

****

**数据挖掘课程2023年春**

**期末大作业**

|  |  |
| --- | --- |
| 院（系）名称 | XXXX |
| 团队名称 | XXXX |
| 团队学生 | XXXX |
| 指导教师 | XXXX |

2023年X月

目录

[第一章 背景 3](#_Toc104844310)

[第二章 XXX 3](#_Toc104844311)

[第三章 XXX 3](#_Toc104844312)

[第四章 收获、体会及建议 3](#_Toc104844313)

# 第一章 背景

文生图模型是一类多模态生成式模型，它可以根据文本描述（prompt）生成与描述匹配的图像，其本质是自然语言到图像之间的映射。近年来，随着DALL-E2、Imagen和Stable Diffusion等文生图模型的发展，高质量的图片被不断生成。然而，如何评价模型生成图片的质量成为生成式模型领域具有挑战性的问题，需要综合考虑语义相关性、图像美观性、图像创造力等多方面。因此，建立一套有效的图片自动评价机制显得十分必要，这样做不仅可以克服人工评价图片成本高、主观性强的缺点，而且可以帮助进一步提升文生图模型性能。

为了解决上述问题，本项目提出了一种多维度图片评价器。该评价器基于对比学习与图片复杂度，从准确性和美观性两个维度出发，采用相对评价的方法比较两张图片中哪张质量更高。

# 第二章 原理

2.1 对比学习

对比学习（Contrastive Learning）是一种基于对比思想的判别式表示学习方法，该方法通过比较不同样本之间的相似性来学习模型。在对比学习中，数据一般被分为三类：锚点（anchor）、正样本（positive example）与负样本（negative example）。锚点指所比较样本之间的固定参照点，通过设置锚点，模型可以更准确地两个样本之间的相似性；正样本指与锚点相似的样本；负样本指与锚点不相似的样本。具体而言，对比学习通过减小特定损失函数，在向量表征空间中将正样本与锚点之间的距离拉近，将负样本与锚点之间的距离拉远，从而很好地实现正负样本的分类。

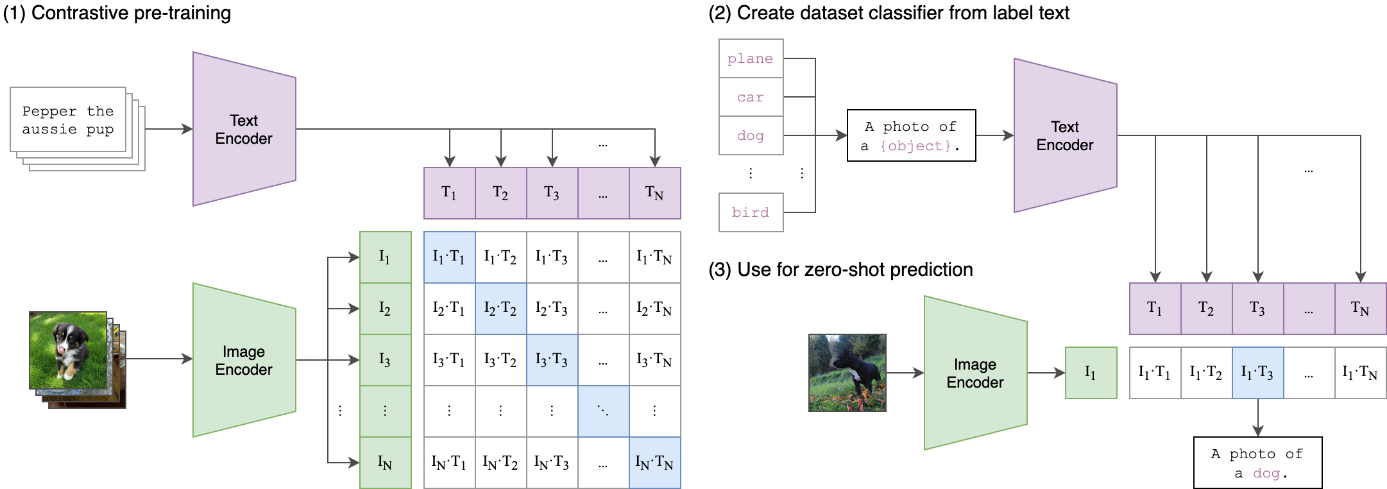
InfoNCE Loss是一种常用的对比学习损失函数，其公式如下：

其中为向量和之间的余弦相似度，为样本总数，代表正样本，代表锚点，为温度系数，是用于控制模型对负样本区分度的超参数。

2.2 CLIP模型

CLIP（Contrastive Language-Image Pre-Training）是OpenAI团队在2021年发布的用于匹配图像和文本的预训练神经网络模型。该模型采用对比学习的思想，同时对文本和图像进行处理，将其在向量空间表示，使得相似的文本和图像向量在向量空间中更加接近。

如图所示，CLIP主要由文本编码器（Text encoder）和图像编码器（Image encoder）组成。文本编码器基于Transformer模型，用以捕捉文本的上下文信息，将输入的文本序列转化为一个固定维度的向量；图像编码器用以捕捉图像中的不同特征信息，将输入的图像转化为一个固定维度向量。之后，为了便于比较图片向量与文本向量，CLIP将它们映射到联合多模态空间（Joint multimodal sapce）。最后，通过计算图片向量与文本向量之间的余弦相似度，CLIP使用对比学习训练模型。



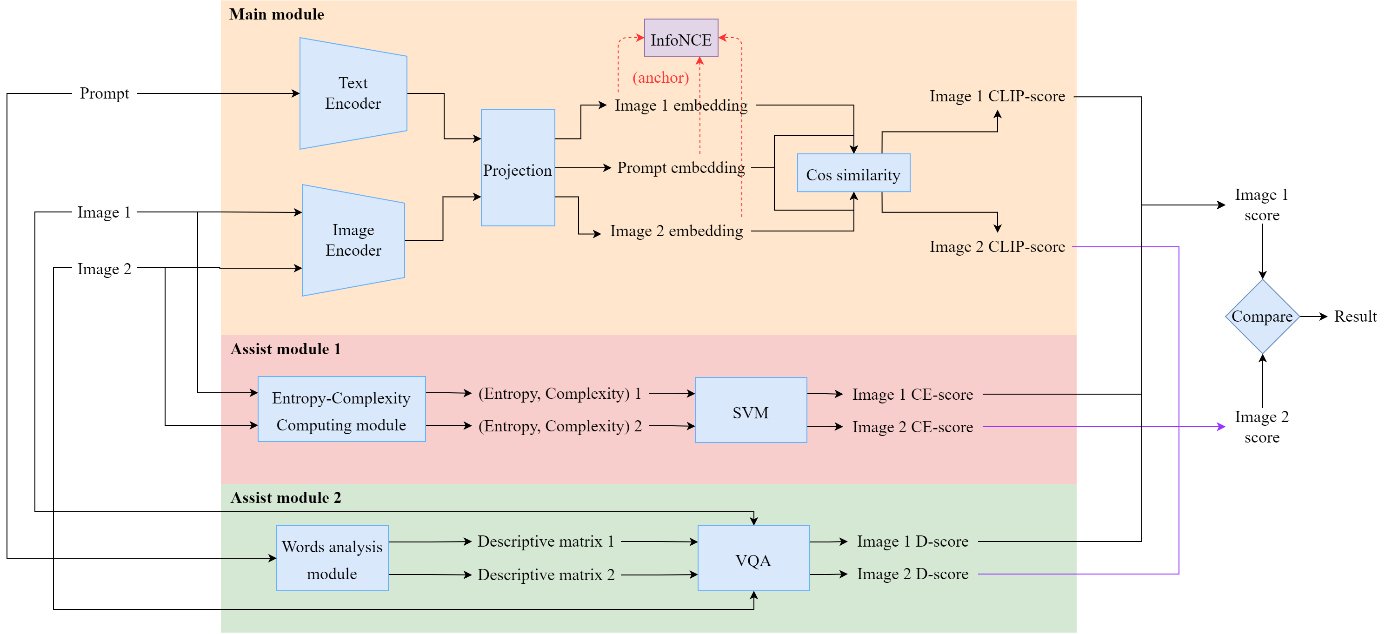
2.3 复杂度衡量

熵（Entropy）可以用于计算一张图片的混乱程度，以描述图片的美观性和和谐性。对于一种概率分布，其香农熵（Shannon entropy）和归一化香农熵（Normalized Shannon entropy）分别为：

对于概率分布，其统计复杂度（Statistical complexity）则表述为如下公式：

其中为归一化常数，为概率分布与均匀分布之间的JS散度：

# 第三章 方法



3.1 整体框架

如图所示，本项目提出的多维度图片评价器由一个主体模块和两个辅助模块组成。主体模块从整体准确性维度出发，基于对比学习的思想，采用CLIP预训练模型对两张待比较图片Image 1和Image 2进行打分，得到CLIP-score作为图片质量的主要评价依据。辅助模块1从图片美观性维度出发，基于熵和复杂度，采用SVM模型对Image 1和Image 2进行打分，得到CE-score；辅助模块2从图片细节准确性维度出发，采用VQA模型对Image 2和Image 2进行打分，得到D-score。辅助模块得分将作为图片质量的次要评价依据，仅对主体模块得分进行小幅度修正。最终得分为CLIP-score、CE-score和D-score的加权和，计算公式如下：

3.2 基于对比学习的主体模块

对于文生图模型图片质量评价任务，我们将好图片视为对比学习中的正样本，将坏图片视为负样本，将描述文本prompt视为锚点。我们希望判断一张图片是否符合prompt，即在对比学习中将好图片与prompt之间的距离拉近，坏图片与prompt之间的距离拉远。

主体模块中采用CLIP网络加以InfoNCE Loss反向传播进行微调，具体步骤如下：

首先，我们使用CLIP的Text Encoder与Image Encoder将输入图片与prompt编码为低维特征向量

之后，将图片向量和文本向量传入两层512×512的线性层中，通过线性变换将CLIP所学习到的特征表示进一步转化为更具判别性的表示，帮助模型更好地区分不同类别之间的差异，以适应判别文生图的任务。

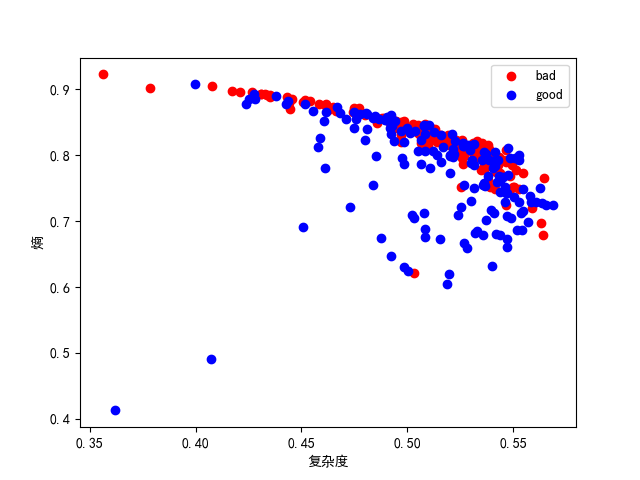
根据编码完成的图片样本嵌入向量与prompt嵌入向量计算InfoNCE Loss中，通过减小InfoNCE Loss，在向量表征空间中将正样本与锚点之间的距离拉近，将负样本与锚点之间的距离拉远，从而实现正负样本的分类，对比仅使用Clip进行对比学习只是用点积而言，我们的方法更具有解释性。

3.3 基于熵-复杂度的辅助模块1

辅助模块1首先计算图片的熵和复杂度。排列熵（Permutation entropy）是一种基于时间序列中的序数模式出现的概率分布的非线性复杂度度量。对于一张大小的图片，其排列熵计算方法如下：

* 首先类似于卷积的滑动窗口操作，使用大小为的滑动窗口采集到个形状的向量，之后将这些向量依次展开形成列向量。
* 对于每个列向量，将其映射为序数排名序列。如列向量，其序数排名序列为。
* 统计上述得到的个序数排名序列在所有可能的序数排名序列出现的概率，得到概率分布。其中，对于长度为的序列，其所有可能的序数排名序列有个。
* 计算的归一化香农熵，即为图片的排列熵。
* 计算的统计复杂度，作为图片的复杂度。

训练数据中好图片和坏图片在复杂度-熵平面中的分布如图所示。



从图中可知，坏图片的熵和复杂度与好图片相比更大，说明坏图片除了其语义关联性较差外，整体画面较为混乱、不协调，好图片可能更加简洁。

之后，采用SVM在训练数据的复杂度-熵平面中进行训练，将好图片和坏图片尽可能区分开。在验证或测试时，使用训练好的SVM模型预测两张待比较图片属于好图片的概率，将概率作为图片的CE-score。

3.4 基于对象分析的辅助模块2

辅助模块2首先对prompt进行词法分析处理，提取出其中的对象、颜色、对象关系等信息，形成一个描述矩阵，如下表所示

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Color | Number | Existence | Object 1 | … | Object n |
| Object 1 |  |  |  |  |  |  |
| … |  |  |  |  |  |  |
| Object n |  |  |  |  |  |  |

描述矩阵的大小为，可看作的矩阵与的矩阵相拼接得到。其中为prompt中的对象数量，为颜色等属性的数量，的矩阵描述了对象之间的关系。

之后将描述矩阵和图片输入到VQA模型中，根据描述矩阵和问题模板构建问题进行问答。对于对象，构造模板：“Can you find any {Object}?”；对于颜色属性，构造模板：“Is the {Object} {Color}?”；对于对象间的关系，构造模板：“Is {relation}?”。

根据描述矩阵的形式，将答案为“Yes”的概率构成得分矩阵。最后，取得分矩阵的均值作为图片的D-score。

# 第四章 实验

4.1 实验分析

本实验数据集采用智源FlagAI所提供的文生图数据，其中训练集共6040对图片，我们使用8：2比例随机划分训练集与验证集进行训练；测试集共有673对图片，最终通过融合模型给予一对图片好与坏的相对比较。

实验中我们对主干网络进行十轮次的微调，初始学习率设置为10e-5,每轮次乘系数0.5进行衰减以避免因学习率过大导致微调震荡。

4.2 实验结果

按照整体框架部分中评分标准对测试集图片进行得分对比输出result.csv, 所得的[FlagEval/README\_zh.md at master · FlagOpen/FlagEval · GitHub](https://github.com/FlagOpen/FlagEval/blob/master/imageEval/README_zh.md)中给出的

# 第五章 总结

5.1 优点：

我们的模型具有多尺度判断能力且可解释性强，主干网络中对比学习方法可以将正样本对与负样本对尽可能区分，熵的判断可以对图片风格进行量化比较，视觉问答模型对三个重要的实体（存在性，颜色，数量）进行判定，最终加权得到得分进行比较使其不仅做到准确度高，而且流程均可解释。

5.2 缺点：

局限于设备原因我与模型复杂度较高，们的模型在运行时需要较长的时间来完成推理或训练，长时间的运行可能会限制模型的实际运用与部署。模型没有经过全面的超参数搜索或权重组合优化过程，导致模型在当前状态下的性能尚未达到最优水平。

# 第六章 收获、体会及建议

在构建多维度图片评价器的过程中，我们有如下的收获与体会：

* 我们了解了文生图模型的发展历程、前景以及挑战。
* 我们深入了解并学习了对比学习方法，并掌握了一些针对大模型进行微调训练的手段，这有助于提高模型的准确率。
* 在完成一项任务中，我们可以从不同的角度进行尝试，以提高模型的可解释性。
* 通过对数据特征进行探索和分析，我们深入了解了数据科学的实践过程，为今后的项目积累了宝贵的经验。

针对本课程实验，我们希望能提供测试数据的标签，以便更好地评估模型效果。