建立量化气候变化图像真实性的评价指标

阅读报告

一 文献信息

1）论文题目：Establishing an Evaluation Metric to Quantify Climate Change Image Realism

2）作者：Sharon Zhou, Alexandra Luccioni, Gautier Cosne,
Michael.Bernstein, Yoshua Bengio

3）发表时间：2019

4）发表途径：33rd Conference on Neural Information Processing Systems (NeurIPS 2019),
Vancouver, Canada

二 研究问题及意义

随着控制任务的成功，生成模型越来越多地应用于人道主义应用。公众对气候变化的认识和关注与其对我们物种和环境的威胁程度不相称，人们很难在心理上模拟气候变化的复杂和概率效应。而生成图像模型缺乏强有力的模式比较评价方法，现有的评价产出质量和多样性的方法有着很强的局限性，特别是对条件模型的局限性。本文从人类评估角度出发，提出了量化生成模型学习模式真实性的方法。作者着重于对一个条件生成模型的评估，该模型说明了气候变化引起的洪水的后果，以鼓励公众对这一问题的兴趣和认识。

三 研究思路

文中首先研究了现有的两种主要的生成模型评估方法：核心初始距离（KID）、初始分数（IS）和 Fréchet 初始距离（FID）等自动化度量，它们都旨在评估生成的样本在分布水平上的视觉质量和样本多样性。还有一种以人为中心的度量，如 HYPE（人眼感知评估），它使用人类来评估图像的真实性。各有优缺点。进而提出了扩展现有自动化度量的方法，但它不能评估不同模式之间的视觉效果，都不令人满意。

文中试图寻找评价模型生成的不同风格的真实感最有效的方法，并能与人类感知现实相关的自动化方法。作者分析了 Pearson 提出的自动风格排序方法和 hype 风格之间的相关系数，以确定与人类感知现实最相关的方法。

四 实验方法

文中用样式级别构建任务：对于每个给定的样式向量，跨多个基于相同模式的样本进行聚合。这种样式级聚合避免了对单个样本进行评估，使用众包人工评估来调整样式级别评估的 HYPE 度量。并将 HYPE 风格与各种自动化指标进行比较。这些指标使 FID 和 KID 适应样式比较任务。对于每个度量，使用不同的初始层进行实验。

分析了 Pearson 提出的自动风格排序方法和 HYPE 风格之间的相关系数时引入测度 r，测度 r 有支持度[-1,1]，其中值 1 和-1 分别表示强正相关和负相关，而值为 0 则表示弱相关。为了确定分数的可分性，我们还使用 25 个重复计算了 r 上 95%的引导置信区间。对于每个复制 i，我们使用带有替换的采样图像计算 HYPE 样式和自动评分，并从中计算 r_i。

1. 人类评估：HYPE-Style

原本 HYPE-Style 是指评估者对半真实和半生成的图像进行校准和过滤，并且被给予无限的时间来标记图像的真实与否。对于每一幅图像，我们计算平均错误率，它对应于判断图像真实性的人类评价者的比例。值越高表示图像越逼真。为了在条件生成下进行样式内比较，有必要对 Hype 做出修改，以两种方式限制该过程：（1）要求对每个样式和图像组合进行多次求值，以便在给定图像中对样式进行比较；（2）确保求值器不会看到从给定输入图像生成的多个样式。

计算分数时，规定按照风格将图像聚合到组中，并计算每个组中所有人类评价者标签的微平均值。具体来说，对于每个样式和图像 x，我们有多个人类标签 I_x^s，标记为“real”（1）和“generated”（0），基于人类对其真实性的判断，我们为每个样式 s 计算，对该特定样式的图像求和 $\text{HYPE-Style}=\sum_i I_i^s$。因此，在生成的图像上得分越高，表示错误率越高，对人类平均来说更真实。人为的评估，虽然更精确和可靠，但执行每种风格是昂贵和耗时的。但可由此作为判断寻找与人类最相关的自动化方法。

2. 自动化样式分析方法

采用 FID 和 KID 来计算单个样式中实际分布和生成分布之间的距离，并将它们用作样式得分。总共评估了 8 种自动化方法{FID, KID}×{pool 1, pool 2, pre-aux, pool 3}。

	pool 1	pool 2	pre-aux	pool 3
FID	0.103 (0.53, 0.153)	0.146 (0.099, 0.193)	**0.433 (0.390, 0.476)**	0.407 (0.366, 0.448)
KID	0.010 (-0.041, 0.061)	0.034 (-0.015, 0.083)	0.432 (0.389, 0.475)	0.367 (0.322, 0.412)

表 1

如表 1 所示，使用 pre-aux 嵌入的 FID 和 KID 在与人类评分相关的方面都超过了其他指标，具有中等的相关。在比较不同层之间的性能时，KID 和 FID 相互跟踪，其中 pre-aux 嵌入首先出现，然后是 pool 3、pool 2 和 pool 1。

作者发现，在辅助分类头（pre-aux）之前的 768 维初始 v3 层表现优于 pool 3 层和其他早期的池层{pool 1，pool 2}。直观地说，前辅助层的最大特点是丰富的层，但仍然是辅助类中的常规层。这种规则化将鼓励图层编码有更一般的特征，这些特征对数据集的影响较小。 在不同域的任务中比数据集更能发挥作用。并且在 FID 和 KID 中，pre-aux 层对 pool 3 的选择是一致的，pre-aux 层的得分分别为 0.433 和 0.432，而在 pool 3 中，FID 和 KID 的得分分别为 0.407 和 0.367。

五 实验结论

本文提出了一个在生成模型上评估不同风格的人类评价指标。作者还评估了 8 种不同的自动化方法，发现在辅助分类器之前使用辅助分类优于使用最后一个池化层的广泛使用的方法，前者更能与人类对这项任务的感知相关联。虽然所评估的自动化方法中没有一种能够完全接近独立使用的人类评估风格，这个工作仍然是评估多模态跨域映射的风格级别属性的初步尝试。在这一领域，使用主流的自动化评估指标仍然是一个难题，作者计划探索更好的方法。

六 启发思考

不管是什么专业技术，它们最大的价值是受用于民。本文的作者利用深度学习的概念应用于模型真实性的评估，用人类评估为基准，寻找最接近人类感知的自动化方法。最终目的是能够生成由气候变化引起的极端天气现象的最真实图像，进而描述气候变化支持的其他灾难性事件，让人们认识到气候改变后的自然景象，提高警醒。工程人才应该具备专业素养的同时兼具人文关怀。

所评估的自动化方法中没有一种能够完全接近独立使用的 HYPE 风格，但通过比较得出最接近的自动化方式，作者依然在继续改进。在探索一个新问题的时候，往往无法一下子得出最优模型，此时，需要在学习他人成果的基础上，对比更新自己的研究，继续不断优化模型。