Knowledge Sharing via Domain Adaptation in Customs Fraud Detection

Sungwon Park[1], **Sundong Kim**[2*], **Meeyoung Cha**[2,1*]

[1]School of Computing, Korea Advanced Institute of Science and Technology
[2]Data Science Group, Institute for Basic Science
psw0416@kaist.ac.kr, sundong@ibs.re.kr, mcha@ibs.re.kr

Abstract

Knowledge of the changing traffic is critical in risk management. Customs offices worldwide have traditionally relied on local resources to accumulate knowledge and detect tax fraud. This naturally poses countries with weak infrastructure to become tax havens of potentially illicit trades. The current paper proposes DAS, a memory bank platform to facilitate knowledge sharing across multi-national customs administrations to support each other. We propose a domain adaptation method to share transferable knowledge of frauds as prototypes while safeguarding the local trade information. Data encompassing over 8 million import declarations have been used to test the feasibility of this new system, which shows that participating countries may benefit up to 2–11 times in fraud detection with the help of shared knowledge. We discuss implications for substantial tax revenue potential and strengthened policy against illicit trades.

Introduction

Customs administrations manage an astronomical amount of trades. Amongst their tasks is the risk management and detection of irregularities and illicit consignments from import declarations. These tasks are critical as import tariffs account for a substantial proportion of the total tax revenue. The detection process has traditionally used rule-based algorithms, which is gradually changing to machine learning algorithms (Mikuriya and Cantens 2021). The World Customs Organization (WCO) has been leading such data initiatives by assisting customs offices in 175 countries with their digital transformation process (Weerth 2009).

Several deep learning algorithms have been proposed and tested at a regional level (Vanhoeyveld, Martens, and Peeters 2020; Kim et al. 2020). However, these advanced models only benefit customs offices that have the capacity to build deep learning models and train them. Many developing and low-income countries do not have such data infrastructure. One method to bootstrap model learning is to utilize shared logs of frauds across custom offices within geographical proximity, as illicit patterns are likely similar in those regions. Sharing knowledge can facilitate data initiative and strengthen policy against illegal trade activities. However, data sharing has not been considered due to privacy concerns, as trade information entails critical industry and business relationships.

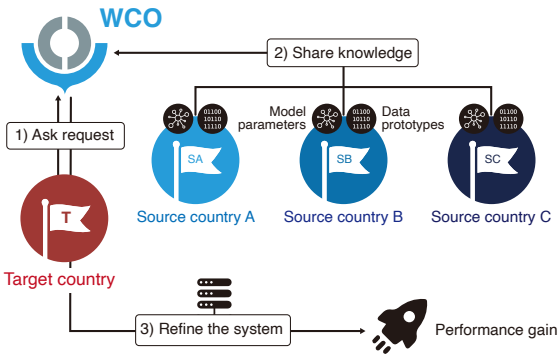

Figure 1: A theoretical model framework describing to facilitate knowledge sharing across multiple countries.

This research proposes first-of-a-kind collaborative customs fraud detection that enables information sharing across regional boundaries. We use techniques in domain adaptation and propose a new data prototyping method to extract transferable knowledge from import declarations. Figure 1 illustrates the workflow. Data donors or source countries share logs of speculative trades, for instance, 1,000 import declarations after removing identifiable information like importer names. Information about the vast majority of normal trades is not shared. Furthermore, instead of the raw data, its embedding is shared in the form of prototypes. At the receiving end, target countries access the accumulated knowledge of fraud patterns from the memory bank and refine their local knowledge via soft attention. Target countries that typically lack infrastructure will likely observe new fraud transactions thanks to the model. This information will loop back to the central memory bank to strengthen the combined knowledge and be shared with all contributing members.

Our memory bank design is called DAS (Domain Adaptation for Sharing Knowledge in Customs), and it uses contrastive learning and clustering to extract meaningful fraud characteristics. Experiments based on multi-year, million-scale import declarations show that all participating countries—including data contributors and recipients—can

*Corresponding authors

benefit from the transferred knowledge with an estimated 2 to 11 times increment in the total tax raised from fraud detection, far extending the capability to manage risk in the studied region.

In practice, DAS can be adopted among countries under bilateral and regional trade agreements to accelerate market opening and pursue higher levels of trade freedom. Globally more countries are starting to share tariff rates, relieve trade barriers, and adopt common policies. For countries initiating trade agreements, implementing this system will facilitate active knowledge sharing and derive tangible efforts on strengthening border security.

Related Work

Customs Fraud Detection Algorithms. Earlier efforts on customs fraud detection utilized rule-based algorithms and random selection algorithms (Hua, Li, and Tao 2006). While some customs offices have adopted machine learning (Filho 2015), many offices in developing countries still report their reliance to rule systems and expert knowledge (Goldberg et al. 2009). Recent studies applied off-the-shelf algorithms, including the ensembled SVM in customs fraud detection (Vanhoeyveld, Martens, and Peeters 2020). State-of-the-art models, for example, the Dual Attentive Tree-aware Embedding (DATE) model, employ gradient boosting and attentions to generate transaction-level embeddings and provide interpretable decisions (Kim et al. 2020). Some newer models utilize concept drift to better represent the changing trade patterns over time (Kim et al. 2022; Mai et al. 2021).

Domain Adaptation Techniques. Domain adaptation aims to learn universal representations that are domain invariant. Representative techniques include latent distribution alignment between the source and target domains (Tzeng et al. 2017; Hoffman et al. 2018; Long et al. 2017). Contrastive learning is used to extract discriminative features between classes (Kang et al. 2019; Thota and Leontidis 2021), and the memory module is used to augment target features using incremental information (Asghar et al. 2020; Zheng and Yang 2019; Liu et al. 2020). A long-standing problem in domain adaptation is negative transfer, which refers to the abnormal scenarios when the source domain data causes reduced learning performance in the target domain due to a large discrepancy in data distributions (Wang et al. 2019; Zhang et al. 2020). Regularization and adaptive source selection methods have been proposed to mitigate this problem (Liu et al. 2020; Abuduweili et al. 2021). Most domain adaptation techniques assume that the source and the target data can be accessed concurrently, which may not be practical for customs under multi-national administrations.

Method

Problem Statement

A fraud detection system f determines whether an import declaration is fraudulent based on its transaction instance $\mathbf{x}$ that is a vector describing information like the imported product and price. f guides customs officers on which transaction to inspect, and customs officers obtain its fraud label y after manual inspection. We denote the inspected import transaction instance and its corresponding fraud label in the custom transactions dataset as $(\mathbf{x}, y) \in \mathcal{D}$. Our task is to design a fraud detection system $f(\cdot|\mathcal{D}_t)$ for a target country t that has limited data logs of import transactions $\mathcal{D}_t$. Let the inspected transactions of the source country s be $\mathcal{D}_s$, then we want to pick out transferable embedded information that will be stored at memory bank M. This shared knowledge can be used to improve the fraud detection system at the target country using $f(\cdot|\mathcal{D}_t, M)$. The source country s is assumed to possess richer logs than the target country t such that $|\mathcal{D}_s| \gg |\mathcal{D}_t|$.

System Overview

The proposed system employs multiple strategies to ensure safety in data transferability. First is the sampling of abnormal, fraud-like transactions to construct the source dataset $\mathcal{D}_s$. Note that most of the "normal" import declarations contain critical trading partners and industry information. In contrast, fraud-suspected trades make up a small volume, which the algorithm utilizes. Second, the domain invariant feature of the HS code is enforced, which is a worldwide item notation convention. Additionally, critical information is anonymized or removed, such as country names, importers, declarants, and detailed descriptions of goods. Lastly, we regulate direct data sharing across domains. Here, the knowledge is shared in the form of model parameters and prototypes, representing compact information for a group of semantically similar transactions. These shared prototypes are combined with the local transaction logs to fine-tune the fraud detection model at the target country. Figure 2 depicts the pipeline of the proposed system in two stages:

- **Stage 1** displays the method through which the source country s shares compressed knowledge to the memory bank M. s will pretrain a network $f(\cdot|\mathcal{D}_s)$ using contrastive learning to extract discriminative features from its fraud-suspected logs. The prototype set $\mathcal{C}_s$ is the resulting transferable knowledge that is stored at M.
- **Stage 2** describes the method through which the target country t refines its detection model with the transferred knowledge. Given a pretrained network $f(\cdot|\mathcal{D}_s)$ and the memory bank M in Stage 1, the data representation at t is augmented using compressed knowledge $\mathcal{C}_s$ in M to fine-tune the network $f(\cdot|\mathcal{D}_t)$.

Domain Invariant Feature using HS code

A transaction encoder is used in both stages to embed import declarations by enforcing domain invariant information. This encoder augments each transaction log to meet international standards like the HS6 code, i.e., the six-digit category code of goods. Figure 3 describes the design of this new encoder.

Let $\mathbf{x}$ denote input features of a transaction for a single import declaration. Then its embedding of transaction information, $\mathbf{p}_\mathbf{x}$, and the embedding of product category information (i.e., HS code), $\mathbf{q}_\mathbf{x}$, are obtained by training a fraud detection model.

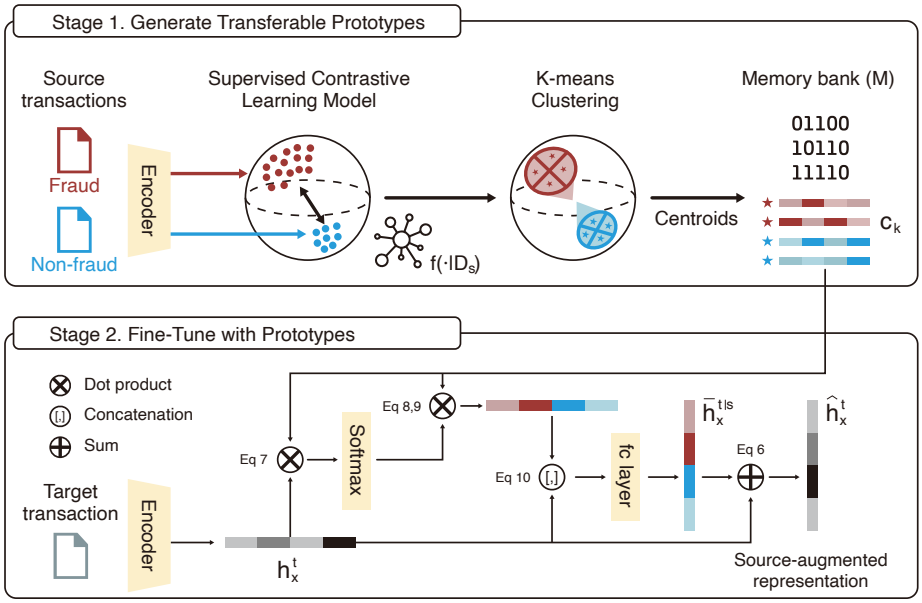

Figure 2: The overall pipeline depicts how the knowledge shared by the source country is added to the transferable memory bank in Stage 1. The subtle difference between the fraud and non-fraud trades is learned from inspecting speculative logs via contrastive learning. Then, the target country can refine its trade representation in Stage 2. The target country will maintain a better fraud detection system with knowledge-enhanced trade representation.

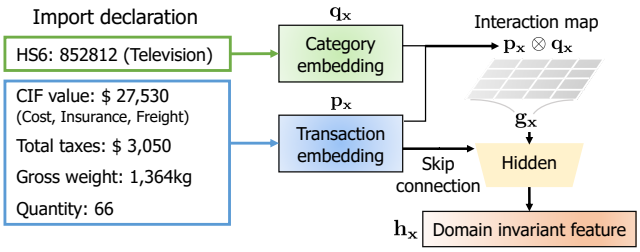

Figure 3: The domain invariant encoder embeds import declarations by increasing the usability across the country.

Any existing detection model can be used. This research uses the state-of-the-art model that is based on a tree-based embedding for interpretable detection (Kim et al. 2020). The generated embedding $\mathbf{p_x}$ may reveal the generic trade characteristics of that region. Hence, we do not use a simple concatenation of these two embeddings but instead augment $\mathbf{p_x}$ to increase its interaction with $\mathbf{q_x}$ by taking their outer product. We use Eq. (1) to model the pairwise correlation between the embedding space of two variables (He et al. 2018):

$$\mathbf{E} = \mathbf{p_x} \otimes \mathbf{q_x} = \mathbf{p_x}\mathbf{q_x}^T, \tag{1}$$

where $\mathbf{p_x}, \mathbf{q_x} \in \mathcal{R}^k$ and $\mathbf{E} \in \mathcal{R}^{k \times k}$. We obtain an interaction vector $\mathbf{g_x}$ by taking a convolutions on $\mathbf{E}$ to learn higher-order relations:

$$\mathbf{g_x} = \text{CNN}(\mathbf{E}). \tag{2}$$

The final representation $\mathbf{h_x}$ of transaction $\mathbf{x}$ is refined via the residual connection (He et al. 2016):

$$\mathbf{h_x} = \text{ReLU}(\phi_a([\mathbf{p_x}, \mathbf{g_x}])), \tag{3}$$

where ϕ_a is a learnable function.

Stage 1. Generate Transferable Prototypes

The encoder projects every transaction $\mathbf{x}$ in $\mathcal{D}_s$ to the domain-invariant space $\mathbf{h_x}$ that is compatible with knowledge transfer. Then the objective of Stage 1 is to learn discriminative features between fraud and non-fraud transactions. We use contrastive learning to pretrain the network $f(\cdot|\mathcal{D}_s)$ and extract prototype set $\mathcal{C}_s$ from the source representation space.

Pretrain with Supervised Contrastive Learning. The use of contrastive learning ensures that positive instances belonging to the same class (i.e., frauds or non-frauds) are pulled together in normalized embedding space while simultaneously pushing apart data instances from different classes (Kang et al. 2019; Han et al. 2020; Chen et al. 2020). We assume source countries own enough inspected logs to leverage the label information effectively. We adopt supervised contrastive learning (SCL) loss to pretrain the network $f(\cdot|\mathcal{D}_s)$ in a fully-supervised manner (Khosla et al. 2020).

$$\mathcal{L}_{SCL} = \sum_{i=1}^{N} -\frac{1}{N_{y_i} - 1} \sum_{j=1}^{N} \mathbf{1}_{i \neq j} \mathbf{1}_{y_i = y_j} \Bigg[\log \frac{\exp(\text{sim}(\mathbf{h_{x_i}}, \mathbf{h_{x_j}})/\tau)}{\sum_{k=1}^{N} \mathbf{1}_{i \neq k} \exp(\text{sim}(\mathbf{h_{x_i}} \cdot \mathbf{h_{x_k}})/\tau)} \Bigg], \tag{4}$$

where N is the batch size, and τ is the sharpening temperature.

Compared to using a simple cross-entropy loss, SCL loss is known to increase the transfer learning performance by maximizing the discrepancy between separate classes (Kang et al. 2019; Khosla et al. 2020). In the experiment, we verify that the pretrained network using SCL loss shows non-trivial performance improvement.

Compress Knowledge by Clustering. Vanilla transfer learning (i.e., passing only model parameters) may suffer from the catastrophic forgetting problem, in which the fine-tuned model tends to "forget" the source dataset's discriminative feature during fine-tuning (Kirkpatrick et al. 2017). To avoid this fallacy, we adopt the memory bank concept (Wu et al. 2018) to enhance the model capacity by storing additional representations from the source dataset $\mathcal{D}_s$ in the memory bank. We compress knowledge as a prototype, a representative embedding for a group of semantically similar instances. This structure also ensures that shared knowledge no longer contains individual trade information. After pretraining, knowledge of frauds and non-frauds is clustered separately and condensed to generate transferrable prototypes. We apply *K*-means clustering to extract the centroid set as prototypes for each class (i.e., $\mathcal{C}_s = \mathcal{C}_{\text{frauds}} \cup \mathcal{C}_{\text{non-frauds}}$). These extracted prototypes $\mathcal{C}_s$ are stored in the memory bank M, allowing knowledge transfer from the source country s. Recipients can utilize any subset of the transferred knowledge for their classification task and detect previously unrecognized illicit trades in $\mathcal{D}_t$.

Multi-Source Memory. The memory bank M can be expanded to adapt a multi-source scenario. For each source country, we pretrain the model independently using $\mathcal{L}_{SCL}$ and extract the prototype sets. The memory bank M can now contain an ensemble of multiple source countries' prototype sets as follows:

$$M = [\mathcal{C}_{s_1}, \mathcal{C}_{s_2}, ..., \mathcal{C}_{s_m}], \tag{5}$$

where m is the number of source countries.

Stage 2. Fine-Tune with Prototypes

Given the memory bank M from s, the target country t can fine-tune its fraud detection system $f(\cdot|\mathcal{D}_t)$ using $\mathcal{D}_t$. We design a fine-tuning step that is inspired by recent memory-based domain adaptation techniques that use enhancers for augmenting data representation (Asghar et al. 2020; Liu et al. 2020). Following this concept, we let each target representation $\mathbf{h}^t_{\mathbf{x}}$ be refined with the representation $\bar{\mathbf{h}}^{t|s}_{\mathbf{x}}$ that is augmented with the source domain knowledge $\mathcal{C}_s$ in the memory M:

$$\hat{\mathbf{h}}^t_{\mathbf{x}} = \mathbf{h}^t_{\mathbf{x}} + \bar{\mathbf{h}}^{t|s}_{\mathbf{x}}. \tag{6}$$

The process of deriving $\bar{\mathbf{h}}^{t|s}_{\mathbf{x}}$ is described in Eq. 7–10.

Source-Augmented Feature. Source-augmented feature $\bar{\mathbf{h}}^{t|s}_{\mathbf{x}}$ is computed via soft-attention toward the set of selected prototypes in $\mathcal{C}_s$ (Vaswani et al. 2017). Let k-th prototype be denoted $\mathbf{c}_k$. Each target transaction $\mathbf{h}^t_{\mathbf{x}}$ attends to prototype features via attention weights computed by dot product similarity:

$$w_k = \psi(\mathbf{h}^t_{\mathbf{x}} \cdot \mathbf{c}_k) \tag{7}$$

$$\mathbf{h}^{t|s}_{\mathbf{x}} = \sum_{k=1}^{|M|} w_k \mathbf{c}_k, \tag{8}$$

where $\psi(\cdot)$ is a softmax function to normalize the dot product similarity scores across all prototypes.

To regulate negative transfer between two domains (Wang et al. 2019), the network must calibrate how much knowledge to transfer from source to target. We calibrate source knowledge by using a single feed-forward network g that computes additive attention weights between the source and target representations.

$$\mathbf{e}_{\mathbf{x}} = g(\mathbf{h}^t_{\mathbf{x}}, \mathbf{h}^{t|s}_{\mathbf{x}}) \tag{9}$$

$$\bar{\mathbf{h}}^{t|s}_{\mathbf{x}} = \phi_r([\mathbf{e}_{\mathbf{x}} \odot \mathbf{h}^{t|s}_{\mathbf{x}}, \mathbf{h}^t_{\mathbf{x}}]) \tag{10}$$

where ϕ_r is a learnable parameter and $\odot$ is a Hadamard product.

Fine-Tuning with Target Data. The target representation $\hat{\mathbf{h}}^t_{\mathbf{x}}$ is dynamically balanced between the direct transaction feature $\mathbf{h}^t_{\mathbf{x}}$ and the source-augmented feature $\bar{\mathbf{h}}^{t|s}_{\mathbf{x}}$ as in Eq. (6). We fine-tune the target country's fraud detection system $f(\cdot|\mathcal{D}_t)$:

$$\mathcal{L}_{cls} = -\frac{1}{|\mathcal{D}_t|} \sum_{\mathbf{x}, y \in \mathcal{D}_t} H(y, \hat{y}) \tag{11}$$

where $\hat{y} = f(\hat{\mathbf{h}}^t_{\mathbf{x}})$ is the predicted fraud score of transaction $\mathbf{x}$ and H is a binary cross entropy.

Results

This section tests feasibility of the proposed knowledge sharing system in terms of detection performance, accommodating multiple sources, and dependency on the required log size and model components.

Experimental Setting

Datasets We employed import declarations from four partner countries of WCO, whose data had been shared for the research purposes under the non-disclosure agreement. We refer to these countries M, C, N, and T in Table 1. The log size and GDP per capita vary by country, allowing us to test various source and target scenarios. Import declaration data is readily available by customs administrations of any country, and the data format is compatible with each other. We considered the common fields, including numeric variables on the item price, weight, and quantity and categorical variables on the HS6 code, importer ID, country code, and received office. The data also contained manual label information indicating whether each declaration was judged fraud or not, and the amount of tax raised after inspection. For the test purposes, these labels had been generated by inspecting all logs. We report the raw variables of import declarations in Table 2.

Country	M	C	N	T
Duration	4 years	4 years	5 years	5 years
Num. imports	0.42M	1.90M	1.93M	4.17M
Num. importers	41K	9K	165K	133K
Num. tariff codes	1.9K	5.5K	6.0K	13.4K
GDP per capita	$300	$1,500	$2,200	$3,300

Table 1: Statistics of the datasets

Variable	Description
Quantity	Specified number of items
Gross weight	Physical weight of the items
HS code	Item code using harmonised system
Country code	Country from which goods were imported
CIF value	Costs including insurance & freight
Total taxes	Tariffs calculated by initial declaration
Illicit (y)	Target binary variable indicating fraud
Revenue	Amount of tax raised after inspection

Table 2: Overview of the transaction-level import data, in which the description of each variable are provided.

Evaluation Metric Raised tax is one of the critical screening factors because customs import duties make up a substantial proportion of the tax revenue in many countries (Grigoriou 2019). We hence consider the amount of tax raised from inspection as the performance indicator. In all experiments, we assume a 5% inspection rate unless otherwise mentioned. We use the term Revenue@5% to represent the ratio of the expected tax collected by inspecting 5% of the trades sorted by the fraud detection algorithm out of the maximum revenue that could have been raised when the entire log was to be inspected.

Training Details The model f is individually pretrained 10 epochs for every country and fine-tuned for 30 epochs. The sharpening temperature τ in Eq (4) was set as 0.07 by following previous works (Gunel et al. 2020). Smaller SCL loss temperatures benefit training more than higher ones, but extremely low temperatures are harder to train due to numerical instability. The number of prototypes was set to 500 per class, thus $|\mathcal{C}_s| = 1,000$. We vary this count later. We use the final month of each dataset as the test set, and the validation set was chosen as two weeks prior. The model was optimized using Ranger with a weight decay of 0.01 (Wright and Demeure 2021). The batch size and learning rate were set as 128 and 0.005 for the entire training epochs. All models were run five times, and their average values were reported.

Performance Evaluation

The first set of experiments (Exp 1–2) aims to test the effectiveness of domain adaptation for single-source and multi-source scenarios. The next experiments (Exp 3–6) test the model's sensitivity and dependency to the target country's log size, model components, and memory-bank usage.

Exp 1. Single-Source Scenarios We inspect all combinations of the source and target pair as follows. The source country will share knowledge based on the 5% of its fraud-like logs $\mathcal{D}_s$ sorted by $\hat{y}$. This knowledge is used in four different ways: (1) no sharing at all, (2) standard fine-tuning by sharing model parameters of $f(\cdot|\mathcal{D}_s)$ directly with the target country, (3) transfer learning based on the adaptive knowledge consistency technique (Abuduweili et al. 2021), and (4) the proposed prototype-based memory bank. Below is a detailed explanation of the baselines used in our experiments.

- Target Only (No sharing): Measure the performance of a base fraud detection model $f(\cdot|\mathcal{D}_t)$ trained on target data. The tree-based embedding model DATE is used (Kim et al. 2020) for f. DATE classifies and ranks illegal trade flows that contribute the most to customs revenue when caught.
- Vanilla Transfer: Sharing parameters of the fraud detection model $f(\cdot|\mathcal{D}_s)$ from source country s and directly fine-tune using target dataset $\mathcal{D}_t$.
- Adaptive Transfer: Adaptive Knowledge Consistency (AKC) technique (Abuduweili et al. 2021) is used to cope with the risk of negative transfer caused by the discrepancy between the source and target countries. It constrains the mean square error between the pre-trained feature extractor outputs and the target feature extractor outputs by adaptively sampling the target dataset $\mathcal{D}_t$. We perform a hard filter according to the sample importance by sorting fraud score difference between the source and target pre-trained models. The features with a lower 20 % score difference were selected.

At the receiving end, we assume the target country has a weaker infrastructure and utilizes only 1% of trades for training. This assumption is simulated by randomly masking the log labels. The base fraud detection model f uses implementations of the tree-based embedding model (Kim et al. 2020), whose performance is reported in the column indicated as 'Target Only' in Table 3. The Revenue@5% performance for the target country M is the lowest, which may be contributed by the smallest log size. Note that country M has the lowest GDP per capita in Table 1. The efficacy of the detection model increases when utilizing knowledge shared by countries with larger log sizes (i.e., choosing Country T as the source country), although with some exceptions. Since the illicit rates of these countries are different, we only compare increments in performance within each target country.

When knowledge is shared in any form, the tax revenue from fraud detection generally increases. For the target country M, the use of DAS from country N increases the tax revenue 11 times from 0.0466 to 0.5152. The exact benefit differs by country, implying that various factors like the log similarity between countries may play a role. Parameter sharing (i.e., column indicated as Vanilla Transfer) sometimes leads to degraded performance due to negative transfer. Transfer learning via the adaptive knowledge consistency technique (i.e., Adaptive Transfer) no longer shows this limitation. Sharing knowledge in the form of prototypes shows the best performance in terms of tax raised, as indicated in the final column of the table. Note this method also

Test Setup	Target Only	Vanilla Transfer	Adaptive Transfer	Proposed DAS
N → M	0.0466	0.0681	0.1091	0.5152
C → M		0.0869	0.1405	0.5271
T → M		0.2112	0.2458	0.2966
M → C	0.0853	0.1806	0.1843	0.1915
N → C		0.0297	0.1081	0.1794
T → C		0.0719	0.1402	0.2271
M → N	0.1837	0.1018	0.3244	0.6681
C → N		0.2368	0.4043	0.6863
T → N		0.1119	0.3032	0.7286
M → T	0.1541	0.3005	0.3037	0.3181
C → T		0.2844	0.3012	0.3073
N → T		0.2007	0.3033	0.3282

S → T indicates the direction of knowledge flow from the source country S to the target country T.

Table 3: Revenue@5% performance under four different settings: (1) when no knowledge is shared and only the target country's logs are used for detection, (2) when the source country's knowledge is shared via direct parameter sharing, (3) when knowledge is shared via transfer learning to mitigate the negative transfer problem, and (4) when knowledge is further embedded as prototypes and shared as proposed in our system. The ratio of tax raised by DAS is substantial compared to other settings.

increases data protection of the source countries.

Exp 2. Multi-Source Scenarios Accumulating knowledge from more than one country further increases the performance of the fraud detection system. Figure 4 shows the average Revenue@5% for using no source at all, a single source, two sources, and three sources. To compute, we averaged the performance of all possible combinations according to the number of source countries. The additional gain in raised tax is most noticeable for using a single source, yet there is a revenue gain with every added source. When all available source knowledge is used, the target countries benefit nearly 2 to 11 times increase in total tax raised.

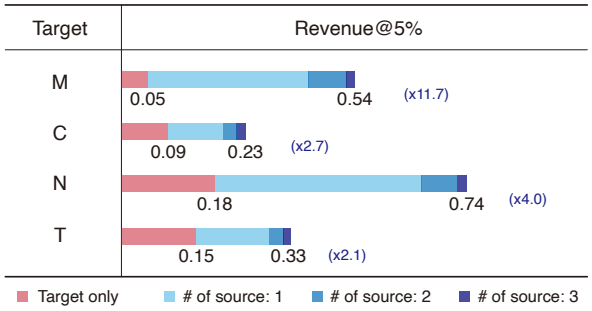

Figure 4: Collected duties are expected to increase 2.1–11.7 times for all tested countries when shared knowledge contributed by multiple sources is used (i.e., blue bars) compared to relying on local knowledge alone (i.e., red bars).

Exp 3. Dependency on Log Size Next, we examine measurable effects such as the log size. Figure 5 shows the average benefit of utilizing a single source versus no shared knowledge over the increasing ratio of log size for each target country. The performance gap is the largest when the log size is 1%, implying DAS can benefit countries with weak infrastructure the most. Yet, the continued benefit is observed for increasing log size from 2% to 10% in the target country, even at a marginal level. The benefit for country T is the smallest, which has the highest GDP per capita. Nonetheless, we emphasize that sharing knowledge increases the capacity to manage risk in the neighboring participating countries, ultimately curtailing illicit trades in that region.

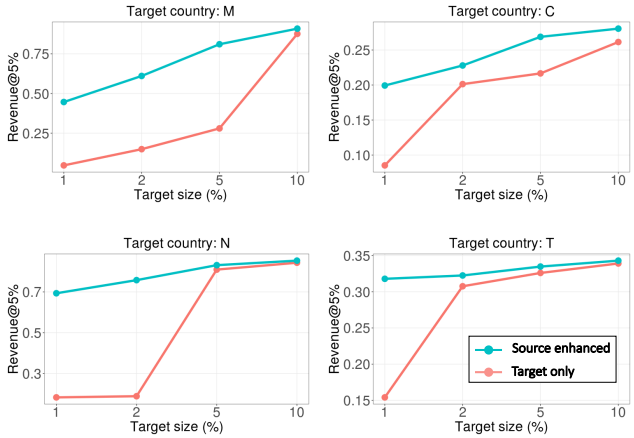

Figure 5: Performance improvement as the log size of the target country increases. The shared knowledge brings the most considerable benefit when the available log size is the smallest (i.e., 1%) and for countries with the weaker economy (i.e., country M).

Exp 4. Dependency on Model Components Table 4 reports the performance degradation when each component of the model is missing. The experiments are done with the target country N with a 1% labeled setting. The full model with all components performs the best, justifying the role of each component. Mainly, domain invariant encoding of DAS contributes to the model performance the most. This experiment considers five scenarios.

- Full model: DAS containing all components.
- Without domain invariant encoding: The model without Eq. 1–3 that simply concatenate between HS6 code and transaction embedding.
- Without contrastive learning: The model without Eq. 4, which uses cross-entropy loss instead of SCL loss for pretraining source dataset.
- Without memory bank: The model without Eq. 8, which fine-tunes the target data without utilizing the information of the source memory bank.
- Without domain calibration: The model without Eq. 10 that only considers the interaction between source feature and target feature.

Setup	M → N	C → N	T → N
Full model	**0.6681**	**0.6863**	**0.7286**
w/o domain invariant encoding	0.4979	0.6668	0.2101
w/o contrastive learning	0.5149	0.6766	0.7018
w/o memory bank	0.6667	0.6564	0.6762
w/o domain calibration	0.6670	0.6657	0.6825

Table 4: Ablation results on country N

Exp 5. Efficacy of Memory Contents Additionally, we discuss the augmentation effect when the source knowledge is transferred through the memory bank. For comparison, we consider a baseline of random memory bank with the same size, inspired by the prior work that randomly injected noise provides practical benefits to deep models (Poole, Sohl-Dickstein, and Ganguli 2014). Table 5 shows that memory bank usage provides non-trivial improvements than using random memory banks. This finding confirms that information shared by source countries is successfully leveraged in the form of prototypes.

	Target country			
Memory setup	M	C	N	T
No memory	0.5183	0.2233	0.6762	0.3190
Random	0.5241	0.2265	0.7126	0.3238
Three countries	**0.5435**	**0.2340**	**0.7395**	**0.3286**

A pretrained network $f(\cdot|\mathcal{D}_s)$ from Eq. (4) is shared together regardless of the memory bank setup.

Table 5: Efficacy of memory-based augmentation

Exp 6. Dependency on Prototype Count Knowledge of frauds is shared as compressed representations called prototypes. The number of prototypes can determine the resolution of transferred knowledge. For instance, one may assume selective fraud patterns may be shared when limiting the prototype count. We test the sensitivity of the model on prototype count $|\mathcal{C}_s|$ by setting it from 10 to as large as the source data size $|\mathcal{D}_s|$. Table 6 shows detection performance remains intact by $|\mathcal{C}_s|$, implying that the model is not sensitive to the choice of hyper-parameters as long as the prototypes are used.

$\lvert\mathcal{C}_s\rvert$	0	10	100	1,000	$\lvert\mathcal{D}_s\rvert$
M → N	0.6667	0.6714	0.6721	0.6681	0.6709
C → N	0.6564	0.6656	0.6727	0.6863	0.6711
T → N	0.6762	0.7379	0.7307	0.7286	0.7314

Table 6: Sensitivity analysis to the number of prototypes for country N. $|\mathcal{D}_s|$ implies that the entire embeddings are transferred instead of the clustered prototypes.

Discussion and Conclusion

This paper presented a first-of-a-kind knowledge sharing system for multi-national customs administrations. Below we discuss the implications of findings in terms of risk management and protection of trade information.

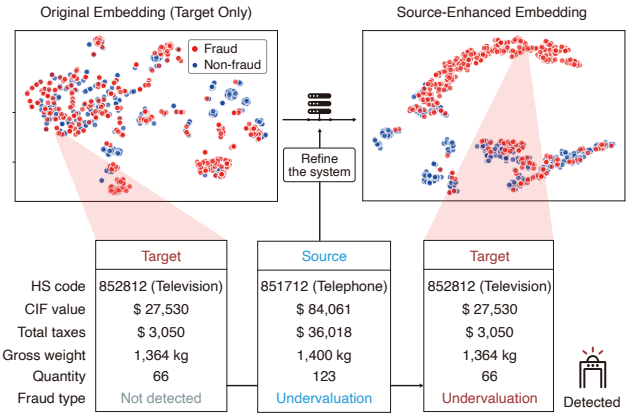

Figure 6: t-SNE plots of the learned embeddings (T → N), when the model is trained only with target-only feature (left) and with source enhanced feature using DAS (right). Fraud cases are successfully detected after receiving prototypes from the source country. Similar fraud examples from the source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

Discussion

DAS has shown substantial revenue potential by utilizing a relatively small fraction of fraud-like logs. Custom offices with weaker infrastructure will likely observe the largest increase in raised tax, whereas the more equipped customs offices will see a smaller gain. However, the benefit needs to be interpreted in two other aspects. First, once the system is set up, the cost of fine-tuning for additional prototypes will be minimal. Hence, any increase in detection performance translates to a potentially substantial additional tax revenue in the target country. Second, empowering countries with weak infrastructure will enable the detection of new fraud patterns that were previously unseen. The new fraud patterns collected from target countries will be shared back to the memory bank. This helps strengthen policy against illicit trades by removing weak spots in the global trade chain and further benefit participating countries (Wang 2018).

Qualitative analysis can be used to show how well the shared information discriminates against illicit transactions. Figure 6 compares embedding results of the target country, without knowledge (i.e., left) and with knowledge (i.e., right). Using the source-enhanced features helps better distinguish frauds from non-frauds. This particular example shows a fraud case that was newly detected by DAS. Without the shared knowledge, this declaration would have been missed. In contrast, the knowledge augmentation placed this declaration more closer to the fraudulent cluster in the embedding.

The proposed method of sharing prototypes is a safe way to transfer knowledge across heterogeneous administrative domains. To our understanding, it is nearly impossible to re-identify information from a memory bank. Here, we show the input data utilized for knowledge building is statistically

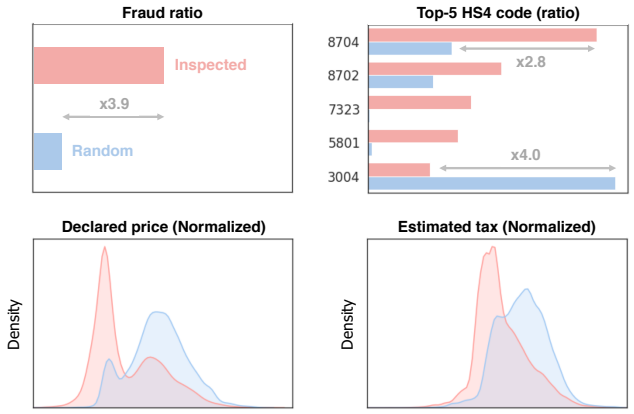

Figure 7: Shared knowledge of frauds is distinct from the mass of the normal trade volume contributed by member offices. Exact figures are hidden.

different from the vast "normal" trades that include critical information about trade partners and prices—information that is confidential and of concern in data sharing.

Figure 7 highlights some of the differences observed for fraud-suspected logs and normal logs. The logs utilized by the model have nearly four times higher fraud ratio than normal logs. In addition, the product categories that frequently appear in fraud-like logs and normal logs have different rank orders, as illustrated in the examples. The average declared price (CIF value) indicates that the initial reporting tariffs written in the declaration form are substantially lower for the inspected logs. Only fraud-suspected logs are used for knowledge building, and the normal data are excluded in the data embedding step.

Concluding Remark

There is an increasing need to share knowledge in data-critical sectors, including customs, medicine, finance, and science. We tested the premise that shared knowledge will facilitate and advance risk management in customs fraud detection. Domain adaptation techniques are one way to build collective knowledge across international administrations. Testing with million-scale data demonstrated that transferable knowledge increases the detection performance for all participating countries. A centrally managed memory bank system that we explored is one such possibility.

As global trade sees dynamic changes due to events like COVID-19, maintaining an agile risk management system has become ever more important. Strengthening the data science capacity at customs administrations and collaboration will be critical as illicit traders continue to use vantage points in the trade network. Starting from this proof of concept, we envision more countries agreeing on the legal support for sharing prototypes outside their administrative domains. When more countries participate in the regional memory bank platform, we expect knowledge sharing will better handle the data challenges. Furthermore, the regional ties will help enable positive trade agreements among the participating countries.

Acknowledgments

This work was supported by the Institute for Basic Science (IBS-R029-C2, IBS-R029-Y4) and the National Research Foundation (No. NRF-2017R1E1A1A01076400) funded by the Ministry of Science and ICT in Korea. We thank the World Customs Organization (WCO), the Capacity Building Directorate, and the partner countries for support in data access. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of WCO.

References

Abuduweili, A.; Li, X.; Shi, H.; Xu, C.-Z.; and Dou, D. 2021. Adaptive consistency regularization for semi-supervised transfer learning. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6923–6932.

Asghar, N.; Mou, L.; Selby, K. A.; Pantasdo, K. D.; Poupart, P.; and Jiang, X. 2020. Progressive memory banks for incremental domain adaptation. In *Proc. of International Conference on Learning Representations*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *Proc. of International Conference on Machine Learning*, 1597–1607.

Filho, J. J. 2015. Artificial intelligence in the customs selection system through machine learning (SISAM). *Receita Federal do Brasil*.

Goldberg, P.; Khandelwal, A.; Pavcnik, N.; and Topalova, P. 2009. Trade liberalization and new imported inputs. *American Economic Review*, 99(2): 494–500.

Grigoriou, C. 2019. Revenue maximisation versus trade facilitation: The contribution of automated risk management. *World Customs Journal*, 13(2): 77–90.

Gunel, B.; Du, J.; Conneau, A.; and Stoyanov, V. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. In *Proc. of International Conference on Learning Representations*.

Han, S.; Park, S.; Park, S.; Kim, S.; and Cha, M. 2020. Mitigating embedding and class assignment mismatch in unsupervised image classification. In *Proc. of the European Conference on Computer Vision*, 768–784.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778.

He, X.; Du, X.; Wang, X.; Tian, F.; Tang, J.; and Chua, T.-S. 2018. Outer product-based neural collaborative filtering. In *Proc. of International Joint Conference on Artificial Intelligence*, 2227–2233.

Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.-Y.; Isola, P.; Saenko, K.; Efros, A.; and Darrell, T. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *Proc. of International Conference on Machine Learning*, 1989–1998.

Hua, Z.; Li, S.; and Tao, Z. 2006. A rule-based risk decision-making approach and its application in China's customs inspection decision. *The Journal of the Operational Research Society*, 57(11): 1313–1322.

Kang, G.; Jiang, L.; Yang, Y.; and Hauptmann, A. G. 2019. Contrastive adaptation network for unsupervised domain adaptation. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4893–4902.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised Contrastive Learning. In *Proc. of Neural Information Processing Systems*, 18661–18673.

Kim, S.; Mai, T.-D.; Han, S.; Park, S.; Khanh, T. N. D.; Soh, J.; Singh, K.; and Cha, M. 2022. Active Learning for Human-in-the-Loop Customs Inspection. *IEEE Transactions on Knowledge and Data Engineering*.

Kim, S.; Tsai, Y.-C.; Singh, K.; Choi, Y.; Ibok, E.; Li, C.-T.; and Cha, M. 2020. DATE: Dual attentive tree-aware embedding for customs fraud detection. In *Proc. of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2880–2890.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13): 3521–3526.

Liu, Z.; Miao, Z.; Pan, X.; Zhan, X.; Lin, D.; Yu, S. X.; and Gong, B. 2020. Open compound domain adaptation. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12406–12415.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2017. Deep transfer learning with joint adaptation networks. In *Proc. of International Conference on Machine Learning*, 2208–2217.

Mai, T.-D.; Hoang, K.; Baigutanova, A.; Alina, G.; and Kim, S. 2021. Customs fraud detection in the presence of concept drift. In *Proc. of the International Conference on Data Mining Workshops*, 370–379.

Mikuriya, K.; and Cantens, T. 2021. If algorithms dream of Customs, do customs officials dream of algorithms? A manifesto for data mobilisation in Customs. *World Customs Journal*, 14(2): 3–22.

Poole, B.; Sohl-Dickstein, J.; and Ganguli, S. 2014. Analyzing noise in autoencoders and deep networks. *arXiv preprint arXiv:1406.1831*.

Thota, M.; and Leontidis, G. 2021. Contrastive domain adaptation. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2209–2218.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7167–7176.

Vanhoeyveld, J.; Martens, D.; and Peeters, B. 2020. Customs fraud detection: Assessing the value of behavioural and high-cardinality data under the imbalanced learning issue. *Pattern Analysis and Applications*, 23: 1457–1477.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Proc. of Neural Information Processing Systems*, 5998–6008.

Wang, F. 2018. Interagency coordination in the implementation of single window: Lessons and good practice from Korea. *World Customs Journal*, 12(1): 49–68.

Wang, Z.; Dai, Z.; Póczos, B.; and Carbonell, J. 2019. Characterizing and avoiding negative transfer. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11293–11302.

Weerth, C. 2009. The structure and function of the World Customs Organization. *Global Trade and Customs Journal*, 4(5): 131–154.

Wright, L.; and Demeure, N. 2021. Ranger21: a synergistic deep learning optimizer. *arXiv preprint arXiv:2106.13731*.

Wu, Z.; Xiong, Y.; Yu, S. X.; ; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3733–3742.

Zhang, W.; Deng, L.; Zhang, L.; and Wu, D. 2020. A survey on negative transfer. *arXiv preprint arXiv:2009.00909*.

Zheng, Z.; and Yang, Y. 2019. Unsupervised scene adaptation with memory regularization in vivo. In *Proc. of International Joint Conference on Artificial Intelligence*, 1076–1082.