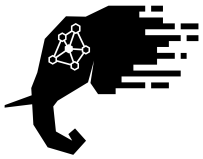

Task 3: Defensive Transformation

After having a lot of fun breaking stuff and trying to bypass the security measures employed by the company in the Red Team under Mike's supervision, you've felt like it's time for tackling a new challenge.

Fortunately, some higher powers at MiNI enterprise felt the same, and in the very morning, just before the state-mandated compulsory coffee break you've received an email, this time from Sandra, the Blue Team Leader.

> *Greetings XYZ,*
>
> *In a galaxy not so far away, our esteemed organization has honed a powerful encoder model using a vast and precious dataset. This model crafts invaluable high-dimensional representations of images, ready to serve various specific downstream tasks. We're set to unleash this mighty force through API endpoints, where users shall pay the toll for each query. In the language of the Force, querying involves users transmitting their visual messages, and in return, the encoder bestows its representations.*
>
> *However, to shield this force-sensitive model from potential dark side endeavors via API queries, we must invoke a protective transformation upon the original encoder representations—a unique spell for every user in the system.*
>
> *Your mission, should you choose to accept it, is to conjure forth a protective transformation that abides by two sacred conditions:*
>
> - *1. The representations post your mystical transformation must retain utility for a solitary user—random disturbances are not the Jedi way.*

- *2. The representations post your enchantment must defy attempts to remap them into the original feature space. The greater the defiance, the stronger the force.*

May the source be with you, always.

Sandra

Young Padawan, will you get this job done?

Endpoints

There's only an evaluation endpoint here.

Task

You need to create a protective transformation, transforming the original representations into a set of protected representations of size in **[32; 2048]**.

For a submission to be accepted, you cannot use just any transformation. Your protected representations must still be useful for downstream tasks, and the accuracy of the classifier trained on protected representations cannot decrease more than **2 p.p**. on downstream tasks.

If it hurt the performance more, the submission would not be accepted.

Datasets

Representation quality evaluation

You will be provided with a DefenseTransformationEvaluate.npz file, which contains a set of representations (a matrix of n_samples x 192) and labels. You can evaluate the influence of your transformations on the representations quality by training a classifier on the clean representations and comparing the accuracy to the classifier trained on representations after your defense transformation.

Submission dataset

The second dataset, DefenseTransformationSubmit.npz contains a matrix of shape **20250×192**, each row being a separate encoded image.

An example code for loading both datasets is provided here.

Submission

The submission should be an *.npz file containing field *representations* being a matrix of size **20250x[your representation's size]** (the representation size has to be in the [32; 2048] interval. Failure to comply will lead to rejection of your submission).

An example code for submission is provided here.

Evaluation

The evaluation will be two-part:

1. Representation quality assessment. We store (privately) a label for each of the samples from the DefenseTransformationSubmit dataset. We train a classifier on your representations and our labels. If the accuracy drops **2 percentage points** below the accuracy on the clean representations, we reject your submission. You may ask, what is the accuracy on the clean data? And you will get no answer to that, hehe
2. If your representations pass the first test, we then try to map them to original representations using [REDACTED] [REDACTED] [ALSO REDACTED]. Your final score is the cosine distance between clean and remapped submission representations.