SCHOEN Adrien

Rapport de réunion n°8

Date de la réunion : 19/03/2021

Date de la prochaine réunion : 26/03/2021

Ordre du jour : Début entrainement

1. Ce que le stagiaire a dit

J'ai évalué les autoencoder par la reconstruction scapy de leurs données générées. Au début je me contentais de voir s'il y avait bien toutes les couches qui étaient reconstruites par scapy (Eth/IP/TCP…) mais ce n'était pas un critère suffisamment discriminant. J'ai donc décidé de regarder non plus au niveau des couches mais au niveau des champs, pour voir si les données générées étaient crédibles avec les données d'entraînement. J'ai créé une fonction qui attribue un score en fonction des différents champs qu'a pu recréer scapy, plus le score est faible, plus le modèle génère des données réalistes.

J'ai ensuite essayé de transvaser mon programme d'entraînement sur mon gros PC. Il a fallu pour ça que je change l'étape de chargement des données. Pour le résultat préliminaire je me contentais de 1000 paquets et je les chargeais en itérant depuis une liste de chemins. Toutefois, j'ai ici plusieurs dizaines de millions de paquets et je ne peux pas me permettre d'instancier une liste.

J'ai donc commencé par essayer de charger les données aléatoirement mais cela ne me permet pas de parcourir l'ensemble des données, je peux retomber sur une image déjà tirée. Il faudrait que je garde en mémoire le nom des fichiers déjà tiré et cela saturerait ma RAM.

Je pense refaire la génération du dataset pour résoudre ces problèmes.

J'ai donc implémenté une lecture par ordre alphabétique des fichiers. Je lis les N premiers, puis je lis les N suivants. Il faut juste que je fasse attention à bien réinitialiser le compteur entre chaque epoch

J'ai essayé d'entraîner le modèle, mais j'ai dû réduire le nombre de données de 83 millions à 100 000 pour avoir un espoir de voir la fin de l'entraînement. J'ai aussi essayé d'implémenter du multiprocessing sur ma carte graphique, mais j'ai appris que c'était uniquement disponible sur Linux.

2. Ce que les encadrants ont ajoutés

Dans la génération, la valeur d'un octet peut influencer l'interprétation que donne scapy aux autres octets. Si le Xieme octet est faux, les X+nieme octet peuvent être bien générés, scapy en aura peut être une interprétation complètement erronée. Il faudrait voir s'il serait possible d'implémenter un processus d'attention dans le modèle.

Les champs que génèrent scapy ne sont pas retranscrits dans l'ordre lors du calcul du score, mais ça pourrait être dû à la fonction de score en elle-même, pas à un défaut de génération. Il faudrait de plus, que ce score évalue la diversité des données générées

Peut-être une autre forme de séquençage à base d'embedding est envisageable

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

Basculer sur Igrida et reprendre l'entraînement sur ce service

Détailler le calcul du score et peut être implémenter des mesures déjà introduites dans l'état de l'art