CapDec一些细节

1. embbedings\_generator.py

负责CLIP文本和图片。如果是文本，prefix是空的，如果是图片，prefix就是图片encode后的结果。默认情况是前者。

默认存放在./data/coco/verified\_split\_COCO\_train\_set\_with\_text\_not\_norm.pkl，里面有三样东西，clip\_embedding、captions、clip\_embedding\_text\_dave，分别指prefix（是图片CLIP后的矩阵）、caption整个的结构以及caption的embeddings（[566747, 640]，即一共566747个text，每个text、CLIP后的embedding长度是640）。具体来看，captions[0]= {'image\_id': 522418, 'caption': 'A woman wearing a net on her head cutting a cake. ', 'id': 681330, 'clip\_embedding': 0}.

1. train.py
2. 构建数据集

Dataset的self.prefixes是clip\_embedding或者clip\_embedding\_text\_dave，这是由use\_image\_embedding\_as\_clipcap决定的。

使用GPT-2，captions\_tokens是经由GPT-2编码得到的，不定长的（要区分其和之前caption的CLIP过后embedding），caption2embedding则就是每个caption的clip\_embedding（从0开始往上），max\_seq\_len最长的captions\_tokens。

默认存放在./data/coco/verified\_split\_COCO\_train\_set\_with\_text\_not\_norm\_tokens.pkl。

Dataloader吐出来的tokens、mask、prefix分别是captions\_tokens、padding后的mask（取0或1，0表示padding，这里额外加了prefix）

Split为train的和为restVal的作为训练集。