Institut Cybersécurité Occitanie

Défis Clés OCCITANIE

La Région Occitanie Pyrénées · Méditerranée

Génération de trafic d'attaque réseau

Approches et représentation du trafic

Gabin Noblet

Journée Scientifique ICO – 12 juillet 2024

LAAS CNRS

CUSTOCY®

cnrs

Thèse encadrée par : Philippe OWEZARSKI, LAAS-CNRS & William RITCHIE, Custocy

Détection d'intrusion réseau

Besoin de **beaucoup de données**

Développement

Benchmark et analyse

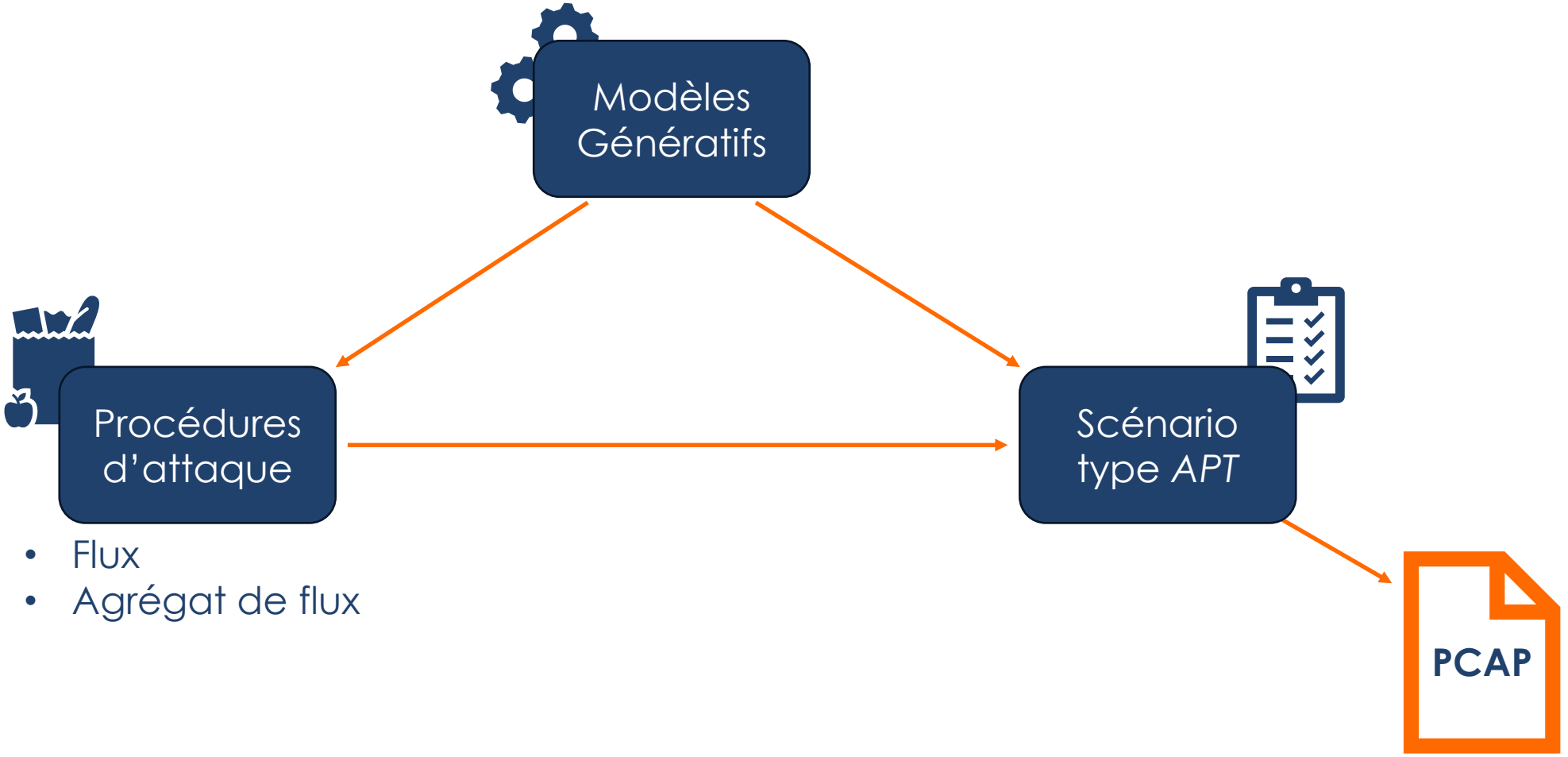
Difficultés de collecte

Données personnelles

Architectures complexes

Labellisation

Étudier la génération de trafic d'attaque réaliste



- Flux
- Agrégat de flux

Génération d'attaques élémentaires

- Choix d'une **représentation de trafic**
  - Caractérisation du trafic adaptée à la problématique
- Choix de **modèle génératif** adapté
  - Architecture de modèle adaptée à la donnée
- Méthodes de **validation** et d'**évaluation** des modèles

Dataset utilisé pour cette étude

Dataset public mis à disposition par la **CTU** (Czech Technical University)

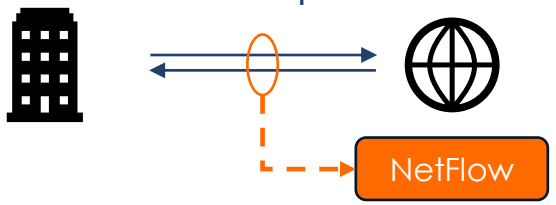
- Composé de flux de ***Command & Control*** (**+500k** flux)
- Trafic **TCP** uniquement
- **Trafic réaliste** d'un botnet type ***TrickBot***

Représentation du trafic

Approches dans la littérature

Ring et al. "Flow-based network traffic generation using Generative Adversarial Networks"

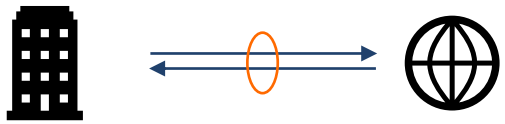
Générateur → Génération de métriques de flux



❌ Echelle du paquet
✅ Notion de flux

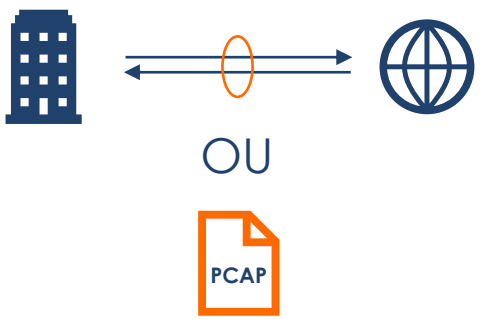
Dowoo et al. "PcapGAN: Packet Capture File Generator by Style-Based Generative Adversarial Networks"

Générateur → Génération d'un ensemble de paquets



✅ Echelle du paquet
❌ Notion de flux

Une approche hybride adaptée au trafic d'attaque



Extraction de listes de paquets par flux (session TCP)

Représentation simplifiée d'un paquet

- Temps inter-paquets
- Sens dans le flux
- Taille de la payload
- Flags TCP

✅ Echelle du paquet
✅ Notion de flux

Une approche hybride adaptée au trafic d'attaque

Exemple flux de *Command & Control* :

IAT (s)	Direction	Payload (bytes)	Flags TCP
0.000	Forward	0	S
0.082	Backward	0	SA
0.001	Forward	0	A
0.022	Backward	388	PA

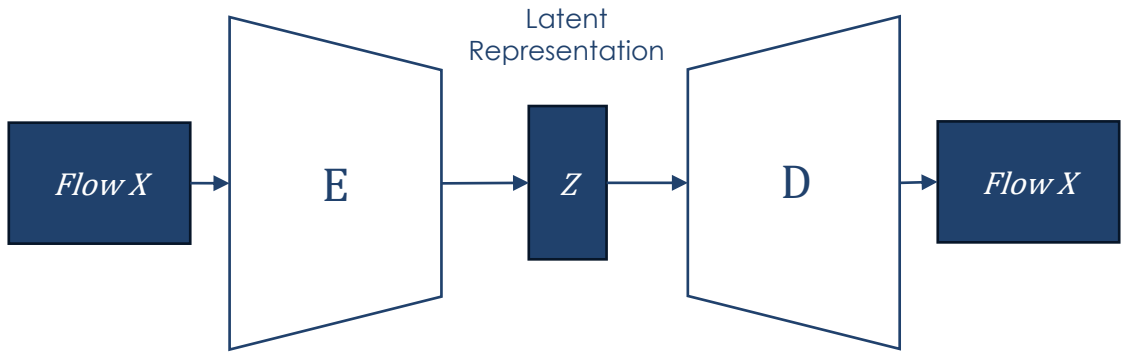
Connexion

Génération de trafic réseau

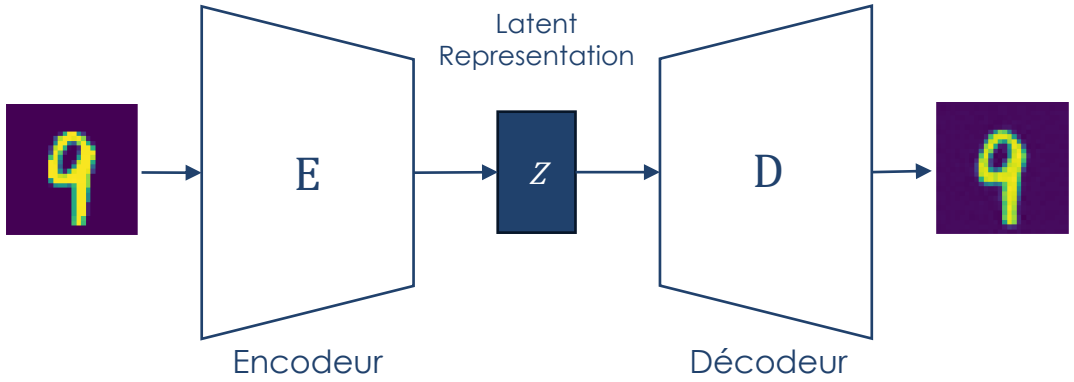
Apprendre une représentation du trafic adaptée aux modèles de génération

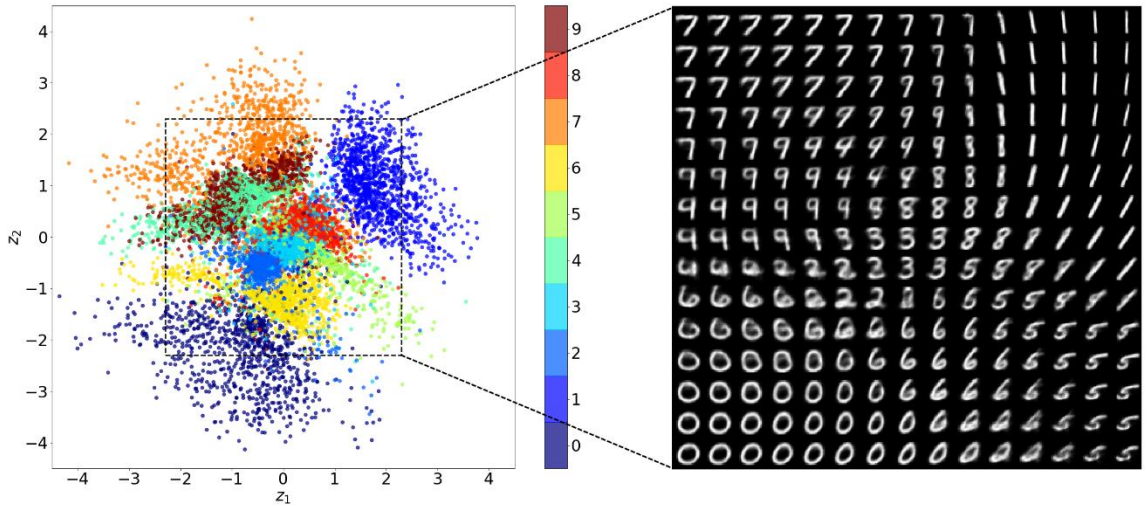
Representation Learning

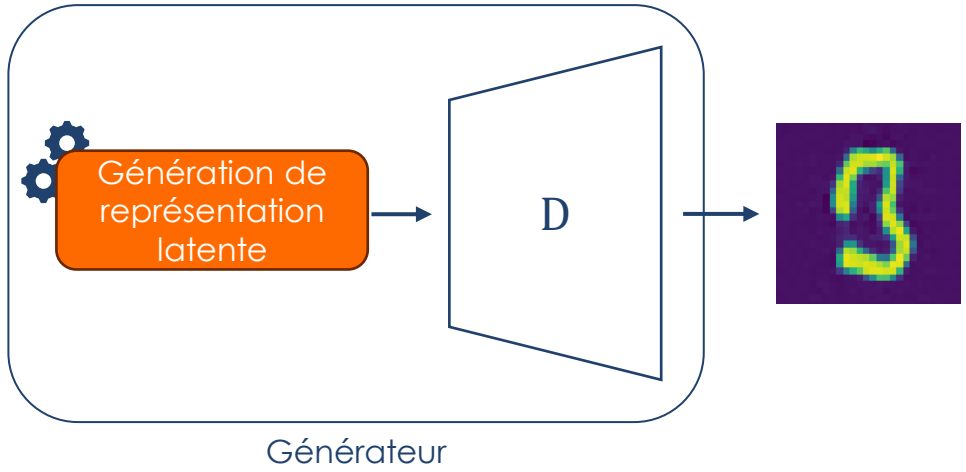
- **Séquences multivariées**
  - **Types** de données **différents**
  - **Non utilisable** directement par un modèle de type **GAN** par exemple
  - **Projection** de la donnée dans un **espace « utilisable » par des modèles génératifs** à l'aide d'un ***Autoencoder***.



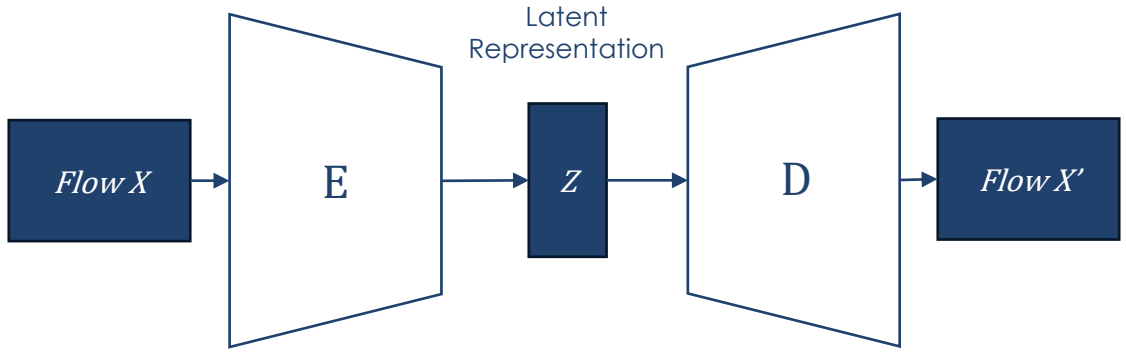
Un mot sur les *Autoencoders*



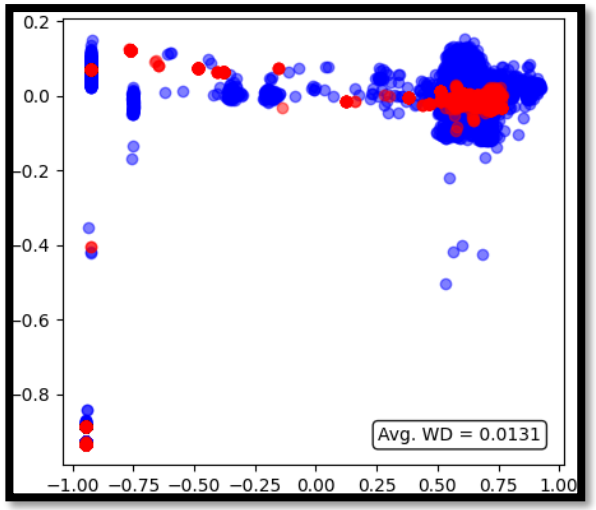




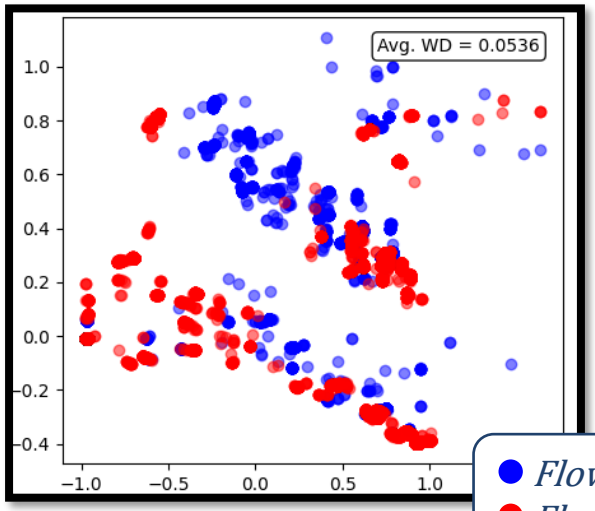
FlowRAE – Modèle seq2vec



IAT



Payload

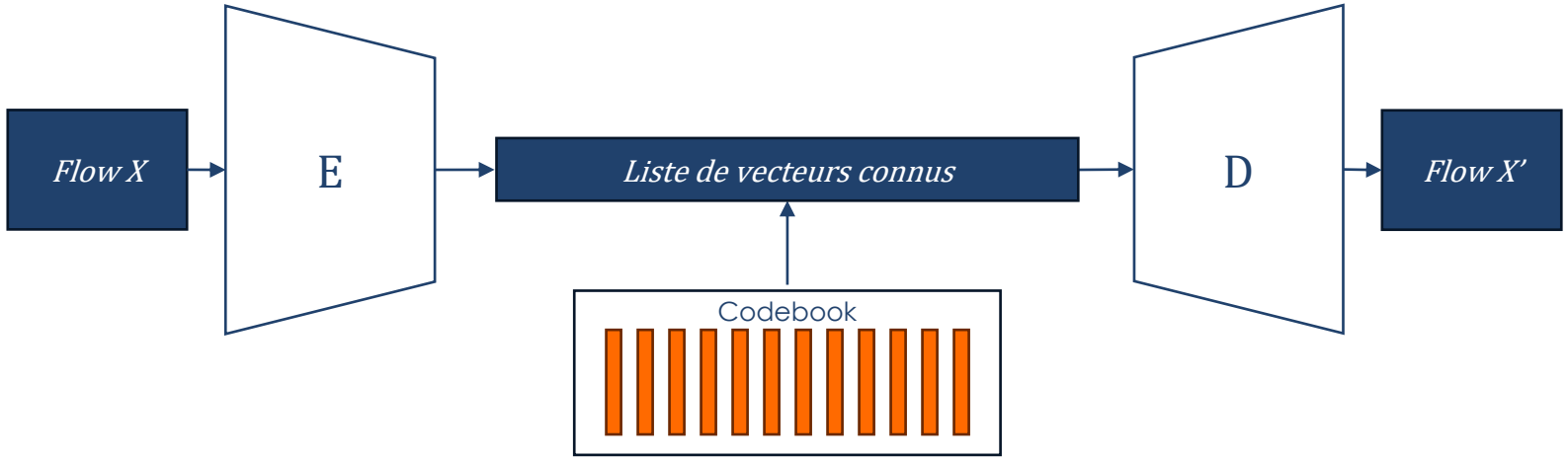


Erreurs de direction : **0,53%**
Erreurs de flags : **0,47%**

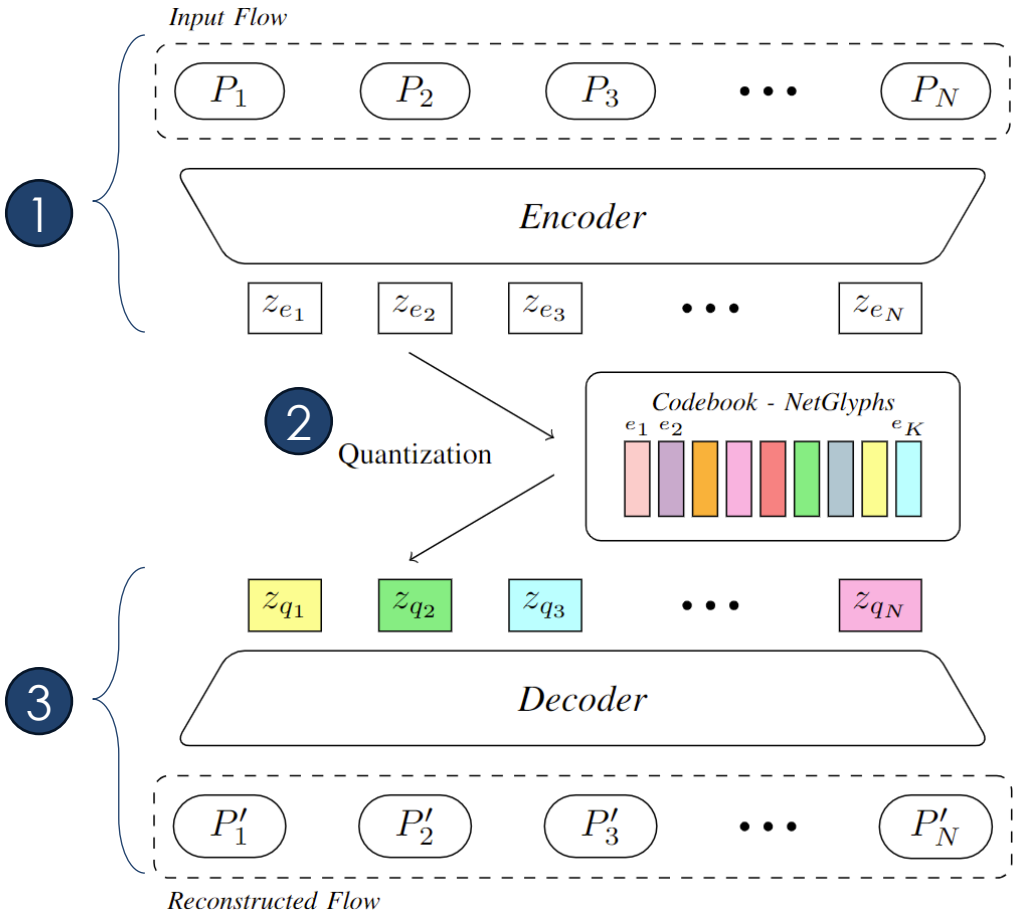
● *Flows X* – Original
● *Flows X'* – Reconstruction

Discrétisation de l'espace latent

- **Apprendre un représentation latente discrète**
  - **Donnée réseau** composée majoritairement de caractéristiques **discrètes** ou **catégorielles**
  - Représentation d'un flux sous forme d'une **séquence de vecteurs discrets** extraits d'un ***Codebook***
  - Forcer le modèle à utiliser la représentation latente

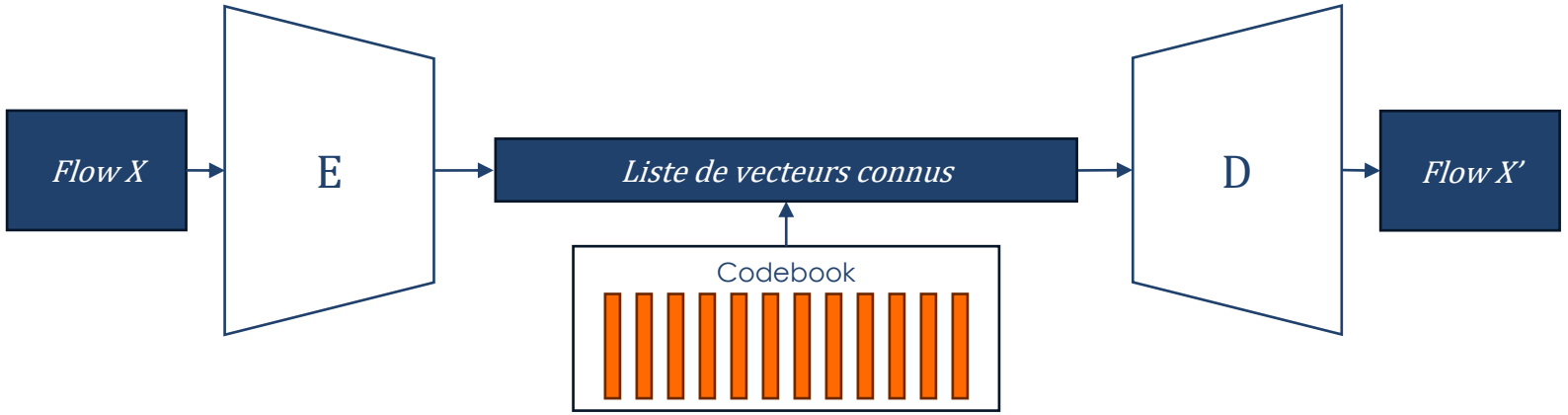


FlowVQ-RAE – Espace latent discret

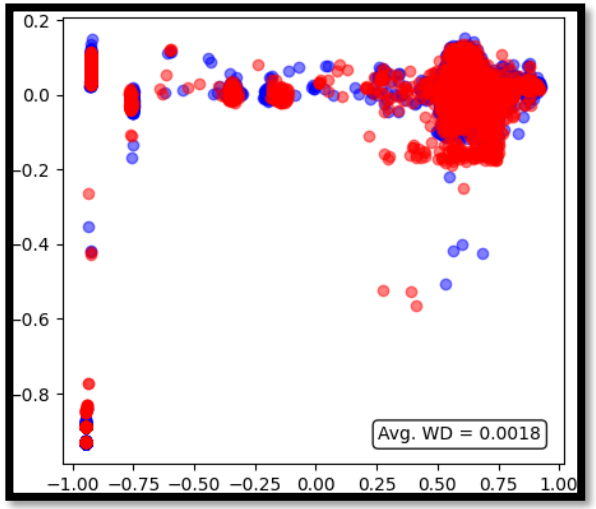


1. **Encoder** : **Flux encodé** en une **séquence Z_e**
2. **Quantization** : **Remplacement** des **vecteurs** de la séquence par les **plus proches vecteurs** du ***Codebook***.
3. **Decoder** : **Flux décodé** à partir de cette **séquence discrète** de « **tokens** » en une **séquence de paquets.**

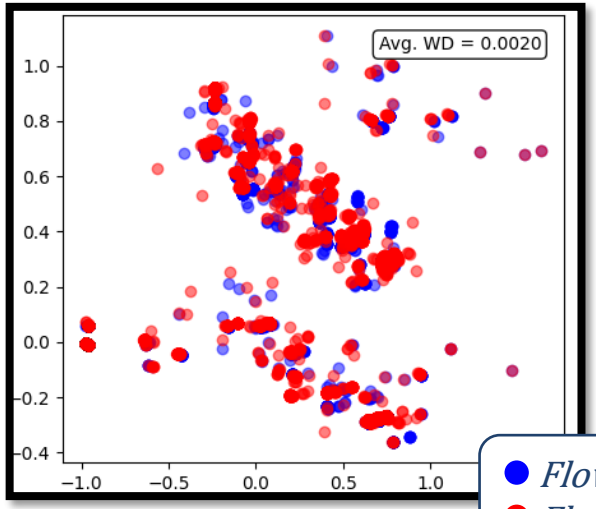
FlowVQ-RAE – Vector Quantization



IAT



Payload



Erreurs de direction : **0%**
Erreurs de flags : **0%**

● *Flows X* – Original
● *Flows X'* – Reconstruction

Génération de trafic réseau

Modèles génératifs

Choix d'un modèle génératif adapté

FlowRAE

- Modèle **seq2vec**
- Représentation **continue**
- **Génération** de **nouveau point** de l'espace latent

GAN

FlowVQ-RAE

- **Discrétisation** d'un flux
- Séquence d'**indices** de vecteur du *Codebook*
- Génération de séquences par **prédiction** du **prochain token**

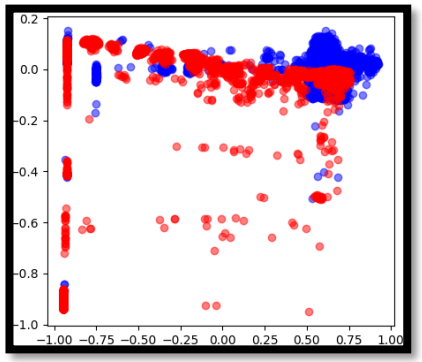
Transformers

Résultats de génération

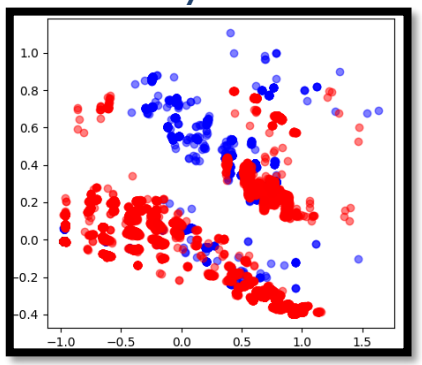
- Original
- Génération

GAN

IAT



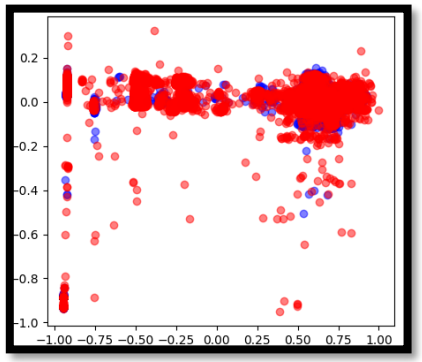
Payload



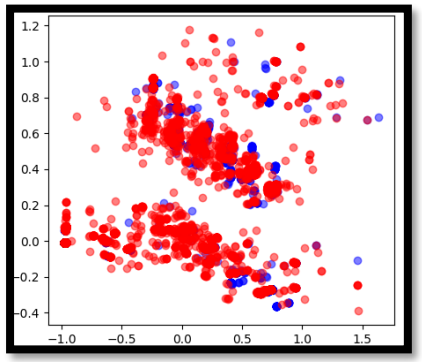
3-way Handshake : **99,99%**
Paquet (PSH + Payload) : **88,6%**

Transformers

IAT



Payload



3-way Handshake : **99,99%**
Paquet (PSH + Payload) : **98,5%**

Conclusion

- Le ***FlowVQ-RAE*** couplé avec un **Transformer** est une **architecture prometteuse** pour la génération de trafic.

- Validation de cette approche avec un **dataset** plus **important** et plus **varié**
  - Contenant **plusieurs classes de trafic**
  - **Génération conditionnée** et identification des classes de trafic
  - Évaluation protocolaire plus approfondie

- **Premier papier scientifique**

MERCI