Project Proposal Template

1. Project Title: Assessing Robustness of CLIP models on Adversarial Attacks

2. Team members (2): Hnin Ookhin (ho2298), Chaewon Park (cp3227)

3. Goal/Objective:

Our goal is to understand the CLIP model's robustness. We will apply adversarial attacks on inference images and assess the model's robustness using classification metrics and explore techniques and ways to improve and analyze robustness.

4. Challenges:

Since we will be referencing the simplified version of CLIP (see references for blog post) and training on a smaller dataset due to timing and resource constraint, there could be discrepancies between the simplified and the original model, and a lower performance could be due to these discrepancies in model size and training dataset.

5. Approach/Techniques:

We will work with a simplified version of CLIP (see references) due to training time and resource constraint. We will use the simplified model as our starting model and train on Flickr dataset which contains image, text pairs.

Adversarial attack is a threat imposed on neural networks to cause mispredictions, and these adversarial examples look just like normal images but have indistinguishable perturbations. They are constructed by assessing sensitivity (gradient) of the loss function in terms of the image and perturbing the pixels which are most likely to maximize the loss function. FGSM and PGD are two of the most well-known adversarial attack methods. FGSM assumes the subspace in which adversarial examples exist are continuous and by adding perturbations to the image in the direction of the gradient's sign, we can easily obtain numerous adversarial examples. PGD is a method proposed by Madry which overcomes the downsides of FGSM. FGSM may not work if the loss function landscape is noisy and moving a step in the direction of the steepest gradient may not lead to the best adversarial example. Hence, instead of taking 'one' step in the gradient direction, PGD takes 'multiple' steps to strategically find the highest loss point.

We will apply both FGSM and PGD attacks on a subset of Imagenet dataset to use as our test dataset. At test time, the model would predict a text description to the given image and we will assume that the image is correctly classified if the classifier word is in the output text. Then, we will perform the standard Top-K classification metric on the test dataset.

We will vary model parameters and modify model layers to see the effects on Top-K classification metric on the test dataset. We will also analyze the network's robustness and see if there are some neurons that contribute to its robustness.

6. Implementation details:

- Type of compute: GPU on Google Cloud
- Software
  - Framework: PyTorch
  - Model reference code: https://github.com/moein-shariatnia/OpenAI-CLIP
  - Original code: we will modify the reference code to add new inference features for classification, as well as modify model parameters to test adversarial performance
- Training Dataset:
  - Flicker 8K: Consists of 8,000 images that are each paired with five different captions which provide descriptions of the images
    - Dataset source: https://www.kaggle.com/datasets/adityajn105/flickr8k
- Adversarial Test Dataset:
  - FGSM Imagenet adversarial: An adversarial dataset made using an FGSM on a subset of Imagenet (1000 images).
    - Dataset source: https://www.kaggle.com/datasets/anirudhyadav9784/adversarial-attack-on-imagenet-dataset
  - PGD Imagenet adversarial: We will need to apply PGD on subset of imagenet to create this test dataset ourselves
    - PGD algorithm: https://github.com/Harry24k/adversarial-attacks-pytorch

7. Demo planned: We will do a live demo of model inference on adversarial images during presentation.

8. References (if any)

- Original CLIP paper: https://arxiv.org/pdf/2103.00020.pdf
- Simple implementation of CLIP blog post: https://towardsdatascience.com/simple-implementation-of-openai-clip-model-a-tutorial-ace6ff01d9f2
- Adversarial Attack code repo: https://github.com/Harry24k/adversarial-attacks-pytorch.git
- PGD attack paper :https://arxiv.org/pdf/1706.06083.pdf
- FGSM attack paper: https://arxiv.org/pdf/1412.6572.pdf