情感分类 (1)：分词

朱梦
主 github：https://github.com/mengzhu0308/mengzhu
副 github1：https://github.com/mengzhu030801/mengzhu01
公众号：梦梦的分享

2020 年 12 月 13 日

人工智能从业人员常把计算机视觉比作 | “人工智能的眼睛”，把自然语言处理 (NLP) 比作“人工智能皇冠上的明珠”。

1 数据预处理

对于计算机视觉，图像数据预处理一般流程如图1所示，可通过 OpenCV 或 Pillow 等图像处理库完成主要流程。因此，图像数据预处理是比较简单的。经过这样的预处理后，输入到卷积神经网络的张量形状通常为 (batch-size, H, W, 3)。

图 1 图像数据预处理流程

对于文本数据，文本数据预处理一般流程如图2所示，下面笔者将通过主要代码讲解。

图 2 文本数据预处理流程

2 分词

```
from jieba import posseg as pseg
text = 'The cat sat on the desk."
text_seged = pseg.cut(text.lower())
```

3 停词

```
term_list = []
stop_flag=('x', 'c', 'u', 'd', 'p', 't', 'uj', 'f', 'r')
for term, flag in text_seged:
    if flag not in stop_flag:
```

```
5           term_list.append(term) if not term.isdigit() else term_list.
                append('num')
```

停词旨在消除标点符号和已经逐渐消亡的词。

4 字符序列转整数向量

```
1  from gensim import corpora
2
3  dictionary = corpora.Dictionary(X)
4
5  def str2id(text, tokenid):
6      out_text = []
7      for term in text:
8          out_text.append(tokenid[term] + 1)
9      return out_text
10
11 str2id(text, dictionary.token2id)
```

5 填充至最大长度

```
1  def sequence_padding(text_list, max_length=None, padding=0):
2  """Numpy函数，将序列padding到同一长度
3  """
4      if max_length is None:
5      max_length = max([len(text) for text in text_list])
6
7      outputs = []
8
9      for text in text_list:
10         text = text[:max_length]
11         pad_width = (0, max_length - len(text))
12         text = np.pad(text, pad_width, mode='constant',
               constant_values=padding)
13         outputs.append(text)
14
15     return np.array(outputs, dtype='int32')
```

6 词嵌入

```
from keras.layers import *
x = Embedding(vocab_size + 1, hidden_dim, input_length=max_length)(x)
```

词嵌入层输入张量的形状是 (batch-size, max-length)，输出张量形状是 (batch-size, max-length, hidden-dim)。

7 实验结果

笔者实现的结果如图3所示，只跑了 5 个周期，效果还是不错的。开发环境是 tensorflow1.14 + keras2.3.1。所有代码已开源在`https://github.com/mengzhu0308/mengzhu`

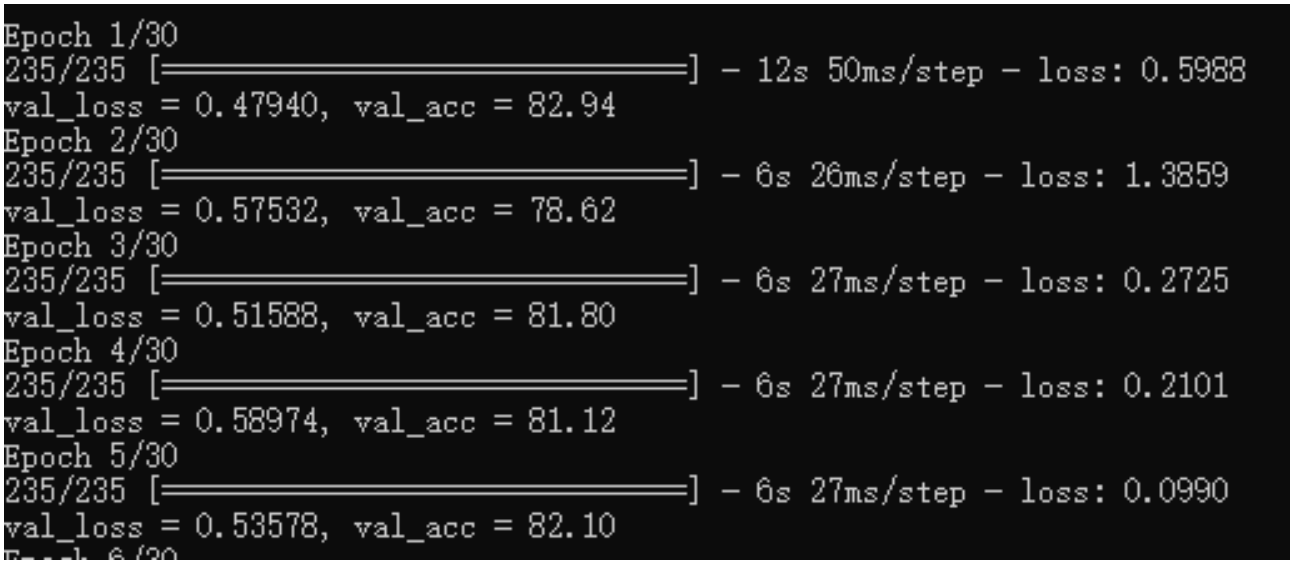

图 3 实验结果