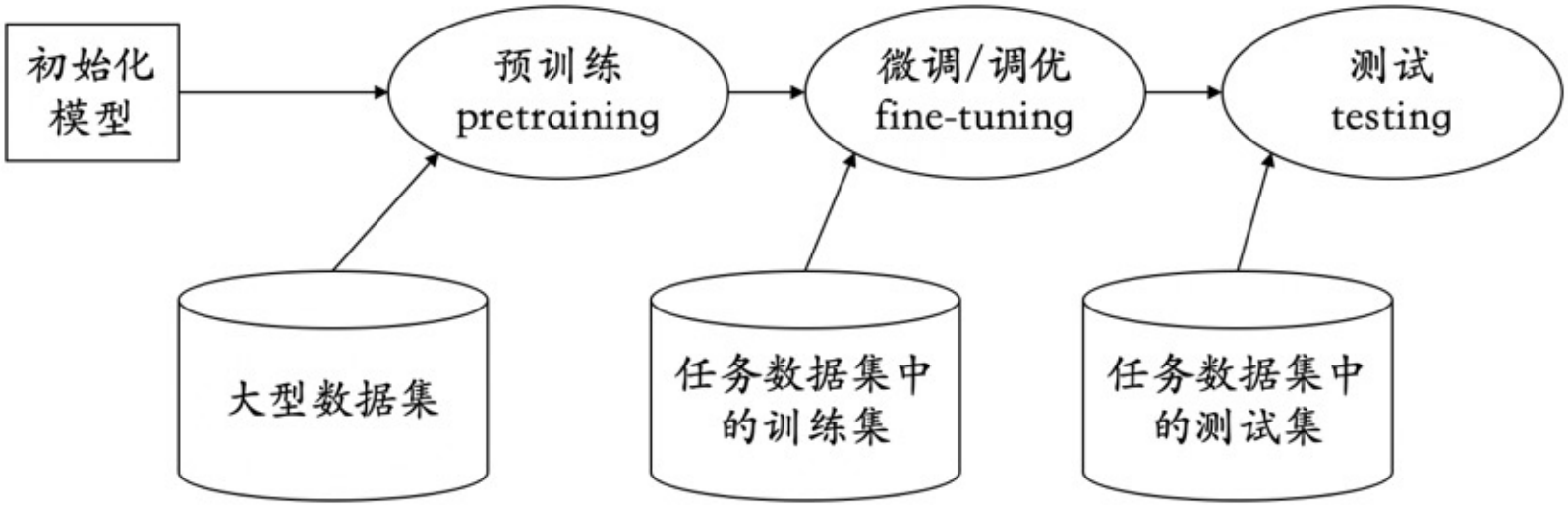
任意搭建Encoders结构，最终输出N+1个向量，记为 C_1，C_2，… C_N，以及C_0



ViT 的图像分类操作


对C_0做Softmax，获取分类得分向量
Softmax
C_0
提取到的图片特征C向量
相当于：Transformer的编码器（多个Encoders堆叠）
包含位置信息的Z向量
CLS (Z_0)
线性变换
flatten
patches

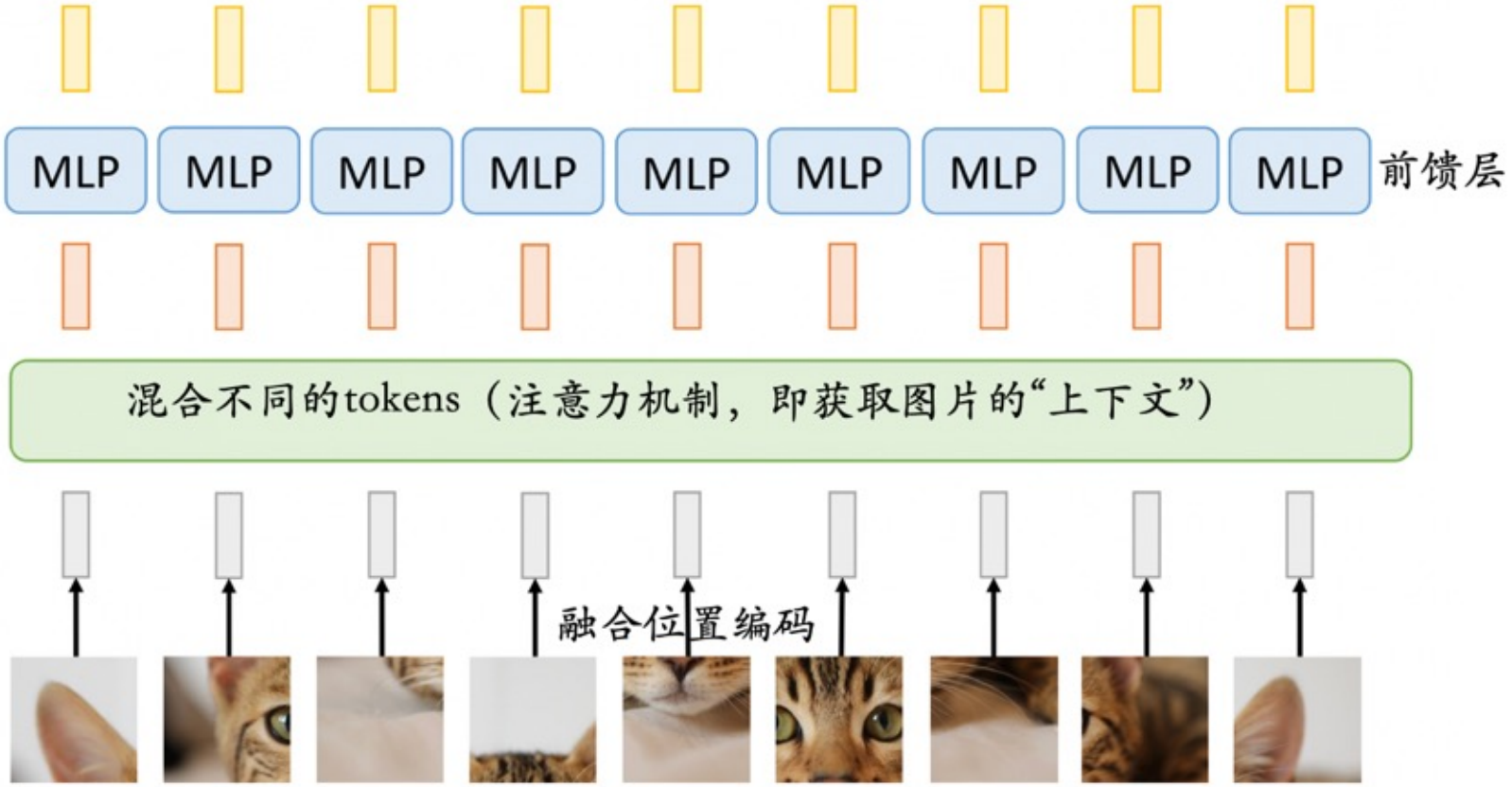
ViT的训练



JFT 300M（谷歌的内部数据集）
ImageNet-21k

ImageNet
CIFAR-10
CIFAR-100
Oxford-IIIT Pets
Oxford Flowers-102

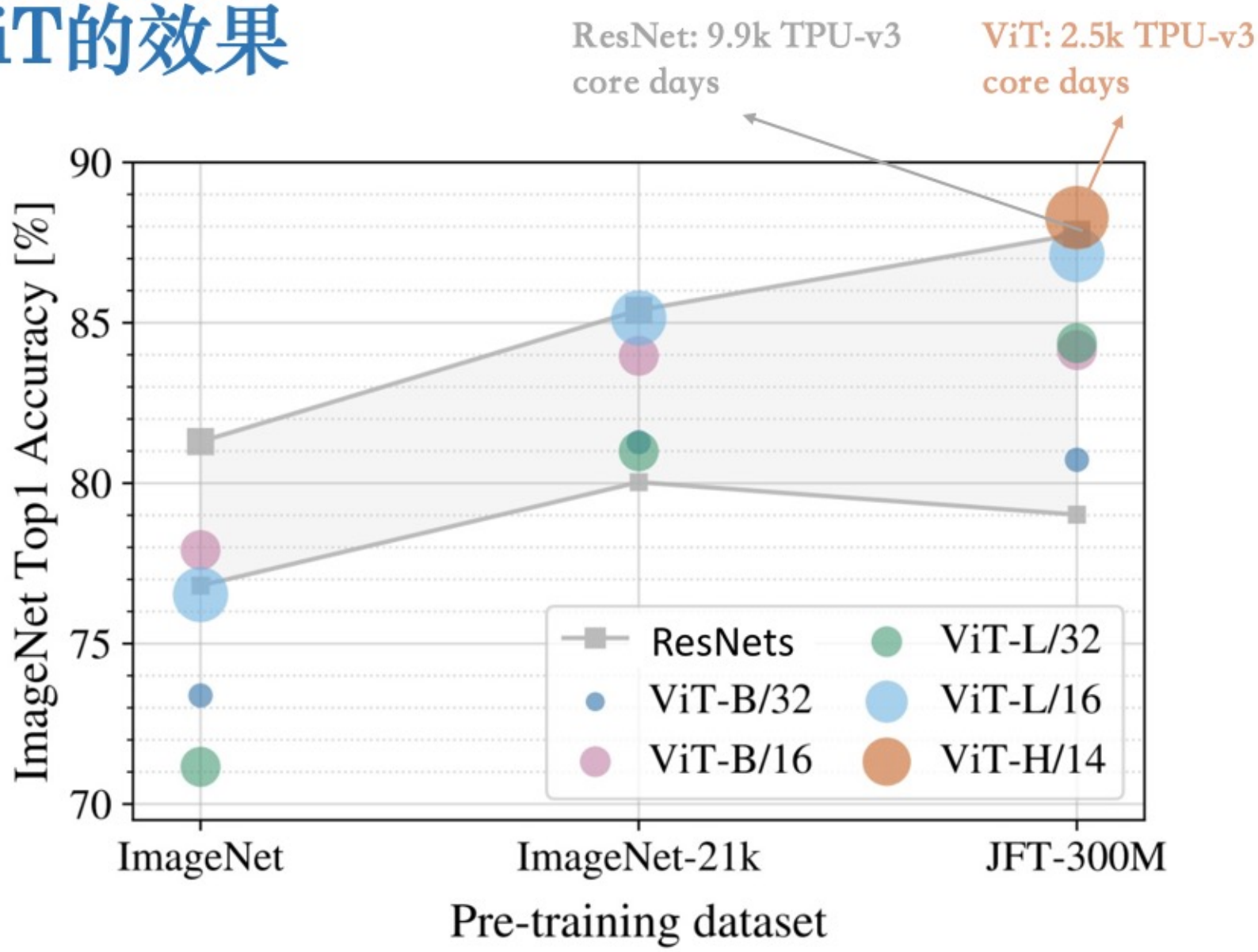
ViT的理解



图片被切分成不连续的patches（可以视为tokens）

类似于一句话的sequence

ViT的效果


ResNet: 9.9k TPU-v3 core days
ViT: 2.5k TPU-v3 core days
ImageNet Top1 Accuracy [%]
90
85
80
75
70
ImageNet
ImageNet-21k
JFT-300M
Pre-training dataset
ResNets
ViT-B/32
ViT-B/16
ViT-L/32
ViT-L/16
ViT-H/14

生成式计算机视觉

生成式人工智能

- 通过深度学习，生成与原始数据相似但不完全相同的新数据。模型会学习给定数据集中的特征和模式，然后以适当的泛化能力生成新内容。
- 生成的内容可包括文本、图像、声音、视频、代码等多种类型（著名产品：ChatGPT）。

生成式人工智能

计算机视觉领域的生成式模型：

- 生成对抗网络（GANs）
- 扩散模型（Diffusion Models）
- 变分自编码器（VAEs）
- 自回归模型（Autoregressive Models）
- 流模型（Flow-based Models，比如RealNVP）

本课内容

生成对抗网络 GANs 的提出人

Generative Adversarial Networks

Ian Goodfellow

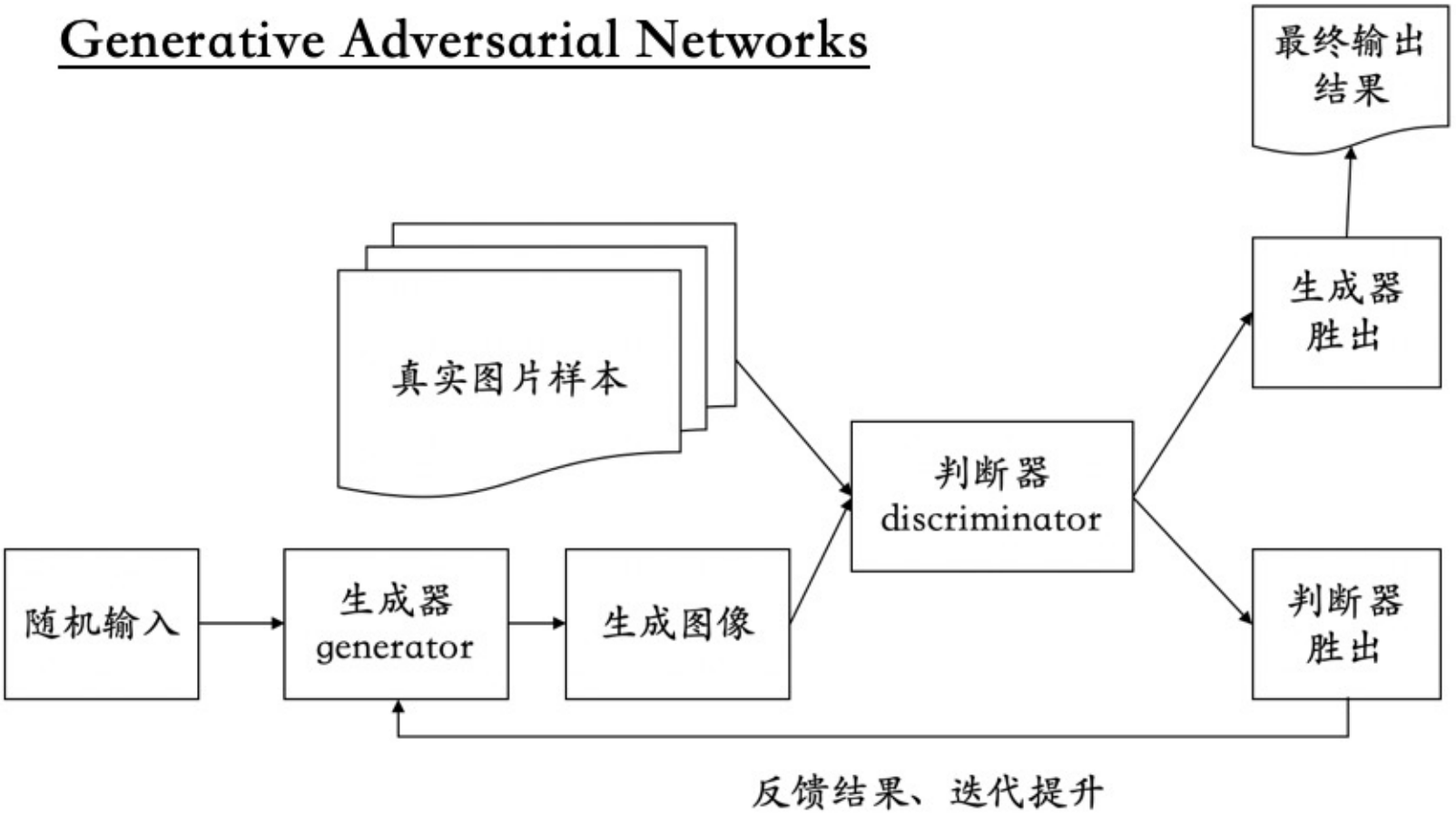
Generative Adversarial Nets

Ian J. Goodfellow[*], **Jean Pouget-Abadie**[†], **Mehdi Mirza**, **Bing Xu**, **David Warde-Farley**, **Sherjil Ozair**[‡], **Aaron Courville**, **Yoshua Bengio**[§]
Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, QC H3C 3J7

Abstract

We propose a new framework for estimating generative models via an adversarial process, in which we simultaneously train two models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G. The training procedure for G is to maximize the probability of D making a mistake. This framework corresponds to a minimax two-player game. In the space of arbitrary functions G and D, a unique solution exists, with G recovering the training data distribution and D equal to $\frac{1}{2}$ everywhere. In the case where G and D are defined by multilayer perceptrons, the entire system can be trained with backpropagation. There is no need for any Markov chains or unrolled approximate inference networks during either training or generation of samples. Experiments demonstrate the potential of the framework through qualitative and quantitative evaluation of the generated samples.

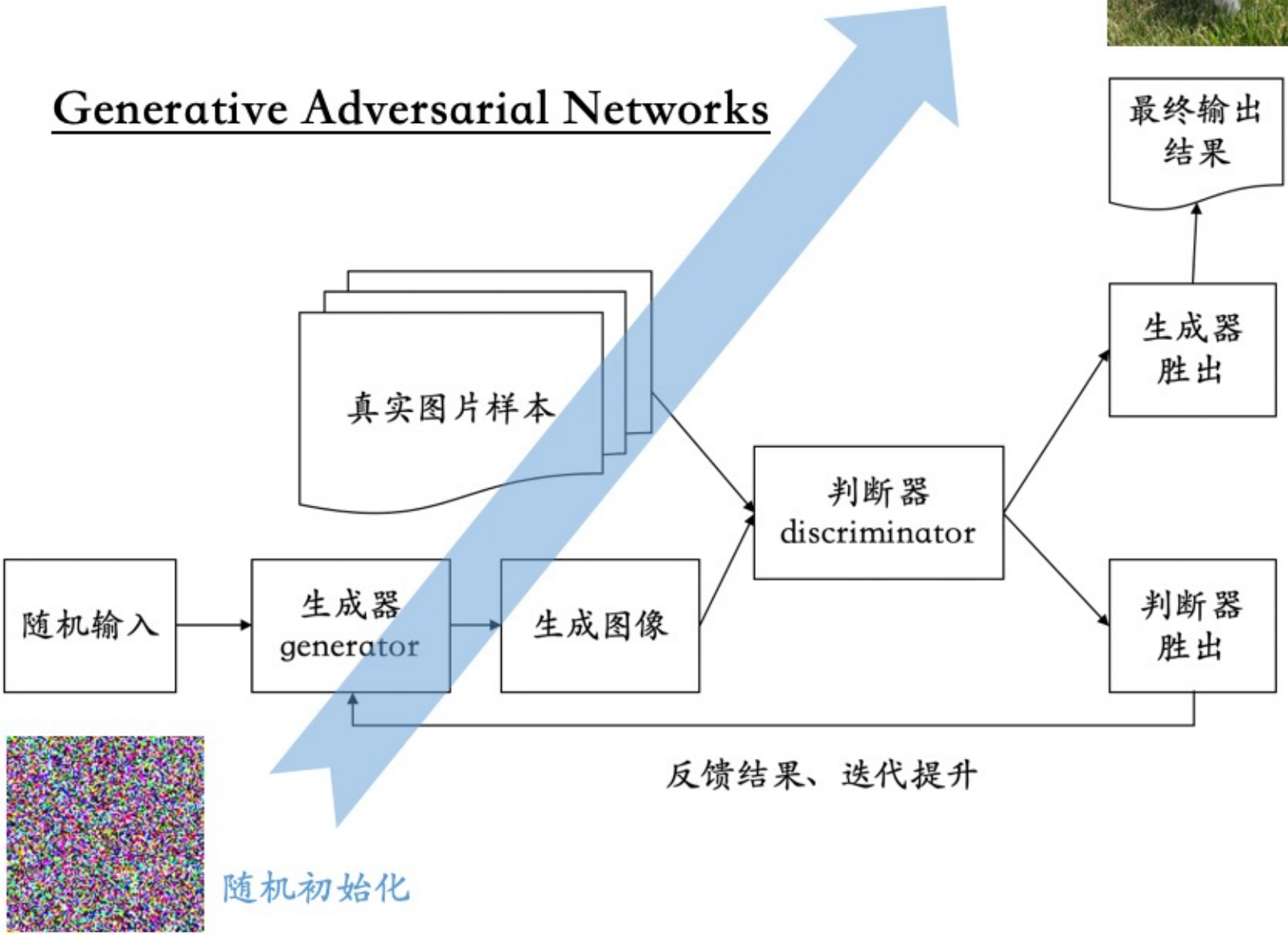
生成对抗网络 GANs


Generative Adversarial Networks
最终输出结果
真实图片样本
生成器胜出
判断器 discriminator
随机输入
生成器 generator
生成图像
判断器胜出
反馈结果、迭代提升

生成对抗网络 GANs

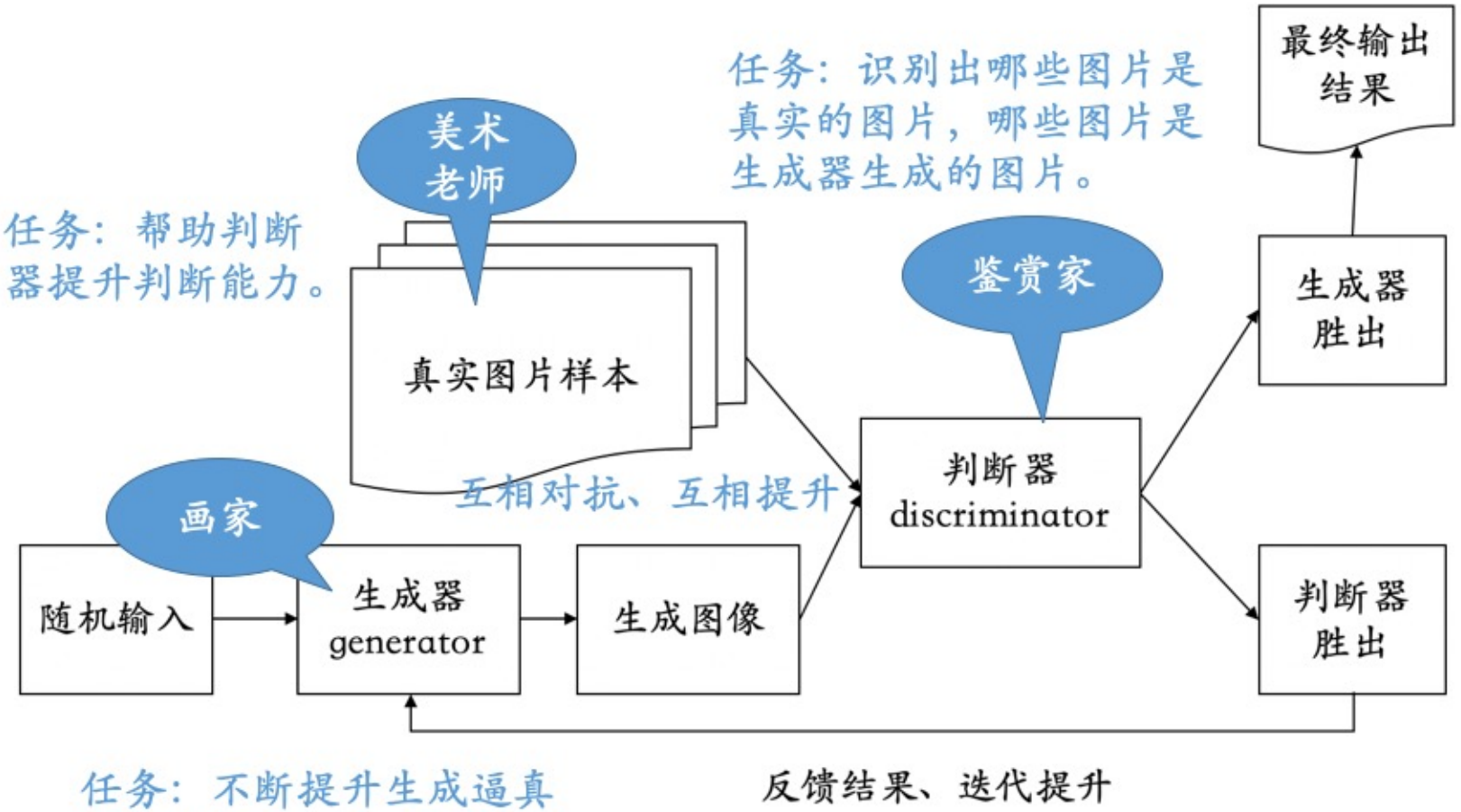
逼真的图片

Generative Adversarial Networks

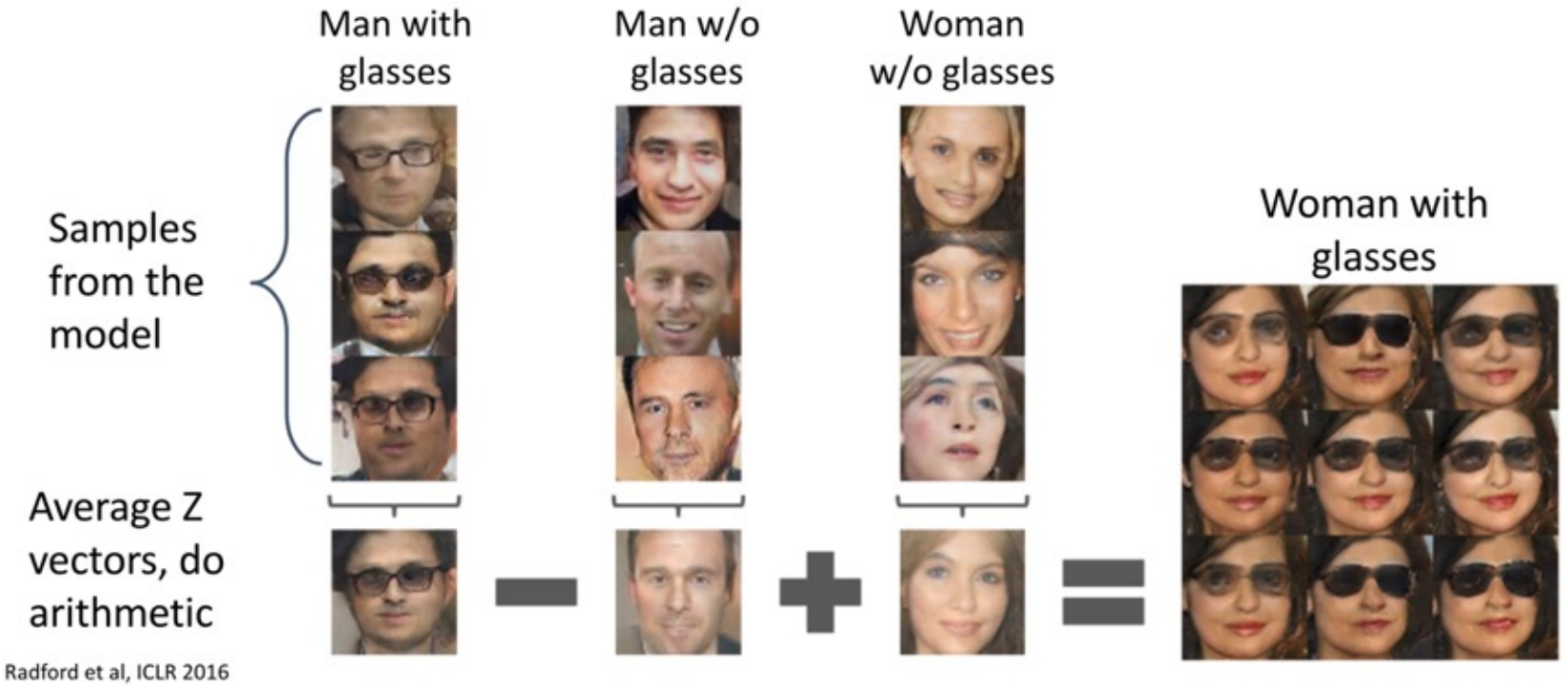


GANs的工作流程

GANs的工作流程


最终输出结果
任务：识别出哪些图片是真实的图片，哪些图片是生成器生成的图片。
美术老师
任务：帮助判断器提升判断能力。
鉴赏家
生成器胜出
真实图片样本
判断器
discriminator
互相对抗、互相提升
画家
随机输入
生成器
generator
生成图像
判断器胜出
反馈结果、迭代提升
任务：不断提升生成逼真图片的能力，力求迷惑判断器，让其无法判断图片是否由生成器生成。

图像矢量变换 Vector Math



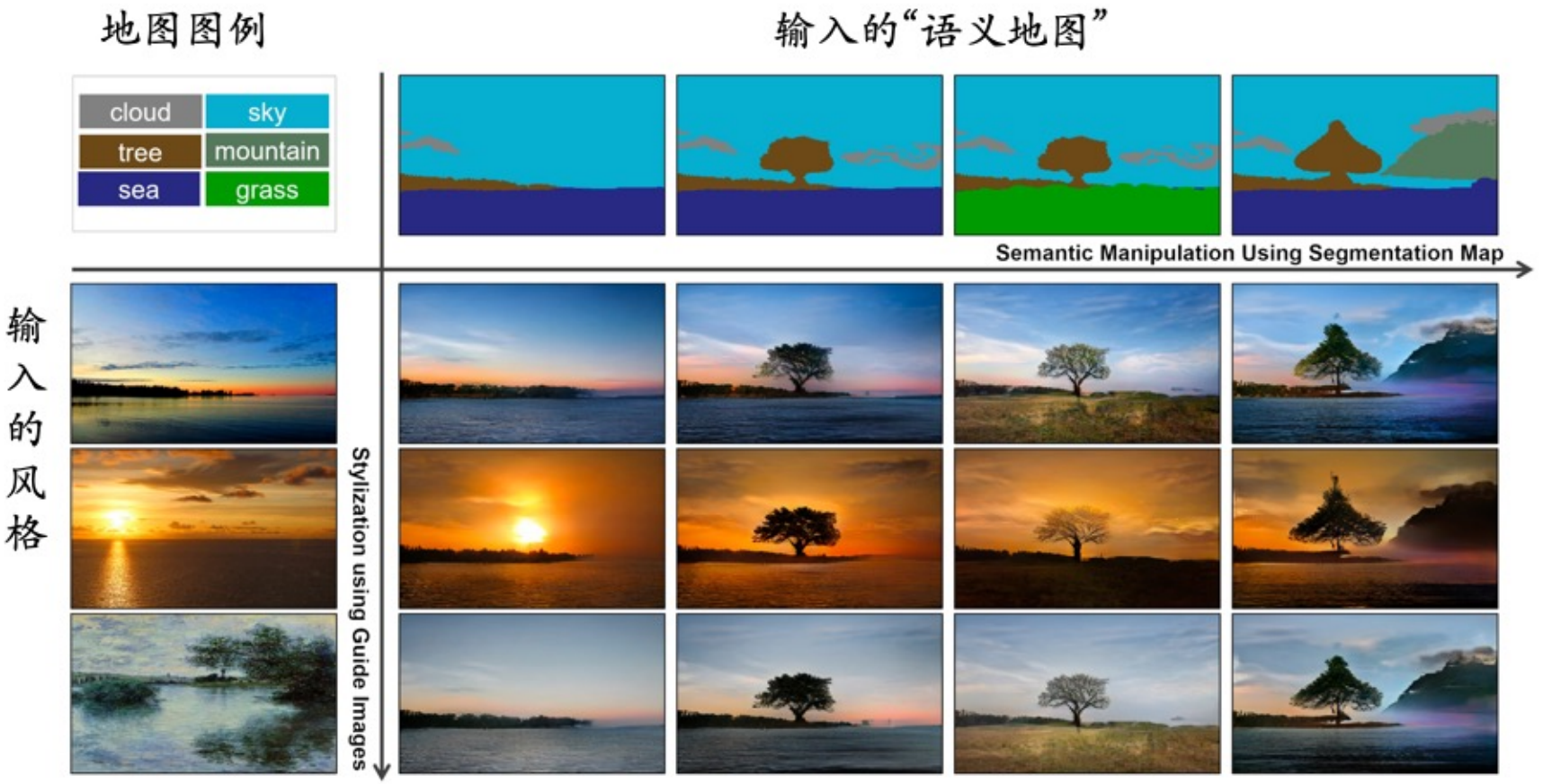
Radford et al, ICLR 2016

“矢量计算”

戴眼镜的男人-不戴眼镜的男人+不戴眼镜的女人=戴眼镜的女人

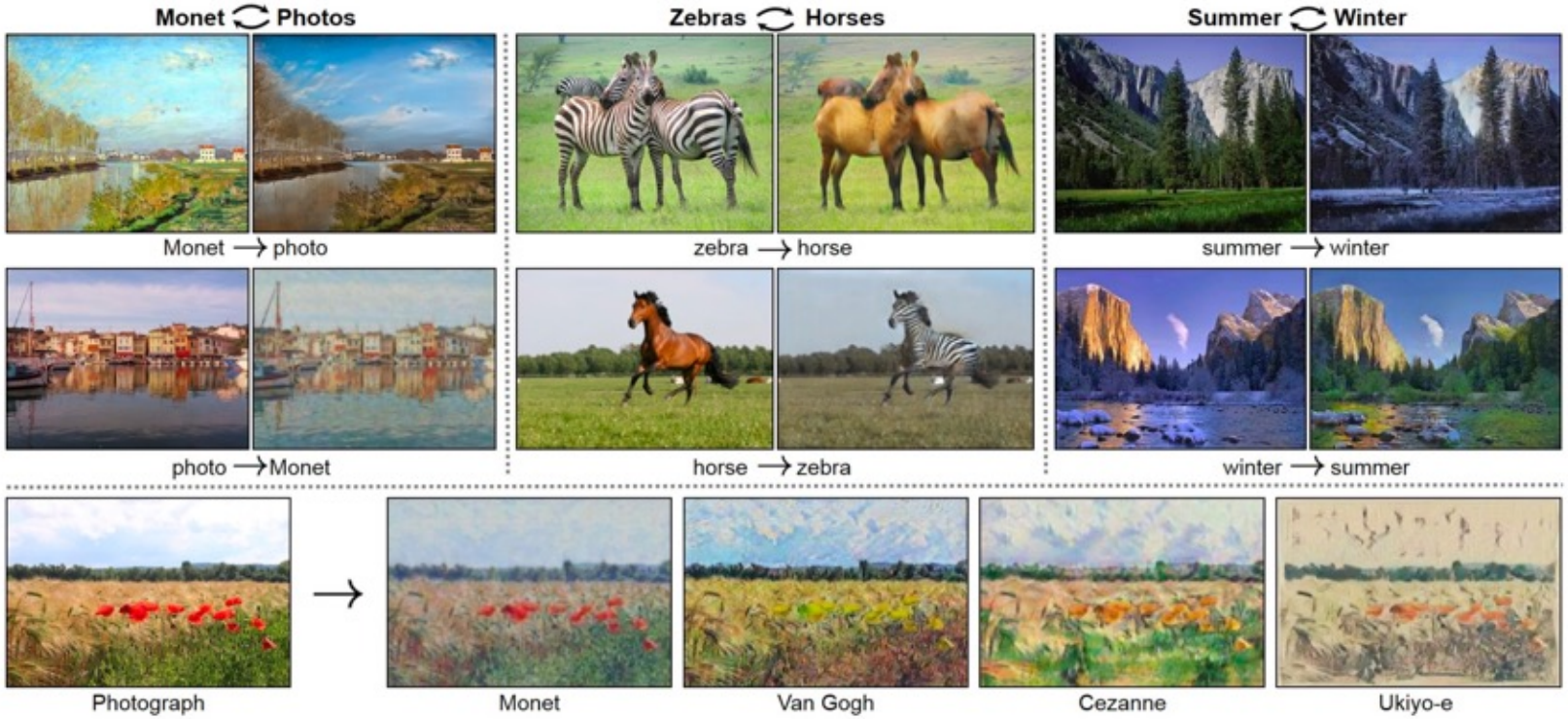
眼镜+不戴眼镜的女人=戴眼镜的女人

语义成图 Label map to image



Park et al, “Semantic Image Synthesis with Spatially-Adaptive Normalization”, CVPR 2019

风格迁移 Style transfer



扩散模型 Diffusion Models

通过模拟随机游走的过程，在每个时间点上不断更新样本分布，从而逐渐逼近目标分布。基本思想：通过一系列迭代过程，从初始状态逐渐演化到目标分布。

应用领域：

- 环境科学和生态学：预测污染物在水体中的传播方式。
- 医学：生物分子在细胞内的传输特性、疾病的扩散机理。
- 公共卫生：预测流行病的传染途径。
- 互联网和新闻媒体：研究信息和舆论的传播规律。
- 金融：研究股票价格等金融变量的波动规律。
- 计算机视觉——生成逼真的图像！

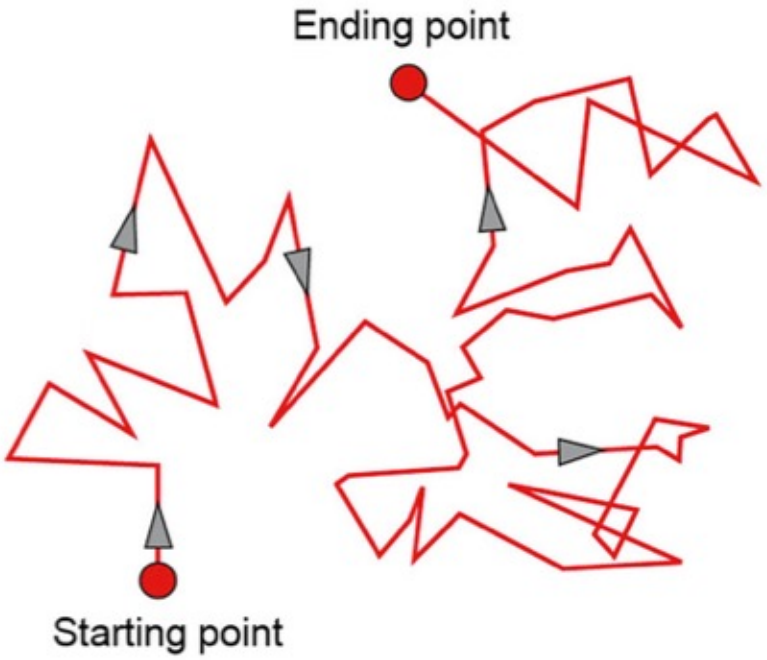
GAN vs Diffusion

	GANs	Diffusion
训练稳定性	训练过程可能不稳定，容易出现模式崩溃和训练振荡等问题。	训练过程中的loss收敛性较好，训练相对平稳。
生成内容	生成某种单类别的图像或模仿图画风格时，效果很好，但难以生成物体要素繁多、空间结构复杂的图像。	可以模拟现实中具有复杂空间逻辑的场景，能够生成更加复杂的逼真图像。
生成速度	速度相对较快。	速度较慢，因为需要迭代更多次，且过程中图像尺寸不变。
使用场景	适用于需要实时反馈，或对特定类别图像生成要求较高的场景。	适用于需要处理多种类别图像或复杂场景的应用。

物理学的扩散过程

初始集中的染色剂分子经过无规则的
热运动逐渐扩散到整个容器。

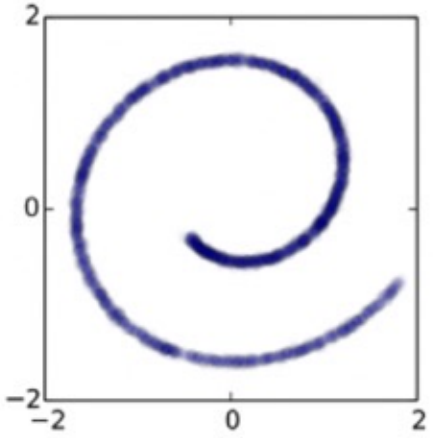
布朗运动

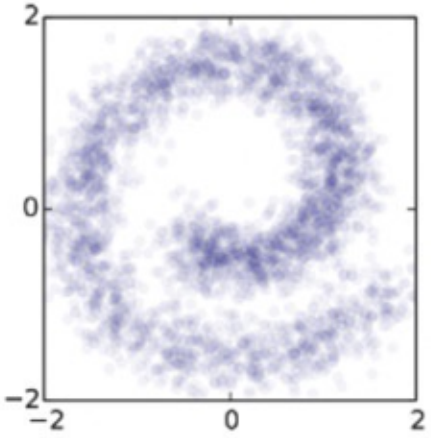


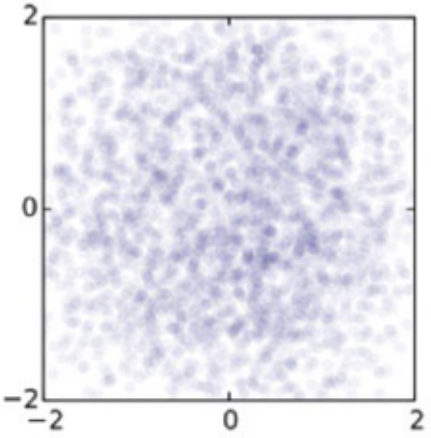
Robert Brown

布朗运动（Brown motion）：悬浮在液体或气体中的微粒所做的持续无规则运动，这种运动是由于微粒受到液体或气体分子从各个方向的不平衡冲撞所引起的。

布朗运动



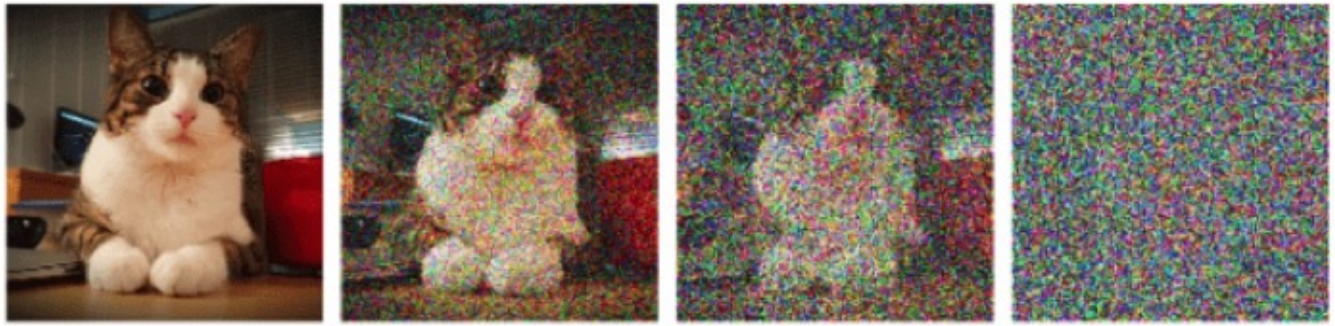




有序的图案

布朗运动

高斯分布（正态分布）



通过模拟中间过程，把图像恢复出来？

Diffusion模型的发展

2015年论文

Deep Unsupervised Learning using Nonequilibrium Thermodynamics

Jascha Sohl-Dickstein JASCHA@STANFORD.EDU
Stanford University

Eric A. Weiss EWEISS@BERKELEY.EDU
University of California, Berkeley

Niru Maheswaranathan NIRUM@STANFORD.EDU
Stanford University

Surya Ganguli SGANGULI@STANFORD.EDU
Stanford University

2020年论文

Denoising Diffusion Probabilistic Models

Jonathan Ho
UC Berkeley
jonathanho@berkeley.edu

Ajay Jain
UC Berkeley
ajayj@berkeley.edu

Pieter Abbeel
UC Berkeley
pabbeel@cs.berkeley.edu

2021年论文

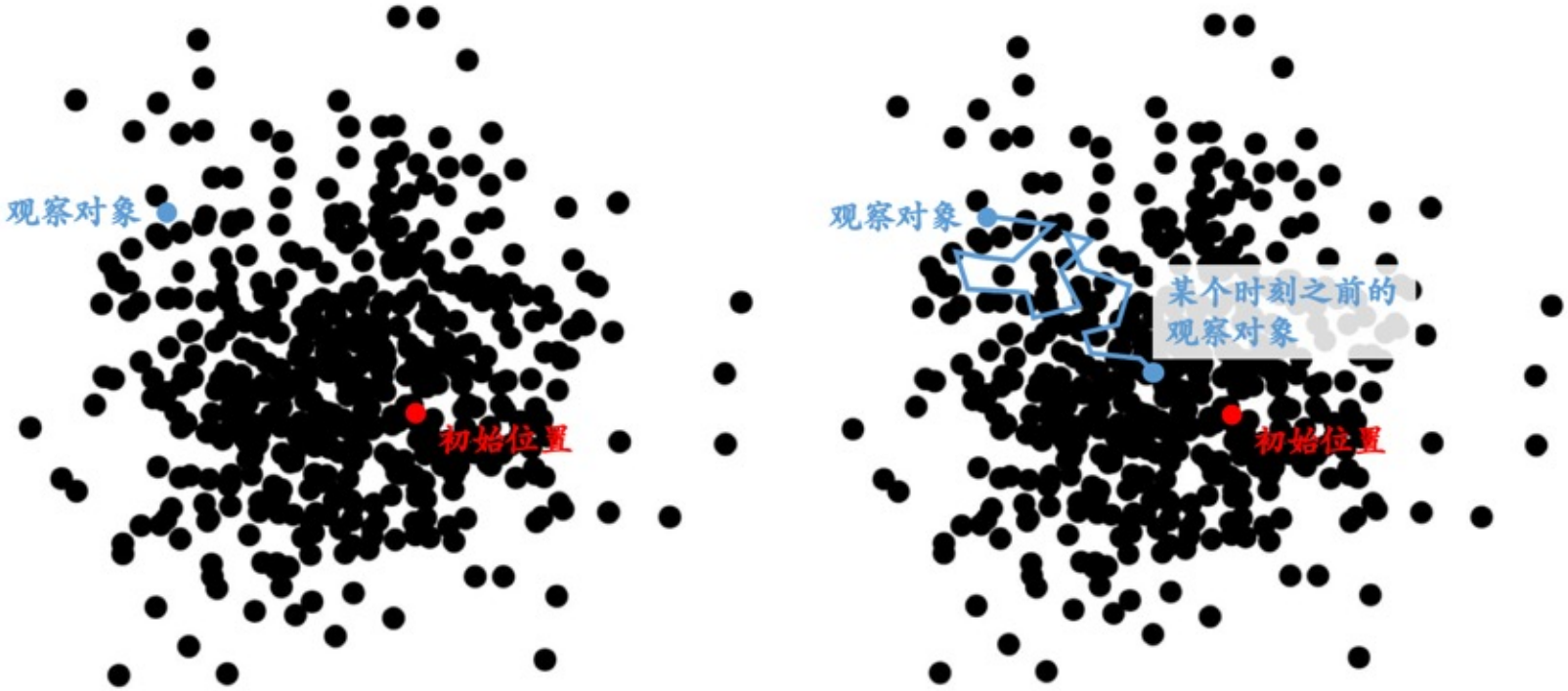
Score-based Generative Modeling in Latent Space

Arash Vahdat*
NVIDIA
avahdat@nvidia.com

Karsten Kreis*
NVIDIA
kkreis@nvidia.com

Jan Kautz
NVIDIA
jkautz@nvidia.com

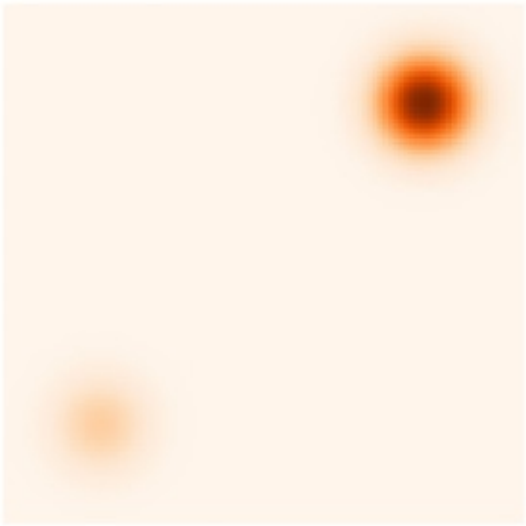
路径回溯



Score function

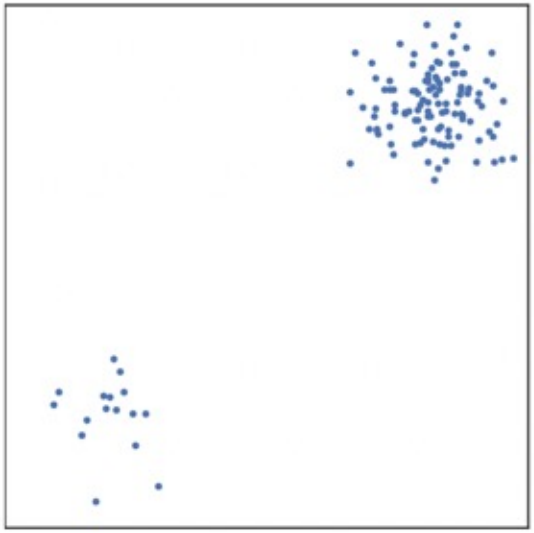
概率分布密度图

$p_{\text{data}}(\mathbf{x})$



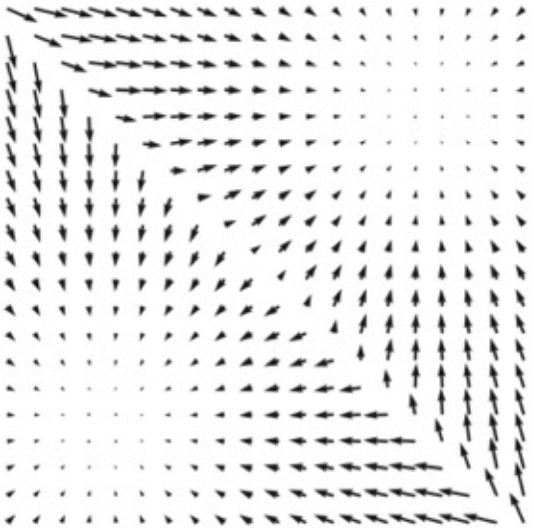
样本分布

$\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$

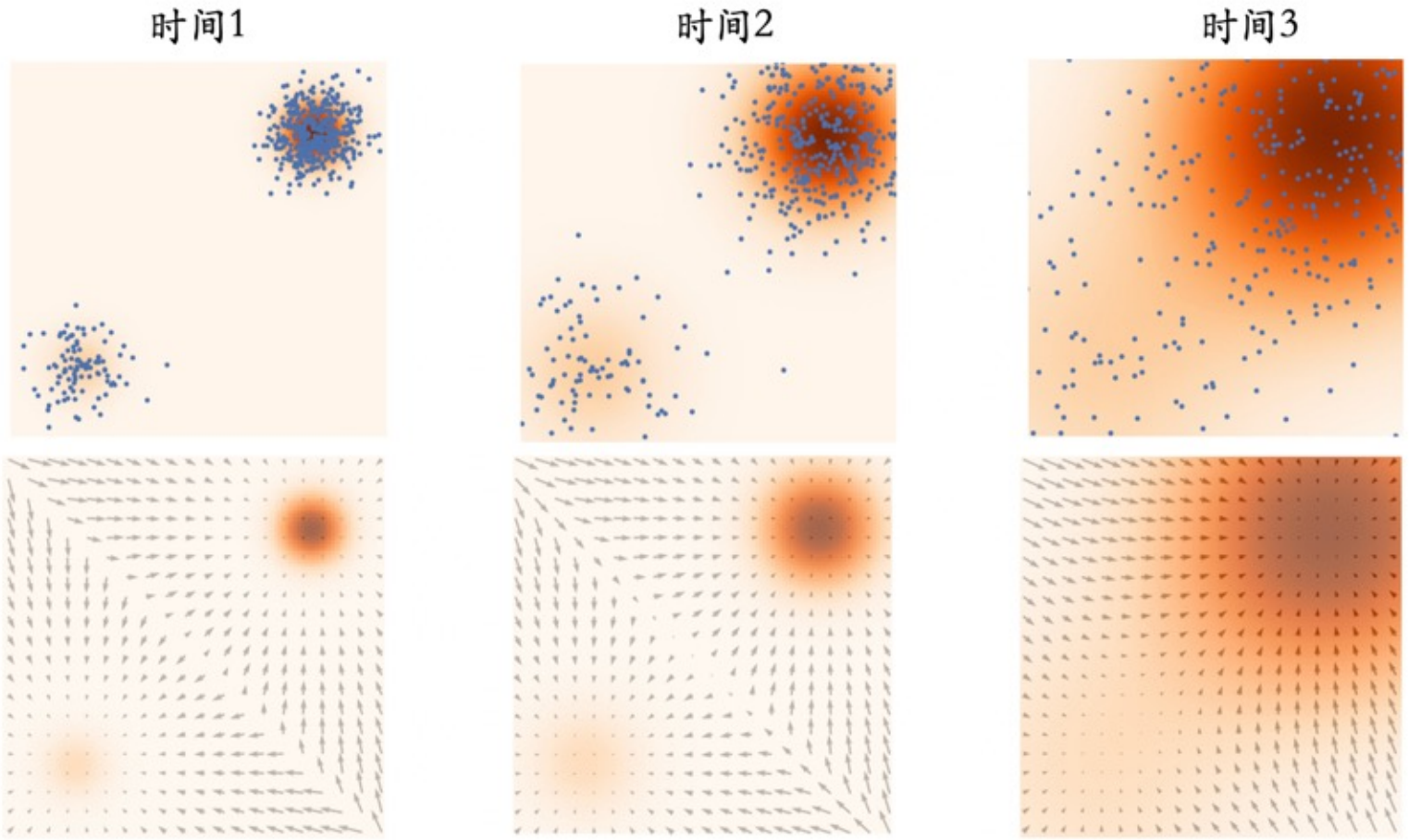


Score function

$\boldsymbol{s}_\theta(\mathbf{x}) \approx \nabla_{\mathbf{x}} \log p_{\text{data}}(\mathbf{x})$



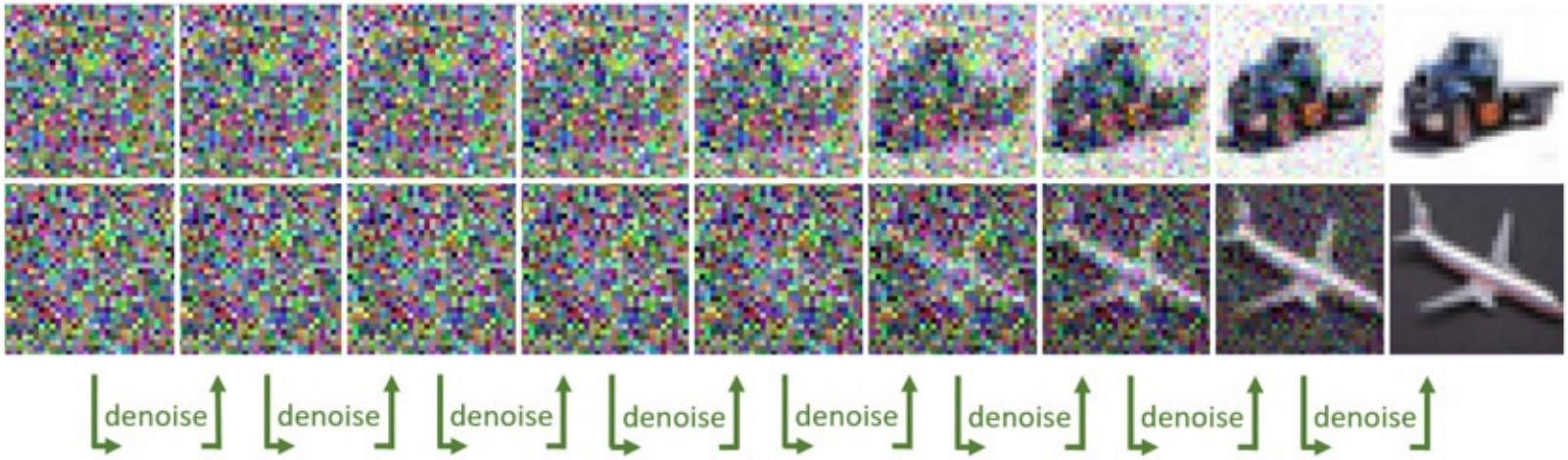
Score function



箭头：让粒子向初始状态“回归”的地图。

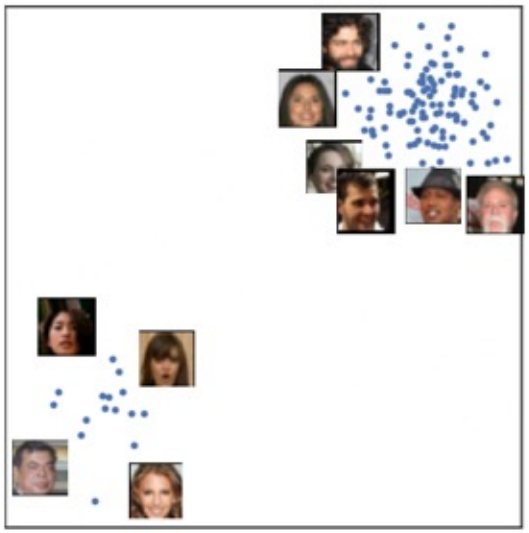
Denoising

按照score function的指引，让粒子（像素点）朝扩散前的初始状态靠拢。

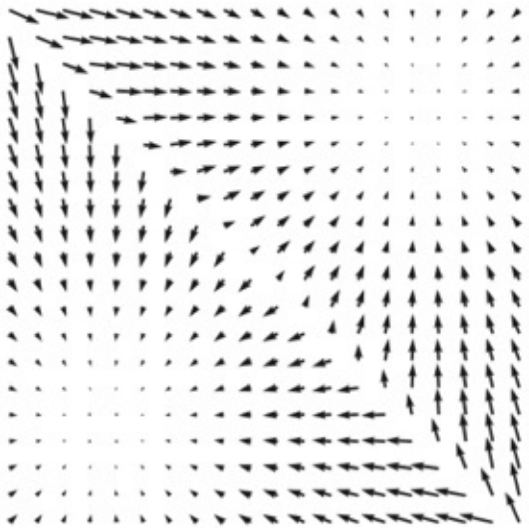


SGMs

基于Score的生成式模型（SGMs），是扩散模型的一种。



score匹配



朗之万
动力模型

数据样本（像素空间）

$\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\} \sim p_{\text{data}}(\mathbf{x})$

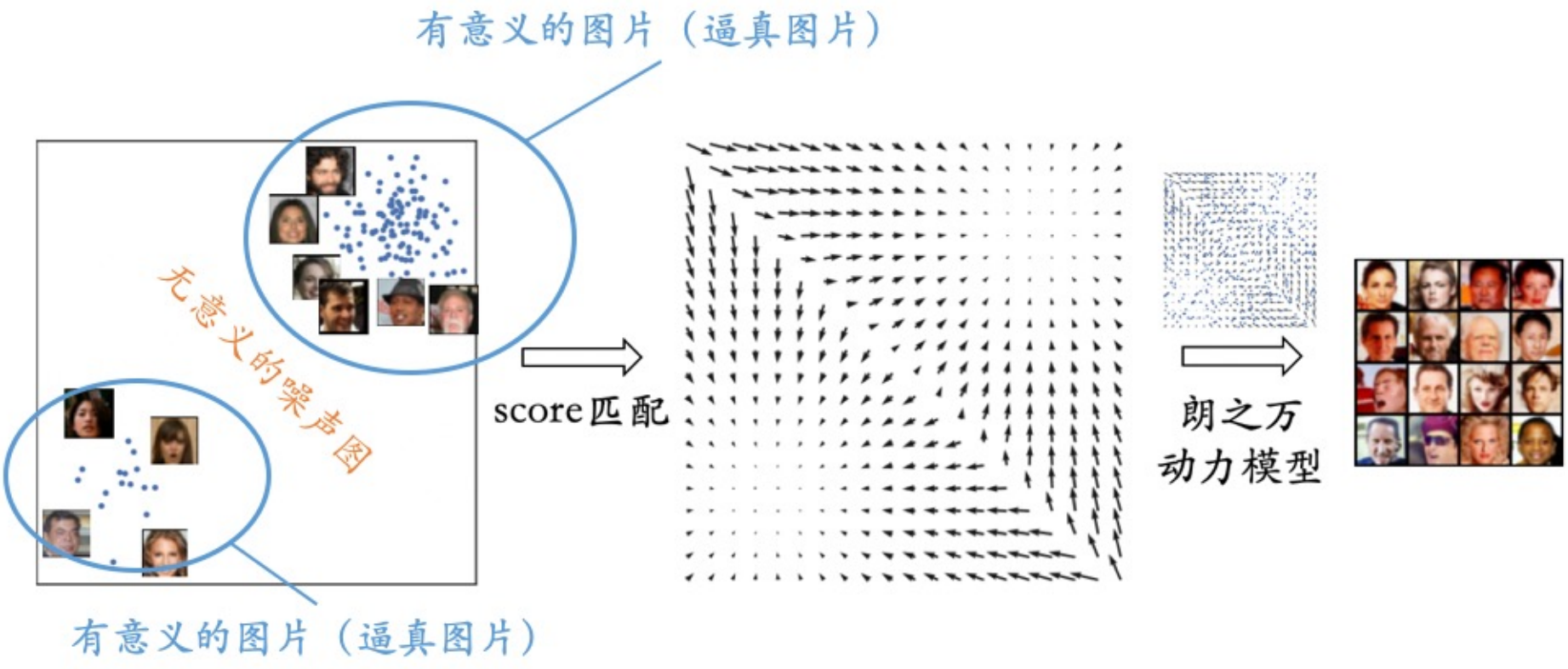
由图片每个点的像素定义
的高维空间

Scores function

$s_\theta(\mathbf{x}) \approx \nabla_{\mathbf{x}} \log p_{\text{data}}(\mathbf{x})$

生成的新样本

SGMs



数据样本（像素空间）

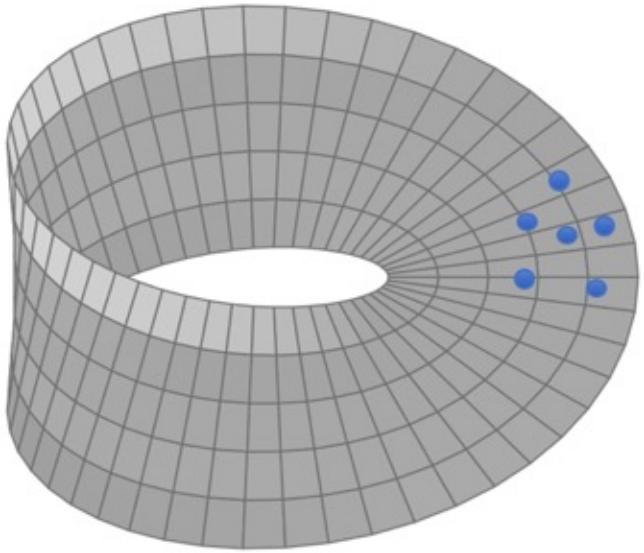
$\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\} \sim p_{\text{data}}(\mathbf{x})$

Scores function

$\boldsymbol{s}_\theta(\mathbf{x}) \approx \nabla_{\mathbf{x}} \log p_{\text{data}}(\mathbf{x})$

生成的新样本

“流形”假说 Manifold hypothesis



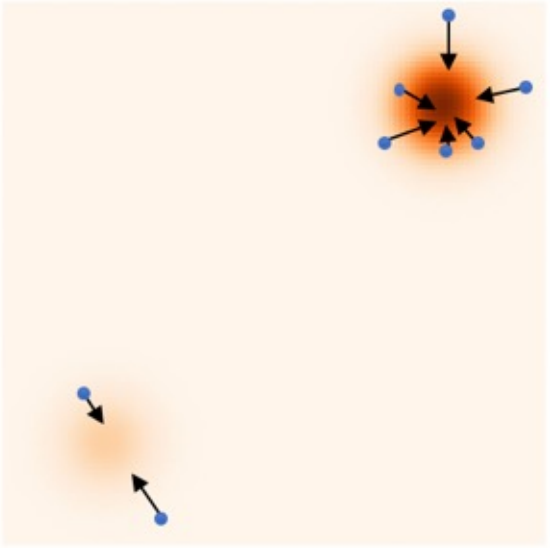
落在manifold内的数据点，被认为是“逼真”的图像。

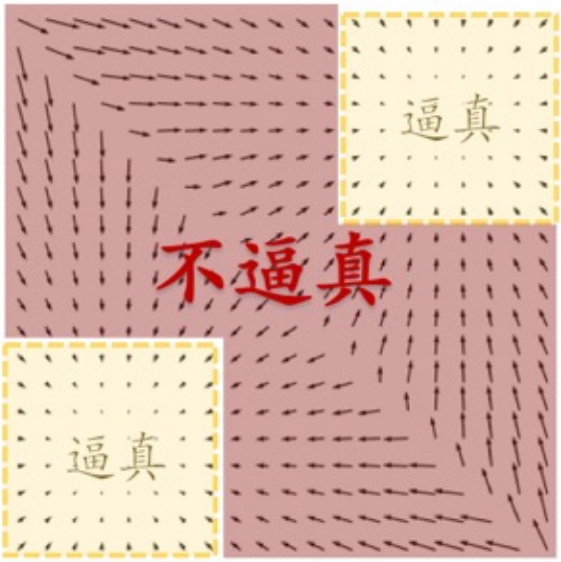
流形假说：在现实世界中的许多高维数据集中，数据点是沿着高维空间内的低维嵌入流形（latent manifolds）分布的。例如，在一个维度为3的数据空间中，有意义的数据点集中分布于一条二维莫比乌斯带上。

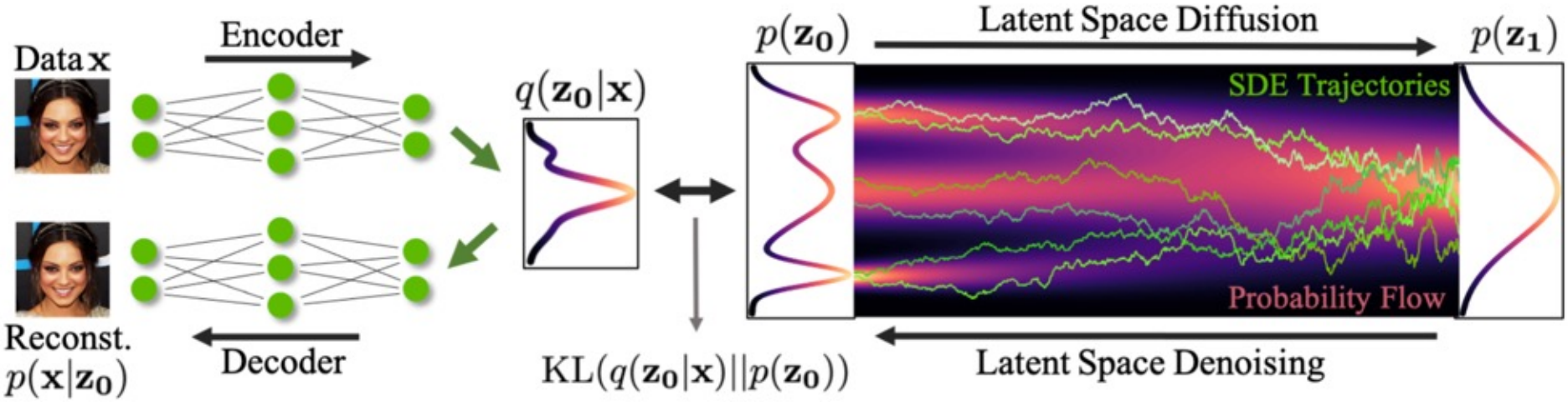
在计算机视觉领域，我们认为像素空间内的大部分区域所对应的图像都是噪声，只有分布在某一低维流形区域的图像是逼真的。

如何得知score function? 深度学习

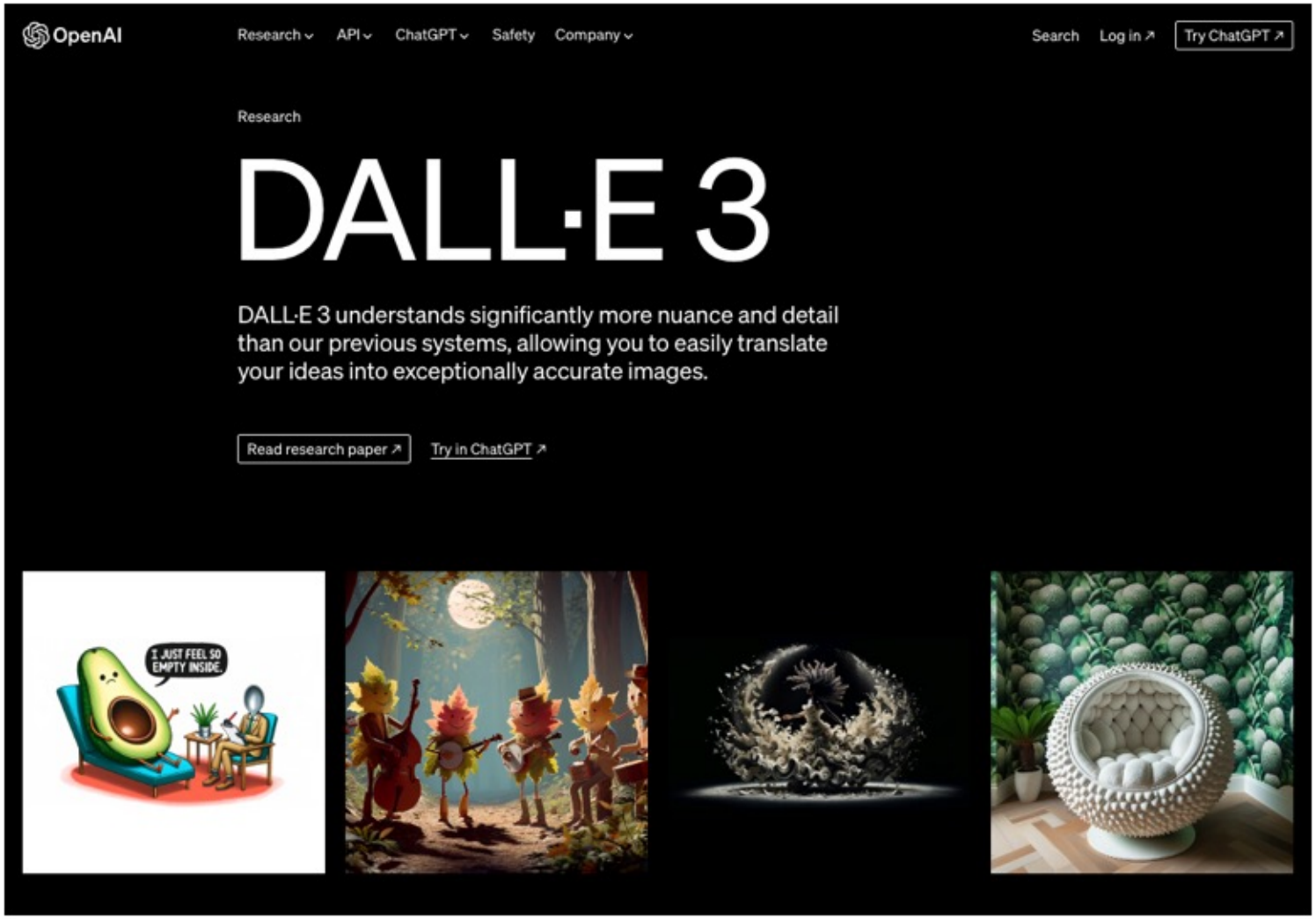
训练神经网络，
输入某个时刻的数据点的位置，
让数据点在像素空间内
向Manifold结构靠拢。







DALL-E


OpenAI
Research
API
ChatGPT
Safety
Company
Search
Log in ↗
Try ChatGPT ↗
Research
DALL·E 3
DALL·E 3 understands significantly more nuance and detail than our previous systems, allowing you to easily translate your ideas into exceptionally accurate images.
Read research paper ↗
Try in ChatGPT ↗
I JUST FEEL SO EMPTY INSIDE.

Sora