October 10, 2019

Edition: 2.1

软件工程课程设计

第二次报告

姜华

软件工程课程设计

Tongji University

School of Software Engineering

Contents

Part 1 循环神经网络

1

1.1 RNN

用于处理序列，处理序列 (例如文本) 的方式是，遍历序列中的所有元素，并保存一个状态，其中包含已查看内容的相关信息。

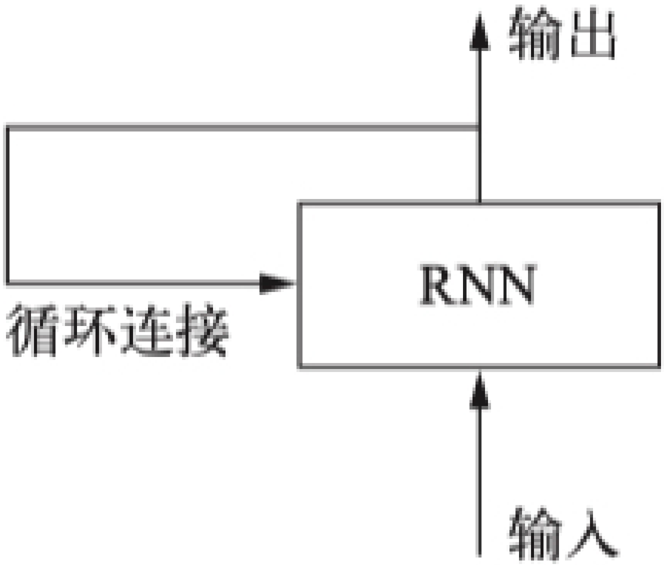


Fig 1.1: RNN

在处理两个不同的独立序列之间，RNN 的状态都会被重置。所以仍然可以将一个序列看作单个数据点，即网络的输入。真正改变的是，数据点不再是在单个步骤中进行处理，相反，网络内部会对序列元素进行遍历。

1.1.1 RNN 的前向传递过程

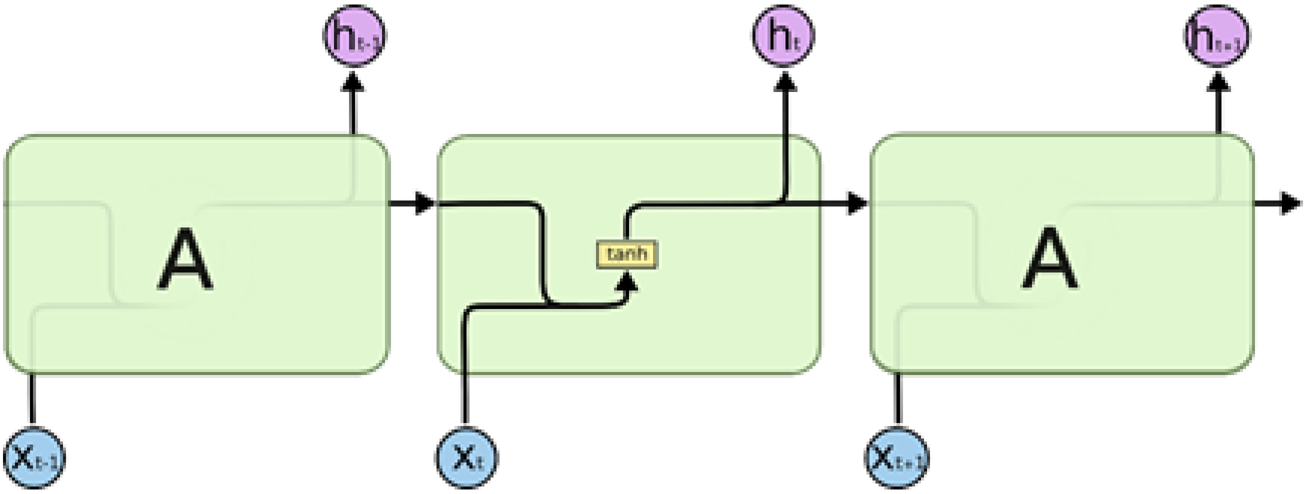


Fig 1.2: RNN 前向传播

如图是三个神经网络，上一个网络的信息（即 ht-1）直接传过来，配合当前网络的输入 Xt，两者结合之后，再通过 tanh 层进行信息压缩，就形成当前网络的输出 ht。(这个 tanh 层就是一个函数，tanh 就是双曲正切函数，可以将输入的值转化为-1 到 1 之间的一个值，通常用于对信息的压缩处理，或者规范化处理。)

1.2 LSTM（长短期记忆网络）

LSTM 层是 SimpleRNN 层的一种变体，它增加了一种携带信息跨越多个时间步的方法。

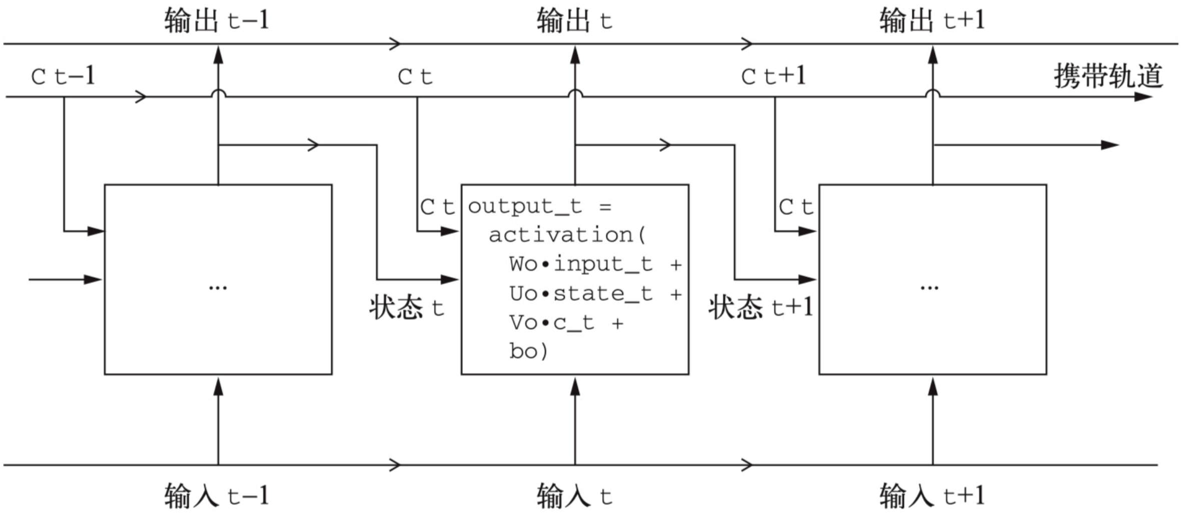


Fig 1.3: LSTM

我们向这张图像中添加额外的数据流，其中携带着跨越时间步的信息。它在不同的时间步的值叫作 Ct ，其中 C 表示携带（carry）。Ct 的计算方法和当前状态有关，从而会影响传递到下一个时间步的状态。

1.3 C-LSTM for text classification

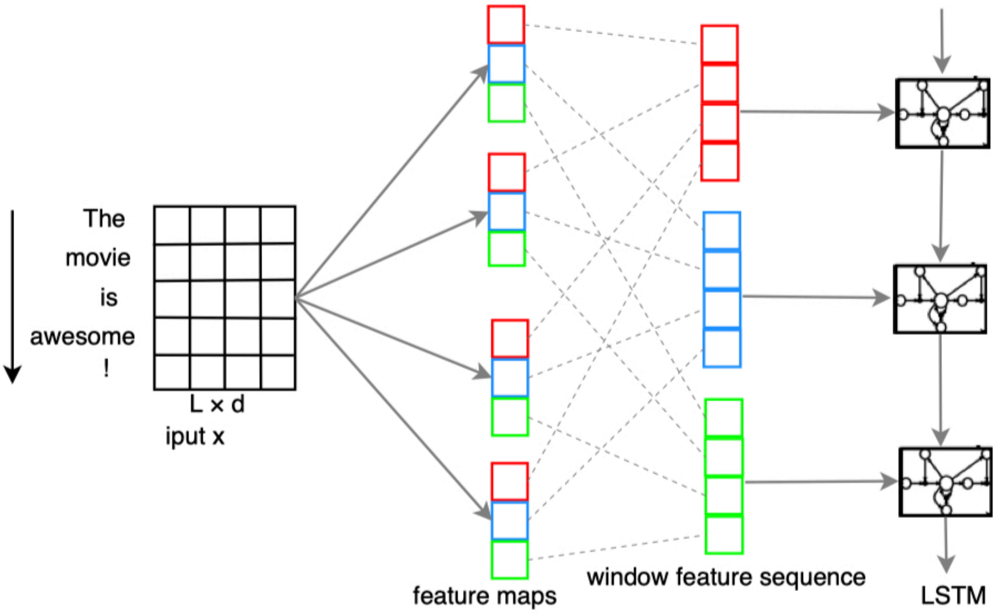


Fig 1.4: C-LSTM

1.3.1 一维卷积神经网络

图中的一维卷积神经网络有四个 filter，每一个 filter 都有一个宽度为 k 的窗口向量，用这个窗口向量对文本序列进行划窗，本图中是划窗 3 次，生成一个特征图 (feature map)，用三种不同的颜色来代表一个 filter 的不同划窗的计算结果，红色，蓝色，绿色的方块代表的计算结果具有位置上的先后关系。把所有 filter 的计算结果按照划窗顺序聚集在一起 (相同颜色的方块放在一起)，然后输入给 LSTM 在 LSTM 的最后一步的输出后加上一 softmax 层，然后进行训练。

Part 2 Experiment

2

2.1 目标

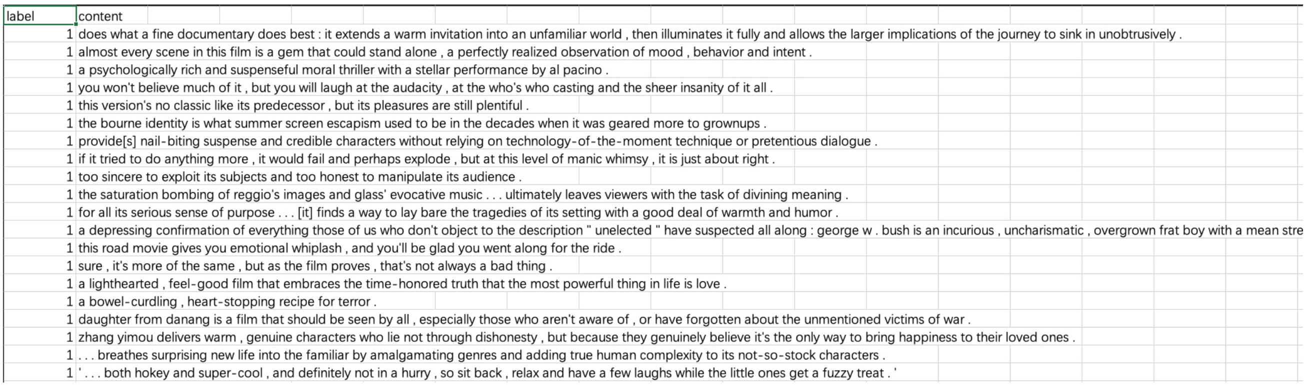
文本分类

2.2 模型

C-LSTM

2.3 数据格式（训练数据和测试数据）

数据存储在 csv 文件中，分为两列，“label” 列和 “content” 列，分别代表一条文本序列的类别和内容，总共有两类



label	content
1	does what a fine documentary does best : it extends a warm invitation into an unfamiliar world , then illuminates it fully and allows the larger implications of the journey to sink in unobtrusively .
1	almost every scene in this film is a gem that could stand alone , a perfectly realized observation of mood , behavior and intent .
1	a psychologically rich and suspenseful moral thriller with a stellar performance by al pacino .
1	you won't believe much of it , but you will laugh at the audacity , at the who's who casting and the sheer insanity of it all .
1	this version's no classic like its predecessor , but its pleasures are still plentiful .
1	the bourne identity is what summer screen escapism used to be in the decades when it was geared more to grownups .
1	provide[s] nail-biting suspense and credible characters without relying on technology-of-the-moment technique or pretentious dialogue .
1	if it tried to do anything more , it would fail and perhaps explode , but at this level of manic whimsy , it is just about right .
1	too sincere to exploit its subjects and too honest to manipulate its audience .
1	the saturation bombing of reggio's images and glass' evocative music . . . ultimately leaves viewers with the task of divining meaning .
1	for all its serious sense of purpose . . . [it] finds a way to lay bare the tragedies of its setting with a good deal of warmth and humor .
1	a depressing confirmation of everything those of us who don't object to the description " unelected " have suspected all along : george w . bush is an incurious , uncharismatic , overgrown frat boy with a mean stre
1	this road movie gives you emotional whiplash , and you'll be glad you went along for the ride .
1	sure , it's more of the same , but as the film proves , that's not always a bad thing .
1	a lighthearted , feel-good film that embraces the time-honored truth that the most powerful thing in life is love .
1	a bowel-curdling , heart-stopping recipe for terror .
1	daughter from danang is a film that should be seen by all , especially those who aren't aware of , or have forgotten about the unmentioned victims of war .
1	zhang yimou delivers warm , genuine characters who lie not through dishonesty , but because they genuinely believe it's the only way to bring happiness to their loved ones .
1	. . . breathes surprising new life into the familiar by amalgamating genres and adding true human complexity to its not-so-stock characters .
1	' . . . both hokey and super-cool , and definitely not in a hurry , so sit back , relax and have a few laughs while the little ones get a fuzzy treat . '

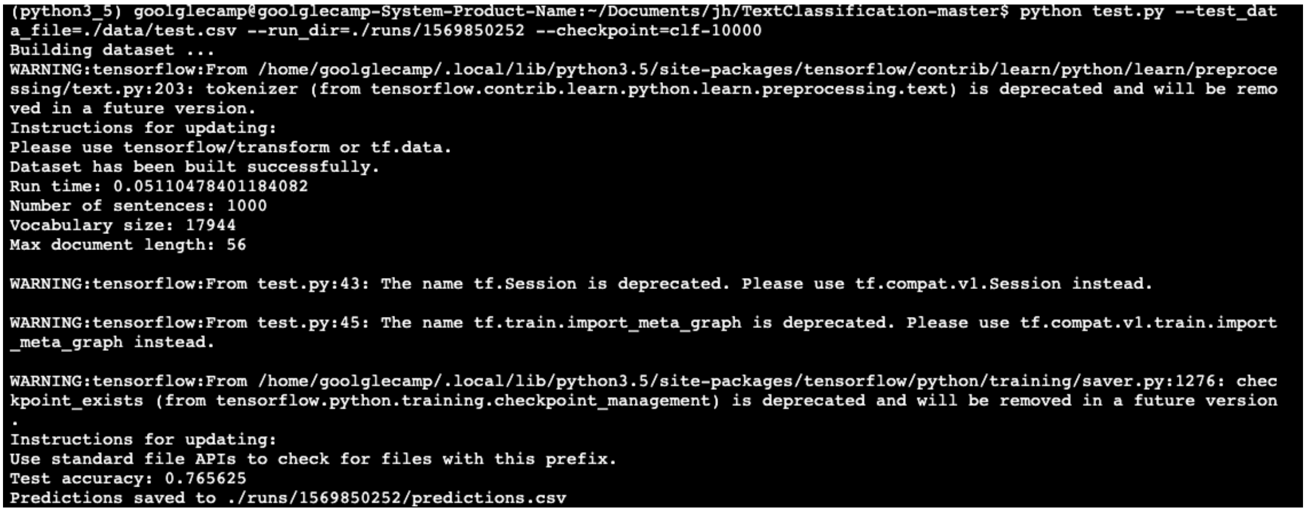
Fig 2.1: C-LSTM

2.4 训练以及验证

- 使用 4500 条 “label1” 的文本和 4500 条 “label2” 的文本进行训练
- 在 450 条 “label1” 的文本和 450 条 “label2” 的文本上进行验证

2.5 Result

在作为验证集的 1000 条文本中获得了 76.6% 的准确率



```
(python3_5) goolglecamp@goolglecamp-System-Product-Name:~/Documents/jh/TextClassification-master$ python test.py --test_dat
a_file=./data/test.csv --run_dir=./runs/1569850252 --checkpoint=clf-10000
Building dataset ...
WARNING:tensorflow:From /home/goolglecamp/.local/lib/python3.5/site-packages/tensorflow/contrib/learn/python/learn/preproce
ssing/text.py:203: tokenizer (from tensorflow.contrib.learn.python.learn.preprocessing.text) is deprecated and will be remo
ved in a future version.
Instructions for updating:
Please use tensorflow/transform or tf.data.
Dataset has been built successfully.
Run time: 0.05110478401184082
Number of sentences: 1000
Vocabulary size: 17944
Max document length: 56

WARNING:tensorflow:From test.py:43: The name tf.Session is deprecated. Please use tf.compat.v1.Session instead.

WARNING:tensorflow:From test.py:45: The name tf.train.import_meta_graph is deprecated. Please use tf.compat.v1.train.import
_meta_graph instead.

WARNING:tensorflow:From /home/goolglecamp/.local/lib/python3.5/site-packages/tensorflow/python/training/saver.py:1276: chec
kpoint_exists (from tensorflow.python.training.checkpoint_management) is deprecated and will be removed in a future version
.
Instructions for updating:
Use standard file APIs to check for files with this prefix.
Test accuracy: 0.765625
Predictions saved to ./runs/1569850252/predictions.csv
```

Fig 2.2: 验证结果

Part 3 下一步工作

3

- 学习使用 C-LSTM 对异常数据序列进行分类

Appendix A

Image Index