Model Extraction Attack For Video Classification

Submitted by-
Team 17

Introduction

- Model extraction is a kind of attack in which the adversary aims to replicate the performance of the victim model while ensuring that minimum queries and access to the victim are required in the process.

- The Problem Statement entasked us with devising novel strategies for the extraction of two video-based victim models - Swin-T and MoViNet A2 under two settings - Black Box and Grey Box.

Assumptions

- The victim model architecture is completely unknown in both black box and grey box setting, and it returns only the predicted class label in response to a query.
- The number of classes in the original dataset on which the victim was trained is well known in advance.
- For the black box setting, it was assumed that no information about the dataset used to train the victim model was known.
- For the grey box setting, the dataset used to train the victim model was known in advance.

Setting A: Black Box

Methodology

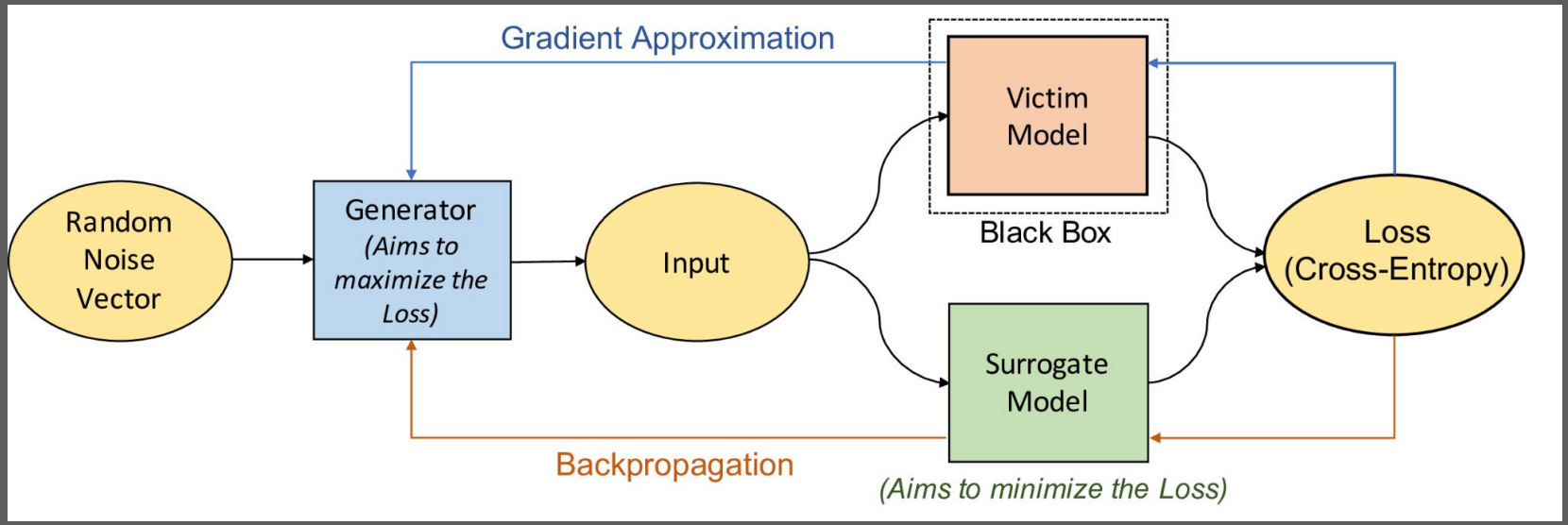

Figure 1: Overall flow diagram of the proposed methodology under Black Box setting

[1] Truong, Jean-Baptiste, et al. "Data-free model extraction." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

Methodology

- To generate 4D tensors, we repeated the 3D tensor across temporal dimensions to make it 4D.
- This was necessary to reduce the computational requirements, given our lack of adequate resources.
- Then, choosing cross entropy as the loss, we trained the surrogate model to minimise the loss while the generator was trained to maximise it.

Experimental Settings

- **Generator**: Vanilla 3D GAN without Discriminator
- **Surrogate Model**: R(2+1)D[2] ResNet
- **Iterations**: 500 iterations, Generator updated 1 step, model updated 5 steps every iteration.
- Used Learning Rate Schedulers to decrease learning rate gradually over the iterations.
- Used SGD as the optimiser for the Generator and ADAM for the surrogate.
- **Loss function**: Cross Entropy.

[2] Tran, Du, et al. "A closer look at spatiotemporal convolutions for action recognition." *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2018.

Results

- For evaluating our surrogate models, we took 50% of the validation split of the Kinetics 400 and 600 datasets, while ensuring equal representation of classes.

- We also benchmarked the victim models (Swin-T and MoViNet A2) on the validation split in order to gain a better insight into the relative performance of our models.

Results

The results we obtained for the black box setting are given below:

Table 1: Metrics obtained in Black Box setting

Setting	Dataset	Top 5 (Victim Model)	Top 5 (Extracted Model)	#Queries made
Black box	Kinetics 400	42.59%	1.25% (P1)	64,000
	Kinetics 600	1.16%	0.83% (P2)	64,000

Challenges we faced

1. The biggest issue we had was the lack of computing resources, due to which our models were severely undertrained.
2. As both the kinetics datasets are quite large in terms of storage requirements, downloading them was also a challenge.
3. We also faced the perilous issue of mode collapse[3] in the video generator. This happens when the generator only produces the videos misclassified by discriminator, leading to reduced variations in the videos generated. Using Conditional GAN might solve this issue, but we lacked the resources to train and implement it.

[3] Thanh-Tung, Hoang, and Truyen Tran. "Catastrophic forgetting and mode collapse in GANs." *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020.

Setting B: Grey Box

Methodology

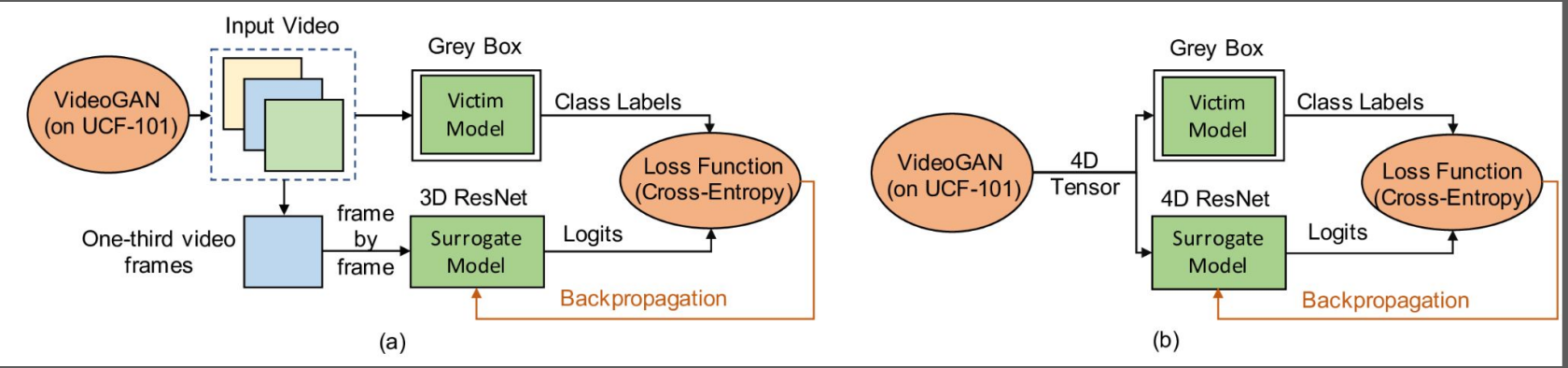

Figure 2: Overall flow diagram of the proposed methodology under Grey Box setting: a) Swin-T as Victim b) MoViNet as Victim

[4] Yan, Wilson, et al. "VideoGPT: Video generation using VQ-VAE and transformers." *arXiv preprint arXiv:2104.10157* (2021).
[5] Gao, Shang-Hua, et al. "Res2Net: A new multi-scale backbone architecture." *IEEE transactions on pattern analysis and machine intelligence* 43.2 (2019): 652-662.

Methodology

- We used a pre-trained GAN[4] (trained on UCF-101 dataset) to supply generated tensors. Given the lesser requirement of computational power, we decided that our strategy should test 3D as well as 4D ResNets, and pick the one most appropriate as per the model being attacked.
- Hence, we found that 3D ResNet (Res2Net 101[5]) was more stable in case of Swin-T extraction. Once again, cross entropy was chosen for the loss function.

[4] Yan, Wilson, et al. "VideoGPT: Video generation using VQ-VAE and transformers." *arXiv preprint arXiv:2104.10157* (2021).
[5] Gao, Shang-Hua, et al. "Res2Net: A new multi-scale backbone architecture." *IEEE transactions on pattern analysis and machine intelligence* 43.2 (2019): 652-662.

Experimental Settings

- **Swin-T Victim**
  - **Generator**: Video Generator pre trained on UCF101
  - **Data**: 1/3rd video frames generated by GAN with corresponding labels
  - **Surrogate Model**: Res2Net 101
  - **Optimiser**: SGD
  - **Epochs**: 10 epochs with 500 iterations in each epoch
  - Learning rate schedulers were used for each setup
  - **Loss function**: Cross Entropy

- **MoViNet A2 Victim**
  - **Generator**: Video Generator pre trained on UCF101
  - **Data**: Videos Generated by GAN with corresponding labels
  - **Surrogate Model**: R(2+1)D ResNet
  - **Optimiser**: ADAM
  - **Epochs**: 58 epochs with 100 iterations in each epoch
  - Learning rate schedulers were used for each setup
  - **Loss function**: Cross Entropy

Results

- As in the case of black box setting, we took 50% of the validation split of the Kinetics 400 and 600 datasets, while ensuring equal representation of classes for evaluating our surrogate models,

- We also benchmarked the victim models (Swin-T and MoViNet A2) on the validation split in order to gain a better insight into the relative performance of our models.

Results

The results we obtained for the grey box setting are given below:

Table 2: Metrics obtained in Grey Box Setting

Setting	Dataset	Top 5 (Victim Model)	Top 5 (Extracted Model)	#Queries made
Grey box	Kinetics 400	42.59%	1.29% (P1)	43,200
	Kinetics 600	1.16%	0.83% (P2)	3,200

Challenges we faced

- Here as well, we were limited by our lack of computational resources and this prevented us from using our model to its full potential.

- To avoid out-of-memory (OOM) errors on the GPU, we had to limit the number of frames taken per video to 32 for Swin-T and 16 for MoViNet when running the evaluation scripts on Google Colaboratory.

Concluding Remarks

The problem statement was quite novel and practical, and aimed to understand an important aspect of model privacy. The lack of literature especially for video-based models called for detailed understanding of various aspects of the implementation pipeline.

We are confident that our approach would result in substantially better results if given adequate computational and temporal resources.

Thank You

Presented By

Team 17