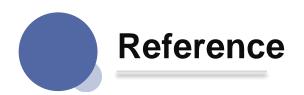


Report

Priceless lab

汇报人: Lilian

2022/8/23



A Variational Information Bottleneck Approach to Multi-Omics Data Integration

Changhee Lee University of California, Los Angeles Mihaela van der Schaar University of Cambridge University of California, Los Angeles Alan Turing Institute

多组学数据集成的变分信息瓶颈方法

文献来源: 2021/艾伦图灵研究所+剑桥大学

Abstract & introduction

来自多种组学技术的数据整合在生物医学研究中变得越来越重要。由于组学平台的非一致性和技术限制,对多个组学的这种综合分析(本文称为 views)涉及**从各种视图缺失模式的不完整观察中学习。**

由于来自不同数据源(例如 TCGA)的实验设计或组成的限制,集成样本通常具有一个或多个完全缺失的组学,具有各种缺失模式。

丢弃缺少组学的样本会大大减少样本量(尤其是在集成许多组学层时),简单的均值插 补会严重扭曲数据的边缘分布和联合分布。

这是一个挑战,因为:

- i) 需要正确处理观察到的视图内部和之间的复杂交互, 以获得最佳预测能力;
- ii)需要灵活集成具有各种视图缺失模式的观察结果。

在本文中,我们**将多组学数据集成建模为从不完整的多视图观察中学习,其中我们将来自每个组学数据的观察称为**view (例如, DNA 拷贝数和 mRNA 表达)。

为解决缺失视图问题,**我们的目标模型要求不仅可以学习与目标任务相关的复杂视图内和视图间交互,而且可以灵活地集成观察到的视图,而不管它们的视图缺失模式如何**。

本文提出了一种用于不完全多视图观测的深度变分信息瓶颈(IB)方法。

将IB框架应用于观察到的视图的边缘和联合表示,以关注与目标相关的views内和联合的信息。我们的方法由四个网络组件组成:一组视图特定编码器、一组视图特定预测器、一个专家乘法(PoE)模块和一个多视图预测器。更具体地说,为了灵活地集成观察到的视图,无论视图丢失模式如何,我们将联合表示建模为边缘表示上的PoE,这将被多视图预测器进一步利用。

最重要的是,通过将联合表示建模为边缘表示的产物,我们可以有效地从具有各种视图 缺失模式的观察视图中学习。在真实数据集上的实验表明,优于最先进的基准测试。

r

method & results

Incomplete Multi-View Observations.

多视图观察不完整。大多数现有方法都需要多阶段培训,即根据**插补方法**构建完整的多视图观察结果,然后培训多视图模型,或者**依靠辅助推理步骤生成缺失的视图**。

作者在[20-Multimodal generative models for scalable weakly-supervised learning.2018.]中介绍了一种生成模型,该模型利用潜在因子分解来实现交叉视图生成,而不需要多阶段训练机制或额外的推理步骤。

然而,已有方法是以**纯无监督**的方式训练的。因此,虽然与视图重建相关的信息将在学习表示中很好地捕获,但与目标任务相关的信息可能会丢失。



method & results

Incomplete Multi-View Observations.

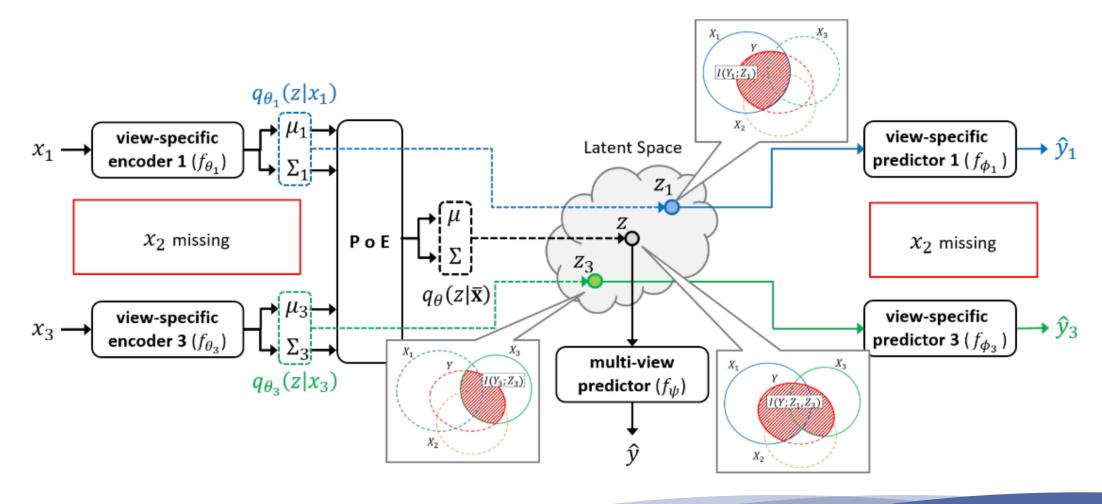
信息瓶颈(IB)原则是一种信息论方法,它定义了任务相关表示的直观概念,即在具有简明表示和提供良好预测能力之间的基本权衡。最近研究了IB框架,以解决多视图问题。作者在[24-Deep multi-view information bottleneck.2019]中使用辅助层将来自每个视图的边缘表示组合成联合表示,在该联合表示上应用IB原理。然而,这项工作无法处理训练和测试期间缺失的视图。

在[25Learning robust representations via multi-view information bottleneck. 2020]中,IB原则被扩展到两个视图的无监督设置,以学习两个视图共有的鲁棒表示。因此,这不适用于多组学数据整合,其目标是以监督的方式灵活整合通常包含共同和互补信息的不完整视图(即来自不同组学层的观察结果)。



method & results

DeepIMV



four network components

- 一组 V 视图特定的编码器,参数化为 $\theta = (\theta_1, \dots, \theta_V)$,即 每个编码器随机地 将来自每个单独视图的观察映射到一个公共潜在空间;
- 专家乘积 (product-of-experts, PoE) 模块,将边缘潜在表示整合为公共空间中的联合潜在表示;
- 由ψ参数化的多视图预测器,其提供基于联合表示的标签预测;
- 由 $\phi = (\phi_1, \dots, \phi_v)$ 参数化的一组V个视图特定预测器,每个视图特定预测器基于由相应的视图特定编码器确定的边缘潜在表示来提供标签预测。

在本文的其余部分, 我们将经常使用 f 来表示确定性映射, 并使用 q 来表示随机映射。

Toward Task-Relevant Representations

在包含边缘和联合的共同潜在空间中寻找任务相关表征观察的各个view对于解决不完整的多视图问题至关重要。为此,我们应用 IB 原则,因为它旨在通过丢弃与目标任务无关的尽可能多的输入信息来学习与任务相关的表示,从而鼓励预测器对过度拟合具有鲁棒性.

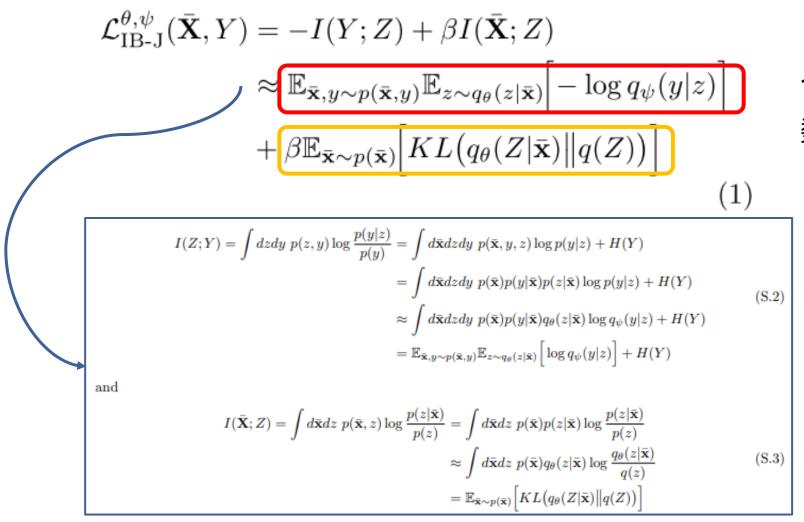
定义Z为公共潜空间,边际表示(marginal representation) $Z_v \in \mathcal{Z}$ 是Xv的随机编码, $q_{\theta_v}(Z_v|X_v)$

它由v ∈ V的第v个视图特定编码器定义。

类似地,我们认为联合表示 $Z \in Z$ 是 $\bar{\mathbf{X}}$ 的随机编码,它由组合V视图特定编码器的输出的编码器块 $q_{\theta}(Z|\bar{\mathbf{X}})$ 定义。

然后,给定从 $q_{\theta}(z|\bar{\mathbf{x}})$,得出的潜在表示 $\mathbf{z} \in \mathbf{z}$,多视图预测器根据定义为 $q_{\psi}(Y|z)$ 的分布来估计目标标签Y。

为了学习用于预测目标Y的观察视图的联合方面,我们基于以下损失对z应用IB原理:



β≥0是为平衡两 个信息量而选择的系 数。

Product-of-experts

ARTICLE _____ Communicated by Javier Movellan

Gatsby Computational Neuroscience Unit, University College London, London

Training Products of Experts by Minimizing Contrastive Divergence

Geoffrey E. Hinton hinton@cs.toronto.edu

WC1N 3AR, U.K.

2002

输入:

包括图像数据集x以及M个输入方式(多模态)

目标:

单一模型可以学习在可能模式的任意子集 \mathcal{Y} 上捕获图像分布 $p(x \mid \mathcal{Y}), \forall \mathcal{Y} \subseteq \{y_1, y_2 \dots y_M\}$ 在本文中,我们考虑了四种不同的模式:

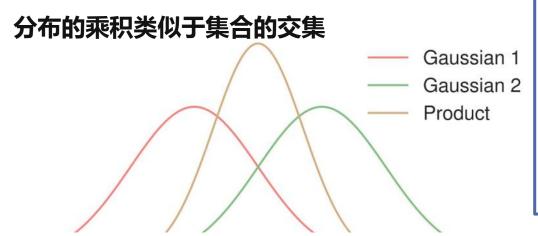
- 文本 text
- 语义切分 segmentation
- 草图 sketch
- 风格参照 style reference

学习基于任何子数据 $\forall \mathcal{Y} \subseteq \{y_1, y_2 \dots y_M\}$ 的图像分布 $p(x \mid \mathcal{Y})$ 是一项挑战

因为它需要一个生成器同时去建模 2^M 个分布

Product-of-experts modeling

每个输入模态都增加了合成图像必须满足的约束,满足所有约束的图像集是满足单个约束的图像集的交点。



可以通过将它们的概率分布相乘然后重新规范化 来组合相同数据的多个潜在变量模型。这种组合单个 "专家"模型的方式使得很难从组合模型中生成样本, 但很容易推断出每个专家的潜在变量的值,因为组合 规则确保不同专家的潜在变量在给定数据时是有条件 独立的。

即**联合条件概率分布** $p(x\mid y_i,i_j)$ 与单条件概率分布 $p(x\mid y_i)$ 和 $p(x\mid y_j)$ 的乘积成正比

使用Product-of-experts来对潜在(latent)条件分布进行建模 $p(z\mid \mathcal{Y}) \propto p'(z) \prod_{y_i \in \mathcal{Y}} q(z\mid y_i)$

- p'(z) 是先验概率分布
- 每个 $expert\ q(z\mid y_i)$ 是编码器预测的单一模态的分布

Product-of-Experts for Incomplete Views

有一个问题很突出。我们应该如何设计联合表示,使其将观察到的视图的边缘表示与任意视图缺失模式集成在一起?我们的解决方案是使用专家乘积 (PoE) 将联合后验(the joint posterior) $q_{\theta}(Z|\bar{\mathbf{X}})$ 分解为边际后验(the marginal posteriors) $q_{\theta_v}(Z|X_v)$ 的乘积,对于 v \in V。形式上,联合后验可以定义如下[20]:

$$p(z|\bar{\mathbf{x}}) \approx C \cdot p(z) \prod_{v \in \mathcal{V}} q_{\theta_v}(z|x_v) \stackrel{\text{def}}{=} q_{\theta}(z|\bar{\mathbf{x}})$$
 (2) C是归一化常数。

$$\mathcal{L}_{\text{IB-M}}^{\theta_v,\phi_v}(X_v,Y_v) = -I(Y_v;Z_v) + \beta_v I(X_v;Z_v)$$

$$\approx \mathbb{E}_{x_v,y\sim p(x_v,y)} \mathbb{E}_{z\sim q_{\theta_v}(z|x_v)} \Big[-\log q_{\phi_v}(y|z) \Big]$$

$$+ \beta_v \mathbb{E}_{x_v\sim p(x_v)} \Big[KL \Big(q_{\theta_v}(Z_v|x_v) \big\| q(Z_v) \Big) \Big]$$
(3)

Algorithm 1 Pseudo-code for training DeepIMV

```
Input: \mathcal{D} = \{\bar{\mathbf{x}}^n, y^n\}_{n=1}^N, learning rate \eta, mini-batch
size n_{mb}, coefficients \alpha, \beta, \{\beta_v\}_{v=1}^V, and \mu_0, \Sigma_0
Output: DeepIMV parameters (\theta, \phi, \psi)
Initialize parameters (\theta, \phi, \psi) via Xavier Initializer
repeat
     Sample a mini-batch: \{\bar{\mathbf{x}}^n, y^n\}_{n=1}^{n_{mb}} \sim \mathcal{D}
     for n=1,\cdots,n_{mb} do
          for v \in \mathcal{V}^n do
               \mu_v^n, \Sigma_v^n \leftarrow f_{\theta_v}(x_v^n)
               z_v^n \leftarrow \mu_v^n + \epsilon \cdot \Sigma_v^n where \epsilon \sim \mathcal{N}(0,1)
               \hat{y}_v^n \leftarrow f_{\phi_v}(z_v^n)
          end for
          \mu^n, \Sigma^n \leftarrow PoE((\mu_v^n, \Sigma_v^n)_{v \in \mathcal{V}^n})
          z^n \leftarrow \mu^n + \epsilon \cdot \Sigma^n where \epsilon \sim \mathcal{N}(0,1)
          \hat{y}^n \leftarrow f_{\psi}(z^n)
     end for
     \hat{\mathcal{L}}_{\text{IB-J}}^{\theta,\psi} = \frac{1}{n_{mb}} \sum_{n=1}^{n_{mb}} \left( - \sum_{c=1}^{C} y^n[c] \log(\hat{y}^n[c]) \right)
                       + \beta \cdot KL(\mathcal{N}(\mu^n, \Sigma^n)||\mathcal{N}(\mu_0, \Sigma_0)))
     \hat{\mathcal{L}}_{\text{IB-M}}^{\theta,\phi} = \frac{1}{n_{mb}} \sum_{n=1}^{n_{mb}} \sum_{v \in \mathcal{V}^n} \left( - \sum_{c=1}^C y^n[c] \log(\hat{y}_v^n[c]) \right)
                       + \beta_v \cdot KL(\mathcal{N}(\mu_v^n, \Sigma_v^n)||\mathcal{N}(\mu_0, \Sigma_0)))
     (\theta, \phi, \psi) \leftarrow (\theta, \phi, \psi) - \eta \nabla_{(\theta, \phi, \psi)} (\hat{\mathcal{L}}_{\text{IB-I}}^{\theta, \psi} + \alpha \hat{\mathcal{L}}_{\text{IB-M}}^{\theta, \phi})
until convergence
```

Training

我们通过最小化边际和联合IB损失的组合来训练整个网络——特定视角编码器预测器 (θ, φ) 和最终预测器 (ψ):

$$\mathcal{L}_{\text{Total}}^{\theta,\phi,\psi}(\bar{\mathbf{X}},Y) = \mathcal{L}_{\text{IB-J}}^{\theta,\psi}(\bar{\mathbf{X}},Y) + \alpha \sum_{v \in \mathcal{V}} \mathcal{L}_{\text{IB-M}}^{\theta_v,\phi_v}(X_v,Y_v)$$



在整个实验过程中,我们在两个真实世界的多组学数据集上评估了不同的多视角学习方法,这两个数据集分别由癌症基因组图谱(TCGA)和癌细胞系百科全书(CCLE)收集,在整合多组学观察结果的背景下,分别用于预测癌细胞的1年死亡率和药物敏感性。

将 DeepIMV 与 2 个基线和 6 个最先进的多视图学习方法进行比较:

Table 1: Comparison table.

Methods	Task Oriented	$\begin{array}{c} {\bf Incomplete} \\ {\bf Views} \end{array}$		
Base1	✓	X		
Base2	✓	X		
GCCA	×	X		
DCCA	×	X		
DCCAE	×	X		
MVAE	×	✓		
CPM-Nets	✓	✓		
MOFA	×	✓		
DeepIMV	✓	✓		

简单地连接来自多个视图的观察 (表示为 Base1)

后期集成方法,它将在每个视图上训练的单个预测器的预测集成为一个集合(表示为 Base2)

为了将我们的实验集中在综合分析上,并克服高维多组学数据中的"维数灾难",我们在每个视图上使用Kernel-PCA(具有多项式核)提取低维表示(即100个特征)。

TCGA Dataset

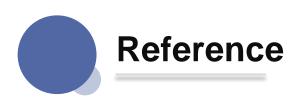
我们根据对 7,295 个癌细胞系(即样本)的多项组学的综合观察分析 1 年死亡率。数据由来自3个不同组学层的每个细胞系的4个不同视图的观察结果组成:(view1) mRNA表达,(view2) DNA甲基化,(view3) microRNA表达,以及(view4) 反相蛋白阵列。在7,295样本中,3,282样本具有具有各种视图缺失模式的不完整多视图观测值:视图1、视图2、视图3和视图4的平均缺失率分别为0.10、0.24、0.13和0.74。

Methods	1 View		2 Views		3 Views		4 Views	
	complete	incomplete	complete	incomplete	complete	incomplete	complete	incomplete
Base1	0.660 ± 0.04	0.675 ± 0.02	0.722 ± 0.03	0.739 ± 0.02	0.750 ± 0.02	0.765 ± 0.02	0.766 ± 0.02	0.781 ± 0.01
Base2	$0.711 {\pm} 0.02$	0.717 ± 0.02	0.746 ± 0.01	0.766 ± 0.00	0.767 ± 0.02	0.775 ± 0.01	$\bf0.783\!\pm\!0.02$	0.790 ± 0.01
GCCA	0.680 ± 0.02	0.650 ± 0.03	0.737 ± 0.02	0.737 ± 0.03	0.764 ± 0.01	0.769 ± 0.02	$0.783 {\pm} 0.01$	0.792 ± 0.01
DCCA	0.702 ± 0.01	0.638 ± 0.03	0.745 ± 0.03	0.761 ± 0.02	0.758 ± 0.02	0.775 ± 0.01	0.776 ± 0.02	0.784 ± 0.01
DCCAE	0.623 ± 0.04	0.605 ± 0.04	0.747 ± 0.03	0.763 ± 0.01	0.774 ± 0.02	0.775 ± 0.01	0.776 ± 0.02	0.778 ± 0.02
MVAE	0.592 ± 0.05	0.589 ± 0.04	0.677 ± 0.02	0.674 ± 0.02	0.731 ± 0.02	0.730 ± 0.01	0.774 ± 0.01	0.781 ± 0.01
CPM-Nets	0.700 ± 0.02	0.709 ± 0.01	0.748 ± 0.02	0.761 ± 0.02	0.766 ± 0.01	0.771 ± 0.01	0.781 ± 0.01	0.788 ± 0.01
MOFA	0.681 ± 0.03	0.646 ± 0.01	0.732 ± 0.01	0.734 ± 0.01	0.756 ± 0.01	$0.764 {\pm} 0.02$	0.781 ± 0.02	0.785 ± 0.02
DeepIMV	0.701 ± 0.02	$0.724 {\pm} 0.02$	$0.757{\pm}0.02$	$\bf0.772 {\pm} 0.01$	$0.776 {\pm} 0.01$	$0.791 {\pm} 0.01$	$0.783 {\pm} 0.01$	$0.801{\pm}0.01$

Methods	Mean Impt.	MVAE Impt.
Base1	$0.765 {\pm} 0.02$	0.771 ± 0.01
Base2	0.775 ± 0.01	$0.784{\pm}0.01$
GCCA	0.769 ± 0.02	0.774 ± 0.01
DCCA	0.775 ± 0.01	$0.784 {\pm} 0.02$
\mathbf{DCCAE}	0.775 ± 0.01	0.773 ± 0.02
\mathbf{MVAE}	0.730 ± 0.01	_
CPM-Nets	0.771 ± 0.01	_
\mathbf{MOFA}	$0.764 {\pm} 0.02$	_
$\mathbf{DeepIMV}$	$0.791 {\pm} 0.01$	_

Poe允许编码器专门分析其相应的视图,并建立不同的专业知识。此外,引入特定于视图的预测器和边际IB损失鼓励特定于视图的编码器关注(可能是互补的)每个视图的任务相关信息,从而简化与Poe分解相关联的训练。我们观察到,这些组件明显有助于DeepIMV的性能提升

Methods	1 View	2 Views	3 Views	4 Views
MoE	0.629 ± 0.03	0.691 ± 0.02	0.736 ± 0.02	$0.768 {\pm} 0.01$
MoE with marginal IBs	0.712 ± 0.01	$0.766 {\pm} 0.01$	$0.786 {\pm} 0.01$	0.790 ± 0.01
PoE	$0.655 {\pm} 0.04$	0.719 ± 0.03	0.755 ± 0.03	0.783 ± 0.02
PoE with marginal IBs	$0.724{\pm}0.02$	$0.772 {\pm} 0.01$	$0.791 {\pm} 0.01$	$0.801 {\pm} 0.01$



RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY | 3



Objective assessment of stored blood quality by deep learning

Edited by Donald Geman, The Johns Hopkins University, Baltimore, MD, and approved July 14, 2020 (received for review January 30, 2020)

August 24, 2020 117 (35) 21381-21390 https://doi.org/10.1073/pnas.2001227117

▼ 779 **₹** 20







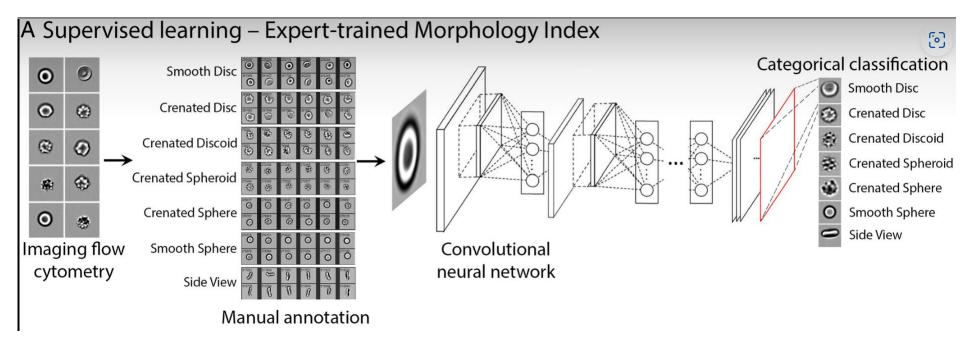


通过深度学习客观评估储存的血液质量

文献来源: August 24, 2020/PNAS/麻省理工学院和哈佛大学博德研究所

开发了一种策略,通过使用**成像流式细胞术和深度学习来评估红细胞的质量,从而避免人类主观性。**

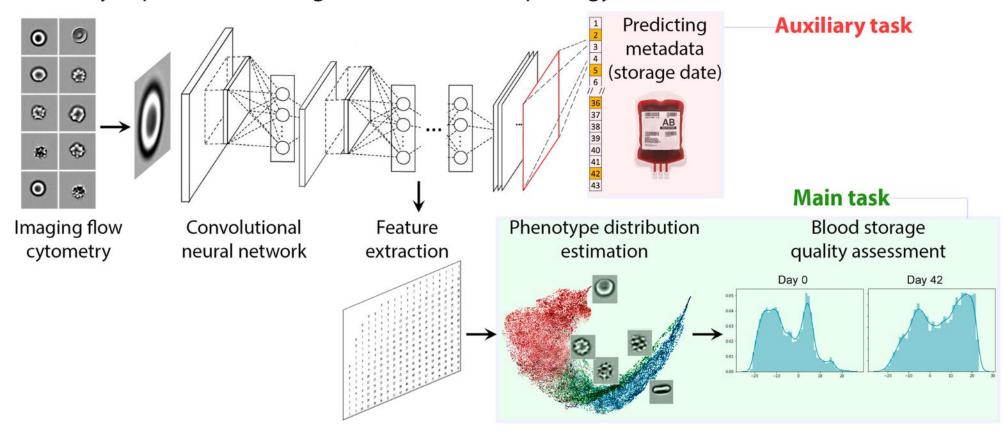
我们通过训练计算机,使用健康和不健康形态的示例图像,成功地实现了传统专家评估的自动化。然而, 我们注意到专家对~18%的细胞存在分歧,因此我们没有依靠专家的视觉评估,而是从随时间采样的红细胞 图像中客观地教授了深度学习网络降解表型。虽然需要使用不同样本进行培训来创建和验证临床级模型, 但这样做可以消除主观评估并促进研究。基于时间的深度学习策略也可能被证明对其他生物学进展有用, 例如发育和疾病进展。

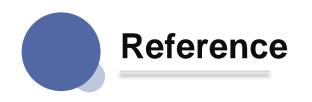


这个完全监督的模型能够近似人类注释者将细胞分类为专家定义的七个形态学类别之一。

研究了一种基于弱监督学习的替代训练策略,其中神经网络独立于专家定义的视觉类别学习红细胞的形态特性。基本策略是训练网络预测一种辅助但具有生物学意义的特性:对每个细胞进行采样的血液单元的储存持续时间(图1B)。

B Weakly supervised learning – Self-learned Morphology Index





Fabrication

| Push |

After adhesive has cured |

Barbed needle |

Thermal-degradable adhesive |

RESEARCH ARTICLE | ENGINEERING | 6

On-demand anchoring of wireless soft miniature robots on soft surfaces

Ren Hao Soon , Ziyu Ren , Wenqi Hu , and Metin Sitti Authors Info & Affiliations

Edited by Ruike Renee Zhao, Stanford University, Stanford, CA; received May 11, 2022; accepted July 25, 2022 by Editorial Board Member John A. Rogers

August 15, 2022 119 (34) e2207767119 https://doi.org/10.1073/pnas.2207767119

无线软微型机器人在软表面上的按需锚固

文献来源: 2022.8 /PNAS/斯坦福大学

提出了一种无线机构,可以通过远程加热精确控制,以 实现按需放针和机械联锁。

通过外部射频 (RF) 基于场的加热精确控制,以实现按需针释放 锚固机构的概念说明与现有软磁机器人集成,以实现在开放和封闭 环境中的锚固。



Explore content Y About the journal Y Publish with us Y

nature > articles > article

Article | Published: 10 August 2022

Warm springs alter timing but not total growth of temperate deciduous trees

Cameron Dow, Albert Y. Kim, Loïc D'Orangeville, Erika B. Gonzalez-Akre, Ryan Helcoski, Valentine

Herrmann, Grant L. Harley, Justin T. Maxwell, Ian R. McGregor, William J. McShea, Sean M. McMahon, Neil

Pederson, Alan J. Tepley & Kristina J. Anderson-Teixeira

<u>Nature</u> **608**, 552–557 (2022) | <u>Cite this article</u> **2654** Accesses | **2** Citations | **112** Altmetric | <u>Metrics</u>

温暖的泉水改变了温带落叶树的节律,但不是总生长

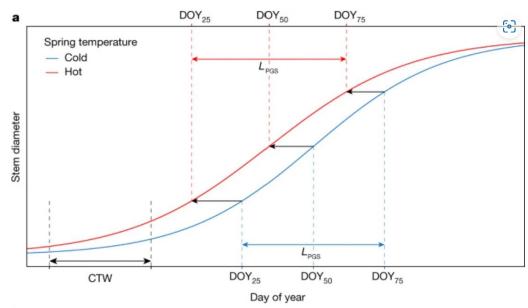
文献来源: 2022.8 /nature/麻省理工学院和哈佛大学博德研究所

地球的森林在吸收人类产生的二氧化碳方面发挥着重要作用,从而为缓解全球变暖做出了重要贡献。本文主要研究由于全球变暖导致生长季节的延长是否会影响木材生长时螯合的碳量。

每年,森林从大气中去除的碳相当于人类活动产生的碳的四分之一左右,**量化森林对全球碳预算的贡献对于评估未来几年气候变化的后果至关重要。目前,全球变暖已经在推进春季叶子出现的时间。**这导致更早地开始 CO2吸收,因此可能促进树干生长,木材生产和碳固存。

本文作者评估北美东部温带落叶林的生长情况,结果表明, 春季树木生长较早并不一定意味着木材产量的年度碳固存显着增加。

使用遥感数据证实,今年早些时候,在两片标志性的落叶林 (弗吉尼亚州弗兰特罗亚尔附近的史密森尼保护生物学研究所森 林和马萨诸塞州彼得舍姆的哈佛森林)发生了春季的叶子出现。 然后,作者探索了树干径向生长(茎周的生长),使用来自近 500棵树的数据,每周评估增量变化,这些数据收集了6年(哈佛 森林)或10年(史密森尼保护生物学研究所森林)。结果显示, 正如预期的那样,在温暖的泉水中,茎秆生长更早开始的相应趋



b		Response to warmer spring T			
		SCBI		Harvard forest	
Variable	Definition	RP	DP	RP	DP
Timing of gro	wth				
DOY ₂₅	Day of year at which 25% of growth is achieved	\downarrow	1	1	1
DOY ₅₀	Day of year at which 50% of growth is achieved	1	1	1	1
DOY ₇₅	Day of year at which 75% of growth is achieved	_	1	1	1
L _{PGS}	Peak growing season length (DOY ₇₅ - DOY ₂₅)	-	-	-	\uparrow
Growth rate					
g_{max}	Maximum growth rate	-	-	\downarrow	1
Annual growt	h				
ΔDBH	Annual growth from dendrometer band	-	-	1	-
RWI	Ring width index from tree-ring chronologies	Mixed	Mixed	-	Mixed



nature

Explore content > About the journal > Publish with us >

nature > articles > article

Article Open Access Published: 20 July 2022

A DNA origami rotary ratchet motor

Anna-Katharina Pumm, Wouter Engelen, Enzo Kopperger, Jonas Isensee, Matthias Vogt, Viktorija Kozina,

Massimo Kube, Maximilian N. Honemann, Eva Bertosin, Martin Langecker, Ramin Golestanian [™], Friedrich

C. Simmel

& Hendrik Dietz

✓

Nature **607**, 492–498 (2022) Cite this article

30k Accesses **2** Citations **583** Altmetric Metrics

一种DNA折纸旋转棘轮电机

文献来源: 2022.7 /nature/慕尼黑工业大学物理系和慕尼黑生物医学工程研究所

电力在各种应用中正日益取代化石燃料,电动机正在迅速从其最广泛使用的竞争对手 - 内燃机和肌肉中夺取市场份额。电力也可能是在微米和纳米尺度上产生机械功的首选动力源,因此需要开发高性能的电纳米电机。

DNA已成为构建纳米级物体的多功能材料,并且在水性环境中从DNA构建更大,更复杂的结构方面 取得了巨大进展。

