知识蒸馏

- 讲讲故事

  - 青出于蓝而胜于蓝，咱们的故事可以挑战别那么大，出于蓝而近似蓝就好了
  - 看起来有点像本是同根生（没有相煎），那故事的主人公就得是俩模型了
  - 老大很强，老二很弱，那咱们是不是得让老二向老大学习，看看人家咋学的
  - 但是同时老二也不能只向老大学，也得学学标准答案（老大也可能出错）

知识蒸馏

- 何为蒸馏
  - 现在谁家不整个大模型，条件好了吃喝都不差钱了，大模型一般都效果好
  - 但是应用可能麻烦点，费资源，可能下游任务设备一般般，那咋整
  - 那你就用小一点的模型呗，比如resnet152用不了那咱们就用resnet18也行
  - 但是现在咱们要耍无赖，既要用小的18层的也要让他效果尽可能进阶152的

知识蒸馏

- 何为蒸馏

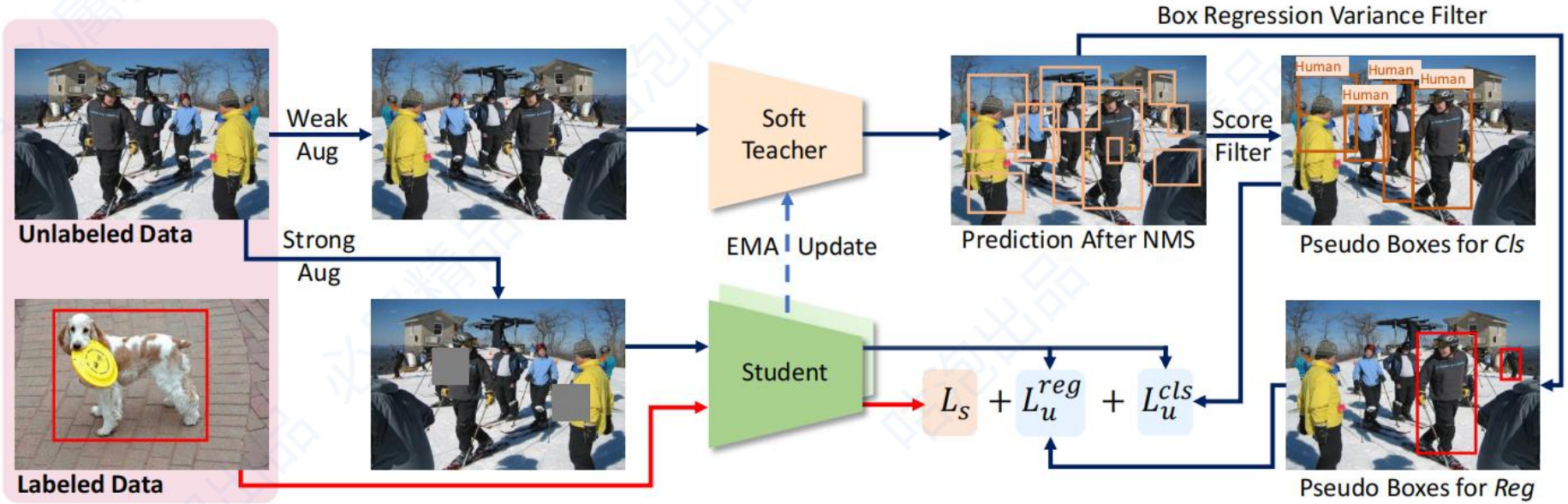
  - 那怎么能达到这个效果呢？这里面就得涉及到一些玄学了（无法证明）
  - 模型参数量越大，效果一定越好吗？不一定，越来越平稳的曲线，有上限
  - 模型的参数量相同，训练策略不同，得到的结果也可能完全不同
  - 那么我们就得想想能不能利用点不同的训练策略让咱们模型既小又好

知识蒸馏

- 蒸馏需要啥呢

- 先来回顾下咱们之前唠过的半监督问题（以物体检测为例）

- T模型生成一些伪标签，然后把这些标签交给S模型来进行学习



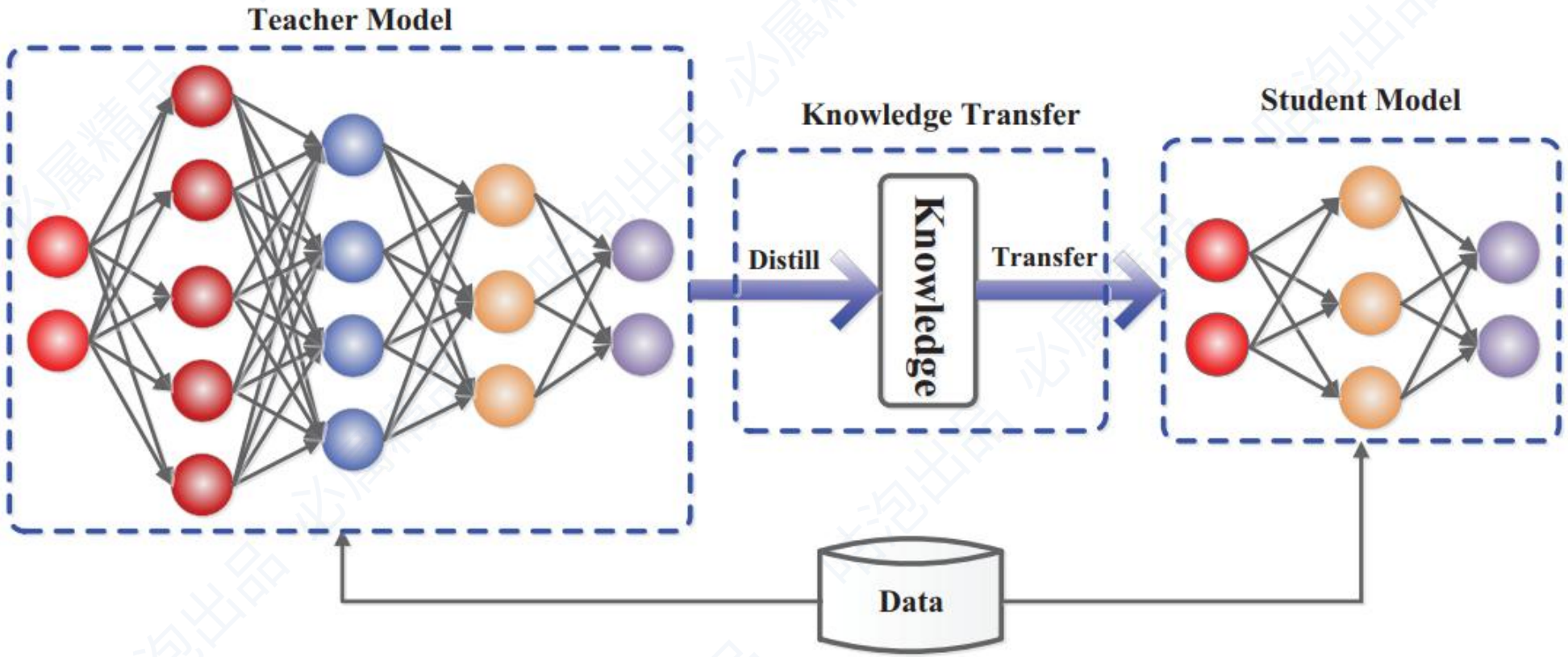
知识蒸馏

- 那么蒸馏要学什么呢？
  - 这里咱们也需要有俩模型（一大一小），可以用不同层数，甚至完全不同
  - 比如resnet不同层数，yolo中s m l x，也可以yolo和mask-rcnn（理论可行）
  - 其中大的模型叫T，它是训练好的，效果也挺好的，然后给它冻住
  - 小的模型叫S，它是要来训练的，得让它学到T里面的知识，尽可能接近T

知识蒸馏

- 基本思想

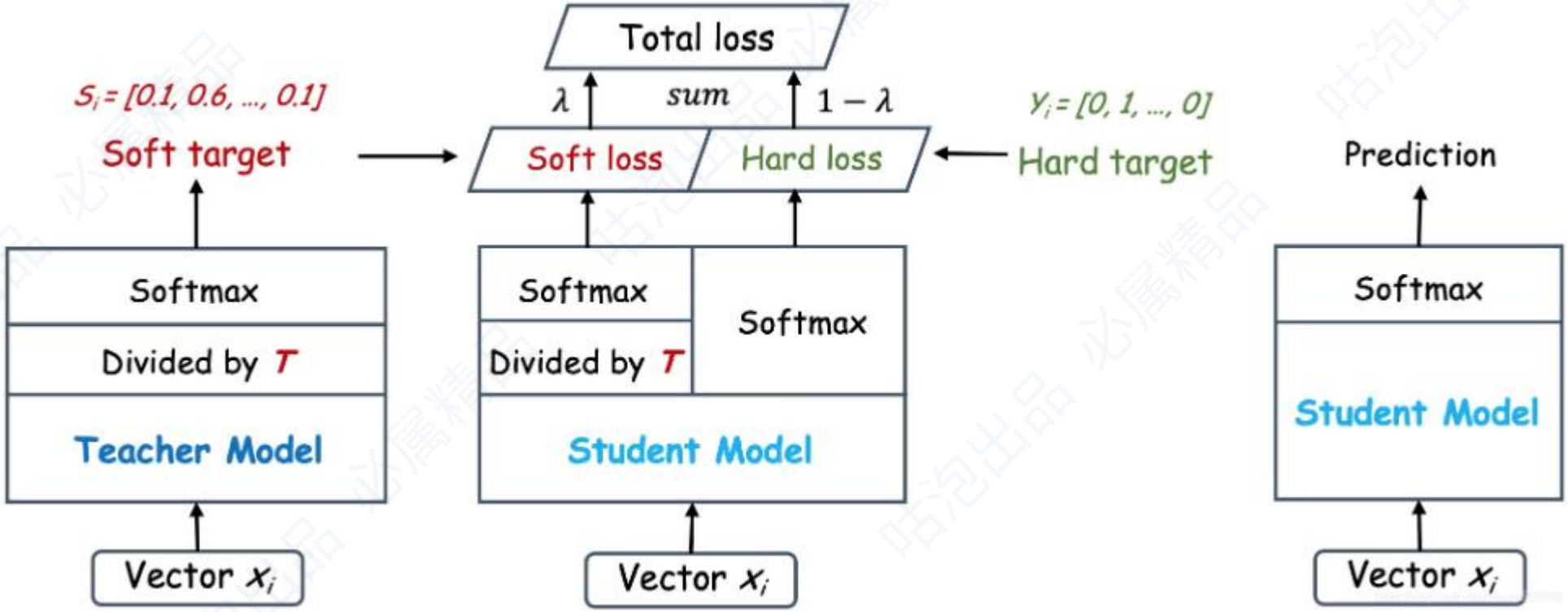
- 大概就是这个图的意思，老师把会的东西通过一种表现形式交给学生



知识蒸馏

- 那么蒸馏要学什么呢？

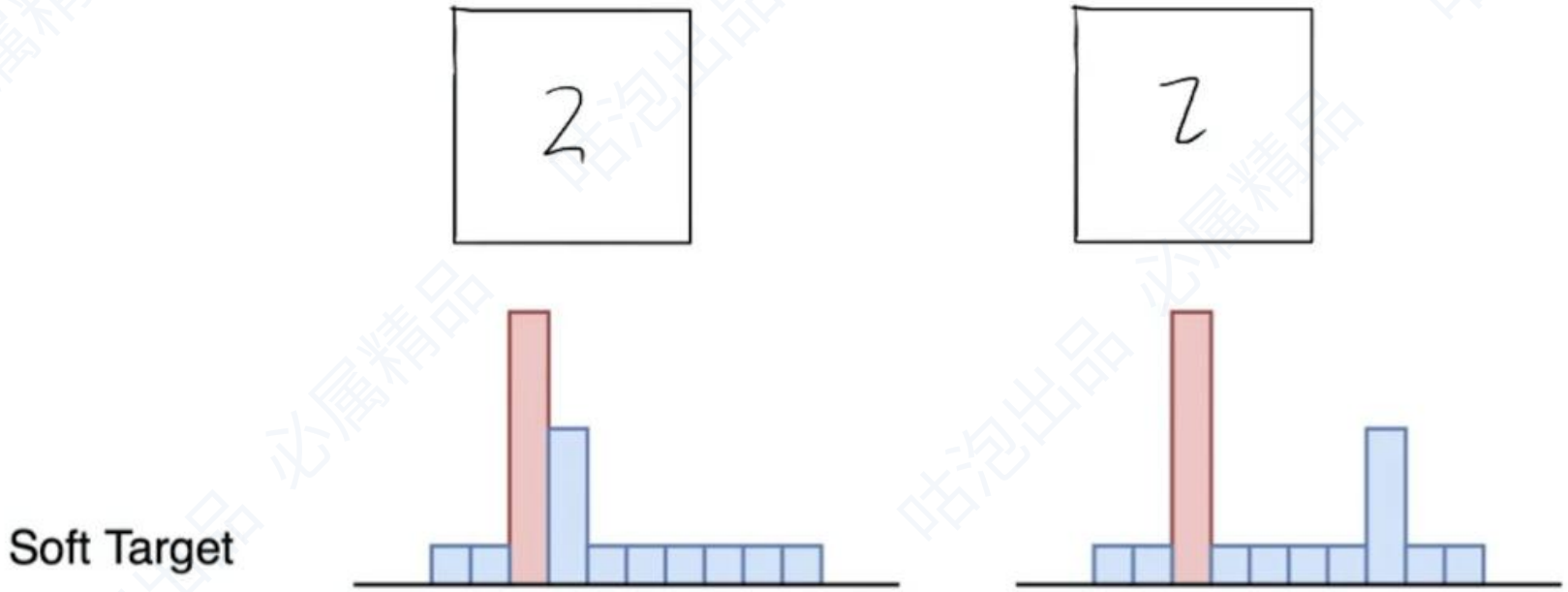
- T模型教S学的不是最终的一个结果，而是解题过程，也就相当于分布



知识蒸馏

- soft target有啥用

  - 左边的2更像3，右边的2更像7，这时候得让模型知道像谁，但是是谁
    （我爸去网吧抓我，回来揍我一顿就hard，给我讲别人家孩子咋咋滴就soft）



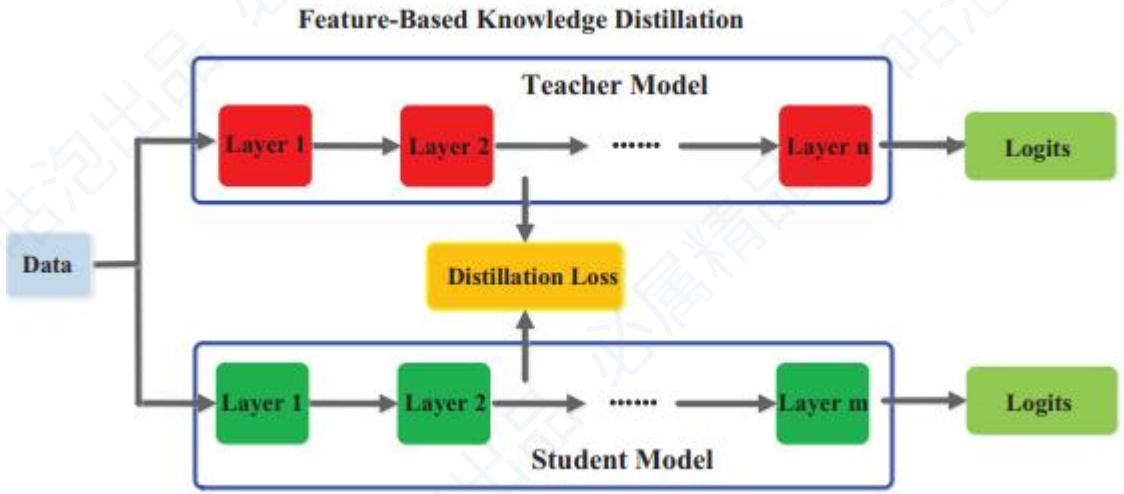
GPT-2

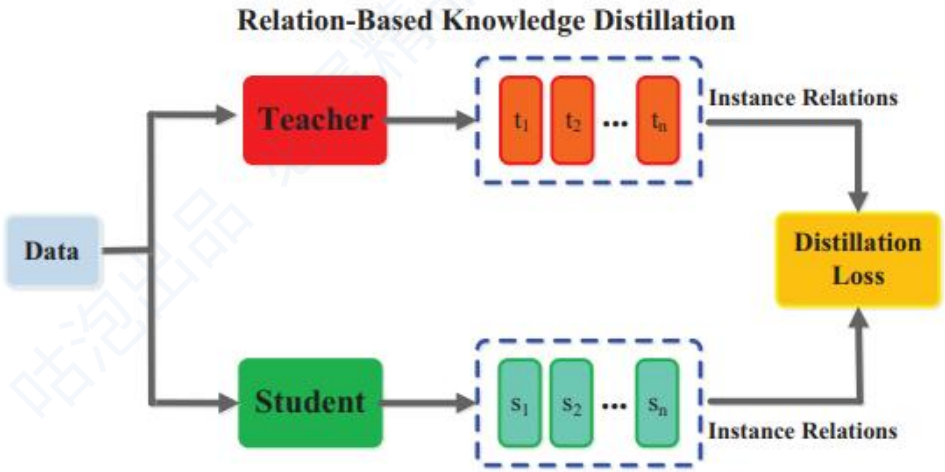
- Temperature的作用
  - 温度就是说对预测结果进行概率重新设计
  - 默认温度为1就相当于还是softmax
  - 温度越高相当于多样性越丰富（雨露均沾）
  - 温度越低相当于越希望得到最准的那个

```
>>> import torch
>>> import torch.nn.functional as F
>>> a = torch.tensor([1,2,3,4.])
>>> F.softmax(a, dim=0)
tensor([0.0321, 0.0871, 0.2369, 0.6439])
>>> F.softmax(a/.5, dim=0)
tensor([0.0021, 0.0158, 0.1171, 0.8650])
>>> F.softmax(a/1.5, dim=0)
tensor([0.0708, 0.1378, 0.2685, 0.5229])
>>> F.softmax(a/1e-6, dim=0)
tensor([0., 0., 0., 1.])
```

知识蒸馏

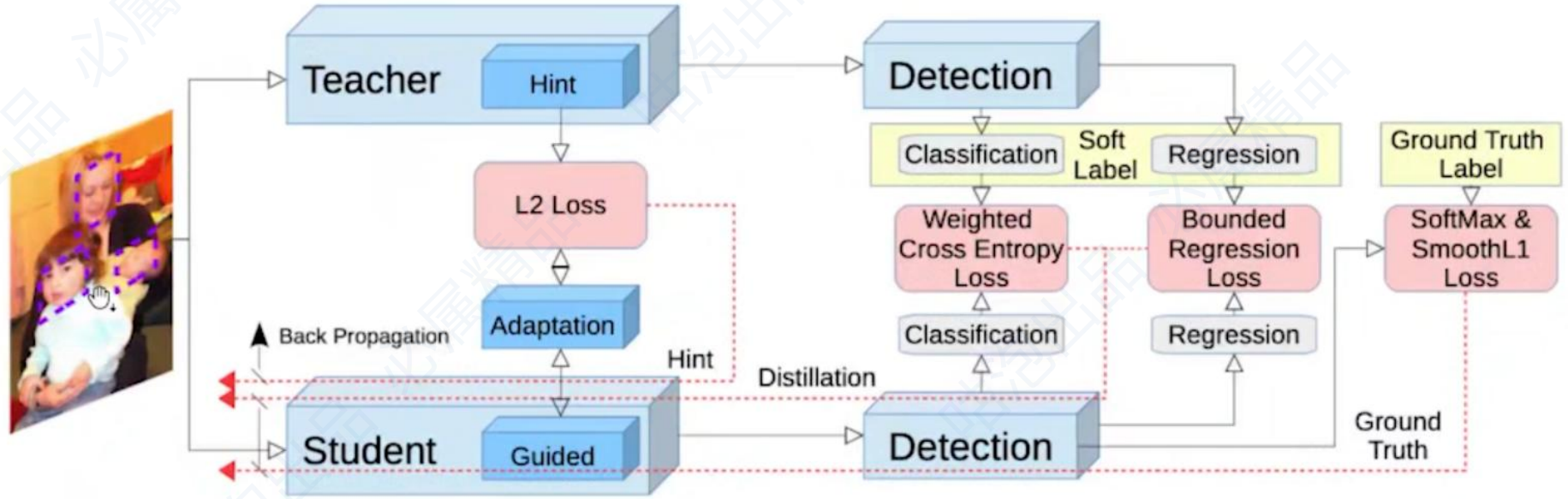
- 不仅可以在输出结果上
  - FeatureBased就是加在特征上，但是一般还需要通过额外卷积或者FC
  - RelationBased就类似对比学习，Batch里面兄弟几个之间的差异也要类似





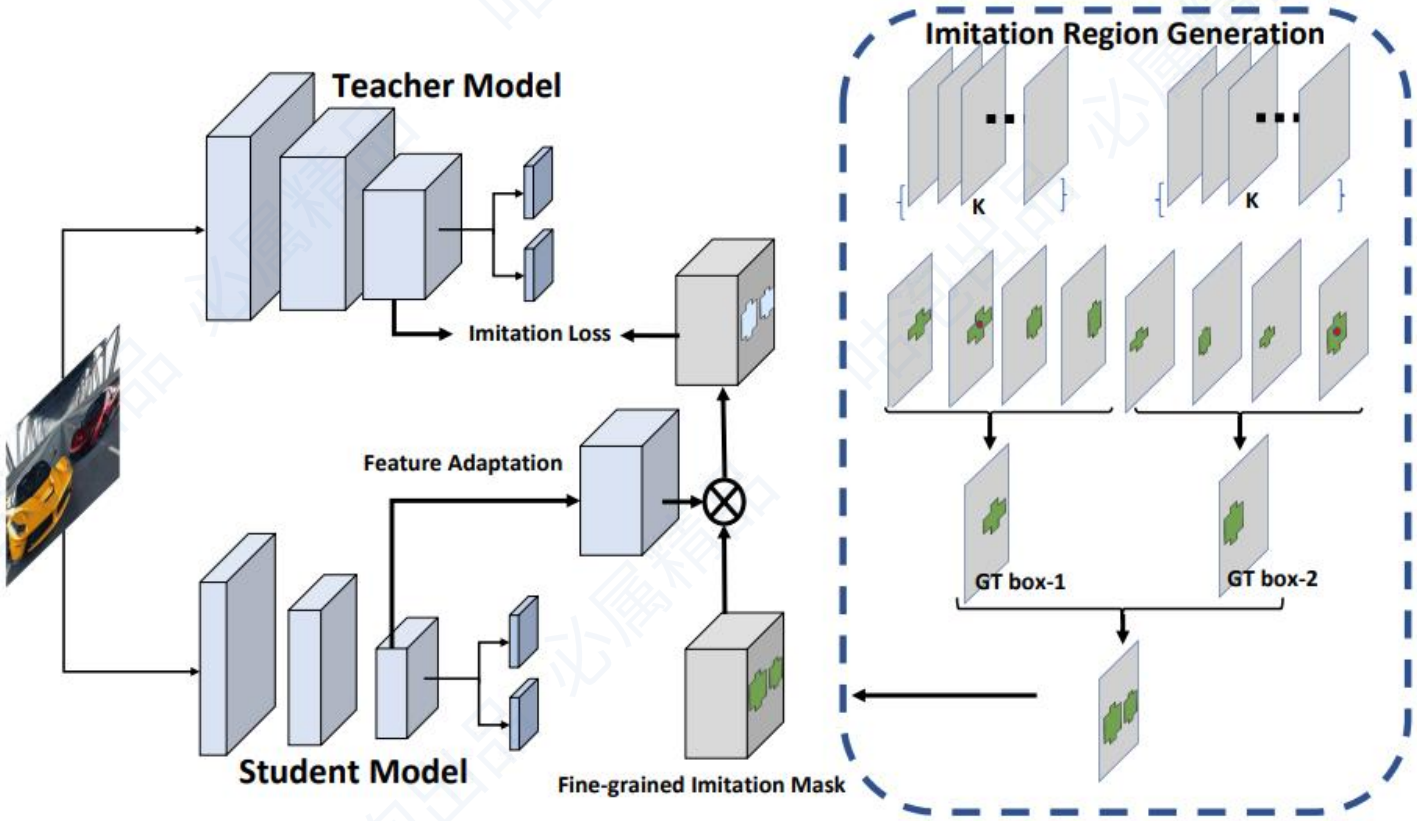
知识蒸馏

- 在物体检测领域应用
  - Backbone上要尽可能一致，分类和回归预测结果也要类似
  - 其实类似半监督任务，T输出伪标签，让S来进行学习



知识蒸馏

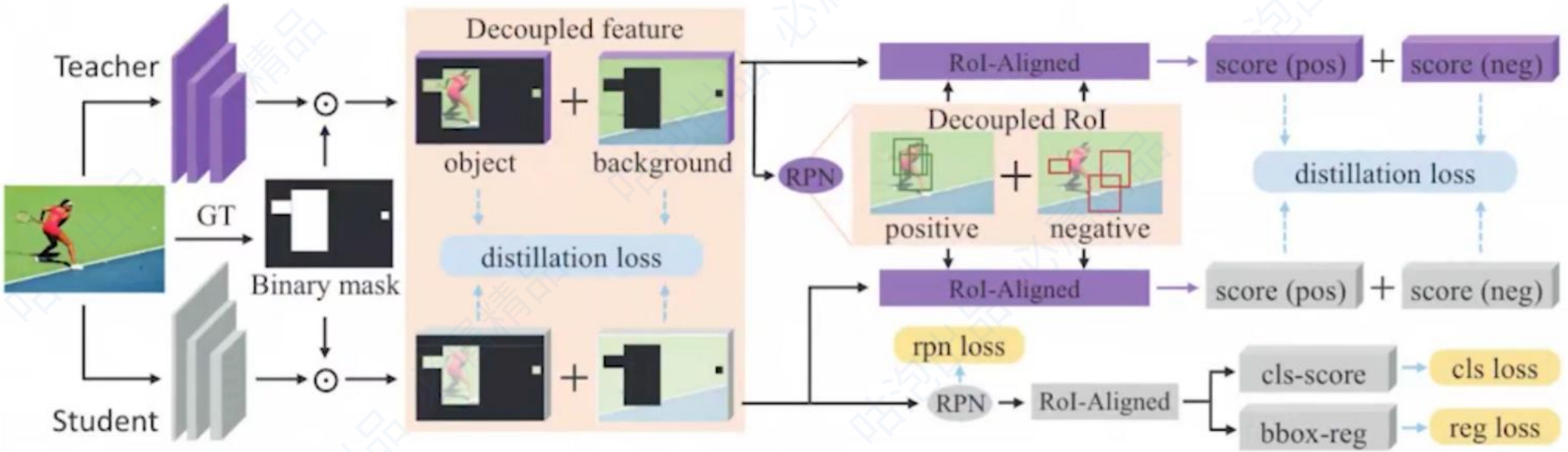
- 在物体检测领域应用
  - backbone如何做的更好?
  - 设计MASK机制区别对待
  - 只取前景区域来进行计算
  - 相当于正样本Anchor参与



知识蒸馏

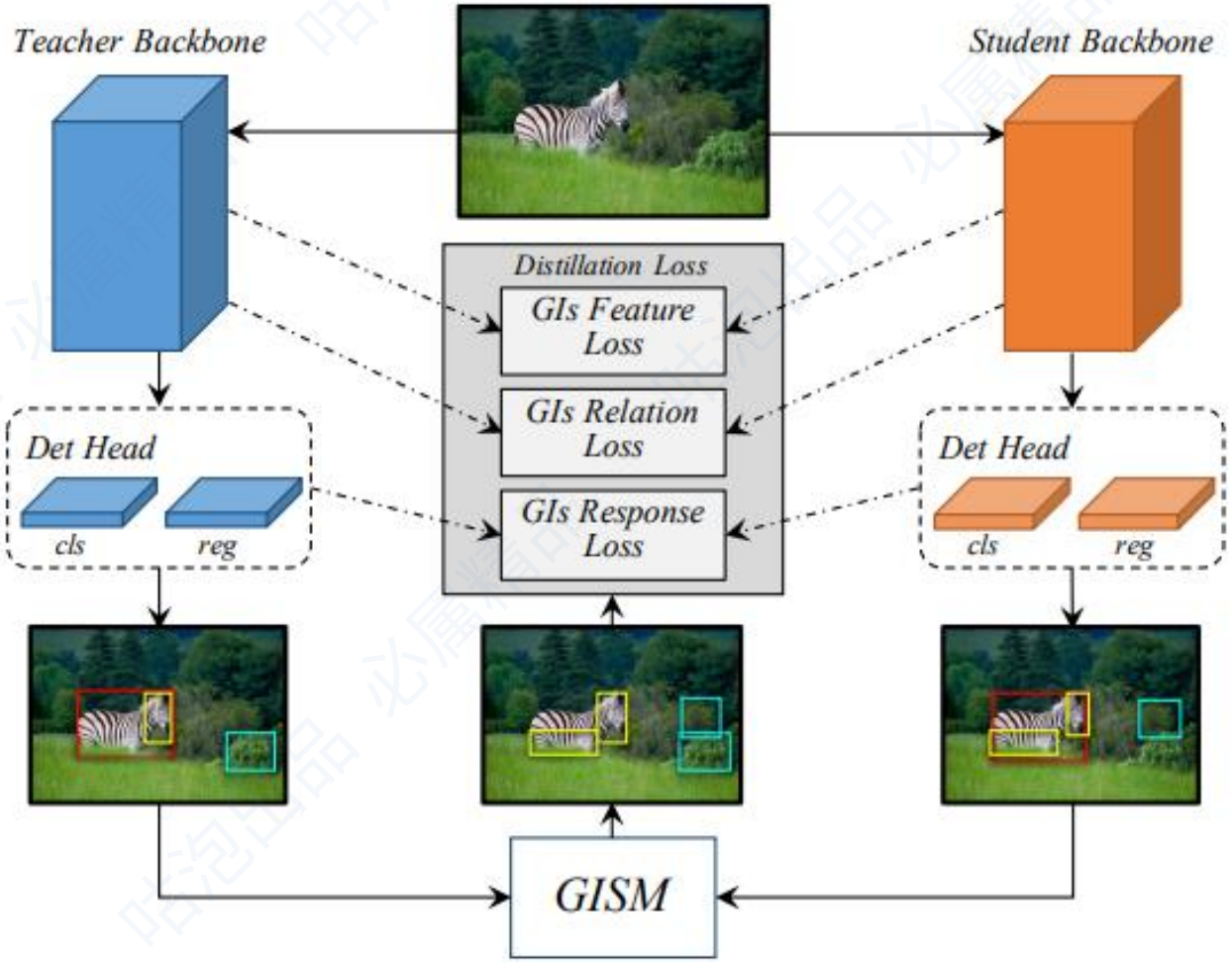
- 在物体检测领域应用

  - 只考虑前景有点不太合适，能不能对背景进行加权呢？（以及正负样本）



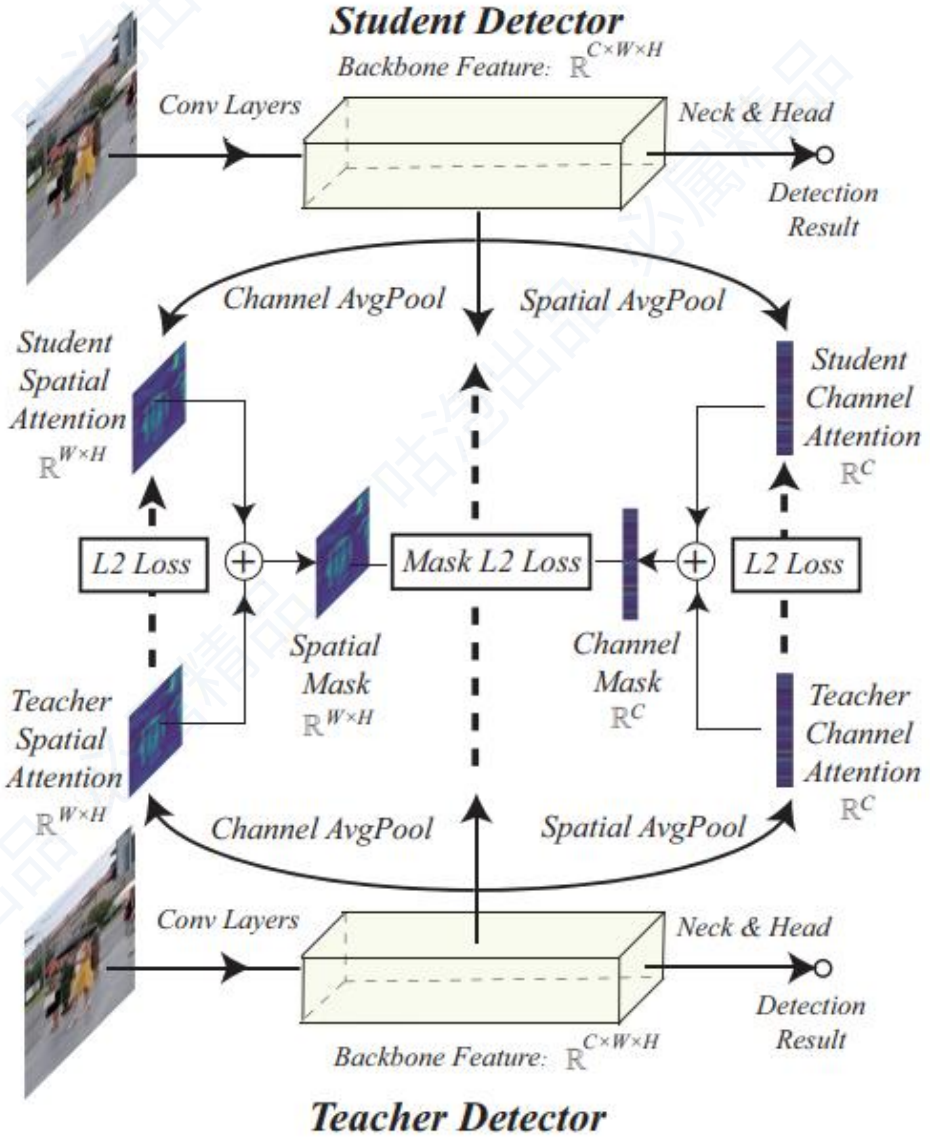
知识蒸馏

- 在物体检测领域应用
  - 重点区域不应该只考虑前背景
  - 差异的地方才是空点考虑的
  - 同样设计MASK机制
  - 这回MASK重点在结果差异上



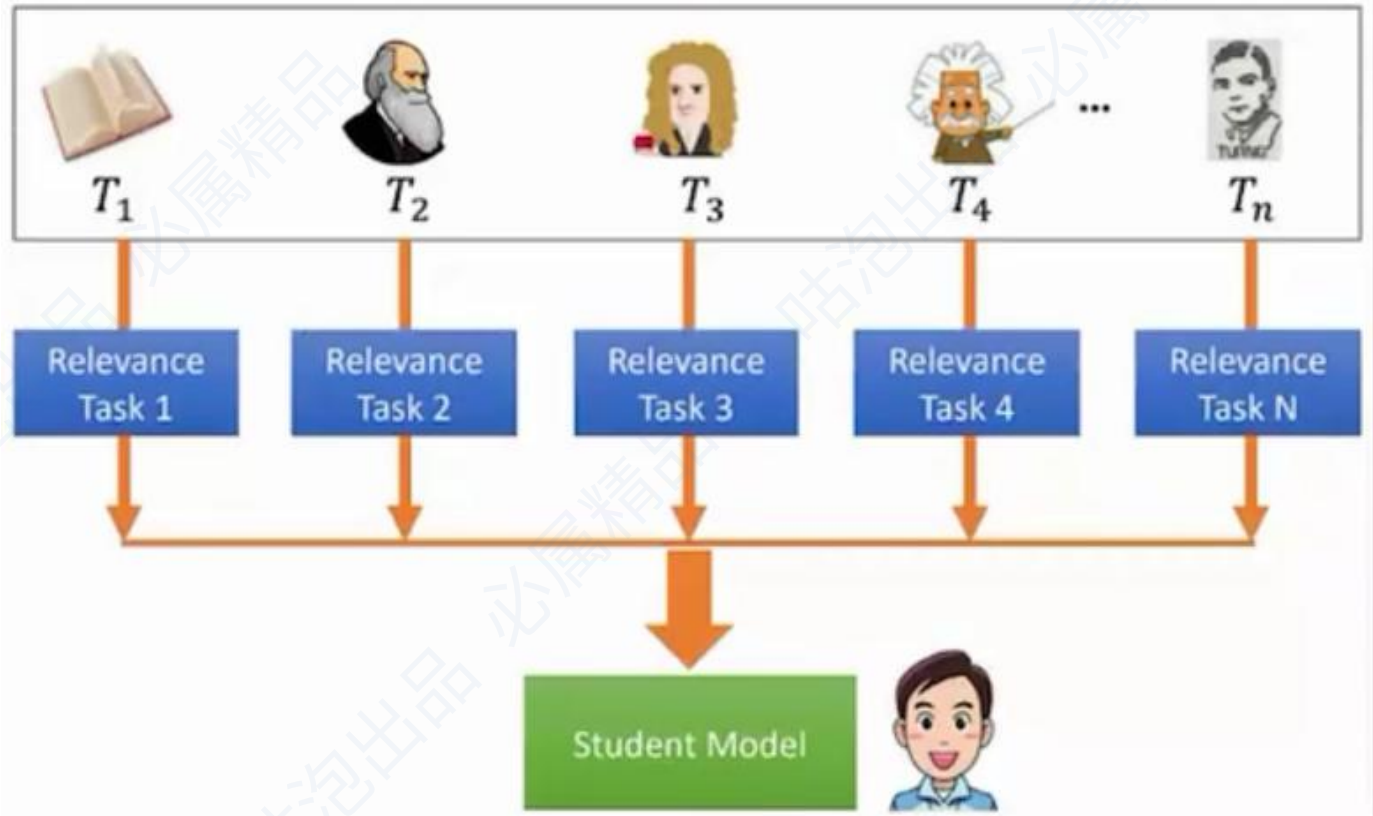
知识蒸馏

- 在物体检测领域应用
  - 人算不如天算，一顿设计一定对吗？
  - 把mask这东西交给Attention来做吧
  - 首先空间和通道维度计算权重，蒸馏下
  - 再将mask结果应用到特征LOSS计算上



知识蒸馏

- 拓展分析-多T RL
  - 一定是一对一教学吗
  - 能不能来多个多对一呢
  - 如何选择不同阶段的T模型呢
  - 设计权重？如何融合呢？



知识蒸馏

- 拓展分析-多T RL

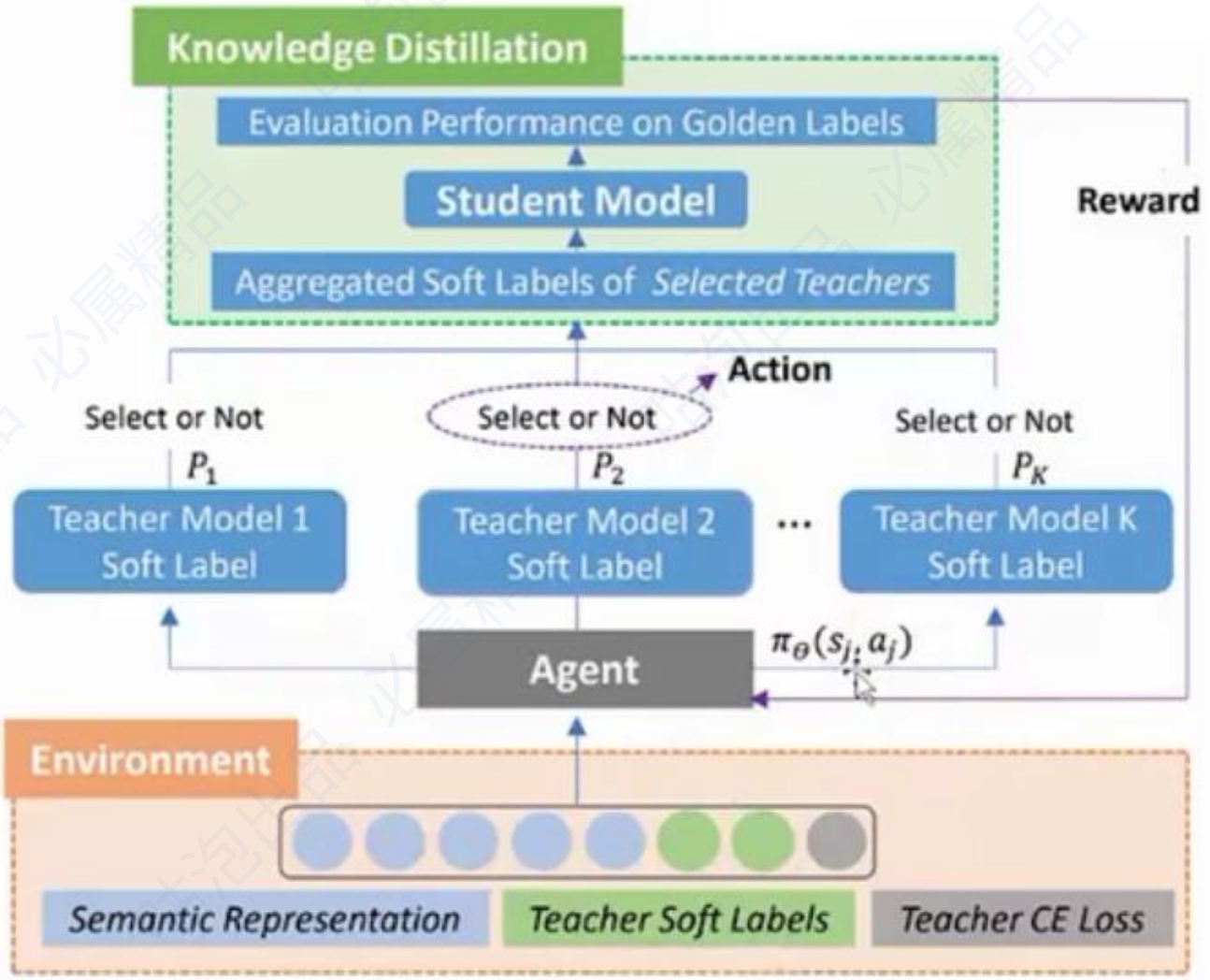
  - 大模型上效果好的，你小模型一定能学的明白吗？

  - 可能在模型学习的过程中也得根据S模型的学习情况来选择路线

Teacher	Student	MRPC	MNLI-mm
		Acc	Acc
BERT-Base	N / A	83.3	83.8
RoBERTa-Base	N / A	88.5	86.8
BERT-Base	$BERT_3$	74.0	75.3
RoBERTa-Base	$BERT_3$	73.0	71.9

知识蒸馏

- 拓展分析-多T RL
  - Agent来学习如何选择策略(老师)
  - Action中0/1就是不选/选择
  - 根据选择情况来让S模型去学
  - 将S的评估结果当作奖励值反馈



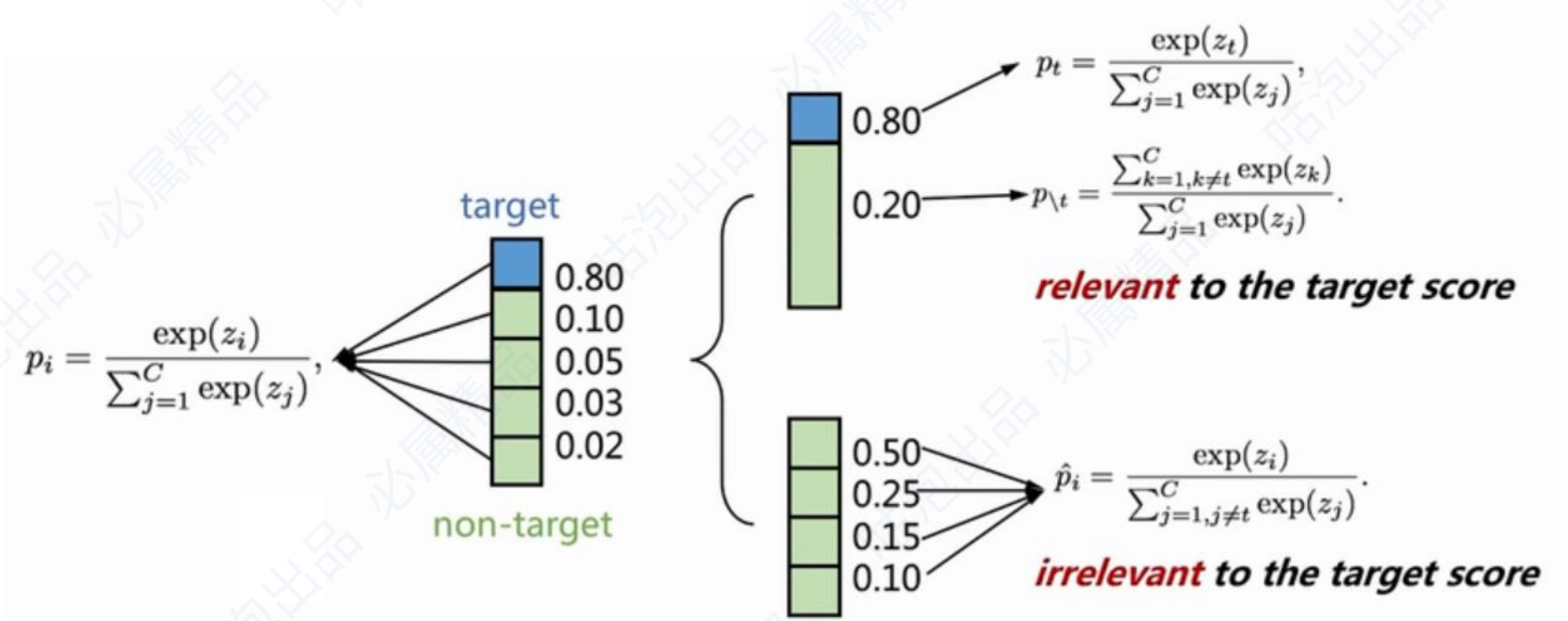
知识蒸馏

- 拓展分析-解耦
  - 你现在喜欢干啥，玩新游戏？认识新朋友？学一项新技术？
  - 怀旧的技术哪个吸引你了最近？曾记否咱们V7说过的make vgg great again
  - 那蒸馏这么多新技术，它也想make谁谁谁great一下子？还真有
  - 咱们刚才都说了结果上(logits)算损失不如特征层面上算损失，难道。。。

知识蒸馏

- 拓展分析-解耦

- T要教给S两件事：1.这件事难度有多大(置信度)2.啥导致了这件事难(错误分布)



知识蒸馏

- 拓展分析-解耦

  - TCKD：Target Class KD(例如左图0.99，右图0.70)；NC：Non-target Class

easy to fit

hard to fit

知识蒸馏

- 拓展分析-解耦
  - 其中TCKD是置信度，NCKD是错误分布
  - 看起来应该是TCKD这哥们来捣乱的
  - 基本结果就是他俩一起用可能还不如一个
  - 他们之间存在着啥矛盾？能不能调节下

student	TCKD	NCKD	top-1	Δ
ResNet32×4 as the teacher				
ResNet8×4			72.50	-
	✓	✓	73.63	+1.13
	✓		68.63	-3.87
		✓	74.26	+1.76
ShuffleNet-V1			70.50	-
	✓	✓	74.29	+3.79
	✓		70.52	+0.02
		✓	74.91	+4.41
WRN-40-2 as the teacher				
WRN-16-2			73.26	-
	✓	✓	74.96	+1.70
	✓		70.96	-2.30
		✓	74.76	+1.50
ShuffleNet-V1			70.50	-
	✓	✓	74.92	+4.42
	✓		70.62	+0.12
		✓	75.12	+4.62

知识蒸馏

- 拓展分析-解耦

  - TCKD适用于难度较大的数据集，越离谱它可能越有用

  - 随机数据增强与随机改变一些标签的情况下是有效果提升的

student	TCKD	top-1	Δ
ResNet8×4		73.82	-
	✓	75.33	+1.51
ShuffleNet-V1		77.13	-
	✓	77.98	+0.85

Table 2. Accuracy(%) on the CIFAR-100 validation. We set ResNet32×4 as the teacher and ResNet8×4 as the student. Both teachers and students are trained with AutoAugment [5].

noisy ratio	TCKD	top-1	Δ
0.1		70.99	-
	✓	70.96	-0.03
0.2		67.55	-
	✓	68.03	+0.48
0.3		64.62	-
	✓	65.26	+0.64

Table 3. Accuracy(%) on the CIFAR-100 validation with different noisy ratios on the training set. We set ResNet32×4 as the teacher and ResNet8×4 as the student.

知识蒸馏

- 拓展分析-解耦

  - 来分析一下，他俩一起为啥会出现问题，矛盾点是什么呢?
  - 预测的准，没利用错误分布
  - 预测不准，利用错误分布
  - 都没预测准，利用有啥用呢

$p_t^{\mathcal{T}} = 0.99$
easy to fit, weight = 0.01

$p_t^{\mathcal{T}} = 0.70$
hard to fit, weight = 0.30

$$\mathrm{KD} = \mathrm{TCKD} + (1 - p_t^{\mathcal{T}})\mathrm{NCKD}.$$

知识蒸馏

- 拓展分析-解耦
  - 好聚好散
  - 分别权重
  - 各自损失
  - 最后叠加