

课程学习价值大不大？

Do we really need Curriculum Learning (CL)?

正方：课程学习价值大

辩手：陆东 王志强 耿甜甜 陈明辉

反方：课程学习价值不大

辩手：蒋希 蒋沁言 林秋实 刘柱

时间：2021.01.18

主席：郑锋

记录人：杨金字

计时人员：郑浩，肖晓宇

正一立论：

课程学习：多个方法也不亏，多视角多维度考虑问题，问题在于是不是真的需要？

是什么：从易到难

从哪来：人类受教育

有啥用：generalization capacity; convergence rate

哪里用：cv, nlp, rl, nas

课程：一个样本，一个系列的分布，

课程学习：逐渐改变样本的分布，从简单样本数量很多概率大到样本总体数量增加和熵增，最终和原始数据一致。

核心：Score function + pacing function

Difficulty ranking: loss 的值作为样本的难易程度的衡量；从某个 epoch 学到了之后没用在用到，简单的样本会在靠前的地方学到；

increment of hard exam: 确定如何设置步长，在某一步达到某一步的增长；


algorithm:

训练集排序；

根据 pacing function，每一步需要用到多大的数据（每步建立新数据集）。

effectiveness:

To guide: 简单样本 loss 比较平滑，梯度方向上小，收敛速度快；

to denoise: 先学 clean data，包含 noise 会被认为是 hard examples；鲁棒和泛化；

convergence speedup。

反四质询：

2009 开始出现 CL: yes

Start small 挖掘样本的难易程度，打分鉴定难易，pacing function: yes

loss 刚开始是大的，后来是小的: no

金字塔是课程学习吗: no

困难性采样属于课程学习范畴吗: no

正常的学习过程中，有隐式的学习过程: yes

平时的深度网络训练中是 CL 吗: no

cl 显式地规定了样本的难易程度: yes

CL>随机的学习>反 CL: yes（简单的对的知识好于简单的错的知识）

应用有限吗: no

反一立论：

CL 是什么：核心是 **data**，对样本如何输入模型进行的操作；
两个部分：**difficulty measure / score function**；**training schedule**.
首先评估简单到难，然后输入模型。

价值：应用+理论+效果三个层面的价值低。

应用：**limitation of environment** 应用环境局限；

固定数据集，打分系统对整个数据集进行打分，数据集不可变更；

困难程度评估和调度方法种类繁多，针对不同任务有不同方法；

在分类任务中应用最广泛，但效果一般。

理论：缺少严格的理论支撑；

困难的样本是更有价值的，需要更多时间和资源的，课程学习抛弃了这一点；

有文章指出当我们的模型直接面对整个训练集的时候效果是更好的（NLP）；

调度算法起的作用更大，难易程度信息起的作用有限；

效果：性价比低。

增大了整个模型训练的开销，打分函数开销极大；

有可能其他的模型是更有效的，可替代；

效果不稳定。

正四质询：

定义 2009 年文章：**yes**

关键词是课程：**yes**

课程的定义是什么：数据是有一个难易的；

容易到难是一个课程：**yes**

（1985 年才是 **starting point**）

未来的潜在价值为什么不是讨论范围之内的：

学术价值如何衡量：当前已经体现的价值；

数据集固定是必需吗：**yes**

改变数据集会影响样本的难易程度：**yes**

不同任务需要不同的方法这是一个缺点：**yes**

课程学习是一个具体的范式或工具吗：是一个方法而不是一个具体的工具

图像分类并不是最主要最广泛的领域：有文献支撑

价值体现在哪方面，如何定义价值：？

为什么困难样本需要更多资源：模型针对困难样本需要更多次的调整。

正二辩驳：

（驳）应用场景有限：多场景，加速收敛，有利于提升鲁棒性；

（辩）深度网络是 **data-driven** 的，需要大量数据；

（驳）课程学习缺少理论支撑：文献支撑，理论角度证明了理想的课程有利于网络学习，有进一步挖掘的空间；

（驳）困难样本的资源消耗：如何定义困难样本，是 **noisy data** 还是难分类的样本？

（辩）反课程学习在干净数据集上会有更好的表现；

（辩）没有增加训练的复杂度。

反三质询：

在 NLP 中有很多应用，其实 NLP 的应用是比较少的，是否同意：？

score function 开销很大：（正驳）把 **score function** 加入到 **loss** 的训练中，不需要很大开销；（反驳）还是要把 **loss** 返回到打分器，还是需要很多算力；

pacing function 的作用是不可替代的吗：（正）**pacing function** 的作用有限；

反二辩驳：

（辩）算力，样本的难易程度是自然属性，全连接结构导致难以迁移。

score function/pacing function: 在时间有限的情况下还有必要挑选这两个函数吗？能保证效果吗？样本的难易程度并不是必要的，去掉这一点后还是课程学习吗？课程学习是否真正提升了收敛速度？

反课程学习优于随机的学习；

方法合理吗，如何评估 **label** 是对是错？和其他数据增强的方法对比，优越性在哪？

正三质询：

算力换来了收敛速度的提升和降噪：（反驳）时间有限怎么办；（正驳）方法合适的情况下没有消耗更多时间；

噪声数据的 **label** 是标错的，比干净数据更难：（反驳）模型不知道对错，随机初始化条件下 **loss** 本质上没有偏差；（正驳）不一定是随机初始化；

普通情况下复杂程度不起作用，**pacing function** 作用更大也是 **CL** 的价值：（反驳）只有 **pacing function** 不算是 **CL**；（正驳）只有 **pacing function** 也是它的价值；

难的样本包含更多信息，先把模型带到一个正确的区域是有效的：（反驳）梯度优化的方向有偏差，更多的学习了简单样本，困难样本学习的资源偏少；（正驳）困难样本本身就很难分，先引导到正确的方向再学习。

反三小结：

强调论点：

- 数据集固定；耗费额外的算力；只有极端情况下 **CL** 才有效；
- 理论矛盾：**hard sample** 不一定是 **noisy data**，**hard sample** 定义是更难分类的数据；（**nlp** 本身就有自然顺序，不可打乱）；为什么 **pacing function** 是有效的，反直觉；
- 耗费的算力巨大且没办法做迁移；数据增强的方法更有效；不能应用在 **unstable data** 上；去找合适的 **function** 非常困难。

正三小结：

- **CL** 来源于人类智慧，符合人类直觉；
- 加快收敛，抗噪，提升性能；
- 不可避免的数据噪声，数据服从正态分布，先给简单数据学起，抑制尾部噪声；
- 应用场景：越来越庞大的数据集，嘈杂的 **label**，需要 **CL** 的思路来处理数据；应用场景非常多。

自由辩论：

（正）反方将 **CL** 定义狭隘化了，应该扩大化为对数据动态地学习；

（反）**learning rate** 由大变小是 **CL** 吗

（正）不是；没有 **score function/pacing function**

（反）

（正）**CL** 是数据层面，**lr** 不是数据层面的

（反）**dropout** 也不是

（正）有 **CL** 的思想

（反）有 **cl** 的思想都是 **cl** 吗

（正）不是

（反）讨论的是 **cl** 的价值，不是他的思想

（正）重要性采样不是 **cl**；**dropout** 有 **cl** 的思想在，不能单纯从数据层面考虑课程学习

（反）隐式学习：由简单到复杂就是 **cl** 吗

（正）定义是什么

（反）满足两个条件；重要性算不算调度策略；

（正）没有用到 **pacing** 不算 **cl**

（反）每次采样用到了 **score**

（正）重要性就是难易程度吗

(反) **cl** 有难易程度和按顺序输入, 不能泛化; 是否承认 **cl**>随机>反 **cl**?

(正) 有条件的情况下明显优于其他方法;

干净数据有更好的效果, 应用是有局限性;

(正) 反 **cl** 在 **clean data** 效果更好, **cl** 在 **noisy data** 更好;

反 **cl** 也是有价值的吗;

(正) 是的, 不是 **cl** 有局限性就否定他的价值;

在某些情况下无法解释, 无法证明 **cl** 比反 **cl** 更好;

(正) 反直觉不是否定 **cl** 的理由;

反直觉需要给出更多解释;

(正) 需要给出更多解释是否证明有更高的价值;

(反) 他的价值体现在它有缺陷吗?

(正) 所有的方法都有缺陷;

(反) **cl**>随机>反 **cl**?

(正) 先学习难的也是违反直觉的

(反) **cl** 是一种正则化方法, 只是为了加快收敛, 舍近求远;

(正) 为什么必须要用 **cl**? 类比预训练, 收敛速度慢和不可用只在图像分类上。nlp 上很成功, 有文献支撑;

(反) 图像上没有大规模应用;

(正) 反直觉的问题: 反 **cl** 是有价值的吗

cl 应该优于反 **cl**

(正) 反 **cl** 有时候更优越, 都有价值, 适用场合不一致, 关键是打分的设置;

(反) 先学简单的会更好

(正) 先学难的在去应对正常的挑战, 是更有效的。(考驾照)

(反) 反 **cl** 先学到更困难的样本, 反而更好, 这是反直觉的; 课程学习的出发点是从易到难, 而不是对数据做一个排序; 反 **cl** 有效恰恰是 **cl** 无效的证明。

(正) **pacing function** 扩展了 **cl** 的应用范围; 针对不同的 **task** 进行不同的课程的分类, 定义的多样性恰恰是它应用多样性的证明;

需要去搜索合适的函数;

(正) 正证明了他的价值, 举例 **NAS**;

必须在原数据集上打分, 无法迁移, **NAS** 是可迁移的;

(正) **transfer teacher**

(反) 最后打分还是需要原数据集: 分类任务下 **cnn** 可以迁移到 **transformer** 吗

(正) 同任务可以迁移, **cl** 和具体架构没有关系

(反) 卷积层难易程度有相似性, **cl** 的缺陷就是难易程度是自然属性, 与架构无关

(正) 数据集的熵不可量化, 这是机器学习的固有缺陷。

观众提问:

(felix) 样本的难易程度会有变化吗? 衡量方式一致吗?

迁移性的问题: 不同类的模型可以迁移吗? 难易程度是样本本身属性还是优化方法的问题?

(qiushi) 样本的难易程度会有变化;

样本难易程度的定义

困难样本的学习

反四总结:

课程学习价值不大: **cl** 强调样本的难易程度; **lr** 没有有意识地区分难易, 不算是 **cl**; 学习的 **epoch** 数来定义难易程度; 可迁移性极弱导致增加了时间和成本;

反直觉: 大多数任务下, 应该先学一些简单的再学难的, 但反 **cl** 优于 **cl**, 这是反直觉的, 需要更多地解释;

文献举例: 课程学习没有起到作用; **limited time** 时反 **cl** 优于 **cl**; **mixup** 效果会更好; 缺乏足够的解释和说明;

与现有结论矛盾: **svm** 先学习边界; 人脸识别, **reid** 先学难的点, **loss** 先大后小;

（驳）有缺陷不证明研究价值大；应用局限。

正四总结：

足以跟迁移学习，**active learning**，**rl** 并列的方向；

有理论支撑；

cl 和 **hem** 是可结合的，两方可以配合达到更好的效果；

研究趋势：有巨大的研究潜力；

跟数据增强有研究共性，怎样去衡量一个数据集的好坏/难易程度：**complexity+diversity**；

与自监督，数据增强，大数据集的发展有密切关联；

跟数据的量级相关，大数据集会有 **label noise** 的问题，**cl** 可以有武之地。