PICARD@OMD
9 déc. 2021

ibS 데이터사이언스그룹
Data Science Group

Partage de connaissances via l'adaptation du domaine dans la détection de la fraude douanière

Meeyoung Cha, Sundong Kim

Data Science Group, Institute for Basic Science

https://ds.ibs.re.kr/

Programme

Modèle de détection de la fraude basé sur l’apprentissage profond

- DATE : « Dual Attentive Tree-aware Embedding » pour la détection de la fraude douanière (KDD’20)

Concept de détection collaborative de la fraude

- Partage de connaissances via l’adaptation du domaine pour la détection de la fraude douanière
  (en cours d’évaluation)

Détection de la fraude douanière

« Améliorer l'efficacité du processus de dédouanement en fournissant une liste d'articles à inspecter en priorité »

Type de fraude	Motivations illicites
Sous-évaluation des marchandises	Éviter la perception des droits de douane ad valorem ou dissimuler des flux financiers illicites des exportateurs
Fraude au classement dans le SH	Obtenir l'application d'un taux de droit réduit ou échanger des marchandises prohibées en contournant les restrictions
Manipulation du pays d'origine	Obtenir un taux de tarif préférentiel dans le cadre d'un accord de libre-échange

Approches antérieures
FNN & CNN (KCS+KISTI, 2018)
Ensembles SVM (Belgique, 2020)
DATE (2020)
Dérive conceptuelle (2021)
Adaptation de domaine (2021)

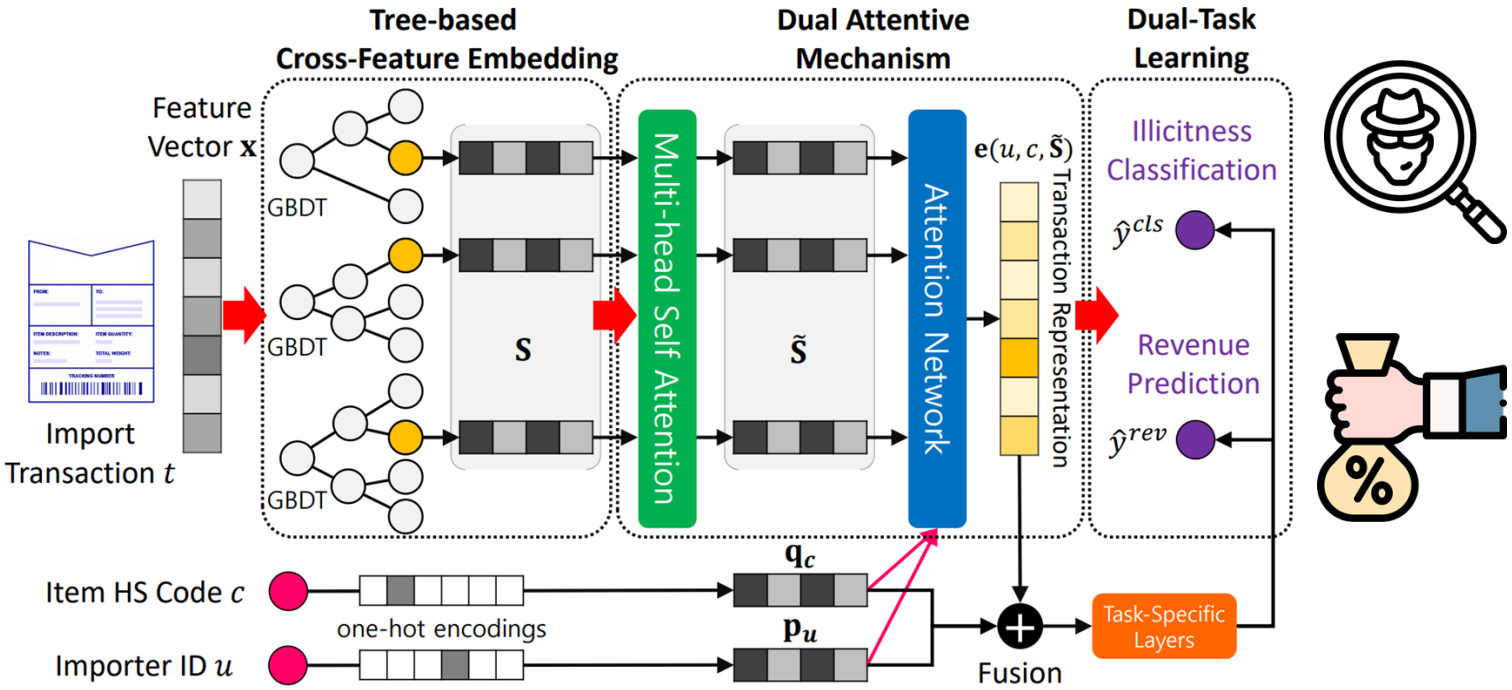
Flux de travaux – Processus de dédouanement

« Comment concevoir un modèle de détection de la fraude durable ? »

Illustration du processus de dédouanement

Modèle DATE

« Modèle d'attention avancé en arborescence avec apprentissage double tâche »



Kim et coll. SIGKDD 2020, http://bit.ly/kdd20-date

Résultats d'évaluation – DATE

- Données utilisées : échanges d'un pays à l'importation
- Formation : 2013-2016
- Tests : 2017
- Taux moyen de flux illicites : 2,2 % (2017)

Indicateurs d'évaluation

Précision, mémoire Recettes

Model	n = 1% (Selecting top 1%)			n = 2%		
	Pre.	Rec.	Rev.	Pre.	Rec.	Rev.
Price	2.75%	1.23%	15.17%	2.23%	1.99%	20.64%
Importer	11.43%	5.10%	4.36%	9.41%	8.39%	7.56%
IForest	5.61%	2.50%	14.30%	6.19%	5.52%	23.14%
GBDT	90.01%	40.15%	24.59%	66.16%	59.04%	38.89%
GBDT+LR	90.95%	40.40%	27.18%	72.94%	65.09%	44.22%
TEM	88.72%	39.59%	39.48%	74.70%	66.43%	58.48%
DATE$_{CLS}$	**92.66%**	**41.33%**	44.97%	**80.79%**	**72.05%**	67.14%
DATE$_{REV}$	82.25%	36.63%	**49.29%**	79.93%	71.22%	**68.48%**

Expériences/tests
Effets sur la durée de la formation
Performance sur les sous-groupes testés
Robustesse sur les entrées corrompues

Disponibilité du code

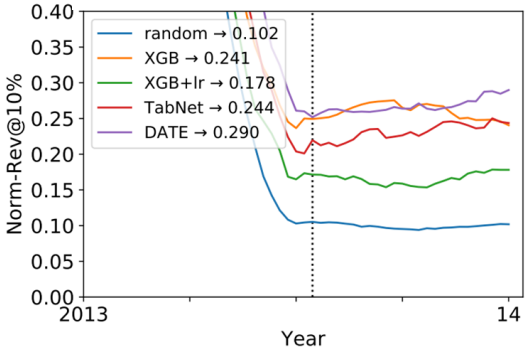
Codes disponibles à l'adresse :
https://bit.ly/customs-fraud-detection

Paquet logiciel pour simuler le système de dédouanement en ligne avec des bases de données propriétaires :

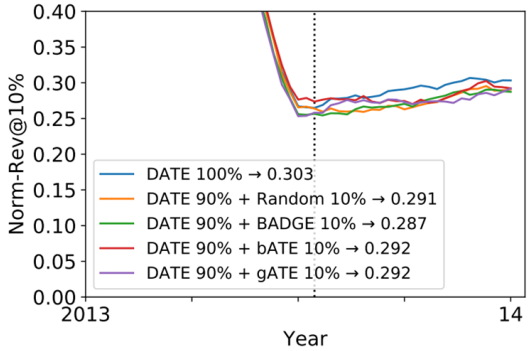
- La version actuelle prend en charge la simulation de sélection douanière avec diverses stratégies
- Les codes sont compatibles avec les déclarations d'importation synthétiques incluses dans le référentiel
- Les données synthétiques fournies ont leurs propres modèles de fraude

```
(py37) $ export CUDA_VISIBLE_DEVICES=3 && python
↪  main.py --data synthetic --semi_supervised 0
↪  --batch_size 512 --sampling hybrid
↪  --subsamplings DATE/bATE --weights 0.9/0.1
↪  --mode scratch --train_from 20130101
↪  --test_from 20130201 --test_length 7
↪  --valid_length 28 --initial_inspection_rate
↪  100 --final_inspection_rate 10 --epoch 10
↪  --closs bce --rloss full --save 0 --numweeks
↪  100 --inspection_plan fast_linear_decay
```

Explorations pures des données synthétiques



Résultats hybrides sur les données synthétiques



Programme

Modèle de détection de la fraude basé sur l’apprentissage profond

- DATE : « Dual Attentive Tree-aware Embedding » pour la détection de la fraude douanière (KDD’20)

Concept de détection collaborative de la fraude

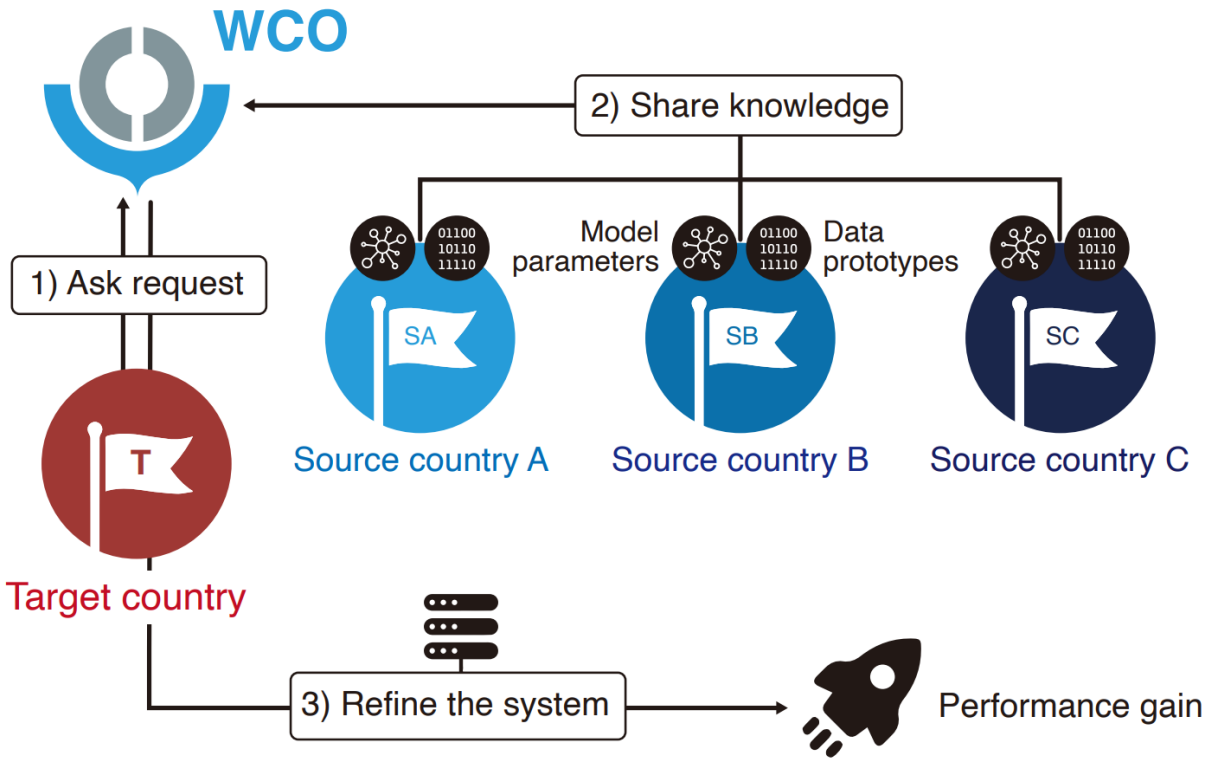
- Partage de connaissances via l’adaptation du domaine pour la détection de la fraude douanière

Motivation pour la détection collaborative de la fraude

- Les pays qui manquent de données fiables rencontrent des difficultés pour repérer les échanges potentiellement illicites.
- Les pays Membres cherchent à se soutenir mutuellement et à bâtir un consortium.
- Cependant, d’après la loi, les données originales ne peuvent pas être mises à disposition à l’extérieur des pays.
- Comment créer une synergie entre les pays et quelle serait une solution réaliste ?

Démonstration de faisabilité

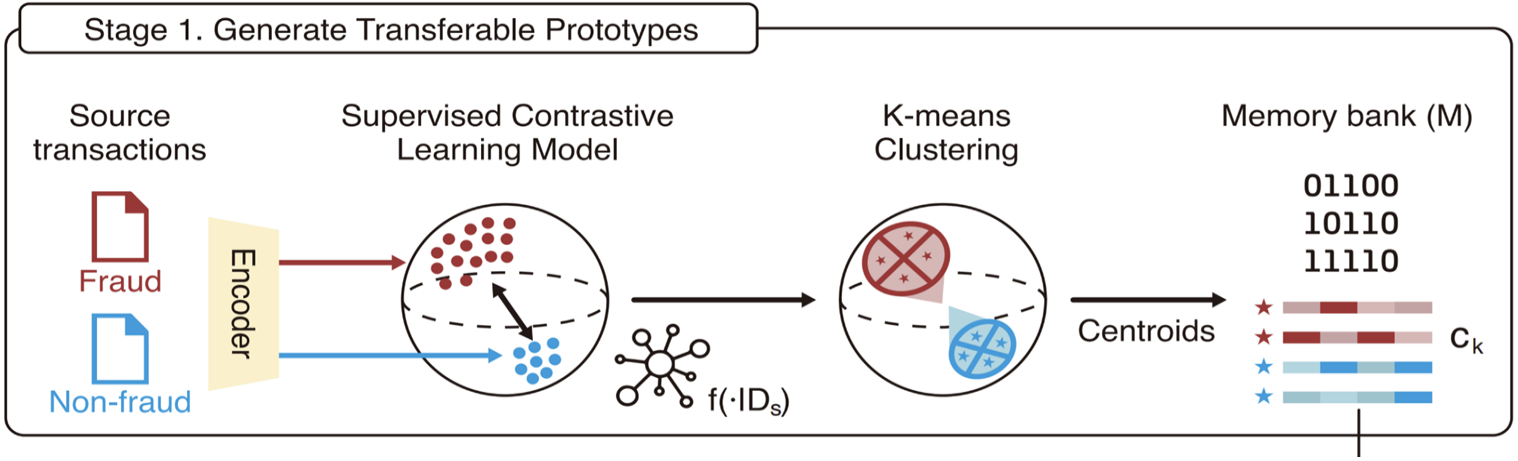
« Faciliter le partage des connaissances entre les Membres »



Cadre de partage des connaissances

Étape 1 : Générer des prototypes transférables (= fraudes représentatives, non identifiables)

- Préparation avec perte contrastive (= trouver l'essence des fraudes et de l'absence de fraude)
  - Maximiser les différences entre les transactions frauduleuses et non frauduleuses
- Compresser les connaissances par agrégation (= partageables et compactes, jusqu'à 1 000 prototypes)

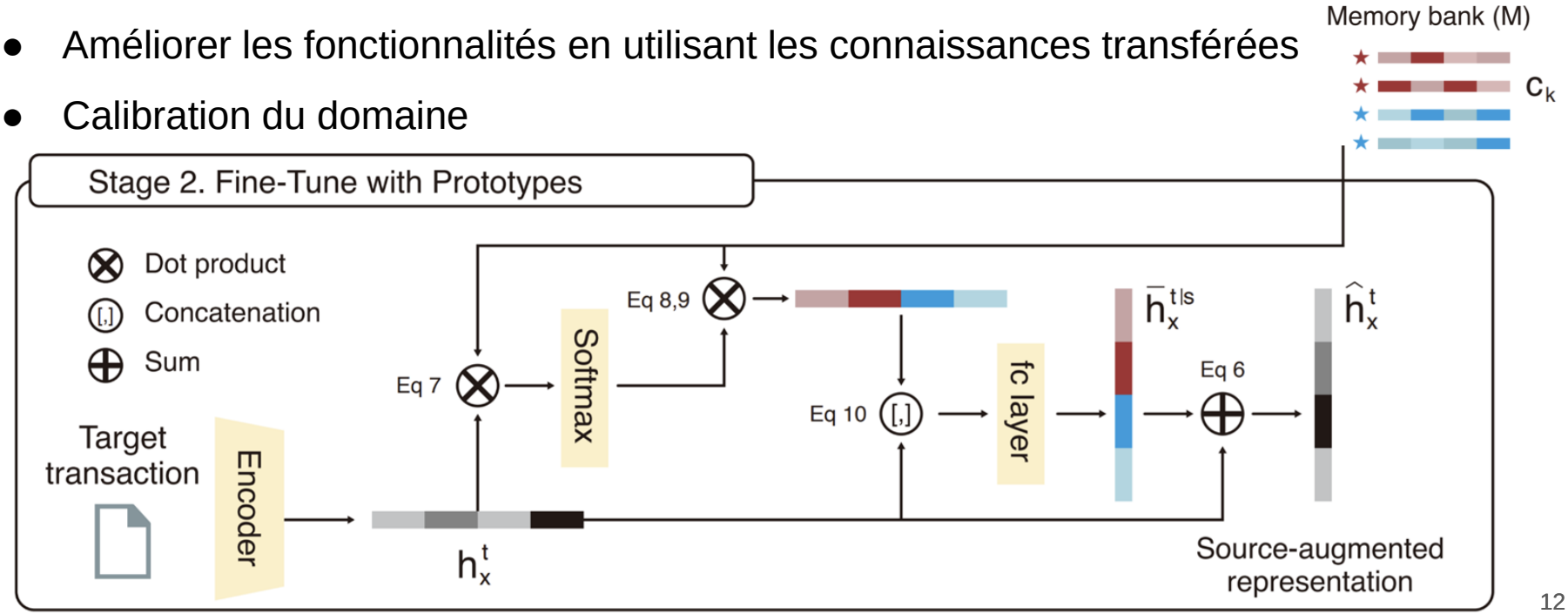


Cadre de partage des connaissances

Étape 2 : Affiner les prototypes

(= affiner l'ensemble de données en utilisant les prototypes du pays source)

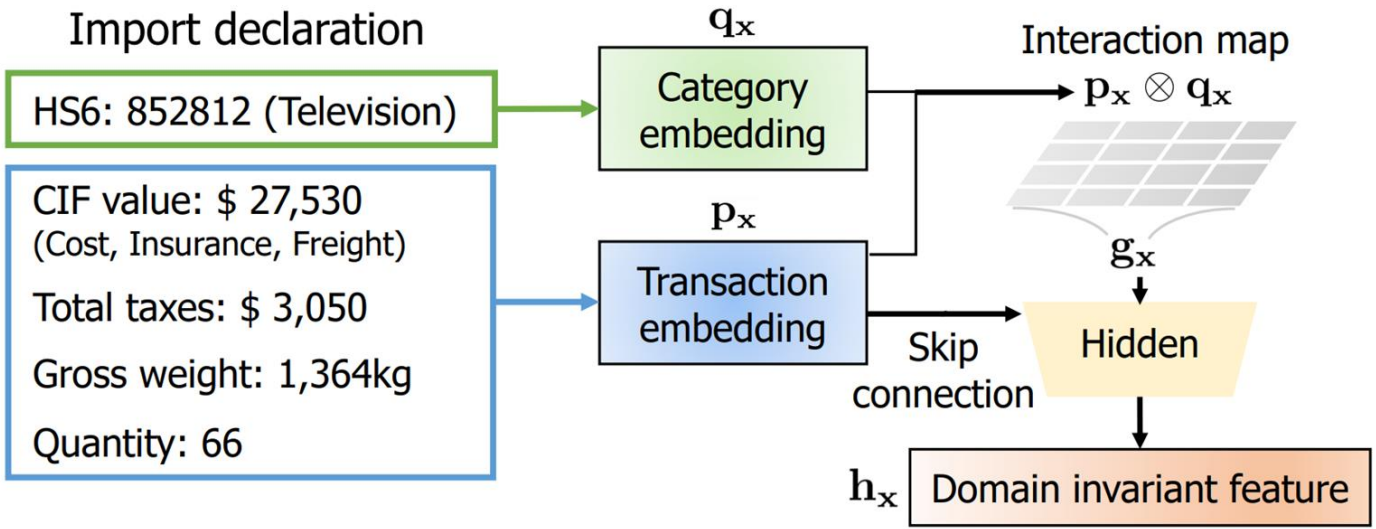
- Améliorer les fonctionnalités en utilisant les connaissances transférées
- Calibration du domaine



Module de base – Structure d'encodage

Encodeur sans modification du domaine pour extraire les caractéristiques

(= convertir les informations sur les marchandises déclarées dans un format identifiable par une machine, après avoir supprimé les informations liées au pays)



Principaux résultats

« La performance a augmenté de manière significative avec l'adaptation de domaine

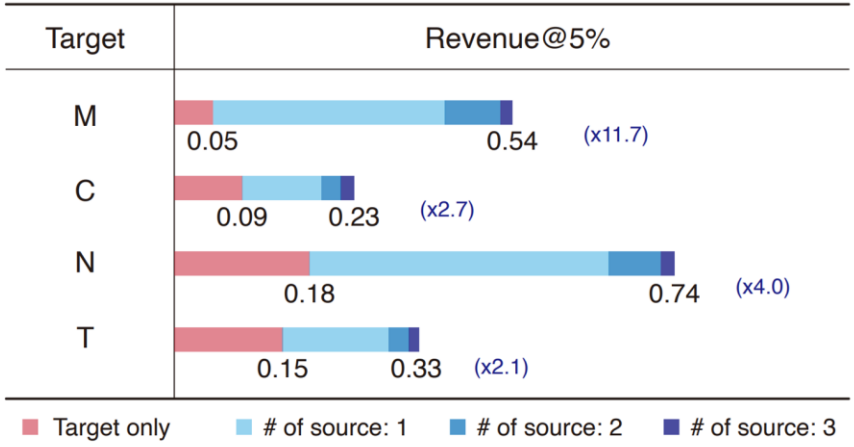


Figure 4: Collected duties are expected to increase 2.1–11.7 times for all tested countries when shared knowledge contributed by multiple sources is used (i.e., blue bars) compared to relying on local knowledge alone (i.e., red bars).

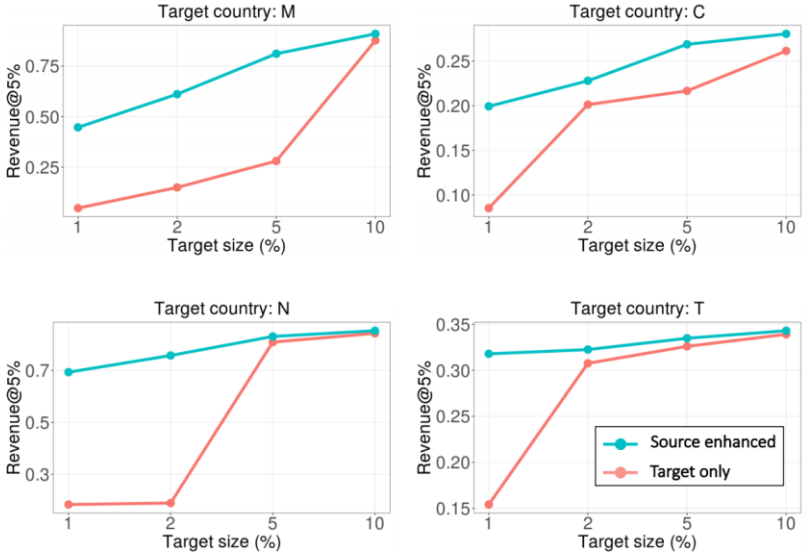


Figure 5: Performance improvement as the log size of the target country increases. The shared knowledge brings the most considerable benefit when the available log size is the smallest (i.e., 1%) and for countries with the weaker economy (i.e., country M).

Analyse qualitative

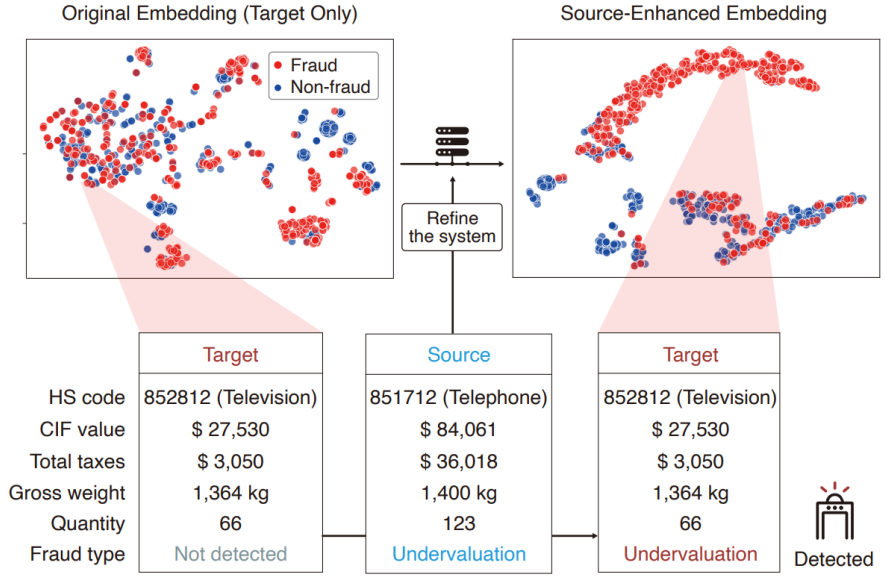


Figure 6: t-SNE plots of the learned embeddings (T $\rightarrow$ N), when the model is trained only with target-only feature (left) and with source enhanced feature using DAS (right). Fraud cases are successfully detected after receiving prototypes from the source country. Similar fraud examples from the source country help flag this case.

Résumé

- Modèle de détection de la fraude basé sur l’apprentissage profond
  - Modèle d’apprentissage double tâche pour atteindre les deux objectifs
  - Paquet logiciel pour simuler le système de dédouanement en ligne
- Concept de détection collaborative de la fraude
  - Partage de connaissances via l’adaptation du domaine
  - Fournir des connaissances pertinentes aux participants