

NLP数据增强方法总结：EDA、BT、MixMatch、UDA

夕小瑶的卖萌屋

以下文章来源于平安寿险PAI，作者孙梦轩



平安寿险PAI

平安人寿人工智能研发团队的官方账号，在这里，我们怀揣促进交流碰撞、与AI共成长...



一只小狐狸带你解锁 炼丹术&NLP 秘籍



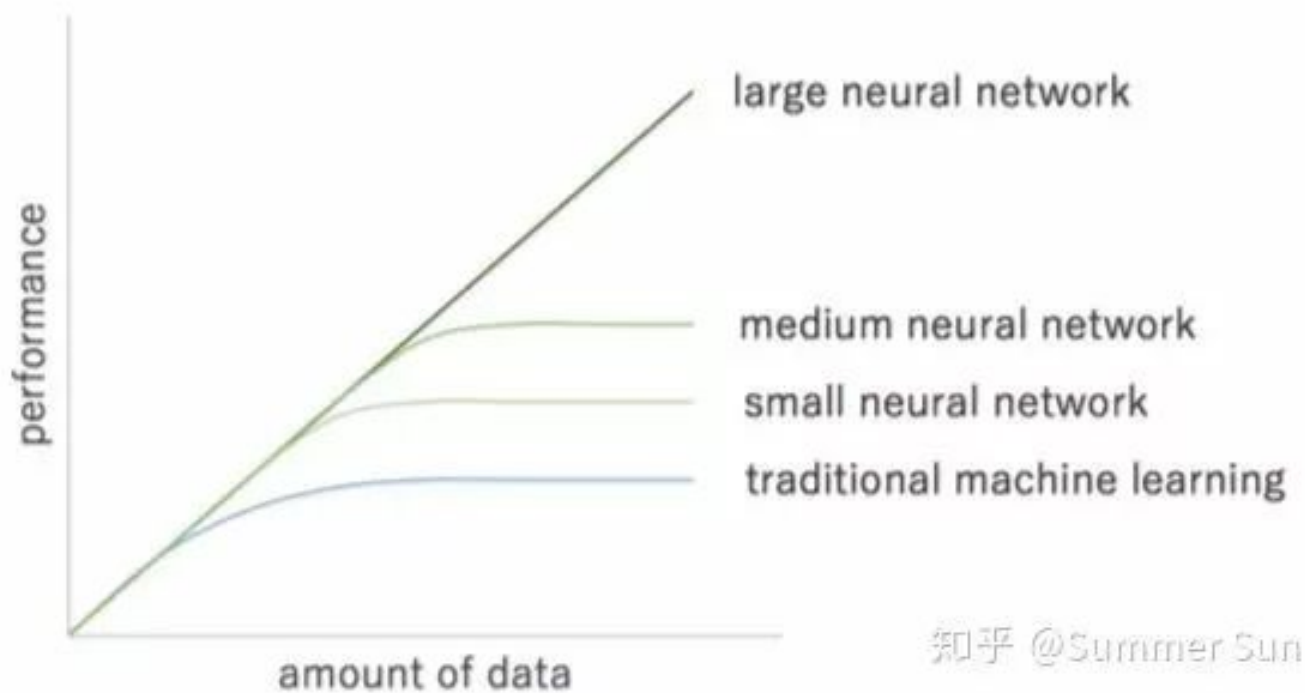
老板，只有几百条数据没法用神经网络呀，可以批点经费标数据吗？

没钱，EDA, BT, UDA了解一下



喵喵喵？

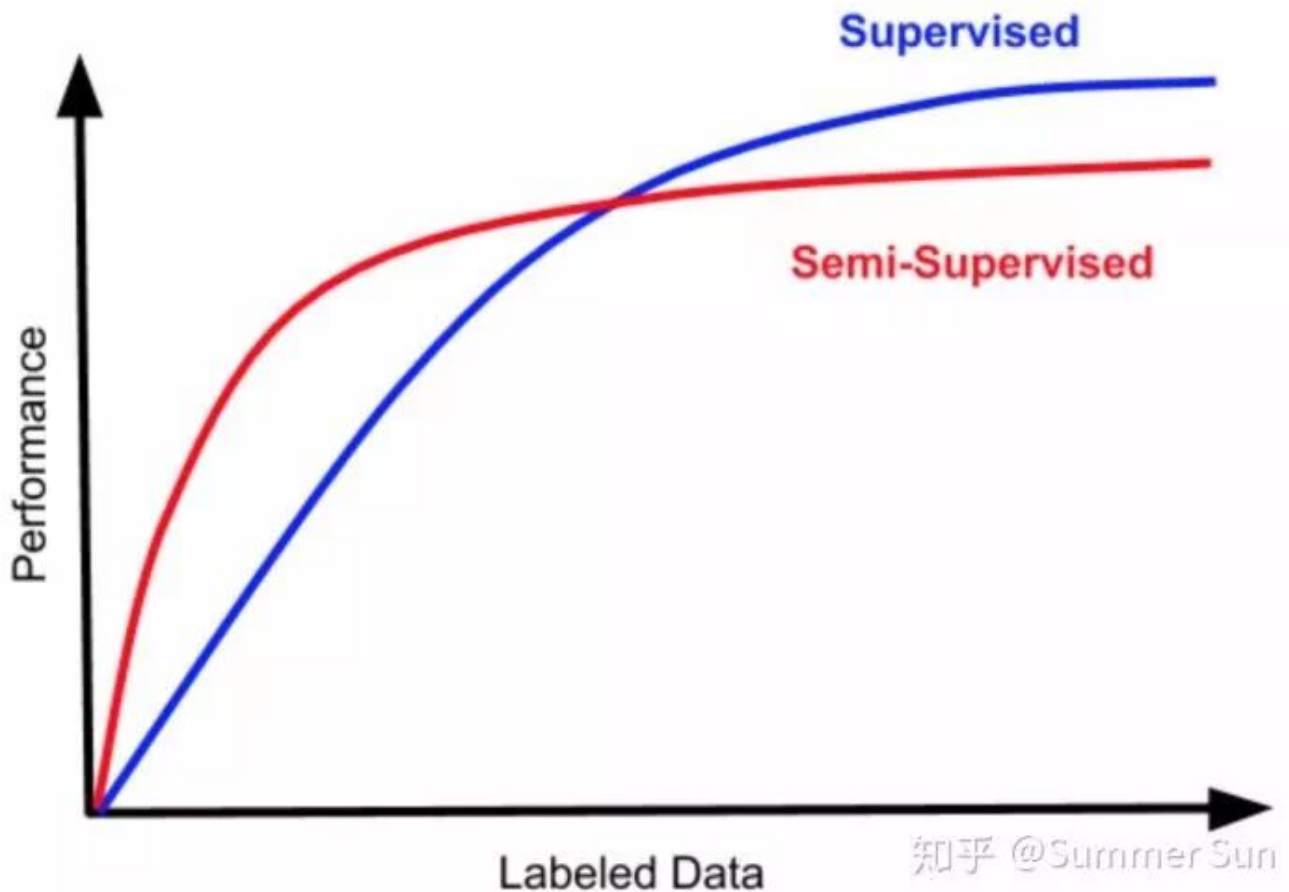
1 数据增强的背景和应用场景



随着AI技术的逐步发展，更好的神经网络模型对数据规模的要求也逐步提升。而在分类任务中，若不同类别数据量相差很大，模型则会出现过拟合现象，严重影响预测的正确性。

从广义上来讲，有监督模型的效果相对半监督或无监督学习都是领先的。但是有监督模型需要获取大量的标注数据，当数据需求达到十万、百万甚至更多时，人工标注数据昂贵的代价已经让很多人望而却步。

如何利用有限的标注数据，获取到更多的训练数据，减少网络中的过拟合现象，训练出泛化能力更强的模型？**数据增强**无疑是一种强有力的解决方法。



数据增强起初在计算机视觉领域应用较多，主要是运用各种技术生成新的训练样本，可以通过对图像的平移、旋转、压缩、调整色彩等方式创造新的数据。虽然，‘新’的样本在一定程度上改变了外观，但是样本的标签保持不变。且NLP中的数据是离散的，这导致我们无法对输入数据进行直接简单地转换，换掉一个词就有可能改变整个句子的含义。因此本文将重点介绍**文本数据增强的方法和技术**，以快速补充文本数据。

2 传统文本数据增强的技术

现有NLP的Data Augmentation大致有两条思路，一个是加噪，另一个是回译，均为有监督方法。**加噪**即为在原数据的基础上通过替换词、删除词等方式创造和原数据相类似的新数据。**回译**则是将原有数据翻译为其他语言再翻译回原语言，由于语言逻辑顺序等的不同，回译的方法也往往能够得到和原数据差别较大的新数据。

Easy Data Augmentation for Text Classification Tasks (EDA) 提出并验证了几种加噪的text augmentation 技巧，分别是**同义词替换 (SR: Synonyms Replace)**、**随机插入 (RI: Randomly Insert)**、**随机交换 (RS: Randomly Swap)**、**随机删除 (RD: Randomly Delete)**，下面进行简单的介绍：

2.1 EDA

(1) 同义词替换 (SR: Synonyms Replace) : 不考虑stopwords, 在句子中随机抽取n个词, 然后从同义词词典中随机抽取同义词, 并进行替换。

Eg: “我非常喜欢这部电影” —> “我非常喜欢这个影片”, 句子仍具有相同的含义, 很有可能具有相同的标签。

(2) 随机插入(RI: Randomly Insert): 不考虑stopwords, 随机抽取一个词, 然后在该词的同义词集合中随机选择一个, 插入原句子中的随机位置。该过程可以重复n次。

Eg : “我非常喜欢这部电影” —> “爱我非常喜欢这部影片”。

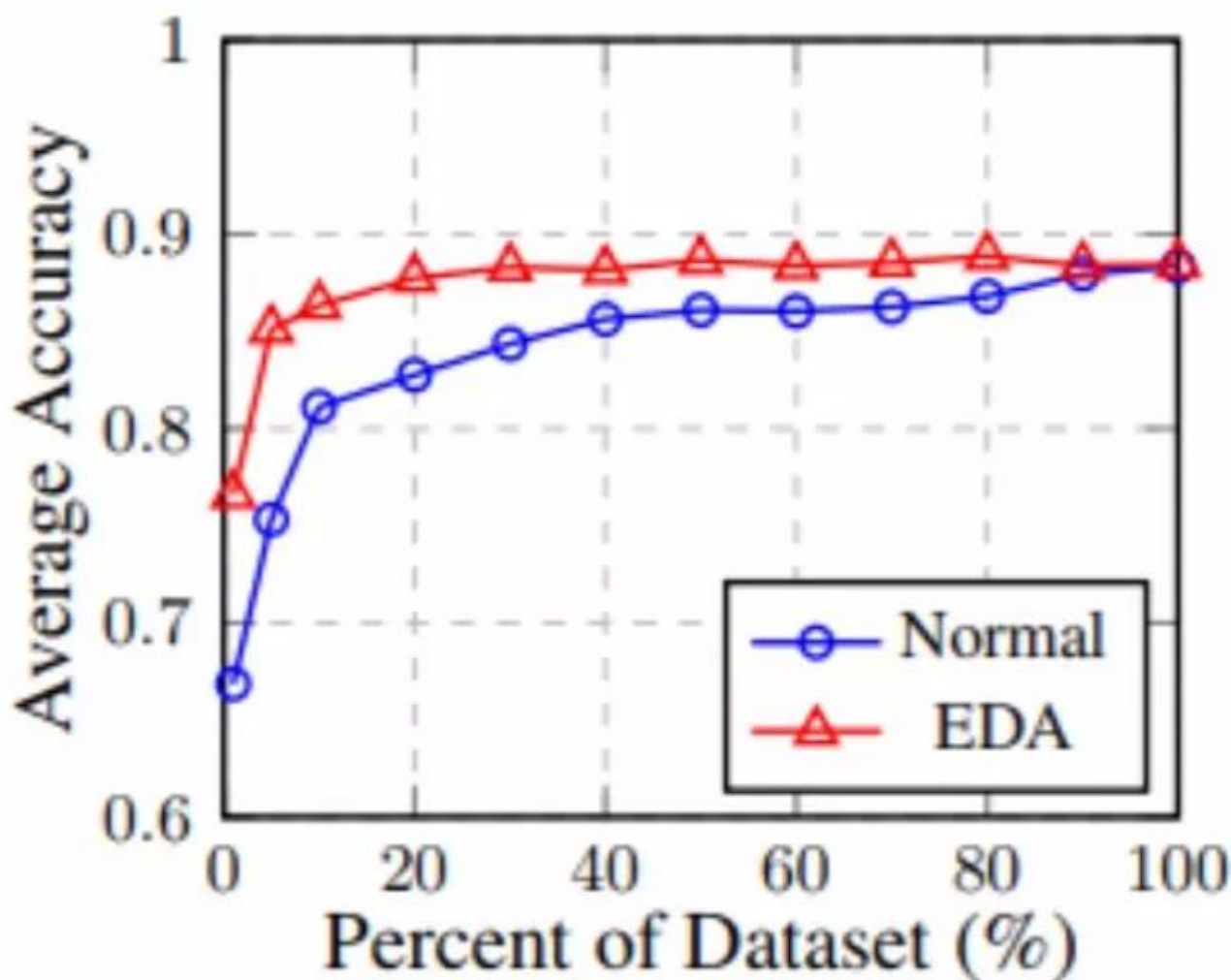
(3) 随机交换(RS: Randomly Swap): 句子中, 随机选择两个词, 位置交换。该过程可以重复n次。

Eg: “如何评价 2017 知乎看山杯机器学习比赛?” —> “2017 机器学习?如何比赛知乎评价看山杯”。

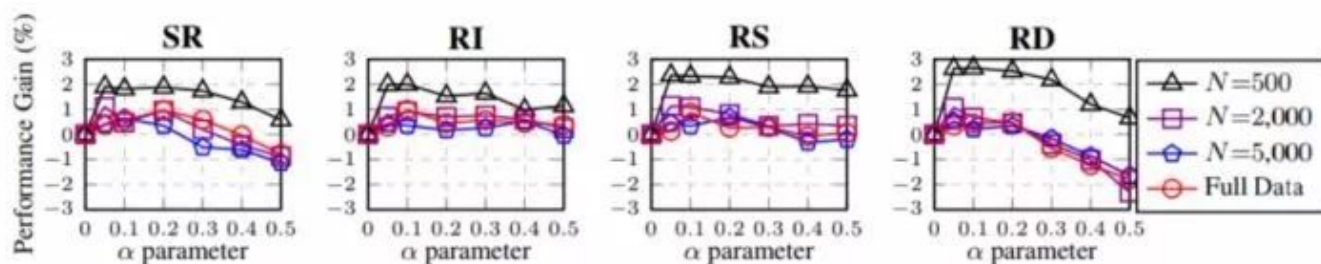
(4) 随机删除(RD: Randomly Delete): 句子中的每个词, 以概率p随机删除。

Eg: “如何评价 2017 知乎看山杯机器学习比赛?” —> “如何 2017 看山杯机器学习 ”。

这四种方法的效果如何呢? 在英文的数据上效果很可观。经过上述四种操作, 数据增强后的句子可能不易理解, 但作者们发现模型变得更加鲁棒了, 尤其是在一些小数据集上。效果如下图:



每一种方法也在作者的结果中展示了还不错的效果：



上图是针对不同训练集大小的五个文本分类任务的EDA操作的平均性能增益。 α 参数粗略地表示“每次扩充改变的句子中单词的百分比”，纵轴是模型增益。

我们可以看到，当 $\alpha = 0.1$ 时，模型提升就能达到很好的效果。**训练数据越少，提升效果效果越明显**。过多的数据增强数据实际上对模型的提升有限，甚至在RD和SR两种方法上还会严重损害效果。

总的来说，传统的文本数据增强的方法在小批量数据中都有较好的表现效果，但**4种方法的缺点**也不能被忽视：

- **同义词替换SR**有一个小问题，同义词具有非常相似的词向量，而训练模型时这两个句子会被当作几乎相同的句子，但在实际上并没有对数据集进行有效的扩充。
- **随机插入RI**很直观的可以看到原本的训练数据丧失了语义结构和语义顺序，而不考虑停用词的做法使得扩充出来的数据并没有包含太多有价值的信息，同义词的加入并没有侧重句子中的关键词，在数据扩充的多样性上实际会受限较多。
- **随机交换RS**实质上并没有改变原句的词素，对新句式、句型、相似词的泛化能力实质上提升很有限。
- **随机删除RD**不仅有随机插入的关键词没有侧重的缺点，也有随机交换句式句型泛化效果差的问题。随机的方法固然能够照顾到每一个词，但是没有关键词的侧重，若随机删除的词刚好是分类时特征最强的词，那么不仅语义信息可能被改变，标签的正确性也会存在问题。

2.2 回译

在这个方法中，我们用机器翻译把一段中文翻译成另一种语言，然后再翻译回中文。

Eg: “周杰伦是一位华语乐坛的实力唱将，他的专辑卖遍了全球。” —> “Jay Chou is a strength singer in the Chinese music scene, his albums are sold all over the world.” —> “周杰伦是中国音乐界的优秀歌手，他的专辑畅销全世界。”

这个方法已经成功的被用在Kaggle恶意评论分类竞赛中。反向翻译是NLP在机器翻译中经常使用的一个数据增强的方法，其本质就是**快速产生一些翻译结果达到增加数据的目的**。

回译的方法往往能够增加文本数据的多样性，相比替换词来说，有时可以改变句法结构等，并保留语义信息。但是，回译的方法产生的数据依赖于翻译的质量，大多数出现的翻译结果可能并不那么准确。如果使用某些翻译软件的接口，也可能遇到账号限制等情况。

3 深度学习数据增强技术

3.1 半监督 Mixmatch

半监督学习方法的提出是为了更好地利用未标注的数据，减轻对于大规模标注数据集的依赖；如今也证明了这是一种强有力的学习范式。

在这篇论文中，作者们把当前不同任务中的做法为半监督学习做了统一，得到了一种新的算法——**MixMatch**。它的工作方式是通过 MixUp 猜测数据扩增方法产生的无标签样本的低熵标签，并把无标签数据和有标签数据混合起来。

作者们通过实验表明 MixMatch 在多种不同的数据集、多种不同的有标签数据规模中都能以很大幅度领先此前的所有方法。比如，在 CIFAR 数据集上、只有 250 个标签的情况下，作者们把错误率降低到了之前方法的 1/4，在 STL-10 数据集上也降低到了之前方法的一半。

作者们也展示了 MixMatch 可以在差分隐私的使用目的下，在准确率和隐私保护之间取得好得多的平衡。最后，作者们进行了对照实验，分析了 MixMatch 方法中的哪些组件最为关键。

3.2 无监督数据增强UDA

由EDA结果可知，传统的数据增广方法有一定的效果，但主要针对小数据量，对于渴求大量训练数据的深度学习模型，传统的方法效果始终有限。而 **Unsupervised Data Augmentation (UDA)** 无监督数据扩增方法的提出，为大量数据缺失打开了一扇大门。

MixMatch 算法除了使用普通的数据增广，还有一个秘诀是 Mixup 增广术。而 UDA 的成功，得益于**对特定任务使用特定目标的数据增强算法**。

与常规噪声比如高斯噪声、dropout 噪声相比，针对不同任务使用不同数据增强方法能够产生更有效的数据。这种方法能够产生有效、真实的噪声，且噪音多样化。另外以目标和性能为导向的数据增强策略可以学习如何在原始标记集中找出丢失的或最想要的训练信号（比如图像数据以颜色为目标进行数据增强）。

下图展示了UDA训练时的目标和结构，为了使用标记和未标记的所有数据，对有标签的数据训练时加入了cross entropy loss 函数。对未标记数据，与Mixmatch使用 l2 loss 不同，UDA对增广后未标记的数据预测结果使用KL散度。**Targeted data augmentation 特定目标的数据增强**则包括了back translation回译、autoaugment(图像)、TFIDF word replacement。其中回译是从英文转法文再译回英文，IDF是从DBPedia语料中获取。

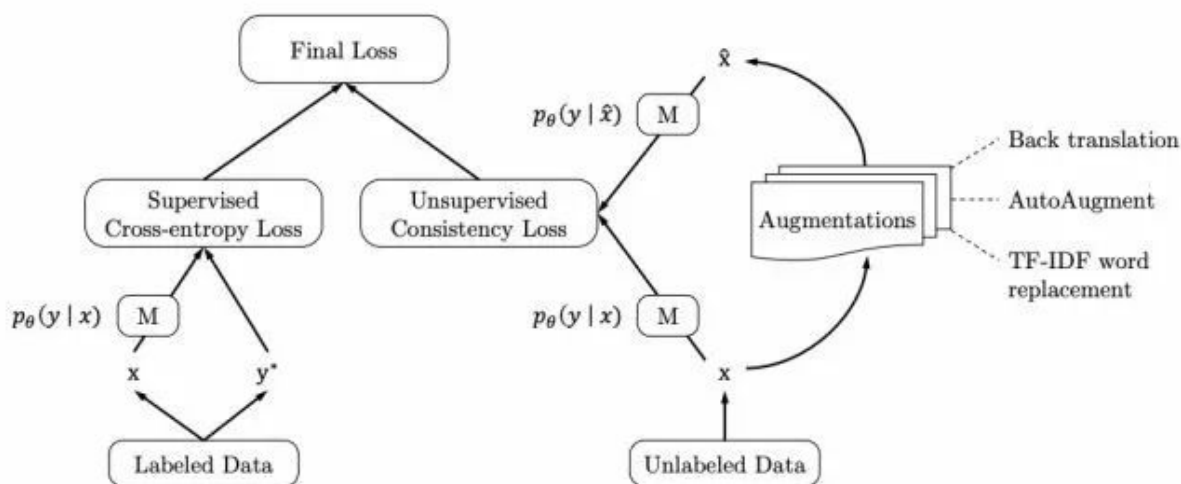


Figure 1: Training objective for UDA, where M is a model that predicts distribution $p_{\theta}(y | x)$ given x , and y^* is the ground-truth label.

知乎 @Summer Sun

作者在文本的处理方式上选用了回译和关键词提取两种方式，回译的方式可以帮助丰富数据的句式和句型，而tfidf方法优化了EDA的随机处理词策略，根据DBPedia先验知识和实际预料的词频确定关键词，再根据确定好的关键词替换同义词，避免无用数据和错误数据的产生。

另外，UDA优秀的另一个重要的突破是采用了Training Signal Annealing (TSA) 方法在训练时逐步释放训练信号。

当收集了少量的标注的数据和大量未标记的数据时，可能会面临标记数据和未标记数据相差很大的情况。比如标记的数据都和保险相关，但未标记的数据都是热点新闻。因为需要采用大量的未标记数据进行训练，所需的模型会偏大，而大模型又会轻松的在有限的有监督数据上过拟合，这时TSA就要逐步的释放有监督数据的训练信号了。

作者对每个training step 都设了一个阈值 η_t ，且小于等于1，当一个标签例子的正确类别P的概率高于阈值 η_t 时，模型从损失函数中删除这个例子，只训练这个minibatch下其他标记的例子。

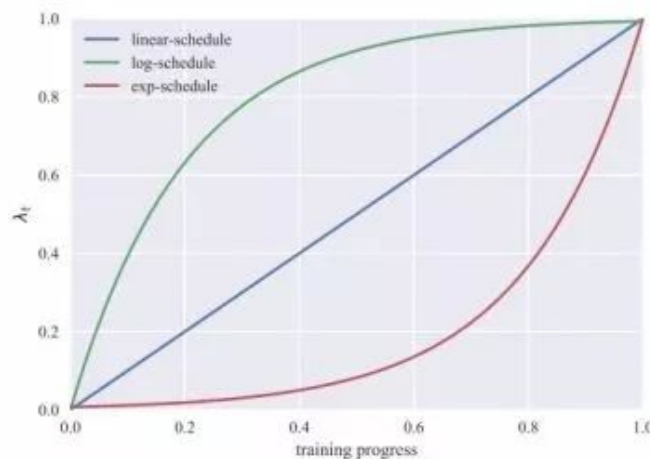


Figure 2: Three different schedules of TSA where λ_t is increased from 0 to 1. We simply set $\eta_t = \frac{1}{k} + \lambda_t * (1 - \frac{1}{k})$ so that η_t goes from $\frac{1}{k}$ to 1.

如上图展示了3种TSA的方式，这3种方式适用于不同数据。exp模式更适合于问题相对容易或标注量较少的情况。因为监督信号主要在训练结束时释放，且可以防止模型快速过拟合。同理，log模式适合大数据量的情况，训练过程中不太容易过拟合。

那么UDA效果如何呢？作者的实验结果显示，这种无监督方法创造的数据在多个任务上都有很好的表现：①在IMDb数据集的分类测试中，UDA只使用20个标签就得到了比此前最好的方法在25,000个有标签数据上训练更好的结果；②在标准的半监督学习测试（CIFAR-10，4000个标签；以及SVHN，1000个标签）中，UDA击败了此前所有的方法，包括MixMatch，而且把错误率降低了至少30%；③在大规模数据集上，比如在ImageNet上，只需要额外增加130万张无标签图像，相比此前的方法，UDA也可以继续提升首位和前五位命中率。

4 数据增强技术实践

利用eda和回译的方法扩增数据我们已经写入相关项目：

<https://github.com>

可以通过pip安装调用

```
1 pip install textda
```

```
1 from textda.data_expansion import *
2
3 print(data_expansion('生活里的惬意，无需等到春暖花开'))
```

output:

```
['生活里面的惬意，无需等到春暖花开',  
'生活里的等到春暖花开',  
'生活里无需惬意，的等到春暖花开',  
'生活里的惬意，无需等到春暖花开',  
'生活里的惬意，并不需要等到春暖花开',  
'生活无需的惬意，里等到春暖花开',  
'生活里的惬意，等到无需春暖花开']
```

4.1 某翻译软件回译：

原句：生活里的惬意，无需等到春暖花开

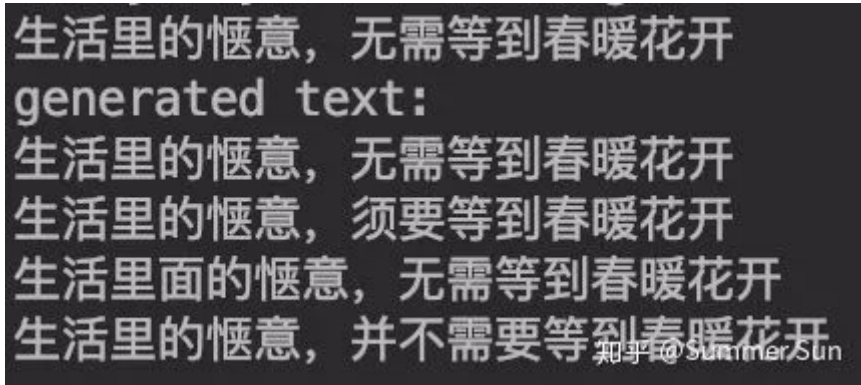
中—>英—>中：生活的舒适，无需等到春天开花

中—>日—>中：生活的舒适，无需等到春天的花朵

中—>德—>中：生活的舒适，无需等到春天开花

中—>法—>中：生活的舒适，无需等待春天的花朵

4.2 EDA 产生的数据：



生活里的惬意，无需等到春暖花开
generated text:
生活里的惬意，无需等到春暖花开
生活里的惬意，须要等到春暖花开
生活里面的惬意，无需等到春暖花开
生活里的惬意，并不需要等到春暖花开

4.3 textda对不平衡文本分类的效果提升

此处以情感正中负文本3分类结果为例：

- 最初训练文本：neg1468, pos 8214, neu 712
- 测试文本：neg1264, pos 1038, neu 708
- 分类模型：fastText文本分类器训练模型

由下图的confusion matrix 可知模型整体加权 f1值为 0.749

Confusion matrix, without normalization					
	precision	recall	f1-score	support	
neg	0.88261	0.79114	0.83438	1264	
neu	0.74384	0.42655	0.54219	708	
pos	0.67233	0.95279	0.78836	1038	
micro avg	0.76113	0.76113	0.76113	3010	
macro avg	0.76626	0.72350	0.72164	3010	
weighted avg	0.77746	0.76113	0.74978	3010	

利用textda的方法将数据扩充至 neg:7458 , pos:8214 , neu:3386

当数据趋于平衡，f1值上升到0.783，将近4个百分点

Confusion matrix, without normalization					
	precision	recall	f1-score	support	
neg	0.85425	0.83465	0.84434	1264	
neu	0.75264	0.50282	0.60288	708	
pos	0.74731	0.93738	0.83162	1038	
micro avg	0.79203	0.79203	0.79203	3010	
macro avg	0.78474	0.75829	0.75961	3010	
weighted avg	0.79347	0.79203	0.78316	3010	

由此可见数据增强方法在处理数据不平衡的分类任务上可以提高模型的性能。

5 数据增强的拓展

5.1 其他数据增强方法

数据增强方法还有很多，且在文本、语音、图像上的方法都各有不同。

(1) 音频：

- 噪声增强
- 随机相同类型抽取拼接
- 时移增强
- 音高变换增强
- 速度调整

- 音量调整
- 混合背景音
- 增加白噪声
- 移动音频
- 拉伸音频信号

(2) 图像:

- 水平翻转垂直翻转
- 旋转
- 缩放 放大缩小
- 裁剪
- 平移
- 高斯噪声
- 生成对抗网络 GAN
- AutoAugment

(3) 文本其他数据增强方法:

- 语法树结构替换
- 篇章截取
- seq2seq序列生成数据
- 生成对抗网络 GAN
- 预训练的语言模型

无论是文本、语音还是图像，数据增强虽然有不同的方法，但这些方法本质上是相似的：传统直观的方法是对不同信号的裁剪、拼接、交换、旋转、拉伸等方式，**采用深度学习模型的方法主要为生成和原数据相类似的数据。**

5.2 防止过拟合其他方法

在深度学习中，为了避免出现过拟合（Overfitting），通常输入充足的数据量是最好的解决办法。当数据无法达到模型的要求或者添加数据后模型由于某类数据过多导致过拟合时，以下方法也可以发挥一些作用：

- **Regularization:** 数据量比较小会导致模型过拟合，使得训练误差很小而测试误差特别大。通过在Loss Function 后面加上正则项可以抑制过拟合的产生。缺点是引入了一个需要手动调整的hyper-parameter。

- **Dropout**: 这也是一种正则化手段，不过跟以上不同的是它通过随机将部分神经元的输出置零来实现。
- **Unsupervised Pre-training**: 用Auto-Encoder或者RBM的卷积形式一层一层地做无监督预训练，最后加上分类层做有监督的Fine-Tuning。
- **Transfer Learning (迁移学习)**: 在某些情况下，训练集的收集可能非常困难或代价高昂。因此，有必要创造出某种高性能学习机 (learner)，使得它们能够基于从其他领域易于获得的数据上进行训练，并能够在对另一领域的数据进行预测时表现优异。

6 总结和展望

训练机器学习或深度学习模型时，良好的数据往往是影响模型的效果最重要的因素之一。而数据不足时数据增强是一个常用的方法。

文本数据增强从对原数据词的变动到句子的变动到段落的变动都有不同的方法，为了保证能够真实提高数据的质量，有以下几个点尤为重要：

(1) 增加的数据要保证和原数据一致的语义信息。

新增后的数据和原数据拥有一样标签的同时，更需要保证有一样的语义信息。单独随机去掉某个词的方式很可能会改变整句的含义（比如去掉一个否定词）。

(2) 增加的数据需要多样化。

从替换词、句式、句型等方面都需要有新的数据以增强模型的泛化能力，单独交换词的方式较为局限。

(3) 增加的数据要避免在有标签数据上过拟合。

当大量的数据在少量的有标签数据上过拟合时，模型虽然可能会出现很高的f1值，但真实的预测效果会相差很多。保证多样化的数据还要保证数据的质量。

(4) 增加的数据和原数据保持一定的平滑性会更有价值，提高训练效率。

生成的数据更接近于真实数据可以保证数据的安全性，大噪音产生的数据和原始数据的标签很可能不同。尤其在某些序列模型中，文本数据的通顺程度严重影响模型的预测。

(5) 增加数据的方法需要带着目标去选择。

对数据缺失的需求明确才能更快的找到理想的数据，对某些关键词的同义词需求较多可以偏重替换词的方式，对句式缺失较多可以偏重回译或者句式语法结构树变换的方式。

对于小数据的情况，使用文本回译或EDA中的简单方法可以达到效果的提升；但想要使用大批量的数据训练神经网络模型，EDA或者回译的方式产生的文本可能并不能满足需求。

而UDA这种无监督数据增强技术，无论对于小数据量或大数据量数据，都可以找到带有目标性的方法获得增强后的平滑的数据，甚至有时效果高于有监督方法训练的模型。

综上，**数据增强的方法**可以作为我们训练nlp模型时一个快速解决数据不平衡或数据缺失的强有力的工具。

可能喜欢

- 显存不够，如何训练大型神经网络？
- 如何让BERT拥有视觉感知能力？两种方式将视频信息注入BERT
- Stanford CS224n追剧计划（附追剧计划详细攻略）
- 如何扩充知识图谱中的同义词
- 一则漫画带你图解强化学习
- 从why到how，从词典匹配到预训练详解中文分词
- 深度神经网络为何会有灾难性遗忘？如何进行有效的持续学习？
- 模型训练太慢？显存不够用？混合精度训练了解一下
- 万万没想到，我的炼丹炉玩坏了



[1] Wei, J. W. and Zou, K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196 ,2019.

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. arXiv:1905.02249 [cs.LG], 2019.

[3] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le.Unsupervised Data Augmentation. arXiv e-prints, page arXiv:1904.12848, Apr 2019.

[4] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le.
Autoaugment: Learning augmentation policies from data. arXiv preprint arXiv:1805.09501,
2018.



夕小瑶的卖萌屋



关注&星标小夕，带你解锁AI秘籍
订阅号主页下方「撩一下」有惊喜哦