Project Selection Task: Listen to See

Deadline for the Task: 7th May 2021 11:59PM
In this task, you are supposed to build a simple image classification model. (Reference)

You are provided with the iNaturalist12k Dataset containing 12,000 images split into 10,000 train and 2,000 validation samples. The images from the dataset belong to 10 classes. You can **download the dataset from here:**
https://storage.googleapis.com/wandb_datasets/nature_12K.zip

Further to make ur task simpler, we provide a starter code (obviously with portions missing from it). There are indicators for the areas where you need to fill code at the **TOP** of each script. Reference links are also provided at the corresponding locations.
Link to code:
https://drive.google.com/drive/folders/1j2cTZIQFXGdQPDm8TZceqCIeRsteIZOG?usp=sharing

Once again, we reiterate, there are **no prerequisites** for this project. All the learning happens after you get selected into the project team.
Note: We are **NOT** going to evaluate you based on the accuracy or performance of the model. Hence, it's completely fine if you are unable to train the model (due to compute reasons, or whatever it be). We are specifically looking to gauge your understanding and willingness to learn/explore.

Apart from the above classification task, we expect you to explore and **answer the following questions.** Once again there is no right or wrong answer to these questions and it's basically to understand your thought process.

1. What are some of the foreseeable problems you can think of which this project might face? (Brownie points: if you could read about possible solutions!)
2. Image Captioning (providing textual description for a large scene) is going to be a major component in this project. What do you think the input and output for such a model will be? Also, suggest a suitable loss function for training these models. (Even a guess is fine, although it would be cool if you can justify it :P)
3. The generation text would have to be converted into human-like speech. To train such a model, what would be the ideal form for representing audio and justify the same. Given this representation, suggest a loss function to compare the predicted audio with the ground truth.

In case of any doubts feel free to contact any of the Group Admins in the **Listen To See** whatsapp group.

HAPPY SOLVING!!!