1.这两周我主要工作是精读了该篇论文.论文题目是jasper:an end-to-end convolutional neural acoustic Model.该篇论文是收到wav2letter的卷积方法的启发,构建了一个比较深且能够扩展的模型,配合上一个精心设计的残差连接设计,比较有效的正则化,和一个强大的优化器.相互配合上在LibriSpeech上取得不错的效果.

2.论文就主要讲四点主要内容,第一点是介绍它的高效的端到端的卷积神经网络模型,第二点是为了配合它的模型所精心选择的激活函数和标准化的方式,本论文为该模型选择了relu和batch normzation,并且通过试验证实选择relu和batch normzation的有效性.第三点是介绍了一个训练优化模型的优化器,叫做NovoGrad,是著名的优化器Adam的一种变体,具有占用内存小的特点.最后通过最后的实验,在LinriSpeech上取得的一个state-of-art的不错的错误率.

3.首先,就是要讲一下这个jasper的具体的结构.左边的图是它的标准结构.因为它是一个声学模型,首先需要提取声频数据的特征.这里最下面用的是mel-filterbank特征.从20ms的窗口计算10ms的重叠,输出每一帧字符的概率分布.

jasper有一个块结构:就比如该图中有B个块,每一个块中还有R个子块.每一个子块都有相应的操作:一个一维的卷积,一个批处理,再加上relu和dropout.一个块中的子块都有相同数量的输出通道.然后就是有一个残差操作,在一个大块中的输入和通过一个残差连接连接到最后一个子块上.这个残差连接首先会通过一个1*1的卷积投影,一考虑不同的数量的输入和输出.然后通过batch norm批处理和归一化层.该批处理化的输出加上最后一个子块的批处理的输出,最后通过relu和dropout产生整个大块的输出.jasper使用这种块结构的主要目的是为加速让gpu推断.

而后就是说,除了它中间的块外.它两边也是有一些其他的块,这些块也都是卷积块,一个预处理块和三个后处理块.

一个10*5的jasper的结构如该表所示,从上到下以此运转,每个块的输出通道会逐渐的变大.

最后讲一下,这个右边的是jasper的变体,叫做dense Residual,它相对于普通的jasper来说,就是加了更多的残差连接,把每一个块的输出都连接加到后面的块的输入中.

4.这一部分讲到标准化和激活的选择.这个实验比较了不同的normalization:batch norm, weight norm,lyer norm.不同的线性激活的单元:rely,clipped rely,leaky rely.不同的门控线性单元:GLU,GAU(gated activation)实验的对象包含一个大的jasper模型和一个小的jasper模型.在小的模型上发现,使用layer norm和GAU上表现最好.在大的模型上实验发现,使用batch norm和relu的效果远高于其他的方法.之后要使用大一点的模型居多,所以说改论文应用batch norm 和relu体系.

右上方图表,在批处理的期间,该论文使用序列掩码从平均值和方差计算排除填充值,还在卷积运算前使用了mask,使批处理中隐蔽的均值和方差.该表格显示在卷积前使用Mask可以得到好的表现,但是如果同时使用Mask和批处理的话效果不是很好.

5.论文还论证了使用残差连接的必要性,分别测试了基础版本的residual和dense,还有进一步的jasper的变体denseNet和denseRet,它们两非常相似,都将每个块的输出连接到后面的块的输入.该表格表明,这四种方法表现效果差不多,使用DenseRnet的表现最好,但是下面两个的成长参数需要更深的模型,所以还是选择dense residual.

6.最后,因为该论文提出的是一个声学模型,如果要将它用入测试的话,还是需要选择一个语言模型用于测试.该论文使用的是神经Transformer-XL模型进行实验.该图标展示这个Transformer-XL LM 复杂度和这错误率有很强的相关性.

7.最后的最后,该文章引进了一个新的优化器.叫做Novograd.他是我们平时常用的Adam优化器的变体,都有使用两个力矩来优化权重.只是它在计算第二个力矩的时候是按层计算的而不是按权重来计算的.他与Adam相比.它减少了内存的消耗.

它在每一个时间步t,NovoGrad计算随机梯度g_t^l,然后计算每一个层的二阶矩v_t^l.再计算一阶剧m_t^l之前,先用二阶矩v_t^l对梯度g_t^l进行缩放.如果说用到了L2正则化,就将权重衰减d*w_{t-1}加在后面,最后,利用学习率a_t计算新的权重.

结果:

最后是在不同数据集上的实验结果,首先是在著名的LIbriSpeech上的结果,Jasper以优异的表现的效果,在Test-clean子集上实现了state-of-art的效果.然后是在WSJ(华尔街日报)和Hub500上进行了测试结果,总体说,jasper的水准在一线水平.总体来说,jasper对于训练和推断是非常高效的,是可以作为一个很好的参考方法.也许可以在此基础上可以探索一下更加复杂的正则化、损失函数、语言模型和优化策略可能是我们之后可以探索的方向.