Ani-GIFs: Human Action Recognition - Domain Generalization

Following my work on Domain Adaptation in Images, we decided to tackle the more challenging problem of Domain Adaptation in Videos. However after reviewing some of the literature in this topic and considering applications of our work, we realised that a more realistic situation would be a Domain Generalization problem. After going through the literature we realised that there is very little research output in this domain primarily due to the fact that there is no dataset available for the task of video domain generalization. Another challenge is to clearly identify what these distinct domains could be. We created the 1st Video action recognition dataset with two distinct domains curated specifically for domain generalization. For the real domain, we used the Kinetics-600 dataset which is a benchmark dataset for human action recognition. The second domain we identified was synthetic domain. We reached this decision by reviewing the recent increase in popularity of synthetic training environments such as Carla and OpenAI's gym.

Our approach began by collecting the synthetic counterpart for all the classes in the Kinetics dataset. We wrote a script to scrape videos using the keywords "cartoon", "animated" and "graphics". We decided to work on GIFs instead of videos because GIFs are shorter length videos which condense the same spatial data in a lesser number of temporal timesteps and hence easier to train on with resource limitations. We used a python script using selenium for scraping these GIFs from the Internet. The next step after scraping GIFs using all 3 keywords was to clean and filter the data. The filtering process was done by us(4 graduate students) ourselves and we decided not to outsource this task. We did not want to sacrifice on the quality of the dataset by introducing bias. This task involved looking at each GIF personally and comparing them to the GIFs in the real domain for the same class. We made sure there was consistency in features between both domains.

The experiments involved running baselines on the dataset using a traditional action recognition model. We then suggested two techniques that could be used to improve performances across different domains. For baselines, we used a pretrainied I3D model, training on Kinetics. We used the synthetic domain as our test domains. Since this model was trained on full length videos and not GIFs, the model was not able to align the temporal shifts in the source videos and target GIFs. We followed this experiment by retraining the entire model on the AniGIFs dataset. We trained the model on the real domain and tested on the animated domain and vice versa. This gave us an effective baseline and highlighted the challenge of domain shift.

To provide a solution to the problem of spatio-temporal alignment, we extended existing generalization techniques for images in a spatio-temporal setting. We used two methods, Pseudo Labelling and Augmentation. Both these methods are well known generalization techniques in the literature and have performed very well when used on images. The challenge here was temporal alignment. For Pseudo labelling, we start with a model trained on a domain. At test time, we classify GIFs from the unseen domain and if the model confidence is above a certain value, we add that training example with its predicted label to the training set and retrain further.

The second approach we used was Augmentations. This is a widely used method of generalization in images and we extended this work to videos. We start with a set (T) of identity transformations and after a certain number of iterations, we randomly sample transformations from a set of predetermined transformations provided by the PIL library. We iteratively sample transformations from this set and apply them to all the frames of our input GIF and calculate the prediction error. At the end of this process, we select the transformation that led to the highest errors on a training batch and add them to the set (T). The model now will train GIFs not only on the identity transformation, but also on this newly determined transformation that the model is vulnerable to.

This project led to the creation of the 1st video human action recognition dataset specific for Domain Generalization. We hope that this dataset acts as a benchmark which facilitates research in this field which has recently gained popularity due to its real world application. We add novelty in terms of extending currently accepted generalization techniques for images to a temporal setting. This also showed the challenge that the spatio temporal setting poses as compared to just a spatial domain shift in images. It also highlights the need for better generalization algorithms.