作业：

**1.思考在自然语言处理中如何发掘模型的可解释性**

目前，深度学习在NLP领域中得到了广泛的应用，如BERT、XLNET等大型预训练语言模型，在自然语言处理下游任务中，只需要输入少量领域专用数据进行微调，就能够取得很好的效果，但是和计算机视觉一样，深度学习应用于自然语言处理领域同样存在可解释性问题。我认为，自然语言处理领域的可解释性，是模型输出的结果应该符合领域知识的预期，可以用语言学的术语和业务领域的语言进行解释，具体可以总结为三点。

1）基于语言特征

语言模型对输入词汇的“理解”，应该与人类的语言理解相近。例如人类认为，“兴奋”和“伤心”是一对反义词。语言模型中输出的词向量也应该通过向量计算，得到“兴奋”和“伤心”是反义词。当然这只是最简单的语言特征。但是，让模型的输出接近人类的语言输出结果，是更加有利于人类对模型内部的理解的，能够通过语言学知识对模型作出部分解释。

2）与领域知识有明确的因果关系

可解释性和应用场景强相关。应用场景的知识，决定着NLP的需求边界、任务类型、结果预期等内容，与场景知识之间有明确的因果关系，才是可解释的。

例如，在快递客服场景之中，针对“寄件延误”的投诉分类，消费者输入的文字可能有（“寄件的人怎么还不来啊”，“2个小时都还没过来寄件”），但模型都应该给出一致的结果（“取件延误”）。难点不在于判断是或否，而在于将消费者的意图识别为寄件延误，因此我认为，在具体业务中，意图识别的可解释性很重要。

3）应该追求处理结果的可解释性，而非NLP算法过程的可解释性

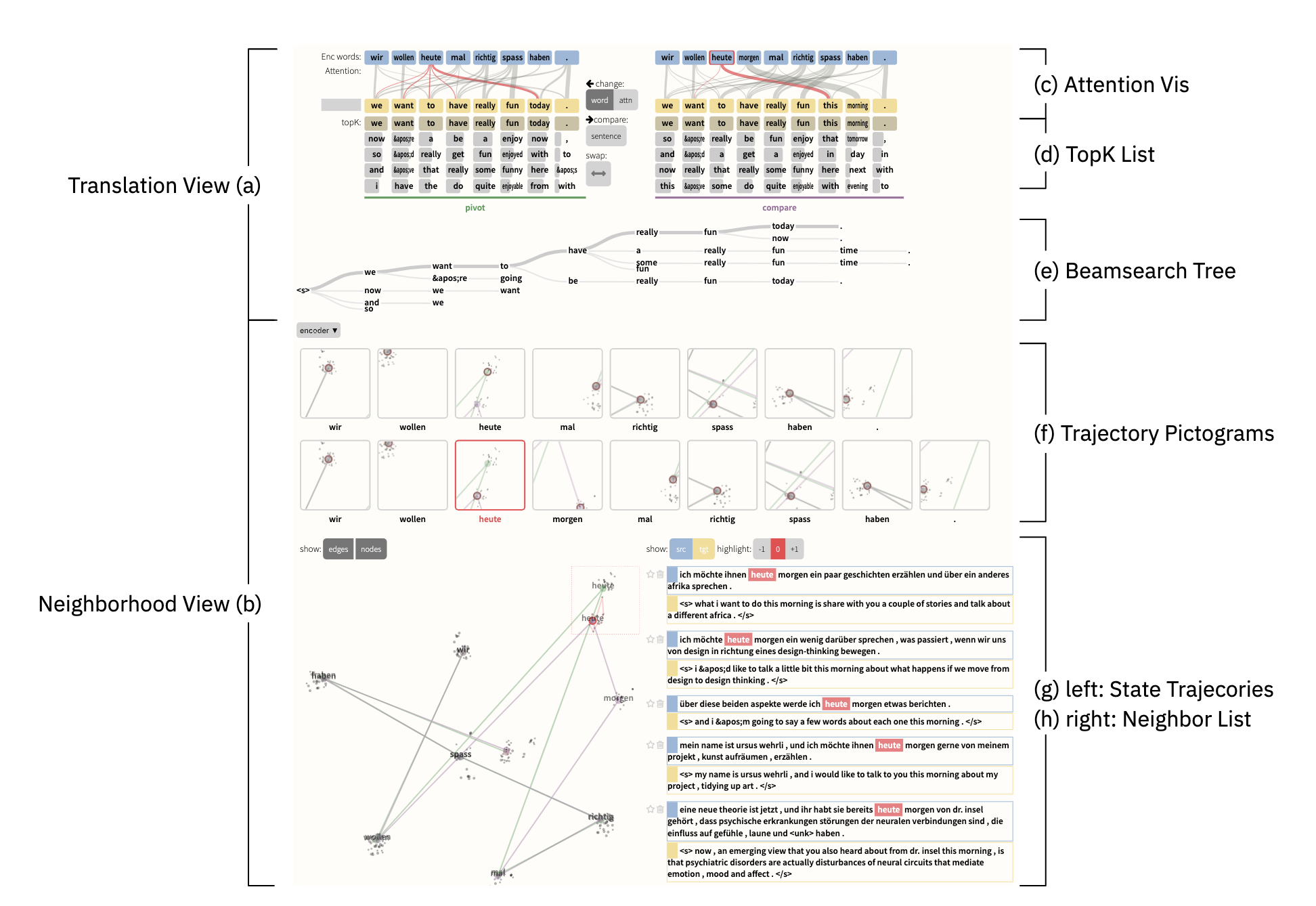
NLP算法可解释性，重点主要在于中间过程所产生的特征的相关性、具体参数值对结果的影响等。这些工作对算法的改进无疑有帮助。但对不懂算法的业务人员来说，要从动辄上亿的参数中，找到与具体业务知识的因果性或相关性，是非常困难的。

若NLP算法过程的可解释性得不到保证，应该优先保证结果可解释性。具体方案为，通过对训练语料的控制，尽可能的缩小每个模型的覆盖范围，把不确定性限制在每个模型内部。换句话说，不能追求对大模型的拟合过程和具体参数值的解释，而应该将一个问题，拆分成多个小模型来表示，提高小模型与领域知识之间的因果关系。

**2.在Seq2Seq和注意力机制中如何可视化模型细节**

**SEQ2SEQ-VIS: A Visual Debugging Tool for Sequence-to-Sequence.** Models Hendrik Strobelt\*, Sebastian Gehrmann\*, Michael Behrisch, Adam Perer, Hanspeter Pfister, Alexand

引用论文地址：<https://arxiv.org/abs/1804.09299>



通过工具对seq2seq可视化

1) 翻译视图（Translation View）

E：编码器以蓝色显示。

D：解码器以黄色显示。

Attention Vis：注意力通过加权的二分连接显示。为了减少视觉混乱，对注意力图进行了修剪。

TopK List：每个时间步的前K个预测。每个词的概率采用条形图编码，黄色高亮显示它最终选择结果。使用了（beam search）。

2) 邻域视图（Neighborhood View）

Seq2seq模型在每个阶段产生高维向量，例如编码器状态，解码器状态或上下文状态，很难直接解释，但我们可以通过查看产生类似向量的样例来估计它们的含义。

G反映的是预先计算好状态的数据集的最近邻域的状态轨迹（采用t-SNE或MDS投影）。为了便于理解较长轨迹，F视图中每一个小窗口都是G视图中的切分，重点关注每一个词。H反映的是最近邻域列表。

**3.对抗样本能否运用到自然语言处理模型中**

**Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey**. Wei Emma Zhang, Quan Z. Sheng, Ahoud Alhazmi, Chenliang Li. arXiv 2019.

引用论文地址：<https://arxiv.org/pdf/1901.06796.pdf>

手机屏幕截图

描述已自动生成

针对文本处理神经网络的攻击类别

1. 白盒攻击

在白盒攻击中，攻击需要访问模型的全部信息，包括模型的体系结构、参数、损失函数、激活函数、输入和输出数据。白盒攻击通常近似于针对特定模型和输入的最坏情况的攻击，其中包含一组扰动。这种对手策略通常非常有效。

举例：FGSM攻击。

FGSM是最早的图像攻击方法之一。在攻击文本DNN方面取得了大量的后续工作。作者使用FGSM的概念来近似对文本分类任务有重要贡献的文本项的贡献，而不是使用FGSM中的成本梯度符号，而是考虑了大小。作者提出了三种攻击方式：插入攻击、修改攻击和删除攻击。

具体地说，他们使用反向传播计算每个训练样本的成本梯度∆xj(f，x，c’)，其中f是模型函数，x是原始数据样本，c’是目标文本类。然后，他们找出包含梯度幅度最高的维度的字符，并将这些字符命名为热点字符，其中包含足够多的字符且出现频率最高的短语被选为热训练短语(HTP)。

在插入策略中，通过在对原始类别有显著贡献的短语附近插入几个目标类别c’的HTP来制作对抗性例子。作者进一步利用外部资源，如维基百科和伪造的事实，选择有效和可信的句子。

在修正策略中，作者确定了热点样本。使用类似的识别HTP的方法将短语(HSP)添加到当前分类中。然后，他们用常见的拼写错误或视觉上相似的字符替换HTP中的字符。

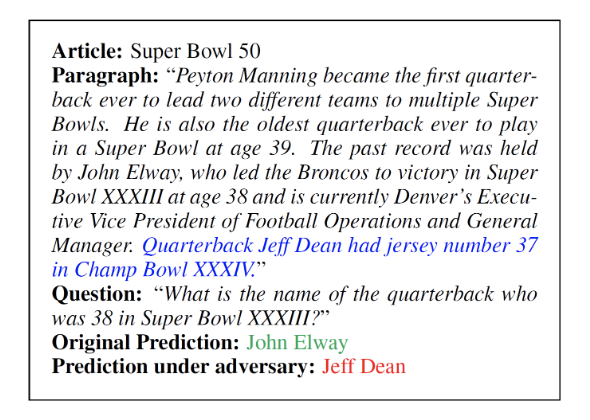
在移位策略中，HSP中不重要的形容词或副词被去掉。这三种策略及其组合在CNN文本分类器上进行了评估。

1. 黑盒攻击

黑盒攻击不需要了解神经网络的细节，但可以访问输入和输出。这种类型的攻击通常依赖于启发式算法来生成敌意示例，而且它更实用，因为在许多现实应用中，DNN的细节对攻击者来说是一个黑匣子。

举例：Concatenation Adversaries攻击。

作者提出在段落末尾加上分散注意力但无意义的句子。这些分散注意力的句子不会改变段落和问题答案的语义，但会使得神经网络得到错误的结果。



Concatenation Adversaries攻击

4.复现Kaggle心脏病数据集冠军kernel，理解所用的模型可解释性技巧

写成博客提交即可，markdown和word均可

1）数据清洗和整理后，作者使用RandomForest模型进行训练，并使用可视化决策树方法，对训练完成后的模型进行可视化操作。

实例代码：

from sklearn.datasets import load\_iris

iris = load\_iris()

# Model (can also use single decision tree)

from sklearn.ensemble import RandomForestClassifier

model = RandomForestClassifier(n\_estimators=10)

# Train

model.fit(iris.data, iris.target)

# Extract single tree

estimator = model.estimators\_[5]

from sklearn.tree import export\_graphviz

# Export as dot file

export\_graphviz(estimator, out\_file='tree.dot',

feature\_names = iris.feature\_names,

class\_names = iris.target\_names,

rounded = True, proportion = False,

precision = 2, filled = True)

# Convert to png using system command (requires Graphviz)

from subprocess import call

call(['dot', '-Tpng', 'tree.dot', '-o', 'tree.png', '-Gdpi=600'])

# Display in jupyter notebook

from IPython.display import Image

Image(filename = 'tree.png')

2）调用混淆矩阵，计算模型的准确率和召回率，并绘制ROC曲线，用以评估该RandomRorest模型的分类性能

1. 计算各个数据特征对模型二分类的贡献（正面或者负面）。

手机屏幕截图

描述已自动生成

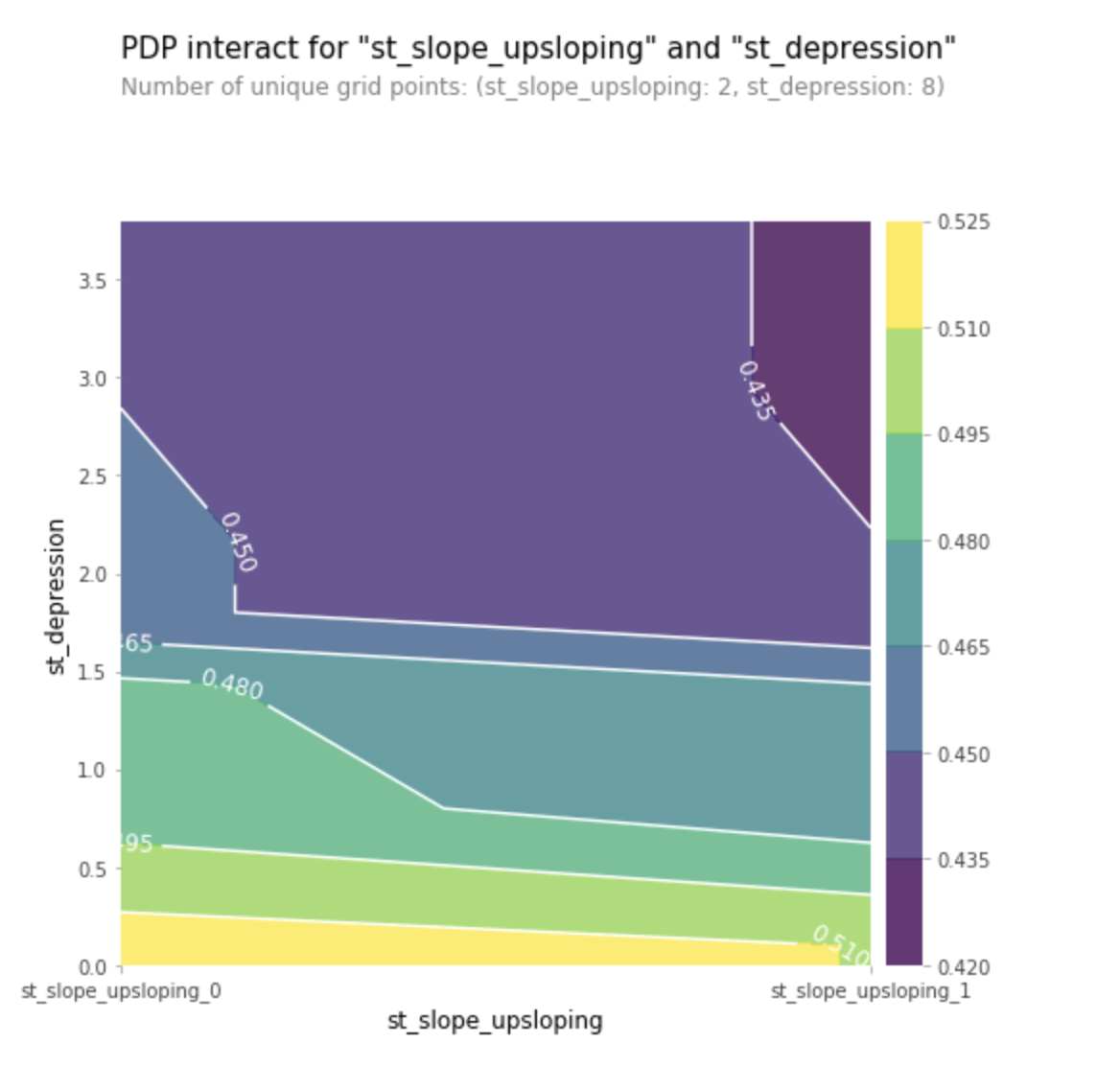
1. 改变单个变量（重要特征的特征值），并查看其对模型输出结果的影响。例如使用部分依赖图仔细了解主要血管的数量特征值的改变对模型的影响。

图片包含 白色, 厨房, 黑色, 灯光

描述已自动生成

主要血管的数量对模型的输出呈现负相关关系，即血管越多越不容易得心脏病。

1. 检查两个变量共同作用对模型的影响，并绘制可视化热图



横轴，展现的是st\_slope\_upsloping特征数值;纵轴，展现的是st\_depression特征数值，二者一同作用对模型输出的影响，可以发现两个特征共同作用都是正相关的，即更容易得心脏病。

1. 计算SHAP均值，并绘制特征SHAP均值直方图

手机屏幕截图

描述已自动生成

手机屏幕截图

描述已自动生成

1. 绘制特征SHAP分布图

手机屏幕截图

描述已自动生成

Feature value反映了特征的数值分布（没做归一化操作，每个特征的数值分布不同，只能直观用热图反映），SHAP value反映了对模型输出的影响（正面或者是负面）

1. 绘制模型对测试对象的分析因子图

手机截图图社交软件的信息

描述已自动生成

预测得心脏病概率为0.36，基线为0.58.红蓝条，表示了主要影响因素对输出概率的正负方面影响（长度表示影响程度，区间为0到1）