Bosch Model Extraction

Team Number 6

March 2022

1 Black Box Attacks

For the given paradigm we make use a synthetic data generator model to generate videos, which can be used to train the student model. We expand upon this generation process and querying process in the following subsections.

1.1 Synthetic Data Generation

Due to limited access to class-conditioned Video GANs and difficulty in training a model from scratch, we attempt to reuse a pre-trained BigGAN model trained on the ImageNet dataset. Specifically, the BigGAN is fine-tuned to generate fake samples which can fool the black-box teacher classifier to predict the required class. We sample a random latent variable, learnable class embeddings which are then passed onto the BigGAN generator and the image, repeated to create a video is passed onto the teacher classifier. Keeping the generator, classifier frozen, we train the embeddings to minimise the predicted teacher class logit and sampled class (the classes are sampled randomly and hence we do not assume prior information about the dataset). This allows for fine-tuning in very few steps. Fig. 1 represents a collection of synthetic images for the class "air drumming (class index 1)". While the image does not look visually meaningful, the output probability of the teacher model for the given image > 0.9. To reiterate, we do not require prior information about the dataset to fine-tune the generator and the above example image is to indicate our generation approach. Therefore, the synthetic samples represent the training distribution of the teacher model and hence can be used to distill the student model.

1.2 Student Model

We propose two student networks each to distill the convolution-based MovieNet and transformer-based Swin-T teacher models respectively. For the former, we use a pretrained resnet50 model and extract per-frame features which are then weighted pooled using a linear layer. For the later, we use a pretrained VIT-T model and extract per-frame features which are then mean-pooled to predict the required class.

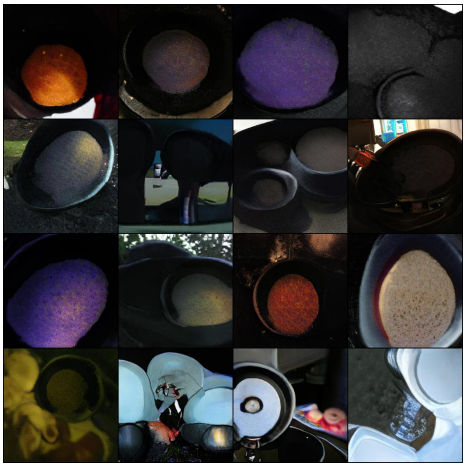

Figure 1: Synthetic images generated by the fine-tuned BigGAN for the class "air drumming".

1.3 Training

We use the generated samples per class, student models defined above and perform standard knowledge distillation. Fig. 2 represents the overall pipeline of our training procedure. Table. 1, describes the obtained results for the black-box model extraction.

Model	Accuracy
MovieNet	12.43
Swin-T	7.8

Table 1: Black-Box Model Extraction Results.

2 Grey Box Attacks

We follow the technique of distillation to perform model extraction. For both the given tasks we make use of an auxiliary irrelevant dataset ie: Moments in Time. We pass each video from Moments in Time (MIT) to the teacher model (SwinT / MoviNet v2) and save the corresponding logits. These logits are then used to distill information into the student model. Once we have a distilled model we further fine-tune it using the 5% training data that was provided for each task.

2.1 Lottery Ticket Hypothesis

Given a larger model, as the data-size decreases, it is expected that the model performance decreases . Therefore, sparsity can be understood as a regularization technique to avoid over-fitting. One of the popular ways of obtaining

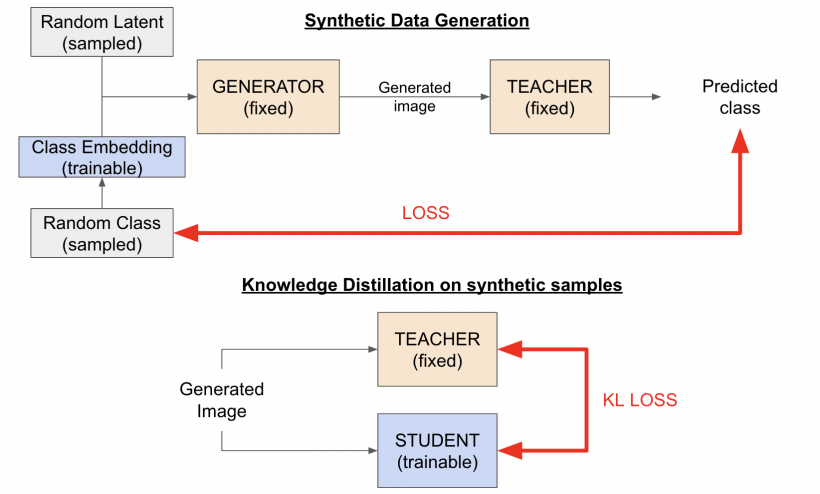

Figure 2: Black Box Model Extraction pipeline.

a sparse network of similar architecture but lower capacity is via pruning. J Frankle et al 2018 formulated the lottery ticket hypothesis (LTH) and proposed iterative magnitude based pruning (IMP). They demonstrated that dense networks can be magnitude pruned to obtain highly sparse sub-networks ("winning tickets") which can match or potentially outperform the original dense network. This approach also induces an inductive bias specific to the task to be learned, which leads to a better network when compared to a dense network of similar size trained from scratch. Recent work has also indicated that the number of samples required to achieve zero generalization error is proportional to the number of the non-pruned weights in the hidden layer. Therefore, we make use of IMP to fine sparse networks during the fine-tuning step on 5% data.

2.2 Training

We reuse the student models defined in the previous section and perform distillation using the auxillary dataset Moments in Time. After distillation, we perform IMP to finetune the model and identify a winning ticket on the 5% data. Fig. 3 represents the overall pipeline of our training procedure. Table. 2 summarizes our results on the gray box setting.

Model	Accuracy
MovieNet	33.5
Swin-T	17.3

Table 2: Gray-Box Model Extraction Results.

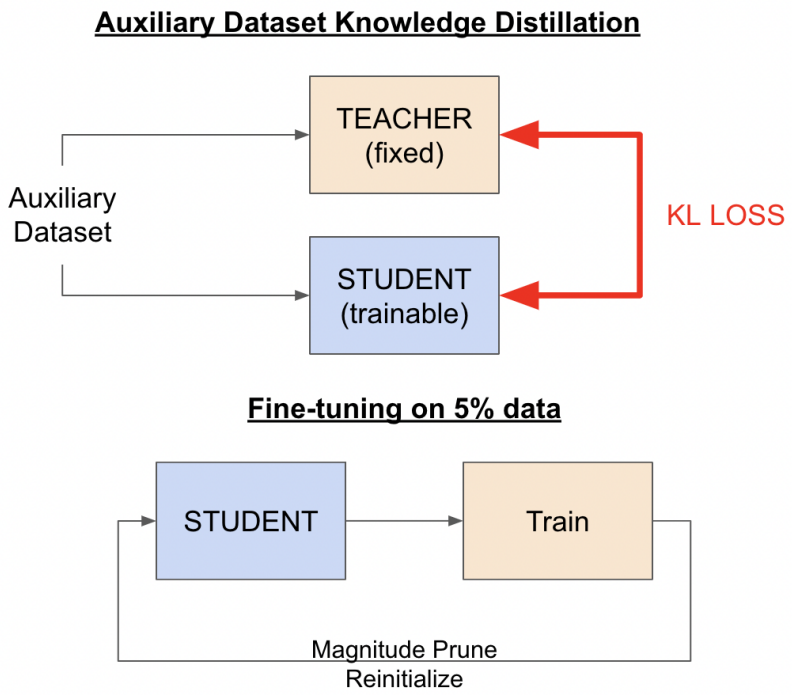

Figure 3: Gray Box Model Extraction pipeline.