- NLP方向总结-transformer
  - Transformer的原理
    - 1. Transformer为何使用多头注意力机制？（为什么不使用一个头）
      答：多头可以使参数矩阵形成多个子空间，矩阵整体的size不变，只是改变了每个head对应的维度大小，这样做使矩阵对多方面信息进行学习，
      但是计算量和单个head差不多。
    - 4.为什么在进行softmax之前需要对attention进行scaled（为什么除以dk的平方根），并使用公式推导进行讲解
      答：假设 Q 和 K 的均值为0，方差为1。它们的矩阵乘积将有均值为0，方差为dk，因此使用dk的平方根被用于缩放，因为，Q 和 K 的矩阵乘积的均值本应该为0，方差本应该为1，这样可以获得更平缓的softmax。当维度很大时，点积结果会很大，会导致softmax的梯度很小。为了减轻这个影响，对点积进行缩放。
    - 6.为什么在进行多头注意力的时候需要对每个head进行降维？
      答：将原有的高维空间转化为多个低维空间并再最后进行拼接，形成同样维度的输出，借此丰富特性信息，降低了计算量
    - 7.大概讲一下Transformer的Encoder模块？
      答：输入嵌入-加上位置编码-多个编码器层（每个编码器层包含全连接层，多头注意力层和点式前馈网络层（包含激活函数层）
    - 11.简单讲一下Transformer中的残差结构以及意义。
      答：encoder和decoder的self-attention层和ffn层都有残差连接。反向传播的时候不会造成梯度消失。
    - 12.为什么transformer块使用LayerNorm而不是BatchNorm？LayerNorm 在Transformer的位置是哪里？
      答：多头注意力层和激活函数层之间。CV使用BN是认为channel维度的信息对cv方面有重要意义，如果对channel维度也归一化会造
      成不同通道信息一定的损失。而同理nlp领域认为句子长度不一致，并且各个batch的信息没什么关系，因此只考虑句子内信息的归
      一化，也就是LN。
    - 17.Transformer的并行化提现在哪个地方？
      答：Transformer的并行化主要体现在self-attention模块，在Encoder端Transformer可以并行处理整个序列，并得到整个输入序>列经过Encoder端的输出，但是rnn只能从前到后的执行，并行计算和建模长距离依赖方面表现出优势。
    - 18.Decoder端可以做并行化吗？训练的时候可以，但是交互的时候不可以
    - 19.序列长度越长，效率越低的原因是？
      self-attention的计算需要存储一个$T\times T$注意力分布矩阵，T表示序列长度
  - MacBERT
    - https://zhuanlan.zhihu.com/p/333202482
    - 中文自然语言预训练模型
  - 论文
    - LLaMA： Open and Effcient Foundation Language Models
      - LLaMA： Open and Effcient Foundation Language Models
        - 定义
          - 发布的模型版本：70亿，130亿，300亿，650亿参数
            - 70亿参数大概13GB，130亿大概25GB，300亿大概61GB，650亿大概120GB
          - 650亿的模型和PaLM-5400亿参数的模型性能相当
        - 模型
          - 训练数据（表1）包括：CommonCrawl，C4爬取的数据，Github代码，维基百科，Stack Exchange，ArXiv论文
            - 共1.4T token
          - transformer架构
          - 预归一化
            - 对每个transformer子层的输入进行归一化，而不是对输出进行归一化。我们使用Zhang和Sennrich（2019）介绍的RMSNorm归一化函数。
          - SwiGLU
            - 用SwiGLU激活函数取代ReLU非线性
          - 旋转嵌入
            - 删除了绝对位置嵌入，取而代之的是在网络的每一层添加Su等人（2021）介绍的旋转位置嵌入（RoPE）
          - AdamW优化器
          - xformers加速
          - 使用2048个A100 GPU（80GB）训练21天
    - GPT-4 Technical Report
      - GPT-4技术报告
        - 定义
          - 对比模型GPT-3.5，GPT-4接收视觉输入，GPT-4接受视觉输入
          - 使用带有人类反馈的强化学习（RLHF）对模型的行为进行微调，以产生更符合用户意图的响应。
        - 模型
          - GPT-4损失预测
            - 使用一个小于GPT4的1万倍的小模型，使用拟合的比例法高度准确地预测了GPT-4的最终损失，参考论文：Scaling Laws for Autoregressive Generative Modeling
          - 有些任务，模型越大，性能越差，但是GPT4性能确很好（图3）
            - https://github.com/inverse-scaling/prize
            - 缩放定律
              - 随着参数数量、使用的计算量和数据集大小的增加，语言模型会变得可预测地更好
            - 逆缩放定律
              - 有些任务的趋势相反：随着语言模型的整体测试损失的改善，任务性能变得单调，可以预见地恶化
            - 看了下，都是奇葩问法
    - GPT-4 System Card
      - GPT-4系统卡
        - 本文主要说的是GPT-4在安全方面的工作。
        - 训练的2个阶段
          - 预训练阶段
            - 使用来自互联网的大型文本数据集，以预测下一个单词
          - 微调阶段
            - 使用一种叫做人类反馈强化学习（RLHF）的算法，以产生人工标注者喜欢的输出
            - 基于规则的奖励模型（RBRMs）
              - 在GPT4的系统安全上很重要
            - 规则的奖励模型和最开始的奖励模型的奖励按一定的权重相加
        - 模型
          - 专家团队包括： 虚假信息、化学、生物风险、网络安全、核风险、经济学、人机交互、法律、教育和医疗方面有专长的人
          - GPT-4在遵循用户意图的能力方面比以前的模型有很大的改进。
          - 安全测试
            - Hallucinations幻觉
              - 即测试模型产生的虚假信息，产生与某些来源有关的无意义或不真实的内容。
              - GPT-4的得分比GPT-3.5模型在避免开放领域的产生幻觉高19个百分点
            - 有害的内容
              - 是违反人为设定策略的内容，或可能对个人、团体或社会构成伤害的内容
            - 刻板印象，贬低，歧视，偏差的世界观
              - 并且使用策略禁止模型和产品用于高风险的政府决策（如执法、刑事司法、移民和庇护），或用于提供法律或健康建议
            - 和平和安全，非常规武器
            - 隐私测试
              - 不泄露隐私
            - 网络安全
              - 网络攻击和新型漏洞识别发现和利用
            - GPT4越狱（图10）
              - 利用特殊的提示，即特殊的角色扮演的方法，让模型逃离约束
    - Zero-Shot Information Extraction via Chatting with ChatGPT
      - 通过与ChatGPT的聊天进行zero-shot的信息提取
        - 定义
          - 主要测试了：实体关系三元组提取、命名实体识别和事件提取，测试英文和中文，六个数据集
        - 模型（图1）
          - 可以限制chatgpt生成json格式的输出，方便解析
          - 2阶段对话生成(表2，3,4)
            - 第一阶段：找到句子中的元素类型
              - 找给给出句子可能包含的实体，关系，事件类型
                - eg：实体识别
                  - 过滤出句子中现有的实体类型，并给出所需的类型列表。
                - eg：事件提取
                  - 获取事件类型的文本分类问题
            - 第二阶段：多个QA回合
              - 分别根据实体类型，关系类型，事件类型，找出句子中对应的词