SCHOEN Adrien

Rapport de réunion n°14

Date de la réunion : 30/04/2021

Date de la prochaine réunion : 07/04/2021

Ordre du jour : Début analyse séquentielle

1. Ce que le stagiaire a dit :

Notre étude est actuellement face à 2 problèmes et une question.

Problèmes :

- Si l'on veut générer du trafic réaliste, il faut que l'on prenne en compte l'échelle des flux. Les communications peuvent parfois s'étaler sur plusieurs paquets et certaines informations contenues dans un paquet peuvent ne se comprendre qu'au regard des autres paquets dans le flux.

- La traduction du payload en image n'est peut-être pas la manière la plus simple et la plus efficace de représenter l'information. Lorsque l'on convertit les octets en pixel et que l'on effectue des convolutions successives sur ces pixels, on peut perdre l'information mutuelle entre les octets. On risque de perdre les relations de structure. Pour préserver ces relations de structure, on peut envisager de voir ce qui a été fait du côté du NLP.

Question:

-Que faire du trafic chiffré, beaucoup de paquets sont chiffrés et je ne vois pas vraiment comment générer cette forme de paquet.

Cette semaine j'ai fait de l'étude bibliographique pour voir comment l'état de l'art répondait à ces problèmes.

J'ai découvert beaucoup d'études et j'ai résumé certaines d'entre elles. Il y a des points assez communs dans ces études. Souvent on ne convertit pas les paquets en image ni même en suite numérique, mais on réalise un embedding en regroupant plusieurs octets entre eux (on considère souvent des unités de 2 octets). Beaucoup de papiers utilisent aussi un mécanisme d'attention pour transmettre des dépendances sur de longue séquences, mais le problème de la taille reste souvent un point de discussion, certaines études fixent une taille standard pour les séquences.

Des méthodes d'anonymisation sont souvent employées dans les papiers. On remplace les valeurs de certains champs sur lesquels on ne souhaite pas entraîner le modèle par des valeurs aléatoires. C'est souvent le cas des champs IP

2. Ce que les encadrants ont ajouté :

Il faut essayer de garder une distance critique par rapport à ce que l'on lit dans la littérature. On ne trouvera sans doute pas de papier qui correspond à nos attentes, mais on peut essayer de prendre des idées dans un peu de chaque papier

La question sur le trafic chiffré est légitime, mais elle est un peu prématurée, on devrait d'abord se concentrer sur du trafic simple à générer (même si l'on doit grandement réduire le dataset) et une fois que l'on aura obtenus des résultats convaincants, on pourra envisager d'élargir notre étude à d'autre forme de trafic.

Il faudrait aussi commencer à revoir le format de présentation, peut-être préparer un léger powerpint pour ne pas à avoir à chercher tout le temps

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

Continuer la recherche bibliographique et les résumés d'études

Décrire un modèle de génération globale avec l'analyse séquentielle

Préparer une présentation pour la prochaine réunion