刘渤　学号：16308073

读书笔记：Deep Learning for Event-Driven Stock Prediction

文章从事件驱动的角度提出了股票市场预测的一种新的深度学习方法。（限制点：**事件驱动**）

总的来说，这种方法的流程是提取新闻事件、事件嵌入和 NTN 训练、深度卷积神经网络预测。

（1）对于新闻事件的提取，（方法）：文章使用了 Open IE 技术和依存句法分析来获取结构化事件：对于给定的新闻句子，首先用 ReVerb 的方法提取候选元组，然后用 ZPar 的方法进行语法分析主谓宾得到元组，通过比较这两者对应元组的相符性选定最终的主事件元组。（意义：）原有的 NLP 技术对于新闻事件的特征获取没有结构性关联，比如对于事件的主体和客体的辨别，而使用事件的结构表征能够清晰地表示事件的逻辑关系。

（2）对于事件嵌入，（方法：）文章使用了神经张量网络实现单词嵌入到事件嵌入：首先使用 skip-gram 算法从语料库学习 d 维单词表示，并用它们的平均值作为事件表示对应的参数，然后通过给定的式子（参考文章 fig.2）得到最终事件嵌入的输出元组。（意义与区别：）单纯地使用结构表征存在稀疏度增大的问题，限制了预测能力。使用事件嵌入转化成密集连续的向量空间能够使得相似的时间对应的向量相似，即使单词不同；同时与作者之前提出的多元关系学习分布式表示任务相比，之前的方法对于每一种关系类型都会训练出一个矩阵或向量，而事件类型很难训练一个特定的模型，所以关系只能表示成一个向量，与事件参数维度一致；其次，之前的方法参数是可以互换的，但是事件嵌入是不可以的，所以必须分别使用张量对两者分别进行学习最终进行语义合成，得到元组。

（3）在得到事件嵌入元组之后，这些样本（元组）就可以进行 **NTN 训练**了，（方法：）通过每一次迭代随机更换事件参数，得到破坏的事件元组进而得到更新后事件元组集合，遍历集合中的事件元组，计算出损失，在损失大于 0 时进行更新，损失为 0 时继续处理下一事件元组。

（4）对于股票的最终预测，（方法：）对于长中期事件，使用了卷积神经网络：将时间排序的事件嵌入每天平均得到单个输入单元作为模型的输入，使用窄卷积运算（窗口：3）组合相邻事件，并在卷积层顶部使用最大池化层保留最有用的局部特征；而对于短期事件，则只需要平均获得事件向量就行了，然后将这三者结合，使用单隐藏层和单输出层的正反馈神经网络：将三者结合的特征向量作为输入最终得到二分类输出。（意义：）股票的预测会受到一段时间新闻事件的影响，使用卷积网络考虑了历史事件同时提取最有代表性的全局特征，将两者的趋势联系起来。

基于以上的分析，文章进行了实验以及评估：

对于研究表明：从新闻标题预测比内容预测更好，所以实验仅从标题中提取特征；（限制点：**新闻标题**）

实验是对比实验：通过对 Luss、Ding、WB-NN、WB-CNN、E-CNN、EB-NN、EB-CNN（参考文章），

在整体指数上：证明了事件嵌入比文字嵌入更加适合作为特征进行股票预测；事件嵌入比结构表征效果更好，也证实了低维稠密向量缓解稀疏问题对于预测的重要性，甚至重要于提取结构信息；CNN 效果比 NN 好，因为考虑了长期历史的事件影响同时提取了最优代表性的特征向量。

在公司股价上：文章的模型都体现出更好的性能，同时对于财富排名较低的公司，相较于两种经典的算法在预测这些公司是效果不佳，文章的算法因为考虑了长中期影响，即使缺少短期影响，仍然能够取得很好地预测结果。

在利润上：文章遵循了之前的一种常见的模拟模型：如果预测某一支股票的价格第二天会上涨（下跌），就以开盘价买下（卖出）10000 美元的股票，在持有的一天发现能赚取 2% 以上（比卖空价低 1%以下）的利润就卖出（买入），否则就以收盘价卖出（归还）。通过这种模拟，文章证明了算法获取的利润更高，因为如果当天没有公司的新闻，之前的算法就无法预测，考虑到他们用不了长中期新闻，虽然不会降低 Acc 和 MCC 但是会降低实际利润。

文章还提出了一种适合他们模型预测的**交易策略**，当预测上涨概率达到某一个阈值时进行买入或卖出会获得最高的利润。

感想：

（1） 首先是对于事件特征提取的一个启发：单纯的 NLP 文本特征提取确实很大程度受限于事件关系的限制，无法准确地表达逻辑关系使得一些词汇相似度极高而逻辑完全不一样的事件被归为同一类型的可能性会增大，这对于后续的模型干扰影响很大；文章给出的事件结构确实给了我对于这部分一个很大的启迪，分清结构关系以及位置不可替换的性质很大程度提高了事件表达的特征。

（2） 事件嵌入的提出也让我接触到特征稀疏度对于模型预测的影响：在我们处理一些结构特征的时候，可能稀疏度的原因会给我们的模型带来影响，而将其转化为低维稠密向量使得它成为更加高效而有代表性的特征，会对相似的事件（即使没有单词相同）有更强的识别能力。

（3） CNN 增强了模型的鲁棒性：在考虑一些长时间影响的模型上使用 CNN 可以达到有效地预测，因为它可以对长期或短期的事件影响进行提取并池化保持有用特征。

（4） 在文章中也学习到了一个股票预测的简单的模拟模型，对于自己可能今后会用到的模拟有一定的启发，同时也学到了一些做实验的控制方法。

改进方案：

（1） 首先这是一篇从事件驱动的角度去分析股票的趋势的文章（受限于事件驱动），正如其后面提到的，对于股票的影响，不仅受金融新闻的影响，还会受到之前价格以及成交量、人们心理因素的影响，综合考虑这些因素进行加权或比对预测可能会更好一些。

（2） 然后因为读了另一篇文章，对于 fake news 也有一定的了解，在文章实验中单纯使用标题进行特征提取可能会受到一定的误导（受限于新闻标题），所以对于文章内容还有新闻质量的检测也是有一定的改进空间。

（3） 在对于事件结构的提取的时候，如果能够提取更多的信息会对于预测有更加好的影响，例如状语和定语，从状语上我们可以学习到事件的影响程度，而从定语上我们可以找到更多相似的词源匹配，提高我们对于特征的识别。

（4） 在文章的模拟模型中，可能不是很贴合实际的市场操作，所以对于模拟的模型可能需要更加完善的改进。

（5） 预测模型可能存在更加有效地网络来优化原有的模型。