【LLM论文阅读】DMPO 与 S-DPO：LLM推荐排序模型的直接偏好优化研究

原创 方方 方方的算法花园 2024年10月25日 10:31 北京

写在前面

如今在推荐系统领域，创新的技术和方法不断出现，致力于增强推荐的准确性和效果。随着研究的开展，人们逐渐察觉到语言模型的建模损失在推荐排序任务中不太适应。DMPO 和 S-DPO 便是致力于解决这一问题的两种直接偏好优化方法，它们的论文发表时间颇为相近，一篇是 2024 年 5 月，另一篇是 6 月，两者思路虽有相似之处，但并非完全一样。本文会先分别对这两篇论文进行简要解读，最后在文末对这两种技术进行简要对比。

论文概况

1 DMPO：直接多偏好优化

1. 论文名称：Finetuning Large Language Model for Personalized Ranking 微调大型语言模型以实现个性化排序
2. 论文链接：https://arxiv.org/pdf/2405.16127
3. Github：https://github.com/BZX667/DMPO
4. 论文作者所在机构：Cashcat、北航、上海大学、复旦大学等
5. 一句话概括：直接多偏好优化（DMPO， Direct Multi-Preference Optimization）框架同时最大化正样本的概率和最小化多个负样本的概率。

2 S-DPO：Softmax-DPO

1. 论文名称：On Softmax Direct Preference Optimization for Recommendation 关于推荐的 Softmax 直接偏好优化
2. 论文链接：https://arxiv.org/pdf/2406.09215v2
3. Github：https://github.com/chenyuxin1999/S-DPO
4. 论文作者所在机构：新加坡国立大学、中国科学技术大学、北海道大学
5. 一句话概括：基于 LM 的推荐器量身定制的直接偏好优化，将排名信息注入到 LM 中，以帮助基于 LM 的推荐器区分偏好项和负项。

DMPO 论文解读

1 DMPO 论文创新点

1. 提出 DMPO 框架
为了缩小大型语言模型（LLMs）与推荐任务之间的差距，提出直接多偏好优化（DMPO）框架。该框架通过**同时最大化正样本的概率和最小化多个负样本的概率，增强了 LLMs 对推荐任务中正负样本比较关系的建模能力。**

2. 性能显著提升

在三个真实世界的公共数据集（“Movielens - 1M”、“Amazon Movies and TV” 和 “Amazon Video Games”）上的少样本场景中，DMPO 显著优于以前基于 LLM 的方法和传统方法，提高了推荐系统的性能。

3. 具有泛化能力

DMPO 在跨域推荐中表现出优越的泛化能力，通过在 “Amazon Movies and TV” 和 “Amazon Video Games” 之间进行跨域实验，结果表明 DMPO 相比基于 LLM 的跨域方法和传统跨域方法有显著改进。

4. 可解释性

DMPO 是一个可解释的推荐系统。模型通过计算候选中每个token的生成概率，并为重要token分配更高的概率，从而解释了为什么选择某些候选进行项目推荐。

2 DMPO 主要方案

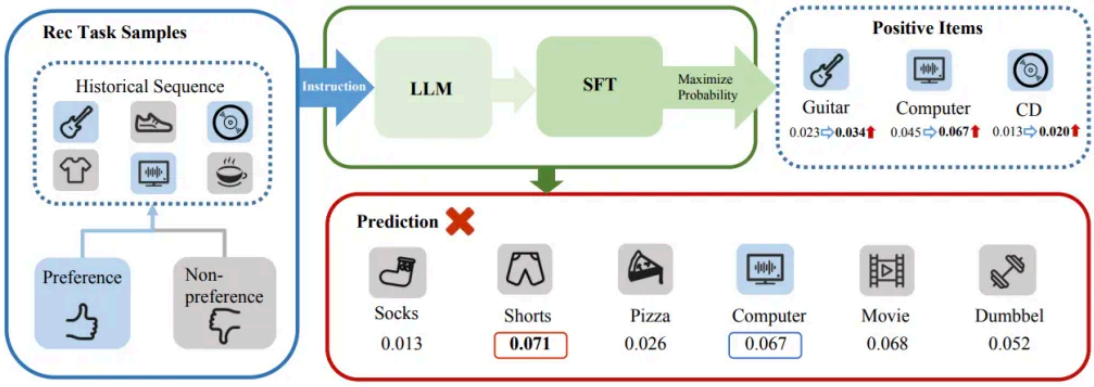

Figure 1: When aligning the LLMs with recommendation tasks, applying Supervised Fine-Tuning (SFT) to LLM initially helps maximize the probability of generating each token in positive items. However, this method overlooks the potential benefits of comparing positive and multiple negative samples. As a result, although the model trained solely with SFT may improve the probability of positive items during training, it may overestimate the probability of unseen negative items during testing, leading to incorrect prediction.

SFT 的作用及局限

（1）初始帮助

对 LLM 应用监督微调（SFT）最初有助于最大化正项中每个token的生成概率。例如，从图中的数据来看，像 “Guitar” 等正项在经过 SFT 训练时，其相关概率值可能会有所提升。

（2）存在局限

然而，这种方法忽略了比较正样本和多个负样本的潜在好处。当模型仅用 SFT 训练时，虽然在训练过程中可能提高了正项的概率，但在测试过程中可能会高估未见过的负项的概率，从而导致错误预测。从图中给出的预测结果可以看到，对于一些负样本如 “Computer Movie Dumbbel”“Socks Shorts Pizza” 等，模型可能会做出不准确的预测。

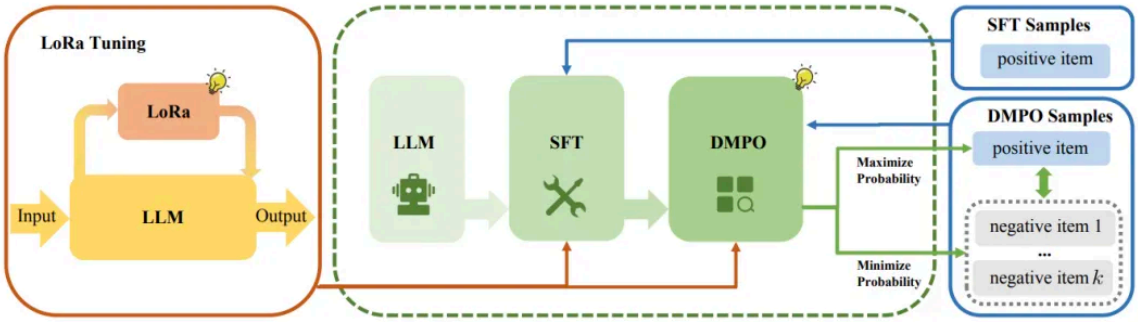

Figure 2: We first performed SFT on the base LLM model and then proceeded with DMPO. SFT samples and DMPO samples were constructed as inputs for training. Both SFT and DMPO were trained using LoRA. DMPO aims to maximize the probability of positive samples while minimizing the probability of multiple negative samples simultaneously.

DMPO 的训练过程

（1）SFT 阶段

首先在基础的大型语言模型（LLM）上开展监督微调（SFT）。这一步骤是整个优化过程的起始点，为后续的 DMPO 操作奠定基础。SFT 的核心目标在于最大化正确答案里每个token的概率。通过这种方式，使模型能够更好地适应推荐任务的基本要求，初步学习到正确答案的一些特征和模式。

（2）DMPO 阶段

- **样本构建**：在 SFT 的基础上进行 DMPO。此过程中构建了 SFT 样本和 DMPO 样本用作训练输入。这些样本包含了与推荐任务相关的各种信息，如用户的历史交互记录、正样本和负样本等。
- DMPO 同样使用 LoRA 进行训练。
- DMPO 同时最大化正样本的概率并最小化多个负样本的概率。这一目标与推荐任务的本质紧密相关，通过这种方式，模型能够更准确地区分用户喜欢和不喜欢的物品，从而提高推荐的准确性。例如，在面对用户的历史交互数据和一系列候选物品时，DMPO 能够更好地判断哪些物品更有可能被用户喜欢（正样本），哪些不太可能被喜欢（负样本），进而优化推荐结果。

Table 1: An instruction tuning sample for DMPO. "Roman Holiday" is the positive item, and the other movies serve as negative items. The order of positive and negative samples placed in the instruction input is randomly altered across different samples to prevent the LLMs from making predictions based on the input sequence.

Task Input	
Instruction:	You are an assistant working on movie recommendations. Here is the user's history of movies they have watched: <Waterloo Bridge>, <Rear Window>, <Forrest Gump>. Rank the likelihood of the user watching the two movies (1) <Roman Holiday> and <Iron Man> (2) <Harry Potter> and <Roman Holiday> (3) <Roman Holiday> and <Spider Man> (4) <The Lion King> and <Roman Holiday>
Task Output	
DMPO Label:	(1) ["<Roman Holiday>, <Iron Man>", "<Iron Man>, <Roman Holiday>"]. (2) ["<Roman Holiday>, <Harry Potter>", "<Harry Potter>, <Roman Holiday>"]. (3) ["<Roman Holiday>, <Spider Man>", "<Spider Man>, <Roman Holiday>"]. (4) ["<Roman Holiday>, <The Lion King>", "<The Lion King>, <Roman Holiday>"].

S-DPO 论文解读

1 S-DPO 论文创新点

1. 指出问题

率先指出基于语言模型（LM）的推荐系统中广泛使用的语言建模损失不适用于排名任务，且未能充分利用用户偏好数据，从而阻碍了推荐性能。

2. 提出 S-DPO 损失函数

提出了 Softmax-DPO（S-DPO），它是一种为基于 LM 的推荐器量身定制的直接偏好优化（DPO）损失的替代版本。该函数在用户偏好数据中融入了多个负样本，将排名信息注入到 LM 中，以帮助基于 LM 的推荐器区分偏好项和负项。

3. 理论联系与挖掘硬负例能力

从理论上建立了 S-DPO 与负采样的 softmax 损失之间的联系，凸显了多个负样本的关键作用，并发现其具有挖掘硬负例的副作用。这不仅提升了性能，还加速了训练过程，确保了其在推荐任务中的卓越能力。

4. 实验验证有效

通过在三个真实世界数据集上进行的大量实验，证明了 S-DPO 在有效建模用户偏好、提高推荐性能以及缓解 DPO 的数据似然下降问题方面的优越性。

2 S-DPO 主要方案

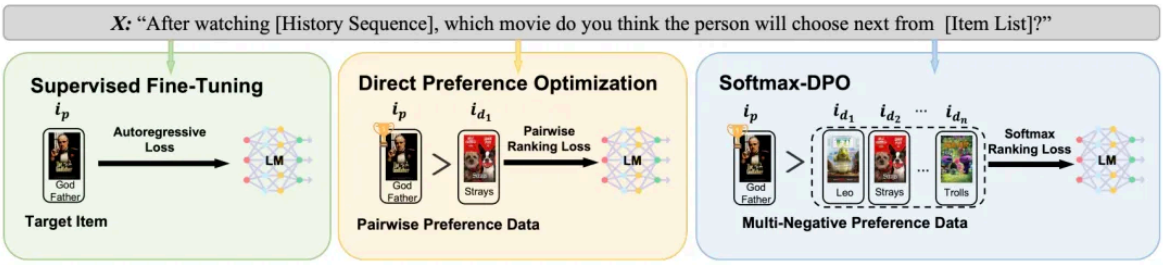

Figure 1: Framework of S-DPO. Different from existing methods which fine-tune LMs with a language modeling loss without tailoring for recommendations, S-DPO proposes to explicitly instill ranking information into LMs. To take one step further, S-DPO incorporates multiple negatives in user preference data and generalizes pairwise DPO loss to softmax ranking loss.

与现有方法对比

（1）现有方法的局限性：现有方法在使用语言模型（LMs）进行推荐时，通常使用语言建模损失进行微调，而没有针对推荐任务进行专门定制。这种方式没有充分考虑用户的偏好数据，特别是没有有效地利用负样本信息来优化模型以更好地进行个性化排名。

（2）S-DPO 的创新点：S-DPO 提出了一种不同的方法，明确地将排名信息注入到 LMs 中。它不仅仅关注正样本，而是通过考虑多个负样本来更好地理解用户的偏好，从而提高推荐系统的性能。

S-DPO 的两个主要阶段

（1）监督微调（Supervised Fine - Tuning）：这是 S-DPO 的第一步，其目的是注入领域知识并提高 LM 遵循指令的能力。在此阶段，推荐任务的数据被构建为基于文本的对，然后基于这些对微调 LMs。例如，对于用户 u，推荐任务提示x_{u}包含用户的历史交互H_{u}、候选项目集 c 以及顺序推荐任务的描述，这个提示x_{u}与候选集中偏好项目i_{p}的标题e_{p}配对形成对数据(x_{u}, e_{p})，然后利用这些对数据通过语言建模损失来微调基于 LM 的推荐器。

（2）偏好对齐（Direct Preference Optimization Softmax - DPO）：

构建偏好数据：在这个阶段，与现有方法只构建正样本对不同，S-DPO 为每个语言提示同时配对正样本和随机采样的多个负样本，以此构建基于文本的偏好数据。例如，对于一个推荐任务提示，不仅有偏好的项目，还有多个不被偏好的项目与之相关联，形成更全面的偏好信息表示。

扩展 DPO 损失：基于这些构建好的偏好数据，S-DPO 将传统基于成对数据的 DPO（Direct Preference Optimization）损失与 Bradley - Terry 偏好模型扩展到 Plackett - Luce 偏好模型。Plackett - Luce 偏好模型能够更好地处理推荐任务中的相对排名问题，从而将原本针对成对偏好的 DPO 损失泛化到能够处理多个项目相对排名的 softmax 排名损失，使得模型能够更好地学习到用户的偏好顺序，进而提高推荐的准确性。

DMPO 与 S-DPO对比

二者的不同从损失函数中可以看到：

$$\mathcal{L}_{\mathrm{S-DPO}}(\pi_\theta;\pi_{\mathrm{ref}}) = -\mathbb{E}_{(x_u,e_p,\mathcal{E}_d)\sim\mathcal{D}}\left[\log\sigma\left(-\log\sum_{e_d\in\mathcal{E}_d}\exp\left(\beta\log\frac{\pi_\theta(e_d|x_u)}{\pi_{\mathrm{ref}}(e_d|x_u)}-\beta\log\frac{\pi_\theta(e_p|x_u)}{\pi_{\mathrm{ref}}(e_p|x_u)}\right)\right)\right] \tag{11}$$

$$\mathcal{L}_{\mathrm{DMPO}}(\pi_\theta;\pi_{ref}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}$$
$$\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w\mid x)}{\pi_{ref}(y_w\mid x)}-\frac{1}{k}\sum_{i=1}^{k}(\beta\log\frac{\pi_\theta(y_{l\,i}\mid x_i)}{\pi_{ref}(y_{l\,i}\mid x_i))}\right)\right]$$

- S-DPO 对多个负样本进行了综合性的处理，具体的操作步骤为先计算正样本与负样本之间的对数概率差异，之后对所有的负样本进行求和，再对求和的结果取指数，最后再取对数等一系列复杂的操作来对损失进行衡量；
- DMPO 则是先计算正样本的对数概率差异，然后计算多个负样本的对数概率差异的平均值，最后将计算得到的正样本对数概率差异与负样本对数概率差异的平均值相减，以此来衡量损失。

整体看来，二者在对 DPO 进行多负样本的改进方面都采取了相应的措施。DMPO 将原始的 DPO 损失中的负样本部分进行了简单的替换，即采用多个负样本取平均值的方式。而 S-DPO 则采用了更为通用的 Plackett-Luce 模型来对偏好分布进行建模，并且还进行了详细的理论推导。最终，将 S-DPO 损失与 Sampled Softmax 建立起了紧密的联系。

语言模型 5　推荐系统 4　LLM与推荐 15　LLM论文阅读 13　#LLM学习 12

语言模型 · 目录

下一篇 · 【LLM论文阅读】LLM4Rerank-多维度重排框架