

# 大模型分享

## 1.什么是大模型

定义：具有大规模参数和复杂计算结构的机器学习模型，数千万到数十亿个参数，通常是由神经网络构成

作用：大模型的设计是为了利用海量的数据来应对更加复杂的内容，提高模型的表达能力和预测性能

## 2.介绍一下大模型的发展历程



## 3.大模型的应用领域

**自然语言处理 (NLP)：**大型模型在NLP领域的应用最为突出。它们可以用于语言建模、文本分类、情感分析、问答系统、命名实体识别、文本生成（包括机器翻译、摘要生成、对话系统等）等任务。

**强化学习：**在强化学习领域，大型模型通常用于学习环境的表示和策略的生成。例如，AlphaGo和AlphaZero是由大型深度神经网络驱动的强化学习系统，它们在围棋、象棋等游戏中取得了令人瞩目的成绩。

**医疗保健：**在医疗保健领域，大型模型可以用于医学影像分析、病理学诊断、基因组学研究等任务。它们可以帮助医生更快速、更准确地进行诊断和治疗。

**金融领域：**在金融领域，大型模型可以用于风险管理、信用评分、市场预测等任务。它们可以分析海量的金融数据，并从中发现隐藏的模式和规律。

**游戏开发：**在游戏开发领域，大型模型可以用于游戏智能体的设计和训练，以及游戏环境的生成和优化。

## 4.介绍一下当前比较火的大模型

- chatgpt
- gemini
- Llama
- grok (xAi完全开源)
- chatGlm，清华大学的计算机科学与技术系、人工智能研究院、智能技术与系统国家重点实验室等单位联合开发

- Qwen/2 (阿里的千问), 文言一心, 百川等
- AIGC方向
- DALL·E 3、Midjourney、Stable Difusion、sora

## 5.介绍大模型的基础架构

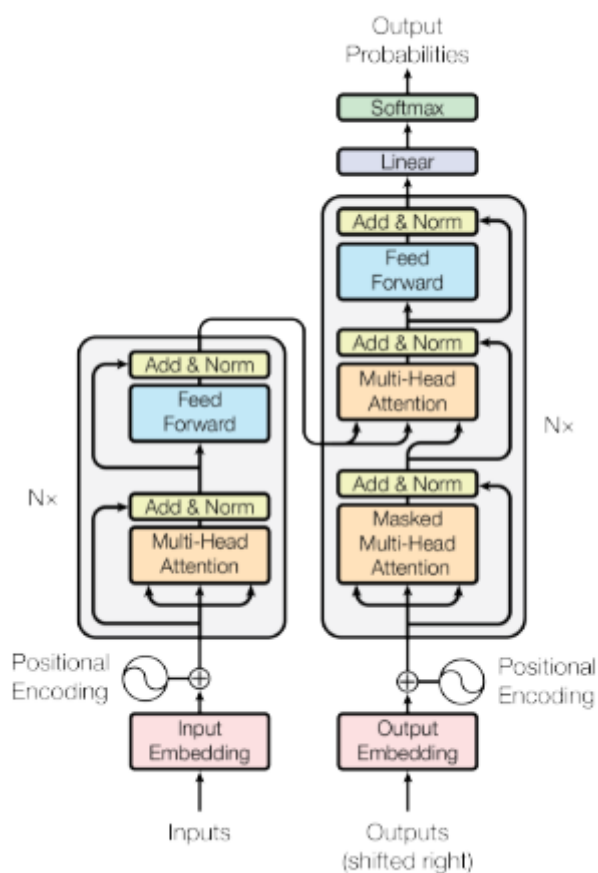
上述大模型都采用了transformer架构或transformer变体,

采用了Transformer的预训练模型, 如BERT、GPT等, 在NLP领域引起了重大突破, 表现好的令人惊讶

在transformer之前, nlp领域一直使用的是RNN及其变体

由于RNN特殊的网络结构, 处理长序列时容易出现梯度消失或梯度爆炸的问题, NLP领域一直没有什么突破, 直到Google在2017年的[Attention Is All You Need](#)论文中提出注意力机制, 才有后续的大模型

transformer架构的核心机制就是注意力机制



自注意力机制:

自注意力机制是一种用于处理序列数据的机制, 它能够在序列中的每个位置上计算该位置与其他位置之间的关联程度, 并根据这些关联程度来加权组合序列中的信息

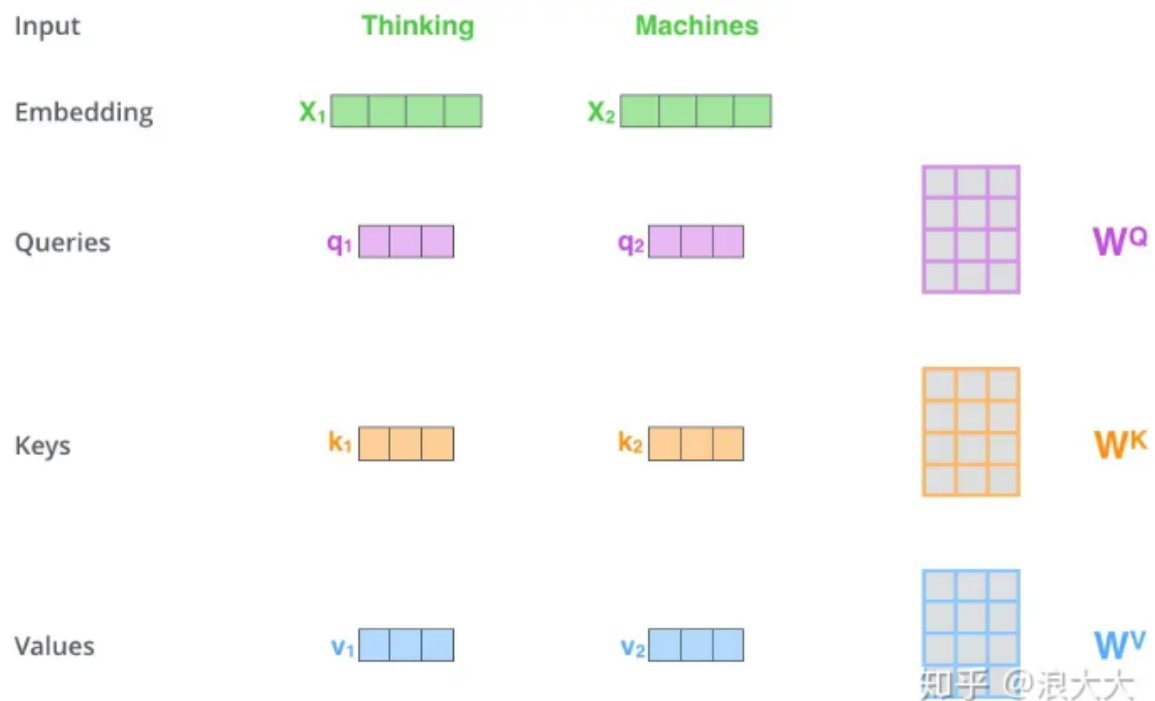
自注意力的基本概念:

Q (Query): 查询是你想要了解的信息或者你想要从文本中提取的特征。

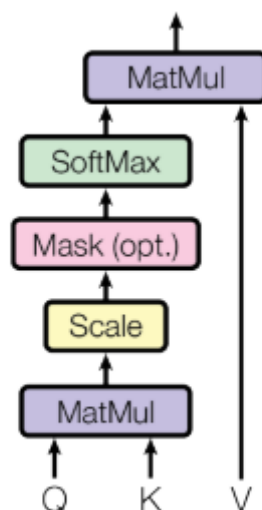
K (key): 键是文本中每个词语的表示

V (Value): 值是与每个词语相关的具体信息或特征

这三个向量是通过词嵌入与三个权重矩阵即  $W^Q, W^K, W^V$ , 相乘后创建出来的



自注意力的计算过程



step1: 将输入数据进行位置编码得到词向量

step2: 词向量与 $w_q, w_k, w_v$ 三个权重矩阵相乘的得到 $Q, K, V$

step3: 计算 $Q, K$ 的点积, 然后Scale, 这里就是除以 $\sqrt{d_k}$

step4: softmax进行归一化处理, 将每个值映射到0-1之间, 并且相加=1

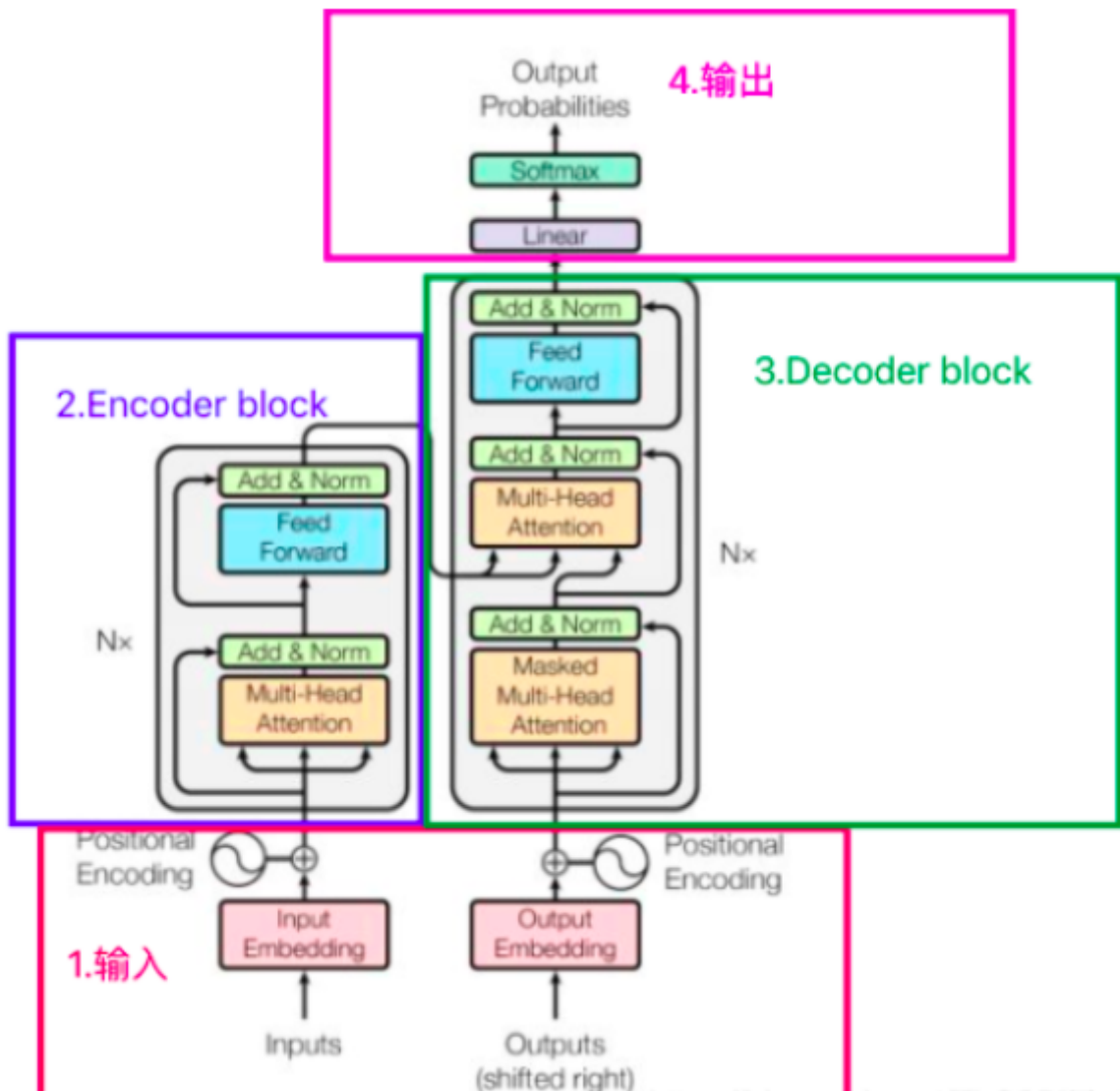
step5: 与 $V$ 进行加权求和

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

多头自注意力:

就是有多组 $QKV$ , 然后将每个注意力头输出的结果进行concat, 再与一个附加权重矩阵 ( $W_o$ ) 相乘,  $W_o$ 在模型中是与 $W_q, W_k, W_v$ 联合训练的

回过头来看transformer结构



## 6. 训练过程

大模型的训练与训练一般的神经网络类似

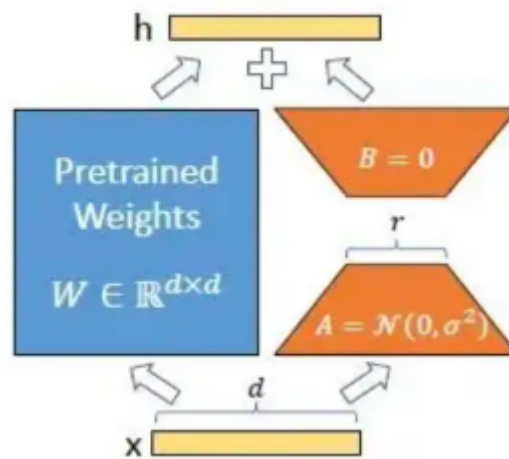
1. 数据集的准备
2. 构建模型，定义好激活函数和损失函数，选好优化器
3. 开始训练

不过大模型的训练非常消耗算力和时间，一般使用大模型都是那预训练模型来做微调

## 7. 下面分享一下大模型的微调方法

参考: <https://segmentfault.com/a/1190000044015741#item-2-3>

1. LoRA, 大语言模型的低阶自适应。LoRA 的基本原理是冻结预训练好的模型权重参数，在冻结原模型参数的情况下，通过往模型中加入额外的网络层，并只训练这些新增的网络层参数



2. P-tuning v2，它是在prefix-tuning方法和P-tuning v1方法上改进而来

Prefix-tuning 微调方法在模型中加入 prefix，即连续的特定任务向量，微调时只优化这一小段参数


P-tuning v1 微调方法是将 Prompt 加入到微调过程中，**只对 Prompt 部分的参数进行训练，而语言模型的参数固定不变**


这里简单介绍一下Prompt

提示工程（Prompt Engineering），也称为上下文提示，是一种通过不更新模型的权重/参数来引导LLM行为朝着特定结果的方法。这是与AI有效交流所需结果的过程。提示工程可以用于各种任务，从回答问题到算术推理乃至各种应用领域，理解提示工程，能够帮助我们了解LLM的限制和能力。


举个例子：

在chatGPT中，用户通常使用 prompt 来与大语言模型进行交互，请求回答问题、生成文本、完成任务等。模型会根据提供的 prompt 来生成一个与之相关的文本，尽量符合用户的要求。


 您  
你现在是一名医生


 ChatGPT  
我可以扮演一名医生的角色，提供关于医学、健康和医疗的建议和信息。有什么可以帮助你吗？

 您  
我牙疼

 ChatGPT  
牙疼可能是许多问题的症状，包括蛀牙、牙龈炎、牙髓炎等。首先，你可以尝试用温盐水漱口，这有助于缓解疼痛和减轻炎症。如果牙疼持续或加剧，最好尽快预约看牙医，以确诊问题并获得适当的治疗。牙医可以检查你的牙齿，并根据需要给予治疗建议。同时，避免食用过硬或过热的食物，以免加重疼痛。

没有使用prompt

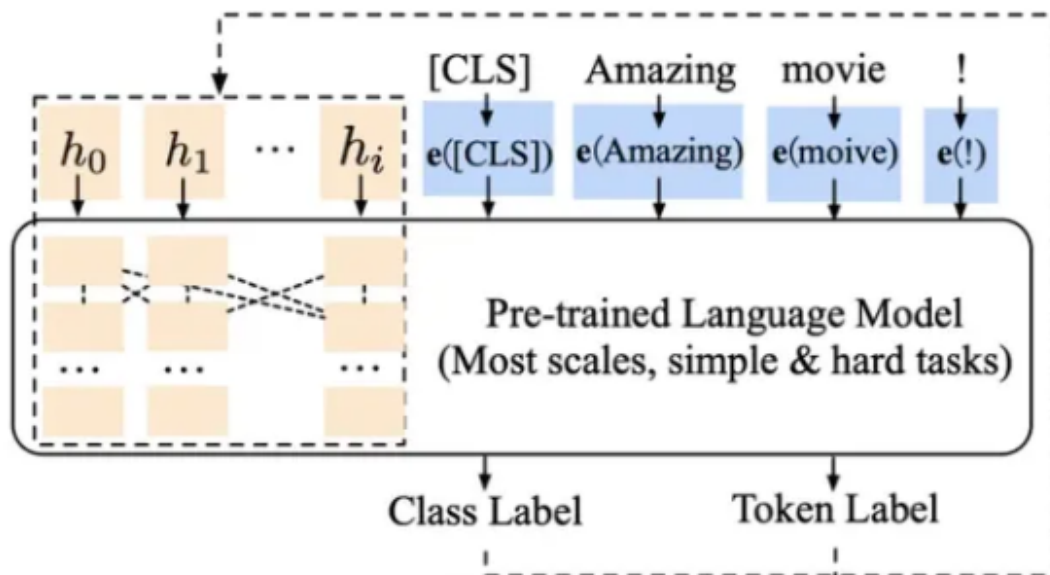
 您  
我牙疼

 ChatGPT  
牙疼确实很难受。你可以尝试用温盐水漱口缓解一下，或者用含有酚妥拉明成分的药物暂时缓解疼痛。但如果疼痛持续或者非常严重的话，最好尽快去看牙医，以免病情加重。



原文链接：[https://blog.csdn.net/2401\\_82469710/article/details/138614802](https://blog.csdn.net/2401_82469710/article/details/138614802)

P-tuning v2就是结合并改进了Prefix-tuning和v1方法



### 3. Freeze

这个方法就是冻结网络部分层，使其不参与梯度更新，只有未被冻结的网络层可以更新

## 8. 例子

以chatGlm3-6b为例

项目地址: <https://github.com/THUDM/ChatGLM3>

### 1. 安装依赖

```
jieba>=0.42.1
ruamel_yaml>=0.18.6
rouge_chinese>=1.0.3
jupyter>=1.0.0
datasets>=2.18.0
peft>=0.10.0
deepspeed==0.13.1
mpi4py>=3.1.5
```

### 2. 整理数据集格式

微调模型的对话能力，而非工具能力，你应该按照以下格式整理数据

```
[
  {
    "conversations": [
      {
        "role": "system",
        "content": "<system prompt text>"
      },
      {
        "role": "user",
        "content": "<user prompt text>"
      },
      {
        "role": "assistant",
        "content": "<assistant response text>"
      },
      // ... Muti Turn
    ]
  }
]
```

```

    {
      "role": "user",
      "content": "<user prompt text>"
    },
    {
      "role": "assistant",
      "content": "<assistant response text>"
    }
  ]
}
// ...
]
```

以 AdvertiseGen 为例

```

{"conversations": [{"role": "user", "content": "类型#裙*裙长#半身裙"}, {"role": "assistant", "content": "这款百搭时尚的仙女半身裙，整体设计非常的飘逸随性，穿上之后每个女孩子都能瞬间变成小仙女啦。料子非常的轻盈，透气性也很好，穿到夏天也很舒适。"}]}
```

### 调整配置文件

微调的配置文件位于 config 目录下，包括以下文件

1. ds\_zero2 / ds\_zero3.json: deepspeed 配置文件。
2. lora.yaml / ptuning.yaml / sft.yaml

配置好后就可以开始微调训练了

LLaMA Factory 是一款**支持多种 LLM**微调方式的工具，包括**预训练**、指令监督微调和奖励模型训练等。它支持 LoRA 和 QLoRA 微调策略，广泛**集成了业界前沿的微调方法**

[https://github.com/hiyouga/LLaMA-Factory/blob/main/README\\_zh.md](https://github.com/hiyouga/LLaMA-Factory/blob/main/README_zh.md)

Unsloth: 一个开源的大模型训练加速项目，使用 OpenAI 的 Triton 对模型的计算过程进行重写，**大幅提升模型的训练速度，降低训练中的显存占用**

<https://my.oschina.net/HuggingFace/blog/11106128>