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 VQ-VAE et GAN

1. Ce que le stagiaire a dit :

J'ai entraîné le VQ-VAE sur l'ensemble du dataset, mais les résultats ne sont pas très bons. Le modèle génère des paquets que scapy n'arrive à reconstruire que moyennement bien (Jusqu'à la couche IP, mais pas la couche 4).

Pour les GAN, j'ai récupéré les codes des GAN et WGAN-GP que j'ai adaptés à nos données, j'ai vérifié que l'entraînement marchait. J'ai bien aperçu le point où le discriminateur et le générateur oscillaient sur leur point d'équilibre. J'ai, toutefois, fait ça assez vite et, pour ne pas me compliquer la tâche, j'ai mis des couches denses au lieu de couches de convolution.

Je pense qu'on pourrait essayer de retrouver les résultats de PAC-GAN en re-générant la base d'entraînement notamment en supprimant certaines parties qu'on ne souhaite pas faire générer par l'intelligence artificielle.

On peut aussi essayer de considérer le paquet (ou sa réduction) comme une séquence et essayer d'appliquer les résultats obtenus sur les séries temporelles (LSTM, RNN, Transformer) ou en NLP.

2. Ce que les encadrants ont ajouté :

Il faudrait voir du côté de la recherche en NLP pour essayer de déterminer des relations entre les champs des paquets. Si on considère un paquet comme une séquence d'octets, cette séquence peut être découpée en différents champs. Pour faire un parallèle avec le NLP, c'est comme si notre paquet était constitué de mots et que les valeurs que prenaient ces mots s'influençaient les unes et les autres.

Exemple : La phrase : « La Voiture est rouge » est constituée de quatre mots. Supposons que nous voulions générer la valeur du quatrième mot : « rouge ».

Le fait que le deuxième mot soit « Voiture » augmente la probabilité que le quatrième mot soit « rouge » (une voiture rouge, c'est logique et fréquent). Supposons maintenant que le deuxième mot soit « maîtresse », la probabilité que le quatrième mot soit « rouge » deviendrait moindre. On a bien ici une influence de la valeur du 2^{e} champ sur la valeur du 4^{e} champ.

Le NLP a énormément progressé ces dernières années et peut être pourrait-on essayer d'en exploiter les avancées en essayant de faire le rapprochement entre nos paquets et des phrases.

Au contraire, l'embedding Word2Vec, lui, nous donne une relation des valeurs au sein des champs. Avec word2vec, j'obtiens une mesure de la proximité contextuelle des éléments d'un même champ. Par exemple j'obtiens que le mot « Voiture » est proche contextuellement du mot « Moto », mais je n'obtiens pas d'informations sur les relations des champs entre eux.

On peut plutôt essayer d'élargir le contexte à des séquences avec un algorithme SequTo Sequ qui utilise des LSTM.

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

Terminer de coder les deux GAN en remplaçant les couches denses par des couches de convolutions

Reprendre la segmentation des paquets et déterminer quel paquet veut-on générer avec du Machine learning.

Commencer à voir des travaux de génération de séquences temporelles en NLP