



南京大學
NANJING UNIVERSITY

人工智能导论

深度学习与大模型

(Deep Learning & Large Models)

郭兰哲

南京大学 智能科学与技术学院

<https://www.lamda.nju.edu.cn/guolz/IntroAI/fall2025/index.html>

Email: guolz@nju.edu.cn

大纲

- 神经元模型到前馈神经网络
- 参数优化：BP算法
- 深度学习
- 计算机视觉与卷积神经网络**
- 自然语言处理与循环神经网络
- 多模态学习

计算机视觉(Computer Vision)

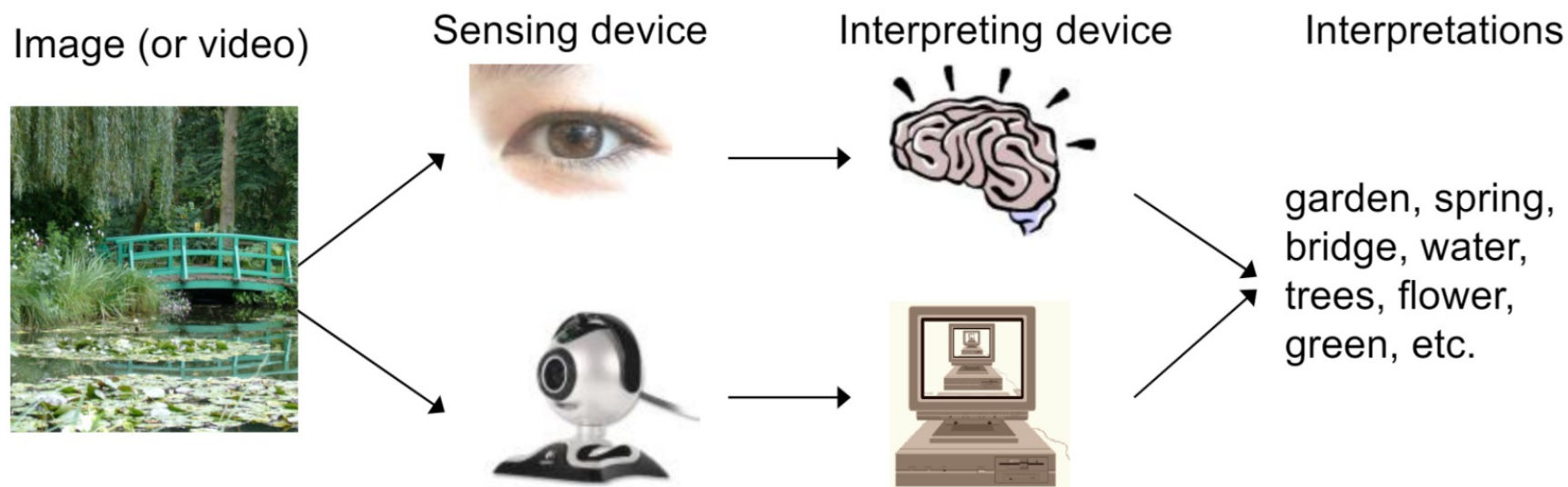
什么是计算机视觉

计算机视觉有哪些任务

如何用人工智能技术解决计算机视觉任务

计算机视觉(Computer Vision)

计算机视觉是一门研究如何对**图像**或者**视频**进行高层理解的交叉学科

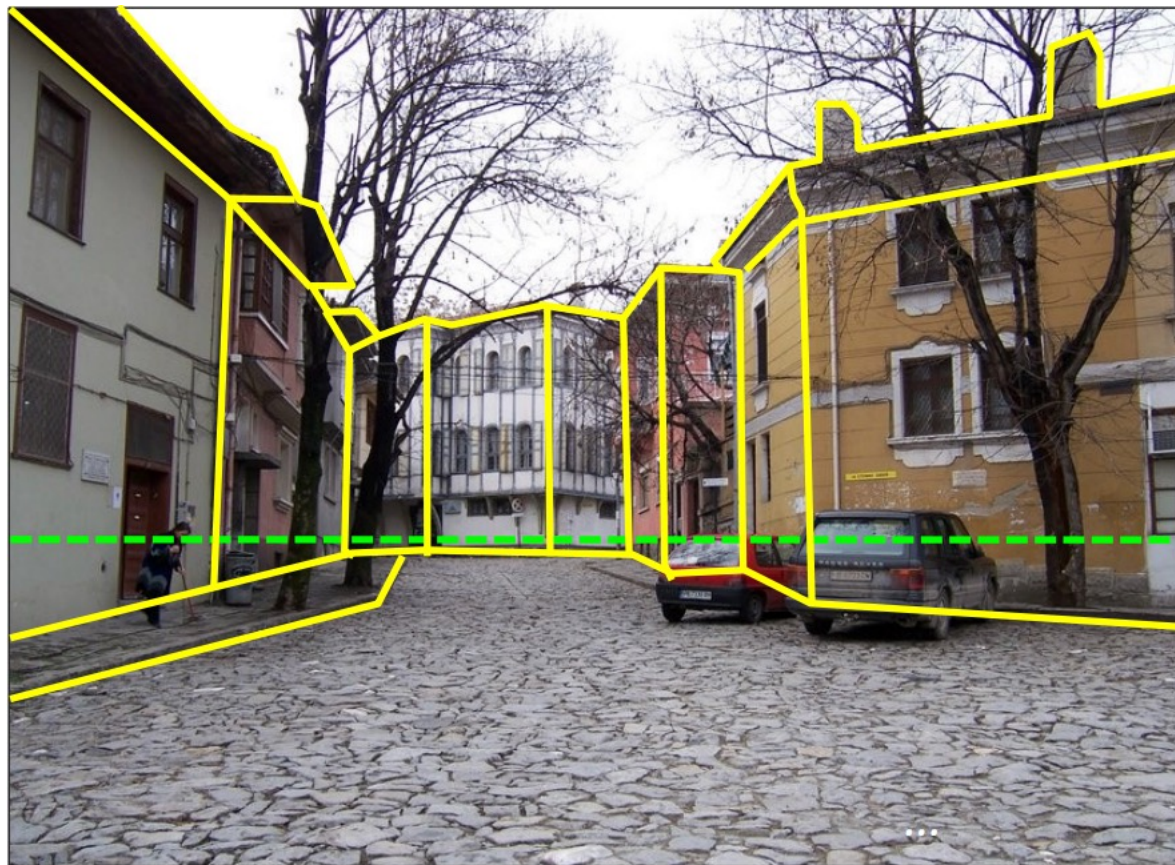


人工智能视角：赋予机器“看”的智能，即，用机器自动实现人类视觉系统的功能，图像或视频的获取、处理、分析和理解等诸多任务

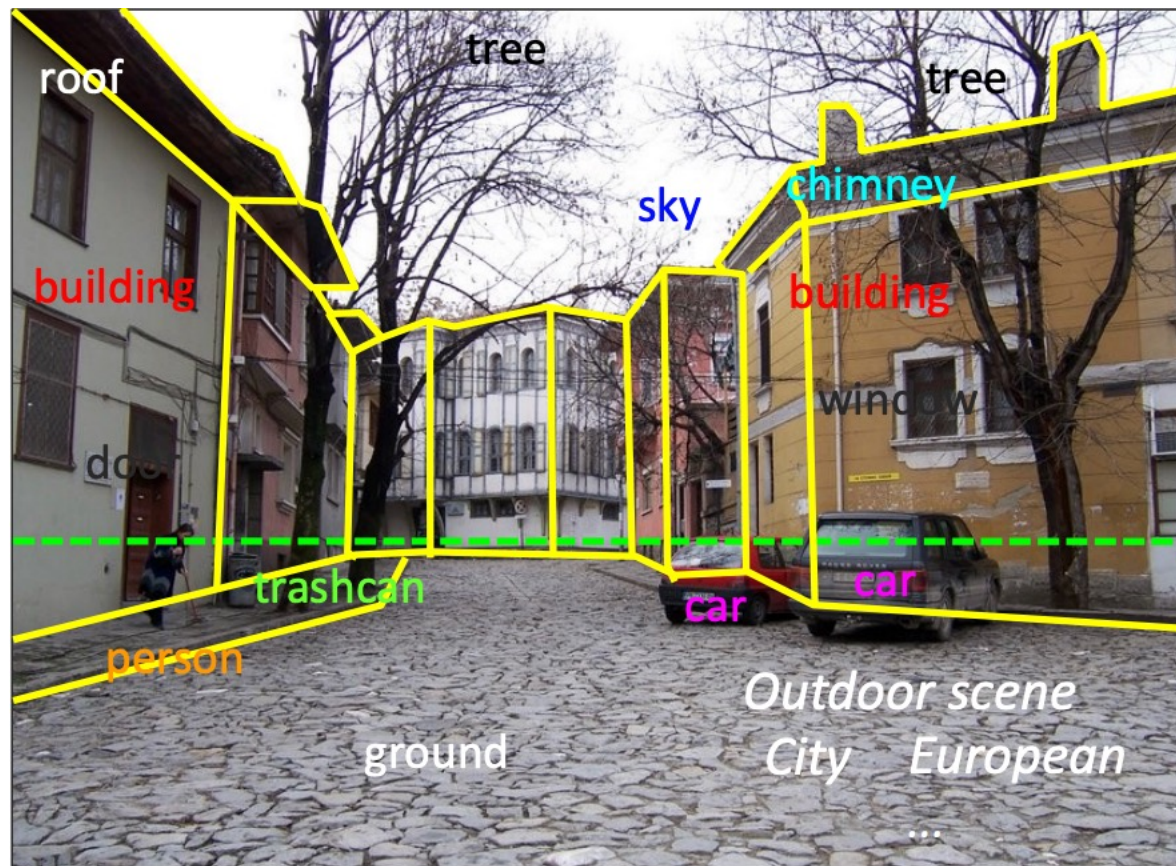
计算机视觉(Computer Vision)



计算机视觉(Computer Vision)



计算机视觉(Computer Vision)



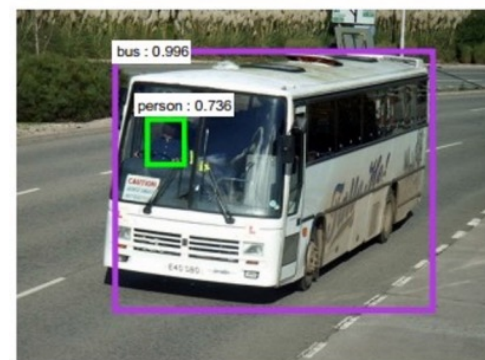
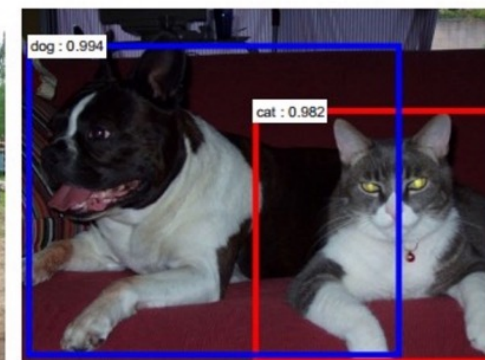
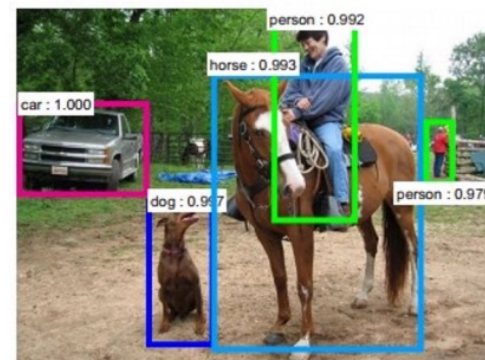
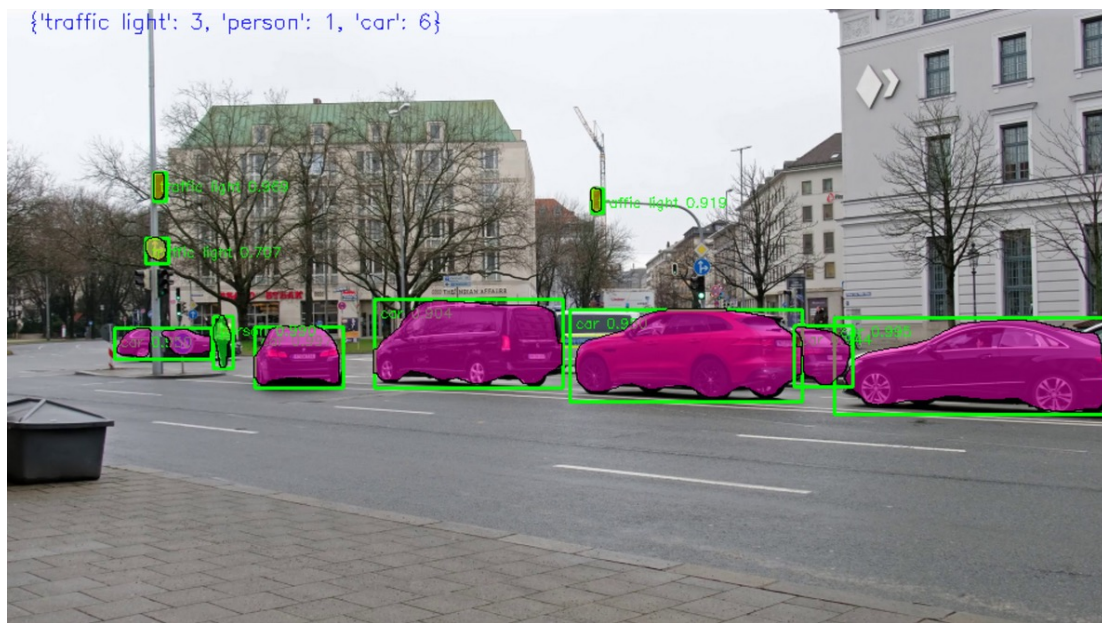
图像分类(Image Classification)

图像分类：将图片划分到某个特定的类别中



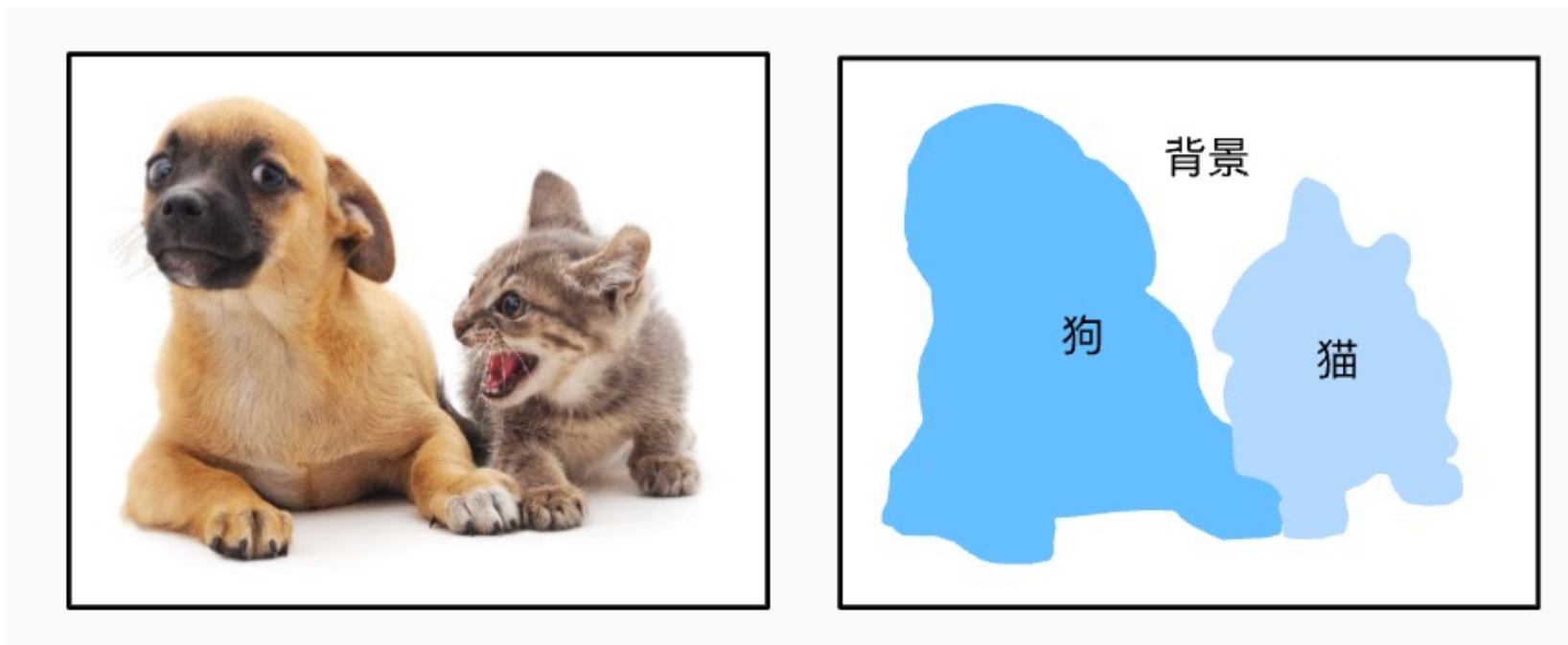
目标检测(Object Detection)

目标检测：检测图像中某个类别的物体的任务，一般来说，就是在图像中寻找我们感兴趣的目標，并给出其位置



语义分割(Semantic Segmentation)

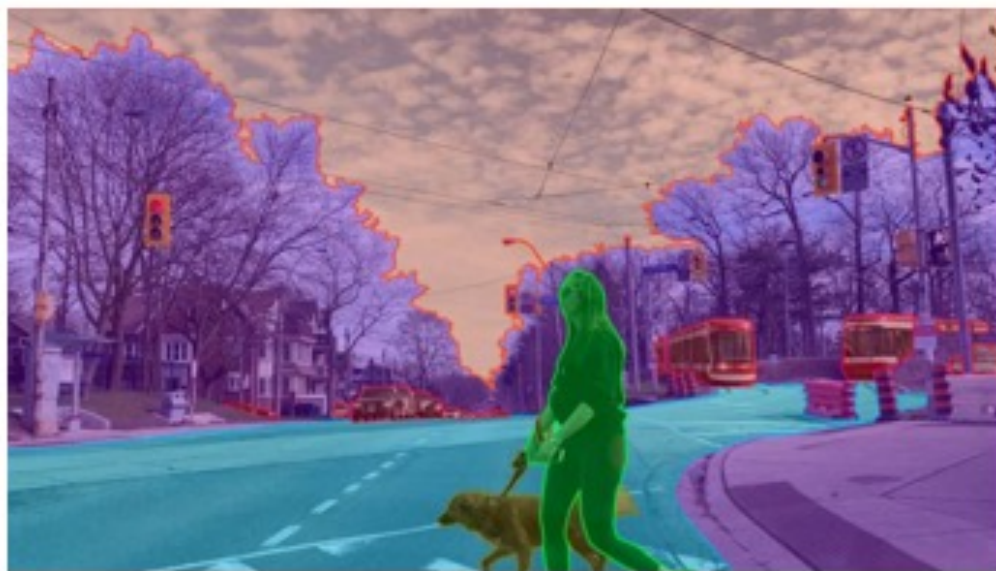
语义分割：关注如何将图像分割成不同类别的区域，与目标检测不同，语义分割可以识别并理解图像中每个像素的内容，其语义区域的标注和预测是像素级的



实例分割(Instance Segmentation)

实例分割：检测和划定出图像中出现的每个不同的感兴趣的对象，研究如何识别图像中各个目标实例的像素级区域。

与语义分割不同，实例分割**不仅需要区分语义，还要区分不同的目标实例**



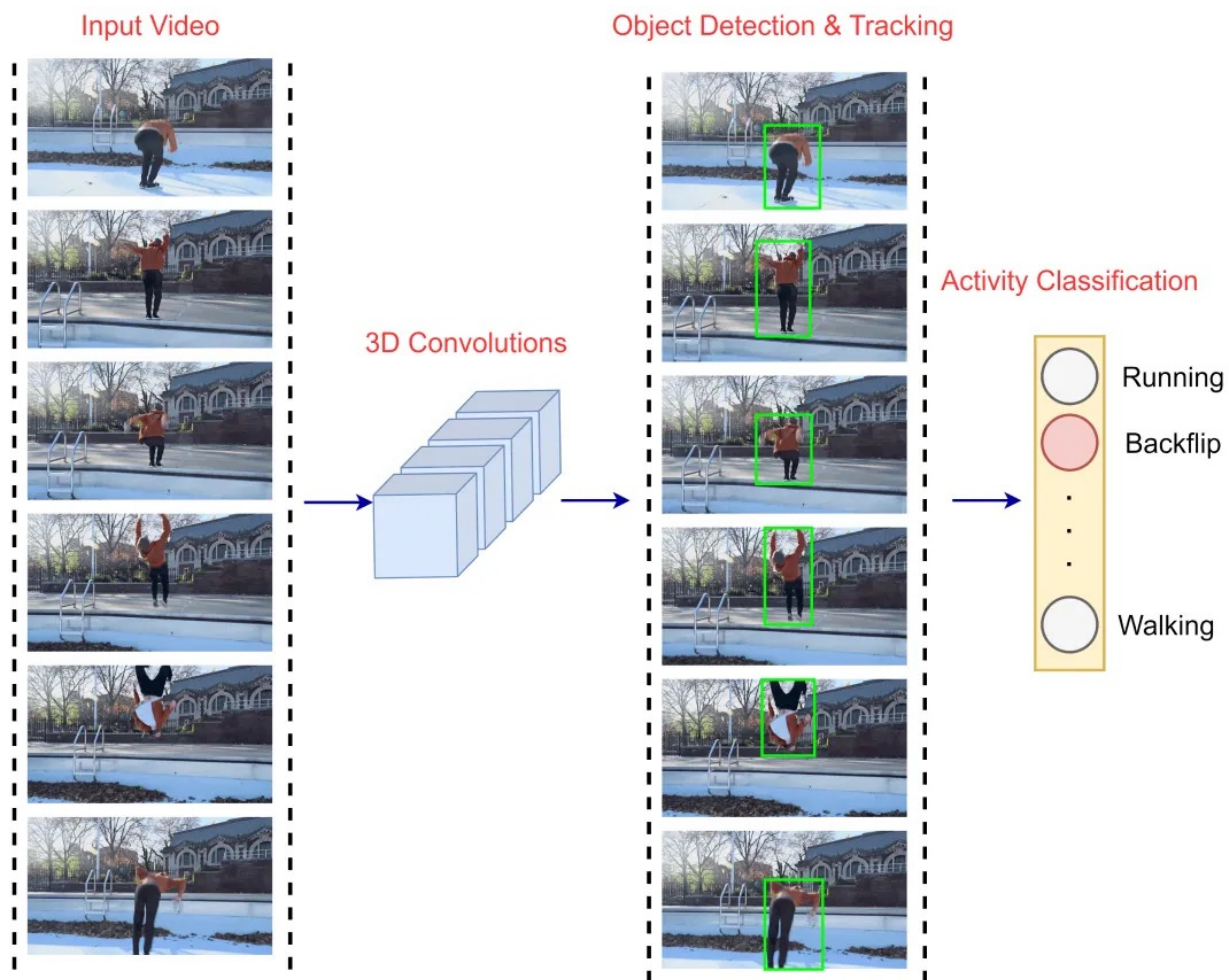
语义分割



实例分割

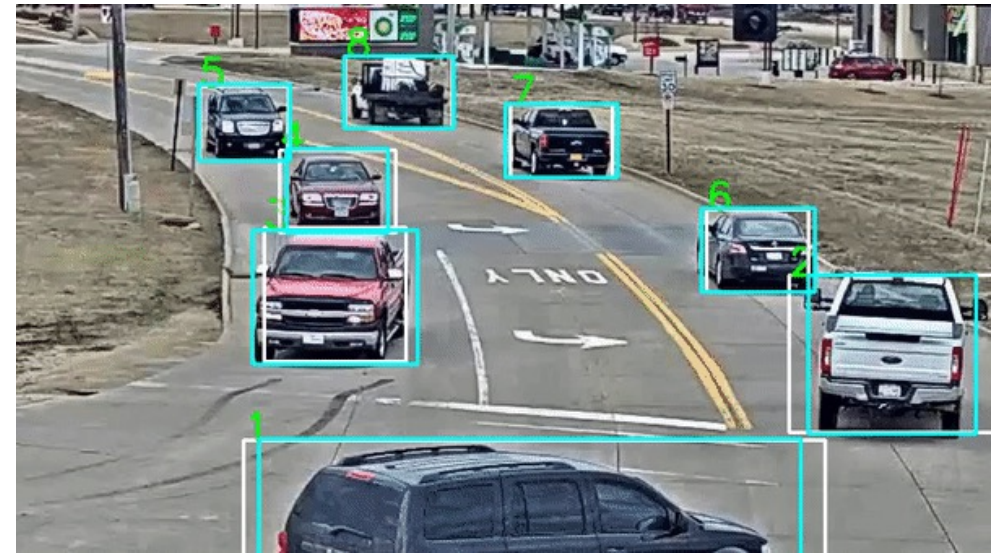
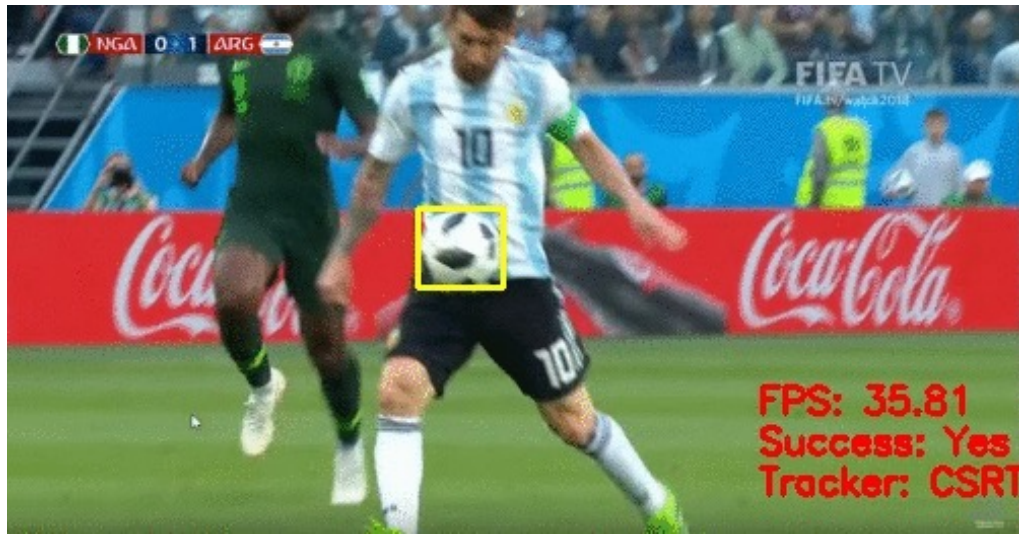
视频分类(Video Classification)

视频分类是给定一个视频，为其划分到指定的类别中



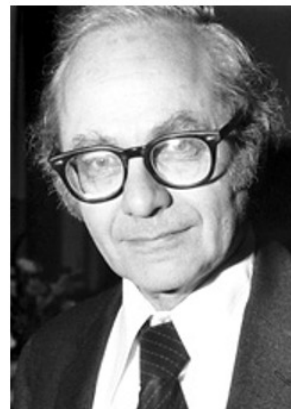
目标跟踪(Object Tracking)

目标跟踪：根据视频序列第一帧中目标的初始状态（中心位置和比例），自动获得该物体在后续视频帧中的状态



计算机视觉概述

- 20世纪50年代：研究生物视觉工作原理

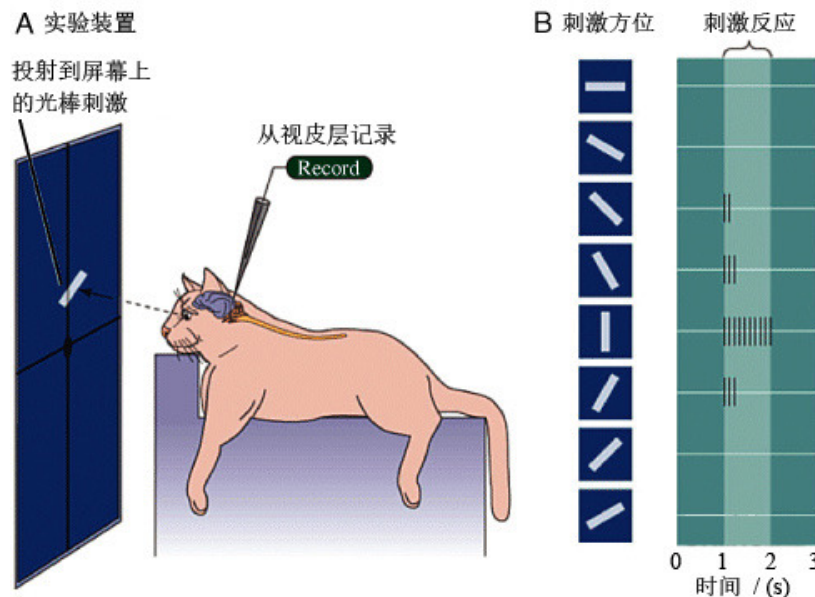


D.H.Hubel



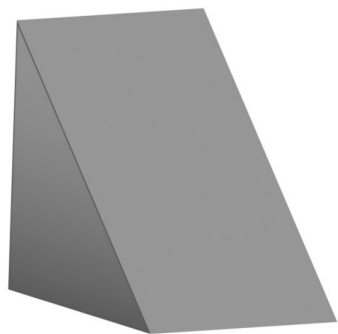
T.N.Wiesel

- Hubel和Wiesel的一些研究成果。他们从生理学的角度来分析猫（据说猫和人类的大脑比较相近）的视觉皮层系统，发现视觉通路中的信息分层处理机制，并提出了感受野的概念，也因此获得了诺贝尔生理学 and 医学奖

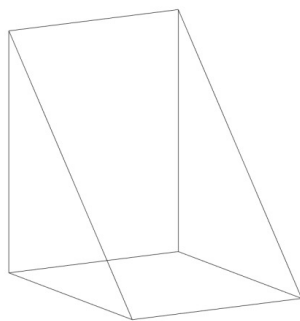


计算机视觉概述

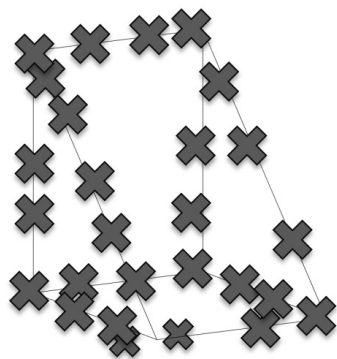
- 20世纪60年代：计算机视觉开始发展
- 第一位计算机视觉博士，Larry Roberts，在1963发表的论文"Machine perception of three-dimensional solids"中，将物体简化为几何形状（立方体、棱柱体等）来加以识别



(a) Original picture



(b) Differentiated picture

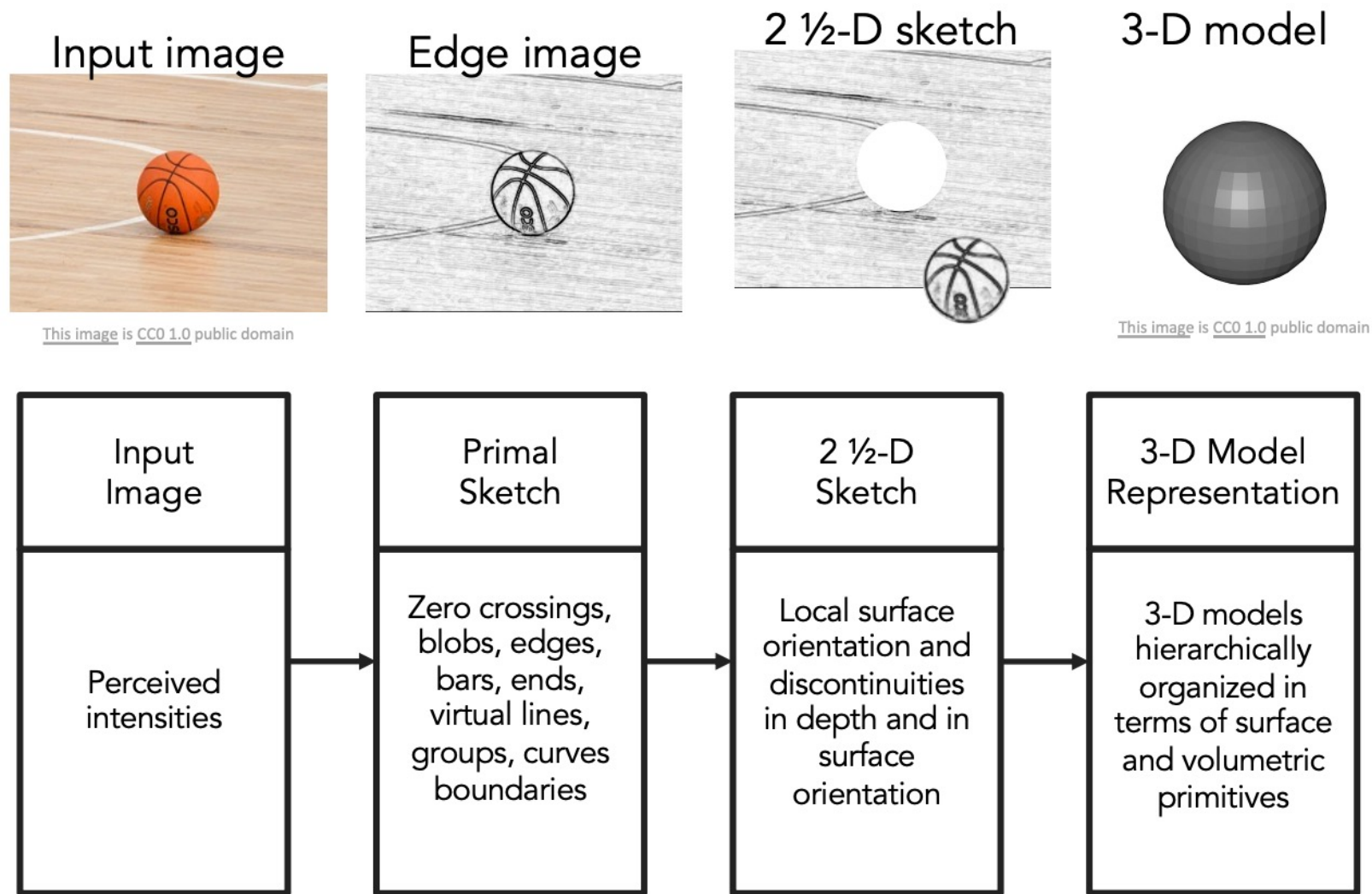


(c) Feature points selected

20世纪60年代早期，在Marvin Minsky领导下MIT的人工智能实验室成立，同一时期John McCarthy在斯坦福也建立人工智能实验室。1966年MIT人工智能实验室试图解决计算机视觉领域的理论问题，虽然实际中结果没有达到预期，但从此计算机视觉领域开始蓬勃发展。

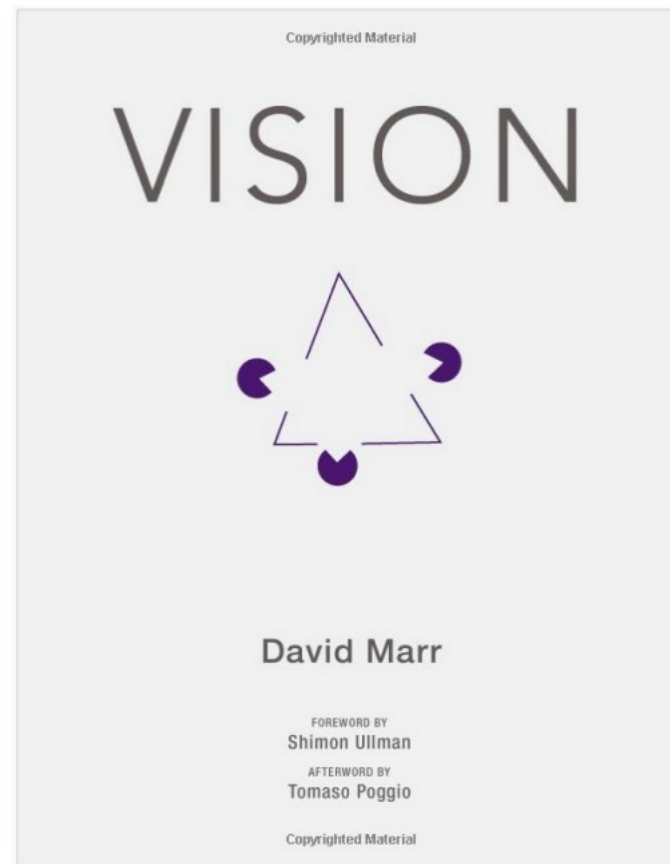
计算机视觉概述

- 20世纪70-80年代
- MIT的心理学教授David Marr在计算机视觉理论方面做出了大量贡献，融合了心理学、神经生理学、数学等多门学科，提出了有别于前人的计算机视觉分析理论，影响了计算机视觉领域的未来二十年的发展。



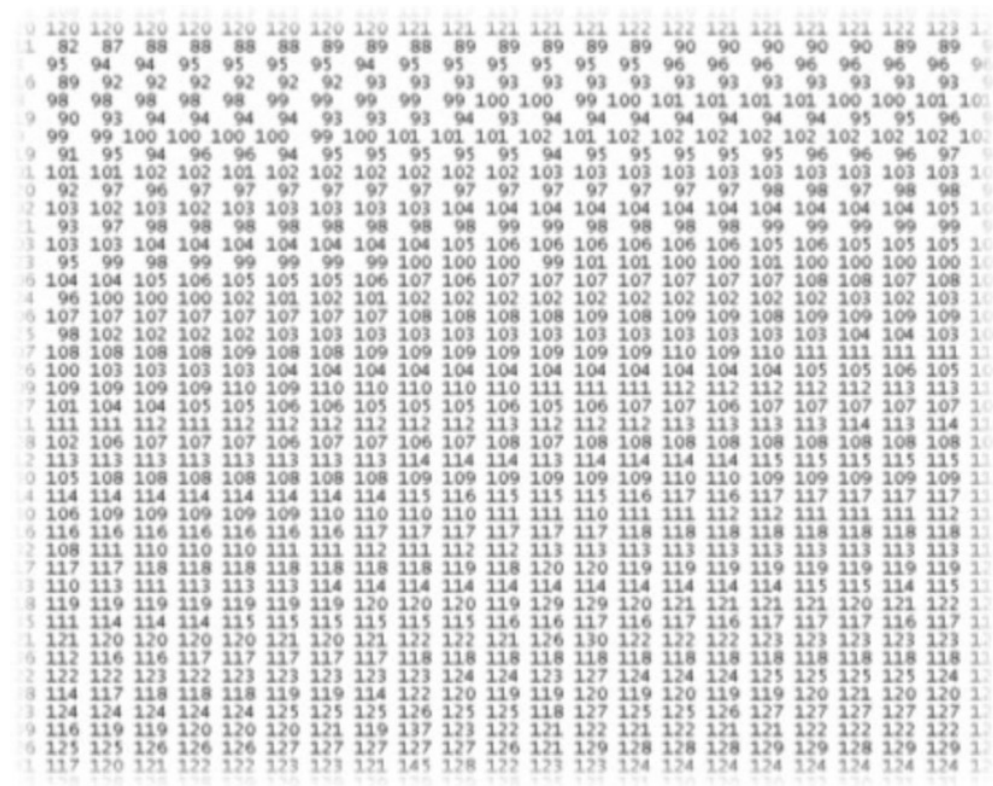
计算机视觉概述

- 20世纪70-80年代
- MIT的心理学教授David Marr在计算机视觉理论方面做出了大量贡献，融合了心理学、神经生理学、数学等多门学科，提出了有别于前人的计算机视觉分析理论，影响了计算机视觉领域的未来二十年的发展。



1982年

视觉任务的挑战



- 数字图像由一个个像素(pixel)组成, 每个像素的亮度、颜色等属性在计算机内表示为一个或多个数字
- 如果是黑白图像(灰度图像), 每个像素由一个字节表示, 0-255的数值表示亮度, 如果是彩色图像, 通常用RGB三个字节表示

视觉任务的挑战



Illumination



Object pose



Clutter



Occlusions



**Intra-class
appearance**



Viewpoint

视觉任务的挑战:intra-class variation



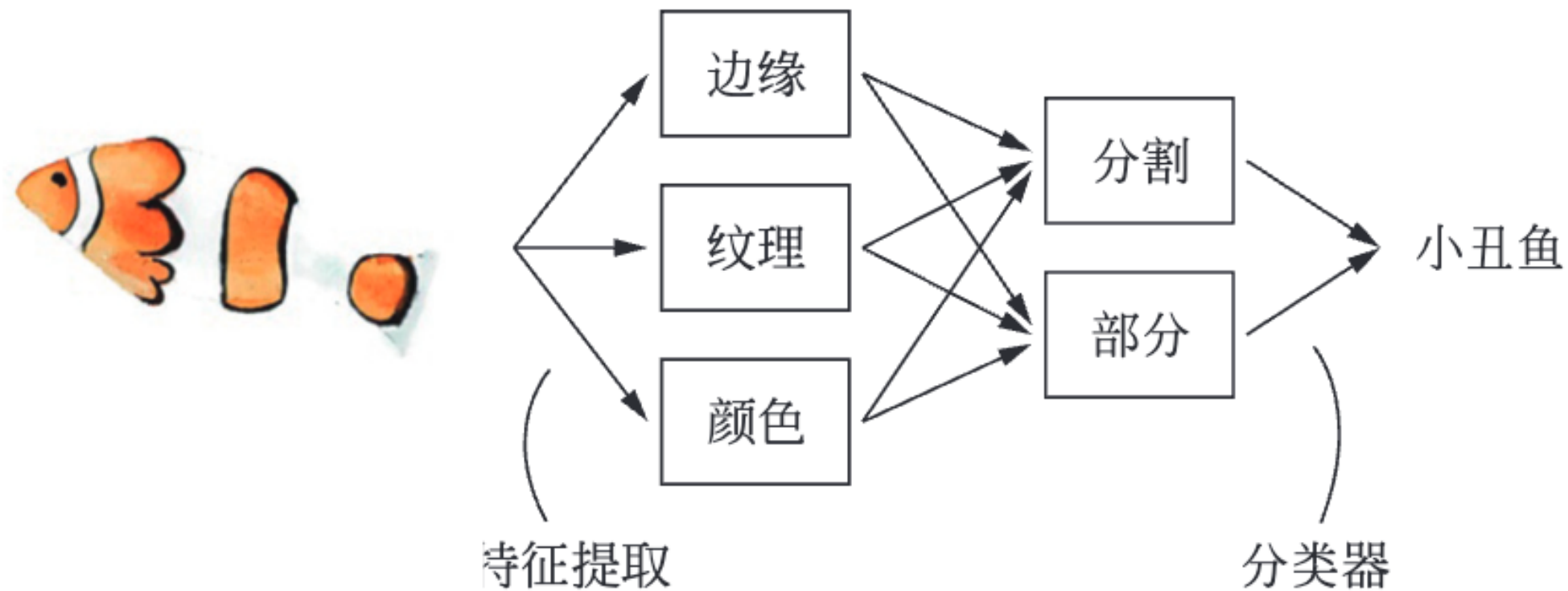
计算机视觉的关键技术

- 马尔视觉理论的核心是从图像或视频中重建物体的“几何”形状模型，围绕此目标，研究人员在此后数十年开展了大量工作
- 到2000年前后，多视几何(Multiple View Geometry)、摄像机内外参数标定、分层三维重建等方法和技术得到长足发展，形成了计算机视觉的“几何”分支，并延续至今，近年来在虚拟现实、增强现实等领域得到了广泛应用
- 另一分支便是基于机器学习，通过将计算机视觉中的问题，定义为从输入图像/视频直接到求解标签的函数拟合问题，采用数据驱动的方法来求解待拟合的函数

基于浅层模型的方法

- 步骤1: **图像预处理**: 用于实现目标对齐、几何归一化、亮度或颜色校正等处理
- 步骤2: **特征设计与提取**: 从预处理的图像中提取描述图像内容的特征
- 步骤3: **特征汇聚或特征变换**: 对提取的特征进行统计或降维处理, 从而得到更利于后续分类或回归任务的特征
- 步骤4: **分类器或回归器的设计与训练**: 采用机器学习算法训练学习器, 决策树、SVM、LR等

基于浅层模型的方法



图像分类的流程

特征设计与提取方法

- 人工设计特征本质上是一种**专家知识驱动**的方法，即研究者通过自身对所研究问题或目标的理解，设计某种流程来提取觉得“好”的特征
- 例如：人类识别早期，研究人员普遍认为应用**面部关键特征点的相对距离、角度或器官面积**等就可以区分人脸，但后来证明这些特征并不好
- 目前，多数人工设计的特征有两类，即**全局特征**和**局部特征**
 - 全局特征：建模图像中全部像素或多个不同区域像素蕴含的信息
 - 局部特征：从一个局部区域内的少量像素中提取信息

各种算子

Roberts算子

$$S_x = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, S_y = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

Prewitt算子

$$S_x = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}, S_y = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix}$$

Sobel算子

$$S_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, S_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

Canny算子

Laplacian算子

...

各种算子



原始图片



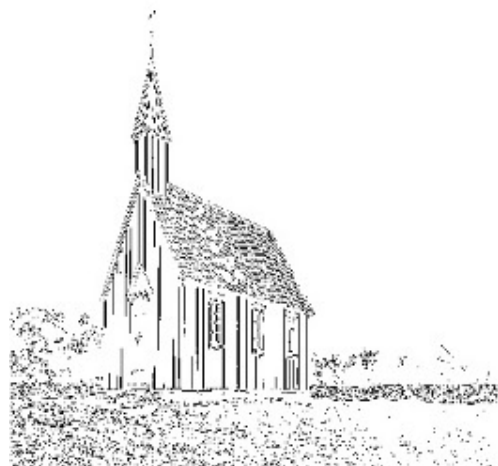
Roberts算子



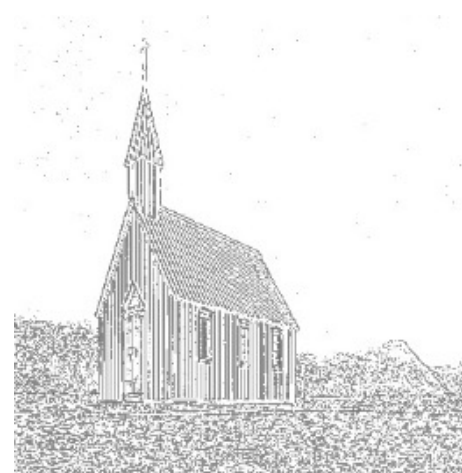
Prewitt算子



Sobel算子

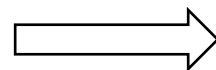
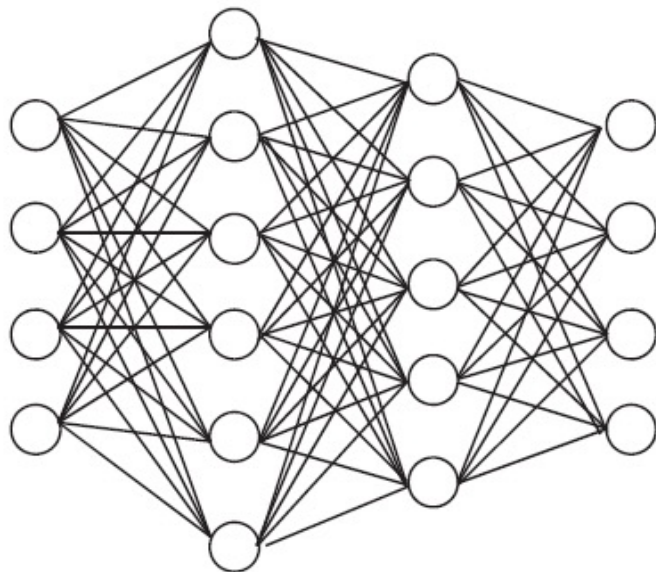
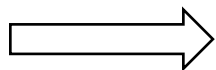


Canny算子



Laplacian算子

全连接前馈网络处理图像的问题



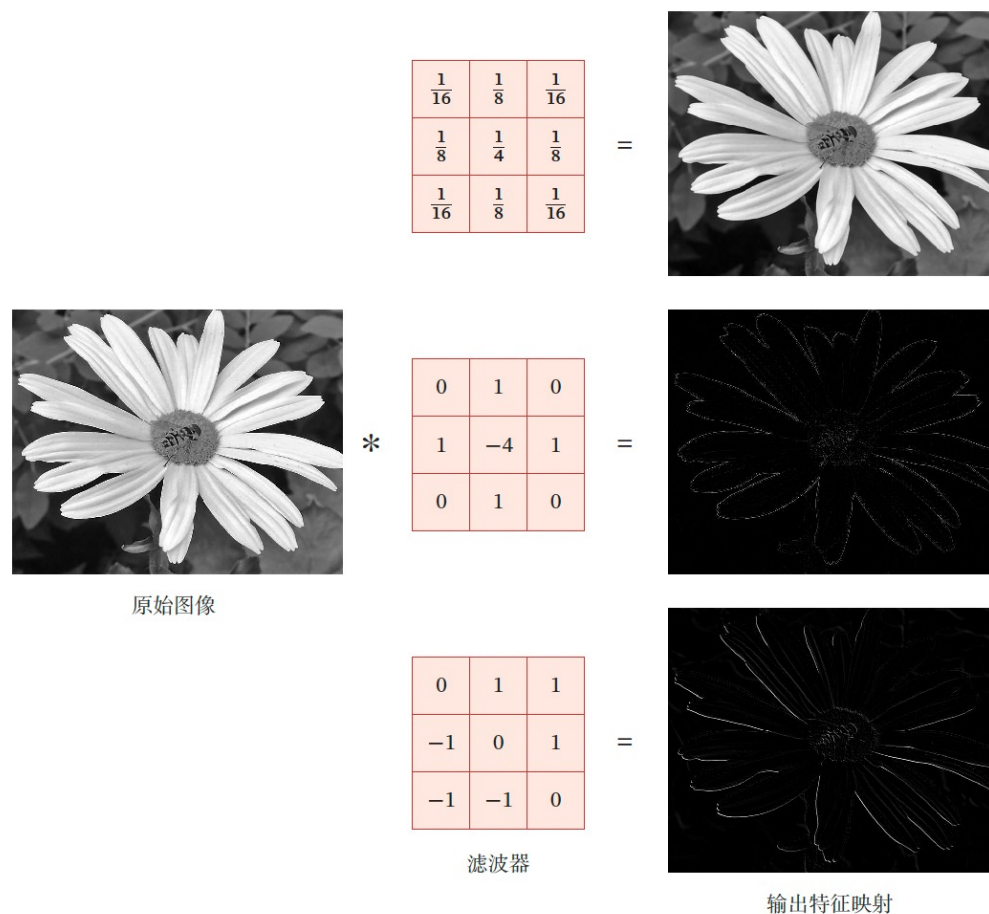
Cat
Dog

- 使用全连接前馈网络处理图像的问题：

- **参数太多**：如果输入数据为 1000×1000 的图像，需要转换为 $1,000,000$ 维的向量，如果第一个隐藏层与输入维度一致，则输入层到第一个隐藏层的待训练参数为 10^{12} 个
- **局部不变性特征**：自然图像中的物体都具有局部不变形特征，比如尺度缩放、平移、旋转不影响语义，前馈网络很难提取局部不变性特征

卷积神经网络(Convolutional Neural Network, CNN)

- **卷积层**: 每个卷积层包含多个特征映射, 每个特征映射通过一种**卷积滤波器**提取一种数据的特征 (**特征提取**)



卷积神经网络(Convolutional Neural Network, CNN)

- **卷积层**：每个卷积层包含多个特征映射，每个特征映射通过一种**卷积滤波器**提取一种数据的特征（**特征提取**）

1 _{x1}	1 _{x0}	1 _{x1}	0	0
0 _{x0}	1 _{x1}	1 _{x0}	1	0
0 _{x1}	0 _{x0}	1 _{x1}	1	1
0	0	1	1	0
0	1	1	0	0

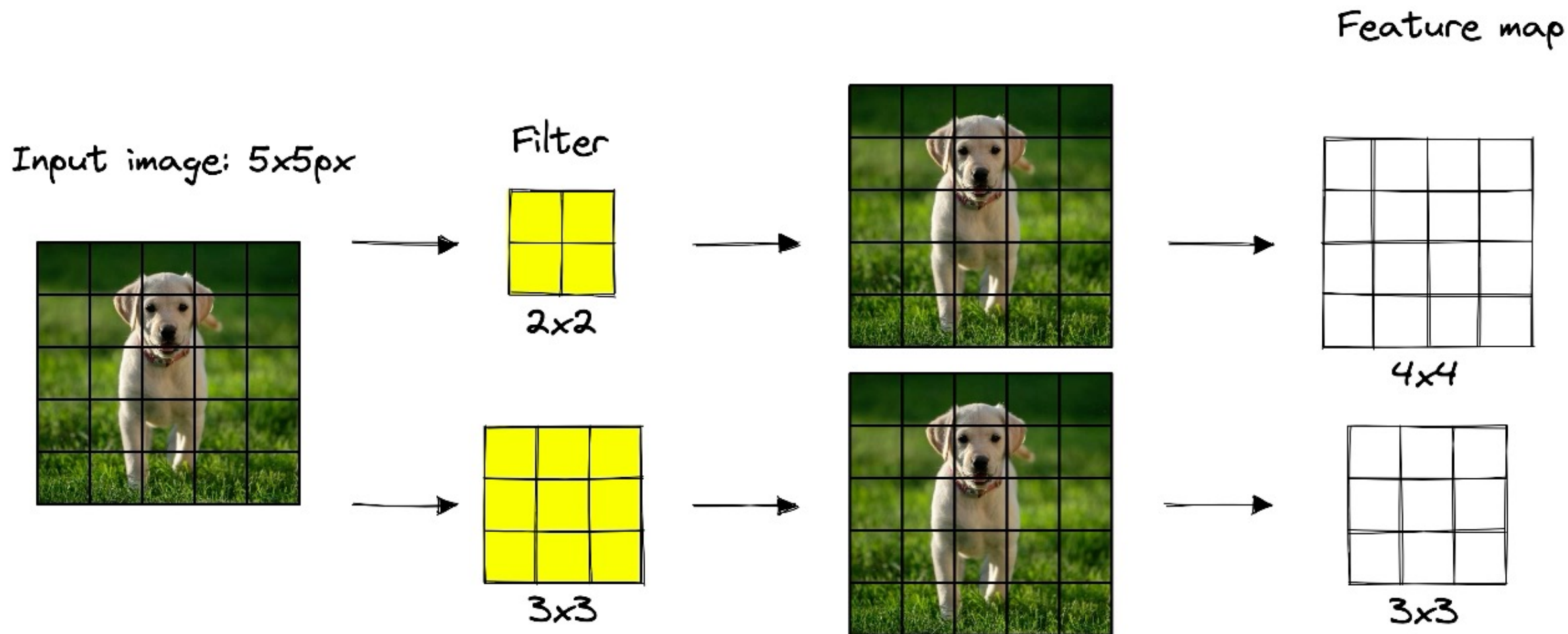
Image

4		

Convolved
Feature

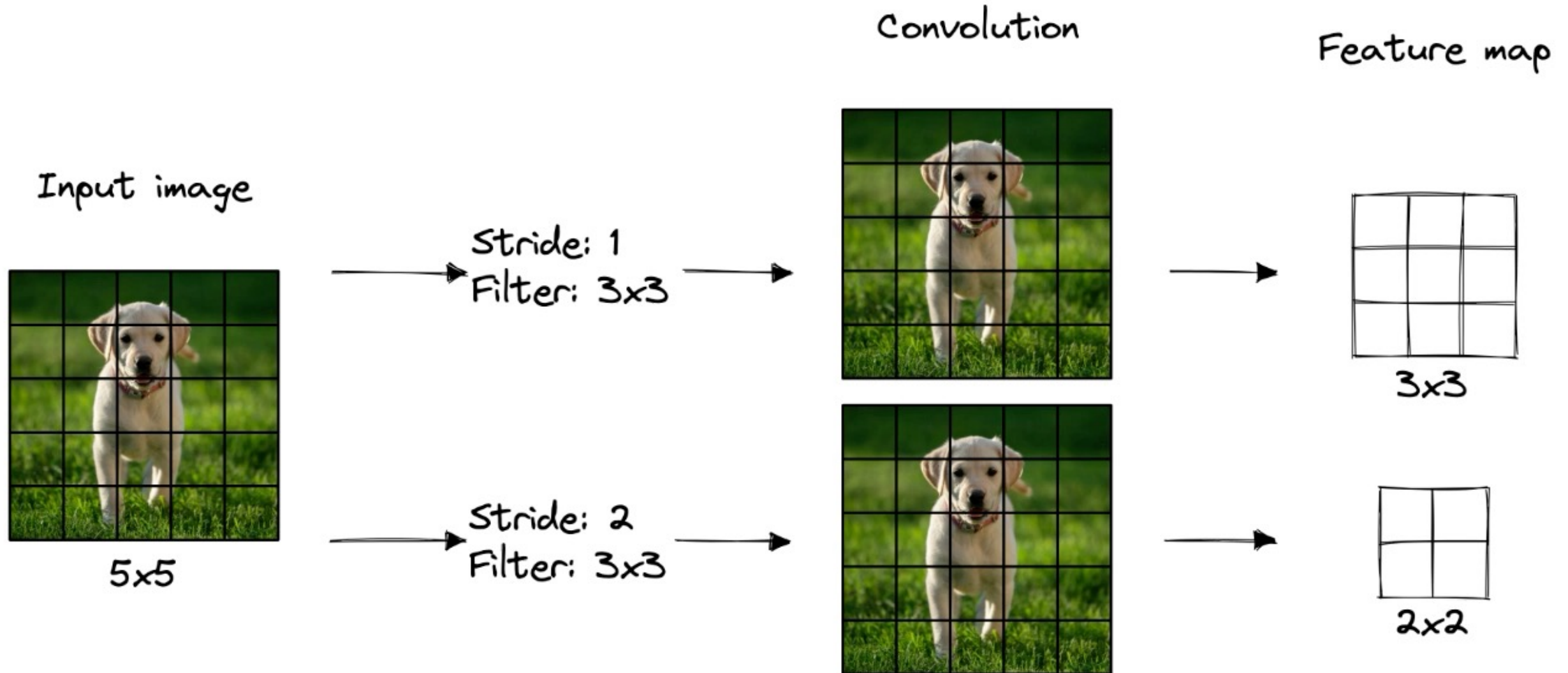
卷积神经网络(Convolutional Neural Network, CNN)

- **卷积层**: 每个卷积层包含多个特征映射, 每个特征映射通过一种卷积滤波器提取一种数据的特征 (**特征提取**)



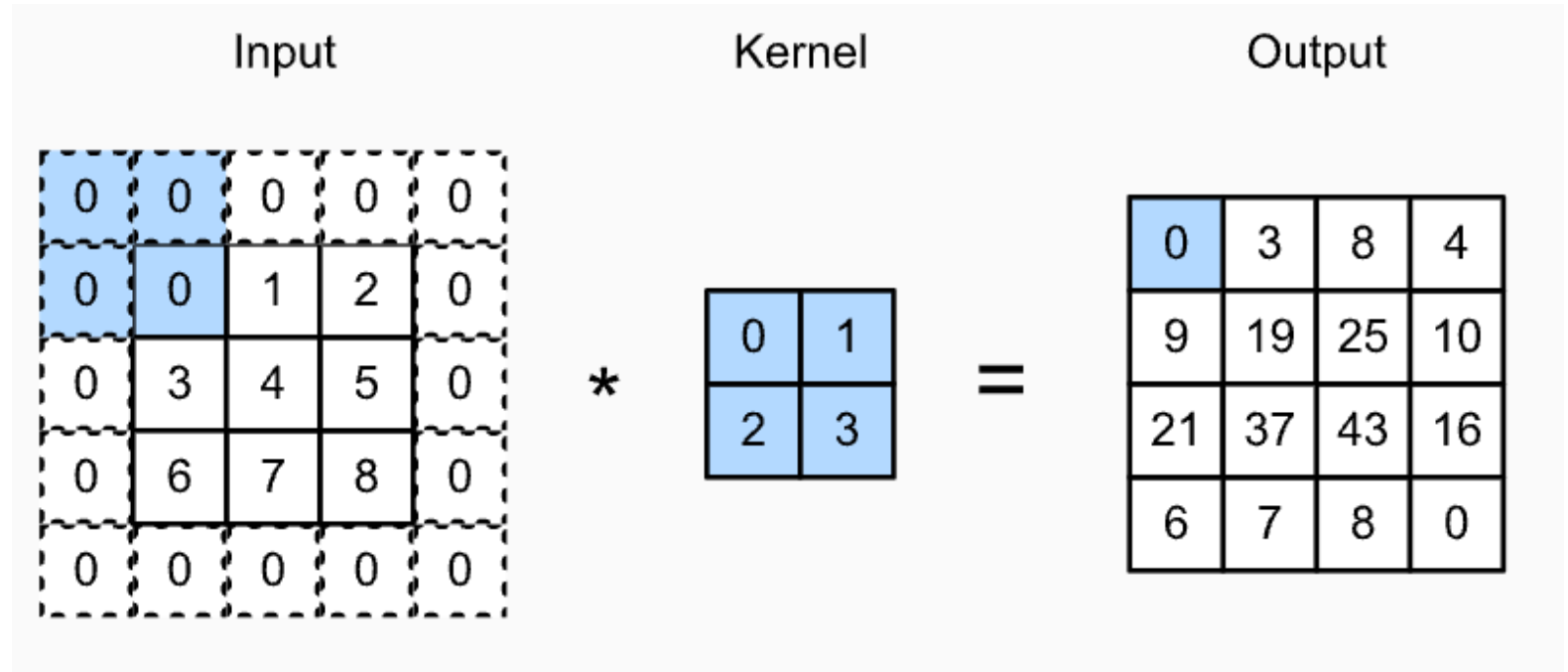
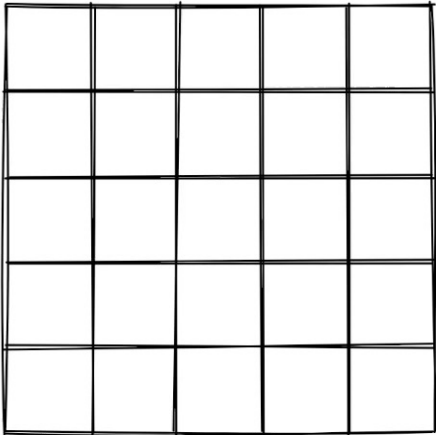
卷积神经网络(Convolutional Neural Network, CNN)

- 步长(Stride)



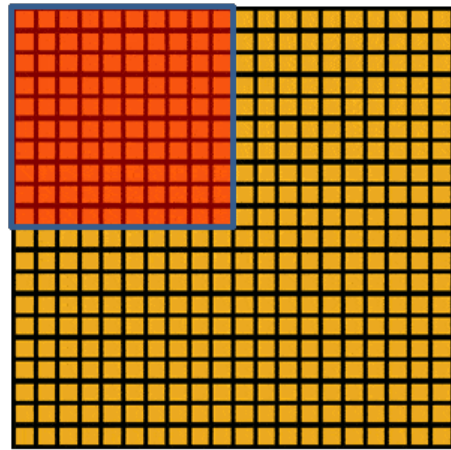
卷积神经网络(Convolutional Neural Network, CNN)

- 填充(Padding)

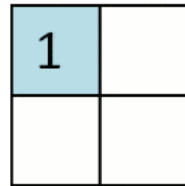


卷积神经网络(Convolutional Neural Network, CNN)

- **采样层**：亦称“池化(Pooling)层”，其作用是基于局部相关性原理进行亚采样，从而在减少数据量的同时保留有用信息（降低参数量级）



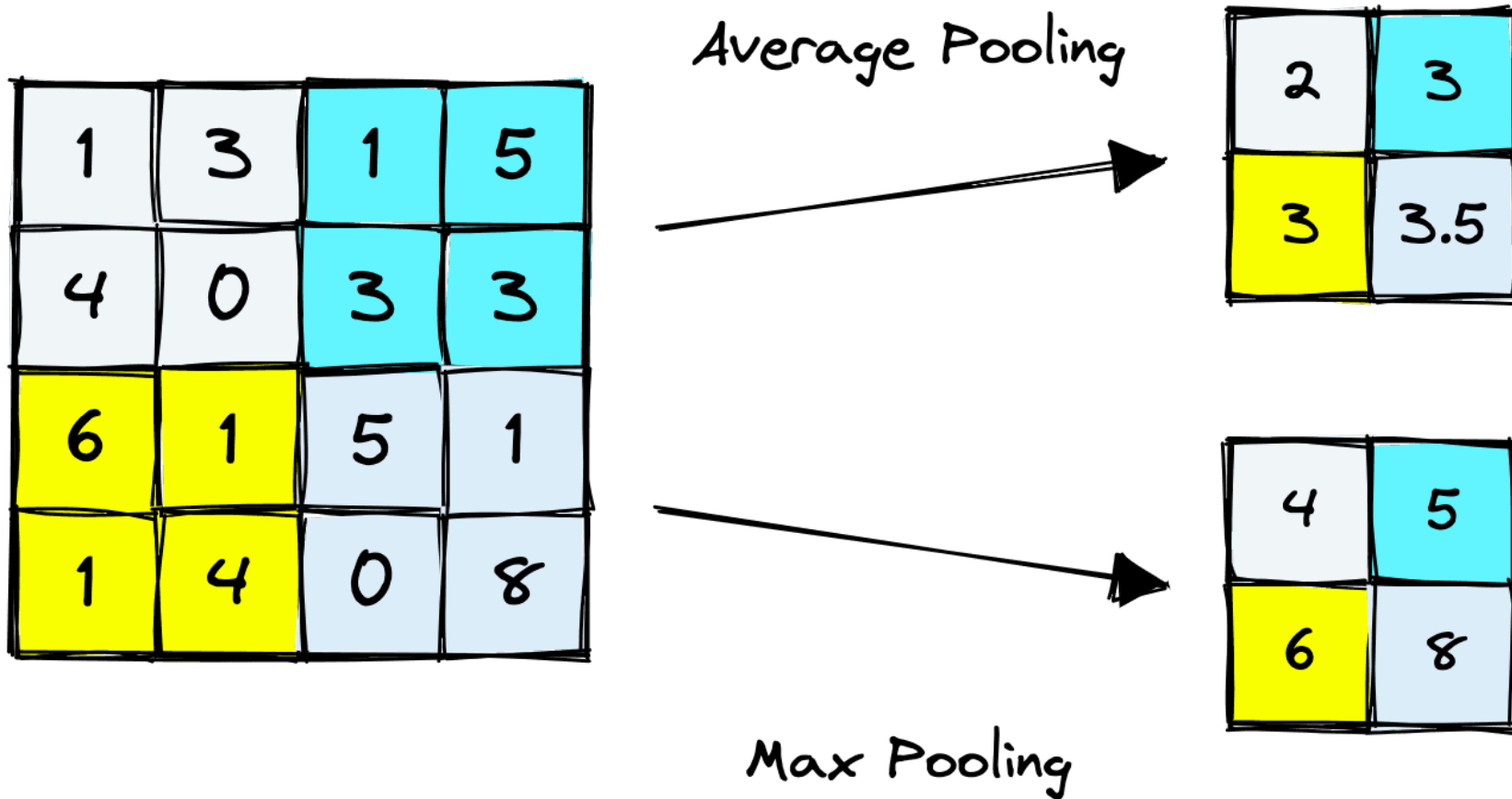
Convolved
feature



Pooled
feature

卷积神经网络(Convolutional Neural Network, CNN)

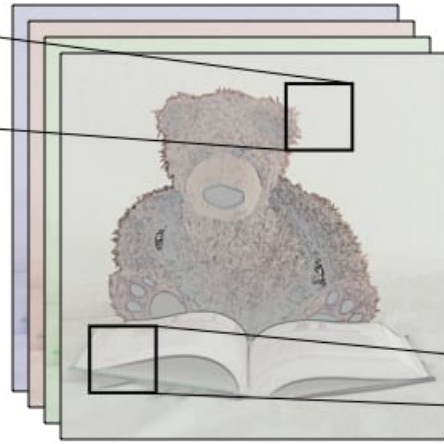
- **采样层**：亦称“池化层”，其作用是基于局部相关性原理进行亚采样，从而在减少数据量的同时保留有用信息（降低参数量级）



卷积神经网络(Convolutional Neural Network, CNN)



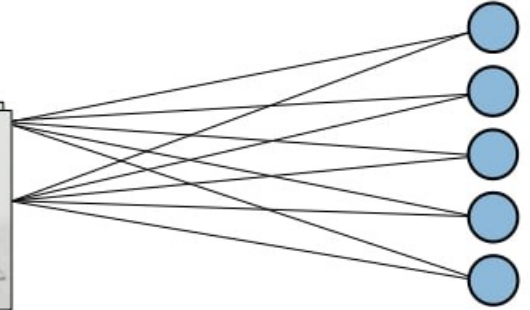
Input image



Convolutions



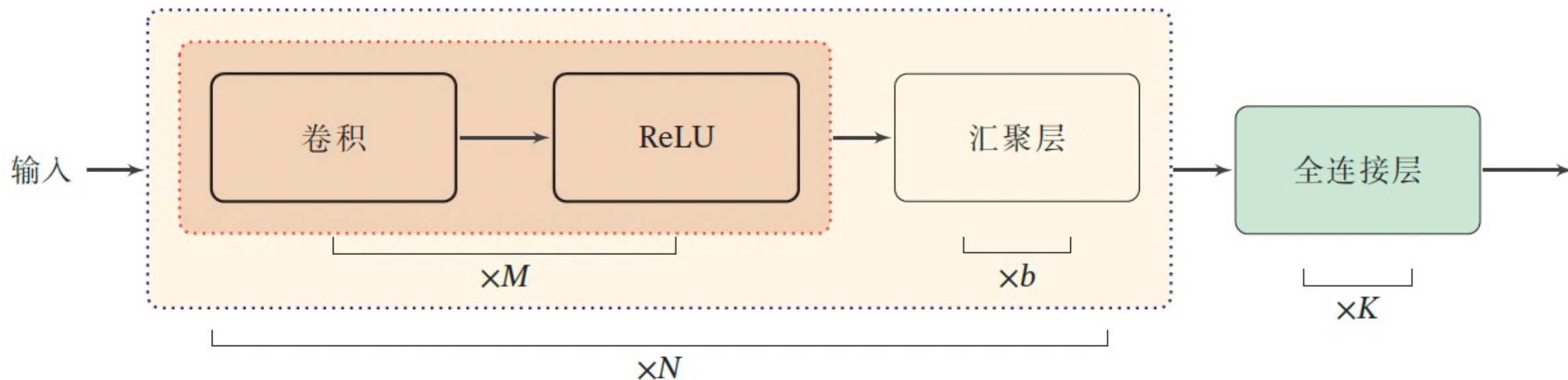
Pooling



Fully Connected

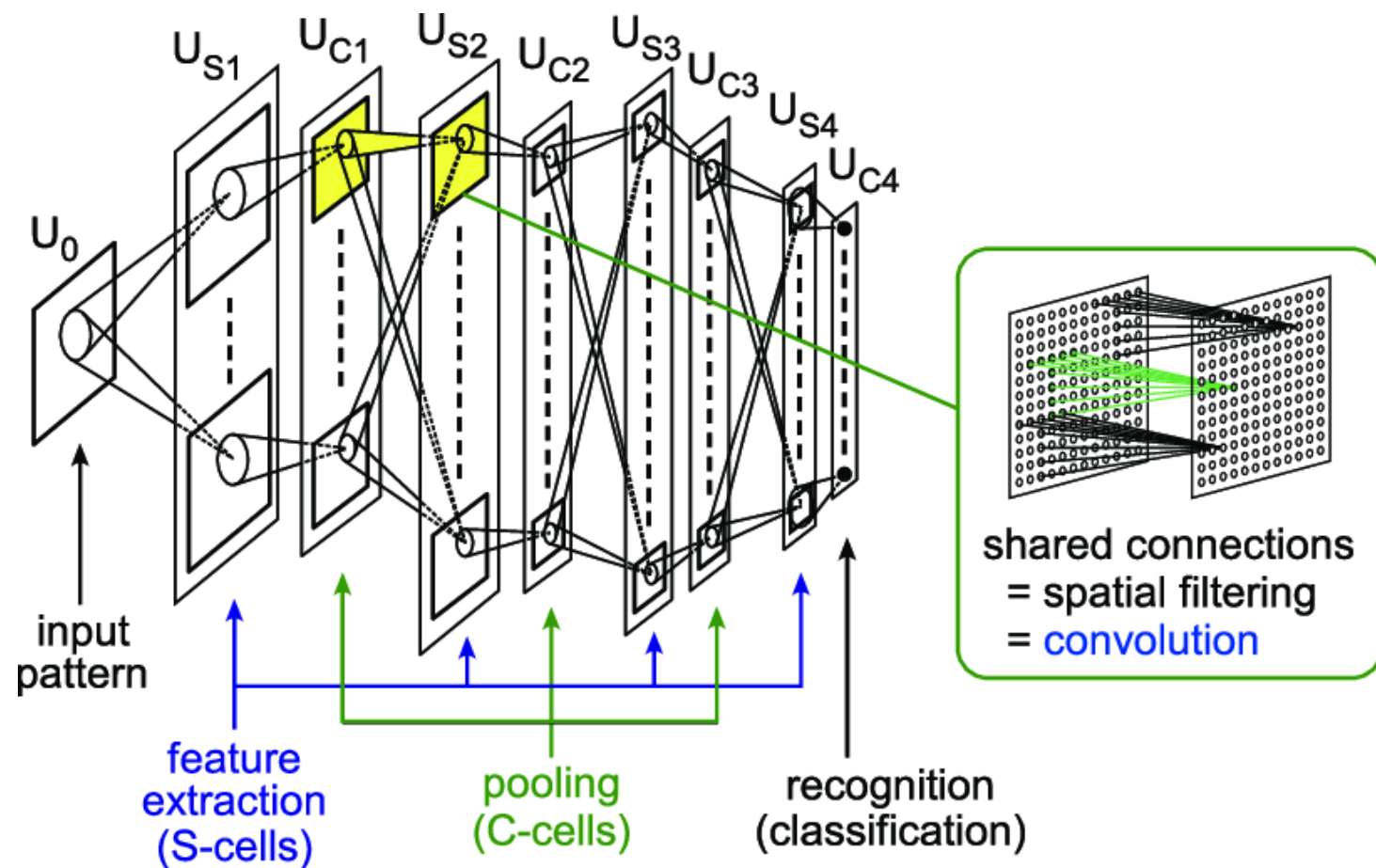
卷积神经网络(Convolutional Neural Network, CNN)

- 卷积神经网络：一般由卷积层、采样层和全连接层构成的前馈神经网络，具有局部连接、权值共享、汇聚等特性，使得卷积神经网络具有一定程度的平移、缩放和旋转不变性，同时与前馈神经网络相比，参数更少



卷积神经网络(1980-)

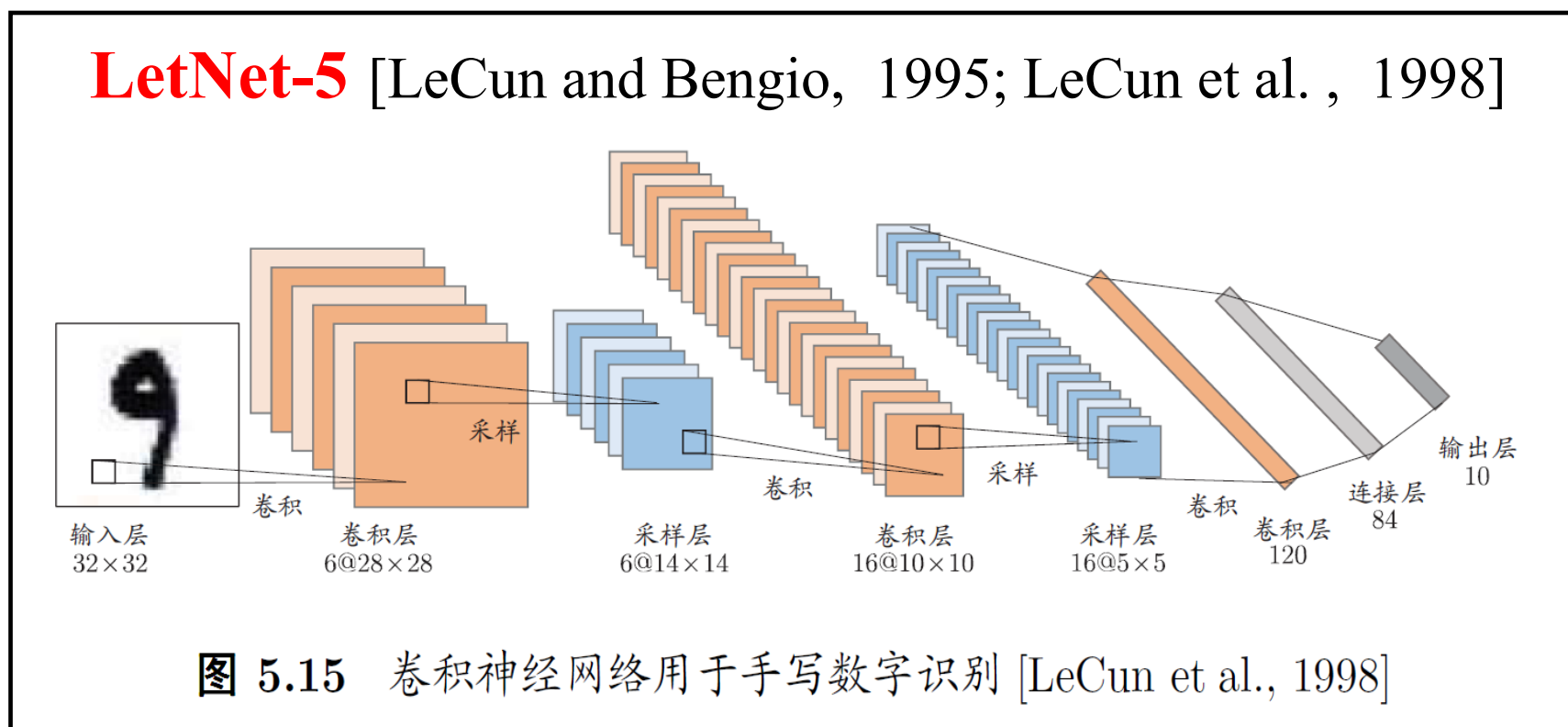
Neocognitron



Kenji Fukushima

卷积神经网络(1980-2010)

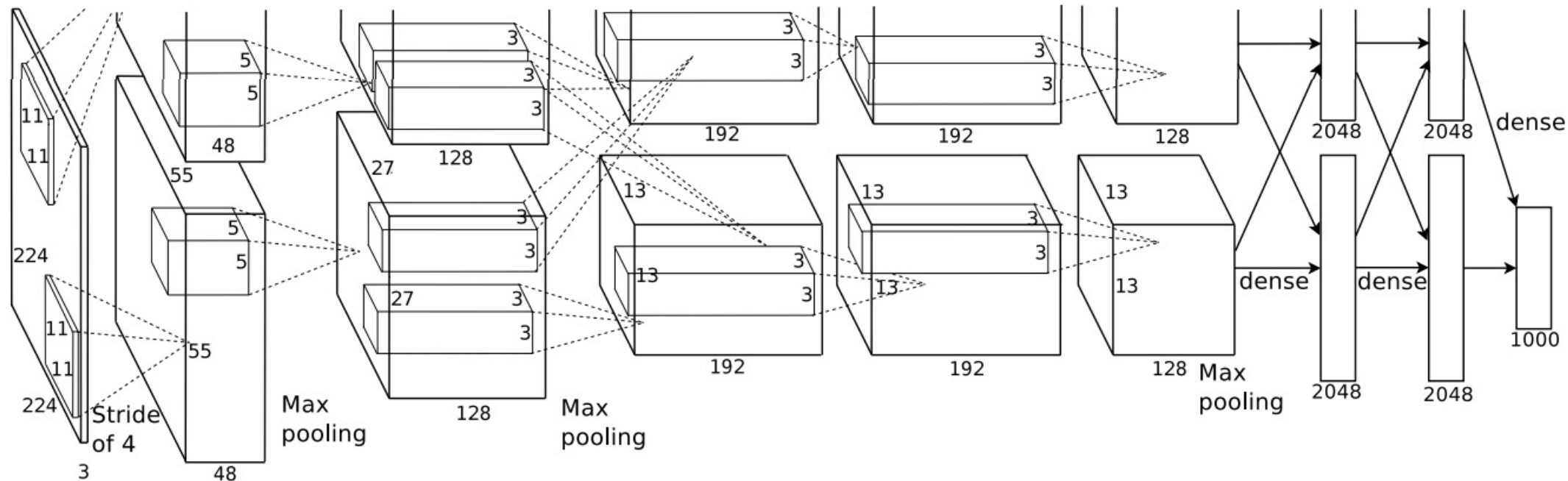
在20世纪80年代末和90年代初，Yann LeCun和他的团队进一步开发了CNN，引入了专门为手写数字识别设计的LeNet-5架构，应用于美国支票识别业务



Yann LeCun

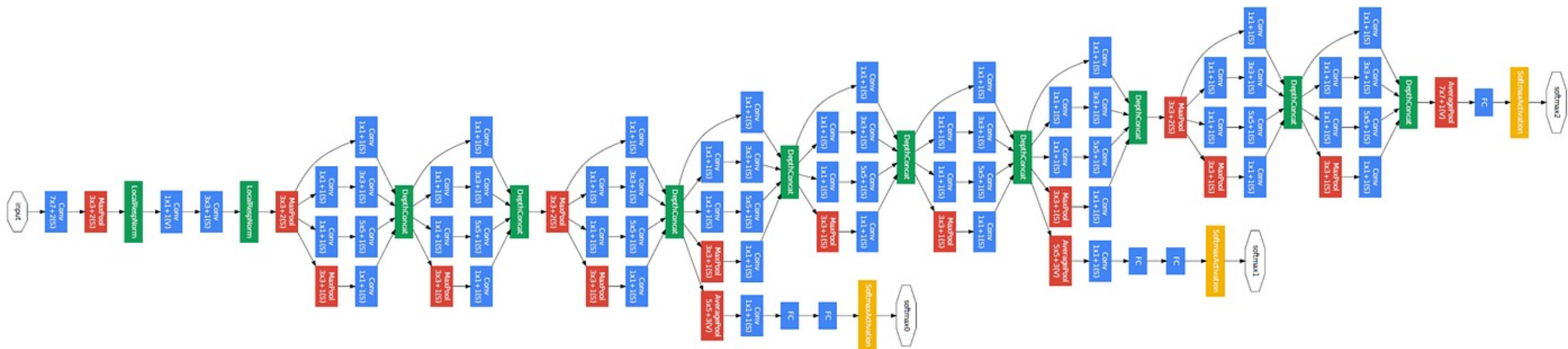
卷积神经网络-AlexNet

- 2012 ILSVRC winner (top 5 error of 16% compared to runner-up with 26% error)
- 第一个现代深度卷积网络模型
- 首次使用了很多现代深度卷积网络的一些技术方法
- 使用GPU进行并行训练，采用了ReLU作为非线性激活函数
- 使用Dropout防止过拟合，使用数据增强



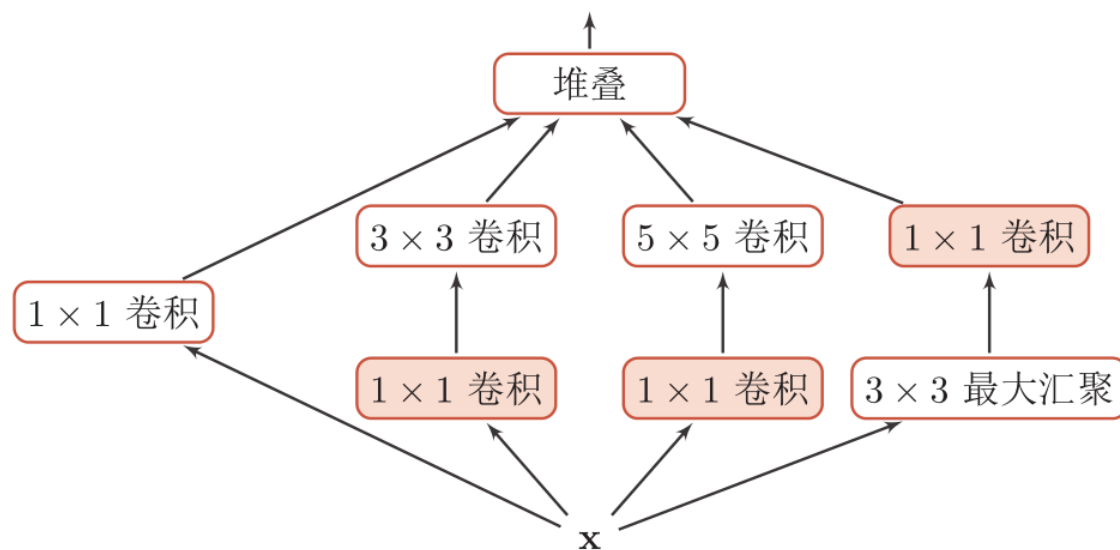
卷积神经网络-InceptionNet

- 2014 ILSVRC winner (22层)
- 参数: GoogLeNet: 4M VS AlexNet: 60M
- 错误率: 6.7%
- Inception网络是由有多个inception模块和少量的汇聚层堆叠而成



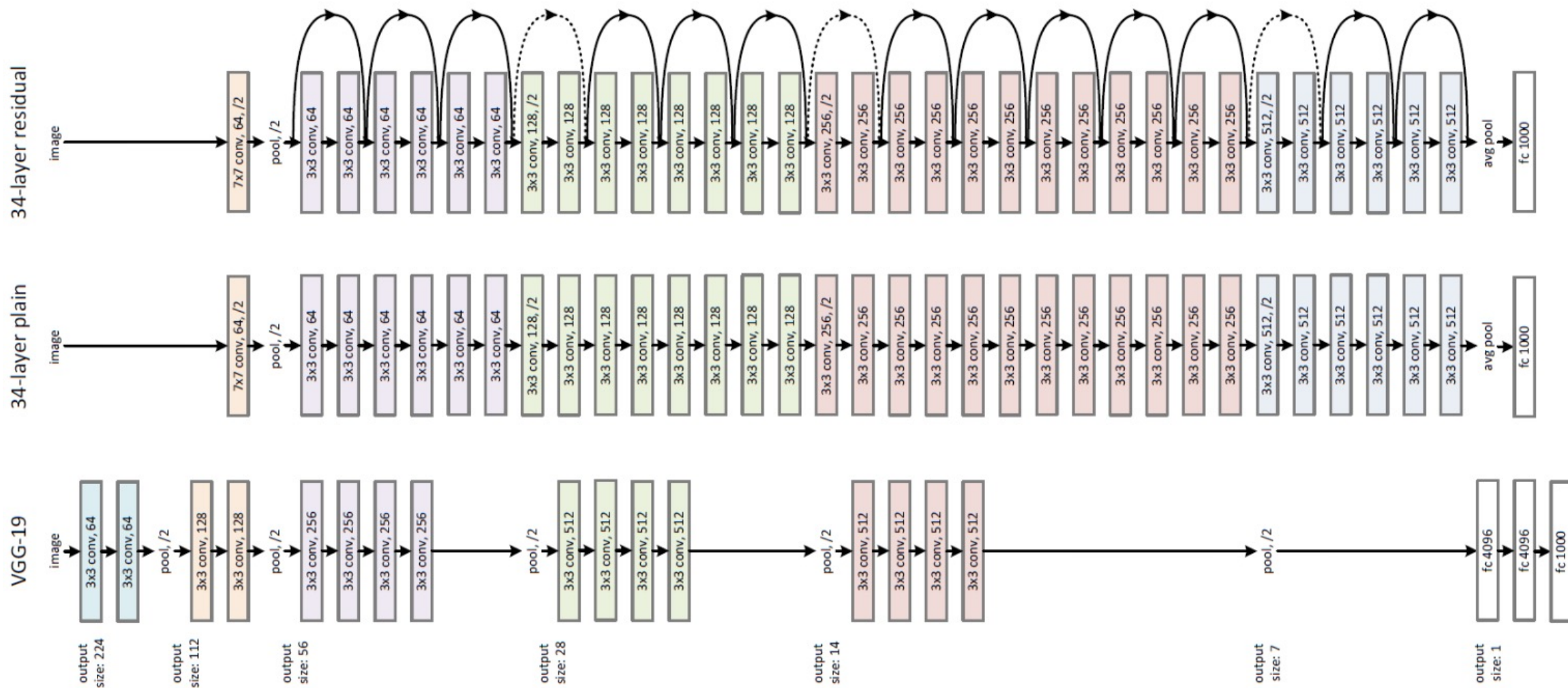
卷积神经网络-InceptionNet

- 在卷积网络中，如何设置卷积层的卷积核大小是一个十分关键的问题
- 在Inception网络中，一个卷积层包含多个不同大小的卷积操作，称为Inception模块
- Inception模块同时使用 1×1 、 3×3 、 5×5 等不同大小的卷积核，并将得到的特征映射在深度上拼接（堆叠）起来作为输出特征映射



卷积神经网络-ResNet

- 2015 ILSVRC winner (152层)
- 错误率: 3.57%



实例：CIFAR-10 图像分类

<https://www.kaggle.com/c/cifar-10>



CIFAR-10 - Object Recognition in Images

Identify the subject of 60,000 labeled images

231 teams · 4 years ago

[Overview](#) [Data](#) [Discussion](#) [Leaderboard](#) [Rules](#)

Overview

Description

Evaluation

CIFAR-10 is an established computer-vision dataset used for object recognition. It is a subset of the **80 million tiny images dataset** and consists of 60,000 32x32 color images containing one of 10 object classes, with 6000 images per class. It was collected by Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton.

实例：ImageNet狗品种识别

<https://www.kaggle.com/c/dog-breed-identification>



扩展阅读

- 《动手学深度学习》：<https://zh.d2l.ai/>
- Fei-Fei Li's Lecture: <http://cs231n.stanford.edu/>
- 常用数据集：
- 图像分类：ImageNet、CIFAR、MNIST
- 目标检测：COCO数据集
- 语义分割：Pascal VOC2012
- 视频数据集：Kinetics-700
- ...

大纲

- 神经元模型到前馈神经网络
- 参数优化：BP算法
- 深度学习
- 计算机视觉与卷积神经网络
- 自然语言处理与循环神经网络**
- 多模态学习

自然语言处理(Natural Language Processing)

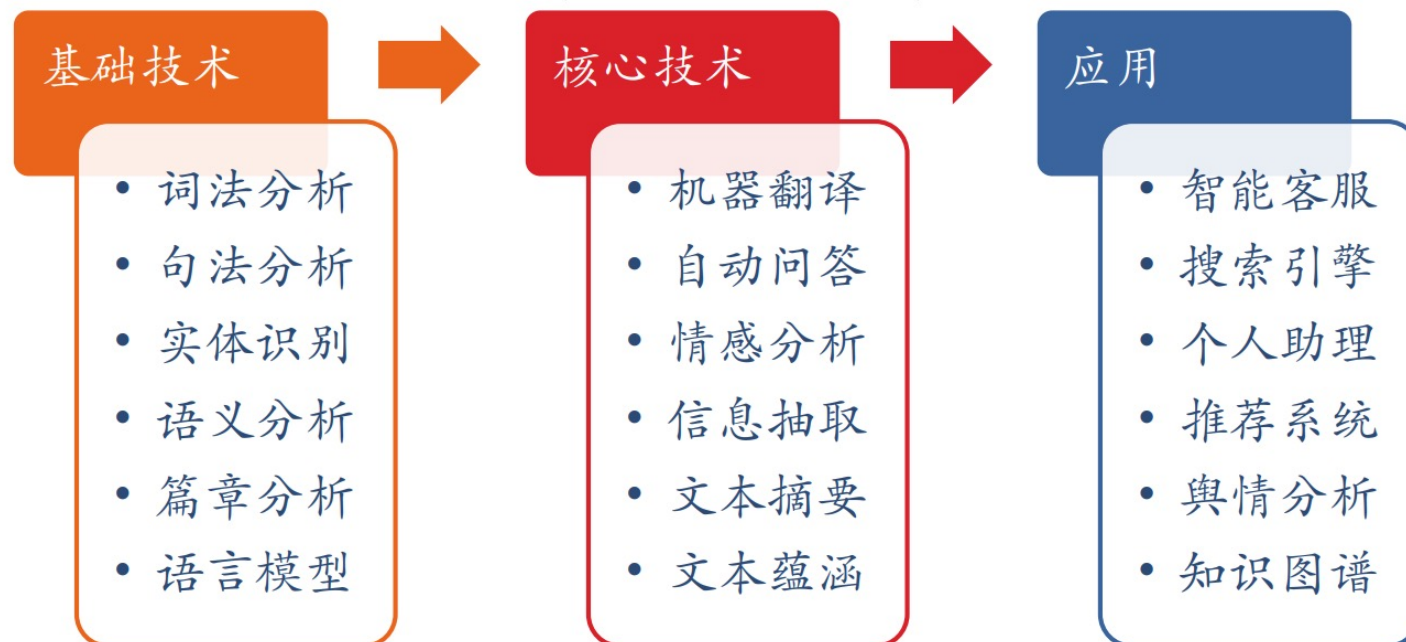
什么是自然语言处理

自然语言处理有哪些任务

如何用人工智能技术解决自然语言处理任务

自然语言处理(Natural Language Processing)

- 自然语言：**人类语言**，区别于人工语言(比如程序语言)
- 自然语言处理：对词、句子、篇章进行分析，对内容进行理解，并在此基础上支持一系列技术，如翻译、问答等



自然语言处理的难点：歧义性

- 组合歧义：
 - 两个/人/一起/过去、个人/问题
 - 从马/上/下来、马上/就来
 - 南京市长江大桥
- 句子级歧义：
 - 白天鹅在水里游泳
 - 该研究所获得的成果

自然语言处理的难点：歧义性

灵魂八问

配钥匙师傅：你配吗？

食堂阿姨：你要饭吗？

算命先生：你算什么东西？

快递小哥：你是什么东西？

上海垃圾分拣阿姨：你是什么垃圾？

滴滴司机：你搞清楚你自己的定位了么？

理发师傅：你自己照照镜子看看你自己，觉得还行么？

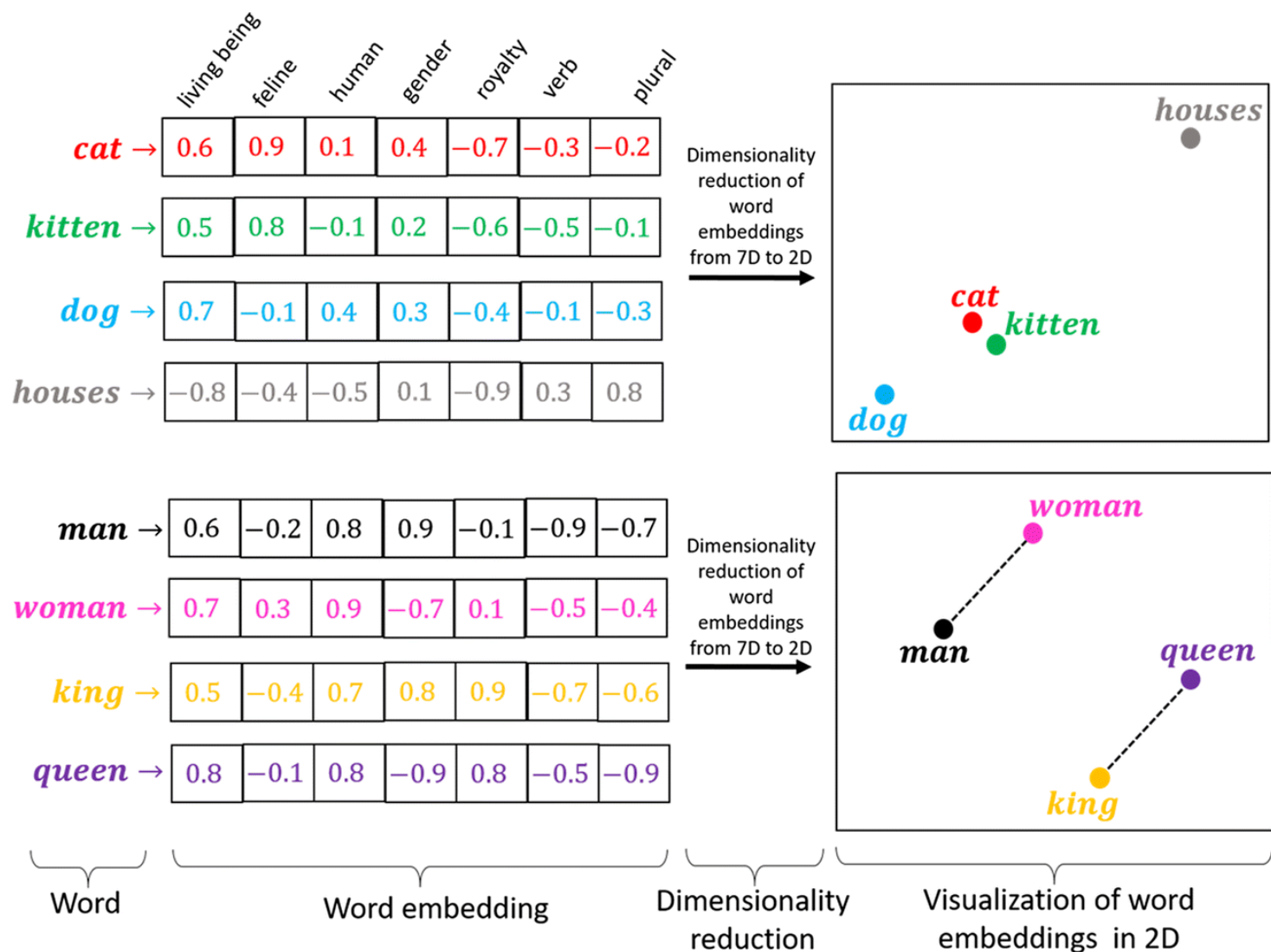
小区保安：你是谁？你从哪里来？要到哪里去？

词的表示

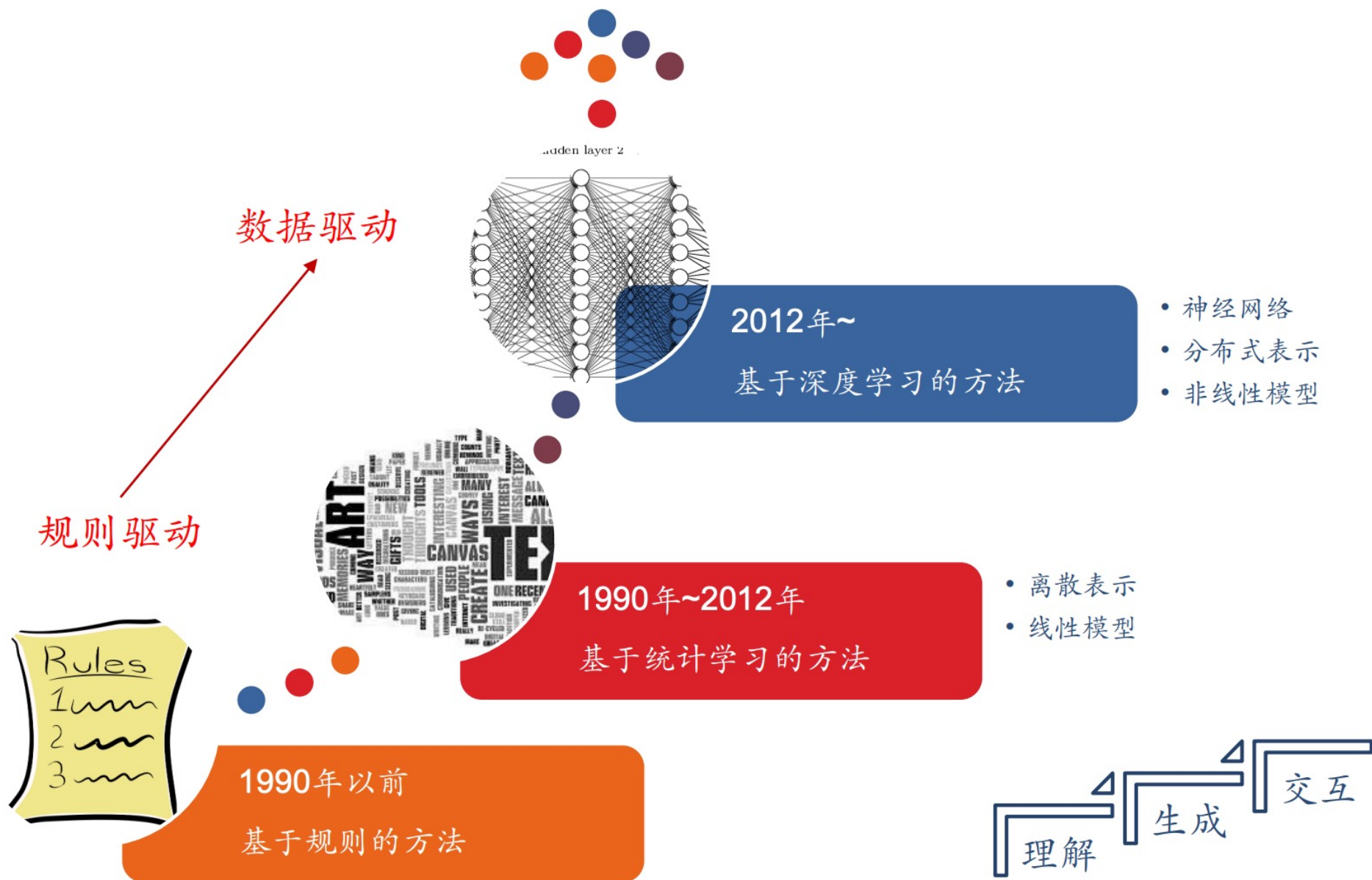
- 传统表示方法：独热向量(One-Hot Vector)
 - 每个词变成一个长向量
 - 南 = [0, 0, 0, 0, 1, 0, 0]
 - 京 = [0, 0, 0, 1, 0, 0, 0]
- 向量的维数通常很大：每一维对应一个词，总维数对应字典中词的数目
- 难以看出词之间的关系：词之间的距离与语义相似性无关

词的表示: Word Embedding

- 为不同的词学习一个 embedding vector
- 语义相近的词在 embedding 空间内具有相似的距离
- 代表性算法: Word2Vec



自然语言处理的发展历史



基于规则的方法

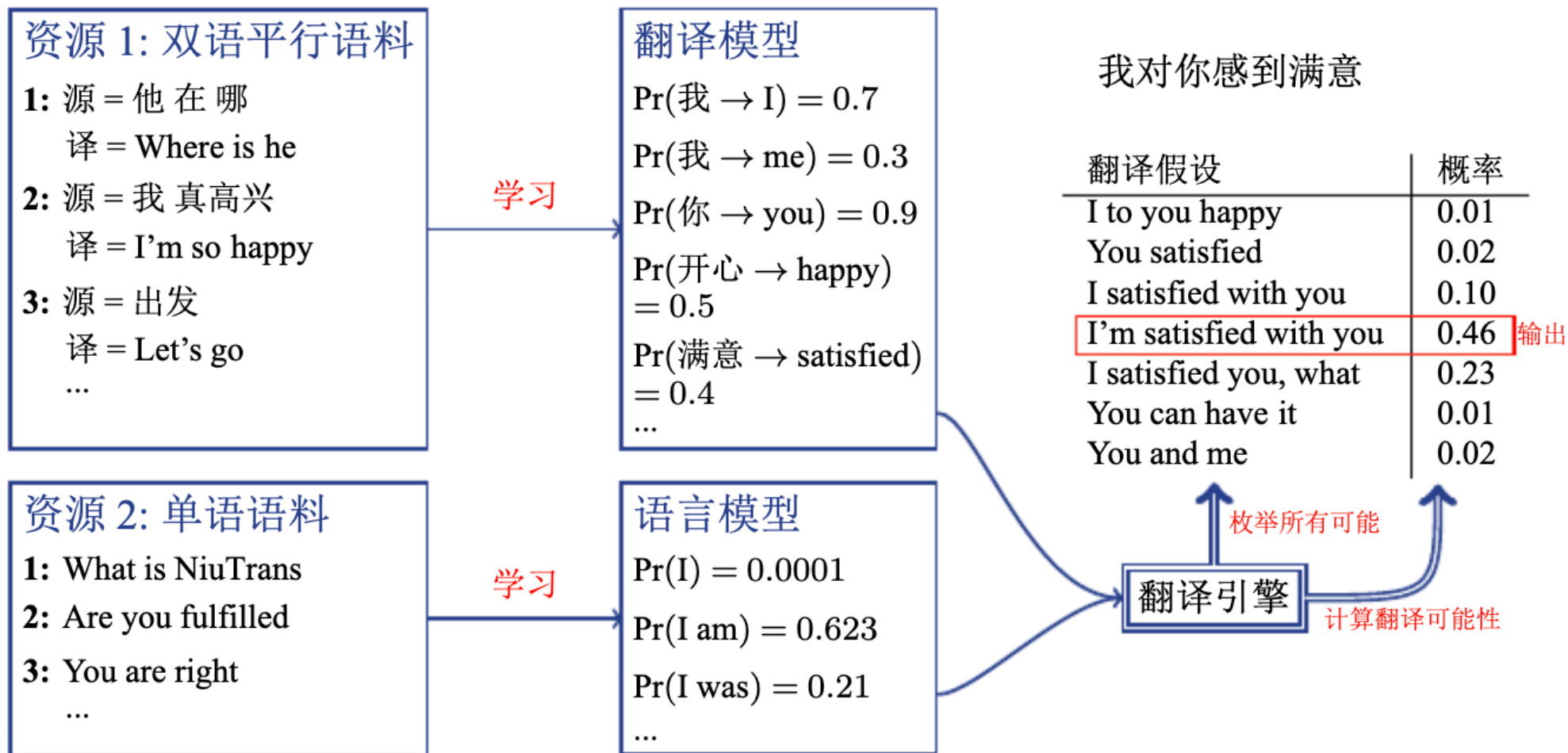
资源：规则库

- 1: **If** 源 = 我, **then** 译 = I
- 2: **If** 源 = 你, **then** 译 = you
- 3: **If** 源 = 感到 满意,
then 译 = be satisfied with
- 4: **If** 源 = 对... 动词 [表态度]
then 调序 [动词 + 对象]
- 5: **If** 译文主语是 I
then be 动词为 am/was
- 6: **If** 源语是主谓结构
then 译文为主谓结构



规则 5

数据驱动的方法



翻译模型从双语平行语料中学习翻译知识，得到各种单词的翻译及其概率

语言模型从单语语料中学习目标语言的词序列生成规律，来衡量目标语言译文的流畅性

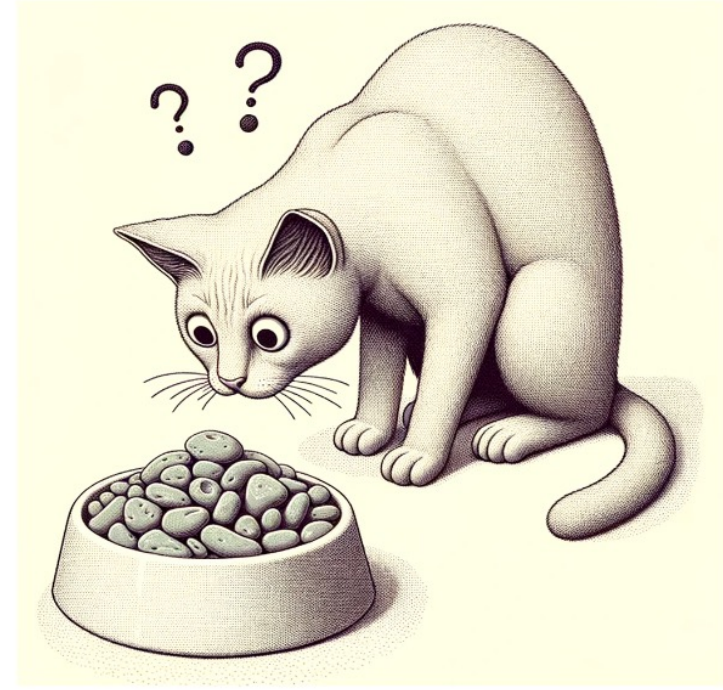
语言模型(Language Model)



The cat drank a bowl of diet coke.



The cat drank a bowl of milk.



The cat drank a bowl of rocks.

Which Sentence is most likely?

语言模型(Language Model)

- 南京大学在南京
 - 南京大学在苏州
 - 南京在南京大学
-
- 英语: You go first
 - 中文: 你先走 vs 你走先

如何学习语言模型？

- 给定文字序列 (word sequence)
 - w_1, w_2, \dots, w_n
 - 语言模型可以估计 $P(w_1, w_2, \dots, w_n)$
- Chain rule
 - $P(w_1, w_2, \dots, w_n) = P(w_n | w_1, \dots, w_{n-1}) P(w_1, \dots, w_{n-1})$
- $P(\text{南京大学}) = P(\text{南})P(\text{京}|\text{南})P(\text{大}|\text{南京})P(\text{学}|\text{南京大})$

N元语法(N-Gram)

- 序列越长，计算和存储多个词共同出现概率的复杂度指数增加
- 近似：引入马尔科夫假设

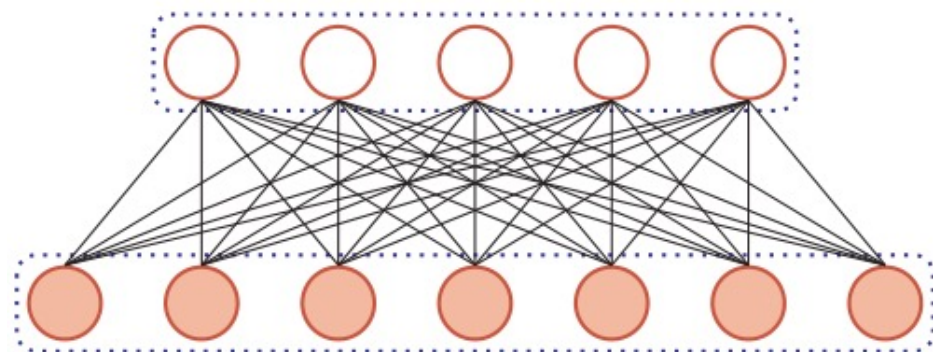
$$P(w_i | w_1, \dots, w_{i-1}) \approx P(w_i | w_{i-n+1}, \dots, w_{i-1})$$

一阶：P(南京大学) = P(南)P(京|南)P(大|京)P(学|大)

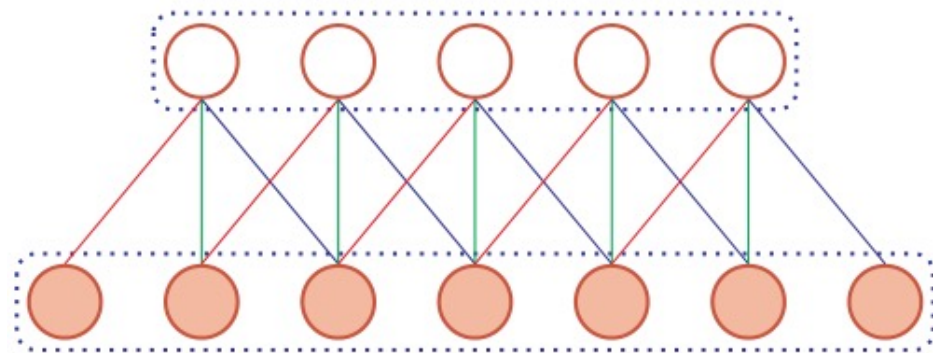
二阶：P(南京大学) = P(南)P(京|南)P(大|南京)P(学|京大)

前馈神经网络处理序列数据

- 连接存在层与层之间，每层的节点之间是无连接的
- 不适用于自然语言等序列数据



(a) 全连接层

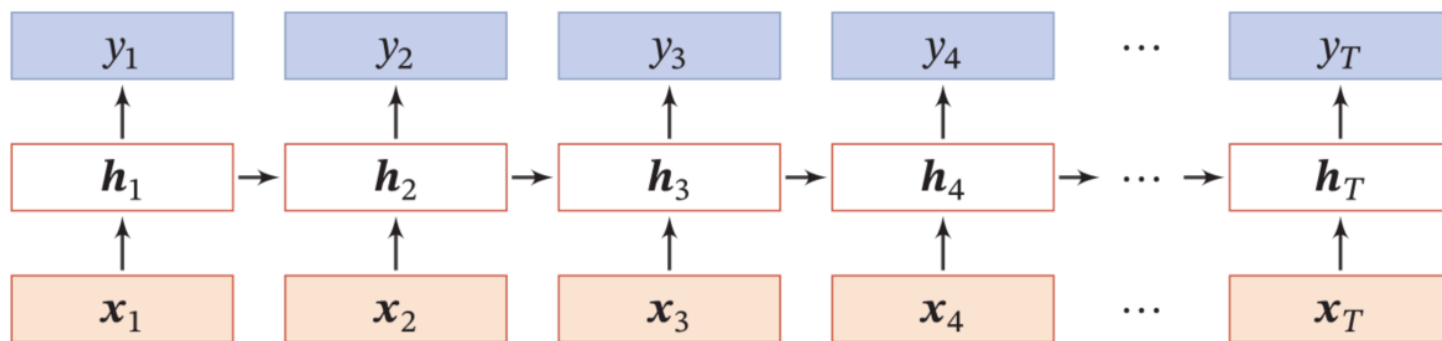
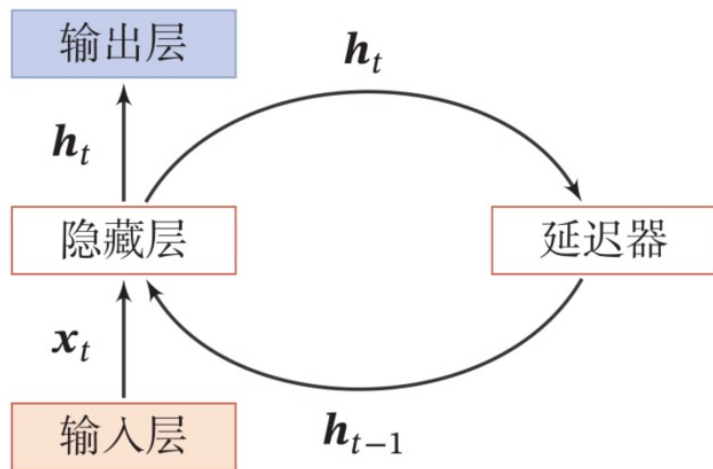


(b) 卷积层

循环神经网络(Recurrent Neural Network, RNN)

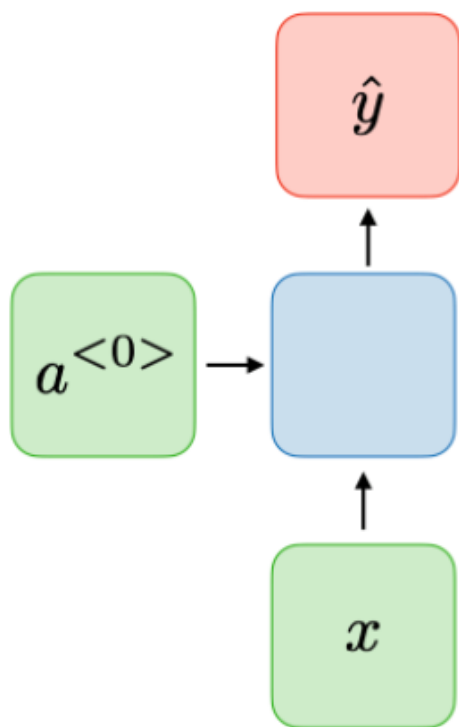
循环神经网络通过使用带自反馈的神经元，能够处理任意长度的时序数据

$$h_t = f(h_{t-1}, x_t)$$

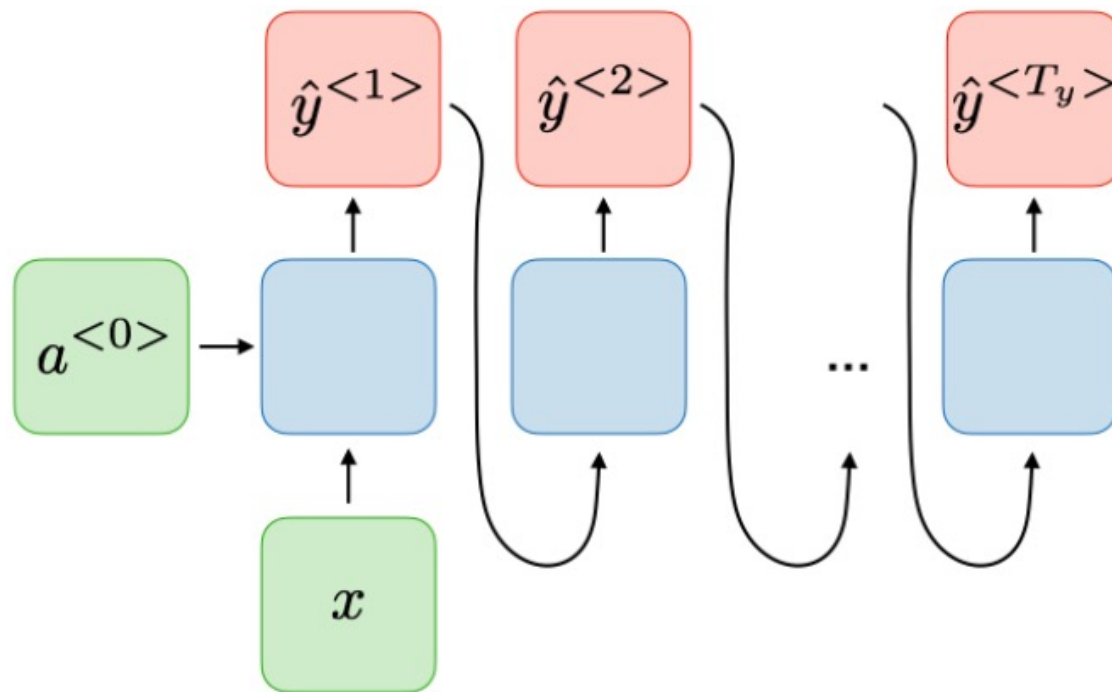


循环神经网络(Recurrent Neural Network, RNN)

One-to-one
传统神经网络

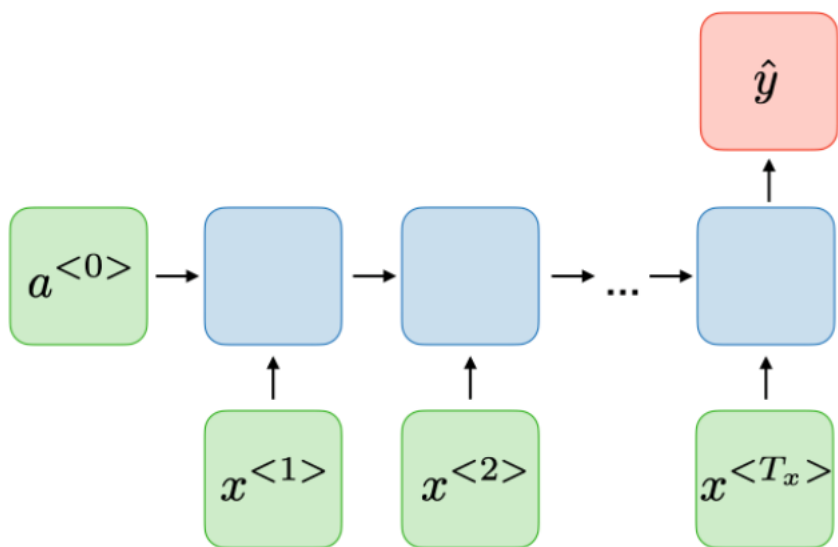


One-to-Many
序列生成

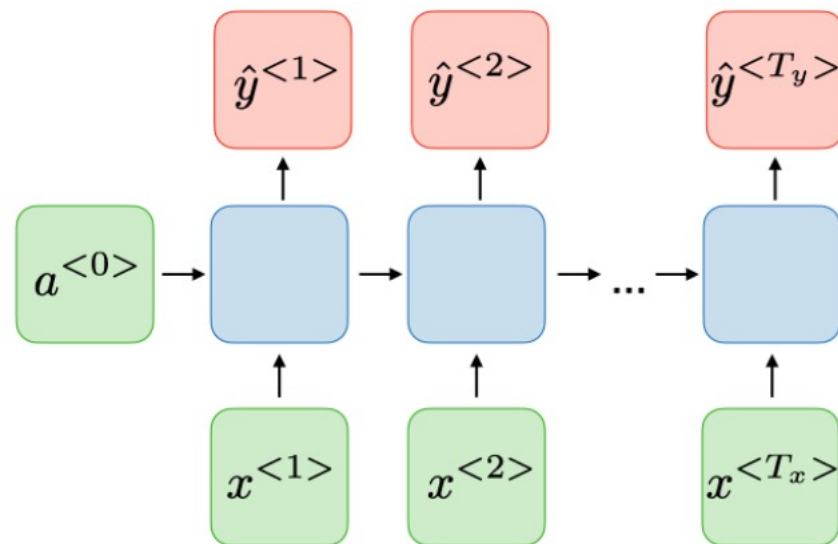


循环神经网络(Recurrent Neural Network, RNN)

Many-to-one
序列分类

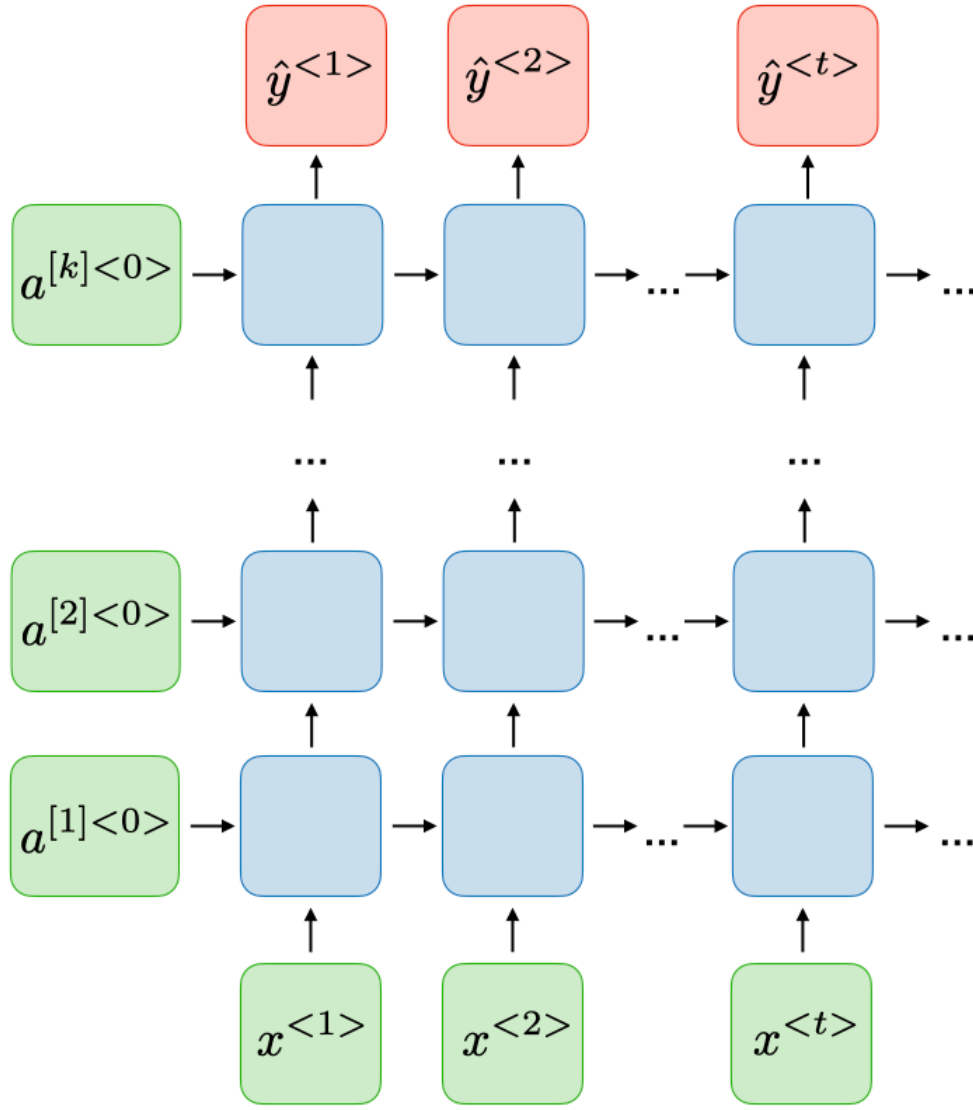


Many-to-Many
语言模型、机器翻译

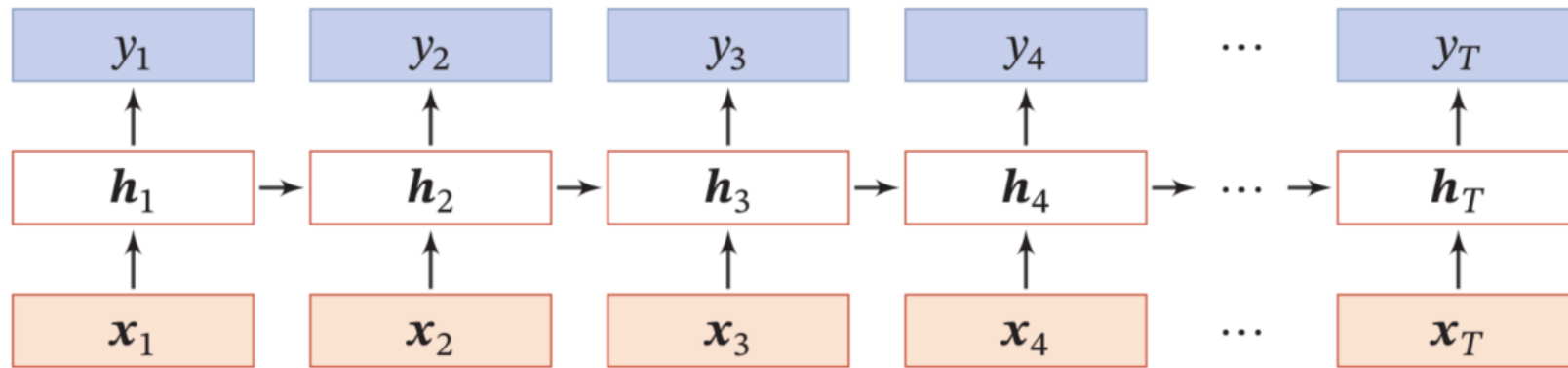


循环神经网络(Recurrent Neural Network, RNN)

Deep RNN



循环神经网络(Recurrent Neural Network, RNN)

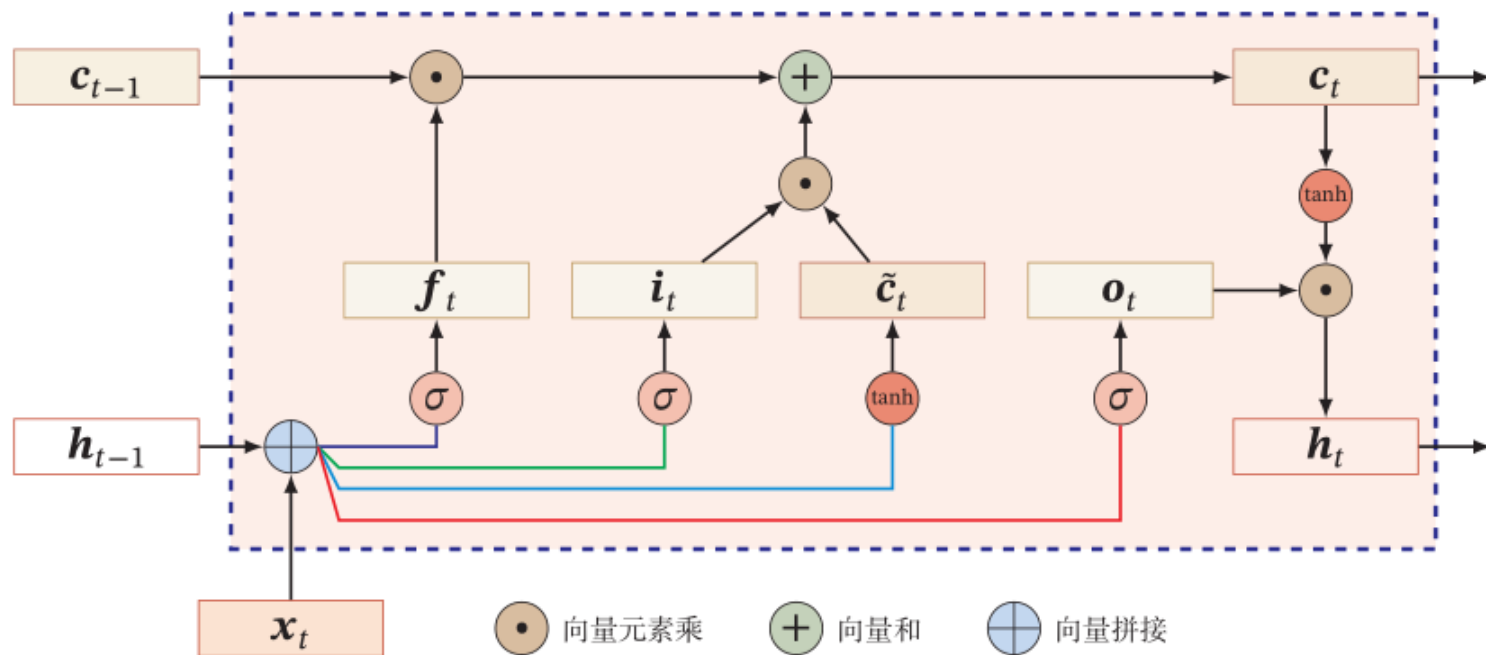


RNN只能记忆前一个状态，缺乏长期的记忆

如何解决？

LSTM (Long Short-Term Memory)

长短期记忆(LSTM)网络(1997): Jürgen Schmidhuber发明LSTM, LSTM使用门控机制来控制信息流, 允许它们捕获序列数据中的长期依赖关系



给模型一个“笔记本”（细胞状态），三个控制机制，决定什么时候该“忘”、什么时候该“记”、什么时候该“查”

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i),$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f),$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o),$$

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t,$$

$$h_t = o_t \odot \tanh(c_t),$$

Who invented deep residual learning?

Jürgen Schmidhuber (28 Sep 2025)
Pronounce: [You_again Shmidhoobuh](#)
Technical Report IDSIA-09-25, IDSIA

AI Blog
[@SchmidhuberAI](#)
juergen@idsia.ch

Who invented deep residual learning?

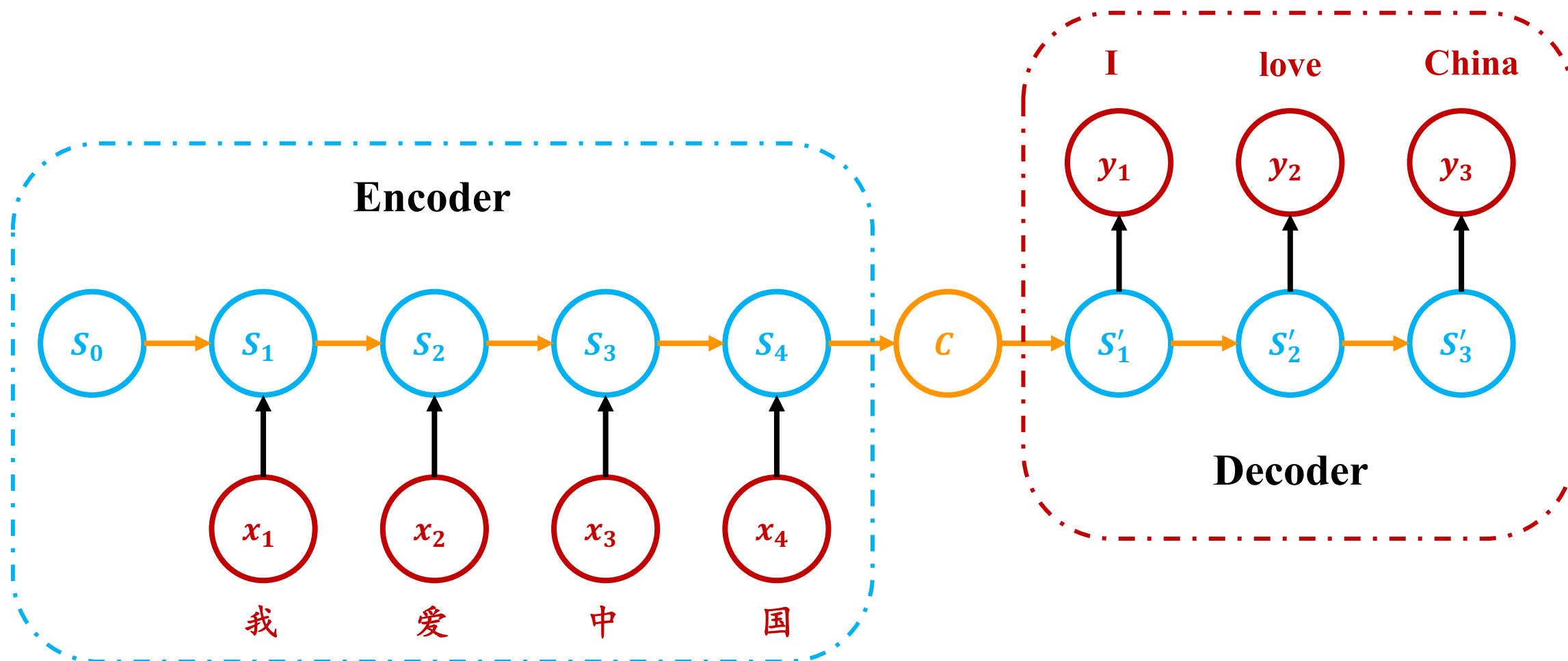
Modern AI is based on [deep artificial neural networks](#) (NNs).^[DLH] As of 2025, the most cited scientific article of the 21st century is an NN paper on *deep residual learning with residual connections*.^[MOST25,25b] Who invented this? Here is the timeline of the evolution of deep residual learning:

- ★ 1991: recurrent residual connections (weight 1.0) solve the vanishing gradient problem
- ★ 1997 LSTM: *plain* recurrent residual connections (weight 1.0)
- ★ 1999 LSTM: *gated* recurrent residual connections (gates initially open: 1.0)
- ★ 2005: unfolding LSTM—from *recurrent* to *feedforward* residual NNs
- ★ May 2015: very deep Highway Net—gated *feedforward* residual connections (initially 1.0)
- ★ Dec 2015: ResNet—like an open-gated Highway Net (or an unfolded 1997 LSTM)

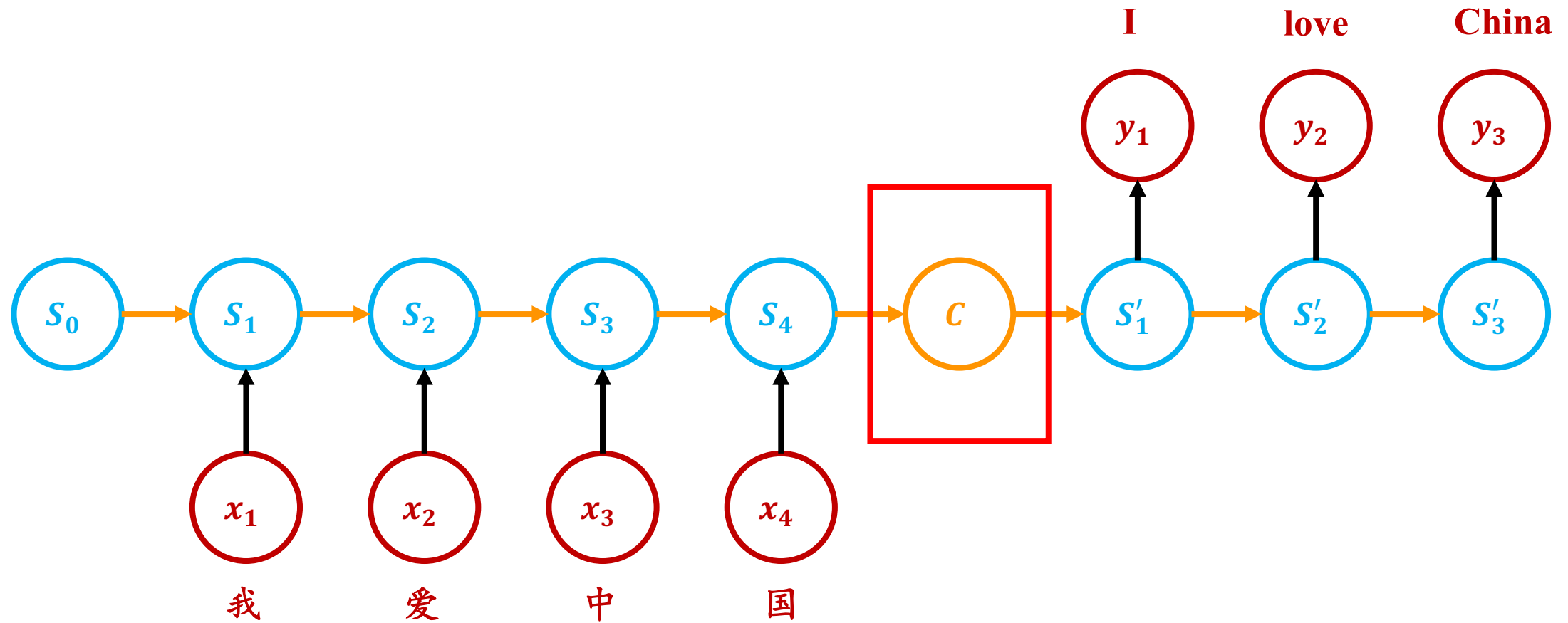
1991: recurrent residual connections solve the vanishing gradient problem

Seq2Seq

- Seq2Seq Model (2014): Ilya Sutskever等人提出Seq 2Seq模型，该模型使用编码器-解码器架构将输入序列映射到输出序列，被广泛用于机器翻译、语音识别等任务

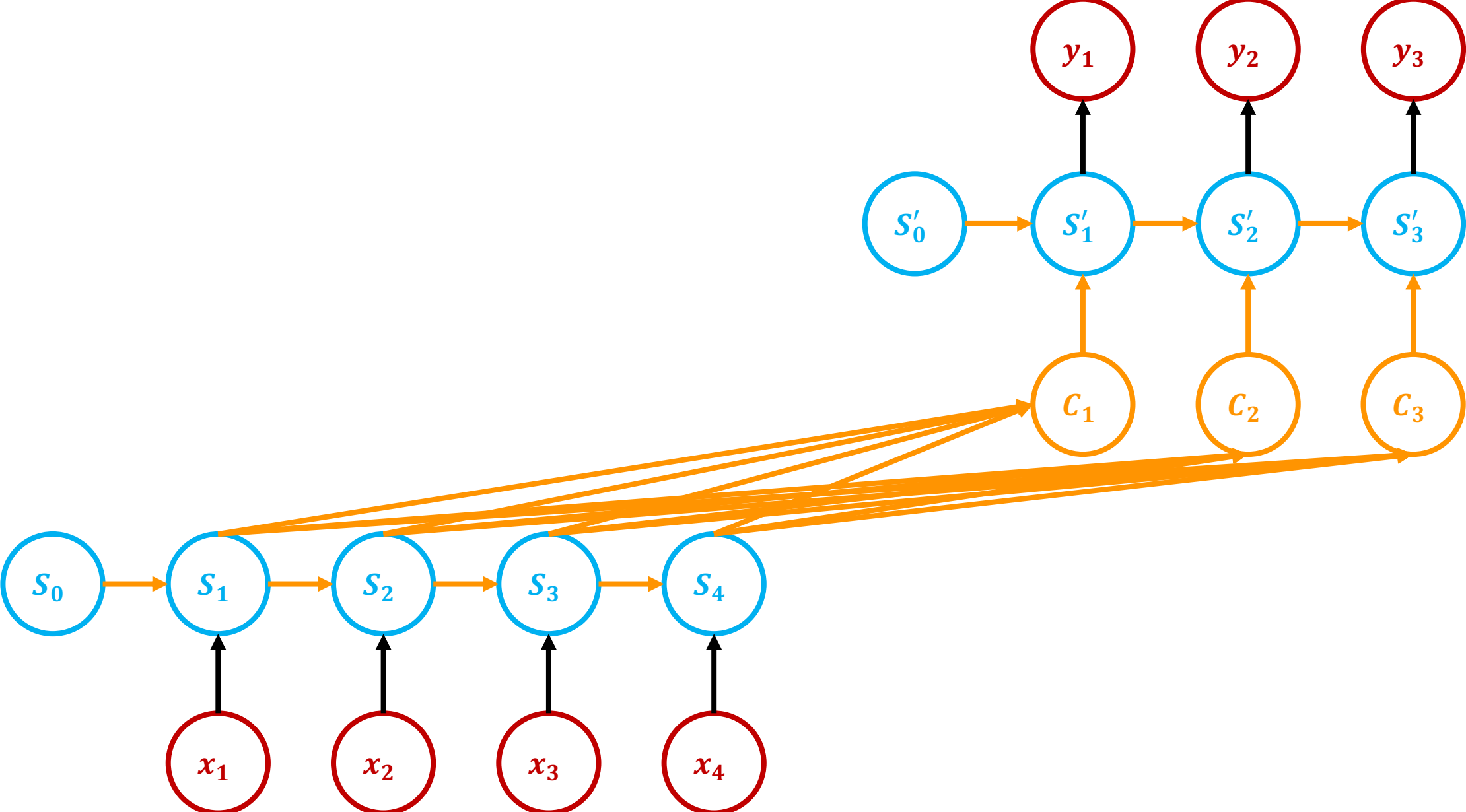


Seq2Seq的局限



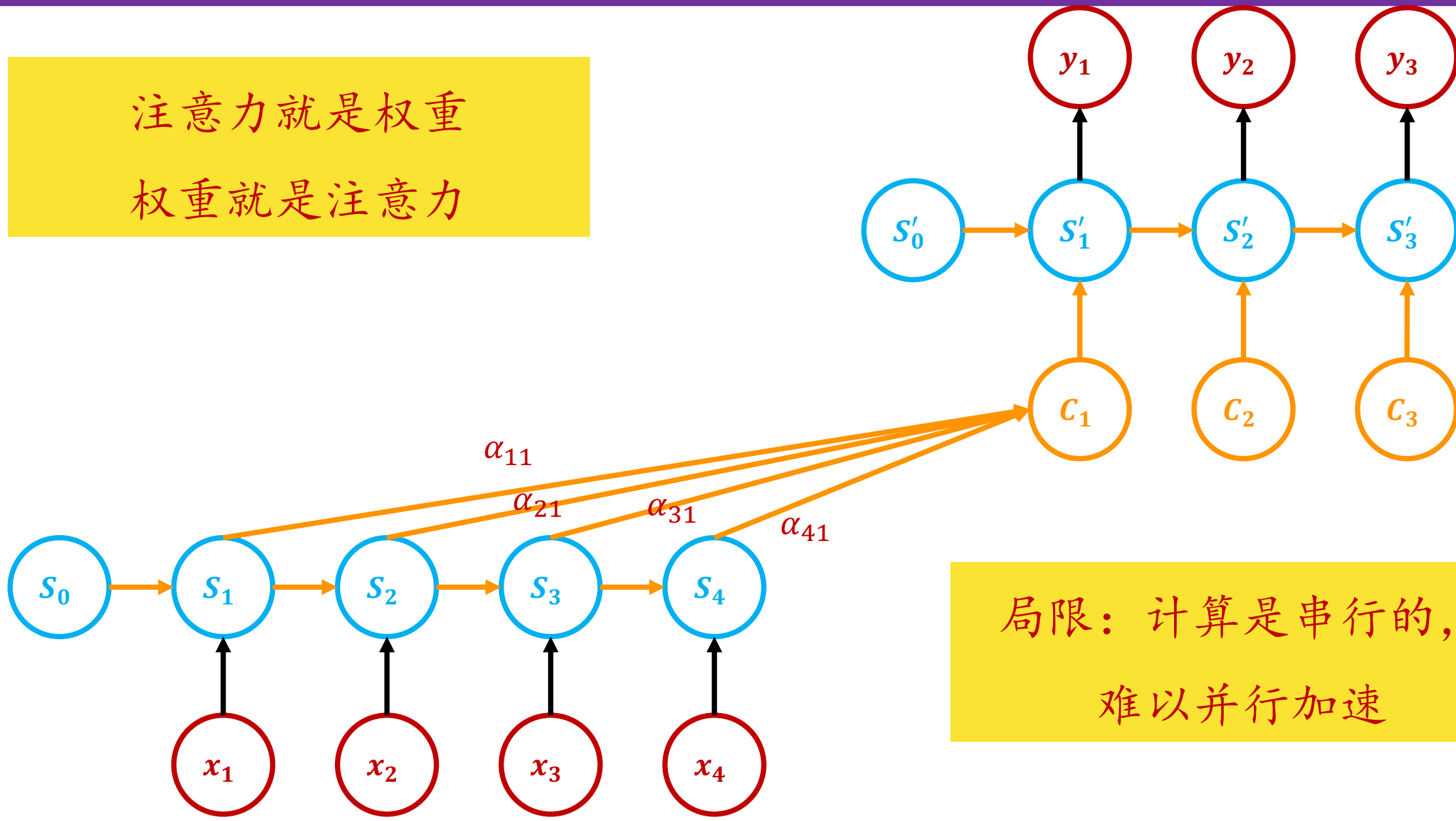
所有的内容都被压缩至一个向量表示，这样有什么局限，怎么改进？

Attention



Attention

注意力就是权重
权重就是注意力



局限：计算是串行的，
难以并行加速

Self-Attention

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

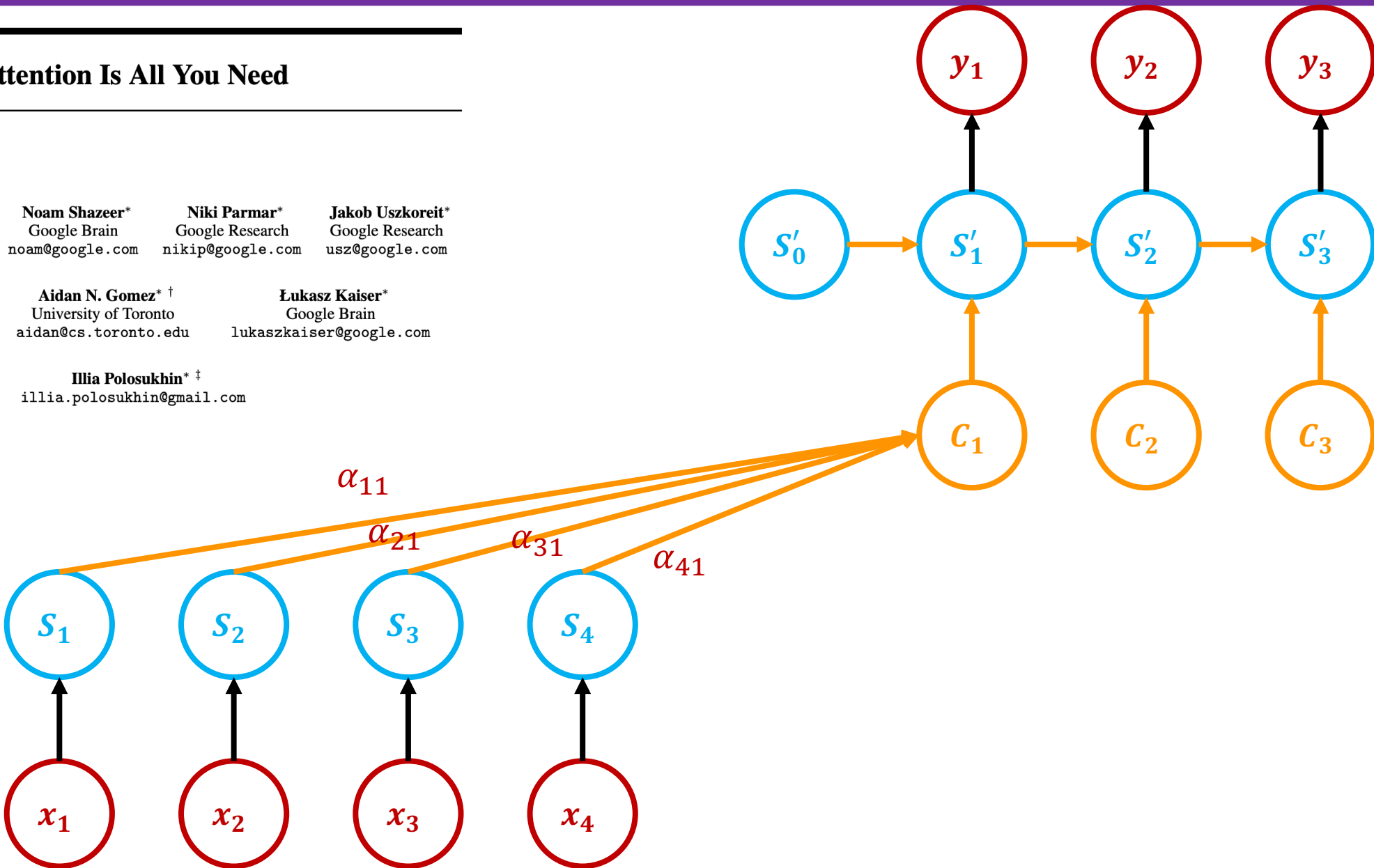
Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

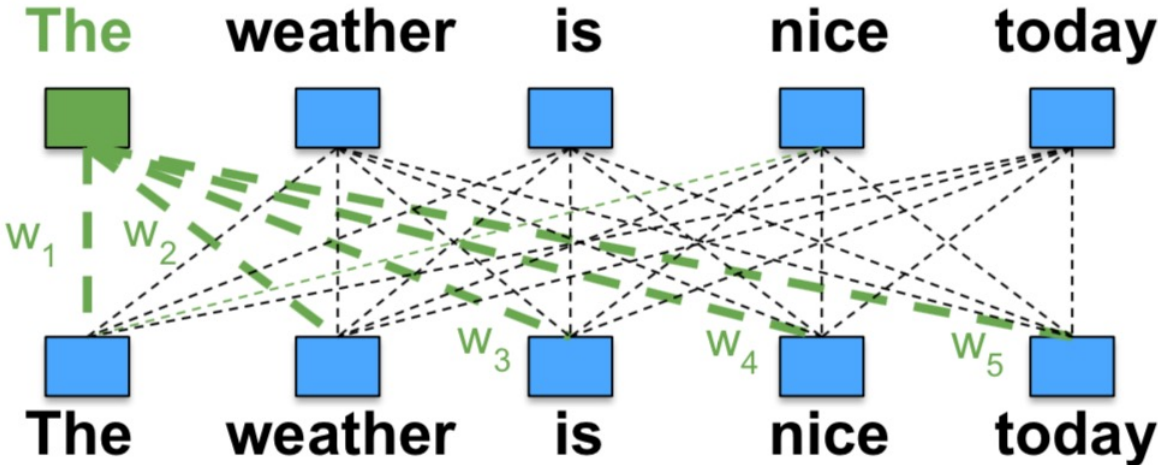
Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaizer@google.com

Ilia Polosukhin* ‡
illia.polosukhin@gmail.com



Self-Attention



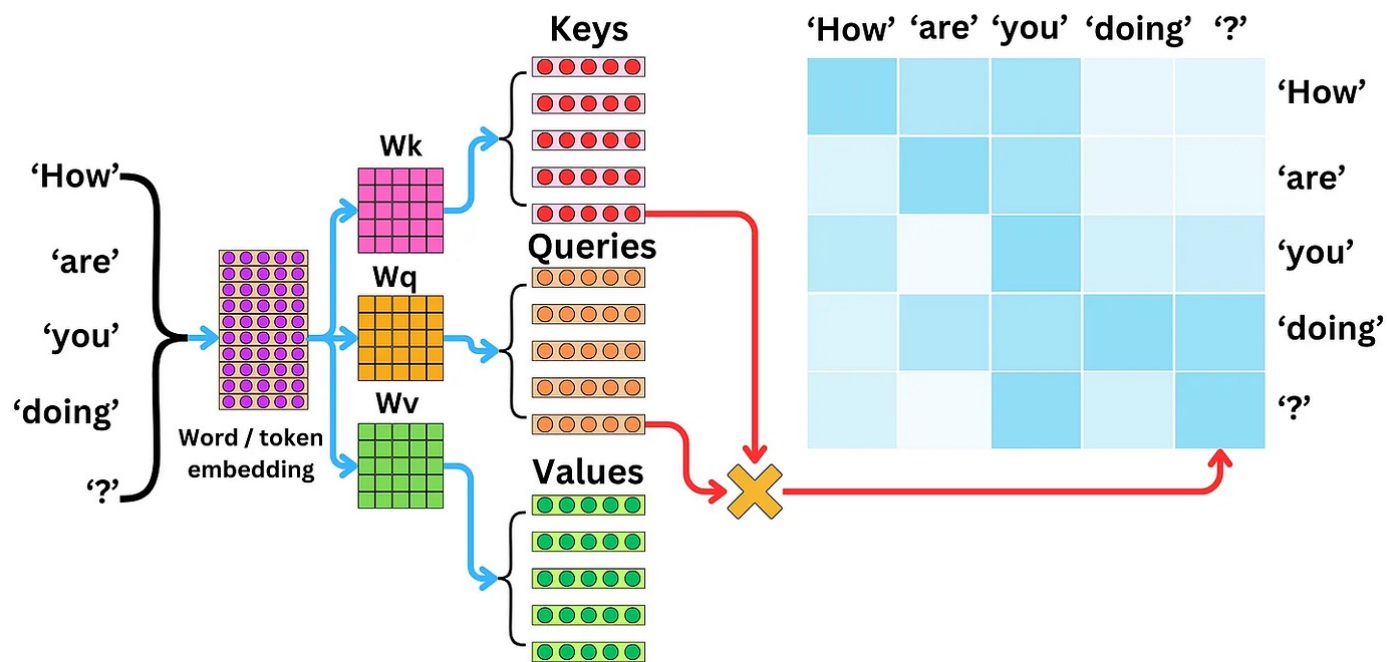
$$W_1, W_2, W_3, W_4, W_5 = \text{softmax} \left(\begin{array}{|c|c|c|} \hline 0.6 & 0.2 & 0.8 \\ \hline \end{array} \times \begin{array}{|c|c|c|c|c|} \hline 0.6 & 0.2 & 0.9 & 0.4 & 0.4 \\ \hline 0.2 & 0.3 & 0.1 & 0.1 & 0.1 \\ \hline 0.8 & 0.1 & 0.8 & 0.4 & 0.6 \\ \hline \end{array} \right)$$

The The weather is nice today

$$\begin{array}{|c|} \hline 1.8 \\ \hline 2.3 \\ \hline 0.4 \\ \hline \end{array} \text{The} = W_1 \times \begin{array}{|c|} \hline 0.6 \\ \hline 0.2 \\ \hline 0.8 \\ \hline \end{array} \text{The} + W_2 \times \begin{array}{|c|} \hline 0.2 \\ \hline 0.3 \\ \hline 0.1 \\ \hline \end{array} \text{weather} + W_3 \times \begin{array}{|c|} \hline 0.9 \\ \hline 0.1 \\ \hline 0.8 \\ \hline \end{array} \text{is} + W_4 \times \begin{array}{|c|} \hline 0.4 \\ \hline 0.1 \\ \hline 0.4 \\ \hline \end{array} \text{nice} + W_5 \times \begin{array}{|c|} \hline 0.4 \\ \hline 0.1 \\ \hline 0.6 \\ \hline \end{array} \text{today}$$

Transformer(2017)

- Transformer: 采用注意力机制处理远程依赖关系，直接建模全局相关性，通过计算注意力权重捕获中远距离元素之间的关系



$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V$$

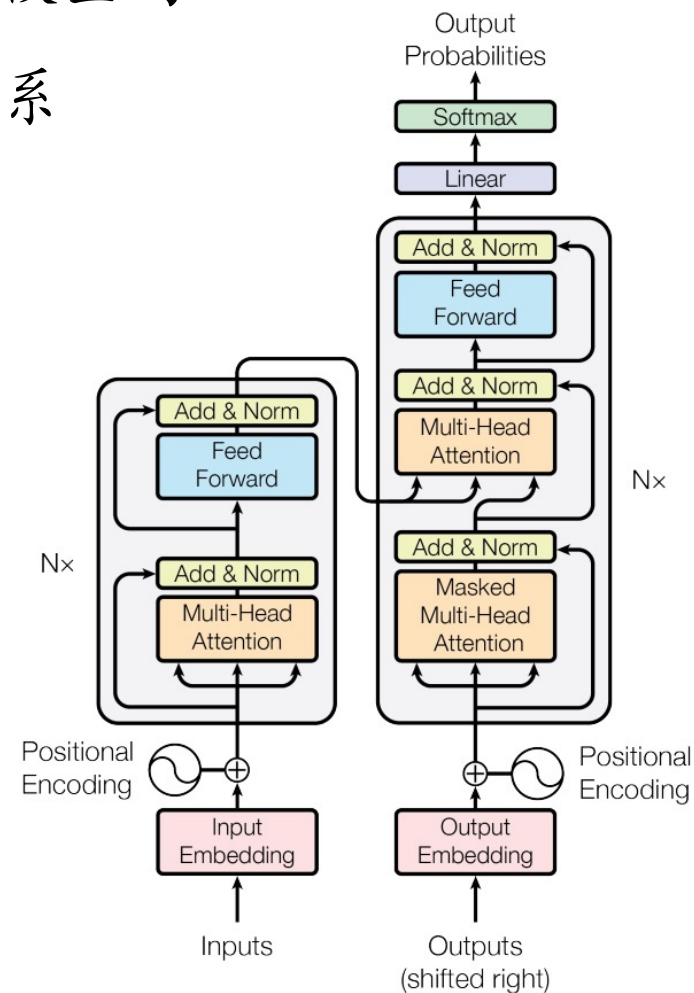
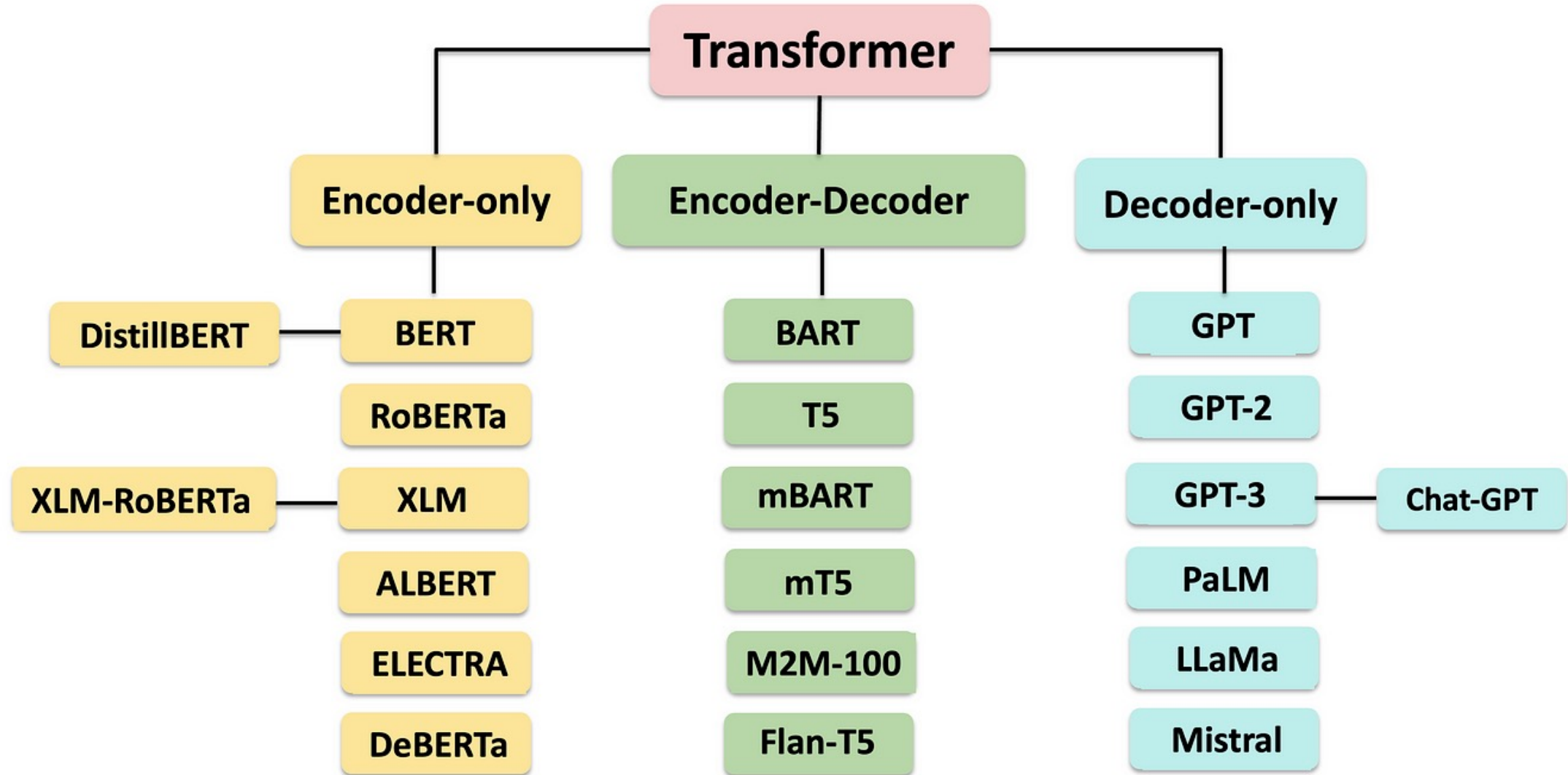


Figure 1: The Transformer - model architecture.

Transformer(2017)



Language model: A brief history

Pre-Transformers

- N-gram model
- A Neural Probabilistic Language Model [[Bengio 2003](#)]
- Word embeddings
- RNN & LSTM
- ...

Post-Transformers

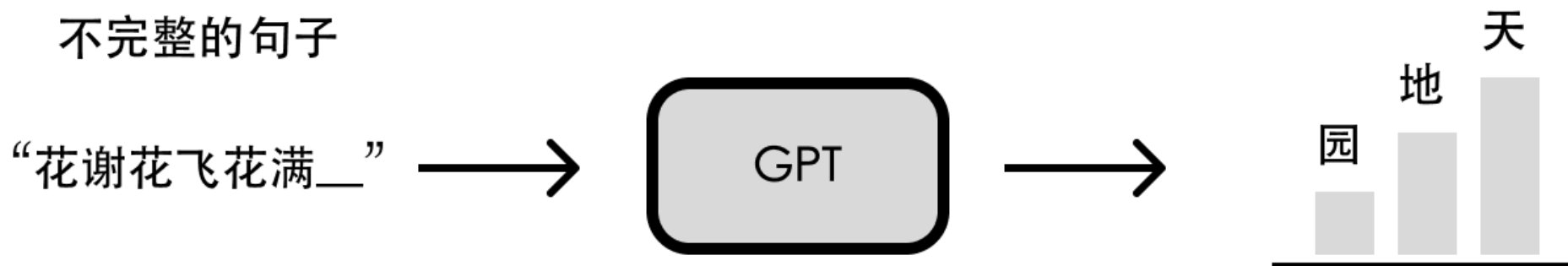
- Bert
- GPT
- GPT-2
- T5
- GPT-3
- ...

GPT-1.0(GPT=Generative Pre-trained Transformer)

不完整的句子

“花谢花飞花满__”

GPT-1.0(GPT=Generative Pre-trained Transformer)



- GPT 初代所做的事：**根据上文补全下文**
 - 从大规模的文本语料中，将每一条文本随机地分成两部分，只保留上半部分，让模型学习下半部分学习到底该填写什么

GPT-2

- 基本想法：给模型出其它的语言任务，应该也会对模型训练有极大的帮助
- 句子打乱顺序再排序、选择题、判断题、问答题、寻找文中的错别字、把预测单字改成预测实体词汇等等，都可以制定数据集添加在模型的预训练里
- GPT-2在GPT初代的基础上，又添加了多个任务，比如机器翻译、问答、文本摘要等等，扩增了数据集和模型参数，又训练了一番

GPT-2

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%

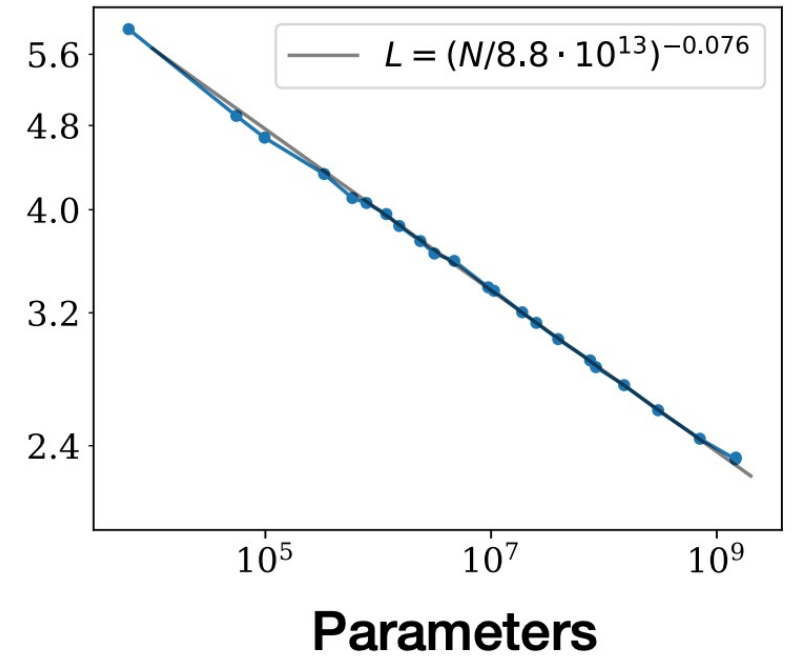
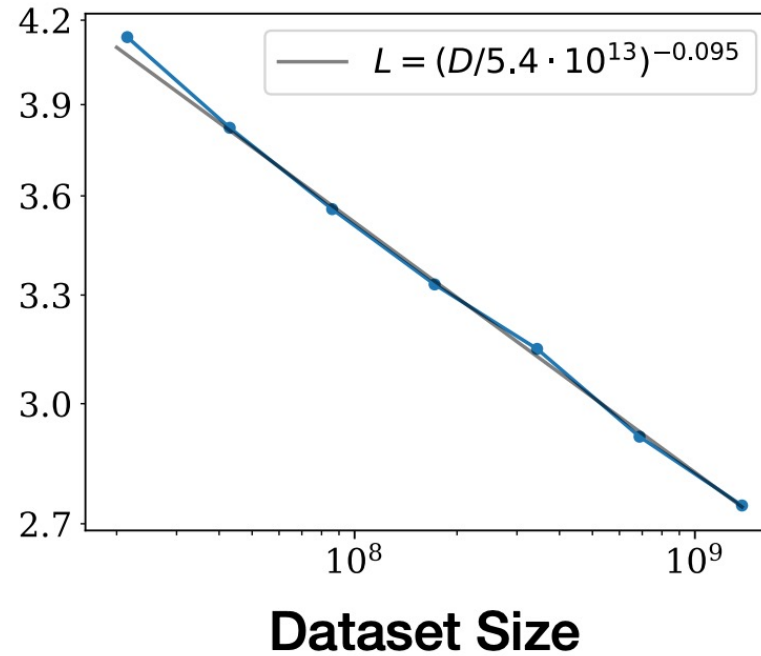
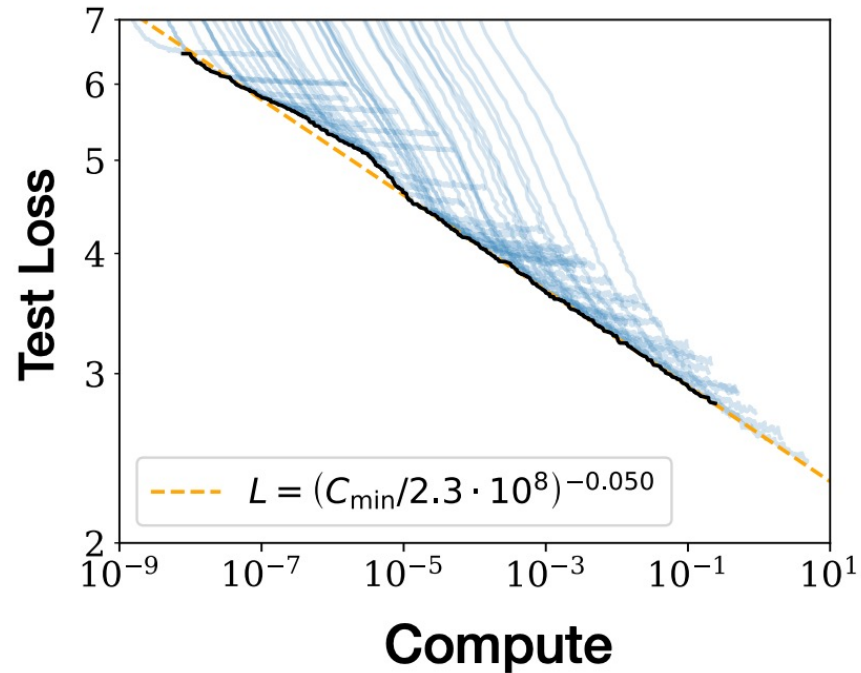
GPT-3

大数据+大参数量+大计算力

模型名	参数量	训练数据
Original GPT (GPT-1)	1.17 亿	4.5 GB文本
GPT-2	15 亿	40 GB文本
GPT-3	1750亿	570 GB文本

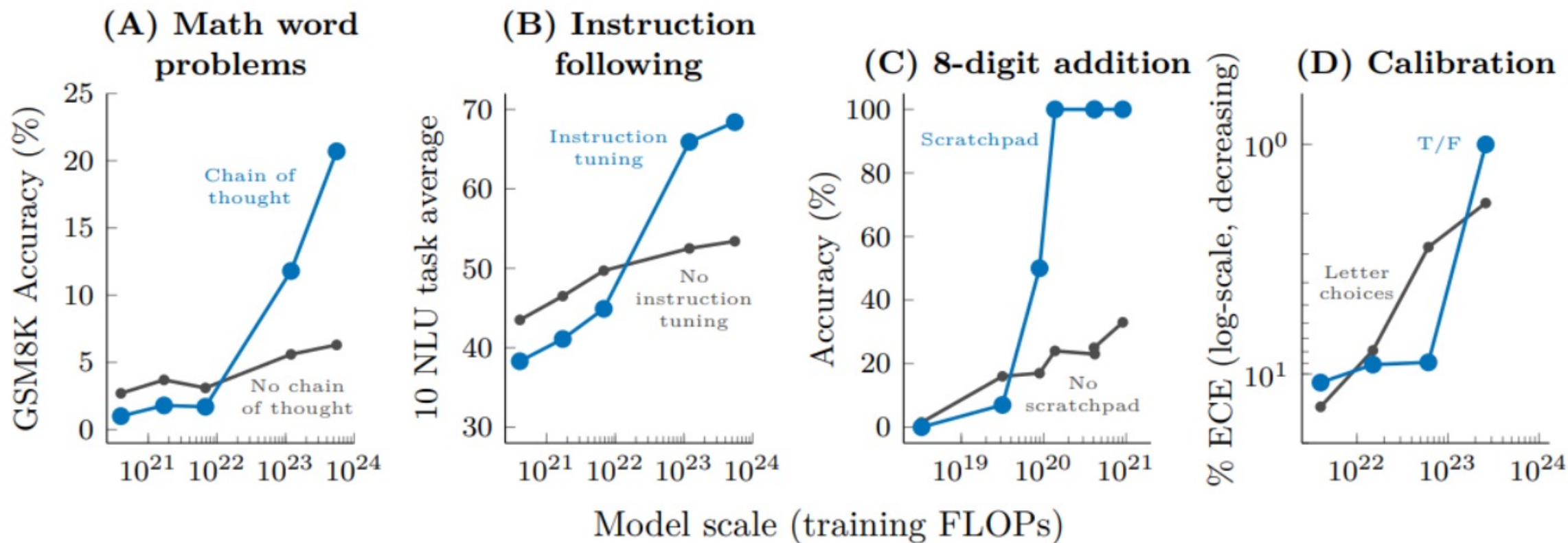
Scaling Law

More data, More compute, Larger models



Emergent ability (涌现能力)

Some ability of LM is not present in smaller models but is present in larger models



ChatGPT

ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.

November 30, 2022
13 minute read



We are excited to introduce ChatGPT to get users' feedback and learn about its strengths and weaknesses. During the research preview, usage of ChatGPT is free. Try it now at chat.openai.com.

ChatGPT Blog: <https://openai.com/blog/chatgpt/>

ChatGPT

□ Phase 1: 预训练(Pre-Training)

- 掌握通用知识

□ Phase 2: 后训练：有监督微调(Supervised Finetuning)

- 与特定任务对齐

□ Phase 3: 后训练：强化微调(RLHF)

- 与人类偏好对齐


ChatGPT

- 有监督微调

输入一个问题 (prompt)

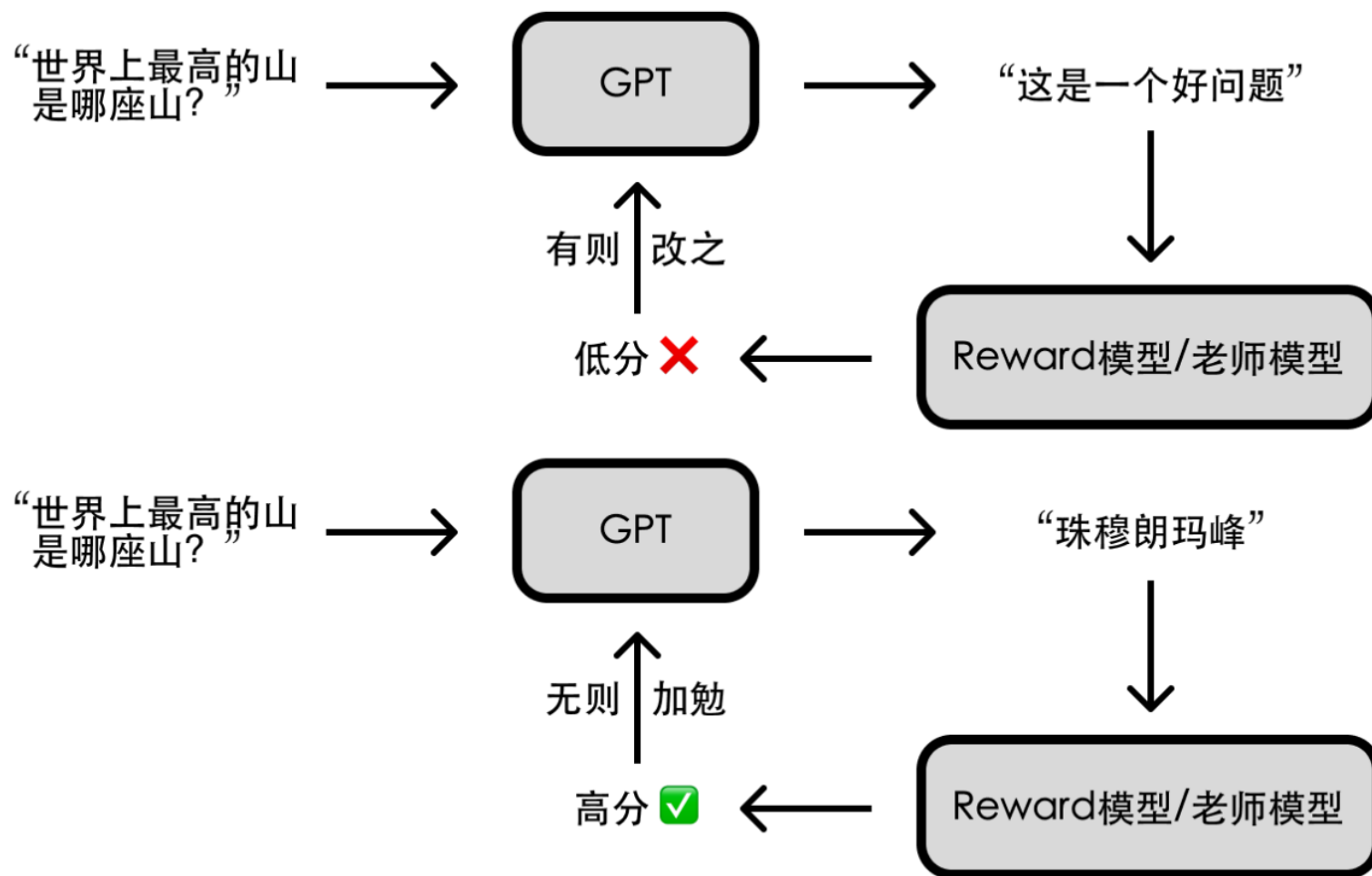
“世界上最高的山
是哪座山？”



“你能告诉我么”
“珠穆朗玛峰” 
“这是一个好问题”

ChatGPT

- 强化学习

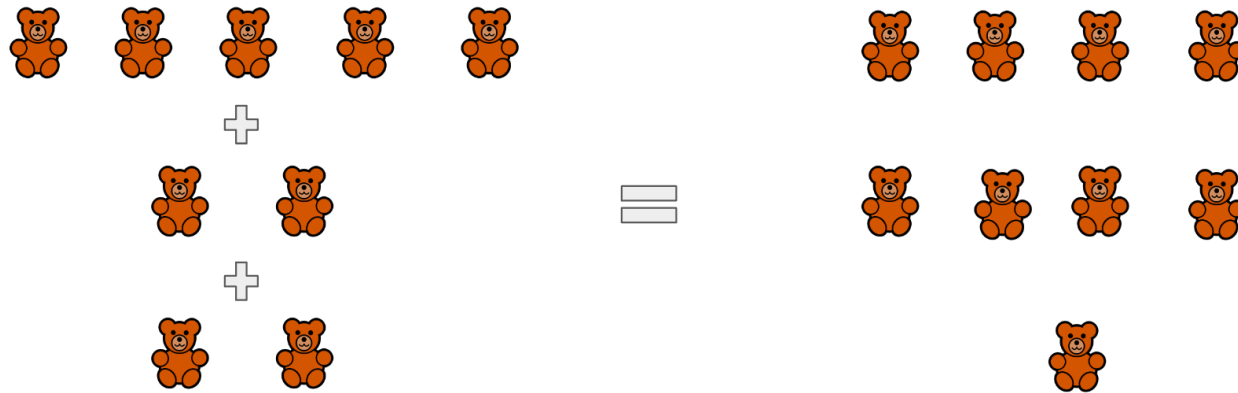


Prompting Language Models



Prompting Language Models

Q: Shawn has 5 toys. For Christmas, he got 2 toys each from his mom and dad. How many toys does he have now?



**A: The
answer is 9
toys**

In-Context Learning: 0-shot/Few-Shot/Instructions

Instruction: Solve the following math problem

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

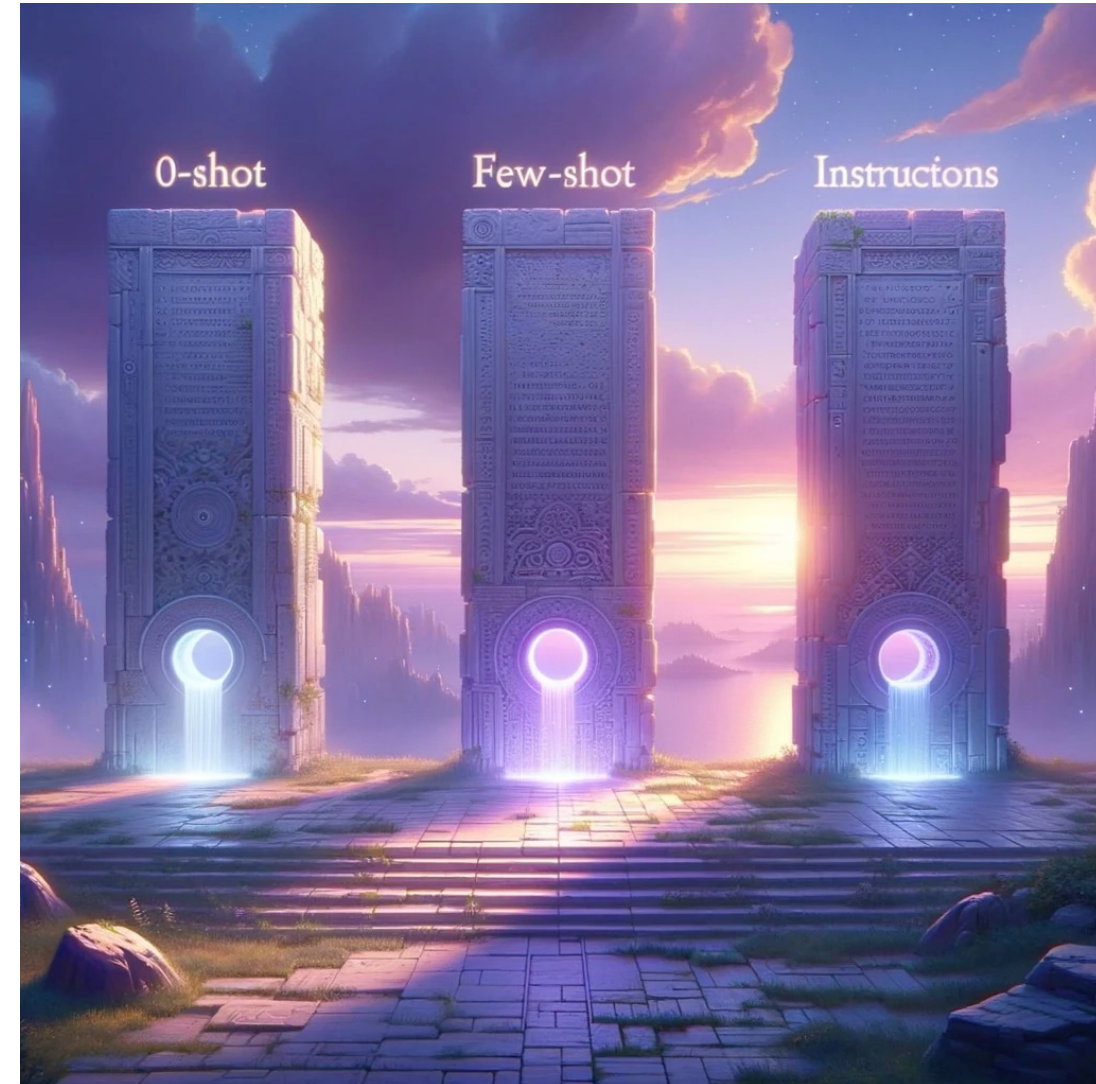
A: *The answer is 5 cars.*

Prompt

Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?

A:

Test Example



In-Context Learning: 0-shot/Few-Shot/Instructions

No Prompt

Prompt

Zero-shot
(0s)

skicts = sticks

Please unscramble the letters into
a word, and write that word:
skicts = sticks

1-shot
(1s)

chiar = chair
skicts = sticks

Please unscramble the letters into
a word, and write that word:
chiar = chair
skicts = sticks

Few-shot
(FS)

chiar = chair
[...]
pciinc = picnic
skicts = sticks

Please unscramble the letters into
a word, and write that word:
chiar = chair
[...]
pciinc = picnic
skicts = sticks

Chain-of-Thought Prompting

Direct Prompt

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

A: *The answer is 5 cars.*

Chain-of-Thought Prompt

Standard prompt:

- $Q \rightarrow A$

Chain-of-thought prompt:

- $Q \rightarrow$ Reasoning Process, A

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

Thought (T): There are originally 3 cars. 2 more cars arrive. $3 + 2 = 5$.

A: *The answer is 5 cars.*

Chain-of-Thought Prompting

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

Zero-Shot CoT Prompting

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The answer is 8. **X**

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4. **✓**

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) 8 **X**

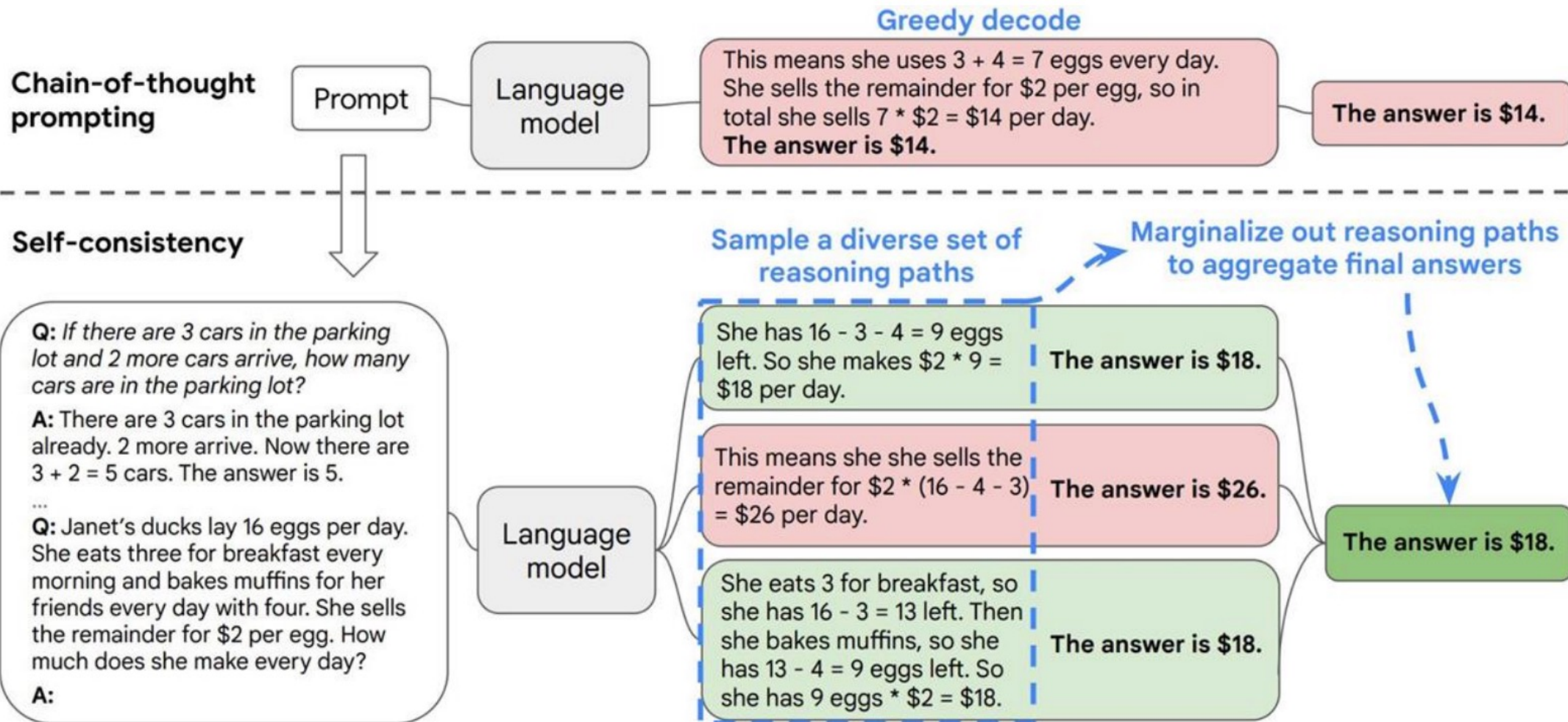
(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

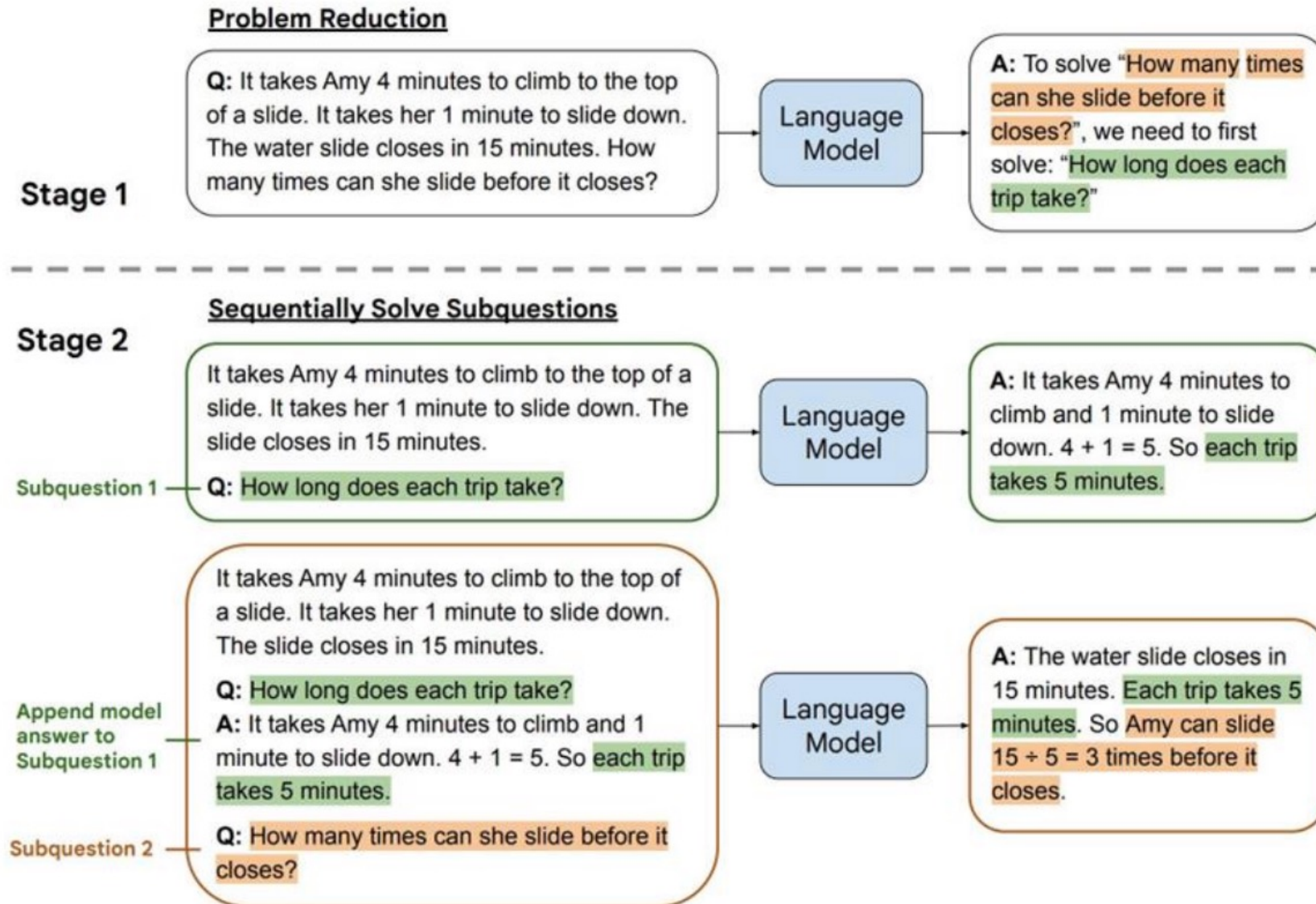
A: **Let's think step by step.**

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. **✓**

Self-Consistency Prompting



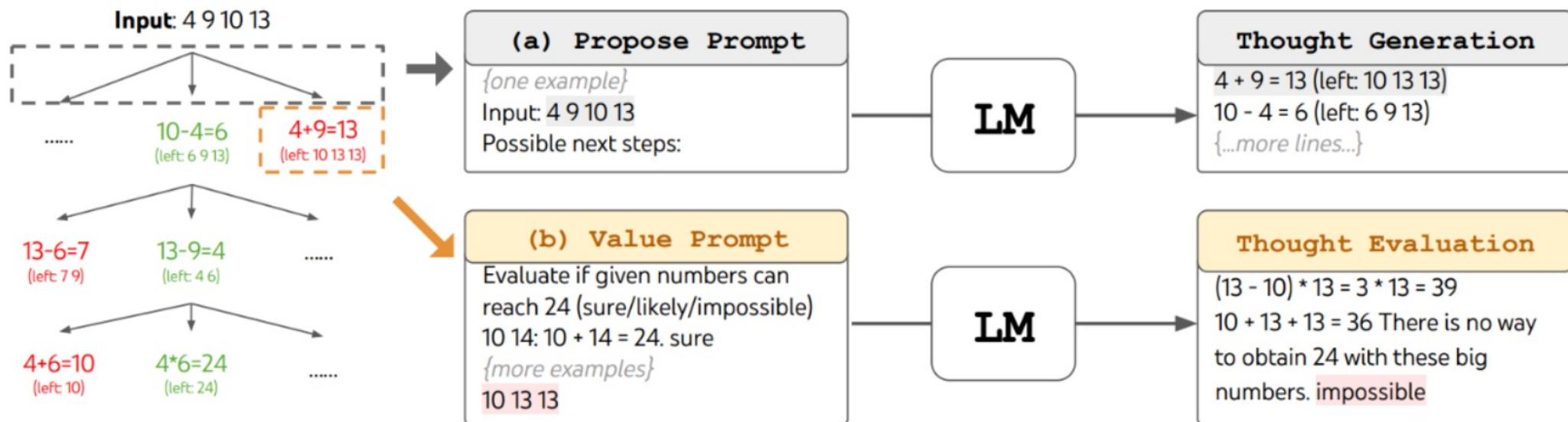
Least-to-Most Prompting



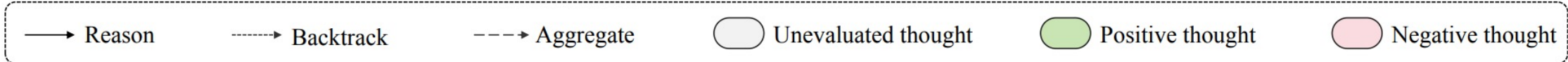
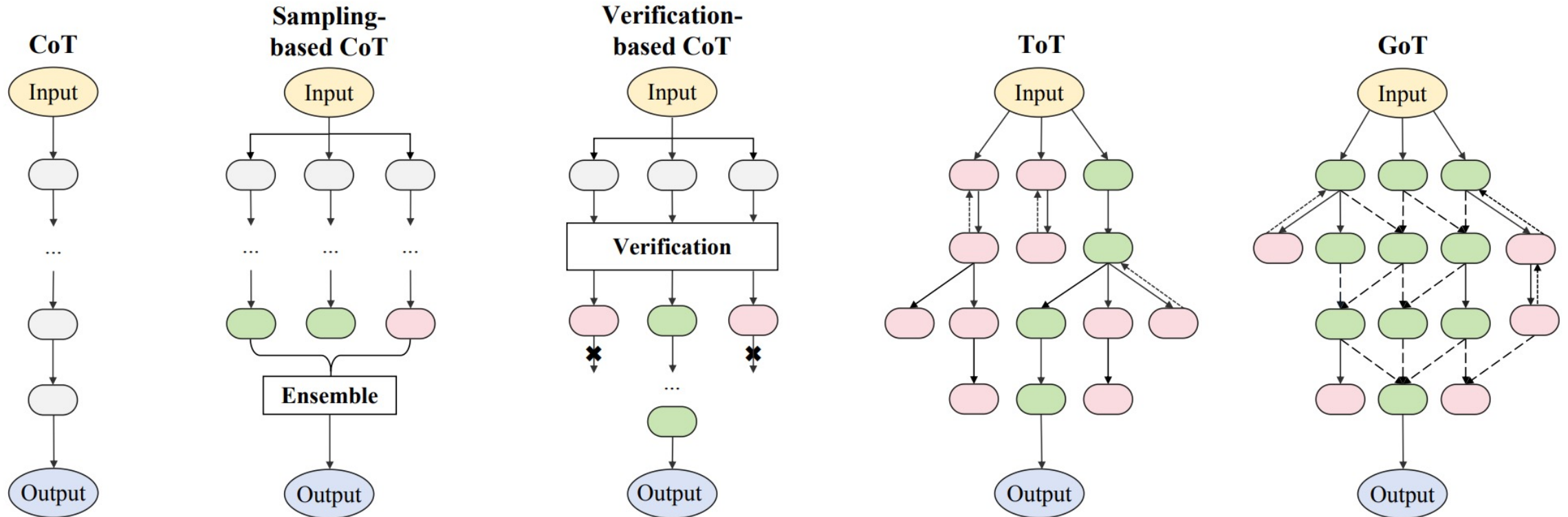
Tree of Thought

4.1 Game of 24

Game of 24 is a mathematical reasoning challenge, where the goal is to use 4 numbers and basic arithmetic operations (+-*/) to obtain 24. For example, given input “4 9 10 13”, a solution output could be “(10 - 4) * (13 - 9) = 24”.



Prompting Techniques



Train your own LLMs: Supervised Fine-Tuning

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?



A: The answer is 5 cars.

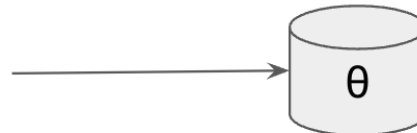
Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?



A: The answer is 39 pieces.

Train/Fine-tune

Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?

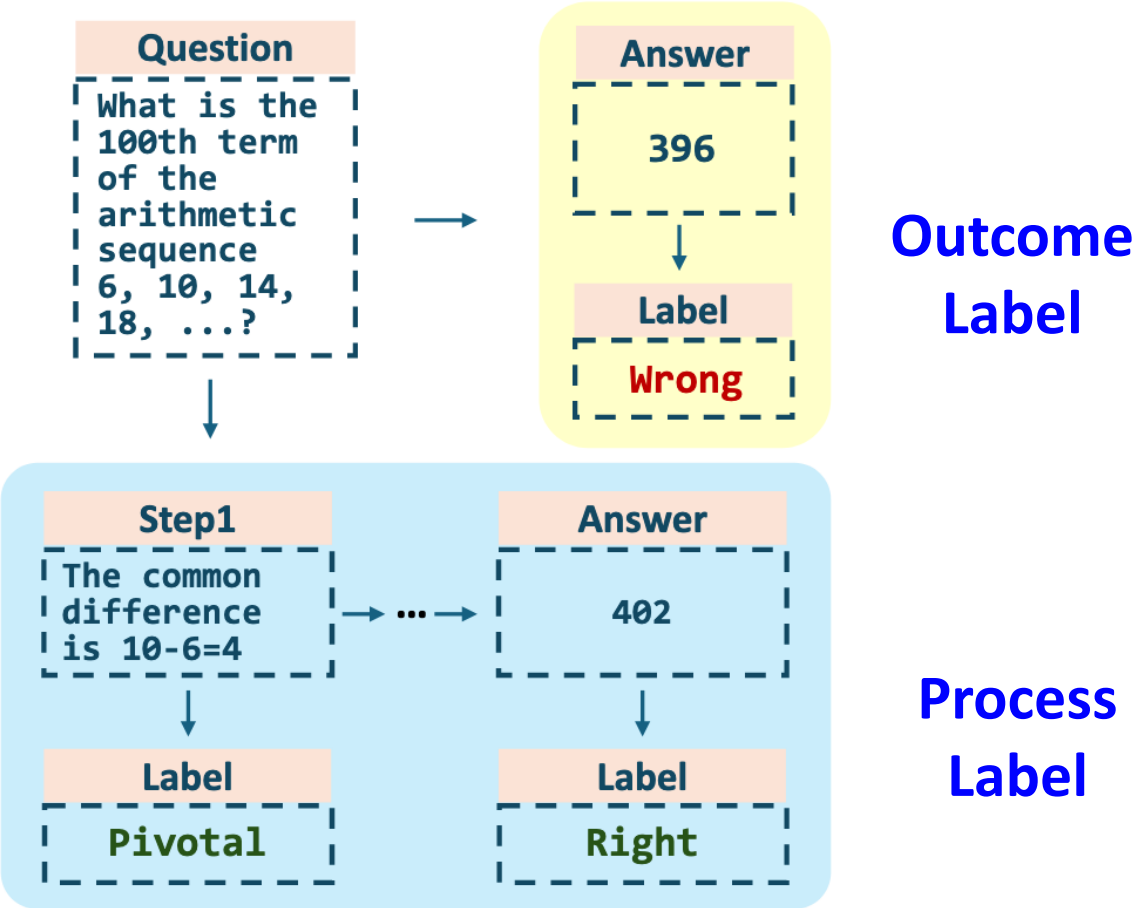


A: The answer is 9 toys

Test

Supervised Fine-Tuning with COT/Rationale

思考过程的标注获取代价很高



The denominator of a fraction is 7 less than 3 times the numerator. If the fraction is equivalent to $\frac{2}{5}$, what is the numerator of the fraction? (Answer: 14)

Let's call the numerator x.

So the denominator is $3x-7$.

We know that $\frac{x}{3x-7} = \frac{2}{5}$.

So $5x = 2(3x-7)$.

$5x = 6x - 14$.

So $x = 7$.

Pretraining + Prompting

- **Fine-tuning (FT)**
 - + Strongest performance
 - - Need curated and labeled dataset for each new task (typically 1k-100k ex.)
 - - Poor generalization, spurious feature exploitation
- **Few-shot (FS)**
 - + Much less task-specific data needed
 - + No spurious feature exploitation
 - - Challenging
- **One-shot (1S)**
 - + "Most natural," e.g. giving humans instructions
 - - Challenging
- **Zero-shot (OS)**
 - + Most convenient
 - - Challenging, can be ambiguous

**Stronger
task-specific
performance**



**More convenient,
general, less data**

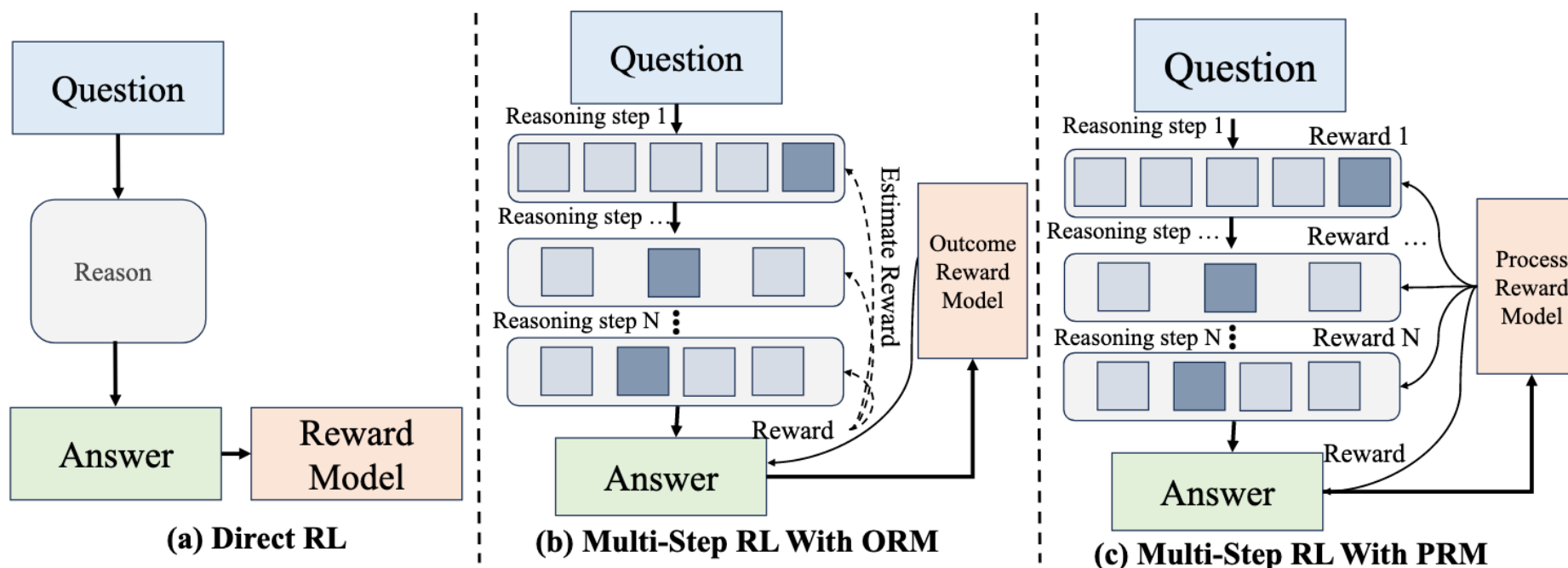
SFT vs RLFT

□ 监督微调 (Supervised Fine-Tuning, SFT)

- 直接基于所构建的推理数据进行监督微调，根据数据标注可以分为过程监督和结果监督

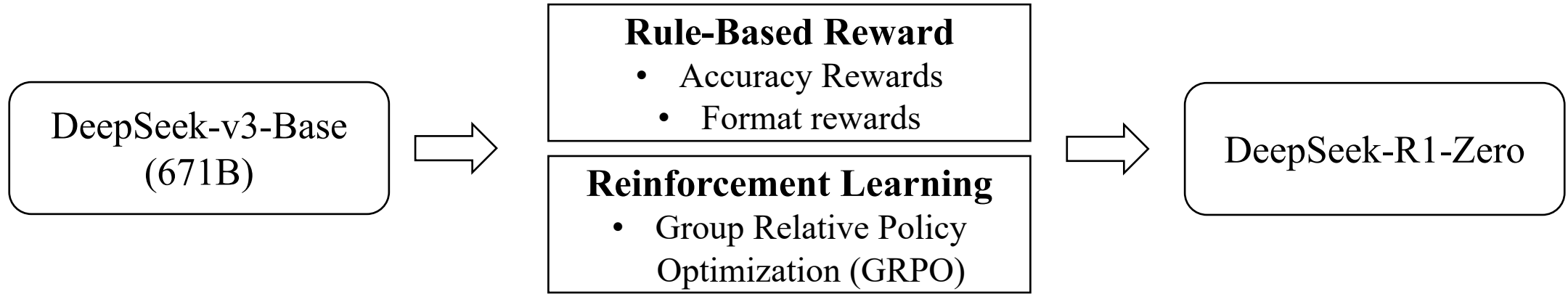
□ 强化微调 (Reinforcement Fine-Tuning, RLFT)

- 设计reward对模型进行强化学习微调



SFT vs RLFT

DeepSeek-R1 Zero: 无需SFT, 纯RL驱动



Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a+x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a+x}} = x$, let's start by squaring both ...

$$(\sqrt{a - \sqrt{a+x}})^2 = x^2 \implies a - \sqrt{a+x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a+x \implies a^2 - 2ax^2 + (x^2)^2 = a+x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$$

...

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be ...

We started with the equation:

$$\sqrt{a - \sqrt{a+x}} = x$$

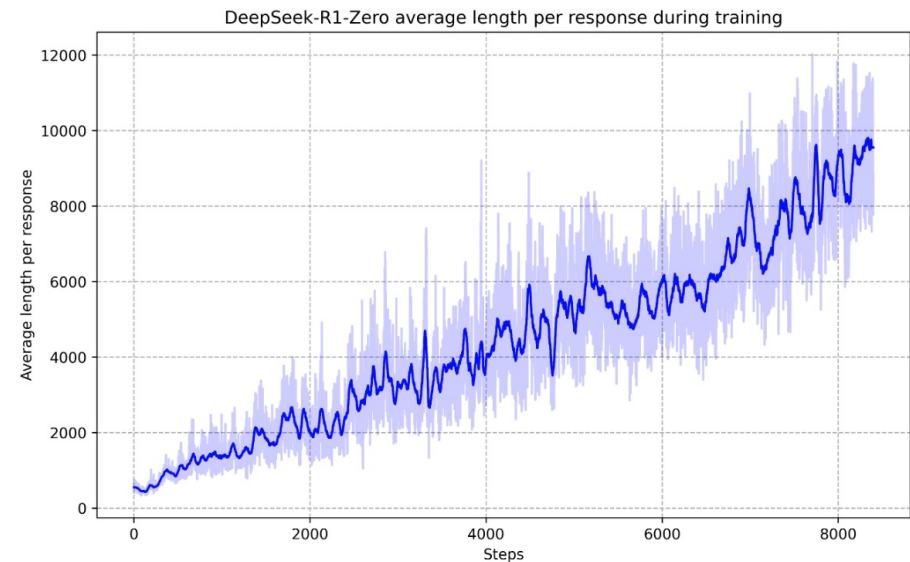
First, let's square both sides:

$$a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$$

Next, I could square both sides again, treating the equation: ...

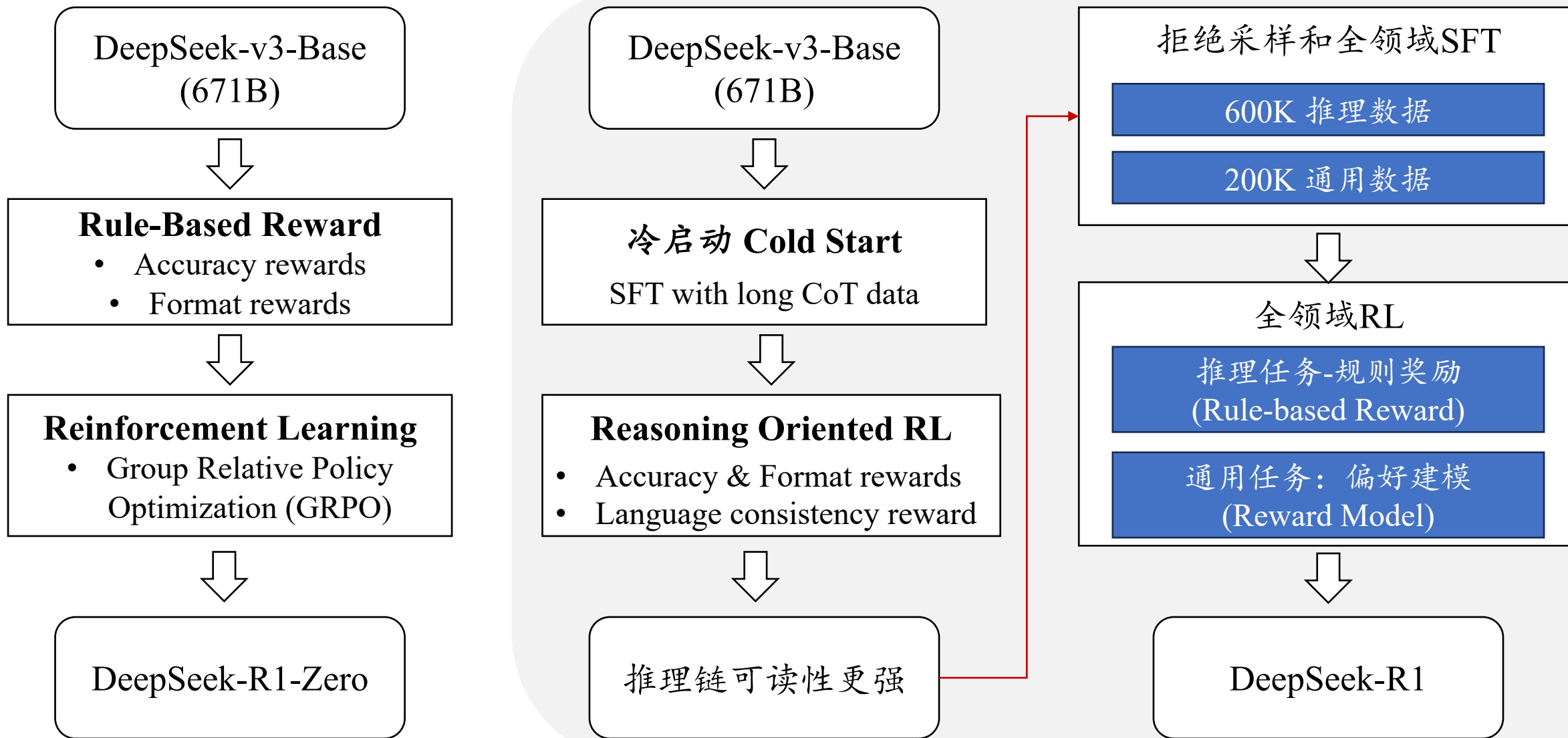
...

Aha Moment



RL驱动下自然涌现出Long-COT推理能力

SFT vs RLFT



SFT vs RLFT

□ 对于形式化推理任务，例如math、code，更容易设计Rule-based reward

□ SFT vs RL

- 直接利用SFT更多是去拟合数据中的pattern，难以学习到背后的规律；
- RL则是通过鼓励模型在最大化reward的过程中学习到推理背后的规律，获得的泛化性和推理表现上界更高
- SFT规范模型输出格式，使得后续的RL可以获得更高收益 [1]
- 小模型RL不容易出现Long CoT [2]

[1] SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training. <https://arxiv.org/pdf/2501.17161>

[2] Demystifying Long Chain-of-Thought Reasoning in LLMs. <https://arxiv.org/pdf/2502.03373>

Some jargon words (学术黑话)

- **Scaling Law**
 - 模型越大、数据越多、计算越猛，性能越强——只要钱够烧，效果不会差
- **Emergent Ability**
 - 模型小的时候不会，一大了突然会了——像是“突然通灵”，让人又爱又怕。
- **Instruction vs. Prompt**
 - Prompt 是“暗号”，Instruction 是“明令”——从“你猜我想干嘛”到“我现在要你做这个”
- **Zero-shot / Few-shot**
 - 不教就会/教几次就会——像天赋型选手
- **CoT (Chain-of-Thought)**
 - “一步一步想”，让模型不再蒙答案，而是先思考再作答
- **ICL (In-context Learning)**
 - 模型“假装学会了”，不更新参数，只靠输入示范“临时学艺”
- **Pre-training and Fine-tuning**
 - 先“读万卷书”，再“精读一本”——粗养+精调的组合拳
- **Alignment**
 - 从“会说”到“说得对”，再到“说得你想听”——是技术也是哲学难题
- **Hallucination**
 - 模型一本正经“瞎编”的时刻——“我不懂，但我会瞎说”

大纲

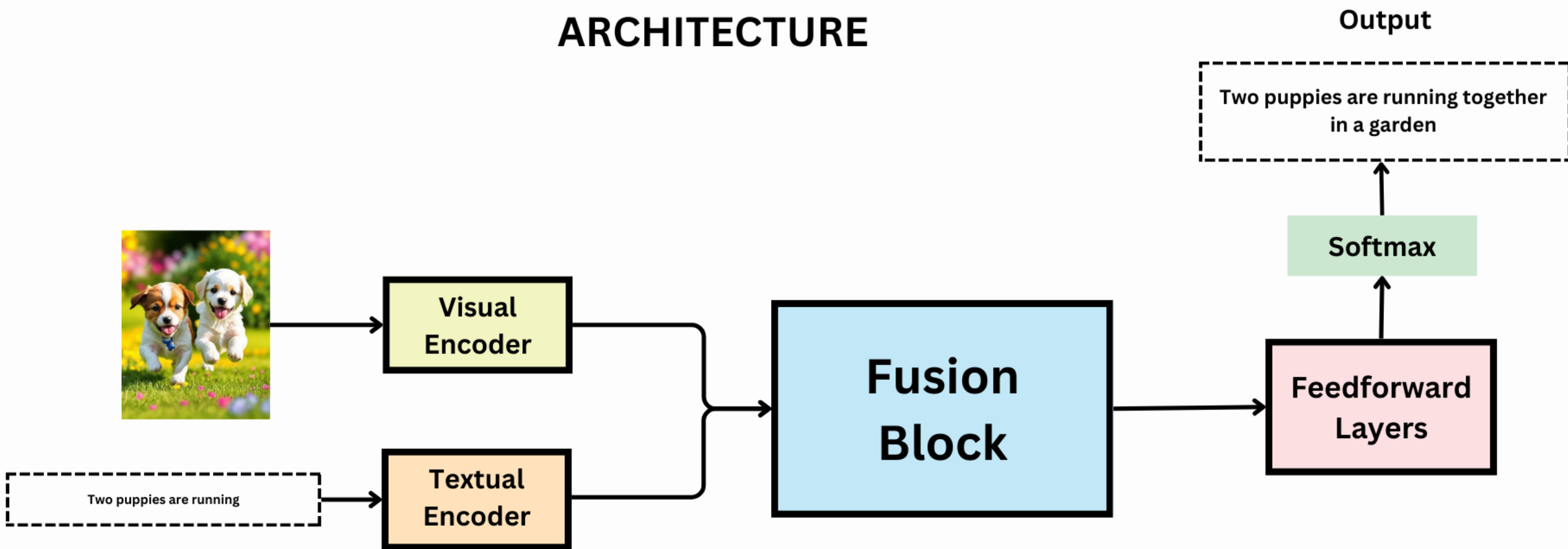
- 神经元模型到前馈神经网络
- 参数优化：BP算法
- 深度学习
- 计算机视觉与卷积神经网络
- 自然语言处理与循环神经网络
- 多模态学习**

多模态任务(Multi-Modal Tasks)

- 人的感知途径：视觉、听觉、触觉、味觉、嗅觉
- 人工智能的模态：图像、语音、文本、动作...
- 图像内部也有模态：可见光、红外光、遥感、X线...
- 典型任务：
- 文生图、图生文、语言指导的机器人操作、图文联合推理...

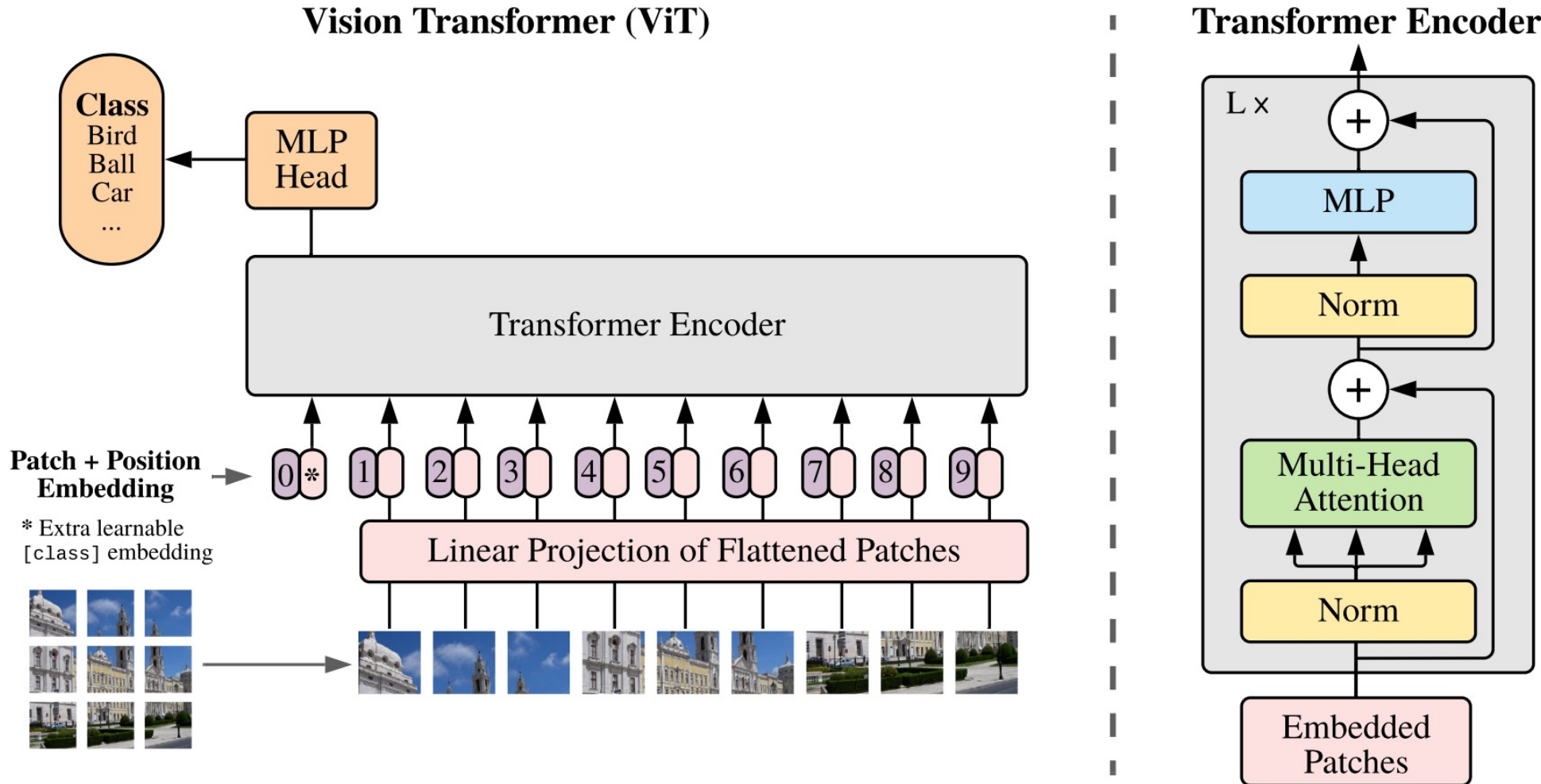
Vision-Language Model

VISION LANGUAGE MODEL ARCHITECTURE



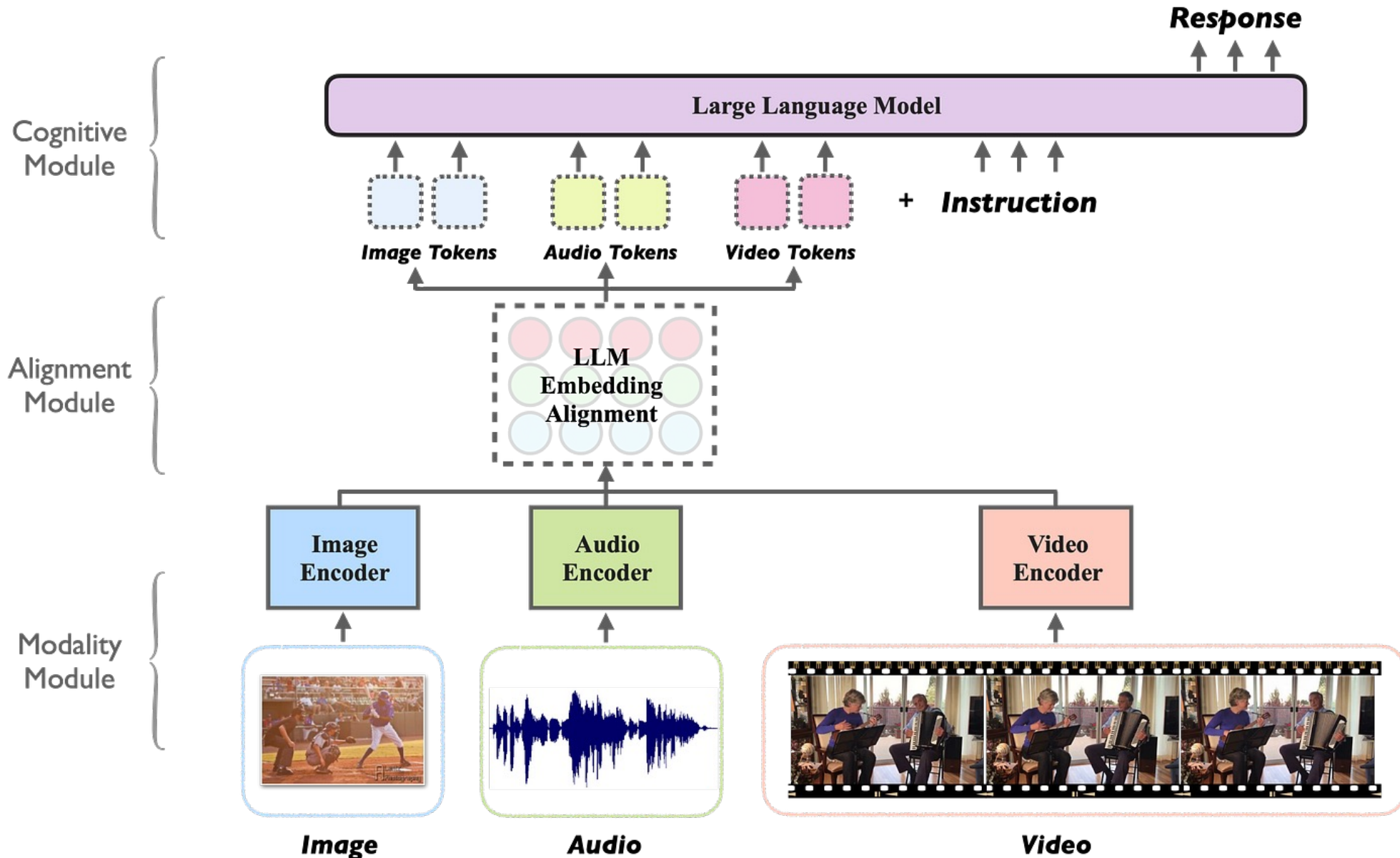
关键问题：表示、对齐、融合、翻译、共同学习

Vision Transformer (ViT)



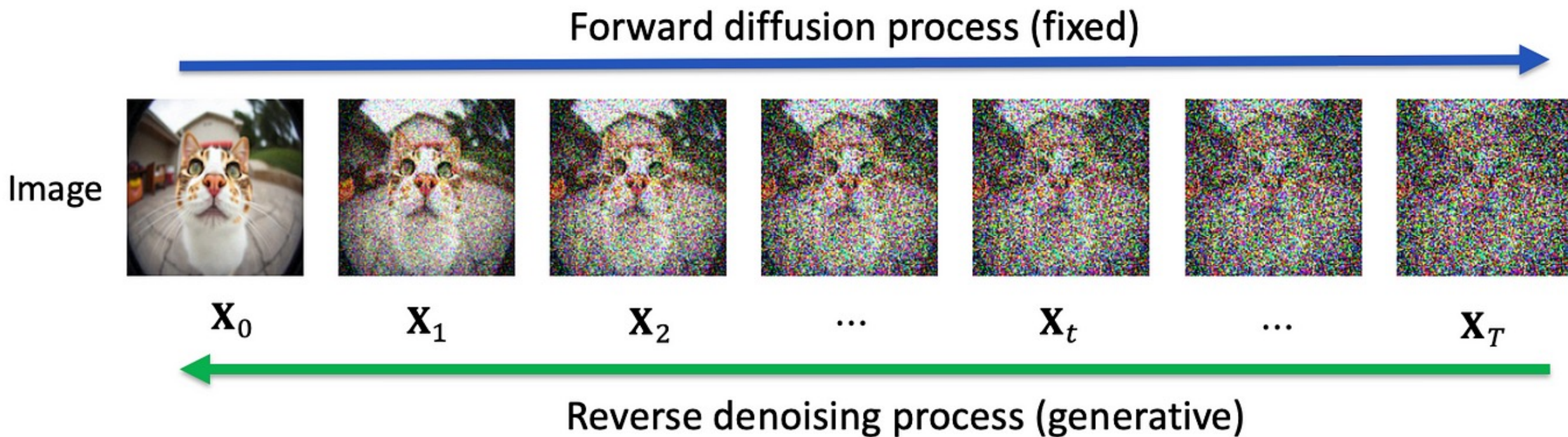
An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021.

MLLM: Multi-Modal Large Language Models



扩散模型(diffusion model, 2015)

- 去噪过程：这些模型逐步添加噪声（前向过程），并学习反向（后向过程），有效地去噪以产生样本
- 马尔可夫链：这两个过程都被构造为马尔可夫链，每个向前的步骤都会添加高斯噪声，模型会学习如何反向去除高斯噪声
- 训练目标：目标是最小化每一步预测和实际噪声之间的差异



Text-to-Image

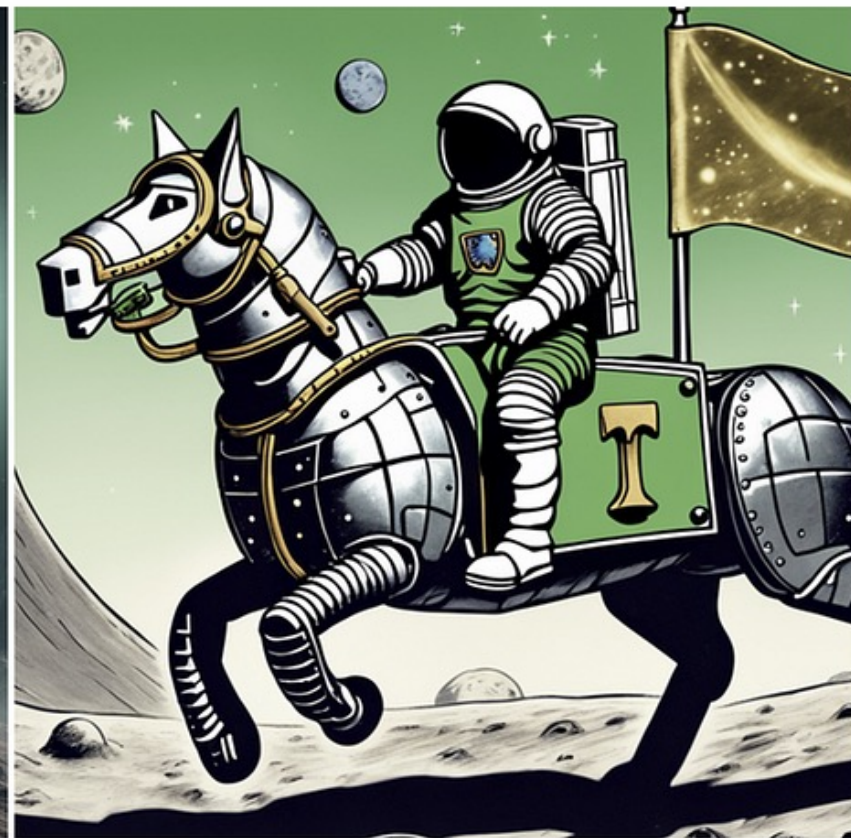
DALL-E 3



MIDJOUREY 5.2



STABLE XL



Text-to-Video

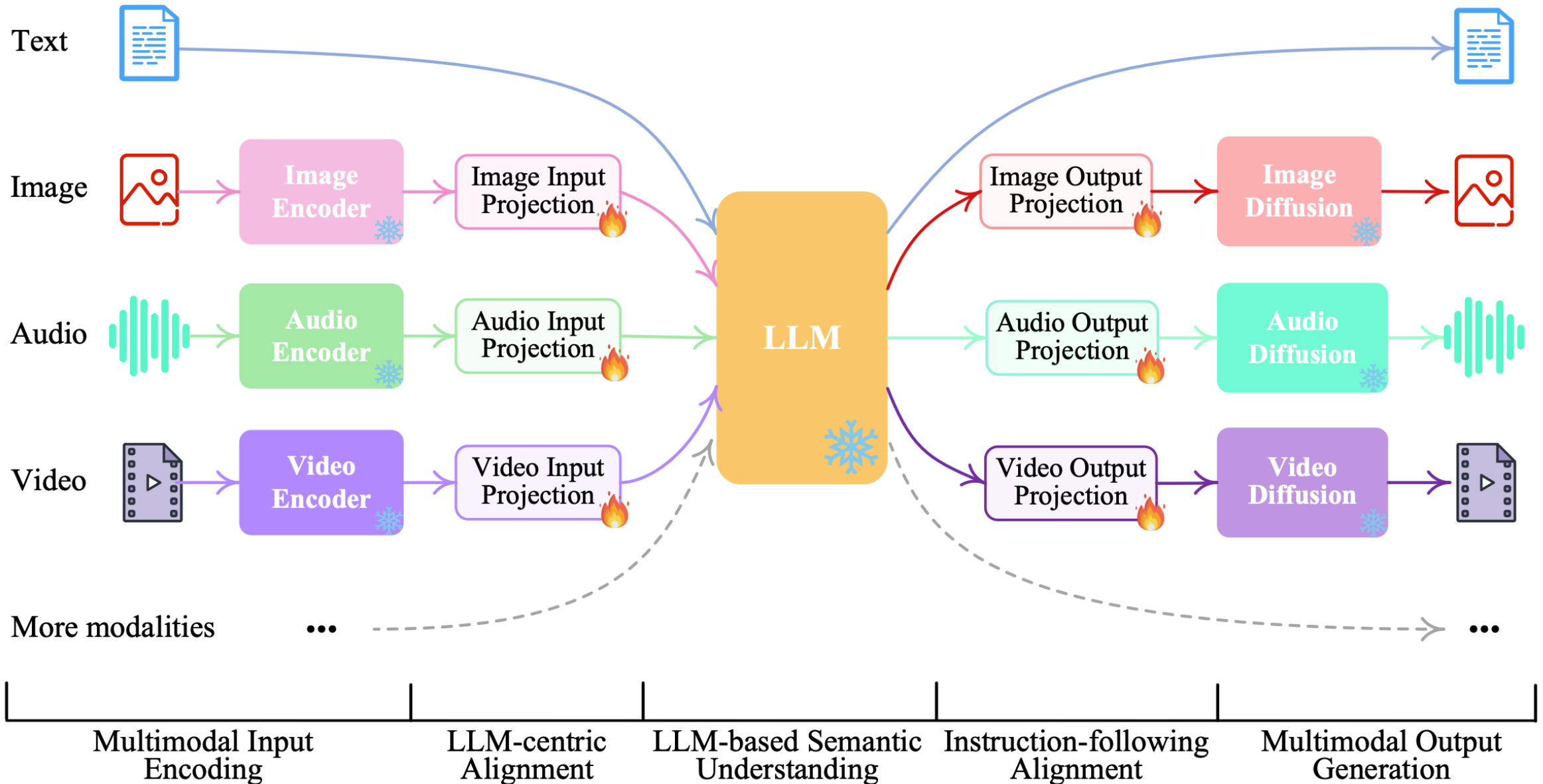


Prompt:
**A cartoon kangaroo
disco dances**



Prompt:
**A litter of golden retriever puppies
playing in the snow. Their heads pop
out of the snow, covered in.**

Multi-Modal Large Language Models



总结

- 掌握MP神经元与多层感知机
- 能够熟练推导BP算法并代码实现
- 了解深度学习的成功的原因(表示学习)、面临的挑战及训练Tricks
- 了解用于处理图像数据的卷积神经网络
- 了解用于处理序列数据的循环神经网络, 知道Transformer核心机制
- 知道多模态模型是什么