# 浏览式阅读

个人感觉这篇论文不是很适合作为本次课题的理解材料。例如，视觉问答任务与文本生成

部分，我们的项目要求为将image的语义信息提取出来，得到language/text，即image->language的转换。而本论文的视觉问答任务是对image->language进行了升级、复杂化（视觉问答任务不仅要将image中摘要文本提取出来，而且还要根据给出的限定question得出对应的答案），其重点不在于image->language的转换。

## 1 自己的总结、评价以及应用

应用：将image中包含的语义信息提取出来，实现image->language的转换。

## 2 文章的主要问题（abstract、疑问句中）

本文研究多模态数据的文本摘要生成，即基于原始输入信息为多模态数据，但是输出结果是基于原始信息的一段总结性文本。

本文的主要研究工作主要由三个部分组成：

（3）本文借鉴摘要的要素，即时间、地点、事件和主体等概念给出了 QABS

模型，该模型首先使用预训练的视觉问答模型对多模态信息中的视觉信息提取出

上述要素信息，之后通过将获取到的要素信息同原始多模态数据中的文本信息应

用注意力机制进行交互。

## 3 结论（abstract以及conclusion中）

重点概念有：Attention机制、LSTM、image->language转换等。

## 4 思路脉络（小标题中的关键句）

视觉问答

目前多模态摘要生成大多依据视觉问答等工作展开，一般使用分开的编码模型分别对文本信息和视觉信息作出理解，之后或者基于融合或交互等方式理解。

Attention机制

注意力机制可以被理解为一种加权机制，

总结来说，Attention 机制相当于是通过某个“外部”查询值 Q 通过对键值对

(K, V) 查询后，对我们感兴趣的值 V 作出加权表示，通过这样的加权表示可以在

不显著增加参数的情况下，记忆大量的原始信息————如相比于 LSTM 等 RNN

结构，是基于循环编码的机制，在足够长的某一个时刻 t 后将很难直接对初始的信

息 v1保留，而基于注意力的机制则可以在某个时刻仍然可以将 v1加以使用，只要对应的权重 α1相比其他时刻权重较大，这显然可以增加模型的记忆能力。

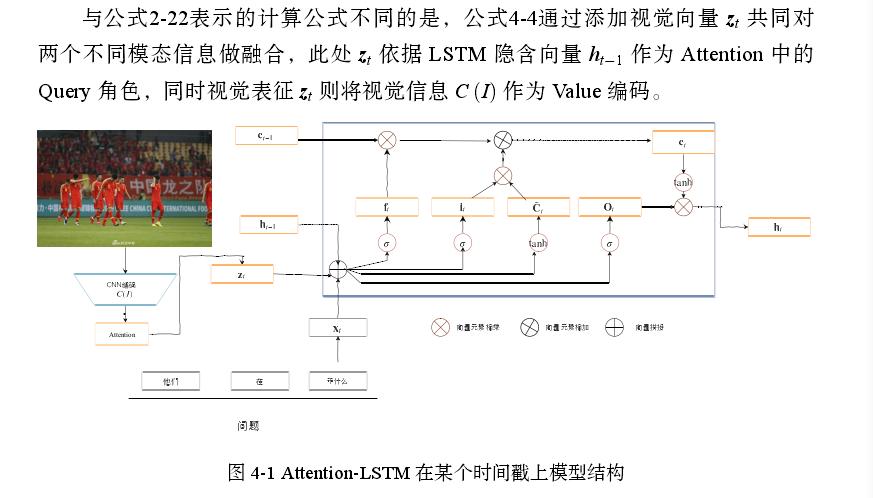
4.1 多模态信息融合

在多模态数据中，通常由于包含文本信息与视觉信息，如何利用这些信息作为模型的输入是一个重要的问题。本部分将介绍常用的视觉信息和文本信息交互的模型，这些模型通常使用注意力机制对两种或多种不同的模态信息进行融合。

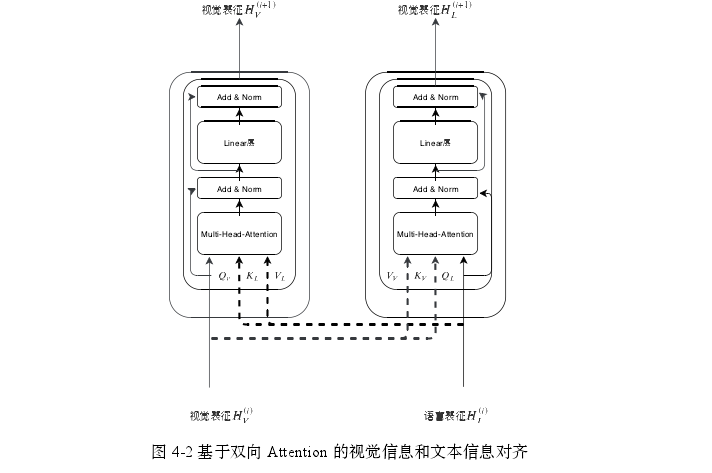
4.1.2 基于双流模型的模态融合方法

由于 CNN 和 RNN 等网络组件分别在图像和文本等模态应用广泛，所以大部分方法均分别对图像和文本使用 CNN 以及 RNN 等，之后使用融合机制将两种不同空间上的信息融合在一起。

在视觉问答问题中，一般文本通过 RNN 后得到的表示作为 Query，同时将 CNN 编码后的视觉特征作为 Value，并结合 LSTM 利用当前隐含状态 ht对视觉信息中的不同空间部位给予不同的关注，模型希望利用 RNN 信息在编码问题时就可以对图像中的不同部位作出关注，具体来说 Attention-LSTM 建立在 LSTM 基础之上。



上一节提到的基于文本查询视觉表征，或者相反的，基于视觉表征查询文本等都可以视为一种模态对另一种模态的注意力查询，这里主要对双向注意力做介绍，通过双向注意力将注意力机制同时应用到两个模态中，从而实现模态对齐。文献[40]通过使用物体检测模型对视觉信息作出编码，同时使用 Transformer对文本信息进行编码，在这之后对两种模态的表征又通过使用 Transformer 中的自注意力机制实现对信息的融合。



如图4-2中所述，传统的 Transformer 中使用自注意力模型，即 Q,K,V 均为同样维度的矩阵，模型通过引入共同注意力层实现对模态的融合，更改公式2-27中的 Q,K,V 分别对应如下：（此处省略公式）

该模型通过 Transformer 对文本和视觉特征都进行编码，并通过大规模的语料库进行自监督预训练学习，主要通过两个不同的任务来学习：（1）使用目标检测网络在获得图像的 ROI 区域后，借鉴 BERT 中有关掩码语言模型的思路，随即对视觉信息和文本信息以 20% 的概率掩码，通过使用相对熵作为损失函数用以重构被掩码的部分；（2）通过将文本、图像形成一个二元组，预测两种模态之间信息是否是对齐的，如果是对齐的，该种情况下模拟 BERT 中两种句子间的关系情况；该任务是一个分类任务。模型使用大量数据集进行自监督学习之后，在模型迁移时，虽然视觉信息和文本信息任然是两个独立的部分，但是两种模态已经使用注意力机制进行了融合，在下游迁移任务中可以独立使用。在下游任务如视觉问答（VQA），图像标注以及视觉常识推理（VCR）等获得较好的结果。

4.2 基于问答的多模态融合摘要生成模型改进研究

目前多模态数据在学术界主要按照如下的几个方向来研究，具体来说视觉任务中的如视觉问答系统（VQA-Visual Question Answering）[13]，图像标注（Image Caption）等任务将理解到的视觉信息用自然语言的形式表达出来，这些任务对图像理解提出了更高的要求，许多类型的多模态模型基于上述结合视觉和文本信息的任务作出预训练等。

4.2.1 视觉问答任务与文本生成

视觉问答任务通常给定一个三元组，如表示成(Image, Question, Answer) 等，通常对于 Image 采用 CNN 做表征，对于文本信息如给出的 Question 和 Answer 等则采用基于 RNN 等表征方法。



如图4-3中，给定任意一张图片和某个自然语言描述的问题，VQA 系统给出置信度最高的答案，图中 playing-soccer 是正确答案。鉴于观察到多模态新闻内容中，配图与内容的关联性一般较大，考虑通过图像信息得到上述的要素或通过图像信息得到上述要素相关的信息，可能有助于对原始摘要的理解。在 VQA 任务中，通过描述视觉问答我们理解视觉问答可以分为分类模型或者是生成模型，此处假设有多模态数据 I 包含文本数据 L 以及以图像存在的视觉信息 V:

I = {L = (w

1

, w

2

, · · · , w

T

) , V = (v

1

, v

2

, · · · , v

N

)} (4-10)

公式4-10中，I 代表某个的样本的表示，语言部分使用 L = (w

1

, w

2

, · · · , w

T

) 表

示，其中 w 表示单词；视觉部分使用 V = (v

1

, v

2

, · · · , v

N

)，为了简化数据研究的对象，此处的 v 代表属于这个样本的某张图像。则 VQA 任务希望通过使用某个网络，设为 f(v

i

|question

i

; Θ

VQA

) 给出上面所列举的关键要素的可能预测答案，设高置信度答案列表为 f(v

i

|question

i

; Θ

VQA

) =

{a

i1

, a

i2

, . . . , a

ik

}，为了确保上述答案序列满足高置信度结论要求答案类别满足如

下公式：

k

j=1

p(a

ij

) ≥ γ

confidence

(4-11)

公式4-11代表对给定的某个问题，由 VQA 模型获得的答案序列应概率和该满足大于等于给定的超参数 γconfidence，举例来说，如图4-3中，给定问题要素为：“这些人在干什么”，给出了 5 个高置信度答案分别为“运动”，“踢足球”，“足球”，“观看”，“棒球”，公式4-11中，对于不同的置信度有不同作用，降低 γconfidence会使得集合 {a

i1

, a

i2

, . . . , a

ik

} 更加可靠，而增加该阈值则会引入更多信息，但也可能引入更多噪声。