Modeling and Simulation 建模与仿真, 2025, 14(1), 116-121
Published Online January 2025 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2025.141012

Hans汉斯

基于排序损失的跨模态检索优化研究

蒋彩兰

上海理工大学管理学院，上海

收稿日期：2024年12月8日；录用日期：2025年1月1日；发布日期：2025年1月9日

摘　要

跨模态检索通过一种模态(如文本或图像)来检索另一模态的数据，传统的跨模态检索方法主要依赖模态对齐与相似性度量，以实现多模态间的特征匹配。本文创新性地提出了一种基于排序的跨模态检索方法，通过引入排序损失来优化跨模态检索过程，使得与查询相关性高的项目在结果中排名靠前，从而实现跨模态检索。实验结果表明，引入排序损失可显著提升跨模态检索性能，尤其在文本与图像匹配中表现出色，为后续研究提供了新的方法视角和坚实的技术基础。

关键词

跨模态检索，排序损失，相似性度量

Research on Optimization of Cross-Modal Retrieval Based on Ranking Loss

Cailan Jiang

Business School, University of Shanghai for Science and Technology, Shanghai

Received: Dec. 8th, 2024; accepted: Jan. 1st, 2025; published: Jan. 9th, 2025

Abstract

Cross-modal retrieval aims to retrieve data in one modality (such as text or images) based on another modality. Traditional cross-modal retrieval methods primarily rely on modality alignment and similarity measures to achieve feature matching across multiple modalities. This paper presents an innovative sorting-based cross-modal retrieval method that optimizes the cross-modal retrieval process by introducing ranking loss, allowing items with higher relevance to the query to be prioritized in the results, thereby enhancing cross-modal retrieval effectiveness. Experimental results demonstrate that the introduction of ranking loss significantly enhances the performance of cross-modal retrieval, particularly excelling in text-image matching tasks. This work provides a

文章引用: 蒋彩兰. 基于排序损失的跨模态检索优化研究[J]. 建模与仿真, 2025, 14(1): 116-121.
DOI: 10.12677/mos.2025.141012

new methodological perspective and a solid technical foundation for future research in the field.

Keywords

Cross-Modal Retrieval, Ranking Loss, Similarity Measure

1. 引言

跨模态检索[1]是一种信息检索技术，旨在通过一种模态(如文本、图像、音频等)来检索另一种模态的数据。例如，用户可以通过输入文本描述来检索相关的图像，或者通过图像来找到相关的文本内容。著名的跨模态模型有 CLIP [2]、DCCA [3]和 CMPM [4]等。CLIP (Contrastive Language-Image Pretraining) [2]是 OpenAI 提出的一个基于对比学习的跨模态模型，利用大规模图像和文本数据的对比信息来实现模态间的对齐，使得文本描述和对应图像的特征在共享空间中紧密相连。DCCA (Deep Canonical Correlation Analysis) [3]则通过深度学习方法对不同模态数据进行对齐，以最大化模态间的相关性，从而在跨模态检索任务中实现有效匹配。CMPM (Cross-Modal Projection Matching) [4]通过将不同模态的数据映射到公共特征空间，以提高跨模态检索的精度和鲁棒性，尤其适用于图像和文本的检索任务。

大多数跨模态检索方法关注的是通过模态对齐和特征相似性来进行匹配。通常是将不同模态的数据投影到共同的表示空间，常用余弦相似度[5]或欧式距离[6]来衡量文本和图像的相似性，希望最大化相关文本–图像对的相似度来使相关的特征向量更接近，从而在查询时通过计算相似度来实现检索。

本文提出了一种基于排序损失的优化方法，用于提高跨模态检索的效果。与传统的基于相似度度量(如余弦相似度)的方法不同，本文使用排序损失来替代相似性度量，使用排序损失来优化不同模态之间的相对排序关系。具体而言，我们将文本和图像表示映射到一个共享的特征空间，并使用排序来驱动模型学习。在训练过程中，通过最小化排序损失，使模型更倾向于将与查询(文本/图像)语义上更接近的候选项(图像/文本)排在前面，从而有效优化跨模态检索的精确度。这种基于排序的优化策略能够克服传统方法中对相似度计算的依赖，从而提升跨模态检索的性能。

2. 基于排序损失的跨模态检索构架

传统的跨模态检索方法通常基于相似度度量，例如余弦相似度或欧几里得距离，将图像和文本映射到一个共享的特征空间中，然后计算它们的相似性。为了提高跨模态检索的效果，通常会利用对比学习方法，通过最大化相关模态对的相似度，最小化无关模态对的相似度。

与传统的跨模态方法不同，本文通过最小化的排序损失进行跨模态检索优化，使用排序损失替代相似性度量，从而实现跨模态检索的优化。本文的思路是训练一个排序模型，使得图片–文本配对的相关性分数符合真实的相关性标签，从而提高跨模态检索的排序效果。本文提出的跨模态检索构架如图 1 所示。

给定一个图像 I 和 n 个文本描述 $\{T_i\}_{i=1}^{n}$，每个图像–文本对 (I,T_i) 之间有相关性标签 $s_i \in \{0,1\}$，其中 $s_i=1$ 表示相关，$s_i=0$ 表示不相关。

使用映射函数 $g(\cdot)$ 和 $h(\cdot)$ 分别将图像和文本嵌入到共享表示空间，得到图像嵌入和文本嵌入：

$$h_{img}=f(I),\ h_{text}=g(T_i),\ i=1,2,\cdots,n \tag{1}$$

将图像嵌入 h_{img} 与每个文本嵌入 h_{text} 进行特征拼接，得到联合特征向量：

$$v_i = \left[h_{img}; h_{text} \right],\ i = 1,2,\cdots,n \tag{2}$$

其中[·;·]表示向量拼接操作。

将拼接特征 v_i 输入排序模型 f_θ，得到图像与每个文本的预测相关性分数 $\hat{s}_i$：

$$\hat{s}_i = f_\theta\left(v_i\right),\ i = 1,2,\cdots,n \tag{3}$$

为了优化跨模态检索，本文引入了排序损失函数 $\mathcal{L}$，通过最小化预测相关性分数 $\hat{s}_i$ 与真实相关性标签 s_i 之间的差距，确保相关文本的分数高于不相关文本，从而提升排序模型的准确性：

$$\mathcal{L} = \sum_{i=i}^{n} \ell\left(\hat{s}_i, s_i\right) \tag{4}$$

通过最小化损失函数，不断更新参数 θ，排序模型 f_θ 能够逐步学习到更精确的跨模态相关性关系。最终，该方法通过优化排序，使得与查询更相关的项目在检索结果中排名靠前，提升跨模态检索的性能。

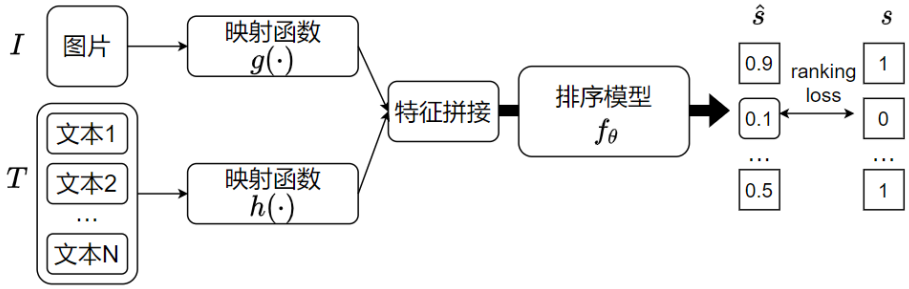

Figure 1. Cross-modal retrieval framework based on ranking loss
图 1. 基于排序损失的跨模态检索构架

3. 映射函数

对于本文使用的文本和图像的映射函数，本文统一采用 CLIP [2]，以实现图像和文本的高效嵌入。使用 CLIP 对图像和文本进行嵌入表示如图 2 所示。CLIP [2]模型通过对比学习的方式，训练出一个共享的表示空间，使得图像和文本可以在同一特征空间中进行有效的对比。

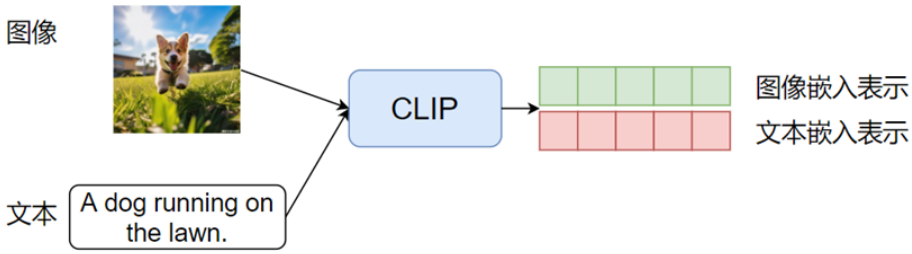

Figure 2. Mapping images and text to the embedding representation space using CLIP
图 2. 使用 CLIP 将图像和文本映射到嵌入表示空间

CLIP [2]包含两个主要组件：图像编码器和文本编码器。图像编码器将输入图像 I 经过预处理后，使用卷积神经网络(CNN)或视觉变换器(ViT)将其嵌入到共享特征空间。该编码器能够提取图像的高层次特征，如颜色、形状和纹理等。文本编码器将输入的文本描述 T_i 进行嵌入，通常使用 Transformer 架构[7]。文本编码器能够理解文本的上下文和语义信息，从而将其映射到同一特征空间。

4. 排序模型

本文选择了 Context-Aware Ranker [8]作为排序模型。Context-Aware Ranker 是一类考虑上下文信息的排序模型，常用于需要复杂排序的任务中，特别是信息检索、推荐系统等领域。传统的排序模型往往只依赖于查询和目标项之间的相关性评分，而 Context-Aware Ranker 则会额外利用上下文信息(如用户行为、历史记录、时间等)来优化排序效果，以便更好地预测用户的真实意图。

在具体实现上，Context-Aware Ranker 可以采用不同的架构。常见的方法包括：

1) 序列模型：利用 RNN、LSTM、Transformer 等模型捕捉用户交互序列中的上下文信息。

2) 特征增强：通过将用户行为、历史偏好等上下文特征引入模型输入，增强传统的排序模型。

3) 多任务学习：将上下文信息作为一个辅助任务，提升主任务的效果。

这些方法都能帮助模型更精准地理解用户需求，从而提升排序质量。在信息检索领域，Context-Aware Ranker 通常通过提高 NDCG、Recall 等指标来证明其有效性。

5. 实验

5.1. 数据集

跨模态检索实验使用的数据集是 Flickr30k [9]。该数据集包含 31,000 张从 Flickr 网站收集的图像，每张图像由人工注释者提供五个文本描述。

本文实验设计了两个任务：文本检索图像和图像检索文本。利用 CLIP [2]将 Flickr30K [9]数据集中的图像和文本转换为嵌入向量，然后进行难例搜索，即为每个查询(图像或文本)随机选取 200 个候选项，根据查询与候选项的余弦相似性，挑选出与查询最相似的 50 个样本，最终得到 51 个候选项作为检索目标。每个图像(文本) – 文本(图像)对被标记为相关(1)或不相关(0)。

在数据划分方面，使用 1000 个样本用于验证，1000 个样本用于测试，其余用于训练。表 1 列出了 Flickr30K 数据集的统计信息。

Table 1. Dataset statistics on Flickr30K
表 1. Flickr30K 数据集的统计信息

数据集	Queries	LabeledObjs	Features	LabelLevls
训练集	29,000	609,000	1024	2
验证集	1000	21000	1024	2
测试集	1000	21000	1024	2

5.2. 实验设置

本文选用了跨模态检索模型 CLIP [2]、DCCA [3]和 CMPM [4]作为基线方法。同时，还采用了四种排序损失的方法，包括 LambdaRank [10]、NeuralNDCG [11]、ListNet [12]和 RankNet [13]。LambdaRank 是一种基于梯度的排序学习方法，它由 RankNet 演化而来，但引入了“lambda”梯度的概念。NeuralNDCG 是一种基于神经网络的排序模型，专门设计用于直接优化 NDCG (Normalized Discounted Cumulative Gain)。ListNet 是一种列表级排序学习方法，它将整个结果列表视作训练单元，而不是仅仅关注对或单项。RankNet 是基于神经网络的对比排序模型，通过学习成对样本的相对排序来优化模型。这四个方法基于 CLIP [2]的预训练嵌入进行排序，以实现跨模态检索。

根据网格搜索的结果，本文配置了排序模型 Context-Aware Ranker 的超参数。具体而言，初始全连接

层的维度设置为 64，编码器层数为 2，每个编码器配备 4 个注意力头，编码器隐藏层维度设为 128。

本文在测试集上进行了 5 次独立实验，并对结果进行了平均处理。在训练过程中，学习率设为 0.001，批处理大小为 32，并使用 Adam 优化器对各组件进行优化，训练总共进行了 100 轮迭代。

5.3. 评价指标

Recall@K 是一种常用于信息检索和排序任务的评价指标，衡量模型在前 K 个返回结果中成功找到相关项的比例。在检索任务中，Recall@K 计算了用户在前 K 个候选项中找到至少一个相关项的概率。其计算公式为：

$$\text{Recall@}K = \frac{\text{找到的相关相数}}{\text{总相关项数}} \tag{5}$$

在实际应用中，Recall@K 更适用于多项相关的情况，例如在推荐系统或搜索引擎中，通过计算 Recall@K 可以评价模型在推荐前 K 个结果中捕获用户兴趣的能力。通常选取 K = 1、5 或 10 来分析模型在不同深度上的检索效果。因此，本文采用 Recall@1、Recall@5 和 Recall@10 作为评估指标。

5.4. 实验结果

表 2 展示了在跨模态检索任务中不同方法的实验结果。可以看出，基于 CLIP 的四个基线方法，包括 LambdaRank [10]、NeuralNDCG [11]、ListNet [12]和 RankNet [13]，在两个任务的 Recall@10 指标上显著超越了 CLIP。这一结果尤为明显，尤其是在图像检索任务中，RankNet 的 Recall@10 比 CLIP 高出 5.13%，进一步证明了将排序策略应用于跨模态检索的有效性。

此外，这四个基线方法在文本检索任务中的 Recall@5 表现也显著优于两个经典的跨模态检索方法 DCCA [3]和 CMPM [4]。特别是 NeuralNDCG 在文本检索中的 Recall@5 超越 DCCA 达 158.7%，表明相比传统的跨模态方法，基于排序损失的优化能够显著提升检索效果，展现了本文方法在性能与竞争力方面的显著优势。

通过这些实验结果，可以清晰地看出，基于排序的方法在跨模态检索中具有显著的提升潜力。这不仅证明了排序模型在处理复杂检索任务时的有效性，还为未来的研究提供了重要的方向。

Table 2. Performance comparison (R@K(%)) on Flickr30K
表 2. 不同方法在 Flickr30K 数据集上的性能比较

loss	文本检索			图像检索		
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
LambdaRank	20.02	75.70	91.74	12.48	52.14	74.66
NeuralNDCG	23.14	80.20	92.25	13.52	57.04	76.54
ListNet	21.82	80.14	92.16	13.82	57.50	76.42
RankNet	20.02	76.84	92.06	13.22	55.24	77.46
DCCA	12.60	31.00	43.00	16.70	39.30	52.90
CMPM	29.10	56.30	67.70	37.10	65.80	76.30
CLIP	29.46	83.46	90.30	16.16	58.56	73.68

6. 结论

本文提出了一种将排序引入跨模态检索的方法，以提升跨模态检索性能。该方法通过将图像和文本

转换为嵌入表示，并将其输入排序模型，利用排序损失训练模型，从而对图像或文本的相似性进行排序，检索出与查询(图像或文本)最相关的项目(文本或图像)。

在跨模态检索实验中，结果显示基于 CLIP 的四种排序损失方法(LambdaRank、NeuralNDCG、ListNet 和 RankNet)在任务中表现优异。这些实验不仅验证了排序方法在提升检索性能方面的潜力，也展示了其在图像与文本匹配任务中的优势。此外，在标题检索中，排序损失方法在 Recall@5 上显著超越了 DCCA 和 CMPM，证明了使用排序损失用于跨模态检索优化的有效性，相比于传统的跨模态检索方法更加有效，显示了本文方法的强大竞争力。

与传统的跨模态检索方法相比，引入排序损失来优化跨模态检索过程，可以显著提升检索的效果和准确性。近年来，许多学者提出了新颖且更具竞争力的排序损失函数，这些损失函数同样可以应用于跨模态检索的优化，从而进一步提升检索效果，因此本文的方法除了有效性，还具有广泛适用性，具有巨大潜力与优势。

综上所述，本文的研究为跨模态检索领域提供了新的视角和方法论。通过结合排序损失与跨模态特征学习，不仅提升了检索精度，也为未来的研究奠定了坚实基础。

参考文献

[1] Wang, K., Yin, Q., Wang, W., Wu, S. and Wang, L. (2016) A Comprehensive Survey on Cross-Modal Retrieval.

[2] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., *et al*. (2021) Learning Transferable Visual Models from Natural Language Supervision. 2021 *International Conference on Machine Learning*, Online, 13-16 December 2021, 8748-8763.

[3] Yan, F. and Mikolajczyk, K. (2015) Deep Correlation for Matching Images and Text. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 3441-3450. https://doi.org/10.1109/cvpr.2015.7298966

[4] Zhang, Y. and Lu, H. (2018) Deep Cross-Modal Projection Learning for Image-Text Matching. In: *Lecture Notes in Computer Science*, Springer, 707-723. https://doi.org/10.1007/978-3-030-01246-5_42

[5] Zhang, C., Cheng, J. and Tian, Q. (2020) Multi-View Image Classification with Visual, Semantic and View Consistency. *IEEE Transactions on Image Processing*, **29**, 617-627. https://doi.org/10.1109/tip.2019.2934576

[6] Wang, Z., Gao, Z., Yang, Y., Wang, G., Jiao, C. and Shen, H.T. (2024) Geometric Matching for Cross-Modal Retrieval. *IEEE Transactions on Neural Networks and Learning Systems*, 1-13. https://doi.org/10.1109/tnnls.2024.3381347

[7] Vaswani, A. (2017) Attention Is All You Need. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[8] Pobrotyn, P., Bartczak, T., Synowiec, M., Białobrzeski, R. and Bojar, J. (2020) Context-Aware Learning to Rank with Self-Attention.

[9] Young, P., Lai, A., Hodosh, M. and Hockenmaier, J. (2014) From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. *Transactions of the Association for Computational Linguistics*, **2**, 67-78. https://doi.org/10.1162/tacl_a_00166

[10] Burges, C.J.C., Ragno, R. and Le, Q.V. (2007) Learning to Rank with Nonsmooth Cost Functions. In: *Advances in Neural Information Processing Systems* 19, The MIT Press, 193-200. https://doi.org/10.7551/mitpress/7503.003.0029

[11] Pobrotyn, P. and Białobrzeski, R. (2021) Neural NDCG: Direct Optimization of a Ranking Metric via Differentiable Relaxation of Sorting.

[12] Cao, Z., Qin, T., Liu, T., Tsai, M. and Li, H. (2007) Learning to Rank. *Proceedings of the* 24*th International Conference on Machine Learning*, New York, 20-24 June 2007, 129-136. https://doi.org/10.1145/1273496.1273513

[13] Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., *et al*. (2005) Learning to Rank Using Gradient Descent. *Proceedings of the* 22*nd International Conference on Machine Learning*, New York, 7-11 August 2005, 89-96. https://doi.org/10.1145/1102351.1102363