学士学位论文

高分遥感影像的图像描述生成

作者姓名：潘钊滢

指导教师：付琨

中国科学院空天信息创新研究院

学位类别：工学学士

专　　业：电子信息工程

学院（系）：中国科学院大学电子电气与通信工程学院

2021 年 6 月

Image Caption Generating of High-resolution Remote Sensing Images

A thesis submitted to

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Bachelor of Engineering

in Electronic Information Engineering

By

Pan Zhaoying

Supervisor: Professor Fu Kun

School of Electronic, Electrical, and Communication Engineering of

University of Chinese Academy of Sciences

June, 2021

中国科学院大学

学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。本人完全意识到本声明的法律结果由本人承担。

作者签名：

日　　期：　2021.05.30

中国科学院大学

学位论文授权使用声明

本人完全了解并同意遵守中国科学院大学有关保存和使用学位论文的规定，即中国科学院大学有权保留送交学位论文的副本，允许该论文被查阅，可以按照学术研究公开原则和保护知识产权的原则公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：　　导师签名：

日　　期：2021.05.30　　日　　期：　2021.06.01

摘 要

得益于遥感技术的发展，遥感影像的获取越来越便捷，其分辨率也不断提高。为了更高效地运用这些海量的遥感影像数据，新的图像处理方法需要被引入遥感影像的处理。对于遥感影像来说，其关键信息在于图像中的目标以及其相对位置关系。而目前对高分遥感影像的研究主要集中在局部的图像信息上，例如目标识别、语义分割等，这样处理得到的数据仍然由大量图像组成，并且只利用了遥感影像中很少一部分的图像信息。而深度学习中的图像描述方法可以同时反映图像目标及其位置关系。并且由图像信息生成的文本信息更简洁精炼，实现了大量遥感影像信息的提取。因此将图像描述生成模型应用于遥感影像是一个应用潜力巨大的方向。本文即从这个出发点展开研究和实验。

本文主要研究内容是将图像描述生成模型应用于高分遥感影像数据集，产生遥感图像描述，从而实现对高分遥感影像的快速理解与信息提取，提高对遥感影像数据的利用率。此外，本文使用了不同模型，并将其应用于不同的数据集上。通过定性与定量比较，得到实际应用效果较好的模型。

关键词：高分遥感图像；遥感影像；深度学习；图像描述生成

Abstract

Thanks to the development of remote sensing (RS) technology, the acquisition of remote sensing images has become more and more convenient, and the resolution of RS images has also been continuously improved. In order to use these massive amounts of remote sensing image data more efficiently, new image processing methods need to be introduced into remote sensing image processing. For remote sensing images, the key information is the targets in the image and their relative position relationship. The current research on high-resolution remote sensing images is mainly focused on local image information, such as object detection, semantic segmentation, etc., while the processed data is still composed of a large number of images, and only a small part of the image information in the RS image is used. The image caption method in deep learning can reflect the information of targets and their position relationship at the same time. Besides, the text information generated from the image information is more concise, and a large amount of remote sensing image information can be extracted. Therefore, applying the image caption generating model to RS images is such a direction with great application potential. This article starts research and experiments from this starting point.

The main research content of this thesis is to apply the image caption generating model to high-resolution RS image datasets to generate image descriptions, so as to realize the rapid understanding and information extraction of high-resolution RS images, which can greatly improve the utilization of RS image data. In addition, this thesis uses different models and applies them to different datasets. Through qualitative and quantitative analysis, models with a better practical application are obtained.

Key Words: High-resolution remote sensing images; Remote sensing images; Deep learning; Image caption

目 录

插图索引

表格索引

中英文对照表及部分缩略词

中文名词	英文名词	英文缩略词
图像描述	Image Caption	
卷积神经网络	Convolutional Neural Network	CNN
自注意力	Self-attention	
注意上的注意	Attention on Attention	AoA
神经网络	Neural Network	NN
前馈神经网络	Feed Forward Network	FFN
循环神经网络	Recurrent Network	RNN
强化神经网络	Reinforcement Network	
卷积层	Convolutional Layer	
特征图	Feature Map	
池化层	Pooling Layer	
最大池化	Max Pooling	
全连接层	Full-connected Layer	
残差网络	Deep Residual Network	ResNet
长短期记忆网络	Long Short Term Memory	LSTM
词向量	Word Vector	
独热编码	One-hot Decoding	One-hot
词嵌入	Word Embedding	
强化学习	Reinforcement Learning	RL

中文名词	英文名词	英文缩略词
策略梯度	Policy Gradient	
自批判序列训练	Self-critical Sequence Training	SCST
多头注意力	Multi-head Attention	
Sydney 图像描述数据集	Sydney-Captions Dataset	Sydney-Captions
UCM 数据集	UCM Dataset	UCM
UCM 图像描述数据集	UCM-Captions Dataset	UCM-Captions
RSICD 数据集	Remote Sensing Image Caption Dataset	RSICD
双语评估辅助工具	Bilingual Evaluation Understudy	BLEU
显式排序的翻译评估指标	Metric for Evaluation of Translation with Explicit Ordering, Alignment	METEOR
有序块	Chucks	
面向召回率的摘要评估辅助工具	Recall-Oriented Understudy for Gisting Evaluation	ROUGE
文本摘要	Text Summarization	
最长公共子序列	Longest Common Subsequence	LCS
F 度量	F-measure / F1 Score	
跳跃二元组	Skip Bigram	
基于共识的图像描述评估	Consensus-based Image Description Evaluation	

中文名词	英文名词	英文缩略词
词频逆文本频率指数	Term Frequency Inverse Document Frequency	TF-IDF
语义命题图像标题评估	Semantic Propositional Image Caption Evaluation	SPICE
概率上下文无关文法	Probabilistic Context-Free Grammars	PCFG
句法依赖树	Syntactic Dependencies Trees	
场景图	Scene Graphs	

符号对照表

符号	意义
Sigmoid(·)	Logistic 函数，S 型曲线
Softmax(·)	归一化指数函数
tanh(·)	双曲正切函数
ReLu(·)	线性整流函数
W	线性变换矩阵
b	偏移项
$[W_1, W_2]$	矩阵拼接符，将W_1, W_2拼为一个矩阵
$CNN(\cdot)$	卷积神经网络运算
$LSTM(\cdot)$	长短期记忆网络运算
$FFN(\cdot)$	前馈神经网络计算
f_{att}	注意力函数
f_{mh-att}	多头注意力函数
$LayerNorm(\cdot)$	层规范化
P	精度（Precision Rate）
R	召回率（Recall Rate）
$\otimes$	克罗内克积，用于元组匹配
$\odot$	哈达马乘积，逐元素相乘

第1章　绪论

1.1　选题背景和意义

遥感是一门对地观测的综合技术，通过人造卫星等平台上的探测仪器对地球表面实施探测。由于遥感探测能在短时间内对大范围地区进行高精度观测，因此遥感技术广泛应用于军事、测绘、环境监测等领域，成为了获取大范围地理信息的关键技术。为了使遥感技术得到的数据物尽其用，需要提高对这样的大量数据的处理速度，因此遥感影像的处理技术受到了人们的广泛关注。由于深度学习在自然图像的处理上取得了非常突出的成就，将自然图像的处理模型和方法迁移到遥感影像也变成了很自然的想法和研究方向。

遥感影像的图像描述生成，是指通过计算机视觉和自然语言处理结合的方法，让计算机使用人类语言将遥感影像的内容表达出来。理想的语言表达应该包含图像中的主要目标以及这些目标的属性和相对位置关系。随着图像描述生成的快速发展，将图像描述生成和遥感影像处理领域融合研究的遥感图像描述生成是一个潜力巨大的方向，其研究成果可用于遥感影像检索、军事情报辅助等领域。

我们可以将高分遥感影像分为高光谱分辨率影像和高空间分辨率影像。两者的区别如图 1.1[1]所示，

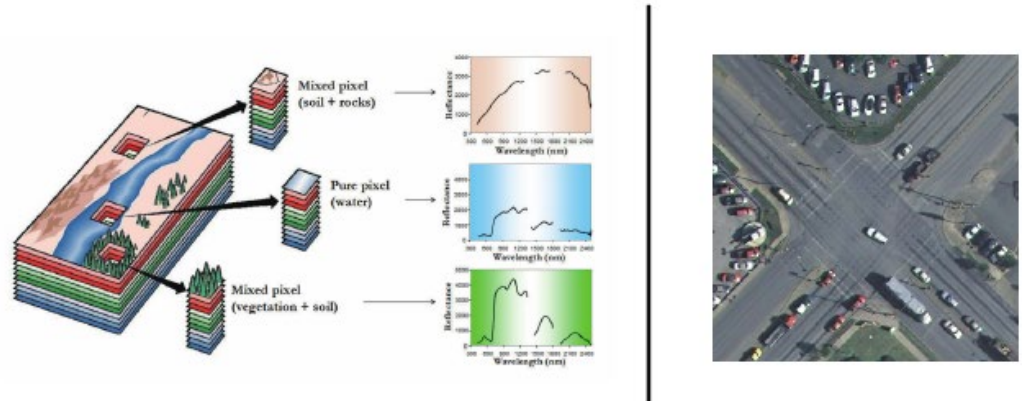


图 1.1 高光谱分辨率遥感影像和高空间分辨率遥感影像

Figure 1.1 High spectral resolution RS image and high spatial resolution RS image

高光谱分辨率的遥感影像同时包含了地物的地理坐标和光谱波段信息。由于地物光谱反射特征不同，它们各有一个近似连续的光谱曲线。在高光谱分辨率遥感的图像采集过程中，在每一个波段都获取不同物体在该波段上的反射率信息，

将得到的反射率信息综合起来就可以得到丰富的地表信息。获取到的影像可以看作是一种多维的数据，除了空间的两维，还有丰富的波段信息，每个波段都可以看作是一个维度。高光谱遥感影像具有光谱连续、图谱合一的特点，因此可以实现对地物信息的精确分类。

而高空间分辨率遥感指的是所得的遥感图像的空间分辨率高，这种遥感能以米级甚至亚米级的空间分辨率观测地球表面，获得的高空间分辨率遥感影像包含大量目标，其边缘信息、空间结构与表层纹理特征也十分清晰，易于分辨。在这两种高分遥感影像中，高空间分辨率遥感图像获取简单，处理方式简单，包含信息丰富，适合作为遥感图像描述研究的主体。

本文在高分辨率遥感影像上应用了四种自然图像的图像描述模型，涵盖了最为经典的 Show and tell 模型，引入注意力机制的经典模型 Show, attend and tell，解决长期依赖问题的 Transformer 模型，以及引入注意力机制最新结构的 Attention on Attention 模型，这些模型的具体网络结构将在第 3 章中详细讨论。

1.2 国内外本学科的研究现状

由于应用于自然图像的图像描述生成（Image caption）发展得比较成熟完善，对遥感图像描述生成具有指导作用，因此本文首先介绍在自然图像上生成图像描述的研究进展和现状，随后将介绍近年来在遥感图像数据集上的一些相关工作。自然图像描述任务早期的解决思路是利用图像处理算子提取图像特征，再经过支持向量机分类等技术，得到图像中各类目标物体，再根据这些目标利用条件随机场等规则生成对应的图像描述。这种思路的具体例子有 BabyTalk[2]，Every Picture Tells a Story[3]等。这种思路依赖于图像特征提取与语言生成的规则，得到的结果并不理想。

随后，RNN 和编解码器[4]在机器翻译领域取得了成功，其思路是使用编码器 RNN 处理翻译的源语言文本，从文本生成中间隐层变量，再使用解码器 RNN 读入中间隐层变量，并生成目标语言的翻译文本。来自 Google 团队的 *Vinyals et al.* 受机器翻译的启发，将编解码器的解决思路引入图像描述领域，将编码器改为 CNN 以实现对图像的编码，再使用特殊的 RNN——LSTM 读取图像特征向量并产生图像描述，这种做法即为图像描述生成领域经典的 Show and tell 模型[5]。同

期还有类似的思路，使用 VGG 和 RNN 完成图像描述生成的 NeuralTalk[6]，由于其数据集没有开源因此没有得到广泛应用。

这两篇论文提出的编解码框架为图像描述生成领域提供了一条开山之路，目前大多数方法都离不开编解码模型。但 Show and tell 模型也存在显著的缺点，其图像特征仅在 LSTM 处理的初始时间片输入，在生成长语句时，随着生成单词越来越多，LSTM 生成新单词与原图像的依赖性越来越低，同时随着网络层数增加，其梯度消失的可能性也大大增加。因此为了解决这些问题，研究者们针对 CNN 或 LSTM 做了一些改进。*Jia et al.*[7]提出了一种 LSTM 的变体，称为 gLSTM（Guiding LSTM），采用了三种不同的语义信息，用于指导不同时刻单词的生成，有利于长语句的生成[7]。此外还有 att-CNN 的提出，作者提出了高层语义信息的概念，并将其理解为多标签分类问题[8]。具体实现是使用 CNN 对图像的多个区域处理，以得到多标签的预测结果。再将多标签预测结果经过最大池化层，作为图像的高层语义信息再输入 LSTM。

此外，注意力机制的引入是随编解码器后的又一重要突破。注意力机制使得生成单词时可以动态地关注输入图像特征的不同区域。*Xu et al.*[9]在 Show, attend and tell 中提出了在卷积网络得到的图像特征中结合注意力机制的方法，引入了上下文信息的概念。在编码器阶段，抛弃了原有的使用全连接层提取图像特征的方法，图像特征换用较低的卷积层的输出，保留了图像的空间信息。而注意力机制可以通过加权来动态选择图像空间特征的区域。在解码器阶段，增加了上下文信息的向量作为输入，该向量是注意力机制作用下得到的区域的特征表达。注意力机制的引入有效解决了图像特征仅在开始阶段传入 LSTM 以及仅传入全局特征的问题。

但注意力机制也可能会引入一些不必要的信号，从而影响图像描述生成的结果。Knowing When to Look [10]中提出了基于哨兵机制的自适应的注意力方法，模型可以动态选择是根据语言模型还是图像显著区域来产生单词[10]。哨兵机制使用一个概率S作为哨兵，在生成单词时计算单词是否属于视觉词汇，即该单词与图像的关联性是否足够大，从而可以选择是否从图像产生单词，或者是根据语言模型的先验知识产生。

在这些方法中无一例外地使用了卷积神经网络作为基本单元，而随着距离的增加，这些模型关联两个任意输入或输出的信号所需要的操作数就越大，使得学

习远距离之间的关系变得困难。因此 Google 在 2017 年提出了用于机器翻译的 Transformer 模型，是注意力机制和自注意力（Self-attention）机制的结合，把任意两个位置的两个单词之间的距离转换为 1，有效地解决了长期依赖的问题。Transformer 是第一个完全依靠自注意力机制来计算其输入和输出表示的转导模型，无需使用序列对齐的 RNN 或卷积[11]。在自然语言处理领域中 Transformer 被首次提出，并应用于机器翻译任务，随后研究者们将其迁移到图像描述生成领域，也取得了不错的结果。

近些年同样出现了许多基于注意力机制做出提升的模型。*Huang et al.*提出了“注意上的注意”（Attention on Attention, AoA）模块[12]，将 AoA 模块应用于编码器和解码器，称为 AoA 网络。AoA 模块使用经典注意力的结果和当前的上下文信息生成信息向量和注意门，两者相乘后得到其关注向量。通过这样的机制，模型可以检测给定特征向量和注意力向量的相关性，从而提升了注意力机制的性能[13]。

随着图像描述领域的发展日渐成熟，深度学习中的强化学习也被引入其中。使用强化学习来训练图像描述生成模型，称为 Self-critical sequence training（SCST）方法[14]。该模型在每个单词生成时使用采样的方法来生成句子，再和基准比较，再使用策略梯度来更新模型，即抑制分数在基准下的句子生成，激励生成分数在基准上的句子。LSTM 通过贪婪解码（Greedy decoding）来生成一个基准句子，即在每次生成单词时，选择最终得到概率最大的单词；同时用到束搜索（Beam search），即每次都选取前N个概率最大的句子。

虽然上述方法在自然图像的描述生成中取得了成功，但遥感图像的描述生成仍然是一个富有挑战的领域，在数据集的构建和模型的迁移复现领域仍然有很大一片空白。*Qu et al.*首先提出了深度多模态神经网络模型[1]，通过 CNN 提取图像特征，并用 RNN 生成句子。此外 *Qu* 标注了两个已有的高分遥感图像数据集，使之成为图像描述生成任务可用的数据集。*Shi et al.* [15]同样提出了一个遥感图像描述框架，仍然是使用 CNN 提取图像特征，但使用预定义模板生成句子。

1.3 本文的主要研究内容

本文的研究内容是实现对高分遥感图像生成对应的图像描述，针对该任务，

本文完成了以下工作：利用 Pytorch 深度学习框架，在两个高分遥感图像数据集 UCM、RSICD 上实现了四种自然图像描述生成模型的迁移与复现；记录了不同模型在这两个遥感图像数据集上的评价指标，定量对比不同模型和不同数据集的训练结果；使用训练好的模型预测高分遥感图像，定性展示遥感图像描述生成的训练结果。

1.4 本文的章节安排

本文第一章介绍了遥感图像描述生成的选题背景，随后介绍自然图像描述生成的发展历程以及近些年在遥感图像数据集上的相关工作，最后介绍了本文任务和章节安排。

本文第二章介绍了深度学习领域的基本概念，以及图像描述生成领域涉及的重要结构，包括神经网络、卷积神经网络、循环神经网络、词向量四部分。

本文第三章介绍了图像描述生成领域的四种模型结构，包括 Show and tell 模型，Show, attend and tell 模型，Transformer 模型和 AoA 模型。

本文第四章介绍了遥感图像描述领域的三个数据集，以及图像描述领域的五种评价指标，随后基于其中的两个数据集开展算法复现的实验，用五种评价指标评价训练后得到的模型，并对模型的测试结果做了一些分析。

本文第五章总结了本文的研究工作，随后对该研究领域可能的改进方向做了讨论。

第2章　图像描述生成领域基本模型和重要概念

2.1　神经网络

神经网络（Neural Network, NN）是一种人工设计的模型，以模仿生物神经的结构和功能。神经网络由大量神经元模型互相连接得到网络模型，其神经元模型之间可以进行信号的传递和计算。神经网络可以基于外界信息改变内部连接和计算结构，即具有自适应性，相比于传统的基于规则的编程算法更灵活，因此被广泛应用于机器学习领域。

通俗来讲，神经网络是由多层神经元模型组成的复杂网络，神经元之间互相依靠权重运算连接，每一层神经元的值依靠于上一层神经元的输出，同时又成为下一层神经元的输入。受到外部信息的输入刺激时，神经网络会经由神经元之间的运算处理输入信息，根据神经网络输出与外部信息给出的参考输出，神经网络会更新内部神经元之间的参数，以调整输出使得与参考输出更相近。此过程也称为神经网络的训练。

一般来说，我们需要关注神经网络的结构、激励函数和学习规则。结构指定了其变量以及拓扑关系，例如最常见的变量是神经元对下一级的权重和各个神经元结构的激励值；激励函数则定义了神经元激励值如何计算得到。一般情况下神经元模型的激励值由另一些神经元的输出加权计算得到；而学习规则指定了神经网络在外部信息刺激下训练参数时，神经网络中的参数如何调整，例如使用评价指标计算参考输出和实际输出之间的差值，调整后的参数应使得该差值更小。其中针对不同网络，我们一般更关注其结构的设计。

根据网络架构分类，神经网络可以分为前馈神经网络、循环神经网络以及强化神经网络。

2.2　卷积神经网络

2.2.1　卷积神经网络概述

卷积神经网络（Convolutional Neural Network, CNN）是一种经典的前馈神经网络，由卷积层、池化层和全连接层组成。二维卷积的结构使得卷积神经网络可以对图像区域进行计算，卷积层和池化层则可以对大量图像数据进行降维操作，

减少数据量，经过降维后的数据可以在全连接层得到较为高效的处理。因此卷积神经网络在图像处理等领域表现出了其巨大的潜力。典型卷积神经网络如图 2.1 所示，

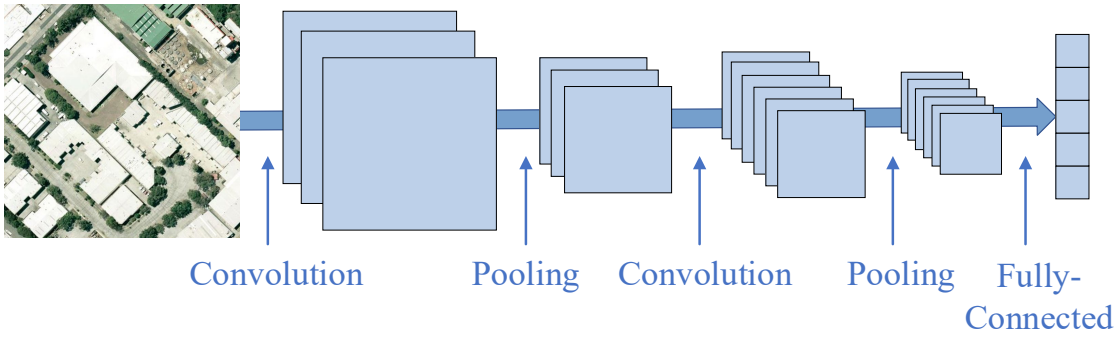


图 2.1 卷积神经网络结构

Figure 2.1 Structure of Convolutional Neural Network

第一层卷积层（Convolution layer）通过在图像上滑动不同的卷积核，将卷积核和图像各个区域进行卷积运算后得到多维的数据块。卷积核是一个$n \times n$的矩阵，n一般大于 1，且小于图像矩阵维度，卷积核在图像上滑动，每次计算得到当前覆盖区域与卷积核的卷积结果，也被称为当前区域的特征值。每个卷积核对输入图像做完运算后得到一张维数更小的数据矩阵，多个卷积核操作后，得到第一层卷积层输出即为一个多维数据块，每一个数据矩阵又被称为特征图（Feature map）。第一层之后的卷积层同样是类似的操作，将输入的数据矩阵与卷积核做卷积得到新的特征图。每一层到下一层之间神经元相互连接，通过与之相连的神经元的数值与对应权重，可以算出下一层的输入值。

卷积核的意义类似于图像模式匹配，如果某个区域的数据与卷积核相似度很高，则输出也会得到一个较高的特征值。使用多个卷积核，可以理解为使用多个图像模式匹配图像，从而提取出图像的局部特征。

池化层（Pooling layer）的操作是非线性形式的降采样，通过采样窗口的滑动来降低数据的维度。由于卷积核通常维数不高，卷积后得到的数据矩阵非常庞大，另外，降采样对图像特征和图像识别影响不大，因此使用池化层快速降低矩阵大小，大大地减少了运算量和参数数量，数据量的减少也在一定程度上抑制了过拟合。一般池化层采用最大池化（Max pooling），对每个池化窗口内的数据取最大值作为输出。给出一个最大池化的示例如图 2.2 所示：

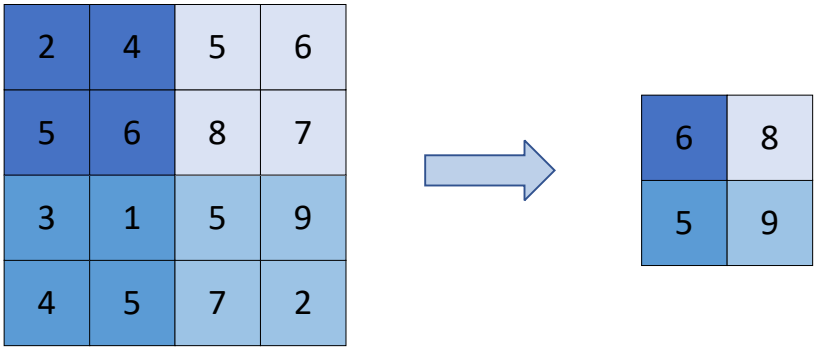


图 2.2 最大池化示意图

Figure 2.2 Example of Max Pooling

全连接层（Fully-connected layer）位于整个 CNN 的最后，经过卷积层和池化层降维后的数据被送入全连接层。全连接层与前一层的所有神经元都有连接，输入的特征图在全连接层经过运算后输出为一串预测向量。全连接层一般的运算操作是将输入特征图乘以一个矩阵再加上一个偏移量，矩阵和偏移量的参数和之前各层间的权重都是需要训练的对象，通过学习规则不断优化这些参数。

一般在全连接层后还有一个输出层，对结果进行类别划分，简单来说就是将神经网络处理得到的向量值映射到[0,1]的实数空间上。二分类问题一般使用Sigmoid分类器，多分类问题一般使用Softmax分类器。Sigmoid函数式如下所示，

$$\delta(z) = \frac{1}{1+e^{-z}} \qquad \ldots(2-1)$$

Softmax函数式如下所示，Softmax分类器与Sigmoid分类器相比，它可以把一个含任意实数的k维向量$\vec{z}$映射到一个k维实向量$\vec{\sigma}$，其中$\vec{\sigma}$每一个元素范围都在(0,1)之间，且所有元素之和为 1。

$$\sigma(\vec{z})_j = \frac{\exp z_j}{\sum_{k=1}^{K} \exp z_k}, for\ j = 1, \ldots, K. \qquad \ldots(2-2)$$

2.2.2　残差网络介绍

卷积神经网络的经典模型包括LeNet、AlexNet、VGG、GoogLeNet和ResNet，接下来以ResNet为例做简要介绍。残差网络（Deep residual network, ResNet）的提出是 CNN 发展历程中里程碑式的事件，在它之前，卷积神经网络的性能随着层数增加而提升，层数越多，网络可以提取的图像特征也越复杂。但层数增加的同时，神经网络反而出现了退化现象，梯度爆炸或者梯度弥散的问题变得棘手，

网络准确率反而出现下降，这种现象阻止了神经网络性能的进一步提升。而ResNet引入了残差学习，解决了深度学习中的模型退化问题[16]。

首先介绍残差单元，其主要思想是在网络中增加了直连通道，允许输入信息直接包含在输出结果中。残差网络单元结构如图 2.3 所示，

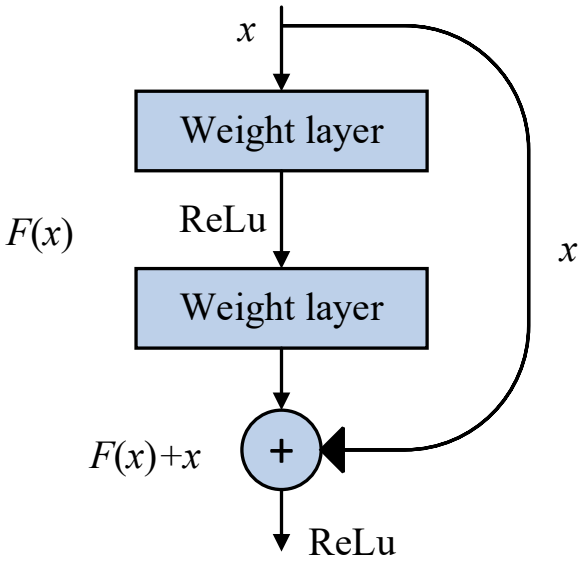


图 2.3 残差网络单元结构

Figure 2.3 Structure of ResNet Unit

令输入为x，一般卷积神经网络学习到的特征记为$H(x)$，由于残差网络将输入直接连到输出结果中，因此我们希望残差网络学习得到$F(x) = H(x) - x$，通过这种方法，残差单元无需学习完整的$H(x)$，而只需要学习上一级网络输出的残差$H(x) - x$，训练残差网络所需的运算量有所下降，且由于输入直连，信息丢失问题也得到了改善，因此残差网络的性能得到了很大提升。

ResNet网络中用到两种残差单元[16]，两种残差单元的结构如图 2.4 所示，分别由不同的卷积网络组成。

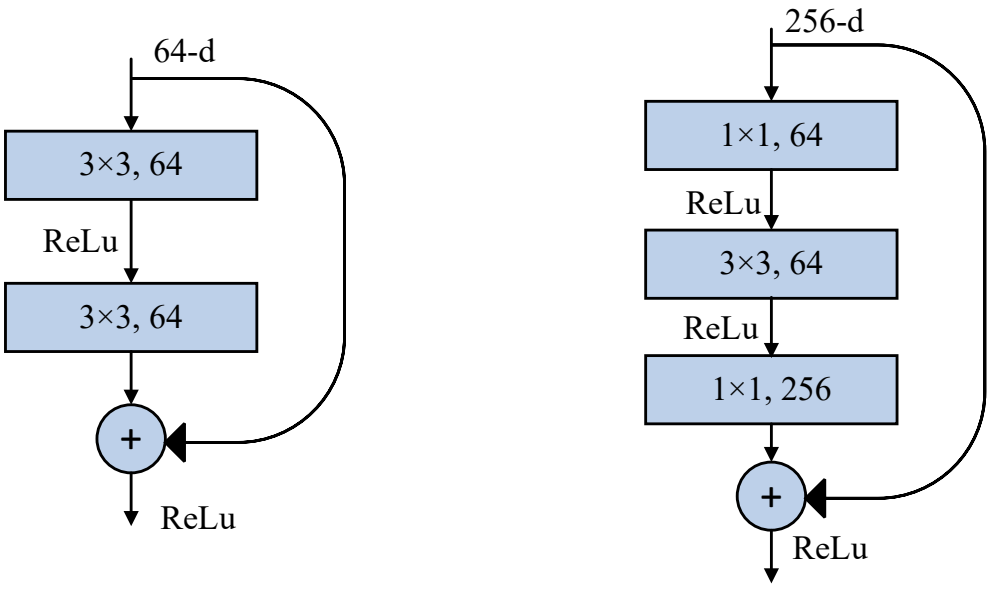


图 2.4 残差网络的两种典型单元结构

Figure 2.4 Structure of Two Typical Units of ResNet

ResNet网络根据其层数分为 18-layer，34-layer，50-layer，101-layer 和 152-layer[16]，这些不同层数的ResNet网络都是由上图两种残差单元组成的，图 2.5 展示了这些ResNet内部的残差单元配置。

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\3\times3,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\3\times3,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\3\times3,256\end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1\,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1\,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1\,512\end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times23$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1\,2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1\,2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1\,2048\end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				

图 2.5 不同层数残差网络的结构

Figure 2.5 Structure of ResNet with Different Numbers of Layer

2.3 循环神经网络

长短期记忆网络（Long Short Term Memory, LSTM）是一种具有长短期信息记忆能力的神经网络[17]，属于循环神经网络（Recurrent Neural Network, RNN）的

一种。循环神经网络主要与前馈神经网络（Feed Forward Network, FFN）相区分，RNN 在处理t时间片的数据时会将$t-1$时间片的隐节点同样作为当前的输入，即具有循环的结构，而 FFN 不具有这样的结构。因此在时序问题的处理上，循环神经网络有较为明显的优势。但循环神经网络中普遍存在长期依赖问题，随着时间增加，之前时间片的信息影响越来越小，因此循环神经网络在长时间信息的记忆上表现不佳。

随后 LSTM 的提出解决了长期依赖问题。传统 RNN 的输出取决于权重、偏置和激活函数，而 LSTM 引入了门的概念，用于不同时间片数据的控制，LSTM 中包含三个门，分别是遗忘门，输入门和输出门[17]，RNN 和 LSTM 的内部结构分别如图 2.6 和图 2.7 所示：

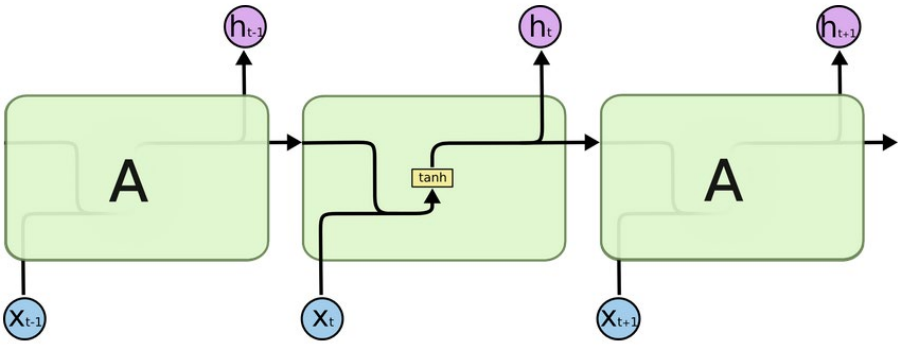


图 2.6 循环卷积网络结构

Figure 2.6 Structure of RNN

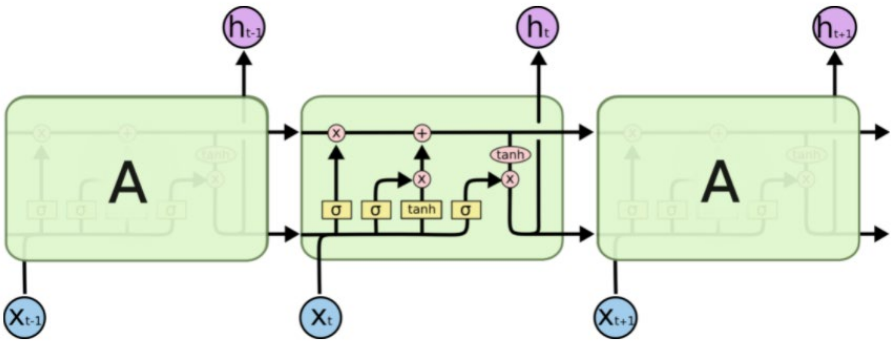


图 2.7 长短期记忆网络结构

Figure 2.7 Structure of LSTM

令时间片t时输入的数据记为x_t，上一层隐节点的输出记为h_{t-1}，通过图 2.7 左侧第一个Sigmoid激活函数后记作f_t，如下式计算：

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \quad \ldots(2-3)$$

f_t被称为遗忘门，是决定上一层单元状态中哪些信息应当被抛弃的变量。f_t是一个元素都在[0,1]范围内的向量，作为上一级单元状态C_{t-1}的权重，控制C_{t-1}中哪些部分参与计算当前单元状态C_t。其中$W_f[h_{t-1}, x_t]$意义是分别对h_{t-1}和x_t应用线性变换矩阵，由于展开写两项会增加一个矩阵参数，为了简并表达，用一个W_f代替，实际上该项可以写成$W_{f1}h_{t-1} + W_{f2}x_t$，以下计算式同理。

使用Sigmoid和tanh函数计算单元状态更新值$\widetilde{C_t}$和输入门i_t，其计算如下所示，

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad \ldots(2-4)$$

$$\widetilde{C_t} = \tanh(W_C[h_{t-1}, x_t] + b_C) \quad \ldots(2-5)$$

输入门i_t决定了单元状态更新值$\widetilde{C_t}$哪些部分参与当前单元状态C_t的计算，与遗忘门f_t一同作为加权值，根据上一层单元状态和这一层的更新值，得到当前层的单元状态：

$$C_t = f_t C_{t-1} + i_t \widetilde{C_t} \quad \ldots(2-6)$$

这一步在图 2.7 中表现为最上端的水平线，只涉及一些简单的加权运算，因此 LSTM 这种单元状态的概念使得信息保持变得更容易。

图 2.7 最右侧线路的Sigmoid和tanh函数计算得到当前单元状态C_t和输出门o_t，输出门o_t控制当前单元状态C_t在隐节点h_t中的输出，其计算如下所示，

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad \ldots(2-7)$$

$$h_t = o_t \cdot \tanh(C_t) \quad \ldots(2-8)$$

到此该层完成了数据的处理，输出单元状态C_t和隐节点h_t到下一层计算。LSTM 即通过这样的单元组合而成。

2.4 编解码器结构

图像描述生成任务通常使用编解码器作为模型的基本结构，其改进一般基于编解码器继续优化其内部结构。编码器通常使用卷积神经网络，将输入的图像转换为一个向量，即为图像的特征向量。解码器通常使用循环神经网络，特征向量送入解码器后，循环神经网络由图像特征向量逐个生成单词，以组成输出的图像描述。

训练时，首先采用预训练的编码器，其参数可以使得编码器较好地输出图像特征向量。固定预训练的编码器参数，对解码器进行训练，以此得到一个效果较好的解码器模型。随后不再固定编码器参数，编码器参数与解码器参数共同参与训练，这样可以提升模型效果。如果从一开始，解码器初始化后，预训练的编码器即参与训练，在解码器的训练过程中会使编码器效果下降，从而使得整个模型的效果打折扣[5]。

2.5 词向量

词向量（Word vector）是一种将文本的词语序列映射成实数向量的方法。由于文本形式的词语是一种非结构化的信息，无法直接参与计算，因此需要将文本词语转换为词向量，也就是用实数将文本词语编码。在深度学习过程中，所有词汇构成词表，以便于模型对文本信息的使用。独热编码（One-hot encoding）是基于规则的自然语言处理研究中常用的处理方法，将每个词编码为一个$1 \times N$的长向量，N是总词表的大小。其中在这个长向量中，每一位都代表了词表中的一个词汇，编码词汇时代表这个词汇的数字置 1，其他位的数字均置 0，以此区分不同的词汇。One-hot 编码虽然简单，计算量小，但不能反映词汇之间的关系，而且由于数据稀疏，在高维情况下距离计算将会变得非常困难。

词嵌入（word embedding）相比于 One-hot 编码更进一步，它将词汇表示为一个特征向量，向量中的值分别代表了这个词汇与不同特征的相关程度。根据特征向量我们可以得知两个词汇的相似程度，例如表 2.1 中 Apple 和 Orange 在几个特征上的数值相近，因此可以看作是一类词汇。

表 2.1 词嵌入示例

Table 2.1 Example of Word Embedding

	Man	Woman	King	Queen	Apple	Orange
Gender	-1.00	1.00	-0.95	0.97	0.00	0.01
Royal	0.01	0.02	0.93	0.95	-0.01	0.00
Age	0.03	0.02	0.70	0.69	0.03	-0.02
Food	0.09	0.01	0.02	0.01	0.95	0.97

词嵌入的具体实现方法有 Word2Vec 模型[18]、GloVe（Global Vectors for Word Representation）[19]模型，为了得到表现较好的词嵌入模型，需要先在文本数据集上进行大量训练，实际使用中可以直接从网络上下载预训练好的模型，应用于各种需要对词汇编码的任务。

2.6 本章小结

第二章介绍了图像描述生成领域基本模型和重要概念。神经网络是利用深度学习开展图像描述生成任务的基础，而神经网络中的 CNN 和 LSTM 是图像描述生成的基石，后续的模型介绍将以这两个网络为基础而展开。词向量是涉及到自然语言的深度学习任务中必不可少的一个环节，同样也广泛应用于各种图像描述模型中。

第3章　图像描述生成模型结构

3.1　概览

基于第 2 章介绍的基本概念，本章进一步介绍了一些图像描述生成的模型结构。这些模型已经在自然图像的描述生成上取得了较好的成果，对遥感图像的描述生成具有指导作用。后续将采用这些模型在遥感图像数据集上开展实验和研究。

本章以四个模型为例介绍其组成结构，分别是 Show and tell 模型[5]、Show, attend and tell 模型[9]、Transformer 模型[11]、AoA 模型[12]。

3.2　Show and tell 模型

受机器翻译的编解码器结构启发，谷歌团队在 2014 年提出了将编解码器应用于图像描述生成领域的方法，命名为 Show and tell 模型[5]。该模型使用 CNN 作为编码器，将输入图片的特征提取出来，通过词嵌入模型产生一个固定长度的向量作为图像特征，又使用 LSTM 作为解码器，利用图像特征生成描述性的语句。训练模型时，令输入的图片记为I，其真实的句子描述使用 One-hot 编码为$S = (S_0, \ldots, S_N)$，其中S_0, S_N 分别表示句子开始和结束的特殊字符，其训练的处理过程如图 3.1 所示，

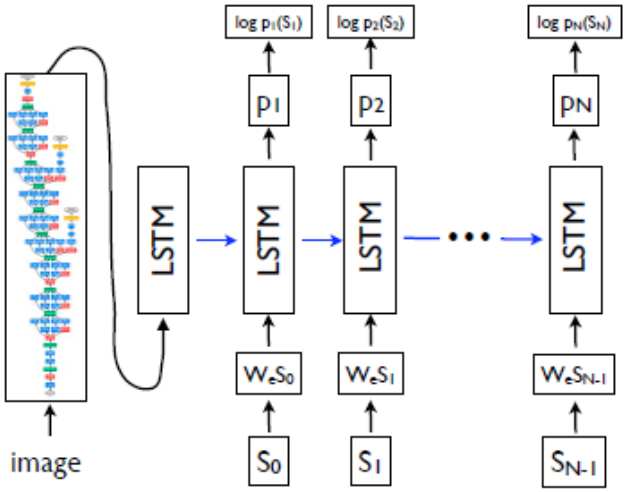


图 3. 1 Show and tell 模型结构

Figure 3.1 The Structure of Show and tell Model

$$x_{-1} = CNN(I) \qquad \ldots(3-1)$$

$$x_t = W_e S_t, t \in \{0, \dots, N-1\} \quad \dots(3-2)$$

$$p_{t+1} = LSTM(x_t), t \in \{0, \dots, N-1\} \quad \dots(3-3)$$

在训练过程中，图像和真实描述标签分别通过 CNN 和词嵌入编码矩阵W_e映射到相同的空间，得到图像的特征向量和真实描述的词向量。这两者共同输入 LSTM 以训练 LSTM 的模型参数，区别是图像特征向量x_{-1}仅在$t=-1$时输入 LSTM，而x_t将在每个时间片依次输入 LSTM。所有 LSTM 单元在同一时间片共享同样的参数，并逐个输出概率向量p_{t+1}，根据p_{t+1}的值可以在词表中得到最大概率所对应的单词，单词序列即组成输出的图像描述。CNN 和 LSTM 具体计算细节可见 2.2 节、2.3 节，此处不再赘述。

3.3 Show, attend and tell 模型

Show, attend and tell 模型顾名思义是 Show and tell 模型的改进版本，引入了注意力机制从而提高了模型的表现。Show, attend and tell[9]仍然沿用了编解码器结构，其编码器仍然选择 CNN，解码器选用 LSTM，此外加入了注意力机制，其处理流程如图 3.2 所示

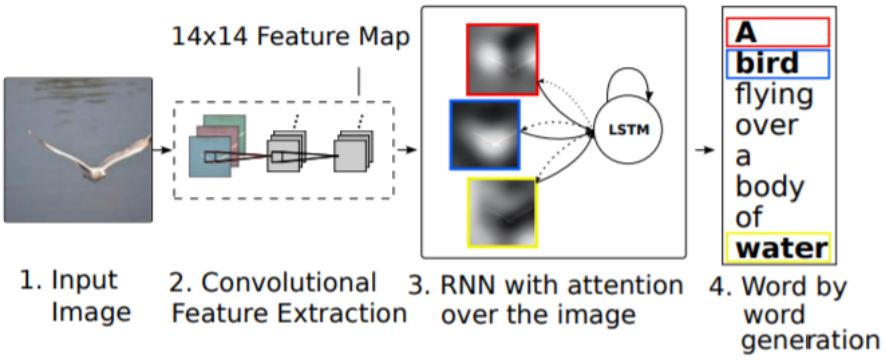


图 3.2 Show, attend and tell 模型处理流程

Figure 3.2 The procedure of Show, attend and tell Model

在编码器 CNN 中，使用了较低层网络输出的特征以更好地描述图像的局部特征，将图像的特征向量记为$\vec{a}$。

$$\vec{a} = CNN(I) \quad \dots(3-4)$$

图像的特征向量$\vec{a} = \{a_1, \dots, a_L\}$，经过注意力模块后生成上下文向量$\vec{z} = \{z_1, \dots, z_C\}$，其中$C$是描述语句长度。由于上下文向量是依次生成的，因此用z_t表

示t时间片的上下文向量，计算如下所示

$$z_t = \sum_{i=1}^{L} \alpha_{t,i} a_i \quad \ldots(3-5)$$

α_t是图像区域的权重，表示上下文向量应当对图像的哪部分区域着重注意，$\alpha_{t,i}$由前一时间片的隐变量h_{t-1}计算得到，首先使用注意力打分函数f计算每个区域的得分$e_{t,i}$，再通过指数函数计算不同区域的加权值$\alpha_{t,i}$，注意力打分函数f可以有不同的选择。

$$e_{t,i} = f(a_i, h_{t-1}) \quad \ldots(3-6)$$

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{k=1}^{L} \exp(e_{t,k})} \quad \ldots(3-7)$$

得到上下文向量z_t后，将其与上一时间片隐变量h_{t-1}、上一时间片输出的单词y_{t-1}输入 LSTM 网络，经过计算得到当前时间片隐变量h_t

$$f_t = \sigma\left(W_f[y_{t-1}, h_{t-1}, z_t] + b_f\right) \quad \ldots(3-8)$$

$$i_t = \sigma(W_i[y_{t-1}, h_{t-1}, z_t] + b_i) \quad \ldots(3-9)$$

$$\widetilde{C_t} = \tanh(W_C[y_{t-1}, h_{t-1}, z_t] + b_C) \quad \ldots(3-10)$$

$$o_t = \sigma(W_o[y_{t-1}, h_{t-1}, z_t] + b_o) \quad \ldots(3-11)$$

$$C_t = f_t C_{t-1} + i_t \widetilde{C_t} \quad \ldots(3-12)$$

$$h_t = o_t \cdot \tanh(C_t) \quad \ldots(3-13)$$

LSTM 具体计算细节可见 2.3 节。最后得到的h_t通过全连接网络后生成当前时间片对应的单词y_t，所有时间片运算结束后得到图像描述$\vec{y} = \{y_1, \ldots, y_C\}$。

3.4　Transformer 模型

在这些方法中无一例外地使用了 CNN 作为基本单元之一，而随着距离的增加，关联两个任意输入或输出的信号所需要的操作数就越大，使得学习远距离之间的关系变得困难。因此 Google 在 2017 年提出了用于机器翻译的 Transformer 模型[11]，是注意力机制和自注意力机制的结合，把任意两个位置的两个单词之间的距离转换为 1，有效地解决了长期依赖的问题。Transformer 是第一个完全依靠自注意力机制来计算其输入和输出表示的转导模型，无需使用序列对齐的 RNN

或卷积。

Transformer 由编解码器结构组成，编码部分和解码部分均由一系列更小的编解码器组成，其个数以 6 为例，内部结构如图 3.3 所示，

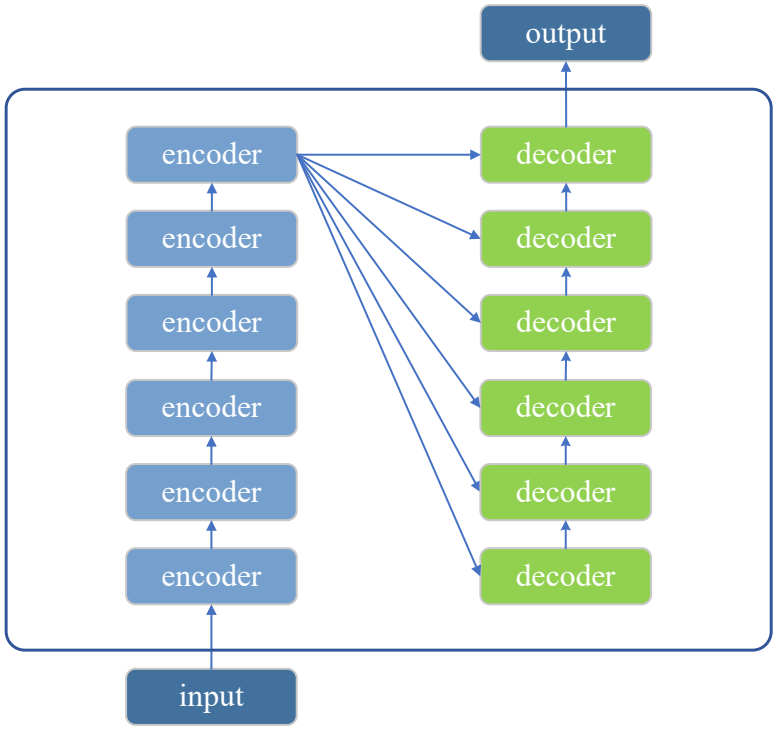


图 3.3 Transformer 网络结构

Figure 3.3 Structure of Transformer Network

所有的编码器结构相同，但权重不会共享。每个编码器由自注意力模块和前馈神经网络模块组成，其组成结构如图 3.4 所示，

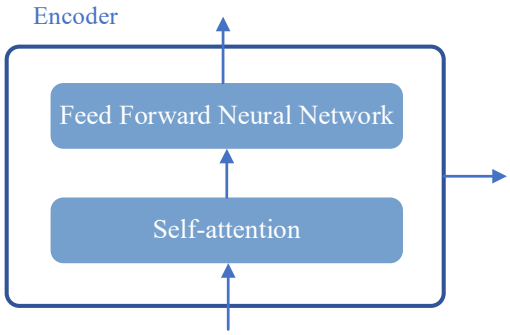


图 3.4 Transformer 编码器结构

Figure 3.4 Structure of Encoder in Transformer

解码器由自注意力模块、注意力模块和前馈神经网络模块组成，相比于编码器，解码器在两层之间多了注意力模块。解码器如图 3.5 所示

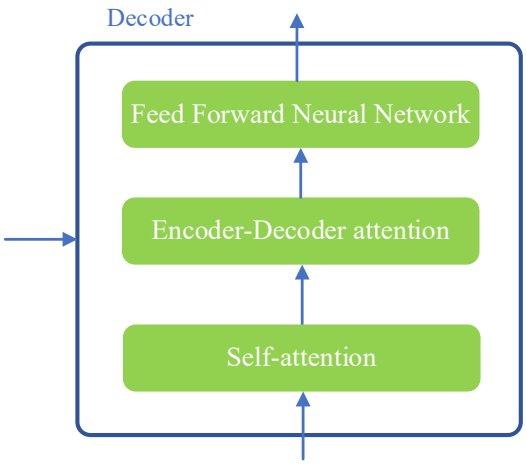


图 3.5 Transformer 解码器结构

Figure 3.5 Structure of Decoder in Transformer

令输入的图像特征向量为$\vec{a} = \{a_1, \dots, a_L\}$，经过自注意力模块后变成$\vec{z} = \{z_1, \dots, z_C\}$，自注意力模块即不使用其他模块的隐变量，$\vec{Q}$、$\vec{K}$、$\vec{V}$均来自于模块输入$\vec{a}$。

$$\vec{Q} = W^Q \vec{a} \quad \dots(3-14)$$

$$\vec{K} = W^K \vec{a} \quad \dots(3-15)$$

$$\vec{V} = W^V \vec{a} \quad \dots(3-16)$$

$$\vec{z} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\vec{V} \quad \dots(3-17)$$

其中d_k是$\vec{K}$矩阵的维度，将计算结果除以d_k是为了使模型更稳定，Softmax将会使计算权重标准化，各项均为整数且和为 1。

如果使用多头注意力（Multi-head attention）模块，那么在训练中会使用多组W_i^Q、W_i^K、W_i^V，得到多组$\vec{z}_i$。将这些$\vec{z}$矩阵拼接，再乘以权重矩阵W^O得到最后的矩阵。

随后$\vec{z}$输入前馈神经网络得到小编码器的输出$\vec{r}$

$$\vec{r} = FFN(\vec{z}) \quad \dots(3-18)$$

为了避免梯度消失的问题，Transformer 模型使用了残差网络的结构，前馈神经网络的输入包括自注意力模块的输出$\vec{z}$，也包括原始输入$\vec{a}$。前馈神经网络的输出$\vec{r}$，与输入$\vec{a}$维度相同，因此这些小编解码器可以堆积使用。最后输出的向量经过全连接层得到概率分布向量，根据词表即可得到当前生成的单词。

3.5 AoA 模型

如果我们将注意力模块打包成一个函数$f_{att}(\vec{Q},\vec{K},)$，其中$\vec{Q}$（Queries）代表查询队列，$\vec{K}$（Keys）和$\vec{V}$（Values）的每一项可以组成键值对$k_i, v_i$，注意力模块首先计算$\vec{Q}$和$\vec{K}$之间的相似度，然后计算$\vec{V}$上的加权得分得到$\hat{V}$，其中$\hat{v_i}$即为$q_i$的关注向量。计算式如下所示，

$$e_{i,j} = f(q_i, k_j) \quad \ldots(3-19)$$

$$\alpha_{i,j} = \frac{\exp(e_{i,j})}{\sum_k \exp(e_{i,k})} \quad \ldots(3-20)$$

$$\hat{v_i} = \sum_j \alpha_{i,j} v_i \quad \ldots(3-21)$$

在 Show, attend and tell 模型中，注意力机制模块的查询队列$\vec{Q}$和$\vec{V}$即为图像特征向量，$\vec{K}$即为隐变量序列，最后输出的是加权后得到的上下文向量。

经典的注意力模块存在一个无法忽视的问题，即当$\vec{Q}$和$\vec{K},\vec{V}$无关时，仍然会生成加权向量，这样产生的加权向量可能会含有误导信息。为了改进注意力模块，*Huang et al.*提出了 AoA 模块，通过计算关注向量和查询队列之间的相关性来解决这样的问题[12]。

AoA 模块提出了信息向量i和注意力门g的概念，利用注意力结果$\hat{v_i}$和当前查询q计算得到

$$i = W_i[q, \hat{v_i}] + b_i \quad \ldots(3-22)$$

$$g = \sigma(W_g[q, \hat{v_i}] + b_g) \quad \ldots(3-23)$$

随后 AoA 将注意力门g与信息向量i逐元素相乘，得到关注向量$\hat{i}$，即为 AoA 模块的注意力输出结果。

$$\hat{i} = i \odot g \quad \ldots(3-24)$$

注意力机制模块和 AoA 模块的结构对比如图 3.6 所示

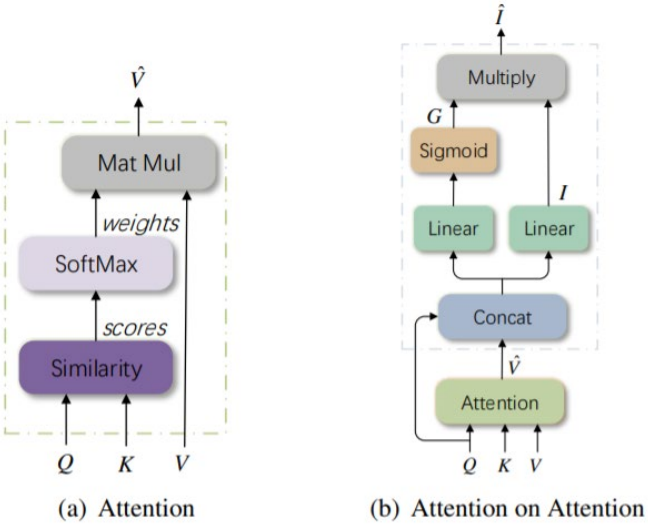


图 3.6 注意力模块和 AoA 模块的对比

Figure 3.6 The Comparison of Attention Module and AoA Module

将 AoA 模块应用于图像描述的编解码，得到 AoA 网络。在编码器中，仍然使用 CNN 得到图像的特征向量$\vec{a} = \{a_1, \dots, a_L\}$，接着输入 AoA 模块以优化图像特征的表达。用$AoA^E$表示编码器中的 AoA 模块，其输出结果再经过残差连接和层规范化，得到最后结果A'

$$A' = LayerNorm\left(\vec{a} + AoA^E(f_{att}, W^{Q_e}\vec{a}, W^{K_e}\vec{a}, W^{V_e}\vec{a})\right) \quad \dots(3-25)$$

其中W^{Q_e}，W^{K_e}，W^{V_e}是线性变换矩阵，f_{att}是经典注意力函数，可以替换为多头注意力函数f_{mh-att}，使用切片分别计算注意力再连接。由于该优化过程不会改变特征向量的维度，所以可以对优化模块进行堆叠而达到更好的优化效果。

解码器中的 AoA 模块记为AoA^D，将优化后的特征向量A'和上一时间片的隐变量h_t得到输出c_t

$$c_t = AoA^D(f_{att}, W^{Q_d}h_t, W^{K_d}\vec{a}, W^{V_d}\vec{a}) \quad \dots(3-26)$$

将当前时间片输入词语的词向量和视觉向量$\bar{a} + c_{t-1}$一并输入 LSTM，其中$\bar{a} = \frac{1}{L}\sum_{i=0}^{L} a_i$，得到

$$x_t = [W_e\Pi_t, \bar{a} + c_{t-1}] \quad \dots(3-27)$$

$$h_t, C_t = LSTM(x_t, h_{t-1}, C_{t-1}) \quad \dots(3-28)$$

其中W_e是词嵌入矩阵，而Π_t是当前时间片t输入词语ω_t的 One-hot 编码，h_t, C_t分别是当前时间片t时 LSTM 的隐变量和单元状态。最后利用隐变量h_t得到当前的单词y_t。

3.6 本章小结

第三章选取并介绍了四种图像描述生成的模型，这些模型是自然图像描述生成领域比较有代表性的模型，在自然图像数据集上也取得了比较好的结果，适合选用这些模型在遥感数据集上开展实验。本章重点介绍了这些模型的结构，第四章将基于这些模型开展在遥感图像数据集上的训练和测试。

第4章　实验设计与结果分析

4.1　实验环境

使用 Linux 命令查看实验环境配置，表 2.1 中展示了实验环境的硬件配置和软件环境。

表 4.1 实验环境

Table 4.1 Experimental Environment

硬件环境	CPU	Intel(R) Xeon(R) Gold 6126 CPU @ 2.60GHz
	GPU	GeForce RTX 2080 Ti
软件环境	操作系统	Ubuntu 16.04 LTS
	Python 版本	2.7.12
	CUDA 版本	10.0.130
	深度学习框架	Pytorch
	Torch 版本	1.0.0

4.2　遥感影像描述数据集简介

对于自然图像来说，图像描述的常用数据集主要有 MSCOCO、Flickr8k 、Flickr30k。遥感影像领域的数据集有 Sydney-Captions、UCM-Caption、RSICD。下面以遥感图像描述的数据集为例展开介绍。

4.2.1　Sydney-Captions 数据集

首先介绍 Sydney 数据集，它是仅含标签的高分遥感影像数据集。Sydney 数据集是从 Google Earth 上获取的悉尼地区的特大幅卫星图像上剪切得到的。全幅卫星图像的分辨率为 18000×14000，空间分辨率为 0.5 m/pixel [20]。随后 *Zhang et al.*将卫星图像分割成 1008 个互不重叠的子图，其子图尺寸均为 500×500。从这些子图中选出分类明确的图片，共选出 613 张分为 7 类的图片，其分类标签分别是工业区、海洋、草地、河流、机场、住宅区和高速路[20]，命名为 Sydney 数据集。该数据集中每张图像对应一种场景类别标签，是目前常用的高分遥感影像数据集。

全幅卫星图像如图 4.1 (a)所示，7 个场景的典型图如图 4.1 (b)所示。

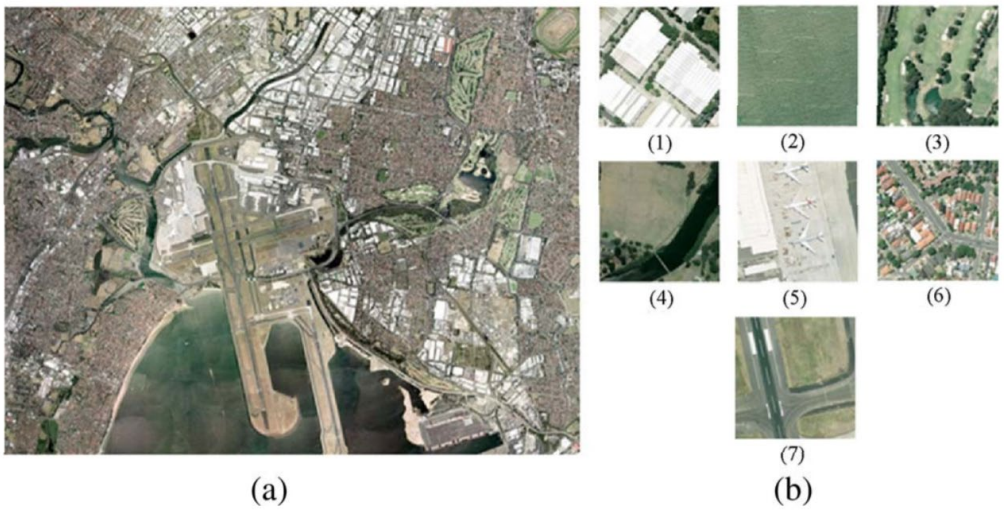


图 4.1（a）UCM 数据集使用的全幅卫星图像；（b）UCM 数据集的 7 个典型场景；（1）工业区，（2） 海洋，（3） 草地，（4） 河流，（5） 机场，（6） 居民区，（7） 飞机跑道.

Figure 4.1 (a) The Whole Satellite Image Used in UCM; (b) Seven Typical Scenes of UCM; (1) Industrial, (2) ocean, (3) meadow, (4) river, (5) airport, (6) residential, and (7) runway.

Qu 对 Sydney 数据集进行了图像描述的标注，构建了 Sydney-Captions 数据集。Sydney-Captions 数据集总共包含有 613 张高分遥感影像，图像分辨率为 500×500，空间分辨率为 0.5m/pixel，每张图片对应 5 句话文本描述，共计 3065 条文本[21]。下面给出一个 Sydney-Captions 数据集的标注实例，该图像是数据集内 imgid 为 452 的实例。

(1). An industrial area with many white buildings and some roads go through this area .
(2). This is an industrial area with some different white buildings .
(3). An industrial area with many white buildings and some roads go through this area .
(4). Some roads with plants on the roadside go through the industrial area .
(5). There are some white buildings in the industrial area with some roads go through .

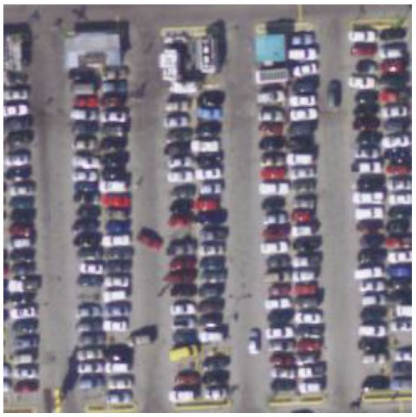
图 4.2 Sydney-Captions 数据集的一个示例

Figure 4.2 An Example from Sydney-Captions Dataset

4.2.2　UCM-Captions 数据集

UC Merced（UCM）数据集是遥感影像的场景分类问题中的另一个常用数据集，该数据集由美国地质调查局提供的卫星图像组成，包含了洛杉矶、迈阿密、休斯顿、波士顿、西雅图等美国城市的高分遥感影像，经过裁剪、分类后组成 UC Merced 数据集[22]，又称为 UCM 数据集。UCM 数据集共包含 21 类常见的场景，包括：机场、港口、油罐、网球场、海滩、棒球场、十字路、立交桥、住宅区、农田、森林、河流等场景。每类场景中含一百张图片，整个数据集一共有 2100 张高分遥感影像，每张图像的分辨率为 256×256，图像空间分辨率为 1ft/pixel [22]。

同样地，*Qu* 基于 UCM 数据集构建了 UCM-Captions 数据集，总共包含 2100 张高分遥感影像，每张图片对应 5 句话文本描述，共计 10500 条文本[21]。下面仍然给出一个 UCM-Captions 数据集的标注实例，该图像是数据集内 imgid 为 1585 的实例。



(1).Lots of cars are very closed to each other in the parking lot .
(2).Lots of cars parked in lines in the parking lot .
(3).Many cars are densely arranged in the parking lot .
(4).It is a parking lot with many cars densely arranged .
(5).Many cars densely arranged in the parking lot .

图 4.3 UCM-Captions 数据集的一个示例

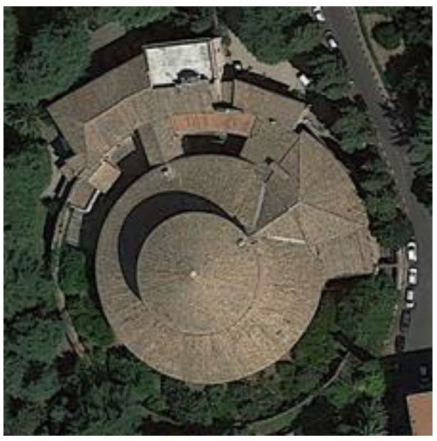
Figure 4.3 An Example from UCM-Captions Dataset

4.2.3　RSICD 数据集

RSICD（Remote Sensing Image Caption Dataset）数据集是由 *Lu et al.*从 Google Earth，百度地图，MapABC，Tianditu 收集了 1 万多幅遥感影像，并将这些图像以各种分辨率固定为 224×224。最后得到遥感影像的总数为 10921 张[23]，每个图像对应有五个句子描述。目前在遥感影像的描述生成领域，RSICD 数据集是最大的数据集，其包含目标种类多样，数据集的描述标签也非常丰富。根据遥感影像的尺度变化和旋转不变性，作者还提供了一些指令对 RSICD 数据集进行综合注释。由于图像是从飞机或卫星上捕获的，因此该数据集进行标注时没有方向性

的概念，而是使用“附近”等词来代替[23]。

该数据集中的样本图像具有较高的类别内多样性和较低的类别间不相似性。RSICD 中共计 24333 条文本，文本中一共有 3323 个不同的词汇。作者为了使句子更丰富，当没有五个不同的句子来描述同一张图像时，通过随机复制现有句子，将句子库扩展到 54605 个句子[23]，图 4.4 展示了名为 church_197 的图像和对应描述。



(1). many green trees are around a circle building near a road .
(2). many green trees are around a circle building near a road .
(3). many green trees are around a circle building near a road .
(4). many green trees are around a circle building near a road .
(5). many green trees are around a circle building near a road .

图 4.4 RSICD 数据集中复制句子的示例

Figure 4.4 An Example of Duplicate Sentences from RSICD

给出一个 RSICD 中图像的标注实例图 4.5，其原图像名为 playground_318，

(1). the playground consists of a red track and a green football field .
(2). two buildings and the road around the playground .
(3). there are a few trees around a plastic playground .
(4). here are the green grass and the tree that has just germinated .
(5). the playground consists of a red track and a green football field .

图 4.5 RSICD 数据集的一个示例

Figure 4.5 An Example from RSICD Dataset

要注意的是 RSICD 图像并不是严格按序号命名，而是以以标签加数字的形式命名，例如前述的 church_197。

4.3　图像描述生成的评价指标

4.3.1　BLEU

BLEU（Bilingual Evaluation Understudy, 双语评估辅助工具）标准最初应用于机器翻译领域，是用于分析翻译结果与参考翻译之间 n 元组相关性的评价指标，[24]属于基于精度的评价方式，直观意义是翻译结果与人类给出的参考翻译越接近，BLEU 得分越高。

对于图像I_i，翻译模型生成的翻译结果为c_i，而人工标注的参考翻译集合设为$S_i = \{s_{i1}, \dots, s_{im}\}$。翻译语句均用$n$元组$\omega_k$表示，$\omega_k$是$n$个有序单词序列，$n$一般取 1 到 4，$n$元组$\omega_k$在语句$s_{ij}$中出现的次数记为$h_k(s_{ij})$，在待评价的翻译语句$c_i$中记为$h_k(c_i)$.

首先计算全局n元组精度，其直观意义是参考翻译与翻译结果n元组重复的次数与翻译结果中n元组个数之比，引入最小值函数是为了将数值限定使之不超过 1。

$$P_n(C,S) = \frac{\sum_i \sum_k \min\left(h_k(c_i), \max_{j\epsilon m}\, h_k(s_{ij})\right)}{\sum_i \sum_k h_k(c_i)} \quad \dots(4-1)$$

其中k指的是长度为n所有可能的n元组的集合数。但仅使用n元组精度会让模型倾向于生成评分高的短句子，因此需要引入简洁性惩罚，待评价翻译句子小于参考翻译时，惩罚生效，最后计算得到的总分将会降低。

$$BP(C,S) = \begin{cases} e^{1-\frac{l_S}{l_C}}, & if\ l_C \le l_S \\ 1, & else \end{cases} \quad \dots(4-2)$$

l_C是待评价的翻译语句c_i的总长，l_S是有效人工标注的翻译语句的总长。如果一个待评价翻译对应多个参考翻译，则取简洁性惩罚最小的参考语句。

最后使用指数计算 BLEU 分数，ω_n对于所有n都是常量$\frac{1}{N}$，

$$BLEU_N(C,S) = BP(C,S)\exp\left(\sum_{n=1}^{N} \omega_n \log P_n(C,S)\right) \quad \dots(4-3)$$

同样地，BLEU 可以用于图像描述生成任务的评估，在模型生成的描述语句与数据集的参考语句之间分析相关性，得到的 BLEU 值越大，说明模型效果越好。

4.3.2 METEOR

METEOR（Metric for Evaluation of Translation with Explicit Ordering, 显式排序的翻译评估指标）同样是机器翻译领域的标准之一，其出现一定程度上弥补了BLEU 的某些缺点。随着召回率进入研究者视野，METEOR 作为基于召回率的评价指标应运而生，相比于BLEU，METEOR可以衡量分段上的准确性。METEOR基于一元组精度的加权调和平均与召回率，其计算需要依赖于一组事先给定的校准（Alignment），记为m，通俗理解即为一元组之间的映射集。而m来自于 WordNet的同义词库，通过最小化语句中的有序块（Chucks, 记为ch）得到[25]。METEOR计算式意为最佳翻译结果与参考翻译之间一元组的精度和召回率的调和平均数。

首先算出精度P_m和召回率R_m，其中$|m|$是待评价翻译中能与参考语句匹配的一元组数量，$h_k(c_i)$、$h_k(s_{ij})$与 BLEU 中提到的相同，分别为n元组ω_k在待评价翻译与参考翻译中出现的次数，

$$P_m = \frac{|m|}{\sum_k h_k(c_i)} \quad \ldots(4-4)$$

$$R_m = \frac{|m|}{\sum_k h_k(s_{ij})} \quad \ldots(4-5)$$

使用加权调和平均计算F_{mean}，α是预设常量，

$$F_{mean} = \frac{P_m}{\alpha P_m + (1-\alpha)R_m} \quad \ldots(4-6)$$

除了对一元组一致性的分析，METEOR 的惩罚值设计使得它可以更好地衡量分段准确性，为此 METEOR 引入了块的概念，块是待评价翻译和参考翻译中相邻的一元组集合，相邻映射越长，块的总数越少，如果完全一致即只有一个块。惩罚值计算如下，ch是块的数量，m是与校准映射的一元组的数量，惩罚值越大，最终 METEOR 得分越低。其中γ, θ为预设常量。

$$Pen = \gamma \left(\frac{ch}{m}\right)^{\theta} \quad \ldots(4-7)$$

最终计算 METEOR 如下所示，

$$METEOR = (1 - Pen)F_{mean} \quad \ldots(4-8)$$

其中α, γ, θ均为默认参数，是预先从另外的数据集上训练得到的，过多的超参数

也成为了 METEOR 的局限之处。

4.3.3　ROUGE

ROUGE（Recall-Oriented Understudy for Gisting Evaluation, 面向召回率的摘要评估辅助工具）是一个设计用于文本摘要（Text summarization）算法评估的评价标准[26]，其重点考虑的是翻译的充分性，对翻译的流畅度则不甚敏感。ROUGE 一般认为有四类，分别是 ROUGE-N，ROUGE-L，ROUGE-W，ROUGE-S。

ROUGE-N 是 ROUGE 系列的第一个标准，其根据待评价句子，计算n元组在待评价句子出现的次数与参考句子中出现的次数之比，即n元组召回率，式中$h_k(c_i)$、$h_k(s_{ij})$分别为n元组ω_k在待评价句子与参考句子中出现的次数，引入最小值函数是为了将数值限定在 1 以下，

$$ROUGE_n(c_i, S_i) = \frac{\sum_j \sum_k \min\left(h_k(c_i), h_k(s_{ij})\right)}{\sum_i \sum_k h_k(s_{ij})} \quad \ldots(4-9)$$

ROUGE-L 是基于最长公共子序列（Longest Common Subsequence, LCS）的评价标准，将待评价句子与参考句子之间最长公共子序列的长度记作$l(c_i, s_{ij})$，首先计算召回率R_l和精度P_l

$$R_l = \max_j \frac{l(c_i, s_{ij})}{|s_{ij}|} \quad \ldots(4-10)$$

$$P_l = \max_j \frac{l(c_i, s_{ij})}{|c_i|} \quad \ldots(4-11)$$

ROUGE-L 通过计算 F 度量（F-measure, or F1 score）得到

$$ROUGE_L(c_i, S_i) = \frac{(1+\beta^2) R_l P_l}{R_l + \beta^2 P_l} \quad \ldots(4-12)$$

其中$\beta = \frac{P_l}{R_l}$，也可以取作常量，一般取 1.2，其取值会影响 ROUGE-L 评分更偏向精度或者是更偏向召回率。

ROUGE-W 是基于 ROUGE-L 的改进版，基于加权最长公共子序列计算，其评分可以关注到连续匹配的情况并赋给高分。设加权最长公共子序列为$l_w(c_i, s_{ij})$，其计算过程中引入加权函数f，通过改变f可以对不同的连续匹配赋不同的分。f对于任何正整数x, y需要满足

$$f(x+y) > f(x) + f(y) \quad \ldots(4-13)$$

考虑最后 ROUGE-W 分数正则化的需求，我们一般考虑取具有相似形式逆函数的函数，例如$f(k) = k^2, f^{-1}(k) = k^{\frac{1}{2}}$。ROUGE-W 计算式如下，

$$R_w = f^{-1}\left(\frac{l_w(c_i, s_{ij})}{f(|s_{ij}|)}\right) \quad \ldots(4-14)$$

$$P_w = f^{-1}\left(\frac{l_w(c_i, s_{ij})}{f(|c_i|)}\right) \quad \ldots(4-15)$$

$$ROUGE_w(c_i, S_i) = \frac{(1+\beta^2)R_w P_w}{R_w + \beta^2 P_w} \quad \ldots(4-16)$$

ROUGE-S 则基于跳跃二元组（Skip bigram）的最小单位，跳跃二元组是一对有序单词，允许这样的两个有序单词中间有其他单词，而不必要求必须是连续的，其中可以设置最多允许跳过多少个单词。这使得 ROUGE-S 的优势在于更灵活地匹配单词组。ROUGE-S 计算过程如下，其中待评价句子中跳跃二元组总个数记为$f_k(s_{ij})$，

$$R_s = \max_j \frac{\sum_k \min\left(f_k(c_i), f_k(s_{ij})\right)}{\sum_k f_k(s_{ij})} \quad \ldots(4-17)$$

$$P_s = \max_j \frac{\sum_k \min\left(f_k(c_i), f_k(s_{ij})\right)}{\sum_k f_k(c_i)} \quad \ldots(4-18)$$

$$ROUGE_s(c_i, S_i) = \frac{(1+\beta^2)R_s P_s}{R_s + \beta^2 P_s} \quad \ldots(4-19)$$

其中在四种 ROUGE 里，最常用的是 ROUGE-N。由于 METEOR 也涉及了召回率计算，因此在本文实验中选用 ROUGE-L。ROUGE 评价得到的分数越高，说明待评价句子越好。

4.3.4 CIDEr

CIDEr（Consensus-based Image Description Evaluation, 基于共识的图像描述评估）是一个专门为图像描述任务提出的评价指标，是对n元组进行词频逆文本频率指数（Term Frequency Inverse Document Frequency, TF-IDF）权重计算得到的[27]。

n元组ω_k的 TF-IDF 权重可记为$g_k(s_{ij})$，计算式如下所示

$$g_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_{\omega_l \in \Omega} h_l(s_{ij})} \log\left(\frac{|I|}{\sum_{I_p \in I} \min(1, \sum_q h_k)}\right) \quad \ldots(4-20)$$

其中$h_k(s_{ij})$是n元组在待评价描述语句中出现的次数。Ω是所有n元组构成的词表，$\sum_{\omega_l \in \Omega} h_l(s_{ij})$是词表中$n$元组出现的总次数，此处由于计算结果中含$k$，因此换用$l$作为$n$元组的角标，意义上并无区别。$I$是数据集中图像的数量，$I_p$是任一描述中出现了$n$元组$\omega_l$的图像集合，$h_k(s_{pq})$则是$n$元组$\omega_k$在这些描述里出现的次数。

直观来看，如果n元组ω_k频繁出现在该图像的参考描述中，该n元组就会在系数项（即 TF 项）得到更高的权重，而在对数函数（IDF 项）中，那些在所有图像中都频繁出现的n元组会得到更低的分数，两项是相互制约的关系。这种方法将出现频率高但和图像内容没有太大关系的单词权重降低，提高了重要图像信息的权重。

利用 TF-IDF 权重可以求得 CIDEr-n 如下

$$CIDEr_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{g^n(c_i) \cdot g^n(s_{ij})}{\|g^n(c_i)\| \|g^n(s_{ij})\|} \quad \ldots(4-21)$$

其中$g^n(c_i)$是由$g_k(c_i)$生成的向量，对应所有长度为n的n元组，$\|g^n(c_i)\|$是向量$g^n(c_i)$的大小，$g^n(s_{ij})$和$\|g^n(s_{ij})\|$对应是$g_k(s_{ij})$生成的向量和向量大小。最终得到 CIDEr：

$$CIDEr(c_i, S_i) = \sum_{n=1}^{N} \omega_n CIDEr_n(c_i, S_i) \quad \ldots(4-22)$$

其中$\omega_n = \frac{1}{N}$，一般N取 4 即可。

4.3.5　SPICE

SPICE（Semantic Propositional Image Caption Evaluation，语义命题图像标题评估）是基于图像的语义表示的评价指标。SPICE 将待评价描述语句与参考描述用概率上下文无关文法（Probabilistic Context-Free Grammars, PCFG）解析成句法依赖树（Syntactic dependencies trees），然后利用基于规则的方法把依赖树映射成场景图（Scene graphs），最后根据场景图得到的元组计算 F 度量得到 SPICE 值[28]。

$$P(c_i, S_i) = \frac{\left|T\left(G(c_i)\right) \otimes T\left(G(S_i)\right)\right|}{\left|T\left(G(c_i)\right)\right|} \quad \ldots(4-23)$$

$$R(c_i, S_i) = \frac{\left|T\left(G(c_i)\right) \otimes T\left(G(S_i)\right)\right|}{\left|T\left(G(S_i)\right)\right|} \quad \ldots(4-24)$$

$$SPICE(c_i, S_i) = F_1(c_i, S_i) = \frac{2P(c_i, S_i)R(c_i, S_i)}{P(c_i, S_i) + R(c_i, S_i)} \quad \ldots(4-25)$$

其中$G(\cdot)$表示映射为场景图的操作，$T(\cdot)$表示将场景图转换为多个元组集合的操作，$\otimes$运算是元组匹配。SPICE 通过这种方法只对图像中的目标、属性和关系做 F 度量，更适合于图像描述任务。

4.4 实验流程

实验基于 UCM 和 RSICD 数据集尝试复现 Show and tell、Show, attend and tell、AoA、Transformer 四个模型，并完成在同一数据集上不同模型的评估。实验中采用的优化方法是首先使用预训练的编码器，只对解码器的参数进行训练，而编码器的参数固定；20 个 epoch 之后开启编码器的训练，使编解码器协同训练以达到更好的效果。

实验中记录了训练损失、验证损失，以及 4.3 节中提到的评价指标（BLEU_1–BLEU_4，METEOR，ROUGE-L，CIDEr，SPICE），以便实验后作图展示训练过程中的变化。模型训练完成后，运行测试实验，记录根据测试图片生成的图像描述，分析模型直观的训练结果。

实验中记录了不同模型在这两个数据集上的训练时长以做参考，其中模型在 UCM 数据集上训练了 50 个 epoch，在 RSICD 上训练了 30 个 epoch。如表 4.2 所示，不同模型在同一数据集上训练时长近似：UCM 数据集上一般训练 2 小时 40 分钟左右，RSICD 数据集上一般训练 10 个小时左右。训练时长和训练数据集的类型有比较高的相关性，可以此为依据定性评估实验是否正常完成。

表 4.2 四个模型在两个数据集上的训练时长

Table 4.2 Training Time for Four Models on Two Dataset

	UCM	RSICD
show and tell	2:41:10	9:27:24
show, attend and tell	2:37:44	9:58:02
transformer	2:54:09	10:02:20
AoA	2:45:14	10:05:25

4.5 实验损失曲线分析

训练损失展示了模型训练过程中的参数拟合过程，理论上是一条下降且收敛的曲线，由于训练损失记录步长较小，局部的波动是正常的情况，实际上训练损失应该是一条拟合后的曲线，其趋势为下降且收敛；模型训练过程中也会在验证集做测试并记录验证损失，验证损失曲线同样应为一条平滑下降且收敛的曲线，如果训练损失收敛，验证损失不收敛，表示模型此时可能出现了过拟合的问题。

对于两个数据集的所有模型，均在每一步记录训练损失，每 100 步记录验证损失以及模型得分，将损失值和评价指标等数值存入 TensorBoard 可读的文件。训练结束后使用 TensorBoard 读取数据并使用 Matlab 作图，对于 UCM 数据集，各模型的损失曲线如图 4.6 所示。

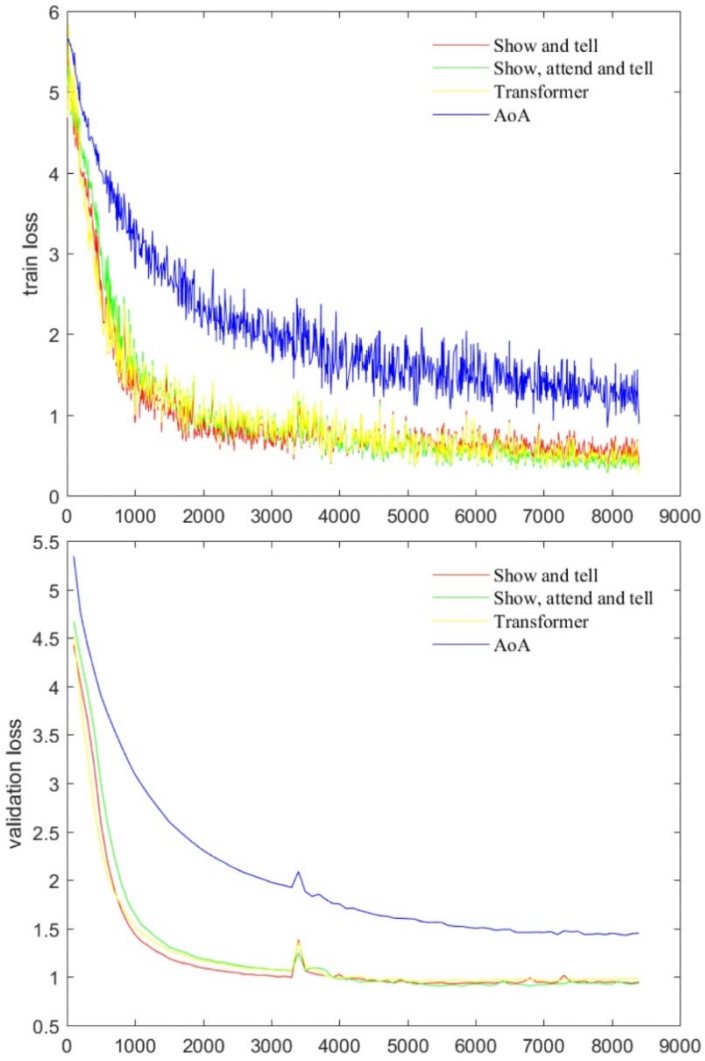


图 4.6（a）四个模型在 UCM 数据集上的训练损失曲线；（b）四个模型在 UCM 数据集上的验证损失曲线

Figure 4.6 (a) The Training Loss Curve of Four Models on UCM; (b) The Validation Loss Curve of Four Models on UCM

对于 RSICD 数据集，各模型的损失曲线如图 4.7 所示。

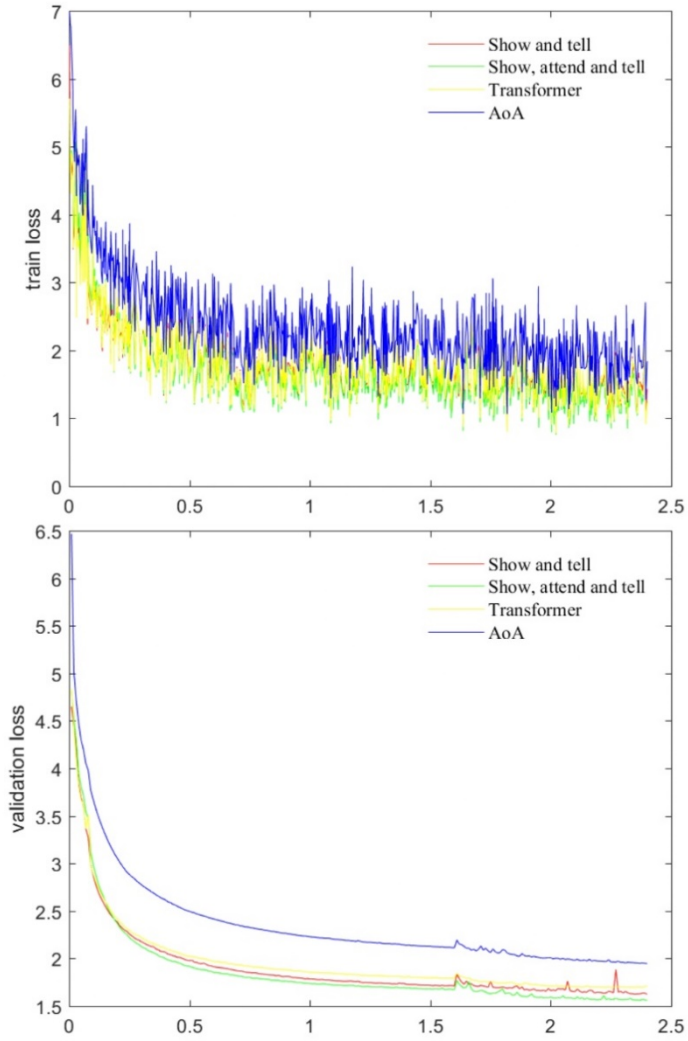


图 4. 7 四个模型在 RSICD 数据集上的训练损失曲线；（b）四个模型在 RSICD 数据集上的验证损失曲线

Figure 4.7 (a) The Training Loss Curve of Four Models on RSICD; (b) The Validation Loss Curve of Four Models on RSICD

可以看到，四个模型在两个数据集上的训练损失和验证损失曲线均表现出下降收敛的趋势，证明了训练的有效性。其中对于两个数据集，AoA 模型的损失曲线均明显高于其他三个模型。

4.6　模型评价指标得分

训练过程中记录了 BLEU_1, BLEU_2, BLEU_3, BLEU_4，METEOR，ROUGE-L，CIDEr，SPICE 的变化数据，仍然使用 TensorBoard 得到这些评价指标的变化曲线。其中 CIDEr 收敛值高于其他指标，因此将 CIDEr 曲线单独列出，其余指标画在一张图内。

UCM 的 CIDEr 曲线变化如图 4.8 所示，可以看到 UCM 数据集上 CIDEr 值呈增长趋势，一般稳定于 3 左右，四个模型的稳定值差别不大。其中曲线在训练进行到 20 个 epoch（即图中 3.4k 的位置）有明显的极低值点和波动，这是由于编码器在此时开始参数的训练，随后编解码器共同训练会带来性能的提升。可以看到 20 个 epoch 之后除去 Show and tell 模型增长不明显外，其他三个模型的 CIDEr 值均有较为明显的增长。

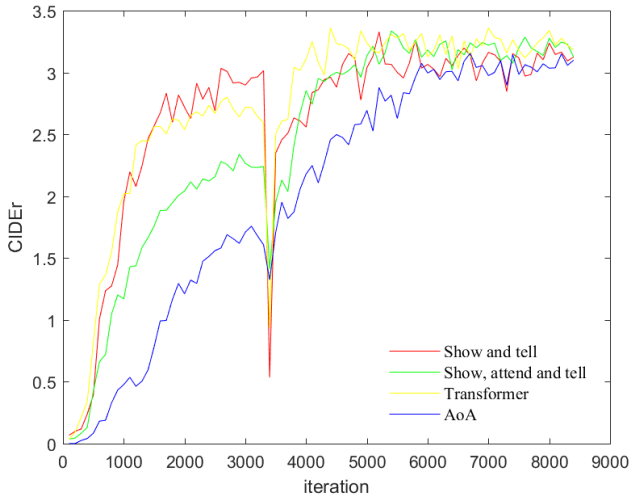


图 4.8 四个模型在 UCM 数据集上的 CIDEr 变化曲线

Figure 4.8 The Curve of CIDEr of Four Models on UCM

UCM 的其他指标变化如图 4.9 所示，经观察可得，这些评价指标一般从大到小为：BLEU_1, ROUGE_L, BLEU_2, BLEU_3, BLEU_4, SPICE, METEOR。其中 BLEU_1, ROUGE_L, BLEU_2, BLEU_3, BLEU_4 稳定在 0.6-0.8 区间内，SPICE 和 METEOR 相较于这些指标更低一些，在 0.4-0.5 之间。不同模型的评分曲线均呈上升后收敛的趋势，其中波折之处均在训练了 20 个 epoch 左右，其波动是编解码器协同训练带来的影响。可以看到编码器参数训练的加入使得曲线波动更剧烈，但也明显地提高了其得分。在基于 UCM 的四个模型之中，从各评价指标的稳定值看，Show and tell 和 AoA 模型得分较好。综合 4.5 节中对损失曲线的分析，在 UCM 数据集上，Show and tell 模型综合表现应最好，但与其他模型相差不大。

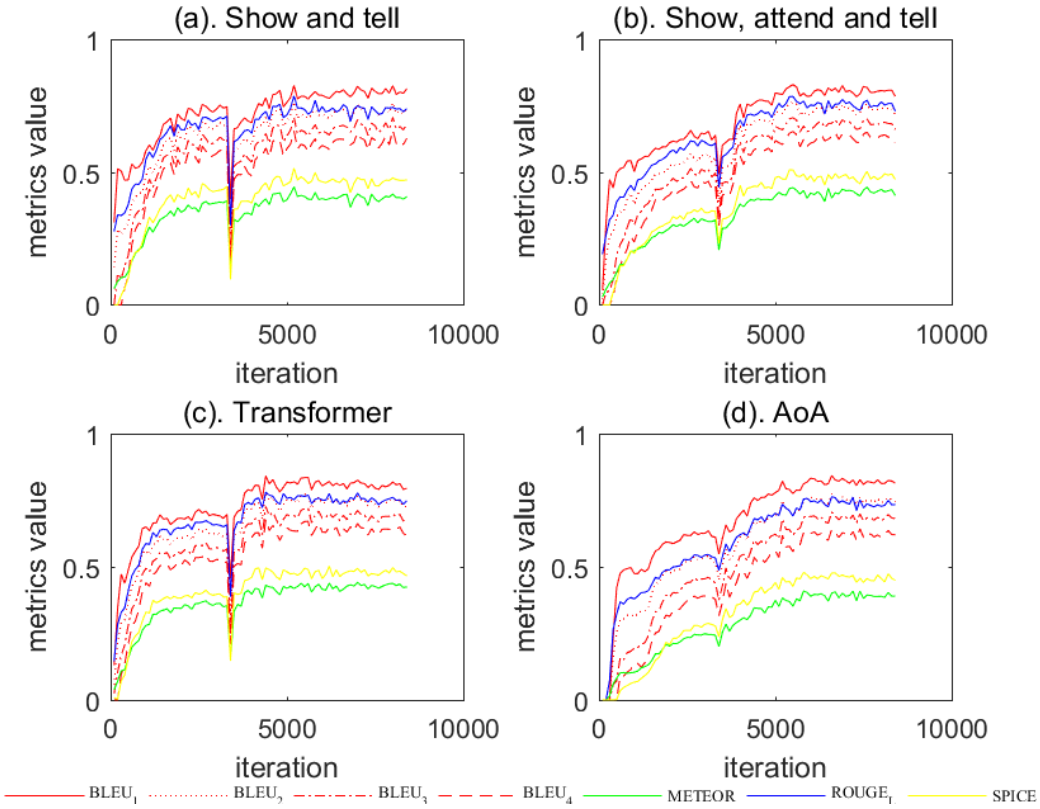


图 4.9 四个模型在 UCM 数据集上的 BLEU_1 - BLEU_4、ROUGE_L、METEOR、SPICE 变化曲线

Figure 4.9 The Curve of BLEU_1 - BLEU_4, ROUGE_L, METEOR and SPICE of Four Models on UCM

四个模型在 RSICD 数据集上 CIDEr 得分曲线如图 4.10 所示，可以注意到经过 30 个 epoch 的训练后，Show and tell 模型和 Show, attend and tell 模型的 CIDEr 可以达到 2.4 左右，Transformer 模型的 CIDEr 值达到 2.3 左右，AoA 模型的 CIDEr 值达到 2.2 左右。同样地我们可以注意到编码器参数训练的加入对四个模型的评分均有一定的增幅作用。从数值上看，Show, attend and tell 模型和 Transformer 模型略胜一筹。

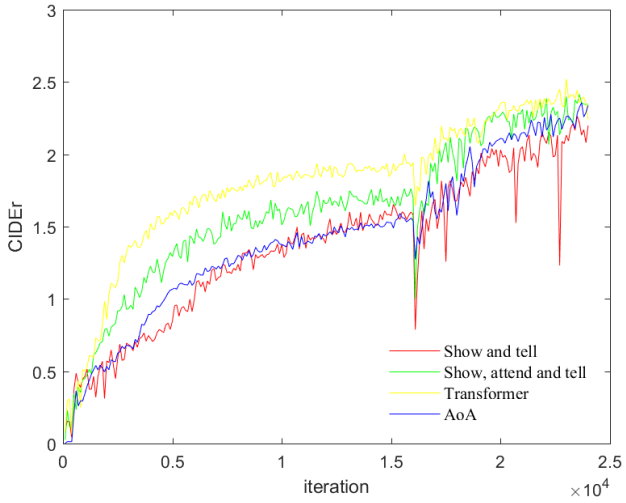


图 4. 10 四个模型在 RSICD 数据集上的 CIDEr 变化曲线

Figure 4.10 The Curve of CIDEr of Four Models on RSICD

四个模型在 RSICD 上的其他指标曲线如图 4.11 所示，类似于 UCM，这些指标的大小关系仍然相对固定，从大到小分别是 BLEU_1, ROUGE_L, BLEU_2, BLEU_3, BLEU_4, SPICE, METEOR。这些评价指标均呈现上升且收敛的趋势，但与 UCM 不同的是，SPICE 和 METEOR 不再明显小于其他指标。所有的这些指标较为均匀地分布在 0.3-0.65 的区间内。对于 RSICD 的其他指标，在 20 个 epoch 处（即图中 16k 处）编码器参数训练加入后产生了极低值点，且从 20 个 epoch 后四个模型的曲线更波动，但数值得到了提升。从训练过程的最终数值上来看，Show, attend and tell 模型和 Transformer 模型略胜一筹。

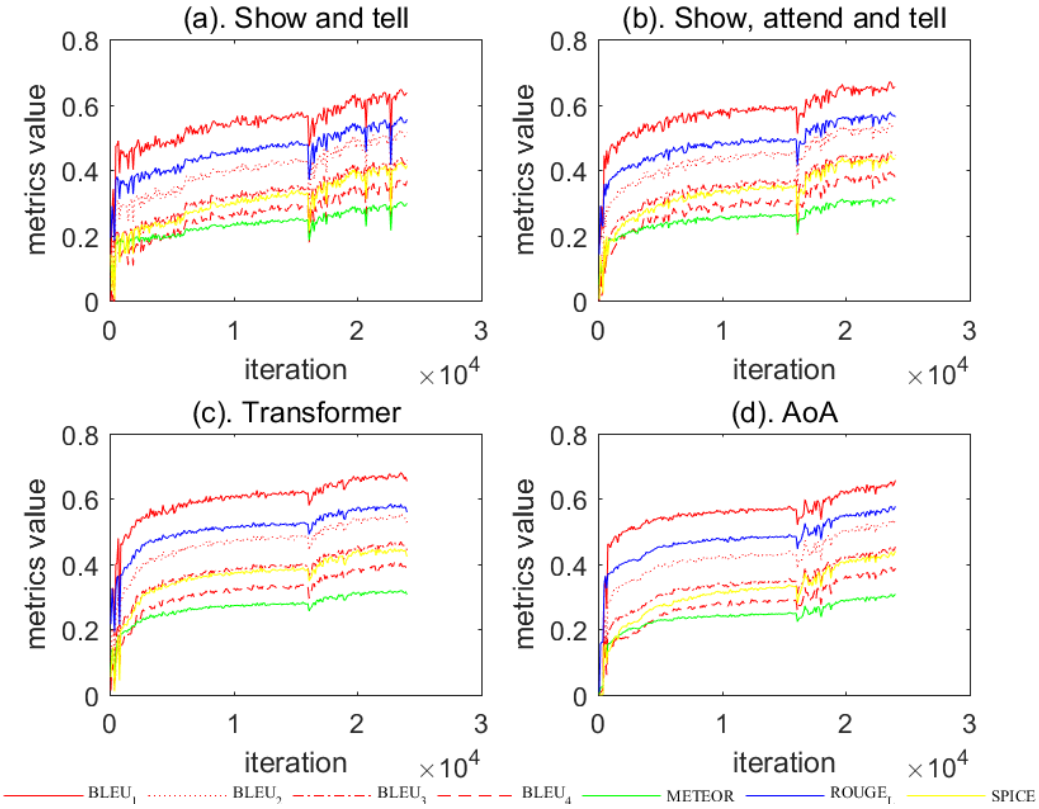


图 4. 11 四个模型在 RSICD 数据集上的 BLEU_1 - BLEU_4、ROUGE_L、METEOR、SPICE 变化曲线

Figure 4.11 The Curve of BLEU_1 - BLEU_4, ROUGE_L, METEOR and SPICE of Four Models on RSICD

4.7　模型预测结果可视化

4.7.1　UCM 数据集上预测结果测试

完成模型的训练后，分别使用 UCM 和 RSICD 数据集中的测试集对训练后的模型进行测试，得到测试图片的描述。这一节将对比不同模型对同一测试图片的描述结果，以直观表征模型的训练结果并定性验证 4.6 节中的推测。

首先选取 UCM 数据集中目标丰富或含有重要细节的图片作为测试例，最终选了 UCM 数据集中第 692 张图片和第 895 张图片，记录数据集中图片和所给描述如图 4.12，

(1). There are lots of houses arranged in lines and a road goes through them .
This is a dense residential area .
(2). It is a dense residential area with lots of houses arranged in lines .
(3). There are lots of houses arranged in lines and a road goes through them .
(4). Lots of houses arranged in lines and a road goes through them .
(5). There are lots of houses arranged in lines and a road goes through them .

图 4.12 UCM 测试用例 1

Figure 4.12 Test Case No.1 from UCM

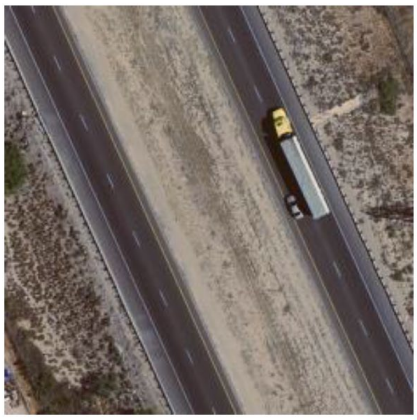
然后从测试结果中取出四个模型对图 4.12 的描述语句，记录在表 4.3 中，可以看到四个模型产生的描述语义与图片相符，Show and tell 模型和 Show, attend and tell 模型产生的句子内容相似且较为准确，Transformer 模型产生的句子与其他三个不甚相同，其描述主要关注了房屋以及屋顶的颜色属性，而对公路等背景关注甚少。AoA 产生的描述与前两个模型相似，但在宾语更加精确到了 area。

表 4.3 四个模型对 UCM 测试用例 1 图片生成的描述

Table 4.3 The Captions Generated from Test Case No.1 from UCM by Four Models

Model	Caption
Show and tell	There are lots of houses arranged neatly and a road goes through them
Show, attend and tell	Lots of houses arranged neatly and a road goes through them
Transformer	There are some buildings with grey roofs
AoA	Lots of houses arranged neatly and a road goes through this area

接着随机取出 UCM 测试集中另一张图片，在文件夹中编号为 895，其原图和数据集真实描述如图 4.13 所示：



(1). There are two white straight freeways parrallel forward .
(2). There are two white straight freeways closed together with with cars on them .
(3). Two straight freeways parrallel forward with with some cars on them .
(4). Two white straight freeways closed together with some plants beside them .
(5). Two straight freeways parrallel forward some plants beside them .

图 4.13 UCM 测试用例 2

Figure 4.13 Test Case No.2 from UCM

取出四个模型对图 4.13 产生的描述，记录在表 4.4 中。从第 894 张图片的测试结果看，四个模型同样较为准确地反映了图片内容，但句子均有些简略，其中 AoA 模型的结果相比于其他三个模型结果更精确一些，具体到了高速路上的车辆。

表 4.4 四个模型对 UCM 测试用例 2 图片生成的描述

Table 4.4 The Captions Generated from Test Case No.2 from UCM by Four Models

Model	Caption
Show and tell	There are two straight freeways in the desert
Show, attend and tell	There are two straight freeways in the desert
Transformer	There are two straight freeways in the desert
AoA	There are two straight freeways with some cars on the roads

在 UCM 数据集上，通过抽取图片直观分析模型预测结果，得到如下结论：四个模型都能比较好地反映图片内容，AoA 稍好一点但相差不大。得到的结论与 4.6 节中根据评价指标的推测相符。

4.7.2　RSICD 数据集上预测结果测试

再选取 RSICD 数据集中名为“sparseresidential_115.jpg”和“bridge_200.jpg”的图片，记录数据集中图片和给定的描述如图 4.14 和图 4.15，以及四个模型产生的描述结果表 4.5 以及表 4.6。

sparseresidential_115.jpg 中目标较多，包含了房屋、草地、树木和池塘，其原图和给定描述如图 4.14 所示，

(1). six houses with dark gray roof in the middle .
(2). a square pool with blue water and gray border in the middle .
(3). a parking lot with white pool besides the house in it .
(4). a residential surrounded by many green plants and a gray road in side .
(5). a building is surrounded by some green trees and meadows .

图 4.14 RSICD 测试用例 1

Figure 4.14 Test Case No.1 from RSICD

四个模型对图 4.14 产生的预测结果如表 4.5 所示，可以看到，Show and tell 模型可以识别并表达房屋、树木和池塘，AoA 模型只描述了房屋和树木。而 Show, attend and tell 模型和 Transformer 模型则准确识别并描述了这四类目标。

表 4.5 四个模型对 RSICD 测试用例 1 图片生成的描述

Table 4.5 The Captions Generated from Test Case No.1 from RSICD by Four Models

Model	Caption
Show and tell	a building with a swimming pool is surrounded by many green trees
Show, attend and tell	a building with a swimming pool is surrounded by many green trees and meadows
Transformer	a building with a swimming pool is surrounded by many green trees and meadows
AoA	a building is surrounded by many green trees

bridge_200.jpg 同样包含了较多种类的目标，包括了房屋、树木、桥和河流，其图片和给定描述如图 4.15 所示，

(1). the small bridge is next to some factory buildings with red roofs .
(2). there is a bridge over a river and some house arranged orderly along the river .
(3). a car is steering on the bridge across the river between two factory mills .
(4). the brighe connects the two sides of a dark narrow river which has buildings with red roofs and blue ones .
(5). some buildings and green plants are in two sides of a river with a bridge over it .

图 4.15 RSICD 测试用例 2

Figure 4.15 Test Case No.2 from RSICD

四个模型对图 4.15 产生的预测结果如表 4.6 所示，可以看到 Show and tell 模型表达了桥、河流和树木，AoA 模型描述了房屋、树木和河流，但 AoA 模型生成的描述不甚通顺。而 Show, attend and tell 模型和 Transformer 模型则识别到了所有目标，并生成了准确的描述。

表 4.6 四个模型对 RSICD 测试用例 2 图片生成的描述

Table 4.6 The Captions Generated from Test Case No.2 from RSICD by Four Models

Model	Caption
Show and tell	a bridge is over a river with some green trees in two sides of it
Show, attend and tell	many green trees and some buildings are in two sides of a river with a bridge
Transformer	many buildings and some green trees are in two sides of a river with a bridge
AoA	many buildings and green trees are in two sides of a river with a river

综上所述，这四个模型在 UCM 和 RSICD 数据集上都能较为准确地完成图像描述生成的任务，其中在 UCM 数据集上四个模型表现相差不大，均能较好地完成描述任务，而在 RSICD 数据集上，Show, attend and tell 模型和 Transformer 模型表现更好，相比于其余两个更完善。

4.8 图像描述正确错误分析

虽然在 4.7 节中抽取并评估了模型的有效性，但浏览测试集的测试结果时注

意到了一些错误。接下来将展示并分析 UCM 数据集测试结果的两个错误。

首先注意到 AoA 模型的测试结果中有一句描述的结构显得异常：

A baseball diamond with sand and sand and sand

寻找数据集内对应的图片如图 4.16 所示，

(1). This is a baseball diamond .
(2). It is a small baseball diamond with sands and grass .
(3). It is a regular baseball diamond .
(4). A regular baseball diamond compose of manicured lawns and sand .
(5). A regular baseball diamond compose of manicured lawns and sand .

图 4.16 UCM 错误分析用例 1

Figure 4.16 Error Analysis Case No.1 from UCM

对照测试图片和生成的描述，可以发现 AoA 模型生成的描述语义大致正确，但生成了冗余的词汇。推测此处是由于 AoA 模型在长句子的生成方面欠佳，存在长期依赖问题，导致句尾句子结构混乱。为了验证此推测，提取有利于长句子生成的 Transformer 模型对这张图片的预测结果：

A regular baseball diamond compose of manicured lawns and sand

可以看到 Transformer 模型生成的长句子准确地反映了图片内容，且未出现结构混乱，因此可以验证之前长期依赖问题的推测。

此外注意到四个模型对 UCM 测试集中的飞机图片描述出现了较大偏差，由于 UCM 数据集中的图片类别间差别较大，而类别内差别较小[29]，因此选取其中一张测试图片以及其真实描述作为代表，如图 4.17 所示，

(1). There are many small airplanes at the airport .
(2). Many different kinds of airplanes are stopped at the airport .
(3). Seven small airplanes are stopped neatly at the airport .
(4). Many white airplanes stopped at the airport .
(5). Many different airplanes are stopped neatly at the airport .

图 4.17 UCM 错误分析用例 2

Figure 4.17 Error Analysis Case No.2 from UCM

对于这张图片，四个模型预测的描述分别如表 4.7 所示，可以看到与原图片和真实描述有较大出入，除了这张图片，其他几张飞机图片预测结果也是类似。推测这种情况是由于 UCM 数据集图片过少，关于飞机的图像训练不佳，没有达到理想的效果。

表 4.7 四个模型对 UCM 错误分析用例 2 图片生成的描述

Table 4.7 The Captions Generated from Error Analysis Case No.2 from UCM by Four Models

Model	Caption
Show and tell	A white storage tank is on the ground
Show, attend and tell	It is a small baseball diamond
Transformer	It is a small baseball diamond
AoA	There are some white white sand beach and some white white sand

至于 RSICD 数据集，其测试集图像数量较大，随机抽查了各个部分，没有发现有明显问题的测试描述。RSICD 比较常见的问题如 4.7 节中所述，在目标丰富的图像中，可能会漏掉一些目标，但总体语义是较为准确的。其原因在于 RSICD 图像数量多，数据集标注描述丰富，类间描述也更加多样性[29]，训练得到的模型更符合实用要求。而 UCM 数据较少，训练好的模型很可能不足以在实际应用中推广使用。

4.9 本章小结

本章主要记录和分析了实验流程和实验结果。首先叙述了工作环境的硬件及

软件配置，随后介绍了三个遥感图像描述的数据集，以及图像描述的评价指标，接着记录了实验流程以及实验得到的数据和曲线，最后基于训练结果进行分析。

本章结论是四个模型分别在两个数据集上都达到了目标要求，生成的描述与图片相符；模型之间有细微差别，对于 UCM 数据集，四个模型效果均能达到要求，对于 RSICD 数据集，Show, attend and tell 模型和 Transformer 模型表现稍好一些。虽然这些模型在 UCM 和 RSICD 数据集上的评价指标数值相差较小，但通过抽取测试和错误例分析，可以得到在 RSICD 上训练后得到的模型效果更好，训练好的模型更适合实际应用。因此出于实际应用的考虑，应当采用在 RSICD 数据集上训练的 Show, attend and tell 模型或 Transformer 模型。

第5章　总结与展望

5.1　论文工作总结

图像描述是建立在计算机视觉和自然语言处理这两大热门方向之上的交叉方向，被研究者们寄予厚望。图像描述任务可以同时反映图像中的目标信息以及它们之间的相对位置信息，因此该技术可以用于对图像信息的全面快速提取。但该领域目前的主要工作大多是基于自然图像数据集开展的，遥感影像数据集上的研究为数不多。随着遥感技术的发展，如何更好地从大量遥感数据中更有效地提取数据，成了一个亟待解决的问题。本文的主要工作是将自然图像上应用的图像描述生成模型用于遥感影像数据集，产生对遥感影像的图像描述，并完成模型的复现与评估。主要结论如下：

1. 在 UCM 和 RSICD 数据集上实现了 Show and tell，Show, attend and tell，Transformer，AoA 等模型的复现。
2. 通过实验损失函数证明了训练的有效性，并通过多种评价指标定量比较各个模型。
3. 将不同模型的测试描述比对，定性验证前述结论。
4. 测试训练得到的模型，将得到的测试描述与真实标签比对，定性验证模型的有效性。
5. 研究了不同模型的少数错误描述，讨论了实验中可能存在的问题。

5.2　进一步工作展望

本文完成了在遥感影像数据集上完成图像描述生成的任务，产生的描述基本符合原图内容。为进一步优化训练结果，可以从以下方向入手：

1. 换用不同的 CNN 网络或调整学习率等参数，可以对实验结果进行小幅度的优化，或者是部分牺牲结果精度以得到训练时长的缩短；
2. 本实验中 beam size 设置为 1，可以尝试改变 beam size 的值来优化训练结果；
3. 可以利用强化学习优化模型的训练。

参考文献

[1] QU B, LI X, TAO D, et al. Deep semantic understanding of high resolution remote sensing image; proceedings of the 2016 International conference on computer, information and telecommunication systems (Cits), F, 2016 [C]. IEEE.

[2] KULKARNI G, PREMRAJ V, ORDONEZ V, et al. BabyTalk: Understanding and Generating Simple Image Descriptions [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2891-903.

[3] FARHADI A, HEJRATI M, SADEGHI M A, et al. Every Picture Tells a Story: Generating Sentences from Images, Berlin, Heidelberg, F, 2010 [C]. Springer Berlin Heidelberg.

[4] CHO K, VAN MERRIëNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [J]. 2014.

[5] VINYALS O, TOSHEV A, BENGIO S, et al. Show and Tell: A Neural Image Caption Generator [J/OL] 2014, arXiv:1411.4555[https://ui.adsabs.harvard.edu/abs/2014arXiv1411.4555V.

[6] KARPATHY A, FEI-FEI L J A E-P. Deep Visual-Semantic Alignments for Generating Image Descriptions [J/OL] 2014, arXiv:1412.2306[https://ui.adsabs.harvard.edu/abs/2014arXiv1412.2306K.

[7] JIA X, GAVVES E, FERNANDO B, et al. Guiding Long-Short Term Memory for Image Caption Generation [J/OL] 2015, arXiv:1509.04942[https://ui.adsabs.harvard.edu/abs/2015arXiv150904942J.

[8] WU Q, SHEN C, LIU L, et al. What value do explicit high level concepts have in vision to language problems? [J/OL] 2015, arXiv:1506.01144[https://ui.adsabs.harvard.edu/abs/2015arXiv150601144W.

[9] XU K, BA J, KIROS R, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention [J/OL] 2015, arXiv:1502.03044[https://ui.adsabs.harvard.edu/abs/2015arXiv150203044X.

[10] LU J, XIONG C, PARIKH D, et al. Knowing when to look: Adaptive attention via a visual sentinel for image captioning; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2017 [C].

[11] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. 2017.

[12] HUANG L, WANG W, CHEN J, et al. Attention on attention for image captioning; proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, F, 2019 [C].

[13] 李志欣, 魏海洋, 张灿龙, et al. 图像描述生成研究进展 [J]. 2021: 1.

[14] RENNIE S J, MARCHERET E, MROUEH Y, et al. Self-critical Sequence Training for Image Captioning [J/OL] 2016, arXiv:1612.00563[https://ui.adsabs.harvard.edu/abs/2016arXiv161200563R.

[15] SHI Z, ZOU Z J I T O G, SENSING R. Can a machine generate humanlike language descriptions for a remote sensing image? [J]. 2017, 55(6): 3623-34.

[16] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2016 [C].

[17] HOCHREITER S, SCHMIDHUBER J J N C. Long short-term memory [J]. 1997, 9(8): 1735-80.

[18] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [J]. 2013.

[19] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation; proceedings of the Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), F, 2014 [C].

[20] ZHANG F, DU B, ZHANG L J I T O G, et al. Saliency-guided unsupervised feature learning for scene classification [J]. 2014, 53(4): 2175-84.

[21] 屈博. 高分遥感影像的语义理解 [D]; 中国科学院大学 (中国科学院西安光学精密机械研究所), 2017.

[22] YANG Y, NEWSAM S. Bag-of-visual-words and spatial extensions for land-use classification; proceedings of the Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems, F, 2010 [C].

[23] LU X, WANG B, ZHENG X, et al. Exploring models and data for remote sensing image caption generation [J]. 2017, 56(4): 2183-95.
[24] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation [Z]. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, Pennsylvania; Association for Computational Linguistics. 2002: 311–8.10.3115/1073083.1073135
[25] BANERJEE S, LAVIE A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments; proceedings of the Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, F, 2005 [C].
[26] LIN C-Y. Rouge: A package for automatic evaluation of summaries; proceedings of the Text summarization branches out, F, 2004 [C].
[27] VEDANTAM R, LAWRENCE ZITNICK C, PARIKH D. Cider: Consensus-based image description evaluation; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2015 [C].
[28] ANDERSON P, FERNANDO B, JOHNSON M, et al. Spice: Semantic propositional image caption evaluation; proceedings of the European conference on computer vision, F, 2016 [C]. Springer.
[29] YUAN Z, ZHANG W, FU K, et al. Exploring a Fine-Grained Multiscale Method for Cross-Modal Remote Sensing Image Retrieval [J]. IEEE Transactions on Geoscience and Remote Sensing, 2021: 1-19.

致 谢

行文匆匆到了致谢部分，心头千言万语提笔却不知从何说起。本科这四年来，不乏痛苦迷惘，而今日看来，又不足为人道，就像一杯浓茶，入口虽苦，细品却有回甘。作为一个后劲不足的学生，本科对我来说过得些许艰难，但在这里我也有幸结识了许多优秀的同伴。虽然过程艰难，但国科大自由的环境、优秀的师资、充沛的资源，无一不让我受益无穷，让我站在了更高的平台去看世界。在此真诚地感谢我的学校中国科学院大学，是她手把手带我走上了科研道路。

关于我的毕业设计，首先要感谢我的导师付琨老师，以及同部门的孙显老师。我与二部结缘于大二的科研实践，两年的科研实践以及毕业设计期间，两位老师给了我无限的鼓励和自由的发展空间，对我一个小小的本科生也非常关心呵护。我在二部参与了不同方向的实践，在两位老师的指导下，我确定了未来的深造方向。他们对工作和教育的热情洋溢也深深感染了我，让我坚定了我的科研梦。在此感谢付老师和孙老师对我的栽培之恩。

在我的毕业设计期间，袁志强师兄给了我非常多的指导和帮助，跟着袁师兄学习的这几个月让我受益匪浅。袁师兄治学严谨，科研能力出众，工作勤奋，是我的好榜样。他对我无私的指导让我学到了很多东西，没有袁师兄就没有这篇论文的顺利完成。以及感谢田雨师兄给我的帮助和指导，毕业设计开题之时得到了田师兄的很多帮助，在毕设过程中师兄也一直关心我的进度和实验情况。感谢实验室里的师兄师姐，他们对我非常热情友善，是我的良师益友，在这里的几个月我感到发自内心的充实和快乐。

此外还要感谢我的父母和哥嫂，他们对我学业的支持是我坚强的后盾，在家人的关心和爱中我才能跋涉至今。感谢我的男朋友马金戈对我的陪伴和支持，本科期间他的乐观在我迷惘的时候鼓舞了我，申请季我们相互支持，最终一同守得云开见月明。

再次感谢我的老师家人和朋友们，希望我不坠青云之志，不负国科大之名。

2021 年 6 月