一、文献信息

（1）论文题目：Cumulo: A Dataset for Learning Cloud Classes

（2）作者：Valentina Zantedeschi, Fabrizio Falasca, Alyson Douglas, Richard Strange, Matt J. Kusner, Duncan Watson-Parris

（3）发表途径：NeurIPS 2019

（4）发表时间：第一版 2019 年，第二版 2020（本文笔记依据第二版）

二、问题意义

云在气候系统中起着至关重要的作用。它是所有降水的来源，对地球的辐射收支影响重大。云的任何变化都会影响环境，这些变化又会反馈云的形成和行为。未来气候预测中最大的不确定性来源之一就是对云的建模和理解不同云类型如何与气候系统相互作用的局限性。降低这种不确定性的第一步就是在高时空分辨率下准确分类云类型。现有的两种分类方法，SCP 分类和 CloudSat 云雷达，都具有一些明显的局限性。为了克服这些，本文介绍了一个用于训练和评估全球云分类模型的基准数据集 Cumulo。它由一年 1 千米分辨率的 MODIS 高光谱图像和 CloudSat 云标签的像素宽度“轨迹”组成。将这些互补的数据集整合在一起是机器学习社区能够开发新技术的首要前提，这将极大地造福于气候社区。本文作者们在一个月的 Cumulo 数据中应用了深度生成模型架构，并首次提出了由主动和被动卫星传感器组合而成的全球高分辨率时空云分类。他们的结果在物理上是合理的。

三、思路方法

Cumulo 将中分辨率成像光谱仪（MODIS）的全球 1km 分辨率图像与 CloudSat 数据的精确测量特性相结合。它包含一年 1354×2030 像素的 MODIS 高光谱图像和 Cloudsat 云标签的像素宽度“轨迹”，对应于 8 个世界气象组织（WMO）属。

本文提出的数据集包含 105120 个地理定位和高光谱图像，并提供不同来源的通道组合。每个卫星图像是在给定的时间 t（每五分钟一张）和给定的位置 l（每个像素与经纬度对相关联）获取的。

（1）MODIS AQUA Calibrated Radiances 的选定辐射通道完全捕获云分类所需的物理特性，并用于训练；$\mathbf{X}^{t,l}$表示来自 MODIS AQUA 校准辐射的 13 个训练频道

（2）MODIS AQUA Cloud Product 通道是用于描述云物理特性的检索功能，可用于验证；$\mathbf{V}^{t,l}$来自 MODIS AQUA Cloud Product 的八个验证通道，它们提供通过将红外发射和应用于 MODIS 原始波段的太阳反射技术相结合而获得的物理和辐射云特性；

（3）MODIS Cloud Mask 可检测云的存在；$\mathbf{C}^{t,l}$是由 MODIS Cloud Mask 导出的云掩模，确定的云像素标记为 1，其他像素标记为 0；

（4）2B-CLDCLASS-LIDAR 提供沿卫星轨道不同高度发现的云类型和其他云特征。$\mathbf{L}^{t,l}$是源自它的覆盖标签掩码，指示每个像素在不同高度识别的最常见云类型。

与世界气象组织八个属相对应的可能云类型有 “层云（St）”、“层积云（Sc）”、“积云（Cu，包括浓积云）”、“雨云层（Ns）”、“高积云（Ac）”、“高层云（As）”、“深对流（积雨云，

Dc)”以及“卷云和卷层云（Ci)”。2B-CLDCLASS-LIDAR 变量定义在多达 10 个不同的垂直层上。每一层对应于给定时间和位置的一个确定云簇。由于识别云的类型和数量不可避免地随着空间和时间的变化而变化，因此每一层不是在固定高度间隔，而是其数量和厚度随像素的变化而变化。

其中存在一些通道仅在白天可用的情况，因为它们直接或间接依赖于日光辐射。一般来说，由于手工艺品造成的缺失值用最接近的（时间和空间上的）可用值填充。

为了展示该数据集，作者还使用可逆流动生成模型（IResNet）来提供基线性能分析。由于 Cumulo 模型的标签数比较少，因此在像素级进行实验比需要全标签掩码或注释的通用语义分割模型获得的分类性能更好。实验从第一个月数据的（2008 年 1 月）提取 3x3 的像素片，并预测每片的标签。当注释可用时，模型使用每片识别到的最常见云类型作为目标。这里要收集两组片数据：一组在卫星轨道上的任何注释像素周围采样的带标签数据；另一组从未注释区域随机选择的未带标签数据。同时本文还使用了可用的云屏蔽来限制对具有高云覆盖率的片的分类。同时还使用可用的云屏蔽来限制对具有高云覆盖率的像素块的分类。作者已经事先训练好了一个混合可逆残差网络，这个网络使得他们可以利用标记集和未标记集，并学习一种表示法，而且类分布还可以进一步细分为细粒度类。因为实际上，云的类型并不局限于研究得很好的 WMO 属，在云群落中，能够识别更多种类的云是一个尚待解决的问题。因此作者留出了分类的进一步细化空间。该模型结合了一个深度生成函数和一个线性分类，同时通过最大化像素片及其标签上的联合对数可能性来训练。

本文将每个张量$\mathbf{X}^{t,l}$分割成 3x3 像素的非重叠片。如下式：

$$\mathbf{X}^{t,l} = \begin{bmatrix} \mathbf{X}^{t,l}_{(0,0)} & \mathbf{X}^{t,l}_{(0,1)} & \cdots & \mathbf{X}^{t,l}_{(0,W)} \\ \mathbf{X}^{t,l}_{(1,0)} & \mathbf{X}^{t,l}_{(1,1)} & \cdots & \mathbf{X}^{t,l}_{(1,W)} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{X}^{t,l}_{(H,0)} & \mathbf{X}^{t,l}_{(H,1)} & \cdots & \mathbf{X}^{t,l}_{(H,W)} \end{bmatrix}$$

实验目的为了学习各片与云分类标签之间的映射关系。至于目标值，作者保留了与数据块相关的标签中最常出现的云类型。

在学习过程中，作者部署他们训练好的混合可逆残差网络，这个网络可通过分解得到张量 $\mathbf{X}$ 数据分布的一种潜在表示。

$$p_\theta(\mathbf{X}) = \sum_{\mathbf{z}} p_\theta(\mathbf{X}|\mathbf{z})p(\mathbf{z})$$

这种网络的特殊之处是它包含了一系列与任何输入互逆的映射关系，并且每个图像 $\mathbf{X}$ 和每个潜在点 z 之间存在直接互逆的关系。因此，可以通过最大限度增大图像 X 与他们目标标签 Y 的联合似然性来最大限度优化参数 θ，抽象为下式：

$$
\begin{aligned}
p_\theta(\mathbf{y}, \mathbf{X}) &= p(\mathbf{y} \mid \mathbf{X}) p(\mathbf{X}) \\
&= p(\mathbf{y} \mid \mathbf{z}) p(\mathbf{z}) \left| \det \left(\frac{d\mathbf{z}}{d\mathbf{X}} \right) \right| \\
&= p(\mathbf{y} \mid \mathbf{z}) p(\mathbf{z}) \prod_{i=1}^{D} \left| \det \left(\frac{d\mathbf{h}_i}{d\mathbf{h}_{i-1}} \right) \right|
\end{aligned}
$$

因为网络没有提供监督，所以上式只能对已标签的集合进行估计。对于无标签集，需要最小化标签的熵以提高预测准确性。因此最终的目标函数如下式：

$$
\begin{aligned}
\max \sum_{(x_k, y_k) \in \mathcal{L}} (\log p(y_k | z_k) + \log p(z_k)) + \sum_{x_k \in \mathcal{U}} \left(\sum_{y \in \mathcal{Y}} \log p(y | z_k) + \log p(z_k) \right) + \\
+ \sum_{j=1}^{D} \log \left| \det \left(\frac{dh_j}{dh_{j-1}} \right) \right|
\end{aligned}
$$

四、实验结论

作者提供了一个可以使用 Cumulo 执行的任务——在全球和每日范围内对云进行半监督分类——并对其进行基准性能分析。实验从第一个月数据的（2008 年 1 月）提取 3x3 的像素片，并预测每片的标签。本文将标记后的数据分为 70%训练，10%验证和 20%测试。同时使用可逆流动生成模型（IResNet）来提供基线性能分析。结果如下

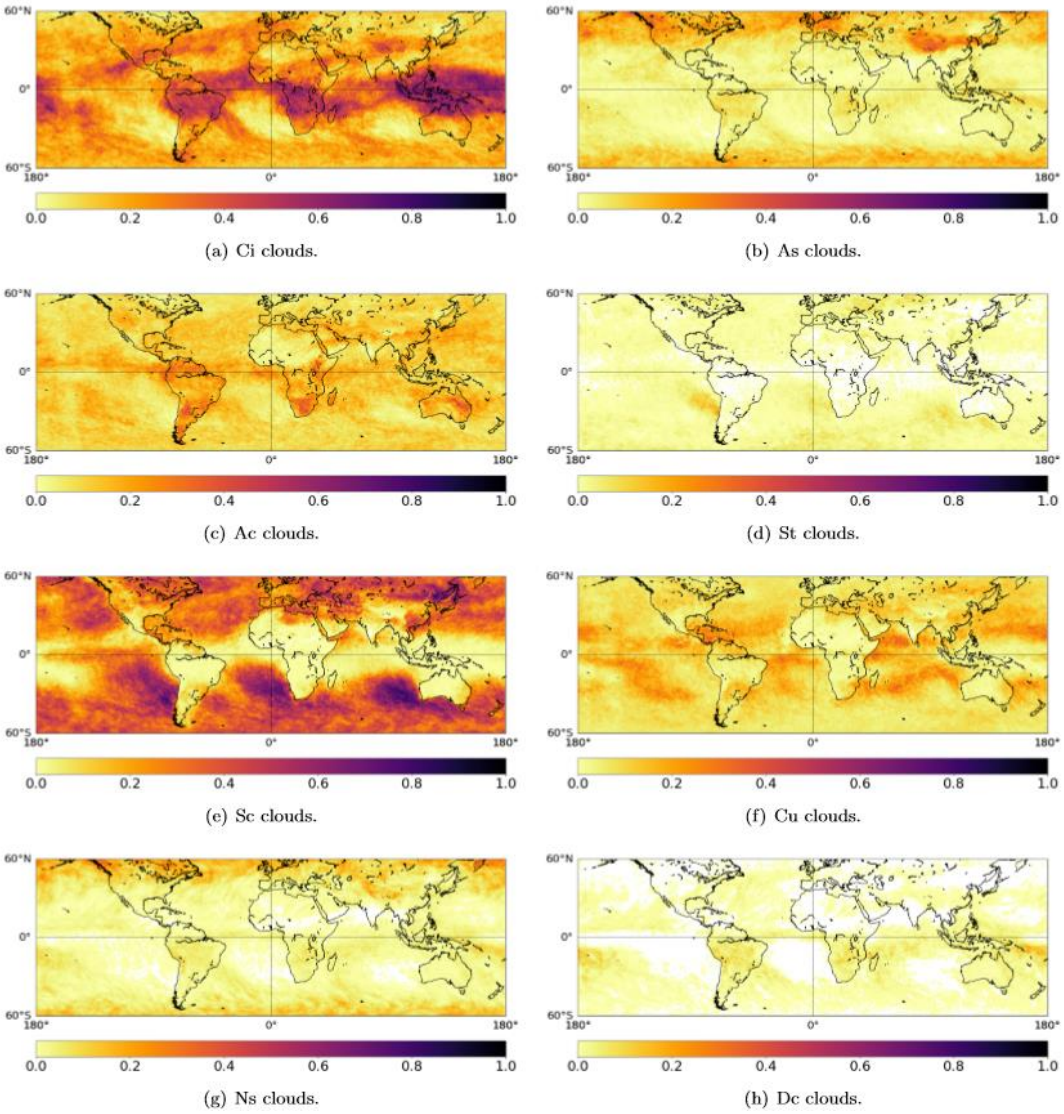
（1）对于在验证集上具有最佳平均准确度的模型（已训练网络），实验测试了检验分类准确度、F1 分数和联合指数上交集（包络每类数据以及平均值）

	Ci	As	Ac	St	Sc	Cu	Ns	Dc	Mean
Accuracy (%)	81.30	84.50	88.29	97.73	88.90	92.40	90.92	98.84	90.36
F1 score	0.68	0.43	0.45	0.58	0.80	0.40	0.47	0.58	0.55
IoU index	0.52	0.28	0.29	0.41	0.66	0.25	0.31	0.41	0.39

表 1 测试集上的 IResNet 分类结果。

（2）2008 年 1 月由 IResNet 预测的出现的云类别。

其中，（e）图层积云的预测与其主要出现在大洋的上升流区域的空间特点是一致的。深对流更接近赤道区域；卷云（高）更广泛地分布在全球。卷云的最常出现地点大致在热带辐合带上方，并且与深对流在空间上相关，这与 Mace 等人的观点一致。同时，实验的这些热图结果与 Sassen 等人对于一年和两年内 CloudSat 标记云类的出现情况的报告非常相似。这些都可以侧面反映该数据集具有其合理性。



(a) Ci clouds. (b) As clouds.

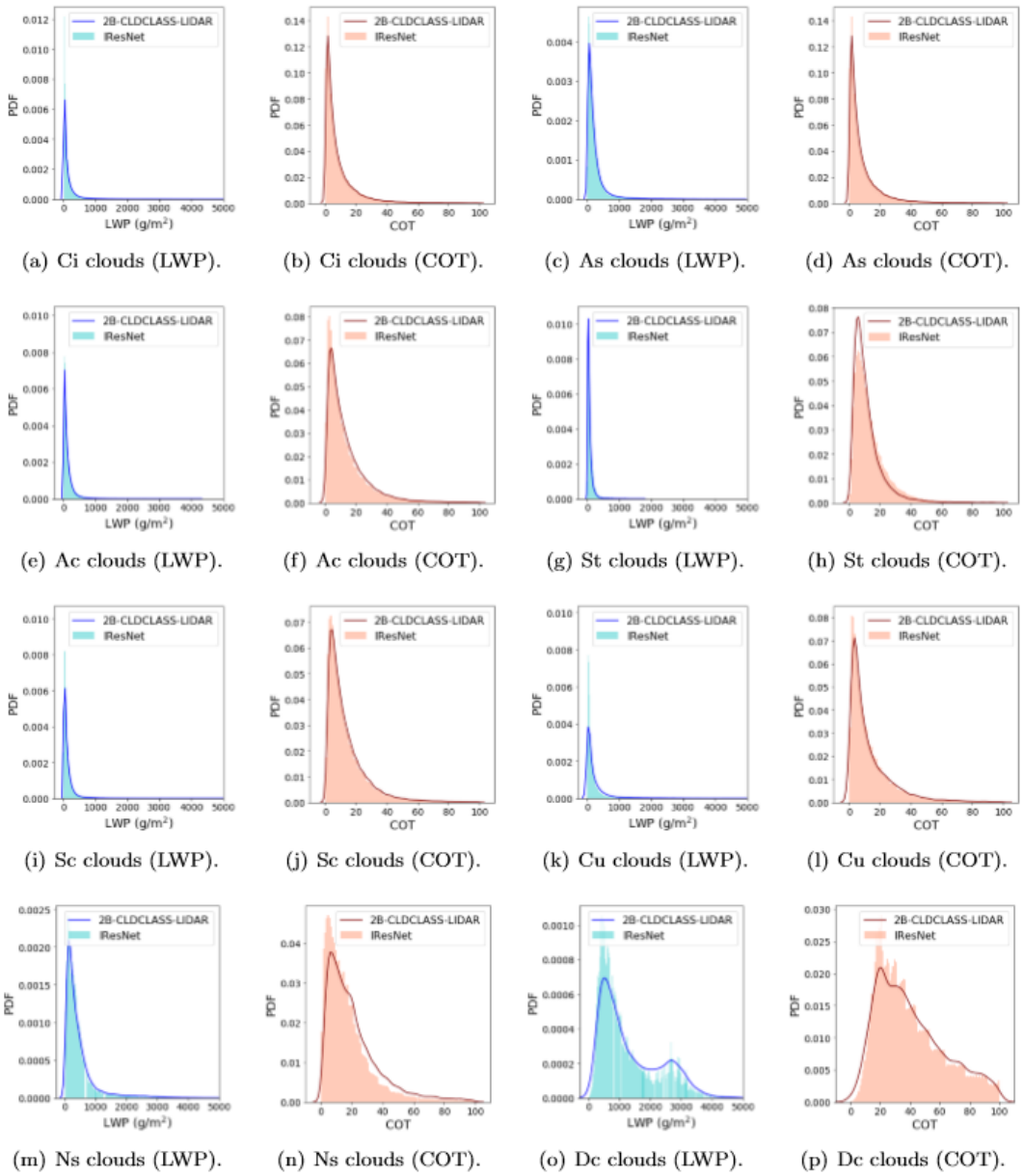
(c) Ac clouds. (d) St clouds.

(e) Sc clouds. (f) Cu clouds.

(g) Ns clouds. (h) Dc clouds.

（3）附加基于物理的评估

作者还考虑了预测等级的液态水路径（LWP）和云光学厚度（COT）变量的分布以及 CloudSat 给出的地面真实度。LWP 是给定点上整个大气柱中的液态水总量，COT 是测量云层底部和顶部之间厚度的单位。这两个变量的预测分布和地面真实值之间的差异是最小的。同时本文还通过 Kullback-Leibler 散度和 Wasserstein 距离对预测分布和 CloudSat 分布进行了定量比较。结果表明尽管 LWP 和 COT 变量中的地面真值存在较大差异（见表 4），但“深对流”类与最大精度（98.84%）相关。



(a) Ci clouds (LWP). (b) Ci clouds (COT). (c) As clouds (LWP). (d) As clouds (COT).

(e) Ac clouds (LWP). (f) Ac clouds (COT). (g) St clouds (LWP). (h) St clouds (COT).

(i) Sc clouds (LWP). (j) Sc clouds (COT). (k) Cu clouds (LWP). (l) Cu clouds (COT).

(m) Ns clouds (LWP). (n) Ns clouds (COT). (o) Dc clouds (LWP). (p) Dc clouds (COT).

2008 年 1 月 IResNet 预测的云类液态水路径（LWP）和云光学厚度（COT）的概率分布

五、启发思考

机器学习是我们现在都绕不开的一个重要议题，也是我们对各方面研究的一个重要方法。本文提出的数据集也提出了新的挑战，比如，在单个云类别中，模型仍可以区分中尺度上物理特性稍有不同的各种云组织，这就意味着每个给定类别都存在子类别。对于气候和机器学习社区而言，提出一种新颖的无监督模型，直接访问能观察到的粗标签的细粒度类，也是一条重要的新研究路线。对我而言，这篇论文比较不同的是，他提出的是新的数据集，而不是针对某个问题的解决算法，而事实上在机器学习中，数据集是非常重要的基础资料。这给了我一定研究上的启发，现有的数据集的问题在哪里，如何合理的组织出有效的数据集，如何评估数据集，这都是需要思考的问题。