VMAF 视频质量评估

Dezeming Family

2023 年 7 月 4 日

DezemingFamily 系列文章和电子书**全部都有免费公开的电子版**，可以很方便地进行修改和重新发布。如果您获得了 DezemingFamily 的系列电子书，可以从我们的网站 [https://dezeming.top/] 找到最新的版本。对文章的内容建议和出现的错误也欢迎在网站留言。

目录

一 介绍

VMAF 的全称是：Visual Multimethod Assessment Fusion，视频质量多方法评价融合。这项技术是由美国 Netflix 公司开发的一套主观视频质量评价体系。

首先有三个问题：

- Netflix 会员将如何评价这段视频的质量——差、一般还是优秀？
- 哪一个视频剪辑看起来更好——用编解码器 A 编码还是用编解码器 B 编码？
- 对于这一集，在 1000 kbps 的速率下，是用高清分辨率编码更好，还是用 SD 编码更好？

这些都是我们在努力为 Netflix 会员提供最佳体验时问自己的问题。几年前，我们意识到，仅仅依靠“金眼 (golden eyes)”是无法有效回答这些问题的。专家观看无法在内容、编码配方和编码管道的整体输出之间进行扩展。虽然可以大规模部署现有的视频质量指标，如 PSNR 和 SSIM，但它们无法准确捕捉人类感知。因此，我们开始了开发一种自动化方法的旅程，以回答“Netflix 会员将如何评价这种编码的质量?”这就是 VMAF 的诞生。

视频多方法评估融合，简称 VMAF，是一种将人类视觉建模与机器学习相结合的视频质量度量。该项目最初是我们的团队与南加州大学的郭教授进行的研究合作。他的研究小组之前曾致力于图像的感知指标，我们一起致力于将这些想法扩展到视频。随着时间的推移，我们与其他研究合作伙伴合作旨在提高与人类主观感知相关的 VMAF 准确性，并扩大其范围以涵盖更多用例。2016 年 6 月，我们在 Github 上开源了 VMAF[4]，还发布了第一个 VMAF 技术博客。

在 Netflix 之外，视频社区发现 VMAF 是一种有价值的质量评估工具。由于行业采用，该项目受益于研究人员、视频相关公司和开源社区的更广泛贡献。

- VMAF 已集成到第三方视频分析工具中（例如，FFmpeg、Elecard StreamEye、MSU 视频质量测量工具和 arewecompressedyet），将其与 PSNR 和 SSIM 等更成熟的指标并列。
- 在诸如 NAB 的行业贸易展和会议中，Video@Scale 使用 VMAF 分数来比较各种编码技术的质量和效率。
- 视频质量专家组（VQEG）是一个由视频质量评估专家组成的国际联盟。在最近的洛斯加托斯、克拉科夫和马德里 VQEG 会议上，VMAF 在多次讨论中得到了评估。

其他研究小组已经交叉验证了 VMAF 的感知准确性。Rassool（RealNetworks）报道了 $4K$ 内容的 VMAF 和 DMOS 分数之间的高度相关性。Barman 等人（金斯顿大学）对游戏内容的几个质量评估指标进行了测试，得出的结论是 VMAF 在预测主观得分方面最好。Lee 等人（延世大学）将质量度量应用于多分辨率自适应流，并表明 VMAF 和 EPSNR 与感知质量的相关性最高。在 Gutiérrez 等人（南特大学）的研究中，VMAF 和 VQM 是表现最好的质量指标，其中为 HD 和 UHD 内容生成 MOS 分数。

我们还阅读了一些研究，其中声称 VMAF 的表现不如预期。我们邀请行业和研究人员评估最新的 VMAF 模型，并鼓励他们与我们分享可能改进下一个 VMAF 版本的反例和角落案例。我们还在后面的章节中给出了使用 VMAF 的最佳实践，以解决一些问题。

VMAF 可以用作更好的编码决策的优化标准，我们已经看到其他公司为此应用 VMAF 的报告。

二 在 Netflix 中，VMAF 的作用

2 1 Codec Comparisons

传统上，编解码器比较共享相同的方法：为多个视频序列计算 PSNR 值，每个视频序列根据一组测试条件以预定义的分辨率和固定的量化设置进行编码。随后，构建速率-质量曲线，并计算这些曲线之间的平均差（BD 速率）。这样的设置适用于编解码器中的微小差异，或用于评估同一编解码器内的工具。对于我们的用例——视频流——使用 PSNR 是不合适的，因为它与感知质量的相关性很差。

VMAF 填补了这一空白，可以捕捉编解码器之间更大的差异，以及缩放伪像，以一种与感知质量更好相关的方式。它使我们能够比较真正相关的区域中的编解码器，即在可获得速率质量点的凸包上。比较不同编解码器和/或不同配置之间的凸包给出了在实际重要的速率质量区域中两个选项的 Pareto 前沿的比较。我们团队最近在编解码器比较方面的一些工作发表在一个关于基于镜头的编码的技术博客上，以及 2018 年图片编码研讨会和 SPIE 数字图像处理应用 XLI 上的学术论文中。

2 2 Encoding Decisions

VMAF 被用于我们的整个生产线，不仅用于衡量我们编码过程的结果，还用于指导我们的编码达到尽可能好的质量。VMAF 如何在编码中使用的一个重要例子是我们的动态优化器，其中每个单独镜头的编码决策由每个编码器选项的比特率和质量测量来指导。VMAF 分数在这个优化过程中至关重要，以获得准确的质量测量，并选择凸包上的最终分辨率/比特率点。

2 3 A/B Experimentation

不同业务领域的研究人员——例如电视用户界面团队和流媒体客户端团队——正在不断创新，以提高流媒体质量。有了 VMAF，我们有了一个工具，可以运行全系统的 A/B 测试，并量化对会员视频质量的影响。例如，研究人员更改自适应流算法或部署新的编码，运行实验，并比较新旧算法或编码之间的 VMAF。

这一指标非常适合评估实验中的质量，因为它在内容上的一致性和反映人类对质量感知的准确性。例如，85 的 VMAF 分数将意味着所有标题的“良好”质量，而不是使用比特率（相同的比特率可以指示标题之间的不同质量）。

三 在对 VMAF 工作中的改进

3 1 Speed Optimization

当我们于 2016 年 6 月在 Github 上首次发布 VMAF 时，它的核心特征提取库是用 C 编写的，控制代码是用 Python 编写的，主要目标是支持算法实验和快速原型设计。根据用户的要求，我们很快添加了一个独立的 C++ 可执行文件，它可以更容易地部署在生产环境中。2016 年 12 月，我们在 VMAF 的卷积函数中添加了 AVX 优化，这是 VMAF 中计算量最大的运算。这使得 VMAF 的执行时间加快了约 3 倍。最近的一次是在 2018 年 6 月，我们添加了帧级多线程，这是一个长期的功能（对龙舌兰酒的特别呼吁）。我们还引入了跳帧的功能，允许在 N 帧中的每一帧上计算 VMAF。这是第一次可以实时计算 VMAF，即使是在 4K 中，尽管精度略有损失。

3 2 libvmaf

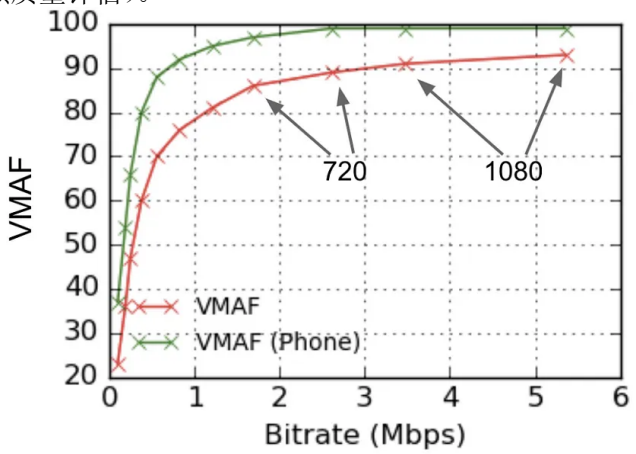
在 FFmpeg 社区的帮助下，我们将 VMAF 打包到一个名为 libvmaf 的 C 库中。该库提供了一个接口，可将 VMAF 测量值纳入您的 C/C++ 代码中。VMAF 现在作为滤波器包含在 FFmpeg 中。FFmpeg-libvmaf 过滤器现在是在压缩视频比特流上作为输入运行 VMAF 的一条方便的路径。

3 3 Accuracy Improvement

自从我们开源 VMAF 以来，我们一直在不断提高其预测准确性。随着时间的推移，我们已经修复了在基本度量和机器学习模型中发现的许多不理想的情况，从而产生了更准确的总体预测。例如，修改基本度量以产生与亮度掩蔽的改进的一致性；更新场景边界处的运动分数以避免由于场景变化而引起的过冲；当外推到高 QP 区域中时，QP-VMAF 单调性现在得以保持。显然，VMAF 模型的准确性在很大程度上也取决于它所训练的主观分数的覆盖率和准确性。与之前的数据集相比，我们收集了一个范围更广的主观数据集，包括更多样的内容和源伪像，如胶片颗粒和相机噪声，以及更全面地覆盖编码分辨率和压缩参数。我们还开发了一种新的数据清理技术，以消除原始数据中的人为偏见和不一致性，并在 Github 上开源。新方法使用最大似然估计来基于可用信息联合优化其参数，并消除了明确的受试者拒绝的需要。

3 4 Viewing Condition Adaptation

VMAF 框架允许根据特定观看条件训练预测模型，无论是在手机上还是在超高清电视上。当我们开源 VMAF 时发布的原始模型是基于这样的假设，即观众坐在类似客厅的环境中的 1080p 显示器前，观看距离是屏幕高度（3H）的 3 倍。这是一种通常适用于许多场景的设置。然而，在将该模型应用于手机观看时，我们发现它并不能准确反映观看者对手机质量的感知。特别是，相对于屏幕高度（> 3H），由于较小的屏幕尺寸和较长的观看距离，观看者感知到的高质量视频具有较小的显著差异。例如，在手机上，与其他设备相比，720p 和 1080p 视频之间的区别较小。考虑到这一点，我们培训并发布了 VMAF 手机型测试标准（专门用于手机视频质量评估）。



上面显示了默认型号和手机型号的示例 VMAF 比特率关系。可以理解为，当在手机屏幕上观看时，相同的失真视频将被感知为具有比在高清电视上更高的质量，并且使用手机型号，720p 和 1080p 视频之间的 VMAF 分数差异更小。

最近，我们添加了一个新的 4K VMAF 模型，该模型预测了 4K 电视上显示的视频的主观质量，并从 1.5H 的距离观看。1.5H 的观看距离是普通观众欣赏 4K 内容清晰度的最大距离。4K 模型与默认模型相似，因为这两个模型都捕捉到 1/60 度/像素的临界角频率下的质量。然而，4K 模型假设了更宽的视角，这会影响受试者使用的中央凹与周边视觉。

3 5 Quantifying Prediction Uncertainty

VMAF 是针对一组具有代表性的视频类型和失真进行训练的。由于基于实验室的主观实验的规模限制，视频序列的选择并没有覆盖感知视频质量的整个空间。因此，VMAF 预测需要与置信区间（CI）相关联，该置信区间表示训练过程的固有不确定性。为此，我们最近引入了一种方法，将 VMAF 预测分数与 95% 置信区间相结合，该方法量化了预测在区间内的置信水平。通过使用完整训练数据对预测残差进行自举来建立 CI。从本质上讲，它使用“带替换的重采样”对预测残差训练多个模型。每个模型都会引入一个略有不同的预测。这些预测的可变性量化了置信水平——这些预测越接近，使用完整数据的预测就越可靠。

四 基本原理

VMAF 是一种 Full reference 的视频质量评估方法，主要包括三种指标 [5, 7]: 视觉信息保真度 (VIF: visualquality fidelity) 、细节损失指标 (DLM: detail loss measure) 、时域运动指标/平均相关位置像素差 (Tl: temporal information) 。其中 VIF 和 DLM 是空间域的，一画面之内的特征。TI 是时间域的，多顿画面之间相关性的特征。这些特性之间融合计算总分的过程使用了训练好的 SVM 来预测。

VMAF 基于 SVM 的 nuSvr 算法，在运行的过程中，根据事先训练好的 model，赋予每种视频特征以不同的权重。对每一帧画面都生成一个评分，最终以均值算法进行归总（也可以使用其他的归总算法），算出该视频的最终评分 [1]。

4 1 视觉保真度 (VIF)

视觉信息保真度指标来源于论文《Image Information and Visual Quality》(2006 年 TIP 论文)，该指标认为人眼看到的图像是图像通过 HVS 过滤出来的信息，HVS 本身就是一个失真通道，即人类视觉失真通道，而失真图像只是比原始图像在经过 HVS 之前又多经过了一个图像失真通道，故可以使用信息论的知识将人眼提取的信息与从原始图像提取的信息进行比较，得出最终评测结果。它是一种基于自然场景统计模型 (NSS：natural scene statistics)、图像失真和人类视觉失真建模的新判据。

VIF 模块是基于 NSS 自然场景图像统计学原理计算，《Image Information and Visual Quality》写着，自然场景仅针对现实生活的图片，但是对于计算机合成图片，不适用。所以使用该工具测评时，需要注意片源类型。

4 2 细节损失指标 (DLM)

细节损失指标是另一个图像质量指标，来源于论文《Image Quality Assessment by Separately Evaluating Detail Losses and Additive Impairments》(2011 年 TMM)。该算法分别评估细节损失 (Detail Loss Measure, DLM) 和附加损伤 (Additive Impairment Measure, AIM)。

细节损失是指影响内容可视性的有用视觉信息的损失，附加损伤是指多余的视觉信息，即在测试图像中出现的分散观众对有用内容的注意力的信息，从而导致不好的观看体验。

4 3 时域运动指标/平均相关位置像素差 (TI)

时域运动指标/平均相关位置像素差是一种衡量相邻帧之间时域差分的算法。这个最为简单，仅仅计算像素亮度分量的均值作差即可得到该值。

参考文献

[1] https://netflixtechblog.com/vmaf-the-journey-continues-44b51ee9ed12

[2] https://zhuanlan.zhihu.com/p/94223056

[3] https://www.fengnayun.com/news/content/78346.html

[4] https://github.com/Netflix/vmaf

[5] https://blog.csdn.net/CrystalShaw/article/details/115952476

[6] https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652

[7] https://zhuanlan.zhihu.com/p/54950132