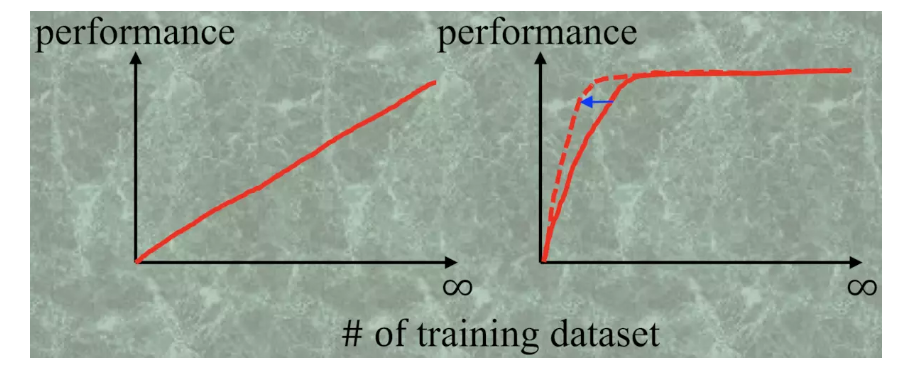
问题1：深度学习真的是数据越多越好吗？  
经过大量的测试我们发现，实际情况并不是像我们想象的这样（左图）而是如下右图所示存在一个临界值，当数据量超过临界值时模型的表现随着数据量的增多并没有太多的提升



问题2：既然如此如何快速的寻找临界值呢？

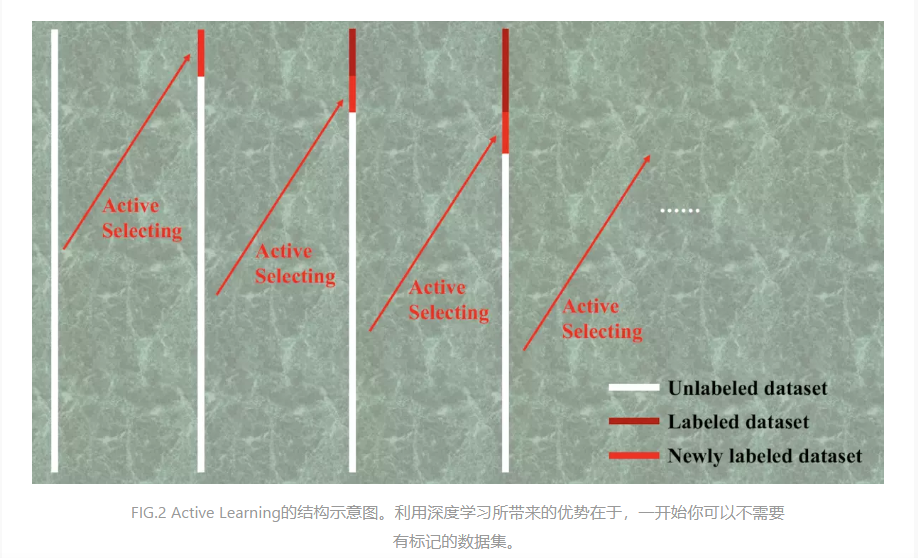
解决方案就是主动学习（Active Learning），去主动学习那些比较“难的”，“信息量大的”样本（hard mining）。关键点是每次都挑当前分类器分类效果不理想的那些样本（hard sample）给它训练，假设是训练这部分hard sample对于提升分类器效果最有效而快速。问题是在不知道真正标签的情况下怎么去定义HARD sample？或者说怎么去描述当前分类器对于不同样本的分类结果的好坏？

论文《[Fine-tuning Convolutional Neural Networks for Biomedical Image Analysis: Actively and Incrementally.](http://openaccess.thecvf.com/content_cvpr_2017/papers/Zhou_Fine-Tuning_Convolutional_Neural_CVPR_2017_paper.pdf)》给出了一个解决方案

由于深度学习的输出是属于某一类的概率（0～1），一个很直观的方法就是用“熵（entropy）”来刻画信息量，把那些预测值模棱两可的样本挑出来，对于二分类问题，就是预测值越靠近0.5，它们的信息量越大。还有一个比较直观的方法是用“多样性（diversity）”（这里所指的多样性与UDA利用无标签数据进行训练的方式十分相似）来刻画labeled data和unlabeled data的相似性

问题3：如何使用主动训练进行训练呢？

使用主动学习进行训练大体的流程如下图:



问题4：多样性具体的含义

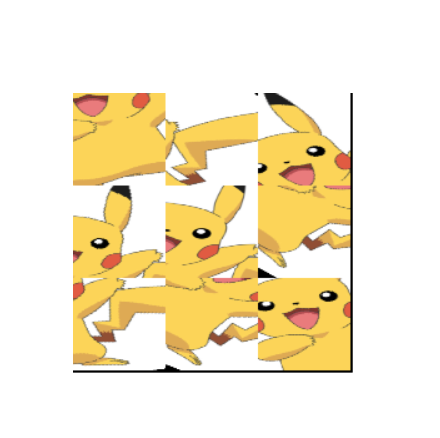
经过data augmentation后的patches，从CNN出来的预测值应该相对是一致的，因为它们的truth应该还是一致的。比如一张猫的图像，经过数据扩充，得到的那些个patch所对应的truth也应该都是猫。对于来自同一幅image的patch集，如果它们的分类结果高度不统一了，那么这个image就是Important的，或者hard sample

注：

由于在annotation之前不知道label，所以我们不能知道网络的预测正确还是错误，但是我们可以知道预测统一还是不统一。所以比如一幅猫的图，如果网络的预测很统一都是狗，那么我们也认为这是一个easy sample，不去active select它的。

结合data augmentation的优点是我们可以知道哪些patch对应什么image，比较容易控制。这样就可以在一个image內算diversity了，每个image对应一个矩阵，大小是一样的，非常的简洁，也容易控制计算量。

问题5：数据增强带来的小问题



即便是很好的分类器，也很难讲裁剪不当的一部分进行分类，但是对于这种情况来说，大部分的数据还是可以被识别出来的。所以不应当被当作hard sample。因此我们再计算相对性时应该保证大方向一致即可。

具体实验：

1. 实验设置：938张图，共五类，batch=16,epoch=20,使用ImageNet预训练的Alexnet
2. 实验结果

|  |  |  |
| --- | --- | --- |
| 方式 | Train | Test |
| 有监督 | 99 | 97 |
| UDA\_EXP | 75 | 75 |
| UDA\_linear | 76 | 66 |
| UDA\_log | 97 | 91 |

Active Learning:

|  |  |  |
| --- | --- | --- |
| epoch | Select data | Test |
| 1 | 723 | 94 |
| 2 | 785 | 96 |
| 3 | 819 | 92 |
| 4 | 868 | 97 |
| 5 | 878 | 95 |
| 6 | 891 | 97 |
| 7 | 893 | 97 |
| 8 | 893 | 97 |