**Unsupervised Data Augmentation for Consistency Training**

**代码**：https://github.com/google-research/uda

**1.概述**

尽管取得了很大的成功，但是深度学习通常在小规模的有标签的训练中表现不佳。在这些场景中，数据增强在缓解对更多标记数据的需求方面显示出了很大的希望，但迄今为止，它主要应用于有监督的设置，并取得了有限的收益。

在这项工作中，我们建议在半监督学习环境中，将数据扩充应用于未标记的数据。我们的方法，称为无监督数据扩充或UDA，鼓励模型预测一个未标记的例子和一个扩充的未标记的例子之间的一致性。与以前使用随机噪声（如高斯噪声或漏失噪声）的方法不同，UDA有一个小的变化，它利用了由最先进的数据增强方法产生的更难和更真实的噪声。这个小小的转变导致了六种语言任务和三种视觉任务的实质性改进，即使标记集非常小。

**半监督学习**

利用未标记的数据来提高深度学习已经成为解决这一昂贵过程的一个重要研究方向。在这个方向上，半监督学习是最有前途的方法之一，最近的工作可分为三类：（1）通过图卷积和图嵌入基于图的标签传播，（2）将预测目标建模为潜在变量，（3）一致性/平滑性增强（smoothness enforcing）。其中，最后一类方法，即基于平滑强制的方法，在许多任务中都表现出良好的效果。

**平滑度强制方法**

平滑度强制方法只是将模型的预测调整为对应用于示例的小扰动（标记或未标记）不太敏感。给定一个观察到的例子，平滑强制方法首先创建一个扰动版本（例如，通常通过添加人工噪声，如高斯噪声或衰减），并强制两个例子的模型预测相似。直观地说，一个好的模型应该对任何不改变例子性质的小扰动保持不变。在这个通用框架下，这类方法在扰动函数上有很大的不同，即扰动示例是如何创建的。

在我们的论文中，我们建议使用在监督学习中发现的最先进的数据增强方法作为平滑性实施框架中的扰动函数。我们表明，更好的增强方法会带来更大的改进，并且它们可以用于许多其他领域。我们的方法，称为无监督数据增强或UDA，最大限度地减少了原始示例上的模型预测与由数据增强生成的示例之间的KL差异。尽管数据增强已经被广泛研究并导致了显著的改进，但它主要应用于有监督的学习环境中。另一方面，UDA可以直接将最新的数据增强方法应用于非监督数据，这些数据在数量上更大，因此有可能比标准的监督数据增强工作得更好。

**贡献**

首先，我们提出了一种称为TSA的训练技术，当可用的未标记数据比标记数据多得多时，它可以有效地防止过度拟合。

其次，我们证明了目标数据增强方法（如AutoAugment）比其他非目标增强方法有显著的改进。

第三，我们结合了一组NLPs的数据扩充，证明了我们的方法是有效的，并补充了表示学习方法，如BERT。

第四，在视觉和语言任务方面，我们的研究显示，与之前的方法相比，我们的研究在性能上有了显著的飞跃。

最后，我们发展了一种方法，使得UDA可以应用于有标记和无标记数据不匹配的类分布。

**2. Unsupervised Data Augmentation (UDA)**

我们首先制定我们的任务，然后提出建议的方法UDA。本文以分类问题为中心，用x表示输入，用y（x）或简单的表示其ground-truth预测目标。我们感兴趣的是学习一个模型来根据输入x预测 ，其中表示模型参数。最后，我们将使用L和U分别表示标记的和未标记的示例集。

**2.1背景：监督数据增强**

数据增强的目的是在不改变标签的情况下，通过对一个例子应用转换，创建新颖、逼真的训练数据。形式上，设为增广变换，从中可以在原实例x的基础上画出增广实例。

为了使增广变换有效，要求从分布中提取的任何示例与x共享相同的基本真值标签，即。给定一个有效的增广变换，我们可以简单地最小化增广例子的负对数似然。

有监督数据增广可以等价地看作是在原有监督集上构造一个增广标记集，然后在增广集上训练模型。因此，增广集需要提供额外的归纳偏差才能更有效。因此，如何设计增广变换就变得至关重要。

尽管取得了很有希望的结果，但数据增强通常被认为是“蛋糕上的樱桃”，它提供了稳定但有限的性能提升，因为这些增强到目前为止仅应用于一组通常较小的标记示例。基于这一局限性，我们开发了UDA来将有效的数据增强应用于未标记的数据，这些数据通常数量较大。

**2.2 Unsupervised Data Augmentation**

半监督学习中最近的一项工作是利用未标记的示例来增强模型的平滑性。这些工作的一般形式可归纳如下：给定一个输入x，通过注入一个小噪声，计算出给定x的输出分布和一个扰动版本。噪声可以应用于x或隐藏状态，也可以用于改变计算过程。 最小化两个预测分布之间的散度度量。

此过程强制模型对扰动不敏感，因此对于输入（或隐藏）空间的变化更平滑。

在这项工作中，我们对现有的平滑性/一致性增强工作提出了一个简单的扭曲，并扩展了以前使用数据增强作为扰动的工作。我们建议使用针对不同任务的最新数据增强作为一种特殊的扰动形式，并在未标记的示例上优化相同的平滑度或一致性强制目标。具体来说，根据VAT，我们选择最小化未标记示例上的预测分布和扩展未标记示例上的预测分布之间的KL差异：

其中是一个数据增强变换，是当前参数的一个固定副本,表示梯度没有通过传播。这里使用的数据增强变换与监督数据增强中使用的增强变换相同，例如文本的后向翻译和图像的随机裁剪。由于在训练过程中动态运行反向翻译成本很高，因此我们离线生成增强的示例。为每个未标记的示例生成多个扩展示例。

为了同时使用标记示例和未标记示例，我们将标记示例上的交叉熵损失和等式1中定义的一致性/平滑性目标加上权重因子作为我们的训练目标，如图1所示。在形式上，目标的定义如下：

通过最小化一致性损失，UDA允许标签信息从有标签的示例传播到无标签的示例。我们将大多数实验的设置为1，并对监督目标和非监督目标使用不同的批大小。我们发现，在一些数据集上，使用较大的批处理大小来实现无监督目标可以获得更好的性能。

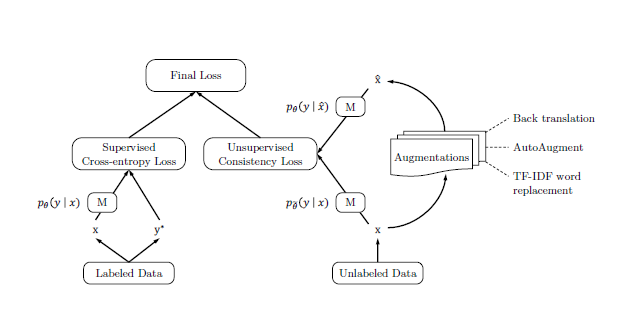


图1:UDA的训练目标，其中M是一个预测给定x的y分布的模型。

与传统的扰动（如高斯噪声、漏失噪声）和简单的增强（如仿射变换）相比，我们认为针对每项任务的数据增强可以作为更有效的“噪声源”。具体地说，使用目标数据增强作为扰动函数有几个优点：

有效扰动：在有监督学习中获得良好性能的数据增强方法具有生成共享相同的真实增强示例的优势. 因此，鼓励原始未标记示例上的预测与扩展未标记示例上的预测之间的平滑性或一致性是安全的。

多样性扰动：数据增强可以生成多样性的示例集，因为它可以在不改变其标签的情况下对输入示例进行大的修改，而高斯或伯努利噪声等扰动只会进行局部更改。在一组不同的增广示例上鼓励平滑性可以显著提高样本效率。

有针对性的归纳偏差：不同的任务需要不同的归纳偏差。如AutoAugment所示，可以直接优化数据增强策略，以提高每个任务的验证性能。这种面向性能的增强策略可以学习找出原始标记集中缺失或最需要的归纳偏差。我们发现，虽然AutoAugment优化了模型在有监督学习环境下的性能，但AutoAugment发现的增强策略在半监督学习环境下运行良好。

正如我们将在消融研究中所显示的，注入定向诱导偏倚的各种有效增强是导致显著性能改善的关键组成部分。

**2.3不同任务的扩充策略**

如第2.2节所述，数据扩充可以定制，以提供特定于每个任务的缺失归纳偏差。在本节中，我们将讨论用于不同任务的三种不同的增强，并讨论数据增强的多样性和有效性之间的权衡。我们利用最近在数据增强方面的进展，并应用以下增强策略：

用于图像分类的自动增强。对于图像分类，AutoAugment使用强化学习直接基于验证性能搜索图像增强操作的“最优”组合，其性能明显优于任何手动设计的增强过程。我们使用所发现的增强策略，并使用AutoAugment的opensource在CIFAR-10、SVHN和ImageNet上进行实验。我们还对CIFAR-10和SVHN使用Cutout[14]，因为它可以与AutoAugment组合以获得更好的性能。

文本分类的反向翻译。反向翻译[55，15]可以在保留原句语义的同时生成不同的释义，并且已经被证明可以显著提高QANet在问答中的性能[66]。因此，我们采用反向翻译系统对情绪分类数据集（包括IMDb、Yelp-2、Yelp-5、Amazon-2和Amazon-5）的训练数据进行解释。我们发现释义的多样性比释义的质量和效度更重要。因此，我们采用温度可调的随机取样，而不是光束搜索来产生。更具体地说，我们使用WMT 14语料库训练英法和法英翻译模型，并对每个句子而不是整个段落进行反向翻译，因为wmt14中的平行数据用于句子级翻译，而情感分类语料库中的输入示例是段落。如图2所示，后译句产生的释义非常多样，语义相似。

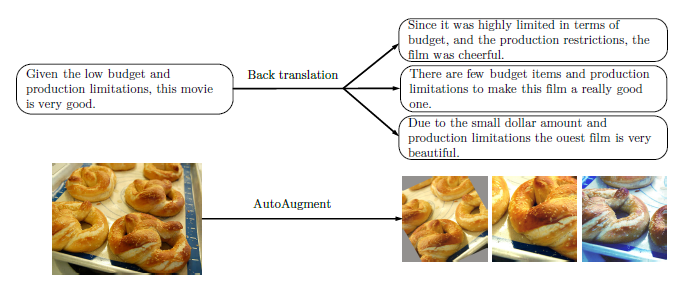


图2：使用反向翻译和自动扩充的扩充示例

基于TF-IDF的文本分类替换。虽然反向翻译很好地保持了原句的整体语义，但不能保证它会保留某些词。然而，在DBPedia上，任务是预测Wikipedia页面的类别，在确定类别时，一些关键字比其他关键字更具信息性。因此，我们提出了一种基于TF-IDF的词替换增强方法，该方法在保留TF-IDF得分较高的关键词的同时，替换通常TF-IDF得分较低的非信息性词。我们请读者参阅附录B了解详细说明。

**2.4多样性和数据扩充有效性之间的权衡**

尽管如第2.2节所述，最先进的数据增强方法可以生成多种有效的增强示例，但多样性和有效性之间存在取舍，因为多样性是通过改变原始示例的一部分来实现的，自然会导致改变地面真值标签的风险。我们发现在数据增强方法的多样性和有效性之间进行权衡是有益的。

在图像分类中，AutoAugment根据监督设置下的验证集性能进行优化，自动寻找多样性和有效性之间的最佳点。

对于文本分类，我们调整随机抽样的温度。一方面，当我们使用温度为0时，随机抽样解码退化为贪婪解码，并生成完全有效但完全相同的释义。另一方面，当我们使用温度为1时，随机抽样会产生非常多样但几乎不可读的解释。我们发现，将Softmax温度设置为0.7,0.8或0,9可以获得最佳性能。

**3附加训练技巧**

在本节中，我们将介绍在不同场景中应用UDA的其他技术。首先，为了使模型能够在不过度拟合的情况下在更多未标记数据上进行训练，我们在第3.1节中介绍了一种称为训练信号退火的技术。然后，为了使预测过平时的训练信号更强，我们在第3.2节中提出了三种直观的方法来增强预测。最后，为了将UDA应用于域外未标记数据，我们在第3.3节中介绍了一种简单的域相关数据过滤方法。

**3.1训练信号退火**

由于获取未标记数据要比获取标记数据容易得多，在实际应用中，我们经常遇到未标记数据量与标记数据量之间存在较大差距的情况。

为了使UDA能够尽可能多地利用未标记的数据，我们通常需要一个足够大的模型，但是一个大的模型可以很容易地过拟合有限大小的监督数据。为了解决这一难题，我们引入了一种新的训练技术，称为训练信号退火（TSA）。

TSA的主要直觉是，随着模型在越来越多的未标记示例上进行训练，逐步释放标记示例的训练信号，而不会过度拟合它们。具体来说，对于每个训练步骤t，我们设置一个阈值，其中K是类别数。

当标记样本的正确类别的概率高于阈值时，我们从损失函数中移除该样本，只训练小批量中的其他标记样本。形式上，给定一小批标记的示例B，我们将监督目标替换为以下目标：

其中是指示函数，只是重整化因子。有效地，阈值用作上限，以防止模型在模型已经有信心的示例上过度训练。当我们在训练过程中逐渐从退火到1时，模型只能缓慢地接受标记实例的监督，这在很大程度上缓解了过拟合问题。假设T是训练步骤的总数，t是当前的训练步骤。为了说明未标记数据和标记数据的不同比率，我们考虑了三个特定的schedule，如图3所示。

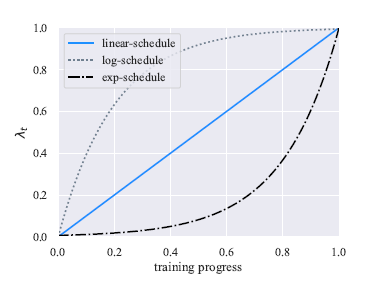


图3:TSA的三个schedule。设，当从0增加到1时，从1/k增加到1。

log-schedule训练开始时阈值增加最快:

linear-schedule -随着训练的进行而线性增加:

exp-schedule —训练结束时，阈值的增加最快：。

直观地说，当模型容易过拟合时，例如当问题相对容易或标记实例的数量非常有限时，exp调度是最合适的，因为监督信号大多在训练结束时释放。遵循类似的逻辑，当模型不太可能过度拟合时（例如，当我们有大量的标记示例或当模型使用有效的正则化时），log-schedule可以很好地服务。

**3.2锐化预测**

我们观察到，在问题困难且标记实例数量很少的情况下，未标记实例和增广未标记实例的预测分布在类别上趋于平坦。因此，来自KL散度的无监督训练信号相对较弱，因此受到有监督部分的控制。例如，在ImageNet上，当我们使用10%的标记集时，未标记示例上的预测分布要比标记示例上的分布尖锐得多。因此，我们发现有助于提高在未标记示例上生成的预测分布，并使用以下三种直观技术：

基于置信度的掩蔽-我们发现掩蔽当前模型不自信的示例是有帮助的。具体地说，在每个小批量中，一致性损失项仅在最大概率（即）大于阈值的示例上计算。

熵最小化-熵最小化[17]将增广示例上的预测分布正则化，使其具有低熵。为了使用这项技术，我们在总体目标上增加了一个熵项。

Softmax温度控制-在计算原始示例的预测时，我们调整Softmax温度。具体来说，计算为，其中表示logits，表示温度。较低的温度对应于更尖锐的分布。

在实际应用中，我们发现基于置信度的掩蔽和softmax温度控制相结合对于标记数据量非常小的设置是最有效的，而熵最小化对于标记数据量相对较大的情况是很有效的。

**3.3领域相关性数据过滤**

理想情况下，我们希望使用域外未标记数据，因为它通常更容易收集，但域外数据的类分布通常与域内数据的类分布不匹配。由于类分布不匹配，使用域外未标记的数据比不使用它会损害性能。为了获取与当前任务相关的域数据，我们采用了一种常见的域外数据检测技术。我们使用在域内数据上训练的基线模型来推断大型域外数据集中的数据标签，并挑选出模型最有信心的示例（在类之间平均分布）。具体来说，对于每一个类别，我们都会根据在该类别中的分类概率对所有域外的示例进行排序，并选择概率最高的示例。

**4.实验**

我们将UDA应用于各种语言和视觉任务。具体来说，我们在第4.1节中展示了六个文本分类任务的实验。然后，在第4.2节中，我们将UDA与其他半监督学习方法在标准视觉基准、CIFAR-10和SVHN上进行了比较。最后，我们在第4.3节的ImageNet上评估UDA，并在第4.4节中提供TSA和增强方法的消融研究。我们在此仅提供比较实证结果所需的信息，并请读者参阅附录D和实施细则。

**4.1文本分类实验**

数据集。我们对包括IMDb、Yelp-2、Yelp-5、Amazon-2、Amazon-5和DBPedia[40、70]在内的六种语言数据集进行了实验，其中DBPedia包含用于分类的Wikipedia页面，而所有其他数据集都是关于不同域上的情感分类的。在我们的半监督设置中，我们将IMDb、Yelp-2和Amazon-2等二元情感分类任务的监督实例数设置为20。对于五向分类数据集Yelp-5和Amazon-5，我们使用2500个示例（即每个类500个示例）。最后，尽管DBPedia有14个类别，但问题相对简单。因此，我们将每个类的训练示例数设置为10。对于未标记的数据，我们使用DBPedia的整个训练集，以及IMDb的训练集和未标记集的连接。我们获得了Yelp评论和Amazon评论的大数据集[41]作为Yelp-2yelp-5、Amazon-2和Amazon-5.3的未标记数据

实验设置。我们采用了BERT[13]中使用的Transformer模型[60]作为我们的基线模型，因为它在许多任务上都有很好的性能。然后，我们考虑了四种初始化方案。具体地说，我们使用（a）随机变压器，（b）BERTBASE，（c）bertrage或（d）BERTFINETUNE初始化我们的模型：bertrage在域内未标记数据上进行微调。后一种微调策略的动机是ELMo[49]和ULMFiT[26]表明，在特定于域的数据上微调语言模型可以提高性能。我们不在DBPedia上进行BERTFINETUNE的进一步实验，因为在DBPedia上对BERT进行微调在我们的初步实验中不会产生比bertrage更好的性能。这可能是因为DBPedia在Wikipedia域中，BERT已经在整个Wikipedia语料库中接受了训练。在这四种情况下，我们比较了有无UDA的性能。

主要结果。文本分类的结果如表1所示，其中有三个关键的观察结果。

首先，无论采用何种模型初始化方案，UDA都能持续提高性能。

最值得注意的是，即使BERT对域内数据进行了进一步的微调，UDA仍然可以在IMDb上将错误率从6:50%显著降低到4:20%。这一结果表明，UDA提供的好处是对表征学习的补充。

其次，与使用全监督数据训练的SOTA模型相比，UDA具有更小的监督实例数量，能够提供体面甚至有竞争力的性能。

特别是在二元情感分类任务中，UDA仅用了20个有监督的例子，在IMDb上的训练效果优于以前的SOTA，在Yelp-2和Amazon-2上的训练效果非常接近。

最后，我们还注意到，五类情感分类任务比它们的二元分类任务要困难得多，并且UDA（每类500个标记示例）和BERT（在整个监督集上训练）之间仍然存在明显的差距。这表明未来还有进一步改进的空间。

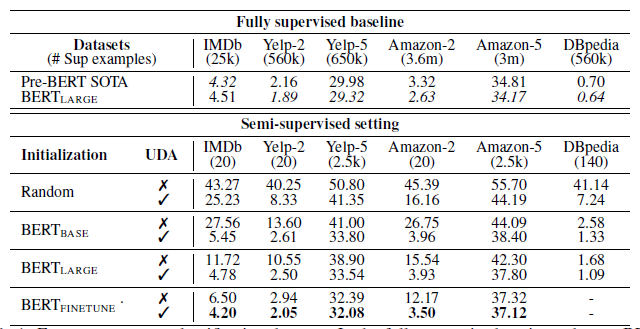


表1：文本分类数据集的错误率。在完全监督的设置中，预BERT SOTAs包括用于Yelp-2和Yelp-5的ULMFiT[26]、用于Amazon-2和Amazon-5的DPCNN[29]、用于IMDb和DBPedia的混合增值税[51]。

不同标记集大小的结果。我们也用不同数目的监督例子来评估UDA的性能。如图4所示，UDA导致了所有标记集大小的一致性改进。在大数据体制下，通过IMDb的完整训练集，UDA也提供了强大的增益。在Yelp-2上，有2000个例子，UDA的性能优于之前的SOTA模型（训练了560000个例子）。

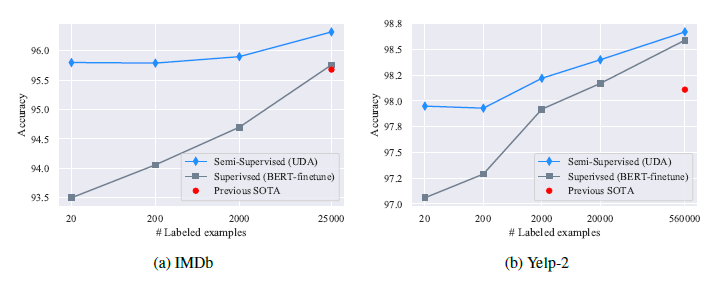


图4:IMDb和Yelp-2上不同数量标记示例的精确度。

**4.2与半监督学习方法的比较**

实验设置。在标准的半监督学习设置之后，我们将UDA与CIFAR-10和SVHN的先前工作进行了比较。我们遵循[46]中的设置，employWideResNet-28-2[67，21]作为我们的基线模型。我们将UDA与伪标号进行了比较，伪标号[36]是一种基于自训练的算法，虚拟对抗训练（VAT）是一种在输入端产生对抗性高斯扰动的算法，-模型，它将简单的输入增强和隐藏的状态扰动相结合，Mean Teacher[58]，它对模型参数和MixMatch进行平滑化，MixMatch是一个并行的工作，它统一了先前关于半监督学习的几个工作。

与现有半监督学习方法的比较。在图5中，我们将UDA与不同大小的标记示例的现有作品进行了比较。UDA比现有的所有方法都有明显的优势，包括MixMatch[3]，这是一个半监督学习的并行工作。例如，对于250个示例，UDA在CIFAR-10上实现8:41的错误率，在SVHN上实现2:85的错误率，而MixMatch在CIFAR-10上实现11:08的错误率，在SVHN上实现3:78的错误率。更有趣的是，当不使用自增强时，UDA与在全监督数据上训练的模型的性能相匹配。特别是，UDA在带有4000个标记示例的CIFAR-10上实现了5:27的错误率，在带有250个标记示例的SVHN上实现了2:85的错误率，这些标记示例与我们的完全监督模型的性能相匹配，而无需自增强，在带有50000个标记示例的CIFAR-10上实现了5:36的错误率，在带有73257的SVHN上实现了2:84的错误率注意UDA和VAT的区别本质上是扰动过程。虽然VAT产生的扰动通常包含真实图像中不存在的高频伪影，但数据增强大多产生多样的真实图像。UDA与VAT的性能差异表明了基于数据增强的扰动方法的优越性。

在附录C中，我们还将UDA与最近提出的方法（包括ICT[61]和mixmixup[19]）进行了比较，这些方法实施的插值平滑度与基于梯度相似性的算法mixup[69]和LGA+VAT[28]相似。UDA优于所有以前的方法，并且减少了超过30%的最新方法的错误率。

采用更先进的体系结构的结果。我们还测试UDA是否可以通过使用Shake Shake（26 2x96d）[16]和pyraminet+ShakeDrop[63]而不是Wide-ResNet-28-2从更高级的架构中获益。如表2所示，当我们采用金字塔网络+抖振时，UDA的错误率为2:7，与完全监督模型的性能匹配而不需要自增强，并且通过混合匹配获得的错误率为4:95，优于最佳半监督学习结果。

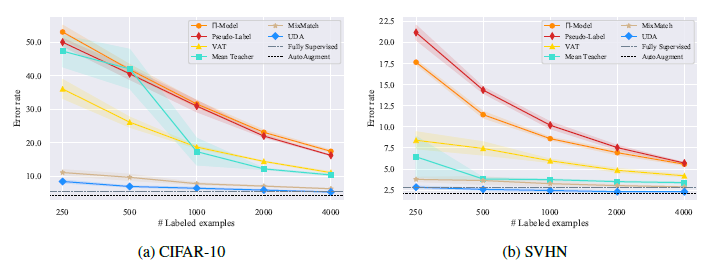


图5:CIFAR-10和SVHN的半监督学习方法与不同标记示例数的比较。文献[3]报道了-模型、伪标签、增值税和平均教师的表现。

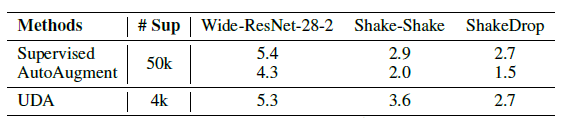


表2:CIFAR-10不同型号的错误率。#Sup表示监督示例的数量。ShakeDrop表示PyramidNet+ShakeDrop。只有4000个标记的示例，UDA与Wide-ResNet-28-2和pyraminet+ShakeDrop的完全监督模型的性能匹配，而没有自动增强。

**4.3 ImageNet实验**

在前面的小节中，我们考虑的所有数据集都有相对较少的训练示例和类。此外，在以往的实验中，我们只使用域内未标记数据，其中未标记数据的类分布总是与标记数据的类分布相匹配。为了进一步测试UDA是否仍能在更大和更具挑战性的数据集上表现出色，我们在ImageNet上进行了实验[12]。我们还开发了一种将UDA应用于域外未标记数据的方法，当我们使用整个ImageNet作为监控数据时，该方法可以提高性能。

实验设置。为了提供一个信息性的评估，我们在两个不同数量的监督示例的设置上进行了实验：（a）我们使用ImageNet的10%的监督数据，同时使用所有其他数据作为未标记数据；（b）其次，我们考虑将ImageNet中的所有图像作为监督数据并获得额外的来自JFT数据集的未标记数据[25，6]。我们使用领域相关数据过滤方法从JFT数据集中过滤出130万张图像。

结果。如表3所示，对于10%监督数据设置，与监督基线相比，UDA将top-1和top-5的精确度分别从55:09%提高到68:66%和77:26%提高到88:52%。与以往的半监督学习（VAT+EntMin）相比，UDA将前5名的准确率从83:39%提高到88:52%。对于表4所示的完整ImageNet设置，与AutoAugment相比，UDA将基线top-1精度从78:28%提高到79:04%，将top-5精度从94:36%提高到94:45%，而未标记数据仅增加1:3M。我们期望更多未标记的数据会有进一步的改进，我们将在未来继续工作。

**4.4消融研究**

在本节中，我们将分析何时以及如何使用TSA以及不同增强方法对研究者和实践者的影响

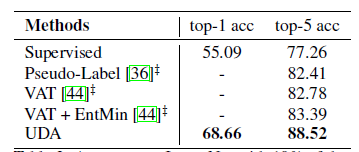


表3:ImageNet上10%标记集的精确度。我们将图像大小设置为224。关于z的结果，见[68]。

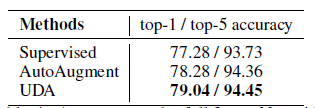


表4：图像大小为331的完整图像网的精度。我们使用ImageNet数据集和来自JFT数据集的1:3未标记图像作为未标记数据。

训练信号退火的烧蚀。我们研究了TSA对两个不同未标记数据量的任务的影响：（a）Yelp-5：在本文的分类任务中，我们有大约6百万个未标记的例子，而只有2:5千个有监督的例子。在本研究中，我们没有通过BERT初始化网络来排除具有预先训练的表示的因素，（b）CIFAR-10：我们有50k个未标记的示例，而有4k个标记的示例。

如表5所示，在Yelp-5上，未标记数据比监督数据多得多，与没有TSA的基线相比，TSA将错误率从50:81%降低到41:35%。更具体地说，当我们选择将监督的训练信号延迟到训练结束时，即exp-schedule导致最佳性能时，就可以获得最佳性能。另一方面，在CIFAR-10上，线性调度是最适合的一种训练信号发布速度，在这种情况下，未标记的数据量并不比监督的数据量大很多。

关于增强方法的消融。有针对性的增强方法，如AutoAugment，已经被证明能显著提高监督学习的性能。在这项研究中，我们希望调查当应用于未标记数据时，目标增广是否有效，以及监督学习中增广的改进是否会导致半监督学习环境的改进。

首先，如表6所示，如果我们将自增强在SVHN上发现的增强策略应用于CIFAR-10（用交换增强表示），则错误率从5:10增加到5:59，这证明了目标数据增强的有效性。此外，如果删除AutoAugment并只使用Cutout，则错误率将增加到6:42。最后，如果我们只使用简单的裁剪和翻转作为增强，错误率将增加到16:17。在SVHN上，不同的增强效果是相似的。这些结果表明了在每一个任务中应用增强方法来注入最需要的归纳偏差的重要性。

我们还观察到，增强方法在有监督学习设置上的有效性转移到我们的半监督设置上。具体来说，在完全监督的学习环境中，Cubuk等人。[9] 同时表明，自增强对截断的改进，截断比基本增强更有效，与半监督环境下的观测结果吻合得很好。在我们对情绪分类的初步实验中，我们还发现，无论是在有监督的学习环境还是无监督的学习环境中，反向翻译都比简单的单词删除或单词替换效果好，尽管单词删除或替换可以提高纯监督基线上的性能。

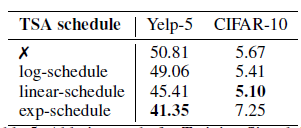


表5:Yelp-5和CIFAR-10上训练信号退火（TSA）的消融研究。显示的数字是错误率。

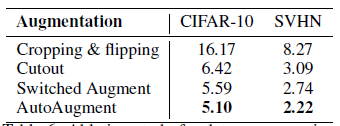


表6：数据增强方法的消融研究。交换增广意味着将AutoAumgent在SVHN上找到的策略应用于CIFAR-10，反之亦然。

**6结论**

本文证明了数据增强和半监督学习是紧密相连的：更好的数据增强可以显著提高半监督学习的效果。我们的方法，UDA，使用高度针对性的数据增强来产生不同的和真实的扰动，并强制模型对于这些扰动是平滑的。我们还提出了一种称为TSA的技术，当有更多的未标记数据可用时，它可以有效地防止UDA对监督数据的过度拟合。对于文本，UDA很好地结合了表示学习，例如BERT，并且在低数据状态下非常有效，在IMDb上仅用20个例子就可以实现最先进的性能。对于vision，UDA在高度基准化的半监督学习设置中减少了30%以上的错误率。最后，UDA可以有效地利用域外未标记的数据，在拥有大量有监督数据的ImageNet上提高性能。