Handover Document: Attention Models for Adversarial Robustness

Code Changes

- **adversary.py**
    - Allow new model types (**parallel_transformers**, **multi_gaze**) to be loaded and evaluated
- **datasets.py**
    - Added batching for ImageNet10 (batching was already implemented for ImageNet and ImagNet100)
- **glimpse.py**
    - Created new retinal warping functions (**warp_func_multi_gaze**, **warp_image_multi_gaze**) that accommodate a unique gaze for each provided image rather than just one, which was required for the multi-gaze model
    - These functions utilize a **bilinear_sampler** so that they can be differentiated with respect to the image gazes
        - Sampler used in place of **tf.gather_nd**, which is not differentiable with respect to its second parameter
- **model_backbone.py**
    - Added **parallel_transformers** model
        - Multi-branch architecture with option to share **resnet** weights
        - Image is first inputted into a **ResNet_CIFAR** model, which outputs the theta parameters defining the affine transformations to be applied[1]
        - Each branch consists of one spatial transformer network (STN) and a **resnet** that the transformed image is then fed into
    - Added **multi_gaze** model
        - Multi-branch architecture with option to share **resnet** weights
        - Image first inputted into a **resnet**, which outputs a set of fixation points
        - Retinal sampling transforms then applied to image, one centered at each fixation point
        - Warped versions of image each fed into a **resnet**
    - Added additional functionality and parameters to **soft_attention_model**
        - Allow model to use either a full **resnet**, a smaller **ResNet_CIFAR**, or a simple CNN
- **trainer.py**
    - Enabled new model types to be trained
    - Utilize batching (current batch size 32) for ImageNet10
    - Added command-line arguments such as number of epochs
- **transformer.py**

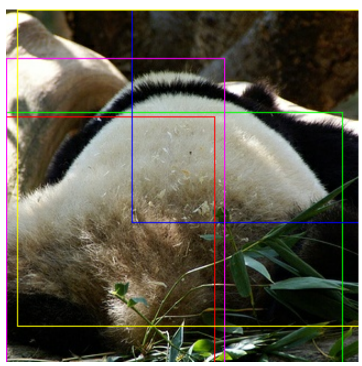
[1] We tried a smaller network with two convolutional layers as well as a full ResNet, but the smaller network produced the same transformations on each image while we did not have success in training the full ResNet

- Have STN output not just transformed image, but also the coordinates of the bounding box associated with the transformation, as well as the center of the box

- **view_images.ipynb**
  - Created Jupyter notebook for saving and viewing model-related images and adversarial perturbations, in addition to investigating the new model types

Experiment Results and Observations

- Results PPT: https://drive.google.com/file/d/14l4TRHk-VcQNXa9ht5Ek_UipsSOpcph8/view?usp=sharing
  - Note that results for the existing Standard ResNet and Retinal Sampling models do not exactly match those in the original paper for ImageNet10
    - Could be due to the introduction of batching
- Parallel Transformers
  - Command-line arguments: **--model=parallel_transformers --dataset=imagenet10 --sampling=0 --coarse_fixations=0 --augment=1 --auxiliary=0 --restricted_attention=1 --shared=0 --epochs=400 --num_transformers=5**
  - Notable **soft_attention_model** parameters: **use_resnet=True, use_full_resnet=False**
  - While the learned bounding boxes are distinct from each other and from image to image, they tend to cling to the sides of the image



    - Appears to be due to many of the theta parameters converging to +/- 1.0
    - Same issue occurs (and to a greater degree) with multi-gaze model
    - Regularization l2=0.001 led to lower standard performance
- Multi-Gaze
  - Command-line arguments: **--model=multi_gaze --dataset=imagenet10 --sampling=0 --coarse_fixations=0 --augment=1 --auxiliary=0 --shared=0 --epochs=400 --num_transformers=5**
  - Notable **soft_attention_model** parameters: **use_resnet=True, use_full_resnet=True, initialize_fixations=True, regularization=0.01**
  - L2 regularization notes
    - Without regularization, all gazes rapidly converge to +/- 160.0 (i.e. corners of the image)

        - Occurs even when **--num_transformers=1** (only one gaze is learned)
      - With regularization l2=0.01, the gazes tend to converge to the range [-10, 10] instead (i.e. all near the center of the image)
        - This convergence may not happen consistently
        - Leads to improved standard and adversarial performance
      - Regularization l2=0.001 leads to worse standard performance than l2=0.01
    - Model gradients are not vanishing or exploding
      - Gradient clipping does not resolve gaze convergence issue
    - Standard and adversarial performance declined when **--shared=1** (i.e. when **resnet** weights were shared between branches)
- Bilinear Sampling
    - Bilinear sampling technique increased adversarial robustness of Retinal Sampling model