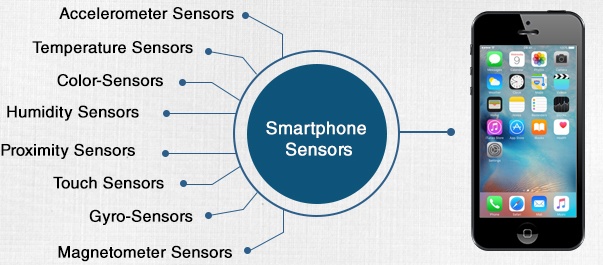
文本-图像多模态概述

1 多模态的概念

定义：一种信息的表现形式和来源可以被称为一种模态。在信息科学领域，信息的载体主要有文字、图像、语音、视频等，这里面的每种载体可被称之为多模态。

同时，除了狭义的定义外，我们可以将其推广到多种不同来源的相同载体，比如可以把不同传感器采集到的信号数据集作为多种模态，视频中的图像和声音分解成两路载体，亦可认定是两种模态[51]。



多模态学习的研究方向：

①多模态表示学习 Multimodal Representation

②模态转化 Translation

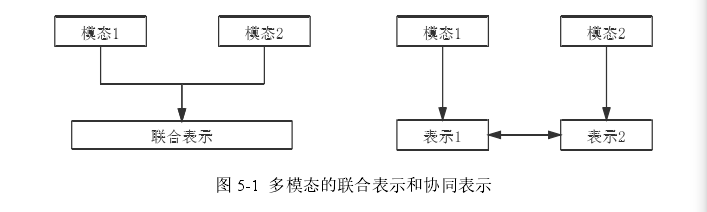
③对齐 Alignment

④多模态融合 Multimodal Fusion

⑤协同学习 Co-learning

* 1. 多模态表示

多模态机器学习（MMML）通过利用机器学习方法实现对多源模态信息的处理和理解[52]。单模态表示学习将信息表示为数值向量或者进一步抽象为更高层次特征向量，而多模态表示学习是指通过利用多模态间互补性，学习到更好的特征表示。主要包括两大研究方向：



（1）联合表示，将多个模态的信息整合，将它们映射到一个统一的多模态向量空间。

（2）协同表示，将多模态中的每个模态分别映射到各自的表示空间，但映射后的向量之间满足一定的相关性约束。这两种表示的简易结构如图 5-1 所示。

1.2 多模态融合

多模态融合的优点主要有三点：

①使得预测结果鲁棒性更强。因为它在观察同一现象时候引入了多种模式，预测趋向可靠；②接触到多种模式使得我们可以捕捉互补的信息，尤其是一些在单模态不可见的信息；

③当缺失其中一种模态时，多模态算法仍然可以运行。

多模态融合任务的难点主要集中在对各独立模态置信水平的判断、模态间的相关性的确定、特征信息的表征、特征信息的降维以及多源异步多模态数据配准研究。随着硬件提升，神经网络方法在多模态融合中逐步被研究人员使用，整体思想是利用联合隐藏层来融合信息。

1.3 多模态关联特征学习

2 如何实现文本-图像跨模态识别与检索

3 技术实现

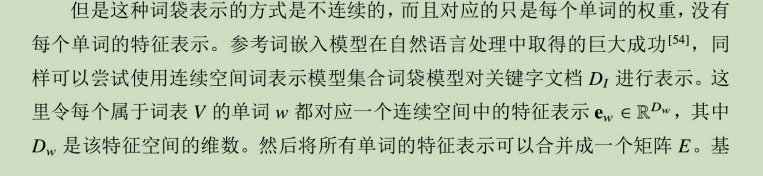
基于 Keras 的图像和文本信息融合分类算法实现（商品图像分类那篇论文笔记）

4 实验设计

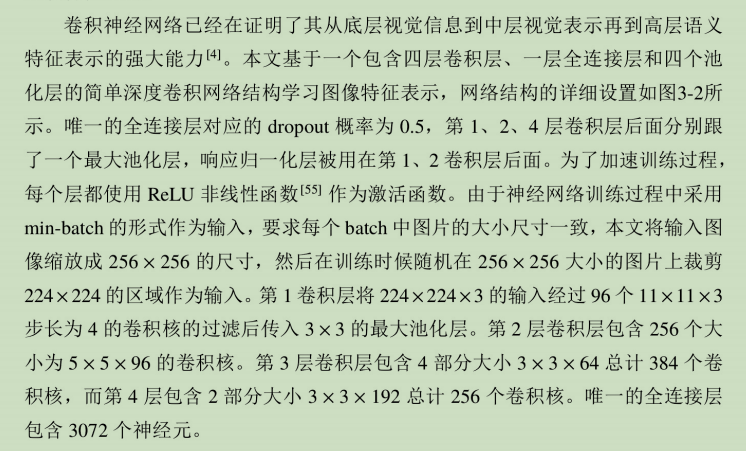
4.1 多模态特征学习

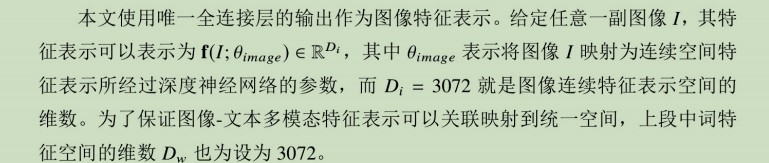
4.1.1 特征表示Network设计

文本特征表示



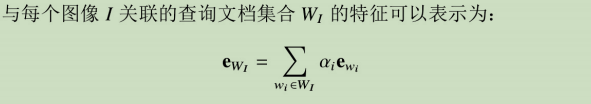
图像特征表示（提取）



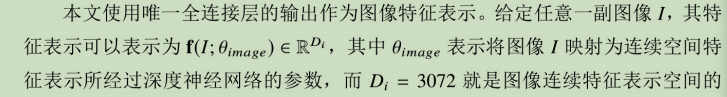


4.1.2 图像-查询文本对齐训练

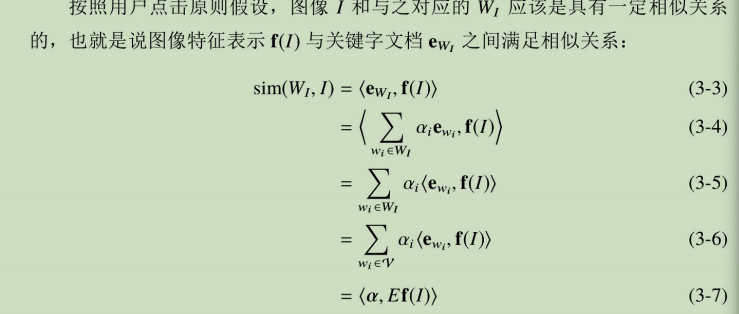
文本特征表示

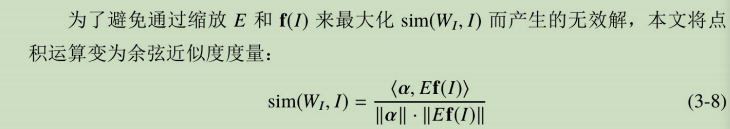


图像特征表示

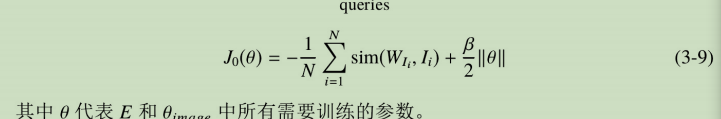


二者的相似关系

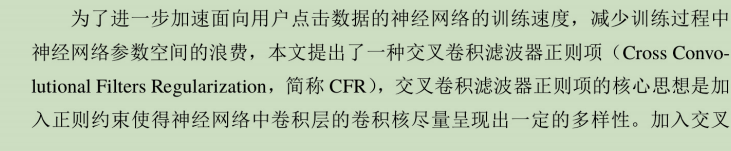


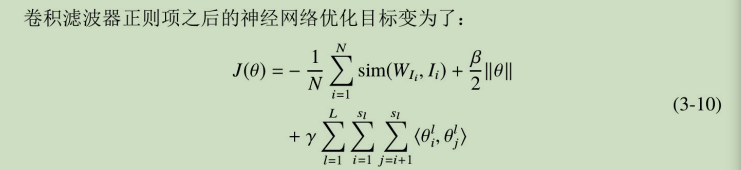


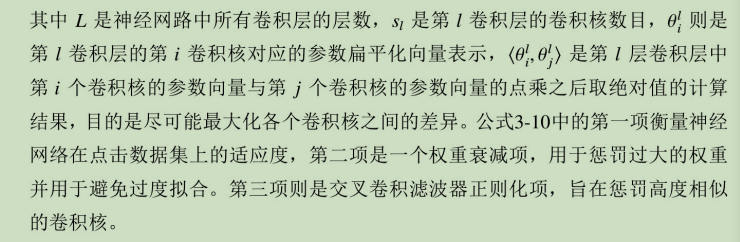
目标优化函数



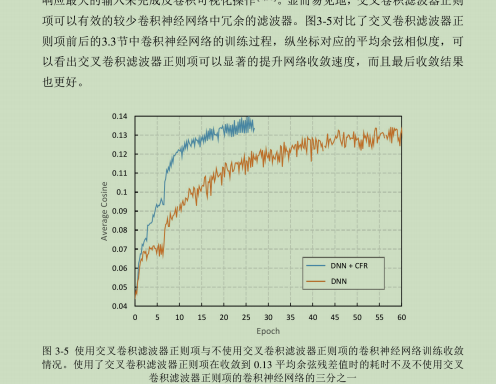
4.1.3 交叉卷积滤波器正则项



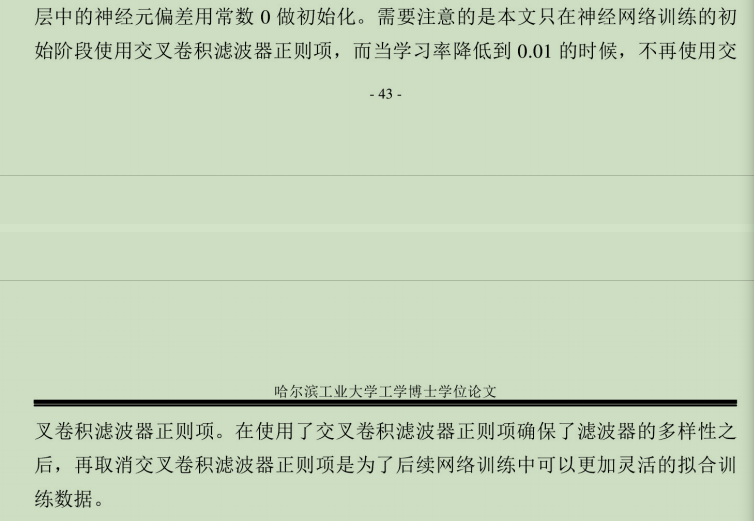




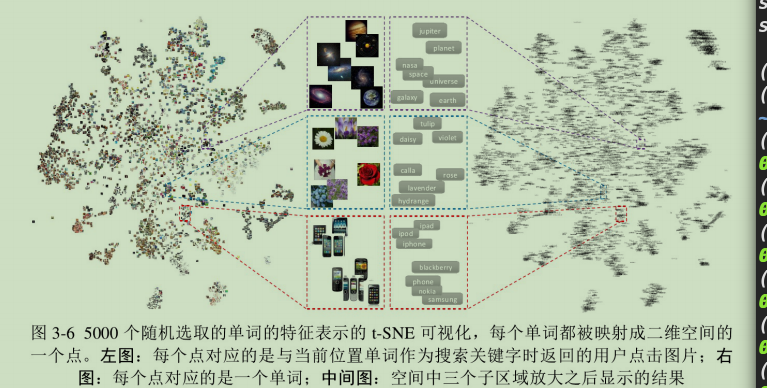
卷积滤波器效果图示

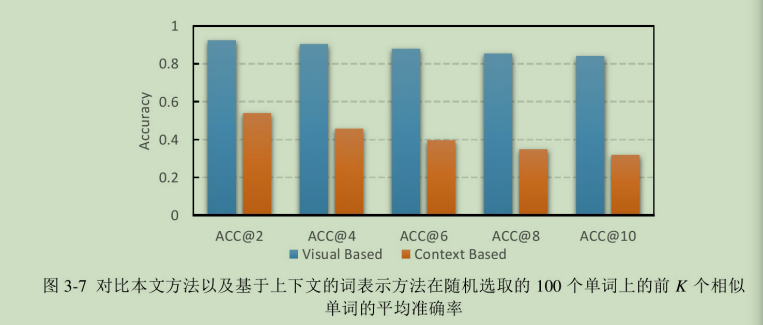


Notes:



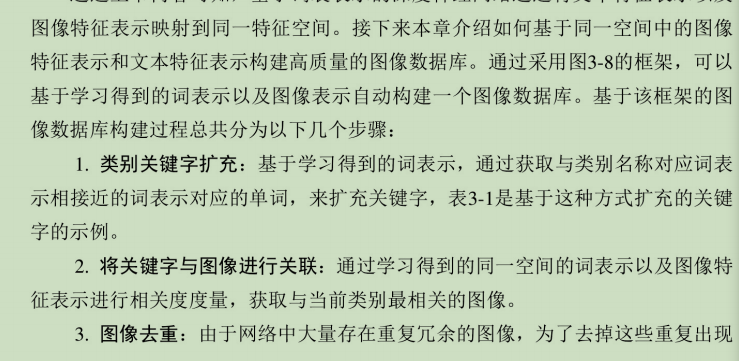
4.1.4 跨模态关联特征分析



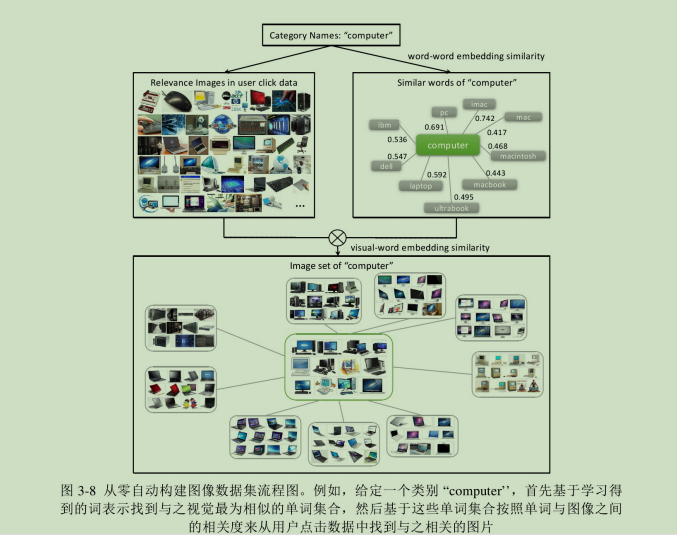


4.1.5 基于多模态关联特征的图像数据库自动构建

基于本章学到的文本与图像特征表示，来构建自己的图像数据库。其基本步骤如下：



主要是类别扩充与去噪两步；



5 基于多模态关联特征的文本-图像跨模态检索

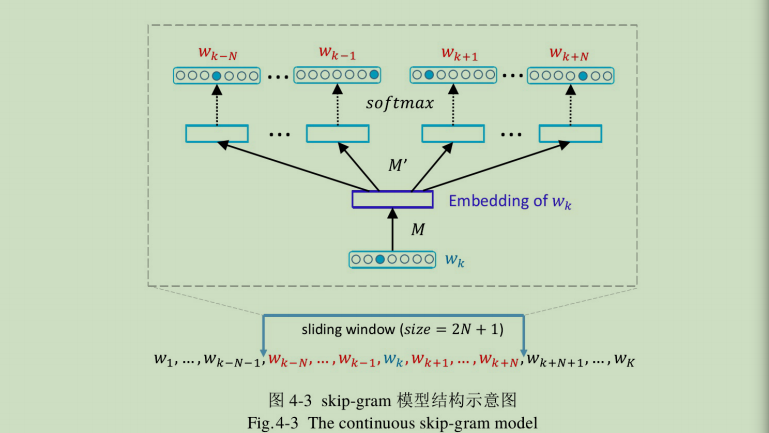
跨模态图像检索任务简介：

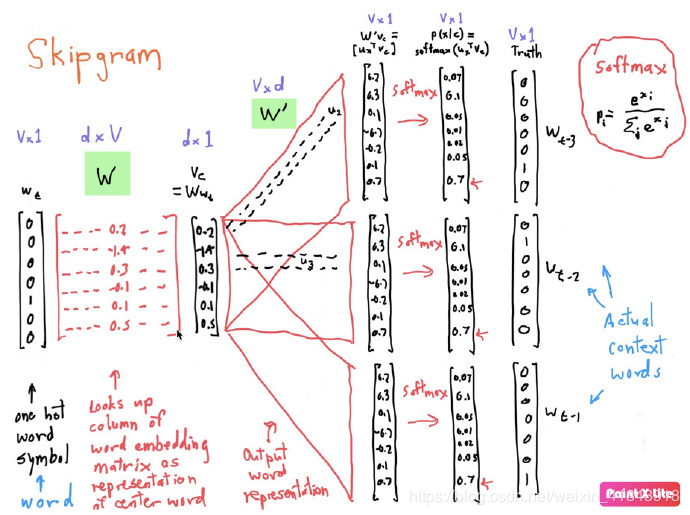
给定用户查询文本关键字，返回与之对应的图像

5.3 基于预训练特征表示的图像检索模型

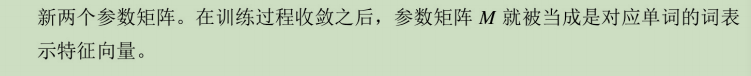
5.3.1 文本特征学习

Word2seq使用skip-gram算法，参考链接：<https://blog.csdn.net/weixin_41843918/article/details/90312339>

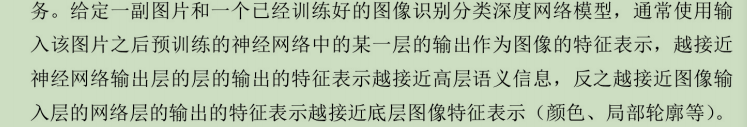




词表示特征向量

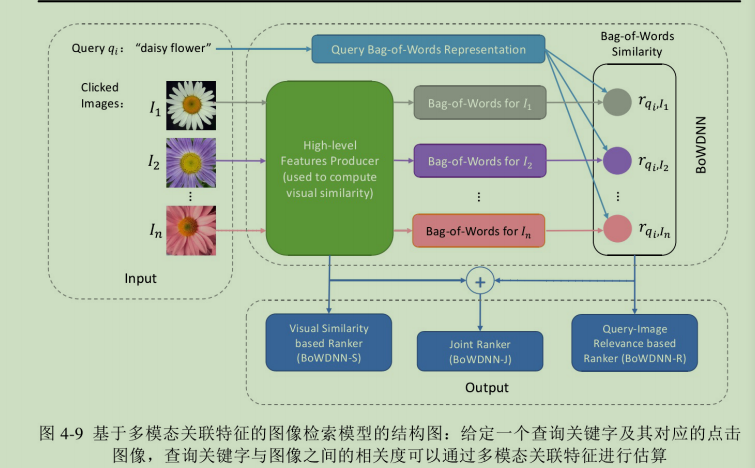


5.3.2 图像特征学习

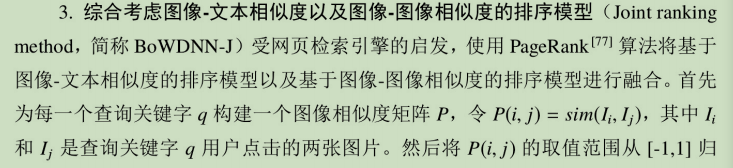


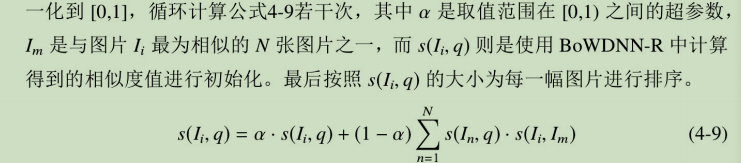
5.4 基于多模态关联特征的图像检索模型

基于多模态关联特征的图像检索模型有三种，三种检索模型是分别基于文本-图像相似度、图像-图像相似度、文本-图像&图像-图像相似度建立的。



经试验验证，综合考虑图像-文本相似度以及图像-图像相似度的排序模型（BoWDNN-J）性能最优。





三种排序模型的性能比较如下表：

