# 半监督consistency\_loss总结

### 1 半监督学习核心问题？

·如何从unlabeled data提炼出合适的信号来指导学习？

·如何平衡labeled/unlabeled data？

·如何选取无标签数据？

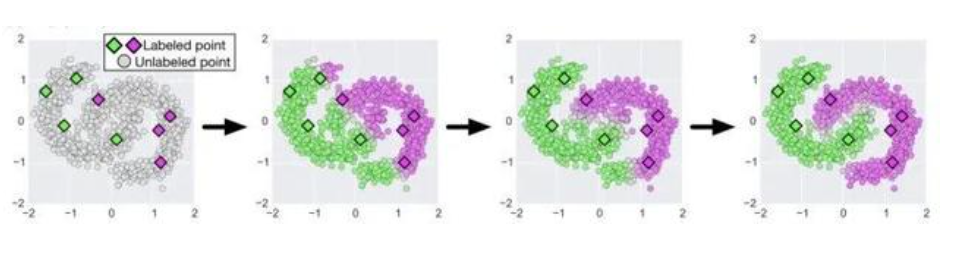
### 2 为什么要利用无标注数据？

通过专家标注的少量有标签数据，结合大量的无标签数据，训练出具备强泛化能力的模型。

### 3 半监督学习基本假设

·平滑假设(Smoothness Assumption)：位于稠密数据区域的两个距离很近的样例的类标签相似，也就是说，当两个样例被稠密数据区域中的边连接时，它们在很大的概率下有相同的类标签；相反地，当两个样例被稀疏数据区域分开时，它们的类标签趋于不同。

·聚类假设(Cluster Assumption)：当两个样例位于同一聚类簇时，它们在很大的概率下有相同的类标签。这个假设的等价定义为低密度分离假设(Low Sensity Separation Assumption)，即分类决策边界应该穿过稀疏数据区域，而避免将稠密数据区域的样例分到决策边界两侧。(熵最小化的思想就是基于低密度假设，半监督学习的目的就是让模型做出低熵预测)

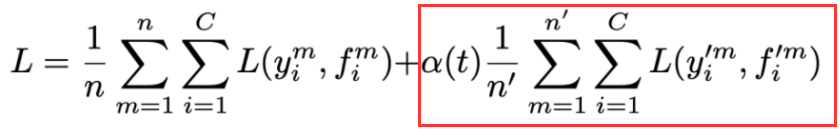


·流形假设(Manifold Assumption)：将高维数据嵌入到低维流形中，当两个样例位于低维流形中的一个小局部邻域内时，它们具有相似的类标签。

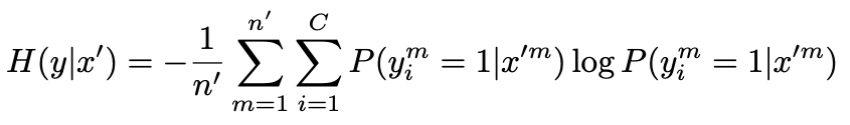
流形假设的主要思想是同一个局部邻域内的样本数据具有相似的性质，因此其标记也应该是相似。这一假设体现了决策函数的局部平滑性。和聚类假设的主要不同是，聚类假设主要关注的是整体特性，流形假设主要考虑的是模型的局部特性。在该假设下，未标记的样本数据就能够让数据空间变得更加密集，从而有利于更加标准地分析局部区域的特征，也使得决策函数能够比较完满地进行数据拟合。

### Pseudo-Label (ICML 2013)

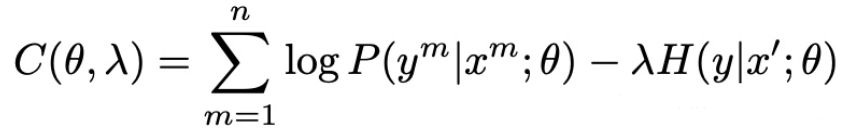
对于无标注数据，选择概率最大的作为pseudo-label(hard-label,为啥不考虑用soft-label)，在训练过程中加入时变参数调整无标注数据的权重，整体Loss为:



熵可以度量class overlap，无标签数据的条件熵为:



通过引入无标注数据的Loss隐式的减少了class overlap。



早期先进行训练，不释放无标注数据的信号(早期模型效果不佳，过早引入产生的信号误差过大)。

缺点：如果class较多，产生预测pseudo-label的值较小或者与第二大概率值较为接近，依然取最大概率值对应的label作为标签明显会引入比较大的错误信号，所以使用hard-label作为伪标签不具有可靠性。

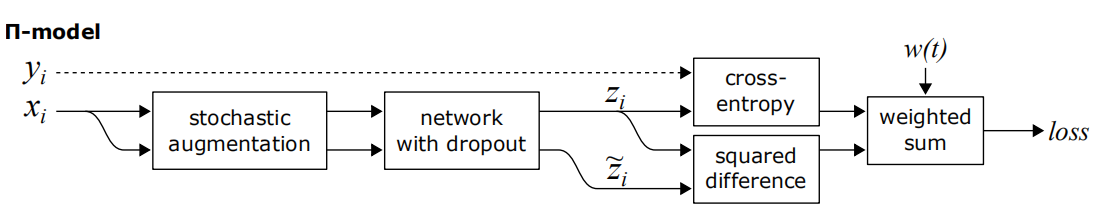
## **ΠModel & Temporal ensembling Model：Temporal ensembling for semi-supervised learning(2017)**

**主要思想：**

·基于平滑性假设提出了一致性正则(consistency regulaization),即作者认为网络在输入数据的附件空间应该是光滑或者平坦的，如果输入数据发生了微小的变化，模型的输出也应该保持不变，一致性正则的思路就是无标注样例在增广后经过分类也应该输出相同的类分布。

·混合网络的集成效果比单独网络的效果好，训练目标总是网络结构的某个特定子集，集成网络可以看作是从更大的网络集合中隐式的找到这种子网络。(要达到模型学习效果所需要的网络往往没有那么复杂，我们总是通过学习复杂网络来找到这个网络)

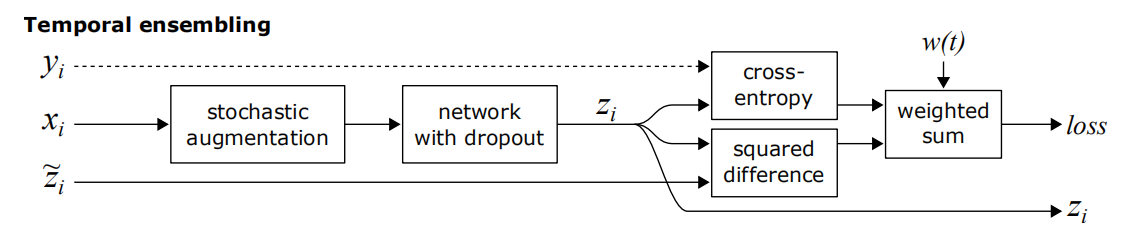
**双模型**：



**基本思路**：

对于每一个训练样本，在训练阶段参与两次前向运算(两次的augmentation和扰动是不同的)，损失由有监督部分的交叉熵损失和一致性正则部分的MSE损失组成，实验证明，引入一致性正则确实增强了模型的鲁棒性。

**时序模型**：



**基本思路**：

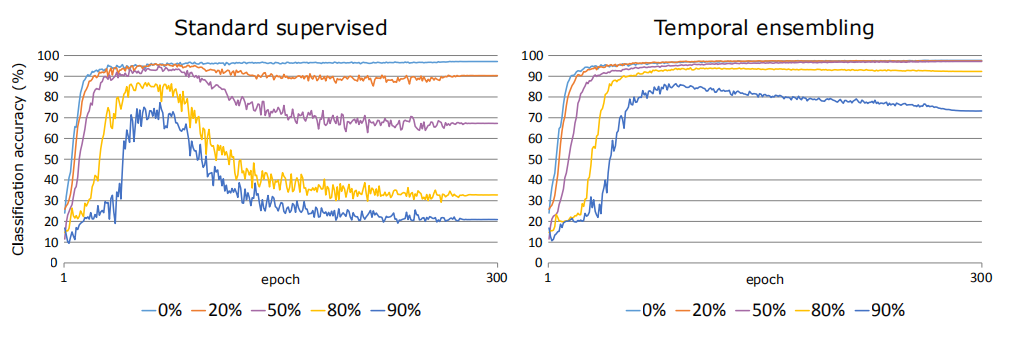
双模型的改进版，基本思路相同，但是对于每一个训练样本，只进行一次前向运算，无标签预测则改为历史预测结果的EMA加权平均，所以相较于双模型带来了两个好处：

·计算量减少了一半(每次只需要一次前向计算，但需要记更多参数，时间换空间)。

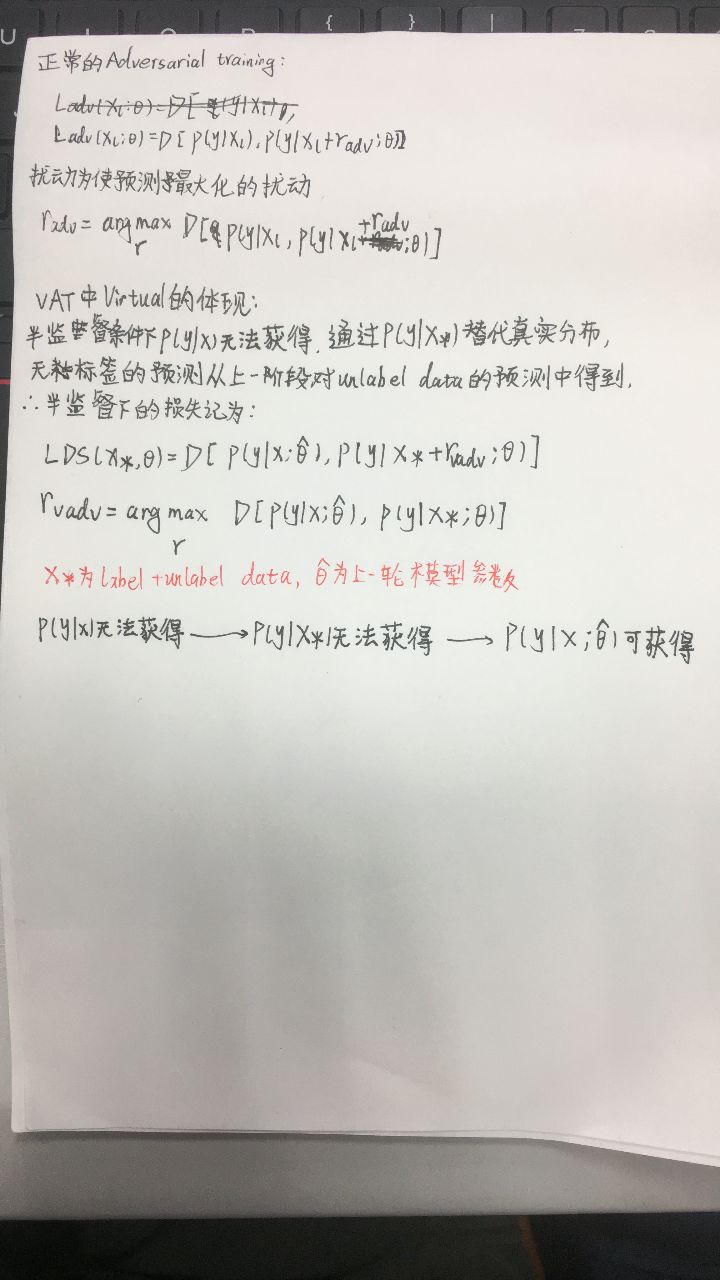
·降低了以单次预测作为无标签预测结果的单次噪声影响。

错误标签容忍：

将时序模型与普通的监督学习相比，对错误标签的容忍程度更高，实验通过将一定比例的样本标签进行随机打乱来产生错误标签，时序模型对错误标签的容忍程度非常高(有标签数据训练鼓励模型学习特殊函数，而引入无标签数据使得映射函数和决策边界更加的光滑)。

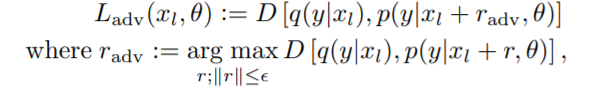


## **VAT：Virtual Adversarial Training: a Regularization Method for Supervised and Semi-supervised Learning(2017)**



**主要思想**:

VAT相比时序模型比较随意的创造扰动，选择在模型变化最快的方向上施加噪声，



**Radv的近似(还没读完)**：

## **Mean Teacher：Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results(2018)**

**主要思想**：

·相比时序模型，research gap主要体现在MT选择对历史模型参数做EMA，而不是对预测结果做EMA，所以他可以在每个step对teacher model进行更新，而不是每一个epoch，相比时序模型更灵活，feedback更快。

·有两种不同的思路提升正则的质量，①谨慎选择扰动，而不是随意使用加法和乘法噪声。（随机扰动的方法会让模型对某些特定方向的微小扰动表现得很敏感、脆弱，这些方向被称作adversarial direction。）②谨慎的选择teacher model，而不是随意的拷贝student model。

VAT选择的前者，MT则是后者，两者的方法切入点不同，可以互补。

## **MixMatch: A Holistic Approach to Semi-Supervised Learning(2019)**

本文总结了近年来的几种方法，作者认为使半监督work的原因主要有三点:

①熵最小化——鼓励模型输出对于未标注数据的自信输出

②一致性正则——鼓励模型在受到扰动时输出相同的输出

③常规正则化——鼓励模型更好的生成并且避免训练时过拟合

**一致性正则：**

监督学习中常用的正则化技术是数据增广，应用于假定为类语义不受输入变化影响，例如，图像分类中，引入弹性变换或者加入噪声，能够在不改变标注的前提下改变像素内容，通俗来讲，它能够通过产生无限接近的新的模拟数据来拓展数据集大小。一致性正则的思路就是利用一个分类器应该输出相同的类分布在未标注样例中即使他已经被增广了。



计算两个增广的L2损失，MT将其中一个用模型参数的EMA值代替了，缺点就是他们运用了特定域的数据增强策略，VAT通过在输入增加了一个加性扰动最大化改变了输出类分布。

**熵极小化**：

在许多半监督学习方法中，一个常见的基本假设是分类器的决策边界不应通过边缘数据分布的高密度区域。实现这个需求的一个办法是分类器输出低熵预测在无标注数据上。

这通常被称为集群假设：如果两个样本在输入分布中属于同一个集群，那么它们很可能属于同一个类。聚类假设相当于低密度分离假设：决策边界应位于低密度区域。这种等价性很容易推断：一个位于高密度区域的决策边界将一个聚类分成两个不同的类，要求不同类的样本位于同一个聚类中，这违反了聚类假设。半监督学习的一致性正则化方法通过鼓励对未标记点u+δ的扰动f（u）=f（u+δ）的不变预测来强化低密度分离假设，当且仅当决策边界穿过低密度路径时，这种一致性和小预测误差可同时满足。

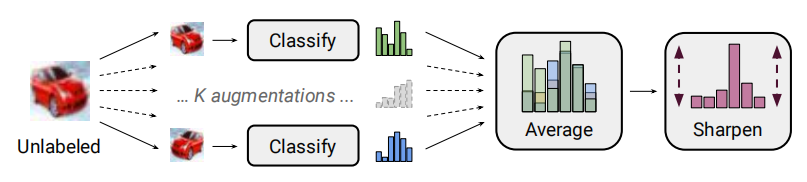
通过最小化无标注数据的条件概率实现(Semi-supervised learning by entropy minimization. In

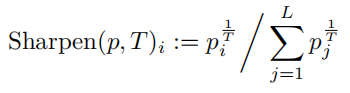
Advances in Neural Information Processing Systems, 2005)。这种熵最小化通过与VAT结合得到更好的结果，Pseudo-Label通过构建硬标签从对无标签数据的高置信度预测和用他们来训练target在标准的交叉熵损失中隐式的实现了熵极小化，MixMatch是通过对无标注数据target分布的“sharpening”函数隐式的实现熵极小化的。

**传统正则化**：

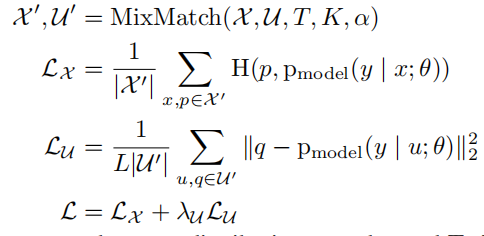
传统正则化就是对模型施加一个约束使得他难以记住训练数据希望他能够更好的推广没见过的数据，我们用权重衰减来惩罚模型参数中的L2范式，我们也用Minxup鼓励曲线表现在样例之间，我们将Mixup用在正则(应用于有标注数据)和半监督方法(应用于无标注数据)中。

**主要思想**：



文章对unlabelled data的处理主要是对每一个无标签样例引入K个不同的数据增广，并经过模型产生预测，取K个预测的算数平均作为伪标签，然后通过sharpening减少数据分布的熵从而鼓励模型产生低熵预测。

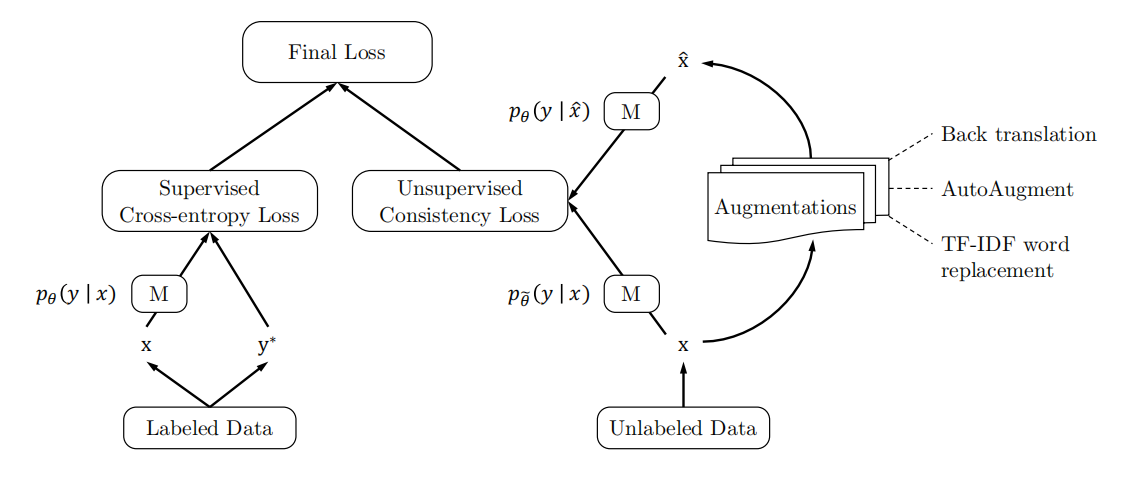
然后要进行MixUp变换，变换会按照一定比例将有标签的数据和无标签的数据进行混合以构成新的样本对。MixUp 变换本身可以视为一种正则化技术。直观来看，它要求，当模型的输入为另外两个输入的线性组合时，其输出也是这两个数据单独输入模型后，所得输出的线性组合。损失计算如下:



**Unsupervised Data Augmentation**

**for Consistency Training（2019）**

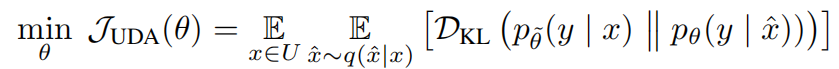
监督学习增广能够等价的看作从原始监督数据集中构建一个增广数据集并在上面训练。因此，增广集需要提供额外的归纳偏差才能更有效。近年，数据增广在多个领域都取得了良好进展，但受限于数据增广一直是作用在小体量有标注数据中，效果有限，所以本文提出了在大量无监督数据中使用数据增广。



最近的无监督数据增广可以总结为两个步骤：

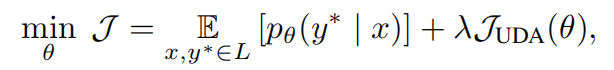
·给定输入x，计算一个输出分布和扰动的输出分布，扰动可以应用在输入，隐藏层，计算过程等方面。

·最小化两个预测分布的散度。



此过程强制模型对扰动不敏感，因此相对于输入（或隐藏）空间的变化更平滑。

**总损失**：



通过优化一致性损失，UDA让模型信息从有标签部分流向无标签部分，并且我们认为相比传统的随机添加扰动，针对任务的特定扰动是一种更有效的扰动，使用针对性的数据增广有几个优点：

· **有效的扰动**——在监督学习中使用数据增广能够帮助产生更多与原数据集共享相同标签的样本，所以鼓励无标签样本和其增广数据具有一致性预测是安全的。

·**多样的扰动**——多种的扰动可以产生不同的输入模拟，(高斯和伯努利扰动只是局部的)，鼓励在多种扰动下的增广数据可以有效的提升样本效率。

·**特定归纳偏好**——不同的目标有不同的归纳偏好，**AutoAugment**中提到数据增强策略可以直接优化，以提高每个任务的验证性能。这种面向表现的增强策略可以找到原始数据集的归纳偏好。