nlp论文 2023 12月4-12月10 R-Drop的正则化方法以及对比损失方法

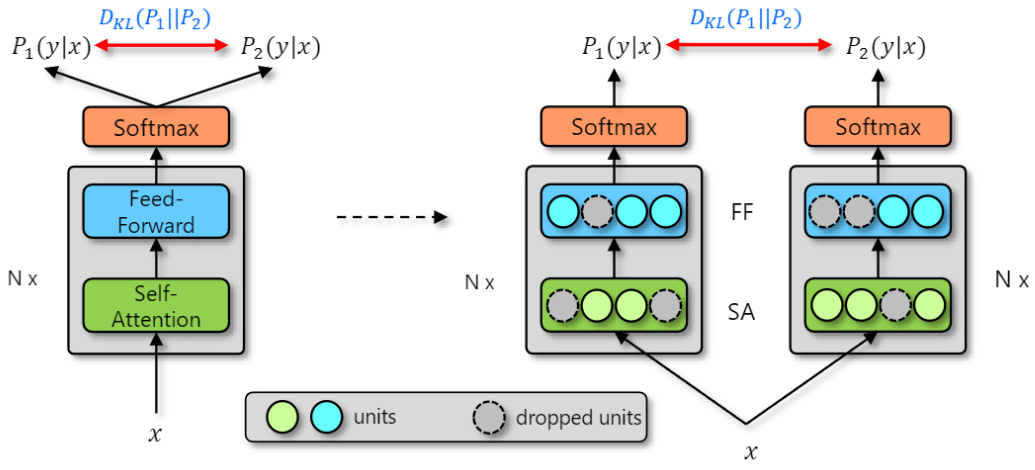
R-Drop: Regularized Dropout for Neural Networks

1.引入的原因

为了改善dropout的缺陷。
dropout：Dropout是一种正则化技术，用于减少神经网络过拟合的风险。过拟合是指模型在训练数据上表现很好，但在新数据上表现较差的情况。Dropout通过在训练过程中随机丢弃（将其权重设为零）一些神经元的输出来降低模型的复杂性。
具体来说，在每一次训练迭代中，Dropout会随机选择一些神经元，并将它们的输出设置为零。这意味着在每次迭代中，模型都在不同的子集上训练，从而使模型更加健壮，能够更好地泛化到未见过的数据。在测试时，不应用Dropout，以充分利用整个网络。
缺点：训练阶段使用dropout丢弃了一部分隐藏层单元，以防止过拟合，并得到了相应的参数。然而，在推断阶段，我们通常使用没有丢弃隐藏层的完整模型进行预测。这就导致了训练阶段和推断阶段使用的模型在隐藏层位置上不一致，即训练的模型和用于预测的模型有差别

2. R-Drop结果特点

R-Drop在dropout上加了Kullback-Leibler，目的是让模型在训练时尽量去掉的是相同的隐藏层而不是像dropout那样去掉不同的隐藏层，这样可以使模型的效果变得更好。



如图所示，以transformer模块为例，输入x,由于去掉不同的隐藏层会得到两个不同的P1 P2 引入Kullback-Leibler就是为了让两个概率尽可能的小使得反过来让去掉的隐藏层是

相同的。

3.dropout与损失函数的区别

（1）**Dropout**：主要从网络结构和训练过程的角度来考虑，通过在训练时随机丢弃一部分神经元的输出，促使模型更加鲁棒，减少对特定训练样本的过拟合。这有助于提高模型的泛化能力，使其在未见过的数据上表现更好。
（2） **损失函数**：主要从目标函数的角度来考虑，通过衡量模型输出与实际标签之间的差异，引导模型朝着更准确的方向更新参数。选择合适的损失函数可以影响模型的训练过程，使其更好地学习任务的关键特征，提高性能。

4 .实验

本文在Machine Translation，Language Understanding，Summarization，Language Modeling，Image Classification的不同数据上都取得了不错的效果。
本文还探讨了损失函数的收敛速度，发现其收敛速度慢一些，于是有探讨了是否可以每k步再使用，结果发现k越大效果越差；还讨论了Kullback-Leibler的Weight α，效果是先随着 α的增加是先增后减。

Supervised Contrastive Learning

数据增强：数据增强操作是通过对原始样本进行一系列随机变换来生成多个不同的样本。这些变换可以包括随机翻转、旋转、颜色扭曲、裁剪等。生成的多个样本被视为同一样本的不同视图，用于进行对比学习。这样可以增加数据的多样性，提高模型的鲁棒性和泛化能力

1.监督vs自监督

1. 监督对比是在有标签数据的情况下进行的对比学习。它利用样本的标签信息来构建正样本对和负样本对。正样本对由同一类别的样本组成，而负样本对由不同类别的样本组成。通过最大化正样本对的相似性，同时最小化负样本对的相似性，来训练模型以更好地区分不同类别的样本。
2. 自监督对比是在无标签数据的情况下进行的对比学习。它利用数据的自身特征进行对比学习，而不依赖于外部的标签信息。通过对同一样本进行不同的数据增强操作，生成多个视图，并将这些视图作为正样本对进行对比学习。自监督对比的目标是使得同一样本的不同视图在特征空间中更加接近，而不同样本的特征则更加分散。

2.交叉熵的缺点

1. 对噪声和标签错误敏感：交叉熵损失函数在训练过程中对噪声和标签错误非常敏感。如果训练数据中存在噪声或标签错误，交叉熵损失函数可能会导致模型学习到

错误的特征表示，从而降低分类准确性。

2. 不具备鲁棒性：交叉熵损失函数在面对输入数据的扰动或变化时，模型的性能可能会下降。例如，当输入图像受到噪声、模糊或JPEG压缩等自然扰动时，交叉熵损失函数训练的模型可能表现出较低的鲁棒性。
3. 对网络权重敏感：研究表明，使用交叉熵损失函数训练的网络在权重固定的情况下，对网络的最后一层进行重新初始化会导致准确率大幅下降。这表明交叉熵损失函数训练的表示学习不够鲁棒，对权重的微小变化非常敏感。

3.对比学习与交叉熵的区别

1. 监督学习 vs 无监督学习：交叉熵损失函数是一种常用的监督学习损失函数，需要标签信息来指导模型的训练。而本文提出的"Supervised Contrastive Loss"是一种结合了对比学习的思想的监督学习损失函数，通过最大化正样本之间的相似性和最小化负样本之间的相似性来训练模型。
2. 正负样本对比：交叉熵损失函数主要关注模型输出与真实标签之间的差异，而"Supervised Contrastive Loss"则更加关注正样本之间的相似性和负样本之间的差异。通过引入对比学习的思想，"Supervised Contrastive Loss"可以更好地学习到样本之间的相似性和差异性。
3. 多个正样本 vs 单个正样本：交叉熵损失函数通常只考虑单个样本与真实标签之间的差异，而"Supervised Contrastive Loss"可以处理多个正样本的情况。它通过对所有正样本进行求和或求平均来计算正样本之间的相似性。

4.对比学习的特点

对比损失函数的工作原理是通过将同一类别的样本映射到相近的表示空间中，同时将不同类别的样本映射到相远的表示空间中，从而实现样本的区分和分类。具体来说，对比损失函数通过计算样本之间的相似度或距离来衡量它们在表示空间中的关系。在训练过程中，对于每个anchor样本，选择与其同类别的positive样本和不同类别的negative样本进行比较