北京航空航天大學
BEIHANG UNIVERSITY

数据挖掘课程 2023 年春

期末大作业

多维度图片评价器

院（系）名称	人工智能研究院
团队名称	主楼挖金
团队学生	修曾琪 张家瑞 魏少杭
指导教师	庄福振

2023 年 5 月

目录

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：

$$\mathcal{L}_{NCE} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\cos(q,k_+)/\tau)}{\sum_{i=1}^{N}\exp(\cos(q,k_i)/\tau)}$$

其中$\cos(q,k)=\frac{q\cdot k}{|q||k|}$为向量$q$和$k$之间的余弦相似度，$N$为样本总数，$k_+$代表正样本，$q$代表锚点，$\tau$为温度系数，是用于控制模型对负样本区分度的超参数。

在此文章[1]中有所证明，减小 InfoNCE Loss，等价于提升锚点和正样本（好图片）

的互信息的下限。换句话说，通过优化 InfoNCE 损失，我们可以将好图片和 prompt 样本的相关联程度进一步提高，这种关联不是简单的匹配，而是从信息论角度来讲的深层次关联匹配。由此，从理论上可以说明我们使用 InfoNCE 来微调 CLIP 模型是正确的。

2.2 CLIP 模型

CLIP[2]（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 Space）。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。

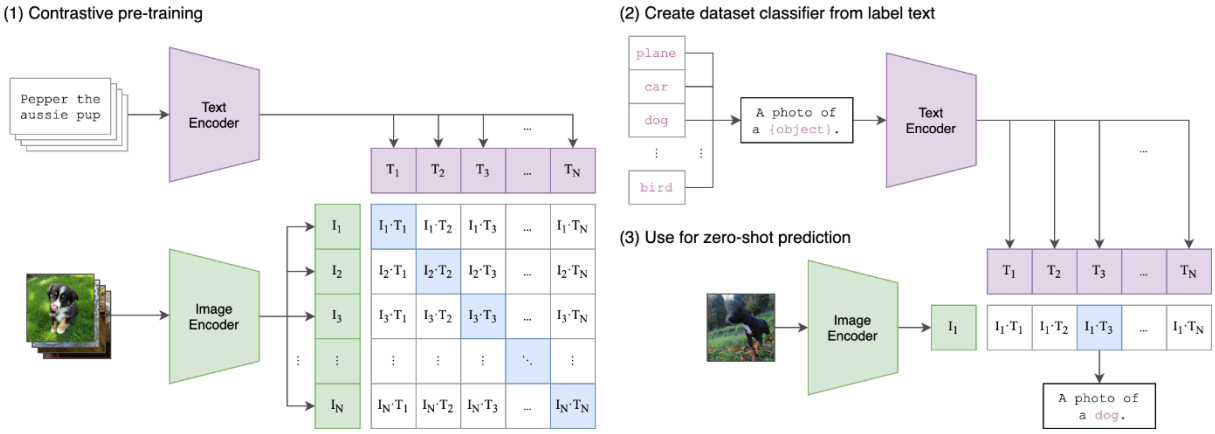


图 2.1 CLIP 框架流程图

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[3]。对于一种概率分布$P = \{p_i, i = 1, \dots, n\}$，其香农熵（Shannon entropy）$S(P)$和归一化香农熵（Normalized Shannon entropy）$H(P)$分别为：

$$S(P) = \sum_{i=1}^{n} p_i \ln \frac{1}{p_i}$$

$$H(P) = \frac{1}{\ln n} S(P)$$

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

$$C(P) = \frac{D(P,U)H(P)}{D^*}$$

其中D^*为归一化常数$-\frac{1}{2}[\frac{n+1}{n}\ln(n+1) + \ln n - 2\ln(2n)]$，$D(P,U)$为概率分布$P$与均匀分布$U = \{u_i = \frac{1}{n}, i = 1, \ldots, n\}$之间的 JS 散度：

$$D(P,U) = S\left(\frac{P+U}{2}\right) - \frac{S(P)}{2} - \frac{S(U)}{2}$$

第三章 方法

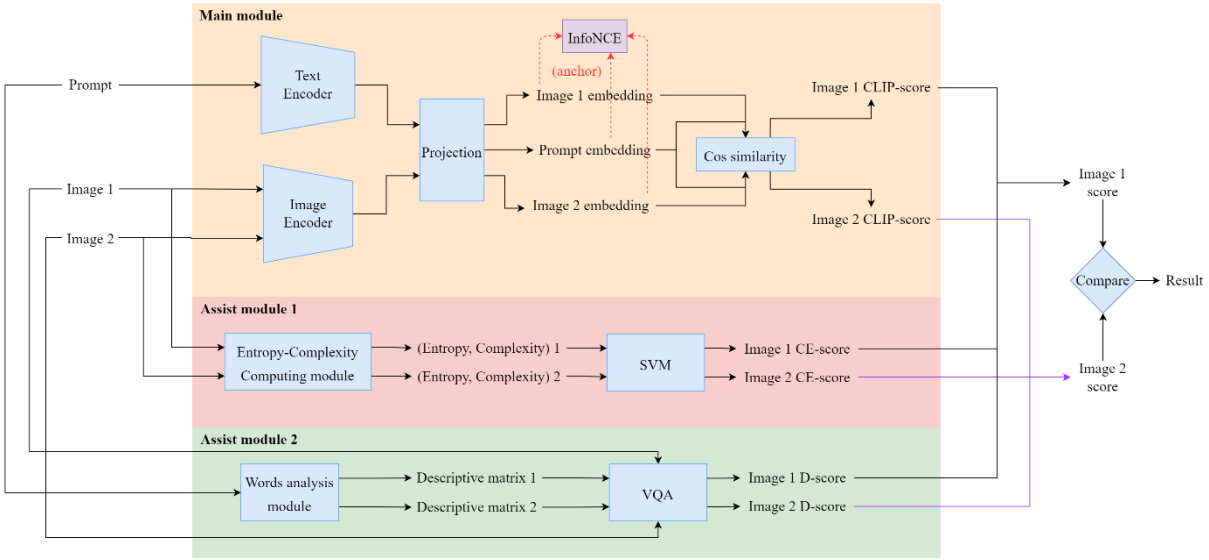


图 3.1 整体框架流程图

3.1 整体框架

如图 3.1 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：

$$final\ score = 0.95 \times CLIP\ score + 0.04 \times CE\ score + 0.01 \times D\ score$$

3.2 基于对比学习与 CLIP 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。如本报告前文所述，InfoNCE 损失优化过程实质上就是提升正样本对（prompt 和正样本）的互信息的下限，这与我们希望拉近正样本对之间的距离的愿望是一致的。

微调阶段——主体模块中采用预训练好的 CLIP 模型，以降低 InfoNCE 损失作为优化目标，通过反向传播进行微调，具体步骤如下：

1.我们使用 CLIP 的 Text Encoder 与 Image Encoder 将输入图片与 prompt 编码为相同长度的低维特征向量，这些向量共处于一个语义空间中。

2.将图片向量和文本向量传入两层 512×512 的可学习的线性层中，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。具体来说，这个可学习的线性变换经过学习能够将原始的语义空间中本质上能够导致好图片与 prompt 相似的高级语义特征提取出来并转入新的特征空间中，而忽视原始语义空间中 prompt 和好图片本质无关的特征。

3.针对线性层输出结果，即编码完成的图片样本特征向量与 prompt 特征向量计算 InfoNCE Loss，通过减小 InfoNCE Loss，在新的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，从而实现好坏图片的区分。相比仅使用 Clip 进行分类的微调任务，我们使用的对比学习微调方法更具有解释性。

推理阶段——我们将微调好的 CLIP 的 Image Encoder 和 Text Encoder，以及附加的可学习线性变换层用于推理。将一个 prompt 和一对图片作为输入，将线性变换层输出的文本特征向量表示分别与这两张图片进行余弦相似度计算，选择相似度更大的图片作为好图片，相似度更小的图片作为坏图片。

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$N \times M$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$d_x \times d_y$的滑动窗口采集到$k = (M - d_x + 1) \times (N - d_y + 1)$个$d_x \times d_y$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[5,2,7,4]，其序数排名序列为[2,0,3,1]。
- 统计上述得到的k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P。其中，对于长度为$d_x \times d_y$的序列，其所有可能的序数排名序列有$n = (d_x \times d_y)!$个。
- 计算P的归一化香农熵$H(P)$，即为图片的排列熵。
- 计算P的统计复杂度$C(P)$，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 3.2 所示。

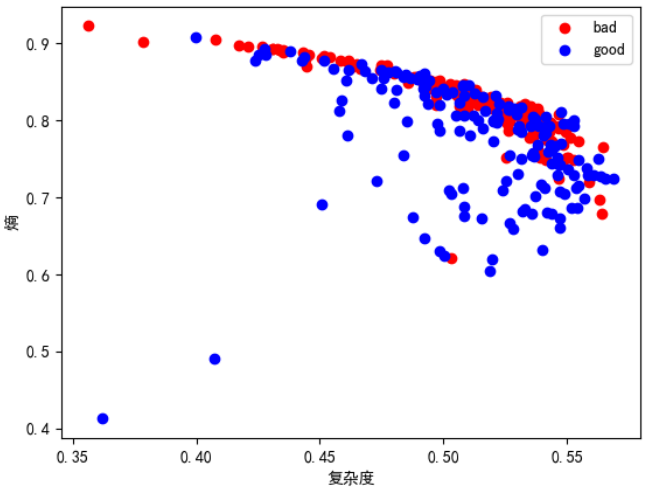


图 3.2 复杂度熵评价图

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，如表 1 所示

表 3.1 prompt 描述矩阵

	Color	Number	Existence	Object 1	…	Object n
Object 1						
…						
Object n						

描述矩阵的大小为$n \times (m+n)$，可看作$n \times m$的矩阵与$n \times n$的矩阵相拼接得到。其中n为 prompt 中的对象数量，m为颜色等属性的数量，$n \times n$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研究院所提供的文生图，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和验证（随机种子设置为 0），每次训练的 batch_size 设置为 16；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

我们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并输出 result.csv，并以 8：2 分验证集（共 1200 对）中进行测试，得到结果准确度为 0.9525，如图 4.1。通过智源研究院项目[5]所提到检验文生图质量特征对我们所输出结果进行进一步验证：

```
--------------------------------------
100%|██████████| 75/75 [27:01<00:00, 21.63s/it]
1137 1143 856
--------------------------------------
1137 / 1200  =  0.9475
1143 / 1200  =  0.9525
```

图 4.1 结果图

4.2.1 实体

我们对模型是否正确分辨出实体的对象、状态、颜色、数量与位置进行举例验证：

1) 对象

选择测试集中婴儿萝卜图像进行对比，如图 4.2 在结果输出中评价第一张图片为 bad，第二张图片为 good，可见模型可以正确分辨出实体。

(a) a-baby-daikon-radish-6.png image1

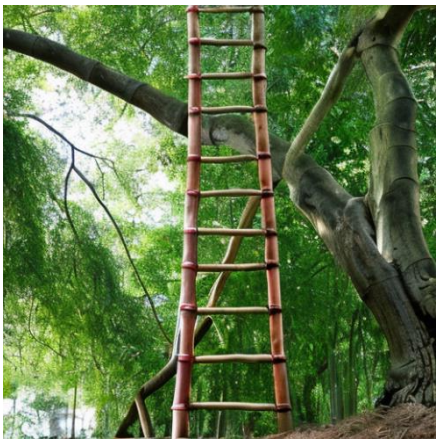
(b) a-baby-daikon-radish-6.png image2

图 4.2 实体对象对比

2) 位置

选择测试集中 prompt 为一把搭在橡树上的竹梯（a bamboo ladder propped up against an oak tree）的两张图像进行对比，如图 4.3 在结果输出中评价第一张图片为 bad,第二张图片为 good，在两张图片均有梯子与橡树情况下，模型可以正确判断物体的位置（against）。

(a) a-bamboo-ladder-propped-up-aga-4 image1



(b) a-bamboo-ladder-propped-up-aga-4 image2

图 4.3 实体位置分析

3) 颜色

选择测试集中 prompt 为树上有方形的蓝色苹果和圆形的黄色叶子（square blue apples on a tree with circular yellow leaves）的两张图像进行对比，如图 4.4 在结果输出中评价第一张图片为 bad,第二张图片为 good，在两张图片均为苹果与黄色叶子条件下，模型可以正确判断物体的颜色。

(a) square-blue-apples-on-a-tree-w-2 image1

(b) square-blue-apples-on-a-tree-w-2 image2

图 4.4 实体颜色分析

4) 数量

选择测试集中 prompt 为四个红酒瓶（four wine bottles）的两张图像进行对比，如图 4.5 在结果输出中评价第一张图片为 bad,第二张图片为 good，在两张图片红酒瓶条件下，模型可以正确判断瓶子数量。

(a) four-wine-bottles-5 image1

(b) four-wine-bottles-5 image2

图 4.5 实体数量分析

5) 状态

选择测试集中 prompt 为一只读漫画书的猫（a cat reading a comic book）的两张图像进行对比，如图 4.6 在结果输出中评价第一张图片为 good,第二张图片为 bad，在两张图片猫的生成效果均不错情况下，模型可以正确判断实体状态。

(a) a cat reading a comic book image1

(b) a cat reading a comic book image2

图 4.6 实体状态分析

4.2.2 风格

我们对模型是否能判断出 prompt 中相应的分格进行分析：

1) 绘画风格

选择测试集中 prompt 为火车素描（a sketch of a train）进行对比，如图 4.7 在结果输出中评价第一张图片为 bad,第二张图片为 good，虽然第一张图片在火车生成效果比第二张图片要更好，但缺少必要的素描绘画风格，导致最终得分低于第二张，可见模型可以正确分辨出绘画风格。

(a) a-sketch-of-a-train image1　　(b) a-sketch-of-a-train image2

图 4.7 绘画风格对比

2) 文化风格

选择测试集中 prompt 为日出时伊斯坦布尔 其中细节是水墨画（Downtown Istanbul at sunrise. detailed ink wash.）进行对比，如图 4.8 在结果输出中评价第一张图片为 good,第二张图片为 bad，两张图片都有必要的水墨画元素，但第二张图显然缺少必要的伊斯坦布尔文化风格，导致最终得分低于第一张，可见模型可以正确分辨出文化风格。

(a) Downtown-Istanbul-at-sunrise image1

(b) Downtown-Istanbul-at-sunrise image2

图 4.8 文化风格对比

4.2.3 细节

我们对图片的细节是否被模型注意到进行了验证，分别分辨实体细节中手部、五官、性别、反常识进行举例验证：

1) 手部

选择测试集中 prompt 为手持篮球（the hands of a single person holding a basketball）图像进行对比，如图 4.9 在结果输出中评价第一张图片为 good,第二张图片为 bad，相比较第一张图片手部构造更为出色，可见模型可以注重区分手部细节。

(a) the-hands-of-a-single-person image1

(b) the-hands-of-a-single-person image2

图 4.9 手部细节对比

2) 五官

选择测试集中 prompt 为一个生气的男人（an angry man）图像进行对比，如图 4.10，

文生图模型均高质量构造出两个男人，但第二张图片男人的表情细节更偏向于生气，在结果输出中评价第一张图片为 bad ,第二张图片为 good，可见模型可以注重区分五官细节。

(a) an angry man image1

(b) an angry man image2

图 4.10 五官细节分析

3) 性别

选择测试集中 prompt 为去上学的男孩（a boy go to school）图像进行对比，如图 4.11，文生图模型构造出两个孩子，但第二张图片性别更偏向于男性，而第一张性别不明，在结果输出中评价第一张图片为 bad ,第二张图片为 good，可见模型可以区分性别细节。

(a) a boy going to school image1

(b) a boy going to school image2

图 4.11 性别细节分析

4) 反常识

选择测试集中 prompt 为四只眼睛的猫（a cat with four eyes）图像进行对比，如图 4.12，在提示词反常识情况下，通过对比两张图片，第二张更偏向于四只眼睛（四只眼瞳），故在结果输出中评价第一张图片为 bad ,第二张图片为 good，可见模型可以区分反常识图片。

(a) a cat with four eyes image1

(b) a cat with four eyes image2

图 4.12 反常识分析

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（文本所含对象的存在性，文本描述的对象颜色，文本提及的对象数量）进行判定，最终加权得到得分进行比较。我们的整个评价器准确度高，而且构建评价器的流程均具有较强的可解释性。

5.2 缺点：

由于设备限制与模型复杂度较高，我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应用与部署。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。

参考文献

[1] van den Oord, A., Li, Y. and Vinyals, O., 2019. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

[2] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. and Krueger, G., 2021, July. Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.

[3] Sigaki, H.Y., Perc, M. and Ribeiro, H.V., 2018. History of art paintings through the lens of entropy and complexity. Proceedings of the National Academy of Sciences, 115(37), pp.E8585-E8594.

[4] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[5] EVA_CLIP: https://github.com/FlagOpen/FlagEval