SCHOEN Adrien

Rapport de réunion n°13

Date de la réunion : 23/04/2021

Date de la prochaine réunion : 30/04/2021

Ordre du jour : Début analyse séquentielle

1. Ce que le stagiaire a dit :

Je suis passé à des GAN convolutifs, mais je pense que le séquençage en image n'est pas vraiment optimal pour notre problème. En effet, même dans le payload, certains champs peuvent dépendre d'autres champs du payload, voir d'autres champs des autres paquets de la séquence de communication.

J'ai régénéré le dataset en me concentrant sur les paquets UDP. J'ai ré-entraîné les modèles sur le payload des paquets IPv4+UDP, en supprimant les 42 premiers octets (l'ensemble des headers). Ce modèle ne nous fournit donc qu'un payload UDP.

L'idée sera donc de fournir à un programme un descriptif du paquet que l'on souhaite générer (IP source, port source, protocole etc.) et de générer le paquet correspondant, en ayant un payload dont les valeurs des champs suive la distribution des valeurs des champs des payload du jeu d'entraînement.

J'ai écrit un programme simple qui me renvoie un paquet UDP avec payload généré, qui prend en entrées les valeurs à inclure dans les headers (IP source, IP destination, adresses MAC etc.)

Mon idée pour continuer dans cette direction serait d'avoir plein de petits modèles différents (peut-être un modèle par protocole applicatif : DNS, HTTPS, HTTP…) qui ne généreront que les payload relatifs à ces applications, et d'avoir un autre modèle qui lui, générerait des flux pour ces applications Un flux serait composé de plusieurs paquets. Je pourrais essayer d'entraîner un modèle à générer des caractéristiques de flux (comme les caractéristiques NetFlow) et à créer des paquets en conséquences

2. Ce que les encadrants ont ajouté :

Il faut continuer à voir ce qu'il se fait du côté de la génération de séquence. Que ce soit pour la génération de caractéristique ou pour le payload, le séquençage des données en image pose deux gros problèmes : d'une part les données traitées doivent être toutes de la même taille et de l'autre, on perd l'information conditionnelle qu'apporte un octet sur un autre.

On peut essayer de voir sur ce qui se fait en NLP et continuer les recherches qui ont été menées cette semaine en ce sens. Par exemple Sequ2Sequ est un modèle d'autoencoder récursif qui encode une séquence x1x2x3….xn dans un espace latent et la reconstruit sur une base de RNN/LSTM.

Générer du payload chiffré n'est pas vraiment intéressant,

La génération de paquets UDP que j'ai présenté ne peut pas être évalué de la même manière qu'avant. Auparavant, on essayait de voir si scapy reconstruisait correctement les paquets. Ca ne peut plus nous servir de critère maintenant car le paquet est directement construit par les informations données par l'utilisateur. Il faudrait essayer de voir si le payload généré pour un type de trafic UDP (DNS par exemple) est cohérent avec notre dataset. Par exemple est ce que les valeurs des différents octets suive la même répartition.

Si on séquence un payload en une image (ce qui est le cas à présent mais ne sera bientôt plus le cas) : on a un nombre d'octet fixe dans notre payload (98) . On pourrait donc essayer de décrire la répartition des valeurs de ces 98 octets via des courbes de distributions ou des diagrammes à moustache pour essayer de voir si les répartitions sont proches de celles du jeu de test.

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

Préparer la présentation de mardi

Reprendre l'étude bibliographique, en ajoutant les études découvertes cette semaine.

Évaluer les distributions de valeur des payload générées.