PICARD@WCO
Dec 9, 2021

데이터사이언스그룹
Data Science Group

Knowledge Sharing via Domain Adaptation in Customs Fraud Detection

Meeyoung Cha, Sundong Kim

Data Science Group, Institute for Basic Science

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

Agenda

Deep Learning-based Fraud Detection Model

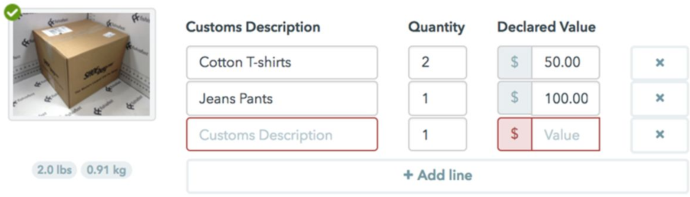
- DATE: Dual Attentive Tree-aware Embedding for Customs Fraud Detection (KDD'20)

Collaborative Fraud Detection Concept

- Knowledge sharing via domain adaptation for Customs Fraud Detection (Under review)

Customs Fraud Detection

"Improve the efficacy of customs clearance process by providing a prioritized list of items to inspect."



Type of frauds	Illicit motives
Undervaluation of trade goods	To avoid ad-valorem customs duty, or conceal illicit financial flows from exporters
Misclassification of HS code	To get a lower tariff rate applied or trade prohibited goods by avoiding restriction
Manipulation of origin country	To get a preferential tariff rate under a free trade agreement

Previous Appraches
FNN & CNN (KCS+KISTI, 2018)
SVM Ensembles (Belgium, 2020)
DATE (2020)
Concept Drift (2021)
Domain Adaptation (2021)

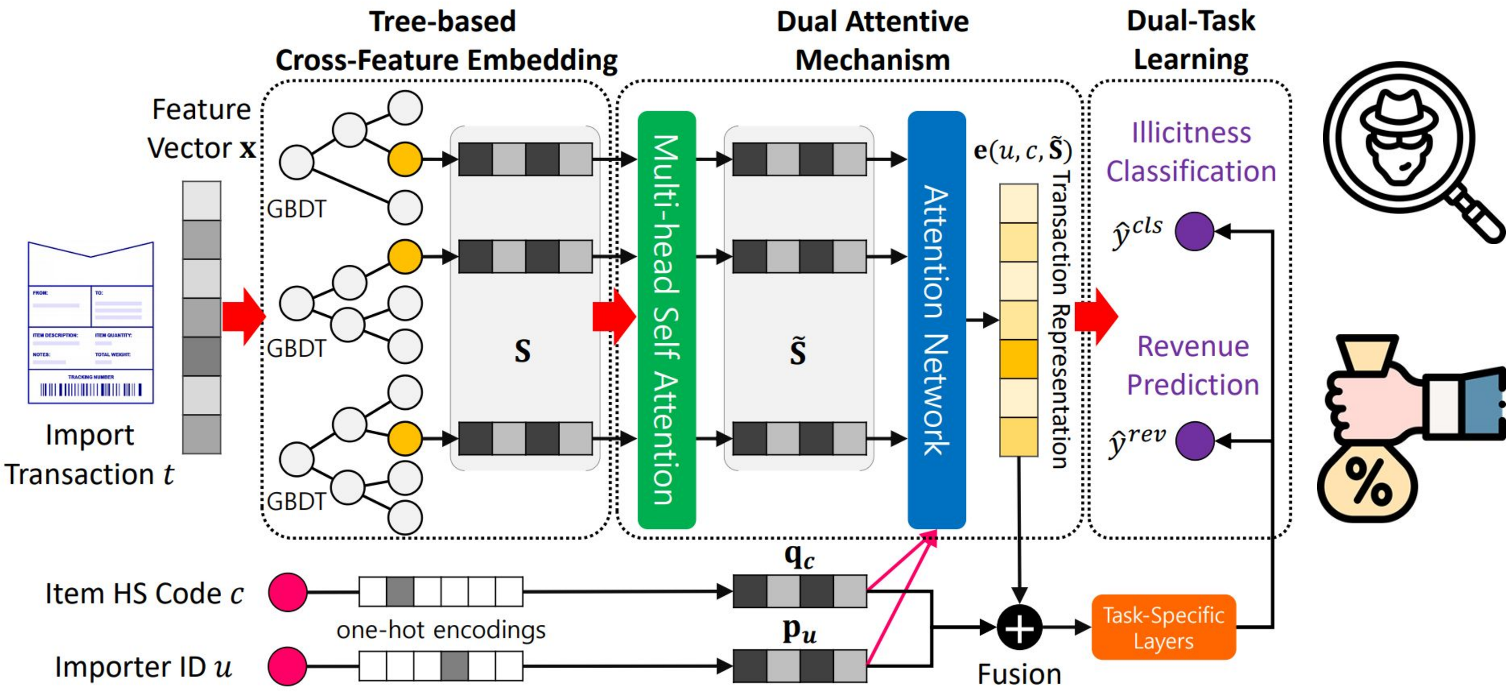
Workflow - Customs Clearance Process

"How to build a sustainable fraud detection model?"

Illustration of the customs clearance process

DATE Model

"Tree-enhanced attention model with dual-task learning"



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

Evalution Results - DATE

- Used data: Import trades of a country
- Training: Y2013 – Y2016
- Testing: Y2017
- Average illicit rate: 2.2% (Y2017)

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%**

Evaluation Metrics

Precision, Recall

Revenue

Supporting Experiments
Effects on training length
Performance on test subgroups
Robustness on corrupted inputs

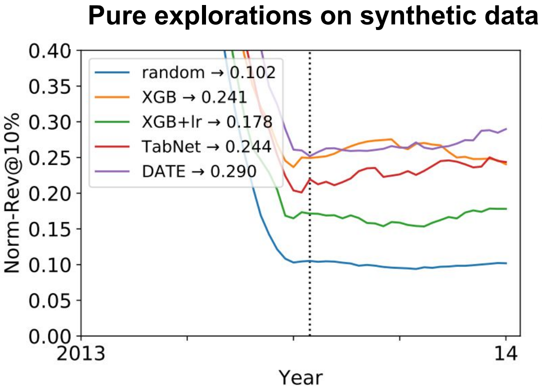
Code Availability

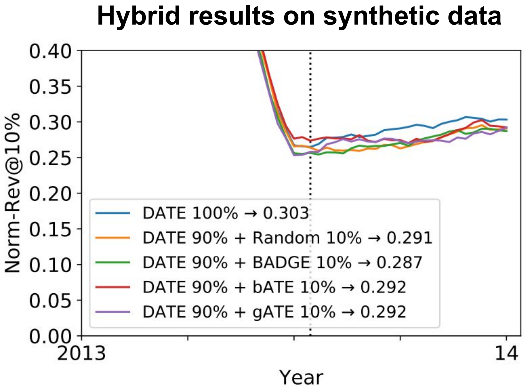
Codes are released at:
https://bit.ly/customs-fraud-detection

Software package for simulating e-clearance system with own datasets:

- Current version supports customs selection simulation with diverse strategies
- The codes run compatible with synthetic import declarations included in the repo
- Provided synthetic data has its fraudulent patterns

```
(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
```





Agenda

Deep Learning-based Fraud Detection Model

- DATE: Dual Attentive Tree-aware Embedding for Customs Fraud Detection (KDD'20)

Collaborative Fraud Detection Concept

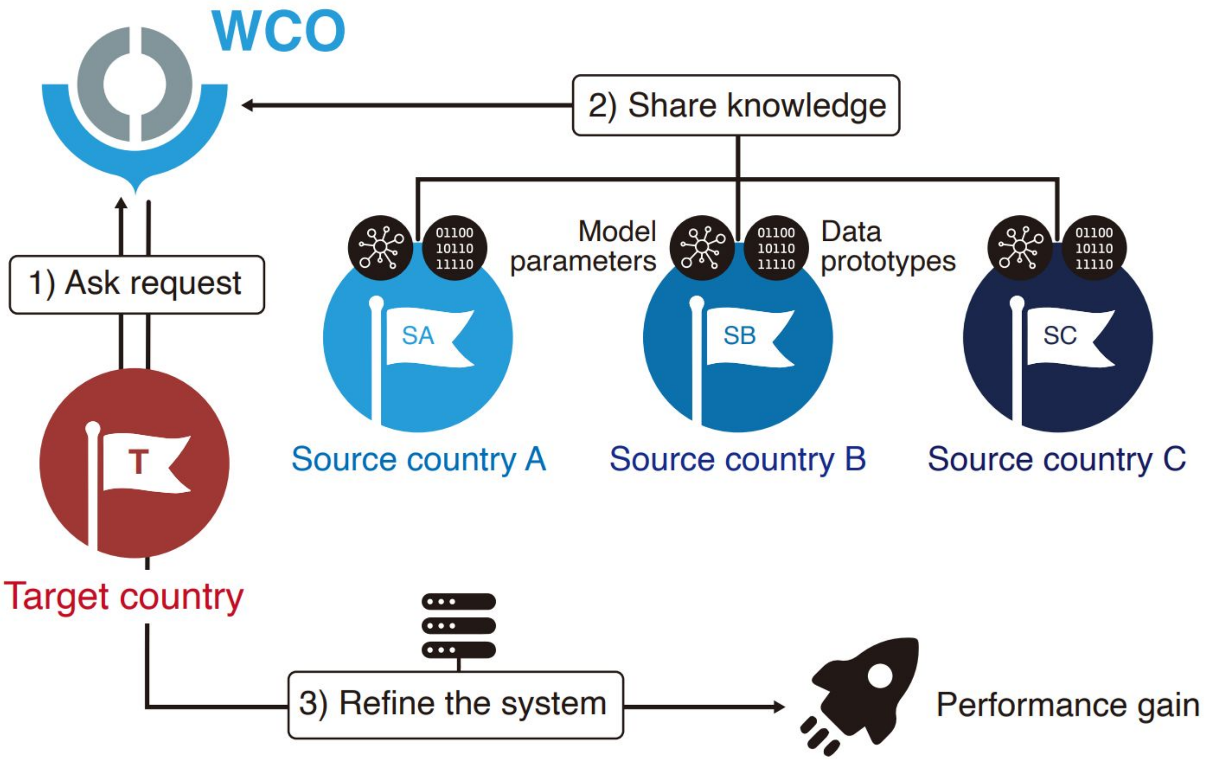
- Knowledge sharing via domain adaptation for Customs Fraud Detection

Motivation to Collaborative Fraud Detection

- Countries with a lack of reliable data struggles with potentially illicit trades.
- Member countries are willing to support each other and build a consortium.
- However, by the law, the original data cannot be accessible outside the countries.
- How to create a synergy between countries, and what can be a feasible solution?

Proof of Concept

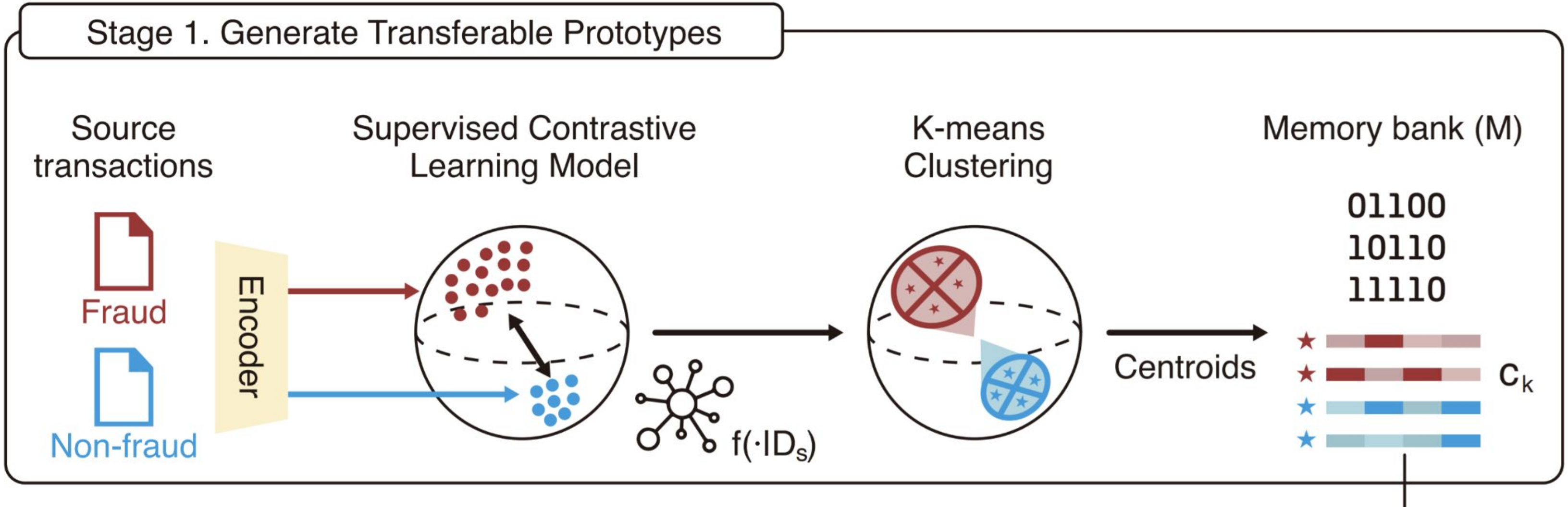
"Facilitate knowledge sharing across members."



Knowledge Sharing Framework

Stage 1: Generate transferable prototypes (= Representative frauds, unidentifiable)

- Pretrain with contrastive loss (= Find the essence of frauds & non-frauds)
  - Maximize the discrepancy between fraud and non-fraud transactions
- Compress knowledge by clustering (= Sharable & compact, up to 1,000 prototypes)

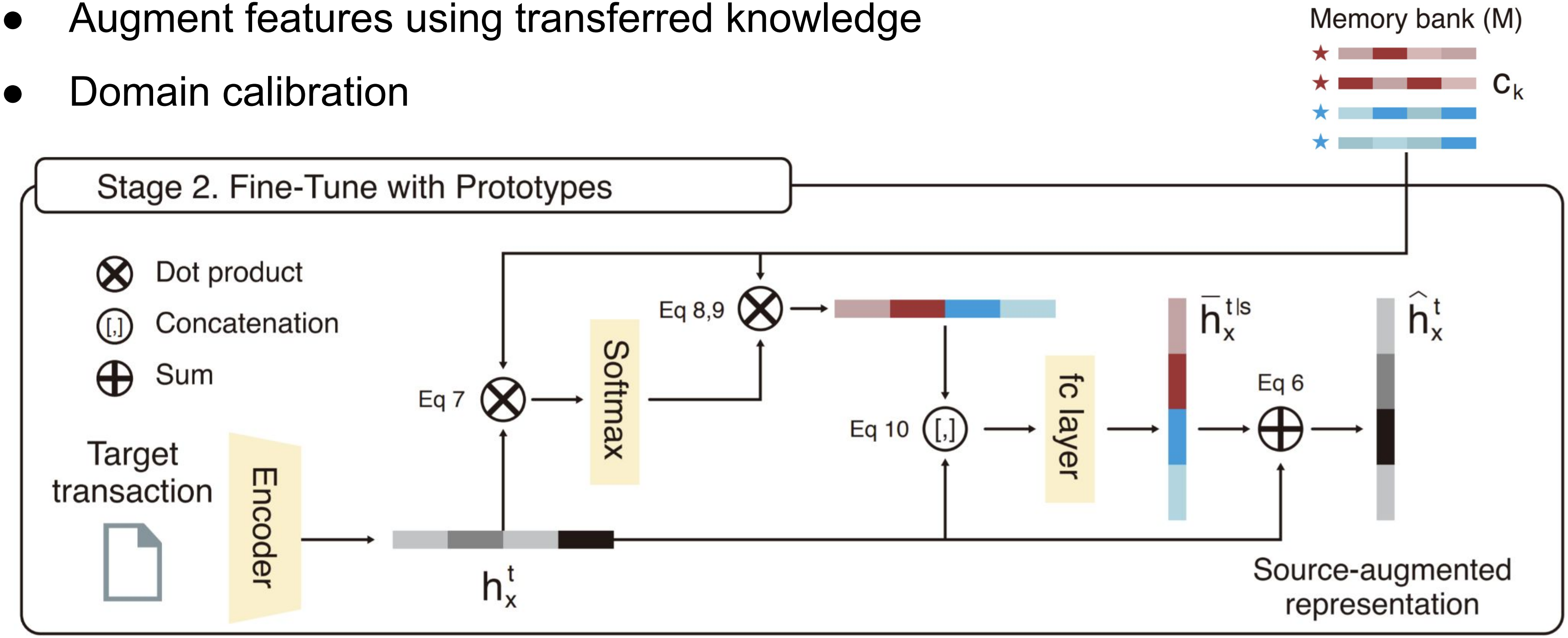


Knowledge Sharing Framework

Stage 2: Fine-tune with prototypes

(= Refine the dataset by using prototypes from the source country)

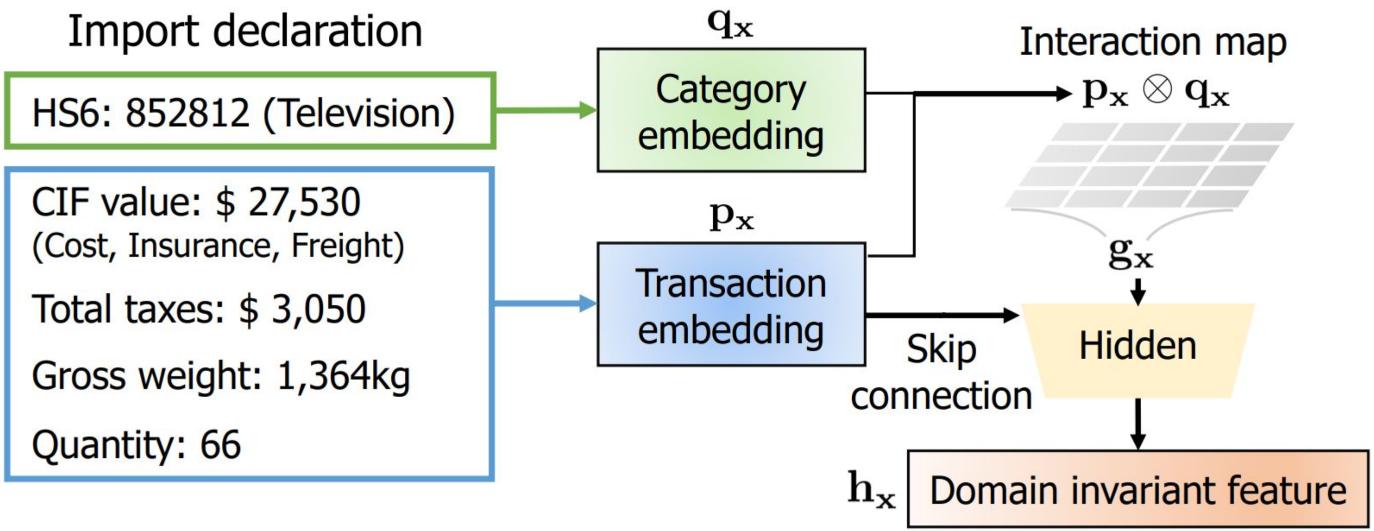
- Augment features using transferred knowledge
- Domain calibration



Base Module - Encoder Structure

Domain-invariant encoder to extract features

(= Transform the information of declared good into a machine-identifiable format, after removing country-specific information)



Main Results

"Performance increased significantly with domain adaptation"

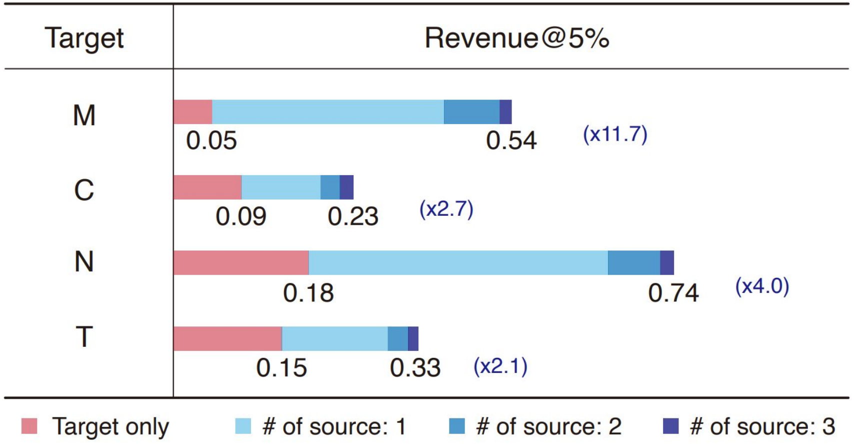


Figure 4: Collected duties are expected to increase 2.1–11.7 times for all tested countries when shared knowledge contributed by multiple sources is used (i.e., blue bars) compared to relying on local knowledge alone (i.e., red bars).

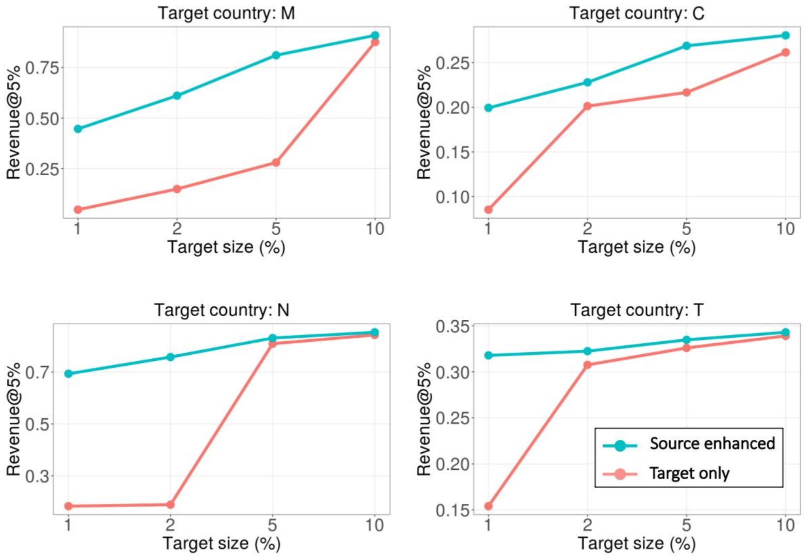


Figure 5: Performance improvement as the log size of the target country increases. The shared knowledge brings the most considerable benefit when the available log size is the smallest (i.e., 1%) and for countries with the weaker economy (i.e., country M).

Qualitative Analysis

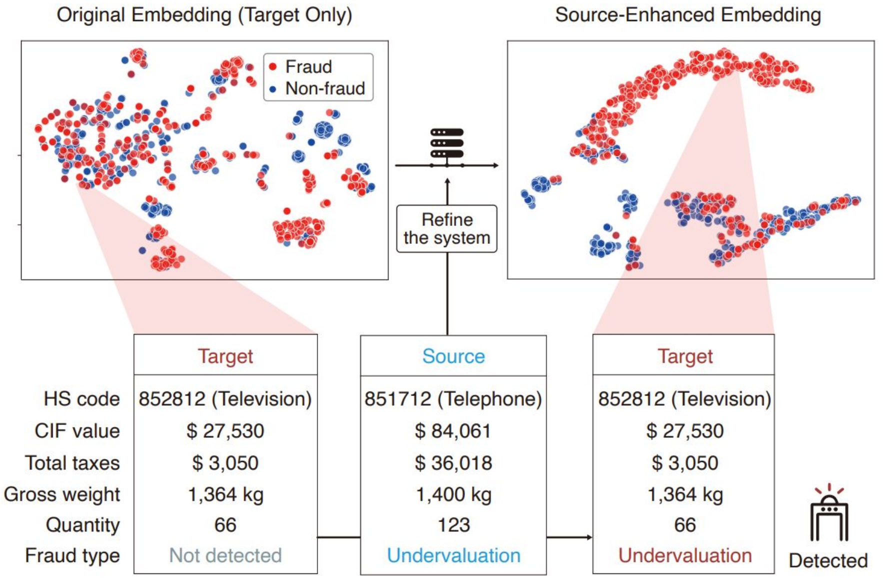


Figure 6: t-SNE plots of the learned embeddings (T → N), when the model is trained only with target-only feature (left) and with source enhanced feature using DAS (right). Fraud cases are successfully detected after receiving prototypes from the source country. Similar fraud examples from the source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

Summary

- Deep Learning-Based Fraud Detection Model
  - Dual-task learning model to meet both objectives
  - Software package for simulating e-clearance system
- Collaborative Fraud Detection Concept
  - Knowledge sharing via domain adaptation
  - Provide suitable knowledge to participants