



AutoML Survey

2019-10-31

王昕





1 引子

机器学习、深度学习中的 Ensemble 和 Stacking 技术在很多领域中都取得了巨大成功，如：图像识别，推荐系统等。

人工设计机器学习算法、DNN 模型，需要经过大量的调试才能得到高性能的模型，占用了开发者大量的精力和时间。

既然设计机器学习算法架构有大量的重复工作，能否将这个过程自动化？





2 AutoML 重要应用样例

1) Automated Feature Engineering

- Data Science Machine, DSM[11]

J. M. Kanter and K. Veeramachaneni, “Deep feature synthesis: Towards automating data science endeavors,” in *IEEE International Conference on Data Science and Advanced Analytics*, 2015, pp. 1–10.

- ExploreKit[12]

G. Katz, E. C. R. Shin, and D. Song, “Explorekit: Automatic feature generation and selection,” in *International Conference on Data Mining*, 2016, pp. 979–984.

- FeatureHub[27]

M. J. Smith, R. Wedge, and K. Veeramachaneni, “FeatureHub: Towards collaborative data science,” in *IEEE International Conference on Data Science and Advanced Analytics*, 2017, pp. 590–600.

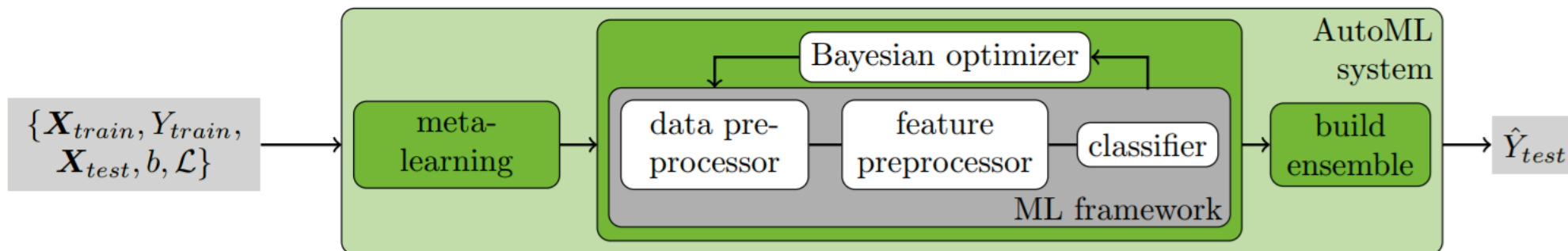




2 AutoML 重要应用样例

2) Auto-sklearn

- 自动学习样本数据: 可以学习样本数据的模样, 自动推荐合适的模型。比如文本数据用什么模型比较好, 比如很多的离散数据用什么模型好。
- 自动调超参
- 自动模型集成: 多个模型组合成一个更强更大的模型。往往能提高预测准确性。



Auto-sklearn WorkFlow





2 AutoML 重要应用样例

2) Auto-sklearn

Auto-sklearn 共有 16 个分类器，基于根据贝叶斯优化找出最佳分类器组合，比如 (0.4 random forest + 0.2 sgd + 0.4 xgboost)。可以认为，Auto-sklearn 是基于 Bayesian Optimizer 的自动化的 Ensemble Learning。

Auto-sklearn 缺点:

- 不支持深度学习，未来可能会推出 AutoNet，类似于谷歌的 Cloud AutoML
- 计算时间较长，往往一个小时以上
- 在数据清洗这块还需要人为参与，目前对非数值型数据不友好





2 AutoML 重要应用样例

3) Neural Architecture Search (NAS)

NIPS 2016 "Learning to learn by gradient descent by gradient descent"

ICLR 2017 "Neural Architecture Search with Reinforcement Learning"

ICML 2017 "Neural Optimizer Search with Reinforcement Learning"

CVPR 2018 "Learning Transferable Architectures for Scalable Image Recognition"
(NASNet)

AAAI 2019 "Regularized Evolution for Image Classifier Architecture Search"





2 AutoML 的目标

- 良好的性能：可以在各种输入数据和学习任务上实现良好的泛化性能；
- 自动化：可以自动完成机器学习工具的配置；
- 高效计算：AutoML 程序可以在有限的预算内返回合理的输出。





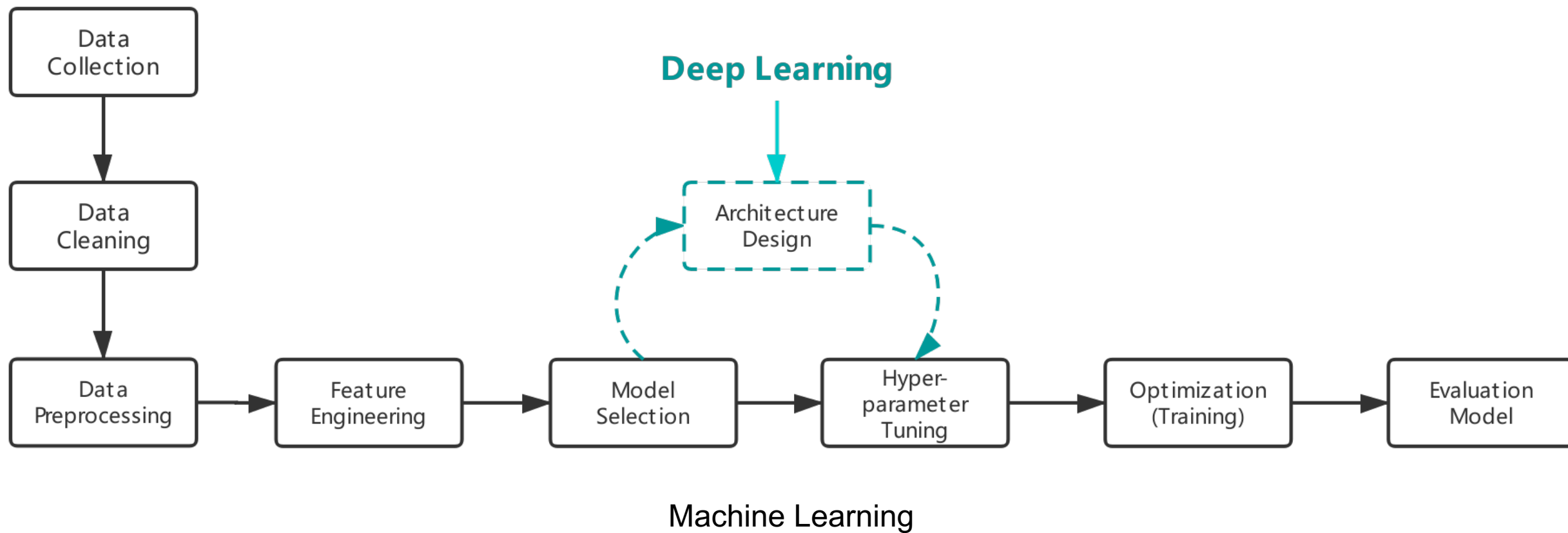
2 AutoML 的目标

Question: AutoML 能否摆脱 No Free Lunch 定理?



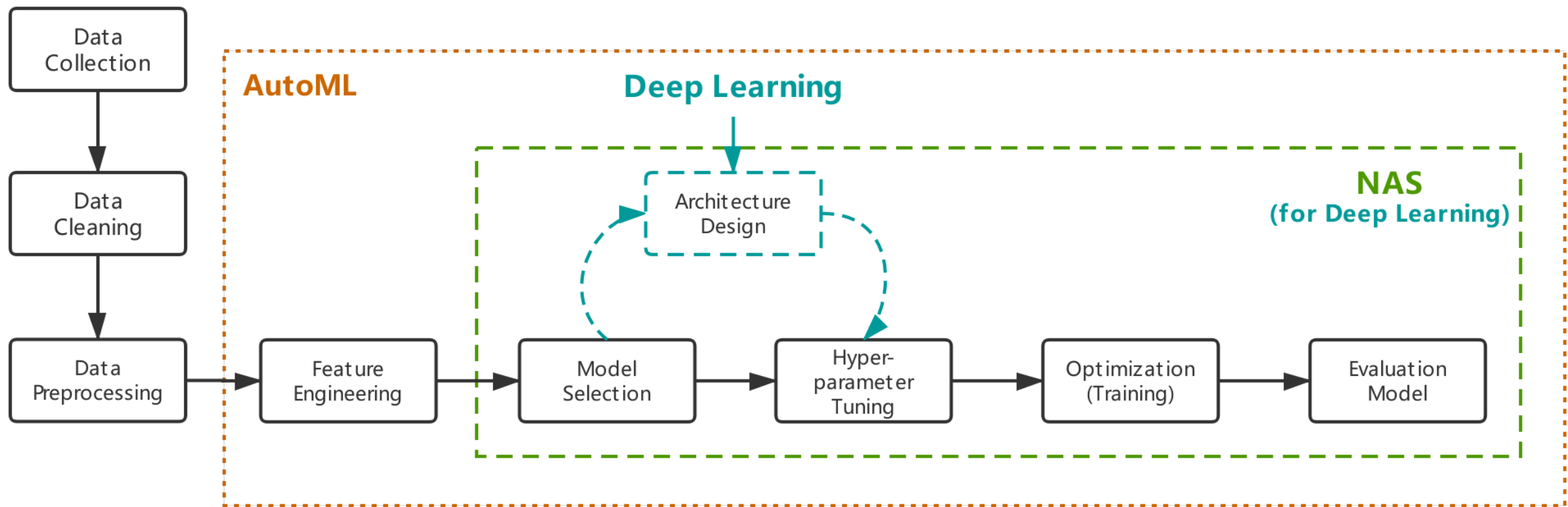
3 对 AutoML 的整体认识与重要概念

1) Workflow - Machine Learning




3 对 AutoML 的整体认识与重要概念

1) Workflow - AutoML



Auto Machine Learning





3 对 AutoML 的整体认识与重要概念

1) Workflow - 基于传统机器学习的 Auto ML

Step 1 Feature Engineering

- Feature Enhancing Methods
- Search Space

Step 2 Model Selection

- Classification Tools
- Search Space

Step 3 Optimization Algorithm Selection

- Optimization Algorithms
- Search Space





3 对 AutoML 的整体认识与重要概念

2) Neural Architecture Search

深度学习可以自动学习出有用的特征，**脱离了对特征工程的依赖**，在图像、语音等任务上取得了超越其他算法的结果。但是，**设计出高性能的神经网络需要大量的专业知识与反复试验，成本极高**。

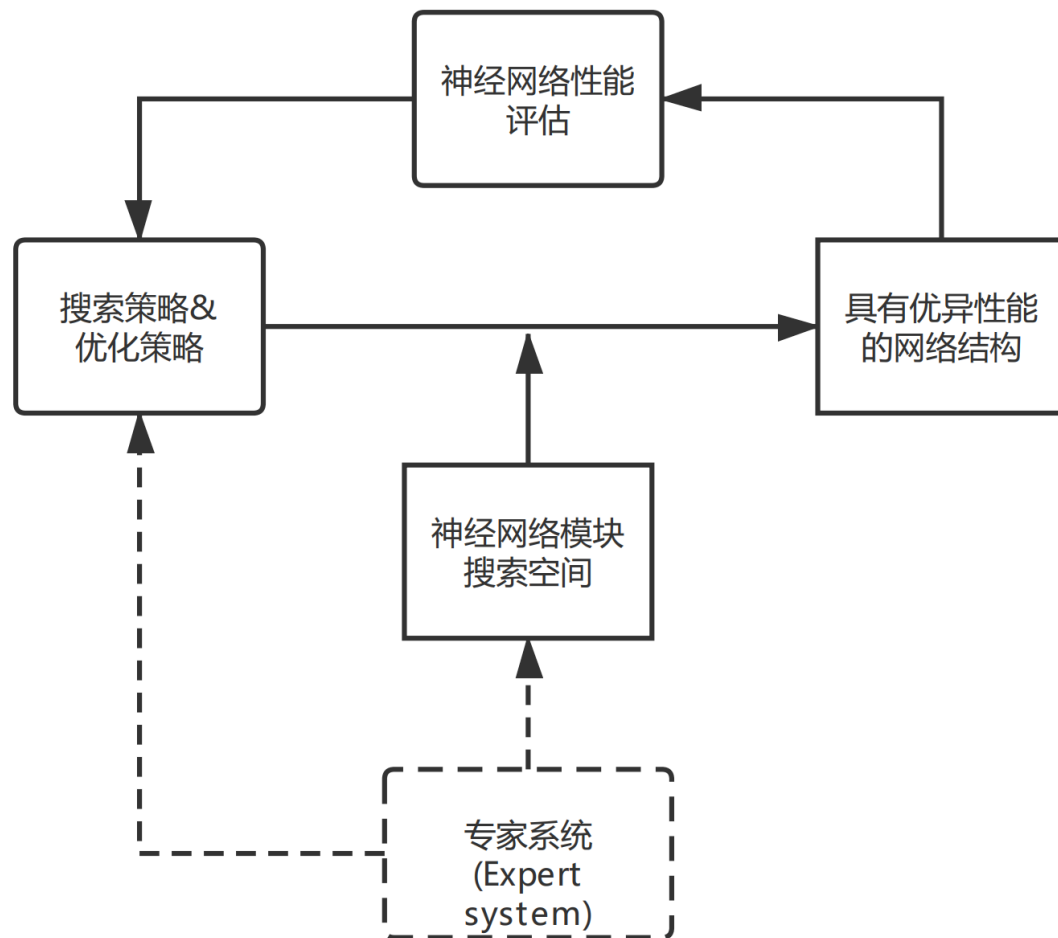
神经结构搜索（Neural Architecture Search, NAS）是一种自动设计神经网络的技术，可以通过算法根据样本集自动设计出高性能的网络结构，在某些任务上甚至可以媲美人类专家的水准，甚至发现某些人类之前未曾提出的网络结构，这可以有效的降低神经网络的使用和实现成本。

NAS的原理是给定一个称为搜索空间的候选神经网络结构集合，用某种策略从中搜索出最优网络结构。神经网络结构的优劣即性能用某些指标如精度、速度来度量，称为性能评估。



3 对 AutoML 的整体认识与重要概念

2) Neural Architecture Search - Workflow





3 对 AutoML 的整体认识与重要概念

2) Neural Architecture Search

- 搜索空间 (Search Space) :

搜索空间原则上定义了可以代表哪些体系结构。结合适用于任务属性的先验知识可以减小搜索空间大小并简化搜索。然而,这也引入了人为修正,但这样做也可能会影响我们发现超越当前人类经验的全新架构。

- 搜索策略 (Search strategy) :

搜索策略说明了如何做空间搜索。它包含了经典的探索-开发 (exploration-exploitation) 之间的权衡。一方面,需要快速找到性能良好的架构;另一方面,避免过早收敛到次优架构 (suboptimal architecture) 区域。

- 性能评估策略 (Performance estimation strategy) :

NAS 的目标通常是找到在未知数据实现高预测性能的架构。性能评估是指评估此性能的过程:最简单的选择是对数据架构执行标准训练和验证,但遗憾的是,这种方法计算成本很高,限制了可以探索的体系结构量。因此,最近的研究大多集中在开发出方法去降低这些性能估计成本。





3 对 AutoML 的整体认识与重要概念

3) 总结

AutoML 是一个非常复杂的问题，也是一个非常活跃的研究领域。

同时，在 AutoML 中有许多上述历史中没有发现的新机会和问题，尤其是在可解释性有待突破。

许多专注于 AutoML 的论文出现在各种会议和期刊上，例如 ICML, NIPS, KDD, AAAI, IJCAI 和 JMLR。

在 AI&Climate 领域，AutoML 能做什么？

