

王耳学推荐 | (五) 多任务学习 MMoE

原创 王耳 sad tom cat 5月10日

来自专辑

王耳学推荐

搞完了毕业论文，就差个答辩。公众号更新？这不是来了嘛(:

STAND PROUD

橋本仁 - STAND PROUD



简易目录

- 引言
- MMoE的介绍
- 总结

引言

最近看到一篇关于多任务学习的文章，是google在2018年的文章《Modeling Task Relationships in Multi-task Learning with Multi-gate Mixture-of-Experts》。因为自己也是刚接触多任务学习，所以将整理的资料做知识梳理，方便后续MMoE的模型介绍。

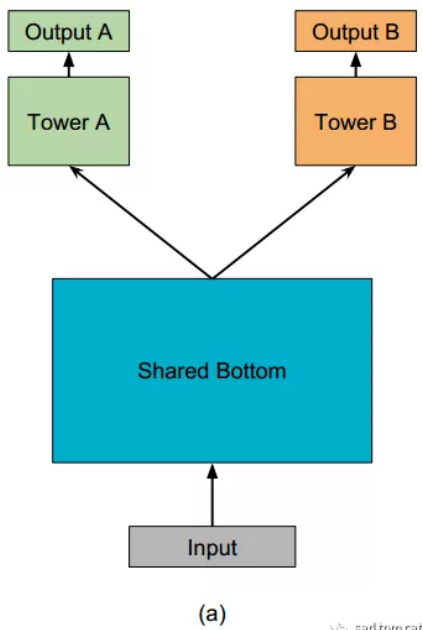
什么是多任务学习？传统的机器学习，深度学习任务是只关注一个损失函数，但是在复杂的业务场景下，常常需要关注多个指标，逐一攻克各个指标难免捉襟见肘。比如在推荐场景下，推荐排序不但关心用户的点击率，也在乎后续的用户对物品的满意度指标。换句话说，好的推荐视频不但要让你点了（点击率），还要你看的爽（观看时长，评分）；好的推荐商品不仅要让你点了（点击率），还要让你买了（转化率）...

因此，多任务学习应运而生。

MMoE的介绍

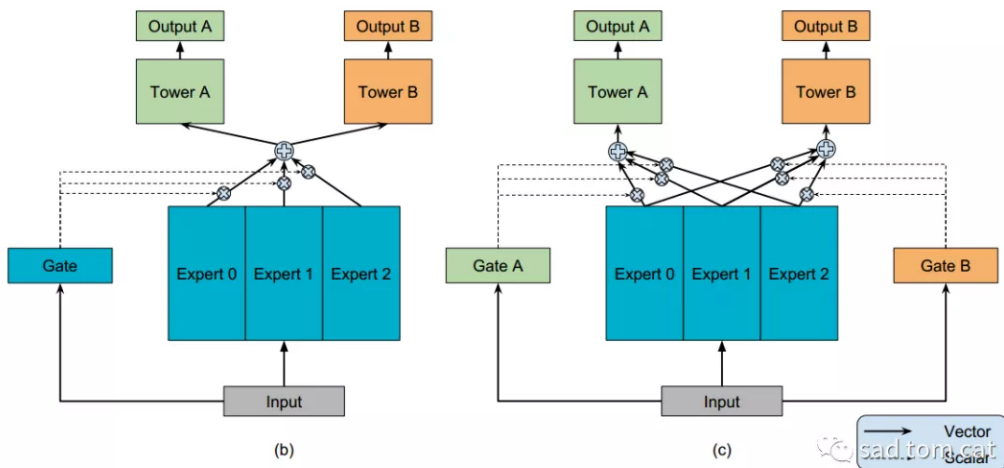
在介绍MMoE之前，需要先介绍**hard parameter sharing**和**soft parameter sharing**作为铺垫。在hard parameter sharing结构中，模型的底部会采用公共的隐藏

层进行训练，方便子任务们在前期可以学到相同的模式。然后，每个子任务根据任务特定场景下的需求，使用特定的网络进行训练。比如，shared bottom是最早应用在多任务学习场景下的网络结构，其结构如下图所示。




为方便理解，shared Bottom层看作是样本的预处理，所有的子任务都共享其参数结果；tower层看作是不同的任务处理，可以是相同的网络结构，或者不同。这样的结构在业界已经得到使用，并且取得较好的结果。美团在2018年发表的技术文章使用的就是hard parameter sharing，其推荐场景下同时对用户点击和下单两个目标进行优化，使用multi-task进行排序，模型效果可以稳定超过原本的XGBoost。

hard parameter sharing的结构设计可以减少过拟合的风险；但与此同时，如果子任务之间的差异过大，会导致公共训练层的参数出现优化困难的情况，甚至会损害效果。针对以上痛点，soft parameter sharing应运而生。准确的说，MMoE与其前身MoE使用的都是soft parameter sharing结构。下图中，(b)是MoE，(c)是MMoE。



MoE中input同时进入Expert层和Gate层：Expert层内细分为N个expert，每个expert都会一个输出结果；Gate层根据input，输出N个标量结果，分别作为Expert层输出的权重值。加权结果拼接后，分别输入至每个子任务中。具体过程，可以用如下公式表示：

$$y = \sum_{i=1}^n g(x)_i f_i(x), \quad (5)$$

where $\sum_{i=1}^n g(x)_i = 1$ and $g(x)_i$, the i th logit of the output of $g(x)$, indicates the probability for expert f_i . 

其中， f_i 是第*i*个expert的输出结果， $g(x)_i$ 是Gate层给第*i*个expert结果的权重。

而MMoE在MoE上进行了细微的改变，不再使用单一一个Gate层控制所有子任务的输入权重，而是让每个子任务单独享有一个Gate层。具体过程，可以用如下公式表示：

$$y_k = h^k(f^k(x)), \quad (6)$$

$$\text{where } f^k(x) = \sum_{i=1}^n g^k(x)_i f_i(x). \quad (7)$$

 sad torn cat

需要注意的是，子任务通过Gate层对Expert层的结果分配不同的权重——权重的大小就包含了子任务对原始输入样本空间的划分结果。MMoE正是通过Gate层的权重分配，来区分子任务之间的关系。

关于shared bottom，MoE和MMoE的改进效果，可具体参考MMoE的文献，在此不做介绍。

总结

用一个不太恰当的比方来概括Shared Bottom，MoE和MMOE：假设高考成绩是一项多任务学习训练过程，只考虑语文，数学和英语三门成绩，三门功课的成绩之和作为训练目标。Input就类比于三门功课所有学习资料之和，那么

- Shared Bottom相当于是在寻找所有学习资料的共通之处，并且将共通之处提炼出来，作为学习三门学课的“不二法门”，并且学完了就去考试（子任务训练）。但是语文，英语需要记忆的东西偏多，数学需要计算的东西偏多，是否有这样的“不二法门”，有待商榷。
- MoE则是将学习资料分成若干份（Expert层），进行系统的学习；并且加上一定的应试技巧（Gate层）。掌握这两个技能后，再应对考试（子任务训练）。
- MMoE更上一层楼，在三门功课上使用了不同的技巧：狂刷数学题，背透必考知识点...最终再应对考试。

<=== to be continued...

参考内容

【1】MMoE论文链接: <https://dl.acm.org/doi/10.1145/3219819.3220007>

【2】一篇关于MTL的综述: 《An Overview of Multi-Task Learning in Deep Neural Networks》 下载链接: <https://arxiv.org/abs/1706.05098>

【3】美团一篇多任务学习的技术文章:
<https://tech.meituan.com/2018/03/29/recommend-dnn.html>

【4】一个基于keras开发的mmoe的demo: <https://github.com/drawbridge/keras-mmoe>