第一页

大家好，这几年深度学习领域这么火热，主要是因为他在两个领域有了重大突破，并且得到了广泛的应用。分别是自然语言处理和图像处理。在自然语言处理领域，神经网络实现了文本语义的理解，可以初步的生成各种类型的文本，例如诗歌，散文，小说等，还可以进行对话；而在图像处理领域呢，则可以实现人脸识别，自动绘图等功能。这一节课呢，我们就通过自然语言处理来一窥神经网络的神奇语言处理能力。

第二页

相信大家都有这样的经历，课余时间大家都在教师里打闹，突然有个同学喊了一句“老师来了”，教室里就突然安静下来了。简简单单的“老师”，“来了”这两个词组合到一起，就有了无比巨大的威力。我们人类天然就有感知语言的能力，那么如何使机器也拥有这种能力，从而实现人工智能呢？

第三页

我们接着再想一下，人类表达世界有几种方式：一种是语言，一种是图像，还有一种是声音；当我们的眼睛看到文字或图像，或者耳朵听到声音的时候，我们就能理解相应的含义，并做出对应的联想。除此之外了，人类还拥有另外一项能力，就是数学能力，数学可以很好的将所有其他学科的知识通用的表达出来，可以说数学是上帝的语言，也是人类之间，以及人类和宇宙对话的通用。所用，科学家想到了用数学了表达语言及图像的含义。

第四页

这里我们用一个简单的例子来看一下，怎么样用数学来表达词汇。有高三数学基础的同学应该都知道，数学有三种基本表达单元：标量、向量、和张量。

我们以“狗”这个单词为例子，如果用标量，我们可以给狗这个词唯一的数字表示1，那么其他的词就用其他数字来表示，从而可以将所有的词划分出来。但是这种方式，存在一个问题，就是一个词它的含义并不是固定的，它可以是多义的，在不同语境下的含义也可能不同。比如说狗可以表示一种动物，狗、猫等等；还有可以描述一个人，说这个人太狗了吧，意思是太怂了，太差劲了。那么如何表征这种多义性呢。向量就是很好的使用这种场景。向量的每个维度代表一种含义的权重，当某个词在这个含义上占比较大，就说明偏向此类的意思。 据说，chatGPT的每个词的向量维度是17\*\*\*

第五页

好了，既然知道向量可能很好的表达语言中的词，那么如何得到每个词的合适的向量表达呢？这节课我们介绍一种方法，word2vec，通过神经网络训练，得到各个词的向量表示。

我们先来将具体的操作方法，首先，我们准备一段预料，例如这里的“我爱杭州电子科技大学”，接着呢，我们用一个滑动窗口来对这段语句进行切分，这里的窗口大小为3，就得到了“我爱杭州”，“爱杭州电子”，“杭州电子科技”，“电子科技大学”,接着，我们将每段文本再拆分，把头和尾的词拿出来作为神经网络的输入，中间的字作为神经网络的输出，也就是用文本的前后两个字来预测中间的字。基于以上方法的神经网络模型叫做CBOW.(continuous bag of words)

第六页

有了上面的数据集之后呢，我们就可以训练神经网络模型了。CBOW模型的结构如这张图所示. 输入就是数据集当中的词的独热向量，假设我们的预料数据集中只包含两个词，那么就用一个六维独热向量来表示，所以输入层的维度为6，隐藏层的维度由我们自行设定，对应着最终训练出来的词向量的维度，我们这里设定为3，即最终得到的词向量的维度为3. 输入的所有单词共享同一个权重矩阵，即这里的v\*n维的矩阵；经过融合得到一个v维的向量，最后经过输出层和softmax层得到输出向量，将输出值与真实值对比计算得到损失值，接着反向传播来修正网络的参数，直到最终网络收敛。

第七页

现在，我们用一个笑例子来演示下具体的计算过程，首先，我们拿到出入向量的表示，我为\*\*\*，杭州为\*\*\*，将这两个向量输入网络，他们分别与权重矩阵相乘，得到隐藏层的输出。

第八页

接着我们将隐藏层的向量融合成一个三维列向量，将其与输出层矩阵相乘，并且经过softmax层处理，得到最终输出；最后计算损失函数，并且不断更新网络的权重系数。

第九页

当神经网络训练好了之后，隐藏层的权重系数就是训练好的词向量，如图所示，权重矩阵Wv\*n的没一列就对应一个词向量，例如爱就是\*\*\*，大学就是\*\*\*

第十页

以上就是词向量训练的基本思想，实际情况下，开发者是将输出的那个单词也拿到输入层，和窗口中的其他单词放在一起输入，然后输出层判别这样的输入组合是否是对的，这样就把预测单词问题变成了一个二分了问题，输出为0和1，分别代表是存在和不是存在这样词语组合。我们在语料库中通过滑动窗口得到的词语都是正确的组合，那么此时预测的输出值都为1，这对网络的训练是不利的。为了得到输出为0的预料，需要用到一种技术，叫做负采样。例如，我爱杭州，需要预测的是中间的爱字，现在把中间变成一个不可能的吃字，那么此时的预测标签就是0；通过这种方式，我们就可以向数据集中添加很多反例，从而使数据集更完备。

最后呢，我也给大家推荐一款非常优秀的自然语言处理python库，Gensim，基于gensim我们可以很容易的训练出我们自己数据集对应的词向量。

好了，以上就是本节课的全部内容，谢谢大家，我也会把ppt放在github上，欢迎大家自取。