**简单神经网络在语言模型中的应用**

**摘要：**

神经网络对非线性特征具有拟合能力，理论上来讲一个二层神经网络可以拟合一切函数；语言模型中的任务往往涉及到简单的分类任务，这种普通的神经网络结构能否用于语言模型中，如何训练神经网络，这值得思考和研究。

1. 前馈神经网络的发展：

神经网络通过对脑内神经元信号传递的模拟，计算输入的加权和与偏置、再通过非线性激活函数（如逻辑回归中常用的sigmoid函数）得到神经元输出，以此实现神经网络基本单元的功能。特别地，神经网络中的激活函数有很多，除sigmoid外，tanh、relu都是不错的选择，但**非线性激活函数为什么必要，**依旧值得思考。

传统的单个感知机模型能很好地实现and和or，但却无法实现xor，原因便在于感知机以线作为决策边界，没有非线性的激活。而xor并不是线性可分函数，因而需要考虑单元分层网络（即添加隐藏层来学习一些有用的表示）。

基于以上，结合神经元的启发和分层网络的思想，前馈神经网络诞生了——又叫多层感知机。单层神经网络用语二元逻辑回归时，我们受simoid启发设计了**softmax策略**，后又发展出多层神经网络，对多分类输出做softmax处理，还可以适当优化结构，**移除偏置单元**。

1. 前馈网络在nlp任务中的应用：

这里我们不研究强大的神经架构，而是从底层建模出发，考虑简单神经网络在语言模型中应用的可能性和方向方法。以文本分类和语言建模两任务为例：

1. 文本分类中的情感分析：

我们完全可以像逻辑回归那样，输入文本二进制特征，输出0/1，即positive or negative。于是，情感特征表示是我们面临的一大问题。

我们可以定义6个输入变量：文本中positive词的数量、negative词的数量、文本中是否有否定词、**文本中第一人称代词和第二人称代词的数量**、文本中是否有叹号、文本字数的对数，这些变量可以一定程度上反映情感特征。接着我们便可以使用一个简单用于分类的前馈神经网络，添加一个隐藏层来允许非线性的交互，但实验之前我们并不能预知它对模型性能的提升产生怎样的影响，因为我们只是做可能性与方向方法的研究。

更进一步，我们可以使用嵌入作为输入特征。然而，文本长度并不总是相同，如果设置一个固定长度的嵌入来表示需要考虑训练数据中文本的最大长度，以此来保证最大限度地适应未来可能不同的输入数据，长则截断短则补零。另一个可能得解决方案是创建一个句子的嵌入，这需要用到所有单词嵌入的平均值、最大值等数据。

当我们输出不想局限于0/1类，而是有更多的情感表示词汇时，就需要用到softmax层来计算最大概率的类别作为输出类别。

1. 神经语言模型：

神经语言模型要实现的主要任务在于语言建模，即根据给定的历史、计算序列中的下一个单词的概率。现今有一些强大的神经网络技术如transformer可以将这一任务实现得很好，但简单的前馈神经语言模型也几乎可以做得一样好。然而，任务要处理任意长度的序列，这和神经网络模型固定神经元个数的输入不一样，为了解决这一难题，我们使用滑动窗口来确保固定长度可以适应模型输入。

神经语言模型会比**N-gram**语言模型在测试数据上有更好的效果，原因在于神经语言模型可以利用嵌入的相似性来概括和预测未见过的搭配，更好地预测下一个词的发生。

1. 神经网络的训练：

问题抽象完成后，我们需要明白神经网络如何训练：即对每一对训练数据（x, y），先通过网络层正向计算找到估计的最大概率类，然后反向计算更新网络中的权重。而衡量真实值与预测值之间差距的损失函数在其中显得尤为重要，因为我们是向着最小化损失函数去做优化和更新的，如在训练中经常使用的用于逻辑回归的交叉熵损失。同样，权重如何更新也很重要，如使用损失函数关于权重的导数（通常由链式法则计算得到）的梯度下降算法。

特别地，反向计算需要每层每个权重关于损失的导数，而损失只在网络最末端计算，因此我们引入依赖于计算图的误差反向传播。通过链式法则，给定计算图和其中所有函数的导数，我们就可以自动计算损失相对于浅层神经元中权重的导数。

**结论：**

神经网络用于语言模型发展前景广阔，即使是简单的模型结构也能适应nlp一些任务的要求，值得深入研究。

**问题：**

1. 在使用嵌入作为输入特征时，除最大长度外，如何衡量最合适的嵌入维度，一定程度上更大是否鲁棒性更强，更小是否更精确？
2. 在神经网络模型处理任意长度序列时，如何选择合适的滑动窗口大小？是否需要根据不同任务特征进行区别？有没有普适的窗口大小可以通用于大部分任务？
3. 在大模型技术的加持下，未来神经语言模型的发展是否进入“瓶颈”？

**参考文献：**

1. Dan Jurafsky, James H. Martin. Speech and Language Processing (3rd ed. draft)
2. 罗浩然,杨青.基于情感词典和堆叠残差的双向长短期记忆网络的情感分析[J].计算机应用,2022,42(4):1099-1107.
3. Andreas Grivas,Nikolay Bogoychev,Adam Lopez.Low-Rank Softmax Can Have Unargmaxable Classes in Theory but Rarely in Practice[Z].arxiv,2022.
4. Deageon Kim.Text Classification Based on Neural Network Fusion[J].TEHNIČKI GLASNIK,2023,17(3):359-366.
5. Yao Gan,Yanyun Fu,Deyong Wang,et al.A novel approach to attention mechanism using kernel functions: Kerformer[J].FRONTIERS IN NEUROROBOTICS,2023,17.
6. Yuyao Li,Xueyi Wen,Xingyu Liu.Research on Keywords Variations in Linguistics Based on TF-IDF and N-gram[J].JOURNAL OF COMPUTING AND INFORMATION TECHNOLOGY,2022,30(3):193-204.
7. Dina Oralbekova,Orken Mamyrbayev,Mohamed Othman,et al.Contemporary Approaches in Evolving Language Models[J].APPLIED SCIENCES,2023,13(23).
8. Jie Song,Yixiao Liu,Yunhua Qu.N-gram Language Model for Chinese Function-word-centered Patterns[J].JOURNAL OF COMPUTING AND INFORMATION TECHNOLOGY,2023,31(1):39-55.