吴强. 多模态检索数据增强方法研究[J]. 智能计算机与应用,2024,14(7):227-230. DOI:10.20169/j.issn.2095-2163.240736

多模态检索数据增强方法研究

吴 强

(东华大学 计算机科学与技术学院, 上海 201620)

摘 要: 多模态检索领域中,基于 Transformer 的多模态模型因其强大的推理能力和高精度的检索效果而备受关注,该模型通常采用预训练和微调的方法进行训练。微调过程使用更大规模数据集可以提高检索精度,构建大规模有标注数据集需要耗费大量人力物力,因此本文提出了 3 个数据增强方法,即概念增强、EDA 搭配回译、图文标签化来扩大数据集规模,从而提升多模态模型检索精度。实验结果表明,使用这 3 种方法和 3 种方法的组合对训练数据进行增强,3 个多模态检索模型的图文检索精度均有所提升。

关键词: 多模态模型; 数据增强; 检索精度

中图分类号: TP399 **文献标志码**: A **文章编号**: 2095-2163(2024)07-0227-04

Research on data augmentation methods for multimodal retrieval

WU Qiang

(School of Computer Science and Technology, Donghua University, Shanghai 201620, China)

Abstract: In the field of multimodal retrieval, Transformer based multimodal models have attracted much attention due to their powerful reasoning ability and high-precision retrieval results. These models are usually trained using pre training and fine-tuning methods. The fine-tuning process using larger datasets can improve retrieval accuracy. Building large-scale annotated datasets requires a lot of manpower and material resources. Therefore, we proposes three data augmentation methods, namely concept augmentation, EDA combined with backtranslation, and image text labeling, to expand the size of the dataset and improve the retrieval accuracy of multimodal models. The experimental results show that using these three methods and their combination to enhance the training data, the image and text retrieval accuracy of the three multimodal retrieval models has been improved.

Key words: multimodal retrieval; data augmentation; retrieval accuracy

0 引 言

随着现代信息技术和互联网的快速发展,人们对多模态信息如图片、视频、文本、语音等的需求日益增长。传统的单一模态检索无法满足人们日益复杂的检索需求,多模态检索应运而生。多模态检索可以同时处理来自不同模态的数据,多模态信息具有高维、非结构化、不确定性强等特点。通过融合多种模态信息,可以更全面地表达信息内容,提高信息检索的效率和精度。

在多模态检索领域,基于 Transformer 的多模态模型采用预训练搭配微调的范式,在预训练过程使用大量无标注的数据集,在微调过程使用有标注的数据集。模型经过微调,可以大幅度提升检索精度。对于模型微调,构建更大规模的数据集能够进一步提高检索精度,但构建大规模的有标注数据集需要耗费大量的人力和物力,如何在不增加数据集的前提下提升检索精度成为一个亟待解决的问题。数据增强的策略与多种深度学习理论有相似之处,如侧重于让正样本对的距离尽可能缩小而将负样本对的距离尽可能扩大的对比学习[1];侧重于同义词替换和随机插入删除交换的对抗学习[2];基于 Transformer 模型的指导生成样本的 GAN 等[3]。数据增强可以通过一系列变换,生成许多新的数据,从而增加数据量。

本文针对构建更大规模的微调数据集来提升检索精度这一问题,提出了 3 个数据增强方法即概念增强、EDA(Easy Data Augmentation)搭配回译、图文标签化,用于扩大微调阶段数据集的规模,进一步提升多模态模型在图文检索任务上的精度。

作者简介: 吴 强(1998-),男,硕士研究生,主要研究方向:多模态检索。Email:15079480472@163.com

收稿日期: 2023-05-07

1 相关工作

1.1 基于 Transformer 的多模态模型

基于 Transformer 的多模态模型有单流和双流两种形式。单流模型将图片和文本的数据融合在一起,使用一个 Transformer 模型处理;双流模型使用两个 Transformer 模型分别处理图片和文本数据,并将处理后的多模态数据输入到 Co-Transformer 模型进行语义融合[4]。

在获取图片特征方面,有 3 种方法:使用 Faster-RCNN(Faster Region-based Convolutional Neural Networks)提取图片目标区域的特征;使用卷积神经网络 CNN(Convolutional Neural Networks)提取图片像素级别的特征;使用计算机视觉模型 ViT(Vision Transformer)获取被分成块类型的图片特征。在获取文本特征方面,多模态模型均采用 BERT(Bidirectional Encoder Representations from Transformers)提取文本的特征。

目前出现了许多基于 Transformer 的多模态模型,如 UNITER(Universal Image-Text Representation Learning)、TERAN(Fine-Grained Visual Textual Alignment For Cross-Modal Retrieval Using Transformer Encoders)、ViLT(Vision-and-Language Transformer)。UNITER 模型提出了预训练时的一些注意点,如在预训练时,如果能够使用一些文本数据进行训练,那么预训练的效果会比没有文本数据预训练时更好;如果将视觉和语言结合起来进行训练,会比仅使用视觉或语言数据进行预训练效果更好[5]。TERAN 适用于图文检索任务,其目标函数直接定义在从架构中输出的区域和单词集上,能够更精确地表示图文的局部信息,同时也能够避免使用全局信息带来的信息损失,在训练过程中通过逐个对齐图片和句子中包含的区域和单词,来获得更精细的匹配结果[6]。ViLT 探索了数据增强方法对模型检索精度的影响,并使用了掩盖文本和对图片进行随机增强的方法,提升模型在图文检索任务上的精度。ViLT 由 Transformer 编码器组成,将图片和文本向量化,预训练任务如图片分类、图片重构、图片问答、文本分类等,学习到多个不同的图片和文本之间的相关性[7]。

图文检索任务是多模态模型下游任务中的一个任务,包括两种类型:以文搜图和以图搜文。以文搜图是使用文本来检索具有相似语义的图片,而以图搜文是使用图片来检索具有相似语义的文本[8]。在图文检索任务中,模型均采用 Recall@K(K = 1, 5,10) 作为评估指标[9]。

1.2 数据增强

数据增强是指通过对现有数据进行一系列变换和扩充来生成新的训练数据的方法,广泛使用在计算机视觉和自然语言处理领域。在计算机视觉领域,常见的数据增强方法包括旋转、缩放、翻转、裁剪、添加噪声等[10]。在自然语言处理领域,数据增强方法大致可以分成 3 类:基于释义的方法、基于噪声的方法和基于采样的方法[11]。基于释义的方法通过同义词替换、回译等方式来生成新的句子,从而扩充数据集;基于噪声的方法则通过添加噪声、随机删除单词等方式来增加数据的多样性;基于采样的方法则是通过从已有数据中随机采样生成新的数据。

2 多模态检索数据增强方法

2.1 概念增强

概念增强将数据集中全体文本的所有名词都提取出来,按照其出现频次由高到低进行排列;其次,从中选择一定比例(60%)的名词;最后,读取这些名词在 WordNet 词典中的释义,并将这些释义以定语从句的方式插入到原始文本中。本文去除了频率最高的前 20%和末尾 20%的专有名词,只保留中间部分的 60%作为名词列表。出现频率最高的 20%名词在文本中频繁出现,而出现频率最低的 20%名词则通常是专有名词。考虑到修饰这些名词可能会影响方法的执行效果,本文避免使用。

对于自行车,本文将其描述为“有两个轮子,一个鞍座和一对脚蹬的交通工具”,而摩托车则描述为“一种具有坚固框架和两个车轮的机动车”,概念增强示例见表 1。通过概念增强,本文为模型提供关于名词更加详细的语义信息,提高其区分不同名词的能力。

表 1 概念增强的示例

Table 1 Example of concept augmentation

原文本	经过概念增强的文本
A racing driver on a red a bicycle is racing hard	A racing driver on a red a bicycle which is two-wheeled small land vehicle is racing hard(对 bicycle 进行增强)
Asian girls in white t-shirts standing around a motorcycle	Asian girls in white t-shirts standing around a motorcycle which is a motor vehicle with a strong frame and two wheels(对 motorcycle 进行增强)

2.2　EDA 搭配回译

EDA 搭配回译的基本思路：在保留文本中名词的情况下，对一张图片对应的 5 个文本依次随机应用 EDA 的其中一个操作，再对得到的文本进行回译操作，从而得到增强后的文本，可以增加文本数据的多样性，同时保留文本的语义信息。

对于原文本“A older Asian man is playing an instrument in front of a young boy on the street”，本文对该文本进行 EDA 搭配回译，EDA 搭配回译的示例见表 2。

表 2　EDA 搭配回译的示例

Table 2　Example of EDA combined with backtranslation

EDA 的操作	经过 EDA 的文本	再进行回译
随机插入	A older Asian man is playing anto（随机插入 to）instrument in front of a young boy on the street	An elderly Asian man was playing an instrument in front of a little boy in the street
随机交换	A older Asian of is playing an instrument in front man（man 和 of 发生交换）a young boy on the street	An elderly Asian was playing an instrument in front of a little boy in the street
随机删除	A older Asian man is playing an instrument in front of a young boy on the（street 被删除）	An elderly Asian man is playing an instrument in front of a little boy
同义词替换	A older Asian man is playing an device（instrument 被替换成 device）in front of a young boy on the street	An elderly Asian man playing with equipment in front of a young boy in the street

2.3　图文标签化

图文标签化首先从一张图片所对应的 5 个文本中提取出名词，选择出现频率最高的两个名词作为图文的标签，分别置于图片上方和文本末尾，实现图片和文本的联合增强。

本文在示例图片和对应的文本上使用图文标签化，增强前后图片和文本的对比效果。图文标签化的示例见表 3，示例图片描述的是一个拿着棒球棍的小女孩在草地上奔跑，从图文对比提取出的标签是 girl 和 bat。

表 3　图文标签化的示例

Table 3　Example of image-text labelization

原文本	增强后的文本	原图	增强后的图片
A little girl wearing a mauve dress and blue sparkly shoes carries a big blue baseball bat	A little girl wearing a mauve dress and blue sparkly shoes carries a big blue baseball bat and the sentence above is girl and bat		

3　实验

为了验证概念增强、EDA 搭配回译、图文标签化 3 种方法的有效性，本文选取多模态模型中常见的 TERAN、ViLT 和 UNITER 作为基准模型，选取 Flickr30K 和 MSCOCO 作为以文搜图和以图搜文任务的数据集进行了相关的实验，实验结果见表 4。

在 Flickr30k 数据集上进行以文搜图任务，对基准模型应用概念增强、EDA 搭配回译、图文标签化，TERAN 模型的 Recall@1 分别提升 1.5%、1.2%、1.1%；Vilt 模型分别提升 0.4%、0.3%、0.2%；UNITER 分别提升 0.5%、0.3%、0.2%。与此同时，在以图搜文任务上，TERAN 模型的 Recall@1 分别提升 0.8%、0.5%、1.4%；Vilt 分别提升 0.1%、0、0.5%；UNITER 分别提升 0.1%、0、0.3%。

在 MSCOCO 数据集上进行以文搜图任务，对基准模型应用 3 个方法，TERAN 模型的 Recall@1 分别提升 1.9%、1.7%、1.5%；Vilt 模型分别提升 1.5%、1.2%、1.1%；UNITER 模型分别提升 1.1%、0.9%、0.8%。在以图搜文任务上，TERAN 模型的 Recall@1 分别提升 0.2%、0、0.3%；Vilt 分别提升 0.1%、0、0.2%；UNITER 分别提升 0.1%、0.1%、0.2%。

表 4 多个模型在 Flickr30k 和 MSCOCO 上图文检索的实验结果

Table 4 Experimental results of multiple models image-text retrieval on Flickr30k and MSCOCO

模型	以文搜图		以图搜文	
	Flickr30k	MSCOCO	Flickr30k	MSCOCO
TERAN	63.1	45.1	79.2	59.3
TERAN +概念增强	**64.6**	**47.0**	**80.0**	**59.5**
TERAN + EDA 搭配回译	**64.3**	**46.8**	**79.7**	59.3
TERAN +图文标签化	**64.2**	**46.6**	**80.6**	**59.6**
TERAN +三种方法组合	**64.9**	**47.3**	**80.7**	**60.1**
ViLT	63.8	42.4	83.5	61.8
ViLT+概念增强	**64.2**	**43.9**	**83.6**	**61.9**
ViLT+ EDA 搭配回译	**64.1**	**43.6**	83.5	61.8
ViLT+图文标签化	**64.0**	**43.5**	**84.0**	**62.0**
ViLT+三种方法组合	**64.4**	**44.0**	**84.0**	**62.1**
UNITER	72.5	50.3	85.9	64.4
UNITER+概念增强	**73.0**	**51.4**	**86.0**	**64.5**
UNITER+ EDA 搭配回译	**72.8**	**51.2**	85.9	**64.5**
UNITER+图文标签化	**72.7**	**51.1**	**86.2**	**64.6**
UNITER+三种方法组合	**73.1**	**51.6**	**86.4**	**64.6**

综上,在 TERAN、ViLT、UNITER 微调过程中应用 3 种数据增强方法,均可以提升模型在图文检索任务上的检索精度,Recall@ 1 有小幅提升。本文将 3 个方法组合并应用,发现组合方法提升检索精度的幅度最大,说明 3 个方法不存在互斥性,可以联合使用。

4 结束语

本文旨在深入研究数据增强方法在提升基于 Transformer 的多模态模型检索精度方面的作用,在基于文本和图文联合的角度提出了针对性的数据增强方法,并使用这些方法增强模型微调的数据集。为了验证概念增强、EDA 搭配回译、图文标签化 3 种方法的有效性,对 3 个多模态模型即 TERAN、ViLT 和 UNITER 在图文检索任务上进行了实验。实验表明,这些增强方法都能提高模型在图文检索任务中的检索精度。

参考文献

[1] WANG D, DING N, LI P,et al. Cline: Contrastive learning with semantic negative examples for natural language understanding[J]. arXiv preprint arXiv:2107.00440, 2021.

[2] PANG T, YANG X, DONG Y, et al. Boosting adversarial training with hypersphere embedding [J]. Advances in Neural Information Processing Systems, 2020, 33: 7779-7792.

[3] 王海文. 基于生成式对抗网络的数据增强方法研究[D]. 南京: 南京邮电大学, 2020.

[4] LU J, BATRA D, PARIKH D, et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks [C]//Proceedings of the 33rd Conference on Neural Information Precessing System. IEEE, 2019:13-23.

[5] CHEN Y C, LI L, YU L, et al. Uniter: Universal image-text representation learning [C]//European Conference on Computer Vision. Cham: Springer, 2020: 104-120.

[6] MESSINA N, AMATO G, ESULI A, et al. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders [J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2021, 17(4): 1-23.

[7] KIM W, SON B, KIM I. Vilt: Vision-and-language transformer without convolution or region supervision [C]// Proceedings of International Conference on Machine Learning. IEEE, 2021: 5583-5594.

[8] 刘颖, 郭莹莹, 房杰, 等. 深度学习跨模态图文检索研究综述[J]. 计算机科学与探索, 2022, 16(3): 489.

[9] PATEL Y, TOLIAS G, MATAS J. Recall@ k surrogate loss with large batches and similarity mixup[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 7502-7511.

[10] 朱晓慧,钱丽萍,傅伟. 图像数据增强技术研究综述[J]. 软件导刊,2021,20(5):230-236.

[11] LI B, HOU Y, CHE W. Data augmentation approaches in natural language processing: A survey[J]. Ai Open, 2022, 3: 71-90.