《火箭发射：一种有效的轻量网络训练框架》

《Rocket Launching: A Universal and Efficient Framework for Training Well-performing Light Net》

团队名称：阿里妈妈事业部

作者：周国睿、范颖、崔润鹏、卞维杰、朱小强、盖坤

摘要

像点击率预估这样的在线实时响应系统对响应时间要求非常严格，结构复杂，层数很深的深度模型不能很好的满足严苛的响应时间的限制。为了获得满足响应时间限制的具有优良表现的模型，我们提出了一个新型框架：训练阶段，同时训练繁简两个复杂度有明显差异的网络，简单的网络称为轻量网络（light net），复杂的网络称为助推器网络（booster net），相比前者，有更强的学习能力。两网络共享部分参数，分别学习类别标记，此外，轻量网络通过学习助推器的 soft target 来模仿助推器的学习过程，从而得到更好的训练效果。测试阶段，仅采用轻量网络进行预测。我们的方法被称作“火箭发射”系统。在公开数据集和阿里巴巴的在线展示广告系统上，我们的方法在不提高在线响应时间的前提下，均提高了预测效果，展现了其在在线模型上应用的巨大价值。

研究背景

响应时间直接决定在线响应系统的效果和用户体验。比如在线展示广告系统中，针对一个用户，需要在几 ms 内，对上百个候选广告的点击率进行预估。因此，如何在严苛的响应时间内，提高模型的在线预测效果，是工业界面临的一个巨大问题。

已有方法介绍

目前有 2 种思路来解决模型响应时间的这个问题：

一方面，可以在固定模型结构和参数的情况下，用计算数值压缩来降低 inference 时间，同时也有设计更精简的模型以及更改模型计算方式的工作，如 Mobile Net 和 ShuffleNet 等工作；
另一方面，利用复杂的模型来辅助一个精简模型的训练，测试阶段，利用学习好的小模型来进行推断，如 KD, MIMIC。这两种方案并不冲突，在大多数情况下第二种方案可以通过第一种方案进一步降低 inference 时间，同时，考虑到相对于严苛的在线响应时间，我们有更自由的训练时间，有能力训练一个复杂的模型，所以我们采用第二种思路，来设计了我们的方法。

研究动机及创新性

火箭发射过程中，初始阶段，助推器和飞行器一同前行，第二阶段，助推器剥离，飞行器独自前进。在我们的框架中，训练阶段，有繁简两个网络一同训练，复杂的网络起到助推器的作用，通过参数共享和信息提供推动轻量网络更好的训练；在预测阶段，助推器网络脱离系统，轻量网络独自发挥作用，从而在不增加预测开销的情况下，提高预测效果。整个过程与火箭发射类似，所以我们命名该系统为“火箭发射”。

训练方式创新

我们框架的创新在于它新颖的训练方式：

1. 繁简两个模型协同训练，协同训练有以下好处：
    a) 一方面，缩短总的训练时间：相比传统 teacer-student 范式中，teacher 网络和 student 网络先后分别训练，我们的协同训练过程减少了总的训练时间，这对在线广告系统这样，每天获得大量训练数据，不断更新模型的场景十分有用。
    b) 另一方面，助推器网络全程提供 soft target 信息给轻量网络，从而达到指导轻量网络整个求解过程的目的，使得我们的方法，相比传统方法，获得了更多的指导信息，从而取得更好的效果。
2. 采用梯度固定技术：训练阶段，限制两网络 soft target 相近的 loss，只用于轻量网络的梯度更新，而不更新助推器网络，从而使得助推器网络不受轻量网络的影响，只从真实标记中学习信息。这一技术，使得助推器网络拥有更强

的自由度来学习更好的模型，而助推器网络效果的提升，也会提升轻量网络的训练效果。

结构创新

助推器网络和轻量网络共享部分层的参数，共享的参数可以根据网络结构的变化而变化。一般情况下，两网络可以共享低层。在神经网络中，低层可以用来学习信息表示，低层网络的共享，可以帮助轻量网络获得更好的信息表示能力。

方法框架：

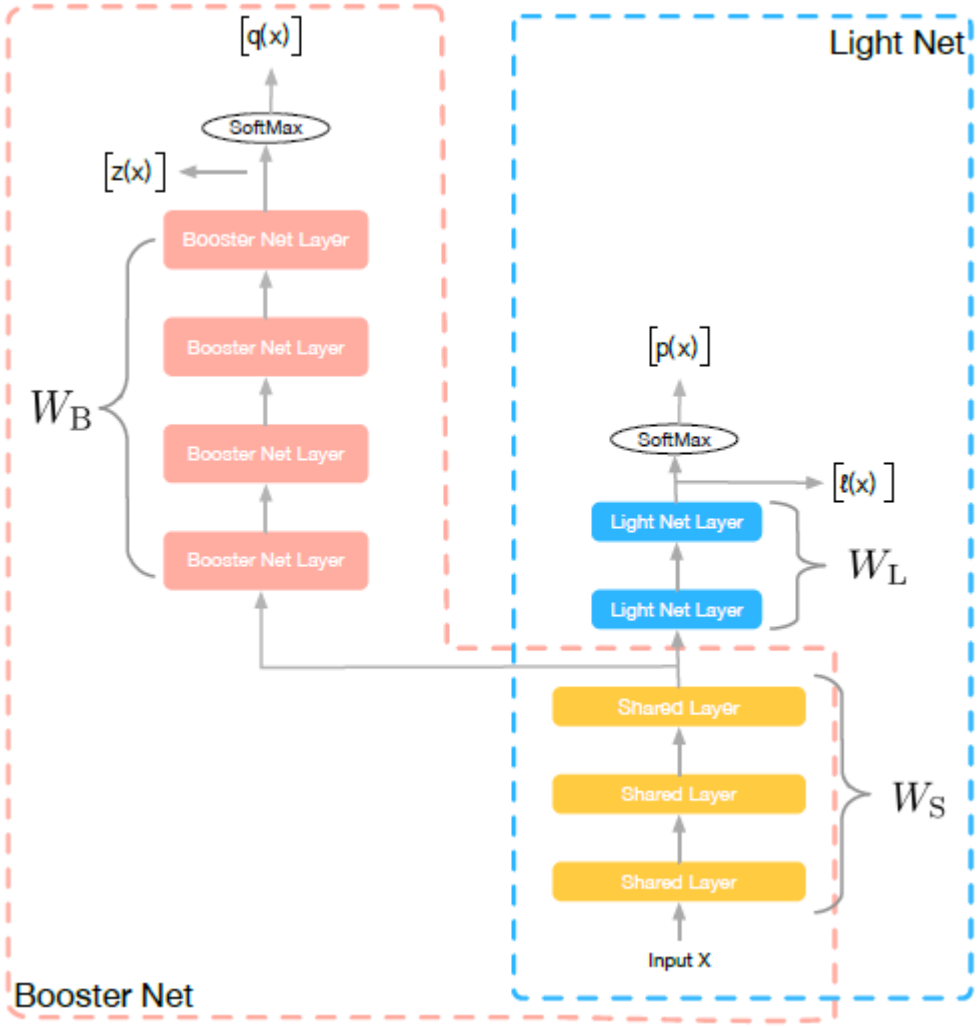


图 1：网络结构

如图 1 所示，训练阶段，我们同时学习两个网络：Light Net 和 Booster Net, 两个网络共享部分信息。我们把大部分的模型理解为表示层学习和判别层学习，表示层学习的是对输入信息做一些高阶处理，而判别层则是和当前子 task 目标相关的学习，我们认为表示层的学习是可以共享的，如 multi task learning 中的思路。所以在我们的方法里，共享的信息为底层参数（如图像领域的前几个卷积层，NLP 中的 embedding）， 这些底层参数能一定程度上反应了对输入信息的基本刻画。

整个训练过程，网络的 loss 如下：

$$\begin{aligned}\mathcal{L}(\boldsymbol{x};\mathbf{W}_{\mathrm{S}},\mathbf{W}_{\mathrm{L}},\mathbf{W}_{\mathrm{B}}) = \mathcal{H}(\boldsymbol{y},\boldsymbol{p}(\boldsymbol{x})) + \mathcal{H}(\boldsymbol{y},\boldsymbol{q}(\boldsymbol{x})) \\ + \lambda\|\boldsymbol{l}(\boldsymbol{x}) - \boldsymbol{z}(\boldsymbol{x})\|_2^2,\end{aligned}$$

Loss 包含三部分：第一项，为 light net 对 ground truth 的学习，第二项，为 booster net 对 ground truth 的学习，第三项，为两个网络 softmax 之前的 logits 的均方误差（MSE），该项作为 hint loss, 用来使两个网络学习得到的 logits 尽量相似。

Co-Training

两个网络一起训练，从而 booster net 会全程监督轻量网络的学习，一定程度上，booster net 指导了 light net 整个求解过程，这与一般的 teacher-student 范式下，学习好大模型，仅用大模型固定的输出作为 soft target 来监督小网络的学习有着明显区别，因为 booster net 的每一次迭代输出 虽然不能保证对应一个和 label 非常接近的预测值，但是到达这个解之后有利于找到最终收敛的解 。

Hint Loss

Hint Loss 这一项在 SNN-MIMIC 中采用的是和我们一致的对 softmax 之前的 logits 做 L2 Loss:

$$\mathcal{L}_{\mathrm{MSE}}(\boldsymbol{x}) = \|\boldsymbol{p}(\boldsymbol{x}) - \boldsymbol{q}(\boldsymbol{x})\|_2^2$$

Hinton 的 KD 方法是在 softmax 之后做 KL 散度，同时加入了一个 RL 领域常用的超参 temperature T:

$$\mathcal{L}_{\text{mimic}}(\boldsymbol{x}) = \|\boldsymbol{l}(\boldsymbol{x}) - \boldsymbol{z}(\boldsymbol{x})\|_2^2$$

也有一个半监督的工作再 softmax 之后接 L2 Loss:

$$\mathcal{L}_{\text{KD}}(\boldsymbol{x}) = \mathcal{H}(\frac{\boldsymbol{p}(\boldsymbol{x})}{T}, \frac{\boldsymbol{q}(\boldsymbol{x})}{T})$$

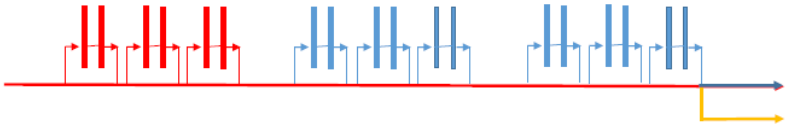
已有的文献没有给出一个合理的解释为什么要用这个 Loss，而是仅仅给出实验结果说明这个 Loss 在他们的方法中表现得好。KD 的 paper 中提出在 T 足够大的情况下，KD 的 Loss $\mathrm{L}_{KD}(x)$ 是等价于 $\mathrm{L}_{mimic}(x)$ 的。我们在论文里做了一个稍微细致的推导，发现这个假设 T 足够大使得 $\mathrm{L}_{KD}(x) = \mathrm{L}_{mimic}(x)$ 成立的情况下，梯度也是一个无穷小，没有意义了。同时我们在 paper 的 appendix 里 在一些假设下 我们从最大似然的角度证明了 $\mathrm{L}_{mimic}(x)$ 的合理性。

Gradient Block

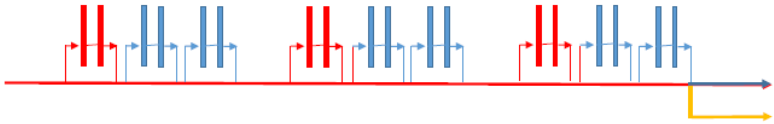
由于 booster net 有更多的参数，有更强的拟合能力，我们需要给他更大的自由度来学习，尽量减少小网络对他的拖累，我们提出了 gradient block 的技术，该技术的目的是，在第三项 hint loss 进行梯度回传时，我们固定 booster net 独有的参数 $\mathbf{W}_b$ 不更新，让该时刻，大网络前向传递得到的 $z(x)$，来监督小网络的学习，从而使得小网络向大网络靠近。

实验结果

实验方面，我们验证了方法中各个子部分的必要性。同时在公开数据集上，我们还与几个 teacher-student 方法进行对比，包括 Knowledge Distillation(KD), Attention Transfer(AT)。为了与目前效果出色的 AT 进行公平比较，我们采用了和他们一致的网络结构宽残差网络（WRN）。实验网络结构如下：



(a) bottom rocket net on wide residual net



(b) interval rocket net on wide residual net

图 2：实验所用网络结构

红色+黄色表示 light net，蓝色+红色表示 booster net。(a)表示两个网络共享最底层的 block，符合我们一般的共享结构的设计。(b)表示两网络共享每个 group 最底层的 block，该种共享方式和 AT 在每个 group 之后进行 attention transfer 的概念一致。

各创新点的效果

我们通过各种对比实验，验证了参数共享和梯度固定都能带来效果的提升

light	booster	base	rocket (no GB)[1]	rocket (no sharing)[2]	rocket (no joint training)[3]	rocket
WRN-16-1(b)	WRN-40-1	8.77	8.50	8.06	8.04	**7.87**
WRN-16-1(a)	WRN-40-1	8.69	8.30	8.23	8.23	**7.85**

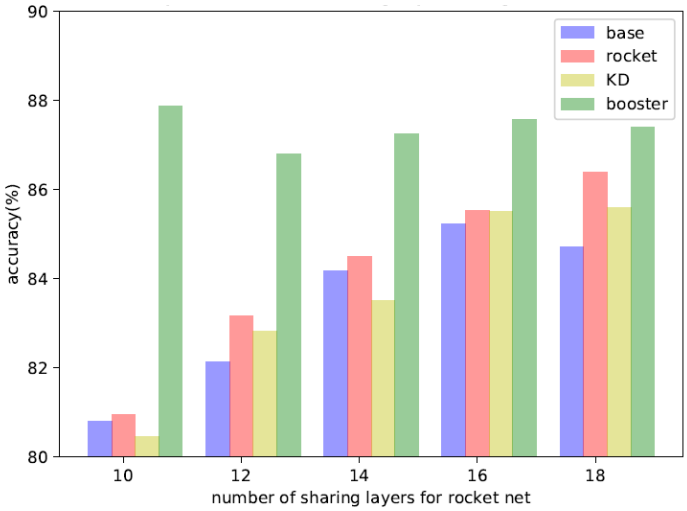
[1] rocket (no GB) means rocket launching without gradient block.
[2] rocket (no sharing) means rocket launching without parameter sharing. [3] rocket (no joint training) means booster net trains first, then light net use some layers of booster to initialize, and use hint loss to learn booster net's logits.

各种 LOSS 效果比较

light	booster	$\mathcal{L}_{\text{mimic}}$	$\mathcal{L}_{\text{MSE}}$	$\mathcal{L}_{\text{KD}}$
WRN-16-1 (b)	WRN-40-1	7.87	8.32	**7.98**
WRN-16-1 (a)	WER-40-1	7.85	8.36	**8.26**

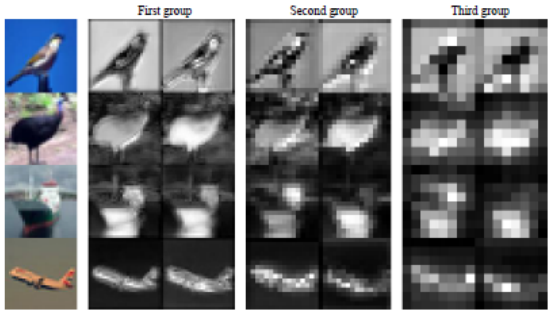
轻量网络层数变化效果图

固定 booster net, 改变 light net 的层数，rocket launching 始终取得比 KD 要好的表现，这表明，light net 始终能从 booster net 中获取有价值的信息。

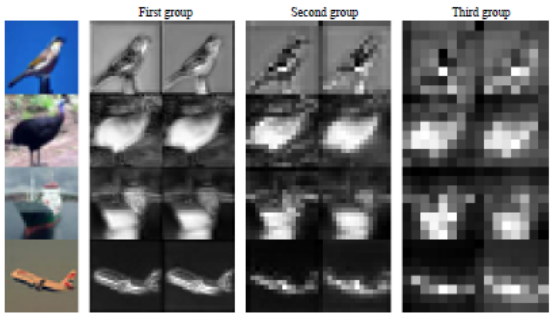


可视化效果

通过可视化实验，我们观察到，通过我们的方法，light net 能学到 booster net 的底层 group 的特征表示。



(a) different group's visualization result on attention transfer



(b) different group's visualization result on rocket launching

公开数据集效果比较

除了自身方法效果的验证，在公开数据集上，我们也进行了几组实验。

在 CIFAR-10 上，我们尝试不同的网络结构和参数共享方式，我们的方法均显著优于已有的 teacher-student 的方法。在多数实验设置下，我们的方法叠加 KD，效果会进一步提升

light	booster	base	rocket (no GB)[1]	rocket (no sharing)[2]	rocket (no joint training)[3]	rocket
WRN-16-1(b)	WRN-40-1	8.77	8.50	8.06	8.04	**7.87**
WRN-16-1(a)	WRN-40-1	8.69	8.30	8.23	8.23	**7.85**

[1] rocket (no GB) means rocket launching without gradient block.
[2] rocket (no sharing) means rocket launching without parameter sharing. [3] rocket (no joint training) means booster net trains first, then light net use some layers of booster to initialize, and use hint loss to learn booster net's logits.

这里 WRN-16-1,0.2M 表示 wide residual net，深度为 16，宽度为 1，参数量为 0.2M.

同时在 CIFAR-100 和 SVHN 上，取得了同样优异的表现

dataset	light	booster	base	AT	KD	rocket	rocket+KD
SVHN	WRN-16-1, 0.2M(b)	WRN-40-1, 0.6M	3.58	2.99	2.31	2.29	**2.20**
CIFAR-100	WRN-16-1, 0.2M(b)	WRN-40-1, 0.6M	43.7	34.1	36.4	33.3	**33.0**

真实应用

同时，在阿里展示广告数据集上，我们的方法，相比单纯跑 light net，可以将 GAUC 提升 0.3%.

我们的线上模型在后面的全连接层只要把参数量和深度同时调大，就能有一个提高，但是在线的时候有很大一部分的计算耗时消耗在全连接层(embedding 只是一个取操作，耗时随参数量增加并不明显)，所以后端一个深而宽的模型直接上线压力会比较大。表格里列出了我们的模型参数对比以及离线的效果对比：

model	# parameters in FC layers	# multiplications in FC layers	inference time of FC Layers	GAUC
base	$576 \times 200 \times 80 \times 2$	131360	7.6 ms	0.632
rocket	$576 \times 200 \times 80 \times 2$	131360	7.6 ms	0.635
booster only	$576 \times 720 \times 360 \times 240 \times 180 \times 90 \times 2$	837900	23.2 ms	0.637

总结

在线响应时间对在线系统至关重要。本文提出的火箭发射式训练框架，在不提高预测时间的前提下，提高了模型的预测效果。为提高在线响应模型效果提供了新思路。目前 Rocket Launching 的框架为在线 CTR 预估系统弱化在线响应时间限制和模型结构复杂化的矛盾提供了可靠的解决方案，我们的技术可以做到在线计算被压缩 8 倍的情况下性能不变。在日常可以减少我们的在线服务机器资源消耗，双十一这种高峰流量场景更是保障算法技术不降级的可靠方案。