硕士学位论文

基于目标检测和图卷积的跨模态检索算法

Cross-modal retrieval algorithm based on target detection and graph convolution

苏林

哈尔滨工业大学

2020 年 6 月

国内图书分类号：TP391.4　　学校代码：10213
国际图书分类号：004.4　　密级：公开

工程硕士学位论文

基于目标检测和图卷积的跨模态检索算法

硕士研究生：苏林

导　　师：邬向前教授

申请学位：工程硕士

学　　科：计算机技术

所在单位：计算机科学与技术学院

答辩日期：2020 年 6 月

授予学位单位：哈尔滨工业大学

Classified Index: TP391.4
U.D.C: 004.4

Dissertation for the Master Degree in Engineering

Cross-modal retrieval algorithm based on target detection and graph convolution

Candidate: Lin Su
Supervisor: Prof. Xiang qian Wu
Academic Degree Applied for: Master of Engineering
Speciality: Computer Technology
Affiliation: School of Computer Science and Technology
Date of Defence: June, 2020
Degree-Conferring-Institution: Harbin Institute of Technology

摘　要

随着数据时代和信息时代的到来。信息和数据越来越成为社会经济发展和人们生活提高的重要推动力。而检索作为高效获取信息方法就显得尤为重要。跨模态检索作为获取跨模态信息的重要方法，社会价值巨大，自然引起越来越多人的关注和研究。随着深度学习和人工智能的发展，跨模态检索也取得了长足发展。其中目标检测和图卷积相结合的深度学习方法越来越引起人们的注意，成为跨模态检索中的一个重要研究方向。本文同样关注这一方向，并且从以下三个方面在该方向上进行了研究。

一、开发多层图卷积中不同层的学习能力，提升多层图卷积的学习能力。在跨模态检索中图卷积的应用往往是连续多层的，这样相比单层可以起到更好的效果。但是以往只使用了最后一层图卷积的输出结果，而没有对多层图卷积其它层进行立体的开发。本文利用跳跃的设计模式，将中间层跳跃过后面的图卷积直接进入下面流程，来实现对多层图卷积不同层的灵活控制，使得能够对不同层进行不同程度的学习和开发。结果表明本文的方法可以整体提升图卷积的特征学习能力，提升跨模态检索的效果。

二、通过多粒度文本特征学习来提高文本部分的特征学习能力。在以往的跨模态检索中，文本部分的特征学习都是使用简单的循环神经网络 GRU 来完成。这样的文本特征学习方式太过简单，无法充分学习到文本信息。我们采用多粒度文本特征学习来代替 GRU。通过多粒度文本特征学习，文本特征学习部分的学习能力得到增强，文本部分可以学到丰富的多种粒度的文本信息，能够将文本中的信息充分学习到。结果表明多粒度文本特征学习增强了文本特征学习能力后提高了检索的抗干扰性，当检索的数据量变大时，检索效果下降的会相对较少。

三、通过混合检索来提升整体检索效果。以往的跨模态检索都是在算法中构建一个检索模型，用一个检索模型的检索能力来进行检索。受混合推荐算法的启发。我们在一个算法框架中构建两个检索模型，让两个检索模型同时起作用，通过一定方式将这两个检索模型整合在一起，使用两个检索模型整体的检索能力来进行检索。由于每个检索模型都有一定的检索能力，当这些模型叠加在一起时，检索能力会相互增强，得到的整体检索效果也就更好。当两个模型的特征向量在特征空间中的态势一致性较高时，增强的效果就会更加明显。由于是两个检索模型共同起作用，因此抗干扰性会增强，当检索的数据规模变大时，检索效果下降的会更少一些。

本文集中从多层图卷积中间层使用、多粒度文本特征学习、混合检索三个方面对目标检测和图卷积相结合的跨模态检索算法进行了立体研究，并且实验证明本文的改进方法和理论创新都在一定程度上提升了检索效果。

关键词：跨模态；检索；图卷积；目标检测

Abstract

With the advent of the data age and the information age. Information and data have increasingly become an important driving force for social and economic development and improvement of people's lives. Retrieval is particularly important as an efficient way to obtain information. As an important method for obtaining cross-modal information, cross-modal retrieval has huge social value, and naturally attracts more and more people's attention and research. With the development of deep learning and artificial intelligence, cross-modal retrieval has also made great progress. Among them, the deep learning method combining target detection and graph convolution has attracted more and more attention, and has become an important research direction in cross-modal retrieval. This article also pays attention to this direction, and conducts research in this direction from the following three aspects.

1. Develop the learning ability of different layers in the multi-layer graph convolution to enhance the learning ability of the multi-layer graph convolution. The application of graph convolution in cross-modal retrieval is often continuous multiple layers, so that it can play a better effect than a single layer. However, in the past, only the output result of the last layer convolution was used, and the three-dimensional development of other layers of the multi-layer image convolution was not used. In this paper, we use the jump design mode to jump the middle layer through the subsequent picture convolution and directly enter the following process to achieve flexible control of the different layers of the multi-layer picture convolution, enabling different levels of learning and development of different layers. The results show that the method in this paper can improve the feature learning ability of graph convolution as a whole and improve the effect of cross-modal retrieval.

2. Through multi-granularity text feature learning to improve the feature learning ability of text part. In the previous cross-modal retrieval, the feature learning of text is accomplished by using a simple cyclic neural network (GRU). Such a learning method of text features is too simple to fully learn the text information. Instead of GRU, we use multi-granularity text feature learning.

Through multi-granularity text feature learning, the learning ability of the text feature learning part is enhanced. The text part can learn rich multi-granularity text information and fully learn the information in the text. The results show that multi-granularity text feature learning enhances the ability of text feature learning and improves the anti-interference ability of retrieval. When the amount of retrieved data is larger, the retrieval effect will decline relatively less.

3. Improve the overall retrieval effect through mixed retrieval. In the past, cross-modal retrieval has always built a retrieval model in the algorithm and used the retrieval capability of the model to carry out the retrieval. Inspired by hybrid recommendation algorithms. We construct two retrieval models in an algorithm framework, make the two retrieval models work at the same time, integrate the two retrieval models together in a certain way, and use the overall retrieval capability of the two retrieval models for retrieval. Because each retrieval model has certain retrieval ability, when these models are superimposed together, the retrieval ability will be enhanced mutually, and the overall retrieval effect will be better. When the state consistency of the feature vectors of the two models is high in the vector space, the enhancement effect will be more obvious. As the two retrieval models work together, the anti-interference will be enhanced. When the data size of retrieval is larger, the retrieval effect will decline less.

This paper focuses on the research on the cross-modal retrieval algorithm combining target detection and graph convolution from three aspects of multi-layer graph convolution middle layer using, multi-granular text feature learning, and hybrid retrieval. And experiments show that the improved methods and theoretical innovations in this paper have improved the retrieval effect to a certain extent.

Keywords: cross-modality; retrieval; graph convolution; target detection

目录

第 1 章 绪　论

1.1 课题研究背景和意义

本课题来源于“司法舆情监控系统”项目。该项目是要实现一个舆情监控系统，可以对互联网上的文字、图片、视频等多媒体信息进行分析，获悉当前社会舆论情况，发现舆论中存在的问题，及时发现不法分子、组织破坏社会舆论的行为，维护社会舆论秩序，保证社会舆论正常进行。跨模态信息检索是对信息进行跨模态的检索和收集，是进行社会舆论监察的基础。

跨模态检索在社会中有着重要的应用价值。人们在生活中经常会遇到一些情况，只拿到了图片、声音等信息，但是不知道具体的信息，此时就需要跨模态检索来帮忙找到相应的信息。例如，在生活中人们可能看到了一个东西却不认识，这时可以用手机拍照通过跨模态检索来知道具体是什么东西，获得相关的介绍信息。听到一首歌不知是什么歌曲，可以通过跨模态检索来查询歌曲的名字，等等。现实生活中我们往往会只有一个东西的某种模态的信息，需要查找具体的实际信息，这就是跨模态检索的使用价值。同样有时我们知道一个事物，想查找相关的视频、图片、音频等信息，来对这个事物有一个全面、生动的了解。这时也需要用到跨模态检索。例如想要了解某种动物，只是文本说明显得太无味，甚至不清楚到底说的是什么。这时查找到这种动物的照片就一目了然了，若再找到这种动物的视频或者音频，听到它的叫声，看到它是如何活动的。那时对这个动物的认识就会生动而深刻。因此跨模态检索在人们的实际生活中有着广泛而重要的应用，可以极大地改善人们的生活。同时当社会中出现重大事件，需要及时获取相关信息时，跨模态检索也具有重要的应用价值。例如，当一种疾病爆发时，我们需要快速获取疾病传染地区的情况，尤其是当地实际的具体情况。这些信息通过文本很难形象具体地传达出来。图片和视频信息则是了解病发地区真实情况的最有效的方式。跨模态检索在此时就扮演了重要角色。一个地区发生了重大事件，对当地造成了重大影响，想了解人们对此事的情感态度。最直接有效的仍然是从相关的图片和视频来获得，人们的动作、反应、面部表情最能体现人们的情感状况。而真实的图片和视频也最能直接反应当地的实际具体情况。这些都是跨模态检索的重要应用场景，也是跨模态检索重要应用价值的具体体现。所以跨模态检索有着重要的实际应用价值。

1.2 国内外研究现状

跨模式检索作为人工智能的一个重要领域，在社会经济中有重大的价值。到目前已经有了大量的这方面的研究工作。大体可以将其分为两类：实值表示学习；二进制表示学习。 对于实值表示学习，学习到的各种模态数据的通用表示都是实值。 为了加快跨模态检索，二进制表示学习方法旨在将数据的不同模态转换成公共汉明空间，在该汉明空间中，跨模态相似性搜索速度很快。 由于表示形式被编码为二进制代码，因此导致了信息丢失，检索精度通常会略有下降。

跨模式检索的主要困难是如何测量不同模态数据之间的内容相似性。子空间学习方法是最流行的一种方法。他们旨在学习由不同模态数据共享的公共子空间，其中可以测量不同模态数据之间的相似性。无监督子空间学习方法使用成对信息来学习跨多模态数据的公共潜在子空间。它们在公共子空间中不同的数据模态之间强制执行成对接近。典型相关分析（CCA）是最流行的无监督子空间学习方法之一，用于建立来自不同模态的数据之间的模态关系。 Rasiwasia等[1]提出了一种用于跨模态多媒体检索的两阶段方法。在第一阶段，CCA 通过最大化两个模态之间的相关性来学习公共子空间。 然后，学习语义空间以测量不同模态特征的相似性。Zhu 等[2]提出了一种针对交叉模式检索问题的贪婪词典构建方法。通过在目标函数中为两种模态包括重构误差项和最大平均差异（MMD）测量，可以保持紧凑性和模态适应性。另一个不受监督的方法是主题模型。主题模型已广泛应用于特定的交叉模态问题。为了捕获图像和注释之间的相关性，潜在狄利克雷分配（LDA）[3]已扩展为学习多模态数据的联合分布。Corr-LDA 使用主题作为共享的潜在变量，这些变量表示多模式数据中互相关的根本原因。Tr-mm LDA 学习了两组独立的隐藏主题，以及一个回归模块，该模块捕获了更一般的关联形式，并允许从另一组线性预测一组主题。Jia 等[4]提出了一种新的概率模型（多模态文档随机字段，MDRF），以学习跨模态的一组共享主题。 该模型在文档级别定义了 Markov 随机字段，该字段允许对更灵活的文档相似性进行建模。

现实中通常将不同类型的数据用于描述网络中相同的事件或主题。例如，用户生成的内容通常涉及来自不同形式的数据，例如图像，文本和视频。 这使得传统方法很难获得多模态数据的联合表示。Ngiam 等[5]受到深度学习进展的启发。应用深度网络来学习多种形式的特征，其重点是学习与嘴唇视频耦合的语音音频的表示形式。然后，玻尔兹曼机[6]成功地学习了多模态数据的联合表示。它首先使用单独的模态友好的潜在模型来学习每种模态的低级表示，然后

融合到高层的深层架构的联合表示中。受使用深度网络的表示学习的启发，Yan 和 Mikolajczyk [7]提出了一种基于深度典范相关分析的端到端学习方案（端到端 DCCA）。Feng 等[8]提出了一种涉及对应自动编码器（Corr-AE）的新型模型，用于交叉模式检索。该模型是通过关联两个单模式自动编码器的隐藏表示而构建的。利用一种新颖的目标来最小化每个模态的表示学习错误和两个模态的隐藏表示之间的相关性学习错误的线性组合，从而对模型进行整体训练。 相关学习误差的最小化迫使模型学习仅具有不同模态中的公共信息的隐藏表示，而表示学习误差的最小化则使得隐藏表示足以重建每种模态的输入。Xu 等[9]提出了一个统一的框架，可以联合建模视频和相应的文本句子。 该框架由语义语言模型，深层视频模型和联合嵌入模型三部分组成。在他们的语言模型中，他们提出了一种依赖树结构模型，该模型将句子嵌入到连续的向量空间中，以保留视觉上扎根的含义和单词顺序。在视觉模型中，他们利用深度神经网络从视频中捕获基本的语义信息。 在联合嵌入模型中，它们将联合空间中的深层视频模型和语义语言模型的输出距离最小化，并共同更新这两个模型。基于这三个部分，该模型能够完成三个任务：自然语言生成、视频检索语言、语言视频检索。

Quadrianto 和 Lampert [10]提出了一种新的度量学习方案（多视图邻域保存投影，Multi-NPP），将不同的模态投影到共享特征空间中，其中欧几里得距离提供了有意义的模态内和模态间相似性。Zhai 等[11]提出了一种新的方法，称为具有全局一致性和局部平滑度的多视图度量学习（MVML-GL）。该框架包括两个主要步骤。第一步，他们寻求全局一致的共享潜在特征空间。在第二步中，通过正则化局部线性回归学习输入空间和共享潜空间之间的显式映射函数。Zhai 等[12]提出一种联合图正则化异构度量学习（JGRHML）算法，以学习用于跨模式检索的异构度量。 他们基于异构度量，通过标签传播进一步学习了高级语义度量。为了预测社交媒体之间的联系，Yuan 等人[13]设计了一个关系生成深层信任网（RGDBN）模型，以学习社交媒体的潜在功能，该模型利用了网络中社交媒体之间的关系。 在 RGDBN 模型中，项目之间的链接是根据它们潜在特征的相互作用生成的。 通过将印度自助餐流程整合到经过修改的 Deep Belief Nets 中，他们学习了潜在的功能，该功能可以最好地嵌入媒体内容和观察到的媒体关系。 该模型能够分析异构数据和同类数据之间的联系，也可用于跨模态检索。Wang 等[14]提出了一种基于模态特定特征学习的新型模型，称为模态特定深度结构（MSDS）。考虑到不同模态的特征，该模型使用两种类型的卷积神经网络将原始数据分别映射到图像和文本的潜在空间表示。特别的，用于文本的基于卷积的网络涉及词嵌入学习，事实证明，这种学习对于提取有意义的

文本特征以进行文本分类十分有效。 在潜在空间中，图像和文本的映射特征形成相关和不相关的图像-文本对，被一对多的学习方案所使用。

Weston 等[15]通过学习图像和注释的联合表示，引入了用于图像注释的可伸缩模型。它针对给定图像学习在注释排名列表顶部的优化精度，并为图像和注释学习低维联合嵌入空间。Lu 等[16]提出了一种用于交叉模式检索的交叉模式排序算法，称为潜在语义交叉模式排序（LSCMR）。他们利用结构化 SVM 学习度量标准，从而可以优化由查询距离引起的数据排名。但是，LSCMR 没有充分利用双向排名示例（双向排名意味着在训练中同时使用了文本查询图像排名和图像查询文本排名示例）。Yao 等[17]提出了一种新颖的排名典范相关分析（RCCA），用于学习查询和图像相似性。RCCA 用于调整 CCA 学习的子空间，以进一步保留点击数据中的偏好关系。

受到深度学习进展的启发，弗罗姆等[18]提出了一种新的深度视觉语义嵌入模型（DeViSE），其目的是利用在文本域中学习到的语义知识，并将其转移到经过训练进行视觉对象识别的模型中。Karpathy 等[19]引入了图像和句子的双向检索模型，该模型为深度神经网络制定了结构化的最大利润目标，该神经网络将视觉和语言数据嵌入到通用的多模态空间中。与以前的将图像或句子直接映射到公共嵌入空间中的模型不同，此模型在更高级的层次上起作用，并将图像（对象）的片段和句子（类型化的依赖关系）类型嵌入到公共空间中。Jiang 等[20]利用现有的图像文本数据库来优化用于交叉模式检索的排序功能，称为深度合成交叉模态学习排名（C2MLR）。C2MLR 考虑从优化成对排名问题的同时增强局部对齐和全局对齐的角度学习多模态嵌入。 特别地，局部对齐（即视觉对象和文本单词的对齐）和全局对齐（即图像级别和句子级别的对齐）被联合利用以最大程度地学习多模态通用嵌入空间。Hua 等[21]开发了一种用于交叉模态检索的新型深度卷积体系结构，名为“使用深度卷积体系结构的交叉模态相关学习”（CMCDCA）。 它由视觉特征表示学习和具有较大余量原理的跨模态相关学习组成。

为了获得更具区分性的公共表示形式，有监督的方法利用标签信息，该信息在公共表示空间中的类之间提供了更好的分隔。有监督子空间学习方法将不同类别的样本强制映射到较远的位置，而相同类别的样本则尽可能地靠近。为了获得更多的判别子空间，一些工作将 CCA 扩展到有监督的子空间学习方法。Sharma 等[22]提出了一种 CCA 的监督扩展，称为广义多视图分析（GMA）。它将线性判别分析（LDA）和边际 Fisher 分析（MFA）扩展到它们的多视图对等物，即广义多视图 LDA（GMLDA）和广义多视图 MFA（GMMFA），并将其

应用于处理跨媒体检索问题。GMLDA 和 GMMFA 考虑了语义类别，取得了可喜的结果。Rasiwasia 等[23]将目前的聚类典范相关性分析（聚类-CCA）用于减少两种数据形式的联合维数。Cluster-CCA 能够学习可区分的低维表示形式，该表示形式可最大程度地提高两种数据模态之间的相关性，同时在学习空间上隔离不同的类。Ranjan 等[24]引入了多标签规范相关分析（ml-CCA），它是 CCA 的扩展，用于通过考虑多标签注释形式的高级语义信息来学习共享子空间。 他们还介绍了 Fast ml-CCA，这是 ml-CCA 的计算有效版本，能够处理大规模数据集。Jing 等[25]提出了一种新颖的基于视图内和视图间监督相关分析（I2SCA）的多视图特征学习方法。它探索了每个视图内以及所有视图之间样本的有用的相关信息。除了基于监督的 CCA 的方法外，Lin 和 Tang[26]提出了一种通用的判别特征提取（CDFE）方法来学习一个通用特征子空间，在该子空间中散布矩阵内和散布矩阵之间的差异最大。

Wang 等[27]提出了一种新的针对交叉模态匹配问题的正则化框架，称为 LCFS（学习耦合特征空间）。它将耦合线性回归，范数和跟踪范数统一为通用的最小化公式，以便可以同时执行子空间学习和耦合特征选择。此外，他们在[28]中将这个框架扩展到两种以上的情况，扩展版本称为 JFSSL（联合特征选择和子空间学习）。主要扩展概括两个方面，一方面他们提出了一种多模态图，以更好地建模数据的不同模态之间的相似关系，这在计算成本和检索性能方面均优于低秩约束。另一方面他们提出了一种新的迭代算法来求解修正后的目标函数，并给出了其收敛性的证明。受到（半）耦合字典学习思想的启发，Zhuang 等[29]将耦合字典学习带入交叉模式检索的监督稀疏编码，这被称为具有多模态检索组结构的监督耦合字典学习（SliM 2）。 它可以利用类信息来共同学习判别式多模态字典以及不同模态之间的映射功能。Liao 等[30]提出了一种非参数贝叶斯上游监督（NPBUS）的多模态主题模型，用于分析多模态数据。NPBUS 模型允许灵活地学习具有各个模态以及在不同模态之间的主题的相关结构。通过合并多模态数据共享的上游监管信息，该模型变得更具区分性。此外，它能够自动确定每个模态中潜在主题的数量。Wang 等[31]提出了一种用于跨媒体检索的有监督的多模态主题互增强建模（M3R）方法，该方法旨在建立一个联合的跨模态概率图形模型，以通过模型因素之间的适当交互来发现相互一致的语义主题。

Wang 等[32]提出了一种正规化的深度神经网络（RE-DNN），用于跨模态的语义映射。 他们设计并实现了一个 5 层神经网络，用于将视觉和文本特征映射到公共语义空间中，从而可以测量不同模态之间的相似性。Li 等[33]提出了一种

深度学习方法来解决带有多个标签的跨媒体检索问题。对提出的方法进行了监督，并可以根据它们共享的接地真实概率矢量建立两种模态之间的相关性。 这两个网络都有两个隐藏层和一个输出层，并且平方损耗被用作损耗函数。Castrejon 等[34]提出了两种方法来规范交叉模态卷积网络，以便中间跨模态表示对齐。这项工作的重点是在模态明显不同（例如，文本和自然图像）并具有类别监督的情况下学习跨模态表示。Wang 等[35]提出了一种有监督的多模态深度神经网络（MDNN）方法，该方法由深度卷积神经网络（DCNN）模型和神经语言模型（NLM）组成，以分别学习图像模态和文本模态的映射功能。它利用了标签信息，因此可以学习针对嘈杂的输入数据的强大映射功能。Li 等[63]提出了一个简单且可解释的推理模型 VSRN，以通过区域关系推理和全局语义推理生成增强的视觉表示。设计了一种解释方法，以可视化和验证所生成的图像表示形式可以捕获场景的关键对象和语义概念，从而使其与相应的文本标题更好地对齐。Frome 等[66]提出了一种特征嵌入框架，该框架使用 Skip-Gram 和 CNN 来提取跨模态的特征表示。然后采用等级损失来鼓励不匹配的图像-文本对之间的距离大于匹配的对之间的距离。Kiros 等[67]使用类似的框架并采用 LSTM 代替 Skip-Gram 来学习文本表示。

大多数现有的实值跨模态检索技术都是基于蛮力线性搜索，对于大规模数据而言这非常耗时。加快相似性搜索的一种实用方法是二进制表示学习，这称为哈希。现有的哈希方法可以分类为单模态哈希，多视图哈希和跨模态哈希。上面提到的方法集中于学习具有同类特征的数据对象的哈希函数。但是，在实际应用中，我们经常从数据对象中提取具有不同属性的多种类型的要素。因此，多视图哈希方法利用包含在不同特征中的信息来学习更准确的哈希码。跨模态散列方法旨在发现数据的不同模态之间的相关性，以实现跨模态相似性搜索。他们将不同的数据模态投影到公共的汉明空间中以进行快速检索。类似地，交叉模态哈希方法可以分类为：无监督方法，基于成对的方法和有监督的方法。关于基于秩的交叉模式哈希的文献很少。根据学习到的哈希函数是线性还是非线性，交叉模态哈希方法可以进一步分为两类：线性建模和非线性建模。线性建模方法旨在学习线性函数以获得哈希码。然而，非线性建模方法以非线性方式学习哈希码。

Rastegari 等[36]提出了一种可预测的双视图哈希（PDH）算法的两种模式。他们制定了一个目标函数，以保持二进制代码的可预测性，并通过应用基于块坐标下降的迭代方法来优化目标函数。Ding 等[37]提出了一种新的哈希方法，称为集体矩阵分解哈希（CMFH）。 CMFH 假定实例的所有模态都生成相同的哈

希码。 它通过一个潜在因子模型从集体矩阵分解中学习统一的哈希码。Song等[38]提出了一种新颖的媒体间哈希（IMH）模型，以将多峰数据转换为公共汉明空间。 该方法探讨了媒体间一致性和媒体内一致性，以得出有效的哈希码，在此基础上学习了哈希函数，可以将新的数据点有效地映射到汉明空间中。Zhu等[39]提出了一种新颖的哈希方法，称为线性交叉模态哈希（LCMH），以实现多媒体搜索的可扩展索引。该方法在训练阶段达到了训练数据大小的线性时间复杂度。关键思想是将每个模态的训练数据划分为k个聚类，然后用到聚类的k个质心的距离表示每个训练数据点，以保留每个模态的内部相似性。 为了保持跨不同模态的数据点之间的相似性，它们将派生的数据表示形式转换为公共的二进制子空间。大多数现有的交叉模态哈希方法学习的哈希函数是线性的。为了捕获更复杂的数据结构，有人研究了非线性哈希函数学习。Wang等[40]基于堆叠式自动编码器提出了一种有效的非线性映射机制，用于多模式检索，称为多模式堆叠自动编码器（MSAE）。映射功能是通过优化新的目标函数来学习的，该目标函数可以有效地捕获来自异构源的数据的模式内和模式间语义关系。MSAE的堆叠结构使该方法能够学习非线性投影而不是线性投影。Wang等[41]提出了一种使用正交正则化（DMHOR）的深度多模态散列，以学习准确而紧凑的多模态表示。该方法可以更好地捕获模态内和模态间的相关性，以学习准确的表示。同时，为了使表示紧凑并且减少代码中的冗余信息，对学习的加权矩阵施加正交正则化器。

Bronstein等[42]提出了第一种交叉模态哈希方法，称为交叉模态相似性敏感哈希（CMSSH）。CMSSH以标准的Boosting方式学习双峰情况的哈希函数。具体而言，在给定两种数据集形式的情况下，CMSSH学习两组哈希函数，以确保如果两个数据点（具有不同形式）相关，则它们对应的哈希码相似且不同。但是，CMSSH仅保留模式间的相关性，而忽略模式间的相似性。Zhen等[43]提出了一种新的多模态哈希函数学习方法，称为共正则哈希（CRH），它基于增强的正则化框架。通过求解DC（凸函数的差）程序来学习哈希码每一位的哈希函数。Hu等[44]提出了一种用于交叉模态检索的多视图哈希算法，称为迭代多视图哈希（IMVH）。IMVH旨在学习区分性哈希函数，以将多视图数据映射到共享的汉明空间中。它不仅保留了视图内的相似性，而且将视图间的相关性合并到编码方案中，在该方案中，它将相似点映射为接近并推开了相异点。跨模态哈希方法通常假定哈希数据驻留在公共汉明空间中。但是，这可能是不合适的，尤其是在模态完全不同时。为了解决这个问题，Ou等[45]提出了一种新颖的关系感知异构哈希（RaHH），它为从多个异构域生成数据实体的哈希

码提供了一个通用框架。与某些将异类数据映射到公共汉明空间中的现有交叉模态哈希方法不同，RaHH 方法为每种类型的数据实体构造一个汉明空间，并同时学习它们之间的最佳映射。 RaHH 框架对数据实体之间的同质和异质关系进行编码，以学习哈希码。Wu 等[46]提出了一种跨模态哈希方法，称为量化相关哈希（QCH），它考虑了域上的量化损失以及域之间的关系。与先前的方法分开量化器的优化和域相关性的最大化不同，此方法同时优化了两个过程。通过最大化跨域的哈希码之间的相关性，可以建立描述相同对象的域之间的基本关系。

Zhen 等[47]提出了一种称为多模态潜在二进制嵌入（MLBE）的概率潜在因子模型，以学习用于跨模态检索的哈希函数。 MLBE 使用一个生成模型来编码跨多个模态的数据对象的内部相似度和外部相似度。 基于最大后验估计，有效地获得了二进制潜在因子，然后将其作为 MLBE 中的哈希码。Zhai 等[48]提出了一种新的跨局部相似性搜索的参数局部多模态散列（PLMH）方法。 PLMH 学习一组哈希函数，以在本地适应每个模态的数据结构。 在输入空间的不同位置学习了不同的局部哈希函数，因此，每个模态中所有点的整体转换是局部线性的，但总体上是非线性的。

为了学习非线性哈希函数，Masci 等[49]介绍了一种基于耦合暹罗神经网络架构的，用于多模式相似性保留哈希的新型学习框架。它利用相似对和不相似对来进行模态内和模态间相似性学习。与大多数现有的跨模式相似性学习方法不同，哈希函数不限于线性投影。通过增加网络中的层数，可以训练任意复杂度的映射。Cao 等[50]提出了相关哈希网络（CHN），一种用于交叉模态哈希的新型混合架构。他们共同学习了适合哈希编码的良好图像和文本表示形式，并正式控制了量化误差。Zhang 等[51]提出了一种多模式散列方法，称为语义相关最大化（SCM），它将语义标签集成到散列学习过程中。该方法使用标签向量获得语义相似度矩阵，并尝试通过学习的哈希码对其进行重构。Wu 等[52]基于字典学习框架。开发一种通过联合多模式字典学习来获取跨不同模态的数据对象的稀疏代码集的方法，该方法称为稀疏多模式哈希（缩写为 SM2H）。在 SM2H 中，首先通过超图对模态内相似度和模态间相似度进行建模，然后通过超图拉普拉斯稀疏编码共同学习多模态字典。基于学习的词典，获取每个数据对象的稀疏代码集，并使用敏感的 Jaccard 度量进行多模态近似最近邻检索。同样，Yu 等[53]提出了一种判别耦合字典哈希（DCDH）方法来捕获多模态数据的底层语义信息。在 DCDH 中，借助类信息学习每种形式的耦合字典。结果，耦合的字典不仅保留了多模态数据之间的内部相似性和相互关联性，而且还包含语义

上具有区别性的字典原子（即，来自相同类别的数据由相似的字典原子重建）。为了执行快速的跨媒体检索，需要学习哈希函数以将数据从字典空间映射到低维汉明空间。

为了捕获更复杂的数据结构，Lin 等[54]提出了一种称为 SePH（语义保留哈希）的两步监督哈希方法，用于跨视图检索。对于训练，SePH 首先将训练数据的给定语义相似度转换为概率分布，并通过最小化 KL 散度在汉明空间中使用将要学习的哈希码对其进行近似。然后，在每个视图中，SePH 都利用核逻辑回归和采样策略来学习从特征到哈希码的非线性投影。对于任何看不见的实例，使用一种新颖的概率方法，利用预测的哈希码及其来自观察到的视图的相应输出概率来确定其统一的哈希码。Cao 等[55]提出了一种新的监督交叉模态哈希方法，即相关自动编码器哈希（CAH），以学习基于深度自动编码器的判别式和紧凑型二进制代码。具体来说，CAH 联合最大化双峰数据揭示的特征相关性和相似性标签中传达的语义相关性，同时通过非线性深度自动编码器将它们嵌入到哈希码中。

北京大学多媒体信息处理研究室彭宇新教授课题组采集并发布了 XMedia 数据集 ，并在半监督跨模态检索等方面做了深入研究；北京交通大学张磊[57]博士和北京邮电大学花妍[58]博士等分别在语义一致的跨模态关联学习方面做了深人研究；浙江大学金仲明[59]博士和北京邮电大学冯方向[60]博士分别在基于深度学习的跨模态检索研究中取得了很好的成果；西安光电精密机械所的李学龙老师课题组在跨模态哈希算法方面做出了突出的贡献。

1.3 本文的主要研究内容和组织结构

1.3.1 主要研究内容

本文主要研究了基于目标检测和图卷积的跨模态检索算法研究。主要包括 3 个方面的研究。

1、多层次图卷积特征信息的研究

对于多层次图卷积，如何开发不同层次图卷积的效果，以充分利用不同图卷积程度的特征来提升检索效果。通过对不同卷积层次的特征进行提取和学习研究，可以提高整体的特征信息丰富度，可以使算法架构灵活地使用不同卷积程度的特征信息，进而提升整体的检索效果。

2、多粒度文本特征学习的研究

以往的跨模态检索的文本特征学习大多是简单的循环神经网络 GRU 等。这

样的文本特征学习方式太过简单，不够充分，文本特征信息无法得到充分的学习。为此我们研究如何能充分挖掘文本特征信息，尤其是学习到不同粒度的文本特征信息。通过多种粒度的文本特征学习来获得丰富的文本信息，同时不同粒度的特征信息之间可以相互补充。因此文本特征学习能力得到了提升，进而提升了检索效果。

3、混合检索的研究

以往的检索算法都是只包含一个检索模型。受混合推荐的启发，我们在一个算法中包含了两个检索模型，这两个检索模型分别单独训练，每个检索模型都有自己的检索能力，检索时对两者的图像特征和文本特征分别取平均来实现两者的整合。通过对两个检索模型进行整合来实现两个检索模型检索能力的相互增强，进而提升整体的检索效果。

1.3.2 本文组织结构

本文章节的组织结构如下：

第一章：绪论部分，主要介绍了本文的课题研究背景和意义，目前过内外的主要研究现状和本文的主要研究内容。

第二章：介绍在基于目标检索和图卷积的跨模态检索中对多层图卷积不同层的研究。重点研究了如何控制对中间层的学习能力的开发。

第三章：主要介绍了跨模态检索任务中对文本进行多粒度特征学习的研究。重点研究了如何通过多粒度文本特征学习来提升文本部分的学习能力。

第四章：主要介绍了跨模态检索任务中混合检索的研究。主要介绍了通过混合检索的方式将多个检索模型整合在一起，实现整体检索能力的增强。

第 2 章 多层图卷积的层次研究

2.1 引言

在跨模态检索中，目标检测和图卷积在图像上的应用是一个重要方向。图像信息十分丰富，不可能学习到图像上所有的语义信息，那样反而会降低效果。目标检测可以将一个图片中最核的信息找到，这样就可以排除不重要信息的干扰，提升信息的价值性。一张图像中各个对象之间往往存在一定的关联，这种关联信息的学习对于跨模态检索十分重要，图卷积恰恰可以学习到各个对象之间的关联，为了充分学习到关联信息，图卷积往往是多层叠加使用的。目标检测和图卷积的联合使用是一个重要的架构设计。本章主要对多层图卷积的中间层学习能力的开发进行了研究，探究如何灵活掌控使用各层图卷积的学习能力来提升检索效果，并对涉及到的技术理论和经典方法进行了介绍。

2.2 相关技术理论

2.2.1 卷积图谱

图卷积的核心思想是将一个个的特征信息看作是图中的节点，用一定的方式计算出特征之间的关联看作是节点之间的边，再根据图中的边，对节点特征信息进行融合推理。

图卷积属于谱卷积的范畴，图上的谱卷积可以定义为信号和滤波器在傅立叶域的乘机如公式（2-1）所示。

$$g_\theta * x = U g_\theta U^T x \qquad (2\text{-}1)$$

其中 U 为归一化普拉斯 $L = I_N - D^{-\frac{1}{2}} A D^{-\frac{1}{2}} = U \Lambda U^T$ 的特征向量矩阵，Λ特是特征值矩阵（对角矩阵）$U^T x$ 是 x 的图傅立叶变换。

可以将 g_θ 理解为一个关于 L 特征值的函数，例如：$g_\theta(\Lambda)$。验证和计算公式（2-1）是一个非常费时，因为和特征向量矩阵 U 相乘的复杂度为 $O(N^2)$。此外，对于大型图，首先计算 L 的特征分解可能非常耗时。为了缓解这个问题（之所以称为缓解，是因为并没有得到彻底解决），我们可以使用近似表达式，式（2-2）中的多项式 $T_k(x)$ 直到 K 阶：

$$g_{\theta'}(\Lambda) = \sum_{k=0}^{K} \theta_k' T_k(\tilde{\Lambda}) \qquad (2\text{-}2)$$

其中 $\tilde{\Lambda}=\frac{2}{\lambda_{\max}}\Lambda-I_N$ 其中的$\lambda_{\max}$指的是 L 的最大特征值，$\theta'\in R^K$是一个切比雪夫系数向量，切比雪夫多项式递归定义如式（2-3）所示。

$$T_k(x)=2xT_{k-1}(x)-T_{k-2}(x) \qquad (2\text{-}3)$$

同时满足初始条件$T_0(x)=1$和$T_1(x)=x$。回到先前的定义，我们可以更新公式为：

$$g_{\theta'}*x\approx\sum_{k=0}^{K}\theta_k'T_k(\tilde{L})x \qquad (2\text{-}4)$$

其中的$\tilde{L}=\frac{2}{\lambda_{\max}}L-I_N$，并且可以得证$(U\Lambda U^T)^k=U\Lambda^kU^T$，需要注意的是表达式是$k$局部的，因为它是 K 阶拉普拉斯多项式，及它只依赖于中心节点（K 阶领域）最大 K 步的节点。计算公式（2-4）所用到的时间复杂度为$O(|\in|)$，和边的数量是线性关系。

2.2.2 Layer-Wise 线性模型

基于图卷积的神经网络模型可由多层卷积层叠加而成，每层跟随一个逐点非线性（point-wise non-linearity）。现在，假设我们将逐层卷积运算限制为$K=1$（参见公式 2-4），即一个线性的*w.r.t* L 的函数，因此是图拉普拉斯谱上的一个线性函数。

通过这种方法，我们可以保持一种丰富的卷积滤波函数，但是并不局限于给出的显式参数化，例如切比雪夫多项式等。我们期待这样一个模型能够解决具有非常宽节点度分布的图（例如社交网络、引用网络、知识图和许多其他真实世界的图数据集）的局部领域结构过拟合问题。此外，这种分层线性模型允许我们构建更深入的模型，这是一种可以提高多个领域建模能力的实践。
在这种情况下，我们将 GCN 网络进一步近似$\lambda_{\max}\approx 2$，神经网络参数将会在大规模的训练之后达到这样预期的估计。因此，公式（2-4）可以简化为：

$$g_{\theta'}*x\approx\theta_0'x+\theta_1'(L-I_N)x=\theta_0'x-\theta_1'D^{-\frac{1}{2}}AD^{-\frac{1}{2}}x \qquad (2\text{-}5)$$

这里有两个自定义的参数θ_0'和θ_1'。这个滤波器可以在整个图中共享，这种形式的连续滤波操作可以对一个节点的k阶邻（k-th）域进行卷积操作，其中k指是神经网络模型中连续滤波操作或者卷积层的数量。在实践中，得益于进一步限制了参数的数量来处理过拟合并且最小化每层操作的数量（例如使用矩阵乘法），得到了表达式（2-6）。

$$g_{\theta'} * x \approx \theta(I_N + D^{-\frac{1}{2}} A D^{-\frac{1}{2}})x \qquad (2\text{-}6)$$

这里将上述的双参数合并为一个单参数$\theta = \theta_0^{'} = -\theta_1^{'}$，需要注意的是$I_N + D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$现在的特征值在$[0,2]$范围内。因此，在深度神经网络模型中重复使用该算子会导致该数值不稳定和梯度消失，为了解决这一问题，我们引入了再归一化技巧如（2-7）所示。

$$I_N + D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \to \widetilde{D}^{-\frac{1}{2}} \widetilde{A} \widetilde{D}^{-\frac{1}{2}} \qquad (2\text{-}7)$$

其中参数满足$\widetilde{A} = A + I_N$，$\widetilde{D}_{ii} = \sum_j \widetilde{A}_{ij}$。我们可以将这个定义泛化到信号$X \in R^{C \times F}$这里的$C$输入通道（每个节点有$C$维的特征向量）和$F$维滤波器的特征映射如下：

$$Z = \widetilde{D}^{-\frac{1}{2}} \widetilde{A} \widetilde{D}^{-\frac{1}{2}} X\theta \qquad (2\text{-}8)$$

这里的$\theta \in R^{C \times F}$是滤波器参数矩阵，$Z \in R^{N \times F}$是信号卷积矩阵。这个滤波计算的复杂度是$O(|\varepsilon|FC)$，这样可以有效实现$\widetilde{A}X$稠密矩阵和稀疏矩阵的乘积操作。

2.2.3 目标检测 Faster-RCNN

Faster-RCNN 如图（2-1）所示，可以看作主要由 Conv layers，Region Proposal Networks，RoI Pooling，Classification 这 4 部分构成。

首先将图像变成统一的 M×N 大小，之后经过 13 个 conv 层、13 个 relu 层，4 个 pooling 层即 Conv layer 层，所有的 conv 层都是：kernel_size=3，pad=1，stride=1；所有的 pooling 层都是：kernel_size=2，pad=1，stride=1。因此最终得到(M/16)×(N/16)的 feature map。之后将 feature map 经过 3×3 卷积后送入 RPN 网络。

RPN 网络实际分为 2 条线，上面一条通过 1×1 的卷积，相当于将各通道的特征进行融合，输出的是 18 通道，这是因为 feature map 上的每一个特征点都被设置 9 个 Anchor 检测框，Anchor 有(1:2，1:1，2:1)三种形状，softmax 对 anchors 进行 positive 和 negative 分类，故有 9×2 通道。先进行 reshape 是为了便于 softmax 计算，之后再 reshape 回来。下面一条用于计算对于 anchors 的 bounding box regression 偏移量，以获得精确的 proposal。将每个 anchor 内的 feature 作为输

入，经过线性回归得到 anchor 和 gound truth 之间的偏移量（平移和缩放量）。而最后的 proposal 层则负责综合 positive anchors 和对应 bounding box regression 偏移量获取 proposals，同时剔除太小和超出边界的 proposals。其实整个网络到了 proposal Layer 这里，就完成了相当于目标定位的功能。

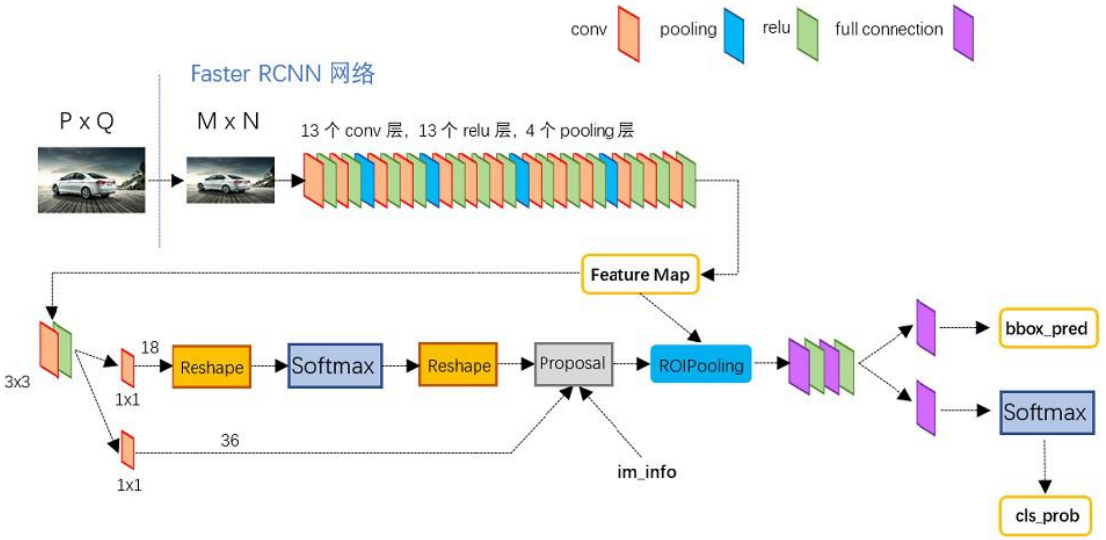


图 2-1 Faster 流程图
Fig.2-1 flow chart of Faster-RCNN

RoI Pooling 层负责收集 proposal，并计算出 proposal feature maps，送入后续网络。RoI pooling 层有 2 个输入：原始的 feature maps 和 RPN 输出的 proposal boxes（大小各不相同），由于 proposal 是对应 M*N 尺度的，所以首先使用将其映射回(M/16)*(N/16)大小的 feature map 尺度；再将每个 proposal 对应的 feature map 区域水平分为固定的网格；对网格的每一份都进行 max pooling 处理。这样处理后，即使大小不同的 proposal 输出结果都是固定大小，实现了固定长度输出，将输出的 feature map 送入下面的网络。

Classification 部分利用已经获得的 proposal feature maps，通过 full connect 层与 softmax 计算每个 proposal 具体属于哪个类别，输出每个类别的概率向量，同时再次利用 bounding box regression 获得每个 proposal 的位置偏移量，用于回归更加精确的目标检测框。至此 fasterRNN 获取了图像中物体的类别信息和位置信息。

2.2.4 编码解码架构

Encoder-Decoder 框架是深度学习中非常常见的一个模型框架，例如在

Image Caption 的应用中 Encoder-Decoder 就是 RNN-RNN 的编码-解码框架；在神经网络机器翻译中 Encoder-Decoder 往往就是 LSTM-LSTM 的编码-解码框架，在机器翻译中也被叫做序列到序列的学习。框架结构如图（2-2）所示.

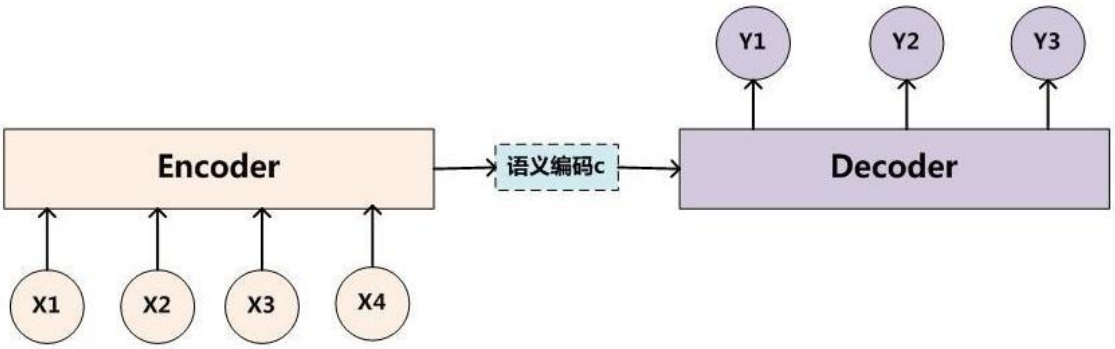


图 2-2 encoder-decoder 流程图
Fig.2-2 flow chart of encoder-decoder

该框架首先将输入看作一个序列，将整个序列作为输入送入循环网络得到语义编码 C，这个语义编码包含了其中的语义信息和各个特征之间的顺序信息。之后语义编码 C 再经过循环网络解码得到输出的特征序列。Encoder 部分本质上是将输入的所有特征信息和顺序信息编码成了一个特征向量，这里是为了脱去信息形式上的特征，而得到深层次的语义信息，使语义剥离了输入特征的形式。Decoder 部分本质上根据这个语义信息 C，在其之上加入另一种信息的形式特征。将语义信息以另一种形式展现出来。Image-Caption 就是先将图像的语义信息提取得到语义编码 C，脱去了图像的特征形式，再根据语义编码 C 加上文本的特征形式得到生成的文本。

2.2.5 Word Embedding

Embedding 是数学领域的有名词，是指某个对象 X 被嵌入到另外一个对象 Y 中，映射 f:X→Y，例如有理数嵌入实数。Word Embedding 是 NLP 中一组语言模型和特征学习技术的总称，把词汇表中的单词或者短语映射成由实数构成的向量上(映射)。

One-hot 是最简单的 Word Embedding，是指将所有词排成一列，对于词 A，只有在它的位置置 1，其他位置置 0，维度就是所有词的数目。One-hot 方法很简单，但是它的问题也很明显：（1）它没有考虑单词之间相对位置的关系。（2）词向量可能非常非常长。

共现矩阵 (Cocurrence matrix)一个非常重要的思想是，我们认为某个词的意思跟它临近的单词是紧密相关的。这是我们可以设定一个窗口（大小一般是5～10），那么在这个窗口内，一个单词与共同出现的其它单词之间就存在共现关系。然后我们就利用这种共现关系来生成词向量。虽然 Cocurrence matrix 一定程度上解决了单词间相对位置也应予以重视这个问题。但是它仍然面对维度灾难。也即是说一个 word 的向量表示长度太长了。

Dristributed representation：低维实数向量，它的思路是通过训练，将每个词都映射到一个较短的词向量上来，可以解决 One hot representation 的问题，。在 word2vec 出现之前，已经有用神经网络 DNN 来用训练词向量进而处理词与词之间的关系了。采用的方法一般是一个三层的神经网络结构（当然也可以多层），分为输入层，隐藏层和输出层(softmax 层)相关或者相似的词，在距离上更接近。自动实现：（1）单词语义相似性的度量；（2）词汇的语义的类比。

Word2Vec，现在最常用、最流行的方法就是 Word2Vec。这是 Tomas Mikolov 在谷歌工作时发明的一类方法，也是由谷歌开源的一个工具包的名称。具体来说，Word2Vec 中涉及到了两种算法，一个是 CBOW 一个是 Skip-Gram。这也是因为深度学习流行起来之后，基于神经网络来完成的 Word Embedding 方法。

Skip-Gram：给定 input word 来预测上下文。首先选句子中间的一个词作为我们的输入词。之后定义 skip_window 的参数，代表从当前 input word 的一侧（左边或右边）选取词的数量。如果设置 skip_window=2，那么 skip_window=2 代表着选取左 input word 左侧 2 个词和右侧 2 个词进入我们的窗口，所以整个窗口大小 span=2*2=4。参数 num_skips，代表从整个窗口中选取多少个不同的词作为 output word，当 skip_window=2，num_skips=2 时，会得到两组 (input word, output word) 形式的训练数据。训练时，剔除高频的停用词来减少模型的噪音，并加速训练，它们对其他词的贡献不大。input word 和 output word 都是 one-hot 编码的向量。最终模型的输出是一个概率分布。

输入的每一个词都是一个 one-hot 表示形式，通常词汇表比较大，如果词汇表有 10000 个词，模型的输入就是一个 10000 维的向量，那么输出也是一个 10000 维度（词汇表的大小）的向量，它包含了 10000 个概率，每一个概率代表着当前词是输入样本中 output word 的概率大小。input word 和 output word 都会被我们进行 one-hot 编码。输入被 one-hot 编码后大多数维度上都是 0（实际上仅有一个位置为 1），所以向量稀疏，如果我们将一个 1*10000 的向量和 10000*300 的矩阵相乘，它会消耗相当大的计算资源，为了高效计算，隐层权

重矩阵看成了一个“查找表”，进行矩阵计算时，直接去查输入向量中取值为1的维度下对应的那些权重值。隐层的输出就是每个输入单词的“嵌入词向量”。经过神经网络隐层的计算，词从一个1*10000的向量变成1*300的向量，再被输入到输出层。输出层是一个softmax回归分类器，它的每个结点将会输出一个0-1之间的值（概率），这些所有输出层神经元结点的概率之和为1。

2.3 算法研究

2.3.1 VSRN 架构

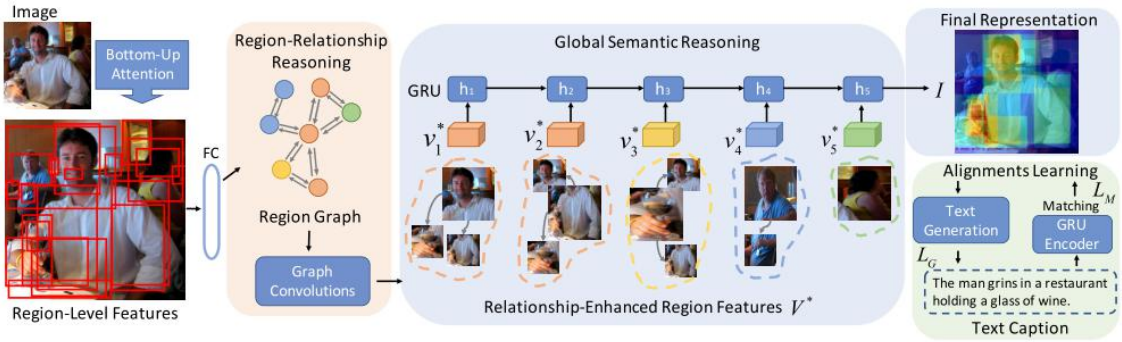


图 2-3 VSRN 流程图
Fig.2-3 Flow chart of VSRN

VSRN[63]是基于目标检测和图卷积的多任务跨模态信息检索的算法。该算法主体结构如图所示。整个算法结构可以分为自下而上的注意力机制得到图像特征表示，区域关系推理，全局关系推理，通过联合匹配和图像描述生成进行学习四个部分。之后用图卷积（GCN）对图像中的各个对象特征进行融合推理学习。

首先是自下而上注意力得到的图像表示，利用自下而上的注意的优势，每个图像可以由一组特征$V=\{v_1,\ldots,v_k\}$，$v_i=R^D$表示，从而每个特征v_i编码一个对象或一个此图像中的区域。作者以ResNet-101为骨干，通过Faster R-CNN 模型实现了自下而上的注意。Peter Anderson 等[61]在Visual Genomes数据集上对其进行了预训练。该模型经过训练后可以预测实例类和属性类，而不是对象类，从而可以帮助学习具有丰富语义含义的特征表示形式。具体地说，实例类包括对象和难以识别的重要内容。例如，“毛茸茸”等属性和“建筑物”，“草”和“天空”之类的属性。使用模型的最终输出，并且每个类别的非最大抑制都在IoU阈值为0.7的情况下进行。然后，作者将置信度阈值设置为0.3，并选择所有类别检测概率大于此阈值的所有图像区域。选择具有最高类别检测置信度

得分的前 36 个 ROI。所有这些阈值均设置为与[61，62]相同。对于每个选定的区域 i，在平均池化层之后提取特征，从而得出 2048 维的 f_i。然后，使用公式（2-9）应用全连接将 f_i 转换为 D 维嵌入。

$$v_i = W_f f_i + b_f \qquad (2\text{-}9)$$

然后，构造 $V = \{v_1, \ldots, v_k\}$，$v_i = R^D$ 表示每个图像，其中 v_i 编码该图像中的对象或显着区域。

受基于深度学习的视觉推理的启发，作者建立了区域关系推理模型，通过考虑图像区域之间的语义相关性来增强基于区域的表示。 具体来说，测量嵌入空间中图像区域之间的成对亲和力，以使用公式（2-10）构建它们之间的关系。

$$R(v_i, v_j) = \varphi(v_i)^T \phi(v_i) \qquad (2\text{-}10)$$

$\varphi(v_i) = W_\varphi v_i$ 和 $\phi(v_j) = W_\phi v_j$ 是两个嵌入。权重参数 W_ϕ 和 W_φ 可以通过反向传播来学习。然后，一个全连接关系图 $G_r = (V, E)$，其中 V 是检测到的区域的集合，而边缘集 E 由亲和力矩阵 R 描述。 R 是通过使用公式（2-10）计算每对区域的亲和力边缘而获得的，这意味着如果两个图像区域具有很强的语义关系并且高度相关，则将存在一个具有高亲和力得分的边缘来连接两个图像区域。作者应用图卷积网络（GCN）在此全连接图上进行推理。每个节点的响应都基于图关系定义的邻居进行计算。我们将残余连接添加到原始 GCN，如公式（2-11）所示。

$$V^* = W_r(RVW_g) + V \qquad (2\text{-}11)$$

式中　W_g ——尺寸为 D×D 的 GCN 层权重矩阵

W_r ——残余结构的权重矩阵

R ——形状为 k×k 的亲和矩阵

按照例程对亲和矩阵 R 进行逐行归一化。输出 $V^* = \{v_1^*, \ldots, v_k^*\}$，$v_i^* \in R^D$ 是图像区域节点的关系增强表示。

基于具有关系信息的区域特征，作者进一步进行全局语义推理，以选择判别信息，并过滤掉不重要的信息，以获得整个图像的最终表示。具体来说，作者通过将区域特征的序列 $V^* = \{v_1^*, \ldots, v_k^*\}$，$v_i^* \in R^D$ 逐一放入 GRU 中来进行这种推理。 在此推理过程中，整个场景的描述将逐渐在存储单元 m_i（隐藏状态）中增长和更新。

在每个推理步骤 i 中，更新门 z_i 分析当前输入区域特征 v_i^* 和最后一步 m_{i-1} 的整个场景描述，以决定该单元对其存储单元进行了多少更新。更新门的计算方

法如公式（2-12）所示。

$$z_i = \sigma_z(W_z v_i^* + U_z m_{i-1} + b_z) \qquad (2\text{-}12)$$

式中 W_z——权重参数

U_z——权重参数

σ_z——sigmoid 激活函数

b_z——偏差

新增内容有助于对整个场景进行描述，其计算方式如公式（2-13）所示。

$$\tilde{m}_i = \sigma_m(W_m v_i^* + U_z(r_i \circ m_{i-1}) + b_m) \qquad (2\text{-}13)$$

式中 σ_z—— tanh 激活函数

$\circ$——逐元素乘法

r_i——重置门

r_i 根据 v_i^* 和 m_{i-1} 的推理来决定要忘记的内容。与更新门类似地计算出 r_i：

$$r_i = \sigma_r(W_r v_i^* + U_r m_{i-1} + b_r) \qquad (2\text{-}14)$$

然后，在当前步骤对整个场景 m_i 的描述是使用先前描述 m_{i-1} 和新内容 $\tilde{m}_i$ 之间的更新门 z_i 的线性插值，如公式（2-15）所示。

$$m_i = (1 - z_i) \circ m_{i-1} + z_i \circ \tilde{m}_i \qquad (2\text{-}15)$$

其中$\circ$是逐元素乘法。由于每个 v_i^* 都包含全局关系信息，因此 m_i 的更新实际上是基于图拓扑的推理，该图拓扑同时考虑了当前局部区域和全局语义相关性。作者将序列 V^* 末尾的存储单元 m_k 作为整个图像的最终表示 I，其中 k 是 V^* 的长度。

为了连接视觉和语言领域，作者使用基于 GRU 的文本编码器将文本标题映射到与图像表示 I 相同的 D 维语义向量空间 $C \in R^D$，后者考虑了句子中的语义上下文。然后，作者共同优化匹配和生成，以了解 C 和 I 之间的对齐方式。

对于匹配部分，作者采用基于铰链的三元组排名损失，重点是使用最有挑战性的错误项来计算损失，即最接近每个训练查询的错误项。我们的损失定义如公式（2-16）所示。

$$L_M = \left[\alpha - S(I,C) + S(I,\hat{C})\right]_+ + \left[\alpha - S(I,C) + S(\hat{I},C)\right]_+ \qquad (2\text{-}16)$$

其中 α 用作余量参数。 $[x]+ \equiv \max(x,0)$。此铰链损失包括两项，一项为 I，一项为 C。$S(\cdot)$ 是联合嵌入空间中的相似函数。作者在实验中使用通常的内积作为相似函数的计算方式。 $\hat{I} = \arg\max_{j \neq I} S(j,C)$ 和 $\hat{C} = \arg\max_{d \neq C} S(I,d)$ 是正对

与(I,C)最接近的错误项。为了提高计算效率，作者在每个小批量生产中都找到了它们，而不是在整个训练集中找到最难分辨的错误项。

对于生成部分，学习到的视觉表示还应该具有生成接近真实字幕的句子的能力。具体来说，我们使用具有注意机制的序列对模型进行排序以实现此目的。我们最大化预测输出句子的对数似然性。损失函数定义为：

$$L_G = -\sum_{t=1}^{l} \log p(y_t \mid y_{t-1}, V^*; \theta) \qquad (2\text{-}17)$$

其中l是输出单词序列的长度$Y=(y_1,\ldots,y_l)$。θ是序列到序列模型的参数。最终损失函数定义如下，以实现两个目标的联合优化。

$$L = L_M + L_G \qquad (2\text{-}18)$$

VSRN 将目标检索和图卷积相结合，先提取图像中的核心关键信息，再对图像中的信息进行融合学习，既包含各个特征信息也包含特征信息之间的关系。这样的结构设计使得图像部分的特征提取和特征学习达到了极好的效果。实验结果证明作者的算法取得了当时的最佳效果。

2.3.2 MLVSRN 架构

MLVSRN 是在 VSRN 基础进行的改进，主要是加入了多层 loss 约束，图卷积（GCN）在使用过程中往往是连续多层使用的，为的是充分进行图卷积学习。但是以往的多层图卷积只考虑了最后一层的输出，用最后一层的输出作为结果。没有对层与层之间的关系进行考虑。多层图卷积每一层都是在前一层的基础上进行的进一步学习，所以每一层学习的特征信息的粒度和侧重都会有所不同，MLVSRN 就是协调了不同层之间特征学习的相互关系。让不同层侧重学习不同的特征信息，各自侧重不同方面，来最终达到整体的效果提升。算法的流程如图（2-4）所示。

整体的框架大体与 VSRN 相似，主要做的改进是在将第 3 层 GCN 输出的特征向量经过 GRU 融合后与文本特征进行 triple loss 学习训练。融合第 3 层用的 GRU 与第 4 层一致，实验证明这样的效果更好，及尽量保持后续的网络结构一致，用的是同一个网络结构，只是在第 3 层多出了一个 triple loss 的学习训练。这样的目的是将 loss 函数提前，如此便可以加大对前 3 层图卷积的学习开发，让前 3 层充分学习特征之间的关系，并进行特征信息的融合，第 4 层就可以专注于在前 3 层的基础上学习更加精微细致的图像关系特征。因为是 2 个 loss 函数，因此在实际训练中前 3 层的参数调整是第 4 层的 2 倍，因为前 3 层已经充分学习了特征信息，因此第 4 层可以再进一步学习更加复杂精微的信息。

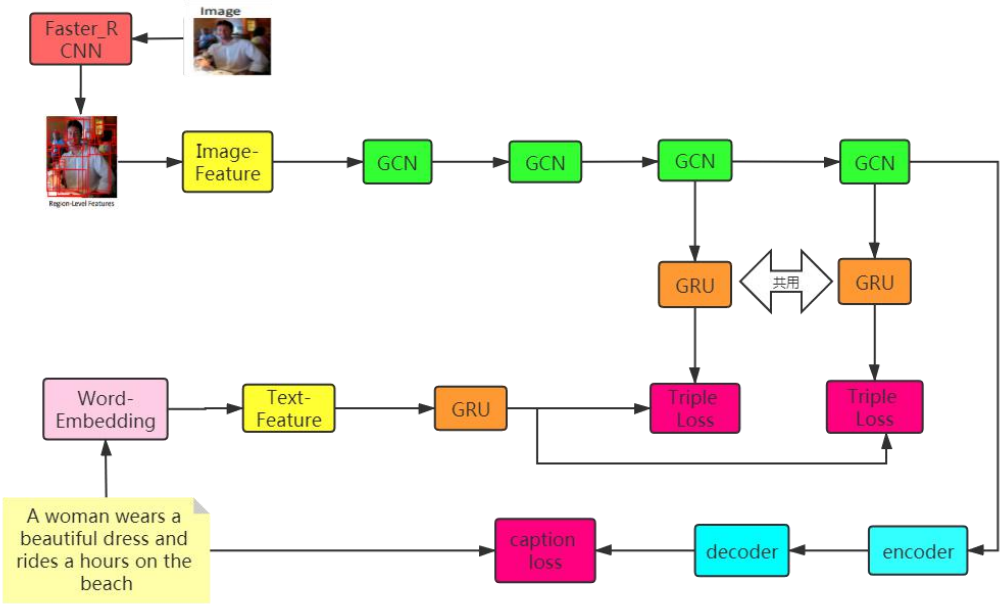


图 2-4 MLVSRN 流程图

Fig.2-4 flow chart of MLVSRN

具体来说，我们通过将区域特征的序列 $V^3=\{v_1^3,\ldots,v_k^3\}$， $v_i^3\in R^D$ 逐一放入 GRU 中来进行信息融合。$V^3=\{v_1^3,\ldots,v_k^3\}$ 表示第 3 层图卷积后的各个对象的特征集合。同时也将 $V^3=\{v_1^4,\ldots,v_k^4\}$， $v_i^4\in R^D$ 经过同一 GRU 进行特征信息的推理和融合。

GRU 融合特征时先计算更新门 z_i， 更新门的计算方法是：

$$z_i=\sigma_z(W_z v_i^3+U_z m_{i-1}+b_z) \qquad (2\text{-}12)$$

之后计算新增内容其计算方式如下：

$$\tilde{m}_i=\sigma_m(W_m v_i^3+U_z(r_i\circ m_{i-1})+b_m) \qquad (2\text{-}13)$$

r_i 根据 v_i^3 和 m_{i-1} 的推理来决定要忘记的内容。 与更新门类似地计算出 r_i：

$$r_i=\sigma_r(W_r v_i^3+U_r m_{i-1}+b_r) \qquad (2\text{-}14)$$

然后，在当前步骤对整个场景 m_i 的描述是使用先前描述 m_{i-1} 和新内容 $\tilde{m}_i$ 之间的更新门 z_i 的线性插值：

$$m_i=(1-z_i)\circ m_{i-1}+z_i\circ\tilde{m}_i \qquad (2\text{-}15)$$

将数据集序列 V^* 末尾的最后一个细胞记忆特征 m_k，当作是第 3 层图卷积输出结果的特征信息进行融合后的整体特征表示 f^3。第 4 层同样的过程，只是输入的 $V=\{v_1^4,\ldots,v_i^4\}$ 是第 4 层图卷积的特征输出。之后将 GRU 融合的第 3 层图卷

积的特征 f^3 和 GRU 融合的第 4 层图卷积的特征 f^4 级联在一起如公式（2-16）所示。

$$f = \left[f^3, f^4\right] \tag{2-16}$$

之后经过全连接层来将特征维度转换为嵌入空间的维度，同时对特征信息进行整合。具体如公式（2-17）所示。

$$I = \sigma(W_f f) \tag{2-17}$$

式中 σ ——激活函数

W_f ——权重参数

f ——图像特征

以上这便是在第 3 层加入损失函数约束的完整过程。后续我们还研究了在第 2 层加入损失函数的效果。在后面的章节会展示出其实际效果。

2.4 实验

2.4.1 实验数据和评价标准

我们在 Microsoft COCO 数据集[64]和 Flickr30K 数据集[65]上进行算法的测试工作。MS-COCO 数据集共包含了 123,287 张图像，其中每张图像带有 5 个文本描述，文本描述的是图像中的主要信息内容。 我们对 MSCOCO 数据集进行拆分，其中包含 113,287 张用于训练的图像，5000 张用于验证的图像和 5000 张用于测试的图像。每个图像带有 5 个标题。 最终的实验结果是通过对 5 倍的 1K 测试图像的结果进行取平均或在完整的 5K 测试图像上进行测试而获得的。Flickr30K 包含了从 Flickr 网站收集的 31783 张图像，每张图像随附 5 条人类注释文本说明。我们使用标准的训练，验证和测试分割，分别包含 28,000 张图像，1000 张图像和 1000 张图像。

对于信息检索中常见的评估指标。本章通过在 K（R @ K）处进行召回率的计算来衡量跨模态检索性能，R@K 代表着在特征空间距离上距离查询最近的前 K 个目标模态特征包含与查询相匹配的的特征信息的概率。

2.4.2 实验具体情况

我们将单词嵌入大小设置为 300，将联合嵌入空间的维度设置为 2048。我们采用与[61，62]相同的设置来设置视觉自下而上的注意模型。 基于 GRU 的全局语义推理的区域顺序由自下而上的注意力探测器生成的类别检测置信度得

分的降序确定。我们使用 Adam 优化器训练了 30 个 epoch。 我们开始的 15 个 epoch 以学习率 0.0002 进行训练，然后将其余 15 个 epoch 的学习率降低到 0.00002。 我们在等式中设置边距α为 0.2。 我们使用的最小批量为 128。对于测试集的评估，我们通过选择在验证集上表现最佳的模型来解决过度拟合问题。根据验证集中的召回总和选择最佳模型。实验环境会对实验结果有很大影响，尤其是 python、pytorch 和其它库的版本对实验结果影响很大。

2.4.3 VSRN 实验结果及分析

首先我们在和 VSRN 相同的数据集（Flick30k 和 coco 数据集）和实验环境下，重新训练了作者的神经网络，代码是作者在 github 上公开的实验代码。首先我们在 flick30k 数据集上重新运行了作者的代码，得到的实验结果如表（2-1）所示。

表 2-1 VSRN 在 flick30k 数据集上的实验

Table2-1 Data of VSRN on Flick30k

Methods	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
$SMlstm_{CVPR'17}$	42.5	71.9	81.5	30.2	60.4	72.3
$VSE{+}{+}_{BMVC'18}$	52.9	79.1	87.2	39.6	69.6	79.5
$SCO_{CVPR'18}$	55.5	82.0	89.3	41.1	70.5	80.1
$SCAN_{ECCV'18}$	67.4	90.3	95.8	48.6	77.7	85.2
VSRN	71.3	90.6	96.0	54.7	81.8	88.2
ours	68.1	88.9	93.7	52.1	78.7	86.1

从跑出来的实验结果可以看出来，我们的实验结果在 flick30k 数据集上与作者论文中所展示的结果大约差距了不到 3 个百分点，实验所用代码、数据集、实验环境皆与作者一致，作者公开说明在后续的研究工作中，对原本在 flick30k 上实现的代码进行了改动，进而导致了公布的代码在 flick30k 数据集上未能达到作者论文中水平。

之后我们在 coco 数据集上重新训练了作者的神经网络，在 coco 数据集上各项指标与作者在论文中公布的实验结果大体一致，这说明我们的实验环境和数据集没有出现问题，作者在 coco 数据集上进行后续研究中改动了一些具体的

代码实现导致了代码在 Flick30k 数据集上效果的下降。在 coco 数据集上的实验结果如下表（2-2）所示。

表 2-2 VSRN 实验

Table2-2 Data of VSRN on coco (1k)

Methods	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
$SMlstm_{CVPR'17}$	53.2	83.1	91.5	40.7	75.8	87.4
$VSE++_{BMVC'18}$	64.6	89.1	95.7	52.0	83.1	92.0
$SCO_{CVPR'18}$	69.9	92.9	97.5	56.7	87.5	94.8
$SCAN_{ECCV'18}$	72.7	94.8	98.4	58.8	88.4	94.8
VSRN	76.2	94.8	98.2	62.8	89.7	95.1
ours	73.0	94.1	97.8	60.3	88.4	94.2

从表（2-3）中可以看到在 5k 的测试集上的效果要远远低于在 1k（将 5k 的测试数据分成 5 个 1k 的测试数据，分别测试后取结果的平均值）上的效果。这可能是因为将 5k 数据同时用来做测试时，由于数据量变大，导致了干扰项变多，因此导致了结果的下降。在以往的众多跨模态检索算法中，均是 5k 的结果远低于 1k 结果。

表 2-3 VSRN 实验

Table2-3 Data of VSRN on coco (5k)

Methods	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
$VSE++_{BMVC'18}$	41.3	69.2	81.2	30.3	59.1	72.4
$SCO_{CVPR'18}$	42.8	72.3	83.0	33.1	62.9	75.5
$SCAN_{ECCV'18}$	50.4	82.2	90.0	38.6	69.3	80.4
VSRN	53.0	81.1	89.4	40.5	70.6	81.1
ours	48.9	78.0	87.4	37.2	68.0	79.2

从 Flick30k 数据集和 coco 数据集上图像检索文本的效果要明显好于文本检索图像的效果。这有两方面的原因，一方面从算法来看算法中对图像特征的学

习十分充分，图像部分的网络架构也十分复杂，而文本部分却只是简单的 GRU 循环神经网络，这样的算法结构设计就导致了文本和图形特征学习的不对称性，出现图像检索文本的效果好于文本检索图像的效果也就很正常。在基于深度学习的跨模态检索领域，对文本特征学习的研究工作没有图像特征学习的研究工作那么多。其次是信息的不对称性，图像信息的丰富度要远远高于文本，一张图像的信息内容远远多于一段文本描述，文本所描述的只是其中的某一部分，无法面面俱到，精致入微。因此这种信息上的巨大差，也导致了图像检索文本和文本检索图像的难度不同，信息丰富度高的检索低更容易，信息丰富度低的检索高的难度会更大。因此我们会看到图像检索文本效果好于文本检索图像的这一普遍现象。

2.4.4 MLVSRN 实验结果及分析

首先我们在 Flick30k 上进行了实验，将 MLVSRN 算法模型得到的实验结果和作者的 VSRN 算法模型的实验结果进行了对比，实验数据如表（2-4）所示。

表 2-4 MLVSRN 在 Flick30k 上实验

Table 2-4 Data of MLVSRN on Flick30k

Methods	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
$SMlstm_{CVPR'17}$	42.5	71.9	81.5	30.2	60.4	72.3
$VSE{+}{+}_{BMVC'18}$	52.9	79.1	87.2	39.6	69.6	79.5
$SCO_{CVPR'18}$	55.5	82.0	89.3	41.1	70.5	80.1
$SCAN_{ECCV'18}$	67.4	90.3	95.8	48.6	77.7	85.2
VSRN(ours)	68.1	88.9	93.7	52.1	78.7	86.1
MLVSRN(ours)	69.6	89.8	94.2	51.7	78.4	86.0

通过表（2-4）我们可以看出来改进后的算法 MLVSRN 在图像检索文本方面各个指标上相比 VSRN 均有所提升。这是因为第 3 层的的特征向量被用于 triple loss 训练，这样导致的前 3 层被增加了一层约束，前 3 层被更加充分地学习，相对于没有加 triple loss，加上 triple loss 后第 3 层图卷积得到的特征向量更利于和文本特征匹配，匹配正确率更高，第 4 层就可以在第 3 层的基础上进行精修，学习精微的特征信息和特征之间的相互匹配，来使得结果进一步提升，

最终达到图像检索文本效果的提升。同时我们可以看到在文本检索图像上效果在各个指标上均有所下降。这是因为 VSRN 的算法结构中，图像特征学习部分十分复杂，充分。而文本特征学习部分时间简单。这就导致了整个算法的效果提升主要依赖于图像特征学习部分的网络结构和参数训练，文本部分起到的作用相对较少。但是 MLVSRN 是在图卷积的第 3 层加上了 triple loss 的训练学习，这样可以更加充分地开发多层图卷积的特征学习能力，提升图像部分特征学习的效果，这样自然可以提升图像检索文本的效果。但是这样更加加剧了图像特征学习与文本特征学习之间的不平衡，就会导致跨膜态检索的效果更加依赖于图像部分参数的训练，相当于图像特征在不断地学习特征关系和特征空间变换来和对应的文本特征信息在空间距离上拉近。而文本特征则相对变动较少。这样导致了图像特征具有很好的适应性来找到想对应的文本特征，因此图像检索文本的效果往往很好。而文本特征相对就显得消极被动，在文本检索图像时效果就相对低些。

之后在 coco 数据集上再次进行了实验，实验数据如表（2-5）和（2-6）所示。

表 2-5 MLVSRN 在 coco (1k)上实验

Table2-5 Data of MLVSRN on coco (1k)

Methods	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
$SMlstm_{CVPR'17}$	53.2	83.1	91.5	40.7	75.8	87.4
$VSE++_{BMVC'18}$	64.6	89.1	95.7	52.0	83.1	92.0
$SCO_{CVPR'18}$	69.9	92.9	97.5	56.7	87.5	94.8
$SCAN_{ECCV'18}$	72.7	94.8	98.4	58.8	88.4	94.8
VSRN(ours)	73.0	94.1	97.8	60.3	88.4	94.2
MLVSRN(ours)	73.2	94.5	98.0	61.1	88.8	94.4

从表（2-5）和（2-6）中的实验数据中可以看到，图像检索文本和文本检索图像任务中 MLVSRN 的效果都比 VSRN 高出一些。范围在 0.2%至 1.4%。而且是每一个指标都有所提升，这说明改进后的效果提升是稳定有效的。在其它图卷积层增加损失函数约束，确实起到了提升效果的作用。R@1 项的增加说明了本章算法对精微信息匹配的提升效果。R@1 是对精微信息特征要求最高的指

标，它要求信息匹配的精确性要求最高。这一指标提升说明了本章所提出的算法在学习精微信息特征和精微特征关系上的效果，增加前 3 层图卷积的约束，开发前 3 层图卷积的学习潜力，来使第 4 层图卷积可以更好地学习精微的特征信息和特征关系以提升整个算法匹配的精确度。

表 2-6 MLVSRN 在 coco (5k)上实验

Table2-3 Data of MLVSRN on coco (5k)

Methods	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
$VSE{+}{+}_{BMVC'18}$	41.3	69.2	81.2	30.3	59.1	72.4
$SCO_{CVPR'18}$	42.8	72.3	83.0	33.1	62.9	75.5
$SCAN_{ECCV'18}$	50.4	82.2	90.0	38.6	69.3	80.4
VSRN(ours)	48.9	78.0	87.4	37.2	68.0	79.2
MLVSRN(ours)	50.7	80.2	88.4	38.8	69.2	79.8

从实验结果提升的角度，coco5k 比 coco1k 效果提升的更多。这是因为 coco5k 是在 5000 张图像 25000 条文本的测试集上进行的测试，coco1k 是在 1000 张图像 5000 条文本的 5 个测试集上进行测试后取平均值得到的。由于 coco5k 比 coco1k 在测试时数据量要大，因此数据的干扰项要更多。在测试集中找到正确匹配对象的难度就会更大。因为本章算法在第 3 层加入损失函数，提高了图卷积前 3 层的特征学习能力，使第 4 层可以学习到更加精微的信息匹配信息，因此在增大了检索规模，增强了干扰程度的情况下，能够取得更好的效果，这是因为学习到了精微的信息匹配信息，对特征信息的分辨能力增强，能够更加容易地分辨出特征信息之间的差异，提升了算法的抗干扰能力，进而提升了跨模态信息检索的匹配效果。

2.5 本章小结

本章我们主要研究了针对多层图卷积，如何充分开发图卷积的学习能力。不再是仅仅使用最后一层输出，而是充分开发中间层的价值，如何利用中间层来提升图卷积效果。本章我们提出了跳跃式结构，将中间层图卷积的输出直接与后面神经网络连接，再进行检索匹配。中间层与最后一层分别进行检索匹配，整个方法包含了两个损失函数，两个匹配。中间层和最后输出层共用图卷积之

后的神经网络，但分开独立进行检索匹配和损失函数计算。我们将这一方法和结构与 VSRN 进行融合提出了 MLVSRN。实验证明 MLVSRN 在各项指标上，相比 VSRN 皆有提升。这说明跳跃式的结构设计可以提升多层图卷积的学习能力，可以灵活对不同层的图卷积进行学习能力开发，提升检索效果。

第 3 章 多粒度文本特征研究

3.1 引言

在以往的跨模态检索研究过程中，人们对文本部分特征学习的研究普遍比较少，只是用简单的循环神经网络 GRU 来实现对文本特征信息的学习，这样的学习方式往往无法充分地学习到文本特征信息，无法满足跨模态检索的需要。有人通过多任务的方式，在跨模态检索任务中加入生成任务，如图像生成文本或文本生成图像来提高学习到的特征信息的质量，使学到的特征向量更利于跨模态检索。但这样依然没有真正提高文本部分的学习能力。为了解决这一问题，我们借鉴 Dong 等[56]的多粒度文本特征学习，使用多粒度文本特征学习方法来提高文本部分的特征学习能力，使文本部分可以学习到丰富的多粒度文本特征信息，解决文本特征学习过于简单，文本特征信息学习不充分的问题，进而改善检索效果。

3.2 文本特征学习

3.2.1 循环神经网络

RNN 的目的是用来处理序列数据。在传统的神经网络模型中，是从输入层到隐含层再到输出层，层与层之间是全连接的，每层之间的节点是无连接的。但是这种普通的神经网络对于很多问题却无能无力。例如，你要预测句子的下一个单词是什么，一般需要用到前面的单词，因为一个句子中前后单词并不是独立的。RNN 之所以称为循环神经网路，即一个序列当前的输出与前面的输出也有关。具体的表现形式为网络会对前面的信息进行记忆并应用于当前输出的计算中，即隐藏层之间的节点不再无连接而是有连接的，并且隐藏层的输入不仅包括输入层的输出还包括上一时刻隐藏层的输出。理论上，RNN 能够对任何长度的序列数据进行处理。但是在实践中，为了降低复杂性往往假设当前的状态只与前面的几个状态相关，

如图（3-1）左侧便是一个简单的循环神经网络如，它由输入层、一个隐藏层和一个输出层组成：如果把上面带箭头的圈去掉，它就变成了最普通的全连接神经网络。x 是一个向量，它表示输入层的值（这里面没有画出来表示神经元节点的圆圈）；s 是一个向量，它表示隐藏层的值（这里隐藏层面画了一个节点，你也可以想象这一层其实是多个节点，节点数与向量 s 的维度相同）；U 是输入层到隐藏层的权重矩阵；o 也是一个向量，它表示输出层的值；V 是

隐藏层到输出层的权重矩阵。那么，现在我们来看看 W 是什么。循环神经网络的隐藏层的值 s 不仅仅取决于当前这次的输入 x，还取决于上一次隐藏层的值 s。权重矩阵 W 就是隐藏层上一次的值作为这一次的输入的权重。

如果我们把图（3-1）左侧展开，循环神经网络也可以画成右侧的样子。

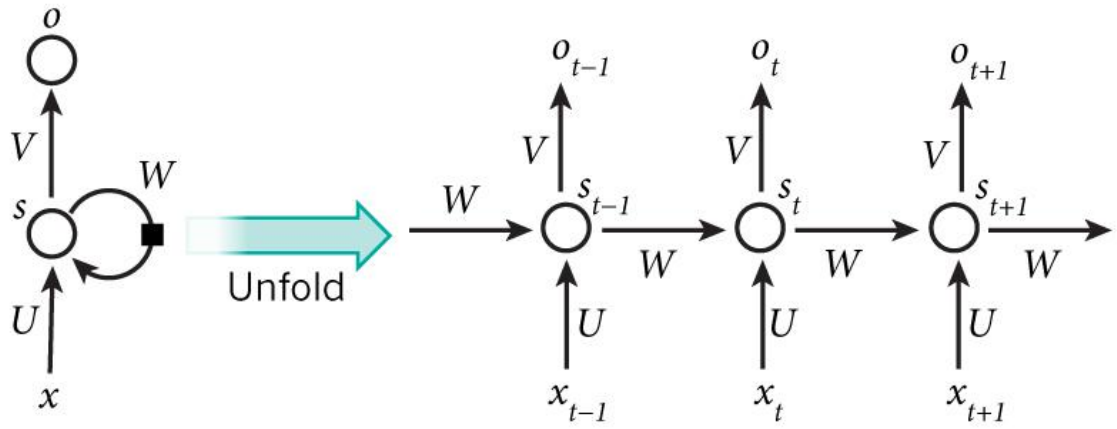


图 3-1 RNN 流程图

Fig.3-1 Flow chart of RNN

这个网络在 t 时刻接收到输入 x_t 之后，隐藏层的值是 s_t，输出值是 o_t。关键一点是，s_t 的值不仅仅取决于 x_t，还取决于 s_{t-1}。我们可以用下面的公式来表示循环神经网络的计算方法：

$$o_t = g(V_{s_t}) \qquad (3\text{-}1)$$

$$s_t = f(Ux_t + Ws_{t-1}) \qquad (3\text{-}2)$$

式（3-1）是输出层的计算公式，输出层是一个全连接层，也就是它的每个节点都和隐藏层的每个节点相连。V 是输出层的权重矩阵，g 是激活函数。式（3-2）是隐藏层的计算公式，它是循环层。U 是输入 x 的权重矩阵，W 是上一次的值作为这一次的输入的权重矩阵，f 是激活函数。从上面的公式我们可以看出，循环层和全连接层的区别就是循环层多了一个权重矩阵 W。可以看出，循环神经网络的输出值，是受前面历次输入值 $x_t, x_{t-1}, \ldots, x_1$ 的影响，这就是为什么循环神经网络可以往前看任意多个输入值的原因。

Long Short Term 网络一般就叫做 LSTM 是一种 RNN 特殊的类型，可以学习长期依赖信息。LSTM 由 Hochreiter & Schmidhuber (1997)提出，并在近期被 Alex Graves 进行了改良和推广。在很多问题，LSTM 都取得相当巨大的成功，并得到了广泛的使用。LSTM 通过刻意的设计来避免长期依赖问题。记住长期的信息在实践中是 LSTM 的默认行为，而非需要付出很大代价才能获得

的能力！所有 RNN 都具有一种重复神经网络模块的链式的形式。在标准的 RNN 中，这个重复的模块只有一个非常简单的结构，例如一个 tanh 层。

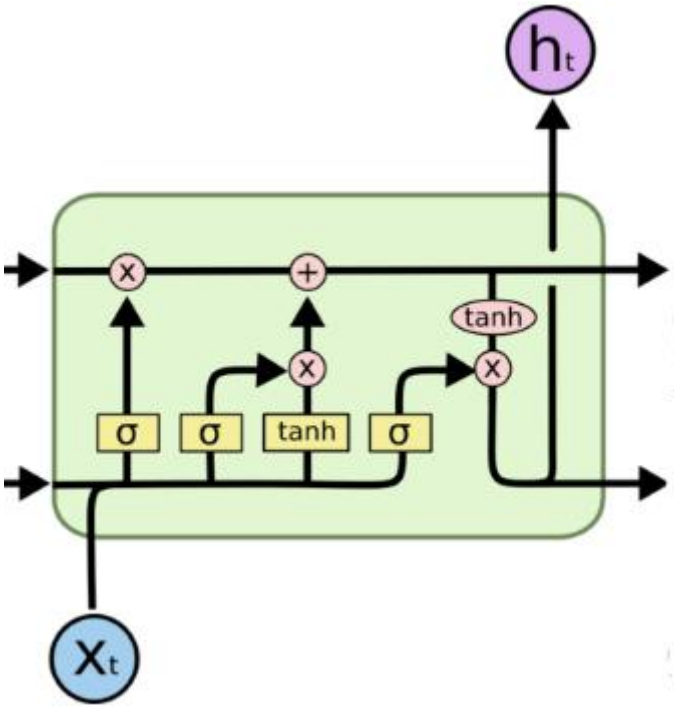


图 3-2 LSTM 流程图
Fig.3-2 Flow chart of LSTM

LSTM 的关键就是细胞状态，水平线在图上方贯穿运行。细胞状态类似于传送带。直接在整个链上运行，只有一些少量的线性交互。信息在上面流传保持不变会很容易。LSTM 有通过精心设计的称作为“门”的结构来去除或者增加信息到细胞状态的能力。门是一种让信息选择式通过的方法，包含一个 sigmoid 神经网络层和一个 pointwise 乘法操作。sigmoid 层输出 0 到 1 之间的数值，描述每个部分有多少量可以通过。LSTM 拥有三个门，来保护和控制细胞状态。在 LSTM 中第一步是决定会从细胞状态中丢弃什么信息。这个决定通过一个称为忘记门层完成。该门的公式表达如式（3-3）所示。该门会读取 h_{t-1} 和 x_t，输出一个在 0 到 1 之间的数值给每个在细胞状态 C_{t-1} 中的数字。1 表示“完全保留”，0 表示“完全舍弃”。

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \qquad (3\text{-}3)$$

下一步是确定什么样的新信息被存放在细胞状态中。这里包含两个部分。第一部分是 sigmoid 层称为“输入门层”，决定什么值我们是将要更新的，如公式（3-4）所示。

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \qquad (3\text{-}4)$$

第二部分是一个 tanh 层创建一个新的候选值向量$\widetilde{C}_t$，该候选值后面会被加入到细胞状态中，如公式（3-5）所示。

$$\widetilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \qquad (3\text{-}5)$$

接下来要将C_{t-1}更新C_t。把旧状态C_{t-1}与f_t相乘，丢弃掉确定需要丢弃的信息，再加上$i_t * \widetilde{C}_t$得到的便是更新后的细胞状态。如公式（3-6）所示。

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \qquad (3\text{-}6)$$

最终，我们需要确定输出什么值。这个输出将会基于细胞状态。首先，通过一个 sigmoid 层如公式（3-7）所示来确定细胞状态的哪些部分将输出。

$$o_t = \sigma(W_o [h_{t-1}, x_t] + b_o) \qquad (3\text{-}7)$$

接着把细胞状态通过 tanh 层进行处理（得到一个在-1 到 1 之间的值）并将它和 sigmoid 门的输出相乘，如公式（3-8）所示。最终输出确定要输出的那部分。

$$h_t = o_t * \tanh(C_t) \qquad (3\text{-}8)$$

GRU 即 Gated Recurrent Unit。为了克服 RNN 无法很好处理远距离依赖而提出了 LSTM，而 GRU 则是 LSTM 的一个变体，当然 LSTM 还有有很多其他的变体。GRU 保持了 LSTM 的效果同时又使结构更加简单，所以它也非常流行。GRU 模型如图所示，GRU 只有两个门，分别为更新门和重置门，即图中的z_t和r_t。更新门用于控制前一时刻的状态信息被带入到当前状态中的程度，更新门的值越大说明前一时刻的状态信息带入越多。重置门用于控制忽略前一时刻的状态信息的程度，重置门的值越小说明忽略得越多。

从直观上来说，重置门决定了如何将新的输入信息与前面的记忆相结合，更新门定义了前面记忆保存到当前时间步的量。如果我们将重置门设置为 1，更新门设置为 0，那么我们将再次获得标准 RNN 模型。为了解决标准 RNN 的梯度消失问题，GRU 使用了更新门（update gate）与重置门（reset gate）。基本上，这两个门控向量决定了哪些信息最终能作为门控循环单元的输出。这两个门控机制的特殊之处在于，它们能够保存长期序列中的信息，且不会随时间而清除或因为与预测不相关而移除。这便是使用门控机制学习长期依赖关系的基本思想。

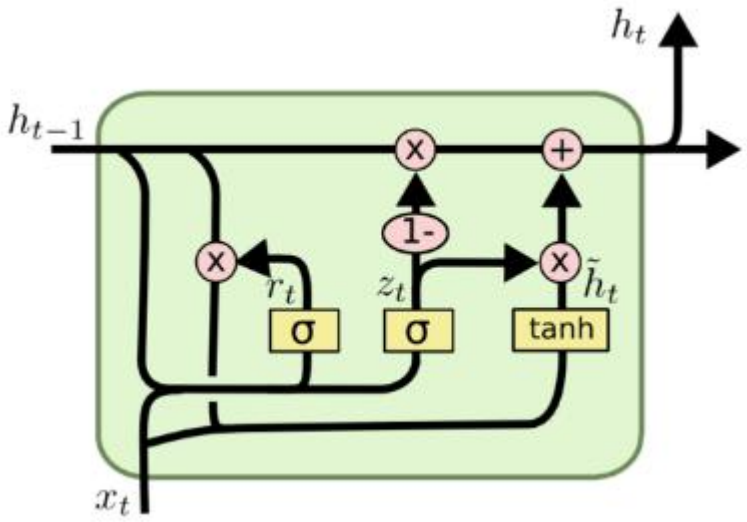


图 3-3 GRU 流程图
Fig.3-3 Flow chart of GRU

重置门的具体计算方式如公式（3-9）所示：

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \tag{3-9}$$

式中 σ —— *sigmoid* 激活函数
W_r ——权重参数
h_{t-1} ——$t-1$ 次的隐含状态
x_t ——t 次输入

更新门的具体如公式（3-10）所示。

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{3-10}$$

其中 z_t 就是计算得到的更新门数值，之后用重置门和第 t 次的输入生成新的候选隐状态。具体计算如公式（3-11）所示。

$$\tilde{h}_t = \tanh(W_{\tilde{h}} \cdot [r_t * h_{t-1}, x_t]) \tag{3-11}$$

其中 $\tilde{h}_t$ 就是候选隐状态，之后用候选隐状态更新当前隐状态，计算机过程如公式（3-12）所示：

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \tag{3-12}$$

更新后隐状态就是 h_t，之后用隐状态 h_t 得到当前 t 次的输出。

$$y_t = \sigma(W_o \cdot h_t) \qquad (3\text{-}13)$$

3.2.2 1D CNN

输入：

1	2	3	4	5	6	7	8

过滤器：

1	1	1	1	1

图 3-4 1d CNN
Fig.3-4 chart of 1d CNN

一维卷积的输入如上所示，输入的数据维度为 8，过滤器的维度为 5。与二维卷积类似，卷积后输出的数据维度为 8−5+1=4。

如果过滤器数量仍为 1，输入数据的 channel 数量变为 16，即输入数据维度为 8×16。这里 channel 的概念相当于自然语言处理中的 embedding，而该输入数据代表 8 个单词，其中每个单词的词向量维度大小为 16。在这种情况下，过滤器的维度由 5 变为 5×16，最终输出的数据维度仍为 4。如果过滤器数量为 n，那么输出的数据维度就变为 4×n。

一维卷积常用于序列模型，自然语言处理领域。

3.3 算法研究

3.3.1 MGVSRN

文本特征的提取方式最普遍常用的是循环神经网络，其中 GRU 使用最多，因为其结构简单，且性能优越。但是单纯的循环神经网络对文本信息的学习能力十分有限，很难充分学习到丰富的文本信息。Dong 等[56]在视频检索任务中提出了多粒度的文本特征学习方法。受其启发，我们用多粒度的文本特征学习方式来改进跨模态图文检索中的文本特征学习部分，提高文本特征学习能力，进而提高跨模态图文检索的效果。具体结构如图（3-5）所示。

我们是在 VSRN 的整体架构上进行的改进，主要的改进部分在文本部分的特征学习部分。首先文本转换为 one-hot 编码后，我们对文本的 one-hot 编码进行第一粒度的整合，方式也比较简单，就是直接取平均。如公式（3-14）所示。

$$f^{(1)} = \frac{1}{n}\sum_{t=1}^{n} c_t \qquad (3\text{-}14)$$

之后我们将 one-hot 编码经过 word-embedding 变成高维特征表示。

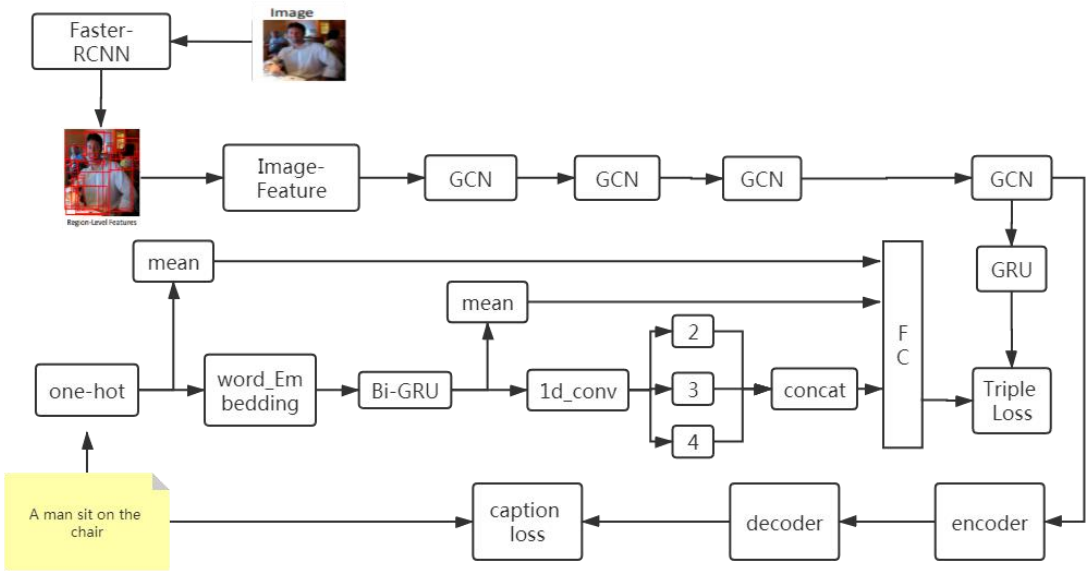


图 3-5 MGVSRN 流程图
Fig.3-5 Flow chart of MGVSRN

word-embedding 是一个矩阵，one-hot 是一个只有 1 维是 1，其余维是 0 的高维向量，两者矩阵相乘可以得到一个特征向量，是 word-embedding 中的某一向量。公式如（3-15）所示。

$$s_t = c_t W_e \tag{3-15}$$

从 word-embedding 得到文本的高维特征表示后，将文本的特征表示经过 Bi-GRU 来进行特征信息的融合。双向递归神经网络可以有效地利用给定序列的过去和将来上下文信息。我们采用双向 GRU（Bi-GRU），它的参数比双向 LSTM 少，因此需要较少的训练数据。Bi-GRU 由两个分离的 GRU 层组成，即前向 GRU 和后向 GRU。前向 GRU 用于以正常顺序编码特征，而后向 GRU 用于以相反顺序编码特征。令 $\vec{h}_t$ 和 $\overleftarrow{h}_t$ 为在特定时间步 $t = 1,\ldots,n$ 时它们对应的隐藏状态。隐藏状态生成如公式（3-16）和公式（3-17）所示。

$$\vec{h}_t = \overrightarrow{GRU}(s_t, \vec{h}_{t-1}) \tag{3-16}$$

$$\overleftarrow{h}_t = \overleftarrow{GRU}(s_{n+1}, \overleftarrow{h}_{t-1}) \tag{3-17}$$

其中 $\overrightarrow{GRU}$ 和 $\overleftarrow{GRU}$ 分别表示前向和后向 GRU，过去的信息分别由 $\vec{h}_{t-1}$ 和 $\overleftarrow{h}_{t-1}$

携带。连接 $\vec{h}_t$ 和 $\overleftarrow{h}_t$，我们得到 Bi-GRU 输出 $h_t = \left[\vec{h}_t, \overleftarrow{h}_t\right]$。向前和向后 GRU 中隐藏向量的大小根据经验设置为 512。因此，h_t 的大小为 1024。将所有输出放在一起，我们得到隐藏向量集合 $H = \{h_1, h_2, \ldots h_n\}$，大小为 $1024 \times n$。通过对沿行维的 H 进行均值池化获得基于 biGRU 的编码。如公式（3-18）所示。

$$f^{(2)} = \frac{1}{n}\sum_{t=1}^{n} h_t \qquad (3\text{-}18)$$

前一层在每个步骤均等地对待 biGRU 的输出。为了增强有助于区分细微差别的文本语义信息，我们在 biGRU 之上构建了卷积网络。特别是，我们采用了最初为句子分类而开发的 1-d CNN。我们 CNN 的输入是由先前的 biGRU 模块生成的隐藏向量集合 H。令 $Conv1d_{k,r}$ 为一维卷积块，其中包含 r = 512 个大小为 k 的滤波器，且 k≥2。将零填充后的 H 放入 $Conv1d_{k,r}$，会生成 n×r 个特征图。通过在特征图上应用 ReLU 激活函数来引入非线性。随着文本中单词个数 n 的变化，我们进一步应用最大池化将特征图压缩为固定长度 r 的向量 c_k。如公式（3-19）所示。

$$c_k = \max\text{-}pooling(\mathrm{Re}LU(Conv1d_{k,r}(H))) \qquad (3\text{-}19)$$

k = 2 的过滤器允许 H 中的两个相邻行彼此交互，而较大 k 的过滤器意味着同时利用更多相邻的行。为了生成多尺度表示，我们部署了多个一维卷积块，其中 k = 2,3,4。它们的输出被 concat 在一起，以形成基于 biGRU-CNN 的编码，如公式（3-20）所示。

$$f^{(3)} = \left[c_2, c_3, c_4\right] \qquad (3\text{-}20)$$

由于 $f^{(1)}$，$f^{(2)}$，$f^{(3)}$ 是通过特定的编码策略在不同级别上依次获得的，因此我们可以合理地假设这三种编码结果彼此互补，并具有一定的冗余度。因此，我们通过级联所有三个级别的输出来获得输入文本的多级编码，如公式（3-21）所示。

$$\phi = \left[f^{(1)}, f^{(2)}, f^{(3)}\right] \qquad (3\text{-}21)$$

实际上，这种级联操作虽然简单，但却是特征组合的一种常见做法。

3.4 实验

3.4.1 实验数据及评价标准

我们在 Microsoft COCO 数据集[64]和 Flickr30K 数据集[65]上进行算法的测试

工作。MS-COCO 数据集共包含了 123,287 张图像，其中每张图像带有 5 个文本描述，文本描述的是图像中的主要信息内容。 我们对 MSCOCO 数据集进行拆分，其中包含 113,287 张用于训练的图像，5000 张用于验证的图像和 5000 张用于测试的图像。每个图像带有 5 个标题。 最终的实验结果是通过对 5 倍的 1K 测试图像的结果进行取平均或在完整的 5K 测试图像上进行测试而获得的。Flickr30K 包含了从 Flickr 网站收集的 31783 张图像，每张图像随附 5 条人类注释文本说明。我们使用标准的训练，验证和测试分割，分别包含 28,000 张图像，1000 张图像和 1000 张图像。

对于信息检索中常见的评估指标。本章通过在 K（R @ K）处进行召回率的计算来衡量跨模态检索性能，R@K 代表着在特征空间距离上距离查询最近的前 K 个目标模态特征包含与查询相匹配的的特征信息的概率。本章共设置了 R@1，R@5，R@10 三种不同的指标。

3.4.2 实验具体情况

我们将单词嵌入大小设置为 300，将联合嵌入空间的维度设置为 2048。我们采用与[61，62]相同的设置来设置视觉自下而上的注意模型。 基于 GRU 的全局语义推理的区域顺序由自下而上的注意力探测器生成的类别检测置信度得分的降序确定。我们使用 Adam 优化器训练了 30 个 epoch。 我们开始的 15 个 epoch 以学习率 0.0002 进行训练，然后将其余 15 个 epoch 的学习率降低到 0.00002。 我们在等式中设置边距 α 为 0.2。 我们使用的最小批量为 128。对于测试集的评估，我们通过选择在验证集上表现最佳的模型来解决过度拟合问题。根据验证集中的召回总和选择最佳模型。实验环境会对实验结果有很大影响，尤其是 python、pytorch 和其它库的版本对实验结果影响很大。

3.4.3 实验结果及分析

首先，我们在 Flick30k 数据集上进行了实验，图像特征已经经过预处理，用 Faster-CNN 提取了特征信息。Faster-RCNN 是在 VG 数据集上训练好后在 Flick30k 数据集上对图像进行特征提取和特征计算。本章采用的数据集和 VSRN 作者采用了同一个数据集。实验结果如表（3-1）所示。

从表（3-1）中的实验结果可以看出来本章提出算法的检索效果在图像检索文本和文本检索图像上都要好于 VSRN 的效果。在图像检索文本任务上，R@1 提升了 0.7 个百分点，R@5 提升了个 1.5 个百分点 R@10 提升了 0.9 个百分点。在文本检索图像任务上，R@1 提升了 0.1 个百分点，R@5 提升了 0.8 个百分点，

R@10 提升了 0.3 个百分点。可以看出 MGVSRN 在各项上都有一定的提升，是整个检索效果整体态势的提高。这里 VSRN 的实验结果是我们用作者公布的代码跑出的实验结果。VSRN 和 MGVSRN 的实验环境、超参数的配置等等都完全一致，只是改进部分的神经网络结构不同。两者是在完全一致的环境和设置下进行的实验对比。可以看出本章的多粒度文本特征学习改善了检索效果。多粒度的文本特征学习可以得到更加丰富的多种粒度的文本特征信息，这种多粒度的文本特征信息提高了跨模态检索的效果。

表 3-1 VSRN 在 Flick30k 上实验

Table3-1 Data of MGVSRN on Flick30k

Methods	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
$SMlstm_{CVPR'17}$	42.5	71.9	81.5	30.2	60.4	72.3
$VSE{+}{+}_{BMVC'18}$	52.9	79.1	87.2	39.6	69.6	79.5
$SCO_{CVPR'18}$	55.5	82.0	89.3	41.1	70.5	80.1
$SCAN_{ECCV'18}$	67.4	90.3	95.8	48.6	77.7	85.2
VSRN(our)	68.1	88.9	93.7	52.1	78.7	86.1
MGVSRN(ours)	68.8	90.4	94.6	52.2	79.5	86.4

之后我们又在 COCO 数据集上对 MGVSRN 进行了实验，实验结果如表（3-2）和表（3-3）所示。从实验结果我们可以看出在图像检索文本任务上效果有所下降尤其是在 R@1 上效果下降明显。我们认为这是多粒度文本特征学习的能力太强导致的。多粒度文本特征学习部分的参数量和神经网络的复杂度要高出简单的循环神经网络很多。这就导致在训练过程中，进行参数调整时，文本特征学习部分也是起到重要作用的一方面。文本特征学习部分的参数学习也会对匹配效果产生很大的影响。同时算法中图像部分的特征学习十分庞大且复杂，有着复杂的神经网络和大量的参数来学习图像特征。这就导致在检索匹配中，图像部分的参数调整会对检索匹配效果产生很大的影响。由于我们是根据特征向量之间的空间距离进行的检索。损失函数 Ranking Triple Loss 是让匹配的文本特征和图像特征的距离的比不匹配的更近。由于文本特征学习和图像特征学习是独立分开的，因此在训练过程中，参数调整可以看作是对空间中图像特征或

文本特征进行空间上的移动，通过调整参数来使得文本特征或图像特征在空间上的位置不断移动，使匹配的文本特征和图像特征离很近，不匹配的文本特征和图像特征距离很远，进而使匹配的特征向量比不匹配的特征向量距离更近。这便是在特征空间中，参数调整对特征向量进行的空间移动。由于检索匹配时是根据文本特征和图像特征之间的空间距离来判断的。而文本部分的学习能力很强，或者说对文本的空间移动能力很强，当需要移动特征向量来提升匹配效果时，文本特征会移动的很多，很活跃。同样图像部分的学习能力也十分强，当需要在空间中移动特征向量来提升匹配效果时，图像特征也移动的很多，很活跃。这样导致的结果就是当特征向量需要移动时，文本特征和图像特征都进行了大量的移动，而导致整体移动过度。及图像特征调整了很多可以让结果变的更好，此时文本特征只需较少移动或不怎么移动即可，但是实际情况是文本特征也移动了很多，这就导致了特征移动过度，反而效果不太好。但是在 Flick30k 上检索效果确很好，这可能和数据集的数据有关。Flick30k 的数据来源是 Flick 社交网站，人们在上面上传的图片和文字有一定的主题性，如饮食、服装会占多数。因此数据集中的数据较为相似，神经网络所学习的数据信息也较为相似，这样数据就限制了神经网络的学习能力的开发。由于很多时候学习的数据很类似，神经网络学习的特征信息也就集中在一些方面，在神经网络学习训练时，这些相似数据可以相互增强效果，因而检索效果会更好。同时由于大量的数据较为相近，在空间中的空间距离差异也较小，因此在调整参数时，需要调整的也比较小，因此也就不容易出现参数调整过度的现象。

表 3-2 MGVSRN 在 coco (1k)上实验

Table3-2 Data of MGVSRN on coco (1k)

Methods	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
$SMlstm_{CVPR'17}$	53.2	83.1	91.5	40.7	75.8	87.4
$VSE++_{BMVC'18}$	64.6	89.1	95.7	52.0	83.1	92.0
$SCO_{CVPR'18}$	69.9	92.9	97.5	56.7	87.5	94.8
$SCAN_{ECCV'18}$	72.7	94.8	98.4	58.8	88.4	94.8
VSRN(ours)	73.0	94.1	97.8	60.3	88.4	94.2
MGVSRN(ours)	70.8	92.9	97.6	60.2	88.2	93.9

我们可以看到在 coco1k 和 coco5k 上整体的效果都有所下降，但是在 coco5k 上文本检索图像的效果反而提升了一些，这并不是提升，而是效果下降的少。从表（3-2）和表（3-3）可以看出来，从 coco1k 到 coco5k 有一个检索效果的下降，这可能是因为测试集的数据量增大导致了检索中的干扰项增多，进而导致了检索效果的下降。但是当文本和图像特征的学习能力都很强时，测试集的数量增大对其产生的影响会减小。因为文本和图像的特征学习能力强，可以学到更加的丰富、精微的特征信息，因此可以学到更具有分辨性的特征信息。因此抗干扰能力也会更强，当测试集的数量增大时，检索效果下降的相对会更少。我们可以看到，在图像检索文本任务上，coco5k 上 MGVSRN 相对于 VSRN 下降的要比 coco1k 的少，这也是测试集增大，而 MGVSRN 由于文本特征学习能力强，增强了抗干扰能力，进而检索效果下降的少的原因。

表 3-3 MGVSRN 在 coco (5k)上实验

Table3-3 Data of MGVSRN on coco (5k)

Methods	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
VSE++$_{BMVC'18}$	41.3	69.2	81.2	30.3	59.1	72.4
SCO$_{CVPR'18}$	42.8	72.3	83.0	33.1	62.9	75.5
SCAN$_{ECCV'18}$	50.4	82.2	90.0	38.6	69.3	80.4
VSRN(ours)	48.9	78.0	87.4	37.2	68.0	79.2
MGVSRN(ours)	47.1	77.0	86.6	37.7	68.1	79.2

3.5 本章小结

本章中我们使用多粒度文本特征学习来提升文本部分的特征学习能力，通过使用多种粒度的文本特征学习来代替原来的简单的循环神经网络 GRU 来使文本部分可以学习到更加丰富、多粒度的文本信息。对文本进行了更加充分的学习，进而提升了检索效果。但文本部分的学习能力过强有时会影响检索效果，因为图像和文本的学习能力都太强时，在空间中移动特征向量的方法，很容易使特征向量移动过度。但是当检索的规模变大时，更强的特征学习能力可以提高检索的抗干扰性，进而在检索效果会更好。

第 4 章 混合检索

4.1 引言

跨模态检索是人工智能领域的一个重要研究方向。在社会生活中应用广泛，有着巨大的应用价值和经济价值。随着深度学习的兴起，跨模态检索也取得了长足发展。跨模态检索是获得与对应的文本（图像）信息相匹配的图像（文本）信息，而推荐是根据以往感兴趣的信息推荐相似的信息。这两者有着相同的地方，都是根据一些信息寻找相似的其它信息。借鉴混合推荐模型的思想，我们在跨模态检索中引入了混合检索。即在一个算法框架中实现两个或多个检索模型，类似混合推荐用平均值来整合多个推荐模型一样。我们通过对检索算法中得到的图像特征和文本特征分别取平均值来整合两个检索模型，实验证明混合检索的方式可以提高检索效果，增强检索的抗干扰能力。

4.2 相关技术理论

4.2.1 Triplet Loss

近来 learning to rank 的思想逐渐被应用到很多领域，比如人脸识别，行人验证等等。learning to rank 中一个非常重要的步骤就是找到一个好的 similarity function，而 triplet loss 是用的非常广泛的一种。

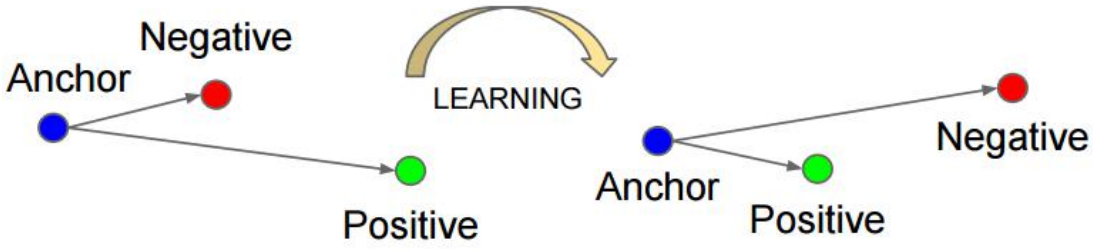


图 4-1 Triple Loss Sketch
Fig.4-1 Sketch of Triple Loss

如图（4-1）所示，triplet 是一个三元组，这个三元组的构成是这样的：从训练数据集中随机选一个样本，将该样本称为 Anchor，然后再随机选取一个和 Anchor (记为 x_a)属于同一类的样本和不同类的样本,这两个对应的样本分别称为 Positive (记为 x_p)和 Negative (记为 x_n)，由此构成一个（Anchor，Positive，Negative）三元组。针对三元组中的每一个元素（样本），训练一个参数共享

或者不共享的网络，得到三个元素的特征表达，分别记为：$f(x_a)$，$f(x_p)$，$f(x_n)$。triplet loss 的目的就是通过学习，让 x_a 和 x_p 特征表达之间的距离尽可能小，而 x_a 和 x_n 的特征表达之间的距离尽可能大，并且要让 x_a 与 x_n 之间的距离和 x_a 与 x_p 之间的距离之间有一个最小的间隔 α。公式化的表示如公式（4-1）所示。

$$\left\|f(x_i^a)-f(x_i^p)\right\|_2^2+\alpha<\left\|f(x_i^a)-f(x_i^n)\right\|_2^2 \qquad \forall(f(x_i^a),f(x_i^p),f(x_i^n))\in T \quad （4\text{-}1）$$

对应的目标函数也就很清楚了如公式（4-2）所示。

$$\sum_i^N\left[\left\|f(x_i^a)-f(x_i^p)\right\|_2^2-\left\|f(x_i^a)-f(x_i^n)\right\|_2^2+\alpha\right]_+ \qquad （4\text{-}2）$$

这里距离用欧式距离度量，+表示[]内的值大于零的时候，取该值为损失，小于零的时候，损失为零。由目标函数可以看出。当 x_a 与 x_n 之间的距离 $<x_a$ 与 x_p 之间的距离加 α 时，[]内的值大于零，就会产生损失。当 x_a 与 x_n 之间的距离 $\geq x_a$ 与 x_p 之间的距离加 α 时，损失为零。

4.2.2 Multi-Layer Perceptron

多层感知机是由感知机推广而来，感知机学习算法(PLA: Perceptron Learning Algorithm)用神经元的结构进行描述的话就是一个单独的。感知机的神经网络如图（4-2）所示。

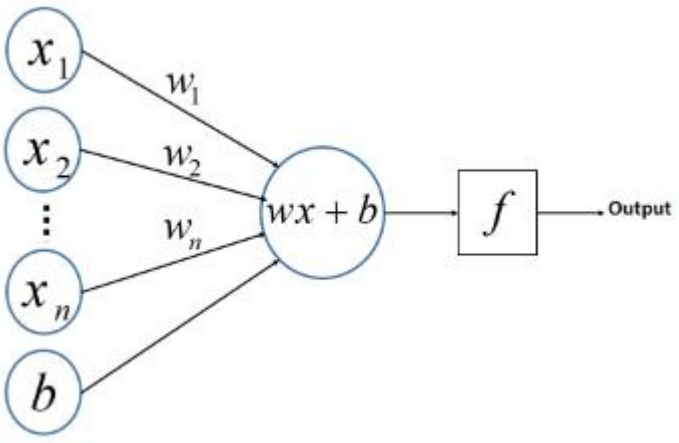


图 4-2 感知机流程图
Fig.4-2 Flow chart of Perceptron

$$u=\sum_{i=1}^{n}w_ix_i+b \qquad （4\text{-}3）$$

$$y=sign(u)\begin{cases}+1,u>0\\-1,u\leq 0\end{cases}\text{从} \qquad （4\text{-}4）$$

从上述内容更可以看出，PLA 是一个线性的二分类器，但不能对非线性的数据并不能进行有效的分类。因此便有了对网络层次的加深，理论上，多层网络可以模拟任何复杂的函数。

多层感知机由感知机推广而来，最主要的特点是有多个神经元层，因此也叫深度神经网络(DNN: Deep Neural Networks)。多层感知机（MLP，Multilayer Perceptron）也叫人工神经网络（ANN，Artificial Neural Network），除了输入输出层，它中间可以有多个隐层，最简单的 MLP 只含一个隐层，即三层的结构。

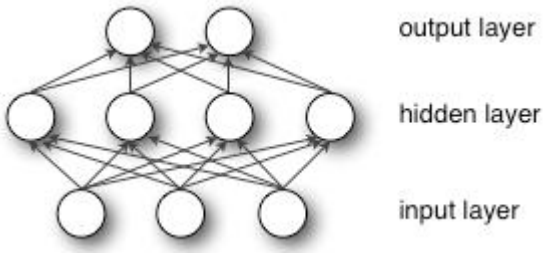


图 4-3 多层感知机流程图
Fig.4-3 Flow chart of Multilayer Perceptron

从图（4-3）可以看到，多层感知机层与层之间是全连接的（全连接的意思就是：上一层的任何一个神经元与下一层的所有神经元都有连接）。多层感知机最底层是输入层，中间是隐藏层，最后是输出层。输入层没什么好说，你输入什么就是什么，比如输入是一个 n 维向量，就有 n 个神经元。隐藏层的神经元怎么得来？首先它与输入层是全连接的，假设输入层用向量 X 表示，则隐藏层的输出就是 $f(W_1X+b_1)$，W_1 是权重（也叫连接系数），b_1 是偏置，函数 f 可以是常用的 *sigmoid* 函数或者 tanh 函数。

$$sigmoid(a)=\frac{1}{1+e^{-a}} \qquad (4\text{-}5)$$

$$\tanh(a)=\frac{e^{a}-e^{-a}}{e^{a}+e^{-a}} \qquad (4\text{-}6)$$

最后就是输出层，隐藏层到输出层可以看成是一个多类别的逻辑回归，也即 *soft* max 回归，所以输出层的输出就是 $soft\max(W_2x_1+b_2)$，x_1 表示隐藏层的输出 $f(W_1x+b_1)$。

MLP 整个模型就是这样子的，上面说的这个三层的 MLP 用公式总结起来就是，函数 G 是 *soft* max 。

$$f(x)=G(b^{(2)}+W^{(2)}(s(b^{(1)}+W^{(1)}x))) \qquad (4\text{-}7)$$

因此，MLP 所有的参数就是各个层之间的连接权重以及偏置，包括 w_1、b_1、w_2、b_2。

4.2.3 反向传播算法

BP 算法是学习过程由信号的正向传播与误差的反向传播两个过程组成。正向传播时，输入样本从输入层传入，经各隐层逐层计算后，传向输出层。若输出层的实际输出与期望的输出（ground truth）不符，则转入误差的反向传播阶段。误差反传是将输出误差以某种形式通过隐层向输入层逐层反传，并将误差分摊给各层的所有单元，从而获得各层单元的误差信号，此误差信号即作为修正各单元权值的依据。这种信号正向传播与误差反向传播的各层权值调整过程，是周而复始地进行的。权值不断调整的过程，也就是网络的学习训练过程。此过程一直进行到网络输出的误差减少到可接受的程度，或进行到预先设定的学习次数为止。

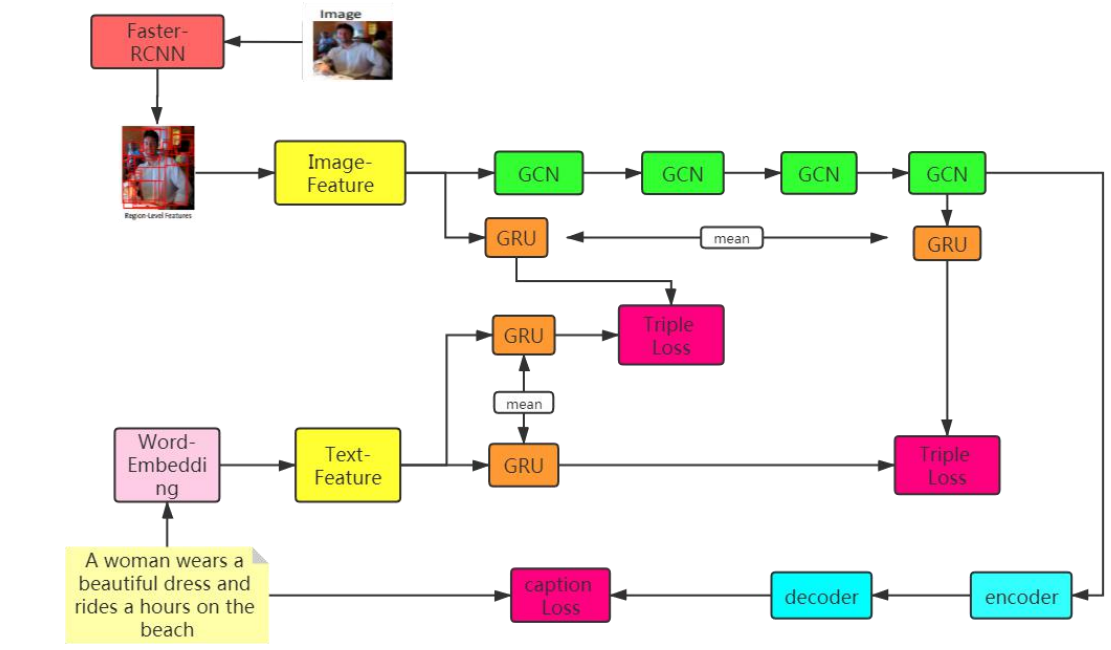
反向传播阶段的误差计算是根据损失函数和求导来算得的。我们学习的目标是让输出层的结果和 ground truth 一致，此时损失函数为 0。我们对参数进行调整的目的就是要减小损失函数，让其近可能地接近 0。我们首先用链式求导来求得多层神经网络中，每一层参数和损失函数之间的导数。这个导数是参数变化量和输出结果变化量之间的比例。因为我们要输出结果下降到 0。因此我们让参数向可以让损失结果减小的方向移动一个量，通过这个方式来使得损失结果减小。这个参数移动方向为梯度下降方向，将参数看作自变量，损失结果看作因变量，自变量向着该方向移动会导致因变量会减小，一般通过导数取负来实现。因为导数为正说明参数增大，损失变大。此时导数取反与步长相乘得到移动量，这个移动量为负，参数加上移动量后会向着变小的方向移动，损失函数结果自然也变小。当导数为负时，说明损失函数结果会随着参数的增大而减小。将导数取反乘上步长后得移动量，此时移动为正。参数加上这个移动量后会向增大的方向移动，损失函数结果会减小。

4.3 算法研究

4.3.1 MLMVSRN

不同的检索模型有不同的检索效果，有的在文本检索图像上效果优越，有的在图像检索文本上效果突出。但是总体来看所有的检索模型都有自身的检索效果，无论是文本检索图像还是图像检索文本，不同检索模型都同时具备着在这两个方面上的检索能力。既然检索模型都具有检索能力，那这种检索能力能

否互相促进。很多跨模态检索的损失函数是 triple ranking loss，这种函数是根据特征向量之间的空间距离远近来对特征进行排名的，排名结果即为检索结果。这个损失函数在训练中的作用就是要让匹配的文本特征和图像特征之间的空间距离要比不匹配的近，及在空间距离上与文本（图像）匹配的图像（文本）要比不匹配的距离文本（图像）更近。经过检索模型计算后得到的特征都会具有一种态势，就是相匹配的特征向量间的空间距离要比不匹配的近，而这样的空间态势可以进行叠加。对两个检索模型得到的图像特征和文本特征分别取平均，这样两个检索模型中的文本特征与图像特征的距离也会取平均。因为每个检索模型都具有让匹配的文本图像距离更近不匹配的距离更远的态势，通过取平均的方式来实现态势的叠加，进而提升整体检索的效果。具体模型结构如图（4-5）所示。



、

图 4-4: MLMVSRN 流程图

Fig. 4-4 The flow chart of MLMVSRN.

我们的混合检索是在 VSRN 的基础上改进而来，其核心思想是将多个检索模型的检索能力进行融合以实现互相增强。这里我们的改进之处在于多出了一个简单的跨模态检索模型。文本部分我们额外用一个 GRU 来整合文本特征，这一部分和原来 VSRN 中文本部分的特征整合一致。但新增的 GRU 用于另一个检索模型的匹配。如公式（4-8）和公式（4-9）所示。

$$s_1 = GRU_1(X) \qquad (4\text{-}8)$$

$$s_2 = GRU_2(X) \qquad (4\text{-}9)$$

其中 $X = \{x_1, x_2, \ldots, x_n\}$，$X$ 代表整个文本的特征集合，x_i 表示第 i 个单词的特征向量表示。GRU_2 是原来的文本特征融合部分，输出的 s_2 是整体文本特征表示，GRU_1 是我们新增的文本特征融合部分，也是作为增加的检索模型的文本特征学习部分，输出的 s_1 将用于新增加的检索模型的匹配。

图像部分我们在 Faster-RCNN 提取对象特征后就对对象特征进行整合，得到整张图片的特征表示，我们通过使用 GRU 来将各个对象特征融合在一起，这些对象特征都未经过图卷积。如公式（4-10）所示。

$$f_1 = GRU_3(V) \qquad (4\text{-}10)$$

其中 $V = \{v_1, v_2, \ldots, v_n\}$ 表示图像特征集合，v_i 代表图像中第 i 个对象的特征向量。经过 GRU_3 得到的 f_1 便是整张图像的特征表示，f_1 用于和 s_1 进行匹配，这样便得到一个检索模型。V 还有一个分支经过 4 层图卷积的学习后得到新的图像特征集合 $V^* = \{v_1^*, v_2^*, \ldots v_n^*\}$，其中 v_i^* 代表经过图卷积后的第 i 个对象的图像特征。公式如（4-11）所示。

$$V^* = GCN(V) \qquad (4\text{-}11)$$

之后，我们对得到的图像特征集合 V^* 进行整合得到整张图片的特征表示，这里我们也采用 GRU 来进行特征融合。如公式（4-12）所示。

$$f_2 = GRU_4(V^*) \qquad (4\text{-}12)$$

f_2 即是对 V^* 进行整合后得到的整张图像的特征。我们将 f_2 与 s_2 进行匹配得到另一个检索模型。这样整个算法框架里就包含了两个检索模型。在测试的时候，或者说实际检索的时候，我们将分别对 s_1, s_2 和 f_1, f_2 取平均，用取平均后的图像特征和文本特征进行检索。这便是多层次混合检索的整个过程。

4.4 实验

4.4.1 实验数据和评价标准

我们在 Microsoft COCO 数据集[64]和 Flickr30K 数据集[65]上进行算法的测试工作。MS-COCO 数据集共包含了 123,287 张图像，其中每张图像带有 5 个文本描述，文本描述的是图像中的主要信息内容。 我们对 MSCOCO 数据集进行拆分，其中包含 113,287 张用于训练的图像，5000 张用于验证的图像和 5000 张用于测试的图像。每个图像带有 5 个标题。 最终的实验结果是通过对 5 倍的 1K 测试图像的结果进行取平均或在完整的 5K 测试图像上进行测试而获得的。

Flickr30K 包含了从 Flickr 网站收集的 31783 张图像，每张图像随附 5 条人类注释文本说明。我们使用标准的训练，验证和测试分割，分别包含 28,000 张图像，1000 张图像和 1000 张图像。

对于信息检索中常见的评估指标。本章通过在 K（R @ K）处进行召回率的计算来衡量跨模态检索性能，R@K 代表着在特征空间距离上距离查询最近的前 K 个目标模态特征包含与查询相匹配的的特征信息的概率。本章共设置了 R@1，R@5，R@10 三种不同的指标。三种指标分别代表了三种不同的程度考察。

4.4.2 实验具体情况

我们将单词嵌入大小设置为 300，将联合嵌入空间的维度设置为 2048。我们采用与[61，62]相同的设置来设置视觉自下而上的注意模型。 基于 GRU 的全局语义推理的区域顺序由自下而上的注意力探测器生成的类别检测置信度得分的降序确定。我们使用 Adam 优化器训练了 30 个 epoch。 我们开始的 15 个 epoch 以学习率 0.0002 进行训练，然后将其余 15 个 epoch 的学习率降低到 0.00002。 我们在等式中设置边距 α 为 0.2。 我们使用的最小批量为 128。对于测试集的评估，我们通过选择在验证集上表现最佳的模型来解决过度拟合问题。根据验证集中的召回总和选择最佳模型。实验环境会对实验结果有很大影响，尤其是 python、pytorch 和其它库的版本对实验结果影响很大。

4.4.3 实验结果及分析

表 4-1 MLMVSRN 在 Flick30k 上实验

Table4-1 Data of MLMVSRN on Flick30k

Methods	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
$SMlstm_{CVPR'17}$	42.5	71.9	81.5	30.2	60.4	72.3
$VSE{+}{+}_{BMVC'18}$	52.9	79.1	87.2	39.6	69.6	79.5
$SCO_{CVPR'18}$	55.5	82.0	89.3	41.1	70.5	80.1
$SCAN_{ECCV'18}$	67.4	90.3	95.8	48.6	77.7	85.2
VSRN(our)	68.1	88.9	93.7	52.1	78.7	86.1
MLMVSRN(ours)	71.1	90.4	94.9	47.8	76.8	84.5

首先我们在 Flick30k 数据集上进行了测试，并将 MLMVSRN 和 VSRN 进行了对比。实验结果如表（4-1）所示。

从表（4-1）中的实验结果可以看出来，在 Flick30k 数据集上。MLMVSRN 图像检索文本的效果要明显好于 VSRN，R@1 提高了 3 个百分点，R@5 提升了 0.5 个百分点，R@10 提升了 1.2 个百分点。这说明混合检索可以明显提高图像检索文本的效果。这可能是因为文本部分得到的两个文本特征差异不大，因为文本部分得到两个文本特征所使用的神经网络结构是一致的，都是 GRU，而且 GRU 的输入是同一个 word-embedding，这就使得我们学到的两个文本特征差异不会很大，所以文本特征从整体态势来看具有一定的一致性。这样两个文本特征取平均相当于是两种态势的叠加，由于一致性较好，因此会有态势上的互相增强，使得取平均后得到的文本特征会更加利于检索。因此图像检索文本的效果会得到提升。

但是文本检索图像的效果反而有所下降，这很可能是因为 Flick30k 数据集上的数据存在主题性的原因。Flick30k 数据集中的数据来自 Flicker 网站，人们在 Flicker 网站上分享个人日常生活信息，但是人们分享的生活信息往往具有一定的主题性。例如：饮食、衣服、以及一些活动等等，这类话题往往是人们分享最多的生活信息，也是人们谈论最多最广泛的内容。由于数据集中的数据具有主题性，因此数据集中会出现大量数据属于同一个主题的现象，而且同一个主题的数据在内容上往往很接近。MLMVSRN 的图像特征学习部分比 VSRN 多出了一个对象特征融合的分支。在 Faster-RCNN 得到图像的对象特征后，直接用对象特征进行融合得到整张图片的特征表示。这个图像的整体特征表示是在对象特征的基础上得来的，包含的都是图像中的内容信息。由于同一主题下大量数据在内容上很接近，因此这一分支得到的图像特征表示就很相似，在特征空间上距离也很近，很难分辨。当这个图像特征表示整合进最终用于检索的图像特征表示后，就导致了最终的图像特征表示也比较相近，难以分辨。因此文本检索图像的效果就出现了下降的现象。

之后我们又在 coco 数据集上进行了训练和测试，实验结果如表（4-2）和表（4-3）所示。从表（4-2）中的实验结果可以看到，我们在 coco1k（1k 是指测试集的数据量）上相比于 VSRN(ours)检索效果有了整体提升。图像检索文本任务上 R@1 提升了 0.8 个百分点，R@5 提升了 0.3 个百分点，R@10 提升了 0.2 个百分点。其中 R@1 提升最多，这说明 MLMVSRN 确实在图像检索文本任务上实现了效果的提升。在 coco1k 上文本检索图像也同样获得了效果的提升，R@1 提升了 1.1 个百分点，R@5 提升了 0.5 个百分点，R@10 提升了 0.4 个百

分点。这是因为 coco 数据集中的数据不存在明显的主题性，不会出现很多数据属于同一主题非常接近的现象。在 coco5k 上，MLMVSRN 的实验结果同样整体好于 VSRN，这说明 MLMVSRN 确实提升了算法的检索效果，混合检索通过将两个检索模型整合在一起，确实可以提升检索的效果。从表（4-2）和表（4-3）可以看出，MLMVSRN 在 coco5k 上的提升幅度要高于在 coco1k 上的提升幅度。首先我们可以看到从 coco1k 到 coco5k 整个的测试结果都有一个明显的下降，这是因为测试集变大的原因。当测试集的数量变大时，在检索的时候面对的检索对象就会变多，干扰项就会变多，因此检索效果会出现下降的现象。但是当用两个特征向量取平均来进行检索时，用于检索的特征向量是由两个检索模型

表 4-2 MLMVSRN 在 coco (1k)上实验

Table4-2 Data of MLMVSRN on coco (1k)

Methods	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
$SMlstm_{CVPR'17}$	53.2	83.1	91.5	40.7	75.8	87.4
$VSE{+}{+}_{BMVC'18}$	64.6	89.1	95.7	52.0	83.1	92.0
$SCO_{CVPR'18}$	69.9	92.9	97.5	56.7	87.5	94.8
$SCAN_{ECCV'18}$	72.7	94.8	98.4	58.8	88.4	94.8
VSRN(ours)	73.0	94.1	97.8	60.3	88.4	94.2
MLMVSRN(ours)	73.8	94.4	98.0	61.2	88.9	94.6

表 4-3 MLMVSRN 在 coco (5k)上实验

Table4-3 Data of MLMVSRN on coco (5k)

Methods	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
$VSE{+}{+}_{BMVC'18}$	41.3	69.2	81.2	30.3	59.1	72.4
$SCO_{CVPR'18}$	42.8	72.3	83.0	33.1	62.9	75.5
$SCAN_{ECCV'18}$	50.4	82.2	90.0	38.6	69.3	80.4
VSRN(ours)	48.9	78.0	87.4	37.2	68.0	79.2
MLMVSRN(ours)	50.1	79.5	88.4	38.2	69.4	80.2

共同产生的，是两个检索模型共同在起作用，因此抗干扰能力就会更强。当检索的数据量增大时对其产生的影响会更小，检索效果下降的就会更少。因此导致了 MLMVSRN 在 coco5k 上的提升幅度要高于在 coco1k 上的提升幅度。

4.5 本章小结

本章我们通过使用混合检索来改善检索效果。我们在一个算法框架内包含了两个检索模型，这两个模型分别独立进行检索匹配训练，在检索时对两个检索模型得到的图像特征和文本特征分别取平均，用取平均后的图像特征和文本特征来进行检索。通过这样的方式我们将两个检索模型整合在了一起。这种检索模型的整合可以使不同检索模型互相增强进而提升检索效果，同时由于一个算法中包含了多个检索模型，是多个检索模型同时在起作用，因此混合检索的抗干扰性更强。

结　论

本文对目标检测和图卷积结合的跨模态检索算法进行了深入研究。包括多层图卷积的中间层研究，引入注意力机制的图像特征融合，使用分合思想的精微匹配问题的研究。

在多层图卷积研究中我们发现，使用跳跃链接，可以使中间层跳过后续的图卷积直接导入之后的神经网络，这种方式可以让我们灵活地控制中间各层的特征学习程度，使得每一层参数的训练程度可以进行灵活控制。通过添加中间层的匹配约束，我们提升了多层图卷积的特征学习能力和跨模态检索效果。在图卷积中间层的开发上，我们还可以使用注意力机制信息提取中间层信息，也可以将中间层特征融合后与最后一层特征融合后衔接在一起。这些方法都可以再进一步研究，来丰富开发中间层图卷积的方法。

在多粒度文本特征学习中，我们用多粒度的文本特征学习来代替原有的循环神经网络 GRU，使文本部分拥有更加强大的文本特征学习能力。可以学习到更加丰富多粒度的文本信息，对文本学习的更加充分，改善了检索效果。但文本部分的学习能力的过强有时也会影响到检索效果。这是因为当图像和文本的学习能力都很强时，在特征空间中文本特征和图像特征都会很灵活地移动来改进匹配效果。当两者都很灵活时，容易出现调整过度而导致的效果稍微下降的现象。但当文本和图像学习能力都很强时，对整个算法抗干扰能力的提升很有帮助。当检索的数据规模变大时，检索中的干扰项就会增多，这时的检索效果往往会有很大的下降，通过增强图像特征和文本特征的学习能力可以提升算法抗干扰能力，减少效果的下降。

在混合检索中，我们在一个算法框架中同时实现了两个检索模型，这两个模型有一部分是共用部分，也有各自独立的神经网络部分，在训练时这两个检索模型分别独立训练。每个检索模型都有着自己的检索能力，在检索时我们将两者的图像特征和文本特征通过取平均整合在一起。这样来实现两个检索模型检索能力的相互增强进而提升检索效果，当两个检索模型得到的特征向量整体态势较一致时，增强效果很更明显。但如果两个检索模型得到的特征向量的整体态势差异较大，那会对检索效果产生一定的影响，因此混合检索中的检索模型应该是具有某种联系的检索模型，来实现检索效果上的相互增强。由于算法中是两个检索模型在起作用，因此整个算法的抗干扰能力会更强。当检索的数据规模增大，干扰项增多时，混合检索的效果下降的相对较少。

参考文献

[1] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet,R. Levy, and N. Vasconcelos, “A new approach to cross-modal mul-timedia retrieval,” in International conference on Multimedia. ACM,2010, pp. 251－260.

[2] F. Zhu, L. Shao, and M. Yu,“Cross-modality submodular dictionary learning for information retrieval,” in International Conference on Information and Knowledge Management. ACM, 2014, pp. 1479－1488.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,”Journal of Machine Learning Research, vol. 3, pp. 993－1022, 2003.

[4] Y. Jia, M. Salzmann, and T. Darrell, “Learning cross-modality similarity for multinomial data,” in International Conference on Computer Vision. IEEE, 2011, pp. 2407－2414.

[5] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng,“Multimodal deep learning,” in International Conference on Machine Learning, 2011, pp. 689－696.

[6] N. Srivastava and R. R. Salakhutdinov, “Multimodal learning with deep boltzmann machines,” in Advances in Neural Information Processing Systems, 2012, pp. 2222－2230.

[7] F. Yan and K. Mikolajczyk, “Deep correlation for matching images and text,” 2015, pp. 3441－3450.

[8] F. Feng, X. Wang, and R. Li, “Cross-modal retrieval with correspondence autoencoder,” in International Conference on Multimedia.ACM, 2014, pp. 7－16.

[9] R. Xu, C. Xiong, W. Chen, and C. J. J., “Jointly modeling deep video and compositional text to bridge vision and language in a unified framework,”in AAAI Conference on Artificial Intelligence, 2015, pp.2346－2352.

[10] N. Quadrianto and C. H. Lampert, “Learning multi-view neighborhood preserving projections,” in International Conference on Machine Learning, 2011, pp. 425－432.

[11] D. Zhai, H. Chang, S. Shan, X. Chen, and W. Gao, “Multiview metric

learning with global consistency and local smoothness,” ACM Transactions on Intelligent Systems and Technology, vol. 3, no. 3, 2012.

[12] X. Zhai, Y. Peng, and J. Xiao, “Heterogeneous metric learning with joint graph regularization for cross-media retrieval,” in AAAI Conference on Artificial Intelligence, 2013, pp. 1198－1204.

[13] Z. Yuan, J. Sang, Y. Liu, and C. Xu, “Latent feature learning in social media network,” in International Conference on Multimedia. ACM, 2013, pp. 253－262.

[14] J. Wang, Y. He, C. Kang, S. Xiang, and C. Pan, “Image-text cross-modal retrieval via modality-specific feature learning,” in International Conference on Multimedia Retrieval, 2015, pp. 347－354.

[15] J. Weston, S. Bengio, and N. Usunier, “Wsabie: Scaling up to large vocabulary image annotation,” in International Joint Conference on Artificial Intelligence, vol. 11, 2011, pp. 2764－2770.

[16] X. Lu, F. Wu, S. Tang, Z. Zhang, X. He, and Y. Zhuang, “A low rank structural large margin method for cross-modal ranking,” in Conference on Research and Development in Information Retrieval. ACM, 2013, pp. 433－442.

[17] T. Yao, T. Mei, and C.-W. Ngo, “Learning query and image similarities with ranking canonical correlation analysis,” in International Conference on Computer Vision, 2015, pp. 28－36.

[18] A. Frome, G.S.Corrado, J.Shlens, S.Bengio, J.Dean, T.Mikolov et al., “Devise: A deep visual-semantic embedding model,” in Advances in Neural Information Processing Systems, 2013, pp. 2121－2129.

[19] A. Karpathy, A.Joulin, and F.Li, “Deep fragment embeddings for bidirectional image sentence mapping,” in Advances in Neural Information Processing Systems, 2014, pp. 1889－1897.

[20] X. Jiang, F.Wu, X.Li, Z.Zhao, W.Lu, S.Tang, and Y.Zhuang, “Deep compositional cross-modal learning to rank via local-global alignment,” in International Conference on Multimedia. ACM, 2015, pp. 69－78.

[21] Y. Hua, H. Tian, A. Cai, and P. Shi, “Cross-modal correlation learning with deep convolutional architecture,” in Visual Communications and Image Processing, 2015, pp. 1－4.

[22] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, “Generalized multiview analysis: A discriminative latent space,” in Computer Vision and Pattern Recognition. IEEE, 2012, pp. 2160－2167.

[23] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal, “Cluster canonical correlation analysis,” in International Conference on Artificial Intelligence and Statistics, 2014, pp. 823－831.

[24] V. Ranjan, N. Rasiwasia, and C. V. Jawahar, “Multi-label cross-modal retrieval,” 2015, pp. 4094－4102.

[25] X.-Y. Jing, R.-M. Hu, Y.-P. Zhu, S.-S. Wu, C. Liang, and J.-Y. Yang, “Intra-view and inter-view supervised correlation analysis for multi-view feature learning,” in AAAI Conference on Artificial Intelligence,2014, pp. 1882－1889.

[26] D. Lin and X. Tang, “Inter-modality face recognition,” in European Conference on Computer Vision. Springer, 2006, pp. 13－26.

[27] K. Wang, R. He, W. Wang, L. Wang, and T. Tan, “Learning coupled feature spaces for cross-modal matching,” in International Conference on Computer Vision. IEEE, 2013, pp. 2088－2095.

[28] K. Wang, R. He, L. Wang, W. Wang, and T. Tan, “Joint feature selection and subspace learning for cross-modal retrieval,” IEEE Transactions on Pattern Analysis and Machine Intelligence, preprint.

[29] Y. T. Zhuang, Y. F. Wang, F. Wu, Y. Zhang, and W. M. Lu, “Supervised coupled dictionary learning with group structures for multi-modal retrieval,” in AAAI Conference on Artificial Intelligence, 2013.

[30] R. Liao, J. Zhu, and Z. Qin, “Nonparametric bayesian upstream supervised multi-modal topic models,” in International Conference on Web Search and Data Mining. ACM, 2014, pp. 493－502.

[31] Y. Wang, F. Wu, J. Song, X. Li, and Y. Zhuang, “Multi-modal mutual topic reinforce modeling for cross-media retrieval,” in International Conference on Multimedia. ACM, 2014, pp. 307－316.

[32] C. Wang, H. Yang, and C. Meinel, “Deep semantic mapping for cross-modal retrieval,” in International Conference on Tools with Artificial Intelligence, 2015, pp. 234－241.

[33] Z. Li, W. Lu, E. Bao, and W. Xing, “Learning a semantic space by deep

network for cross-media retrieval,” in International Conference on Distributed Multimedia Systems, 2015, pp. 199–203.

[34] L. Castrejon, Y. Aytar, C. Vondrick, H. Pirsiavash, and A. Torralba, “Learning aligned cross-modal representations from weakly aligneddata,” in Computer Vision and Pattern Recognition, 2016.

[35] W. Wang, X. Yang, B. C. Ooi, D. Zhang, and Y. Zhuang, “Effective deep learning-based multi-modal retrieval,” International Journal on Very Large Data Bases, vol. 25, no. 1, pp. 79–101, 2016.

[36] M. Rastegari, J. Choi, S. Fakhraei, D. Hal, and L. Davis, “Predictable dual-view hashing,” in International Conference on Machine Learning,2013, pp. 1328–1336.

[37] G. Ding, Y. Guo, and J. Zhou, “Collective matrix factorization hashing for multimodal data,” in Computer Vision and Pattern Recognition.IEEE, 2014, pp. 2083–2090.

[38] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen, “Inter-media hashing for large-scale retrieval from heterogeneous data sources,” in International Conference on Management of Data. ACM, 2013, pp.785–796.

[39] X. Zhu, Z. Huang, H. T. Shen, and X. Zhao, “Linear cross-modal hashing for efficient multimedia search,” in International Conference on Multimedia. ACM, 2013, pp. 143–152.

[40] W. Wang, B. C. Ooi, X. Yang, D. Zhang, and Y. Zhuang, “Effective multi-modal retrieval based on stacked auto-encoders,” in International Conference on Very Large Data Bases, 2014, pp. 649–660.

[41] D. Wang, P. Cui, M. Ou, and W. Zhu, “Learning compact hash codes for multimodal representations using orthogonal deep structure,” IEEE Transactions on Multimedia, vol. 17, no. 9, pp. 1404–1416, 2015.

[42] R. He, W.-S. Zheng, and B.-G. Hu, “Maximum correntropy criterion for robust face recognition,” Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 8, pp. 1561–1576, 2011.

[43] Y. Zhen and D.-Y. Yeung, “Co-regularized hashing for multimodal data,” in Advances in Neural Information Processing Systems, 2012,pp. 1376–1384.

[44] Y. Hu, Z. Jin, H. Ren, D. Cai, and X. He, “Iterative multi-view hashing for

cross media indexing,” in International Conference on Multimedia.ACM, 2014, pp. 527–536.

[45] M. Ou, P. Cui, F. Wang, J. Wang, W. Zhu, and S. Yang, “Comparing apples to oranges: a scalable solution with heterogeneous hashing,” in International Conference on Knowledge Discovery and Data Mining.ACM, 2013, pp. 230–238.

[46] B. Wu, Q. Yang, W. Zheng, Y. Wang, and J. Wang, “Quantized correlation hashing for fast cross-modal search,” in International Joint Conference on Artificial Intelligence, 2015, pp. 3946–3952.

[47] Y. Zhen and D.-Y. Yeung, “A probabilistic model for multimodal hash function learning,” in International Conference on Knowledge Discovery and Data Mining. ACM, 2012, pp. 940–948.

[48] D. Zhai, H. Chang, Y. Zhen, X. Liu, X. Chen, and W. Gao, “Parametric local multimodal hashing for cross-view similarity search,” in International Joint Conference on Artificial Intelligence. AAAI Press,2013, pp. 2754–2760.

[49] J. Masci, M. M. Bronstein, A. M. Bronstein, and J. Schmidhuber, “Multimodal similarity-preserving hashing,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 4, pp. 824–830, 2014.

[50] Y. Cao, M. Long, and J. Wang, “Correlation hashing network for efficient cross-modal retrieval,” CoRR, vol. abs/1602.06697, 2016.

[51] D. Zhang and W.-J. Li, “Large-scale supervised multimodal hashing with semantic correlation maximization,” in AAAI Conference on Artificial Intelligence, 2014, pp. 2177–2183.

[52] F. Wu, Z. Yu, Y. Yang, S. Tang, Y. Zhang, and Y. Zhuang, “Sparse multi-modal hashing,” IEEE Transactions on Multimedia, vol. 16, no. 2,pp. 427–439, 2014.

[53] Z. Yu, F. Wu, Y. Yang, Q. Tian, J. Luo, and Y. Zhuang, “Discriminative coupled dictionary hashing for fast cross-media retrieval,” in Conference on Research & Development in Information Retrieval. ACM,2014, pp. 395–404.

[54] Z. Lin, G. Ding, M. Hu, and J. Wang, “Semantics-preserving hashing for cross-view retrieval,” in Computer Vision and Pattern Recognition,2015, pp.

3864－3872.

[55] Y. Cao, M. Long, J. Wang, and H. Zhu, “Correlation autoencoder hashing for supervised cross-modal search,” in International Conference on Multimedia Retrieval, 2016.

[56] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang.Dual Encoding for Zero-Example Video Retrieval. CVPR 2019.

[57] PUTTHIVIDHY D，ATTIAS H T，NAGARAIAN S S．Topic regression multi-modal latent dirichlet allocation for image annotation[C] IEEE Conference on Computer Vision and Pattern Recognition，San Francisco，2010: 3408-3415．

[58] ZHENG Y，ZHANG Y J，LAROCHELLE H．Topic modeling of multimodal data: an autoregressive approach[C]．in IEEE Conference on Computer Vision and Pattern Recognition，Columbus，Ohio，2014:1370-1377．

[59] JIA Y，SALZMANN M，DARRELL T．Learning cross-modality similarity for multinomial data[C]．IEEE International Conference on Computer Vision，Barcelona，Spain，2011: 2407-2414．

[60] LIAO R，ZHU J，QIN Z．Nonparametric bayesian upstream supervised multi-modal topic models[C]．Proceedings of the international conference on web search and data mining．New York，USA，2014: 493-502．

[61] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang.Bottom-up and top-down attention for image captioning and visual question answering. In CVPR, 2018.

[62] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching.In ECCV, 2018.

[63] Kunpeng Li , Yulun Zhang , Kai Li , Yuanyuan Li and Yun Fu. Visual Semantic Reasoning for Image-Text Matching.In ICCV, 2019.

[64] Tsung-Yi Lin, Michael Maire, Serge Belongie, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In ECCV, 2014.

[65] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL, 2014.

[66] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio,Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In NIPS, 2013.

[67] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. arXiv, 2014.

哈尔滨工业大学学位论文原创性声明和使用权限

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《基于目标检测和图卷积的跨模态检索算法》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：苏林　　日期：2020年6月17日

学位论文使用权限

学位论文是研究生在哈尔滨工业大学攻读学位期间完成的成果，知识产权归属哈尔滨工业大学。学位论文的使用权限如下：

（1）学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；（2）学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）研究生毕业后发表与此学位论文研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。

作者签名：苏林　　日期：2020年6月17日

导师签名：　　日期：2020年6月17日

致 谢

衷心感谢导师邬向前教授对本人的精心指导。他的言传身教使我受益匪浅。邬老师对学术有着很深的理解，有着丰富的科研经验和心得，带我走入了科研的世界，让我感受到了科研的魅力。邬老师经常亲自询问我们的情况，帮助我们解决问题，给予我们指导，使得我们能够在研究中得到快速成长和提高。

在邬老师的指导和帮助下，我渡过了美好而充实的研究生生活。不仅科研能力得到了巨大提升，在生活能力上也得到了巨大成长。邬老师注重科研与运动，科研与文艺，科研与生活并重，使我们可以得到全面的发展与成长，拥有丰富的精神生活。丰富的活动让科研不再单调，让研究生活变得精彩。

感谢邬向前教授，以及实验室全体老师和同窗们的热情帮助和支持！

本课题承蒙国家重点研究与发展计划（2018YFC0832304）基金资助，特此致谢。