SCHOEN Adrien

Rapport de réunion n°7

Date de la réunion : 12/03/2021

Date de la prochaine réunion : 19/03/2021

Ordre du jour : Programmation autoencoders

1. Ce que le stagiaire a dit

J'ai utilisé pytorch pour coder divers auto encoders sur Google colab, que j'ai entraîné à l'aide de 1000 paquets pré-traités issus de ma base de données.

J'ai ensuite utilisé une librairie Git Hub pour coder des modèles plus complexes. Il faut faire attention à adapter la dernière couche pour qu'elle corresponde à nos données d'entrée (La sortie du modèle doit être de la même forme que l'entrée).

J'ai pour l'instant essayé deux modèles complexe, les VQ-VAE et les B-VAE. Les VQ-VAE permettent d'avoir des espaces de représentation discrets, ce qui pourrait être intéressant pour certaines de nos caractéristiques (exemple : protocole)

J'ai essayé de voir différents critères d'évaluation. D'abord l'erreur binaire induite par la reconstruction : combien de bits sont mal reconstruits par le modèle. Mais ce critère ne semble pas adapter à notre problème : en effet il évalue la pertinence de la reconstruction mais pas la plausibilité des échantillons générés.

J'ai alors essayé de voir quelle allure avaient les données directement générées par le modèle. Il n'y avait pas de padding

Cette mauvaise génération peut avoir pour cause le très faible volume de données d'entraînement. Je n'ai pas disposé du dataset entier sur Google colab. J'ai beaucoup plus de données à mon domicile

2. Ce que les encadrants ont ajoutés

Tous les modèles ne sont pas forcément bons à implémenter, il faut savoir pourquoi ils ont été développés et voir si leurs avantages relatifs correspondent à nos besoins.

Si l'initialisation du dataset est trop longue (car la liste des fichiers est trop lourde), on peut essayer de contourner l'étape de dénombrement des fichiers en utilisant intelligemment leurs propriétés de génération (nom du fichier, chemin absolu...)

Comme critère d'évaluation initial, on pourrait essayer de voir si le paquet généré est reconnu comme tel par scapy. On pourrait ensuite voir s'il ne lève pas d'alerte de la part d'un IDS

Reprendre l'étude bibliographique si nécessaire et si je ne comprends pas bien des modèles.

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

Reprendre entraînement des modèles avec le plus de données possible.

Essayer la reconstruction en paquet à l'aide de scapy comme critère d'évaluation

Essayer d'avoir accès au cloud.