论文解读：深度网络模型加速与压缩

王栋

2017-11-20

1 知识蒸馏

知识蒸馏（knowledge distillation），一个通俗的解释就是：针对某一个学习任务，假定现在已经根据数据学习得到了一个训练好了的模型（teacher model），现在要求结合原始训练数据或者别的数据，从该 teacher model中学习得到一个学生模型（student model），同时要求该student model可以在某些指标上优于其teacher model（当然，作为trade-off，该student model可能会在别的指标上略逊色于其teacher model），这样的一个学习得到student model的过程就叫做知识蒸馏。现在对于知识蒸馏的研究，仅仅停留在图像分类的任务上，绝大多数的实验都是基于ImageNet、cifar-10、cifar-100、MNIST等等数据集展开的，所以也就可见，知识蒸馏这种模型压缩技术最大的一个弊端便是应用面过于局限——目前仅仅只有将其应用到图像分类的研究，至于是否能将其迁移到回归任务、半监督等等学习任务，还有待进一步的研究。

1.1 NIPS2014: Do Deep Nets Really Need to be Deep?

motivation 该文章对于 Do Deep Nets Really Need to be Deep 这样一个open的问题展开了工作，实际上的实验思路是：通过知识蒸馏的方式，得到一个较浅的student model，然后再对该student model进行评估，看是否得到了与teacher model而言comparable的效果，那么自然而然地，如果浅层的student model接近甚至超过teacher model，便提供了网络未必真的需要很深的证据。本文工作提供了经验性的证据：浅的网络也有能力学习到深层网络所学习得到的函数，在有些情况下，这些浅层的网络和较深的网络有些相同数量的参数（浅层网络与深层网络具有相同的数量，那么势必得比深层的网络要宽）。

In this paper we provide empirical evidence that shallow nets are capable of learning the same function as deep nets, and in some cases with the same number of parameters as the deep nets.

method

1. Mimic Learning via Regressing Logits with L2 Loss：从针对分类任务的teacher model中，让student model通过回归任务来学习参数。一般情况下，对于分类任务，都是使用softmax形式的输出，结合cross-entropy来构建loss（teacher model采用这种形式），而student model的学习，是将teacher model的logits输

出作为回归目标使用L2来构建loss，其中logits指的是teacher model中输出的，在未输入softmax函数之前的数据，公式上表示则有

2. Speeding-up Mimic Learning by Introducing a Linear Layer: 当在student model与teacher model有差不多相同数量参数的情况下进行对比试验的时候，由于student model层数较少，所以student model的隐层必然比较宽，也就是输出神经元较多，也就遇到这样的一个问题：student model的隐层参数数量较大，是一个极大的矩阵。虽然这个问题不是致命的，也可以学习得到较好的student model，但是模型的收敛速度很慢。针对上述问题，可以通过矩阵的低秩分解来加速学习，同时减少参数的数量。公式上表示则有：

Q & A

- Q: 纵然teacher model已经是训练完成了的，但是依旧无法避免其面对有些样本，不能work，i.e,分类错误，那么将这样的"标注"不正确的样本喂给student model不是不合逻辑吗?

  A: 事实上，知识蒸馏的一个根本目的是让student model使用尽可能少的参数，去尽可能地学习teacher model已经学习得到的函数，也就是说让student model尽可能地去模拟(mimic) teacher model，所以teacher model预测错误，那么让student model学习这种错误的预测。

- Q: 相比在原始数据上直接训练student model，为什么这种基于知识蒸馏的训练模式会得到准确考更高的student model?

  A: 模型压缩一定程度上起到了一个正则化的作用；而如果直接在原始数据上学习student model，一般过拟合的情况则会比较严重

conclusion 该论文从实验的角度来讲，更多的是验证性实验，对"深度网络的深度对结果的影响"展开了研究，模型压缩不过是其实验过程中附带出来的方法，或者说实验的方法本质上是模型压缩，但是该实验的本质目的并非是研究模型压缩。该文章通过验证一个参数较少的、较浅的student model可以通过知识蒸馏的方式来学习得到teacher model所学到的函数（但是可能学习起来比较困难），来间接地证明了深度网络中深度并非是必须的，但是或许可以让学习过程变得容易一点。（不过，不得不说的，在ResNet中，又提出了随着深度的不断增加，深度网络训练变得越来越难）

1.2 Distilling the Knowledge in a Neural Network

motivation 已经有工作演示了可以将一个繁杂的集成系统中的知识进行压缩，来得到一个更容易部署的小模型。所以，该文章旨在使用一种不同的压缩方法来进一步发展这种知识蒸馏的压缩方法。其中，所谓的"知识"这一抽象概念，在脱离了任何实例的情况下，本质上就是一个学习好了的从输入向量到输出向量的映射。

Caruana and his collaborators have shown that it is possible to compress the knowledge in an ensemble into a single model which is much easier to deploy and we develop this approach further using a different compression technique.

method　回顾1.1中的方法：结合logits与L2来构建loss训练student model，而本文则是使用**soft targets**来构建loss，其中soft targets指的是logits经过softmax函数处理后得到的概率形式，同时，又引入超参数**T（温度）**来使得soft targets更加soft，公式上的表示则有

$$f(x) = x^2$$

其中，xxxxxxxxxxxxxxxx。对于训练样本，可以使用原始的训练数据，也可以使用分离的"transfer set"，但是当使用原始训练数据时，除了teacher model提供的soft targets，还可以提供更多的信息——ground-trth。所以，直观上来看，我们可以使用两种目标函数的加权平均来构建全局的loss。第一种目标函数就是构建student model输出的soft与teacher model输出的soft targets之间cross-entropy loss，其中两者的产生都需要相同的T（温度）；第二种目标函数则是student model输出的soft与原始训练数据集中的ground-truth之间cross-entropy loss。

When the correct labels are known for all or some of the transfer set, this method can be significantly improved by also training the distilled model to produce the correct labels. One way to do this is to use the correct labels to modify the soft targets, but we found that a better way is to simply use a weighted average of two different objective functions. The first objective function is the cross entropy with the soft targets and this cross entropy is computed using the same high temperature in the softmax of the distilled model as was used for generating the soft targets from the cumbersome model. The second objective function is the cross entropy with the correct labels. This is computed using exactly the same logits in softmax of the distilled model but at a temperature of 1. We found that the best results were generally obtained by using a condiderably lower weight on the second objective function. Since the magnitudes of the gradients produced by the soft targets scale as 1/T 2 it is important to multiply them by T 2 when using both hard and soft targets. This ensures that the relative contributions of the hard and soft targets remain roughly unchanged if the temperature used for distillation is changed while experimenting with meta-parameters

对于训练集中的一个样本，loss构建

Q & A

- Q: 既然引入了T，那么teacher model要与student model一起训练吗?

  A: 不需要，对于teacher model中的softmax，直接引入T改变原本的输出就可以，并未引入可学习参数。

- Q: T的引入有什么作用?

  A: 可以带来更多的信息;

conclusion　不使用0 1 这种类别标注的hard targets，而是使用soft targets，这样可以带来一定的正则化效果。