arXiv:2304.08359v1 [cs.LG] 17 Apr 2023

热门人工智能基准的能效考量

Raphael Fischer、Matthias Jakobs、Katharina Morik

多特蒙德工业大学，人工智能组，44227 多特蒙德，德国
firstname.lastname@tu-dortmund.de

摘要

人工智能的发展需要更加注重资源利用，并走向可持续发展。这要求清晰地评估和报告能效权衡，例如牺牲较快的运行时间以换取更高的预测性能。尽管已有一些研究能效的方法被提出，但热门方法和数据集的全面结果仍然缺乏。在这项工作中，我们通过提供热门人工智能基准的实证见解，尝试填补这一信息空白，共进行了100项实验。我们的发现表明，不同数据集各自具有独特的能效格局，也显示出不同方法高效运行的可能性各不相同。

引言

尽管近期有人呼吁让人工智能（AI）更具可持续性（Strubell、Ganesh 和 McCallum，2020），许多研究仍然将预测质量置于资源意识之上。因此，该领域的创新大多伴随着巨大的计算开销（Schwartz 等人，2020）。在深度机器学习（ML）中，为了获得精度很高、能耗足迹又相对较小的模型，通常必须进行资源消耗巨大的训练（Tan 和 Le，2019；Howard 等人，2017）。虽然明确报告并比较这类能效权衡至关重要，但迄今为止，相关研究大多只考察了单一应用场景（Patterson 等人，2021；Garc´ıa-Mart´ın 等人，2019；Fischer 等人，2023）。即使对于常用的基准数据集，也很难找到有关能耗和能效的简明信息。

本文通过拓展近期发表的效率评估研究（Fischer 等人，2023），为机器学习可持续性研究作出了贡献。其方法论的部分内容借鉴了欧盟成熟的能源标签体系（欧盟委员会，2019），并以一种同样创新的方式传达此类人工智能信息。让非专业人士理解复杂的模型属性，是迈向可信人工智能的重要一步，尽管这并不容易（Morik 等人，2022；Brundage 等人，2020）。原论文研究了用于 ImageNet 分类的深度计算机视觉模型，而我们成功地将他们的方法论应用于范围更广的基准测试数据和来自 Scikit-learn（Pedregosa 等人，2011）与 OpenML（Vanschoren 等人，2014）的方法。为了支持这些新的应用，我们对现有框架的实现进行了调整和进一步改进；所有代码目前可在 https://github.com/raphischer/sklearn-energy-efficiency 获取。基准测试在科学研究和工业领域常用于概念验证测试，我们希望这项研究能让未来人工智能系统的构建更加注重资源使用。

版权所有 © 2023，人工智能促进协会（www.aaai.org）。保留所有权利。

探究机器学习的能源效率

我们首先说明评估机器学习任务能源效率所需的方法论。这主要基于 Fischer 等人的研究，其中还对相关内容进行了更详细的技术说明。本文将着重介绍关键概念，并说明我们如何将其扩展应用于本研究。

表征机器学习实验

由于人工智能的使用场景差异极大，量化效率并非易事。因此，Fischer 等人将任何 X 需要进行效率测试的机器学习实验界定为由配置和环境组成。前者包含待解决任务的信息（例如推理、训练、鲁棒性测试、...），以及数据集和模型的信息（连同所有超参数）。环境则由执行实验所用的硬件和软件来界定。还可以将碳排放纳入考量，方法是把当地能源结构（假设其保持不变）作为环境的一部分（Patterson 等人，2021年）。

用于评估任务解决效率的相关属性称为指标 （例如准确率、模型大小、功耗）。实验配置决定适用的指标 $m_i \in M_X$。显然，使用不同配置运行实验可以比较其效率：对于固定的任务、数据集和环境，可以考察不同模型的性能。研究特定模型在不同环境中部署时的效率也可能很有意义。请注意，某些配置和环境可能无法实现。例如，数据的选择通常会限制适用的模型，而某些模型也可能需要特定的软件或硬件。

计算模型效率

给定实验 X 的能源效率可通过一个三步流程进行评估。首先，从实验日志文件中汇总每项指标 $m_i \in M_X$ 的数值$\mu_i(X)$。对于某些指标（例如准确率），这很简单；但其他指标（如功耗）则需要依赖特定硬件的专用软件工具（Henderson 等人，2020）。

随后，将指标值与合理且针对具体应用的参考模型 X^*进行比较，并映射到指数刻度 上。使用这种相对刻度，不同指标的数值便可进行比较：

$$\iota_i(X^\circ) = \left(\frac{\mu_i(X)}{\mu_i(X^*)}\right)^{\sigma_i} \tag{1}$$

对于参考模型（即 $X = X^*$），所有指标的指数得分均为一，而其他模型则相对于参考模型进行评估。为了便于直观比较，指数值越高应始终表示性能越好，这通过常数 σ_i 实现。它用于指示第 i 个指标是数值较高（ $\sigma_i = +1$，例如准确率）还是较低（$\sigma_i = -1$，例如功耗）更理想。例如，准确率比参考模型高 20% 的模型，其指数得分为 1。2；而运行时间指数为 0。9，则表示执行速度慢了 10%。请注意，为了获得可靠的指数值，必须针对每种环境分别测量参考性能。

为了让效率信息更容易理解，每个指数刻度被划分为五个评分区间，从而为每个数值分配一个A - E 评级。随后，可以通过（加权）中位数将各项指标评级汇总为综合模型评级。对指标进行加权，可以降低高度相关指标（例如不同的准确率分数）的影响，或针对特定使用场景微调效率评估（例如内存限制严格的情况）。与欧盟能耗标签系统类似，针对具体应用场景的参考模型、权重和评分区间均可由专家确定（欧盟委员会，2019）。通过选择新的参考标准，可以轻松定期更新评级系统，以适应新的技术进步。此外，Fischer 等人表明，混合能源标签能够以更易理解的方式传达效率信息，弥合人工智能专家与知识较少的目标群体之间的差距。

扩展

迄今为止，能源效率仅在基于深度学习的 ImageNet 分类任务中得到研究（Fischer 等人，2023）。因此，所提出的实验特征分析仅限于模型；在此，我们讨论更多样化的机器学习方法，例如 k近邻（kNN）、支持向量机（SVM）、高斯朴素贝叶斯（GNB）、逻辑回归和岭回归（LR、RR）、线性模型的随机梯度下降变体（SGD）、自适应提升集成方法（AB）、（极端）随机森林（RF 和 XRF），以及简单的多层感知机（MLP）（Pedregosa 等，2011）。请注意，术语上的这一变化并不会真正改变所提出的方法论，因为使用这些方法同样会得到特定的模型。

	指标	权重
复杂度	参数	0.4
	模型文件大小	0.3
	浮点运算	0.3
质量	测试 F1 分数	0.5
	测试 Top-1 准确率	0.3
	测试 Top-5 准确率	0.2
资源	功耗	0.6
	运行时间	0.4

表 1：效率指标的类别，以及各指标在确定综合评分时对应的权重。

我们考察了与 Fischer 等人相同的指标，但不包括每轮训练指标，因为这些指标不适用于我们的大多数方法。此外，我们还计算并报告 F_1 分数。由于某些模型不提供单独的类别概率，我们在这些情况下改为报告两次 Top-1 准确率（因为 Top-1 准确率是 Top-5 准确率的紧致下界）。作为对原有工作的扩展，我们提议像表 1 所示，明确对指标进行分类。我们的理由是，每项机器学习实验都具有某种固有的复杂度、质量和资源需求，分别由不同的指标集加以衡量。在分配权重时，我们决定让每个组对最终评分同等重要（即每组内的权重总和为 1）。值得注意的是，功耗获得了最高权重，以优先考虑能源效率。

结果与评估

上述所有方法均部署在来自不同领域和格式的 10 个热门分类数据集上，包括表格数据、图像和文本（完整列表见表2）。首先，我们通过随机搜索寻找特定应用的最优超参数。在 75% 的数据上进行五折交叉验证，为每种方法 × 数据组合测试了 50 组随机赋值。该划分也用于训练最终模型，其余 25% 用作测试数据。我们借助 RAPL 和 psutil 估算功耗，并通过 PAPI 对浮点运算进行分析。执行环境为一台配备 Intel Xeon W-2155、Python 3.8 和 Scikit-learn 1.1.2 的计算机。我们选择了尽可能普通的特定应用参考方法，以便获得相对较差和较好的选项。下文将讨论测试集上的推理效率，并建议通过我们的交互式网页[1]进行更深入的了解。

首先，表2列出了所有基准测试中表现最佳（即效率最高）的方法，并按数据集大小排序。

[1]https://github.com/raphischer/sklearn-energy-efficiency

数据集	方法	功耗		F_1 得分		运行时间		参数	
		[Ws]	索引	[%]	索引	[ms]	索引	[#]	索引
Covertype	B LR	5.734	1.341	72.29	1.032	30.26	1.117	385.0	1.000
MNIST	A RR	4.297	3.891	85.30	0.911	24.73	11.76	7.85e+3	35.37
20 Newsgroups	A LR	5.447	12.21	79.07	1.466	63.23	19.38	2.60e+6	0.000
SpeedDating	B LR	0.371	2.626	84.53	1.002	2.769	1.117	121.0	1.000
LFW People	A MLP	8.978	4.927	23.10	1.753	56.58	16.28	7.78e+5	0.231
音素	一个随机梯度下降	1.046	1.867	75.20	0.871	2.646	1.088	6.000	950.1
LFW 配对	一个随机梯度下降	0.931	2.397	52.90	0.912	7.746	7.903	5.83e+3	23.99
德国信用	一 SGD	0.303	1.218	77.60	1.026	1.625	8.011	21.00	472.5
输血 SC	B 随机梯度下降	2.420	1.082	75.93	1.007	1.014	1.230	5.000	1.000
Olivetti 人脸	B RF	1.930	1.755	93.00	1.081	26.56	0.268	6.81e+4	12.15

表2：每个数据集最高效的机器学习方法，以及各项指标值和指数得分。颜色表示指标或复合模型评级。如果多个模型获得了最佳复合评级，则选择功耗表现最佳的模型。

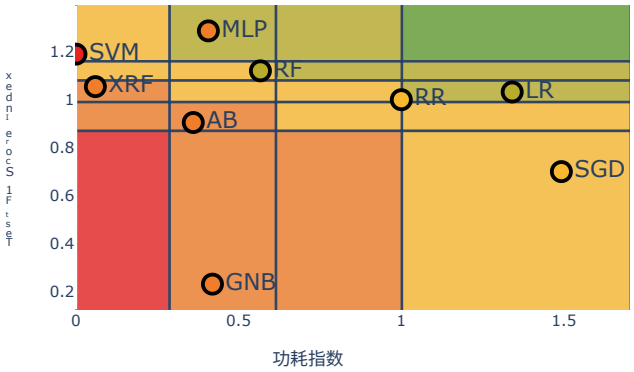

图1：最重要模型指标的散点图（功耗与 F_1 分数）。数值越高，表示性能越好。模型颜色表示复合评级。

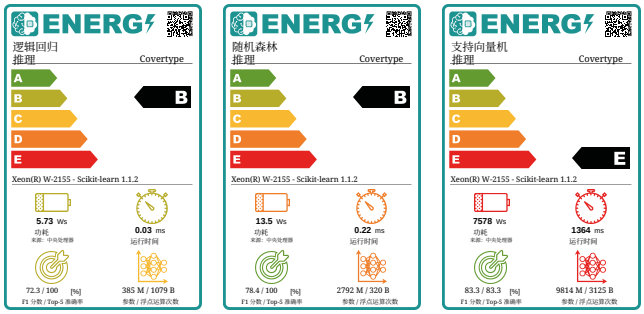

图2：Covertype数据集上效率最高和最低的方法的能耗标签。这些标签通过颜色编码评级，全面展示了不同方法如何权衡各种效率属性。

我们选择复合评级最高的方法；若有多个候选方法，则优先考虑功耗较低者。这里展示了复合评级以及权重最高的指标，同时列出指数评分和相应评级（以颜色标示）。正如预期，没有“面面俱到”的方法，每种方法在效率的某些方面都各有取舍。例如，我们可以看到，在 Covertype 数据集上，逻辑回归较低的功耗和运行时间足以弥补其在其他指标上的平平表现。

让我们更仔细地看看并比较此基准测试中的不同方法。图1展示了一个散点图，呈现模型在功耗与 F_1 分数之间的权衡关系。正如预期，参考方法（RR）位于 (1,1) 坐标处。散点图划分为不同颜色的网格，代表离散评级区间，而点的颜色则表示复合评级。在许多情况下，两者颜色并不一致，这是因为图中仅展示了所考察的八项指标中的两项。另请注意，最大化



图3：最大数据集的方法效率分布。所提出的效率评估方法为所有应用提供了有用的信息。

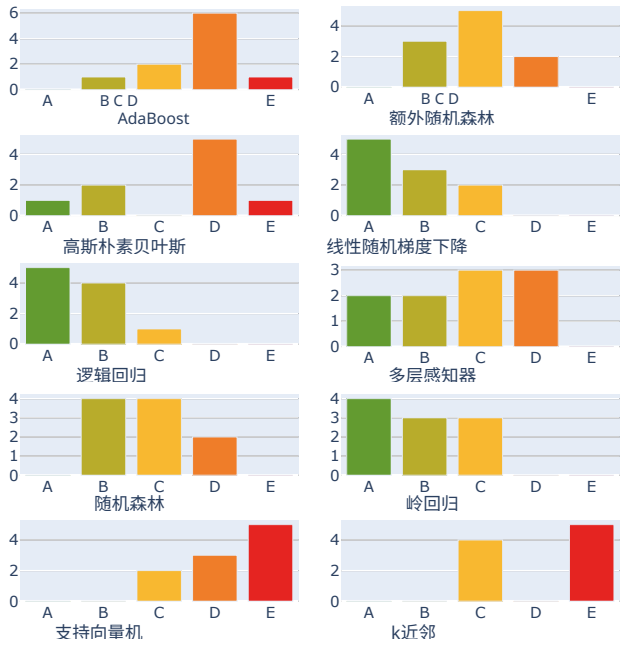

图4：所有十个数据集上各方法的综合评分分布。在我们的实证研究中，线性回归模型似乎效率最高。

功耗指标的索引对应于最小化相应的测量值，这要归功于 σ_i（参见公式 1）。图2展示了所选方法的能耗标签。我们可以看到，尽管随机森林的准确率分数略高，但由于计算更节省资源，逻辑回归获得了相同的复合评级。然而，对于支持向量机而言，其预测质量虽更高，却仍无法抵消巨大的计算开销。

回到更宏观的视角，我们可以考察图3所示的综合评分分布。可以看出，所提出的方法论确实能提供有用的效率概览：每个数据集都有明显表现较好和较差的方法。请注意，这些结果是在所有指标和应用中使用固定评分区间得出的；我们仅为每个数据集选择了不同的参考模型。

我们还可以查看图4所示的各机器学习方法在所有测试数据集上的评分分布。在我们的实证研究中，一些方法（例如线性回归变体）似乎更有可能高效运行。我们认为，这些发现对人工智能系统开发者大有帮助，他们希望为自己的应用场景找到效率最高的模型。

结论

在这项工作中，我们研究了在常见分类基准测试上部署不同机器学习方法的能源效率。为此，我们以近期发表的方法论（Fischer 等人，2023）为基础并加以改进。作为扩展，我们加入了一些指标，并重新设计了加权方式，以评估模型的复合评级。实验展示了每个数据集都有其独特的能效格局，因为每种方法都会在某些性能方面作出权衡。纵观所有基准测试的统计数据，有些方法更有可能获得较高的能效评级。我们希望研究结果能揭示在选择机器学习方法时那些大多尚未得到记录的能效权衡。我们展示了所提方法论的可行性，并希望能帮助开发者在构建未来的人工智能系统时更加关注资源使用。

致谢

本研究由德国联邦教育与研究部和北莱茵-威斯特法伦州资助，属于拉马尔机器学习与人工智能研究所项目的一部分。

参考文献

Brundage, M.; 等. 2020. 迈向可信人工智能开发——：支持可验证声明的机制。欧盟委员会. 2019. 关于家用洗衣机和家用洗烘一体机能源标签的欧盟委员会授权条例 (EU) 2019/2014。Fischer, R.; Jakobs, M.; M¨ucke, S.; 和 Morik, K. 2023. 评估机器学习能源效率的统一框架。载于机器学习与数据库知识发现原理与实践，39–54。瑞士 Cham：Springer Nature Switzerland。Garc´ıa-Mart´ın, E.; Rodrigues, C. F.; Riley, G.; 和 Grahn, H. 2019. 机器学习能耗估算。《并行与分布式计算杂志》，134：75–88。Henderson, P.; 等. 2020. 朝着系统化报告机器学习的能源与碳足迹迈进。Howard, A. G.; 等. 2017. MobileNets：面向移动视觉应用的高效卷积神经网络。Morik, K.; 等. 2022. 是的，我们关心！——通过关怀标签框架认证机器学习方法。《人工智能前沿》，5。Patterson, D. A.; 等. 2021. 碳排放与大型神经网络训练。Pedregosa, F.; 等. 2011. Scikit-learn：Python 中的机器学习。《机器学习研究杂志》，12：2825–2830。Schwartz, R.; Dodge, J.; Smith, N. A.; 和 Etzioni, O. 2020. 绿色人工智能。《美国计算机协会通讯》，63(12)：54–63。Strubell, E.; Ganesh, A.; 和 McCallum, A. 2020. 现代深度学习研究中的能源与政策考量。载于*AAAI*人工智能会议，13693–13696。Tan, M.; 和 Le, Q. 2019. EfficientNet：重新思考卷积神经网络的模型缩放。载于第*36*届国际机器学习会议，6105–6114。Vanschoren, J.; Van Rijn, J. N.; Bischl, B.; 和 Torgo, L. 2014. OpenML：机器学习中的网络化科学。《*SIGKDD*探索通讯》，15(2)：49–60。