PLE-Tencent 理论与实践

# 背景

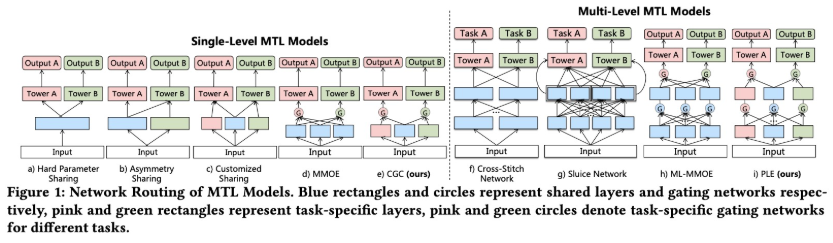
当不同学习任务之间较为相关时，多任务学习可以通过任务之间的信息共享提升学习的效率。但通常情况下，任务之间的相关性并不强，有时候甚至是冲突的，此时应用多任务学习可能带来负迁移(negative transfer)现象，会影响网络的表现。

此前已经有部分研究来减轻负迁移现象，如谷歌提出的MMoE模型。但通过实验发现，多任务学习中往往还存在seesaw phenomenon，即：多任务学习相对于多个单任务学习的模型，往往能够提升一部分任务的效果，同时牺牲另外部分任务的效果。即使通过MMoE这种方式减轻负迁移现象，seesaw phenomenon仍然是广泛存在的。

论文提出了Progressive Layered Extraction(简称PLE)，来解决多任务学习的seesaw phenomenon。

# 多任务学习模型

论文中将MTL模型分为了Single-Level MTL Models和Multi-Level MTL Models，如下：



## Single-Level MTL Models

### a. Hard Parameter Sharing

这是最为常见的MTL模型，不同的任务底层的模块是共享的，然后共享层的输出分别输入到不同任务的独有模块中，得到各自的输出。当两个任务相关性较高时，用这种结构往往可以取得不错的效果，但任务相关性不高时，会存在负迁移现象，导致效果不理想。

### b. Asymmetry Sharing

不同任务的底层模块有各自对应的输出，但其中部分任务的输出会被其他任务所使用，而部分任务则使用自己独有的输出。哪部分任务使用其他任务的输出，则需要人为指定。

### c. Customized Sharing

不同任务的底层模块不仅有各自独立的输出，还有共享的输出。

### d. MMoE

**参考之前的文章：**

### e. CGC

本论文提出的结构，下面将详细介绍。

## Multi-Level MTL Models

### f. Cross-Stitch Network

具体参考：

### g. Sluice Network

具体参考：

### h. ML-MMoE

MMoE模型的多级叠加

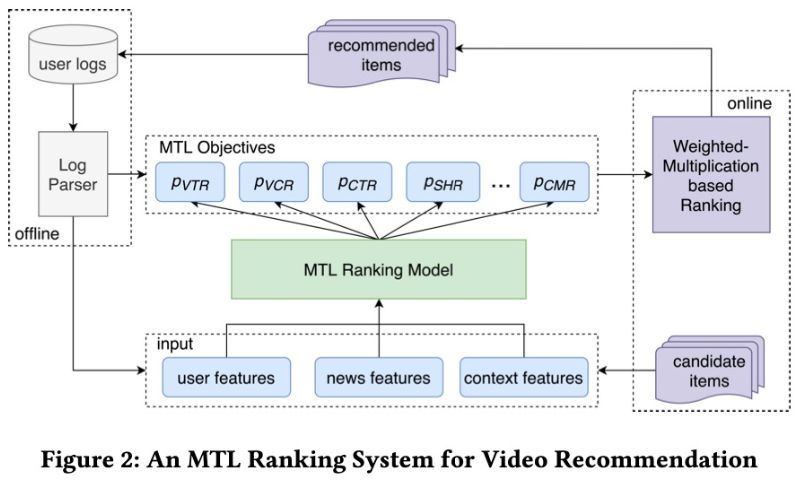
### i. PLE

CGC模型的多级叠加

# PROGRESSIVE LAYERED EXTRACTION

## seesaw phenomenon

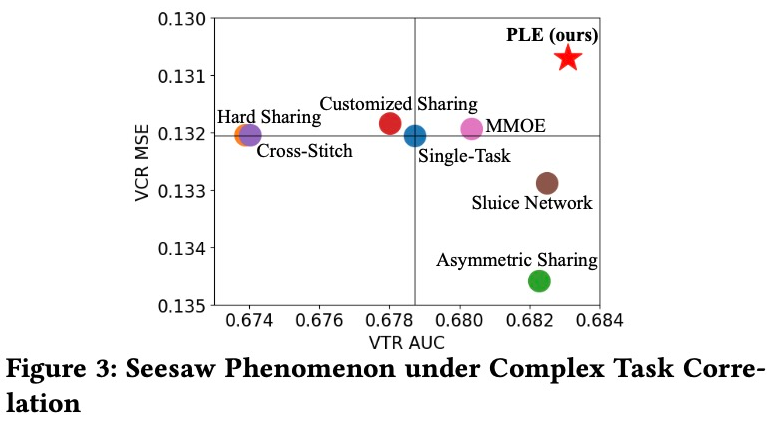
论文主要基于腾讯视频推荐中的多任务学习为例进行介绍，其视频推荐架构如下图：



这里主要关注VCR和VTR两个任务，VCR任务是视频播放完成度，即视频播放时长/视频总时长，这个是回归问题，并以MSE作为评估指标。VTR表示此次观看是否是一次有效观看，即观看时长是否在给定的阈值之上，这是二分类问题（如果没有观看，样本Label为0），并以AUC为评估指标。

两个任务之间的关系比较复杂：首先，VTR的标签是播放动作和VCR的耦合结果，因为只有观看时间超过阈值的播放动作才被视为有效观看；其次，播放动作的分布更加复杂，在存在WIFI时，部分场景有自动播放机制，这些样本就有较高的平均播放概率，而没有自动播放且需要人为显式点击的场景下，视频的平均播放概率则较低。

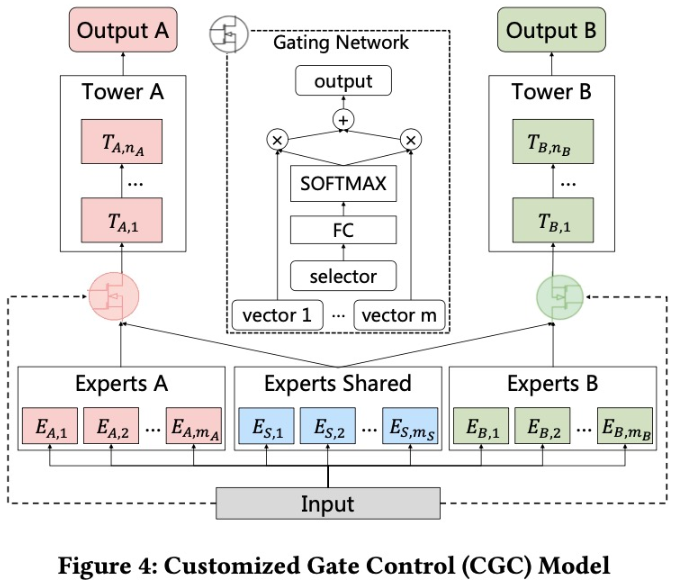
论文对比了上述所有结构的MTL在腾讯视频VCR和VTR两个任务上相对单任务模型的离线训练结果：



可以看到，几乎所有的网络结构都是在一个任务上表现优于单任务模型，而在另一个任务上表现差于单任务模型。尽管MMoE有了一定的改进，在VTR上取得了不错的收益，但在VCR上的收益接近于0。

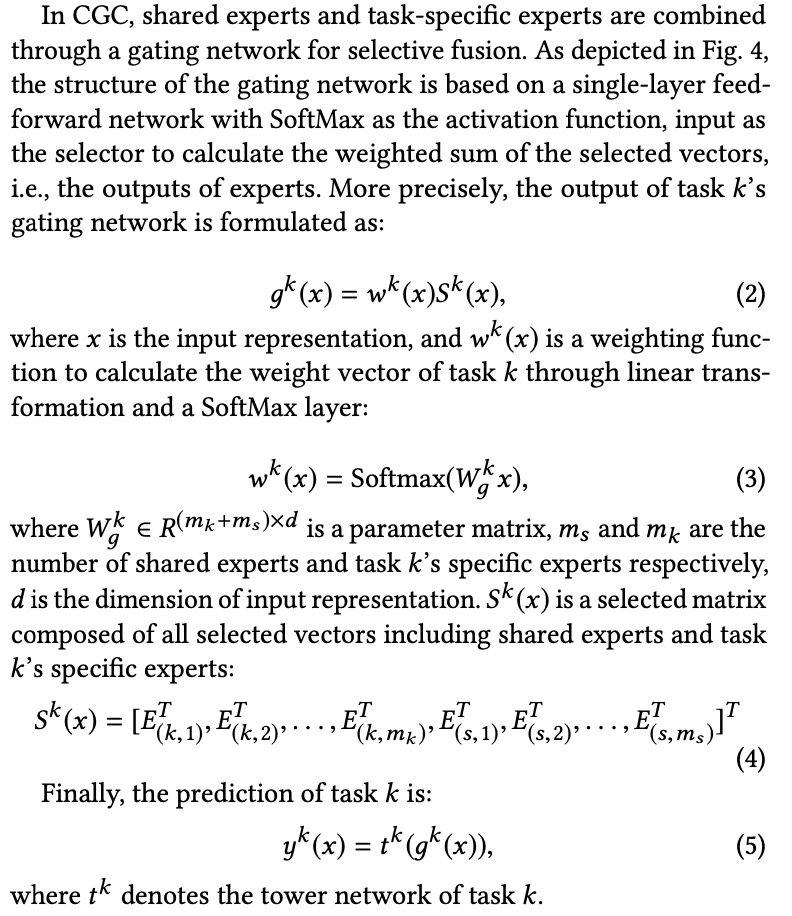
MMoE模型存在以下两方面的缺点，首先，MMoE中所有的Expert是被所有任务所共享的，这可能无法捕捉到任务之间更复杂的关系，从而给部分任务带来一定的噪声；其次，不同的Expert之间也没有交互，联合优化的效果有所折扣。

## Customized Gate Control(CGC)

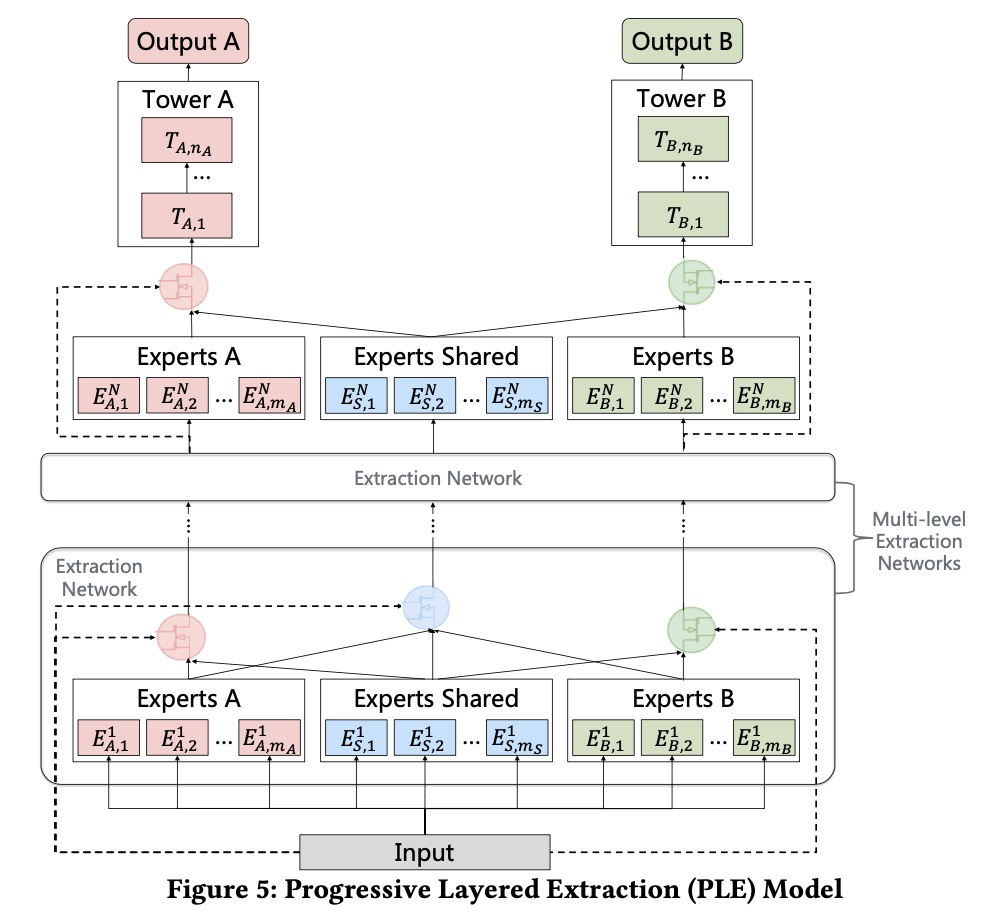


CGC网络结构是Customized Sharing和MMoE的结合版本，每个任务有共享的Expert和独有的Expert。每个Expert有多个sub-network即experts，其数量可以作为参数调节。

以任务A来说，将Experts A里面的多个Expert的输出以及Experts Shared里面的多个Expert的输出，通过类似于MMoE的门控机制之后输入到任务A的上层网络中，计算公式如下：

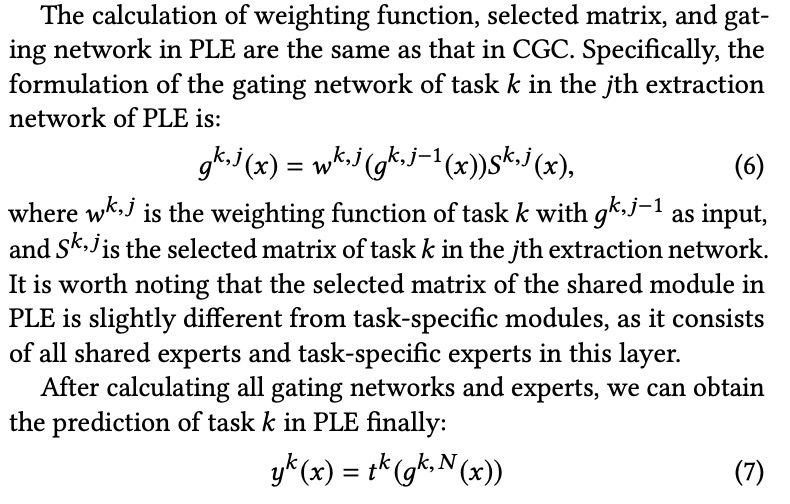


## Progressive Layered Extraction(PLE)



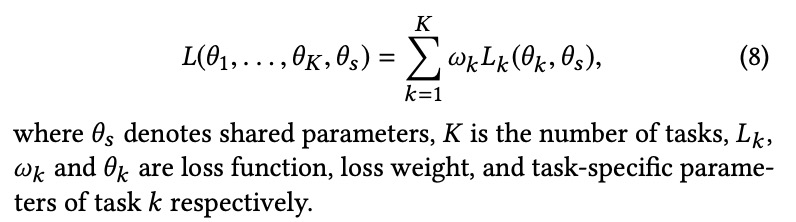
PLE是在CGC基础上考虑了不同Expert的交互，可以看作是Customized Sharing和ML-MMOE的结合版本。

下层模块中增加了多层Extraction Network。在每一层Extraction Network，共享Experts不断吸收各自独有的Experts之间的信息，而任务独有的Experts则从共享Experts中吸收有用的信息，具体计算和CGC一样。

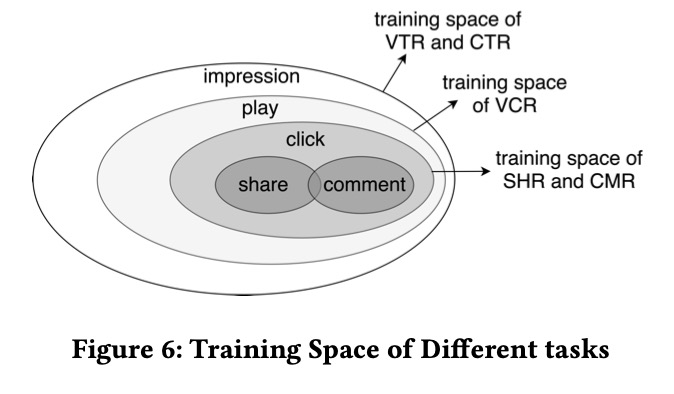


## MTL loss优化

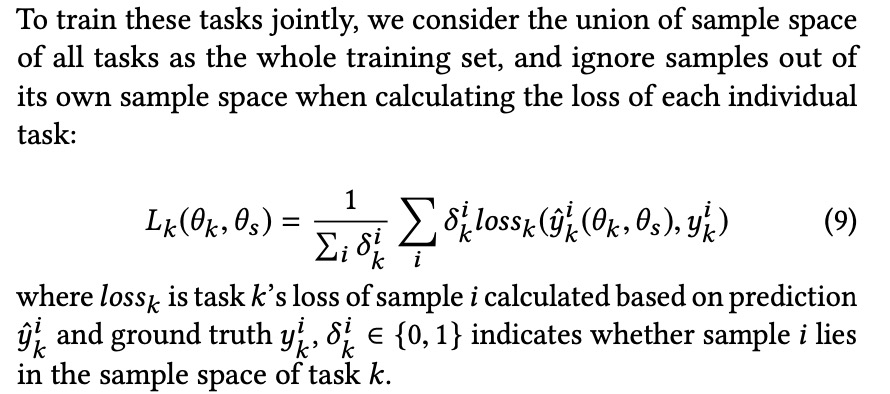
传统的MTL的损失是各任务损失的加权和，如下：



而在腾讯视频场景下，不同任务的样本空间是不一样的，比如计算视频的完成度，必须有视频点击行为才可以。不同任务的样本空间如下图所示：

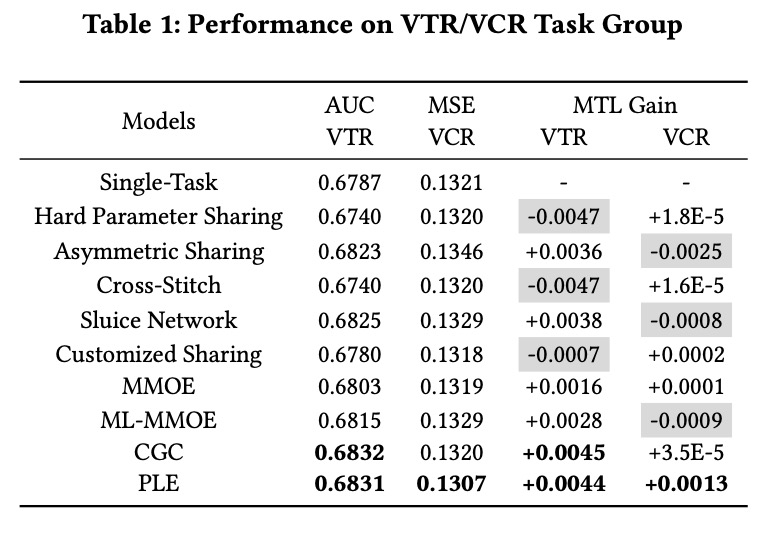


本文是在Loss上进行优化，不同的任务仍使用其各自样本空间中的样本，如下：



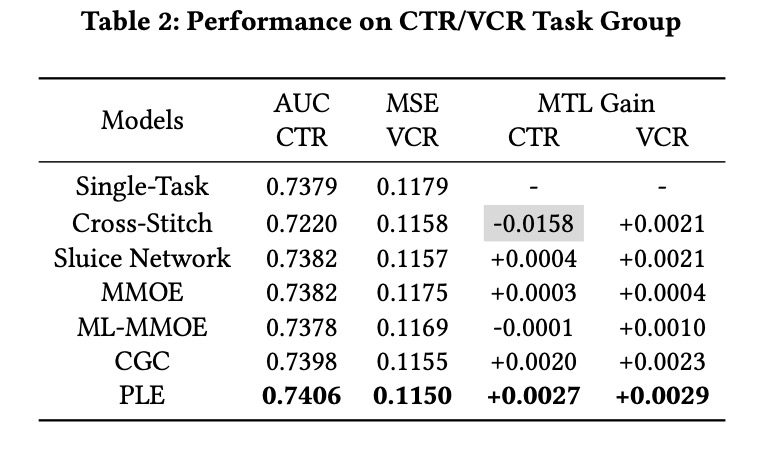
# 实验

## 离线训练结果



VCR和VTR之间的关系复杂，从图中可以看出：

* CGC和PLE的效果均显著优于其他模型，PLE效果最好；
* 许多模型存在seesaw phenomenon，VCR提升但VTR下降，或者，VTR提升但VCR下降；
* MMoE均能提升VTR和VCR，但效果不显著；



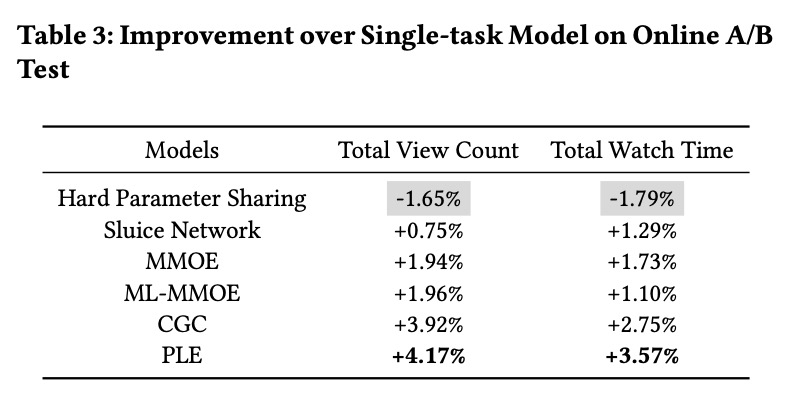
CTR和VCR之间是正相关关系，关系简单，从图中可以看出：

* 除了Cross-Stitch Network，其他模型均在CTR和VCR任务上取得正向效果，没有出现seesaw phenomenon；
* CGC和PLE的效果均显著优于其他模型，PLE效果最好；

## 线上ABtest效果

论文在线上进行了4周的ABtest实验，主要优化目标是VCR和VTR，结果如下：

* MTL模型比单任务模型效果更好；
* CGC和PLE的效果均显著优于其他模型，PLE效果最好；



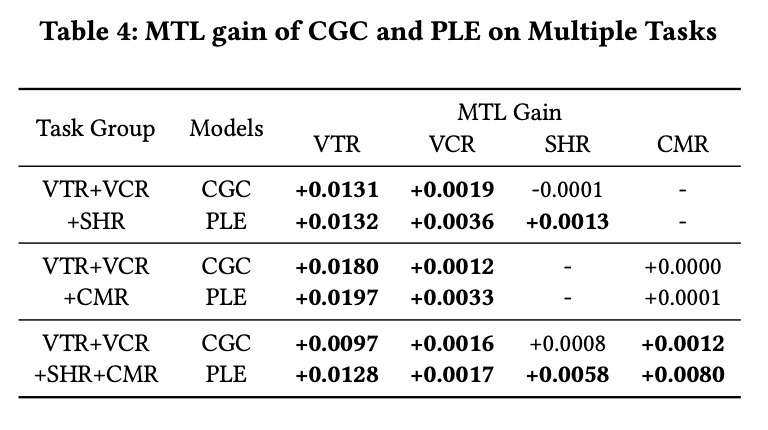
## 更多任务表现结果

除了VCR和VTR，作者还引入了SHR和CMR两个任务到MTL框架中，比较了CGC和PLE的效果，从图中可以看出：

CGC和PLE在多任务学习中的效果均显著优于单任务模型；

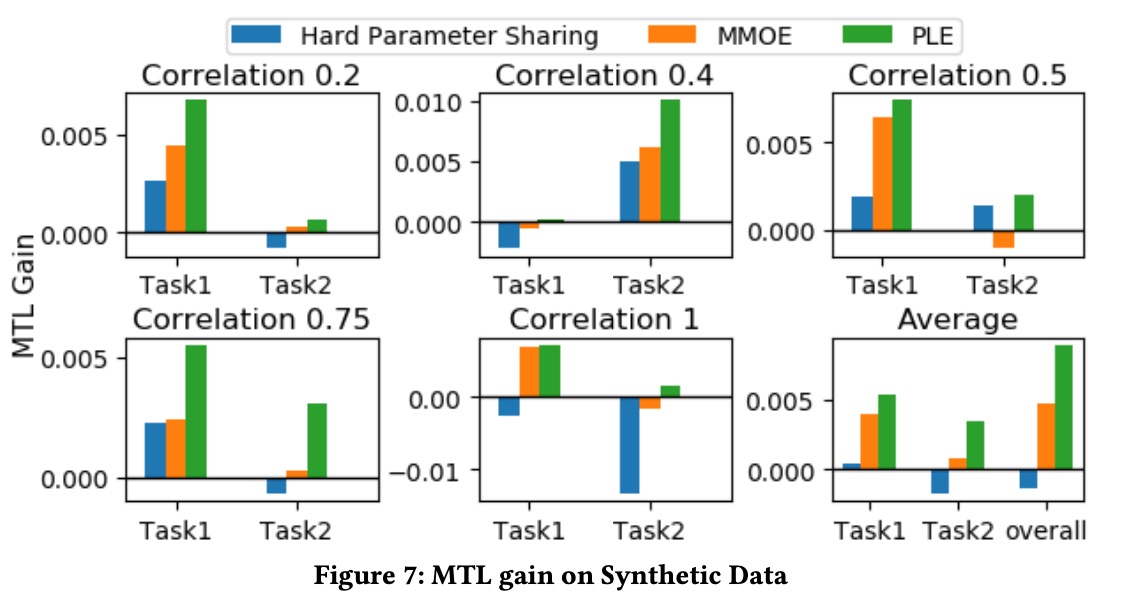
CGC和PLE在超过两个子任务的多任务学习中，可以有效地避免seesaw phenomenon和负迁移；

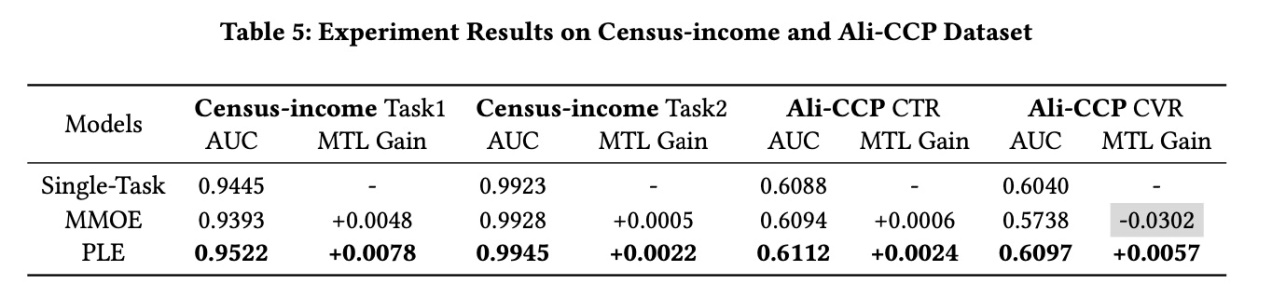
PLE的效果优于CGC；



## 公开数据集表现

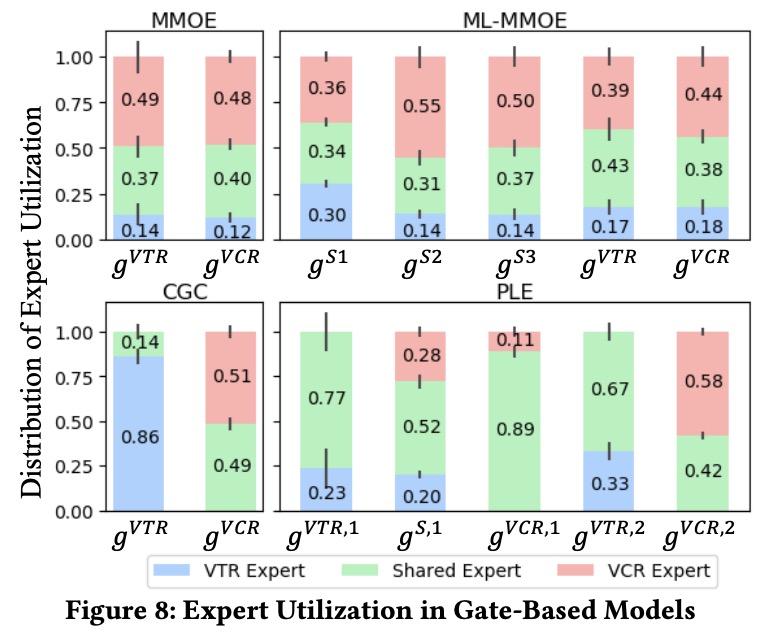
作者在三个公开数据集上比较了Hard Parameter Sharing、MMoE和PLE的效果，从图中可以看出：Hard Parameter Sharing和MMoE均存在seesaw phenomenon，而PLE则表现很好，有效地消除了seesaw phenomenon；





## Expert utilization analysis

为了公平对比，作者采用了single-level的PLE和ML-MMoE，然后可视化了CGC、MMoE、PLE和ML-MMoE的expert的utilization，如图所示：



从图中可以看出：

* CGC里VTR和VCR的expert权重有着显著不同，而MMoE中几乎相似，这也表明CGC效果优于MMoE；
* MMoE和ML-MMoE所有的expert权重几乎不为0，这也表明：没有先验知识的情况下，MMOE and ML-MMOE很难收敛到CGC和PLE的结构，即便理论上存在可能性；
* 与CGC相比，PLE的shared experts对Tower有更大的影响，尤其是在VTR任务中。PLE性能优于CGC，这表明共享更高级的更深层表示的价值。换句话说，为了在任务之间共享某些更深的语义表示，PLE提供了更好的联合路由和学习方案。

# 其他多任务模型

<https://mp.weixin.qq.com/s/c7BdWUM9BFQZfzWP7C_wfQ>