# 从描述到视觉概念的召回[[1]](#footnote-1)

赵淼译

(北京交通大学电子与信息工程学院，电子科学与技术，15650761201)

## 摘要

总结了各领域的代表性方法，并对其优势和局限性进行了探讨。在本文中，我们首先讨论了在早期工作中使用的方法，这些方法主要是基于检索和模板的。然后，我们将主要注意力集中在基于神经网络的方法上，这些方法给出了最先进的状态。基于神经网络的方法根据它们所使用的特定框架进一步划分为子类别。详细讨论了基于神经网络的每个子类别。在此之后，在基准数据集上比较了最先进方法的效果。在此基础上，对未来的研究方向进行了讨论。

## 1 介绍

Image Caption的难度在于一方面要获取图像中的语义信息，另一方面要以合理的形式表现出来。

获得图像中的目标，属性和目标间的关系，本身并不容易，组织成一句话就更为困难。

应用：比如生产自然的人类机器人互动，早期儿童教育，信息检索，以及视力受损的帮助。

早期的研究：基于模板的和基于检索的。这些基于已有的caption和硬编码的语言学架构不够灵活，表达性很受限。

## 2 基于检索的图像caption

一种方案是直接假设在数据库里存在合适的句子描述可以通过检索的方法，获得这种描述，例如度量检索图像和数据库图像的相关性获得排名靠前的图像的描述。

二是对检索到的句子进行重组获得新的句子。例如将数据库图像标注拆解成短语，利用短语获得描述。

缺点明显：难以对一些新的场景和目标的组合进行描述。

## 3 基于模板的方法

这是一个语义上的限制过程，需要一定的视觉概念和语法规则的优化过程。这里类似前面提到的生成式模型。检测好基础词汇然后套模板。

因为在基于模板的框架下的描述生成——工作被严格地限制在可视模型识别的图像内容中，通常只有很少的可视模型可用，通常对覆盖率、创造力和复杂的生成句子的限制是有限制的。此外，与人写的标题相比，使用僵硬的模板作为句子的主要结构会使生成的描述变得不那么自然。

## 4 基于深度学习的方法

这里将这类方法也进行了分类，分成了几个子类。

### 探索机制

由于人类探索的主观性，这里介绍了BLEU [123] , ROUGE-L [124] , METEOR [125] and CIDEr [126] . BLEU, ROUGE-L and METEOR的方法进行测试。这些探索机制多是通过机器翻译机制获得。

BLEU通过不同长度的短语的相似性获得候选句子和参考句子之间的近似性。最大到4，越高的分数说明产生的句子越流畅。

ROUGE-L用来探索机器翻译的流畅性和充分性。这个指标使用了一个候选句子和一组引用句之间最长的公共子序列，来测量它们在句子级别上的相似性。两个句子之间最长的公共子序列只需要序列的单词匹配，而匹配的单词不一定是连续的。因为这个度规自动包含了最长的序列的普通ngram，所以可以自然地捕捉到句子的层次结构。

METOR是自动机器翻译的评价指标。首先执行了候选句子和参考句子之间广义的单个一元组的匹配，然后根据匹配结果获得一个分数。指标包括准确率，召回率和匹配词的联合。所有独立计算的分数中最好的分数被作为候选人的最终评价结果。这个度量的引入是为了解决BLEU度量的弱点，它是基于匹配的ngram的精度而得到的。

CIDER是基于一种共识进行图像caption质量测量的方法。这种机制使用了各个参考句子的普遍规律，使用TFIDF作为权重对n元组进行加权进而评价语法性、显著性、重要性和准确性。

### 数据集比较

三种数据集，Flicker8k，Flicker30k，COCO。

三种数据集的数据集量不同，侧重场景也有所不同。1.主要包含人类和动物。2.人类每天的活动和事项3.复杂的自然场景下收集的，只有训练集和验证集。

1. 赵淼 电信研1706班 电子科学与技术 17120036 15650761201 [↑](#footnote-ref-1)