Google 多任务学习框架 MMoE

原创 阿泽 贪心科技AI 5月29日

来自专辑

paper

基于神经网络的多任务学习已经过成功应用内许多现实应用中，比如说之前我们介绍的阿里巴巴基于多任务联合学习的 ESMM 算法，其利用多任务学习解决了 CVR 中样本选择偏差和样本稀疏这两大问题，并在实际应用场景中取得了不错的成绩。

多任务学习的目的在于用一个模型来同时学习多个目标和任务，但常用的任务模型的预测质量通常对任务之间的关系很敏感（数据分布不同，ESMM 解决的也是这个问题），因此，google 提出多门混合专家算法（Multi-gate Mixture-of-Experts，以下简称 MMoE）旨在学习如何从数据中权衡任务目标（task-specific objectives）和任务之间（inter-task relationships）的关系。所有任务之间共享混合专家结构（MoE）的子模型来适应多任务学习，同时还拥有可训练的门控网路（Gating Network）以优化每一个任务。

MMoE 算法在任务相关性较低时能够具有更好的性能，同时也可以提高模型的可训练性。作者也将 MMoE 应用于真实场景中，包括二分类和推荐系统，并取得了不错的成绩。

1.Introduction

这一节主要介绍一些基础知识和背景，包括多什么是任务学习和多任务学习的挑战。

1.1 MTL

MTL（Multi-Task Learning）有很多形式：联合学习（joint learning）、自主学习（learning to learn）和带有辅助任务的学习（learning with auxiliary task）等都可以指 MTL。一般来说，优化多个损失函数就等同于进行多任务学习（与单任务学习相反）。

本篇文章，包括之前的 ESMM 都是属于带有辅助任务的多任务学习。

MTL 的目标在于**通过利用包含在相关任务训练信号中特定领域的信息来提高泛化能力**。

那么，什么是相关任务呢？我们有以下几个不严谨的解释：

1. 使用相同特征做判断的任务；
2. 任务的分类边界接近；
3. 预测同个个体属性的不同方面比预测不同个体属性的不同方面更相关；
4. 共同训练时能够提供帮助并不一定相关，因为加入噪声有时也可以增加泛化能力。

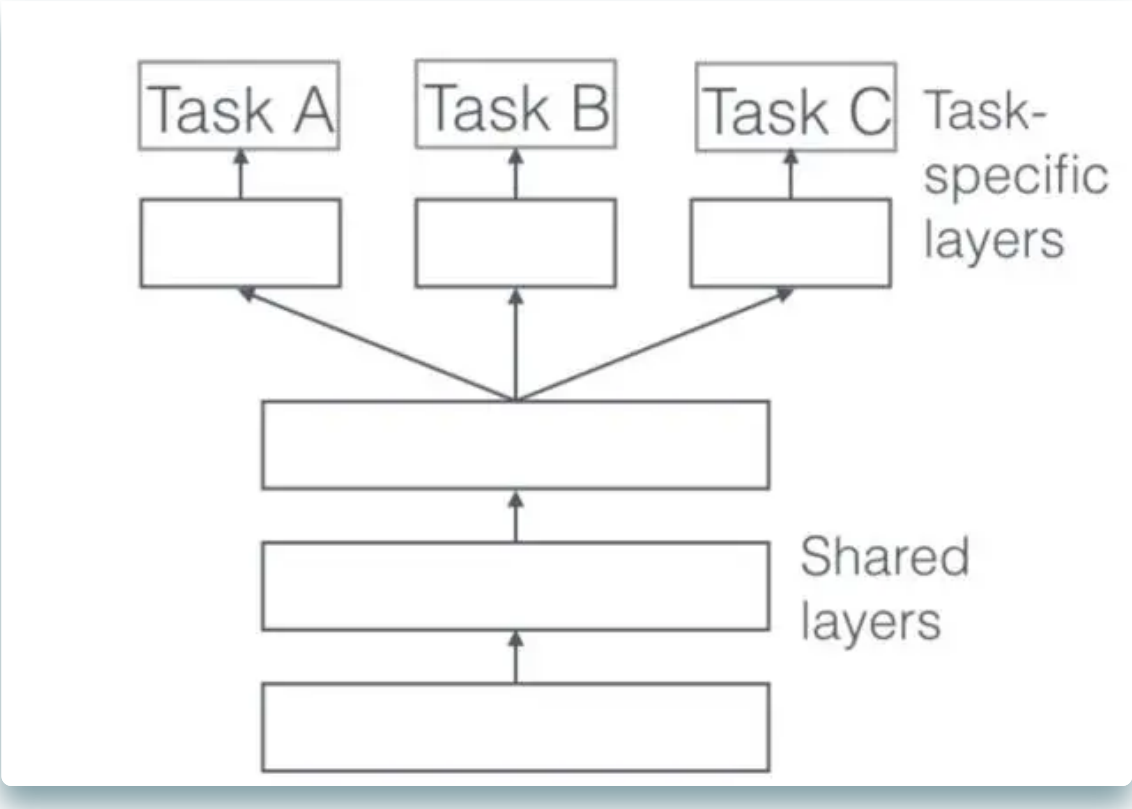
任务是否相似不是非0即1的。越相似的任务收益越大。但即使相关性不佳的任务也会有所收益。

1.1.1 Common form

MLT 主要有两种形式，一种是基于参数的共享，另一种是基于约束的共享。

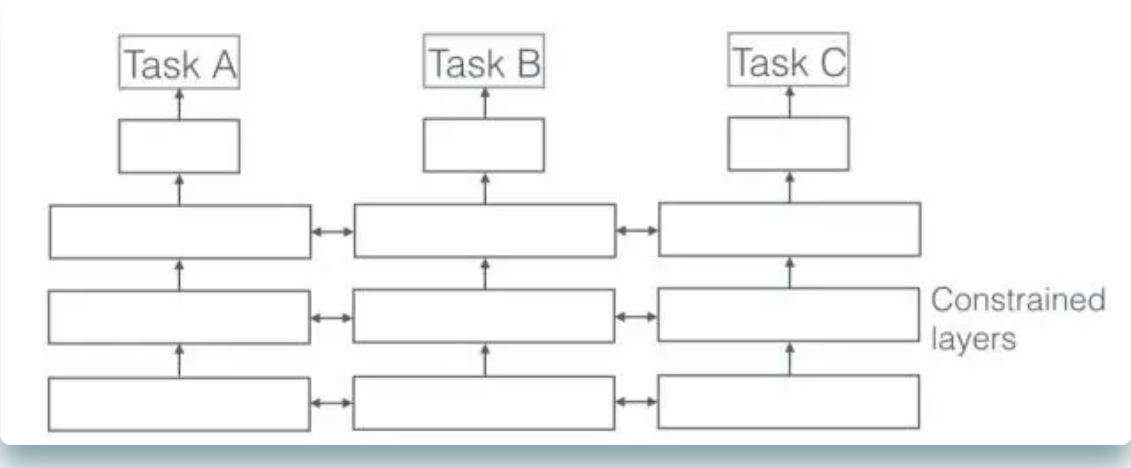
Hard 参数共享

参数共享的形式在基于神经网络的 MLT 中非常常见，其在所有任务中共享隐藏层并同时保留几个特定任务的输出层。这种方式有助于降低过拟合风险，因为同时学习的任务越多，模型找到一个含有所有任务的表征就越困难，从而过拟合某个特定任务的可能性就越小。ESMM 就属于这种类型的 MLT。



Soft 参数共享

每个任务都有自己的参数和模型，最后通过对不同任务的参数之间的差异施加约束。比如可以使用L2进行正则，迹范数（trace norm）等。



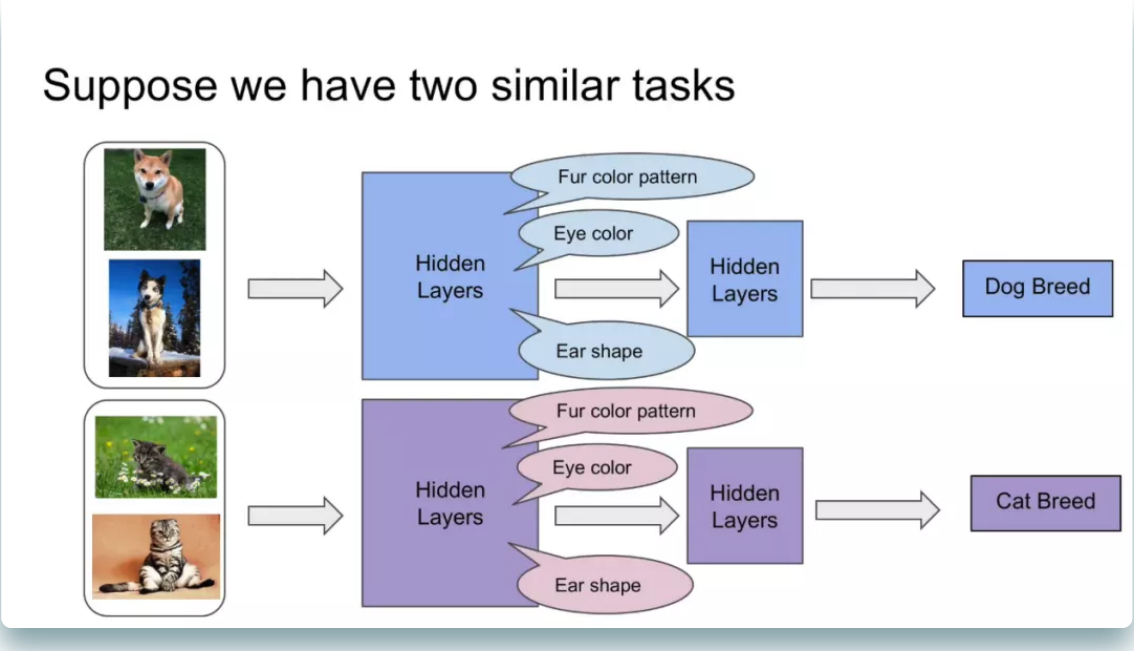
1.1.2 Why MTL work

那么，为什么 MLT 有效呢？主要有以下几点原因：

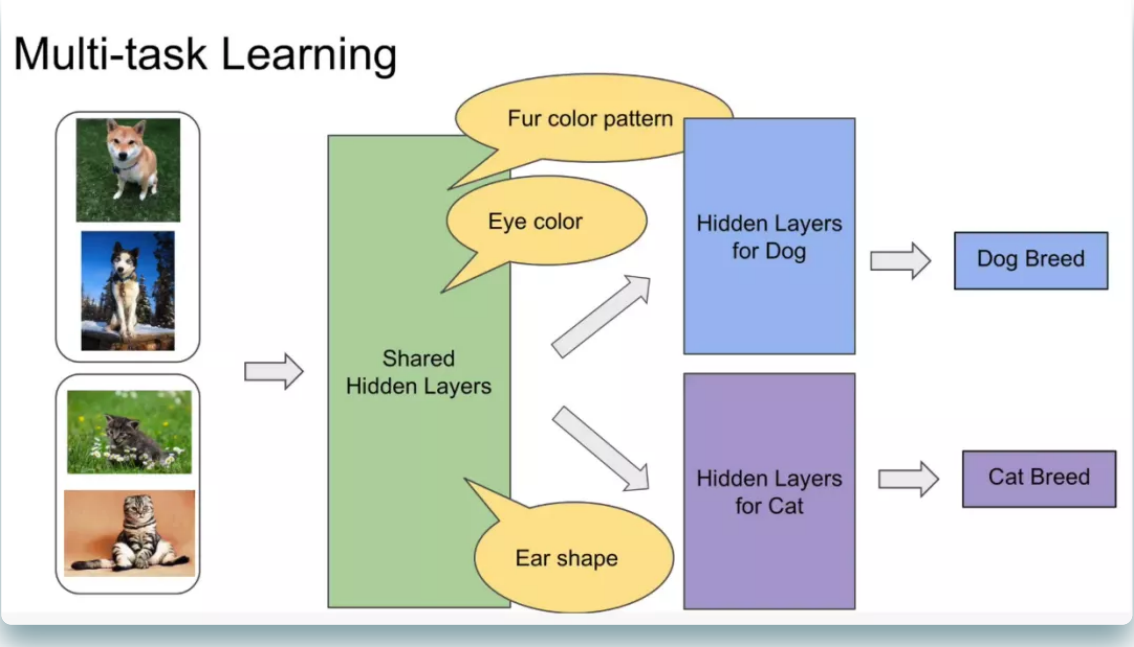
1. 多任务一起学习时，会互相增加噪声，从而提高模型的泛化能力；
2. 多任务相关作用，逃离局部最优解；
3. 多任务共同作用模型的更新，增加错误反馈；
4. 降低了过拟合的风险；
5. 类似 ESMM，解决了样本偏差和数据稀疏问题，未来也可以用来解决冷启动问题。

1.2 Challenge in MTL

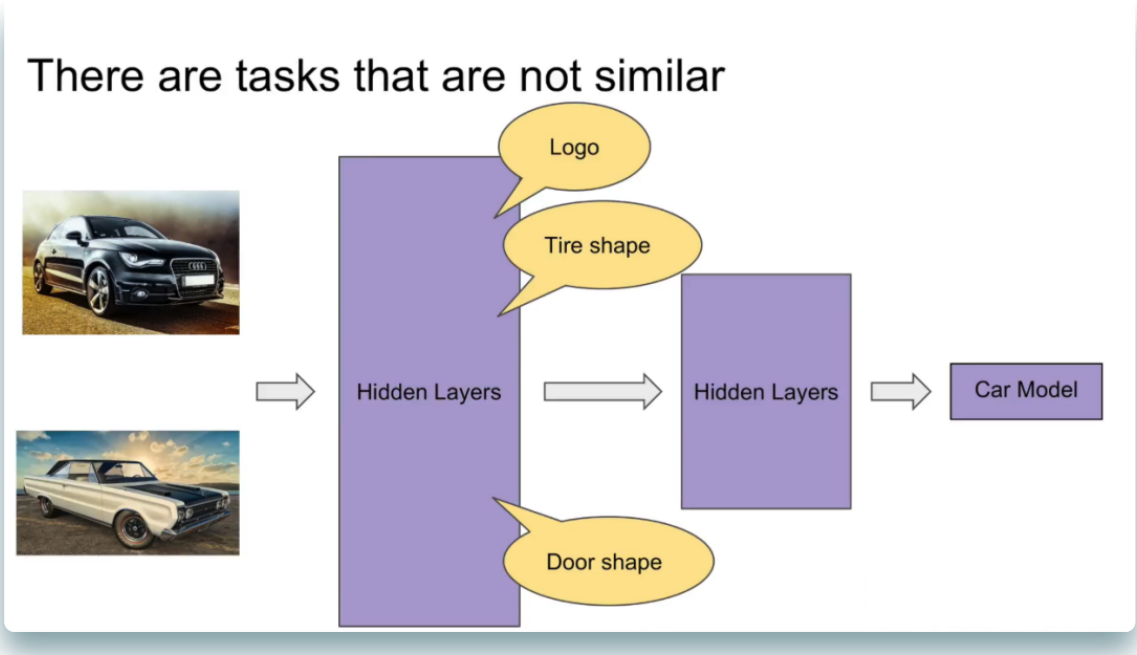
在多任务学习中，假设有这样两个相似的任务：猫分类和狗分类。他们通常会有比较接近的底层特征，比如皮毛、颜色等等。如下图所示：



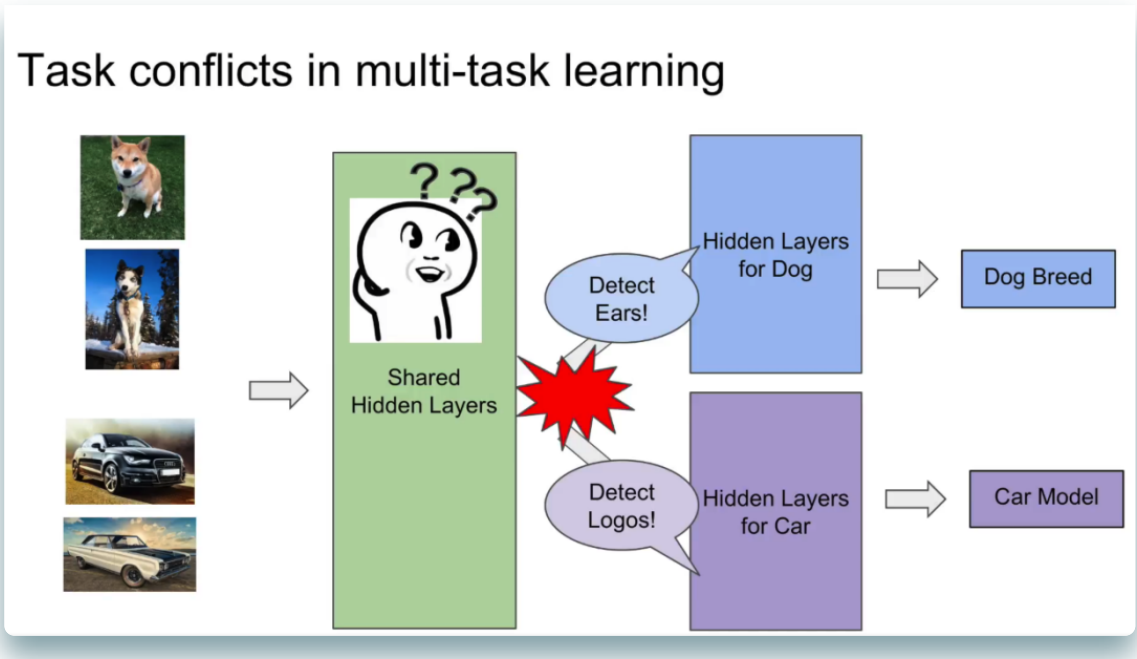
多任务的学习的本质在于共享表示层，并使得任务之间相互影响：



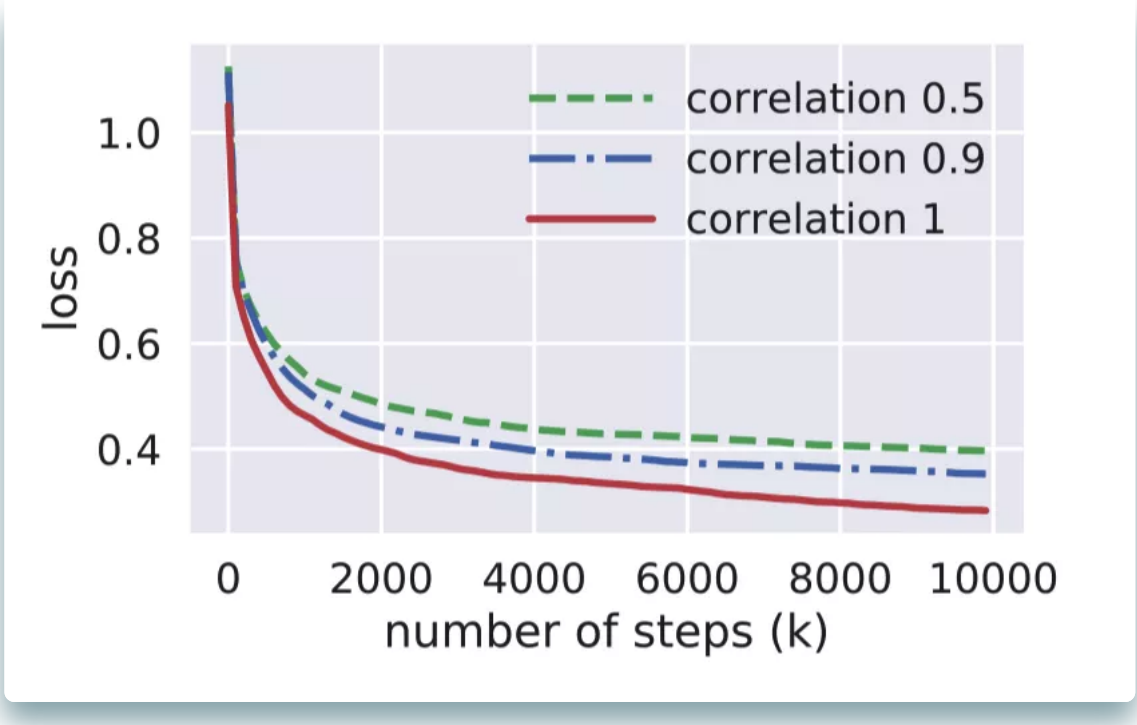
如果我们现在有一个与猫分类和狗分类相关性不是太高的任务，如汽车分类：



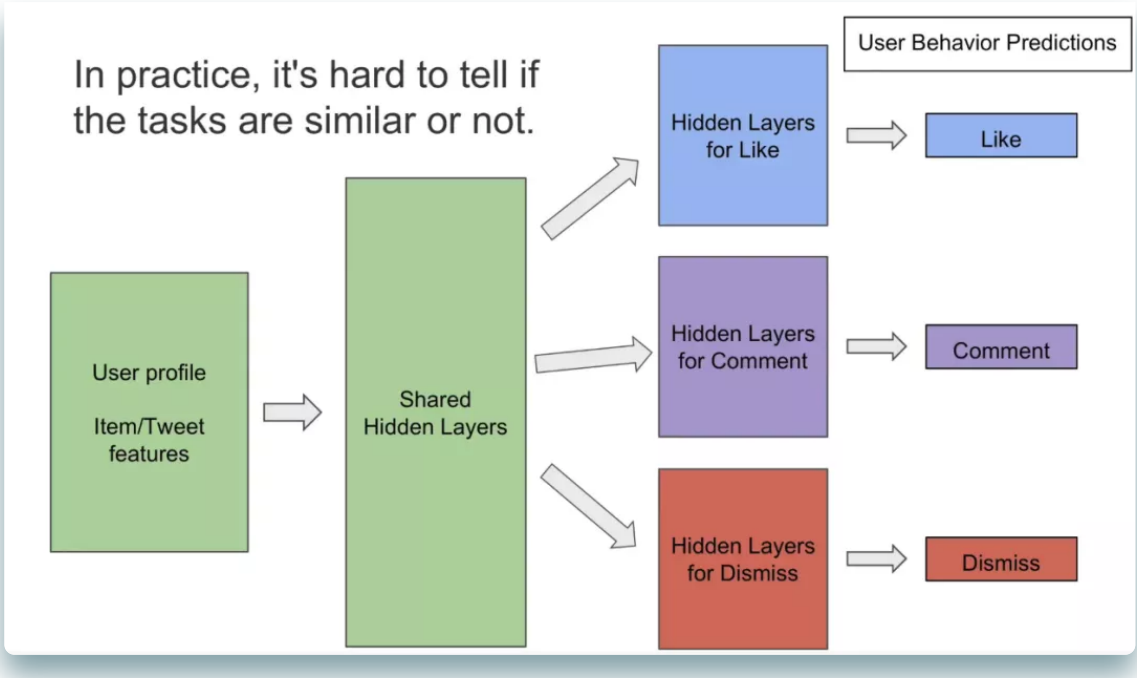
那么我们在用多任务学习时，由于底层表示诧异很大，所以共享表示层的效果也就没有那么明显，而且更有可能会出现冲突或者噪声：



作者给出相关性不同的数据集上多任务的表现，其也阐述了，相关性越低，多任务学习的效果越差：



其实，在实际过程中，如何去识别不同任务之间的相关性也是非常难的：



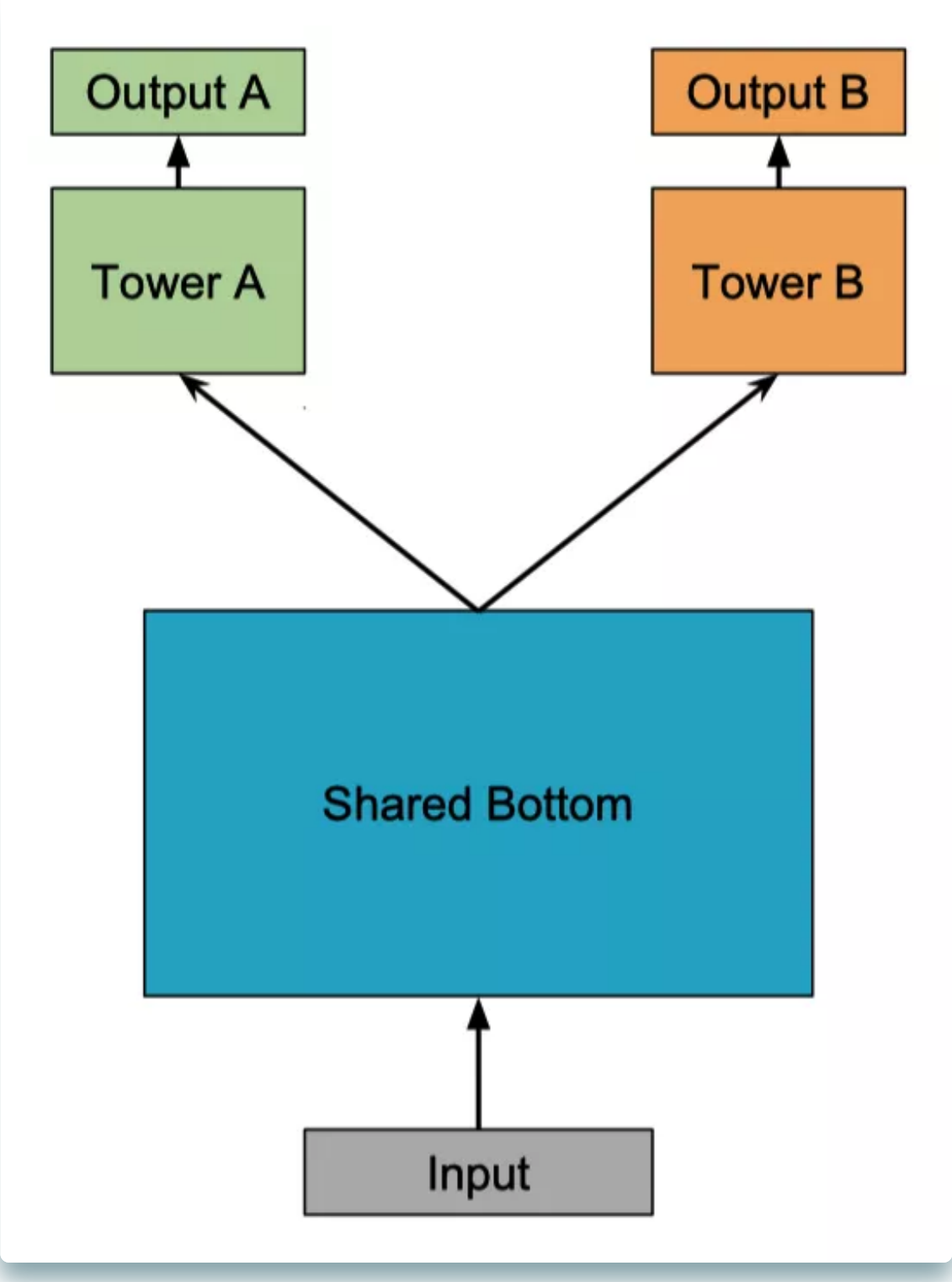
基于以上原因，作者提出了 MMoE 框架，旨在构建一个兼容性更强的多任务学习框架。

2.MMoE

本节我们详细介绍下 MMoE 框架。

2.1 Shared-Bottom model

先简单结下 shared-bottom 模型，ESMM 模型就是基于 shared-bottom 的多任务模型。这篇文章把该框架作为多任务模型的 baseline，其结构如下图所示：



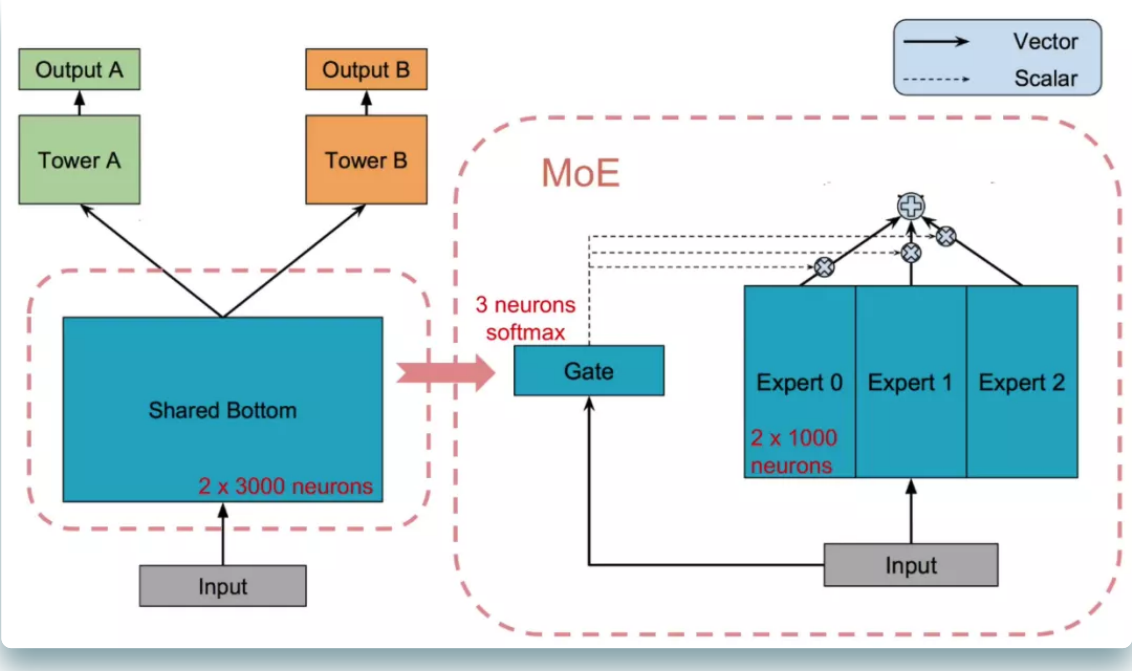
给出公式定义：

$$y_k = h^k(f(x))$$

其中，f 为表征函数，h^k 为第 k 个子网络（tower 网络）。

2.2 One-gate MoE Layer

而 One-gate MoE layer 则是将隐藏层划分为三个专家（expert）子网，同时接入一个 Gate 网络将各个子网的输出和输入信息进行组合，并将得到的结果进行相加。



公式如下：

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

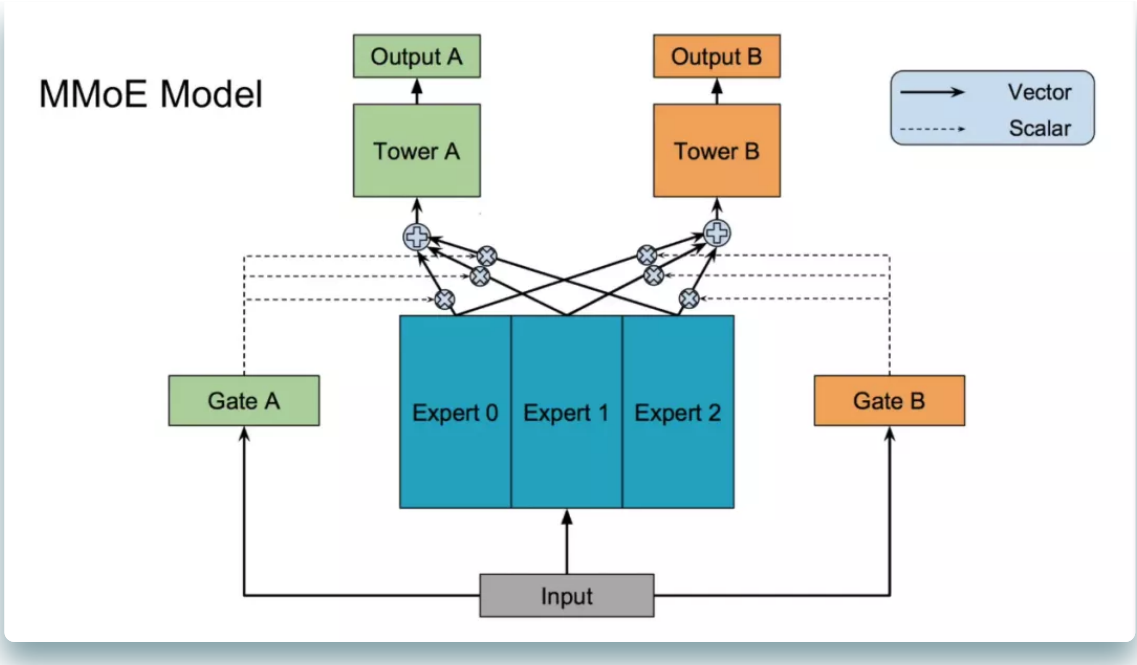
其中，$f_i(x)$ 为第 i 个专家子网的输出；$\sum_{i=1}^{n} g(x)_i = 1$，$g(x)_i$ 为第 i 个 logit 输出，表示专家子网 f_i 的权重，其由 gate 网络计算得出。

MoE 的主要目标是实现条件计算，对于每个数据而言，只有部分网络是活跃的，该模型可以通过限制输入的门控网络来选择专家网络的子集。

2.3 Multi-gate MoE model

MoE 能够实现不同数据多样化使用共享层，但针对不同任务而言，其使用的共享层是一致的。这种情况下，如果任务相关性较低，则会导致模型性能下降。

所以，作者在 MoE 的基础上提出了 MMoE 模型，为每个任务都设置了一个 Gate 网路，旨在使得不同任务和不同数据可以多样化的使用共享层，其模型结构如下：



给出公式定义：

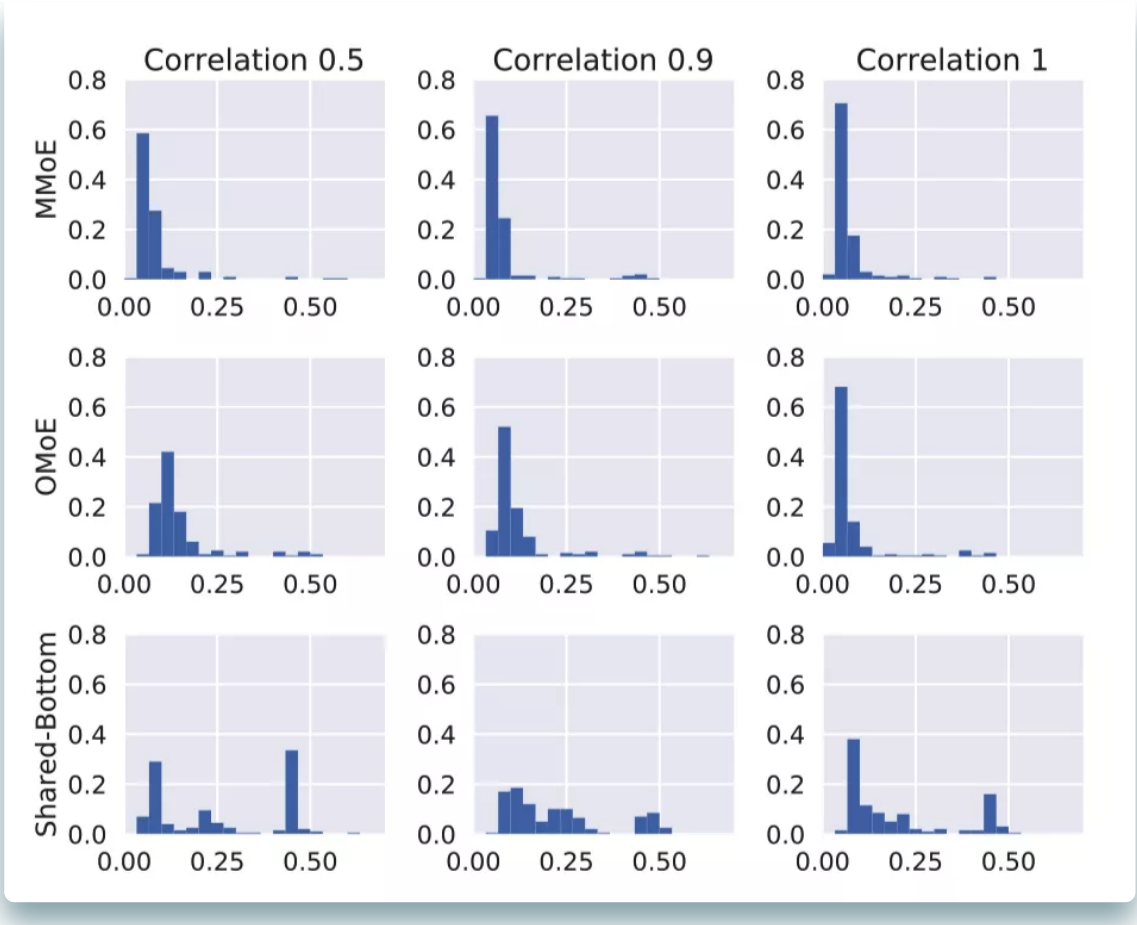
$$y_k = h^k(f^k(x)) \qquad where \qquad f^k(x) = \sum_{i=1}^{n} g$$

这种情况下，每个 Gate 网络都可以根据不同任务来选择专家网络的子集，所以即使两个任务并不是十分相关，那么经过 Gate 后也可以得到不同的权重系数，此时，MMoE 可以充分利用部分 expert 网络的信息，近似于单个任务；而如果两个任务相关性高，那么 Gate 的权重分布相差会不大，会类似于一般的多任务学习。

3.Experiment

简单看下实验。

首先是不同 MLT 模型对在不同相关性任务下的参数分布，其可以反应模型的鲁棒性。可以看到 MMeE 模型性能还是比较稳定的。



第一组数据集的表现：

Group 1	AUC/Income		AUC/Marital Stat	
	best	mean	w/ best income	mean
Single-Task	0.9398	0.9337	**0.9933**	0.9922
Shared-Bottom	0.9361	0.9295	0.9915	0.9921
L2-Constrained	0.9389	0.9359	0.9922	0.9918
Cross-Stitch	0.9406	**0.9361**	0.9917	0.9922
Tensor-Factorization	0.7460	0.6765	0.8175	0.8412
OMoE	0.9387	0.9319	0.9928	0.9923
MMoE	**0.9410**	0.9359	0.9926	**0.9927**

第二组数据集的表现：

Group 2	AUC/Education		AUC/Marital Stat	
	best	mean	w/ best education	mean
Single-Task	0.8843	0.8792	**0.9933**	0.9922
Shared-Bottom	0.8836	0.8813	0.9927	0.9917
L2-Constrained	0.8855	0.8823	0.9923	0.9918
Cross-Stitch	0.8855	0.8819	0.9919	0.9921
Tensor-Factorization	0.7367	0.7256	0.7453	0.7497
OMoE	0.8852	0.8813	0.9915	0.9912
MMoE	**0.8860**	**0.8826**	0.9932	**0.9924**

大型推荐系统的表现：

Metric	AUC@2M	AUC@4M	AUC@6M	R2@2M	R2@4M	R2@6M
Shared-Bottom	0.6879	0.6888	0.6900	0.08812	0.09159	0.09287
L2-Constrained	0.6866	0.6881	0.6895	0.08668	0.09030	0.09213
Cross-Stitch	0.6880	0.6885	0.6899	0.08949	0.09112	0.09332
OMoE	0.6876	0.6891	0.6893	0.08749	0.09085	0.09230
MMoE	**0.6894**	**0.6897**	**0.6908**	**0.08978**	**0.09263**	**0.09362**

Gate 网络在两个任务的不同分布：

4.Conclusion

总结：作者提出了一种新颖的多任务学习方法——MMoE，其通过多个 Gate 网络来自适应学习不同数据在不同任务下的与专家子网的权重关系系数，从而在相关性较低的多任务学习中取得不错的成绩。

共享网络节省了大量计算资源，且 Gate 网络参数较少，所以 MMoE 模型很大程度上也保持了计算优势。

5.Reference

1. Ma J, Zhao Z, Yi X, et al. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 1930-1939.
2. 《KDD 2018 vedio》
3. **《Multi task learning多任务学习背景简介》**