

Kaggle大神们都在用什么语言、框架、模型？这里有一份详细统计

机器之心 2022-03-28 12:47

选自medium

作者：Eniola Olaleye

机器之心编译

编辑：张倩




对于ML学习者和从业者来说，参加竞赛是一个很好的锻炼机会，还能赚取一些零花钱。那么，你知道哪个平台比赛最多，成绩比较好的那些团队都在使用什么架构、什么模型吗？在这篇文章中，一位名叫Eniola Olaleye的数据科学爱好者介绍了他们的统计结果。

↑

367

↓

Posted by u/hcarlens 3 days ago



[News] Analysis of 83 ML competitions in 2021

News

I run mlcontests.com, and we aggregate ML competitions across Kaggle and other platforms.

We've just finished our analysis of 83 competitions in 2021, and what winners did.

Some highlights:

- Kaggle still dominant with a third of all competitions and half of \$2.7m total prize money
- 67 of the competitions took place on the top 5 platforms (Kaggle, Alcrowd, Tianchi, DrivenData, and Zindi), but there were 8 competitions which took place on platforms which only ran one competition last year.
- Almost all winners used Python - 1 used C++!
- 77% of Deep Learning solutions used PyTorch (up from 72% last year)
- All winning computer vision solutions we found used CNNs
- All winning NLP solutions we found used Transformers

统计网站：<https://mlcontests.com/>

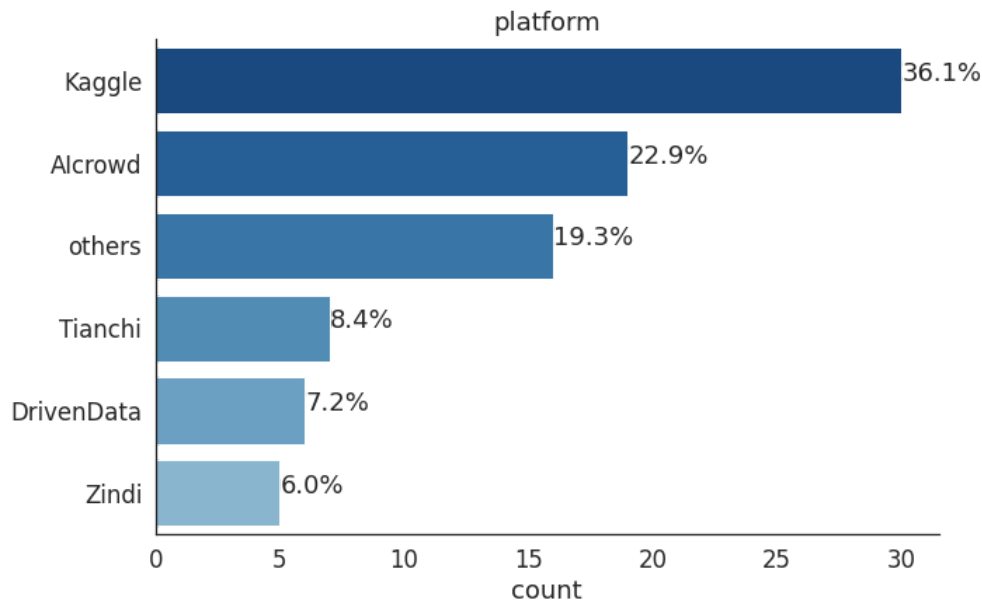
作者得出了几个重要结论：

- 1、在所有竞赛中，Kaggle上的竞赛数量仍然占据1/3，而且奖金数量占270万美元总奖金池的一半；
- 2、在所有比赛中，有67场比赛是在前5大平台（Kaggle、Alcrowd、Tianchi、DrivenData 和 Zindi）上举行的，有8场比赛是在去年只举办了一场比赛的平台上举行的；
- 3、几乎所有的冠军都使用了Python，只有一个冠军使用了C++；
- 4、77%的深度学习解决方案使用了PyTorch（去年高达72%）；
- 5、所有获奖的CV解决方案都使用了CNN；
- 6、所有获奖的NLP解决方案都使用了Transformer。

以下是这次调查的详细信息：

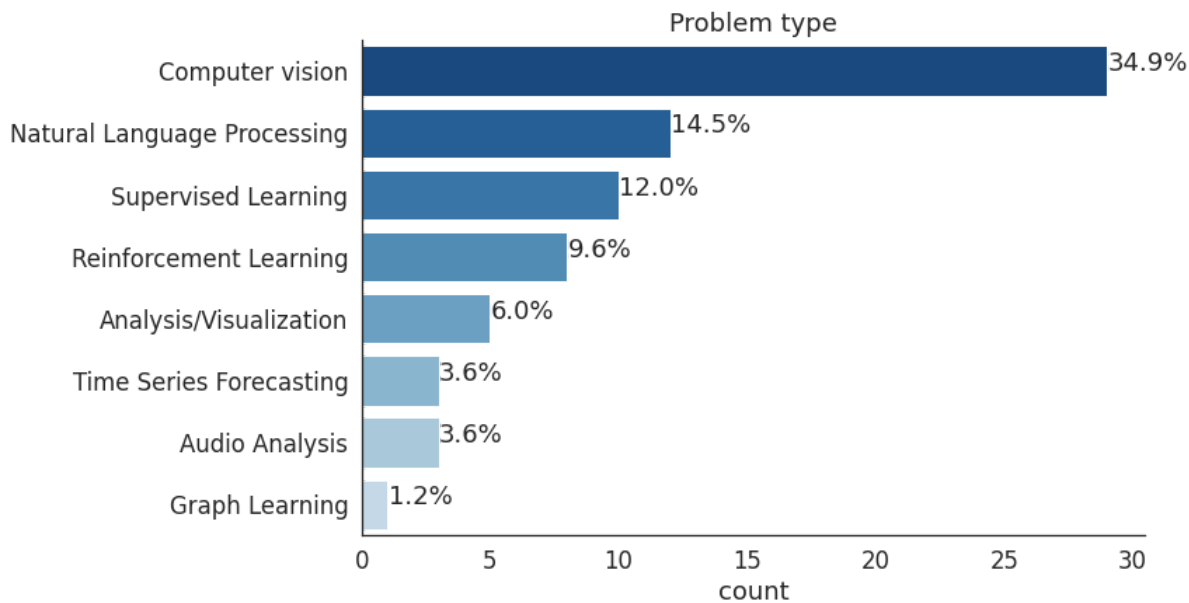
平台类型

在本次调查中，作者总共统计了16个平台上的83场竞赛。这些竞赛的总奖金池超过270万美元，其中奖金最丰厚的比赛是由Driven data举办的Facebook AI Image Similarity Challenge: Matching Track，奖金高达20万美元。

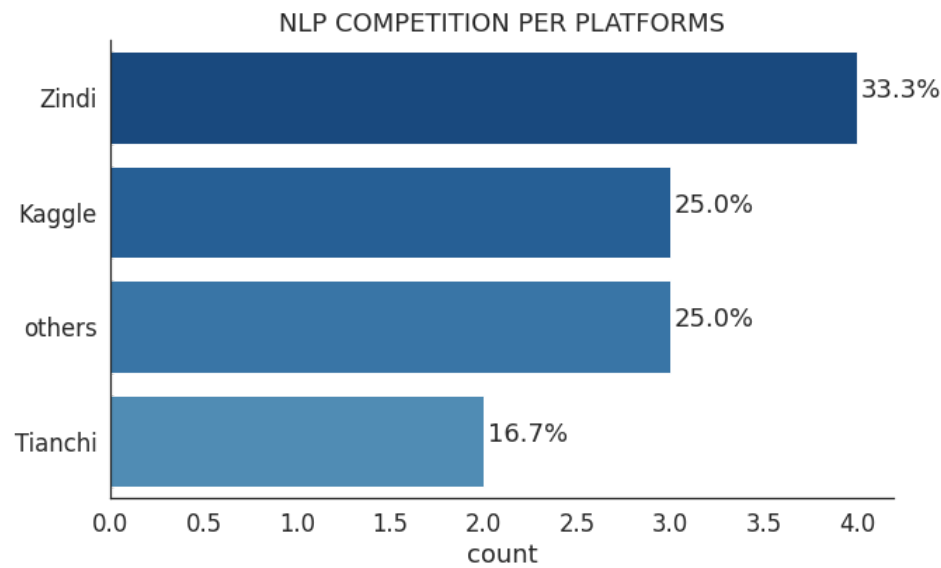


竞赛类型

此次调查显示，2021年最常见的竞赛类型是计算机视觉和自然语言处理。与2020年相比，这部分变化很大，当时NLP竞赛仅占竞赛总数的7.5%。

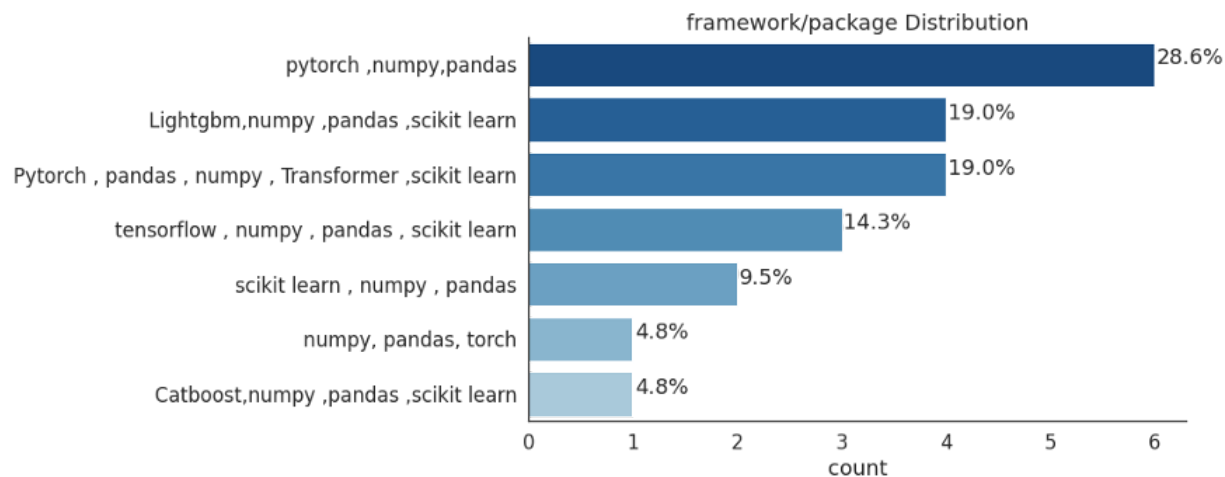
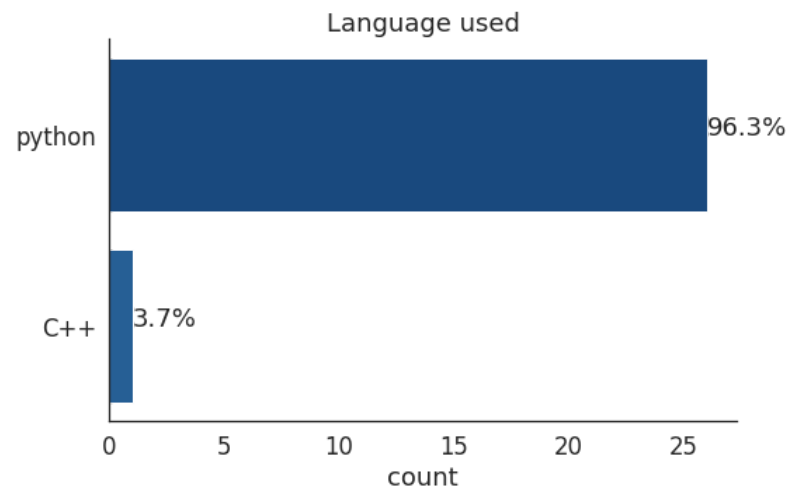


在众多NLP竞赛中，Zindi与AI4D（Artificial Intelligence for Development Africa）合作举办的竞赛数量最多，比赛内容包括将一种非洲语言翻译成英语或其他语言以及针对一种非洲语言进行情感分析。

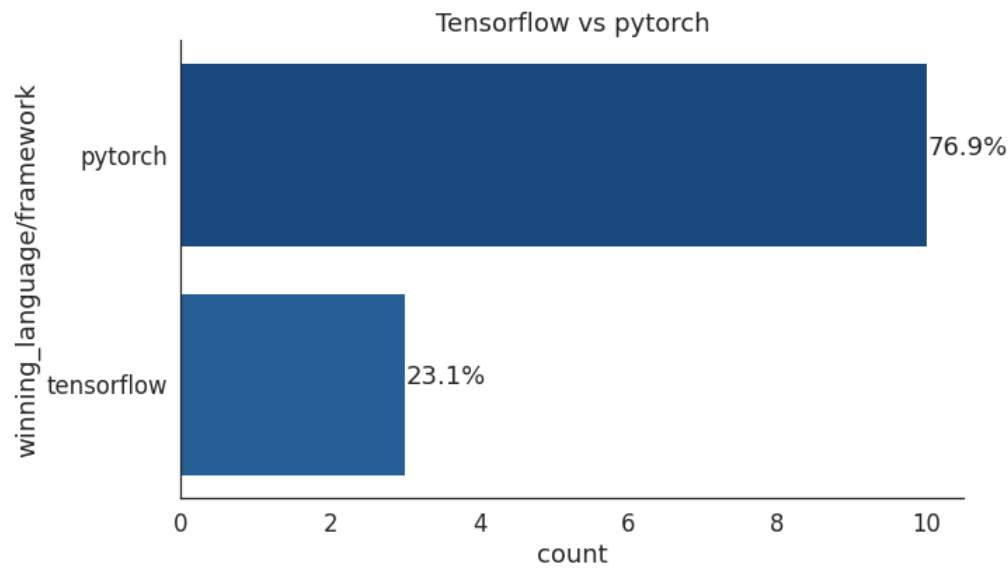


语言与框架

在这次调查中，主流的机器学习框架依然是基于Python的。Scikit-learn非常通用，几乎被用于每个领域。

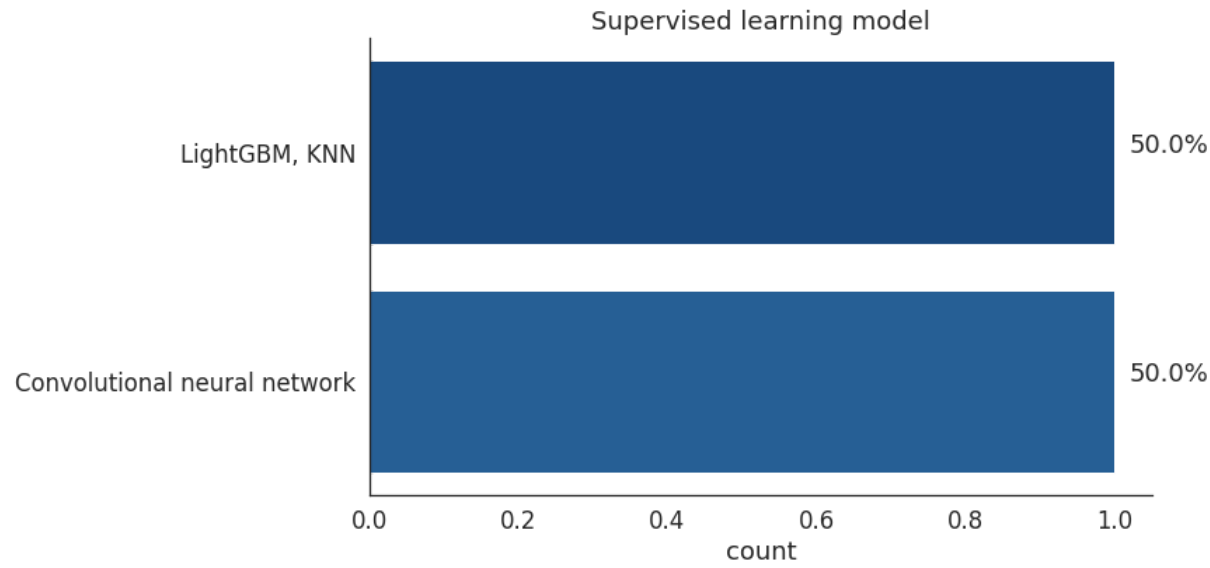


不出所料，两个最流行的机器学习库是Tensorflow和Pytorch。其中，Pytorch在深度学习比赛中最受欢迎。与2020年相比，在深度学习竞赛中使用PyTorch的人数突飞猛进，PyTorch框架每年都在快速发展。



冠军模型

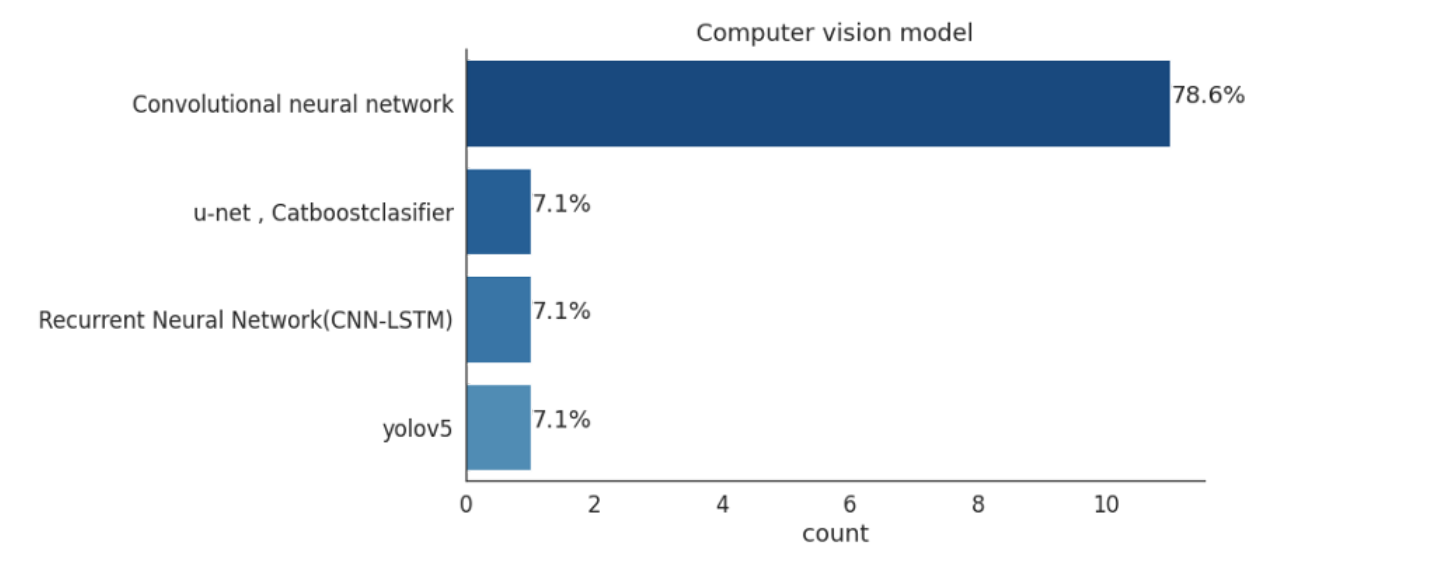
监督学习



在经典机器学习问题中，Catboost、LightGBM等梯度提升模型占据主流。

举个例子，在一个室内定位和导航的Kaggle竞赛中，选手需要设计算法，基于实时传感器数据预测智能手机在室内的位置。冠军解决方案考虑了三种建模方法：神经网络、LightGBM和K-Nearest Neighbors。但在最后的pipeline中，他们只用LightGBM和K-Nearest Neighbours达到了最高分。

计算机视觉



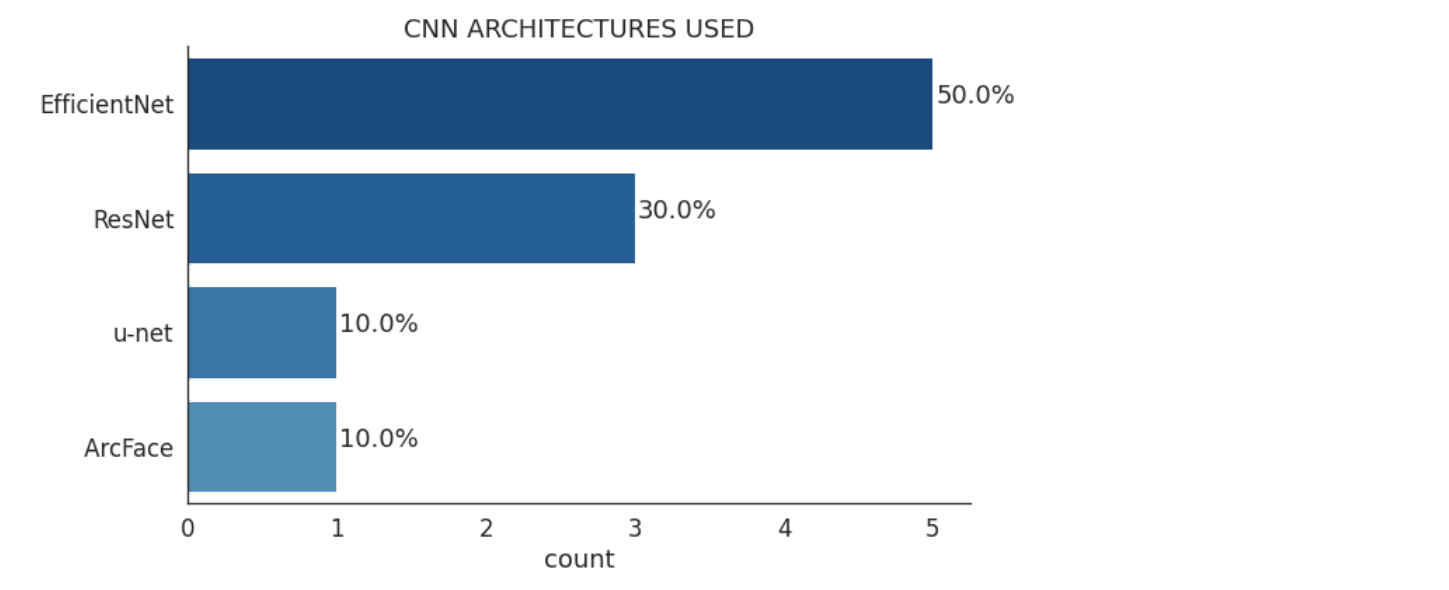
自从AlexNet在2012年赢得ImageNet竞赛以来，CNN算法已经成为很多深度学习问题都在用的算法，特别是在计算机视觉方面。

循环神经网络和卷积神经网络并不相互排斥。尽管它们似乎被用来解决不同的问题，但重要的是这两个架构都可以处理某些类型的数据。例如，RNN使用序列作为输入。值得注意的是，序列并不局限于文本或音乐。视频是图像的集合，也可以用作序列。

循环神经网络，如LSTM，被用于数据具有时间特征的情况（如时间序列），以及数据上下文敏感的情况（如句子补全），其中反馈循环的记忆功能是实现理想性能的关键。RNN还在计算机视觉的下列领域中得到了成功的应用：

- 「日间图片」与「夜间图片」是图像分类的一个例子（一对一RNN）；
- 图像描述（一对多RNN）是根据图像的内容为图像分配标题的过程，例如「狮子猎鹿」；
- 手写体识别；

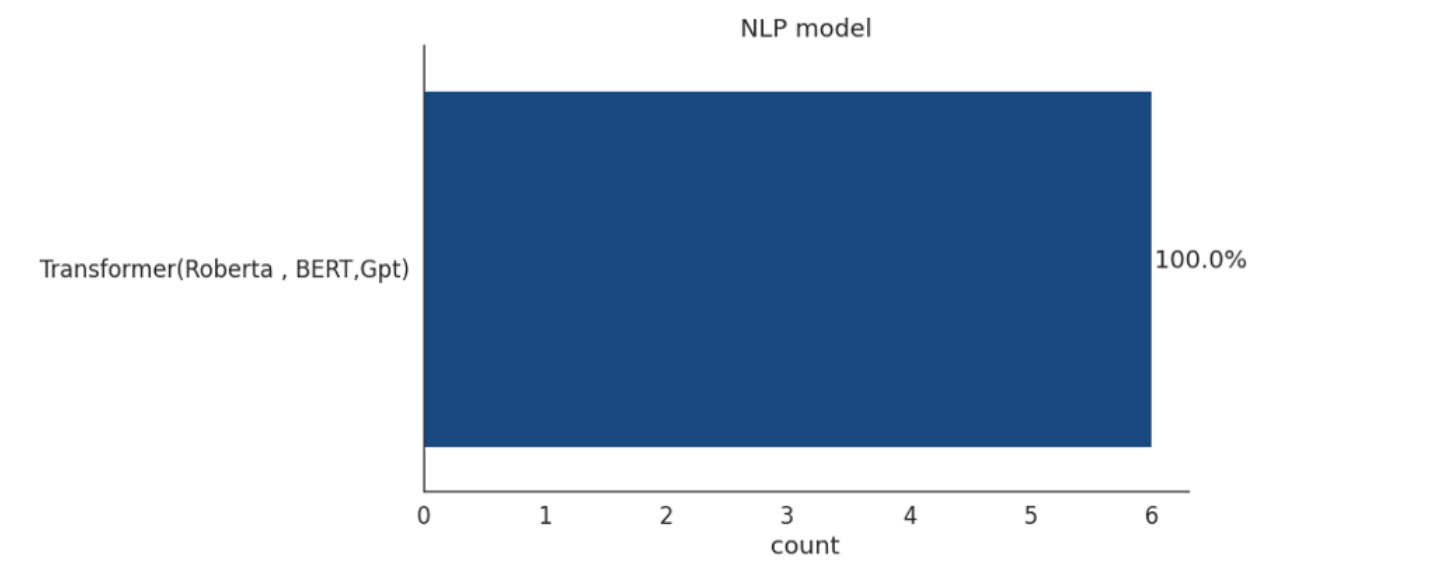
最后，RNN和CNN的结合是可能的，这可能是计算机视觉的最先进的应用。当数据适合CNN，但包含时间特征时，混合RNN和CNN的技术可能是有利的策略。



在其他架构中，EfficientNet脱颖而出，因为它专注于提高模型的准确性和效率。EfficientNet使用一种简单而有效的技术——复合系数（compound coefficient）来放大模型，使用缩放策略创建了7个不同维度的模型，其精度超过了大多数卷积神经网络的SOTA水平。

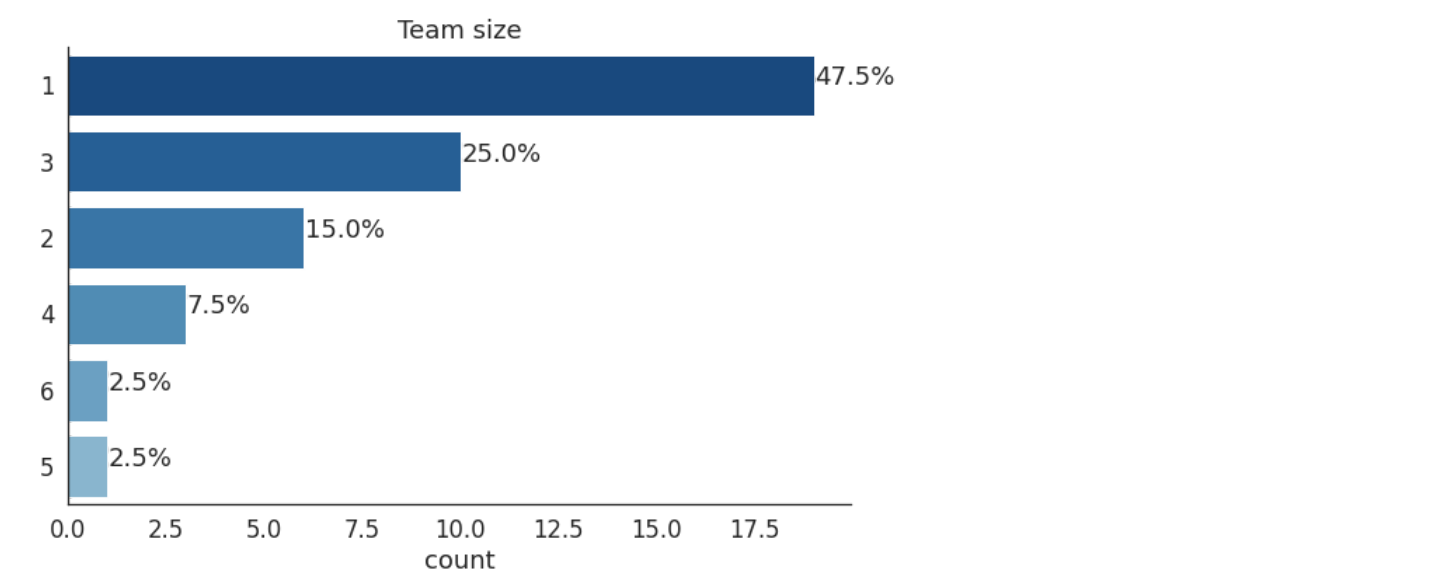
NLP

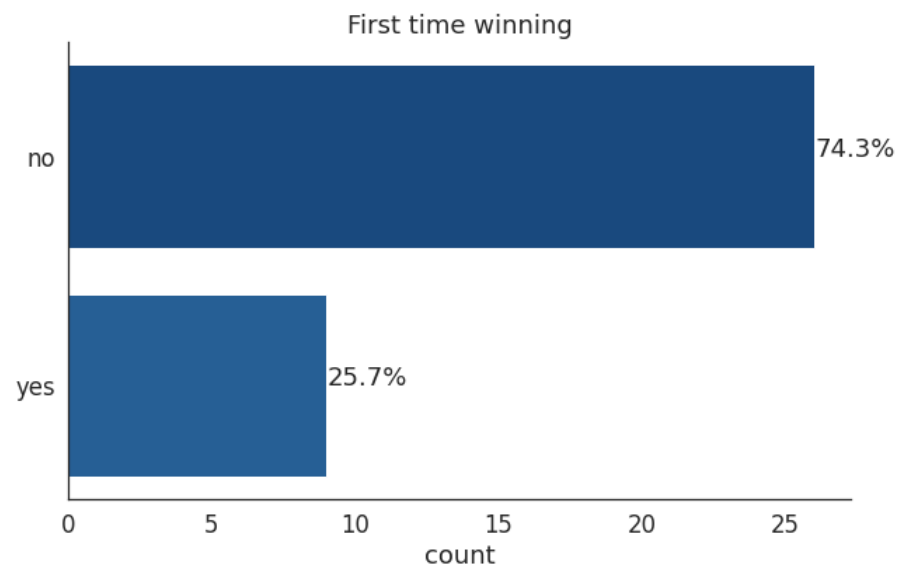
像2020年一样，2021年NLP领域大型语言模型（如Transformer）的采用比例显著增加，创历史新高。作者找到了大约6个NLP解决方案，它们全都基于transformer。



获胜团队情况

作者在数据集中追踪了35场比赛的获胜者。其中，只有9人之前从未在比赛中获奖。与2020年相比，可以看到赢得很多比赛的老参与者一次又一次获胜，只有少数几人首次得奖，在百分比上没有真正明显的变化。

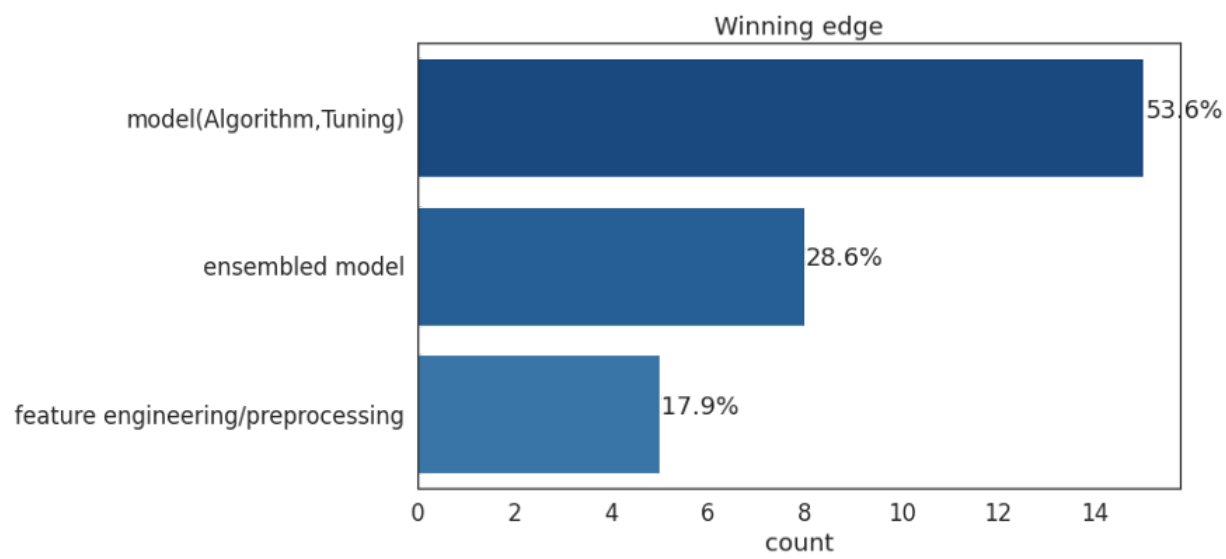


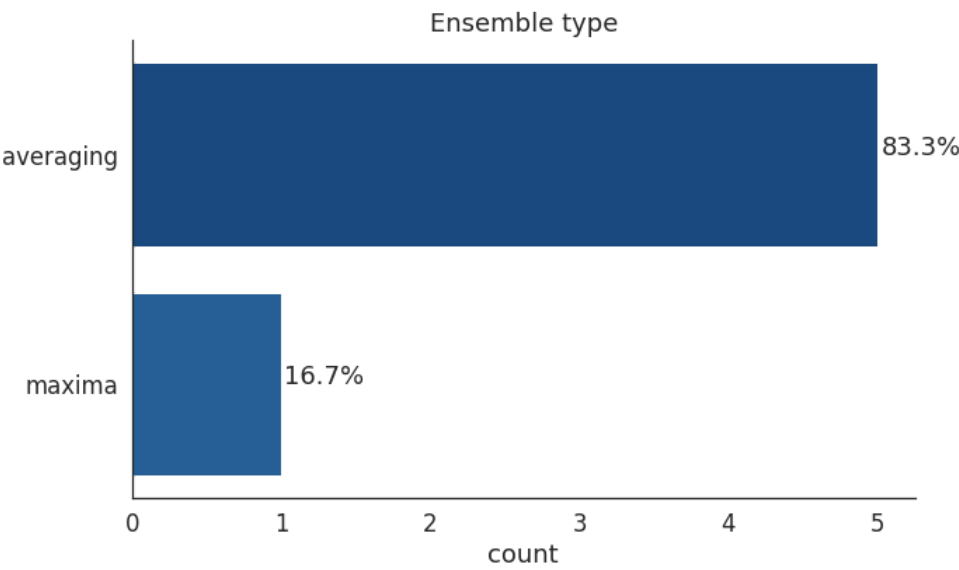


优势方案

在机器学习竞赛的优胜方案中，集成模型成为了首选方法之一。集成方法中最常用的方法是求平均，即构建多个模型并通过将输出和的平均值相加将其组合在一起，从而达到更稳健的性能。

在调整一个模型时，一旦你达到了一个收益率下降的点，通常最好重新开始构建一个产生不同类型错误的新模型，并将它们的预测求平均。

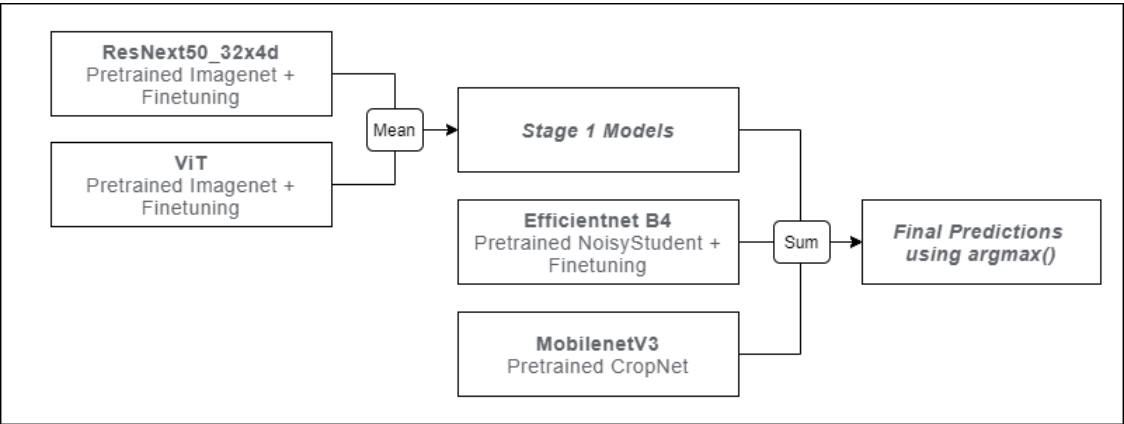




集成方法应用示例

在一个kaggle「木薯叶病分类」比赛中，选手要将木薯叶子图像分类为健康或四类疾病。冠军解决方案包括4个不同的模型CropNet、EfficientNet B4、ResNext50和ViT，并采用了平均方法。

获胜者从ResNext和ViT模型中取类权重的平均值，并在第二阶段将这种组合与MobileNet和EfficientnetB4结合。



原文链接：<https://medium.com/machine-learning-insights/winning-approach-ml-competition-2022-b89ec512b1bb>

参与问卷调查 获取百页报告

2021-2022
年度AI技术趋势报告

TRENDS OF ARTIFICIAL INTELLIGENCE TECH DEVELOPMENT REPORT 2021-2022

研究热点 趋势解读 专家问卷 数据探究

扫描二维码, 立即参与!

文章已于2022-03-28修改

喜欢此内容的人还喜欢

2022年，PyTorch在AI顶会的占比已经上80%了

机器之心

一文梳理视觉Transformer架构进展：与CNN相比，ViT赢在哪儿？

机器之心

大到31x31的超大卷积核，涨点又高效，一作解读RepLKNet

机器之心