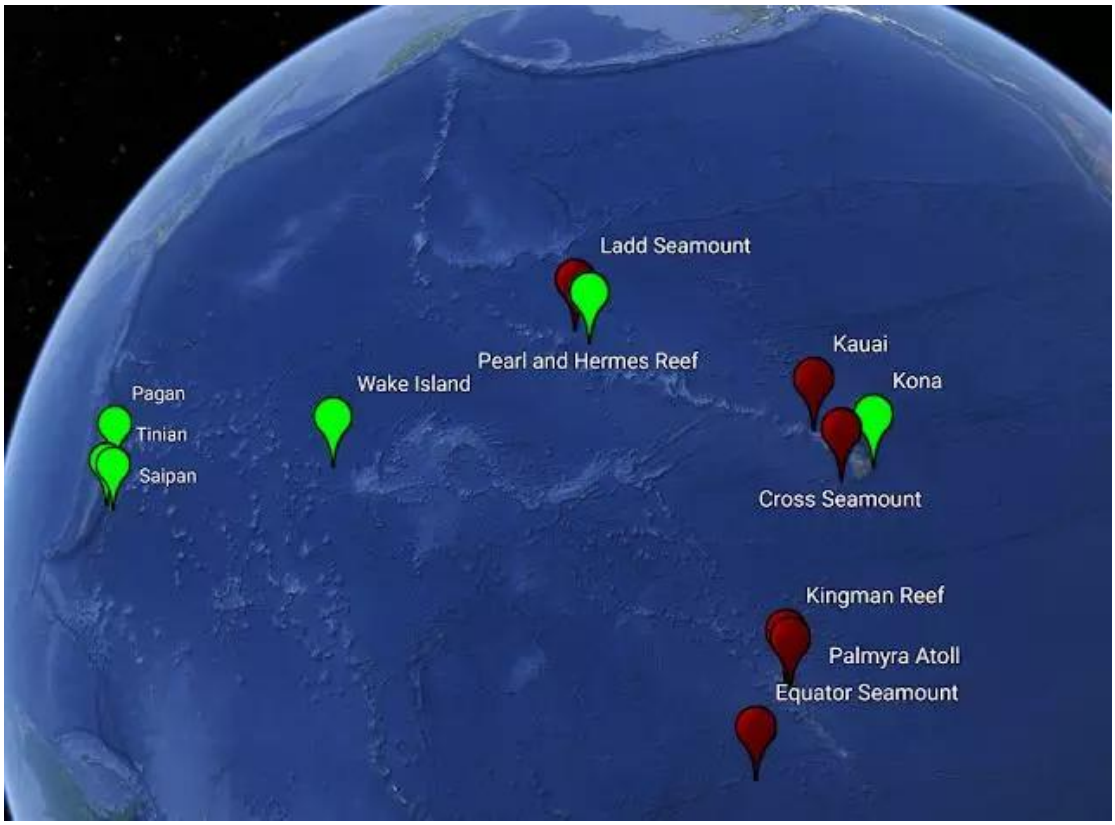
# 利用卷积神经网络对座头鲸进行声学探测

原创：Google谷歌开发者11 月 22 日

文 / Matt Harvey，Google AI Perception 软件工程师

在过去几年中，Google AI Perception 团队开发出音频事件分析技术，并将其应用于 YouTube 上的非语言字幕、视频分类和检索。此外，为了进一步推动社区中的研究，我们还发布了 AudioSet 评估集，并开源了部分模型代码。最近，我们逐渐发现许多保护组织正在收集大量的声学数据，而且我们想知道是否有可能将我们开发的这些技术应用到这些数据中，从而为野生动物监控和保护提供帮助。

作为 AI for Social Good（AI 造福社会）计划的一部分，我们与美国国家海洋和大气管理局 (NOAA) 的太平洋岛屿渔业科学中心合作，开发出相关算法，用于识别 15 年间太平洋多个地点水下录音中的座头鲸叫声。这项研究的结果提供了有关座头鲸出现位置、季节性、日常呼叫行为和种群结构的重要新信息。这一点对于研究偏远的无人岛屿尤为重要，因为科学家之前并未掌握与此类岛屿有关的信息。此外，由于数据集时间跨度很大，因此了解座头鲸发出叫声的时间和地点，将有助于了解多年来座头鲸的分布位置是否发生变化，尤其是在人类海洋活动日渐增多的情况下。该信息在有效减轻对座头鲸的人为影响方面将发挥关键作用。




HARP 部署位置。绿色：目前正在录音的地点。红色：之前录音的地点

## 被动声学监测和 NOAA HARP 数据集

被动声学监测是使用名为水听器的水下麦克风侦听海洋哺乳动物的过程。水听器可用于记录信号，以便能够离线完成探测、分类和定位任务。相较于船基视觉勘察，此方法具有一些优势，包括探测水下动物的能力、更远的探测距离和更长的监控周期。自 2005 年以来，NOAA 已从太平洋岛屿地区 12 个地点的海底水听器收集录音，而该区域是某些座头鲸种群冬季繁衍后代的目的地。

这些数据记录在名为高频声学记录包，或 HARP（Wiggins 和 Hildebrand，

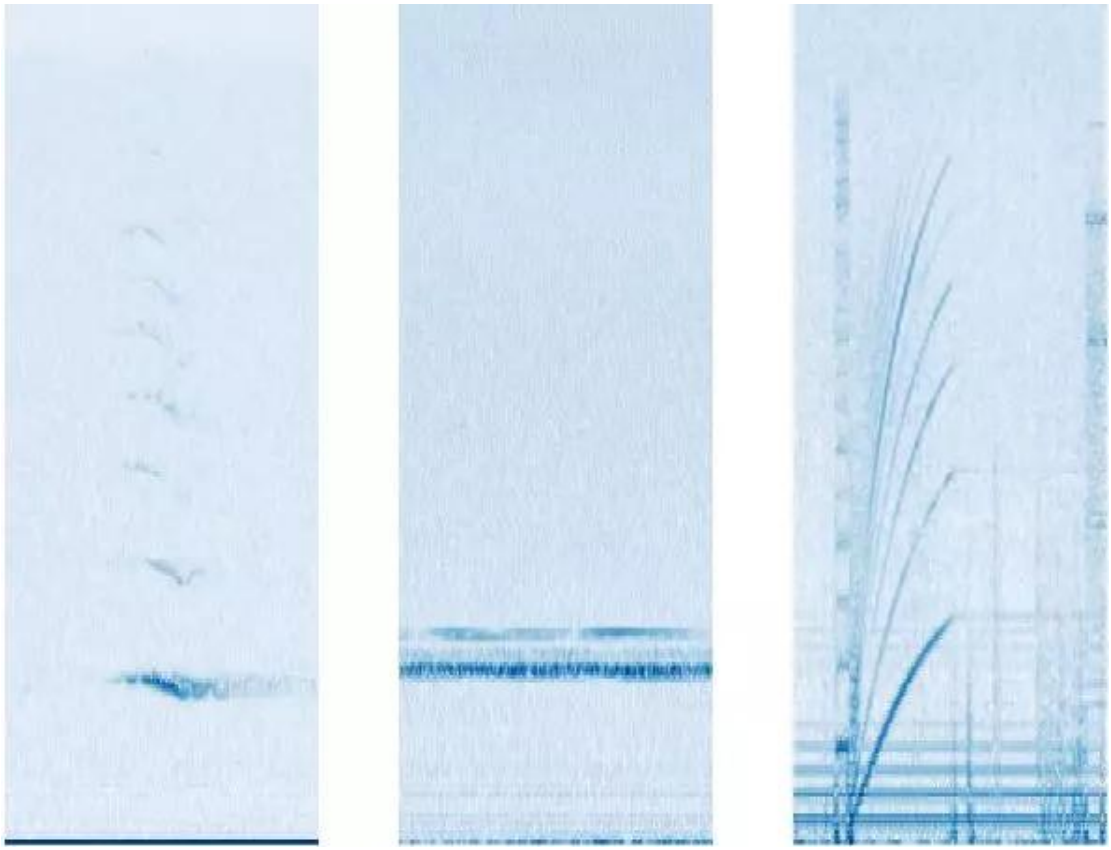
2007 年；点击查看完整

PDF http://cetus.ucsd.edu/Publications/Publications/Wiggins_UT07.pdf）的设备上。NOAA 提供总计约 15 年的音频，或者说从 200 kHz 缩减到 10kHz 后，提供了 9.2 TB 的音频。（由于座头鲸发出的大部分声能量都在 100Hz-2000Hz 的范围内，因此即使使用较低的采样率也几乎不会造成数据丢失。）

从研究的角度来看，在如此大量的数据中识别目标物种是需要完成的第一个重要阶段，这可以为进行更高层次的种群数量、行为或海洋分析提供相关信息。但即便是借助现有的计算机辅助方法，手动标记座头鲸的叫声也非常耗时。

**监督式学习：优化用于探测座头鲸的图像模型**

我们一般会将音频事件探测当作图像分类问题处理，其中图像是指声谱图，即在时频轴上绘制声功率的直方图。



在数据集中找到的音频事件声谱图示例，其中 x 轴为时间，y 轴为频率。左图：一头座头鲸的叫声（特指一个音调单位），中图：来源未知的窄带噪声，右图：来自 HARP 的硬盘噪声

这个示例很好地展示了图像分类器（其目的是区别分类）的功能，因为不同光谱（频率分解）及其时间变化（即不同声音类型的特征）在声谱图中由不同的视觉模式来代表。对于图像模型本身，我们采用 ResNet-50，一个通常应用于图像分类的卷积神经网络架构。该架构已经在非语言音频分类方面取得成功。这是个监督式学习设置，只有手动标记的数据可以用于训练（占整个数据集的 0.2%，在下一部分中，我们会介绍一种利用未标记数据的方法。）

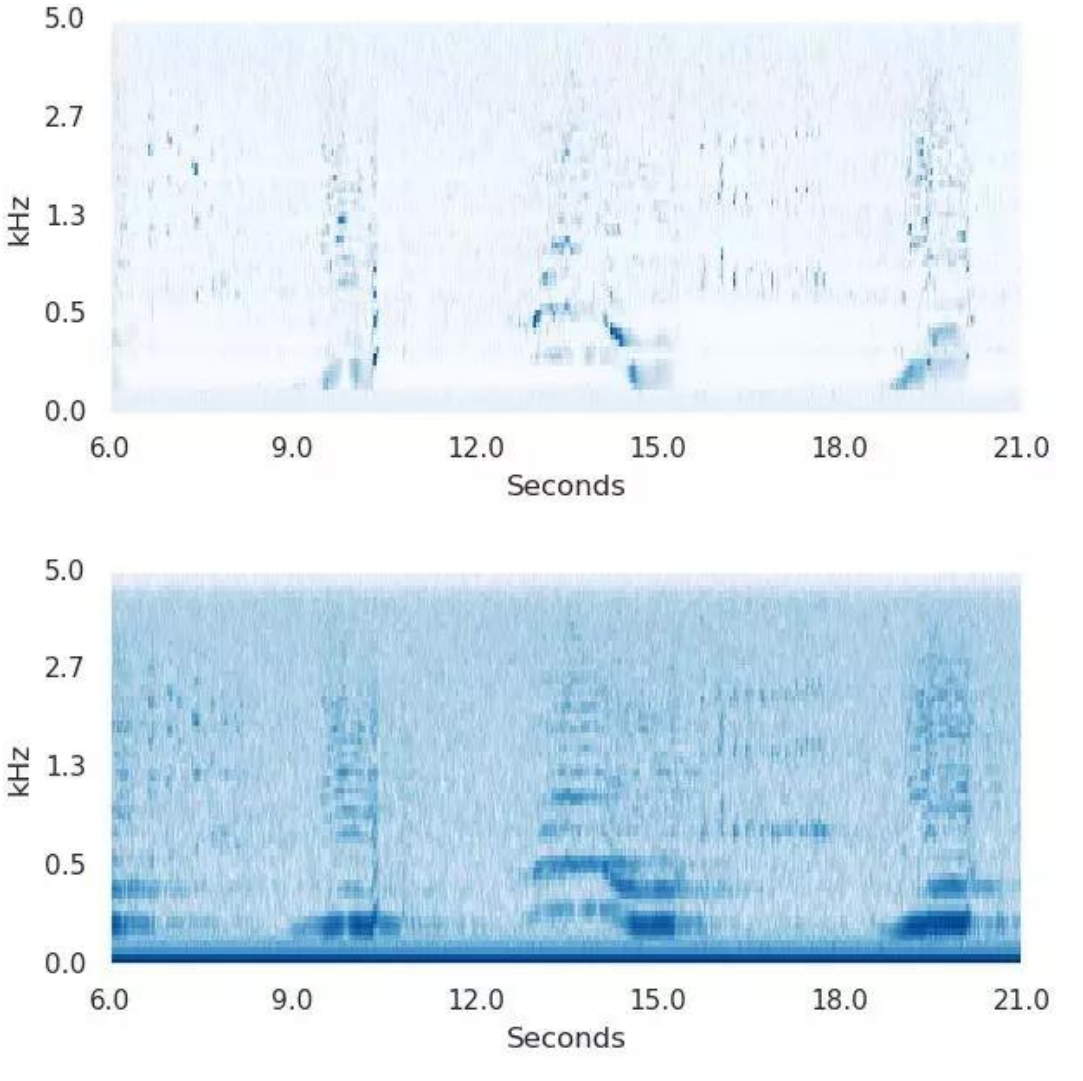
从波形图到声谱图的过程涉及参数的选择和增益调整函数。我们可以遵循

常用的默认选择（对数压缩便是其中之一），但也需要针对特定领域进行调整，以便获得最佳的鲸鱼叫声探测效果。座头鲸发出的声音很多变，但经常会出现持续、调频的音调单位。您可以听听下面这个音频示例：

音频示例来自谷歌开发者 00:0001:15

如果频率完全没有变化，则声谱图中显示的音调单位为水平条。由于座头鲸的叫声经过调频，我们实际看到的是弧线而不是条，但弧线的某些部分近乎水平。

窄带噪声可谓是该数据集面临的特有挑战，而这种噪声通常是由附近的船只和设备自身所发出。在声谱图中，窄带噪声显示为水平线，早期版本的模型会将其与座头鲸的叫声混淆。这促使我们尝试采用通道能量归一化 (PCEN) 方法。该方法可以抑制平稳的窄带噪声。事实证明该方法非常有用，使得鲸鱼叫声探测的错误率降低 24%。




相同 5 个单位的声谱图，声音来源为从上述录音的 0:06 开始截取的座头鲸叫声。上图：PCEN。下图：振幅平方的对数。相对于使用 PCEN 时的鲸鱼叫声，底部深蓝色水平条经过对数压缩后颜色变得更浅

除 PCEN 外，在长时间内进行平均预测也有助于提高查准率。一般的音频事件探测也会获得同样的效果，但对于座头鲸的叫声，查准率的提升度相当大。这可能是因为我们数据集中的叫声是以鲸鱼叫声（可持续超过 20 分钟的结构化单位序列）为主。在一段叫声中的某个单位结束时，另一个单位很可能会在两秒钟之内开始。图像模型的输入涵盖短时窗，但由于叫

声太长，来自较远时窗的模型输出会提供额外信息，这些信息对为当前时窗作出正确预测非常有用。

总体来说，在评估我们 75 秒的音频片段测试集时，该模型可识别某个片段是否包含座头鲸的叫声，其中查准率为 90% 以上，查全率为 90%。但在解释这些结果时应该谨慎小心；训练和测试数据均来自类似的设备和环境条件。即便如此，似乎有望针对部分非 NOAA 来源音频进行初步检测。

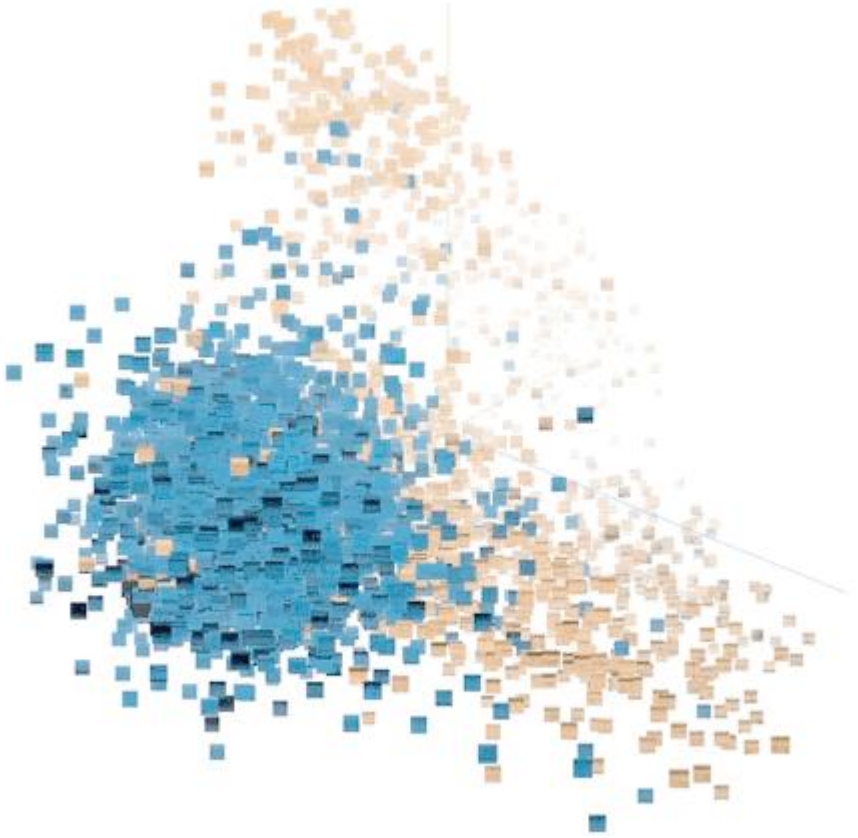
**非监督式学习：用于查找类似叫声单位的表征**

“此数据中哪些是座头鲸的叫声？”，对于这一问题，我们有不同的解决方法。首先取得几个座头鲸叫声的示例，然后针对每个示例，在数据集中寻找更多与之类似的声音。这里类似的定义可以通过我们将其认定为监督式学习问题时所使用的相同 ResNet 来学习。在监督式学习中，我们在 ResNet 输出的基础之上，使用标签学习分类器。而在非监督式学习中，当相应的音频示例在时间上接近时，我们支持一对 ResNet 输出向量在欧几里得距离上相互靠近。借助该距离函数，我们可以检索到更多与给定示例类似的音频示例。将来，对于用于区分不同座头鲸单位类型的分类器而言，这可能是非常有用的输入信息。

为了学习距离函数，我们采用了《音频语义表征的非监督式学习》("Unsupervised Learning of Semantic Audio Representations") 中介

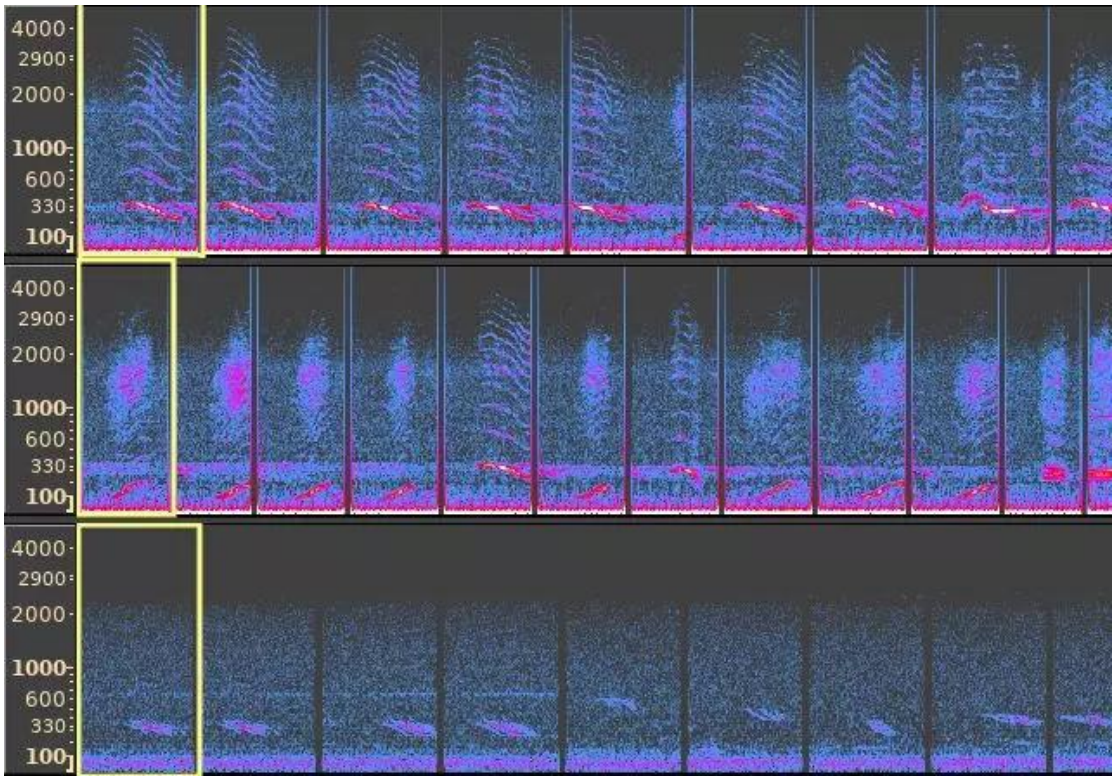
绍的一种方法。该方法依据的观点是，时间上的相近与意义上的相近有关。该方法随机抽取样本，并以三个样本为一组，每组均包含锚点、正值和负值。研究人员对正值和锚点进行采样，以便它们可以差不多同时开始。举例而言，我们应用的样本组包含座头鲸单位（锚点），同一头鲸鱼发出的单位相同的重复叫声（正值），以及来自其他月份的背景噪声（负值）。将这三个样本传递到 ResNet（带有绑定权重）即可将其表示为三个向量。通过一个所学距离函数忠于语义相似性的间隔，使迫使锚点和负值之间距离大于锚点和正值之间距离的损失最小化。

对标记点示例进行主成分分析 (PCA) 使我们可以将结果可视化。座头鲸与非座头鲸之间的距离显而易见。您可以使用 TensorFlow Embedding Projector 自行探索。尝试将 Color by 更改为 class_label 和 site 中的每一个。此外，尝试将 PCA 更改为投影仪中的 t-SNE，以可视化呈现优先保留相对距离而非样本方差。



非监督式表征中的 5000 个数据点示例。（橙色：座头鲸。蓝色：非座头鲸

考虑到单个 “查询” 单位，我们使用嵌入向量之间的欧几里得距离，在整个信息库中检索最近的相邻单位。在某些情况下，我们发现数百个查准度良好的相同单位实例。

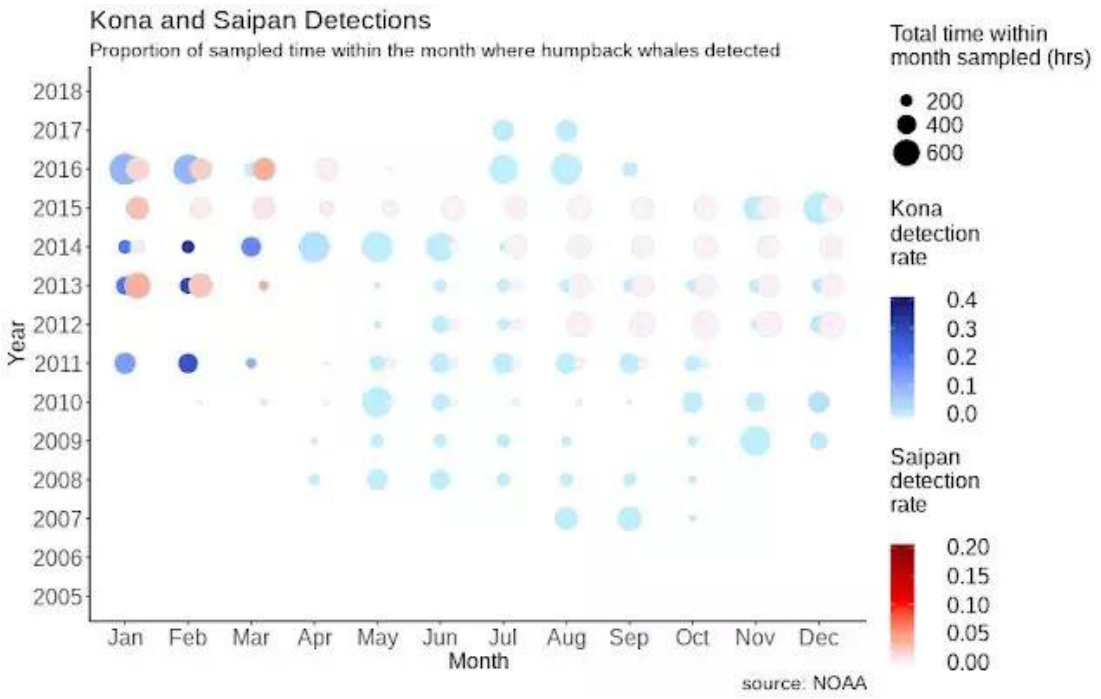



手动选择的查询单位（用方框标记）和使用非监督式表征发现的最近相邻单位

我们计划日后使用这些单位为区分叫声单位的分类器构建训练集。我们还可以利用这些单位扩展用于学习座头鲸探测器的训练集。

**监督式分类器对整个数据集的预测**

我们绘制了按时间和位置分组的模型输出总结图。我们并没有在所有年份对所有地点都进行部署。周期性暂停（例如：开启 5 分钟，关闭 15 分钟）能在有限的电池电量下实现更长时间的部署，但时间安排可能有所不同。为了处理这些可变性来源，我们会考虑探测到座头鲸叫声的取样时间与一个月中总录音时间的比例：




科纳和塞班站以年/月为轴的叫声探测时间密度

显著的季节性变化与已知的模式相符，其中座头鲸种群夏季在阿拉斯加附近进食，然后迁移到夏威夷群岛附近繁衍后代。这是一个很好的模型合理性检查。

我们希望对完整数据集的预测能够为 NOAA 的专家提供相关信息，以便其更深入地分析这些种群的状态，以及鲸鱼受到人为影响的程度。Google 致力于加快机器学习的应用速度，以应对世界上最大的人道主义和环境挑战。而我们也希望除已经取得的成就外，未来我们还能够取得一系列的成功。

**致谢**

我们要感谢 Ann Allen（NOAA 太平洋岛屿渔业科学中心）提供大量的地面实况数据、多轮实用反馈，以及本文中的部分内容。Karlina Merkens（NOAA 附属机构）提供了更深入的实用指导。我们还要感谢 NOAA 太平洋岛屿渔业科学中心全体工作人员，感谢他们收集和分享声学数据。

Google 内部的 Jiayang Liu、Julie Cattiau、Aren Jansen、Rif A. Saurous 和 Lauren Harrell 也为本文的推出做出贡献。我们要特别感谢 Lauren，其负责设计了分析部分的图表，并使用 ggplot 完成了绘制。