# 浏览式阅读

## 1 自己的总结、评价以及应用

本文依托深度学习技术，利用深度图像特征以及深度文本特征将图像与文本之间的语义鸿沟进一步缩小。本文从两个层面上对图像 文本多模态关联学习进行研究：面向图像 文本多模态关联学习的基础问题研究以及基于图像 文本多模态关联的相关应用研究。其中面向图像 文本多模态关联学习的基础问题研究包括图像 文本数据集关联、基于图像 文本相关性的特

征关联学习，而对于图像 文本多模态关联的相关应用，本文主要关注跨模态图像检索以及多模态视觉问答任务。

## 2 文章的主要问题（abstract、疑问句中）

## 3 结论（abstract以及conclusion中）

## 4 思路脉络（小标题中的关键句）

摘 要

但是多模态关联学习由于其输入数据不同模态之间的巨大差异往往被认为是一个很具有挑战的研究课题。而多模态关联学习的关键是在于为不同模态的输入数据建立相互关联关系。本文专注于包括多模态数据对齐以及多模态关联特征学习在内的关于图像 文本多模态关联学习两大基础研究课题，以及与图像 文本多模态关联学习最为相关的两大重要应用：跨模态图像检索以及多模态视觉问答。

首先，针对多模态数据集对齐任务提出了一套新的框架，

其次，提出了一种全新的基于深度卷积神经网络的图像 文本多模态关联特征学习的算法。

之后本文主要关注跟多模态关联学习相关的两大重要应用。对于跨模态图像检索，本文介绍了三种不同的框架。本文讨论的另外一个跟多模态关联学习相关的应用是多模态视觉问答任务。

关键词：多模态学习；图像识别；图像检索；视觉问答；深度卷积神经网络；深度

张量神经网络；注意力模型

1.2.2 多模态特征关联

多模态特征关联要求可以直接对不同模态下的异构特征之间建立联系。

① 等人的工作 中提到相同信息内容的不同形

态下的特征之间存在一定的内在关联，而通过引入典型性关联分析的方法可以将这种异构数据特征映射到一个公共的隐空间中去，在这个公共的隐空间中不同模态的输入数据之间可以直接计算相关度系数。

②另外 等人 还提出了一种在跨

模态特征关联映射过程中尽量保证原本模态的特征空间分布与映射后空间的特征空间分布保持一致的方法。

但是上述这些方法都是通过简单的线性变换的方式将两个不同模态的特征做关联映射，而实际情况中，简单的线性变换、特征值计算的方式很难处理不同模态特征复杂的分布情况。

1.2.3 跨模态图像检索

传统局部图像特征表示一般不包含语义信息，属于图像底层的特征表示，但是文本的特征表示一般都是对关键字的语义信息进行描述。这种底层图像特征与高层语义文本特征之间的差异导致了传统的跨模态图像检索的性能一般较差。

基于图像识别任务作为监督信息学习得到的深度图像特征一般都具有很强的语义信息。通过使用这种基于深度学习的高层图像特征，很大程度上缩短了图像特征表示与文本语义特征表示之间的差异。

1.2.4 多模态关系推理

多模态关系推理是一项更为复杂的任务，这项任务要求模态之间的特征不只是简单的一对一或者一对多的关联关系，而且要求不同模态之间的特征可以经过某些特定的关系表示建立关联。目前关于多模态之间关系的表示的研究工作较少。以视觉问答任务为例，目前的研究方向主要集中在多模态特征的融合，研究内容的重点是图像特征与句子特征的融合。

但是上述这些特征融合的方法都把多模态特征之间的相互关系忽略了，为了解决这一问题最近也有少数研究工作开始将研究的重心转移到多模态关系的建模上。

1.3 本文主要研究内容

1.3.1.1 多模态关联学习基础问题研究

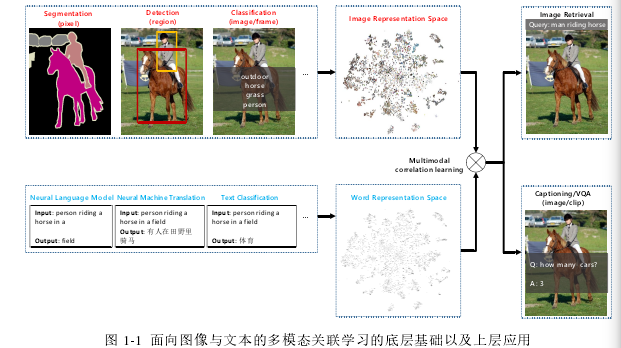
入分析了图像 文本多模态关联学习任务的特征：

图像信息与文本信息之间具有一定互补性。

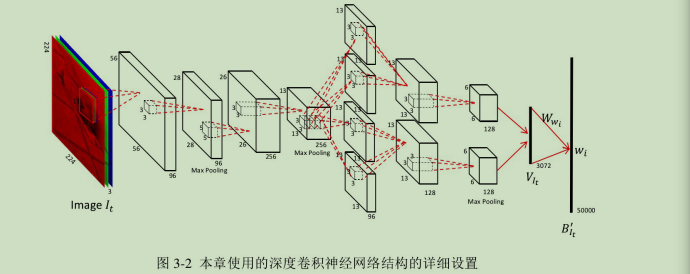
图像信息与文本信息之间具有一定相关性。

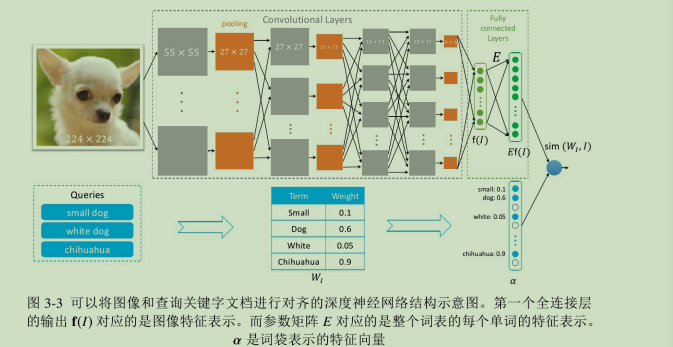
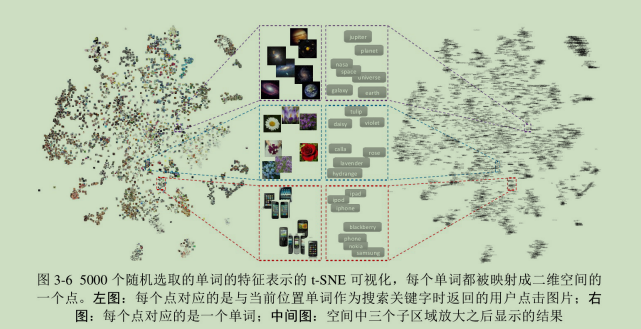
另外基于图像信息与文本信息之间具有一定相关性这个特点，本文提出了一种全新的神经网络结构，并且创新性地使用了搜索引擎用户点击数据作为弱监督数据，学习图像 文本的关联特征，得到一个可以同时表示图像信息和文本信息的公用特征空间，在该特征空间中，不但可以直接对图像 图像、文本 文本这种同模态的特征相关性进行计算，而且可以对图像 文本、文本 图像这种跨模态特征相关性进行直接计算。而且，不同于传统基于人工标注数据集上进行图像识别训练得到的图像特征表示，基于本文方法学习得到的图像特征表示和文本特征表示更加符合真实用户需求分布情况。

1.3.1.2 多模态关联学习相关应用研究



1. 基于图像与文本相关性的多模态关联特征学习



1. 
2. 

结 论

本文依托深度学习技术，利用深度图像特征以及深度文本特征将图像与文本之间的语义鸿沟进一步缩小。

本文从两个层面上对图像 文本多模态关联学习进行研究：面向图像 文本多模态关联学习的基础问题研究以及基于图像 文本多模态关联的相关应用研究。本文的主要研究成果总结如下：

1.提出了一种基于图像与文本信息互补性的从大规模互联网数据中进行图像 文本数据对齐的方法。（数据收集与数据集处理）

2.提出了一种全新的深度卷积神经网络模型，在公共连续特征空间中学习图像特征表示以及文本特征表示。

3.提出并对比了三种不同的跨模态图像检索模型：

4.针对多模态视觉问答任务，提出了一种全新的基于注意力模型的深度张量神经网络来对图像 问句 答案多模态三元组之间的关系以及推理过程进行建模，然后利用基于回归的训练方式将本文提出的方法应用在之前应用较广的视觉问答模型上去。

## 5 难理解点

# 梳理阅读

## 1 解决了什么问题

## 2 用了什么技术/方法

## 3 该技术/方法有什么优点