

复现基于视觉语言关联的盲图像质量评价：多任务学习视角

摘要

本文复现的论文目标是推进盲图像质量评估 (BIQA)，它在没有任何参考信息的情况下预测人类对图像质量的感知。该论文为 BIQA 开发了一种通用的自动化多任务学习方案，以自动确定模型参数共享和损失权重的方式从其他任务中利用辅助知识。具体来说，该论文首先使用文本模板描述所有候选标签组合（来自多个任务），并从视觉文本嵌入的余弦相似度计算联合概率。每个任务的预测可以从联合分布中推断出来，并通过精心设计的损失函数进行优化。通过学习 BIQA、场景分类和扭曲类型识别三个任务的综合实验，验证了所提出的 BIQA 方法受益于场景分类和扭曲类型识别任务，并在多个 IQA 数据集上优于目前的最先进方法，在群体最大差异化竞争中具有更强的鲁棒性，并且更有效地重新对齐来自不同 IQA 数据集的质量注释。

关键词：盲图像质量评价；多任务学习；失真分类

1 引言

作为一项基本的计算视觉任务，盲图像质量评估旨在预测数字图像的视觉质量，而无法访问底层原始质量对应（如果有的话）。在深度学习时代，BIQA（盲图像质量评估）的发展主要表现为解决大量可训练参数与少量人类质量评分（以平均意见得分 MOS 形式）之间的冲突的策略。当合成失真（例如，高斯噪声和 JPEG 压缩）是主要关注的问题时，补丁训练、质量感知预训练和从噪声伪标签中学习是较少（或不）依赖 MOSs 的实用训练技巧。这里的基本假设是：1) 原始质量的图像存在并且是可访问的；2) 视觉失真可以有效和自动地模拟；3) 全参考 IQA 模型是适用的，并提供足够的质量近似。然而，当涉及到真实的相机失真（例如，传感器噪声，运动模糊或两者的组合）时，所有这些假设都不成立。已经探索了一组不同的训练技巧，包括迁移学习、元学习和对比学习。结合多个数据集进行联合训练和识别信息样本进行主动微调的新兴技术也可以被视为应对 BIQA 中数据挑战的方法。

受视觉语言预训练 [1] 的启发，该论文提出了一种通用的、自动化的 BIQA 多任务学习方案，试图回答上述突出的问题。这里的“自动化”是指所有任务的模型参数共享和分配给每个任务的损失权重都是自动确定的。该论文考虑了两个额外的任务，场景分类和失真类型识别，前者在概念上与 BIQA 相冲突，而后者则密切相关。首先使用文本模板总结输入图像的场景类别、失真类型和质量水平。然后，使用对比语言-图像预训练模型 CLIP [1]，这是一种由大量图像-文本对训练的视觉和语言联合模型，以获得视觉和文本嵌入。三个任务的联合

概率可以通过图像嵌入和所有候选文本嵌入之间的余弦相似度来计算。最后联合分布边缘化以获得每个任务的边际概率，并进一步使用边际分布作为权重将离散化的质量水平转换为连续的质量分数。

2 相关工作

2.1 盲图像质量评价

传统的盲图像质量评价要么依赖于自然场景统计形式的手工设计特征，要么依赖于代码本形式的浅层特征学习。深度学习利用了特征提取和质量回归的端到端优化，极大地推进了 BIQA 领域的发展。近年来，BIQA 新范式蓬勃发展，旨在探索下一代 BIQA 的发展方向。代表性工作包括：用于局部质量预测的补丁到图片学习，用于有价值样本识别的主动学习，用于交叉失真场景的统一优化，用于快速适应的元学习 [2] 等等。该论文利用了多任务学习来促进辅助知识转移。

2.2 CLIP 应用

CLIP 在协助广泛的视觉任务方面显示出很大的希望。最初，Radford 等人 [1] 利用了 4 亿图像-文本对来预训练一系列 CLIP 模型，这些模型在广泛的下游视觉任务中表现出显著的零射击转移能力。Zhou 等人 [3] 建议以牺牲语言可解释性为代价，及时调整以提高迁移效率。在进入概念后不久，CLIP 就找到了（开放词汇）语义分割和目标检测的方法。Vinker 等人发现了一种 CLIP 用于对象素描的新用法，对对象语义有很好的理解。与这项研究最接近的是 Wang 等人 [4]，他们假设 CLIP 模型具有固有的质量意识，并通过即时工程采用它们来评估图像质量和美学。但该论文的方法在概念上和计算上都与他们的有很大的不同。该论文是利用 CLIP 在多任务学习环境中通过辅助知识转移来帮助 BIQA。此外，还对 CLIP 模型进行微调，而不是在不牺牲语言可解释性的情况下进行即时调优，以获得更好的质量预测性能。

2.3 多任务学习

多任务学习作为多目标优化的具体案例 [5] 通常通过跨任务共享计算和信息来提高任务准确性、记忆成本和推理时间。现有方法的不同之处主要在于两种设计选择：模型参数共享和损失加权。不需要手动指定共享哪些参数，也不需要学习确定具有组合复杂度的任务特定参数，而是假设 LIQE 图像编码器中的所有参数都是共享的，其容量在端到端优化过程中动态分配给每个任务。对于损失加权，Sener et al. [6] 和 Lin et al. [7] 将多任务学习视为多目标优化，并根据 Karush-Kuhn-Tucher 条件求解最优损失加权。Liu 等人 [8] 实现了一种基于模型不可知元学习的自动加权方案，该方案允许对主要（即最重要）任务进行规范。其他有效的损失加权启发式方法包括学习任务不确定性、梯度归一化和损失下降率。在本文中，由于其概念简单、计算方便和效率高，该论文中采用了该方法。

3 本文方法

3.1 本文方法概述

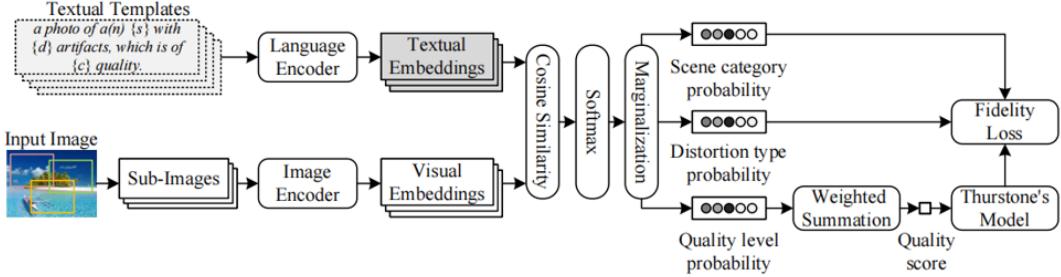


图 1. LIQE 系统框架图

LIQE 系统框架图如图 1 所示。给定图像 $\mathbf{x} \in \mathbb{R}^N$ 可能经历几个阶段的退化，而 BIQA 模型的目标 \hat{q} 是预测 \mathbf{x} 的感知质量，接近其 MOS 值 $q(\mathbf{x}) \in \mathbb{R}$ 。还使用了五个质量等级的李克特量表： $c \in \mathcal{C} = \{1, 2, 3, 4, 5\} = \{“坏”, “差”, “一般”, “好”, “完美”\}$ ，并通过如下方程将 c 与 \hat{q} 联系起来：

$$\hat{q}(\mathbf{x}) = \sum_{c=1}^C \hat{p}(c | \mathbf{x}) \times c \quad (1)$$

其中， $C = 5$ 是质量等级的数量， $\hat{p}(c | \cdot)$ 是要估计的 c 的边际概率。除了 BIQA，还包括一个概念冲突的任务-场景分类和一个密切相关的任务-扭曲类型识别。这里考虑 9 个场景类别： $s \in S = \{“动物”, “城市景观”, “人类”, “室内场景”, “景观”, “夜景”, “植物”, “静物” 和 “其他”\}$ 。一张图像可能包含多个场景标签。这里不区分合成失真和真实失真，而是识别图像中的主要失真： $d \in D = \{“模糊”, “颜色相关”, “对比度”, “JPEG 压缩”, “JPEG2000 压缩”, “噪声”, “过度曝光”, “量化”, “曝光不足”, “空间局部化” 和 “其他”\}$ ，总共有 11 个。“其他”类别包括没有失真的图像（即原始质量）。然后很自然地创建一个文本模板来将来自三个任务的标签放在一起：“一张带有 d 失真的 s 类别的照片，图像质量为 c 。”总共有 $5 \times 9 \times 11 = 495$ 个候选文本描述。

3.2 损失函数定义

给定 $\hat{p}(c, s, d | \mathbf{x})$ ，将其边缘化得到 $\hat{p}(c | \mathbf{x})$ ，并通过 Eq.(1) 计算质量估计 $\hat{q}(\mathbf{x}) \in \mathbb{R}$ 。基于排序的方式，考虑 BIQA 的两两学习排序模型估计。具体来说，对于来自同一 IQA 数据集的图像对 (x, y) ，根据它们的基准 MOSs 值来计算二元标签：

$$p(\mathbf{x}, \mathbf{y}) = \begin{cases} 1 & \text{if } q(\mathbf{x}) \geq q(\mathbf{y}) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

在 Thurstone 的模型下，我们估计 x 被感知比 y 更好的概率为

$$\hat{p}(\mathbf{x}, \mathbf{y}) = \Phi\left(\frac{\hat{q}(\mathbf{x}) - \hat{q}(\mathbf{y})}{\sqrt{2}}\right) \quad (3)$$

其中 $\Phi(\cdot)$ 为标准正态累积分布函数，方差固定为 1。同时采用保真度损失作为统计距离度量：

$$\begin{aligned} \ell_q(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) &= 1 - \sqrt{p(\mathbf{x}, \mathbf{y})\hat{p}(\mathbf{x}, \mathbf{y})} \\ &\quad - \sqrt{(1 - p(\mathbf{x}, \mathbf{y}))(1 - \hat{p}(\mathbf{x}, \mathbf{y}))} \end{aligned} \quad (4)$$

场景分类损失函数：

$$\begin{aligned} \ell_s(\mathbf{x}; \boldsymbol{\theta}) &= \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} (1 - \sqrt{p(s | \mathbf{x})\hat{p}(s | \mathbf{x})}) \\ &\quad - \sqrt{(1 - p(s | \mathbf{x}))(1 - \hat{p}(s | \mathbf{x}))} \end{aligned} \quad (5)$$

失真类别损失函数：

$$\ell_d(\mathbf{x}; \boldsymbol{\theta}) = \left(1 - \sum_{d \in \mathcal{D}} \sqrt{p(d | \mathbf{x})\hat{p}(d | \mathbf{x})}\right) \quad (6)$$

在此基础上，最终损失定义为三个单独损失的线性加权总和：

$$\begin{aligned} \ell(\mathcal{B}, t; \boldsymbol{\theta}) &= \frac{1}{|\mathcal{P}|} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{P}} \lambda_q(t) \ell_q(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) + \\ &\quad \frac{1}{|\mathcal{B}|} \sum_{\mathbf{x} \in \mathcal{B}} (\lambda_s(t) \ell_s(\mathbf{x}) + \lambda_d(t) \ell_d(\mathbf{x})) \end{aligned} \quad (7)$$

加权向量 $\boldsymbol{\lambda}(t) = [\lambda_q(t), \lambda_s(t), \lambda_d(t)]^\top$ 的第 t 次迭代可以根据相对下降率 [9] 自动计算：

$$\lambda_j(t) = \frac{\exp(w_j(t-1)/\tau_2)}{\sum_i \exp(w_i(t-1)/\tau_2)}, w_j(t-1) = \frac{\ell_j(t-1)}{\ell_j(t-2)} \quad (8)$$

4 复现细节

4.1 实验设置

本实验在 6 个 IQA 数据集上进行了实验，其中 LIVE、CSIQ 和 KADID-10k 包含合成失真，LIVE Challenge、BID 和 KonIQ-10K 包含真实失真。从每个数据集中随机抽取 70% 和 10% 的图像，分别构建训练集和验证集，剩下的 20% 用于测试。对于具有合成扭曲的三个数据集，再根据参考图像将训练集、验证集和测试集分开，以确保内容独立性。重复这个过程十次，并计算斯皮尔曼等级相关系数 SRCC 和皮尔逊线性相关系数 PLCC 结果来作为评价指标。

4.2 实验环境搭建

本文的实验环境如下：

Python == 3.12.5

```
PyTorch == 2.5.0  
torchvision == 0.20.0  
此外，还需要安装：  
pip install ftfy regex tqdm  
pip install git+https://github.com/openai/CLIP.git
```

4.3 与已有开源代码对比

对于已有的开源代码，运行时会出现许多小问题，如编译问题、版本问题、路径问题。在本次复现实验过程中，对这些问题逐一进行了修改，最终才得以运行。同时，由于 GPU 容量不足，减少了实验的批量大小，并且尝试减少数据集，在单个数据集上进行训练。

5 实验结果分析

本实验从测试集中选取失真类型以及失真程度不相同的几组图像进行实验。同时还随机选取网络上以及生活中的一些图像，手动施加不同类型及不同程度的失真，来对模型进行实验测试。部分测试结果如图 2 所示。

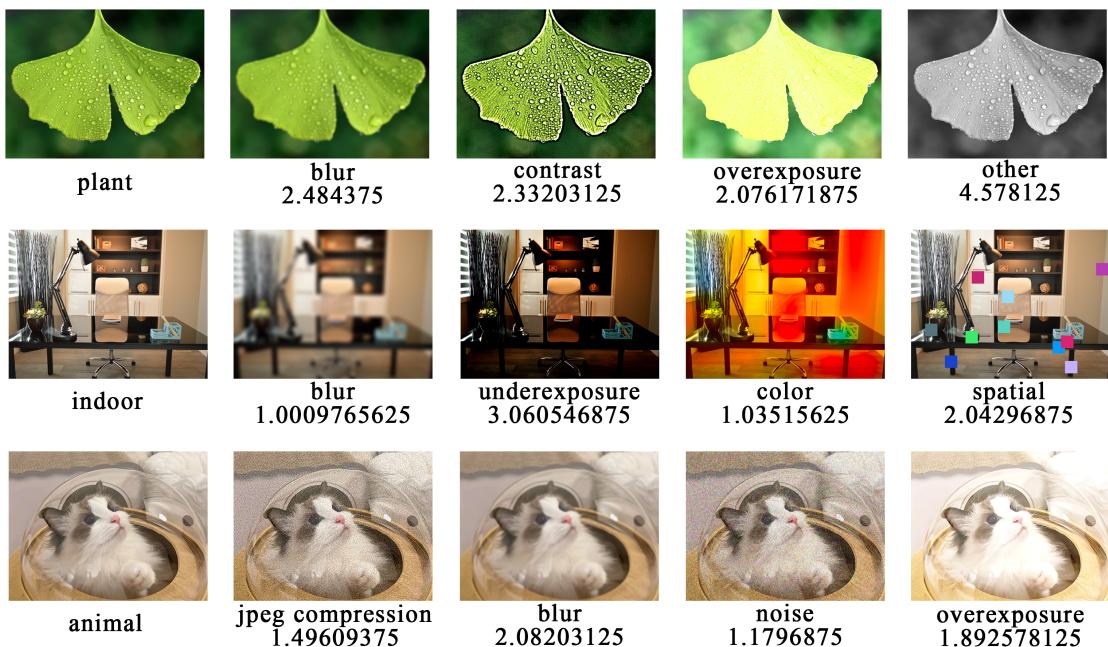


图 2. 实验结果

可以得出实验结果对场景分类、失真类型，以及对图像质量的评价，基本上符合人的主观判断。但是由于场景分类的类别较少，且有些场景可能属于不同的分类，所以有些情况下会出现对场景错误分类的情况。同时，有些图像可能是叠加了多种失真，但是该模型只考虑了一种失真情况，这对失真类型分类具有一定的局限性。在本次实验中还发现，对于尺寸较大的图片来说，该模型得出的图像质量评价分数总体上比人的主观质量评价偏低，这可能是

由于人的主观会偏向于整体的感受，而该模型不能很好地平衡大尺寸图片的整体情况和局部质量。

6 总结与展望

该论文从多任务学习的角度，通过视觉-语言的对应关系，制定了一个盲图像质量评价的方法。训练过程中，在多个 IQA 数据集上同时优化了一对图像和语言编码器，用于盲图像质量评价、场景分类和失真类型识别。同时还设计了三种保真度损失来训练模型，并采用简单高效的动态加权方案自动对多任务损失进行加权求和我们展示了所提出的 LIQE 的有效性，并验证了各种设计选择的合理性。结果还表明，学习到的模型以一种更有感知意义的方式重新排列来自不同数据集的 MOS 值。相信该论文提出的多任务学习视角将为下一代 BIQA 模型以及其他机器视觉应用的计算模型的发展提供一些启示。

参考文献

- [1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, pages 8748–8763, 2021.
- [2] Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. MetaIQA: Deep meta-learning for no- reference image quality assessment. In IEEE Conference on Computer Vision and Pattern Recognition, pages 14131–14140, 2020.
- [3] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. In International Journal of Computer Vision, 130(9):2337–2348, Aug. 2022.
- [4] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. CoRR, abs/2207.12396, 2022.
- [5] R. Timothy Marler and Jasbir S. Arora. Survey of multi- objective optimization methods for engineering. Structural and Multidisciplinary Optimization, 26(6):369–395, Mar. 2004.
- [6] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In Advances in Neural Information Processing Systems, pages 525–536, 2018.
- [7] Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qing-Fu Zhang, and Sam Kwong. Pareto multi-task learning. In Advances in Neural Information Processing Systems, volume 32, pages 12037–12047, 2019.
- [8] Shikun Liu, Stephen James, Andrew J. Davison, and Edward Johns. Auto-Lambda: Disentangling dynamic task relationships. Transactions on Machine Learning Research, 2022.

- [9] Shikun Liu, Edward Johns, and Andrew J. Davison. End-to-end multi-task learning with attention. In IEEE Conference on Computer Vision and Pattern Recognition, pages 1871–1880, 2019.
- [10] Zhang, Weixia, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. "Blind image quality assessment via vision-language correspondence: A multitask learning perspective." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 14071-14081. 2023.