Introduction

In our daily life, we must have encountered this situation that sometimes it is hard to find a specific photo in our album. We might forget the specific location or exacting time of shooting but only remember some key words. We have to look through the whole album to find the photo we want, which is very time consuming. So, we want to develop a model which can help us find the specific photo we want with only a few words or a short description!

Challenges

The evaluation process is pretty slow. For example, we want to evaluate our model on N image-text pairs, the computation complexity is N^2, which is very time-consuming. Currently, we used two pretrained models rather than creating our own model. So we are not sure whether it is acceptable. We might need some creative insights/suggestions. GPU issue on GCP

Insights

We have just finished the setup for two models and we are dealing with the GPU issue on GCP. The fine-tuned CLIP model will have a better performance.

Plan

We might need to spend more time to come up with some creative points. We have changed our basic idea. Our purpose of this project is to find the corresponding photo in the album when provided with a short description. Actually, this goal is easily achieved by using the CLIP model. However, just using the CLIP model will not produce a satisfactory result. As shown in the paper, the zero-shot CLIP R@1 is 68.7% while other pre-trained models R@1 can achieve 76.7%. So we want to fine tune the CLIP model with our own album dataset. The first thing we need to do is to create several descriptions for each photo in the album. To achieve this goal, we used a pre-trained image captioning model to create captions for our photos. Then, we will use these photo-caption pairs to fine tune the CLIP model and the result is supposed to be better than just using CLIP alone.