AI 的历史
现状与发展

ZOMI

Course chenzomi12.github.io

GitHub github.com/chenzomi12/DeepLearningSystem

Talk Overview

1. AI 系统概述

- AI 历史，现状与发展
- 算法与体系结构的进步
- AI 系统的组成与生态
- 大模型对AI系统的挑战

2. AI 芯片

3. AI 编译器

4. AI 推理引擎

5. AI 开发框架

6. 异构集群调度与管理

7. AI 大模型

Talk Overview

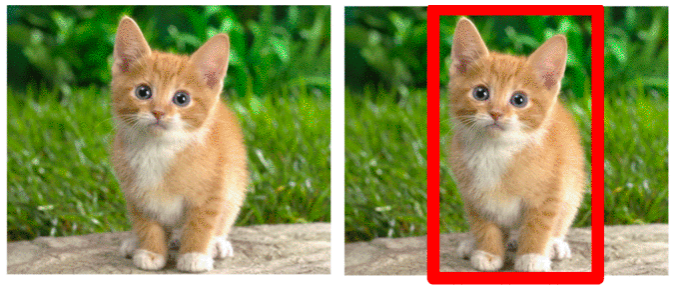
- **AI 历史，现状与发展**

1. The widespread application of AI - AI 的广泛应用
2. Learning methods for AI - AI 的学习方法
3. AI algorithm, model status - AI 算法现状
4. AI algorithm, model trends - AI 算法趋势

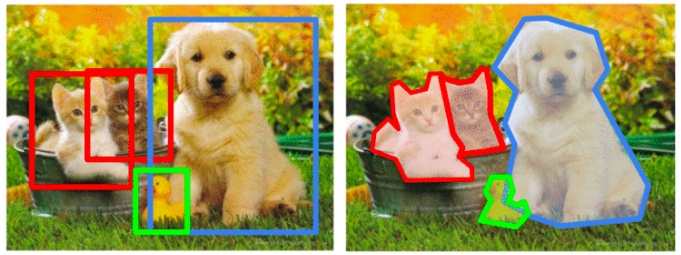
1. AI 广泛应用

AI 的广泛应用：面向领域

CV

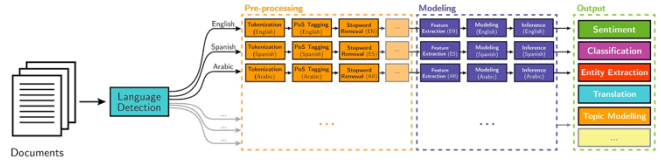


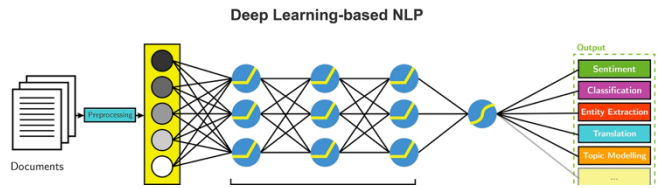
Single object

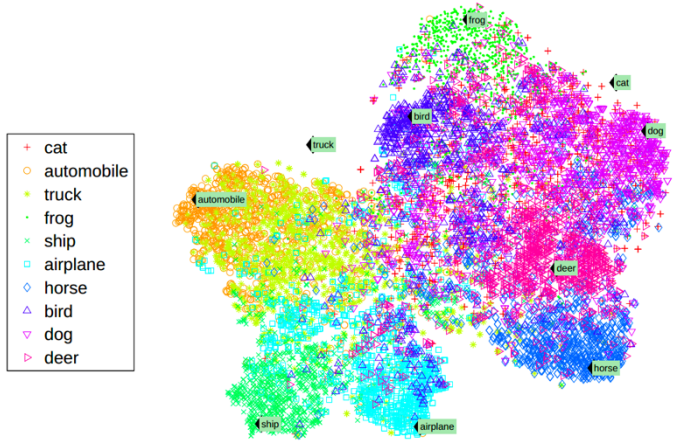


Multiple objects

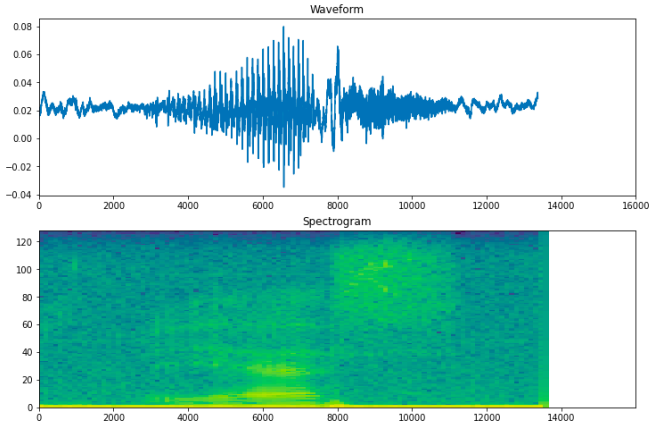
NLP

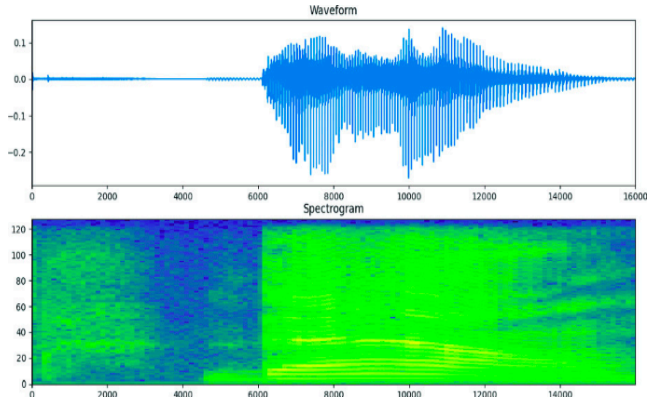






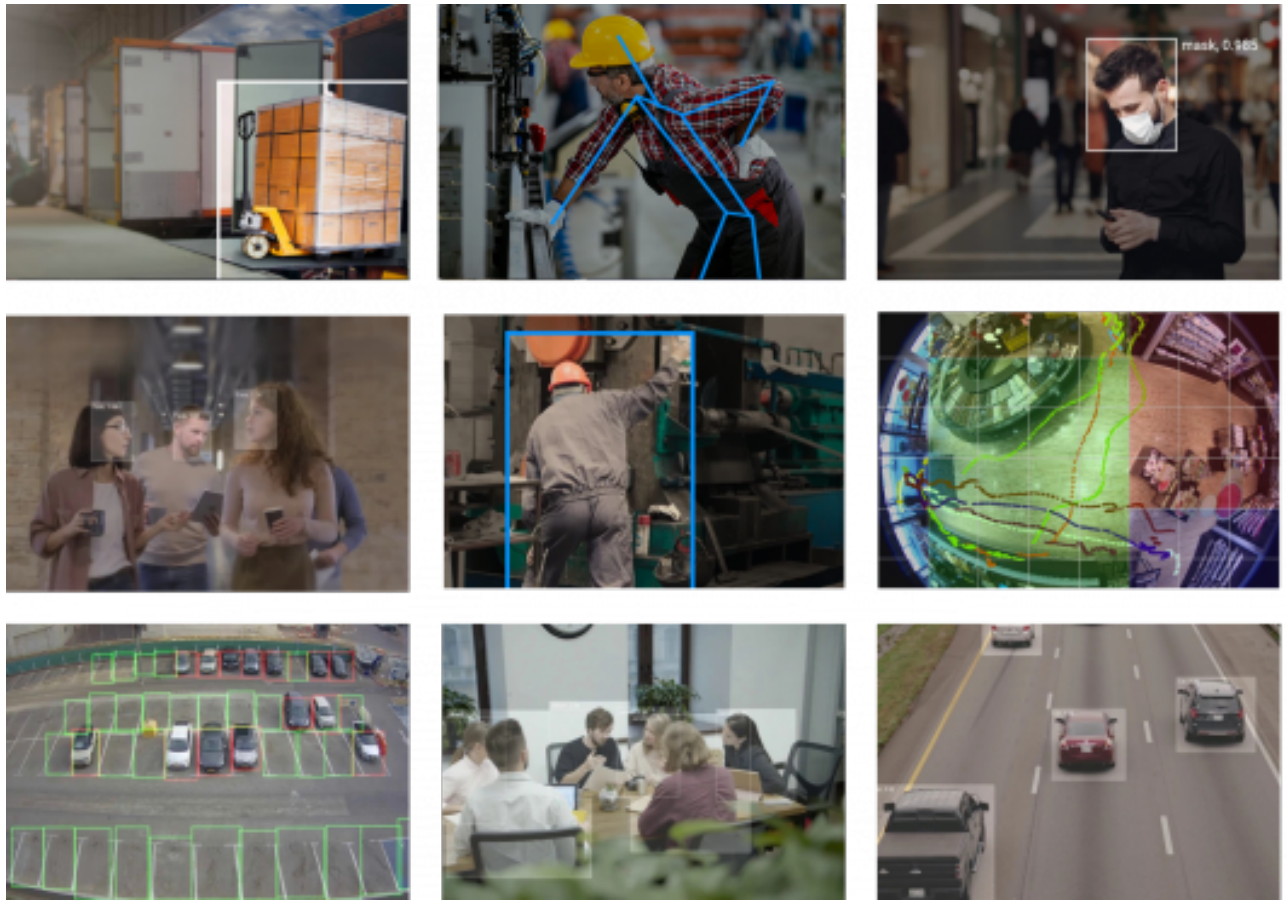
Audio



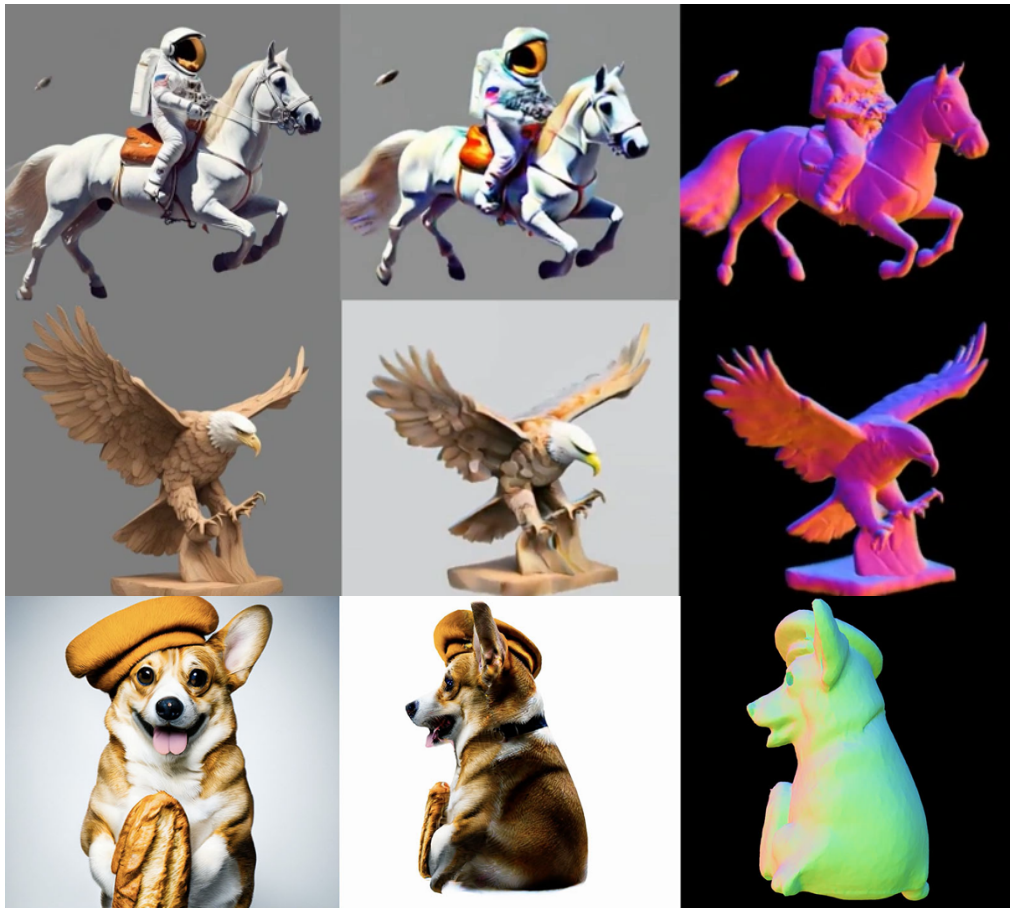


CV 领域应用案例

Classification/Detection/Segmentation

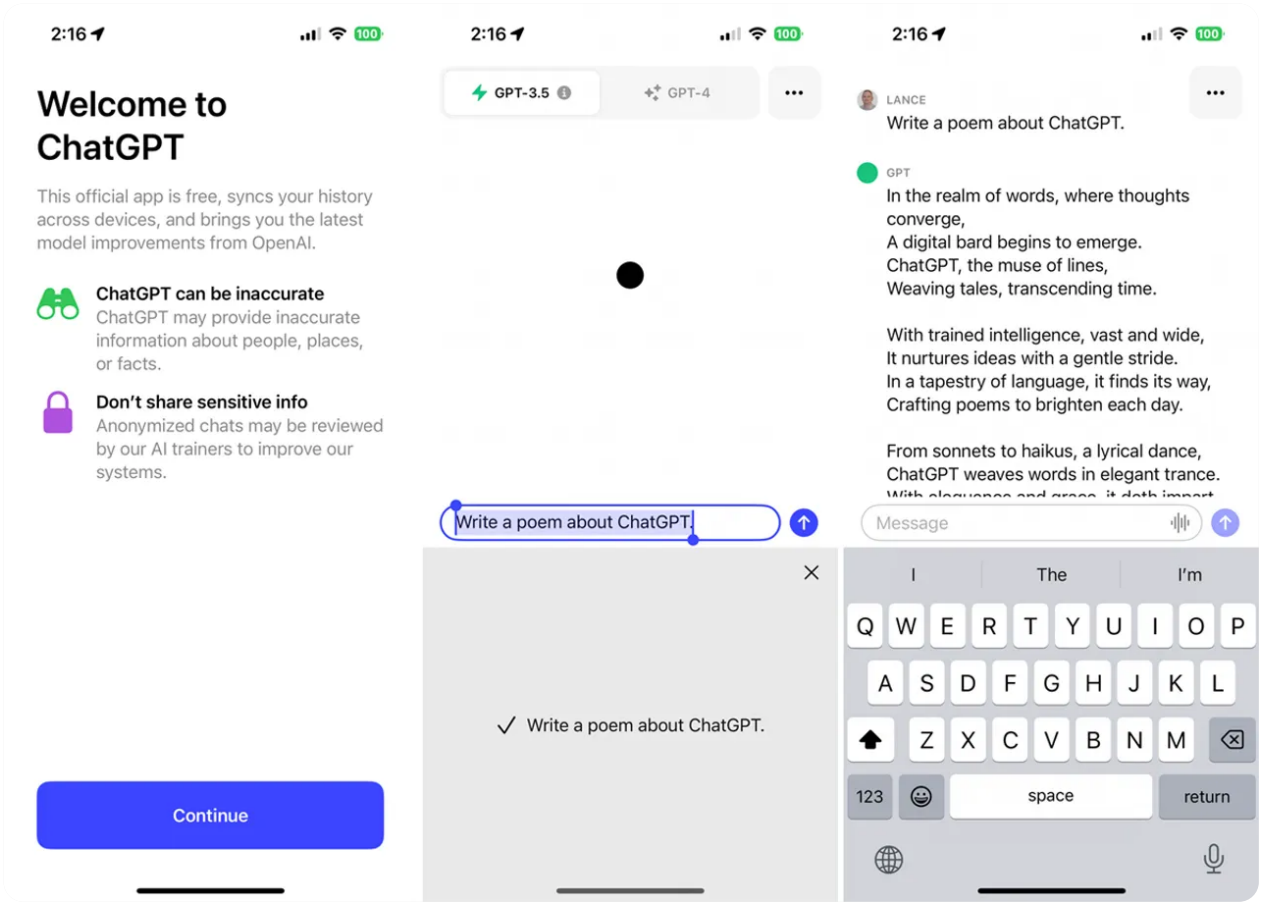


Generation/3D reconstruction



NLP 领域应用案例

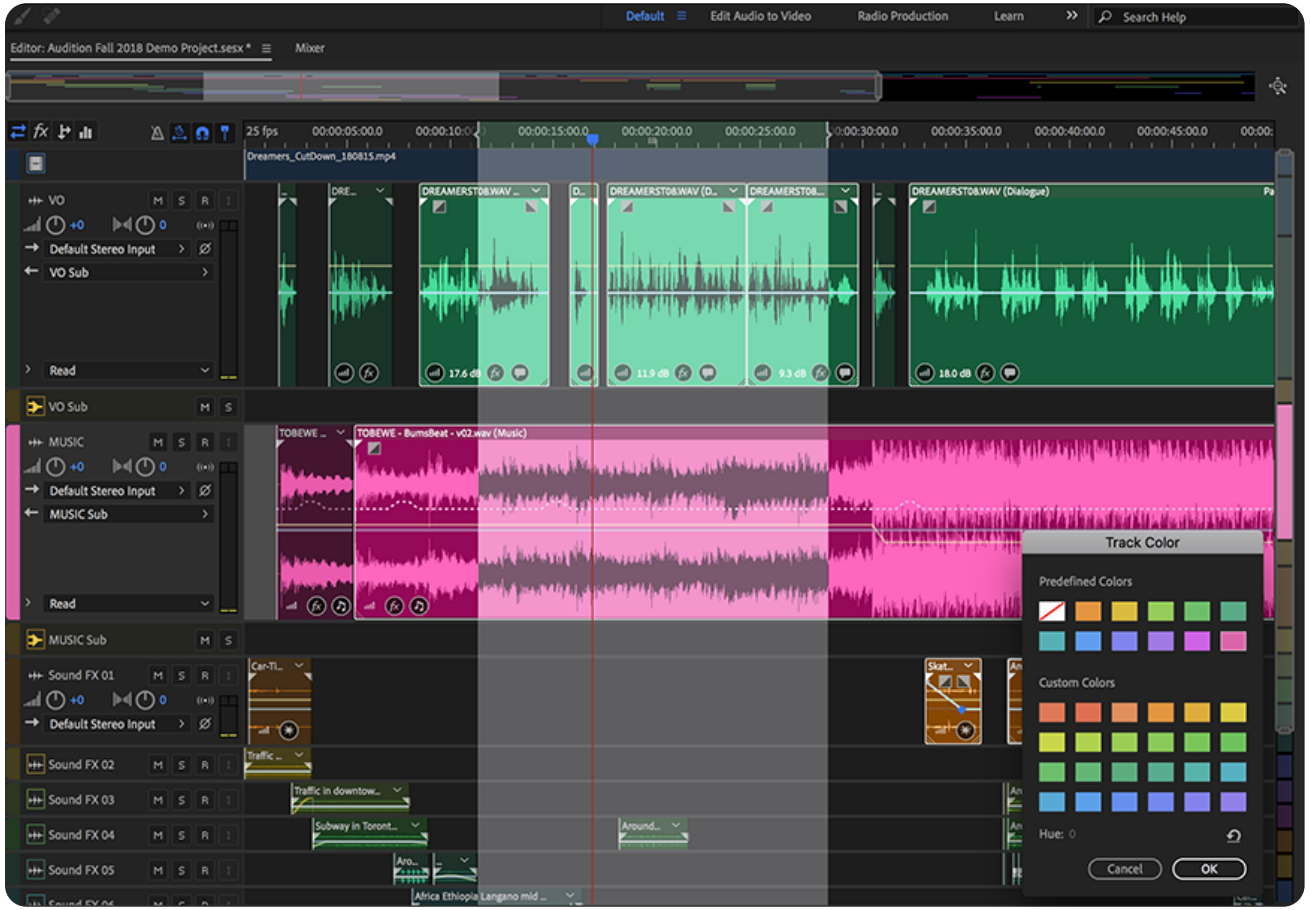
Dialogue/Prompt/Info Search



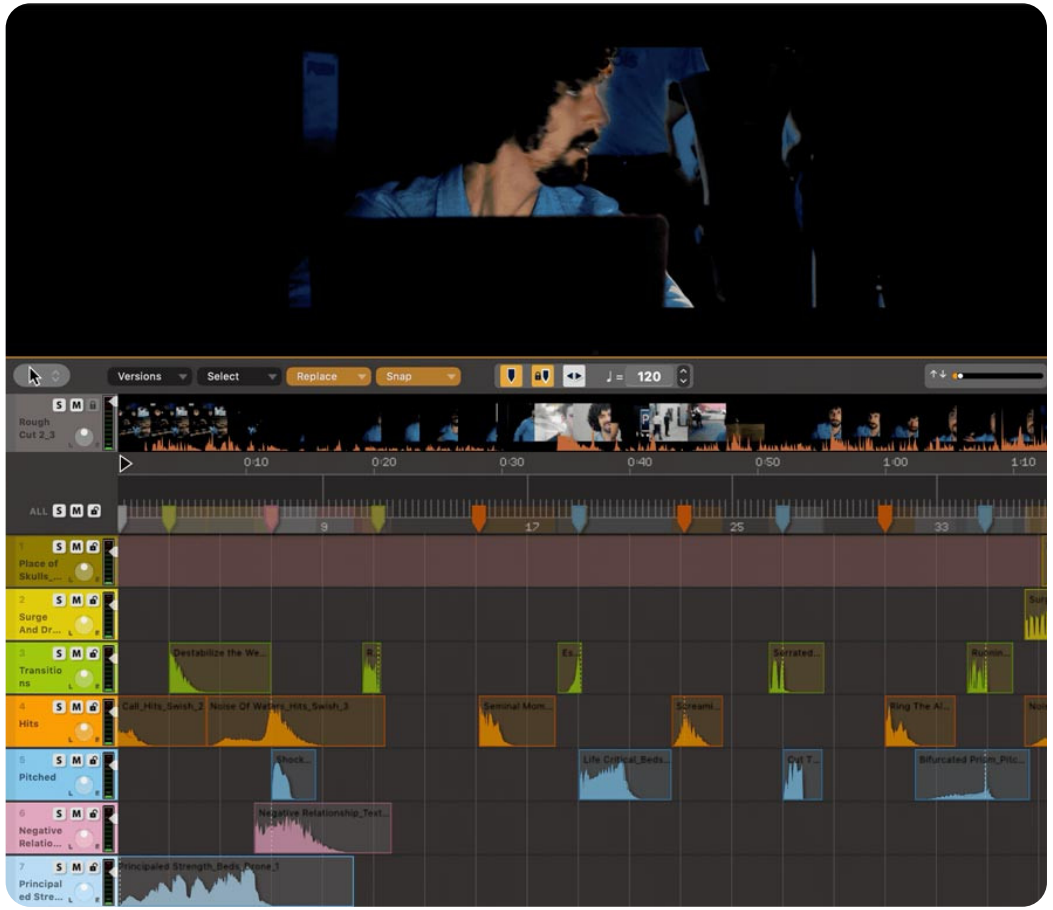
Text Generation/Summary

Audio 领域应用案例

Audio Generator/Audio Editor Improve

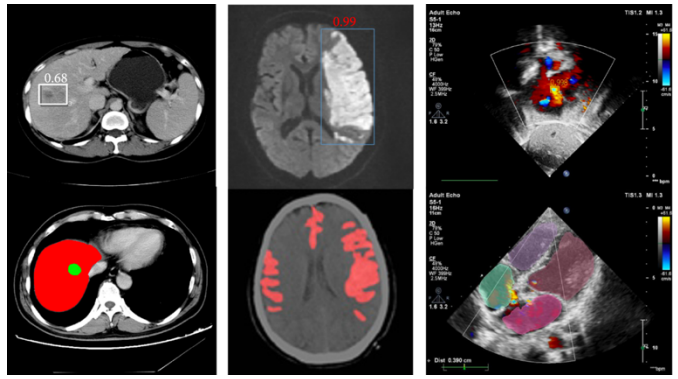


Classification/Alignment

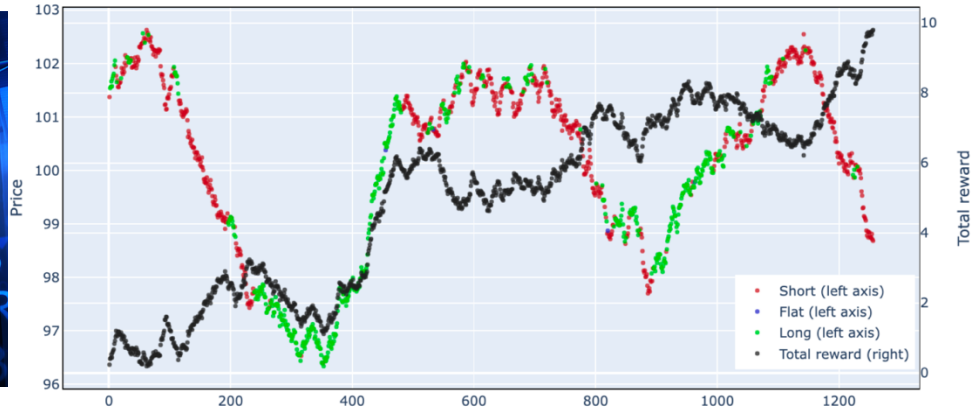


AI 的广泛应用：面向行业

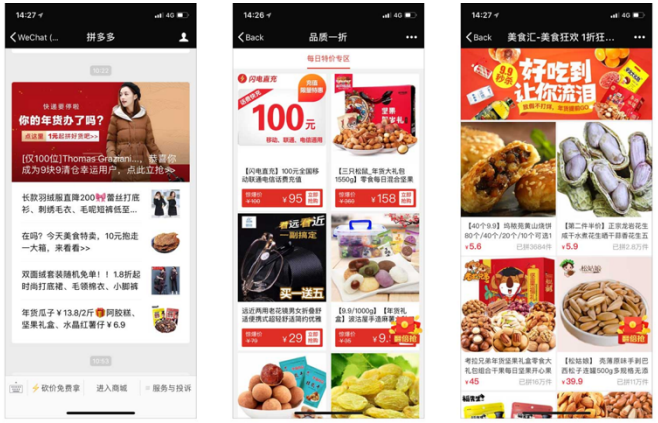
Medical



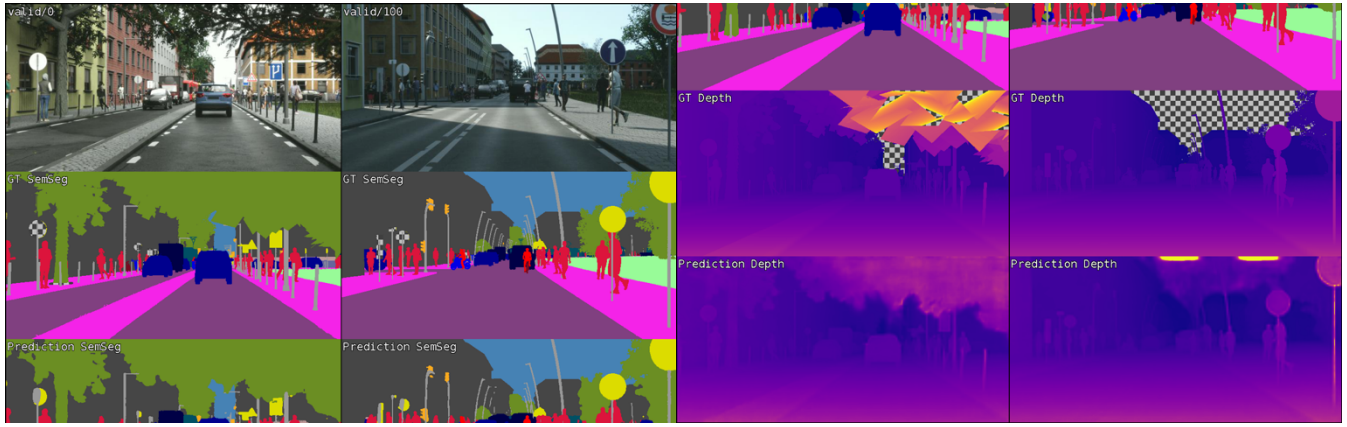
Quantitative trading



Recommend

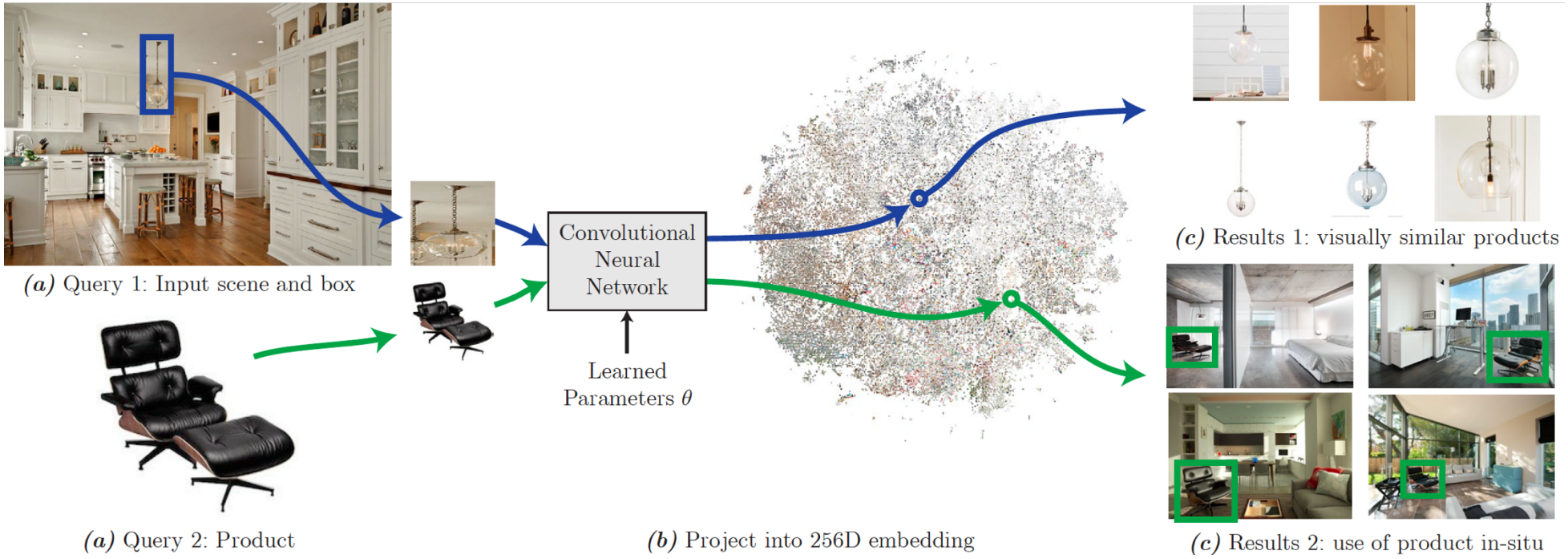


Autonomous driving



AI 的广泛应用：互联网

- 谷歌、百度、微软必应（Bing）等公司通过人工智能技术进行更好的文本向量化，提升检索质量，同时人工智能进行点击率预测，获取更高的利润。



(a) Query 1: Input scene and box

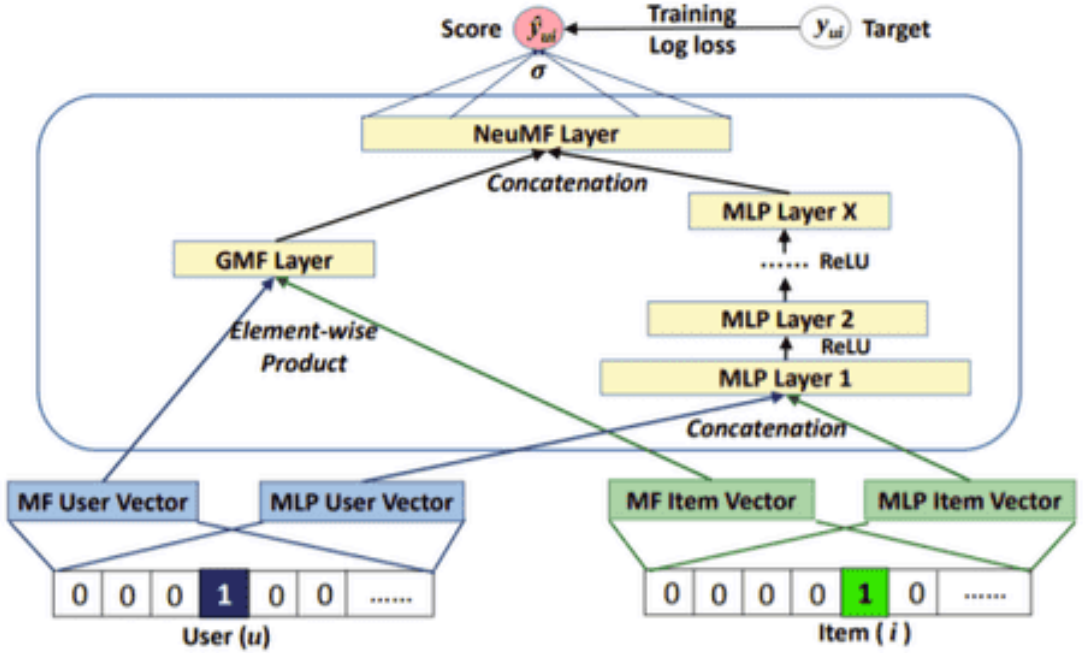
(a) Query 2: Product

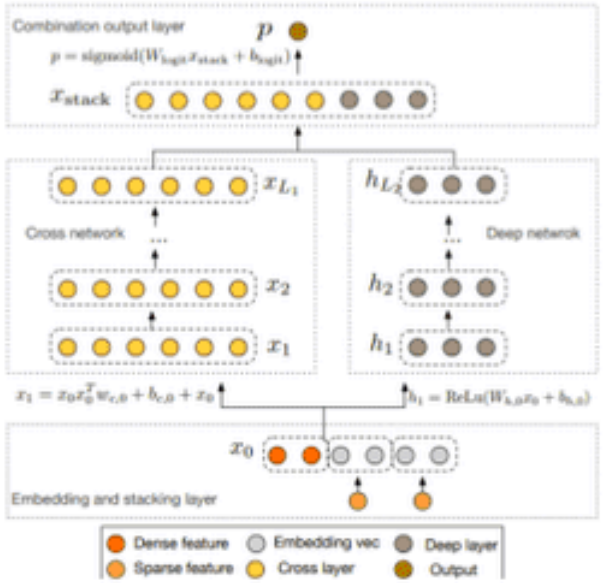
(b) Project into 256D embedding

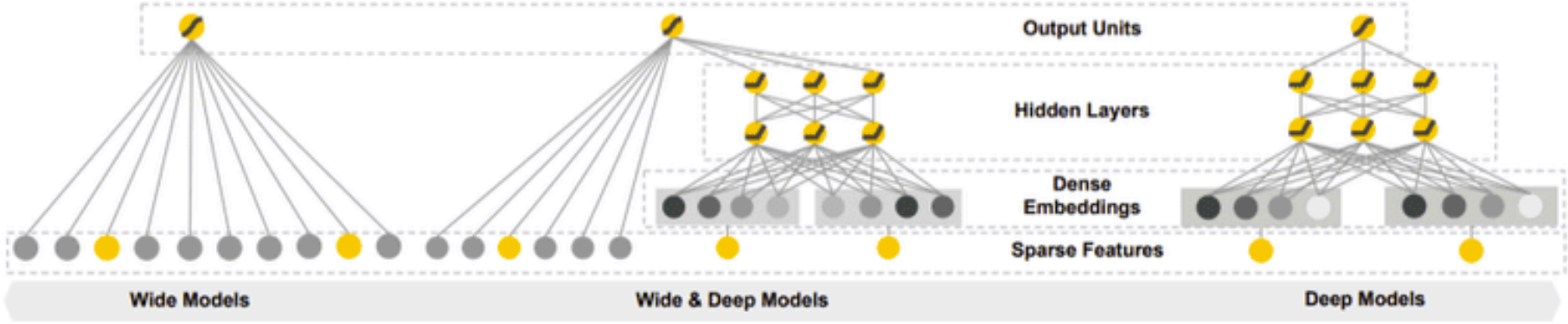
(c) Results 1: visually similar products

(c) Results 2: use of product in-situ

AI 的广泛应用：互联网

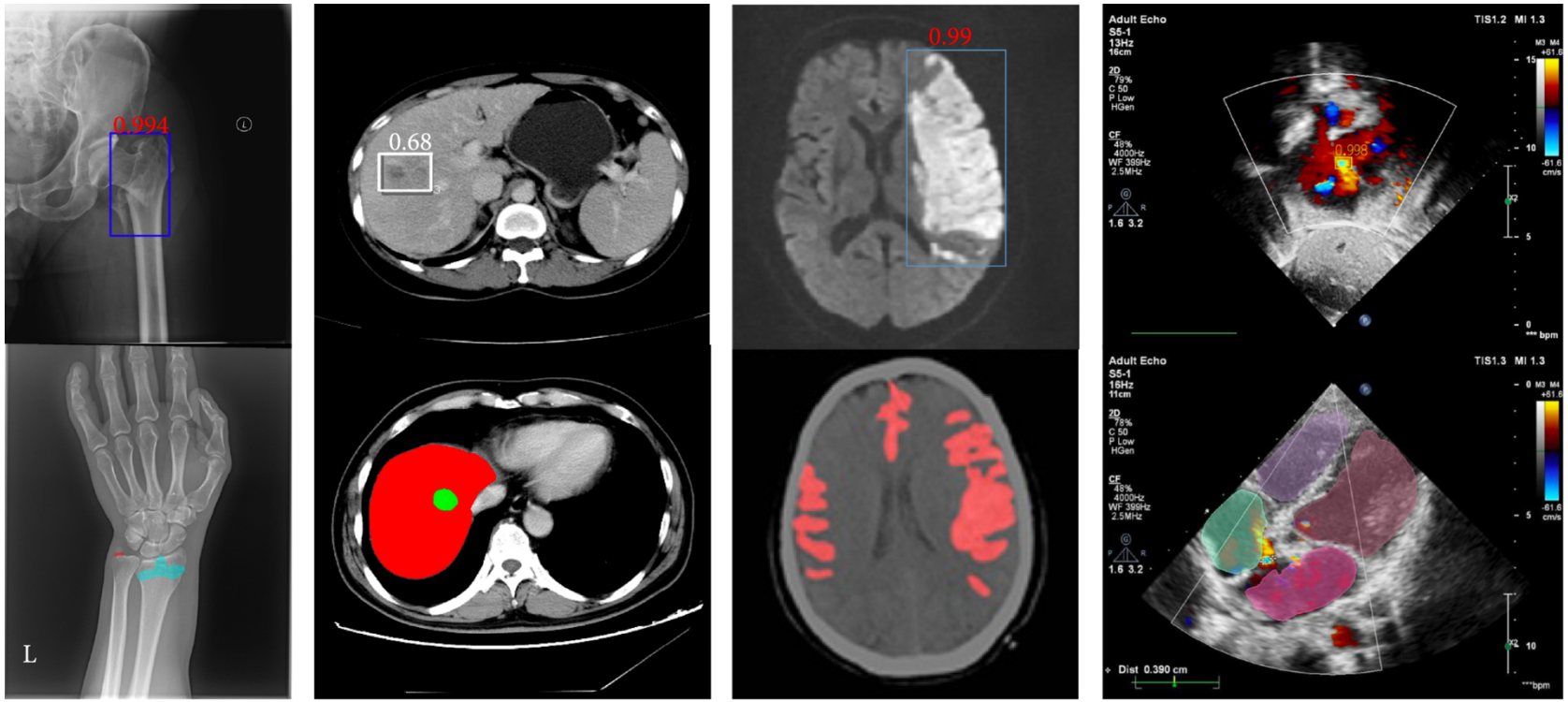

Score
Training
Log loss
Target
NeuMF Layer
Concatenation
GMF Layer
MLP Layer X
ReLU
MLP Layer 2
ReLU
MLP Layer 1
Element-wise Product
Concatenation
MF User Vector
MLP User Vector
MF Item Vector
MLP Item Vector
User (u)
Item (i)


Combination output layer
Cross network
Deep network
Embedding and stacking layer
Dense feature
Sparse feature
Embedding vec
Cross layer
Deep layer
Output


Output Units
Hidden Layers
Dense Embeddings
Sparse Features
Wide Models
Wide & Deep Models
Deep Models

AI 的广泛应用：医疗

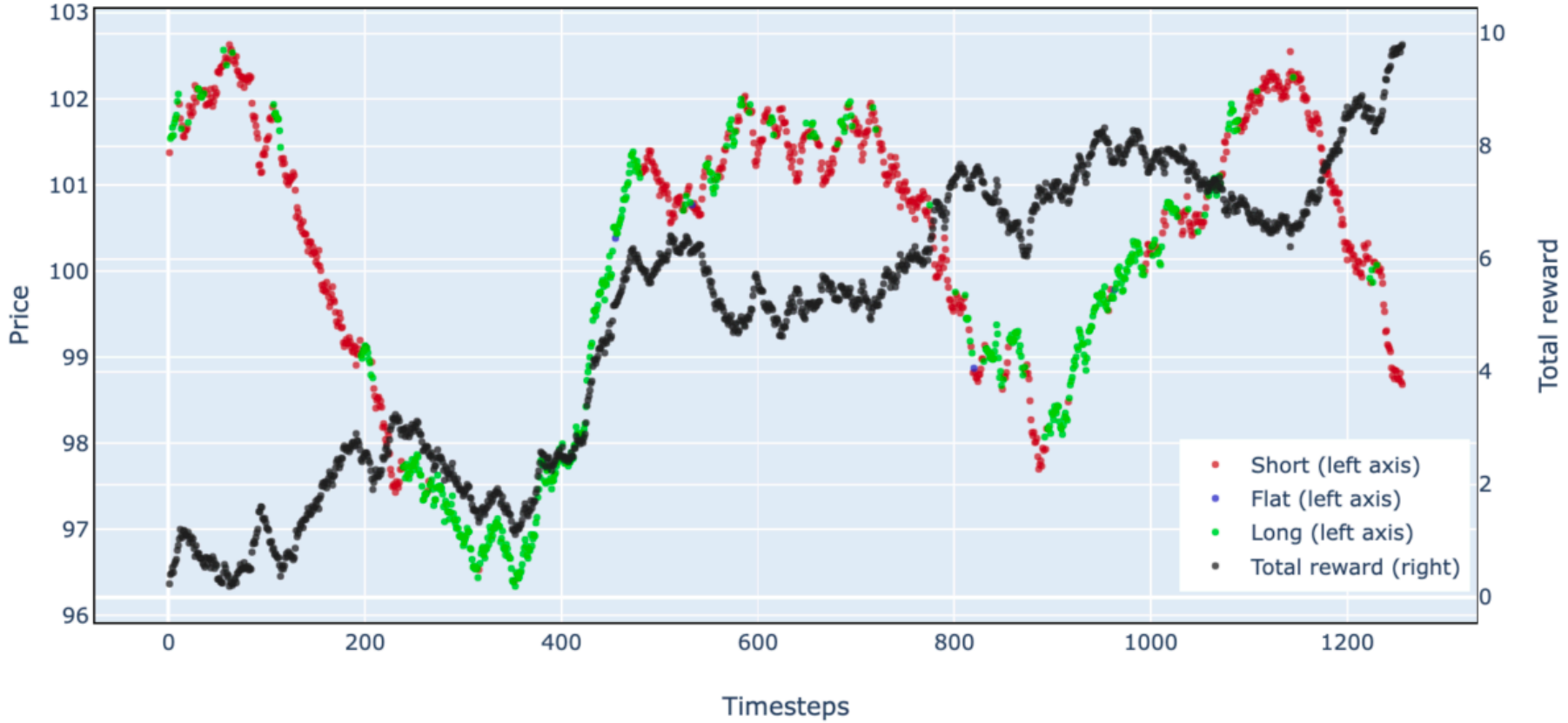
- IBM 沃森（Watson）从海量的医学文献和病历中提取医生临床诊断经验，通过让人工智能模型学习掌握临床诊断方法，辅助医生进行诊断。



Bone X-ray　　Liver CT　　Brain MRI　　Cardiac ultrasound

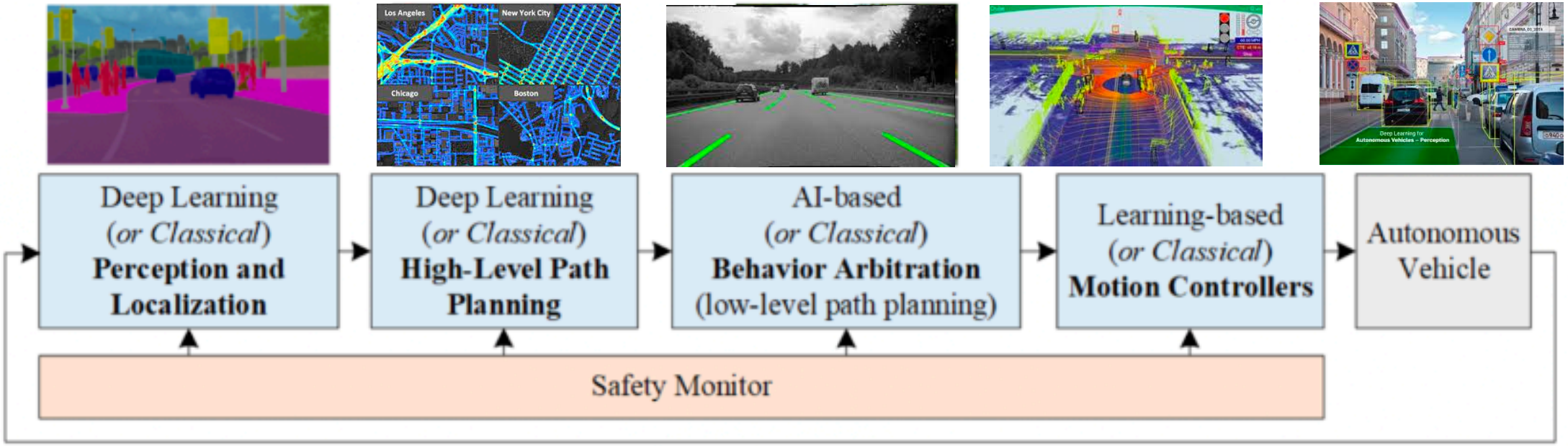
AI 的广泛应用：金融

- 通过反欺诈、关联分析、时序预测、量化交易等 AI 算法可以较早识别风险，并预测未来发展趋势。



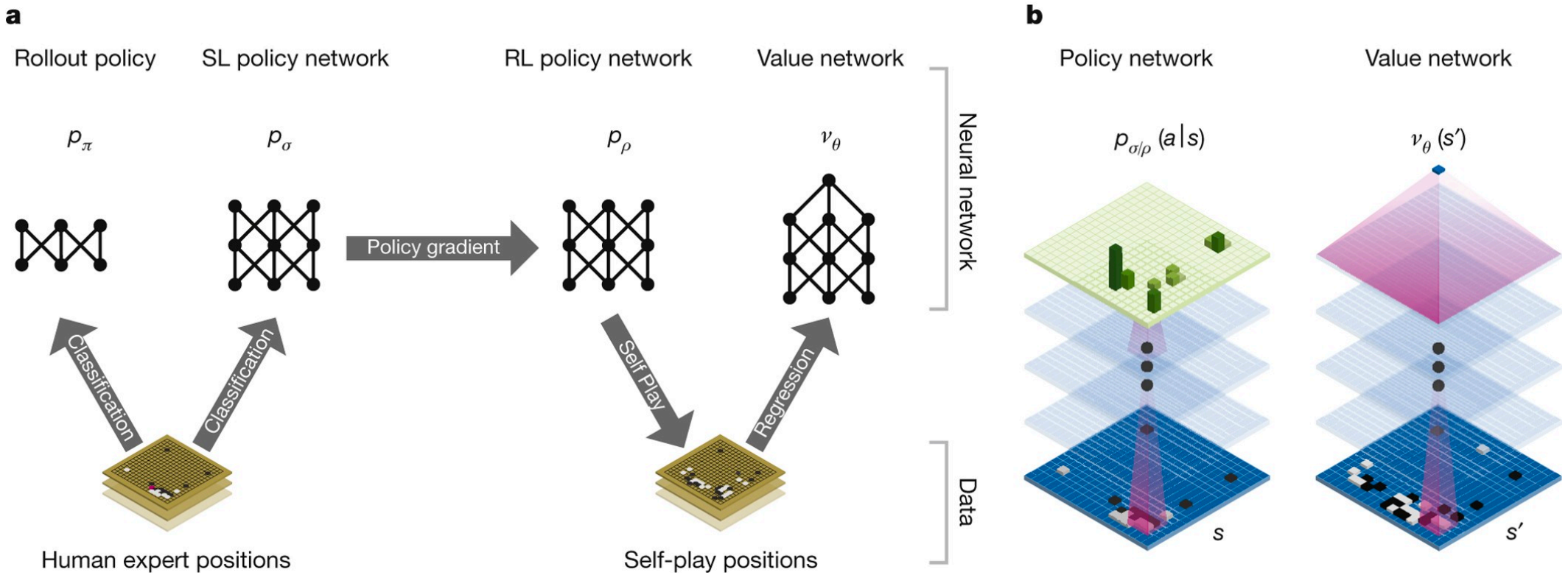
AI 的广泛应用：自动驾驶

- 通过物体检测模型能够进行更好的路标检测，道路线检测进而增强自动驾驶方案。



AI 的广泛应用：游戏

- 在游戏中通过强化学习技术进行对战，设计新的策略，提升游戏体验。



AI 的广泛应用：安防

- 通过AI技术与安防软硬件的结合,实现事前预防、事中相应预警、事后追查、省时省力的安防管控,解决了传统安防只能事后取证,且取证难的痛点。

移动应用


10:39
Search
4G
洛基假扮绿巨人
提问
关注
推荐
热榜
你见过最土的土豪有多土?
刘嘉儿: 今年大年初二 在广州去昆明的动车上 二等座 我看到一位大约30岁的男子 穿着一双我叫不...
51,333 赞同 · 5,105 评论
一个男生的魅力体现哪里?
山雞: 我喜欢上我男朋友，是因为有一次他去给我买奶茶。 付钱的时候，他拿出了他的钱包，钱包很破旧，但是他面色如常，举止自然。 我当时就很佩...
9,196 赞同 · 1,766 评论
怎样低价入住五星级酒店? 我们来帮你
你要知道，其实速8和希尔顿酒店的价格在某些时候差距并不大，这个APP里的很多高级酒店都是平民价，性价比很高
广告 艺龙旅行 · 查看详情
长期运动对外貌的改变有多大?
江离: 大学四年就是我的变化史哈哈哈哈哈哈 三年除了生理期和受，几乎保持了每天至少1km的...
首页
想法
市场
消息
我的


10:35
4G
私家课
我的私家课
法律英语训练营：法律圈...
职问 • 职场新人的在线大学
¥399
试听
理财实战课：用基金定投...
罗元裳 • 7 分钟理财 CEO
¥69 ¥99 特价
试听
职场新人必修课：职业规...
Sophia肖 • 换行CEO，前长...
¥99
试听
听复旦名师妙解传统文化...
傅杰 • 复旦大学中文系博导
¥99
试听
睡眠革命：17 课还你高...
马春树 • 北京大学医学部医...
¥99
试听
互联网运营课：从外行到...
张亮-Leo♂ • 畅销书《从零...
¥48.3 ¥69 特价
试听
首页
想法
市场
消息
我的

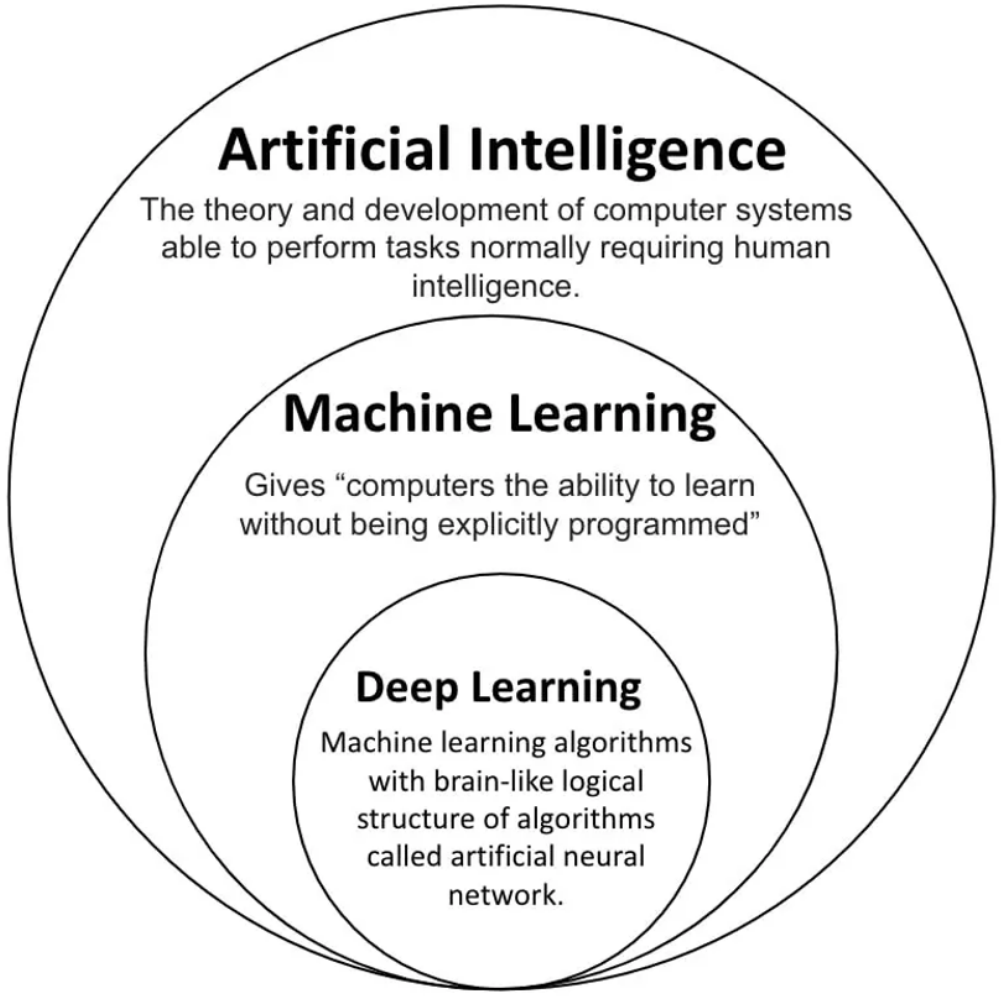

小红书
关注
推荐
减脂
明星
视频
健身塑型
这瘦脸太绝了
紧致下颌线!
瘦脸快来🔥0成本get紧致下颌线 !!告别双下巴
爱运动的元气少女CiCi 7万
挖到一双奶茶色踩屎感增高7cm的老爹鞋
轻描淡鞋 3048
mini vlog | 沉浸式化妆routine
江疏影 8531
深蓝色
+
黑色
啊啊啊发现一家超🔥的宝藏店铺！！！
深远旺仔 5894
脸真的小很多诶！！
首页
购物
消息
我


keep 奖牌
潮流买手店
新趋势
福利社
户外主义
宝藏好店
开学报到妆
甜酷厚底鞋
卡通餐具
笔尖春天
开学穿什么
"买份呵护礼"
点击购买
诗裴丝·3款香型全系列免洗空气感蓬松喷雾干发神器 懒
¥63
蜜桃挂脖运动长袖带胸垫假
直戳心巴巴可爱五公主套装们/发饰分享
首页
购物
消息
我

2. AI 学习方法

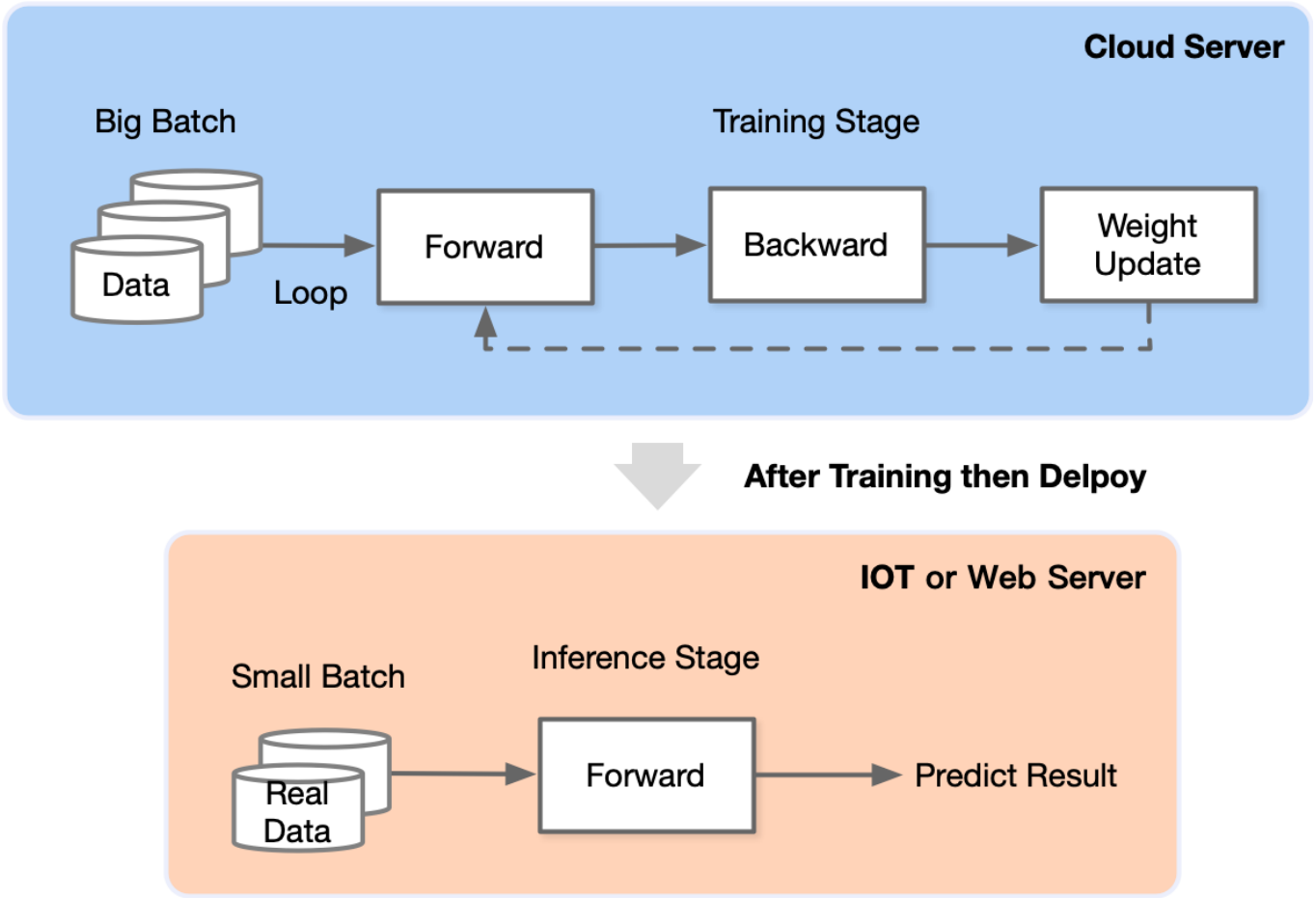
AI 的学习范式

- 机器学习主要研究内容如何“学习”，因此学术圈提出了非常多的“学习”方法和范式，例如迁移学习、强化学习等，都属于 AI 的学习范式。



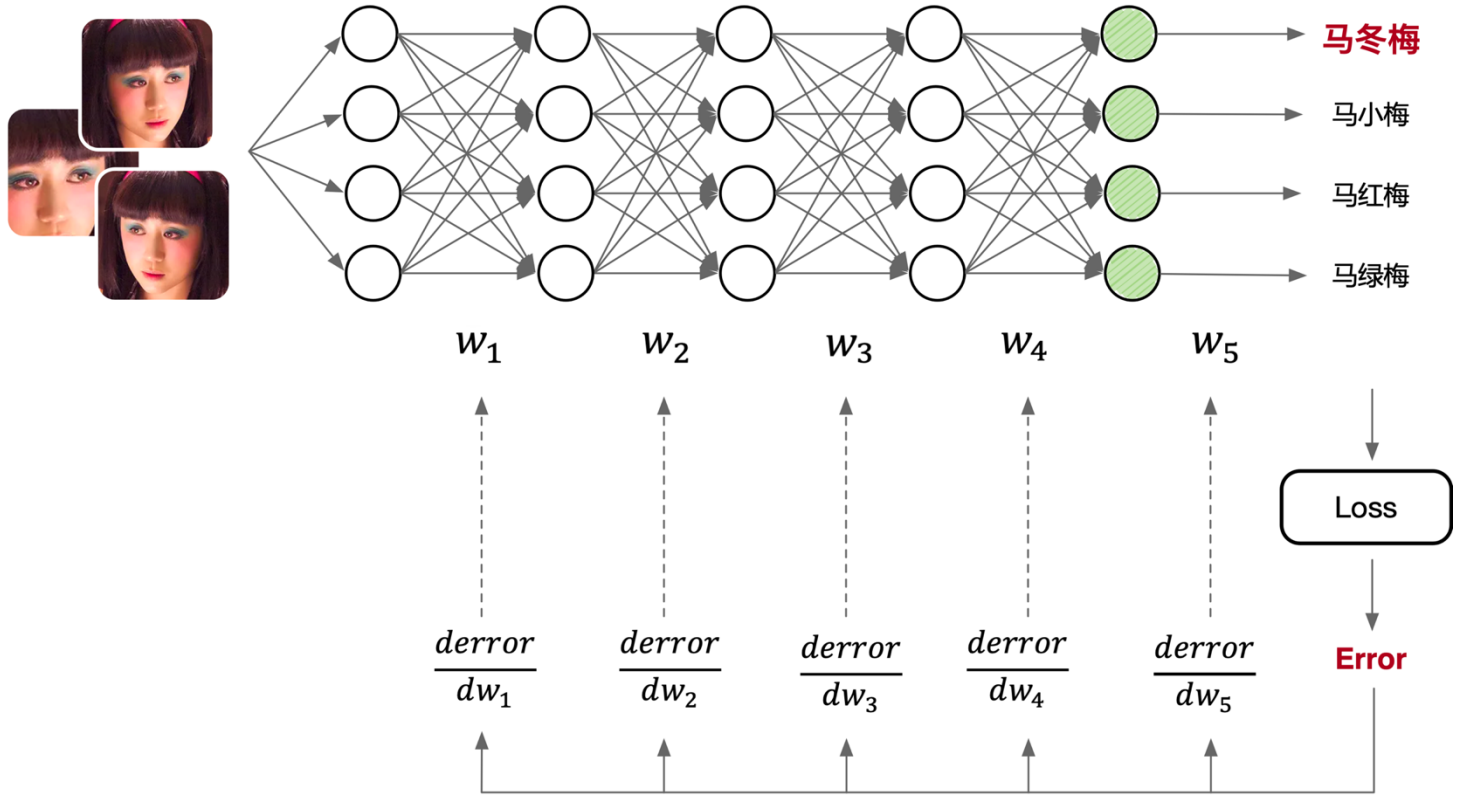
深度学习的基本流程

- **训练**通过设计合适 AI 模型结构以及损失函数和优化算法，将数据集以 mini-batch 反复进行前向计算并计算损失，反向计算梯度利用优化函数来更新模型，使得损失函数最小。训练过程最重要是梯度计算和反向传播。
- **推理**在训练好的模型结构和参数基础上，一次前向传播得到模型输出过程。相对于训练，推理不涉及梯度和损失优化。最终目标是将训练好的模型部署生产环境中。



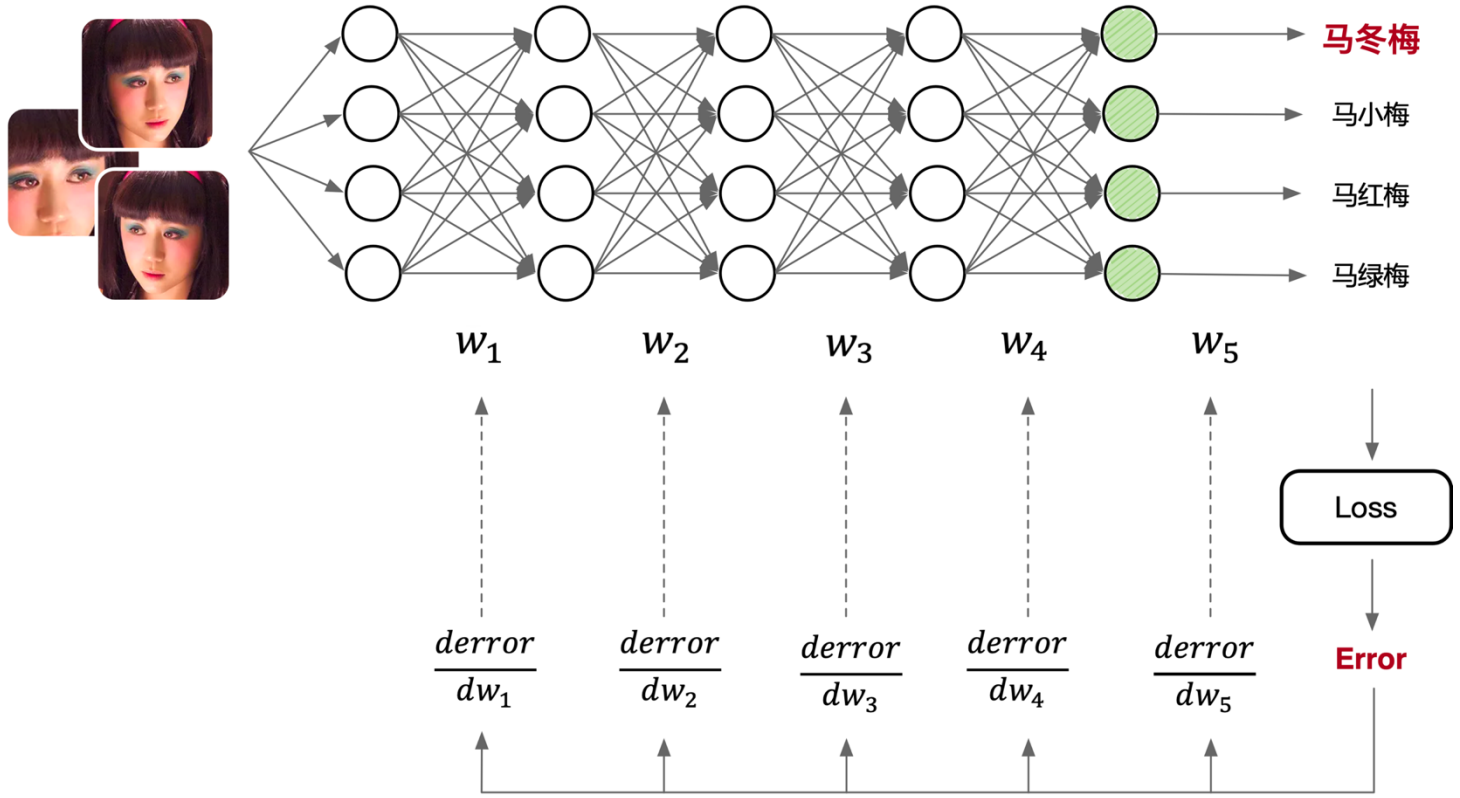
AI 的学习方法：特征与样本

- 确定深度学习模型的输入特征（Feature）与输出标签（Label）数据样本（Sample）：
  - 给模型输入图片，输出是图片类别。用户需要提前准备好模型输入输出数据，进而展开后续的训练。



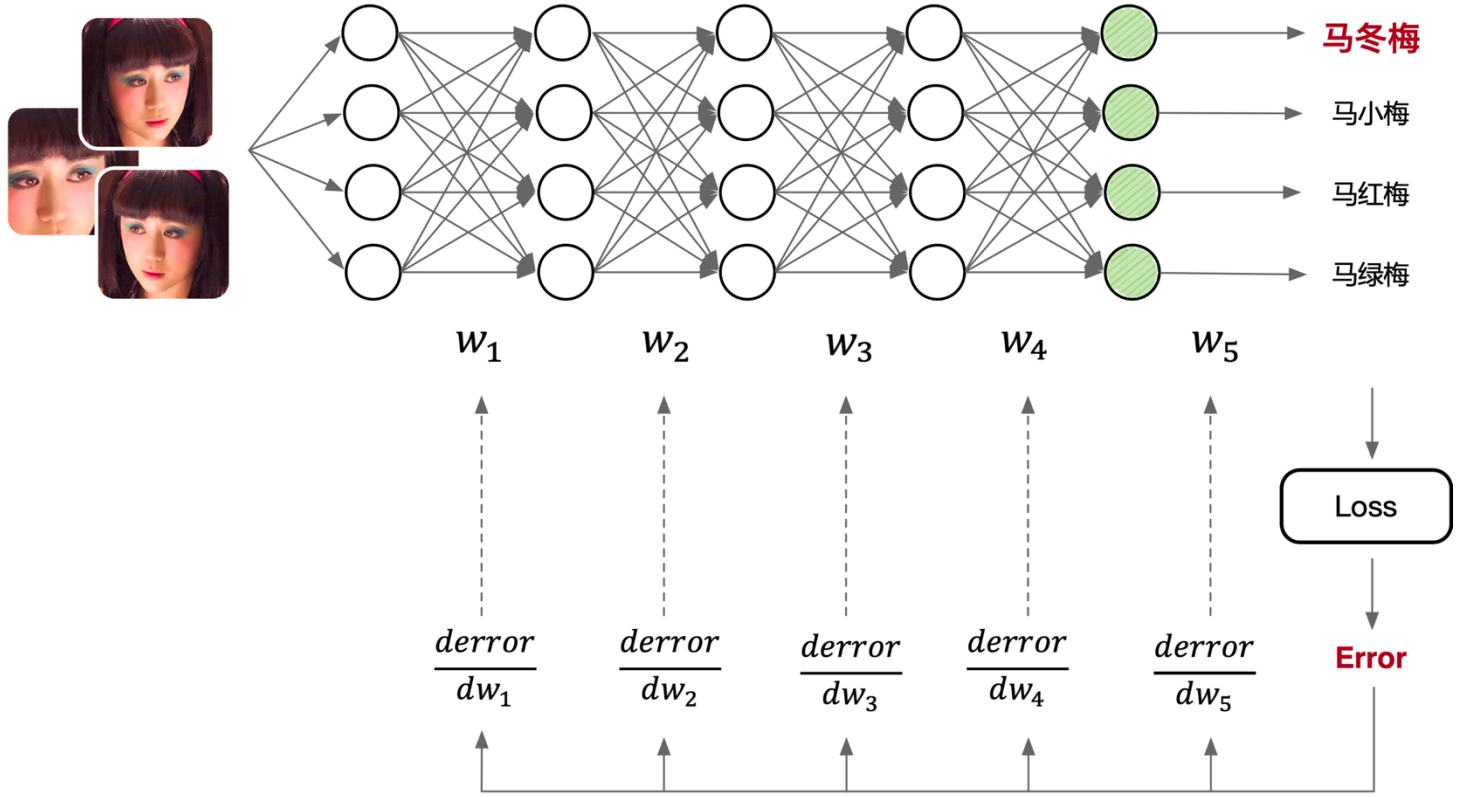
AI 的学习方法：确定神经网络模型

- 设计模型结构：开发者通过 AI 框架开发了神经网络模型结构，实线代表权重、圆代表对输入特征数据计算的操作。其中 w_n 表示权重，即可以被学习和更新的数值。



AI 的学习方法：训练过程 Training

- 训练过程就是根据用户给定的带有标签的数据集，不断通过梯度下降算法，学习出给定数据集下最优的模型权重 w_n 的取值。



AI 的学习方法：训练过程 Training

1. 前向传播（Forward Propagation）：

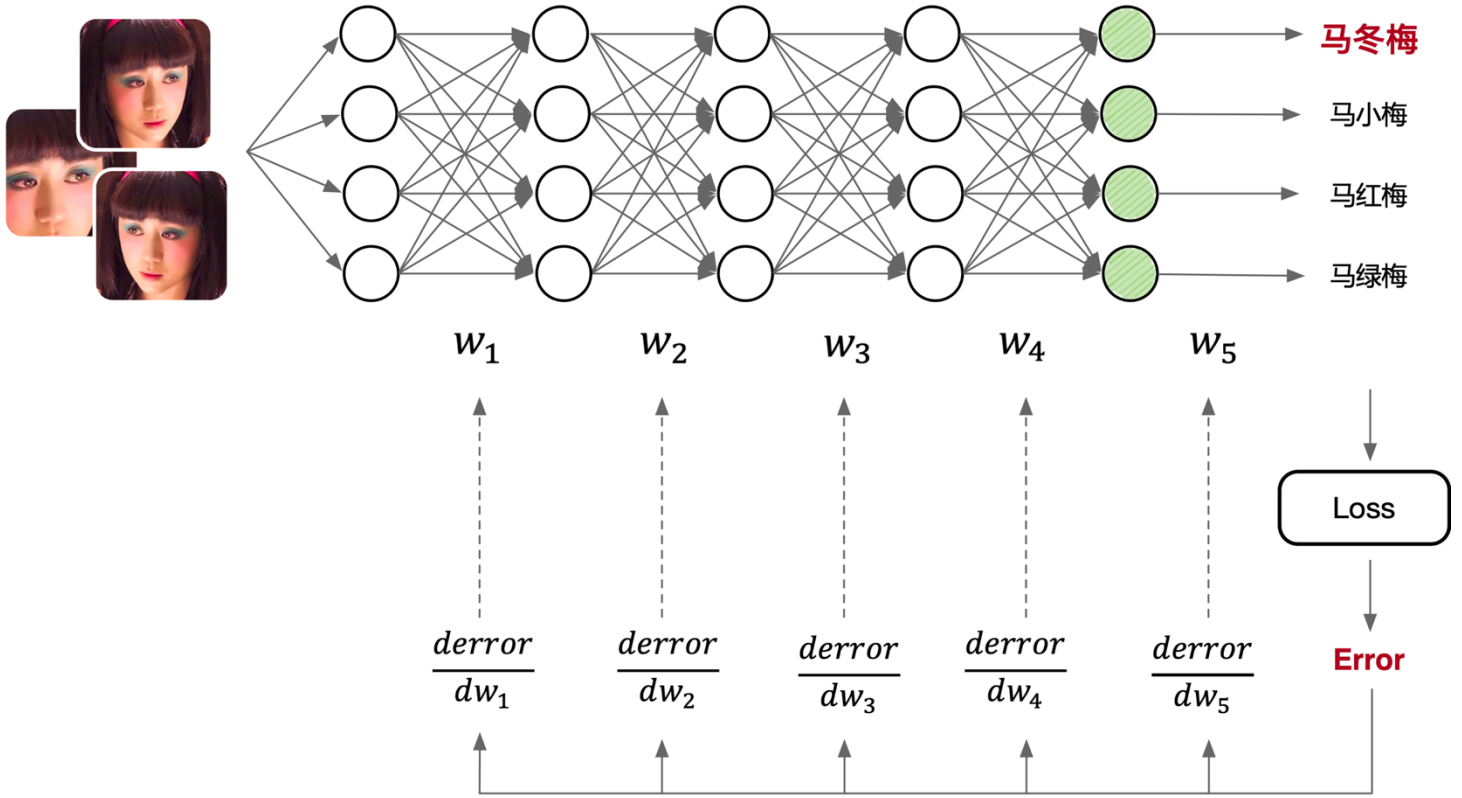
   输入到输出各层计算（如卷积、池化层等），产生输出并完成损失函数计算。

2. 反向传播（Back Propagation）：

   由输出到输入反向完成整个模型中各层的权重和输出对损失函数的梯度求解。

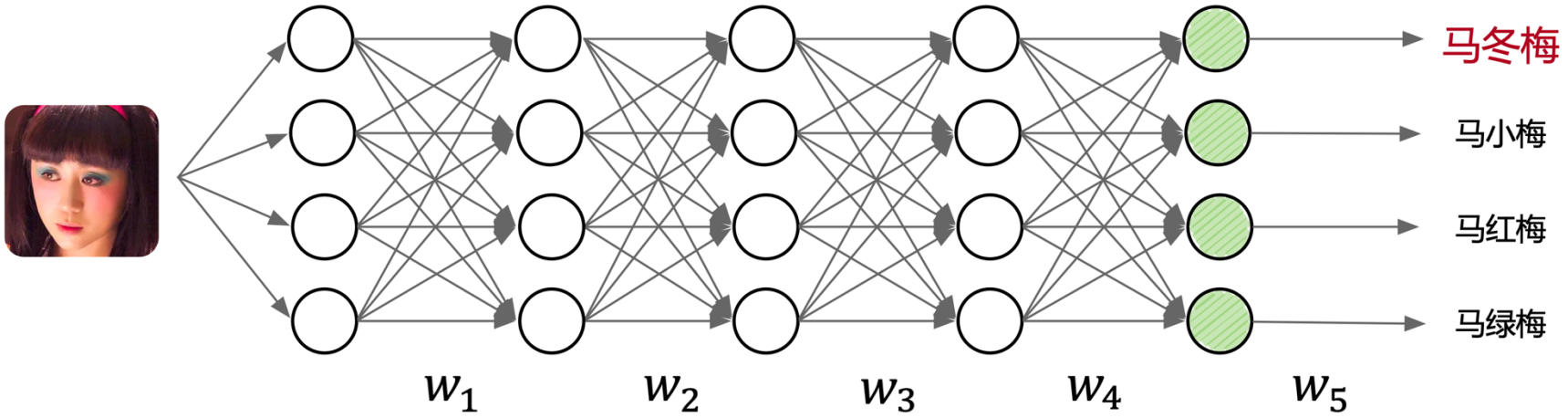
3. 梯度更新（Weight Update）：

   根据指定的指定学习率，对模型权重通过梯度下降算法完成权重的更新。



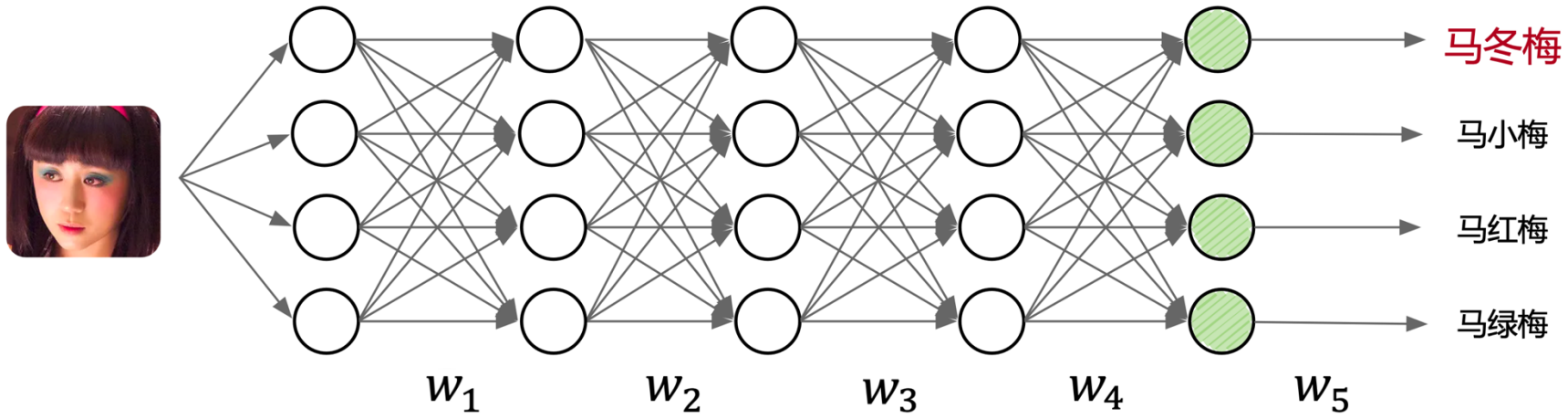
AI 的学习方法：推理过程 Inference

- 当完成模型训练，意味着在给定的数据集上，模型已经达到最佳或者满足需求的预测效果。如果 AI 算法工程师对模型预测效果满意，就可以进入模型部署进行推理和使用模型。推理（Inference）只需要执行训练过程中的前向传播过程。



AI 的学习方法：推理过程 Inference

- 由输入到输出完成整个模型中各个层的矩阵计算（如卷积、池化层等），产生输出。输入人脸图片，输出结果为4个向量 [0.8, 0.04, 0.06, 0.03, 0.07]，向量中的各个维度编码了图像的类别可能概率，其中第一个位置的编码向量类别概率最大，后续应再转换为可读信息。

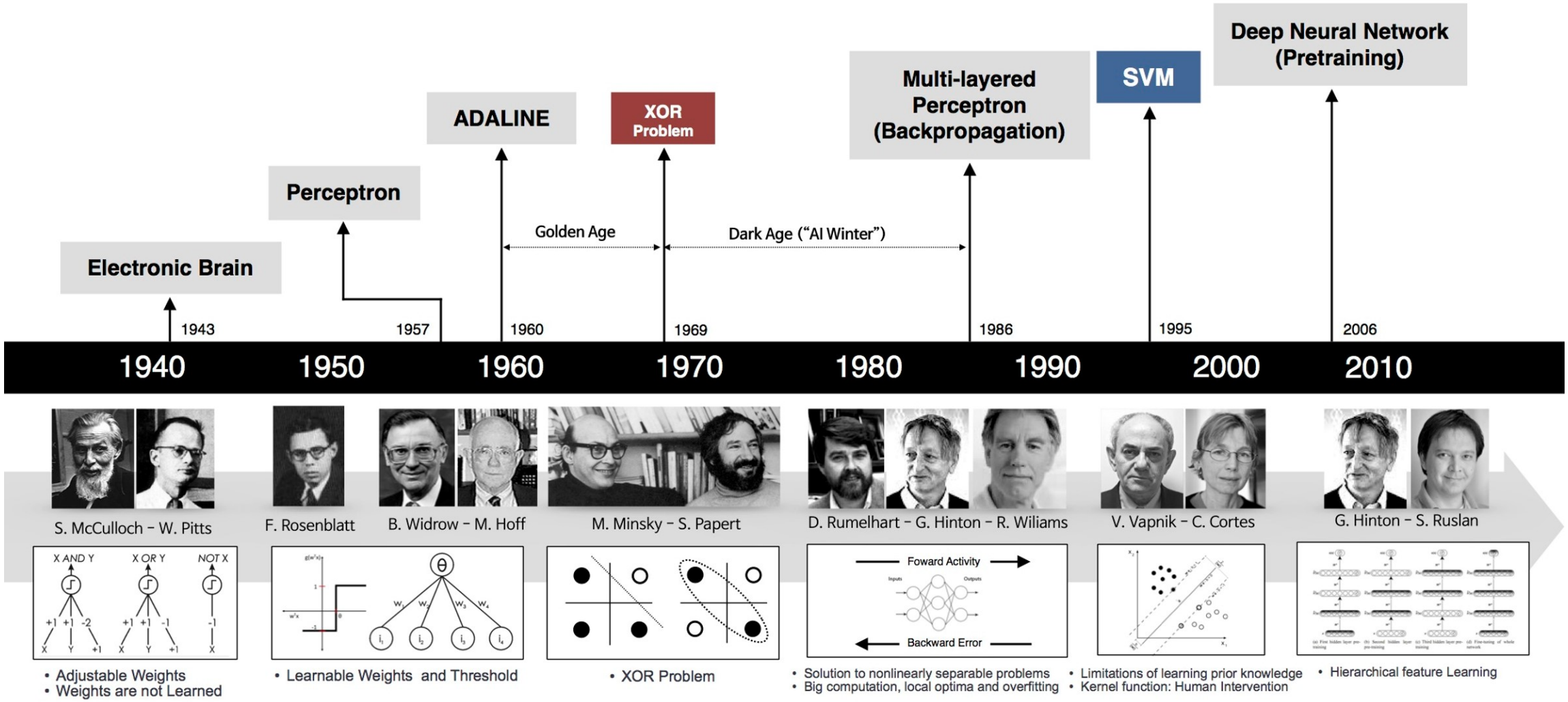


与 AI 系统关系？

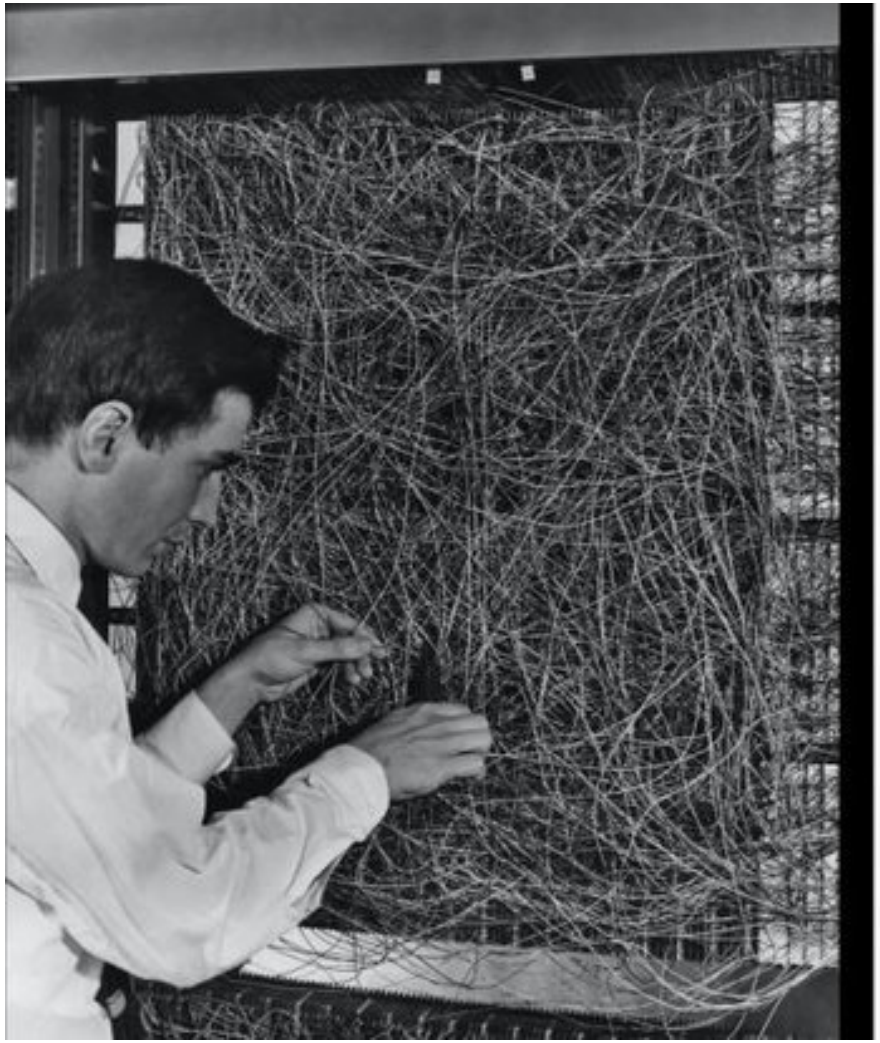
1. AI 系统：围绕训练和推理负载的全生命周期的开发与执行各个环节，提供给算法工程师软件上便捷的模型设计和开发体验，极致的硬件执行性能。
2. 保证安全性、应对海量规模数据、更大模型结构、多租户执行环境、利用 AI 加速器硬件特性、挖掘硬件和集群极致算力。

3. AI 基本理论奠定

AI 的发展历程


Deep Neural Network (Pretraining)
SVM
Multi-layered Perceptron (Backpropagation)
ADALINE
XOR Problem
Perceptron
Electronic Brain
Golden Age
Dark Age ("AI Winter")
1943
1957
1960
1969
1986
1995
2006
1940
1950
1960
1970
1980
1990
2000
2010
S. McCulloch - W. Pitts
F. Rosenblatt
B. Widrow - M. Hoff
M. Minsky - S. Papert
D. Rumelhart - G. Hinton - R. Wiliams
V. Vapnik - C. Cortes
G. Hinton - S. Ruslan
X AND Y
X OR Y
NOT X
Foward Activity
Backward Error
• Adjustable Weights
• Weights are not Learned
• Learnable Weights and Threshold
• XOR Problem
• Solution to nonlinearly separable problems
• Big computation, local optima and overfitting
• Limitations of learning prior knowledge
• Kernel function: Human Intervention
• Hierarchical feature Learning

Frank Rosenblat Perceptron



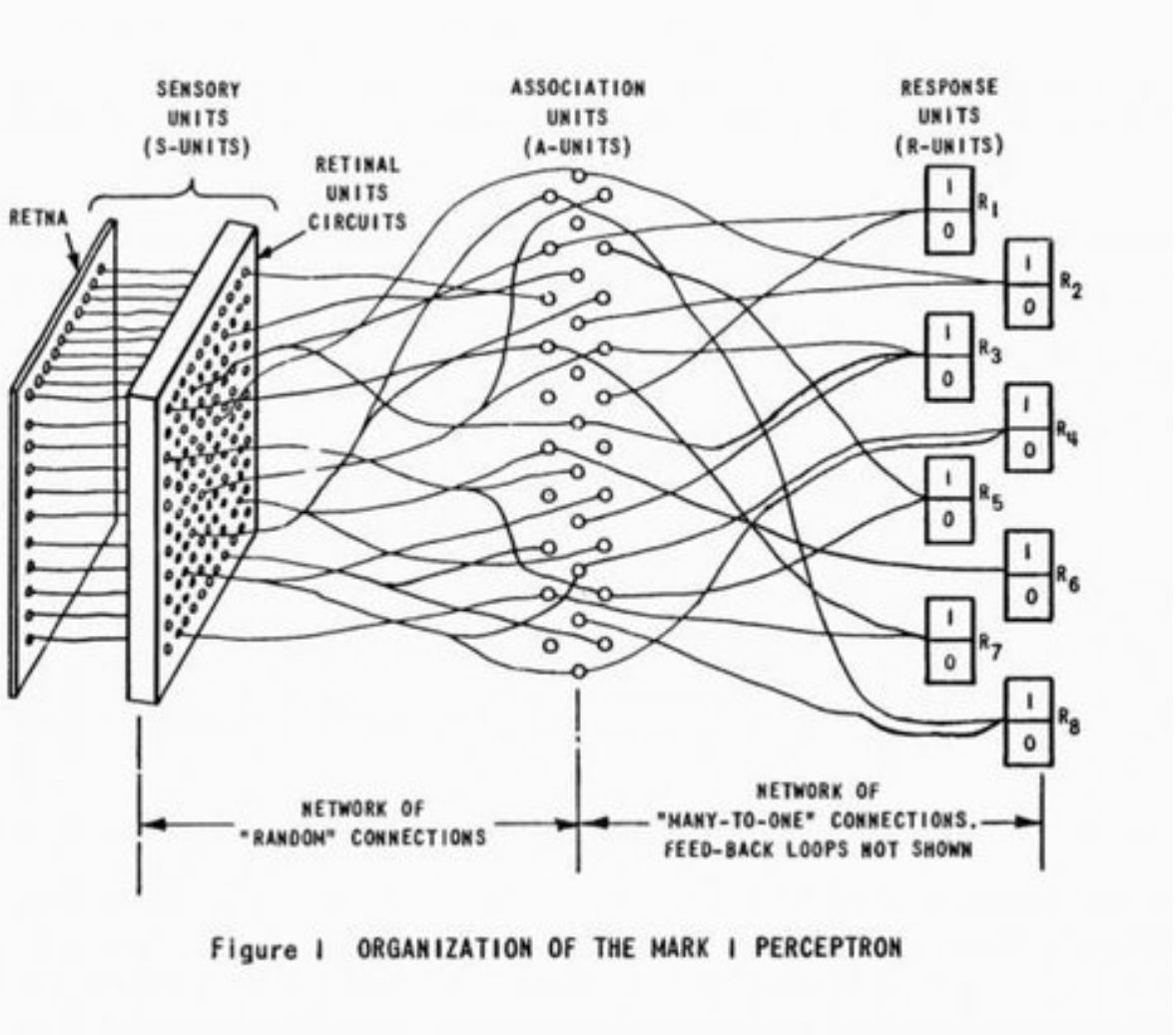


Figure 1 ORGANIZATION OF THE MARK I PERCEPTRON

Geoff Hinton

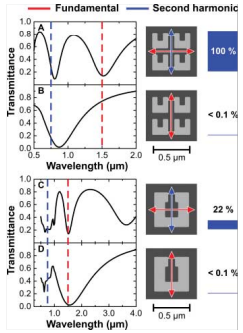


Fig. 3. Theory, presented as the experiment (see Fig. 1). The SHG source is the magnetic component of the Lorentz force on metal electrons in the SRRs.

The setup for measuring the SHG is described in the supporting online material (*22*). We expect that the SHG strongly depends on the resonance that is excited. Obviously, the incident polarization and the detuning of the laser wavelength from the resonance are of particular interest. One possibility for controlling the detuning is to change the laser wavelength for a given sample, which is difficult because of the extremely broad tuning range required. Thus, we follow an alternative route, lithographic tuning (in which the incident laser wavelength of 1.5 μm, as well as the detection system, remains fixed), and tune the resonance positions by changing the SRR size. In this manner, we can also guarantee that the optical properties of the SRR constituent materials are identical for all configurations. The blue bars in Fig. 1 summarize the measured SHG signals. For excitation of the *LC* resonance in Fig. 1A (horizontal incident polarization), we find an SHG signal that is 500 times above the noise level. As expected for SHG, this signal closely scales with the square of the incident power (Fig. 2A). The polarization of the SHG emission is nearly vertical (Fig. 2B). The small angle with respect to the vertical is due to deviations from perfect mirror symmetry of the SRRs (see electron micrographs in Fig. 1). Small detuning of the *LC* resonance toward smaller wavelength (i.e., to 1.3-μm wavelength) reduces the SHG signal strength from 100% to 20%. For excitation of the Mie resonance with vertical incident polarization in Fig. 1D, we find a small signal just above the noise level. For excitation of the Mie resonance with horizontal incident polarization in Fig. 1C, a small but significant SHG emission is found, which is again polarized nearly vertically. For completeness, Fig. 1B shows the off-resonant case for the smaller SRRs for vertical incident polarization.

Although these results are compatible with the known selection rules of surface SHG from usual nonlinear optics (*23*), these selection rules do not explain the mechanism of SHG. Following our above argumentation on the magnetic component of the Lorentz force, we numerically calculate first the linear electric and magnetic field distributions (*22*); from these fields, we compute the electron velocities and the Lorentz-force field (fig. S1). In the spirit of a metamaterial, the transverse component of the Lorentz-force field can be spatially averaged over the volume of the unit cell of size a by a by t. This procedure delivers the driving force for the transverse SHG polarization. As usual, the SHG intensity is proportional to the square modulus of the nonlinear electron displacement. Thus, the SHG strength is expected to be proportional to the square modulus of the driving force, and the SHG polarization is directed along the driving-force vector. Corresponding results are summarized in Fig. 3 in the same arrangement as Fig. 1 to allow for a direct comparison between experiment and theory. The agreement is generally good, both for linear optics and for SHG. In particular, we find a much larger SHG signal for excitation of those two resonances (Fig. 3, A and C), which are related to a finite magnetic-dipole moment (perpendicular to the SRR plane) as compared with the purely electric Mie resonance (Figs. 1D and 3D), despite the fact that its oscillator strength in the linear spectrum is comparable. The SHG polarization in the theory is strictly vertical for all resonances. Quantitative deviations between the SHG signal strengths of experiment and theory, respectively, are probably due to the simplified SRR shape assumed in our calculations and/or due to the simplicity of our modeling. A systematic microscopic theory of the nonlinear optical properties of metallic metamaterials would be highly desirable but is currently not available.

References and Notes

1. J. B. Pendry, A. J. Holden, D. J. Robbins, W. J. Stewart, *IEEE Trans. Microw. Theory Tech.* **47**, 2075 (1999).
2. J. B. Pendry, *Phys. Rev. Lett.* **85**, 3966 (2000).
3. R. A. Shelby, D. R. Smith, S. Schultz, *Science* **292**, 77 (2001).
4. T. J. Yen *et al.*, *Science* **303**, 1494 (2004).
5. S. Linden *et al.*, *Science* **306**, 1351 (2004).
6. C. Enkrich *et al.*, *Phys. Rev. Lett.* **95**, 203901 (2005).
7. A. N. Grigorenko *et al.*, *Nature* **438**, 335 (2005).
8. G. Dolling, M. Wegener, S. Linden, C. Hormann, *Opt. Express* **14**, 1842 (2006).
9. G. Dolling, C. Enkrich, M. Wegener, C. M. Soukoulis, S. Linden, *Science* **312**, 892 (2006).
10. J. B. Pendry, D. Schurig, D. R. Smith, *Science* **312**, 1780; published online 25 May 2006.
11. U. Leonhardt, *Science* **312**, 1777 (2006); published online 25 May 2006.
12. M. W. Klein, C. Enkrich, M. Wegener, C. M. Soukoulis, S. Linden, *Opt. Lett.* **31**, 1259 (2006).
13. W. J. Padilla, A. J. Taylor, C. Highstrete, M. Lee, R. D. Averitt, *Phys. Rev. Lett.* **96**, 107401 (2006).
14. D. R. Smith, S. Schultz, P. Markos, C. M. Soukoulis, *Phys. Rev. B* **65**, 195104 (2002).
15. S. O'Brien, D. McPeake, S. A. Ramakrishna, J. B. Pendry, *Phys. Rev. B* **69**, 241101 (2004).
16. J. Zhou *et al.*, *Phys. Rev. Lett.* **95**, 223902 (2005).
17. A. K. Popov, V. M. Shalaev, available at http://arxiv.org/abs/physics/0601055 (2006).
18. V. G. Veselago, *Sov. Phys. Usp.* **10**, 509 (1968).
19. M. Wegener, *Extreme Nonlinear Optics* (Springer, Berlin, 2004).
20. H. M. Barlow, *Nature* **173**, 41 (1954).
21. S.-Y. Chen, M. Maksimchuk, D. Umstadter, *Nature* **396**, 653 (1998).
22. Materials and Methods are available as supporting material on *Science* Online.
23. P. Guyot-Sionnest, W. Chen, Y. R. Shen, *Phys. Rev. B* **33**, 8254 (1986).

24. We thank the groups of S. W. Koch, J. V. Moloney, and C. M. Soukoulis for discussions. The research of M.W. is supported by the Leibniz award 2000 of the Deutsche Forschungsgemeinschaft (DFG), that of S.L. through a Helmholtz-Hochschul-Nachwuchsgruppe (VH-NG-232).

Supporting Online Material

www.sciencemag.org/cgi/content/full/313/5786/502/DC1
Materials and Methods
Figs. S1 and S2
References

26 April 2006; accepted 22 June 2006
10.1126/science.1129198

Reducing the Dimensionality of Data with Neural Networks

G. E. Hinton* and R. R. Salakhutdinov

High-dimensional data can be converted to low-dimensional codes by training a multilayer neural network with a small central layer to reconstruct high-dimensional input vectors. Gradient descent can be used for fine-tuning the weights in such "autoencoder" networks, but this works well only if the initial weights are close to a good solution. We describe an effective way of initializing the weights that allows deep autoencoder networks to learn low-dimensional codes that work much better than principal components analysis as a tool to reduce the dimensionality of data.

Dimensionality reduction facilitates the classification, visualization, communication, and storage of high-dimensional data. A simple and widely used method is principal components analysis (PCA), which finds the directions of greatest variance in the data set and represents each data point by its coordinates along each of these directions. We describe a nonlinear generalization of PCA that uses an adaptive, multilayer "encoder" network

Alex Krizhevsky

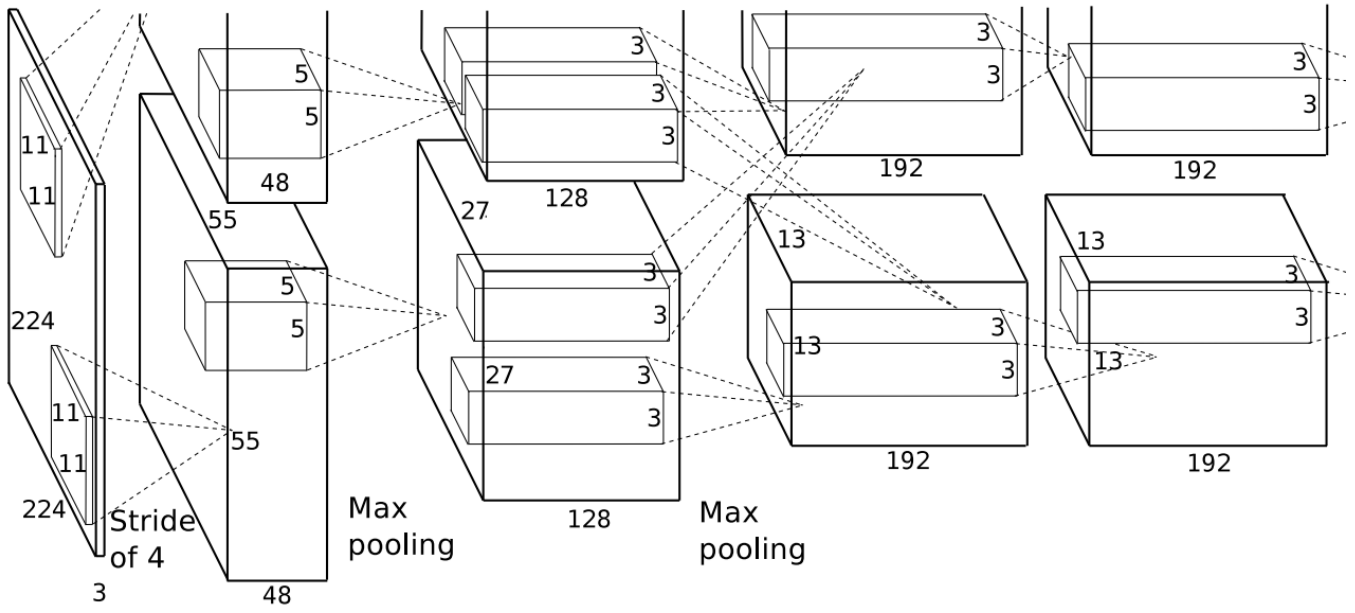


Figure 2: An illustration of the architecture of our CNN, explicitly showing between the two GPUs. One GPU runs the layer-parts at the top of the figure at the bottom. The GPUs communicate only at certain layers. The network's the number of neurons in the network's remaining layers is given by 253,440-4096–4096–1000.

ImageNet

Thank you

把AI系统带入每个开发者、每个家庭、每个组织，构建万物互联的智能世界

Bring AI System to every person, home and organization for a fully connected, intelligent world.

ZOMI

Course chenzomi12.github.io

GitHub github.com/chenzomi12/DeepLearningSystem