SCHOEN Adrien

Rapport de réunion n°11

Date de la réunion : 09/04/2021

Date de la prochaine réunion : 16/04/2021

Ordre du jour : Résultats VAE

1. Ce que le stagiaire a dit

J'ai réalisé l'entraînement du BetaVae sur 20 et sur 50 epochs. Il y a eu un phénomène de surentraînement, il faudra que j'implémente un mécanisme de sortie d'entraînement afin de l'éviter. Même sans ce problème, le modèle entraîné n'arrive pas à bien reconstruire les données d'entrées. Par exemple au lieu de faire un padding de 0, il peut arriver que le modèle réalise un padding de 2 ou de 3. L'évaluation des deux modèles s'est faite dans un notebook et nous avons pu voir un schéma des modèles.

J'ai ensuite travaillé sur l'exploitation des modèles de VQ VAE entraînés. Durant l'entraînement des VQ VAE nous avons déterminé des représentations discrètes pour les distributions de probabilité des données d'entrée. Nous avons ainsi des images en entrée que nous arrivons à projeter sur un embeding. Le problème est que nous n'arrivons pas à faire l'association entre les images et leurs représentations dans l'embedding, on doit donc entraîner un modèle de réseau récursif appelé PixelCNN. Ce réseau va essayer de prédire la représentation sur l'embedding des images. Je n'ai pas réussi à faire cette étape, j'ai des problèmes de dimensions qui ne se correspondent pas entra la sortie de PixelCNN et la sortie de l'encoder

J'ai trouvé un notebook qui me permettrait de résoudre ce problème, étant donné qu'il réalise toutes les étapes de l'entrainement du VQ VAE.

J'ai regardé les différents protocoles des paquets. Peut-être serait-il judicieux de restreindre la génération uniquement aux payloads d'un seul type de paquet.

2. Ce que les encadrants ont ajoutés

On peut continuer dans deux directions : soit on spécialise notre génération sur une partie d'un certain type de payload et on continue la génération de séries d'octets, soit on essaie de regarder à plus haut niveau et on essaie de voir des associations de regles pour pouvoir créér des suites de paquets.

Si l'on continue la génération au niveau des payload, il faut commencer par réduire la diversité du dataset et simplfier les modêles. On pourrait essayer de reprendre ce qui a été fait dans le papier PAC-GAN

On pourrait commencer par se concentrer sur la génération de données UDP. Attention toutefois, certains champs sont aussi à déterminer : notamment le port source. Pour certains de ces champs contextuels, on pourra utiliser le travail réalisé sur IP2Vec

Si l'on souhaite travailler à plus haut niveau, il faudra comprendre l'utilisation et la création des DSL, présentés aux dernieres réunions.

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

Continuer la génération de paquets entiers par VQ VAE en exploitant l'entraînement de PIxel CNN

Reprendre la génération du dataset en se concentrant uniquement sur une seule journée et sur un seul type de paquet.

Refaire le travail réalisé à l'aide d'autoencoder, mais cette fois-ci avec des GAN.