AI学习笔记--Tensorflow--Text Hub

此笔记本（notebook）使用评论文本将影评分为*积极（positive）*或*消极（nagetive）*两类。这是一个*二元（binary）*或者二分类问题，一种重要且应用广泛的机器学习问题。

本教程演示了使用 Tensorflow Hub 和 Keras 进行迁移学习的基本应用。

我们将使用来源于网络电影数据库（Internet Movie Database）的 IMDB 数据集（IMDB dataset），其包含 50,000 条影评文本。从该数据集切割出的 25,000 条评论用作训练，另外 25,000 条用作测试。训练集与测试集是*平衡的（balanced）*，意味着它们包含相等数量的积极和消极评论。

此笔记本（notebook）使用了 tf.keras，它是一个 Tensorflow 中用于构建和训练模型的高级API，此外还使用了 TensorFlow Hub，一个用于迁移学习的库和平台。有关使用 `tf.keras` 进行文本分类的更高级教程，请参阅 MLCC文本分类指南（MLCC Text Classification Guide）。

运行此例子之前需要下载 TensorFlow 的组件。需要在终端执行：

```
→  Tensorflow pip install tensorflow_hub
→  Tensorflow pip install tensorflow_datasets
```

在执行完毕后，开始编辑 Python代码：

```
from __future__ import absolute_import, division, print_function,
unicode_literals

import numpy as np

import tensorflow as tf

import tensorflow_hub as hub
import tensorflow_datasets as tfds

print("Version: ", tf.__version__)
print("Eager mode: ", tf.executing_eagerly())
print("Hub version: ", hub.__version__)
print("GPU is", "available" if
tf.config.experimental.list_physical_devices("GPU") else "NOT
AVAILABLE")
```

并且打印出相应的版本信息：

```
Version:  1.14.0
Eager mode:  False
Hub version:  0.7.0
GPU is NOT AVAILABLE
```

之后开始下载 IMDB 的测试训练集，我们按照 6：4 的比例分配总共 15,000个样本数据，10,000个作为样本数据，并且 5,000个作为测试数据。

```
# 将训练集按照 6:4 的比例进行切割，从而最终我们将得到 15,000
# 个训练样本，10,000 个验证样本以及 25,000 个测试样本
train_validation_split = tfds.Split.TRAIN.subsplit([6, 4])

(train_data, validation_data), test_data = tfds.load(
    name="imdb_reviews",
    split=(train_validation_split, tfds.Split.TEST),
    as_supervised=True)
```