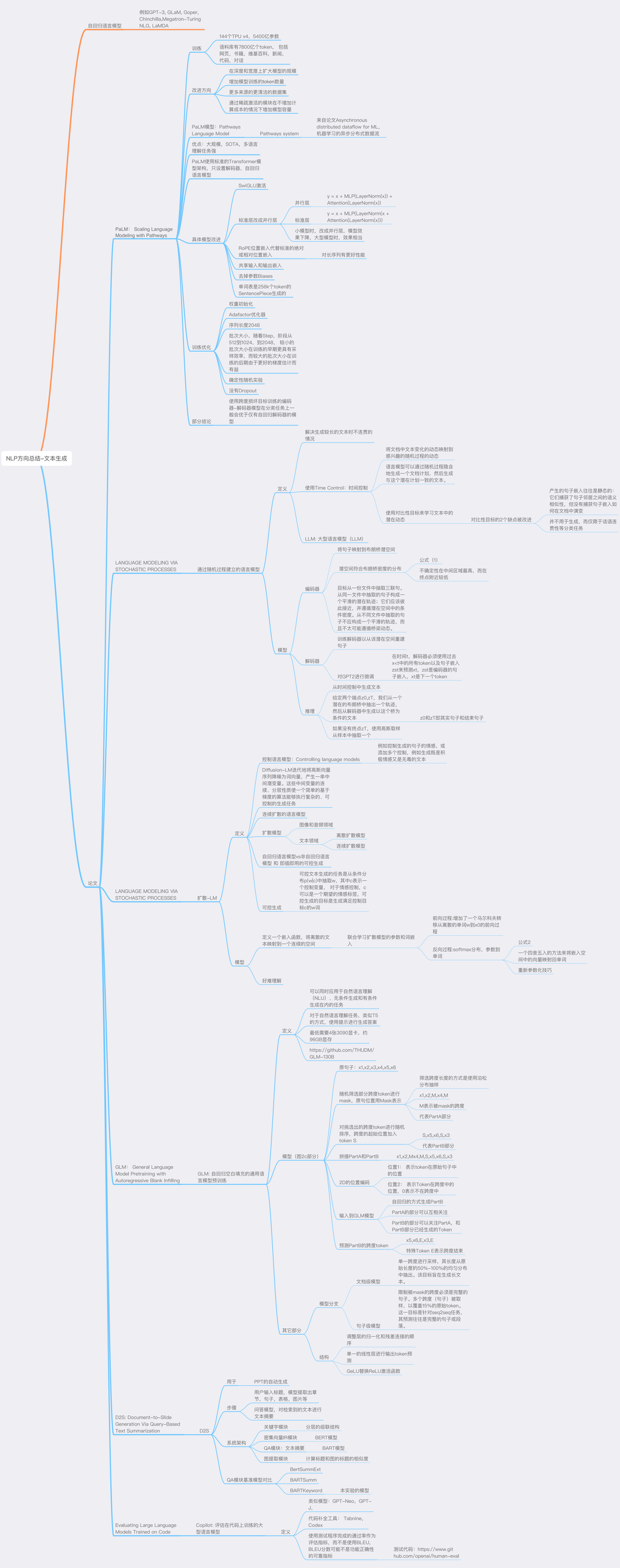
NLP方向总结-文本生成
自回归语言模型
例如GPT-3, GLaM, Goper, Chinchilla,Megatron-Turing NLG, LaMDA
PaLM：Scaling Language Modeling with Pathways
训练
144个TPU v4，5400亿参数
语料库有7800亿个token， 包括网页，书籍，维基百科，新闻，代码，对话
具体模型改进
在深度和宽度上扩大模型的规模
增加模型训练的token数量
更多来源的更清洁的数据集
通过稀疏激活的模块在不增加计算成本的情况下增加模型容量
PaLM模型：Pathways Language Model
Pathways system
来自论文Asynchronous distributed dataflow for ML，机器学习的异步分布式数据流
优点：大规模，SOTA，多语言理解任务强
PaLM使用标准的Transformer模型架构，只设置解码器，自回归语言模型
SwiGLU激活
标准层改成并行层
并行层
y = x + MLP(LayerNorm(x)) + Attention(LayerNorm(x))
标准层
y = x + MLP(LayerNorm(x + Attention(LayerNorm(x)))
小模型时，改成并行层，模型效果下降，大型模型时，效果相当
RoPE位置嵌入代替标准的绝对或相对位置嵌入
对长序列有更好性能
共享输入和输出嵌入
去掉参数Biases
单词表是256k个token的SentencePiece生成的
训练优化
权重初始化
Adafactor优化器
序列长度2048
批次大小，随着Step，阶段从512到1024，到2048，较小的批次大小在训练的早期更具有采样效率，而较大的批次大小在训练的后期由于更好的梯度估计而有益
确定性随机实验
没有Dropout
部分结论
使用跨度损坏目标训练的编码器-解码器模型在分类任务上一般会优于仅有自回归解码器的模型
论文
LANGUAGE MODELING VIA STOCHASTIC PROCESSES
通过随机过程建立的语言模型
定义
解决生成较长的文本时不连贯的情况
使用Time Control：时间控制
将文档中文本变化的动态映射到感兴趣的随机过程的动态
语言模型可以通过随机过程隐含地生成一个文档计划，然后生成与这个潜在计划一致的文本。
使用对比性目标来学习文本中的潜在动态
对比性目标的2个缺点被改进
产生的句子嵌入往往是静态的：它们捕获了句子邻居之间的语义相似性，但没有捕获句子嵌入如何在文档中演变
并不用于生成，而仅限于话语连贯性等分类任务
LLM: 大型语言模型（LLM）
模型
编码器
将句子映射到布朗桥潜空间
潜空间符合布朗桥密度的分布
公式（1）
不确定性在中间区域最高，而在终点附近较低
目标从一份文件中抽取三联句。从同一文件中抽取的句子构成一个平滑的潜在轨迹：它们应该彼此接近，并遵循潜在空间中的条件密度。从不同文件中抽取的句子不应构成一个平滑的轨迹，而且不太可能遵循桥梁动态。
解码器
训练解码器以从该潜在空间重建句子
在时间t，解码器必须使用过去xᵢ<t中的所有token以及句子嵌入zst来预测xt，zst是编码器的句子嵌入，xt是下一个token
对GPT2进行微调
推理
从时间控制中生成文本
给定两个端点z0,zT，我们从一个潜在的布朗桥中抽出一个轨迹，然后从解码器中生成以这个桥为条件的文本
z0和zT即其实句子和结束句子
如果没有终点zT，使用高斯取样从样本中抽取一个
LANGUAGE MODELING VIA STOCHASTIC PROCESSES
扩散-LM
定义
控制语言模型：Controlling language models
例如控制生成的句子的情感，或添加多个控制，例如生成既是积极情感又是无毒的文本
Diffusion-LM迭代地将高斯向量序列降噪为词向量，产生一串中间潜变量。这些中间变量的连续、分层性质使一个简单的基于梯度的算法能够执行复杂的、可控制的生成任务
连续扩散的语言模型
扩散模型
图像和音频领域
文本领域
离散扩散模型
连续扩散模型
自回归语言模型vs非自回归语言模型 和 即插即用的可控生成
可控生成
可控文本生成的任务是从条件分布p(w|c)中抽取w，其中c表示一个控制变量， 对于情感控制，c可以是一个期望的情感标签，可控生成的目标是生成满足控制目标c的w词
模型
定义一个嵌入函数，将离散的文本映射到一个连续的空间
联合学习扩散模型的参数和词嵌入
前向过程:增加了一个马尔科夫转移从离散的单词w到x0的前向过程
反向过程:softmax分布，参数到单词
公式2
一个四舍五入的方法来将嵌入空间中的向量映射回单词
重新参数化技巧1
好难理解
GLM: General Language Model Pretraining with Autoregressive Blank Infilling
GLM: 自回归空白填充的通用语言模型预训练
定义
可以同时应用于自然语言理解（NLU）、无条件生成和有条件生成在内的任务
对于自然语言理解任务，类似T5的方式，使用提示进行生成答案
最低需要4张3090显卡，约96GB显存
https://github.com/THUDM/GLM-130B
模型（图2c部分）
原句子：x1,x2,x3,x4,x5,x6
随机筛选部分跨度token进行mask，原句位置用Mask表示
筛选跨度长度的方式是使用泊松分布抽样
x1,x2,M,x4,M
M表示被mask的跨度
代表PartA部分
对抽选出的跨度token进行随机排序，跨度的起始位置加入token S
S,x5,x6,S,x3
代表PartB部分
拼接PartA和PartB
x1,x2,Mx4,M,S,x5,x6,S,x3
2D的位置编码
位置1：表示token在原始句子中的位置
位置2：表示Token在跨度中的位置，0表示不在跨度中
输入到GLM模型
自回归的方式生成PartB
PartA的部分可以互相关注
PartB的部分可以关注PartA，和PartB部分已经生成的Token
预测PartB的跨度token
x5,x6,E,x3,E
特殊Token E表示跨度结束
其它部分
模型分支
文档级模型
单一跨度进行采样，其长度从原始长度的50%–100%的均匀分布中抽出，该目标旨在生成长文本。
句子级模型
限制被mask的跨度必须是完整的句子。多个跨度（句子）被取样，以覆盖15%的原始token。这一目标是针对seq2seq任务，其预测往往是完整的句子或段落。
结构
调整层的归一化和残差连接的顺序
单一的线性层进行输出token预测
GeLU替换ReLU激活函数
D2S: Document-to-Slide Generation Via Query-Based Text Summarization
D2S
用于
PPT的自动生成
步骤
用户输入标题，模型提取出章节，句子，表格，图片等
问答模型，对检索到的文本进行文本摘要
系统架构
关键字模块
分层的级联结构
密集向量IR模块
BERT模型
QA模块：文本摘要
BART模型
图提取模块
计算标题和图的标题的相似度
QA模块基准模型对比
BertSummExt
BARTSumm
BARTKeyword
本实验的模型
Evaluating Large Language Models Trained on Code
Copilot: 评估在代码上训练的大型语言模型
定义
类似模型：GPT-Neo，GPT-J，
代码补全工具： Tabnine，Codex
使用测试程序完成的通过率作为评估指标，而不是使用BLEU，BLEU分数可能不是功能正确性的可靠指标
测试代码：https://www.github.com/openai/human-eval