SCHOEN Adrien

Rapport de réunion n°5

Date de la réunion : 26/02/2021

Date de la prochaine réunion : 05/032021

Ordre du jour : début du code

1. Ce que le stagiaire a dit

Présentation du code. On génère des batchs d'images avec les labels correspondants. Ces images représentent des morceaux du code binaire des paquets. Dans chaque image une valeur binaire est répétée plusieurs fois sur les lignes et sur les colonnes.

Les paquets sont lus un par un dans différents fichiers choisis aléatoirement. On ne peut néanmoins pas ouvrir toute la base de données de cette façon, il existe une limite au système d'ouverture de fichier.

Il y a un problème avec le dataset, je ne possède pas tous les fichiers. La plupart des attaques ne sont pas présentes.

Il y a un souci lors du découpage des paquets. Le dernier bout est trop petit pour rentrer dans une image ; j'ai essayé d'adapter le nombre de répétition des valeurs binaires, mais je ne suis pas arrivé à obtenir une image de la même taille que les autres. J'ai donc mis un appel à une fonction de redimensionnement des images en attendant.

Il existe d'autre moyen prometteur de séquencer les codes binaires des paquets. On peut notamment rester en une dimension et séquencer ces codes comme des vecteurs.

2. Ce que les encadrants ont ajoutés

Ne pas chercher le code optimal : il faut surtout un code qui marche. Si on arrive pas à ouvrir tout le dataset avec les piles, ne pas hésiter à faire plus simple : diviser les pcap de façon à pouvoir tous les ouvrir assez rapidement.

Il n'y a normalement pas de souci avec le dataset, il faudra sans doute le re-télécharer

On peut utiliser du bourrage pour les morceaux de paquet les plus petits, mais le mieux reste de ne pas diviser les paquets. En effet on risque de devoir entraîner notre GAN à générer des bouts de paquets que l'on ne pourra pas utiliser seul. Il faut que le GAN connaisse l'ordre des différents bouts de paquets (on peut alors penser à un CGAN ou le label serait la place de l'image dans le paquet) mais le plus simple est de se restreindre à une image = un paquet

Il faut assez vite définir une méthode d'évaluation pour les données générées, surtout si on veut essayer plusieurs méthodes de séquençage différentes.

On pourra plus tard essayer de générer des caractéristiques plus haut niveau (comportement utilisateur …) mais pour l'instant il faut se concentrer sur l'analyse bas niveau(payload, paquet, flux).

3. Ce qu'il faut faire pour la prochaine séance

Télécharger à nouveau le dataset et s'assurer de son intégrité

Mettre en place le séquençage 1 image = 1 paquet

Implémenter le padding pour les paquets les plus petits.