# RecSys'24 | Meta:使用LLM的摘要能力提升内容推荐

原创 州懂学习笔记 州懂学习笔记 2024年11月18日 00:03 广东

州懂学习笔记
分享大模型推荐系统相关知识和学习笔记
53篇原创内容

公众号

## RecSys'24 | Meta:使用LLM的摘要能力提升内容推荐

> 标题: EmbSum: Leveraging the Summarization Capabilities of Large Language Models for Content-Based Recommendations
> 地址: https://arxiv.org/pdf/2405.11441
> 机构: UBC & Meta
> 会议: RecSys'24

## 1. 前言

内容推荐系统(如新闻推荐)的一个核心问题是如何基于用户长期的交互行为历史, 去为用户提供更符合用户兴趣的个性化内容。随着近年来大语言模型的发展, 越来越多的方法尝试使用LLM来提升内容推荐。

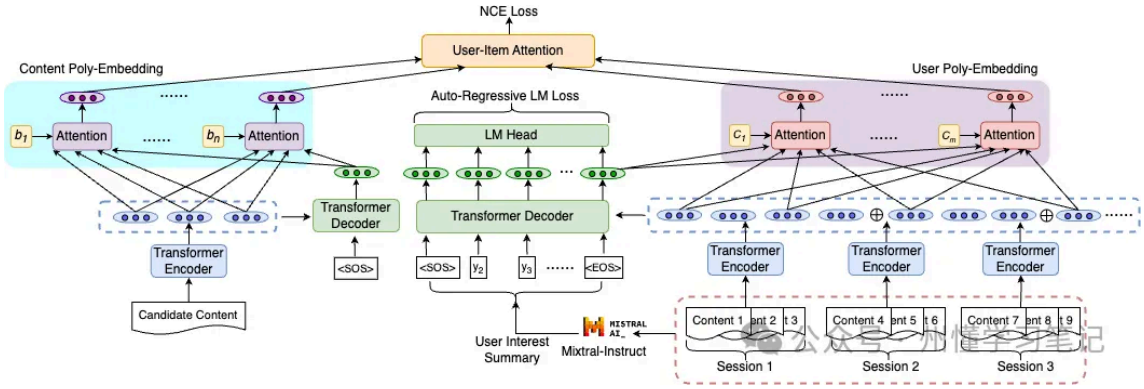
以往使用预训练语言模型做内容推荐的方法, 主要有两个弊端:

- **对用户历史行为内容的用户兴趣捕获能力较弱:** 以往的方法主要是对用户行为的每个内容独立编码后, 再做聚合, 这种方式捕获用户兴趣的能力较弱。
- **难以满足时延的要求:** 也有一些方法将Candidate Item也整合到用户序列建模中, 以更好地对齐用户与Candidate Item, 但这种cross-encoder的方式时延比较高。

为此, 作者提出了新的框架EmbSum, 通过划分session使模型得以处理较长的用户交互序列, 并可离线预先计算用户和候选物品，同时捕捉用户参与历史中的互动。下面详细介绍。

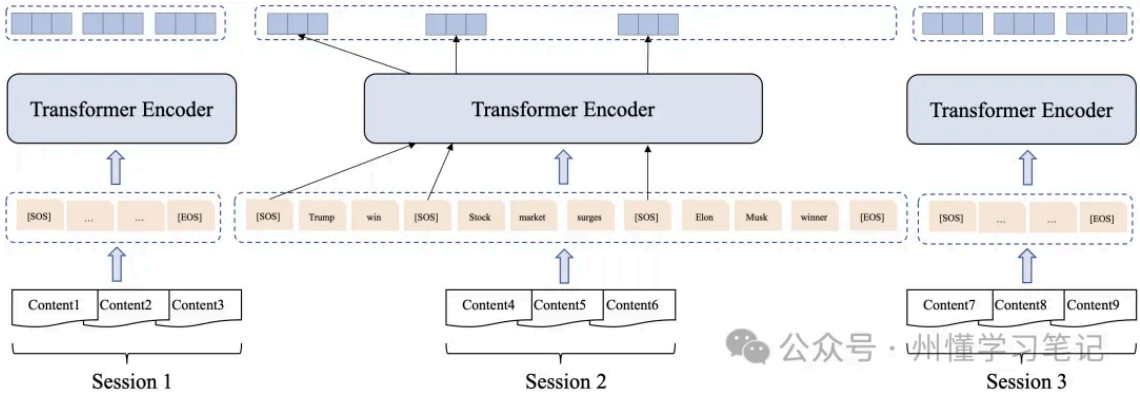
## 2. 方法

作者所提方法EmbSum的整体框架如下图所示:



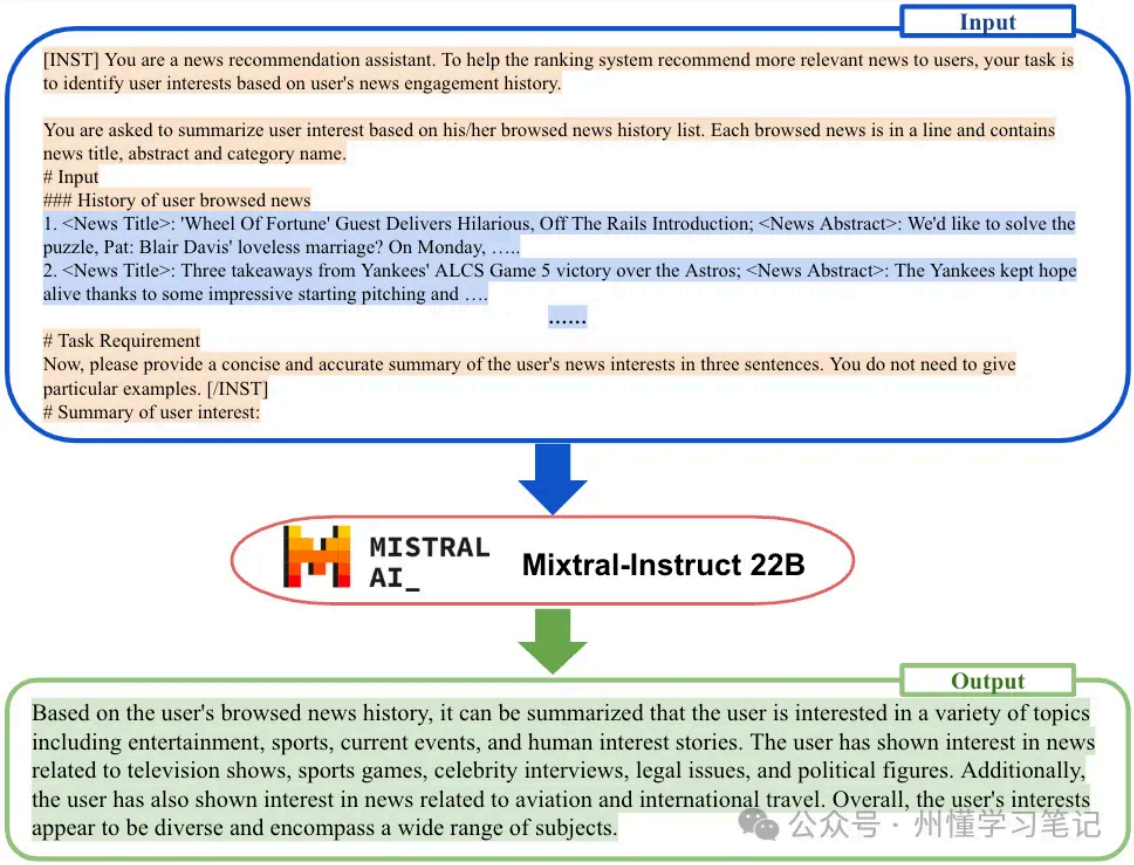

## 2.1 User Session Encoding

对于用户历史交互的行为序列$E_{u_i} = \{e_1, e_2, \ldots, e_k\}$, 长度记为$k$, 作者将其按时间顺序划分成若干个Session $E_{u_i} = \{\eta_1, \eta_2, \ldots, \eta_g\}$, 这里每个Session最多包含$p$个内容$\eta_i = \{e_1, e_2, \ldots, e_p\}$, 每个Session再单独过Transformer Encoder进行编码, 这里作者使用T5-small作为backbone。然后, 作者把每个文本内容开始符号[SOS]所对应的隐层状态作为该文本内容的表征。整体流程如下图所示:

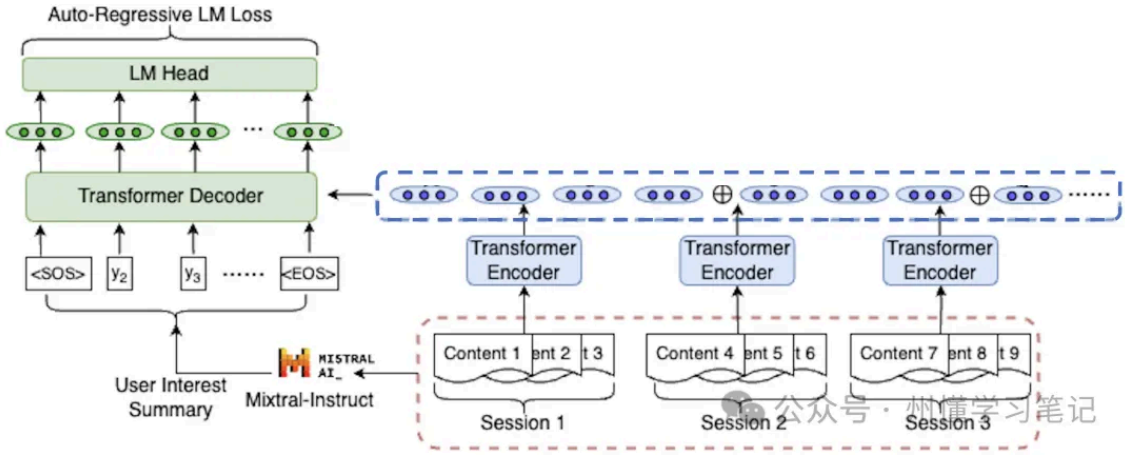



## 2.2 User Engagement Summarization

为了更好的理解&捕获用户兴趣, 作者利用LLM的摘要能力辅助训练。具体地, 首先, 作者使用Mixtral-8x22B-Instruct, 输入用户历史交互内容, 生成一些用户兴趣摘要。这里, 作者给了个求例, 如下图所示:




大模型生成的这些用户兴趣摘要, 会与前面用户Session编码环节Encoder所有token隐含状态一起输入到T5 decoder中, 如下图所示:



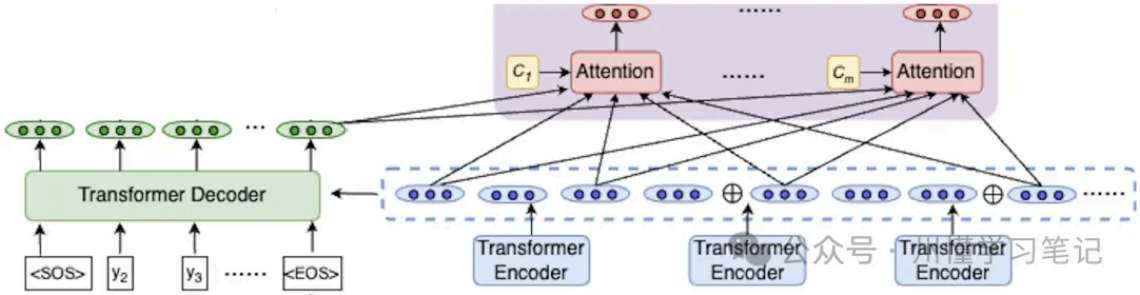

然后再使用自回归训练方式进行训练:

$$\mathcal{L}_{\text{sum}} = -\sum_{j=1}^{|y_j^{u_i}|} log(p(y_j^{u_i}|E, y_{<j}^{u_i}))$$

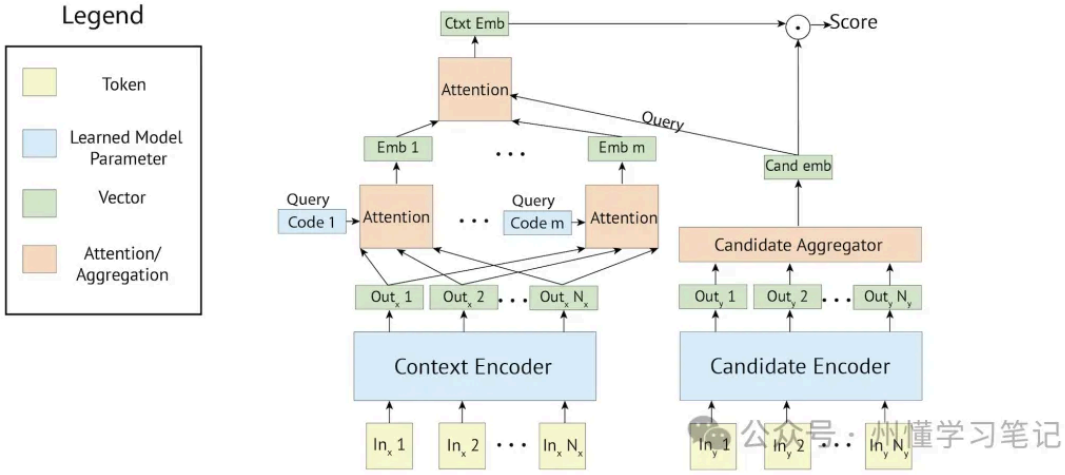
其中, $E$表示前面编码后各token的隐含状态。

2.3 User Poly-Embedding

前面User Session Encoding得到的$k$个$d$维表征是每个Session独立编码的, 还缺少对用户历史内容的全局考虑。为此, 作者将前面自回归训练过程中最后一个Token所对应的$d$维隐含状态作为用户交互内容的全局表征, 再与$k$个$d$维Concat起来得到$Z \in \mathrm{R}^{(k+1)\times d}$的矩阵, 再做poly-attention，如下图所示:




关于poly-attention后,可以参考poly-encoder的原论文的框架图



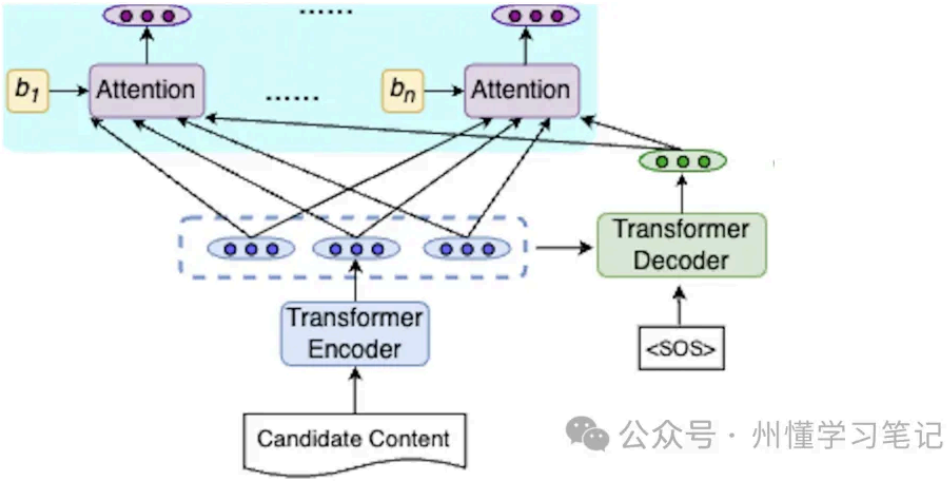

每个用户向量(即对应于上图中的Emb 1 ~ Emb m)的形式化描述如下:

$$\alpha_a = \text{softmax}\left[c_a \tanh(ZW^f)^\top\right]Z$$

其中, $c_a \in \mathrm{R}^{1\times p}$和$W^f \in \mathrm{R}^{d\times p}$为可学习的参数。然后, 再将这$m$个向量拼接起来得到$A \in \mathrm{R}^{m\times d}$, 作为最终的用户表征，论文这里称它们为User Poly-Embedding(UPE)。

2.4 Candidate Content Modeling

在前面Session编码中, 作者直接使用每个文本内容开始符号[SOS]所对应的隐含状态来作为相应Item的表征, 但这里对于候选Item, 作者并没有这么去做, 而是将该候选Item内容过Encoder后的所有隐含状态也像User Poly-Embedding类似处理, 如下图所示:




在这个过程中, 可以得到的$n$个表征, 将它们拼接起来就可以得到矩阵$B \in \mathrm{R}^{n\times d}$, 这个矩阵就作为最终的候选Item的内容表征, 论文这里称它们为Content Poly-Embedding (CPE)

2.5 预测与训练

2.5.1 CTR预测

对于用户$i$和内容$j$, 在前面得到的用户poly embed $A_i$和内容poly embed $B_j$后, 作者使用下面方式预估相关性得分$s_j^i$。

首先, 将用户表征和候选内容表征相乘后再做flatten:

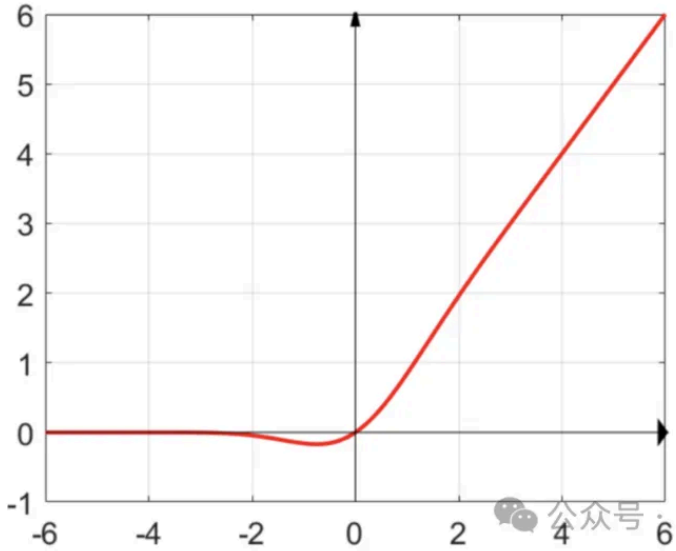
$$K_j^i = \text{flatten}(A_i^\top B_j)$$

这样就得到$K_j^i \in R^{mn}$的向量。然后再做attention:

$$W^p = \text{softmax}(\text{flatten}(A \cdot \text{gelu}(BW^s)^\top))$$

$$s_j^i = W^p \cdot K_j^i$$

这里, $W^s$为可学习参数, 此外, 作者使用了gelu激活函数, 如下图所示, 不过多阐述:






2.5.2 训练

训练上, 作者使用了NCE Loss:

$$\mathcal{L}_{\text{NCE}} = -\log\left(\frac{\exp(s_+^i)}{\exp(s_+^i) + \sum_j \exp(s_{-,j}^i)}\right)$$

其中, $s_+^i$表示正样本, $s^i_$表示负样本。

然后, 最终整体的Loss为:

$$\mathcal{L} = \mathcal{L}_{\text{NCE}} + \lambda \mathcal{L}_{\text{sum}}$$

其中, λ为超参, 作者设置为0.05

## 3. 实验部分

### 3.1 整体效果

| | MIND | | | | Goodreads | | | |
|---|---|---|---|---|---|---|---|---|
| | AUC | MRR | nDCG@5 | nDCG@10 | AUC | MRR | nDCG@5 | nDCG@10 |
| NAML | 66.10 | 34.65 | 32.80 | 39.14 | 59.35 | 72.16 | 53.49 | 67.81 |
| NRMS | 63.28 | 33.10 | 31.50 | 37.68 | 60.51 | 72.15 | 53.69 | 68.03 |
| Fastformer | 66.32 | 34.75 | 33.03 | 39.30 | 59.39 | 71.11 | 52.38 | 67.05 |
| CAUM | 62.56 | 34.40 | 32.88 | 38.90 | 55.13 | 73.06 | **54.97** | 69.02 |
| MINS | 61.43 | 35.99 | 34.13 | 40.54 | 53.02 | 71.81 | 53.72 | 68.00 |
| NAML-PLM | 67.01 | 35.67 | 34.10 | 40.32 | 59.57 | 72.54 | 53.98 | 68.41 |
| UNBERT | 71.73 | 38.06 | 36.67 | 42.92 | 61.40 | 73.34 | 54.67 | 68.71 |
| MINER | 70.20 | 38.10 | 36.35 | 42.63 | 60.72 | 72.72 | 54.17 | 68.42 |
| UniTRec | 69.38 | 37.62 | 36.01 | 42.20 | 60.00 | 72.60 | 53.73 | 67.96 |
| EmbSum (ours) | **71.95** | **38.58** | **36.75** | **42.97** | **61.64** | **73.75** | 54.86 | **69.08** |



### 3.2 消融实验

| | AUC | MRR | nDCG@5 | nDCG@10 |
|---|---|---|---|---|
| **MIND** | | | | |
| Ours | **71.95** | 38.58 | **36.75** | **42.97** |
| wo CPE | 68.17 | 36.49 | 33.72 | 40.20 |
| wo grouping | 71.34 | 38.29 | 36.41 | 42.55 |
| wo UPE | 71.41 | **38.64** | 36.70 | 42.90 |
| wo $\mathcal{L}_{sum}$ | 71.43 | 38.42 | 36.39 | 42.60 |
| **Goodreads** | | | | |
| Ours | **61.64** | **73.75** | **54.86** | **69.08** |
| wo CPE | 60.97 | 72.94 | 54.39 | 68.53 |
| wo grouping | 61.39 | 73.53 | 54.79 | 68.86 |
| wo UPE | 61.35 | 73.55 | 54.67 | 68.81 |
| wo $\mathcal{L}_{sum}$ | 61.50 | 73.55 | 54.74 | 68.91 |



## 3.3 超参影响