传媒

行业评级 **强大于市（维持评级）**

2024年4月18日

AI搜索：怎么看Kimi的空间？

证券分析师：

杨晓峰 执业证书编号：S0210524020001

研究助理：

马梓燕 ：S0210124030014

一、Kimi核心竞争力：长文本能力

1、通过研究Kimi技术核心基础论文《Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context》和《XLNet: Generalized Autoregressive Pretraining for Language Understanding》，发现Kimi在长文本能力上采用Transformer XL模型，使用分段级循环机制和相对位置编码技术，解决了Transform模型存在的问题；在整理能力方面，XL-Net模型模型结合了置换语言建模机制和两流自注意力机制，提高了推理的效率和准确度。

2、对Kimi、文心一言、通义千问和豆包进行搜索实测对比：**在长文本能力的网络资源搜索方面，应用优缺表现不一，Kimi综合表现较好；在长文本能力的本地资源搜索方面，对比可以处理本地文件的kimi和文心一言，kimi搜索较为准确。**

二、工程化能力比较：Kimi VS Perplexity

和海外AI搜索引擎龙头Perplexity对比：**资料检索能力方面，从资料来源方面、答案整理、推理能力来看，Kimi能力范围约为Perplexity免费版与付费版之间。**同时，Kimi展示出较大进步，在不到一个月时间内，资料来源更多元化。

三、用户空间：Kimi的AI搜索市场需求

1、垂类搜索需求增加，逐渐代替传统搜索引擎；而通过内嵌AI搜索功能或开发AI搜索应用，AI搜索获得青睐。

2、Kimi热度逐渐消失后，开始进入自然增长时期，自然增长仍然强劲。类比海外AI搜索应用Perplexity，kimi显示出APP端增长落后于网页端增长的规律。Kimi正在新一轮的广告投放，将流量导向APP下载，后续广告投放有效性还有待确认。

四、算力支持VS商业模式

阿里云对标中国版“微软云”，积极布局与第三方AI大模型的合作；Kimi获得阿里最新一轮参投，可对标海外“OpenAI+微软”模式。据IT桔子显示，目前阿里已投资Minimax、百川智能、零一万物、智谱AI和Kimi等AI创投公司。

Kimi商业化对标海外Pperplexity：目前perplexity的收入主要来自于会员收入，未来可能会引入广告模式。

建议关注：

一、国产AI应用：

1、AI搜索：昆仑万维；

2、AI陪伴：紫天科技、盛天网络；

3、AI出版 ：中国科传、中信出版、中国出版；

4、AI+IP：中文在线、荣信文化、掌阅科技；

5、AI游戏：宝通科技、恺英网络、 巨人网络、神州泰岳、三七互娱、吉比特、完美世界、姚记科技、星辉娱乐。

二、港股互联网公司的布局

1、大模型公司：腾讯控股、阿里巴巴

2、AI内容平台 ：哔哩哔哩、阅文集团、快手

风险提示：AI竞争激烈，AI发展不及预期

目录

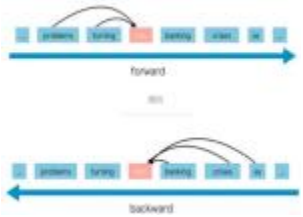
常用的自然语言处理（NLP）技术

Transformer模型

目前最常用的NLP框架

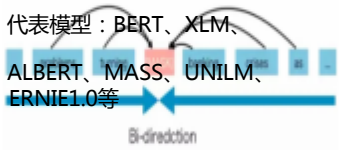
单向特征表示的自回归（AR）预训练语言模型

代表模型：GPT，GPT-2，GPT-3、GPT-J、CTRL等



双向特征表示的自编码（AE）预训练语言模型

代表模型：BERT、XLM、ALBERT、MASS、UNILM、ERNIE1.0等



获取双向信息进行预测，如想要预测位置t的单词，既可以前向获取信息也可以后向获取信息

Kimi模型

Transformer XL模型

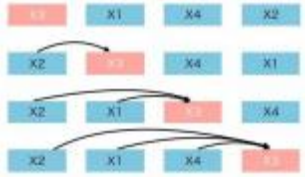
Transformer XL是基于Transformer的神经网络架构，专门用于处理NLP任务中的长文本输入问题，特别是在语言建模和序列建模

利用分段级循环机制和相对位置编码技术，解决了

Transform模型可能涉及到的三个问题：

- 模块之间互相分割
- 时间混乱导致的信息混乱
- 加快模型对同一内容的跑动速度

双向特征表示的自回归预训练语言模型----**XL-Net**



- 结合了自回归（AR）语言建模和自编码（AE）的优点
- XLNet 通过一种称为置换语言建模（Permutation Language Modeling）的机制，使得模型能够学习到双向上下文信息，而不需要依

XLNet 可以通过最大化所有可能的因式分解顺序的对数似然，学习双向语境信息；

用**自回归本身的特点克服BERT 的缺点**；此外，

XLNet 还**融合了当前最优自回归模型 Transformer-XL 的思路。**

资料来源：《Transformer-XL: Attentive Language Models Beyonda Fixed-LengthContext》，戴尔科技集团，aman.ai,华福证券研究所

传统的Transformer受制于输入的长度。这种模式带来了模块割裂、 时间混乱、速度慢的问题 ，Transformer XL模型对此提出了解决办法。

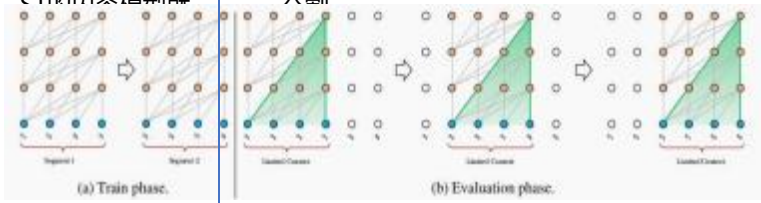
假设一个文本总共有12M的tokens ，我们将其分成三个模块（ S1、 S2、 S3 ），每个部分是4Mtoken。用Transformer去进行模型的跑动 ，模块按照S1、 S2、 S3逐个跑模型。

Transformer

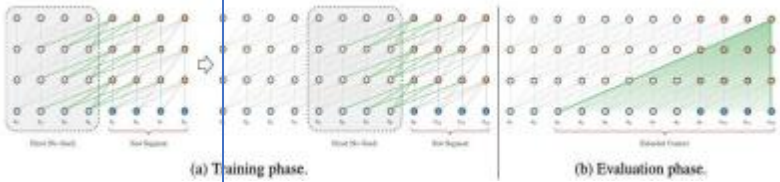
潜在问题

1、每个模块是互相割裂的 ，所以跑S2的时候 ，模块S1的内容模型就

问题1：模块之间互相分割



2、不同模块的跑



3、分三次跑模块的时长较长

问题3：同一内容的跑动速度慢

Transformer XL模型：分段级循环机制和相对位置编码

1、模块之间的互相分割：分段级循环机制（Segment-Level Recurrence），Transformer-XL 在处理新文本段时重用前一个段的隐藏状态，在段与段之间建立了循环连接。这种机制使得模型能够捕捉比固定长度的上下文更长的依赖关系。

将S1模块的信息，通过压缩，将其放置在S2的信息单元中，模型跑动时不仅跑S1的信息，还能跑S2的信息

以此类推，可以将S1、S2、S3模块的所有信息都顺利输入。

2、时间混乱导致的信息混乱：采用相对位置编码（Relative Positional Encoding），为了使状态重用机制有效而不引起时间上的混乱，Transformer-XL引入了相对位置编码，来定义时间偏差，即模型不需要知道每个键向量的绝对位置，而只需要知道它们与查询向量之间的相对距离。

S1模块中的x1、S2模块中的x5以及S3模块的x9都处在模型跑动的第一个位置，在Transformer当中，x5和x1的位置就是同样的。如果所有模块都能跑通的话，会造成大模型并不能很好的识别出顺序。因此，需要给x1、x5和x9去加上一个相对位置，这样就会使得大模型很好的识别出顺序。

3、加快模型对同一内容的跑动速度：由于传统的Transformer跑模型需要分三个模块轮流跑动，所消耗的时间就是3倍的单模 块；但是XL可以重用前一个阶段的表示，不需要从头开始计算，因此大大提升了评估速度。

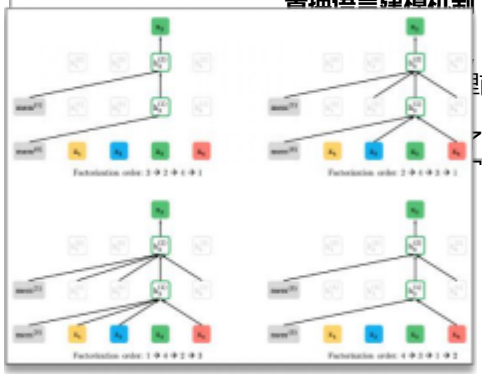
资料来源：《Transformer-XL: Attentive Language Models Beyonda Fixed-LengthContext》，华福证券研究所

XL-Net模型模型结合了置换语言建模机制和两流自注意力机制 ，提高了推理的效率和准确度。置换语言建模机制训练时充分融合了上下文特征，同时也不会造成掩码机制下的有效信息缺失，提高了推理准确度；双流注意力用于置换语言建模机制，需要计算序列的上下文信息（上文信息和下文信息各使用一种注意力机制）在不降低模型的精度的情况下，提高了推理的效率。

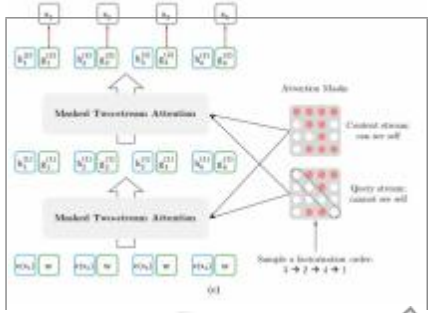
图表 ：置换语言建模机制

BERT等其他模型

在预训练时，采用将需要推理的遮住，使模型结合上文进行推理。一定程度上会导致预训练和微调输入的数据不统一。



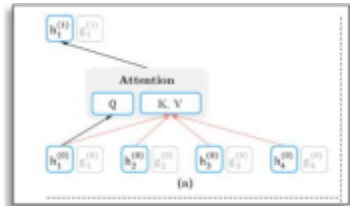
但在预训练时，若要推
理前的内容 ，又可以看到
同时，在微调的时候也
推理准确度的同时，也



两流自注意力机制通过分离内容流（content stream）和查询流（query stream）来完成对目标的遮挡。

因为置换语言建模不会将目标遮挡，但是由于训练推理又不可以直接将需要推理的单词输入，因此引入查询流对要推理的单词进

内容流：进行标准的Transform模型操作



查询流：只保留位置信息，忽略内容信息。使用参数w来代表位置的编码