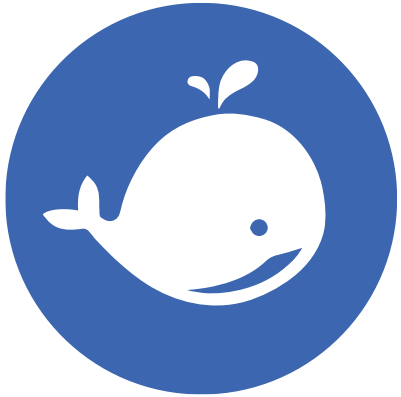


Datawhale AI夏令营 第三期

一行代码上大分

AutoGluon：小白上分神器

分享嘉宾：骆秀韬

Datawhale

/01 一行代码？

完整训练代码

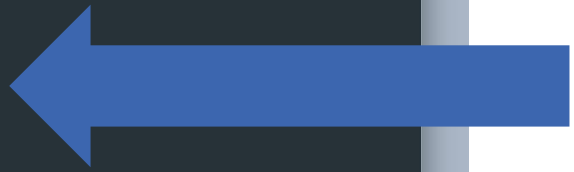
Datawhale

```python
from autogluon.tabular import TabularDataset, TabularPredictor
import pandas as pd

train_data = TabularDataset("./data/train.csv")
test_data = TabularDataset("./data/test.csv")
submit = pd.DataFrame()
submit["uuid"] = test_data["uuid"]
label = "target"

predictor = TabularPredictor(
    label=label,
    problem_type="binary",
    eval_metric="f1",
    ).fit(
    train_data.drop(columns=["uuid"]),
    )

submit[f"{label}"] = predictor.predict(test_data.drop(columns=["uuid"]))
submit.to_csv("submit.csv", index=False)
```



您当前最佳成绩为0.84311提交时间为
（注：以当前所在团队提交的最佳成绩

具体含义：

- 指定预测标签
- 设定问题类型
- 指定评估指标
- 盘它！fit()

AutoGluon 是何方神圣?

AutoML for Image, Text, Time Series, and Tabular Data

官方网站：https://auto.gluon.ai/stable/index.html

Github地址：https://github.com/autogluon/autogluon

沐神倾情推荐

Datawhale

10行代码战胜90%数据科学家?

20.8万 56 2021-05-09 13:30:16 未经作者授权，禁止转载

https://www.bilibili.com/video/BV1F84y1F7Ps

AutoGluon背后的技术

6.6万 47 2021-05-17 07:53:42 未经作者授权，禁止转载

https://www.bilibili.com/video/BV1rh411m7Hb

/02

AutoML
技术杂谈

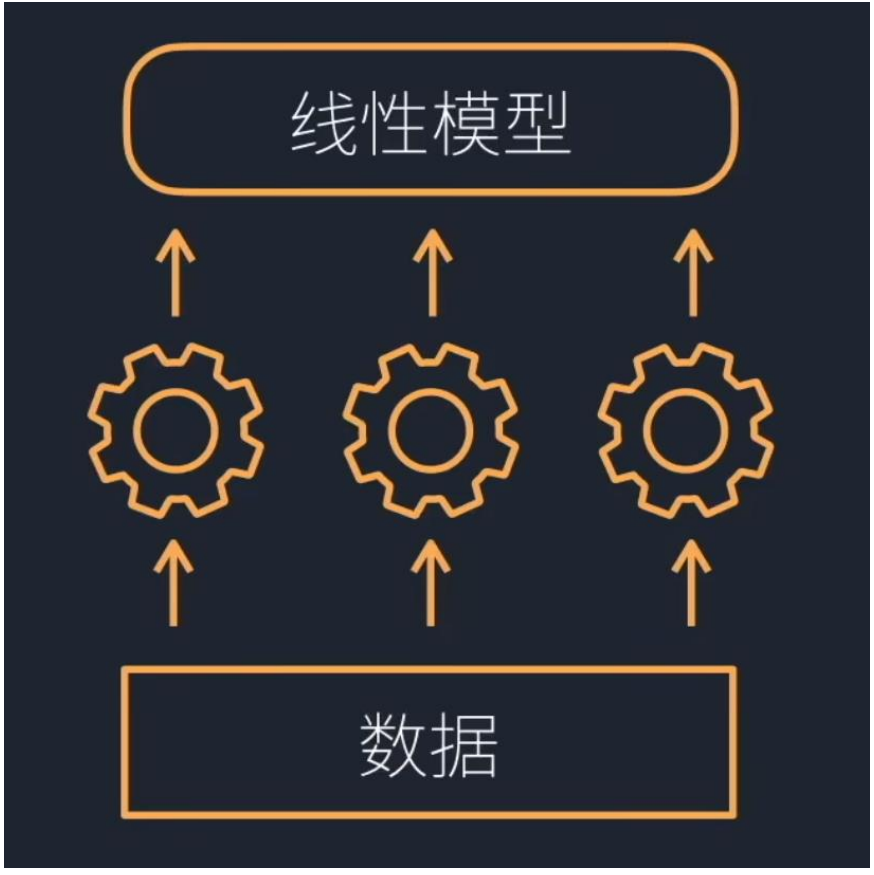
十个臭皮匠能不能打过诸葛亮?

Datawhale

- Stacking
- K-折交叉 Bagging
- 多层 Stacking

Stacking

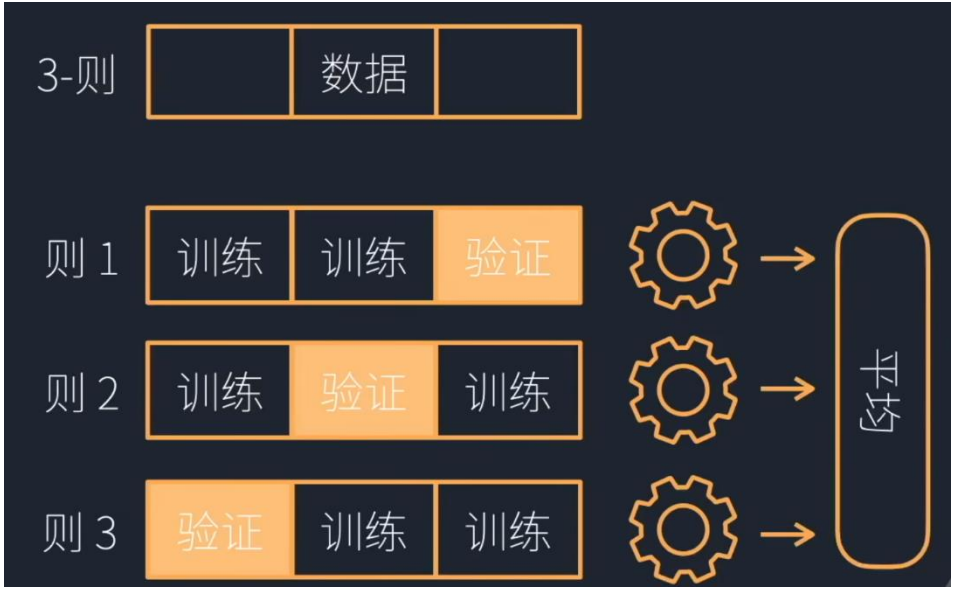
Datawhale



- 在同一个数据集上训练多个不一样的模型
- 这些模型的输出然后进入到一个线性的模型
- 得到最终的输出

K-折交叉 Bagging

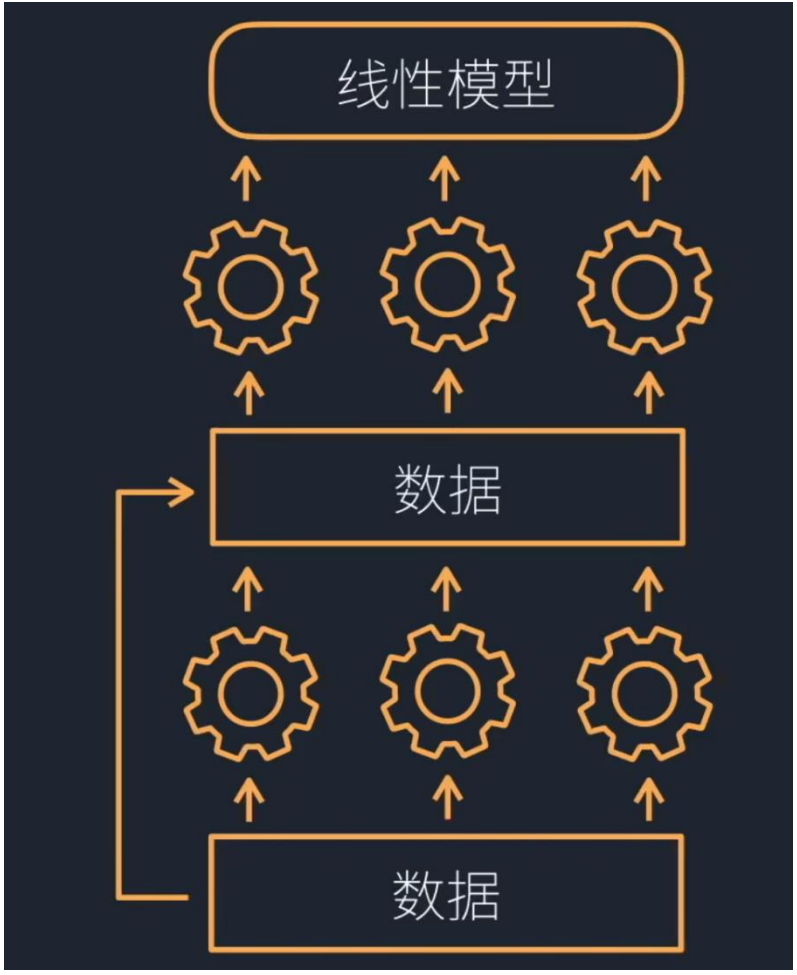
Datawhale



- 不同的初始权重与数据训练多个同类模型
- K-折交叉验证的相同做法
- 整合多个同类模型的结果
  - 回归：取平均值
  - 分类：投票，少数服从多数

多层 Stacking

Datawhale



- 将多个模型的输出和数据合并起来再做一次 Stacking
- 防止过拟合，多层 Stacking 往往配合 K-则交叉 Bagging 使用
- 计算开销剧增！不环保 :(

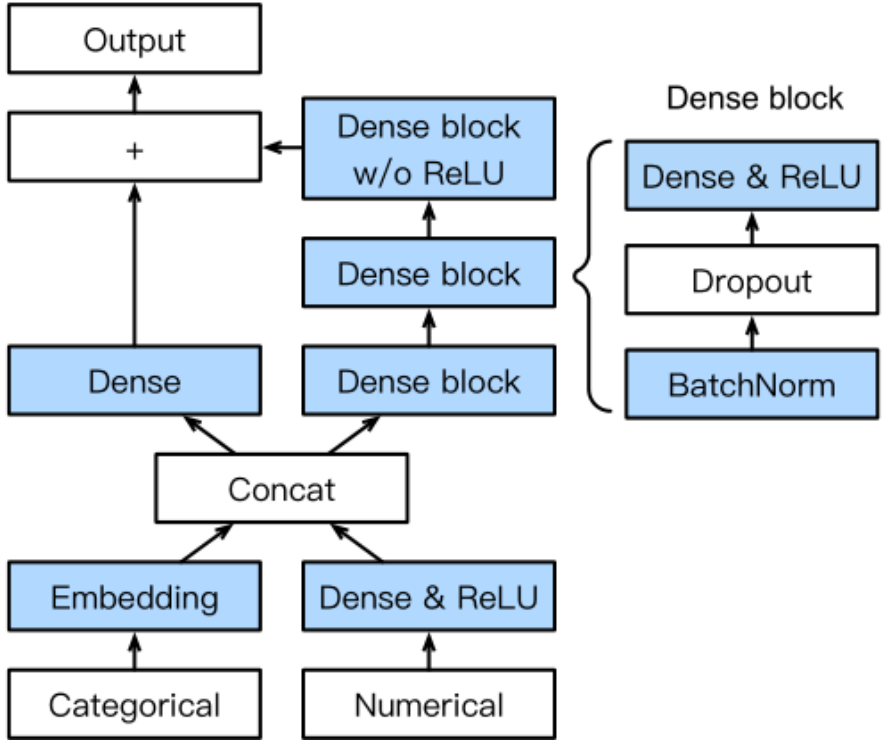
训练策略和神经网络架构

Datawhale

Algorithm 1 AutoGluon-Tabular Training Strategy (multi-layer stack ensembling + n-repeated k-fold bagging).

Require: data (X, Y), family of models $\mathcal{M}$, # of layers L

Preprocess data to extract features
for $l = 1$ **to** L **do** {Stacking}
  for $i = 1$ **to** n **do** {n-repeated}
    Randomly split data into k chunks $\{X^j, Y^j\}_{j=1}^k$
    for $j = 1$ **to** k **do** {k-fold bagging}
      for each model type m in $\mathcal{M}$ **do**
        Train a type-m model on X^{-j}, Y^{-j}
        Make predictions $\hat{Y}_{m,i}^j$ on OOF data X^j
      end for
    end for
  end for
  Average OOF predictions $\hat{Y}_m = \{\frac{1}{n}\sum_i \hat{Y}_{m,i}^j\}_{j=1}^k$
  $X \leftarrow \text{concatenate}(X, \{\hat{Y}_m\}_{m \in \mathcal{M}})$
end for



AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data

- 机器学习的大量工作仍然需要人工干预：特征提取、模型选择、参数调节
- 希望自动化进行机器学习中的特征、模型、优化、评价，甚至部署等环节

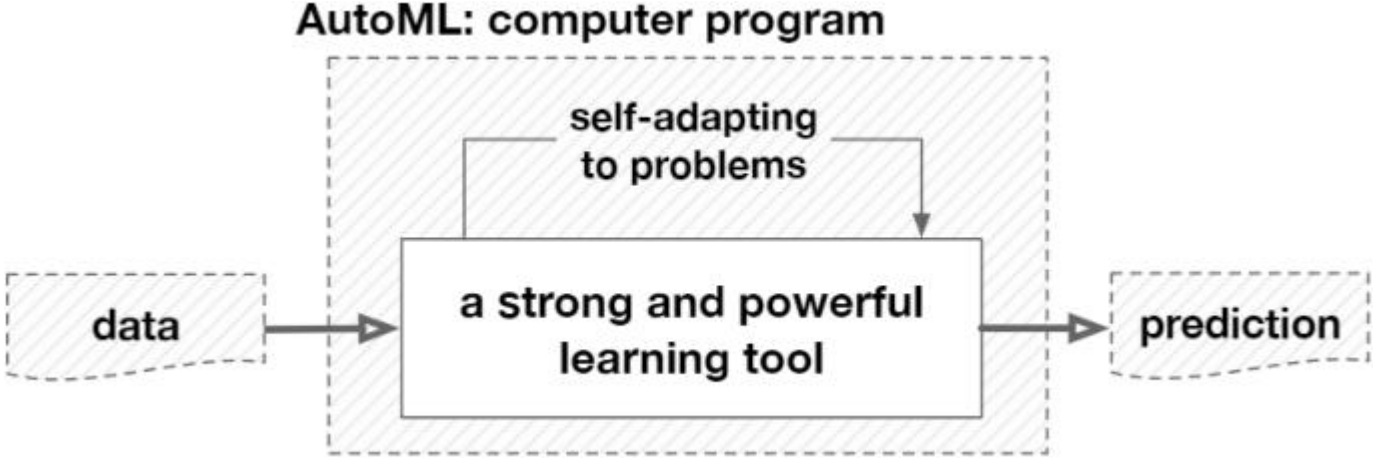
Taking Human out of Learning Applications: A Survey on Automated Machine Learning

AutoML问题的定义

Datawhale

$$\max_{\text{configrations}} \text{performance of learning tools,}$$

$$\text{s.t.} \begin{cases} \text{limited (or no) human assistance} \\ \text{limited computational budget} \end{cases}$$

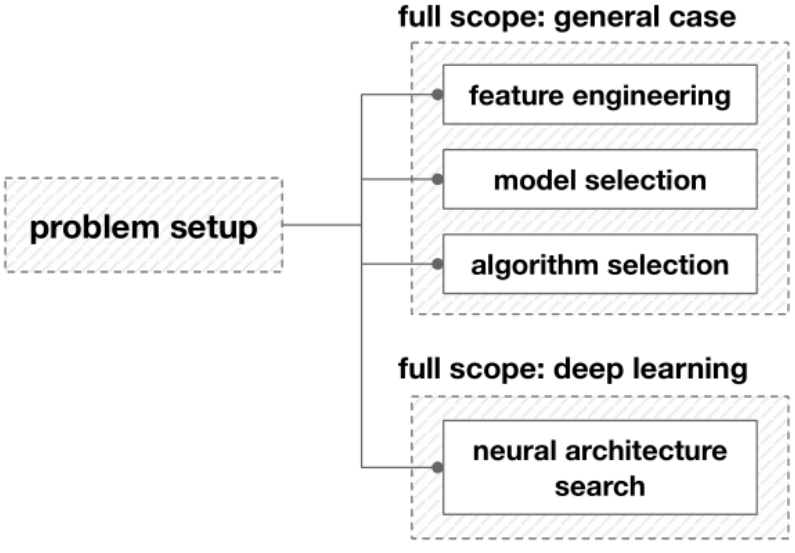


三个核心：

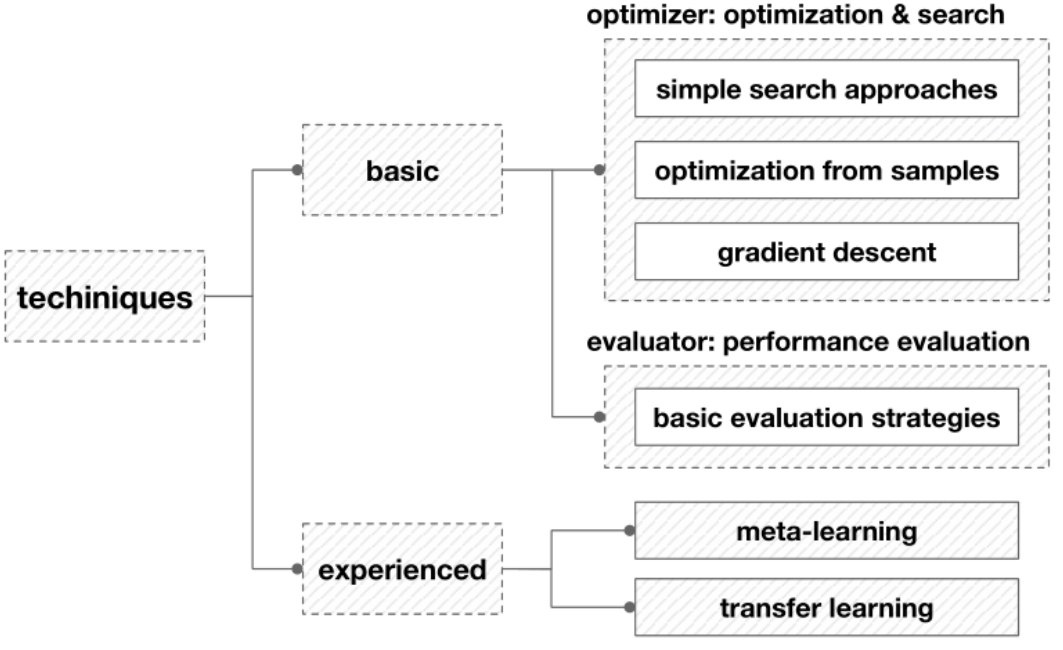
- 更好的训练效果
- 更少的人工参与
- 更低的计算资源

AutoML问题的构成

Datawhale



(a) "What to automate": by problem setup.

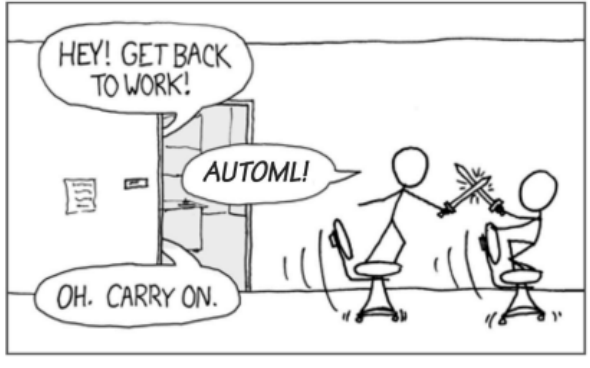


(b) "How to automate": by techniques.

Automated Machine Learning (AutoML)

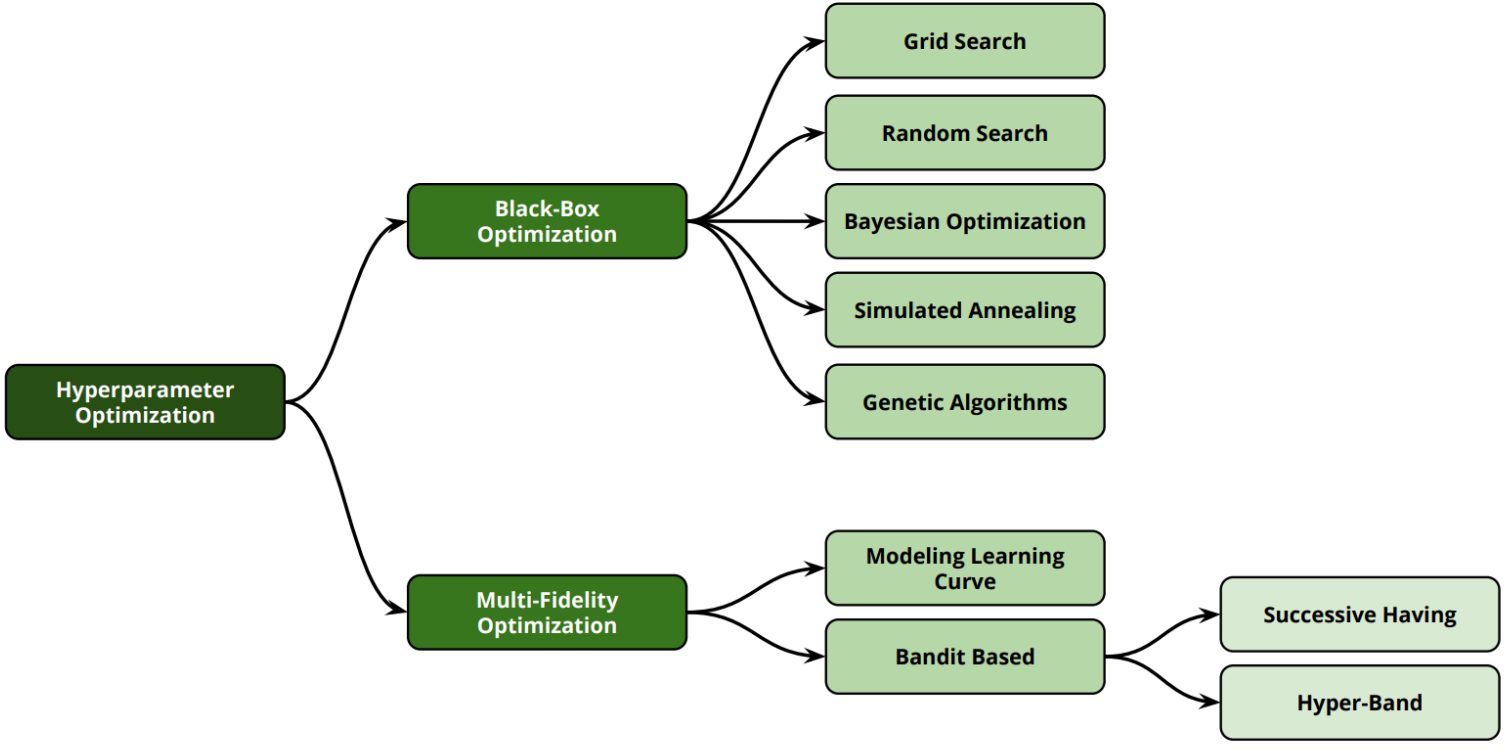
- Automate every step in applying ML to solve real-world problems: data cleaning, feature extraction, model selection…
- Hyperparameter optimization (HPO): find a good set of hyperparameters through search algorithms
- Neural architecture search (NAS): construct a good neural network model

THE DATA SCIENTIST'S #1 EXCUSE FOR LEGITIMATELY SLACKING OFF: "THE AUTOML TOOL IS OPTIMIZING MY MODELS!"



超参数优化方法

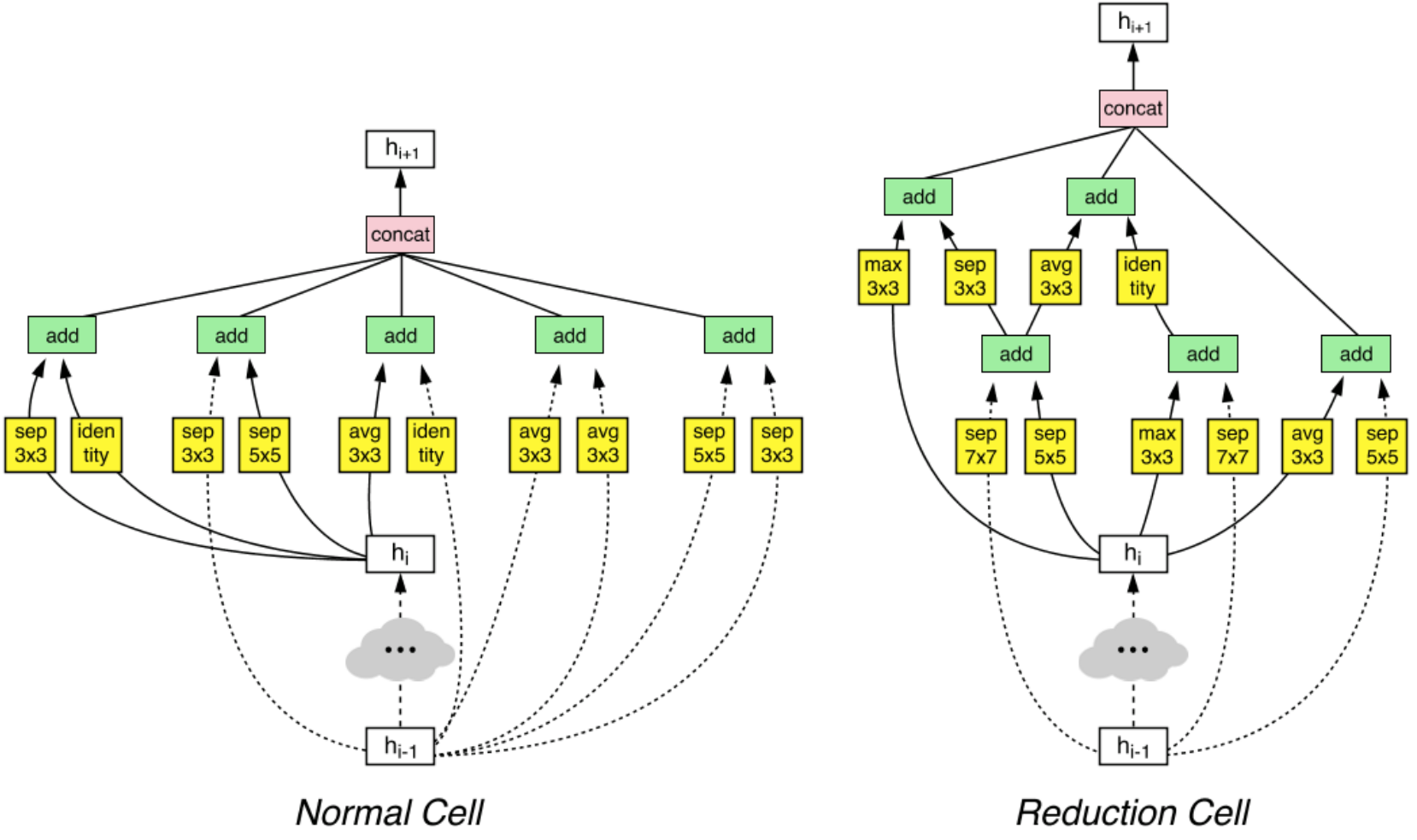
Datawhale



Automated Machine Learning: State-of-The-Art and Open Challenges

神经架构搜索

Datawhale



Learning Transferable Architectures for Scalable Image Recognition

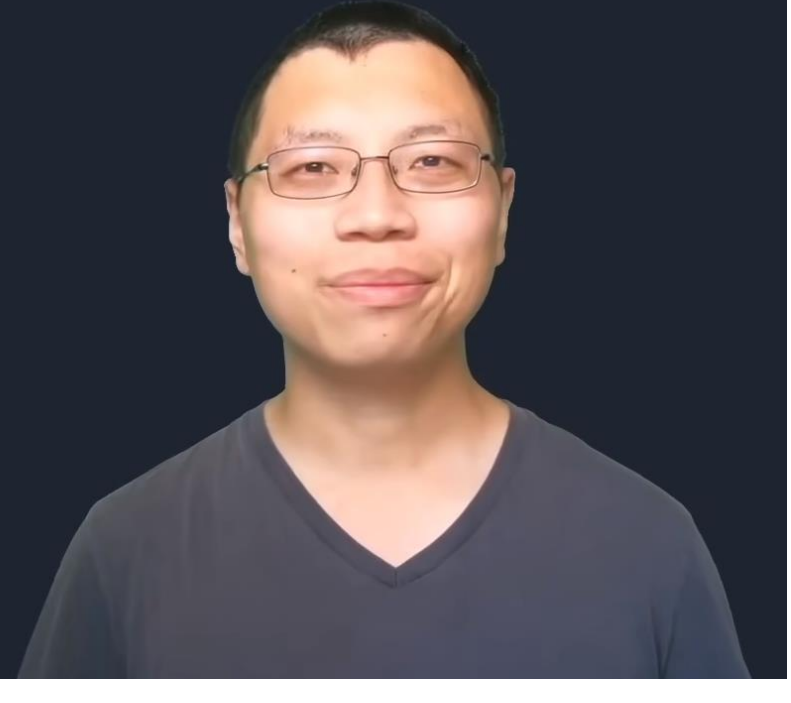
结果对比

跟李沐学AI bilibili

50个数据集

限制4小时运行时间

	冠军
AutoGluon	30
TPOT	5
GCP	7
Auto-sklearn	4
H2O	2
Auto-WEKA	1



for the learner，
和学习者一起成长

Thank you

很高兴跟大家一起在夏令营中度过这个愉快的暑假，我们一起继续向人工智能训练大师之路进发！