Report of vision-language pre-training

First, we divide the GIF data set into a training set and a test set, where the training set includes more than 12W graphic pairs, and the test set includes more than 4W data. Then we extract 16 frames of pictures for each GIF image. If less than 16 frames, we will use the copy method to make up; if more than 16 frames, we will use the method of uniformly extracting frames to extract 16 frames.

Then we use the pre-trained ResNet50 model on ImageNet to extract feature on the 16 frames extracted from all GIF images to obtain the 1000-D feature vector. Input to the traditional Sequence to Sequence model for training for 200 generations, when the model converges, generate the caption of the test video, and submit the results.