**“看图说话”——Image Caption方法的改进和微调**

深度学习第12小组

小组成员：戴子琦 史海秀 汪士栋 吴典儒 宗子睿

**摘要**

随着深度学习技术的发展，图像描述（Image Caption）任务在计算机视觉与自然语言处理交叉领域受到了广泛关注。本报告旨在改进和微调现有的“看图说话”（Image Caption）方法，通过优化Encoder-Decoder架构，引入注意力机制与预训练模型，提升模型在复杂图像语境中的表现能力与泛化能力。

本报告回顾了图像描述任务的研究历程，包括基于模板匹配的方法、基于检索的方法以及近年来流行的Encoder-Decoder框架。在此基础上，本报告的Baseline Model采用卷积神经网络（CNN）和循环神经网络（RNN）组成的Encoder-Decoder框架，通过随机梯度下降（SGD）进行训练，并通过束搜索（Beam Search）等技术进行推理。然而，该模型在处理复杂图像时存在局限性，无法充分关注图像的局部特征。

为了解决这一问题，本报告尝试了基于Transformer的模型，通过缩放点积注意力机制使模型能够选择性关注输入的内容。进一步地，本文使用了Spatial and Channel-wise Attention Transformer（SCA Transformer），在视觉空间中结合了空间和通道维度的注意力，从而更加充分地学习图像特征。

此外，本报告还探讨了预训练模型在图像描述任务中的应用。通过利用在大规模数据集上预训练的CLIP与GPT2模型，小组构建了一个高效的图像描述生成系统。该系统利用CLIP的图像编码器提取图像特征，并通过GPT2作为文本生成器，实现了高质量的图像描述生成。

在实验部分，本报告在Flickr8k和Flickr30k数据集上对提出的模型进行了评估。实验结果表明，与Baseline Model相比，基于Transformer的模型和预训练模型在BLEU、ROUGE、METEOR和CIDEr等评价指标上均有显著提升。特别是在引入SCA模块后，模型在描述主题上的准确性得到了进一步提高。

综上所述，本报告通过改进和微调现有的Image Caption方法，提出了一系列有效的模型结构，并在实验中取得了显著的性能提升

**关键词:** 图像描述；Encoder-Decoder架构；注意力机制；SCA Transformer; CLIP; GPT2

**一、引言**

在当今数字化时代，图像描述（Image Caption）作为计算机视觉与自然语言处理的交叉领域任务，正在经历着深刻的变革与发展。本文聚焦于改进和微调“看图说话”（Image Caption）方法，探索了基于Encoder-Decoder架构与预训练模型的创新应用。

图像描述不仅要求对图像内容进行深入的语义理解，还要能够生成符合自然语言习惯的流畅描述，因而具有重要的理论意义和广泛的应用前景。传统的方法往往依赖于先验的句式模板或图像检索库，但这些方法在灵活性、多样性和泛用性上存在明显的局限性。随着深度学习技术的进步，特别是Seq2Seq模型和Transformer架构的引入，Encoder-Decoder框架在图像描述任务中表现出了巨大的潜力与优越性。

本报告的创新点主要集中在优化Encoder-Decoder架构，特别是通过引入注意力机制与预训练模型，提升了模型在复杂图像语境中的表现能力与泛化能力。我们基于最新的研究成果和方法，对传统的模型进行了改进，探索了多模态Transformer、空间与通道注意力机制等新技术的应用，以期在提升图像描述生成质量的同时，扩展其在实际应用中的适用范围。

**二、文献综述**

图像描述(1mage Caption, IC)是一种结合计算机视觉和自然语言处理的跨模态跨领域的任务。一般地，它将生成与输入的图像内容相关的自然语言描述。因此不仅要求完备的图像语义理解，还要求复杂的符合人类表达的文本输出，具备重要的理论意义和应用前景。

针对图像描述任务的早期研究可以主要分为两类。一类是基于模板匹配的方法。这类方法通过对语法和语义的约束过程来生成图像描述。通常，该方法先检测出图像中特定的一系列视觉概念，然后通过语句模板、特定的语法规则或组合优化算法等将这些概念进行连接以生成描述语句。Yang等人[1]利用四元组作为生成图像描述的语句模板。首先使用检测算法评估图像中的物体和场景，然后利用语言模型预测可能用于构成描述语句的动词、场景和介词，并利用隐马尔可夫模型推理得到最佳的四元组，最后通过填充四元组给出的语句结构生成图像描述。

另一类是基于检索的方法。该类方法预先定义一批与目标图像视觉上相似的图像-文本匹配语料集合，利用相似度匹配的方法检索出1组语句作为该图像的描述语句。Farhadi等[2]提出基于三元组构建语义空间以连接图像和语句。给定查询图像，该方法通过求解马尔可夫随机场将给定图像映射到语义空问，并使用相应的相似度度量措施计算给定图像与现有语句的距离，将与给定图像最相似的语句作为相应的描述。

但这些早期方法需要依赖于先验的句子模板或图像检索池，在准确率不够优秀的同时也存在缺乏灵活性、多样性和泛用性的问题。

自Seq2Seq模型提出以来，Encoder-Decoder框架在机器翻译领域取得了巨大的成功，这也使得研究者开始将这个框架应用到图像描述任务中。该框架中编码器通常由多层卷积神经网络(CNN)构成，对图像进行视觉信息的特征提取。解码器则由循环神经网络(RNN)构成。

Vinyals等人[3]受机器翻译模型的启发提出了神经图像描述生成(neural image caption generator，NIC)方法。NIC将Inception v3作为CNN提取图像特征，用LSTM获得更好的全局语义生成描述。编码器使用新方法进行批量归一化，并将CNN最后的隐藏层输出作为解码器LSTM的输入。

Karpathy等人[4]提出了一种架构更简单的深度视觉语义对齐模型Neural Talk，其基本思想是假定语句的某些部分对应的是图像中特定的未知区域。该模型利用图像区域特征提取器VGGnet、语句双向RNN和对齐2种模态的结构化目标来推断语句片段和图像区域之间的对齐。模型将文本和图像区域被映射到公共的多模态嵌入空间中，然后利用多模态RNN推断出视觉和语义的对齐并生成新的描述语句。

随着Transformer架构的提出，注意力机制，尤其是自注意力机制在自然语言处理中的强势表现获得了广泛的关注，这也吸引了图像描述研究者的兴趣。使用注意力机制能够在利用图像特征引导生成描述时对特征的局部使用有所侧重，这点是常规的Encoder-Decoder框架难以做到的。

Xu等人[5]在NIC模型的基础上把注意力机制应用在图像描述的图像特征中。其基本思想是将编码阶段获取的图像特征进行注意力处理，提出随机性硬注意机制和确定性软注意机制的两种不同方法来产生注意信号。

Chen等人[6]提出SCA-CNN(spatial and channel-wise attention CNN)，在CNN中融合了空间注意和通道注意。在生成图像描述的过程中，SCA-CNN在多层特征图中动态地调整多层特征图和生成语句的上下文，对视觉注意的位置(即多层特征图中关注的空间位置)和内容(即关注的通道)进行编码。

Anderson等人[7]提出了Top-down、Bottom-up机制。该模型编码器改进了目标检测算法Faster R-CNN来进行提取图像的区域特征。Bottom-up机制用来提取图像各区域的特征向量，Top-down的机制用来分配Bottom-up提取的特征对文本描述的贡献度，两者提取到的特征组合得到联合注意力权重。在解码阶段使用一个双层LSTM模型，分别是Attention LSTM和Language LSTM。

Huang等人[8]加入了AoA（Attention on Attention）模块应用在编码器和解码器中。该模块增加另一种注意来扩展了常规注意机制。该模块通过两个线性变换生成一个“信息向量”和一个“注意门”，通过使用逐元素乘法将注意门应用于信息向量来增加另一个注意力，得到“关注信息”，即预期的有用知识。

Yu等[9]首次将Transformer整体结构引入图像描述模型，提出多模态Transformer模型，使模型能在统一的注意力机制中获取信息。

Guo等人[10]对自注意机制进行改进：首先，提出了规范自注意模块，对自注意重新参数化，有益于自注意内部的规范化；其次，提出了几何感知自注意模块，使自注意能够明确有效地考虑图像中物体之间的相对几何关系，从而弥补Transformer模型无法对输入物体的几何结构进行建模的局限性。

随着可用于视觉语言任务的数据集规模的扩大，对训练目标要求的提高，针对具体的数据集与任务需求单独训练模型对研究和商业应用渐渐变得难以负担，预训练模型的出现缓解了这个问题。通过在大规模数据集上对模型架构进行训练，再针对具体的应用场景与任务类别进行微调，已经变为了一种研究与应用的范式。因为预训练模型能在大规模的数据集上进行训练，所以往往在具体的任务中取得更好的效果。这种方法有利于扩展模型的适用范围，推动计算机视觉和自然语言处理技术在更广泛的应用场景中的应用。

Junnan Li等[11]提出BLIP预训练模型，采用Multimodal Mixture of Encoder-Decoder (MED)模型架构，旨在统一处理图像和文本信息，实现视觉-语言任务的理解和生成。采用自助法的思想，引入Captioner和Filter，BLIP实现了对高噪声的网络数据的利用，在多个视觉-语言任务上取得了先进的结果。

Alec Radford等[12]提出clip预训练模型，对于image,text数据对，采用对比学习的范式训练得到image encoder和text encoder。此结构在训练时已经将图像和语言强耦合，所以clip不仅在多个视觉-语言任务上表现出色，且泛化性能非常好，在不少数据集上未经微调也取得有竞争力的结果。

由于图像内容的复杂性, 精确地标注图像数据常常是不切实际的, 因而如何充分利用无标签图像数据成为重要问题。同时, 利用强化学习克服了神经网络梯度消失和梯度爆炸问题后, 将其与无监督学习方法进行模型优化成为自然的思路, 也已经在图像描述生成的任务中获得了良好的效果。基于强化学习的图像描述生成主要针对评估指标进行优化, 能够提高模型的整体评估得分。

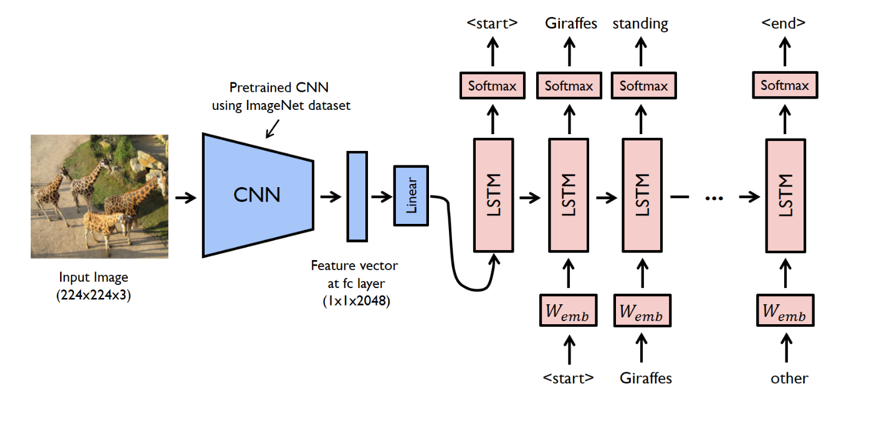
此外, 近年来风格化图像描述(Stylized Image Captioning)的任务开始逐渐收到关注。该任务旨在存准确描述图像内容的前提下, 将语言风格融人描述中, 如积极、消极、幽默和浪漫等语言风格。模型依赖于成对的风格文本语料库, 数据获取代价较高。因此, 擅长无监督学习的生成式对抗网络(Generative Adversarial Networks, GAN)开始成为一个新的热点, 使用GAN来训练风格化描述的模型逐渐增多。

**三、模型结构**

(详细介绍各个模型的提出动机与原理架构，包括数学公式和架构图)

**1. Baseline Model**

（1）模型架构：

 自动描述图像内容需要识别图片并用自然语言表达对象的关系，这意味着我们既需要视觉理解，也需要语言模型。Baseline模型的架构基于Encode-Decoder框架，由一个视觉卷积神经网络（CNN）和一个语言生成循环神经网络（RNN）组成。模型的灵感来自于机器翻译领域的最新发展，特别是使用循环神经网络进行序列建模的方法。下图展示了Baseline模型的架构图：

（2）模型原理：

模型使用以下方法来定义和训练：，其中I是图像，S是其正确的描述。利用链式法则展开句子S的联合概率这里N是特定示例的长度。RNN用于建模条件概率，其中表示固定长度的隐藏状态或记忆。LSTM用于解决RNN中的梯度消失或爆炸问题，其核心是记忆单元c和控制单元行为的“门”。LSTM的门控制和单元更新公式为：

输入门；遗忘门；更新门；单元更新：；记忆输出；下一个词的概率分布

这些公式定义了如何训练模型以最大化给定图像的目标描述句子的可能性。模型通过随机梯度下降（SGD）进行训练，并通过束搜索（Beam Search）等技术进行推理以生成句子。

**2. Transformer based Model**

**2.1 Base Transformer**

(1) Baseline中存在的问题

Baseline已经能够在一定程度上跨多模态进行简单的image caption.但架构中仍存在着如下问题：Baseline中使用了Resnet50和LSTM作为Image Encoder和Caption Decoder. 但该模型框架在为图像生成描述时无法随着时间的推移对特征的使用有所侧重。同时通常将视觉场景作为一个整体来考虑，没有考虑与描述语句部分相关的图像局部内容。此时利用注意力机制或者更进一步地结合Transformer block结构来引导图像描述生成，能够很好地缓解这个问题。

(2) 注意力机制

Transformer 采用缩放点积注意力,使模型能选择性关注输入的内容：

注意力机制函数的三个自变量分别为查询矩阵Q、键矩阵K和值矩阵V. 在自注意力机制中，这三个矩阵通过同一个输入进过线性变换产生，具有并行计算的优势。同时，考虑到每一个自注意力模块捕捉到的信息有限，使用包含h个头的多头注意力机制在不同的特征子空间内提取信息，从而获得更多的知识：

其中Concat代表按特征维度拼接，代表单头注意力机制。

(3) 模型框架改进

Base Transformer保留整体的Encoder-Decoder架构，但针对其中的各个block进行了修改——将Encoder-Decoder架构中的block替换为相应的Transformer block。同样利用Resnet50模型预先提取图片的特征图通过线性映射，得到的结果作为Image Caption的输入，而将文本内容利用embbeding加上位置编码作为Caption Decoder的输入。

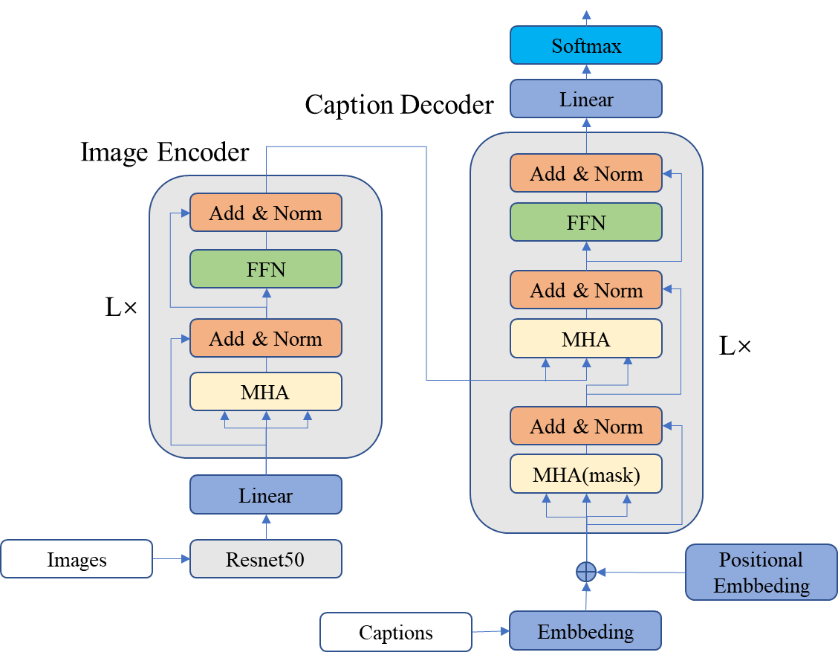


图 Base Transformer架构图

**2.2 Spatial and Channel-wise Attention Transformer**

(1) Base Transformer中存在的问题

Base Transformer架构中的Image Encoder的输入同样使用Resnet50提取出的特征图。这些视觉特征自然是空间的，通道的和多层的。然而Base Transformer架构中的Image Encoder中使用的注意力只有特征维度加权的空间注意力。仅仅使用这种注意力难以全面有效对视觉特征进行关注，从而有时难以精确地发挥出视觉特征的引导作用，甚至会产生较大的偏差。

(2) Spatial and Channel-wise Attention模块(简记为SCA)

按照不同的顺序（例如Channel-Spatial或者Spatial-Channel, 将先进行Channel-wise Attention后进行Spatial Attention的模块简记为C-S顺序）以此在不同方向上对特征图计算相应的注意力得分并进行叠加，从而获得最后的注意力结果。SCA模块中同样包括了自注意力的计算以及多头机制的处理以获得更多的信息。

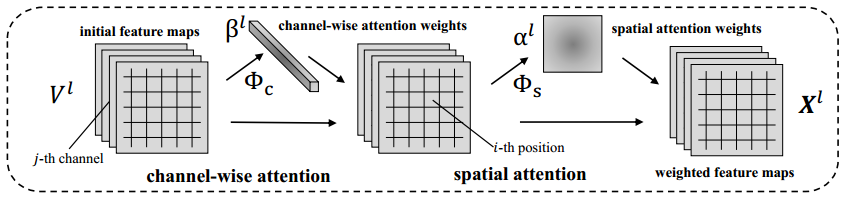


图 SCA具体原理(以C-S顺序为例)

(3) 模型框架改进

在Base Transformer架构中的Image Encoder中引入Spatial and Channel-wise Attention模块(简记为SCA). 它能够在模型学习图像局部特征时结合空间和通道的注意力，同时对视觉注意的位置(即多层特征图中关注的空间位置)和内容(即关注的通道)进行编码，从而充分利用图像的信息引导后续Caption Decoder. 剩余的大致架构与Base Transformer框架相似。

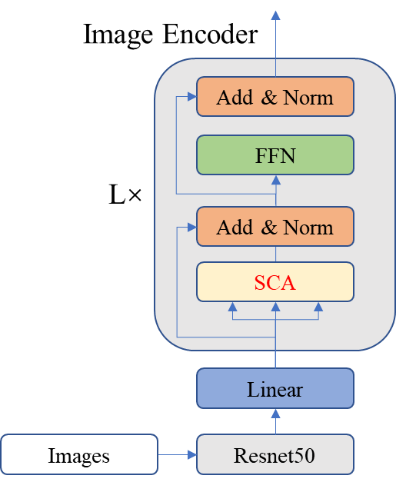


图 SCA Transformer的Image Encoder架构图

**3. Pre-trained Model**

（1）预训练模型

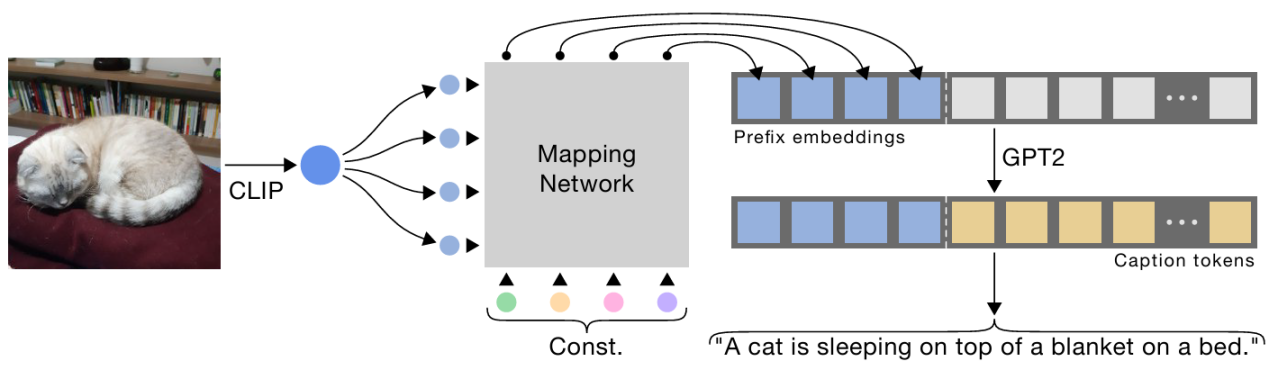
在我们的结构中使用到了两个预训练模型：CLIP与GPT2。

Clip(Contrastive Language-Image Pretraining)是由Open AI开发，采用对比学习的范式训练出的用于理解图像与语义之间关联的深度学习模型。以最大化正样本（对角线）相似度，最小化负样本（非对角线）相似度为目标共同去优化图像与文本的encoder。该模型训练时将图像和自然语言相耦合，模型迁移效果特别好。

GPT2是由OpenAI开发，基于transformer的大语言模型，对语言有很好的表征能力。

（2）模型架构

因为clip强大的图像理解能力，我们使用clip中的image encoder理解图像；因为GPT2强大的语言生成能力，我们使用GPT2作为文本生成器。在两个模型之间再采用映射层（MLP或transformer）进行衔接。模型确切结构如下：将图像输入clip的image encoder，输出固定长度的序列；将结果输入映射层，输出固定长度的嵌入层；将其作为embedding输入GPT2，输出描述图像的内容。

图 预训练模型架构

（3）预训练模型的优势

前述模型针对视觉特征的处理做了改进，但任务的最终目标仍然是需要用语言来描述图形内容，而Clip在训练时就与文本强耦合，在图生文的任务中使用图像编码器会更有针对性。且预训练模型能够在很大的数据集上进行训练，一般具有较强的表达能力。

**四、数据集介绍及处理**

Flickr8k和Flickr30k是两个广泛应用于图像描述生成任务的数据集。Flickr8k包含8000张图像，每张图像由5个独立的描述语句注释 。相比之下，Flickr30k规模更大，包含31783张图像，同样每张图像由5个描述语句注释 。这两个数据集的图像和描述都来源于Flickr网站，因其丰富的内容和多样性，被广泛用于研究视觉和语言处理的交叉领域。

在处理这两个数据集时，我们采用了一系列预处理步骤，以确保数据的一致性和质量。这些步骤包括：

1. **加载数据**：首先，我们读取Karpathy提供的JSON文件，该文件包含图像的信息和对应的描述语句。通过指定数据集的路径，我们可以灵活地选择处理Flickr8k或Flickr30k数据集。
2. **构建词汇表**：我们统计了所有非测试集中的词汇出现频率，并过滤掉那些在数据集中出现次数少于指定阈值（例如5次）的词汇。词汇表中还包含特殊标记如<pad>、<unk>、<start>和<end>，这些标记在后续的描述编码过程中起关键作用。
3. **处理图像和描述**：对于每张图像，我们首先检查其是否能够正常打开。随后，根据指定的每张图像描述数量（默认为5），我们从所有符合长度要求（不超过30个词）的描述中进行选择或随机填充，确保每张图像都有5个描述。在对描述进行编码时，我们在每个描述的开头和结尾分别添加<start>和<end>标记，并用词汇表中的索引替代单词。
4. **保存处理后的数据**：处理完成后，我们将数据分为训练集、验证集和测试集，并分别保存为JSON文件。这些文件包含图像路径和对应的编码描述，便于后续的模型训练和评估。

通过上述步骤，我们能够有效地预处理Flickr8k和Flickr30k数据集，为图像描述生成模型的训练提供高质量的数据支持。这个预处理过程不仅确保了数据的一致性和规范性，还为后续的模型开发打下了坚实的基础。

**五、评价指标介绍**

（数学公式，含义，表格）

1、BLEU (Bilingual Evaluation Understudy)

BLEU是一种常用于机器翻译的评价指标，通过计算生成文本与参考文本之间的n-gram重合度来评估翻译质量。其中N-gram重合度是计算生成文本和参考文本中相同长度的n-gram（常用1到4个词的n-gram）的重合情况。

BLEU的计算方式如下。它通过计算精确度和长度惩罚来得到最后的结果。

精确度（Precision）是生成文本中n-gram与参考文本中n-gram匹配的比例。其公式为。

其中，为各阶N-gram的精度，表示第k个N-gram在待评句中出现的次数，表示第k个N-gram在标准参考句中出现的次数，表示某N-gram在多条标准参考句中出现最多的次数，分子整体表示取N-gram在待评句和参考句中出现的最小次数。由于阶数升高会导致N-gram统计量的精度减小，为平衡各阶统计量的作用，取其加权平均值，再乘以惩罚因子，改进相应的公式。

长度惩罚（Brevity Penalty, BP）是为防止生成文本过短而计算的指标。长度惩罚公式为。

其中，c是生成文本的长度，r是参考文本的长度。

结合精确度和长度惩罚，最终的BLEU公式为。

其中，是n-gram的精确度，是权重。

2、ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE是一组评价指标，主要用于摘要和生成文本的评估，特别关注召回率。常用变体有ROUGE-N、ROUGE-L、ROUGE-W、ROUGE-S。ROUGE-L。本实验主要用的指标是ROUGE-L,它是基于最长公共子序列（LCS）的重合度，考虑了词序，我们将着重介绍这一指标。

ROUGE-L是基于LCS来评价待评句和参考句之间的相似性,以待评句 X 和参考句Y 为例,其公式如下。

其中，, , m和n分别代表X和Y的长度,

3、METEOR (Metric for Evaluation of Translation with Explicit Ordering)

METEOR是一种基于单字召回率和单精度的加权调和平均数针对机器翻译的评价指标，考虑了词形变化和词序。另外，该指标增加了基于 WordNet的同义词库以解决同义词匹配的问题。

其中为结合精确度和召回率计算的调和平均数，公式为：

其中，P是精确度，R是召回率。Penalty是惩罚系数。

4、CIDEr (Consensus-based Image Description Evaluation)

CIDER是将句子看作文档处理,使用 TF-IDF计算其单词权重,从而把句子表示成向量形式,通过计算待评句和参考句的 TF-IDF 向量的余弦距离来计算两者之间的相似性。

其计算方式为

其中,和分别是待评句和参考句集合。

对于不同n值，余弦相似度平均值为：

TF-IDF 的权重g\_n(·）的计算方式如下：

其中,ω表示不同的N-grams,(·)表示ω 在待评句或参考句中出现的次数,Ω是N-grams字表,I是图像集合。

**六、实验设计与结果分析**

**1. 模型结构**

(超参数设置，搜参的过程(如果有))

* 1. **Baseline Model**

1. EncodeCNN:经过预训练的ResNet50
2. DecoderRNN:

batch\_size = 10; embed\_size = 512; hidden\_size = 2048

num\_decoder\_layers = 4

其中batch\_size表示样本批量; embed\_size表示RNN block的输入维度; hidden\_size表示全连接层的维度; num\_decoder\_layers分别表示一个block中Caption Decoder的层数.

**1.2 Transformer based Model**

(1) Base Transformer超参数

使用的注意力机制为Spatial Attention.

batch\_size = 30; embed\_size = 512; d\_model = 512; hidden\_size = 2048

nhead = 8; num\_encoder\_layers = 4; num\_decoder\_layers = 4

其中batch\_size表示样本批量; embed\_size表示Transformer block的输入维度; d\_model表示注意力机制的映射维度; hidden\_size表示全连接层的维度; nhead表示多头注意力机制的头数; num\_encoder\_layers和num\_decoder\_layers分别表示一个block中Image Encoder和Caption Decoder的层数.

此外，我们使用Adam优化器以及阶跃衰减的学习率调整方法。具体而言，将基本学习率设为, 其中t为从1开始的当前epcoh数。经过6个epcohs的迭代后，学习率每3个epochs衰减1/2.

(2) SCA Transformer超参数

使用的注意力机制为Channel-wise Attention & Spatial Attention. 同时在实验中尝试了C-S与S-C的两种不同顺序的模块，最终通过比较实验选择性能较优的注意力顺序为C-S顺序。

此外由于机器设备性能与内存的限制，缩小了批次大小batch\_size, 同时也将各特征层维度进行等比缩小，同时保持层数与注意力头数不变。

batch\_size = 10; embed\_size = 128; d\_model = 128; hidden\_size = 512

nhead = 8; num\_encoder\_layers = 4; num\_decoder\_layers = 4

优化器与学习率的设计与Base Transformer相同。

**1.3 Pre-trained Model**

Clip的基学习器可以是VIT或ResNet；映射层可选择MLP或者transformer。训练时不对clip进行微调，如果对GPT2进行微调，则映射层仅需要MLP即可；如不对GPT2进行微调，则映射层需要transformer，这样就得到四个模型。

四个模型的超参数都如下：

GPT2 embedding长度：Prefix\_lenth=40

映射层网络深度：num\_layers=8

Batch\_size=40; learning\_rate=2e-5

**2. 训练成本**

(模型参数(包含模型总参数量和训练参数量，训练参数量主要针对预训练模型)，训练设备，训练时间

这点建议做成表格(我们训练的设备可能都不一样，所以整成某个模型用了什么设备，有多少参数，训了多少时间这样))

**2.1 Baseline Model**

总参数量：31639743

训练参数量：13193583

训练设备：kaggle P100

训练时间：约9min30s/epoch

**2.2 Transformer based Model**

(1)Base Transformer

总参数量：61625319

训练参数量：38117287

训练设备：2 cores 16GB内存 GPU P100和2 cores 15GB内存 GPU T4×2

训练时间：约8min30s/epoch

(2)SCA Transformer

总参数量：28601319

训练参数量：5093287

训练设备：2 cores 16GB内存 GPU P100和2 cores 15GB内存 GPU T4×2

训练时间：约14min/epoch

**2.3 Pre-trained Model**

**训练设备：RTX 4090**

1. VIT transformer

总参数量：329305089

训练参数量：53587968

训练时间：2min/epoch

（2）VIT MLP

总参数量：307185921

训练参数量：155908608

训练时间：2min/epoch

（3）Resnet transformer

总参数量：333237249

训练参数量：57520128

训练时间：1min40s/epoch

（4）Resnet MLP

总参数量：307677441

训练参数量：156400128

训练时间：1min40s/epoch

**3. 模型效果**

(评价指标，对指标的结果分析。)

见展示PPT中的结果表格和分析。

**4. 典型测试集case分析**

(典型测试集case分析，最后用一张自定义照片跑一下最优模型)

选取测试集中的两张随机图片，利用训练好的各个模型架构中的最优模型结果进行image caption生成, 观察各模型的生成效果并进行分析。



图 Base Transformer Case



图 SCA Transformer Case



1. **总结与展望**

**1、总结**

相比于Baseline, 利用Transformer block使模型在生成Caption时能够利用注意力机制关注相应的局部特征。同时引入SCA进行改进，在视觉空间中结合了空间和通道维度的注意力，能够更加充分地学习图像特征。最终模型在各个指标上均有所提升，同时在除BLUE1外的指标上提升都较明显。说明模型已经能够生成在短句、词序合理以及语句层面上充分完整的描述。引入SCA模块后能够在描述主题上有所帮助。

Clip模型训练时已经将图像与文本强耦合，因此在Image Caption任务中，利用其图像编码器对图像特征进行提取会非常契合任务场景；利用GPT模型强大的语言能力能较好的根据提取的图像特征进行语言上的描绘；经过映射层的连接可以较好地适配两个模型。模型的结构简单，训练成本较低；模型的拟合能力很强，8k的数据量不能完全发挥其潜力；引入更多的数据量能使模型在更宏观的语句、词义以及主题层面获得进一步提升。

2、不足与改进

Transformer based model中对图像特征的初步提取使用了Resnet50，即直接利用了图像的全局特征进行注意力计算，但通常是图片中的关键目标影响到描述生成效果。仅使用图像全局特征进行注意力运算可能难以把控关键对象特征的细节。后续可以尝试利用相应的目标检测模型（例如Faster R-CNN）来替换Resnet50.

尝试将Clip引入图像的特征提取中，作为辅助器提取跟语言更高相关的特征。例如在Clip之前引入Faster R-CNN，让Clip模型对主要的目标分别进行特征提取。

此外，图像描述任务的实际应用中不仅需要提高准确性, 它也需要从描述多样性、可控性和特定领域适用性等方面考虑。而我们的模型在这些方面可能存在一定的不足。后续可以尝试使用和设计这些方面的评价指标，以及探索强化学习和生成学习的Image Caption模型方法。

**八、参考文献**

[1] Yang Yezhou, Teo C L, Daume H, et a1. Corpus-guided sentence generation of natural images[C] //Proc of the 2011 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2011: 444-454

[2] Farhadi A, Hejrati M, Sadeghi M A, et a1. Every picture tells a story: Generating sentences from images [G] //LNCS 6314: Proc of the European Conf on Computer Vision. Berlin: Springer, 2010: 15-29

[3] Vinyals O, Toshev A, Bengio S, et a1. Show and tell: A neural image caption generator[C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE Computer Society, 2015: 3156-3164

[4] Karpathy A, Li Feifei. Deep visual-semantic alignments for generating image descriptions[C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE Computer Society, 2015: 3128-3137

[5] Xu K, Ba J L, Kiros R, et a1. Show, attend and tell: Neural image caption generation with visual attention[C] //Proc of the 32nd Int Conf on Machine Learning. Cambridge, MA: MIT Press, 2015: 2048-2057

[6] Chen Long, Zhang Hanwang, Xiao Jun, et a1. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning[C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE Computer Society, 2017: 6298-6306

[7] Anderson P, He Xiaodong, Buehler C. et a1. Bottom-up and top-down attention for image captioning and visual question answering[C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE Computer Society, 2018: 6077-6086

[8] Huang Lun, Wang Wenmin, Chen Jie, et a1. Attention on attention for image captioning[C] //Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 4633-4642

[9] Yu J, Li J, Yu Z, et a1. Multimodal transformer with multi-view visual representation for image captioning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(12): 4467-4480.

[10] Guo Longteng, Liu Jing, Zhu Xinxin, et a1. Normalized and geometry-aware self-attention network for image captioning[C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE Computer Society, 2020: 10324-10333

预训练

[11] Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. ICML.

[12] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision.

评价指标

[13] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation[C] ∥Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Istanbul, Turkey: Association for Computational Linguistics, 2002: 311-318.

[14] BANERJEE S, LAVIE A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments[C] ∥Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor, Michigan: ACL, 2005: 65-72.

[15] LIN C Y. Automatic evaluation of summaries using n-gram cooccurrence statistics[C] ∥Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. United States: Association for Computational Linguistics, 2003: 71-78.

[16] VEDANTAM R, ZITNICK C L, PARIKH D. CIDEr: Consensus-based image description evaluation[C] ∥IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015). Boston, MA, United States: IEEE Computer Society, 2015: 4566 4575.

**附录**

附录1-代码

附录2-图表