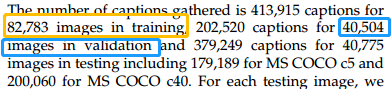
MSCOCO数据集的Karpathy split是什么

# 一、原始数据集

MS COCO数据集（2014版）中IC相关样本的原始划分如下（<https://arxiv.org/pdf/1504.00325.pdf>）：

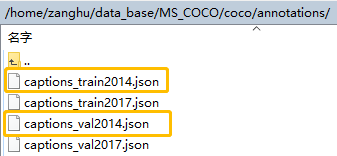


训练集：82,783

验证集：40,504

测试集：50,775

需要注意的是，可能是出于保密原因，测试集并未附带标注信息，如下图所示：



因此，实验时能够使用的其实就是训练集和验证集的约123,000个样本。

# 二、新的划分

Karpthy等人在2015年CVPR论文《Deep Visual-Semantic Alignments for Generating Image Descriptions》（<https://arxiv.org/abs/1412.2306>）中提出了一种新的训练、验证、测试集划分方式：

Step 1：将原始验证集的40,504个样本中，划分为3个子集，名称和样本数如下：

新验证集（val）：5000

新测试集（test）：5000

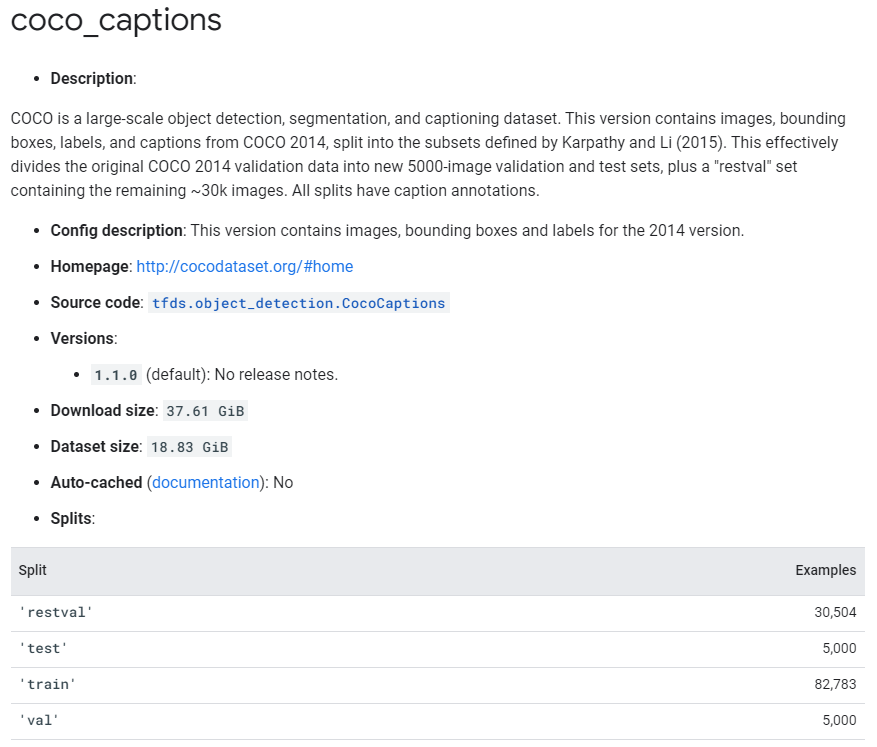
剩余验证集（restval）：30,504

Step 2：之后将上面的“restval”和原始训练集混在一起组成一个新训练集：

新训练集（train）：82,783 + 30,504 = 113,287

以上就是IC领域著名的“Karpathy Split”。

COCO新的数据集详情如下（<https://www.tensorflow.org/datasets/catalog/coco_captions>）：



# 划分结果

由于Karpthy等人在划分数据集时采用了随机抽取的方式，因此为了确保后续实验中的划分结果和Karpthy实验中的划分方式一致，可以使用如下的结果：

1. 划分后的JSON文件，每个样本都已标注为4各类别（train、val、test和restval）之一：

<http://cs.stanford.edu/people/karpathy/deepimagesent/caption_datasets.zip>

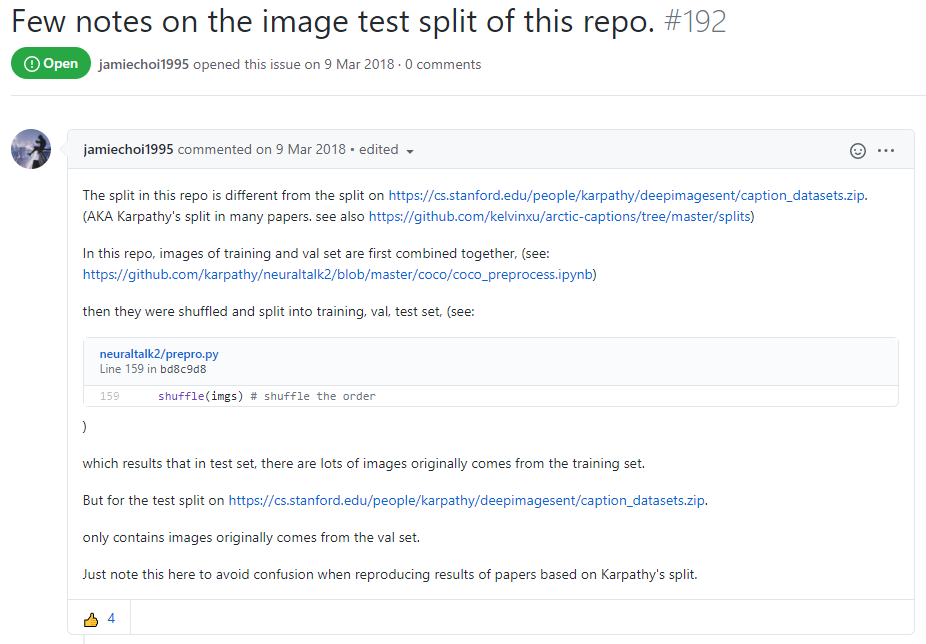
1. 或使用Kelvin Xu等人论文配套代码中的划分：

<https://github.com/kelvinxu/arctic-captions/tree/master/splits>

# 附录A：划分方式

1. Github上面karpathy/neuraltalk2的一个Issue详细论述了划分源码：

<https://github.com/karpathy/neuraltalk2/issues/192>



1. 知乎上的一个回答：MSCOCO数据集的karpathy test split是什么？

<https://www.zhihu.com/question/283314344>

# 参考资料

1. Xinlei Chen et al., Microsoft COCO Captions: Data Collection and Evaluation Server, <https://arxiv.org/pdf/1504.00325.pdf>
2. Karpathy et al., Deep Visual-Semantic Alignments for Generating Image Descriptions, CVPR 2015, <https://arxiv.org/abs/1412.2306>
3. Xu Kelvin et al., Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015, <https://arxiv.org/abs/1502.03044>
4. Github: <https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning>
5. Github: <https://github.com/karpathy/neuraltalk2/issues/192>
6. Tensorflow: <https://www.tensorflow.org/datasets/catalog/coco_captions>
7. 知乎：<https://www.zhihu.com/question/283314344>
8. Twitter: <https://twitter.com/karpathy/status/592193801310973952>