# **一致性训练的无监督数据增强方法**

尽管有很多成功，深度学习仍然在小标注体量数据集表现不好。在这种场景中，数据增强在缓解对标注数据更多的需求上表现出很好的作用，但到目前为止，它主要运用于有监督环境而且取得了有限的收益。我们的工作提出在无监督数据的半监督学习时运用数据增强。

我们的方法鼓励无标注数据和增广无标注数据的模型预测一致，不像以前使用随机噪声，我们使用的是最新的数据增广方法——产生更加难以学习，更加真实的噪声。这种小的转变得到了充分的进步在标注数据非常少的六个语言任务和三个视觉任务中。

半监督学习在最近的研究中可以被分为三类：

·图卷积神经网络

·将预测目标建模为潜在变量

·一致性学习

其中，一致性学习的效果很出众，该方法将模型的预测正则化，使其对应用于样例的小扰动不那么敏感。一个好的模型通常在收到小扰动的时候是不会改变他的本质特征的。

我们的UDA方法，通过最小化原样例和增广样例之间的KL散度执行。

主要贡献：

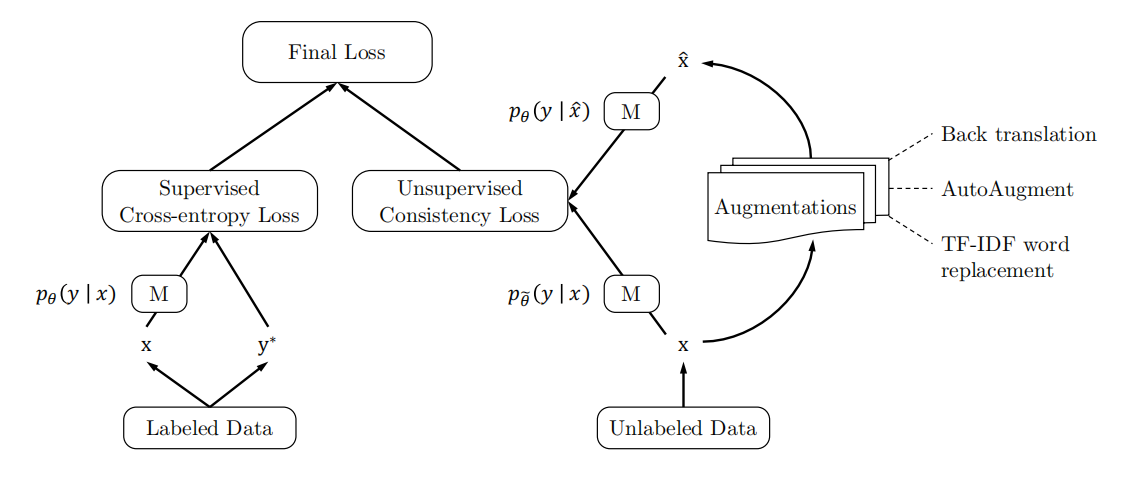
·当有很多未标记数据比标记数据可用时，我们提出了训练技巧TSA能有效防止过拟合。

·我们表明，目标数据增强方法（如AutoAugment[9]）比其他非目标增强方法有显著的提升。

·我们结合了一组NLP的数据增强，并表明我们的方法工作良好，并补充了表示学习方法(如BERT)

·我们的方法在视觉和语言任务都有很好的表现。

·我们开发了一种方法，使UDA甚至可以应用于标记和未标记数据失配的类分布。



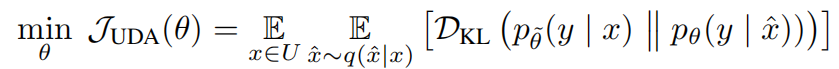
## UDA

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：

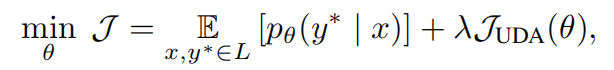
·给定输入x，计算一个输出分布和扰动的输出分布，扰动可以应用在输入，隐藏层，计算过程等方面。

·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。