情感分析 实验报告

计 76 陈之杨 2017011377

1 实验目的

使用神经网络模型实现新浪新闻情感八分类。

2 模型建立

2.1 词向量

笔者使用了预训练的词向量作为模型使用的词向量[1]。训练语料为搜狗新闻，维数为 300。

实验中，可以将词向量设置为可训练以实现词向量的微调（fine-tune），但笔者测试发现这样做对训练结果并没有产生显著的影响，故所有实验中词向量均为不可训练，这样会大大提升训练效率。

此外，由于 300 维的词向量对本次作业来说规模较大，可以使用主成分分析（PCA）对词向量降维。但笔者在测试时同样没有发现显著的影响，故最终实验没有采用该方法。

2.2 批量训练

笔者使用分批训练（Minibatch training）的随机梯度下降（SGD）算法进行模型训练。由于训练集不同新闻的长度不同，笔者使用零填充（Padding）的方法，将所有样本补长至 800 进行训练。

2.3 全连接网络基线

笔者使用了全连接网络作为基线模型。模型结构为：

- 词向量嵌入层
- 向量化层（Flatten）

[1] https://github.com/Embedding/Chinese-Word-Vectors

- 全连接层（size = 50，激活函数为 Relu 函数）
- 全连接层（size = 10，激活函数为 Softmax 函数）

2.4 全连接层的低秩分解

对于 $\mathbb{R}^n \to \mathbb{R}^m$ 的全连接层，转移矩阵的大小为 $n \times m$。但通常神经网络提取出来的特征，我们无须使用满秩的线性变换去处理。将 $n \times m$ 的矩阵分解为两个 $n \times r$ 和 $r \times m$ 的矩阵的乘积 $(r < n, m)$，往往可以大幅度减少模型参数量，但不会明显影响模型的表示能力，也一定程度上抑制过拟合现象。

实验中，笔者将大小为 50 的全连接层的变换矩阵分解为 $n \times 10$ 和 10×50 的两个矩阵的乘积，训练时间减少了约一半，但模型性能没有出现过度的下降。

2.5 卷积神经网络

对于卷积网络模型，我们将 d 维的词向量视为 d 个通道，使用不同大小的卷积核去捕捉不同长度的特征，然后使用池化（Pooling）对每个通道降维，使用一个全连接层进行分类。模型结构如下：

- 词向量嵌入层
- 卷积层（kernel size = $2, 3, 4, 5$）
- 池化层
- 连接层（Concatenate，将不同卷积核提取的特征连接起来）
- 向量化层（Flatten）
- 全连接层（size = 10，激活函数为 Softmax 函数）

2.6 Dropout

Dropout 是一种简单而常用的抑制模型过拟合的方法。在训练过程中，按照一定比率随机将神经元置零，并将正常工作的神经元的输出放缩至正常的输出量级。

在实验中，使用了 Dropout 的卷积网络结构并无区别，只是在池化层后添加了一个 Dropout 层。

2.7 循环神经网络（SimpleRNN, GRU, LSTM）

循环神经网络可以较好地处理可变长的序列信息。由于普通的循环神经网络在处理较长的信息时，由于反向传播时梯度的传递需要多次乘同一个权值矩阵，存在梯度爆炸或梯度消失的问

题，因此笔者还使用了门控循环单元（GRU）和长短时记忆（LSTM）模型进行对比。LSTM 引入了遗忘门、输入门和输出门来控制新信息的进入和老信息的遗忘，避免梯度消失或梯度爆炸的发生。GRU 则是一种常用的 LSTM 的变体。

模型结构为：

- 词向量嵌入层
- 循环层（SimpleRNN/GRU/LSTM，size = 50）
- 全连接层（size = 10，激活函数为 Softmax 函数）

3 实验结果

3.1 超参数

实验中使用了动量加速的随机梯度下降（SGD）优化器进行训练。学习率（Learning rate）为 10^{-3}，L2 正则化权值衰减率（Weight decay）为 10^{-5}，动量加速参数为 0.9，批训练大小（Batch size）为 100。

事实上，笔者简单测试了不同超参数对训练的影响。权值衰减（Weight decay）对抑制过拟合并没有起到肉眼可见的效果，动量加速的效果比起普通的 SGD 也没有显著的加速。只有调整学习率（Learning rate）对训练过程有显著的影响。将学习率设得更大（如 10^{-2}）确实大大提升了模型的训练速度，但训练时也会出现明显的波动（如正确率突然下降到 20%）。考虑到这点，笔者将学习率设定得相对较小以求稳妥。

3.2 模型比较

各模型在测试集上的分类结果如下表所示。

模型	损失函数	准确率	F1-score	相关系数
全连接网络	1.5049	56.64%	0.2154	0.5363
低秩分解的全连接网络	1.6673	55.70%	0.2102	0.5188
卷积网络（无 Dropout）	1.2379	59.16%	0.2665	0.5742
卷积网络（Drop rate = 0.25）	1.1900	60.37%	0.2897	0.5858
卷积网络（Drop rate = 0.5）	1.2146	59.69%	0.2508	0.5804
循环网络（SimpleRNN）	1.5890	47.08%	0.0829	0.4231
循环网络（GRU）	1.6883	47.62%	0.0806	0.4239
循环网络（LSTM）	1.5749	47.76%	0.0808	0.4307

3.3 讨论

总的来看，神经网络在情感分类问题上的效果并不出众，最高也只能达到约 60% 的准确率。F1-score 的值则更低，这主要是因为训练集中情感的分布很不均匀，导致一些很少被预测到的情感严重拉低了分值。

CNN 比全连接网络有明显的优势，尽管 CNN 的参数量远小于全连接网络，但卷积层能够有效提取局部特征，更容易学到有用的信息。此外，适当使用 Dropout 抑制过拟合，可以提高 CNN 的性能。若置零率过高，也会拉低模型收敛的速度。

让人诧异的是，本该最适合处理序列问题的 RNN 反而性能最差，甚至不如全连接网络。笔者推测，这是由于**在情感分析问题中，上下文信息不太重要，而一些感情色彩强烈的词语对分类起到至关重要的作用**。这也可以解释擅长提取局部特征的 CNN 为何有较好的表现。此外，比起标准 RNN，GRU 和 LSTM 虽有性能提升，但并不明显。笔者推测这只是由于 GRU 和 LSTM 有更多的参数。梯度爆炸/消失也许并不是 RNN 性能较差的主要原因。

4 思考题

(1) 实验训练什么时候停止是最合适的？简要陈述你的实现方式，并试分析固定迭代次数与通过验证集调整等方法的优缺点。

理论上说，模型在刚开始出现过拟合时表现效果最好，因此应当在训练集正确率明显开始高于测试集正确率时，或连续若干轮（epoch）测试集正确率不上升时停止训练。

但实际实验时可以发现过拟合并不意味着模型停止学习有用的信息，测试结果通常依然会（相比训练集结果）缓慢上升，并且还会出现明显的波动现象。由于随机梯度下降（SGD）本身就是利用数据集中的噪声对整个训练集的梯度做扰动，以期达到跳出局部最优点或鞍点的效果，出现波动后达到更优的解的现象是合理的。考虑到这点，以及为了公平比较不同模型的收敛速度，我采用了固定训练 100 个 epoch 的方法。固定轮数的缺点在于容易错过效果较好的结果，中途保存历史最优模型的 checkpoint 可以避免这一问题。

(2) 实验参数的初始化是怎么做的？不同的方法适合哪些地方？（现有的初始化方法为零均值初始化，高斯分布初始化，正交初始化等）

由于训练神经网络是一件困难的事，我们目前对优化神经网络没有全面的理解，参数初始化没有一个固定的标准。

参数初始化最重要的一点是破坏模型的对称性。如果同一层的参数被初始化成同一个值，可能导致它们的梯度相同而永远相同，这不利于学习复杂的函数关系。因此，除了偏置参数通常使用零初始化外，其它参数通常都使用各种随机初始化方法。通常参数都采用零均值或高斯分布初始化。但是在 RNN 中，通常采用正交初始化。单位正交矩阵的特征值模长为 1，可以避免训练初始时即出现梯度消失/爆炸的情况。

(3) 过拟合是深度学习常见的问题，有什么方法可以防止训练过程陷入过拟合。

我们可以在模型训练中加入各种正则化方法（Regularization）抑制过拟合。正则化方法是通过人为修改损失函数、数据集或模型结构，希望模型优先学出不容易过拟合的结果。

降低参数数量可以起到抑制过拟合的效果。在模型的表示能力差不多的时候，参数量大的模型更容易学习到训练集中的噪音，导致过拟合。卷积神经网络优于全连接网络的原因之一就在于共享权值，减少参数量，不易过拟合。实验中全连接网络在一个 epoch 后训练集正确率就远远超过了测试集，相比之下卷积网络的过拟合现象就相对较轻。此外，在全连接网络中使用低秩分解，在卷积网络中使用小卷积核叠加而非单层大卷积核等等，都是抑制过拟合的方法。

最常用的正则化方法是在损失函数中添加参数范数惩罚，常用的有 L_1 和 L_2 范数，其目的在于希望模型在相同的损失函数下优先学出较小的参数。L_2 范数惩罚在基于梯度的训练方法中等价于权值衰减（Weight decay），学出来的模型参数较为接近原点。而 L_1 范数惩罚会使学出的参数具备一定的稀疏性（Sparseness）。

除此之外，还可以在数据集进行正则化。例如，数据增强（利用原训练集，人为创造一些数据加入训练集中，增强模型泛化能力）、坐标平滑（给错误标签一个小常数 ϵ 的正确概率，避免模型的预测走向极端）等。

对模型的正则化，通常可以在模型中加入噪声。例如 Dropout 通过随机将神经元置零，使神经元得到充分训练，使模型更加鲁棒。

(4) 试分析 CNN，RNN，全连接神经网络（MLP）三者的优缺点。

全连接网络较为稠密，参数量大，收敛较快，但同时也很容易陷入过拟合。矩阵乘法易于并行，故训练效率高。

卷积网络通过共享权值降低了参数量，卷积运算可以有效地提取局部特征，往往能够得到较高的正确率。但是卷积运算量较大，训练效率比起其它网络要低很多。

全连接网络和卷积网络通常只能处理固定大小的输入,对于序列信息必须通过补零(Padding)达到固定长度，如果训练集序列长短参差不齐，会导致存储空间的浪费，循环网络则专为处理序列信息设计，但由于大量的循环存在，训练速度较慢。

5 总结

通过这次实验，笔者对使用不同的神经网络处理自然语言的分类问题有了基本的认识。情感分类是自然语言处理中的重要问题，但 60% 的正确率显然离臻于艺术的（State-of-the-art）结果还有一定的距离，因此还需要对情感分类问题做进一步的探索。