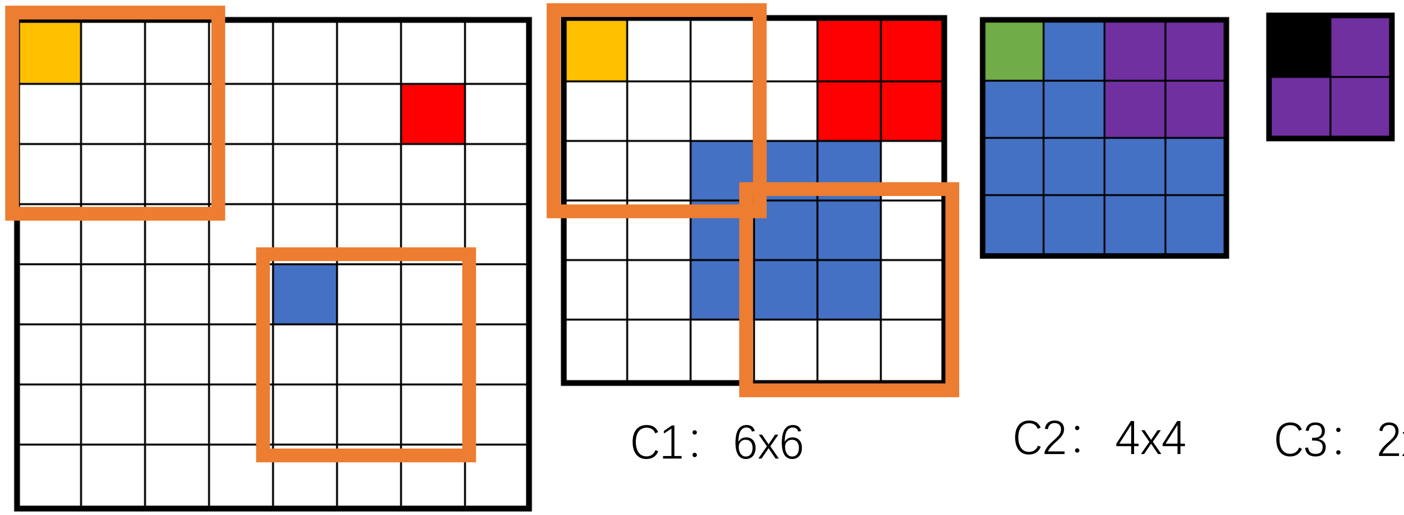
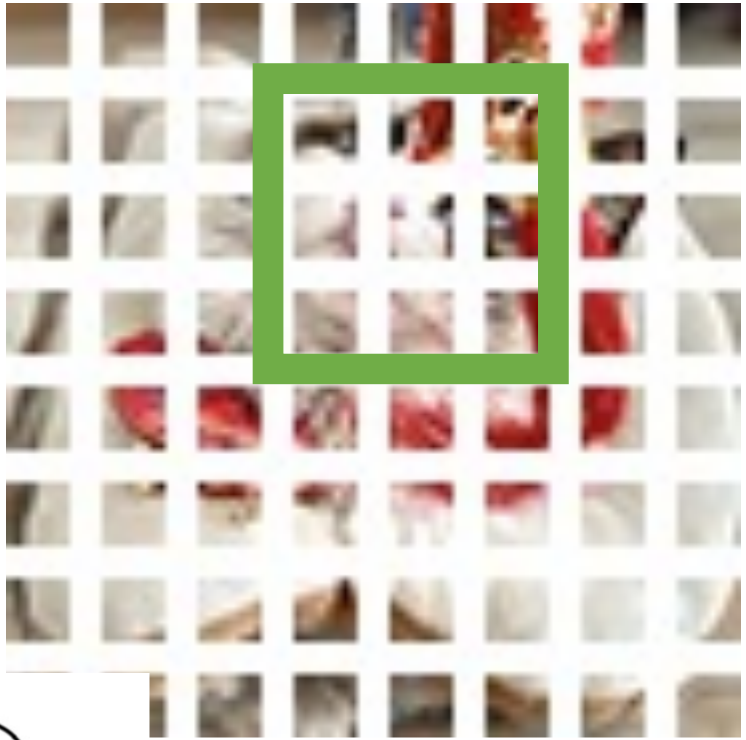
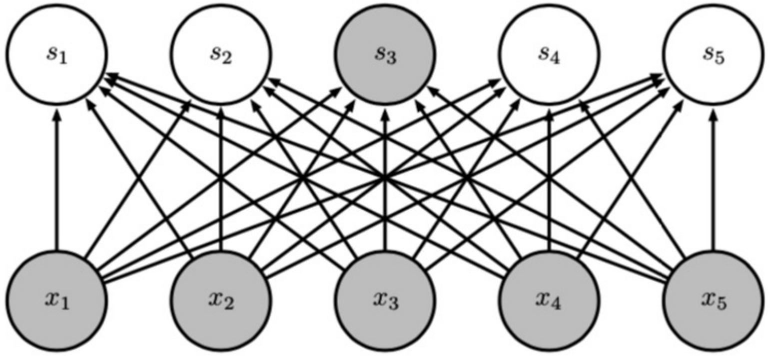
# 0.0 感受野（receptive field）&参数共享

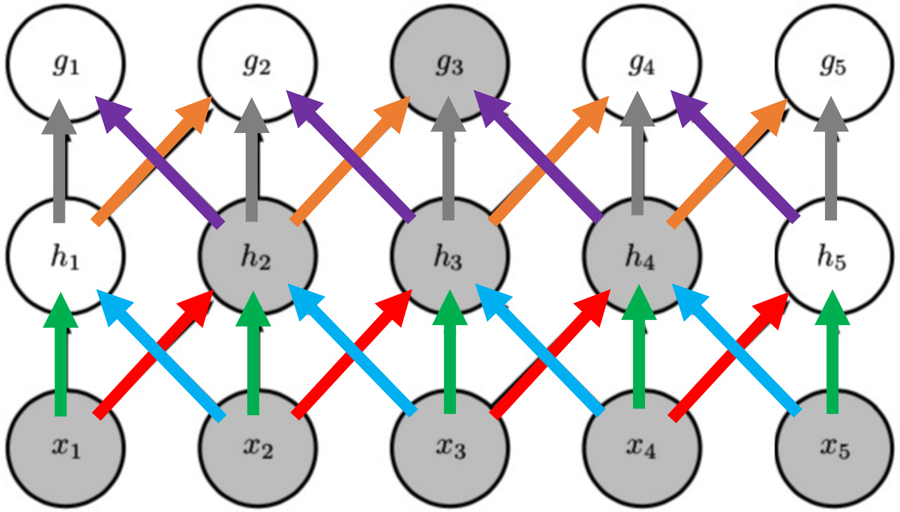


C1：6x6　C2：4x4　C3：2x2

原图：8x8　3x3的卷积核

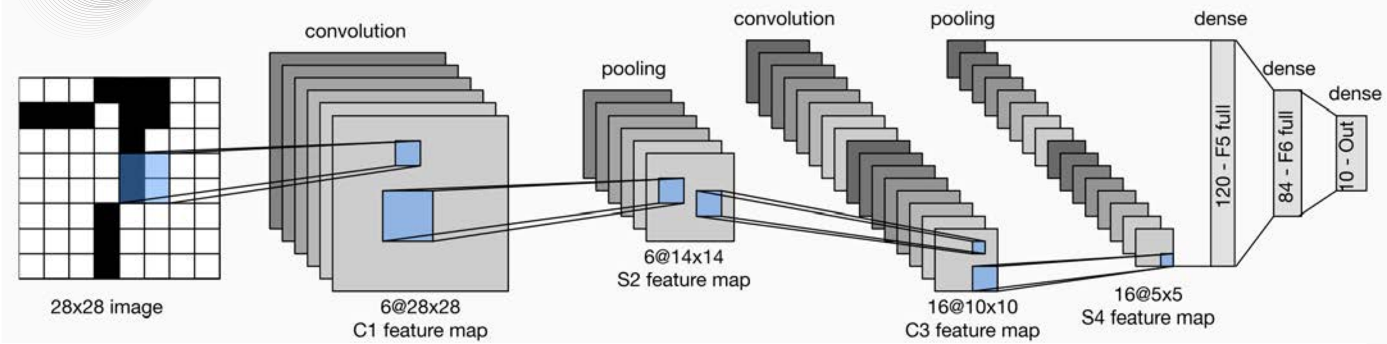




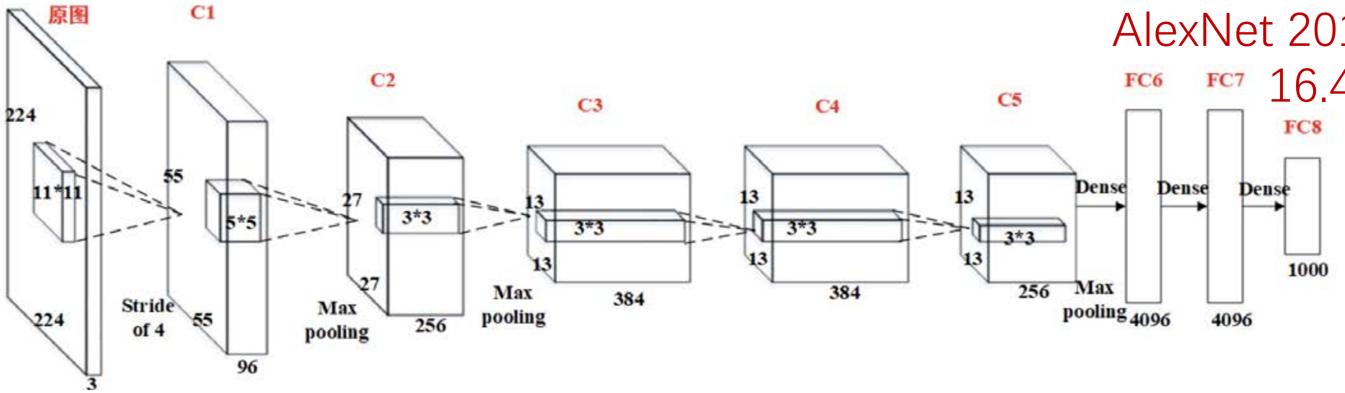



# 0.1 卷积神经网络

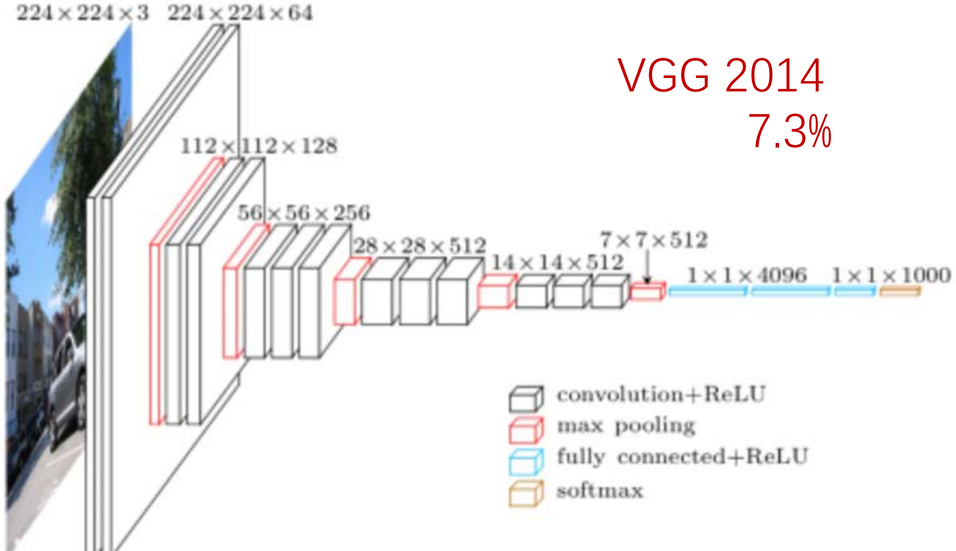
LeNet 1998



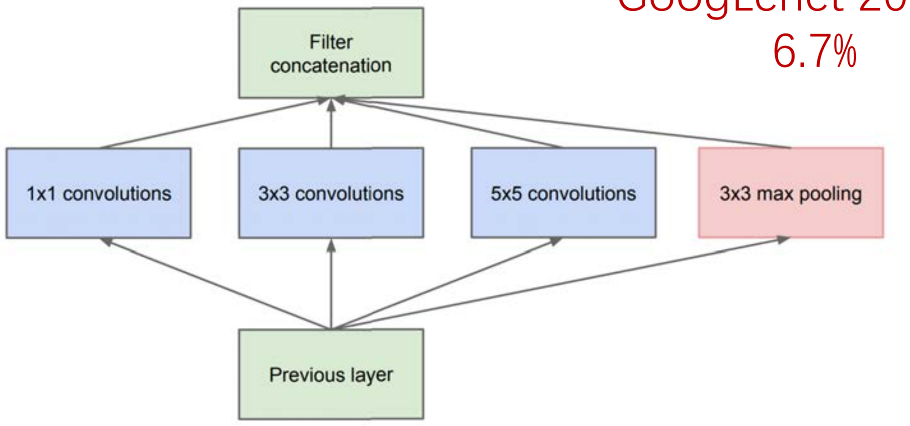

AlexNet 2012
16.4%




VGG 2014
7.3%

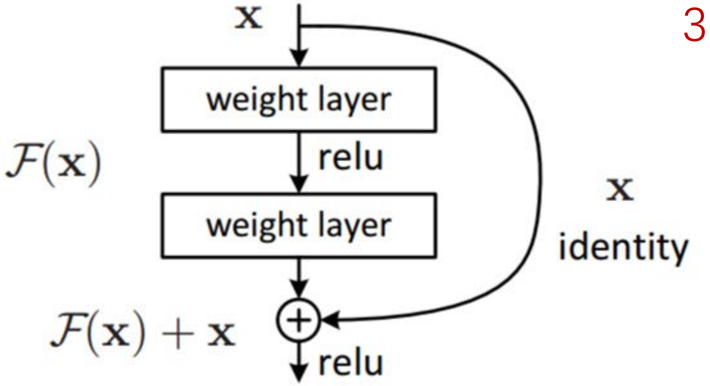



GoogLenet 2014
6.7%

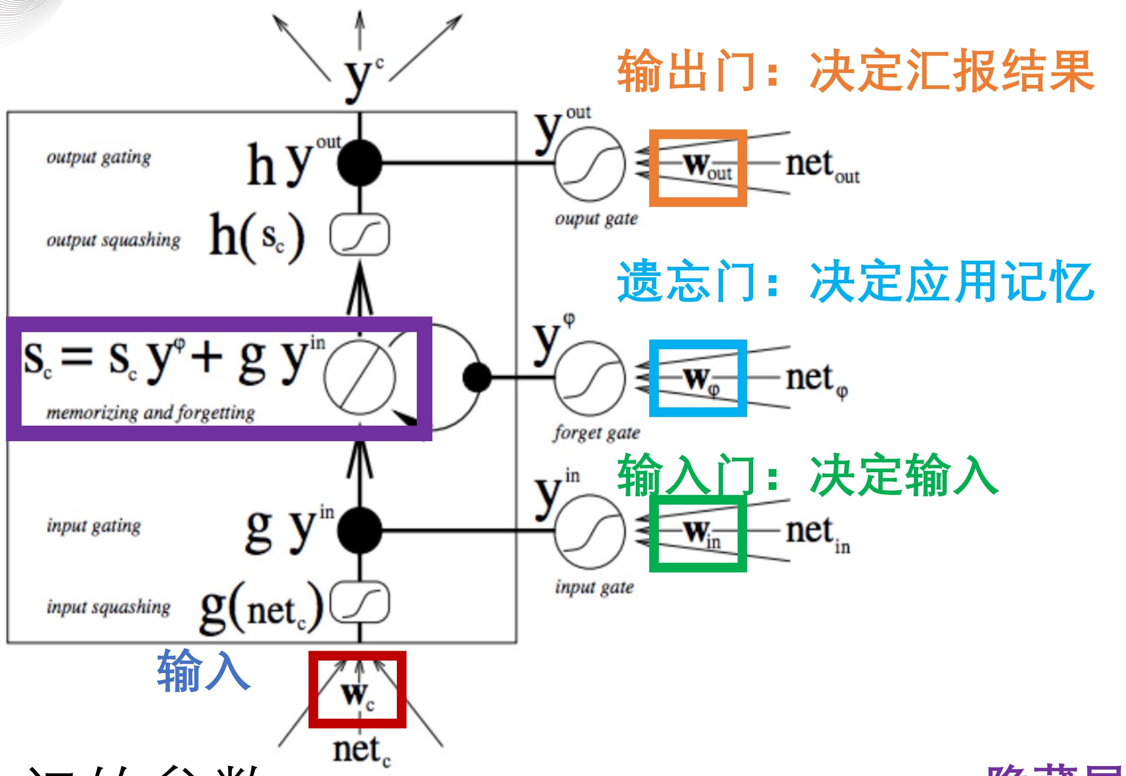



(a) Inception module, naïve version

ResNet 2015
3.6%



# 0.2 手撕LSTM




数据

| | | | | |
|---|---|---|---|---|
| 2 | 4 | 1 | 3 | 4 |
| 1 | 1 | 0 | 1 | 1 |
| 1 | 1 | 1 | 0 | 1 |
| 1 | 0 | 1 | 1 | 1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 输入g | | 2 | 4 | 1 | 3 | 4 |
| 输入门 | | 1 | 1 | 0 | 1 | 1 |
| 遗忘门 | | 1 | 1 | 1 | 0 | 1 |
| 隐藏层 | 0 | 2 | 6 | 6 | 3 | 7 |
| 输出门 | | 1 | 0 | 1 | 1 | 1 |
| 输出 | | 2 | 0 | 6 | 3 | 7 |

- 门的参数
  - 所有参数实际是训练出来的
  - 演示方便我们手动设定
  - $w_c = [1,0,0,0]$
  - $w_{in} = [0,1,0,0], b_{in} = -0.5, f = sigmoid$
  - $w_\varphi = [0,0,1,0], b_\varphi = -0.5, f = sigmoid$
  - $w_{out} = [0,0,0,1], b_{out} = -0.5, f = sigmoid$

# 4.1 强化学习

- 一个**复杂**、**长期**的问题



- 市场、投资组合
- 市场状态、投资决策、回报（多样化）

2
0
2
3

# 新潮神经网络

# 目 录

CONTENT

Part

01

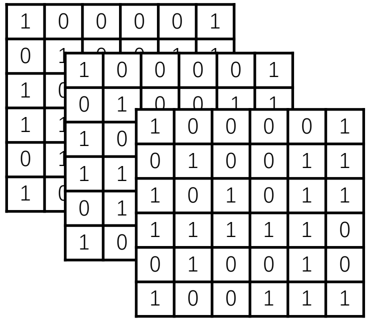
# 天剑绝刀

- 从天剑绝刀到倚天屠龙
- Attention!

# 1.0 CNN主图像 RNN主时序

- CNN：抓图像、文本中的**显著特征**，网络结构**优化**，做的**很深**
- RNN：把握**历史信息**，动态控制**权重分配**，序列结构
- CNN无法把握**序列关系**，难以应对变动长度
- RNN不易理解**高维关系**，不易并行处理
- 任务互相稍有交集，但表现说实话一般般
- 少林功夫加卡拉OK有没有搞头？

# 1.1 图片、文本、语音、经济运行数据、股市数据……




天地玄黄，宇宙洪荒。日月盈昃，辰宿列张。

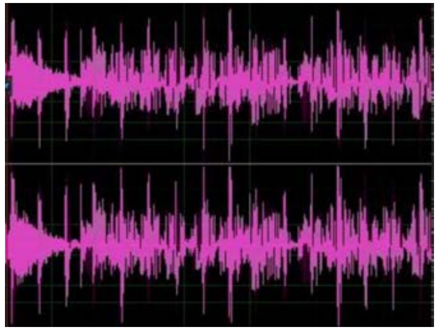
寒来暑往，秋收冬藏。闰余成岁，律吕调阳。

云腾致雨，露结为霜。金生丽水，玉出昆冈。

剑号巨阙，珠称夜光。果珍李柰，菜重芥姜。

海咸河淡，鳞潜羽翔。龙师火帝，鸟官人皇。

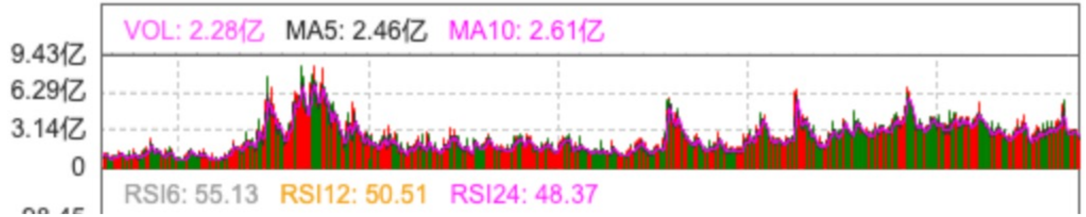
| | | | |
|---|---|---|---|
| -0.4 | 0.37 | 0.02 | -0.34 |
| -0.15 | -0.02 | -0.23 | -0.23 |
| 0.19 | -0.4 | 0.35 | -0.48 |
| -0.08 | 0.31 | 0.56 | 0.07 |
| -0.04 | -0.09 | 0.11 | -0.06 |
| 0.27 | -0.28 | -0.2 | -0.43 |
| -0.02 | -0.67 | -0.21 | -0.48 |
| -0.04 | -0.3 | -0.18 | -0.47 |
| 0.09 | -0.46 | -0.35 | -0.24 |
| 0.21 | -0.48 | -0.56 | -0.37 |



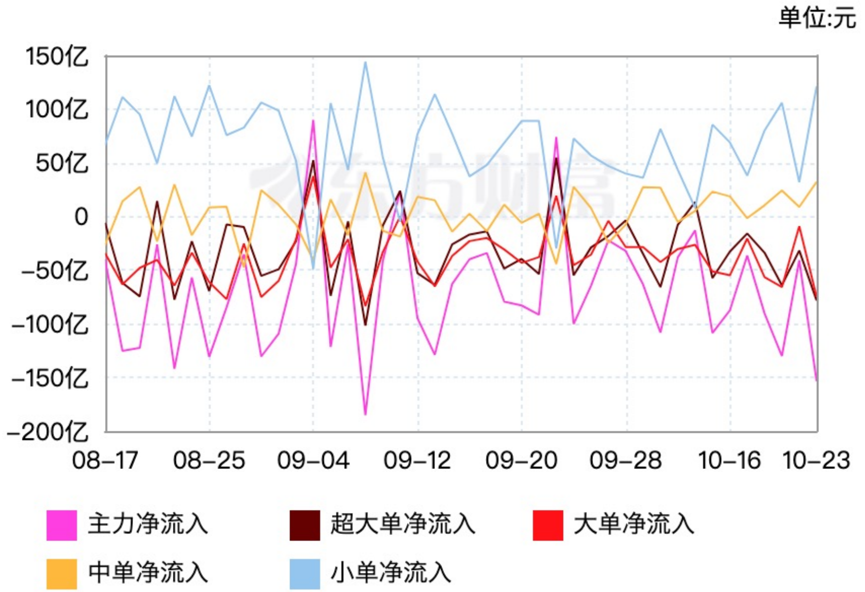
- 本质上说，数据可以被视为一个vector的序列
- 如果我们可以找到一个通用模型
- 变动长度的序列
- 多种多样的输出
- 可以修改的连接
- 那是不是就有个好模型了？

# 1.1 多种多样的任务

- 输入是一个序列的向量（长度可变）
- 输出是固定长度的序列
- 输出是一个值
- 输出是一个不定长度的序列





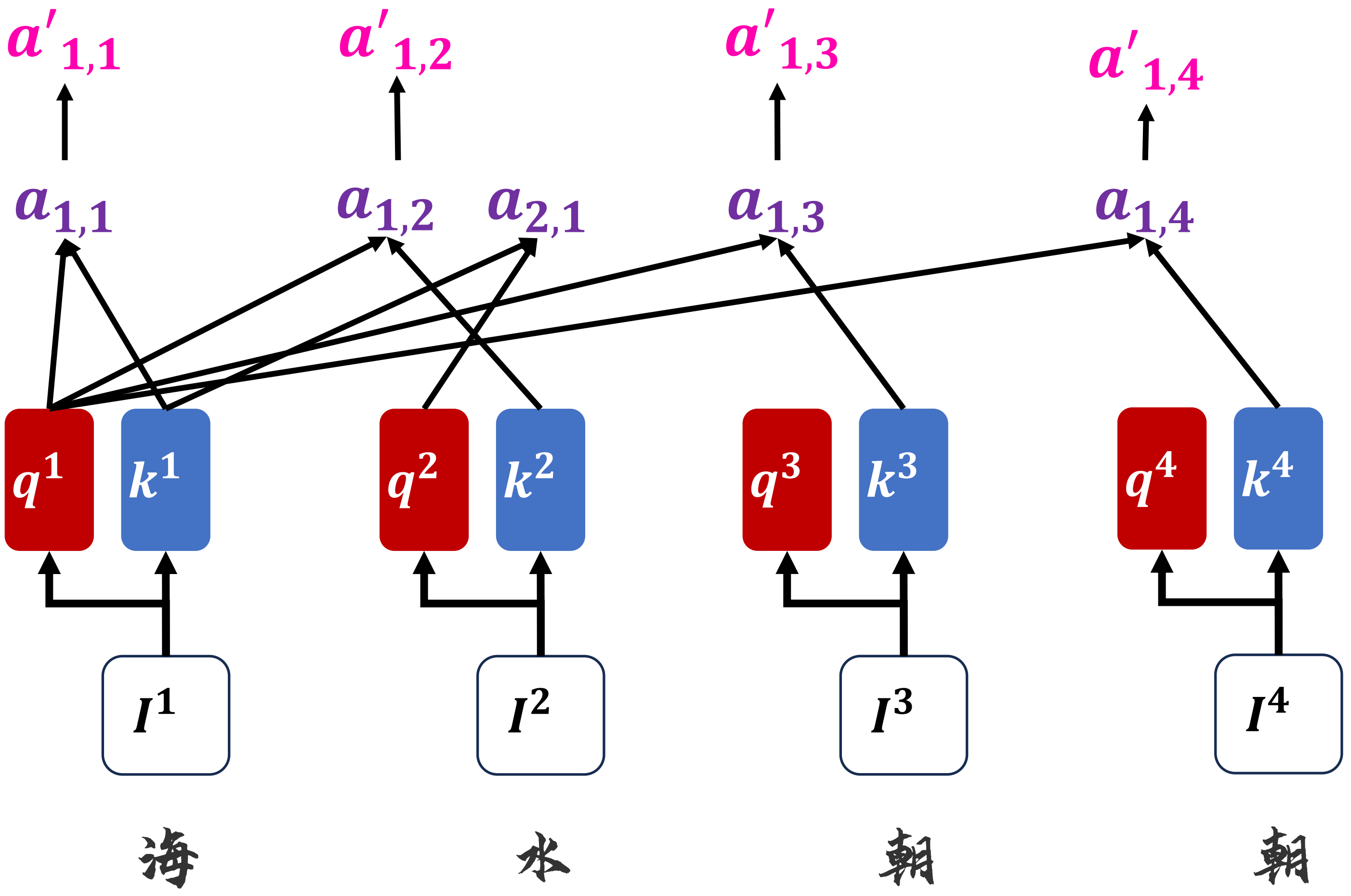
# 1.2 群策群力

海水朝朝朝朝朝朝朝落 浮云长长长长长长长消


马

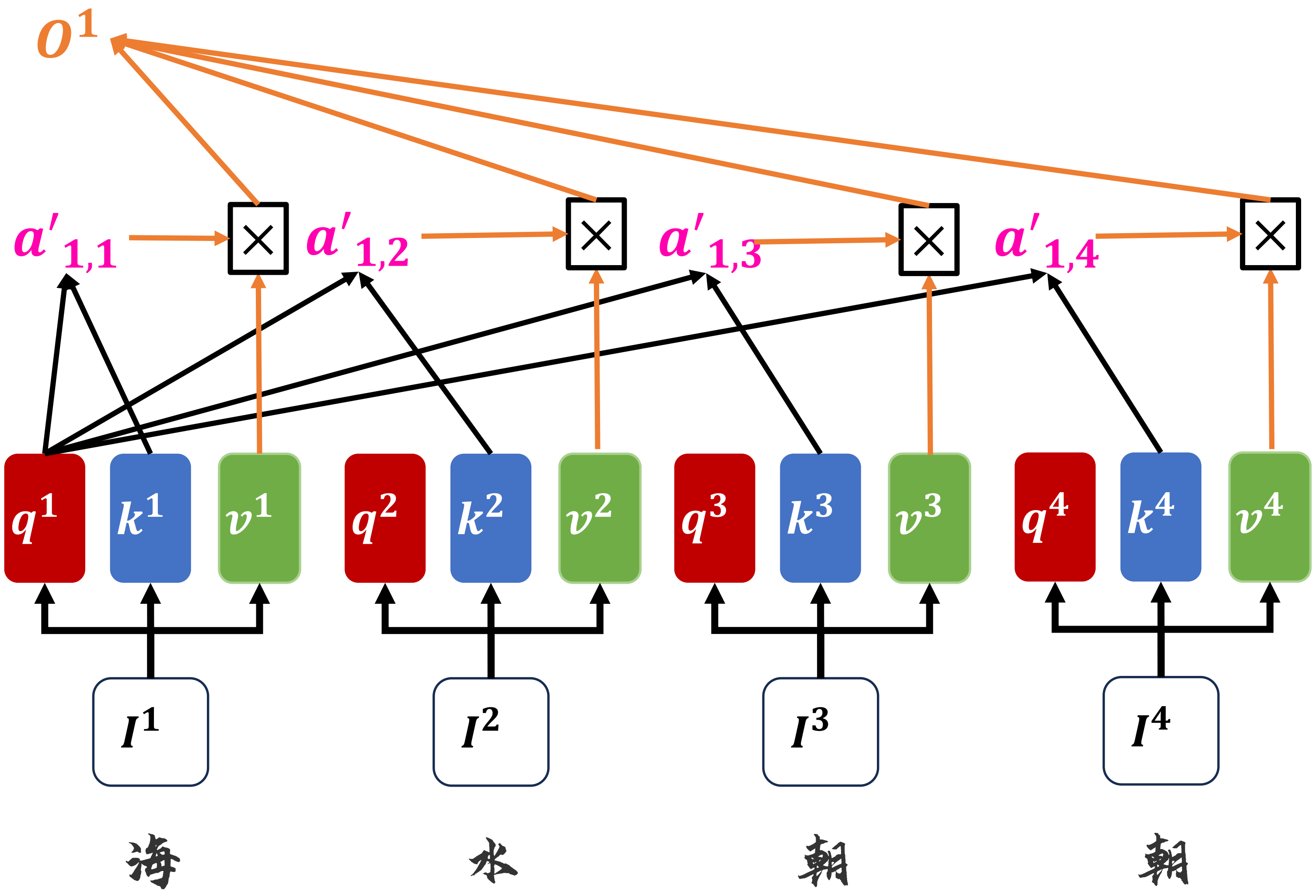
- CNN的计策：
  - 感受野：我们只看**一部分**
  - 参数共享：一部分在不同区域**共享参数**
- RNN的计策：
  - **序列关系**不能丢
  - 但也要“记得要忘记”，用门来控制**权重**
- **结合一下：**
  - **我们尝试使用门来决定感受野**
  - **马看见什么，是马决定的**

# 1.2 出山



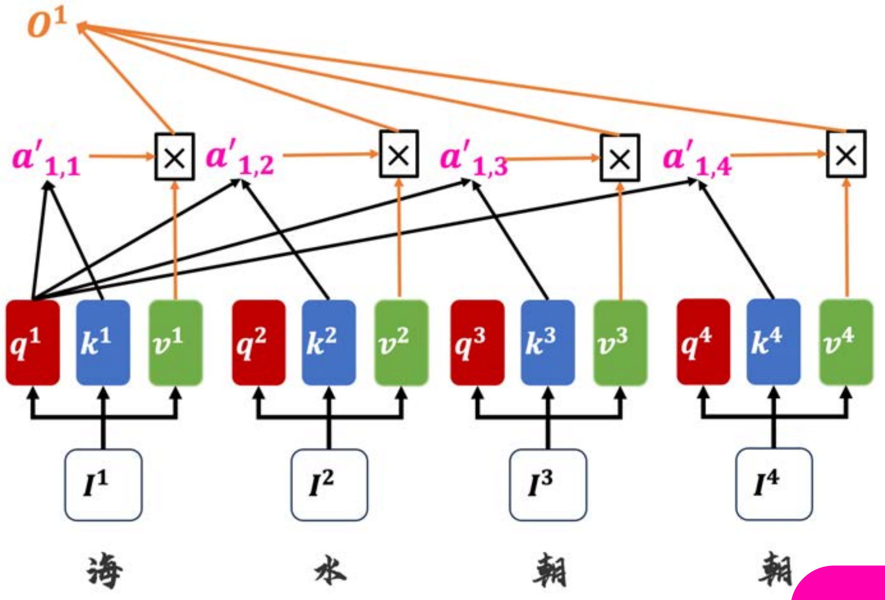

- $\boldsymbol{q}^{\mathrm{i}} = \boldsymbol{W}^{\boldsymbol{q}} * \boldsymbol{I}^{\boldsymbol{i}}$ $\boldsymbol{q}$ $for\ querry$
- $\boldsymbol{k}^{\boldsymbol{i}} = \boldsymbol{W}^{\boldsymbol{k}} * \boldsymbol{I}^{\boldsymbol{i}}$ $\boldsymbol{k}$ $for\ key$
- 参数共享!
- $a_{i,j} = {q^{i}}^{T} k^{j}$ a $for\ attention$
- $a'_{i,j} = Relu(a_{i,j})$
- $Relu(x) = \max(0, x)$

## 1.2 出山

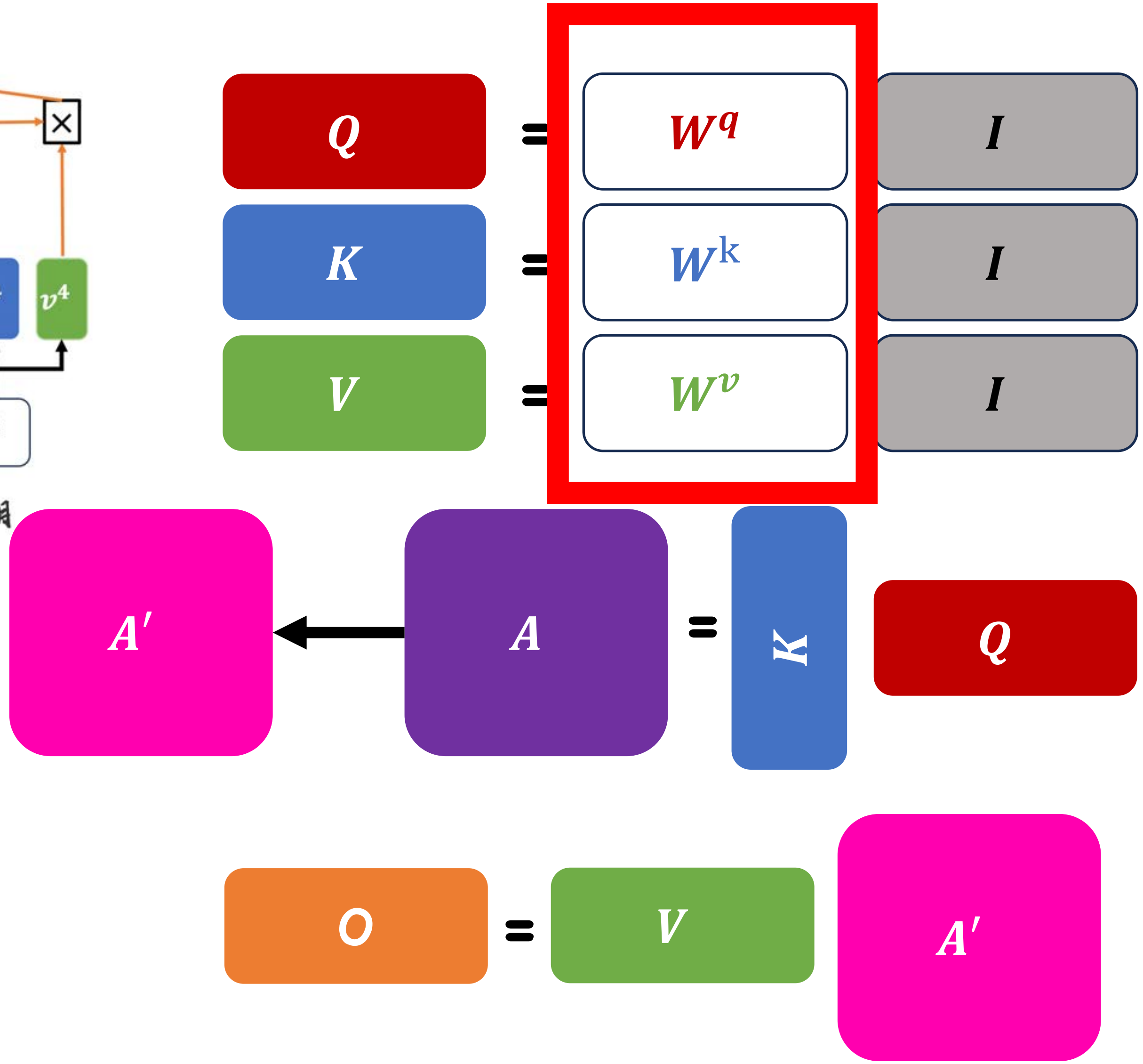



- $q^i = W^q * I^i$ $q$ *for querry*
- $k^i = W^k * I^i$ $k$ *for key*
- 参数共享!
- $a'_{i,j} = Relu\left({q^i}^T k^j\right)$
- $a'$ *for attention*
- $v^i = W^v * I^i$ $v$ *for value*
- $O^i = \sum_{j=1}^{D} a'_{i,j} * v^j$ *Output*

# 1.2 复习下

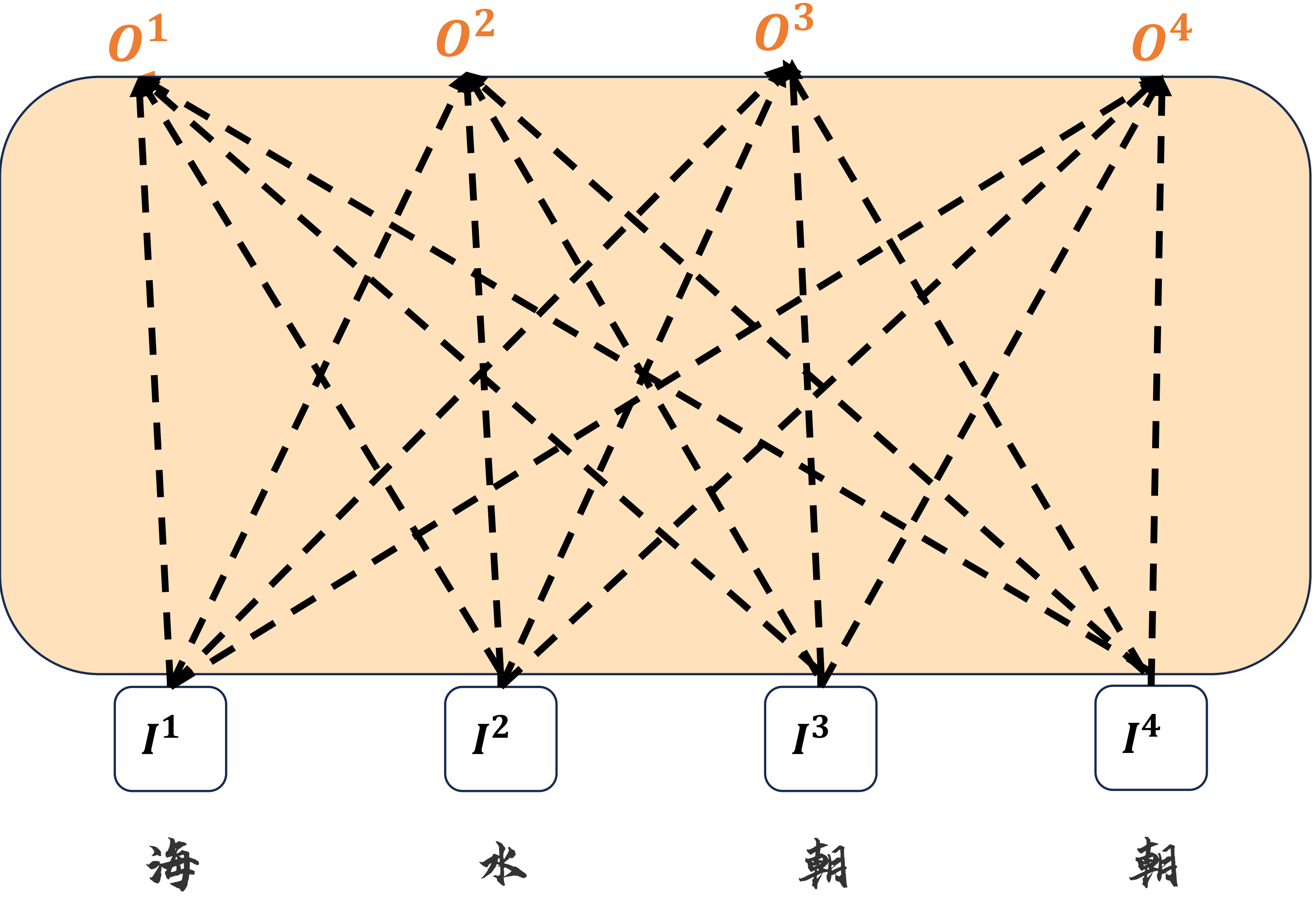



- $q^i = W^q * I^i$ $q$ $for$ $querry$
- $k^i = W^k * I^i$ $k$ $for$ $key$
- $a_{i,j} = {q^i}^T k^j$ a $for$ $attention$
- $a'_{i,j} = Relu(a_{i,j})$
- a′ $for$ $attention$
- $v^i = W^v * I^i$ $v$ $for$ $value$
- $O^i = \sum_{j=1}^{D} a'_{i,j} * v^j$ $O$ $for$ $Output$




只有三个权重矩阵需要被训练

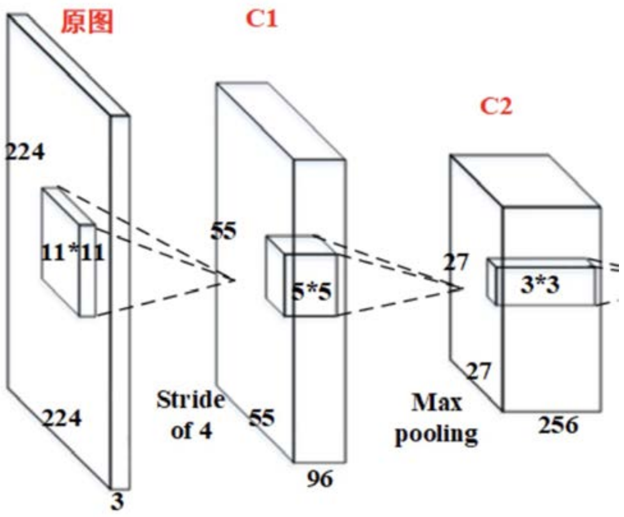
# 1.2 Self Attention 自注意力模型

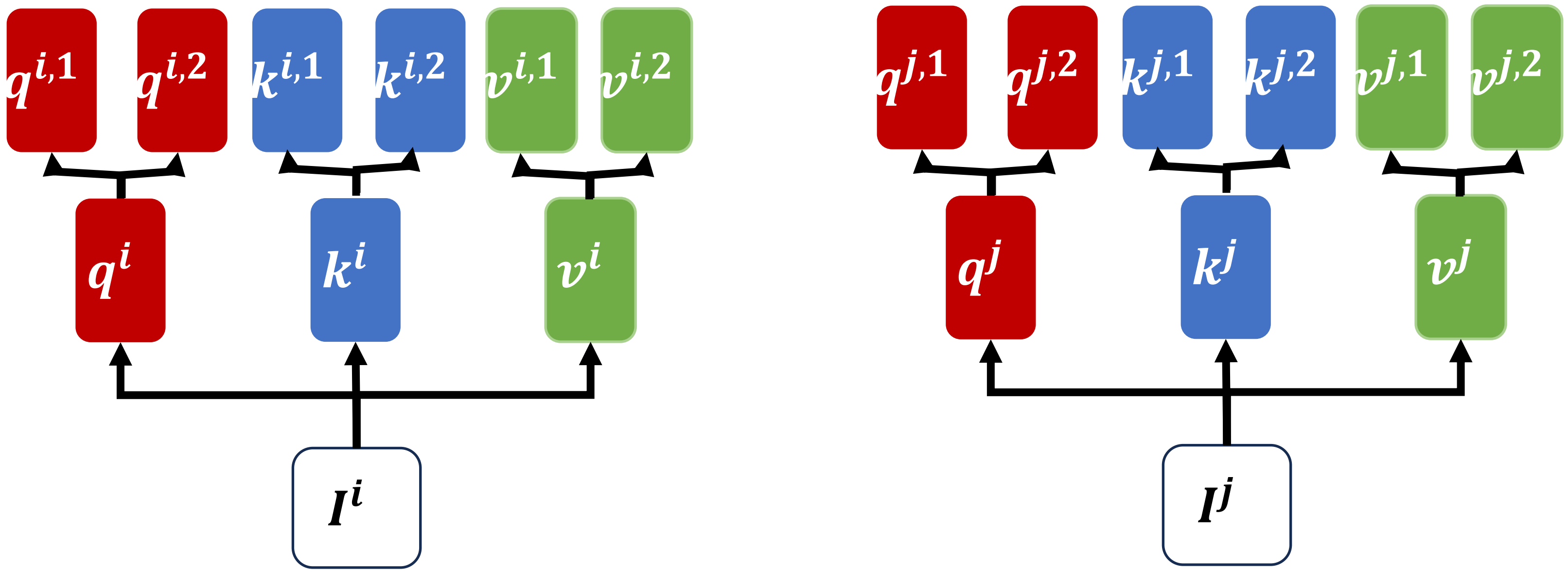


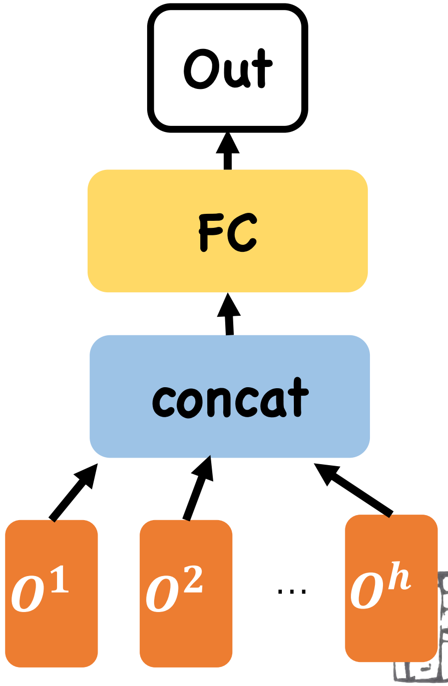

- $q^i = W^q * I^i$ $q$ $for$ $querry$
- $k^i = W^k * I^i$ $k$ $for$ $key$
- $a'_{i,j} = Relu\left({q^i}^T k^j\right)$
- a′ $for$ $attention$
- $v^i = W^v * I^i$ $v$ $for$ $value$
- $O^i = \sum_{j=1}^{D} a'_{i,j} * v^j$ $Output$
- **全局感受野**: 间接“全连接”
- 可以堆叠使用:
  - I、O可能都是隐藏层

# 1.2 Multi head Attention 多头注意力

- CNN又来了：
  - 我的经验是往深里卷，来挖掘更多信息
  - 一套Q、K、V只能有一套信息
  - 能不能多套？
- 当然可以啊，多头Attention！



1.3

# Attention is all you need （2017）

https://arxiv.org/abs/1706.03762

## 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.

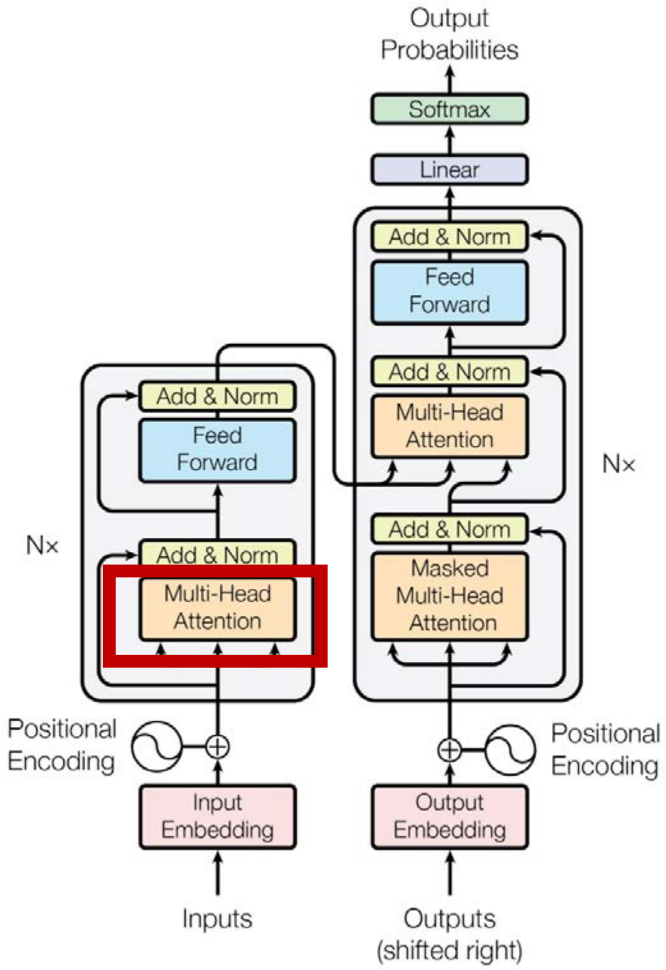
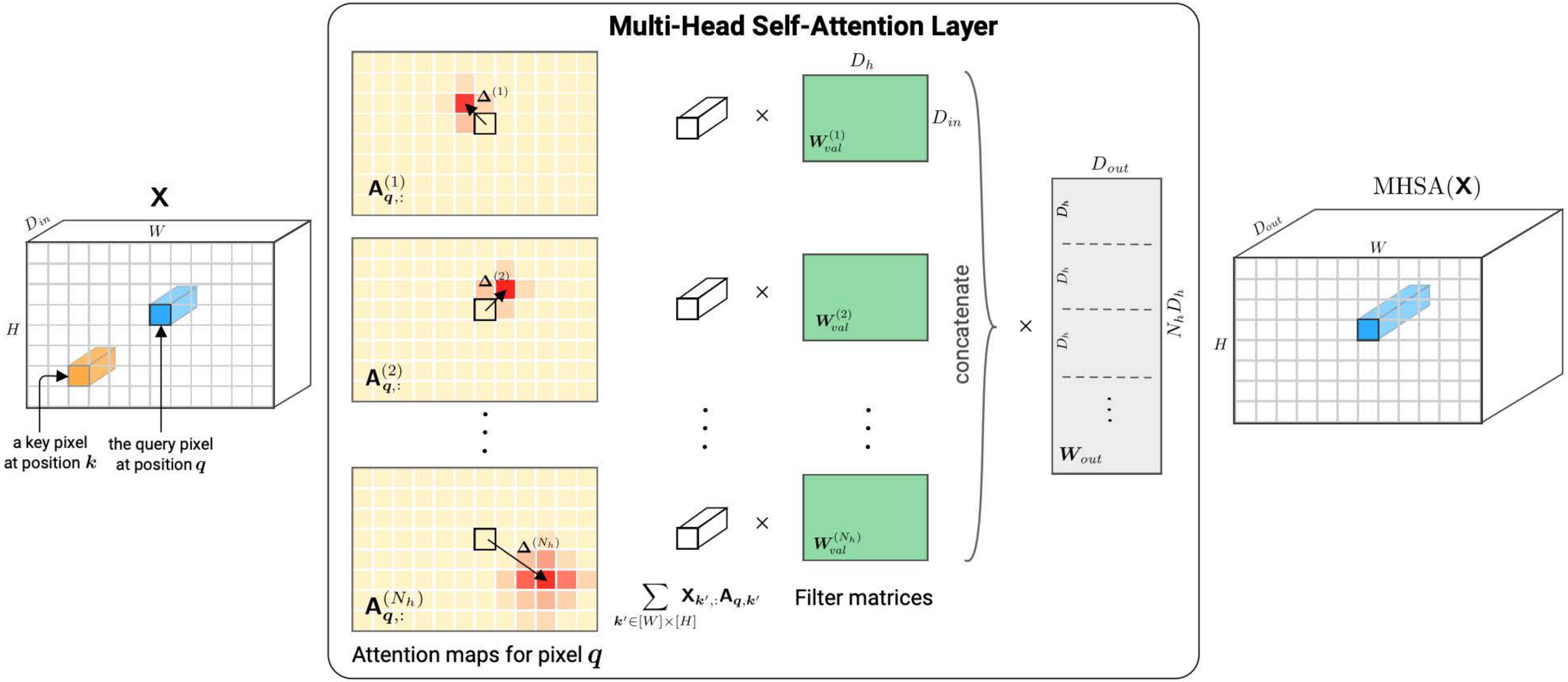




Figure 1: The Transformer - model architecture.

# 1.3 CNN确实是Attention的一个特例（2019)

https://arxiv.org/abs/1911.03584



# 1.3 Attention力量的代价（2020）

https://arxiv.org/abs/2010.11929

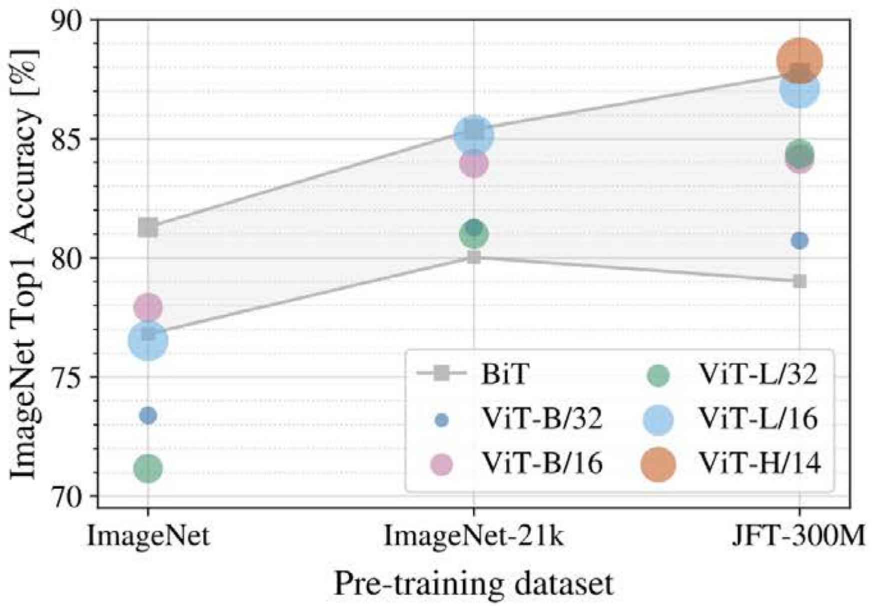



Figure 3: Transfer to ImageNet. While large ViT models perform worse than BiT ResNets (shaded area) when pre-trained on small datasets, they shine when pre-trained on larger datasets. Similarly, larger ViT variants overtake smaller ones as the dataset grows.

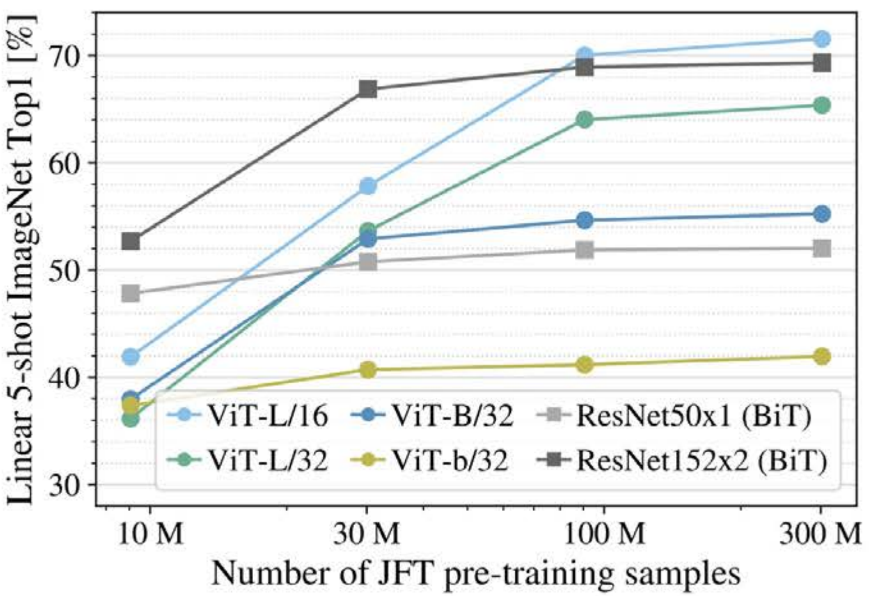



Figure 4: Linear few-shot evaluation on ImageNet versus pre-training size. ResNets perform better with smaller pre-training datasets but plateau sooner than ViT, which performs better with larger pre-training. ViT-b is ViT-B with all hidden dimensions halved.

# 1.4 天剑绝刀 到 倚天屠龙

天剑主守，绝刀主攻。一守一攻，一柔一刚，尽败天下武林。

卧龙生 1964

金庸 1961