System for Artificial Intelligence

Introduction to System for AI

Siyuan Feng
Shanghai Innovation Institute

OUTLINE

01 Intro to Deep Learning

02 Overview of ML System

01

Intro to Deep Learning

Machine Learning

Object recognition → Cat

Text generation (next-token prediction)

The capital of China is → Beijing

Image generation

An AI Infra engineer →

Machine Learning

The capital of China is

An AI Infra engineer

$$x \xrightarrow{F(x)} y$$

Cat

Beijing

Machine learning is a general algorithm to build the function $f(x)$

An Overview of Deep Learning Models

上海创智学院
Shanghai Innovation Institute

- Convolutional Neural Networks
- Recurrent Neural Networks
- Transformers
- Mixture-of-Experts

An Overview of Deep Learning Models

上海创智学院
Shanghai Innovation Institute

- **Convolutional Neural Networks**
- Recurrent Neural Networks
- Transformers
- Mixture-of-Experts

CNNs are Widely Used in Vision Tasks

上海创智学院
Shanghai Innovation Institute

Object Detection

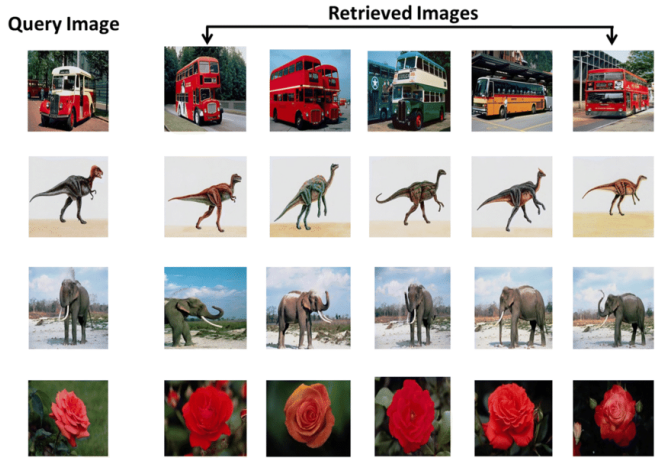


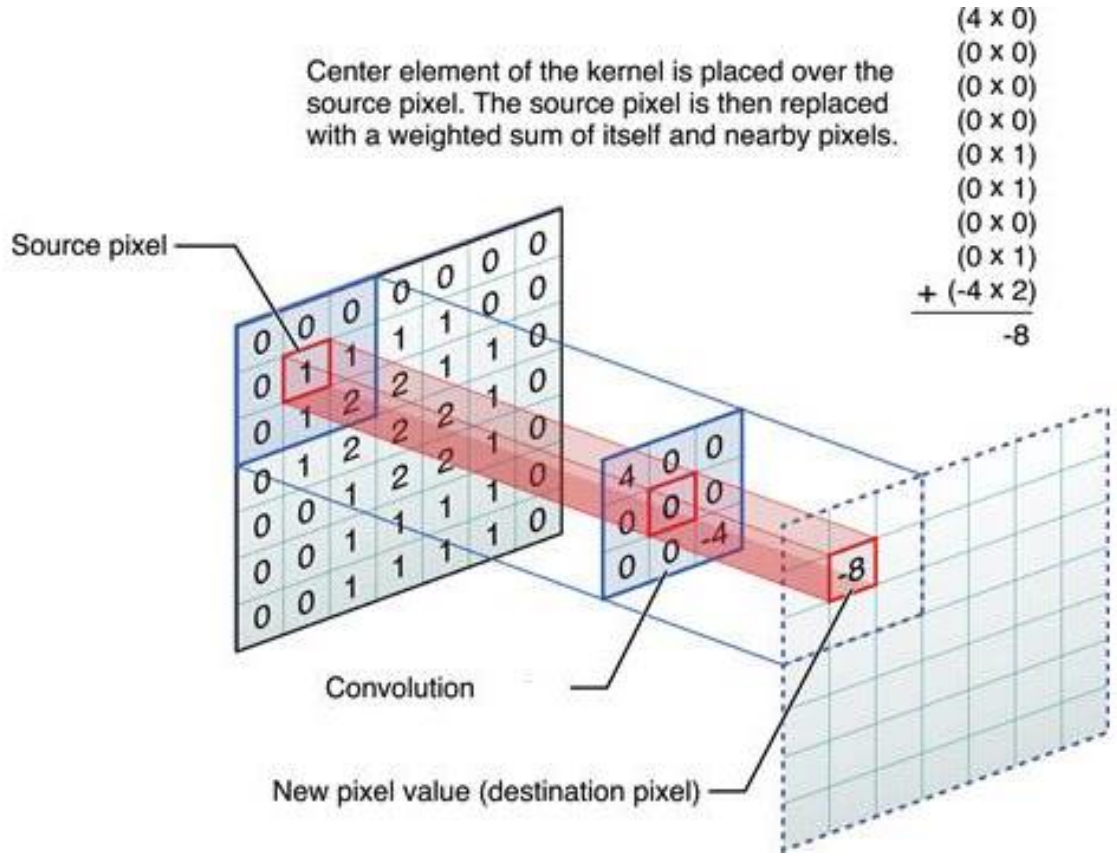
Image Retrieval

Self-Driving

Convolution

上海创智学院
Shanghai Innovation Institute

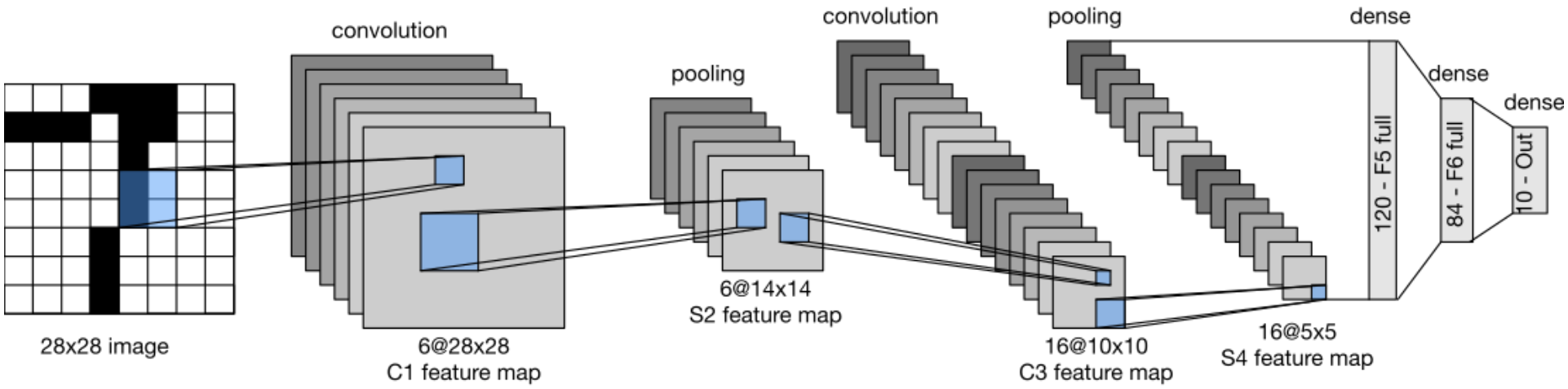
- Convolve the filter with the image: slide over the image spatially and compute dot products



Convolutional neural network

上海创智学院
Shanghai Innovation Institute

- A sequence of convolutional layers, interspersed by pooling, normalization, and activation functions



An Overview of Deep Learning Models

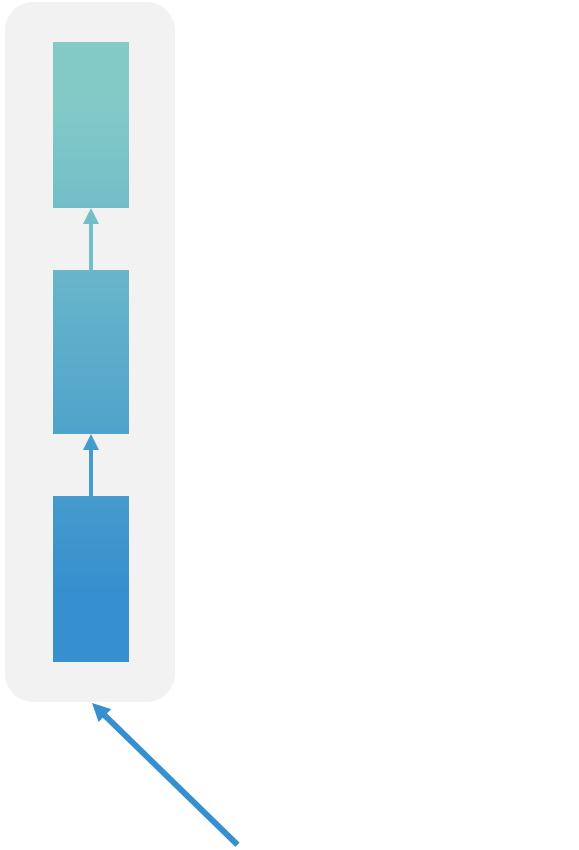
上海创智学院
Shanghai Innovation Institute

- Convolutional Neural Networks
- **Recurrent Neural Networks**
- Transformers
- Mixture-of-Experts

Recurrent Neural Networks: Process Sequences

上海创智学院
Shanghai Innovation Institute

one to one



Vanilla Neural Networks

Recurrent Neural Networks: Process Sequences

one to one

one to many

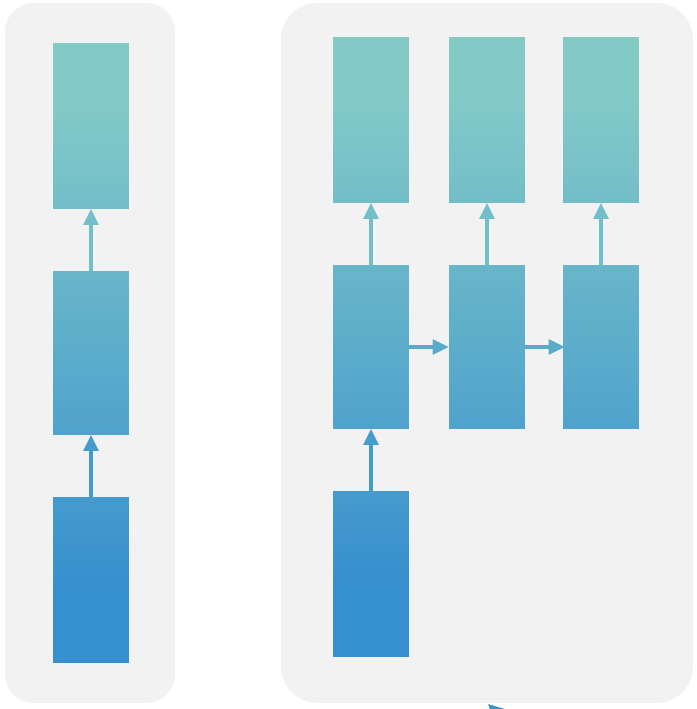
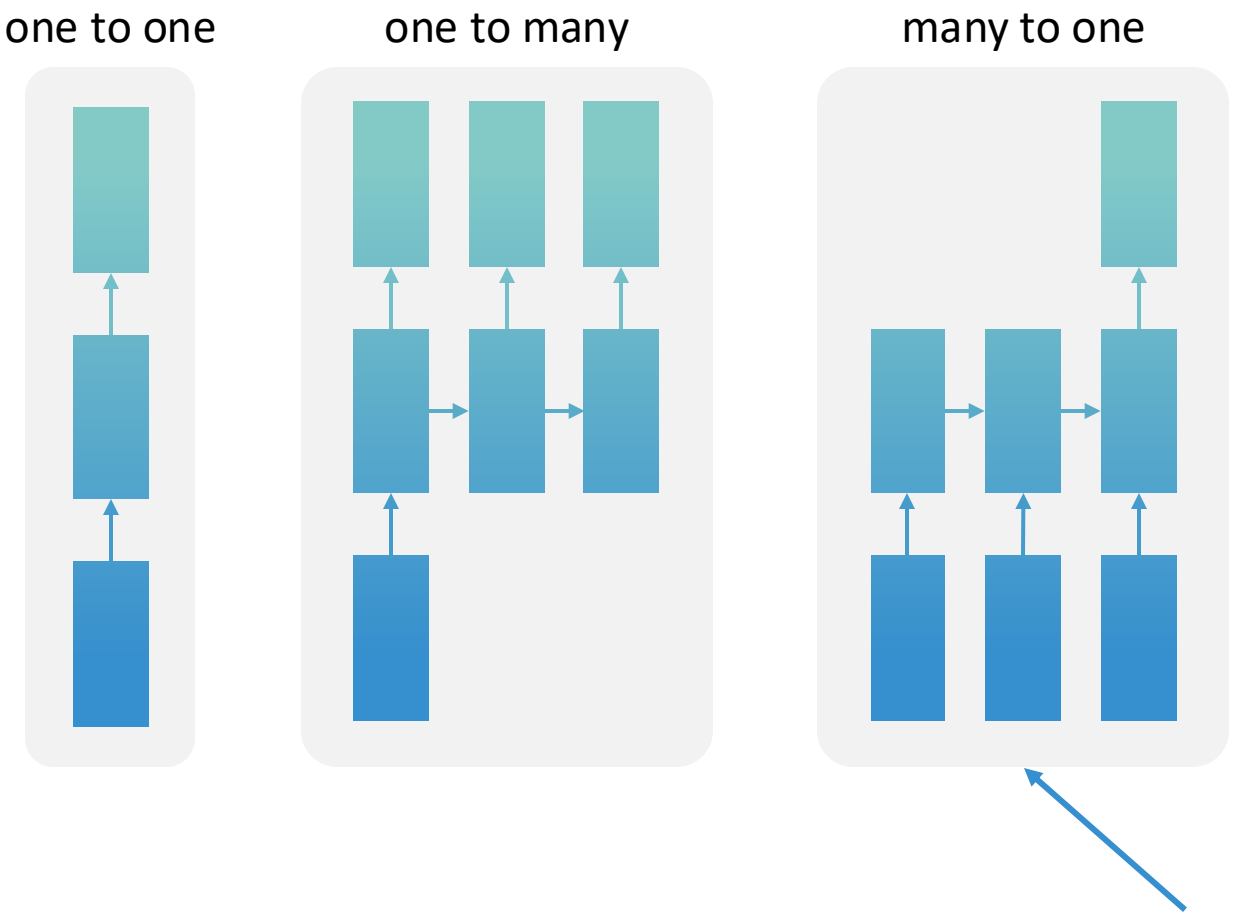


image captioning
Image -> sequence of words

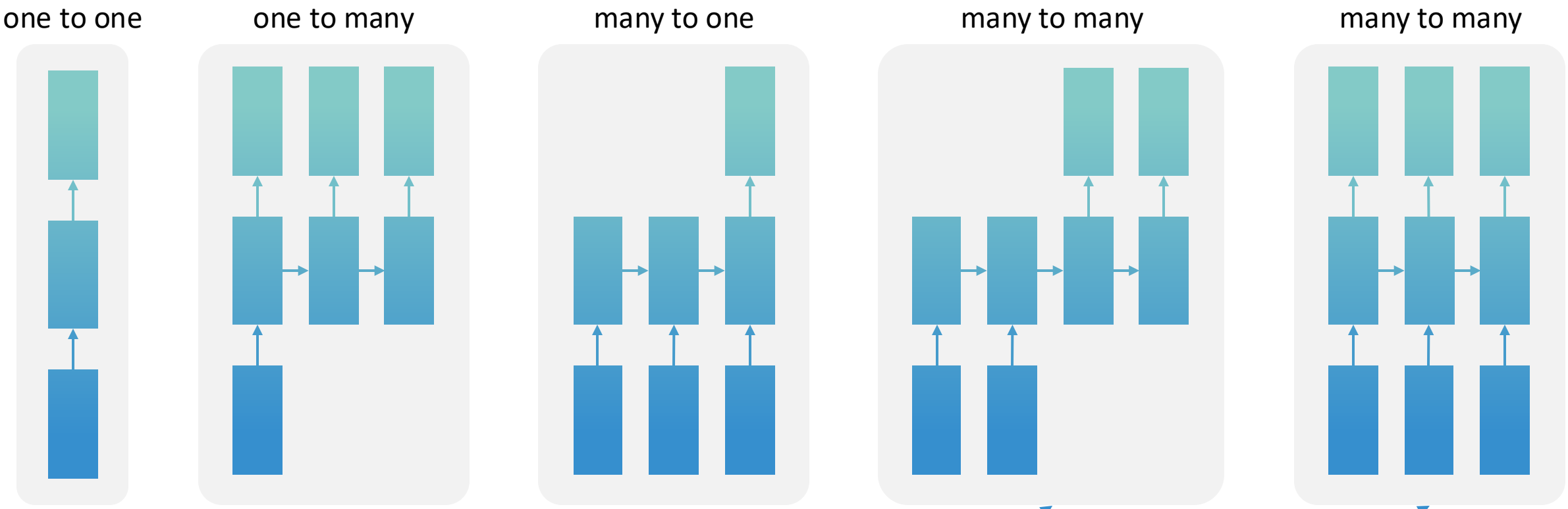
Recurrent Neural Networks: Process Sequences



action prediction
sequence of video frames -> action

Recurrent Neural Networks: Process Sequences

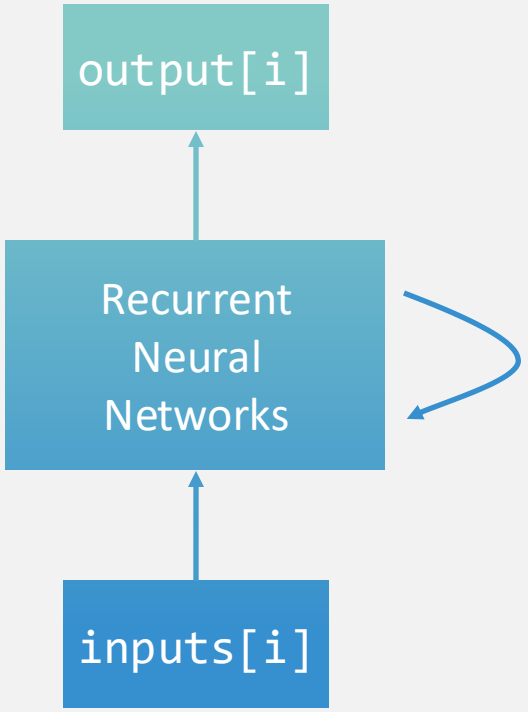
上海创智学院
Shanghai Innovation Institute



Video captioning: sequence of video frames -> sequence of words
Machine translation

Video classification on frames

Recurrent Neural Networks

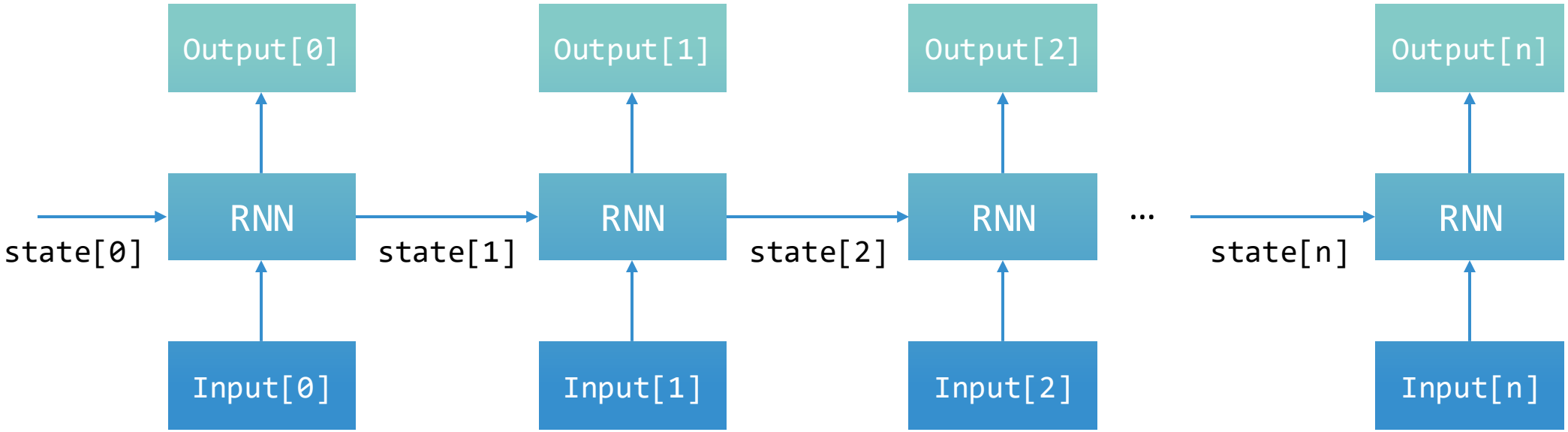


Arbitrary number of outputs

Key idea: RNNs have an internal state that is updated as a sequence is processed

Arbitrary number of inputs

Representing RNNs in Computation Graphs



When do we need RNNs?

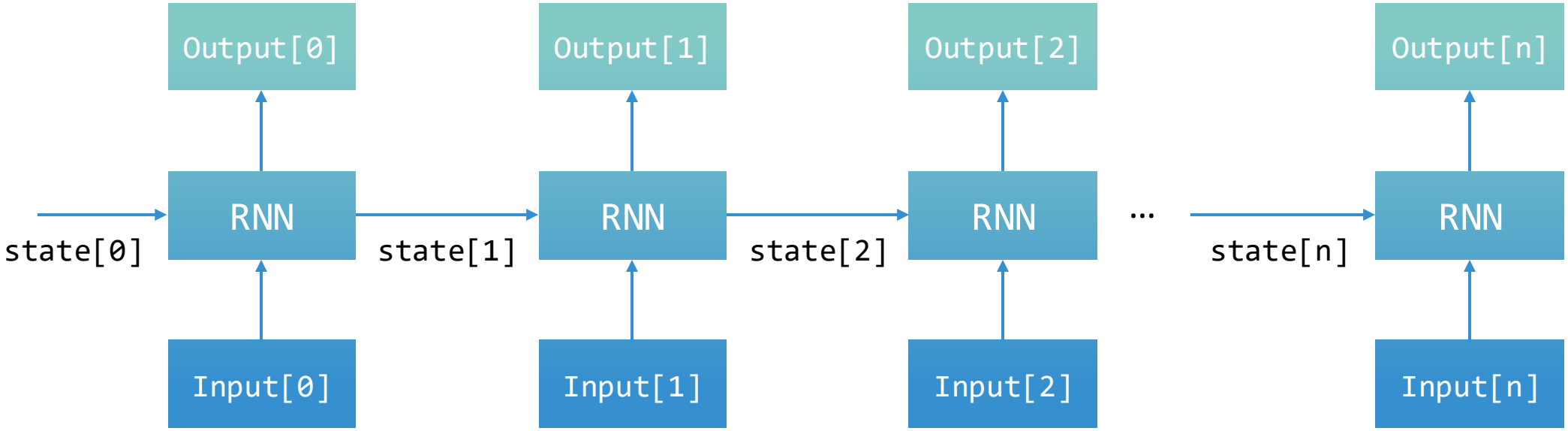
上海创智学院
Shanghai Innovation Institute

- RNNs are designed to process sequences (texts, videos)
- RNNs are extremely useful when you want your model to have internal states when a sequence is processed
- Commonly used in reinforcement learning (RL)

Inefficiency in RNNs?

上海创智学院
Shanghai Innovation Institute

- Problem: **lack of parallelizability.** Both forward and backward passes have O(sequence length) unparallelizable operators
- A state cannot be computed before all previous states have been computed
- Inhibits training on very long sequences



An Overview of Deep Learning Models

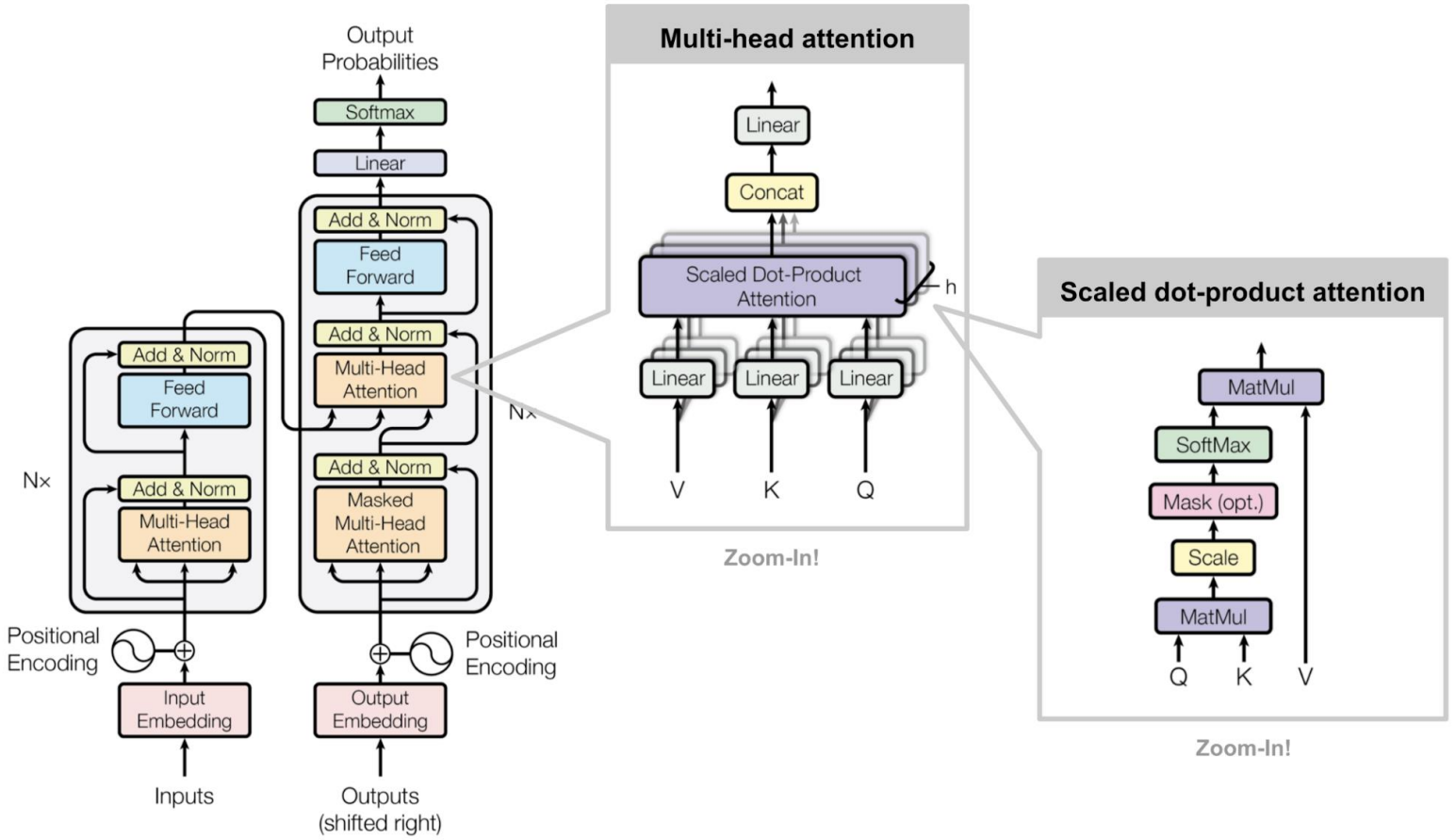
上海创智学院
Shanghai Innovation Institute

- Convolutional Neural Networks
- Recurrent Neural Networks
- **Transformers**
- Mixture-of-Experts

Attention: Enable Parallelism within a Sequence

上海创智学院
Shanghai Innovation Institute

- Idea: treat each position's representation as a **query** to access and incorporate information from a set of **values**



Discussion: Attention vs RNN

上海创智学院
Shanghai Innovation Institute

An Overview of Deep Learning Models

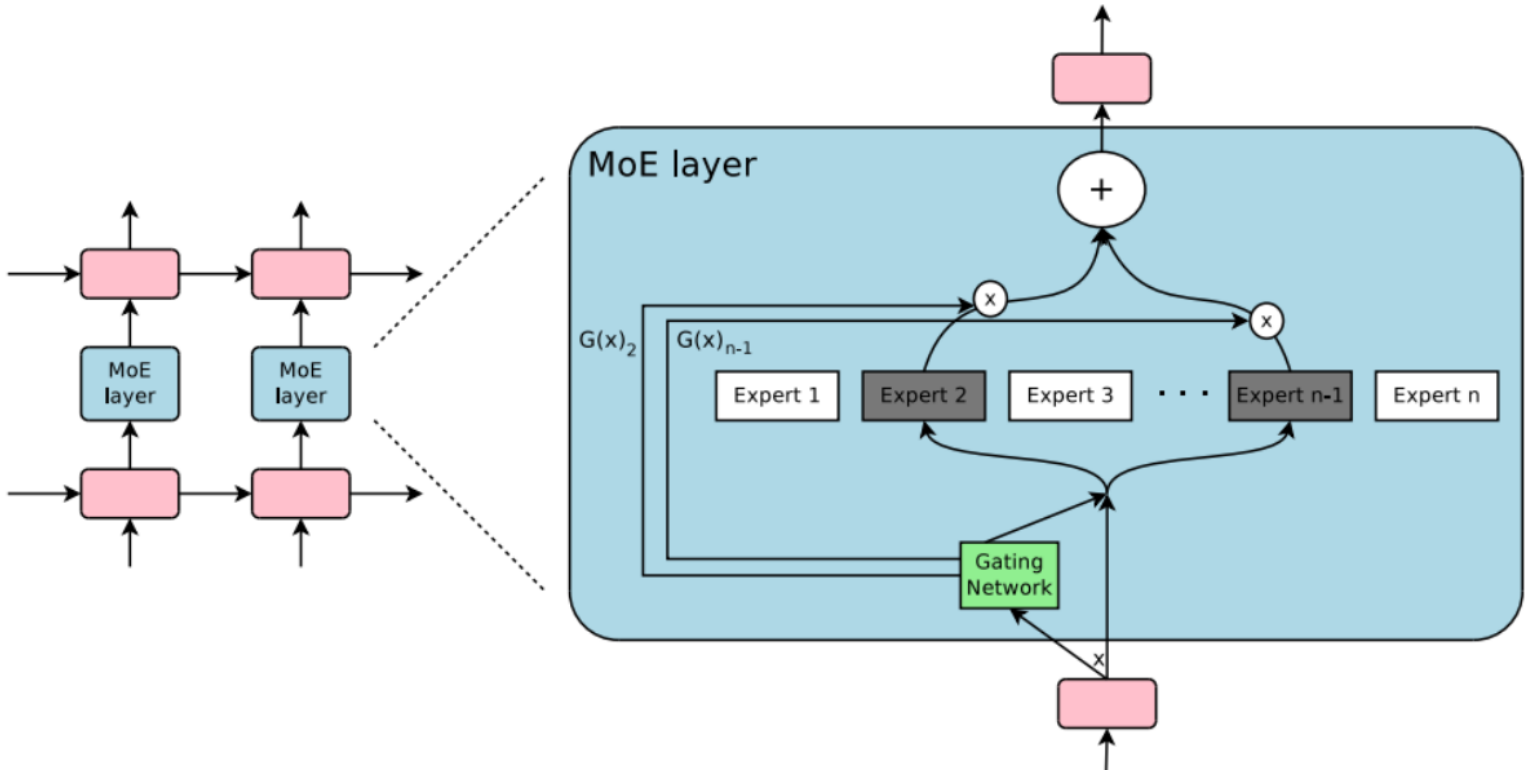
上海创智学院
Shanghai Innovation Institute

- Convolutional Neural Networks
- Recurrent Neural Networks
- Transformers
- **Mixture-of-Experts**

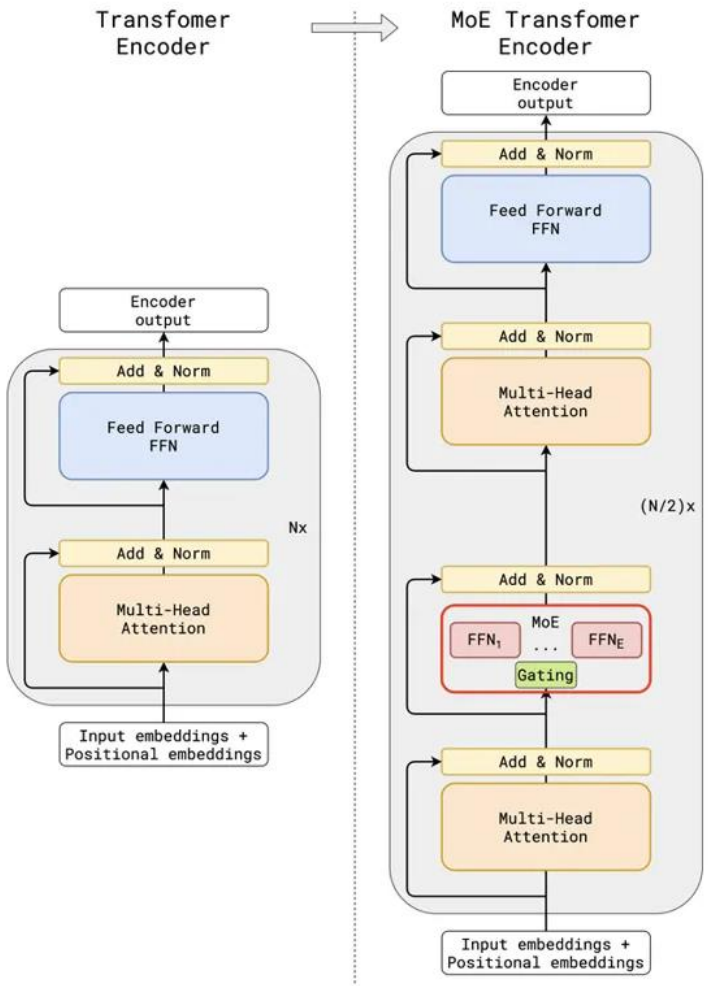
Mixture-of-Experts

上海创智学院
Shanghai Innovation Institute

- Key idea: make each expert focus on predicting the right answer for a subset of cases
- **Actual: a kind of model-level sparsity.**



MoE Large Language Models



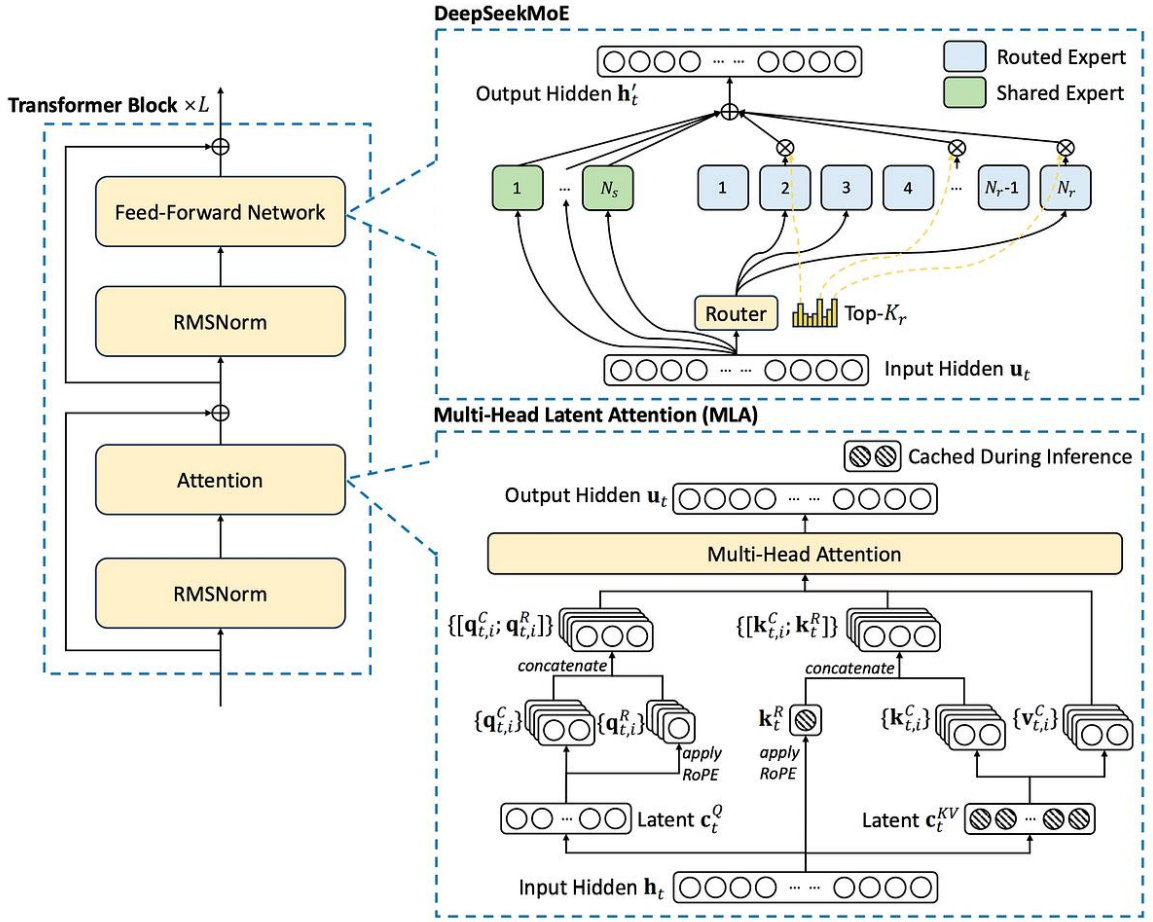


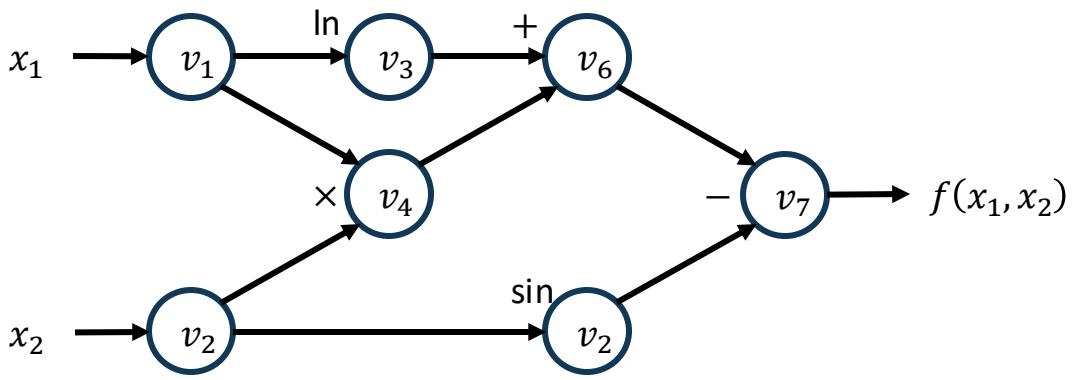
Figure 2 | Illustration of the basic architecture of DeepSeek-V3. Following DeepSeek-V2, we adopt MLA and DeepSeekMoE for efficient inference and economical training.

Computational Graph Abstraction

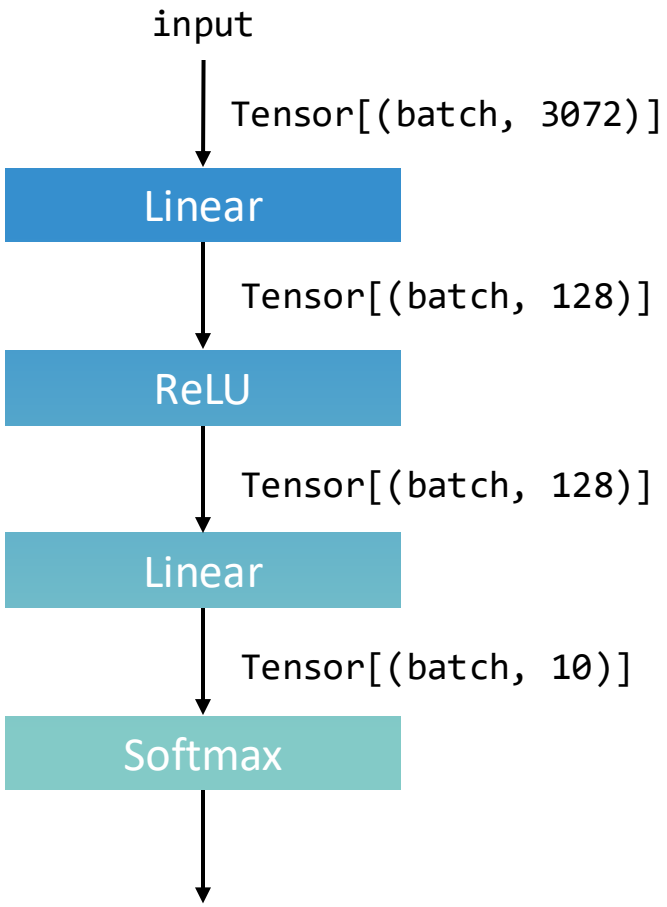
上海创智学院
Shanghai Innovation Institute

- Nodes represents the computation (operation)
- Edge represents the data dependency between operations

$$f(x_1, x_2) = \ln(x_1) + x_1 x_2 - \sin(x_2)$$



Computational Graph for A Classification Model



Tensor is the central data format in deep learning models

02

Overview of ML System

An Overview of Machine Learning Systems

上海创智学院
Shanghai Innovation Institute

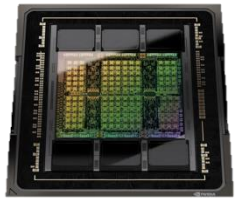
ML Models

Automatic Differentiation

Graph Optimization

Parallelism / Distributed

Hardware Acceleration



NVIDIA GPU

HUAWEI NPU

Mobile devices

Layer 1: Automatic Differentiation

上海创智学院
Shanghai Innovation Institute

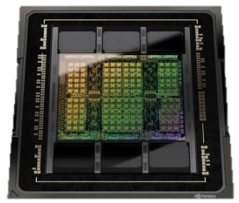
ML Models

Automatic Differentiation

Graph Optimization

Parallelism / Distributed

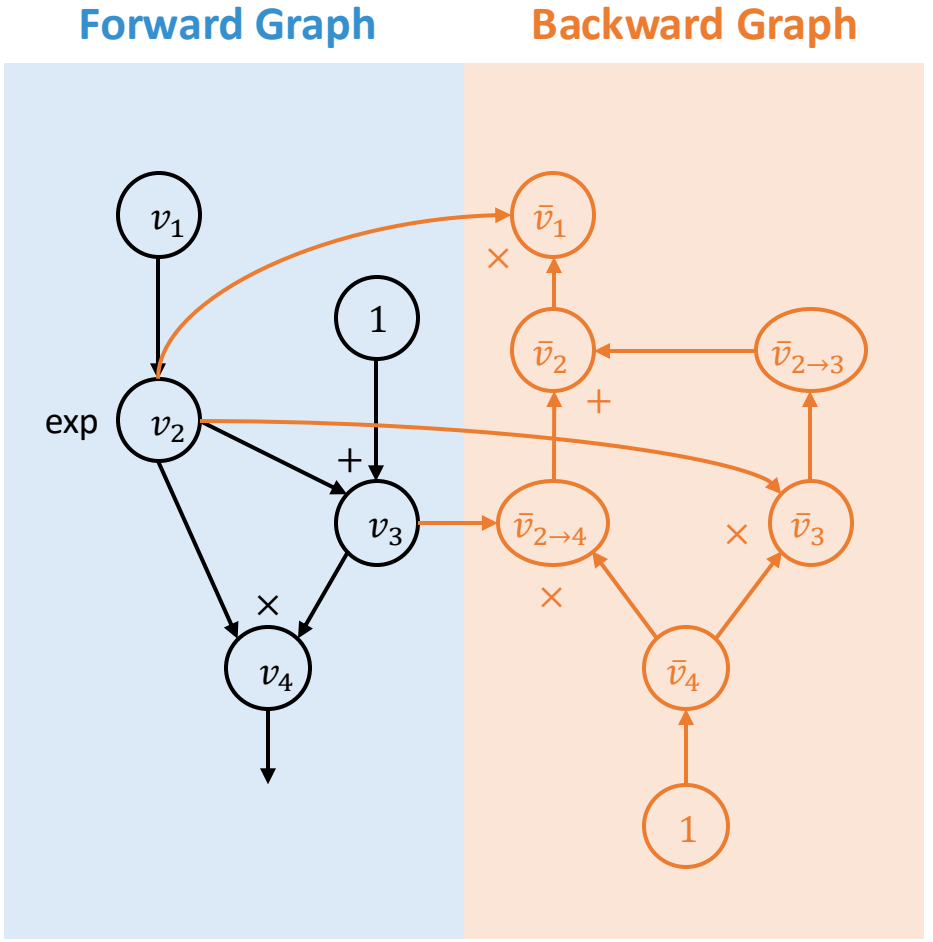
Hardware Acceleration



NVIDIA GPU

HUAWEI NPU

Mobile devices

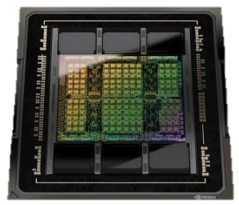


Layer 2: Graph Optimization

上海创智学院
Shanghai Innovation Institute

ML Models

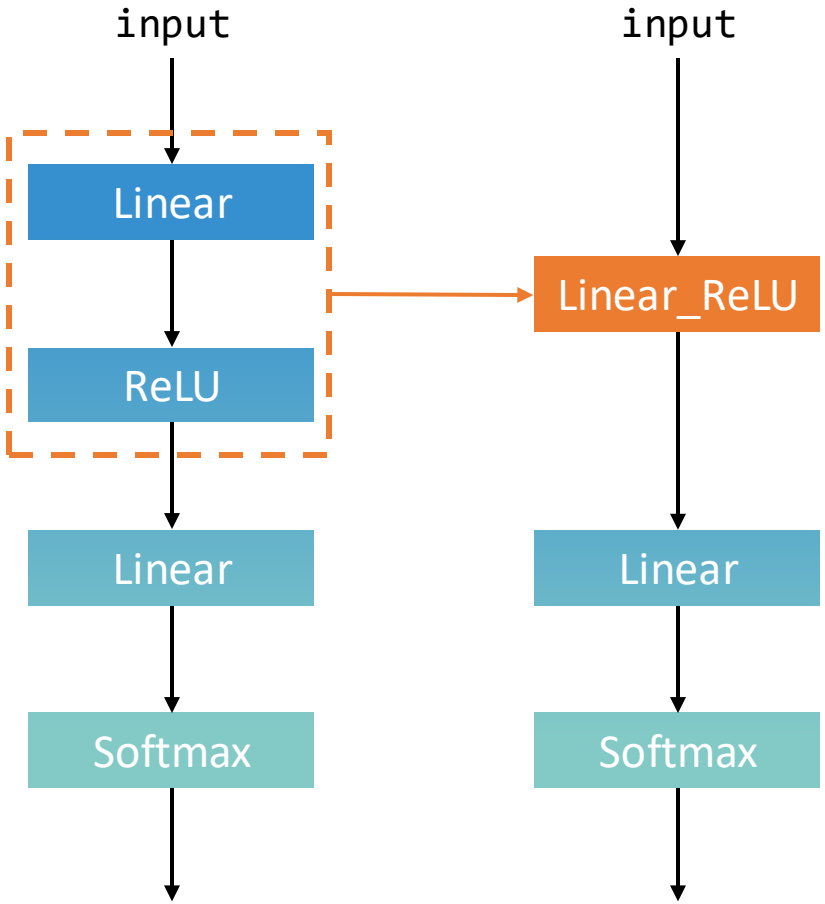
- Automatic Differentiation
- Graph Optimization
- Parallelism / Distributed
- Hardware Acceleration



NVIDIA GPU

HUAWEI NPU

Mobile devices



Discuss: Why does fuse kernels work?

Layer 3: Parallelization

上海创智学院
Shanghai Innovation Institute

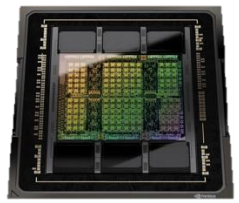
ML Models

Automatic Differentiation

Graph Optimization

Parallelism / Distributed

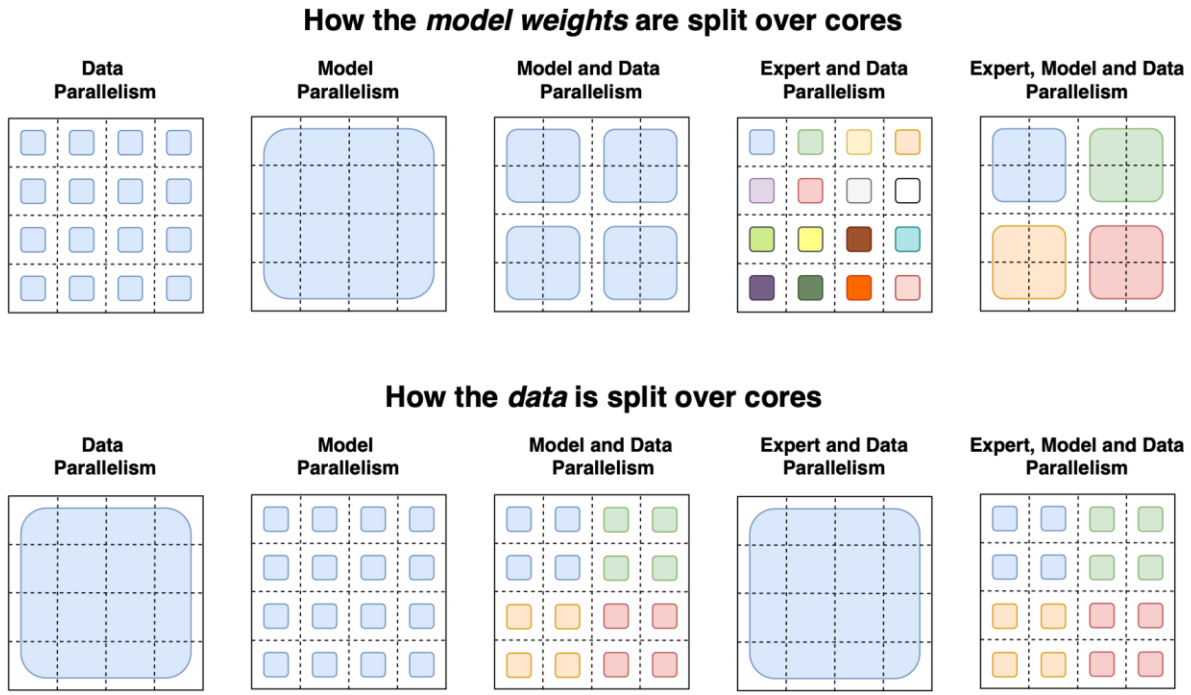
Hardware Acceleration



NVIDIA GPU

HUAWEI NPU

Mobile devices



Layer 4: Hardware Acceleration

上海创智学院 Shanghai Innovation Institute

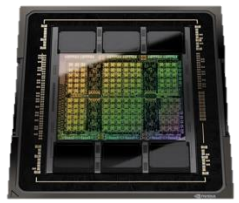
ML Models

Automatic Differentiation

Graph Optimization

Parallelism / Distributed

Hardware Acceleration

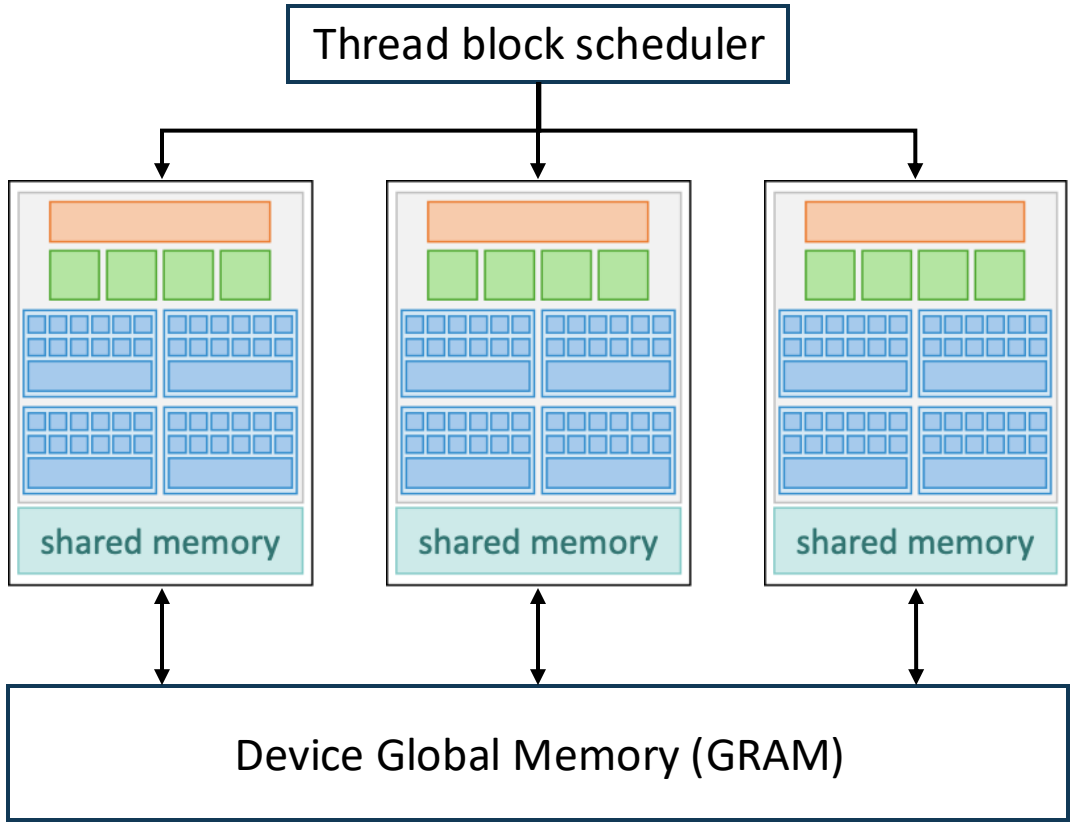


NVIDIA GPU

HUAWEI NPU

Mobile devices

Leverage the computation capability of the hardware backend



Cross-Layer Optimizations

上海创智学院
Shanghai Innovation Institute

ML Models

Automatic Differentiation

Graph Optimization

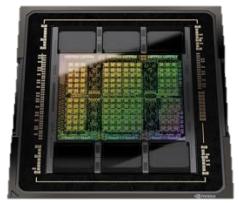
Parallelism / Distributed

Hardware Acceleration

Kernel fusion for training `torch.compile`

Mega-Kernel

Computation-Communication Overlap



NVIDIA GPU

HUAWEI NPU

Mobile devices

Upcoming Lectures

上海创智学院
Shanghai Innovation Institute

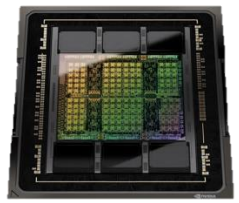
ML Models

Automatic Differentiation

Graph Optimization

Parallelism / Distributed

Hardware Acceleration



NVIDIA GPU

HUAWEI NPU

Mobile devices

- Automatic Differentiation
- General Hardware Acceleration
- CUDA Programming
- Ascend Programming
- ML Compilation
- LLMs general optimizations
- Distributed computing
- LLM pre-training
- LLM serving / inference
- LLM post-training / RL

Acknowledgement

The development of this course, including its structure, content, and accompanying presentation slides, has been significantly influenced and inspired by the excellent work of instructors and institutions who have shared their materials openly. We wish to extend our sincere acknowledgement and gratitude to the following courses, which served as invaluable references and a source of pedagogical inspiration:

- Machine Learning Systems[15-442/15-642], by **Tianqi Chen** and **Zhihao Jia** at **CMU**.
- Advanced Topics in Machine Learning (Systems)[CS6216], by **Yao Lu** at **NUS**

While these materials provided a foundational blueprint and a wealth of insightful examples, all content herein has been adapted, modified, and curated to meet the specific learning objectives of our curriculum. Any errors, omissions, or shortcomings found in these course materials are entirely our own responsibility. We are profoundly grateful for the contributions of the educators listed above, whose dedication to teaching and knowledge-sharing has made the creation of this course possible.

System for Artificial Intelligence

Thanks

Siyuan Feng
Shanghai Innovation Institute