

武汉理工大学

(申请工学硕士学位论文)

面向食谱-食物图像的 跨模态检索方法研究

培养单位：计算机科学与技术学院

学科专业：软件工程

研究生：管子辰

指导教师：李琳 教授

副指导教师：无

2021 年 5 月

分类号_____

密级_____公开_____

UDC_____

学校代码_____10497_____

武汉理工大学

学位论文

题 目_____面向食谱-食物图像的跨模态检索方法研究_____

英 文_____Cross-modal Retrieval on Cooking Recipes and_____

题 目_____Food Images_____

研究生姓名_____咎子辰_____

指导教师 姓名_____李琳_____职称_____教授_____学位_____博士_____

单位名称_____计算机科学与技术学院_____邮编_____430070_____

副指导教师 姓名_____无_____职称_____学位_____

单位名称_____邮编_____

申请学位级别_____硕士_____学科专业名称_____软件工程_____

论文提交日期_____2021 年 3 月_____论文答辩日期_____2021 年 5 月 21 日_____

学位授予单位_____武汉理工大学_____学位授予日期_____

答辩委员会主席_____袁景凌_____评阅人_____教育部盲审_____

教育部盲审_____

2021 年 5 月

独 创 性 声 明

本人声明，所呈交的论文是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得武汉理工大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签 名：_____日 期：_____

学位论文使用授权书

本人完全了解武汉理工大学有关保留、使用学位论文的规定，即学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人承诺所提交的学位论文（含电子学位论文）为答辩后经修改的最终定稿学位论文，并授权武汉理工大学可以将本学位论文的全部内容编入有关数据库进行检索，可以采用影印、缩印或其他复制手段保存或汇编本学位论文。同时授权经武汉理工大学认可的国家有关机构或论文数据库使用或收录本学位论文，并向社会公众提供信息服务。

（保密的论文在解密后应遵守此规定）

研究生（签名）：_____ 导师（签名）：_____ 日期_____

摘要

近年来，分享美食照片和食谱成为一种流行趋势，这种趋势伴随着社交媒体的壮大而不断地成长。人们在社交媒体上正面对着成千上万的海量食物图像、视频和食谱文本，迫切需要食谱-食物图像跨模态检索框架，来进行食物图像和烹饪食谱的双向准确检索。目前流行的框架中存在如下不足。首先，由于大多使用 word2vec 结合 LSTM 对菜谱文本进行编码，然而基于单词级别的菜谱文本编码难以表示菜谱步骤间的顺序关系，对于菜谱而言制作顺序却是最重要的信息。其次，由于 Recipe 1M 数据集庞大，食谱文本和食物图片均来自网络美食分享网站爬取，Recipe 1M 中噪声数量类型众多，流行框架没有专门考虑处理食谱-食物图像噪声问题。最后，三重损失在跨模态检索框架训练中虽然取得了较好的检索质量，但应用到食谱-食物图像跨模态检索时，存在模型训练后泛化能力不足的问题，需要进行改进。

基于上述问题，本文进行如下研究：

(1) 本文从食谱文本表示学习，食物图像表示学习，模态对齐和跨模态学习等方面逐一分析各基线模型。同时，根据各基线模型在 Recipe 1M 数据集上的实验结果，对比和分析各基线模型存在的不足，明确研究内容。

(2) 本文提出了基于句子级食谱文本编码的食谱-食物图像跨模态检索框架 SBNR。该框架分为四个组件。多层句子级别注意力机制组件进行食谱的句子级嵌入学习，食物图像过滤组件处理食物图像中的噪声信息，模态对齐组件强化模态分布对齐，跨模态学习组件使用两种改进的三重损失学习跨模态对齐信息。本文所提出的框架在 Recipe 1M 数据集上的实验结果表明，在 1K 和 10K 测试数据集上，本框架在 R@1、R@5、R@10 和 MedR 四个评价指标上优于各基线模型。从实验结果及样例分析中得出不足，为后续研究工作提供指导作用。

(3) 为进一步改进食谱文本中多层语义的学习和食物图像更细粒度的噪声过滤，本文基于 (2) 中提出的框架和食物图像区域检测，提出了食谱-食物图像噪声筛选打分组件。框架改进部分主要体现在多头注意力机制的使用，增强了食谱文本嵌入，增强学习食谱的多层隐层句子级嵌入信息的能力，称改进后的 SBNR 为 SBNR⁺框架。本文改进后的 SBNR⁺框架在 Recipe 1M 数据集上的实验结果表明，SBNR⁺在 R@1、R@5、R@10 和 MedR 四个评价指标上进一步提升了跨模态检索质量。从实验结果及样例分析中讨论出优势和不足，为将来

研究工作提供了建议。

关键词：跨模态检索，文本编码，食物图像过滤，三重损失，图像检测

Abstract

In recent years, benefiting from the development of social media, there is a growing trend towards sharing food images and recipes. Faced with massive food images, videos and food recipes on social media, there is an urgent need for a cooking recipes and food images cross-modal retrieval framework. The following shortcomings exist in the current popular cross-modal retrieval frameworks. Firstly, most of them use word2vec combined with LSTM to encode recipes, however, word-level recipe embeddings are difficult to represent the sequential relationship between recipe instructions, for which the order is the most important information. Secondly, since the cooking recipes and food images in the Recipe1M dataset are crawled from food sharing websites and popular frameworks do not specifically consider dealing with the cooking recipe and food image noise problem, there exist various types of noise in the dataset. Finally, although triplet loss has achieved good retrieval quality in the cross-modal retrieval frameworks, the generalization ability of the trained model still needs to be improved when applying the triplet loss to perform the recipe-image cross-modal retrieval.

Based on the above problems, this thesis conducts the following studies.

(1) In this thesis, we analyze the baseline models one by one in terms of recipe text representation learning, food image representation learning, modal alignment and cross-modal learning. At the same time, based on the experimental results of each baseline model on the Recipe 1M dataset, the shortcomings of each baseline model are analyzed to demonstrate our proposed approach.

(2) This thesis proposes a cross-modal framework for cooking recipes and food images retrieval (SBNR) based on sentence-level cooking recipe embeddings. SBNR consists of four components. The multi-layer sentence attention networks component performs sentence-level embedding learning of cooking recipes, the food image filtering component handles the noise information in food images, the modal alignment component reinforces the modal distribution alignment, and the cross-modal learning component learns cross-modal alignment information using two improved triplet loss. The experimental results of the proposed framework on the

Recipe 1M dataset show that our framework SBNR outperforms each baseline model on four evaluation metrics, $R@1$, $R@5$, $R@10$ and MedR, on 1K and 10K test datasets. Shortcomings are drawn from the experimental results and sample analysis to provide guidance for subsequent research work.

(3) To further improve the learning of multi-layer sentence-level semantics in cooking recipes and finer-grained noise filtering of food images, this thesis proposes a cooking recipes and food images noise scoring component based on the framework in (2) and food image region detection. The optimization parts mainly lie in the multi-head attention networks, enhanced recipe embeddings, learned multi-layer implicit sentence-level embedding information of cooking recipes. The new framework is called SBNR+. The experimental results of the optimized framework SBNR+ on the Recipe 1M dataset in this thesis show that SBNR+ further improves the cross-modal retrieval quality on four evaluation metrics, $R@1$, $R@5$, $R@10$ and MedR, greatly outperforming each baseline model. Shortcomings are identified from the experimental results and sample analysis, and suggestions are provided for future research work.

Key words: cross-modal retrieval, text embeddings, food image filtering, triplet loss, image detection

目录

摘 要	I
Abstract.....	III
第 1 章 绪论	1
1.1 选题背景	1
1.2 选题意义	2
1.3 国内外研究现状	3
1.3.1 食谱文本表示学习	3
1.3.2 食物图像表示学习	4
1.3.3 食谱-食物图像模态对齐	5
1.3.4 食谱-食物图像跨模态学习	5
1.4 存在的问题	6
1.5 研究内容	8
第 2 章 任务及基线模型	10
2.1 任务描述	10
2.2 Recipe 1M 数据集介绍.....	11
2.3 评价指标	12
2.4 基线模型分析	13
2.4.1 JNE 模型	13
2.4.2 ATTEN 模型	14
2.4.3 ADAMINE 模型.....	15
2.4.4 ACME 模型	16
2.4.5 R2GAN 模型	18
2.5 基线模型实验结果及分析	19
2.5.1 JNE 模型	20
2.5.2 ATTEN 模型	20
2.5.3 ADAMINE 模型.....	20
2.5.4 ACME 模型	21
2.5.5 R2GAN 模型	21
2.6 本章小结	22
第 3 章 基于句子级编码的食谱-食物图像跨模态检索框架	23

3.1 问题描述	23
3.2 整体框架	23
3.3 组件介绍	25
3.3.1 多层句子级别注意力机制组件	25
3.3.2 图像过滤组件	28
3.3.3 模态对齐组件	30
3.3.4 跨模态学习组件	31
3.4 实验	34
3.4.1 数据集和评价指标	34
3.4.2 对比基线	34
3.4.3 实验参数	35
3.4.4 实验结果与分析	35
3.4.5 样例分析	38
3.4.6 实验中的不足	40
3.5 本章小结	41
第 4 章 基于食物图像检测的噪声过滤及检索方法	42
4.1 问题描述	42
4.2 整体框架	42
4.3 组件介绍	44
4.3.1 基于多头注意力机制的组件改进	44
4.3.2 食物图像的区域检测组件 YOLO	48
4.3.3 基于食物图像区域检测的噪声打分筛选组件	49
4.4 实验	52
4.4.1 数据集和评价指标	52
4.4.2 对比基线	52
4.4.3 实验参数	52
4.4.4 实验结果与分析	53
4.4.5 样例分析	57
4.4.6 实验中的不足	59
4.5 本章小结	60
第 5 章 总结与展望	62
5.1 总结	62

5.2 未来展望	63
致谢	64
参考文献	66
攻读硕士学位期间获得与学位论文相关的研究成果	71

第1章 绪论

1.1 选题背景

食物是人类生活的必需品，也是生活体验的基础。民以食为天，食物在人类生活上，健康上，幸福上都有重大影响^{[1][2]}。食品相关研究可为指导人类行为、改善人类健康、了解烹饪文化等多种应用和服务提供支持。随着社交网络、移动网络和物联网（IoT）的快速发展，人们通常会上传、分享和记录食物图像、食谱、烹饪视频和食物日记，网红店探店和美食尝鲜成为 vlog 视频的主流内容，从而产生大规模的食物数据。大规模的食物数据提供了丰富的食物方面的知识，有助于解决人类社会的许多核心问题，例如，食物营养学，食物文化研究，以及合理的广告投放等。因此，食物计算方面的研究走上了历史的舞台。在食物计算中，计算方法被应用于解决医学、生物学、美食学和农学^[3]中与食品相关的问题。

近年来，食物相关研究^[4-6]已成为当今研究的热点，受到各个领域的广泛关注。无论是大规模的食物数据，还是最近计算机科学的突破，都在改变人们分析食物数据的方式。因此，针对不同的以食物为导向的任务和应用，在食物领域开展了大量的工作。食物计算领域有十分多的研究方向，无论是大规模的食物数据，还是最近计算机科学的突破，都在改变人们分析食物数据的方式。因此，针对不同的以食物为导向的任务和应用，在食物领域开展了大量的工作。例如食物识别^[7-12]，食物健康管理^[13]，食物检索，食品文化研究^[14,15]，食品成分及健康分析^[16-19]，食物推荐^[20,21]等。现如今，人们更加关注健康的饮食和营养均衡的食谱。然而，当人们在社交媒体上面对数以亿计的食物食谱和食物图片时，他们可能很难找到自己喜欢的准确而有价值的信息。在这个大背景下，食物图像与菜谱文本的检索便成为了本文的研究方向。食物图像与菜谱的检索从早期单纯的食物图像检索^[22]，食物菜谱检索^[23]，到如今的跨模态食物图像菜谱检索。

跨模态检索^[24-30]旨在检索不同模态的相关条目。由于异构性差异，不同类型的媒体具有不一致的分布和特征表示，这使得跨媒体检索具有相当大的挑战性，从而很难计算两种不同模态之间的相似性。食物图像与食谱的检索任务同样存在这种差异性。同时，跨模态检索的技术发展，也有助于其他食品计算方

向的研究，食品健康等需要根据食物图片计算卡路里和营养信息的操作，由于直接从图片中很难看出食物的大小和重量，所以很难进行预测估计，但是在跨模态检索的辅助之下，可以让计算更加精确。由此可看出，食谱-食物图像的跨模态研究扮演重要的作用。

1.2 选题意义

目前，无论是日常生活中还是网络社交媒体上，都存在着海量食物方面的数据，包括多种形式，例如食谱文本，美食视频，美食图片等。这些数据对科研界和产业界都具有极高的价值。本文进行跨模态食谱-食物图像的研究有着广泛的意义。首先，探索出一个更加合理的食物图像处理方案。找出可以编码食物图像且良好处理食物图像中噪声的手段，并提供一个处理食物图片和菜谱文本匹配对噪声的方案。其次，探索出菜谱文本多层句子级别嵌入，来解决单词级别食谱向量对于菜谱间关系处理的局限性问题。再者，探索出一个强化跨模态对齐的方法，以加强模态间的分布对齐。最后，探索出一个更加适合食谱-食物图像检索工作的三重损失函数，增强模型泛化，提高训练结果的准确度。

(1) 广泛的应用场景

食谱-食物图像的跨模态研究拥有着十分广泛的应用场景。首先，日常生活中，最广泛的应用是根据美食照片进行食谱及相关信息的检索。如今，微博，快手，抖音等各大社交媒体充斥着美食分享，当人们面对这些诱人的美食，总想获得其相关的信息，如何制作，食物的名称是什么，营养成分等。例如，第一次见到川菜鱼香肉丝的人们，可能永远想不到只有猪肉没有鱼肉的菜名叫鱼香肉丝。此时跨模态检索就扮演了重要的作用，可以根据美食视频检索相关菜谱信息^[31]，或利用食物图片检索到所有相关信息。其次，食谱-食物图像的研究也同时可以应用于外卖行业，视频 vlog 制作行业，根据检索到的食物信息，用户可以很快进行点餐，增加用户的使用体验，提高用户粘性。视频博主则可以快速定位美食信息，制作美食探店 vlog 或者美食制作视频。

(2) 促进跨模态检索领域的研究

自从 2017 年 Recipe 1M 数据集^[32]的提出，食谱-食物图像跨模态检索研究进入一个新的阶段，神经网络应用到食谱-食物图像检索领域。2018 年，模态对齐的损失函数从 2017 年的余弦损失升级为三重损失^[33]。随后 2019 年，hard

example mining^[34]三重损失的使用让食谱-食物图像的检索精确度得到了极大提高^[35]。hard example mining 三重损失是借鉴了人脸识别文章的思想^[34]，并进行改进应用到食谱-食物图像跨模态检索任务中，主要是为了解决三重损失无法有选择性地筛选训练数据的问题。改进后的三重损失可以合理分配训练数据的权重，做到智能分配权重筛选训练数据、降低训练量、增强训练效果和模型泛化能力。食谱-食物图像跨模态检索的发展，同时也给其他跨模态领域提供了解决跨模态对齐问题的思路。同样，其他跨模态任务的发展，也促进食谱-食物图像的跨模态检索研究，根据语音文本文字的跨模态对齐的思想^[36]，本文也基于注意力机制^[37]提出了一个食谱信息过滤食物图像噪声的组件。同时，为了筛选更细粒度的食谱-食物图像噪声对，本文又结合食物图像区域检测提出噪声筛选打分组件，也给后续的研究者提供一个食谱文本和食物图像噪声过滤的思路。

1.3 国内外研究现状

近年来，为了提高图像检索的效率，人们采用了深度学习方法。一般来说，现如今的方法^{[32][33][35][38-40]}方法均是使用 CNN（如 VGG16^[41]或 Resnet^[42]）来嵌入食物图像，并使用 LSTM^[43]和 word2vec^[44]来嵌入烹饪食谱。这些方法利用余弦损失^[32]、三重态损失^[33]或 GAN 损失^[40]加强跨模态向量的对齐同时优化匹配对。本小节针对各个流行的食谱-食物图像跨模态检索方法的特点，分类讲解国内外发展现状。

1.3.1 食谱文本表示学习

2017 年，Amaia Salvador 等人^[32]提出了 Recipe 1M 数据集，提供了在跨模态数据上训练检索任务的能力。作者对菜谱使用 word2vec 结合 LSTM 进行菜谱嵌入，将得到的嵌入结果映射到高维度公共空间，并参与后续的训练学习。同年，Jing-jing Chen 等人^[38]提出了 SAN 模型，作者提出了一个深度学习的模型，叠层注意力网络。该模型的亮点之处在于通过从配方中提取成分的相关区域并分配更大的权重来学习对应关系，即通过分割图像并且利用注意力机制筛选有效区域。但大部分情况下，SAN 模型甚至没有传统的 CCA^[45]方法有效。

2018 年 Jing-jing Chen 等人^[39]提出 ATTEN 模型，这个模型将注意力机制同时用在了菜谱的标题，配料，制作步骤，通过不断学习获得注意力权重，来加强对菜谱的信息筛选。同年，Micael Carvalho 等人^[33]提出 ADAMINE 模型。作

者仍然是单词级别的对食谱进行处理，食谱间的顺序问题，食谱间的关系都不能很好的表达在嵌入向量上。其次作者将配料和制作步骤拼接成向量之后，并没有利用食谱标题来嵌入食谱向量，由于菜谱标题很大程度上表明了菜谱的类别和特色，所以在处理过程中重要程度不言而喻。

2019 年 Bin Zhu 等人^[40]提出了 R2GAN 模型，R2GAN 采用了 LSTM 结合层次 LSTM 两种结构对食谱进行处理。对嵌入后的菜谱向量拼接，映射到公共高纬度空间。由于食谱的制作步骤一般长度较长，LSTM 很难以记忆整个制作步骤的信息，R2GAN 采用的层次 LSTM 也是解决此问题的一种方法。同样是 2019 年，Amaia Salvador 等人^[46]提供了一种全新的思路处理食物菜谱和食物图像。在本文中，作者介绍了一个逆向烹饪的系统，它可以根据给定的食物图像重建烹饪食谱。这个系统通过一个新的架构将食物食谱的配料作为一个集合进行预测，在不强制任何顺序的情况下对它们的依赖性进行建模，然后通过同时关注图像和其推断的食物食谱配料来生成烹饪指令。2019 年后，大批研究人员开始着手向生成的方面发展。这确实是一个很不错的想法，既然食物图像菜谱有噪声并且比较难以去除，同时检索速度还十分的缓慢，那不如直接生成结果来的更加高效，这是本文的研究中可以思考和学习的地方。

可以看出针对食谱的处理，各流行的模型框架只是从单词级别编码出发结合 LSTM 学习其中食谱的顺序信息。由于如今句子级嵌入编码的发展，本文认为，对于食谱制作步骤这种十分强调顺序信息的嵌入任务，句子级编码将会比单词级编码性能更强。

1.3.2 食物图像表示学习

2017 年，Amaia Salvador 等人^[32]提出 JNE 模型时，采取的是 Resnet 对食物图像直接处理，将输出的食物图像嵌入编码直接映射到高纬度公共空间。

2019 年 Bin Zhu 等人^[40]提出了 R2GAN 模型，表示过程文本（如跨模态检索的食谱问题）本身就是一个难题，更不用说从食物食谱生成图像了。R2GAN 在生成的图像下，考虑了嵌入和图像空间中的 two-level ranking loss。这些附加组件不仅带来了出色的检索性能，而且生成了接近真实的食物图像，有助于解释食谱的排名。在 Recipe 1M 数据集上，R2GAN 显示了对数据大小的高度可伸缩性，优于所有现有的方法，并生成直观的图像供人类解释搜索结果。可以认为食物菜谱图像的检索也进入了对抗生成时代，首先，对于图像中的噪声问题，

可以采用过滤以及筛选的手段来去除,同样也可以通过对抗生成的方式生成食物图像,这样既解决了图像中大量的噪声问题,也解决了检索时十分缓慢的问题。看起来,对抗生成是一种很美妙的手段,实际上它确实也大大提高了跨模态检索的准确度,是本文需要学习研究的思路。

2020 年 Bin Zhu 等人提出了 CookGAN 模型^[47],从一个新的视角,即图像生成中的因果链,探讨了文本到图像的合成问题。因果关系是烹饪中常见的现象。菜肴的外观根据烹调动作和食物配料而变化。合成的挑战在于生成的图像应该描述物体上的动作的视觉结果。本文提出了一种新的网络体系结构 CookGAN,它模拟因果链中的视觉效果,保留细粒度的细节,并逐步向上采样图像。特别地,提出了一个烹饪模拟器子网络,通过一系列的步骤,基于食材和烹饪方法之间的相互作用,逐步改变食物图像。Recipe 1M 上的实验验证了 CookGAN 能够以令人印象深刻的初始分数生成食物图像。此外,图像在语义上是可解释和可操作的。本文也是使用了 GAN 网络^[48]对食物图像和食谱进行处理研究,对本文的后续研究有一定指导意义。

现如今研究现状下,研究人员除了使用 GAN 生成食物图像外,对食物图像中存在的噪声并没有进行处理,而实际实验中食物图像中噪声类型复杂,多种食物图像噪声会极大影响跨模态检索模型的性能。

1.3.3 食谱-食物图像模态对齐

JNE 模型中,为了强化食谱-食物图像跨模态对齐,作者提出语义正则化组件。语义正则化可以理解为,JNE 框架在高纬度公共空间中,将匹配的食谱-食物图像对均经过共享权重的全连接层,保证框架在不断训练学习时,提供一定的模态对齐能力。

2019 年, Hao Wang 等人^[35]提出的 ACME 模型,使用一个对抗性学习策略来实施模态对齐。文章中使用了 WGAN-GP^[49]来加强模态对齐,当使用对抗生成的损失函数来进行训练学习的时候,将相辅相成的生成器和鉴别器换成食谱向量和图片向量,目的就是希望两者分布一致,这样可以大大提高检索的准确度。可以看出 GAN 对于食物图像跨模态的研究有很大的帮助。

1.3.4 食谱-食物图像跨模态学习

JNE 模型中,使用了余弦损失来进行跨模态学习,余弦损失虽然有一定的

帮助，但是其限制也较大，本文会在后续章节详细分析。

对于 ADAMINE 模型，首先，他们直接在损失中集成语义信息，以细化潜在空间的结构，同时限制要学习的参数数量。其次，他们将模型依赖于双三重损失来适应联合学习目标。再者，提出了一个新的随机梯度下降加权方案，该方案适用于这种双重深度嵌入架构，它在小批量上计算并自动执行自适应挖掘。这篇文章作者提出的双三重损失策略，对后来的研究起到了很大的影响。因为相比于余弦损失只关心匹配对的学习，三重损失同时还关注不匹配对在高维空间的情况，简单来说就是匹配对拉近距离，不匹配对拉远距离。三重损失相比于余弦损失，使得检索结果得到了很大的提高。

ACME 模型中，使用 hard example mining^[34]三重损失策略来学习训练模型。这个策略的有效之处，就在于可以对数据食谱图像匹配对进行有选择性挖掘。三重损失训练强调将匹配对在高维空间拉近距离，而不匹配对在高维空间拉远距离。hard example mining 策略加强了三重损失，它每次不是平等对待所有匹配对和不匹配对，而是去搜寻将某一模态作为 anchor point 后与其最相似高维空间中距离最远的另一个模态的数据以及最不相似高维空间中距离最近的数据，极大增加模型泛化能力，使训练快速收敛，效果十分优秀。

hard example mining 三重损失虽然在跨模态检索任务中取得了较好的成绩，但是由于其学习特性，如果高维空间中存在噪声数据干扰，那么模型的训练梯度将会存在偏差，最终导致模型的检索性能和泛化能力下降。

1.4 存在的问题

虽然近些年，食谱-食物图像跨模态检索任务有了很大的发展，基本框架也是由 2017 年 Salvador 等人提出的 JNE 框架进行改进。使用 CNN (如 VGG16^[41]或 Resnet^[42])来嵌入食物图像，并使用 LSTM^[43]和 word2vec^[44]来嵌入烹饪食谱。随后利用余弦损失^[32]、三重损失^[33]或 GAN 损失^[40]加强跨模态向量的对齐同时优化匹配对。尽管现有的框架在深度学习方面的图像配方检索方面取得了相当大的改进，但基于实验调查和观察，仍然存在很多的不足之处：

(1) 现如今的方法只关注单词级别信息来分析菜谱，而不考虑菜谱中制作步骤的顺序问题。如今的相关工作集中在 LSTM 上，用词嵌入的方法来增强烹饪指令中各个步骤之间的关系。在单词级别的分析中，食谱信息中的一些重要信息将丢失。对于制作顺序这种多句子存在并且互相间关系密切的情况，句子

层面的信息是至关重要的，因为它包含句子关系信息，即做菜顺序和配料的顺序。例如，要做炒鸡蛋，众所周知的做法是先热油再炒鸡蛋，如果顺序相反，先加鸡蛋再加入食用油，便成油煮鸡蛋，鸡蛋本身容易吸油的情况下，油温过低鸡蛋更容易吸油导致成品很腻，如果锅子质量不好还会鸡蛋粘锅焦糊，由此可见菜谱制作步骤的顺序是十分重要的。而且，现有的方法并没有比较充分利用烹饪菜谱的信息，我们需要充分利用菜谱步骤，菜谱标题，菜谱配料信息，注意句子的层次特征。调查文献发现，句子的层次信息以及句子级别的向量处理，BERT^[50, 51]可以良好的进行处理。

(2) 食物图像中存在大量的噪声问题仍需要关注。现如今的通用方法中只考虑了使用 CNN 来对图像编码，但是并没有关注到如果食物图像中存在大量噪声，那么对此进行跨模态检索菜谱的效率是很低下的。这里的噪声问题包括，图像本身存在大量的噪声，例如一张食物图片背景很复杂宽广，而真正有用的食物信息只占了很小的面积如图 1-1 中 (4)，这种情况需要过滤无效的信息，因为无效的图像信息对检索菜谱没有任何帮助。其次，如果本身食物图像和菜谱就不匹配的情况下，如图 1-1 中 (1) 的菜谱是南瓜炖肉而不是披萨，又例如，食物图片是西红柿炒鸡蛋，而对应的菜谱则是韭菜炒鸡蛋，这会对训练带来很大问题，因为本身两者就不是匹配对。如图 1-1 中 (2)、(3) 有关食物的有效区域太小，大部分都是无用的背景信息。我们需要一种有效的手段来筛选这种噪声图像，现如今也缺少这种方面的处理策略。



图 1-1 Recipe 1M 食物图像噪声

(3) 三重损失函数存在可改进的空间。传统的三重损失关注了每一个匹配对，并且分配他们相同的权重去完成训练任务。但是在实际使用中，三重损失如果更有倾向性和选择性，那么训练效率和准确度可以大大提高。通过阅读文献我们认识到，传统的三重损失对已经完成优化的匹配对还进行多次训练学习，

这是一种无意义的行为，如果在数据量很大的情况下，这会很大的增加训练负担。同时，训练网络时，并不是所有的匹配对信息都是有意义的，要有选择的去学习。现在已经有研究人员发现，使用 **hard example mining** 的三重损失效率更高，结果的准确度也更高，因此对于三重损失的优化任务还需要进行。

(4) 食谱-食物图像训练数据本身存在食物图像和食谱文本不匹配的噪声，但并没有研究人员对这部分进行研究。多种类型的食谱-食物图像噪声，会严重影响模型训练时的泛化能力，并且其大量存在于 **Recipe 1M** 数据集中，是不可忽视的。除去在第二点提到的噪声以外，本文后续研究还遇到了一种比较难处理的噪声类型。例如，如果白切鸡的菜谱描述详细，但是对应的食物图像中并没有任何食物的语义信息，也就是没有鸡。又如果菜谱上传者为了帮助大家做菜，在制作步骤中详细写了很多配料相应的价格，虽是善意之举，但影响跨模态检索框架训练。这种噪声通过单纯过滤可能无法很好消除其影响，还需要一个筛选机制去除它们。

1.5 研究内容

针对上述提到的现存缺陷，本文从实体角度出发，提出了基于句子级别和强化噪声过滤的食谱-食物图像跨模态检索模型 (**SBNR** 和 **SBNR⁺**)。本文着重研究食谱-食物图像间的跨模态检索关系，下面将详细介绍本文主要研究内容。

针对菜谱中包含的菜谱标题信息，配料信息以及制作步骤信息充分利用，关注句子级别的嵌入，结合食物图片的噪声过滤和改进，损失函数的改进，基于食物图像识别的噪声筛选打分组件研究，达到强化跨模态对齐效果，增强食谱-食物图像的跨模态检索性能。具体做出如下研究内容：

(1) 首先，明确此次研究的任务。第一，研究对食谱进行句子级向量嵌入的方法。第二，研究过滤食物图像中噪声的方法。第三，探索更适合食谱-食物图像跨模态检索任务的损失函数。第四，研究食谱-食物图像的各种噪声类型，根据噪声类型提出合理的筛选打分组件，构造更加纯净的训练数据集。通过以上四点策略来增强食谱-食物图像跨模态检索的性能。其次，详细阐述实验数据集，对现如今各个流行的食谱-食物图像跨模态检索基线模型，进行深入分析，包括开创性的模型 **JNE**，引入注意力机制的模型 **ATTEN**，使用双三重损失的模型 **ADAMINE**，引入 **hard example mining** 三重损失策略和对抗生成损失强化模态对齐的 **ACME**，以及使用 **GAN** 网络处理食谱-食物图像的模式 **R2GAN**。对

这些模型的优点逐一列举，研究其模型取得良好表现的关键因素。同时，根据各基线模型在 Recipe 1M 上的实验结果详细分析每个模型存在的不足，为本文后续提出基于 (SBNR 和 SBNR⁺)，铺垫良好的研究目标和方向。

(2) 提出基于句子级别食谱文本编码的食谱-食物图像跨模态检索框架 (SBNR 框架)。食谱嵌入方面，提出多层句子级别注意力机制组件。使用句子级编码模型分别菜谱标题、菜谱配料、菜谱制作步骤编码。通过注意力机制筛选包含多层隐层信息的三种嵌入向量，拼接三种嵌入向量，通过全连接层将其映射到高维公共空间。食物图像方面，提出食物图片过滤组件，使用拥有丰富语义的菜谱句子级嵌入向量，结合注意力机制对食物图像中的噪声过滤。模态对齐方面，提出模态对齐组件，利用对抗生成思想，将食谱-食物图像两模态分别作为对抗生成思想中的“生成器”和“鉴别器”。生成器目的是产生以假乱真的数据从而骗过鉴别器，而鉴别器的目的就是识别出生成器生成数据的真假，两者相互对抗，共同进步，对抗中有发展，发展中有对抗，相辅相成。当使用对抗生成的损失函数来进行训练时，就可以将相辅相成的两方替换成食谱向量和食物图片向量，通过学习强化模态分布对齐。跨模态学习方面，提出跨模态学习组件。跨模态学习组件由两种三重损失函数组成，分别是 hard example mining 策略三重损失和食谱-食物图像跨模态检索任务特化三重损失。通过对比基线模型，实际检索效果以及消融实验来阐述本文提出的框架的优缺点，以便后续学习研究。

(3) 为了进一步优化本文提出的框架，在多层句子级食谱编码的语义信息提取，以及食物图像噪声的过滤，本文进行了后续的研究，提出了改进后的 SBNR⁺ 框架。鉴于检索框架的部分性能受限于注意力机制结构不足以捕获更细粒度的语义信息，本文进行了基于多头注意力机制的框架优化。更换原框架中使用的注意力机制，由于多头注意力机制存在多个头部，在训练时初始化为不同的权重并且各自独立训练学习，所以可以允许模型在不同的表示子空间里学习到相关的信息。针对两种类型食谱-食物图像匹配对噪声，提出了基于食物图像区域检测的噪声筛选打分组件。两种类型噪声分别为，食物图像中存在的噪声，常见情况为图像中背景复杂而食物区域较小。食谱-食物图像匹配对并不匹配的噪声，常见情况为匹配对食谱和图像描述不同菜品，或者食物图像中并没有相关食物图像的区域。使用食物图像区域检测手段结合菜谱的语义信息共同过滤食物图像，通过打分筛选策略过滤不利于模型训练的噪声。

第 2 章 任务及基线模型

本章将从任务描述、数据集、评价指标、基线模型分析以及实验结果对比分析这些方面，对食谱-食物图像跨模态检索任务予以讨论。通过分析食谱-食物图像跨模态检索任务中有代表性的模型，引出如今此领域中存在的不足，确定第三章的研究内容。

2.1 任务描述

食谱-食物图像跨模态检索框架的通用框架结构如图 2-1 所示，本文着重于研究食谱-食物图像间双向的跨模态检索研究。对于两个模态的输入，食谱和食物图像两模态中，食谱包括了三种有效信息，食谱标题，食谱制作步骤，食谱配料。所要完成的任务就是分别通过食谱编码嵌入，食物图像编码嵌入得到模态的嵌入向量，通过全连接层的作用，将两模态编码嵌入向量映射到相同维度的潜在向量空间中。再经过有效的损失函数和各种有效的模态对齐手段，保证两模态间具有对应关系的食谱-食物图像嵌入向量，在高维度空间中比不对应的其他食谱-食物图像的欧氏距离更近。例如，有一个西红柿炒鸡蛋的菜谱，以及一张西红柿炒鸡蛋的正面写真，将两模态映射到高维度潜在空间时，我们要采用各种手段保证，西红柿炒鸡蛋的照片和其菜谱是紧密相连在一起的，要比西红柿炒鸡蛋照片和炸鸡柳的菜谱或者是回锅肉的菜谱之间的距离更近。

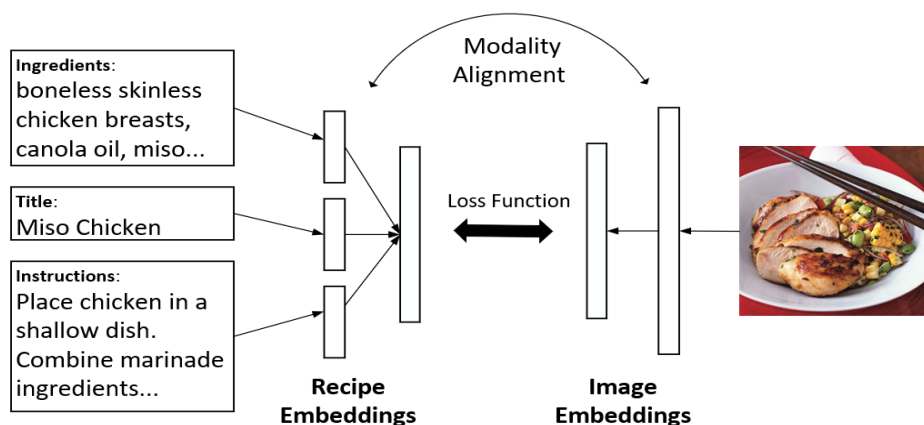


图 2-1 食谱-食物图像跨模态检索通用框架结构

2.2 Recipe 1M 数据集介绍

Recipe 1M^[32]自从2017年由Salvador团队发布之后便成为食谱-食物图像方面研究最常用的数据集^{[32][33][35][39][40][46][47]}。其基本数据构成信息如下表2-1所示：

表 2-1 Recipe 1M 数据构成信息

数据集划分	Recipes	Images
训练数据集	720,639	619,508
验证数据集	155,036	133,860
测试数据集	154,045	134,338
总计	1,029,720	887,06

（1）数据规模与来源

在Recipe 1M^[32]数据集推出之前，几乎所有现成的食品相关数据集只包含食物的分类图像^[52-55]，或只包含食谱配方文本^[56]。直到近几年才有一些数据集被发布，包括食谱和图片。其中Food-101^[57]有101K个图像被平均划分在101个类别中；但是每个类别的配方都是原始的HTML格式，想要使用需要进行额外的处理，并不是很方便。在此后的工作中，陈和Ngo^[58]提出了一个数据集，其中包含110241个图像，标注了353个配料标签和65284个配方，每个配方都有简要介绍、配料列表和制作步骤说明。值得注意的是，数据集只包含中国菜的食谱，这也可以为之后进行中文的食谱-食物图像检索任务提供可能。

尽管上述数据集为研究员们研究食谱-食物图像研究提供支持，但以上提到的数据集不论是在通用性和规模上都仍然有限制。由于学习有效表达的能力在很大程度上取决于可用数据的数据量和质量，Salvador团队创建并公开发布的一个大规模结构化食谱-食物图像数据集，其中包括超过100万份食谱和80万张图像^[32]。与该食谱-食物图像计算领域目前最大的数据集相比，Recipe1M包含的食谱量是^[56]的两倍，图像数量是^[58]的八倍。

（2）数据质量与分布

数据集中的配方平均由9种配料成分组成，由于数据来源于各大美食分享网站，均为一些难度适中的家常美食，大约一半的食谱都有图像，这些食物图像描述了完全做好后的菜肴。Recipe 1M包括大约0.4%的重复菜谱和2%的重复食物图像（不同配方可能共享相同的图像）。排除这0.4%的食谱，20%的食

谱有非唯一的食谱标题。0.2%的食谱共用相同的配料，但相对简单（如意大利面、麦片），中位数为 6 种配料。数据集的作者表示，小心地删除了任何完全相同的重复或食谱共享相同的图像，以避免训练和测试子集之间的重叠。如表 2-1 所示，大约 70%的数据被标记为训练，其余数据在验证集和测试集之间平均分配。

在图 2-2 中，可以很容易地观察到菜谱的数据的分布是重尾的。例如，在已确定的 16k 种独特成分中，常用的 4000 种占了 95%数量。在另一端是冗长的食谱和配料表与食谱，其中包括子食谱。类似的离群值问题也存在于图像中：由于所包含的一些配方集管理用户提交的图像，像巧克力饼干这样的流行配方的图像比平均值多几个数量级。值得注意的是，25%的图像与 1%的配方相关，而所有图像的一半属于 10%的配方；第二层的大小（唯一配方的数量）为 333k。

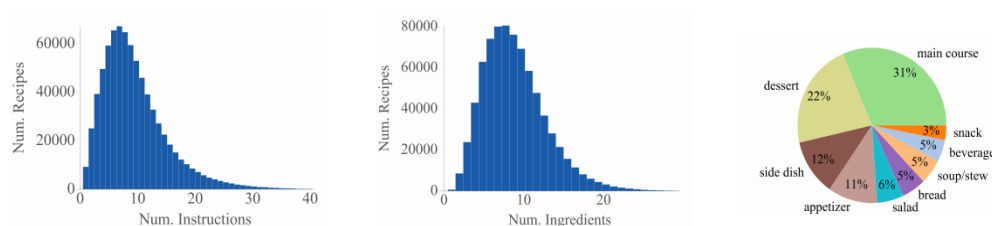


图 2-2 Recipe 1M 中菜谱相关数据展示

2.3 评价指标

评测食谱-食物图像跨模态检索模型的指标分为两个，MedR 和 R@K。MedR 是 medium rank 的缩写，理解为模型在测试子集中取得结果的中位数值，反映了模型整体的命中稳定性。R@K 是 rank at top K 的缩写，理解为模型在测试子集中前 K 个数据对中可以命中正确匹配对的百分比。其中 R@K 通常设置为 R@1，R@5，R@10 三种参数。

具体来说，首先在测试集中取 1000 个食谱-食物图像数据对（食谱和食物图像成对被选择到测试集）的 10 个不同子集和 10000 个食谱-食物图像数据对的 10 个不同子集进行采样。然后，将一个模态中的每一项视为一个查询（例如，一个食物图像），并根据查询嵌入和候选嵌入之间的欧式距离对另一模态中的实例（例如，一个食谱）进行排序。通过欧式距离进行检索，用标准度量来评估跨模式检索任务的性能。对于之前采样的每个测试子集（1k 和 10k），计算

MedR。同评估前 K 的召回率，即匹配项在前 K 个结果中排名的查询百分比。

2.4 基线模型分析

本小节针对主要的基线模型进行介绍和分析，讨论各基线模型在食谱-食物图像跨模态检索任务中的优势。

2.4.1 JNE 模型

2017 年，Amaia Salvador 等人提出 Recipe 1M 数据集^[32]。作为现如今最大公共可用的食谱-食物图像数据集，Recipe 1M 提供了在食谱和食物图像跨模态数据上训练检索任务的能力，也是本文主要讨论的数据集。Amaia Salvador 团队还通过对 Recipe 1M 数据集的研究，提出了影响深远的一个食谱-食物图像跨模态检索框架，如图 2-3 所示。

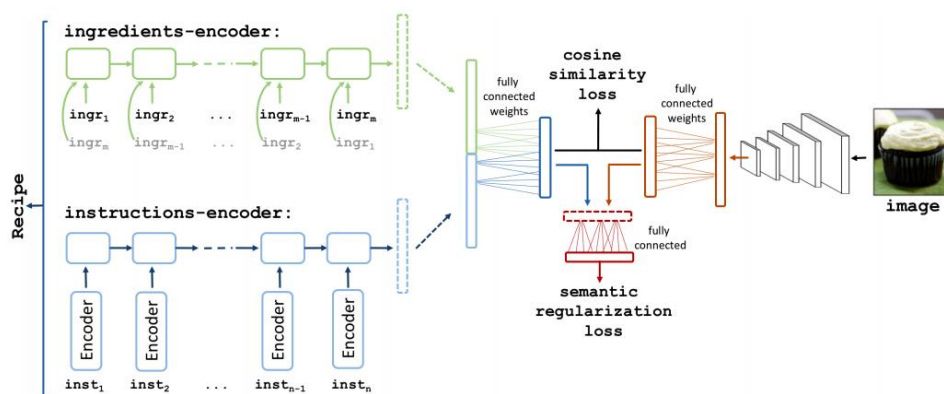


图 2-3 JNE 模型总体结构^[32]

从结构上看，JNE 检索框架与前文提到的，通用食谱-食物图像跨模态检索框架外形相似，原因就是通用的食谱-食物图像跨模态检索框架，就是从 JNE 发展起来的。JNE 框架对于食谱信息通过两个编码器结构进行编码，一个是食谱配料编码器，一个是食谱制作步骤编码器。两编码器通过 word2vec 结合双向 LSTM 对制作步骤和配料进行编码嵌入，作者表示使用双向 LSTM 代替 LSTM 的原因是，菜谱的制作步骤是很长的，平均 208 个单词，仅仅使用单个 LSTM 是无法表示如此长度的句子。作者的这个提示对后续的研究者是一个很大的警告，需要充分考虑菜谱制作步骤的长度问题。食物图像方面，作者使用了 Resnet

网络对食物图像进行嵌入编码。最终得到的食谱和食物图像嵌入编码，经过全连接层映射到高维度的公共空间中。JNE 模型在损失函数方面，使用了两种损失，第一种是余弦损失，第二种是语义正则化损失。

当食谱-食物图像对是匹配对即 $y=1$ 时，余弦损失表示为公式 (2-1)：

$$1 - \cos(\Phi^R, \Phi^v) \quad (2-1)$$

当食谱-食物图像对是不匹配对即 $y=-1$ 时，余弦损失表示为公式 (2-2)：

$$\max(0, \cos(\Phi^R, \Phi^v) - \alpha) \quad (2-2)$$

余弦损失的思想，匹配的食谱-食物图像对相似度在高维度公共空间中不断学习提高，不匹配的食谱-食物图像对在高维度公共空间中低于阈值 α 时，不做处理，余弦相似度在 JNE 模态对齐中起到了至关重要的作用。语义正则化损失可以理解为，JNE 框架在高纬度公共空间中，将匹配的食谱-食物图像对均经过共享权重的全连接层，保证框架在不断训练学习时，提供一定的模态对齐能力。最终的损失函数通过两部分损失完成学习任务，如公式 (2-3)。 λ 表示权重参数，为了一定程度上限制语义正则化损失，作者认为 JNE 框架主要是余弦损失保证模态对齐，语义正则化损失只是辅助作用，不能作为主要学习目标去训练。

$$L(\Phi^r, \Phi^v, c_r, c_v, y) = L_{\cos}((\Phi^r, \Phi^v), y) + \lambda L_{reg}(\Phi^r, \Phi^v, c_r, c_v) \quad (2-3)$$

JNE 模型还进行了一个有趣的验证学习结果语义性的实验，这个实验足以引发后续研究者大量思考。这个实验简单介绍就是，如果对学习得到的食谱嵌入向量或者菜谱嵌入向量，进行加减操作会得到什么结果？举个例子，如果用芝士沙拉图片嵌入向量，减去沙拉图像嵌入向量，加上蛋糕图像嵌入向量会得到芝士蛋糕的图像嵌入向量。这是一个有趣且引人深思的小实验，证明了学习到的向量存在语义特性。

2.4.2 ATTEN 模型

2018 年 Jing-jing Chen 等人提出 ATTEN 模型^[39], 这个模型将注意力机制同时用在了菜谱的标题, 配料, 制作步骤, 通过不断学习获得注意力权重, 来加强对菜谱的信息筛选。图中可以看到, ATTEN 模型将双向 LSTM 改为了双向 GRU。ATTEN 模型也将菜谱的标题加入到了嵌入编码过程中, 这个思路是有效的, 因为一个菜谱中的分类信息大多数就包含在菜谱标题中。例如, 炸油糕, 炸鸡排归类为炸制类食谱, 回锅肉, 鱼香肉丝归类为川菜类食谱等。损失函数以及图像的处理基本上沿用了 JNE 模型的思想。ATTEN 模型的整体结构如图 2-4。

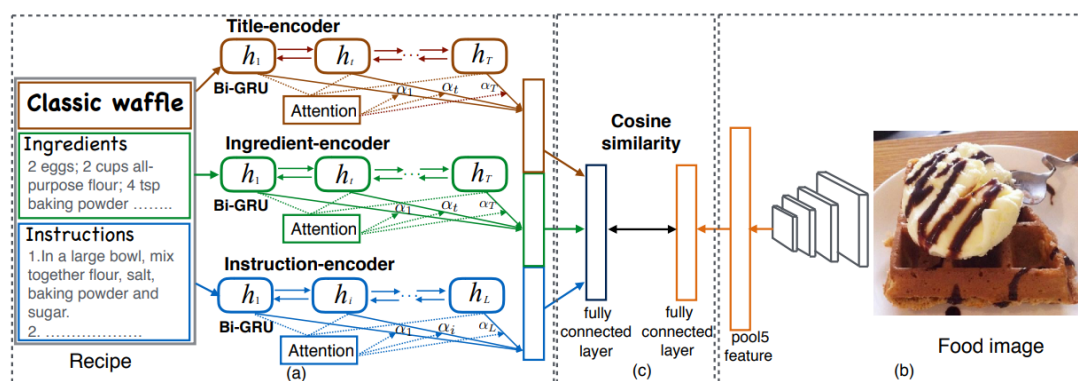


图 2-4 ATTEN 模型总体结构^[39]

2.4.3 ADAMINE 模型

Micael Carvalho 等人提出 ADAMINE 模型^[33]。首先, 将对齐问题建模为一个利用检索和基于类的特征的联合学习框架, 直接在损失中集成语义信息, 以细化潜在空间的结构, 同时限制需学习的参数数量。其次, 将模型依赖于双三重损失来适应联合学习目标。第三, 提出了一个新的随机梯度下降加权方案, 它在小批量上计算并自动执行自适应挖掘。ADAMINE 模型结构如图 2-5 所示。

ADAMINE 侧重点在损失函数部分, 不管是对食谱的处理还是对食物图片的处理, 并没有在 JNE 基础上进行改进。ADAMINE 使用双三重损失, 此损失函数对后来的食谱-食物图像跨模态检索研究起到了很大的影响。相比于余弦损失只关心匹配对的学习情况, 忽视了对不匹配对的处理。三重损失函数的突出之处在于, 它同时还关注不匹配对在高维度公共空间的分布情况。简单来说就是当学习对象是匹配对时, 三重损失函数会将其两者拉近距离, 不匹配对在学习

过程中会被拉远距离。这种做法就避免了余弦损失的缺陷，虽然拉近了匹配对，但是不匹配对却更加靠近的情况。三重损失相比于余弦损失，使得检索结果得到了很大的提高。

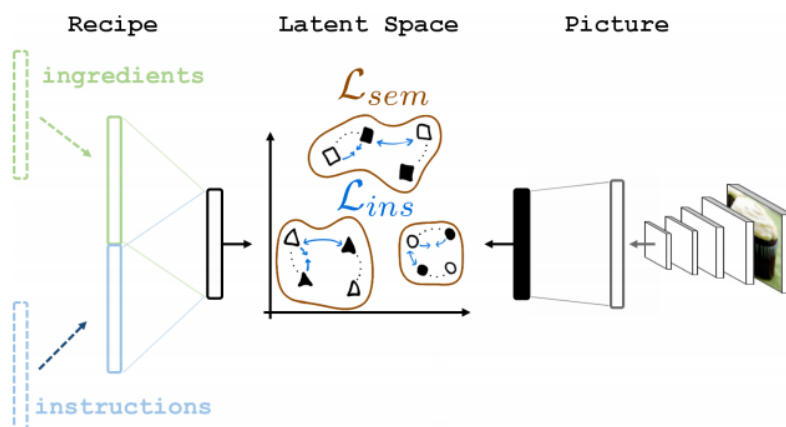


图 2-5 ADAMINE 整体结构图^[33]

2.4.4 ACME 模型

Hao Wang 等人提出了 ACME 模型^[35]，作者研究了烹饪菜谱与食物图像的跨模态检索任务，提出了一种新的框架-对抗性跨模态嵌入模型 ACME 来解决食品领域的跨模态检索问题。ACME 模型的结构如图 2-6 所示。

从图 2-6 可以看到，ACME 模型分为模态对齐模块，权重共享模块，跨模态检索学习模块以及翻译一致性模块。ACME 处理食物图像时使用了前文提到所有模型相同的方式，没有改进。ACME 处理菜谱过程中，没有处理菜谱标题，同样延续了前人的成果使用了 LSTM 模型。

模型中的第二个模块，跨模态检索学习模块是文本的重要创新点。前文提到，食谱-食物图像跨模态检索一路发展下来，从余弦损失到三重损失，都存在缺陷和不足。基础的三重损失对所有学习数据分配相同的权重，这可能会严重阻碍模型的泛化，因为同一食谱对应的多幅食物图像之间可能存在较大的方差。为了解决三重损失所存在的缺陷，ACME 使用了 hard example mining 策略的三重损失。此策略选择极端情况，最大化训练梯度。相比于三重损失，hard example mining 采取优先选择高纬度公共空间中，距中心点最远的正样本实例和最近的负样本实例，目的是希望最不相似的正样本实例与中心点可以通过学

习进行对齐，而与中心点最相似的负样本实例可以通过学习变成不相似。此策略使训练更加有目的性，收敛迅速，效果也十分优秀。

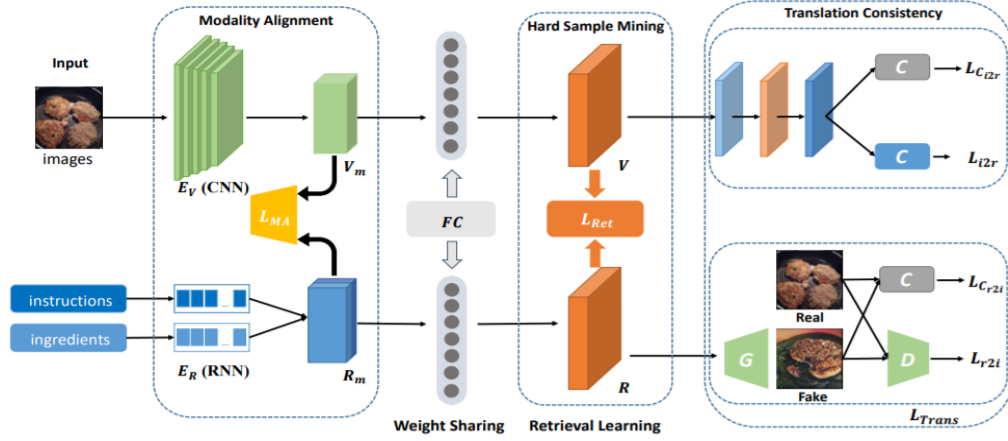


图 2-6 ACME 模型整体结构^[35]

这里 ACME 为了保证跨模态嵌入特征的对齐，使用了对抗生成思想，公式 (2-4) 如下：

$$\begin{aligned} \mathcal{L}_{MA} = & \mathbb{E}_{\mathbf{i} \sim p_{\text{image}}} [\log D_M(\mathbf{E}_V(\mathbf{i}))] + \\ & \mathbb{E}_{\mathbf{r} \sim p_{\text{recipe}}} \left[\log(1 - D_M(\mathbf{E}_R(\mathbf{r}))) \right] \end{aligned} \quad (2-4)$$

$$\min_{\mathbf{E}_V, \mathbf{E}_R} \max_{D_M} \mathcal{L}_{MA}$$

可以看到，公式中通过采用对抗生成思想，来保证食谱和食物图像嵌入向量通过鉴别器 D_m 处理后，促进其两模态的数据分布对齐。

模型中的第三个模块，跨模态翻译一致性模块。此模块使得一种模态的嵌入能够恢复另一种模态中对应实例的一些重要信息，确保学习到的特征表示可以保留跨模态的信息。

模型中的跨模态学习模块公式如 (2-5) 所示，作者结合三种损失，通过设置偏置项来进行学习，平衡各损失的关系。

$$L = L_{Ret} + \lambda_1 L_{MA} + \lambda_2 L_{Trans} \quad (2-5)$$

2.4.5 R2GAN 模型

2019 年 Bin Zhu 等人提出了 R2GAN 模型^[40]，表示过程文本（如跨模态检索的食谱问题）本身就是一个难题，更不用说从食物食谱生成食物图像了。本文研究了一种新的对抗生成网络，称为食物食谱检索对抗生成网络（R2GAN），以探索从过程文本中生成图像以解决检索问题的可行性。

R2GAN 模型如图 2-7 所示，分为四个模块食谱嵌入学习模块，食物图像嵌入学习模块，语义学习模块，GAN 学习模块。食谱嵌入学习模块中，R2GAN 采用了 LSTM 结合层次 LSTM 两种结构对食谱进行处理。对嵌入后的菜谱向量拼接，映射到公共高纬度空间。食物图片嵌入学习模块中，并未出现改进。语义学习模块是借鉴了 JNE 模型中的思想。R2GAN 重点模块是 GAN 学习模块，R2GAN 的新颖性来源于其结构设计，特别是采用了一个生成器和双鉴别器的对抗生成网络，使得从配方中生成图像成为一种可行的思路。

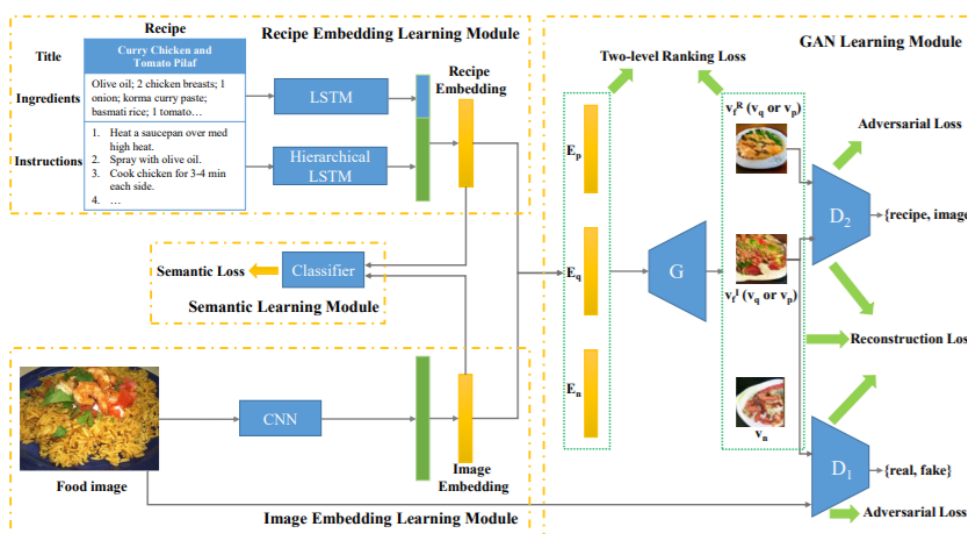


图 2-7 R2GAN 模型整体结构^[40]

此外，在生成的图像下，考虑了嵌入和图像空间中的 two-level ranking loss。这些附加组件不仅带来了出色的检索性能，而且生成了接近真实的食物图像，有助于解释食谱的排名。在 Recipe 1M 数据集上，R2GAN 显示了对数据大小的高度可伸缩性，并生成直观的图像供人类解释搜索结果。

2.5 基线模型实验结果及分析

表 2-2 展示了本章中着重介绍的五个基线模型在 Recipe 1M 数据集上的食物图像→食谱文本的检索实验结果，表 2-3 展示了五个基线模型在 Recipe 1M 数据集上的食谱文本→食物图像的检索实验结果，测试数据均来自于 1K 测试子集。10K 测试子集上的检索实验结果与 1K 测试子集上的实验结果趋势保持一致，所以这里做省略处理。从表中可以清楚看到，ACME 取得了最好的结果。本小节中，根据各基线模型的实验结果逐个分析各基线模型。

表 2-2 食物图像→食谱文本的检索实验结果

评价指标 基线模型	MedR	R@1	R@5	R@10
JNE ^[32]	5.2	25.6	51.0	65.0
ATTEN ^[39]	4.6	25.6	53.7	66.9
ADAMINE ^[33]	2.5	39.8	69.0	77.4
R2GAN ^[40]	2.0	39.1	71.0	81.7
ACME ^[35]	1.0	51.8	80.2	87.5

表 2-3 食谱文本→食物图像的检索实验结果

评价指标 基线模型	MedR	R@1	R@5	R@10
JNE ^[32]	5.1	25.0	52.0	65.0
ATTEN ^[39]	4.6	25.7	53.9	67.1
ADAMINE ^[33]	2.1	40.2	68.1	78.7
R2GAN ^[40]	2.0	40.6	72.6	83.3
ACME ^[35]	1.0	52.8	80.2	87.6

2.5.1 JNE 模型

JNE 框架提出了 Recipe 1M 并且进行了跨模态检索研究，但是由于是开创性的尝试，存在如下三点不足：

(1) 单词级别的 word2vec 操作以及双向 LSTM 对菜谱的学习，是很难以保证菜谱的顺序性信息的。菜谱最重要的部分就是制作步骤，而制作步骤之所以叫做步骤，就是因为其存在严格的顺序性。例如，农家小炒肉制作中，经常步骤是热油，下干辣椒，花椒等爆香，然后加入配菜炒制。如果顺序出现问题，先下干辣椒花椒爆香，随后加入配菜最后加入食用油。这样轻则饭菜糊锅，重则干辣椒花椒被炒干燃烧有安全危险。所以 JNE 在制作步骤中，每一句与每一句的关联没有很好的处理。

(2) Recipe 1M 中食物图像数量庞大，其中存在大量噪声图片。JNE 并没有考虑对噪声图片进行处理，也没有考虑当食谱和食物图像对并不匹配但却在数据集中表示为同一组时，怎样筛选去除。

(3) 发挥很大作用的余弦损失，在高纬度公共空间中，虽然可以一定程度拉近匹配对，但是并没有对不匹配的数据进行拉远操作。余弦损失的训练效果相比后续研究人员的模型来说并不出色，可能导致这一问题的原因是，高纬度公共空间中，不匹配对的数据和噪声数据不做处理时，其总会比拉近的匹配对数据更加靠近，相似度更高。这样的缺陷也导致后续研究者大量使用三重损失这个方法。

2.5.2 ATTEN 模型

虽然 ATTEN 模型跨检索能力相比 JNE 模型提高甚微，但引入注意力机制对于后续研究者起到指导作用。注意力机制在这里的引入，是一个很好的思想，原因是，在训练模型中，经常会遇到 Recipe 1M 数据集中的制作步骤是“混合所有配料”，也就是前文中所提到的沙拉菜谱的通常做法。这时，如果不加区分的使用所有菜谱信息，那么制作步骤会成为很大的干扰项，可能导致两完全不同菜谱，在高维度公共空间，因为“混合所有成分”的制作步骤而距离靠近，影响训练结果。如果在处理时加入了注意力机制，可以通过模型的训练，降低此部分的注意力权重，减小干扰项对跨模态检索的影响。

2.5.3 ADAMINE 模型

ADAMINE 模型使用的双三重损失是一个良好策略，本小节根据其实验结果，分析它的不足之处如下：

(1) ADAMINE 模型仍然是单词级别的对食谱进行处理，食谱间的顺序问题，食谱间的关系都不能很好的表达在食谱嵌入向量中。

(2) ADAMINE 模型将配料和制作步骤拼接成向量之后，并没有利用食谱标题来嵌入食谱向量，由于菜谱标题很大程度上表明了菜谱的类别和特色，所以在处理过程中重要程度不言而喻。

(3) 双三重损失虽然在跨模态检索任务上比较有效，但是对于训练数据给予相同的训练权重，经常会忽视重要数据的训练影响力，这可能会严重阻碍模型的收敛和泛化，因为同一食谱对应的多幅食物图像之间可能存在较大的方差。后续研究员们使用了 **hard example mining** 策略的三重损失，此策略可以有效解决此点缺陷，不过也带来了新的不足，本文会在后续篇章中讨论。

(4) 食物图像处理时并没有考虑噪声问题，而食食物图像噪声和谱-食物图像匹配对噪声对训练结果的影响是不可忽视的。

2.5.4 ACME 模型

ACME 虽然在 Recipe 1M 数据集上取得了极为优异的成绩，验证了其中使用的思路是有效的，但是也存在一些不足和缺陷。

(1) ACME 模态对齐模块中，食谱和食物图像嵌入手段没有改进。如前文一直提到，这样做缺少对制作步骤间顺序性信息的学习，食物图像中可能存在的噪声也并未过滤。ACME 没有在菜谱嵌入时使用食谱标题，这样会缺少一些菜谱的类别信息，而这些类别信息在类似沙拉的菜谱中是尤其重要的，因为沙拉菜谱中配料经常十分相似，制作步骤常出现“混合所有配料”。

(2) ACME 跨模态检索学习模块中，使用了 **hard example mining** 策略的三重损失。考虑如下情况，如果模型学习过程中，选择了距离中心点最近的负样本实例，而此实例是噪声数据，那么训练出的模型将获得一个错误的梯度下降方向，极大的影响了训练结果。

(3) ACME 使用了翻译一致性模块，此模块从文中描述来看，它并没有发挥比较有效的作用。作者也缺少了消融实验来表明此机制的有效性。

2.5.5 R2GAN 模型

从实验结果可以看出, R2GAN 取得了良好的检索质量。R2GAN 使用 GAN 的动机可能有两方面, 以对抗的方式学习兼容的跨模态特征, 通过展示从菜谱生成的食物图像来解释搜索结果。可以认为食物菜谱图像的检索也进入了对抗生成时代。首先, 对于图像中的噪声问题, 可以采用过滤以及筛选的手段来去除, 同样也可以通过对抗生成的方式生成食物图像, 这样既解决了图像中大量的噪声问题, 也解决了检索时十分缓慢的问题。

2.6 本章小结

本章主要介绍了任务描述, 相关数据集和评价指标, 详细讨论各基线模型以及实验结果分析。通过对比各基线模型的优缺点, 来了解食谱-食物图像跨模态检索的发展状况, 明确研究目标和任务。同时, 针对分析出各基线模型不足, 食物图像噪声问题, 食谱文本嵌入问题, 模态对齐问题以及跨模态学习问题, 后续章节本文会针对这些不足进行逐一解决。

第3章 基于句子级编码的食谱-食物图像跨模态检索框架

在第二章中，本文分析讨论了各大流行的食谱-食物图像跨模态检索模型，分析各基线模型的优势，并从实验结果讨论它们的不足。单词级别嵌入编码对于菜谱制作步骤的顺序性信息覆盖不足，食物图片嵌入编码没有考虑噪声问题，模态对齐手段还可以继续优化，三重损失对所有数据分配相同权重进行学习影响模型的泛化能力。针对这些问题，本文提出基于句子级编码的食谱-食物图像跨模态检索框架。此框架旨在通过图像过滤组件，多层句子级注意力网络组件，模态对齐组件以及跨模态学习组件来解决上述问题。同时，为第四章框架优化和基于食物图像检测的噪声筛选打分策略做铺垫。

3.1 问题描述

用 v_i 表示食物图像，用 r_i 表示烹饪食谱， $i=1,2,3\dots$ 。其中 $v_i \in V$ 和 $r_i \in R$ ， V 和 R 分别代表食物图像域和食谱域。本章的任务是通过编码嵌入手段把 v_i 和 r_i 映射到一个共同的高维度潜在空间 L_d ，其中 d 表示维度。在第二章提到的各个流行模型框架中，均完成了这个任务，同时实施各种模态对齐手段强化食谱-食物图像的模态对齐。但是，这些操作忽略了很多重要信息。食物图片中的噪声会严重影响模型的训练，单词级食谱嵌入无法高效捕获制作步骤间的顺序信息，未改进三重损失可能在食谱-食物图像跨模态领域，菜谱食物图像一对多情况下，泛化能力不强。本章首要目标，就是解决上述提到的问题。

3.2 整体框架

如图 3-1 所示，本文提出的食谱-食物图像跨模态检索框架（简称为 SBNR 框架）包含四个部分，分别是图片过滤组件、多层句子级别注意力机制组件、模态对齐组件和跨模态学习组件。我们将菜谱输入到多层句子注意网络组件中，得到菜谱嵌入。在图像过滤组件中，对食物图像进行编码后，利用多层句子注意网络组件中的食谱嵌入对食物图像进行过滤。食物图像和食谱的嵌入信息被输入模态对齐组件，并通过对抗生成思想进行分布对齐。此外，食物图像和食谱的嵌入信息也被输入到跨模态学习组件中，通过改进后的三重态损失函数强

化跨模态对齐。

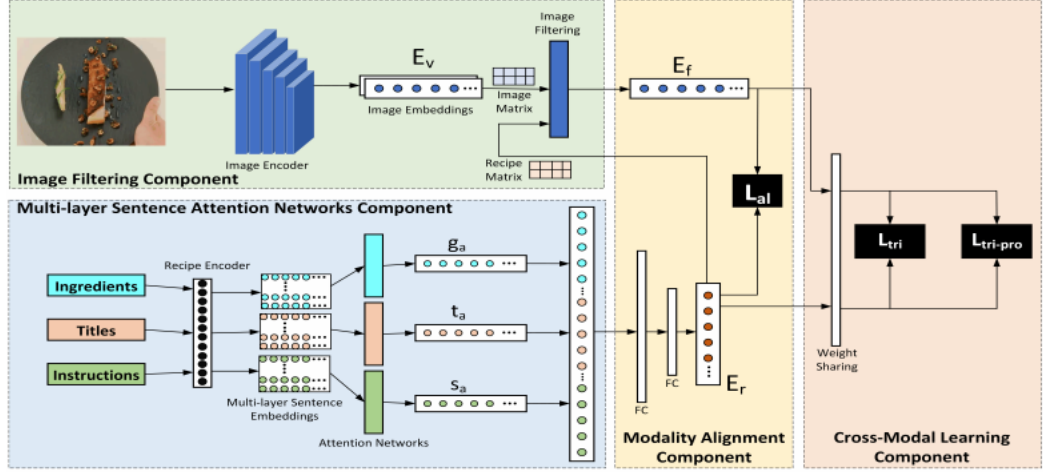


图 3-1 SBNR 框架整体结构

v_i 表示食物图像，用 r_i 表示烹饪食谱， $i=1, 2, 3$ 。其中 $v_i \in V$ 和 $r_i \in R$ ， V 和 R 分别代表食物图像域和食谱域， v_i 和 r_i 组成食谱-食物图像对 (v_i, r_i) 。本章的任务是把 v_i 和 r_i 映射到一个共同的高层次的潜在空间 L_d ，其中 d 是维度。

在多层句子注意网络组件中，本文使用三个注意力网络结合 BERT 来编码 $r_i: F_r(r_i)=E_{ri}$ 。在第 3.3.1 节多层句子级别注意力机制组件中，详细解释如何编码食谱。

在食物图像过滤部分，使用 Resnet-50 将 v_i 编码高纬度潜在空间 $L_d: F_v(v_i)=E_{vi} \rightarrow L_d$ ， E_{vi} 代表 v_i 的食物图像特征嵌入。为了尽量减少食物图像中噪声的影响，如食物图像中不匹配的食谱-食物图像对和无用的占比很大的图像背景噪声信息，通过注意力网络和食谱特征嵌入对食物图像进行过滤操作。将 E_{vi} 和 E_{ri} 输入到注意力网络中，通过由多层句子级别注意力机制组件，提取的含有丰富语义的食谱特征嵌入辅助过滤食物图像，从而减少食物图像中的噪声部分。由食物图像过滤组件输出的特征嵌入表示为 E_{fi} 。在第 3.3.2 节食物图像过滤组件中，详细解释如何处理食品图像中的噪声。

在模态对齐组件中，使用对抗生成思想来确保不同模态的嵌入特征遵循相同的分布。在第 3.3.3 节跨模态对齐组件中解释这种方法。

在跨模态学习组件中，在计算损失函数之前，使高纬度潜在空间的食物图像和食谱特征通过一个全连接层，在该层中，两种模态共享相同的权重^[59]。这

里使用两种三重损失被用来优化本文提出的框架。在第 3.3.4 节跨模态学习组件中，详细说明两种类型的三重损失。

3.3 组件介绍

本节会对本文提出的 SBNR 跨模态检索框架进行详细的介绍，四个组件分成四个小节。3.3.1 详细介绍多层句子级别注意力机制组件如何处理食谱制作步骤，食谱标题，食谱配料。3.3.2 详细介绍食物图像过滤组件的工作原理。3.3.3 详细介绍模态对齐组件如何通过对抗生成思想来进行模态分布对齐。3.3.4 详细介绍跨模态学习组件如何通过两种策略的三重损失来进行模态对齐学习。

3.3.1 多层句子级别注意力机制组件

多层句子级别注意力机制组件整体结构如图 3-2 所示，本框架使用食谱编码器 BERT 对食谱配料、食谱标题和食谱制作步骤分别进行多层句子级别文本编码。设计了三个注意力机制分别提取配料、标题和制作步骤的多层句子级编码间各层关系，输出结果进行拼接组成句子级菜谱文本编码，下文将对该组件进行详细的介绍。

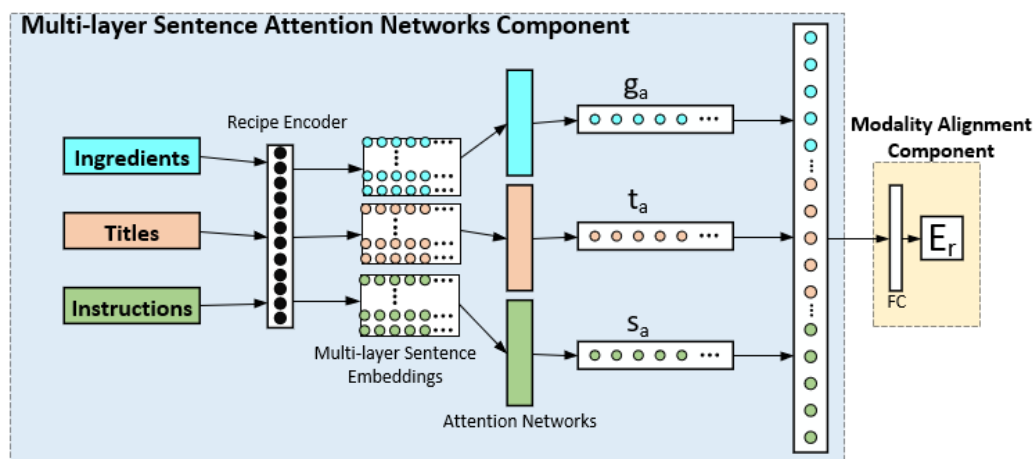


图 3-2 多层句子级别注意力机制组件

对于食谱文本信息，BERT 用于句子级编码可以较好表达句子的语义信息，从而通过语义信息辅助来过滤食物图像的噪声。此组件中的注意力机制使用目的只是用于计算一种合适的多层句子级别编码的权重信息，理解为句子级的表

示学习，从而进行融合来满足过滤食物图像的操作，并不是为了提取食谱句子级别中的类似主谓宾的语义语法成分。

对于一个食谱，有一个食谱的标题 t_i ，食谱的制作步骤 s_i ，食谱的配料 g_i 。可以观察到一个食谱的标题通常都包括分类信息，例如西红柿炒鸡蛋属于炒制类别，戚风蛋糕属于甜品烘焙类别，铁锅炖属于炖菜类别等，同样的食谱的配料顺序和制作步骤顺序也是非常重要，例如，烤制酱香饼时，要先加入油并预热电饼铛，在放入制作好的面团，最后刷上肉酱葱花出锅，但如果顺序错乱，先放入面团刷肉酱烤制再加油，那想必做出来一定是魔鬼料理。当面对制作步骤中只有一句话为“混合所有成分”的食谱时，有关食谱的标题信息和配料信息将显得格外重要，因此需要合理利用食谱中的所有信息，例如在制作沙拉类型的菜品时，不管是鸡肉沙拉，水果沙拉还是减脂营养套餐沙拉，制作步骤都是混合所有成分，这个时候配料和标题更加需要关注。对于 t_i 、 s_i 、 g_i ，分别用 BERT 进行句子级别的编码，可以得到食谱标题嵌入 t_{ei} 、食谱制作步骤嵌入 s_{ei} 和食谱配料嵌入 g_{ei} ，它们的特征嵌入中都包含多层句子级信息。鉴于 Bert 中低层包含表面特征、中层包含句法特征和高层包含语义特征，分别用三个注意网络来处理 t_{ei} 、 s_{ei} 和 g_{ei} 。

如上所述，对于烹饪食谱，BERT 被用来编码 t_i ， s_i ， g_i ，得到 t_{ei} ， s_{ei} ， g_{ei} ，它们都包含多层句子级信息。食谱的标题 t_i 虽然简短，但包含了食品的分类信息。因此，将 t_i 看作一个句子，并使用一个标记器来获得 t_i 的标记。之后，标题标记被转换为分段 id。ID 用于区分不同的句子。最后，利用食谱标题注意力网络从多层隐藏状态中筛选提取食谱标题信息。食谱标题注意力网络的表示如式 (3-1)，(3-2)，(3-3) 所示：

$$u_i = \tanh(W_m^T t_{e_i} + b_m) \quad (3-1)$$

$$\alpha_i = \frac{\exp(\omega^T u_i)}{\sum_n \exp(\omega^T u_j)} \quad (3-2)$$

$$t_{a_i} = \sum_n \alpha_i t_{e_i} \quad (3-3)$$

这里 W_m^T 和 ω^T 是注意力变换矩阵， b_m 是偏置项。输入食谱标题嵌入 t_{ei} ，其中 $i=1, 2, 3, \dots$ ， i 表示批大小。 t_{ai} 表示食谱的标题注意力网络所输出的食谱标题嵌入信息。

众所周知，对于一个食谱来说，食谱的制作步骤是最为重要的部分。排除少数制作步骤为“混合所有成分”的情况，在大多数情况下，统计发现每个制作步骤至少包括十句话去描述详细制作过程。设置特殊的标记来标记句子的开始 ([CLS]) 和句子分隔/句子结束 ([SEP])。后续操作与处理 t_i 时相同。注意力网络的表示如式 (3-4)，(3-5)，(3-6) 所示：

$$u_i = \tanh(W_m^T s_{e_i} + b_m) \quad (3-4)$$

$$\alpha_i = \frac{\exp(\omega^T u_i)}{\sum_n \exp(\omega^T u_j)} \quad (3-5)$$

$$s_{a_i} = \sum_n \alpha_i s_{e_i} \quad (3-6)$$

这里 W_m^T 和 ω^T 是注意力变换矩阵， b_m 是偏置项。输入食谱标题嵌入 s_{e_i} ，其中 $i=1, 2, 3, \dots$ ， i 表示批大小。 s_{a_i} 表示食谱的标题注意力网络所输出的食谱标题嵌入信息。

对于一个食谱的配料信息 g_i ，在食谱配料的出现顺序和它们在制作步骤中，出现顺序进行一对一的对应关系。这样做的目的是，让配料信息可以包含更多的制作步骤的顺序信息。众所周知，做菜时配料的使用是循序渐进的。例如，在制作本地特色的烧茄子时，配料的加入顺序应该严格按照加油，茄子，青辣椒，盐鸡精香醋，西红柿，蒜末，如果出现顺序颠倒，蒜末的香气和半熟的西红柿的香气就会损失，导致菜品样貌欠佳，水分较多味道欠佳。由于配料之间顺序的重要性，为每个配料设置了特殊的标记，以便 BERT 可以学习食谱中配料之间的顺序信息。后续操作与处理 t_i 时相同。配料的注意力网络的表示如式 (3-7)，(3-8)，(3-9) 所示：

$$u_i = \tanh(W_m^T g_{e_i} + b_m) \quad (3-7)$$

$$\alpha_i = \frac{\exp(\omega^T u_i)}{\sum_n \exp(\omega^T u_j)} \quad (3-8)$$

$$g_{a_i} = \sum_n \alpha_i g_{e_i} \quad (3-9)$$

这里 W_m^T 和 ω^T 是注意力变换矩阵， b_m 是偏置项。输入食谱标题嵌入 g_{ei} ，其中 $i=1, 2, 3, \dots$ ， i 表示批大小。 g_{ai} 表示食谱的标题注意力网络所输出的食谱标题嵌入信息。

综合菜谱的标题，配料，制作步骤嵌入信息，可以表示一个菜谱的综合嵌入信息为 (3-10)：

$$E_{r_i} = FC\left(\left[s_{a_i}, t_{a_i}, g_{a_i}\right]\right) \quad (3-10)$$

这里将食谱制作步骤，标题，配料拼接成一个向量，经全连接层 $FC()$ 的处理，食谱嵌入会被映射到一个潜在的高维度空间 L_{1024} ，表示高维度空间的维度是 1024 维。

3.3.2 图像过滤组件

本文提出的食物图像过滤组件如图 3-3 所示。其目标是减少食谱-食物图像中不匹配的图像食谱对和无用的噪声信息。向一个注意力网络中输入从多层句子级别注意力机制组件得到的菜谱编码 E_{ri} ，并结合 Resnet 输出的食物图片编码 E_{vi} 来训练学习，以达到通过多层句子级别网络输出的富含丰富语义信息的菜谱嵌入，来协助过滤食物图片嵌入，从而过滤掉图片中的噪声信息，输出过滤后的食物图片编码 E_{fi} 。注意力机制最初是在一个序列到序列的配置中提出的，在这个配置中解码器学习它应该注意的哪些些部分，并一步一步地解码一个个单词^[60,61]。

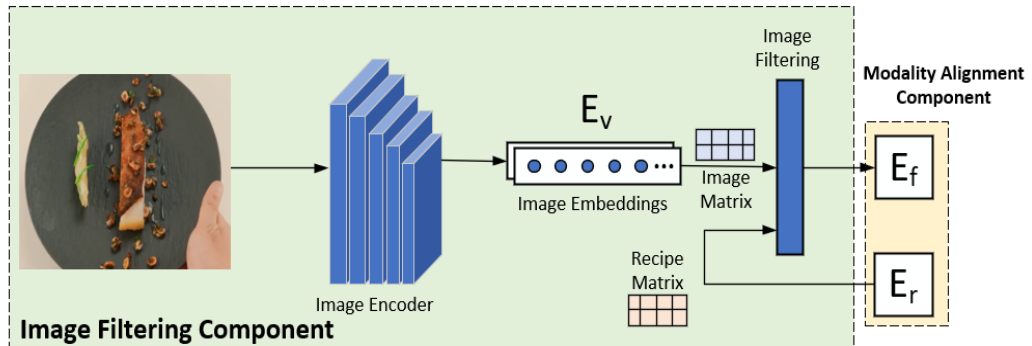


图 3-3 图像过滤组件

在本研究中，利用注意力机制来学习 E_{fi} 与 E_{vi} 之间的权值，以取代解码的目的。这类似于^[37]中的自注意力机制，但不同的是本文从两种不同的模态中学习注意力的权值。图像过滤嵌入 E_{fi} 的定义如式 (3-11)，(3-12)，(3-13) 所示：

$$a_{j,i} = \tanh(W_v E_{v_i} + W_r E_{r_j} + b) \quad (3-11)$$

$$\alpha_{j,i} = \frac{\exp(\omega^T a_{j,i})}{\sum_n \exp(\omega^T a_{j,t})} \quad (3-12)$$

$$E_{f_i} = \sum_n \alpha_{j,i} E_{v_i} \quad (3-13)$$

E_{fi} 是这个图像过滤组件的输出，它表示已经被食谱嵌入联合筛选过的食物图片嵌入。 $\alpha_{j,i}$ 是归一化注意力权重。 E_{fi} 是 E_{vi} 的加权和。 W_v 是食物图像权重矩阵。 W_r 是食谱权重矩阵。 ω^T 是一个变换矩阵， b 是偏置项。基于图像过滤组件，可以从食物图像嵌入中提取有效信息，增强两种模态之间的对齐。在图 3-4 中详细画出其图像过滤组件内部计算的过程。

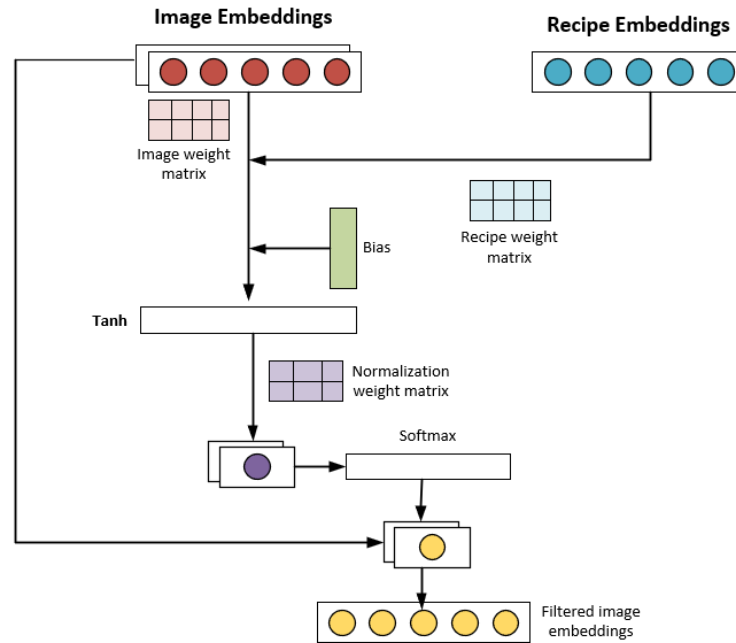


图 3-4 食物图像过滤组件内部计算过程

3.3.3 模态对齐组件

在图 3-5 中展示模态对齐组件的结构,此组件受 WGAN-GP 和 ACME 的启发,如果没有一个强力的模态对齐方法,模型的收敛速度将会较为缓慢。

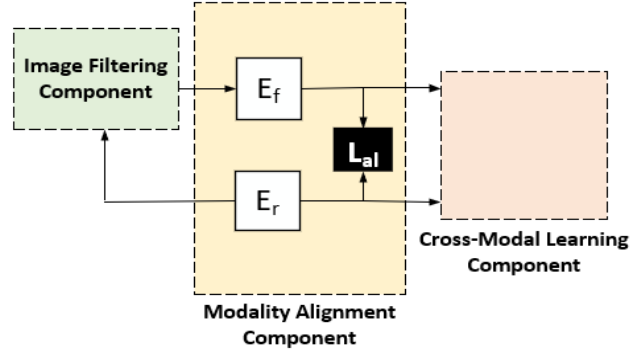


图 3-5 模态对齐组件结构

SBNR 框架使用对抗性损失来加强模态之间的分布一致性。 E_i 和 E_r 被输入到鉴别器 D_s 中, 对齐损失 L_{al} 被定义为等式 (3-14) :

$$L_{al} = \mathbb{E}_{i \sim p_{image}} [\log D_s(E_f(i))] + \mathbb{E}_{r \sim p_{recipe}} [1 - \log D_s(E_r(r))] \quad (3-14)$$

此公式通过如下公式 (3-15) 进行学习优化:

$$\min_{E_f, E_r} \max_{D_s} L_{al} \quad (3-15)$$

在优化过程中, D_s 试图区分 E_r 和 E_f , 但本小节的目标是对齐食物图像和食谱的模态分布, 使 D_s 无法区分出 E_r 和 E_f , 这就是利用了对抗生成的思想。随着训练不断地进行, 当 D_s 确实无法区分出 E_r 和 E_f 的区别时, 可以认为跨模态对齐达到了一个较好的水准, 完成了强制对齐食谱和食物图像跨模态数据对的分布情况。

3.3.4 跨模态学习组件

如图 3-6 所示，跨模态学习组件包括两个损失函数，一个是使用 **hard example mining** 策略的三重损失，一个是根据实际实验结果进行改进的三重损失。本文将在下文分别介绍两者的具体学习过程。

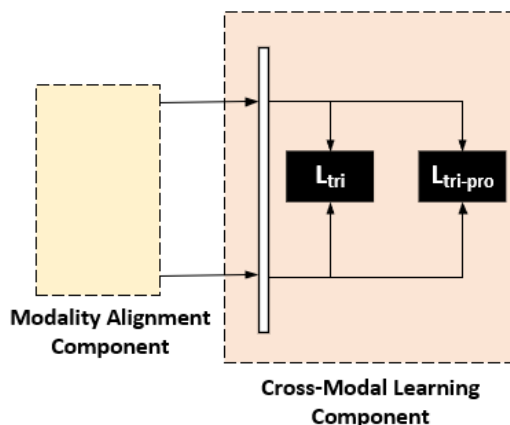


图 3-6 跨模态学习组件结构

首先分析一下，为何需要进行三重损失的改进操作。在训练框架进行训练到达尾声阶段，发现当使用 **hard example mining** 策略的三重损失时，会出现在潜在的高维度空间中噪声数据非常接近于中心点的情况。由于 **hard example mining** 策略，模型会首先选择到这些噪声当作训练数据，无论如何优化食谱-食物图像跨模态检索框架，都不能减少这些噪音的影响。这个问题如图 3-7 所示。在潜在的高维度空间 L_d 中，如果将一个食物图像的中心点表示为 P_{av} ，需要找到两个食谱数据，并应用 **hard example mining** 策略的三重损失。第一点 P_{nr} 是离 P_{av} 最近的负例食谱（食谱-食物图像都是成对输入的，例如西红柿炒鸡蛋的食谱配对着西红柿炒鸡蛋的图片，负例食谱就表示和选择到的食物图像中心点 P_{av} 不为食谱-食物图像匹配对的食谱数据），第二点 P_{pr} 是离 P_{av} 最远的正例食谱。为了方便理解，如图 3-7 展示此处描述的情况，我们可以画一个圆，它的半径是 P_{pr} 到 P_{av} 的距离。总的来说，我们的目的就是希望圆的半径足够小，以至于圆里面几乎不存在负例食谱，而正例食谱又可以落在距离中心点最近的位置，以达到模态对齐的效果。当研究 **Recipe 1M** 数据集时发现，如果 P_n 是噪声，则采用 **hard example mining** 的三重损失函数的训练重点，会在训练后期总

是集中在这些噪声上，因为这些噪声距离中心点足够近并且为负例食谱。无论如何优化它，都不能使它远离中心点 P_{av} 。噪声数据会影响圆内其他并没有距离中心点很近的负例食谱的训练，这就是本小节需要解决的问题所在。改进了的三重损失，使其既能处理 P_n ，又能聚焦于圆内边界的负例食谱 P_{br} ，以达到减少 P_n 对食谱-食物图像跨模态训练框架的影响。

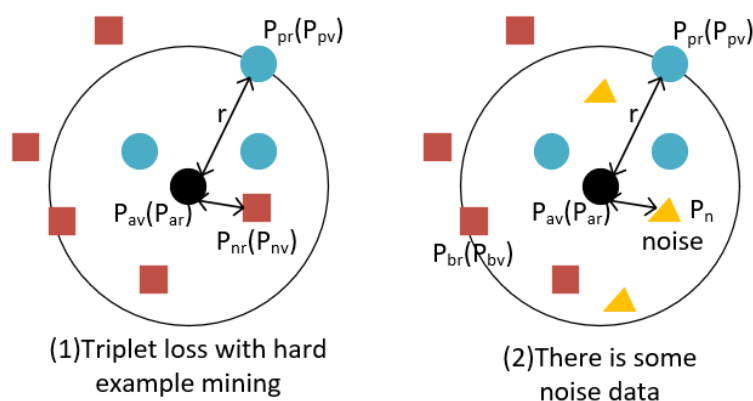


图 3-7 hard example mining 三重损失缺陷

此外，对于烹饪食谱，希望使 E_{vi} 靠近高维度空间 L_d 中的 E_{vj} ，当 i 等于 j 时，并使 E_{vi} 远离其他食谱嵌入 E_{rk} ，当 i 不等于 k 时。在 Recipe 1M 数据集中，有时会有多个食物图像与一个食谱具有多对一的关系。希望一个给定食谱的所有食物图像的特征嵌入在高维度空间 L_d 中靠近这个食谱，同时远离 L_d 中的其他食谱。

总体的损失函数定义如式 (3-16) 所示：

$$L = L_{tri} + \lambda_1 L_{al} + \lambda_2 L_{tri-pro} \quad (3-16)$$

其中 λ_1 和 λ_2 是训练权重系数。 L_{tri} 是采用 hard example mining 策略的三重损失，如式 (3-17) 中所定义。 L_{al} 是对抗生成损失，如式 (3-18) 中所定义。 $L_{tri-pro}$ 是本文改进的三重损失，如等式 (3-18) 中所定义。在总体的损失函数中，设置了 λ_1 和 λ_2 两个训练权重系数，其目的便是平衡两者损失函数在训练过程中发挥的作用。

(1) 采取 hard example mining 策略的三重损失

本小节的任务是使 E_{fi} 和 E_{fj} 对于给定的食谱-食物图像对，如果 i 等于 j ，那么这两个嵌入特征在高维度空间中是靠近的，对于 i 不等于 j 的情况下，食谱-食物图像嵌入特征在高维度空间中是远离的。这里，通过使用三重损失来实现这个思想。为了加快模型在学习时的收敛性，本文结合了 hard example mining。hard example mining 的思想，对于负例数据，本文选择离中心点最近的负样本（负例食谱样本点 P_{nr} 和负例食物图像样本点 P_{nv} ）（食物图像中心点 P_{av} 和食谱的中心点 P_{ar} ），对于正例数据，本文选择距离选定的中心点最远的正样本（正例食谱样本点 P_{pr} 和正例子食物图像样本点 P_{pv} ）。两点之间的距离由欧氏距离 $Ed()$ 计算，误差范围设为 α 。三重损失 L_{tri} 定义如式（3-17）所示：

$$L_{tri} = \min_{E_f, E_r} \left(\sum_n \left[Ed(P_{av}, P_{pr}) - Ed(P_{av}, P_{nr}) + \alpha \right] + \sum_n \left[Ed(P_{ar}, P_{pv}) - Ed(P_{ar}, P_{nv}) + \alpha \right] \right) \quad (3-17)$$

（2）改进的三重损失

传统的三重损失方法关注所有的训练数据，并视所有的训练数据同等重要，如前分析所知，训练中很多数据本身便符合三重损失的要求，并不需要进行训练学习。同时，噪声的存在，将所有数据认为同等重要的做法，对处理噪声是十分不利的。这不仅会大大影响模型的训练速度，而且难以解决高维度空间 L_d 中的关键问题。我们需要寻找一个合适的负样例点来改善三重损失，目的是既能更好地处理噪声 P_n ，又能聚焦于优化处理距离圆内边界最近的负食谱样本点 P_{br} （ P_{bv} 被设置为距离圆内边界最近的负食物图像样本点），这样做可以减少 P_n 噪声对训练的影响。在训练框架的最后训练阶段，我们希望把重点放在距离圆内边界的最接近的样本点上，而不是噪声数据上。改进的三重损失 $L_{tri-pro}$ 的定义如式（3-18）所示：

$$L_{tri-pro} = \min_{E_f, E_r} \left(\sum_n \left[Ed(P_{av}, P_{pr}) - Ed(P_{av}, P_{br}) + \alpha \right] + \sum_n \left[Ed(P_{ar}, P_{pv}) - Ed(P_{ar}, P_{bv}) + \alpha \right] \right) \quad (3-18)$$

这里 P_{av} 是食物图像中心点， P_{ar} 是食谱中心点。 P_{pr} 为正例食谱样本点， P_{pv}

为正例食物图像样本点。样本点之间的距离由欧氏距离 E_d 计算，误差范围设为 α 。

3.4 实验

3.4.1 数据集和评价指标

本文的实验是在 Recipe 1M 上进行的，Recipe 1M 被广泛使用在 [32][33][35][39][40][46][47] 中，是囊括了食谱和食物图像的最大型的食物计算数据集之一。根据实验需要，对数据集进行了处理，得到 276226 个训练数据、59573 个测试数据和 59288 个验证数据。

本章在构造后的测试数据集上，随机抽取 1K 和 10K 数据作为测试数据子集。实验结果是根据 50 次随机抽取 1K 和 10K 测试数据子集，并在其上进行测试而得出。

评价指标使用第二章介绍的 $R@K$ 和 MedR，其中 $R@K$ 设置为 $R@1$ 、 $R@5$ 和 $R@10$ 。这些评测指标被广泛应用于 [32][33][35][39][40][46][47] 模型中。

3.4.2 对比基线

本小节对本章对比的基线模型做简单介绍。介绍如下：

CCA^[45]：典型性相关分析旨在最大化相似对之间的相关性。

SAN^[38]：堆叠式注意网络（SAN）学习一个联合空间，通过两层注意机制增强食物图像及其对应食谱之间的相似性。

JNE^[32]：首先提出 Recipe 1M 数据集，并且在此数据集上，使用成对余弦损失来寻找食物图像和食谱配方跨模态之间的联合嵌入，其模型结构被后续研究人员广泛使用，影响深远。

ATTEN^[39]：ATTEN 对菜谱标题、菜谱制作步骤和菜谱配料成分分别应用注意力机制，对后续研究院有很大启发。从实验中，可以了解到，注意力机制对于食谱-食物图像跨模态检索任务的可行性，筛选食谱信息的有效性。

ADAMINE^[33]：ADAMINE 使用双三重损失和自适应学习策略，来处理食谱-食物图像的跨模态检索问题。三重损失相比于余弦损失学习能力更强，模型泛化能力也更强。双三重损失的使用，有效避免了余弦损失训练时，忽视相距中心点较近的负样本实例计算的情况。

R2GAN^[40]: R2GAN 是建立在跨模态嵌入和对抗生成网络的基础之上, 以完成食谱-食物图像的跨模态检索任务。作者证明了对抗生成思想在食谱-食物图像跨模态检索中的有效性。给后续研究人员带来启发, 面对噪声可以有两路走, 一条路可以选择筛选过滤, 一条路可以选择对抗生成进行生成。

ACME^[35]: ACME 采用了 hard example mining 的三重损失, 结合对抗生成损失强化模态对齐。ACME 采用的改进的三重损失策略发挥了巨大效果, 有选择性的对数据进行学习, 极大提高了模型框架的泛化能力和检索能力。

3.4.3 实验参数

实验中使用 Resnet-50 对食物图像进行编码, 并在 Recipe 1M 上进行预训练, 得到 1024 维高纬度空间的食物图片嵌入向量。食谱标题、制作步骤和配料成分分别由 BERT 进行编码, 编码的结果均是 (12, 768)。对于从食谱中提取有效信息的三个注意力网络, 其注意权重分别设置为 (768, 12) 和 (12, 1)。对于食物图像过滤的注意力网络, 分别设置食物图像和食谱的注意力权重 (1024, 512), 偏置值为 (512)。第二个注意力权重均为 (512, 1)。所有注意力机制中的参数都初始化为标准正态分布的随机数。通过实验发现, 将式(3-16)中的 λ_1 和 λ_2 分别设为 0.005 和 0.01 效果最好。该框架使用数据批次大小为 64 的 Adam 优化器^[62]进行训练。初始学习速率设置为 0.0001, 动量设置为 0.999。

3.4.4 实验结果与分析

(1) 检索质量综合对比分析

通过表 3-1 展示了食物图像→食谱文本的检索实验结果, 表 3-2 展示了食谱文本→食物图像的检索实验结果。在 1K 和 10K 测试数据集的检索任务中, SBNR 框架在所有评价指标中优于所有基线。在 1K 测试数据集上, SBNR 得到的 MedR 为 1.0, 了解到 MedR 最好的结果就是 1.0, 因为其定义就是位于中间位置的数据检索命中的排名, 那么最好成绩也就是一次检索就可完成任务。在 10K 测试数据集上, SBNR 可以得到 MedR 为 7.0。SBNR 的 R@K 也比其他基线模型更优。实验表明, 多层句子注意力网络组件和图像过滤组件能够提取句子级别的嵌入信息, 滤除食物图像中部分噪声信息。模态对齐组件和跨模态学习组件可以加强模态对齐, 为食物图像检索任务找到有效的潜在空间。也就是说, SBNR 框架可以使食谱-食物图像的特征嵌入在高维空间中, 同一类别的

数据对距离大大近于不同类别的数据对，以达到跨模态对齐的目标。

表 3-1 食物图像→食谱文本的检索实验结果

测试集大小	基线模型	食物图像→食谱文本检索			
		MedR	R@1	R@5	R@10
1K	CCA ^[45]	15.7	14.0	32.0	43.0
	SAN ^[38]	16.1	12.5	31.1	42.3
	JNE ^[32]	5.2	25.6	51.0	65.0
	ATTEN ^[39]	4.6	25.6	53.7	66.9
	ADAMINE ^[33]	2.5	39.8	69.0	77.4
	R2GAN ^[40]	2.0	39.1	71.0	81.7
	ACME ^[35]	1.0	51.8	80.2	87.5
	SBNR	1.0	52.7	81.7	88.9
10K	JNE ^[32]	41.9	-	-	-
	ATTEN ^[39]	39.8	7.2	19.2	27.6
	ADAMINE ^[33]	16.5	12.5	31.5	42.2
	R2GAN ^[40]	13.9	13.5	33.5	44.9
	ACME ^[35]	7.0	22.0	45.3	56.6
	SBNR	7.0	22.1	45.9	56.9

从实验结果来看，注意力机制的使用提高了食谱-食物图像跨模态检索框架的检索性能。究其原因，一方面对于食物图像过滤部分，注意力机制通过利用句子级编码的菜谱嵌入，学习可分配的权重，来达到筛选过滤食物图像的目的。另一方面对于多层句子级编码的融合部分，因为各层编码表达的含义并不相同，本文希望利用注意力机制的特性，来智能筛选融合合适的句子级编码信息。本章使用的注意力机制目的并不是对句子的语义进行提取，句子语义的提取由

BERT 完成，注意力机制只是负责了句子级编码融合时的权重分配以及食物图像过滤时的权重分配。

表 3-2 食谱文本→食物图像的检索实验结果

测试集大小	基线模型	食谱文本→食物图像检索			
		MedR	R@1	R@5	R@10
1K	CCA ^[45]	43.0	24.8	24.0	35.0
	SAN ^[38]	42.3	-	-	-
	JNE ^[32]	5.1	25.0	52.0	65.0
	ATTEN ^[39]	4.6	25.7	53.9	67.1
	ADAMINE ^[33]	2.1	40.2	68.1	78.7
	R2GAN ^[40]	2.0	40.6	72.6	83.3
	ACME ^[35]	1.0	52.8	80.2	87.6
	SBNR	1.0	54.1	81.8	88.9
10K	JNE ^[32]	39.2	-	-	-
	ATTEN ^[39]	38.1	7.0	19.4	27.8
	ADAMINE ^[33]	15.2	13.6	32.8	43.4
	R2GAN ^[40]	12.6	14.2	35.0	46.8
	ACME ^[35]	7.0	23.3	47.1	57.9
	SBNR	7.0	23.4	47.3	57.9

可以看出，在 10K 数据子集上，SBNR 框架的结果并没有很好的优于 ACME 模型，这表明 SBNR 还可以有很大的进步空间。同时，检索过程中，数据对的逐个计算欧氏距离所导致的问题就是检索缓慢，这也是本文列出的基线模型共同拥有的问题。这个问题如果面临模型框架落地成实际生活中的应用时，除了检索准确度外，检索速度也是关键问题。框架的检索速度成为框架优化的一个

方向，未来的研究中可以更加关注框架的检索速度。

(2) 消融实验

本节将展示由于 SBNR 的不同组件的作用而带来收益的消融实验。首先，利用 BERT 的最后一个隐藏层的输出的食谱嵌入向量进行实验（BLH），并结合 hard example mining 三重损失。本小节以增量的方式添加更多的组件：首先，对 BERT 多层隐藏层输出的向量进行平均（Ave）来进行实验。在此基础上，使用了三个有注意力偏置值（TAB）和无注意力偏置值（TNB）的注意力网络进行实验。在此基础上，在最优的合成模型上加入食物图像过滤组件（IF）。最后，增加了改进的三重损失（NTL）。

表 3-3 消融实验 **BLH**: BERT 最后一层隐层, **Ave**: BERT 十二层隐层取平均, **TAB**: 有偏置值的多层句子级注意力机制组件, **TNB**: 无偏置值的多层句子级注意力机制组件, **IF**: 食物图像过滤组件, **NTL**: 改进后的三重损失。

评价指标 消融组件	MedR	R@1	R@5	R@10
BLH	2.0	46.1	76.3	83.4
Ave	1.95	47.7	77.5	85.5
TAB	1.7	48.6	78.7	86.2
TNB	1.6	49.5	79.8	86.8
IF+TNB	1.4	50.0	80.3	87.7
IF+TNB+NTL	1.0	52.7	81.7	88.9

在表 3-3 中列出 1K 测试数据子集上每一步的检索食谱时的测试结果。如表 3-3 所示，当不使用注意力偏置值（TNB）时，SBNR 能够在 MedR 中得到 0.1 的改善，在 R@1 中得到 0.9 的改善。在 SBNR 中，图像过滤组件的工作效果良好。IF 组件相比于 TNB 可以在 MedR 中提高 0.2，在 R@1 中提高 0.5。本文还进一步发现，如果使用改进的三重损失，可以获得更好的 MedR（1.0）和 R@1（81.7）。此外，所有的注意力网络对 SBNR 都有较大帮助。

3.4.5 样例分析

(1) 食谱文本→食物图片检索

图 3-8 展示了食谱作为输入的检索结果。在 SBNR 框架中输入了三种不同的食谱（炖肉类、蛋糕类和烤制类的食谱）。top5 的食物图像展现出 SBNR 框架可以很好地捕捉食谱-食物图像两种模态之间的语义信息。当牛肉炖南瓜这个食谱被输入到 SBNR 中时，可以得到的食物图像都包含牛肉和南瓜这两种重要的食材。对于第三个关于烤花椰菜的食谱，在检索结果中，第四个和第五个食物图像分别是炸鸡块和炸羊排。对于检索结果会出现这种情况，本文认为这个其主要原因可能是 SBNR 框架没有给食谱标题和食谱配料中的食物类别信息赋予足够的权重。如果增强配料和标题分类信息在学习过程中的权重，本文认为，至少主要配料是不会出现差错的，而这里检索花椰菜为主要配料的食谱却得到鸡肉和羊肉。同样也可看出，在食谱查询中，烤花椰菜和炸鸡块中都会使用到常见的配料，例如油盐，鸡蛋，面粉等，当出现鸡蛋、盐等油炸食品的基本配料成分时，会比较严重的影响检索结果，因为几乎大部分菜谱都使用了它们。做菜时，主要食材和次要食材在食谱中影响力是完全不同的。例如，制作炸洋葱圈，那么洋葱作为主要食材的训练影响力相比于次要食材面粉蛋液面包糠就需要更加突出。也许，出现在食谱标题中的食物类别信息应该在食谱中得到更多的关注。
















Recipe query	Top5 retrieved images				
Title: Beef Tzimmes with Butternut Squash Ingredients: lean beef chuck, vegetable oil, onion, carrots, butternut squash, honey, ground cinnamon, water... Instructions: Cut beef in 1 1/2- 2 inch pieces and pat them dry. Heat 1 tablespoons oil in a heavy stew pan...					
Title: Chocolate Stout Cake Ingredients: unsalted butter, Guinness stout, pitted prunes, bittersweet chocolate, baking soda, salt, eggs... Instructions: Put oven rack in middle position and preheat oven to 350F. Lightly brush ring pan with melted butter...					
Title: Crispy Curry Roasted Broccoli Ingredients: breadcrumbs, garlic cloves, cilantro, fresh ginger, curry powder, garam masala, salt, broccoli... Instructions: Preheat oven to 400F. Line a baking sheet with foil and grease. In a shallow bowl, combine ...					

图 3-8 食谱文本→食物图片检索结果展示

(2) 食物图片→食谱文本检索

图 3-9 显示了从食物图像作为输入的检索结果。具有复杂背景的食物图像检索是从 Recipe 1M 中选择的，本小节想利用这些食物图像，测试图像过滤组件的工作性能。如图 3-9 所示，即使排在前几位的检索结果肉眼看起来非常相似，但 SBNR 也可以检索到正确项。然而，对于第二个食物图像的检索，菜谱是关于玉米饼卷。虽然本文的框架检索到了正确的食谱，但可以发现检索到的食谱中的第二个食谱是关于冰淇淋。冰淇淋配料中虽然也有“面粉，玉米饼”，但它是一种完全不同于玉米饼卷的食物。从这里也可以看出来，在训练模型过程中的权重分配还可以进一步进行调整，加强食谱类别信息的关注度，做到食谱的标题等类别信息占比更大，保证不同类别的食物可以更容易区分开。





Image query				
Best match	Title: Cold Sesame Noodles Ingredients: soy sauce, rice vinegar, hot pepper flakes, brown sugar, creamy peanut butter, toasted sesame oil... Instructions: In saucepan over medium heat, mix first 8 ingredients together (soy sauce to chicken broth), stir....	Title: Sweet Tortilla Roll-Ups Ingredients: fat - free ricotta cheese, sugar, cinnamon, vanilla, flour tortillas Instructions: Set oven to broil. In a mixing bowl, blend ricotta cheese, sugar, cinnamon and vanilla. Spread...	Title: Alaine's Blue Cheese Dressing Ingredients: mayonnaise, blue cheese, balsamic vinegar, garlic powder... Instructions: In a medium bowl, mix mayonnaise, blue cheese, balsamic vinegar, pepper, and garlic powder...	Title: Caesar Salad Ingredients: egg, kosher salt, garlic clove, garlic powder, olive oil... Instructions: Coddle the egg by bringing a small saucepan of water to a boil and immersing the egg for 60 to 90 second...
Top2 retrieved recipes	Title: Sesame Cold Noodles Ingredients: Chinese egg noodles, sesame oil, tahini, sugar, soy sauce, rice wine vinegar, fresh scallions... Instructions: Cook the noodles in boiling salted water until they are tender but not mushy. Drain and rinse in cold water....	Title: Tex-Mex Ice Cream Sundaes Ingredients: vegetable oil, flour tortillas, sugar, ground cinnamon, cinnamon ice cream, chocolate syrup... Instructions: In a heavy skillet, pour oil to the depth of 1/2-inch. Heat oil over medium heat. When oil is hot....	Title: World's Easiest and Most Amazing Two-Ingredient Dip Ingredients: sour cream, soy sauce Instructions: Mix sour cream and soy sauce together in a bowl until evenly combined.	Title: Caesar Salad Ingredients: romaine lettuce, seasoned croutons, 1/2 cup KRAFT Reduced Fat Parmesan Style Grated Topping... Instructions: Toss lettuce with croutons and grated topping in large bowl. Add dressing; mix lightly...

图 3-9 食物图片→食谱文本检索结果展示

3.4.6 实验中的不足

如表 3-1，表 3-2 所示，在 1K，10K 数据子集上，SBNR 框架相比 ACME 优势并不突出。SBNR 框架还可以继续改进，如下列举可以努力的方向：

(1) 首先，虽然食物图像筛选过滤组件和句子级食谱文本嵌入发挥了很大作用，但是这里有一个需要考虑的问题。本章使用的注意力机制过于简单，难

以关注到更细粒度的噪声信息和多层句子级嵌入信息，无法很好的提取多层隐层句子级食谱文本各层向量间的相互作用的关系。注意力机制方面的改进工作，可以继续进行的。

(2) 其次，实验中遇到的噪声问题还存在多种类型，不匹配的食谱-食物图像对是其中需要关注的重点。单纯使用食谱文本句子级嵌入的语义还不能很好的筛选 Recipe 1M 数据集中的多种噪声，还需要一种手段来处理更细粒度，更多种类的食谱-食物图像噪声。

(3) 再者，由于实验条件受限制，希望本章实验使用的句子级编码器可以进行个性化的食谱文本训练还不能实现。使用预训练的句子级编码器可能导致对食谱文本编码时，丢失部分食谱制作步骤中的顺序信息。

3.5 本章小结

本章提出了基于句子级编码的食谱-食物图像跨模态检索框架（SBNR 框架），主要是解决了食谱-食物图像跨模态检索中的食物图片噪声问题，句子级别食谱文本嵌入问题，食谱-食物图像跨模态对齐问题，其中对齐手段又包括两种策略，第一种是对抗生成思想加强食谱-食物图像跨模态数据分布，第二种是采用两种三重损失策略更好地适应食谱-食物图像跨模态检索问题。最后对 SBNR 框架进行测试，展示出测试结果并与现如今流行的方法进行检索质量分析和样例分析，阐述本文提出框架的优点和存在的不足，为下一章节的改进和新研究任务做铺垫。

第4章 基于食物图像检测的噪声过滤及检索方法

在上一章中,本文就食谱-食物图像跨模态检索问题提出了 SBNR 跨模态检索框架,但也存在着不足之处。Recipe 1M 的数据海量,导致其中包含多种类型的噪声数据。SBNR 最初采用的噪声过滤手段,虽然起到了较好的效果,但还可以继续进行改进。通过更细粒度的手段在检测食谱-食物图像噪声数据对的时候,去除不匹配的食谱-食物图像对。本章将通过句子级别编码表现出的丰富语义,结合食物图像区域检测,提出食谱-食物图像噪声筛选打分组件。对前一章提出的检索框架中注意力机制改进为多头注意力机制,增强 SBNR 框架学习多层句子级食谱文本嵌入向量的能力,进一步改进检索质量,提出改进的 SBNR⁺框架。

4.1 问题描述

分析 SBNR 框架实验结果, SBNR 在以下两种噪声类型的影响下食谱-食物图像跨模态检索效果不佳。第一种,食物图片中食物的有效区域较小,背景信息复杂多变或存在多种类型食物存在于同一张图中。第二种,食谱-食物图像对本身不匹配,食谱描述的食物和食物图像中的信息不相符,或者食物图像中本就不存在食物信息。假设食物图像过滤后的嵌入表示为 E_f , 食谱嵌入表示为 E_r , 食物图像中有效食物区域通过食物图像检测组件识别并表示为 Pos 。本章的研究目的就是通过筛选打分组件,结合食物图像区域检测,食谱-食物图像嵌入向量的语义信息输出匹配对打分 $P_{pair} = NF(E_f, E_r, Pos)$, 来完成训练集噪声筛选过滤功能。针对第三章的 SBNR 框架优化着重体现在多头注意力机制的设计使用,多头注意力机制来处理包含多层隐层句子级信息的食谱数据,输出食谱配料嵌入 s_{me} , 食谱制作步骤嵌入 g_{me} , 食谱标题嵌入 t_{me} 。本章使用多头注意力机制改进 SBNR 框架,完成更细粒度的食物图像噪声筛选打分组件。改进后的框架命名为 SBNR⁺框架。

4.2 整体框架

SBNR⁺框架与未改进的 SBNR 框架整体结构相似,本小节不重复展示。本

章将于 4.3.1 节分别详细介绍如何设计，多头食谱标题注意力机制组件、多头食谱制作步骤注意力机制组件、多头食谱配料注意力机制组件。

本章通过研究食物图像区域检测技术，结合句子级别食谱文本语义和注意力机制的思想，提出基于食物图像区域检测的噪声筛选打分组件，对食谱-食物图像中的噪声对进行筛选过滤，以构造出更适合跨模态框架学习的训练数据集。本章提出的基于食物图像检测的噪声筛选打分组件结构如图 4-1 所示：

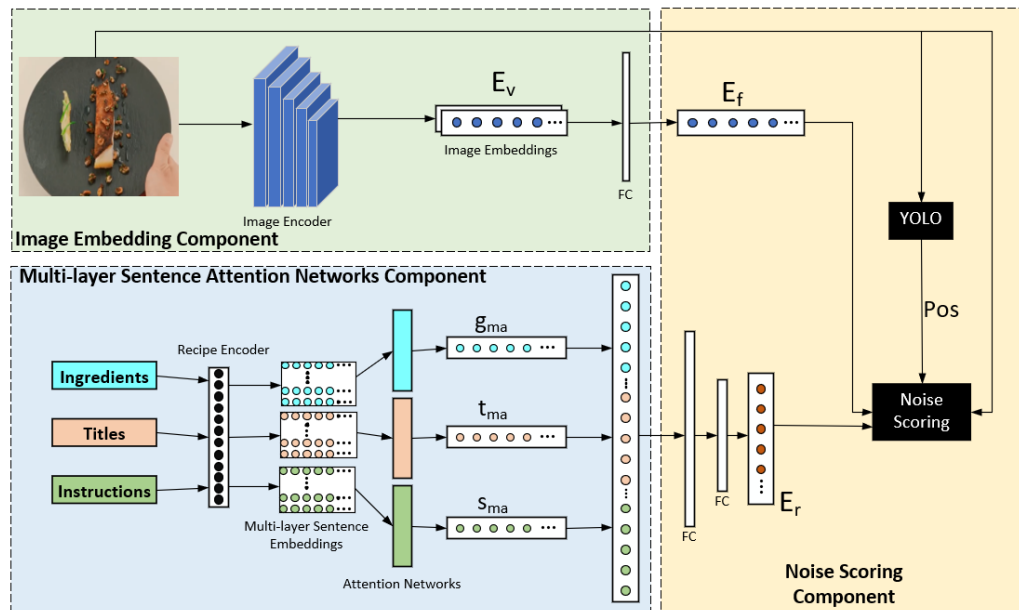


图 4-1 基于食物图像区域检测的噪声筛选打分组件结构

从图 4-1 可以看出，该组件是多层句子级注意力机制组件结合图像嵌入组件合作完成的。食物图像嵌入组件输出食物图像嵌入 E_f ，多层句子级注意力机制组件输出结果经过全连接层 FC，得到食谱文本嵌入 E_r 。该组件将原本未编码的食物图像输入到 YOLO 中，得到食物图像检测区域坐标 Pos。最后，该组件把 E_f 、 E_r 、Pos 和原本未编码的食物图像一起输入噪声打分组件，得到打分结果 P_{pair} 。

在 4.3.2 节中，本章将详细介绍食物图像的区域检测组件 YOLO。

在 4.3.3 节中，本章将详细介绍基于食物图像区域检测的噪声筛选打分组件。

根据噪声筛选打分组件得到的食谱-食物图像匹配对打分结果，本章构造出一个更加纯净的训练数据集。新构造的训练数据集应用于多头注意力机制改进

后的框架，并于 4.4 节中展示实验。

4.3 组件介绍

4.3.1 基于多头注意力机制的组件改进

第三章本文提出了 SBNR 跨模态检索框架，其中使用了四个注意力机制。四个注意力机制中，三个都是为了提取 BERT 输出的拥有多层隐层食谱文本嵌入向量，目的是更细粒度的提取各层隐层包含的信息，低层包含表面特征、中层包含句法特征和高层包含语义特征。但是第三章所使用的所有注意力机制均是基础的注意力机制，虽然取得较好的结果，但是可以继续改进。注意力机制近年来也得到广泛的应用，包括基础的注意力机制，自注意力机制以及多头注意力机制。最为大家熟知的就是那篇《Attention is All You Need》^[37]论文，这篇文章一经提出，自注意力机制成为注意力方面的研究热点，在各任务上也取得不错的效果。

注意力机制可以从人类身上来理解，注意力机制的思想是模仿人类的视觉注意力机制。众所周知，人类在看某些东西的时候，往往因为注意力集中于某一点上，就忽视场景中的其他信息。例如，夜深人静的时，一位食物计算领域的研究员正在手动筛选过滤食物图片，当看到一幅幅美食图片时，他的注意力就集中在了食物上，而食物图片中食物之外的噪声信息可能很难发现。

上一章，便是利用注意力机制的思想来筛选食谱嵌入向量，但发现在此任务上，多头注意力机制相比于普通的注意力机制更有优势。多头注意力机制的优势体现在，允许模型在不同的表示子空间里学习到相关的信息。本章将设计三个多头注意力机制并应用于 SBNR 框架中，提出 SBNR⁺框架。

对处理食谱嵌入向量的多层隐藏层，应用多头注意力机制进行筛选，由于多层隐藏层输出的嵌入向量包含了不同类型的信息，需要合理的筛选对训练有利的信息，通过多头注意力机制的思想，在这里将多层隐层嵌入向量合成为一层嵌入向量。多头注意力机制的思想类似于经常在机器学习中使用的集成思想，通过多个学习模型的集成，可以减少过拟合的情况出现。通过对多头注意力机制中各个注意力模型的矩阵权重初始化不同，又保证被集成的每一个注意力模型单独训练，各个模型参数互不干涉，可以使被集成的多个注意力机制模型，在不同的表示子空间里学习到相关的食谱信息。

针对包含多层隐层信息的食谱标题嵌入向量，多头注意力机制的使用如图 4-2 所示：

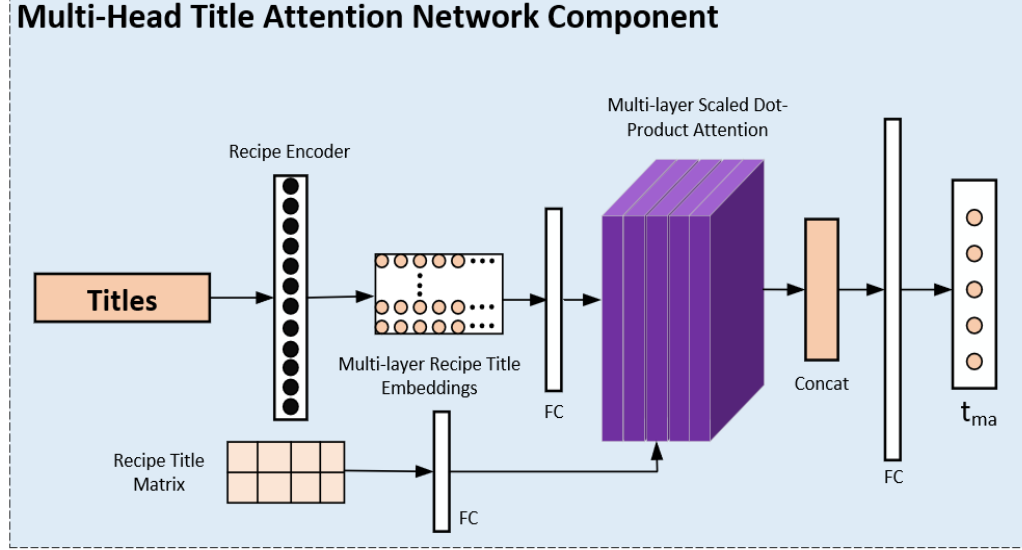


图 4-2 多头食谱标题注意力机制组件

组件输出食谱标题的嵌入向量 t_{ma} 。食谱标题嵌入向量 t_{ma} 的计算过程如下所示。Attention 操作是上文提到的 Multi-Layer Scaled Dot-Product Attention。 t_{mej} 表示第 j 个拥有多层隐层信息的食谱标题嵌入向量， W_i^t 表示其输入多头注意力机制时，注意力机制第 i 个头使用的线性变换参数。 W_i^T 表示食谱标题权重矩阵， W_i^m 同样表示输入注意力机制时，注意力模型第 i 个头使用的线性变换参数。 W_i^{tme} 表示多层隐层信息的食谱标题嵌入向量输入注意力机制时，注意力机制第 i 个头使用的线性变换参数。因为多头注意力机制各个头参数并不共享，所以每个线性变换参数 W 均不同。 $head_{ti}$ 表示多头注意力模型中第 i 个头输出的结果，具体公式如 (4-1) 所示。

$$head_{ti} = \text{Attention}\left(t_{me_j} W_i^t, W_i^T W_i^m, t_{me} W_i^{tme}\right) \quad (4-1)$$

得到 h 个头的多头注意力模型输出后，使用 Concat 操作对 h 个头的输出进行合并。 W^{tma} 表示与合并后的嵌入向量相运算的线性变换参数。 t_{maj} 表示合并第 j 个多层隐层食谱标题嵌入向量后，多头注意力机制输出如公式 (4-2) 所示。

$$t_{ma_j} = \text{Concat}(\text{head}_{t1}, \dots, \text{head}_{sh}) W^{ma} \quad (4-2)$$

注意到 Scaled Dot-Product Attention 组件，此组件便是第三章提及的注意力机制模型，其内部结构如图 4-3 所示。

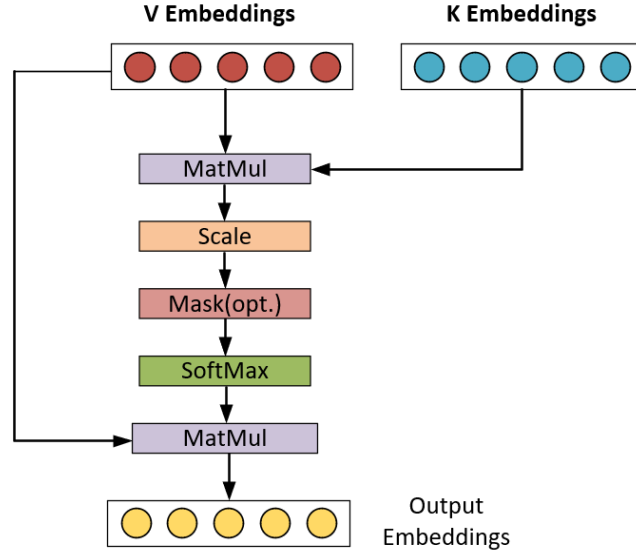


图 4-3 Scaled Dot-Product Attention 内部结构

其中 MatMul 模块表示矩阵乘法，Scale 模块表示向量通过全连接层处理完成维度变换，Mask 模块表示通过输入的 opt 进行设置是否忽略某项，SoftMax 模块计算概率权重。

对 h 个这样的基础注意力机制模型分别训练学习食谱标题信息，然后通过 Concat 组件，对 h 个头部输出的结果进行拼接集成。Multi-Layer Scaled Dot-Product Attention 如公式（4-3）所示：

$$\text{Attention}(v_t, K, V) = \sum_{s=1}^m \frac{1}{Z} \exp\left(\frac{\langle v_t, k_s \rangle}{\sqrt{d_k}}\right) v_s \quad (4-3)$$

其中 Z 表示归一化因子。使用多头注意力机制，处理食谱-食物图像两种模态的信息，将嵌入向量 V 参与嵌入向量 K 的运算，最终模型输出的注意力权重再次与嵌入向量 V 进行矩阵相乘计算，得到输出的嵌入向量。公式中分母的作

用是，起到调节作用，使得上方内积不至于太大。如果分子内积过大，这里使用的 SoftMax 方法就会输出 0 或者 1 两种极端情况，无法起到训练作用。

对食谱制作步骤和食谱配料的操作如同食谱标题相同，使用三个多头注意力机制分别处理食谱三种文本信息。当处理食谱制作步骤时，将其输入到食谱制作步骤注意力机制组件，输出食谱制作步骤嵌入向量 s_{ma} 。同样嵌入操作应用到食谱配料，输出食谱配料嵌入向量 g_{ma} 。这里就不重复绘制相同操作步骤的展示图。

针对包含多层隐层信息的食谱制作步骤嵌入向量，多头注意力机制的处理如式 (4-4) 所示。 s_{mej} 表示第 j 个拥有多层隐层信息的食谱制作步骤嵌入向量， W_i^s 表示其输入多头注意力机制时，注意力机制第 i 个头使用的线性变换参数。 W_s^T 表示实验中使用的食谱制作步骤权重矩阵， W_i^m 同样表示输入注意力机制时，注意力模型第 i 个头使用的线性变换参数。 W_i^{sme} 表示多层隐层信息的食谱制作步骤嵌入向量输入注意力机制时，注意力模型第 i 个头使用的线性变换参数。因为多头注意力机制各个头参数并不共享，所以每个线性变换参数 W 均不同。 $head_{si}$ 表示多头注意力模型中第 i 个头输出的结果。

$$head_{si} = \text{Attention}\left(s_{mej} W_i^s, W_s^T W_i^m, s_{me} W_i^{sme}\right) \quad (4-4)$$

得到 h 个头的多头注意力模型输出后，使用 Concat 操作对 h 个头的输出合并的公式 (4-5)。 W^{sma} 表示与合并后的嵌入向量相运算的线性变换参数。 s_{maj} 表示合并第 j 个多层隐层食谱制作步骤嵌入向量后，多头注意力机制的输出。

$$s_{maj} = \text{Concat}(head_{s1}, \dots, head_{sh}) W^{sma} \quad (4-5)$$

同样针对多层隐层的食谱配料嵌入向量，多头注意力机制的处理公式如下 (4-6) 所示。 g_{mej} 表示第 j 个拥有多层隐层信息的食谱配料嵌入向量， W_i^s 表示其输入多头注意力机制时，注意力模型第 i 个头使用的线性变换参数。 W_g^T 表示实验中使用的食谱配料权重矩阵， W_i^m 同样表示输入注意力机制时，注意力模型第 i 个头使用的线性变换参数。 W_i^{gme} 表示多层隐层信息的食谱配料嵌入向量输入注意力机制时，注意力模型第 i 个头使用的线性变换参数。 $head_{gi}$ 表示多头注意力模型中第 i 个头输出的结果。

$$\text{head}_{gi} = \text{Attention}\left(g_{me_j} W_i^s, W_g^T W_i^m, g_{me} W_i^{gme}\right) \quad (4-6)$$

得到 h 个头的多头注意力模型输出后, 使用 **Concat** 操作对 h 个头的输出合并公式 (4-7)。 W^{gma} 表示与合并后的嵌入向量相运算的线性变换参数。 g_{maj} 表示合并第 j 个多层隐层食谱配料嵌入向量后, 多头注意力机制的输出。

$$g_{ma_j} = \text{Concat}(\text{head}_{g1}, \dots, \text{head}_{gh}) W^{gma} \quad (4-7)$$

根据已经编码嵌入菜谱的三个重要信息, 使用拼接的操作将三者嵌入结果送入第三章的注意力机制进行权重融合后拼接在一起, 并一起送往全连接层 FC, 使得拼接后的菜谱嵌入向量被映射到高纬度公共空间中。具体操作如下公式 (4-8) 所示。

$$E_{r_i} = FC\left(\left[\text{Attention}(s_{ma_i}), \text{Attention}(t_{ma_i}), \text{Attention}(g_{ma_i}) \right]\right) \quad (4-8)$$

4.3.2 食物图像的区域检测组件 YOLO

对于一张给定的食物图片, 已经有很多研究员对食物的检测方法提出很多模型, $WISeR^{[10]}$ 便是其中之一, 它通过改进 **Resnet** 网络结构, 增加了 **SLICE** 卷积模块, 对食物图片进行横向切分, 使 **Resnet** 网络在分层提取食物信息的能力上得到加强。正是因为 **SLICE** 卷积模块的使用, 当处理分层食物例如, 汉堡包, 千层肉饼等拥有多层结构的食物能力较好, 但对于盘装食物图片, 碗装食物图片, 它们并没有明显的分层结构, **SLICE** 卷积模块并不能有效处理这种情况。**WISeR** 只能对食物图像检测做出分类判断, 不能准确框选出图片中的食物区域, 并不能作为研究食物图片检测的出发点。

如今, 目标检测方面的模型方法主要解决有如下三个核心问题:

- (1) 图像中的物体的类别究竟是什么?
- (2) 图像中物体的位置到底在哪?
- (3) 图像中物体的形状大小变化时如何考虑?

基于这三个问题，目标检测中 **two-stage** 检测算法的特点就是识别的错误率低，但是由于分阶段进行，模型的运行速度慢，难以满足实时检测。例如，在餐馆中加装摄像头，希望实时识别并计算用餐顾客的饮食习惯，**two-stage** 很难完成这个任务。**one-stage** 的特点也就很明显了，速度更快，但舍弃了一定的准确度。主流的物体检测模型有 **two-stage** 的 Faster-RCNN^[63] 和 **one-stage** 的 YOLO^[64]。

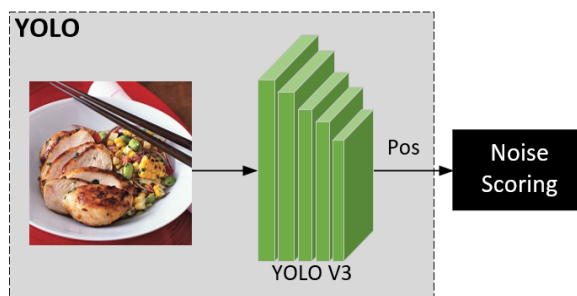


图 4-5 食物区域检测组件结构

本章中，食物区域检测组件基于 YOLO V3 算法，。YOLO V3 在 UEC FOOD-256^[65]数据集上进行训练，使其拥有检测食物图像中食物区域的能力。可以看到图像区域检测组件，其内部结构如图 4-5 所示，图中变量 **Pos** 表示 YOLO V3 算法输出的检测到的食物图像中框选食物的矩形框坐标信息。此组件的输出结果为食物图片中，可以通过矩形结构完整框选食物区域的矩形框坐标数组。该组件输出结果如下图 4-6 所示：



图 4-6 食物图像检测组件检测效果

4.3.3 基于食物图像区域检测的噪声打分筛选组件

在食谱-食物图像跨模态检索研究中，当使用结合 **hard example mining** 策略

的三重损失进行模型训练时，虽然其高维度潜在空间的表现比余弦相似度损失更优，但是在面对庞大数据量的数据集时，噪声信息在三重损失仍然很难处理。上一章提到，本文改进了三重损失，使用了 **hard example mining** 的三重损失结合本文改进后的三重损失，共同学习食谱-食物图像跨模态检索任务，但是训练模型到达后期阶段时，仍然可以看到噪声信息的影响下，模型在一个小范围内摆动。模型由于噪声数据的影响已经达到提升的瓶颈，所以本小节针对噪声对的打分筛选系统进行研究，旨在获得一个更加纯净的食谱-食物图像对训练数据集。

前文提到，**Recipe 1M** 虽然食谱-食物图像数据丰富，但是由于是网络上爬取的用户上传菜谱图像数据，存在较多不必要的噪声信息，并且庞大数据量导致很难进行人工筛选过滤。当针对检索框架表现不佳的食谱-食物图像数据对进行分析时，发现导致检索不佳的数据对包括以下情况：

(1) 食谱和食物图像并无关系，例如，食谱是展示如何制作黄焖鸡，但是配图却展示的是煲仔饭，这种噪声数据对在 **Recipe 1M** 是存在最普遍的噪声类型。由于很多食谱食物图片是一对多的关系，一个食谱对应多个食物图片，而多张食物图片中，也同样存在很多噪声。例如，一个南瓜粥的食谱，对应了多张展示图片，而展示图片中很多只是包含了桌饰和餐具，又或者是包含了复杂的制作步骤过程中的每个步骤阶段展示，这些图片对于食谱-食物图像检索任务来说只是无用的图片信息。

(2) 食谱-食物图片中，无用信息占比过大。例如，一张高分辨率的手机拍摄食物照片中，和主题相关的食物只占据了照片的八分之一到十分之一大小，而图片中其他区域是餐桌装饰和背景等无关信息。或者由于拍摄角度问题，食物并没有正对着摄像头而是倾斜着展示，也导致照片中有效的食物信息范围过小。

(3) 极少一部分食谱-食物图像对难以区分原因是，其制作步骤只有一句话，混合所有食材。这种制作步骤只有一句话的情况，多见于沙拉类型的食谱中，众所周知，沙拉大多是把准备好的食材简单搅拌混合。大多沙拉的配菜都是类似酸奶，牛油果，沙拉酱等常用配料，这就更加导致这部分数据的检索任务比较难完成，因为他们具备很少的个性化信息。

4.3.2 节中，研究了食物图像中食物区域的检测方法，并得到了较高准确度的食物图像检测模型。使用食物图像检测模型的目的，便是获得食物图片中食物的框选位置，根据有效的食物框选面积与整个食物图像面积比，结合图像过

滤组件，通过计算公式（4-9）对 Recipe 1M 中的训练集数据进行打分筛选，构造出一个新的训练数据集。基于食物图像区域检测的噪声打分筛选组件如图 4-7 所示。

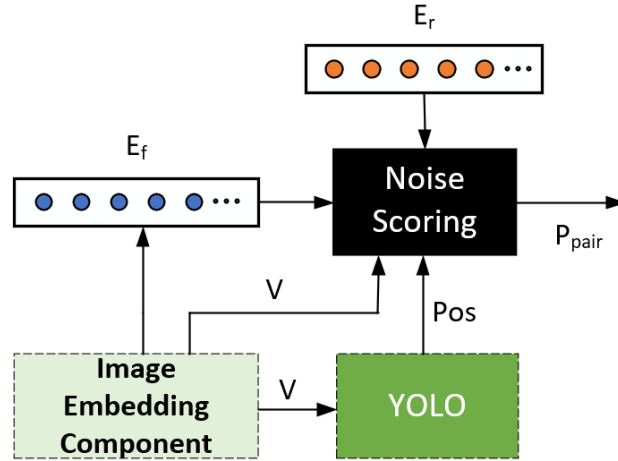


图 4-7 基于食物图像区域检测的噪声打分筛选组件

但有一点需要特别注意，如果一个食物图片中出现多个食物，而检测方法框选出多个有效食物矩形框后，如何处理这些多个矩形框？正是由于这种情况偶尔出现，本文才选择使用食谱语义辅助计算匹配对打分，为的就是防止出现多个选框时，错误计算了选框面积，单一的食物图像面积打分机制出现错误的概率更大。希望语义筛选起到的作用是，筛去食谱-食物图像匹配对中并不匹配的问题。经过实验，确定了实验方法，如果存在多个矩形选框的情况，那么优先框选面积最大的食物区域作为标准进行打分计算。

食谱-食物图像对噪声打分筛选组件的内部计算公式如（4-9）所示：

$$P_{\text{pair}} = \left(\text{Area}(\text{Pos}) / S_v \right) * \lambda_1 + \left(1 - \text{Rank}(E_r, E_f) / \text{BatchSize} \right) * \lambda_2 \quad (4-9)$$

其中 E_r 表示菜谱嵌入向量， E_f 表示过滤后的食物图片向量。 V 表示输入到食谱-食物图像跨模态检索框架中的未编码食物图像， S_v 表示此食物图像的图像面积大小。 Pos 表示由 YOLO 组件识别检测出的食物图像画框的矩形坐标， Area 表示根据矩形坐标计算矩形框面积大小。 P_{pair} 表示经过噪声过滤组件运算输出食谱-食物图像匹配对的打分分值。 Rank 表示在语义测试时，每 BatchSize 个匹

配对进行一次筛选, Rank 是计算在这 BatchSize 个匹配中目标匹配对的排名, 如果语义排序过低, 可以认为这食谱-食物图像匹配对并不是匹配的, 收集数据集时出现差错, 类似西红柿炒鸡蛋菜谱配图是烤羊肉串情况一样。

4.4 实验

4.4.1 数据集和评价指标

本章的实验是在 Recipe 1M 和 UEC FOOD-256 两个数据集上进行。根据实验需要对数据集 Recipe 1M 进行了处理, 本章使用基于食谱文本和食物图像的噪声打分筛选组件对 Recipe 1M 的训练集进行筛选过滤, 得到 50000 个训练数据。验证数据和测试数据与第三章一致。UEC FOOD-256 数据集用于训练 YOLO V3, 使其可以在食物图片中完成食物图像区域检测的任务。

UEC FOOD-256 与 Recipe 1M 不同之处在于, 此数据集被广泛应用于食物图像检测研究中, 因为其提供了食物图片中食物区域的有效边框, 方便各种食物图片检测模型的训练。UEC FOOD-256 数据集包含了 256 种类型, 一共 25088 张食物图片, 其中的每一张食物图片中都包含食物的边框信息, 以此来表示食物图片中食物的具体位置。UEC FOOD-256 中的大多数食物类别都是日本和其他国家流行的食物, 其也被广泛用于食物识别。

本文在构造后的测试数据集上, 随机抽取 1K 和 10K 数据作为测试数据子集。实验结果是根据 50 次随机抽取 1K 和 10K 测试数据子集, 并在其上进行测试而得出。

评价指标使用第二章介绍的 $R@K$ 和 MedR, 其中 $R@K$ 设置为 $R@1$ 、 $R@5$ 和 $R@10$ 。这些评价指标被广泛应用于[32][33][35][39][40][46][47]模型中。

4.4.2 对比基线

本章的对比基线与第三章一致, 但本章将新增一个对比基线, 就是第三章提出的未经改进过的 SBNR 跨模态检索框架。第三章提出的 SBNR 框架作为对比基线, 目的是通过对比实验结果展示第四章框架改进的效果。

4.4.3 实验参数

实验中使用第三章提出的 SBNR 框架结构, 改进其中的三个注意力机制,

设计替换为新的多头注意力机制，每个多头注意力机制分 8 个头部，进行独立的训练学习。所有多头注意力机制中的参数都被独立初始化。在第三章框架的 Resnet-50，模态对齐组件，跨模态学习组件的预训练权重基础上，进行多头注意力机制的后续训练，训练数据集是经过筛选打分组件过滤筛选后的训练数据集，目的是使得模型快速收敛，加快实验进度。该框架使用数据批次大小为 64 的 Adam 优化器^[62]进行训练。初始学习速率设置为 0.0001，动量设置为 0.999。

4.4.4 实验结果与分析

(1) 检索质量综合对比分析

通过表 4-1 展示了食物图像→食谱文本的检索实验结果，表 4-2 展示了食谱文本→食物图像的实验结果。在 1K 和 10K 测试数据集的检索任务中，SBNR⁺框架在所有评价指标中都优于所有基线。在 1K 测试数据集上，SBNR⁺得到的 MedR 为 1.0，因为其定义就是位于中间位置的数据检索命中的排名，那么排名最高也就只有 1.0，也就是说一次检索就可以命中结果，MedR 这个参数在 1K 数据子集的情况下，已经达到最高标准。

在 2019 年时，各个基线模型就已经在 1K 数据子集中 MedR 获得了 1.0 的最高成绩，这个评测指标的数据子集个数 1K 个过小或许已经不适合如今的检索框架。值得一提，ADAMINE 模型的论文中，可能出现了比较明显的小错误，在 1K 数据子集的测试情况下，ADAMINE 模型 MedR 取得了 1.0 的成绩，可以理解为其中位数处必定是 1.0，也就是说 R@1 的结果一定在 50%以上，而其论文中展示的结果来看最好结果 R@1 为 41.6%，那么 MedR 不会是 1.0，这两个实验结果出现了矛盾。在 10K 测试数据子集上，SBNR⁺得到的实验结果 MedR 为 1.0。

本章改进后的框架 SBNR⁺性能 R@K 结果超出其他方法。实验表明，多层句子注意力网络组件，多头注意力机制和筛选打分组件构造出的训练数据集等思路是有效的。句子级的食谱嵌入信息，包含了更多制作步骤间的顺序信息，更多食物制作步骤间的关系信息。

使用的多头注意力机制，相比于先前使用的普通注意力机制训练效果更好。原因可能是多头注意力机制在训练时参数不共享，因此多头注意力机制的每个头部可以通过学习得到数据中不同的信息，从而模型可以在不同的表示子空间里学习到相关的信息。多头注意力机制类似于机器学习中经常用到的集成思想，

通过集成多个注意力机制，避免过拟合等情况的出现，当每个模型之间差异性越大，模型集成后的效果将越好，以此达到良好的学习效果。食物图像中不同区域代表不同的子空间，在多头注意力机制处理菜谱时，菜谱文本子空间可以关注自身的不同子空间区域，比第三章的组件粒度更细致。

表 4-1 食物图像→食谱文本的检索实验结果

测试集大小	基线模型	食物图像→食谱文本检索			
		MedR	R@1	R@5	R@10
1K	CCA ^[45]	15.7	14.0	32.0	43.0
	SAN ^[38]	16.1	12.5	31.1	42.3
	JNE ^[32]	5.2	25.6	51.0	65.0
	ATTEN ^[39]	4.6	25.6	53.7	66.9
	ADAMINE ^[33]	2.5	39.8	69.0	77.4
	R2GAN ^[40]	2.0	39.1	71.0	81.7
	ACME ^[35]	1.0	51.8	80.2	87.5
	SBNR	1.0	52.7	81.7	88.9
	SBNR⁺	1.0	59.3	86.3	92.6
10K	JNE ^[32]	41.9	-	-	-
	ATTEN ^[39]	39.8	7.2	19.2	27.6
	ADAMINE ^[33]	16.5	12.5	31.5	42.2
	R2GAN ^[40]	13.9	13.5	33.5	44.9
	ACME ^[35]	7.0	22.0	45.3	56.6
	SBNR	7.0	22.1	45.9	56.9
	SBNR⁺	4.0	29.2	55.0	66.2

同样可以看出，利用筛选打分组件构造出新的训练数据集，并将框架在其上训练也起到了效果。这里使用的筛选打分组件主要处理两种类型的噪声，第

一种类型是食谱-食物图像匹配对确实是匹配的,但是由于食物图像中有效区域很小,复杂背景区域面积大,训练时不容易捕获其中的食物信息。第二种类型是,食谱-食物图像匹配对并不匹配,可能是由于数据集构成时,数据量庞大没有人工筛选导致的,这种类型的噪声会误导训练模型,使得模型向错误的方向学习。后续小节的检索结果展示中,会继续提出发现的新种类噪声,那就是菜谱中大量出现的食材价格问题,这会影响菜谱的嵌入和跨模态检索任务的实施。

表 4-2 食谱文本→食物图像的检索实验结果

测试集大小	基线模型	食谱文本→食物图像检索			
		MedR	R@1	R@5	R@10
1K	CCA ^[45]	43.0	24.8	24.0	35.0
	SAN ^[38]	42.3	-	-	-
	JNE ^[32]	5.1	25.0	52.0	65.0
	ATTEN ^[39]	4.6	25.7	53.9	67.1
	ADAMINE ^[33]	2.1	40.2	68.1	78.7
	R2GAN ^[40]	2.0	40.6	72.6	83.3
	ACME ^[35]	1.0	52.8	80.2	87.6
	SBNR	1.0	54.1	81.8	88.9
	SBNR⁺	1.0	59.8	86.7	92.8
10K	JNE ^[32]	39.2	-	-	-
	ATTEN ^[39]	38.1	7.0	19.4	27.8
	ADAMINE ^[33]	15.2	13.6	32.8	43.4
	R2GAN ^[40]	12.6	14.2	35.0	46.8
	ACME ^[35]	7.0	23.3	47.1	57.9
	SBNR	7.0	23.4	47.3	57.9
	SBNR⁺	4.0	30.3	55.6	66.5

Recipe 1M 中经常存在一对多的食谱-食物图像匹配对，表现为一个食谱对应多个食物图片，对应的多个图片中也经常出现只有个别食物图片拥有有效食物区域，其他图片中大多是制作步骤中的半成品展示和食谱每一步骤的图文讲解图，这种情况也需要打分筛选组件去除不必要的噪声图片。

从最终的实验结果来看，本文改进后的框架的综合性能是最为出色的，此框架可以使食谱-食物图像的特征嵌入在高维公共空间中，使同一类别的数据对距离大大近于不同类别的数据对，以达到跨模态对齐的目标。

（2）消融实验

本节将展示由于框架的不同模块的作用而带来收益的消融实验。首先，展示还未优化过的原始框架，也就是第三章提出基于句子级食谱文本嵌入编码的食谱-食物图像跨模态检索模型（SBNR）。随后以增量的方式添加更多的组件：首先，在此模型的基础上，改进了其中使用的三个注意力机制，分别设计了食谱配料，食谱标题，食谱制作步骤的多头注意力机制。这三个多头注意力机制的设计为（MTH）。最后，使用食谱-食物图像对噪声筛选打分组件对数据集进行筛选过滤，构筑出一个筛选后的数据集，提供给改进后的框架训练学习，表示为（IRF）。

表 4-3 消融实验 **MTH**：应用三个多头注意力机制，**IRF**：新构造训练数据集。

组件成分	MedR	R@1	R@5	R@10
SBNR	1.0	52.7	81.7	88.9
MTH	1.0	57.5	85.1	91.7
IRF(SBNR ⁺)	1.0	59.3	86.3	92.6

在表 4-3 中列出 1K 测试数据子集上，每一步消融实验的食物图像检索食谱的测试结果。如表 4-3 所示，当设计了新的多头注意力机制时（MTH），第三章提出的框架进一步提高结果，MedR（1.0），R@1（57.5）。本文认为主要原因是，先前的 SBNR 框架的各个组件工作都很出色，但受限于基础的注意力机制的性能，难以细致捕获菜谱文本句子级隐层信息，同时对食物图像过滤组件，普通的注意力机制难以过滤细粒度的噪声。而多头注意力作为如今性能最强的注意力机制之一，激活了 SBNR⁺的性能，表明了本文在其他方面所做的

努力是有效果的，尤其是句子级食谱的嵌入编码工作。当添加了食谱-食物图像对筛选打分组件之后（IRF），产生了锦上添花的效果，MedR（1.0），R@1（59.3），这也就是第四章改进后的SBNR⁺框架。此次操作表明，筛选过滤组件可以小幅度提升模型的泛化能力，它在一定程度上发挥了效果，作为辅助组件，这个过滤思路是值得尝试的。

4.4.5 样例分析

（1）食谱检索食物图片

图 4-8 中展示了食物图像检索结果的结果测试。在本文提出的框架中输入了三种不同的食谱（炖菜类、饮品类和炸制类的食谱）。如图 4-8 所示，top5 的食物图像展现出本文提出的食谱-食物图像跨模态检索框架，可以很好地捕捉食谱-食物图像两种模态之间的语义信息，并且 top1 都是正确结果，但还是存在不足。当蔬菜汤这个食谱被输入到框架中时，得到的食物图像都检索到了炖菜类的食物图像，食谱中的各种语义信息都展现在了检索出的食物图像中。对于第三个食物图像，可以看出已经存在很大噪声干扰，这里的噪声情况就是食物有效区域的不完全展示，和背景干扰信息盘子占据大量图像面积。对于第二个关于鸡尾酒饮品的食谱，在检索结果中，第二张食物图像的检索完美诠释了框架中过滤食物图像中噪声的能力。看到第三张食物图像，可以表明一个关键问题，食物图像过滤组件虽然发挥了重要作用，但是有时会成为其缺点，过滤能力过强导致筛选去除过多的其他信息，使得食物图像的某些主体被过滤，很明显第三张食物图像重点是汉堡，一旁的番茄酱只是陪衬，本框架却将关注点放在了番茄酱中。

对于第四张食物图像这个蛋糕类的食谱出现在了鸡尾酒饮品的菜谱检索中，其主要原因可能是，本文提出的框架还可以继续优化，其给食谱标题和食谱配料中的食物类别信息没有赋予足够的权重。此菜谱中的配料以及部分制作步骤的操作，与糕点类有相似之处，例如配料中的椰浆和南瓜汁，导致了糕点类的食谱出现。在第三个菜谱的检索中，可以看到检索框架还是出色的完成任务，所有检索出的食物图像确实饱含了菜谱语义，也可以展现出框架的食物图像噪声过滤能力。

对于第四章提出的筛选过滤打分组件的作用方面，可以看到虽然检索到的结果中 top1 都是正确的，但是其中不乏存在检索到极大噪声或者毫不相关的信

息，模型的泛化能力还需要继续改进。
















Recipe query	Top5 retrieved images				
Title: Hearty Root Veggie Soup Ingredients: olive oil, leeks, celery, garlic clove, turnips, rutabagas, russet potatoes, carrots... Instructions: Heat oil in heavy large pot over medium-low heat, Add leek, celery and garlic and saute until....					
Title: Puerto Aventuras Cocktail Ingredients: dark rum, peach liqueur, coconut cream pineapple juice, grenadine Instructions: Mix in a blender with ice. Increase or decrease ingredients based on your preference...					
Title: Vitello alla Parmigiano (Veal Parmesan) Ingredients: veal cutlets, flour, eggs, dry breadcrumbs, extra virgin olive oil, marinara sauce, parmesan cheese... Instructions: Preheat the broiler of an oven and place the oven rack 10 from the heating element...					

图 4-8 食谱检索食物图像结果展示

(2) 食物图片检索食谱

图 4-9 显示了从食物图像到食谱的检索结果。具有复杂背景的食物图像检索是从 Recipe 1M 中选择的，目的是想利用这些食物图像，测试本检索框架工作性能。在检索结果中，本文提出的框架都能一击命中正确结果。即使排在前几位的检索肉眼看起来结果非常相似，但本框架也可以检索到正确的结果。对于第二个食物图像的检索，可以看到其食物图像是十分复杂的，既包含了多种食物共存一个图像中的情况，又包含了各种无用的装饰和边框。不过本框架依然可以完成任务，这种性能上的泛化能力，食物图像过滤组件和跨模态学习组件功不可没。根据三个食物图像的检索结果，观察到检索排行第二的菜谱，并不是食物图像中关于炸鱼的描述而是芝士蛋糕。

通过仔细研究数据，本章发现了新的一种噪声类型，Recipe 1M 中的菜谱和食物图片都是从各大用户分享菜谱的网站爬取，从而导致其中存在一种新的噪声类型，制作步骤中包含很多价格信息。菜谱上传的用户细心上传了配料的价格信息，包含在了菜谱制作步骤中，目的是教导学习者购买正确的食材，但是这善意之举对于食谱-食物图像跨模态检索任务，却是噪声信息。后续如果需要继续优化模型，可以从这种新的噪声种类入手，解决大量存在的价格信息的

影响。


Image query			
Best match	Title: White Chocolate Macadamia Cherry Oatmeal Cookies Ingredients: all - purpose flour, baking soda, salt, butter, sugar, old fashioned oats, white chocolate chips... Instructions: Preheat oven to 325F. Line two baking sheets with parchment paper or use silicon baking pads....	Title: Chocolate Chip banana muffins Ingredients: sugar, flour, baking soda, vanilla, sour cream, eggs, butter, bananas, chocolate chips Instructions: Preheat oven to 350F. Line 12 cup muffin pan with cupcake liners or grease with butter well...	Title: Crunchy Fish Dippers Ingredients: firm white fish fillets, eggs, 4 cups honey-flavored multi-grain cereal flakes with oat clusters... Instructions: Preheat oven to 375F. Dip fish in eggs, turning over to evenly coat both sides. Coat fish evenly ...
Top2 retrieved recipes	Title: Banana Rum Coconut Cookies Ingredients: dark brown sugar, mashed bananas, reduced - fat mayonnaise, rum, all - purpose flour... Instructions: Preheat oven to 350. Place first 4 ingredients in a large bowl; beat with a mixer at medium speed until blended...	Title: Healthy Okara Snack Honey Lemon Muffin and Cake Ingredients: okara, eggs, sugar, skim milk, olive oil, lemon juice, honey, dried raspberries Instructions: Use a hand-held mixer and beat the eggs as much as possible. For better results, beat in a double broiler...	Title: PHILADELPHIA 3-STEP Low-Fat Berry Cheesecake Ingredients: low - fat graham crackers, fat free cream cheese, sugar, 1 tsp. lemon zest Safeway 4 ct For \$5.00 thru 02/09.. Instructions: Heat oven to 300F. Sprinkle graham crumbs onto bottom of 9-inch pie plate sprayed with cooking spray...

图 4-9 食物图片检索食谱结果展示

4.4.6 实验中的不足

如表 4-1，表 4-2 所示，在 10K 数据子集上，改进后框架的取得较好结果。不过此框架还可以继续改进。如下列举可以努力的方向：

(1) 首先，虽然食物图像筛选过滤组件和多头注意力机制发挥了很大作用，但是这里有一个需要考虑的问题。当进行噪声的筛选打分操作时，使用到了图像检测模型，而打分公式是根据图像检测模型检测出的食物图像有效区域去进行的。当食物图像中出现多个食物区域时，暂时的策略是选择面积最大的有效食物区域，当选择的这个食物区域并不是匹配对的食物图像时，就出现了手动创造了不必要噪声的操作。

(2) 其次，如图 4-10 所示，食物图像区域检测模型误差是不容忽视，第一张图图像检测模型将一本菜单图像区域错误识别为食物图像区域，第二张图右上角还有两个煎蛋，但图像检测模型并没有框画出此区域这也是一个出现误差的情况。虽然图像检测模型都提供检测区域并且识别的功能，但是检测误差

就存在的情况下，识别检测区域类别的功能误差就更大，导致并没有使用检测模型的识别功能。这些误差会经过后续我们的筛选公式的操作进一步扩大，最后会影响筛选后训练数据集的构建。不过正是由于 Recipe 1M 数据量巨大，才使用了这种筛选方法。数据集的数据量巨大带来的收益就是可以一定程度接受这种误差，但同时数据量巨大的弊端就是复杂情况多，更难以处理食谱-食物图像数据。

(3) 再者，由于实验条件受限制，无法使菜谱嵌入编码模型进行针对菜谱的细致的个性化训练学习，这可能会让菜谱嵌入编码的模型在编码时出现误差，可能因为它编码出的向量中食谱制作步骤间各步骤的关联性和顺序性并不强。同时，检索过程中，数据对的逐个计算欧氏距离所导致的问题就是检索缓慢，这也是列出的基线模型共同拥有的问题。此问题可以成为一个食谱-食物图像跨模态检索模型优化的方向，以供后续研究人员进行研究。进行课题研究最终都是要服务人们生活的，所以很多时候不但要考虑理论上的成功，还需要对实际生活中的是使用进行思考。如果后续要进行模型框架的落地操作，就必须面对用户在使用应用时的等待时间，而逐个计算欧氏距离并进行排序的检索操作过于缓慢，用户拍一张照片等待检索的菜谱，等待了数分钟都没有发回结果，那这个应用便没有竞争力。



图 4-10 食物图像检测组件误差情况展示

4.5 本章小结

本章提出了基于食物图像检测的噪声筛选打分过滤框架，主要是解决了食谱-食物图像跨模态检索中的噪声问题，以及改进注意力机制为多头注意力机

制，增强检索框架学习菜谱文本不同隐层句子级信息。最后对提出的噪声打分筛选框架以及多头注意力机制的性能进行测试，展示出测试结果并与现如今的各基线模型进行对比，阐述改进后框架的优点和存在的不足，为后续研究人员给予启发。

第5章 总结与展望

5.1 总结

随着社交媒体的发展，人们展示生活如此多娇的欲望正在高涨。俗话说，民以食为天，食物在人类生活上，健康上，幸福生活上都有重大影响。食品相关研究可为指导人类行为、改善人类健康。网络上充斥着各种美食分享信息，这海量的食物信息扑面而来，人们才更需要一个高效准确的食谱-食物图像跨模态检索框架。

本文的研究目标是提出一个高准确度的食谱-食物图像跨模态检索框架，通过句子级食谱嵌入，模态对齐，跨模态学习和食谱-食物图像噪声处理协作完成。并且在此框架上，进一步改进和研究，使其泛化能力进一步提高，检索效果进一步增强。本文所作的主要工作如下：

(1) 本文分析了如今流行的各种食谱-食物图像跨模态检索模型，讨论各模型的优势，并根据在 Recipe 1M 数据集上的实验结果分析各模型的不足，明确研究目标。介绍 Recipe 1M 数据集的信息，展示其成为如今最常用的食谱-食物图像跨模态检索任务数据集的原因。

(2) 提出了基于句子级编码的食谱-食物图像跨模态检索框架（SBNR 框架），以逐一解决（1）中分析出各模型的不足和现存问题。此框架分为四个组件，多层句子级别注意力机制组件进行食谱的句子级嵌入，食物图像过滤组件处理食物图像中的噪声，模态对齐组件强化模态分布对齐，跨模态学习组件使用两种改进的三重损失学习跨模态对齐信息。展示了本文提出框架在 Recipe 1M 上的实验结果及分析和样例分析，讨论框架暂存的优缺点，为后续研究工作提供指导作用。

(3) 提出了面向食谱文本的基于食物图像区域检测的噪声筛选打分组件，并对本文第三章提出的跨模态检索框架进一步改进，提出改进后的 SBNR⁺框架。改进部分主要体现在进行更细粒度的多头注意力机制的应用，增强食谱嵌入时学习多层隐层句子级嵌入信息。在 Recipe 1M 数据集上展示实验结果及分析和样例分析，讨论优化后的框架的优势和存在的不足，为未来研究工作提供建议。

通过以上三部分工作，本文提出改进后的食谱-食物图像跨模态检索框架，

在所有如今流行的检测指标 MedR, $R@1$, $R@5$ 和 $R@10$ 上超过其他食谱-食物图像跨模态检索模型。此实验结果极大肯定了本文提出的各组件工作性能, 也为后续研究人员带来一定启发。

5.2 未来展望

本文框架虽然已经取得了良好的实验结果, 但是仔细分析不能良好检索的食物数据, 还是发现了不足和可以继续优化的工作:

(1) 第四章实验展示中, 本文在观察和分析实验结果后, 发现第三种食谱-食物图像任务中普遍存在的噪声类型。此种噪声表现为, 菜谱制作步骤中, 菜谱制作者为了更好教导学习者应该购买的食材, 写入了大量的食材价格和食材重量信息。虽然这是菜谱作者的善意举动, 但是对于跨模态检索任务来说, 会一定程度上影响检索, 是应该考虑处理的噪声情况。

(2) 本文虽然第一章就提及到食谱-食物图像跨模态检索的速度缓慢, 但暂时还并没有在这个方面进行过工作。食谱-食物图像跨模态检索的检索方法是, 使用一个模态的嵌入向量去与另一个模态的一批次内所有嵌入向量计算欧氏距离, 通过排序找到距离最近的嵌入, 简单来说就是找出模态对齐最强的另一个模态嵌入向量。逐个计算距离的做法在实际使用中速度缓慢, 不能够满足用户的日常使用。如果希望食谱-食物图像跨模态检索框架能够应用于实际, 那么检索速度问题不可忽视。

致谢

经历 本科 的离开，我在硕士阶段醒过来。
我读 我写 我思考，学术并不能一丝怠慢。
听见 PYTHON 的启动，我在某年某月检索来。
我想 我等 我期待，LOSS 却不能因此安排。
阴天 傍晚 电脑前，未来有个嵌入在等待。
求导 求梯 求收敛，最优拐几个弯才来。
我遇见谁 会有怎样的卷积，我的论文 它在多远的未来。
我心里苦 来自发量和 BUG，我土下座 祈祷我的优化器。
我往前飞飞过一片论文海，我们也曾在 DEBUG 受伤害。
我看 ACM 入口有点窄，我排着队拿着毕业的号码牌。
总有一天我的程序会写完...

如今武汉理工大学研究生三年转瞬即逝，留下来的感叹和怀念，犹如孙燕姿的《遇见》，既有低沉压抑，却又硕果累累，向往美好的未来。公元 2014 年，风和日丽的下午，武理大学课堂里听闻有一位姓李名琳的教授，执教于武汉理工大学计算机学院。谈起李老师，学子们敬意油然而生，伴随着石楠花的香气遍布整个教室，均拍手评价，李老师夜以继日来教导，学生们脚踏实地学术搞。李老师凭借着深厚的学术功底和负责的教学态度，亲自演绎大制作科幻校园情景剧，《李老师！学生们的膝盖请拿去！》。公元 2018 年，跟随李老师研究学术，让我深刻体会到，李老师有三宝，细心，热情，颜值好。所谓兵者，诡道也，李老师凭实打实的硬实力和湘辣豪爽的性格，既可驰骋国际学术会议的战场上，又可年会聚餐与 IDEA 团队的同学谈笑风生，不禁让人感叹，“师者当如李琳老师”。研究生期间，国际学术会议的指尖离我的手腕只有 0.0001 公分，但是一秒之后这个指尖的主人将被我彻底赶上，因为我决定写一篇论文，虽然我此生写字无数，但拥有李老师指导的文字一定更有说服力。那时，深受国际学术会议全英文撰写高难度任务的阻碍，我犹如泰坦尼克号上的“Jack”，将与我的小论文“Rose”擦肩而过，如果给这个擦肩定一个期限，或许这一别，就是永远。眼看大事不妙，又见李老师脚踩祥云赶到，这畅快的学术指导堪比夏日的冰啤和烧烤，冰块和孜然要多少有多少，喝多的室友在傻笑，他说看见一个 IDEA 团队的学生将要在学术的战场上程风呼啸。

相传过去，李老师大手一挥，创办了如今的 IDEA 团队。这是一个友善和睦，却又奋进拼搏的团队，团队中的学生们都心怀“日复一日淬炼知识，待我把烈日磨成刀，不做只勇无谋的匹夫，我探我自己的学术道，大家均来自全国各地，带着中华民族的骨气，刻苦钻研同时也盼出人头地，八万里踏破铁鞋，还在被春寒料峭打磨，吾乃武汉理工学子，只盼学成后再报效家国。”。正是这股子拼劲，使我们团队一片欣欣向荣，互相帮助。在这里也愿团队中的大家以后人生路上披荆斩棘，让那一切发生的事情都是最好的安排。

写致谢时，正值农历生日，一大早收到家人们的祝福，成年人的情绪波动往往就藏在这日常生活的细节之中。说再多的感谢都是疲软，两座难以征服的大山压在心头，一座叫常回家看看，一座叫让家人过上好日子。而正是这两座大山，才是我不断努力奋斗的动力。

最后，祝一路上遇到的大家，若能如愿，愿能永恒。

参考文献

- [1] Achananuparp P, Lim E P, Abhishek V. Does journaling encourage healthier choices? Analyzing healthy eating behaviors of food journalers[C]//Proceedings of the 2018 International Conference on Digital Health. 2018: 35-44.
- [2] Nordström K, Coff C, Jönsson H, et al. Food and health: individual, cultural, or scientific matters?[J]. Genes & nutrition, 2013, 8(4): 357-363.
- [3] Min W, Jiang S, Liu L, et al. A survey on food computing[J]. ACM Computing Surveys (CSUR), 2019, 52(5): 1-36.
- [4] Ahn Y Y, Ahnert S E, Bagrow J P, et al. Flavor network and the principles of food pairing[J]. Scientific reports, 2011, 1(1): 1-7.
- [5] Chung J, Chung J, Oh W, et al. A glasses-type wearable device for monitoring the patterns of food intake and facial activity[J]. Scientific reports, 2017, 7(1): 1-8.
- [6] Sajadmanesh S, Jafarzadeh S, Ossia S A, et al. Kissing cuisines: Exploring worldwide culinary habits on the web[C]//Proceedings of the 26th international conference on world wide web companion. 2017: 1013-1021.
- [7] Aguilar E, Remeseiro B, Bolaños M, et al. Grab, pay, and eat: Semantic food detection for smart restaurants[J]. IEEE Transactions on Multimedia, 2018, 20(12): 3266-3275.
- [8] Deng L, Chen J, Sun Q, et al. Mixed-dish recognition with contextual relation networks[C]//Proceedings of the 27th ACM International Conference on Multimedia. 2019: 112-120.
- [9] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.
- [10] Martinel N, Foresti G L, Micheloni C. Wide-slice residual networks for food recognition[C]//2018 IEEE Winter Conference on applications of computer vision (WACV). IEEE, 2018: 567-576.
- [11] Min W, Liu L, Luo Z, et al. Ingredient-guided cascaded multi-attention network for food recognition[C]//Proceedings of the 27th ACM International Conference on Multimedia. 2019: 1331-1339.
- [12] Wu H, Merler M, Uceda-Sosa R, et al. Learning to make better mistakes: Semantics-aware visual food recognition[C]//Proceedings of the 24th ACM international conference on Multimedia. 2016: 172-176.
- [13] Fang S, Shao Z, Mao R, et al. Single-view food portion estimation: Learning

- image-to-energy mappings using generative adversarial networks[C]//2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, 2018: 251-255.
- [14] 庞广昌. 中华饮食文化和食品科学探源[J]. 食品科学, 2009(03):10-19.
- [15] Min W, Bao B K, Mei S, et al. You are what you eat: Exploring rich recipe information for cross-region food analysis[J]. IEEE Transactions on Multimedia, 2017, 20(4): 950-964.
- [16] McCrickerd K, Forde C G. Sensory influences on food intake control: moving beyond palatability[J]. Obesity Reviews, 2016, 17(1): 18-29.
- [17] Ofli F, Aytar Y, Weber I, et al. Is saki# delicious? the food perception gap on instagram and its relation to health[C]//Proceedings of the 26th International Conference on World Wide Web. 2017: 509-518.
- [18] Mejova Y, Abbar S, Haddadi H. Fetishizing food in digital age:# foodporn around the world[C]//Proceedings of the International AAAI Conference on Web and Social Media. 2016, 10(1).
- [19] 王来晓. 公众食品营养与食品健康现状调查分析[J]. 世界最新医学信息文摘, 2017(09):215-216.
- [20] Elswailer D, Trattner C, Harvey M. Exploiting food choice biases for healthier recipe recommendation[C]//Proceedings of the 40th international acm sigir conference on research and development in information retrieval. 2017: 575-584.
- [21] 李轩. 引入 Adaboost 概率矩阵分解的糖尿病个性化饮食推荐算法[D]. 吉林大学.
- [22] Farinella G M, Allegra D, Moltisanti M, et al. Retrieval and classification of food images[J]. Computers in biology and medicine, 2016, 77: 23-39.
- [23] Barlacchi G, Abad A, Rossinelli E, et al. Appetitoso: A search engine for restaurant retrieval based on dishes[J]. CLiC it (2016), 2016, 46.
- [24] Wang Y, Lin X, Wu L, et al. Effective multi-query expansions: Collaborative deep networks for robust landmark retrieval[J]. IEEE Transactions on Image Processing, 2017, 26(3): 1393-1404.
- [25] Wang Y, Lin X, Wu L, et al. Lbmch: Learning bridging mapping for cross-modal hashing[C]//Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval. 2015: 999-1002.
- [26] Wu L, Wang Y, Shao L. Cycle-consistent deep generative hashing for cross-modal retrieval[J]. IEEE Transactions on Image Processing, 2018, 28(4): 1602-1612.
- [27] Yu Y, Tang S, Raposo F, et al. Deep cross-modal correlation learning for audio and lyrics in music retrieval[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2019, 15(1): 1-16.

- [28] 冯方向. 基于深度学习的跨模态检索研究[D]. 北京邮电大学.
- [29] 花妍. 具有语义一致性的跨模态关联学习与信息检索[D]. 北京邮电大学, 2015.
- [30] 陈小平. 基于深度模型学习的跨模态检索[D]. 北京邮电大学, 2018.
- [31] Cao D, Yu Z, Zhang H, et al. Video-based cross-modal recipe retrieval[C]//Proceedings of the 27th ACM International Conference on Multimedia. 2019: 1685-1693.
- [32] Salvador A, Hynes N, Aytar Y, et al. Learning cross-modal embeddings for cooking recipes and food images[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 3020-3028.
- [33] Carvalho M, Cadène R, Picard D, et al. Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings[C]//The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018: 35-44.
- [34] Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification[J]. arXiv preprint arXiv:1703.07737, 2017.
- [35] Wang H, Sahoo D, Liu C, et al. Learning cross-modal embeddings with adversarial networks for cooking recipes and food images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 11572-11581.
- [36] Xu H, Zhang H, Han K, et al. Learning alignment for multimodal emotion recognition from speech[J]. arXiv preprint arXiv:1909.05645, 2019.
- [37] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. arXiv preprint arXiv:1706.03762, 2017.
- [38] Chen J, Pang L, Ngo C W. Cross-modal recipe retrieval: How to cook this dish?[C]//International Conference on Multimedia Modeling. Springer, Cham, 2017: 588-600.
- [39] Chen J J, Ngo C W, Feng F L, et al. Deep understanding of cooking procedure for cross-modal recipe retrieval[C]//Proceedings of the 26th ACM international conference on Multimedia. 2018: 1020-1028.
- [40] Zhu B, Ngo C W, Chen J, et al. R2gan: Cross-modal recipe retrieval with generative adversarial network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 11477-11486.
- [41] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
- [42] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

-
- [43] Hochreiter S, Schmidhuber J. Long short-term memory[J]. *Neural computation*, 1997, 9(8): 1735-1780.
 - [44] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. *arXiv preprint arXiv:1301.3781*, 2013.
 - [45] Hotelling H. Relations between two sets of variates[M]//*Breakthroughs in statistics*. Springer, New York, NY, 1992: 162-190.
 - [46] Salvador A, Drozdzal M, Giro-i-Nieto X, et al. Inverse cooking: Recipe generation from food images[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019: 10453-10462.
 - [47] Zhu B, Ngo C W. CookGAN: Causality based Text-to-Image Synthesis[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020: 5519-5527.
 - [48] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. *arXiv preprint arXiv:1406.2661*, 2014.
 - [49] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein gans[J]. *arXiv preprint arXiv:1704.00028*, 2017.
 - [50] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. *arXiv preprint arXiv:1810.04805*, 2018.
 - [51] Jawahar G, Sagot B, Seddah D. What does BERT learn about the structure of language?[C]//*ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*. 2019.
 - [52] Bossard L, Guillaumin M, Van Gool L. Food-101—mining discriminative components with random forests[C]//*European conference on computer vision*. Springer, Cham, 2014: 446-461.
 - [53] Kawano Y, Yanai K. Foodcam: A real-time food recognition system on a smartphone[J]. *Multimedia Tools and Applications*, 2015, 74(14): 5263-5287.
 - [54] Meyers A, Johnston N, Rathod V, et al. Im2Calories: towards an automated mobile vision food diary[C]//*Proceedings of the IEEE International Conference on Computer Vision*. 2015: 1233-1241.
 - [55] Xu R, Herranz L, Jiang S, et al. Geolocalized modeling for dish recognition[J]. *IEEE transactions on multimedia*, 2015, 17(8): 1187-1199.
 - [56] Kusmierczyk T, Trattner C, Nørnvåg K. Understanding and predicting online food recipe production patterns[C]//*Proceedings of the 27th ACM conference on hypertext and social media*. 2016: 243-248.
 - [57] Wang X, Kumar D, Thome N, et al. Recipe recognition with large multimodal food

- dataset[C]//2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE, 2015: 1-6.
- [58] Chen J, Ngo C W. Deep-based ingredient recognition for cooking recipe retrieval[C]//Proceedings of the 24th ACM international conference on Multimedia. 2016: 32-41.
- [59] Peng Y, Qi J. CM-GANs: Cross-modal generative adversarial networks for common representation learning[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2019, 15(1): 1-24.
- [60] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.
- [61] Chorowski J, Bahdanau D, Serdyuk D, et al. Attention-based models for speech recognition[J]. arXiv preprint arXiv:1506.07503, 2015.
- [62] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.
- [63] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(6): 1137-1149.
- [64] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.
- [65] Kawano Y, Yanai K. Automatic expansion of a food image dataset leveraging existing categories with domain adaptation[C]//European Conference on Computer Vision. Springer, Cham, 2014: 3-17.

攻读硕士学位期间获得与学位论文相关的研究成果

- [1] Zan Z, Li L, Liu J, et al. Sentence-based and Noise-robust Cross-modal Retrieval on Cooking Recipes and Food Images[C]//Proceedings of the 2020 International Conference on Multimedia Retrieval (ICMR2020). 2020: 117-125. (CCF B, EI 检索, 第一作者)