2023-2024 学年第一学期《自然语言处理》复习提纲 v2.0

（2024 年 1 月 9 日）

《自然语言处理》课程期考核采用闭卷考试的形式，主要通过单选题、简答题、编程题等多种题型考核学生对深度学习的基本概念、基本理论及相关技术的掌握情况。

第一部分　单向选择题

【考试时，此部分分值占比 20%，共计 20 道小题】

（部分题目故意没有给出答案，目的是希望大家通过自己查找答案加深理解，避免偷懒造成的死记硬背）

1、下列哪个操作能被用于 tensor 的维度顺序交换()

A. torch.reshape()　B. torch.permute()

C. torch.view()　D. torch.squeeze()

2、下列哪些技术能被用于计算两个词向量之间的距离？（ C ）

A. 词形还原　B. 条件随机场

C. 余弦相似度（Cosine Similarity）D. N-grams

3、文本语料库的可能特征是什么？（ D ）

A. 文本中词计数 B. 词的向量标注

B. C. 词性标注（Part of Speech Tag）D. 以上所有

4、给定一个列表 a=[2,3,5,7,8,9],则以下操作正确的是（）

A.　file(lambda x:x**2,a)

B.　filter(lambda x: if x**2,a)

C.　sorted(lambda x:x**2,a)

D. filter(lambda x:x**2,a)答案：D

5、给定 PyTorch 中的张量 a=torch.tensor([[53,77,25,16],[69,50,15,9]])，则 torch.argmax(a,1)的输出结果是以下哪一项？（　）

A. tensor([1, 0, 0, 0])　B. tensor([1, 0]　C. tensor([34, 29])　D. tensor([29, 34, 25, 16])

6、下列哪项是关键词归一化技术？

A. 隐马尔科夫模型　B. 词性标注（Part of Speech）

C. 命名实体识别（Named Entity Recognition）

D. 词形还原（Lemmatization）

答案：D

7、下面哪个是 NLP 应用案例？

A. 从图像中检测物体　B. 面部识别

C. 语音生物识别　D. 文本摘要

答案：D。

8、给定一个 numpy 数组 a=np.array([[15,31,9,25],[33,20,11,18]]),则 np.max(a,1)的输出结果是（）

A. array([33, 31, 11, 25])　B. array([31, 33])

C. array([33])　D. 33

9、给定一个张量 a=torch.arange(9)，则切片操作 a[5:]的输出结果是（ ）

A. tensor([5, 6, 7, 8])　B. tensor([5, 6, 7])

C. tensor([5])　D. tensor([0,1,2,3,4])

10、以下哪种操作不属于 PyTorch 中的数据预处理操作（　）

A. transforms.ToTensor( )　B. transforms.Norm( )

C. transforms.CenterCrop( )　D. transforms.Grayscale( )

11、单个神经元本质上是一个（　）模型

A. KNN　B. 朴素贝叶斯　C. 逻辑回归　D. 集成学习

12、假如有一张文件名为 hpu 的图像(图像类型为.jpg 格式),其所在的绝对路径是 img_dir，则以下图像操作中哪一项操作是正确的？（　　）

A. Image.open('img_dir\hpu.jpg')　　B. torch.open(r'img_dir\hpu.jpg')

C. plt.imread((r'img_dir\hpu.jpg')　　D. np.imread((r'img_dir\hpu.jpg')

13、给定一张彩色图片，将其用 matplotlib（plt）读入后,如果需要颠倒其通道顺序，则以下操作哪一项是正确的:（　　）

A. plt.imshow(img[:])　　B. plt.imshow(img[:,:,-1])

C. plt.imshow(img[::-1])　　D. plt.imshow(img[:,:,::-1])

14、在文本挖掘中，可以使用以下哪项命令完成将文本转换为 tokens，然后将其转换为整数或浮点向量的操作？

A. CountVectorizer　　B. TF-IDF

C. 词袋模型（Bag of Words）　D. NERs

答案：A

解析 CountVectorizer 可帮助完成上述操作，而其他方法则不适用。

15、下列哪种词嵌入支持上下文建模（Context Modeling）？

A. Word2Vec　　B. GloVe

C. BERT　　D. 以上所有

答案：C

解析：只有 BERT（Bidirectional Encoder Representations from Transformer)支持上下文建模。

16、下列哪种嵌入方式支持双向上下文（Bidirectional Context）？

A. Word2Vec　　B. BERT　　C. GloVe　　D. 以上所有

答案：B

解析：只有 BERT 支持双向上下文。Word2Vec 和 GloVe 是词嵌入，它们不提供任何上下文。

17、下列哪种词嵌入可以自定义训练特定主题？

A. Word2Vec　B. BERT　C. GloVe　D. 以上所有

答案：B

18、对于 PyTorch 中的张量 a，假如 a.shape=torch.Size([32, 8])，则以下哪一个输出结果是正确的？（　）

A. a.unsqueeze(1).shape 的输出结果为：torch.Size([1,32, 8])

B. a.view(-1,1,a.size(0),a.size(1)).shape 的输出结果为：torch.Size([1, 1, 32, 8])

C. a.squeeze(0).shape 的输出结果为：torch.Size([8])

D. a.squeeze(0).shape 的输出结果为：torch.Size([32])

19、以下哪一项不属于深度学习中的激活函数？（　）

A. torch.acfun()　B. torch.relu()　C. F.leaky_relu()　D. torch.sigmoid()

20、关于 python 中的生成器、迭代器和可迭代对象，以下哪个说法是正确的（　）

A. 生成器一定是迭代器　B. 迭代器一定是生成器

C. 可迭代对象一定是迭代器　D. 可迭代对象一定是生成器

21、对于一个给定的 token，其输入表示为它的 token 嵌入、段嵌入（Segment Embedding）、位置嵌入（Position Embedding）的总和

A. ELMo　B. GPT　C. BERT　D. ULMFit

答案：C

解析：BERT 使用 token 嵌入、段嵌入（Segment Embedding）、位置嵌入（Position Embedding）。

22、从左到右和从右到左训练两个独立的 LSTM 语言模型，并将它们简单地连接起来

A. GPT　B. BERT　C. ULMFit　D. ELMo

答案：D

解析：ELMo 尝试训练两个独立的 LSTM 语言模型（从左到右和从右到左），并将结果连接起来以产生词嵌入。

23、用于产生词嵌入的单向语言模型

A. BERT　B. GPT　C. ELMo　D. Word2Vec

答案：B

24、给定一张大小为 256*256 的图像，经过一层卷积核大小为 5*5 的卷积层后，假如没有 padding 操作，则其输出的特征图大小为（　）

A. 256*256　B. 254*254　C. 252*252　D. 250*250

25、以下哪种 NLP 模型的准确性最高？

A. BERT　B. XLNET　C. GPT-2　D. ELMo

答案：B

解析：XLNET 在所有模型中都给出了最好的准确性。它在 20 个任务上都优于 BERT，在情感分析、问答、自然语言推理等 18 个任务上都取得了顶尖的结果。

26、在神经网络中，以下哪种技术用于解决过拟合？（ D ）

A.Dropout　B.正则化　C.批归一化　D.所有

27、以下哪种操作不属于 PyTorch 中的数据预处理操作（ B ）

A. transforms.ToTensor( )　B. transforms.Norm( )

C.transforms.CenterCrop( )　D. transforms.Grayscale( )

28、下列哪一项在神经网络中引入了非线性？（ B ）

A.随机梯度下降 ； B.修正线性单元（ReLU） ；C.卷积函数 ；D.以上都不正确

29、以下不是 LSTM 结构的是（　）

A. 输入门　B. 遗忘门　C. 输出门　D. 梯度门

30、长短期记忆网络属于（ B ）网络的一种。

A. 线性回归　B. 深度前馈神经网络　C. 卷积神经网络　D. GAN 网络

31、以下不是池化层的作用的是（ B ）

A. 特征降维，避免过拟合　B. 空间不变形

C. 提取图片的局部特征　D. 减少参数，降低训练难度

32、卷积神经网络中的卷积层的主要作用是（ D ）

A. 大幅度降低参数量级（降维）　B. 用来输出想要的结果

C. 读入数据　D. 负责提取图片中的局部特征

33、对于梯度爆炸，可以用（　）或挤压渐变来解决。

A. 截断　　B. 退化　　C. 扩大范围　　D. 补充

34、深度学习中对于回归问题通常用到以下哪种损失函数（　　）

A. MSE　　B. Cross Entry　　C. 0-1 损失函数　　D. BCE

35、torchvision 提供了一个常用的函数（　　），能将多张图拼接在一个网络中

A. make_grid( )　　B. make_gether( )　　C. gether_grid( )　　D. make_plot( )

第二部分 问答题【此部分分值占比 50%左右】

1、请详细描述循环神经网络（RNN）、长短期记忆网络（LSTM）各自的原理，画出其网络结构，并阐述各自的优缺点。

参考答案：

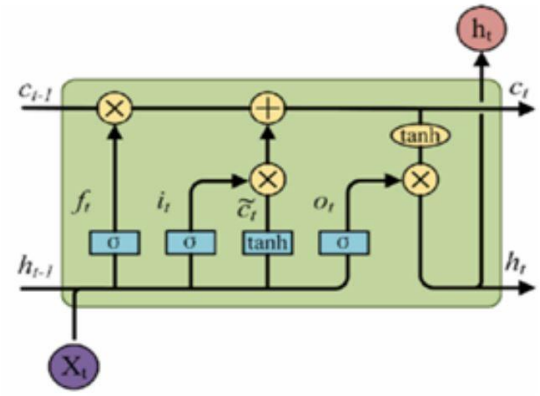
（1）循环神经网络（RNN）

普通的 CNN 只能单独的处理一个个的输入，前一个输入和后一个输入是完全没有关系的。但是某些任务需要能够更好的处理序列的信息，即前面的输入和后面的输入是有关系的。RNN 对具有序列特性的数据非常有效，它能挖掘数据中的时序信息以及语义信息。首先一个简单的循环神经网络由输入层、隐藏层和输出层组成。当前时间步 t 的隐藏状态 Ht 将参与计算下一时间步 t+1 的隐藏状态 Ht+1。而且 Ht 还将送入全连接输出层，用于计算当前时间步的输出 Ot。

RNN 的梯度爆炸和消失问题：RNN 并不能很好的处理较长的序列。一个主要的原因是，RNN 在训练中很容易发生梯度爆炸和梯度消失，这导致训练时梯度不能在较长的序列中一直传递下去，从而使 RNN 无法捕捉到长距离的影响。梯度消失问题是当梯度随着时间的推移传播时梯度下降，梯度值变得非常小，就不会继续学习。

（2）长短期记忆网络（LSTM）

长短期记忆(LSTM)是一种特殊结构的 RNN，主要是为了解决长序列训练过程中的梯度消失和梯度爆炸问题。相比于普通的 RNN，LSTM 能够在更长的序列中有更好的表现。能够解决在 RNN 网络中梯度衰减的问题。LSTM 中引入了记忆单元(cell)，其设计目的是用于记录附加的信息。为了控制记忆元，加入了输入门、输出门和遗忘门。



遗忘门：决定应丢弃或保留哪些信息。来自前一个隐藏状态的信息和当前输入的信息同时传递到 Sigmoid 函数中去，输出值介于 0 和 1 之间，越接近 0 意味着越应该丢弃，越接近 1 意味着越应该保留。

输入门：用于更新细胞状态。首先将前一层隐藏状态的信息 Ht-1 和当前输入的信息 X~t 传递到 sigmoid 函数中去。将值调整到 0~1 之间来决定要更新哪些信息。0 表示不重要，1 表示重要；其次还要将前一层隐藏状态的信息 Ht-1 和当前输入的信息 Xt 传递到 tanh 函数中去，创造一个新的侯选值向量（候选记忆）。最后将 sigmoid 的输出值 It 与 tanh 的输出值(候选记忆)相乘，sigmoid 的输出值将决定 tanh 的输出值中哪些信息是重要且需要保留下来的。

输出门：输出门用来确定下一个隐藏状态的值，隐藏状态包含了先前输入的信息。首先，我们将前一个隐藏状态 Ht-1 和当前输入 Xt 传递到 sigmoid 函数中，然后将新得到的细胞状态 Ct 传递给 tanh 函数；最后将 tanh 的输出与 sigmoid 的输出 Ot 相乘，以确定隐藏状态应携带的信息 Ht。再将隐藏状态作为当前细胞的输出，把新的细胞状态 Ct 和新的隐藏状态 Ht 传递到下一个时间步长中去。

LSTM 和 RNN 相比能解决梯度消失的问题：对于 RNN 来说，对于每一个时间点，记忆单元 cell 里的信息都会被覆盖掉，但是 LSTM 里不一样，它是把原来 memory 里面的值乘上一个值在加上 input 的值放到 cell 里面，它的 memory 和 input 是相加的，所以不像 RNN 在每个时间点都会被覆盖掉，只要前一时刻的信息一被 format 掉影响就消失了，但是在 LSTM 里的影响一直会存在，除非遗忘门 Forget Gate 把 memory 里的信息清洗掉。

2、请阐述自然语言处理中词表示方法之一的独热编码，为什么要使用的热编码？独热编码有什么优缺点。

参考答案：

独热编码，即 One-Hot 编码，又称一位有效编码，其方法是使用 N 位状态寄存器来对 N 个状态进行编码，每个状态都有它独立的寄存器位，并且在任意时候，其中只有一位有效。比如颜色特征有 3 种：红色、绿色和黄色，转换成独热编码分别表示为（此时上述描述中的 N=3）：001, 010, 100。（当然转换成 100, 010, 001 也可以，只要有确定的一一对应关系即可）。

在机器学习算法中，一般是通过计算特征之间距离或相似度来实现分类、回归的。一般来说，距离或相似度都是在欧式空间计算余弦相似性得到。对于上述的离散型颜色特征，1、2、3 编码方式就无法用在机器学习中，因为它们之间存在大小关系，而实际上各颜色特征之间并没有大小关系，比如非要说红色>绿色是不合理的。所以，独热编码便发挥出了作用，特征之间的计算会更加合理。

独热编码的优点：

避免了顺序关系：独热编码将每个类别表示为一个独立的向量，避免了引入类别之间的顺序关系。

节省存储空间和计算成本：独热编码生成的向量是稀疏的，仅有一个元素为 1，其余为 0。这种稀疏表示适用于大多数机器学习算法，节省了存储空间和计算成本。

可处理不同数量的类别：独热编码可以处理不同数量的类别，因为每个类别都被表示为固定长度的向量，与其他类别的数量无关。

独热编码的缺点：

维度灾难：对于具有大量不同类别的特征，独热编码可能导致维度灾难，即数据的维度迅速增加。这可能使得模型更难以训练，尤其是在数据量有限的情况下

不适用于高基数特征：当类别数量非常大时，独热编码可能变得不切实际。

可能导致冗余性：当特征之间存在相关性时，独热编码可能导致冗余性，因为每个类别的信息都是相互独立的。

3、什么是数据规范化（Normalization），我们为什么需要它？

答：Normalization 的中文翻译一般叫做“规范化”，是一种对数值的特殊函数变换方法，也就是说假设原始的某个数值是 x，套上一个起到规范化作用的函数，对规范化之前的数值 x 进行转换，形成一个规范化后的数值。

规范化将越来越偏的分布拉回到标准化的分布,使得激活函数的输入值落在激活函数对输入比较敏感的区域,从而使梯度变大,加快学习收敛速度,避免梯度消失的问题。

按照规范化操作涉及对象的不同可以分为两大类：

一类是对第 L 层每个神经元的激活值 进行 Normalization 操作，比如 BatchNorm/ LayerNorm/ InstanceNorm/ GroupNorm 等方法都属于这一类；

另外一类是对神经网络中连接相邻隐层神经元之间的边上的权重进行规范化操作，比如 Weight Norm 就属于这一类。

有了这些规范目标，通过具体的规范化手段来改变参数值，以达到避免模型过拟合的目的。

4、全连接神经网络和循环神经网络有什么区别？循环神经网络 (RNN) 有哪些应用？

参考答案：

全连接神经网络信号从输入到输出沿一个方向传播。没有反馈回路；网络只考虑当前输入。它无法记住以前的输入（例如 CNN）。

循环神经网络的信号双向传播，形成一个循环网络。它考虑当前输入和先前接收到的输入，以生成层的输出，并且由于其内部存储器，它可以记住过去的数据。

RNN 可用于情感分析、文本挖掘等，可以解决时间序列问题，例如预测一个月或季度的股票价格。

5、请阐述自然语言处理中的 word2vec 模型？

参考答案：

word2vec 是一群用来产生词向量的相关模型。这些模型为浅而双层的神经网络，用来训练以重新建构语言学之词文本。网络以词表现，并且需猜测相邻位置的输入词，在 word2vec 中词袋模型假设下，词的顺序是不重要的。训练完成之后，word2vec 模型可用来映射每个词到一个向量，可用来表示词对词之间的关系，该向量为神经网络之隐藏层。

word2vec 模型主要有 Skip-Gram 和 CBOW 两种，从直观上讲，Skip-Gram 是给定 input word 预测上下文，而 CBOW 是给定上下文，来预测 input word。总体上说，skip-gram 的训练时间更长，对于一些出现频率不高的词，在 CBOW 中的学习效果就不如 Skip-Gram，skip-gram 准确率更高。

CBOW 模型中 input 是 context（周围词）而 output 是中心词，训练过程中其实是在从 output 的 loss 学习周围词的信息也就是 embedding，但是在中间层是 average 的，一共

预测 V(vocab size)次就够了。skipgram 是用中心词预测周围词，预测的时候是一对 word pair，等于对每一个中心词都有 K 个词作为 output，对于一个词的预测有 K 次，所以能够更有效的从 context 中学习信息，但是总共预测 K*V 词。

6、如何解决梯度消失和梯度爆炸问题？

（1）梯度消失：根据链式法则，如果每一层神经元对上一层的输出的偏导乘上权重结果都小于 1 的话，那么即使这个结果是 0.99，在经过足够多层传播之后，误差对输入层的偏导会趋于 0。可以采用 ReLU 激活函数有效的解决梯度消失的情况，也可以用 Batch Normalization 解决这个问题。关于深度学习中 Batch Normalization 为什么效果好？（2）梯度爆炸根据链式法则，如果每一层神经元对上一层的输出的偏导乘上权重结果都大于 1 的话，在经过足够多层传播之后，误差对输入层的偏导会趋于无穷大可以通过激活函数来解决，或用 Batch Normalization 解决这个问题。

第三部分　编程题【此部分分值占比 30%左右】

此处不提供具体的编程题复习例题，编程题考核内容为课堂所讲案例、实验案例或布置的课程作业中截取的某一部分或其灵活变种。（比如如何设计一个网络模型，深度学习训练过程中的关键代码等等）

再次强调：大家在考试时，不用完全照抄上面的参考答案，消化后提炼重点部分阐述。另外，由于考试需要拉开成绩，故所给的复习提纲并未涵盖所有考试内容。平时给大家布置的作业和课堂上的编程案例也需要复习。