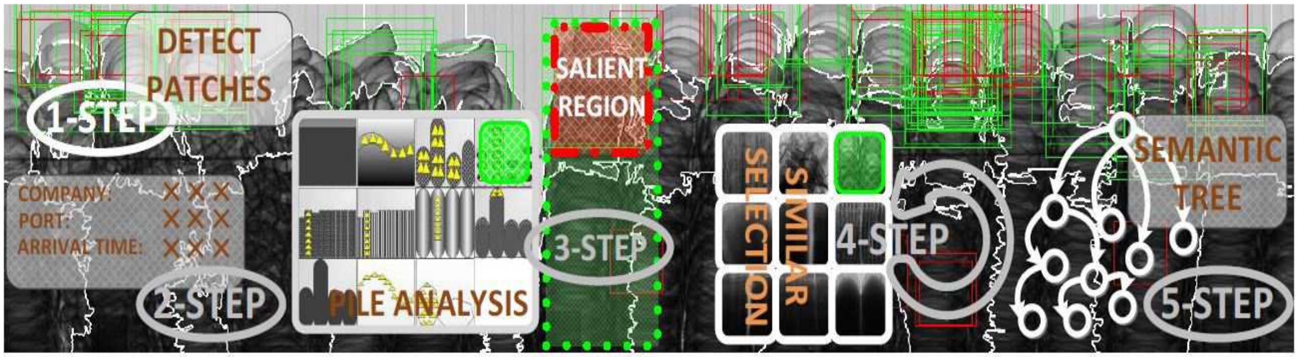
A comprehensive solution for deep-learning based cargo inspection to discriminate goods in containers

Jiahang Che
NUCTECH company limited
chejiahang@nuctech.com

Yuxiang Xing*
Tsinghua University
xingyx@mail.tsinghua.edu.cn

Li Zhang
Tsinghua University
zli@mail.tsinghua.edu.cn



Abstract

In this work, we attempt to classify commodities in containers with HS(harmonized system) codes, which is a challenging task due to the large number of categories in HS codes and its hierarchical structure based on a product's composition and economic activity. To tackle this problem, in this paper we propose an ensemble model which incorporates fine-grained image categorization, data analysis on cargo manifests, and human-in-the-loop paradigm. By employing deep learning, we train a triplet network for fine-grained image categorization. Then, by investigating massive information from cargo manifests, unreasonable predictions can be filtered out. With human-in-the-loop embedded, human intelligence is integrated to justify the resulted HS codes. Moreover, a HS code semantic tree is built to trade off specificity and accuracy.

1. Introduction

Globalization creates unprecedented opportunities in pursuit of economic prosperity. Nevertheless, enormous increases in international trade poses serious challenges to large container inspection which is exploited by customs to verify commodity descriptions to ensure goods in compliance with regulations as well as fence off improper entry. Traditionally, inspectors analyze inspection images and make decisions. This exposes weakness in several respects: low efficiency, heavy reliance on the experience of inspectors and so forth. Therefore, an intelligent fine-grained classification for X-ray images in container inspection system is urgently needed to serve as an assistant tool for inspectors.

Among generic image classifications, fine-grained image classification is more challenging because of the subtle inter-class variation and large intra-class variation. In recent years, fine-grained image classification has captured intense interest and remarkable progress has been made in this field. Especially with deep learning methods dominating image classification tasks, performance on standard dataset, Caltech-UCSD Birds-200-2011, has increased to 85% without the human-in-the-loop [8]. Besides, the accuracy on the well-known 102 Oxford flower dataset has raised up to 90% [5].

As to container inspection systems, there are two essential obstacles to overcome in container inspection tasks.

One obstacle comes from the X-ray imaging process. X-rays from an accelerator source pass through a container and the penetrated/remaining X-ray photons are collected by detectors in the other side of the container. The signals formed by the detectors are further processed to generate an digital radiographic image used for inspection. Therefore, unlike visible light images, the X-ray images are mostly obscured

*Corresponding Author

by overlapped objects along X-ray pathes. Edges of various goods could be mingled together on the X-ray image and hardly detectable. Moreover, once goods in the container are rearranged or X-rays are projected from another perspective angle, radiographic images will differ greatly [22].

Second difficulty we are facing lies in the HS (harmonized system [1]) classification that is well-used in customs declaration. HS comprises 21 sections, 96 chapters and further divided into more than 5000 heading and subheadings. Additionally, due to its organization by component material and economic activity, commodities with identical materials may be designated with different HS codes. For instance, fresh potatoes locate in '07019000' and frozen ones in '07101000'. For a brief introduction to HS code and HS classification please refer to section 3.4.

Our task can be identified as fine-grained HS classification. Namely, we intend to classify traded goods in a container with HS codes for customs inspection purpose. To solve this problem and overcome the difficulties described above, in this paper, we propose an ensemble model which is composed of three models. The first model is for image categorization. By leveraging deep learning, we train the triplet network to build up a feature dataset. By investigating the scopes of main product of companies generated from cargo manifests, a second model is built to provide a probability describing the likelihood of HS classes for the given company. Finally, user responses are taken into account to correct inaccurate prediction. Further, a HS code semantic tree is built up for balancing specificity and accuracy.

In the next section, we review related work on fine-grained image classification. We explain our ensemble model including four parts in detail: image categorization model, data analysis model, user response model and HS code semantic tree in the third section. Finally, we demonstrate our experimental results.

2. Related Works

Fine-grained image classification refers to the problem of classifying images into subcategories within a common entry level category. It is an extremely challenging task due to subtle inter-class variation and large intra-class variation. Recently, it has drawn a lot of attention and some standard datasets such as Caltech-UCSD Birds-200-2011 [20] ,Oxford flowers [12] and butterflies [9] has been collected to facilitate research. Great efforts have been devoted in early works on designing the feature extractors [6, 10] and the overall performance is strongly dependent on these handcrafted features. By employing deep neural network, manual feature extraction is replaced and great progress on fine-grained image classification has thus been made [23, 7, 25, 8]. Among them, a prevailing two-stage framework is applied to deal with the problem of fine-grained image classification. With localizing representative regions, features are extracted to train classifiers subsequently [24, 11].

Additionally, a human-computer method is firstly introduced in [2]. With well pre-designed 20 questions, this framework provides the possibility of incorporating any object recognition algorithm with human expertise, which drives up the accuracy to 95% for the dataset Caltech-UCSD Birds-200-2011. As an extension of [2], with two heterogeneous forms of information as user responses and by applying localized part and attribute detectors, the total amount of human effort is further reduced [18]. Later, a novel human-in-the-loop fine-grained categorization system is developed [19] based on perceptual similarity rather than expert-driven vocabulary, which reduces reliance on the expert-defined terminology so that it is flexible to be applied in other domains. A new image representation called bag-of-FLHs is built in [5] and a kernel function is then used for classification to get 92% accuracy on Oxford flowers dataset.

Considering the special character of our problem, it would be of little help to employ conventional methods directly in generic fine-grained image classification to tackle the problem of HS image classification. HS code of commodities are determined by a set of factors including the material of which goods are composed, its function and its forms. As a result, even the commodities, who have the identical composition, may vary in HS codes. This forces us to resort to other valuable information accompanying with images. An ensemble model is thus designed to solve our task.

3. Technical Details

Our ensemble model is based on a probability framework which incorporates fine-grained image classification with data analysis on cargo manifests and user responses analysis. More precisely,

$$p(c|x, I_{3rd}, U_t) \propto p(c|x)p(c|I_{3rd})P(c|U_t, x). \qquad (1)$$

Here, x denotes a radiographic cargo image, I_{3rd} covers the information collected from cargo manifests, U_t includes a series of user responses, and c stands for the predicted HS-code. Besides, the function p is a conditional probability distribution which characterizes how likely goods in the container correspond to the predicted HS code with the given condition. This formula (1) indicates that our ensemble model combines three models: image categorization model, data analysis model, and user response model, to produce convincing results. The terms on the right side correspond to these three models respectively, which will be further discussed in section 3.1, 3.2, 3.3.

Moreover, we build a HS code semantic tree to trade off between accuracy and specificity. Once users consider the predicted probability of the 8-digit HS code is lower than expected, they can trace back level by level and refer to their parent nodes to obtain a more confident output(as shown Figure 5). For more details, please refer to section 3.5.

3.1. Image categorization model

At training stages, we aim to build a feature dataset which contains a list of representative features for each class. Followed is the fast objection detection algorithm used at testing stages.

3.1.1 Build a feature dataset

For clarity it is necessary to figure out the meaning of representative features. It is reasonable for us to assume that representative features of the same class group together whereas those of distinct classes are well separated.

Before deep learning is rapidly expanding its influence in computer vision, a traditional way to extract features is to use Fisher kernel framework [13]. Even though this approach has made much progress in image categorization at that time, it is still a shallow learning and involves a vast amount of parameters. It takes approximately 20 hours of CPU 2.4GHz to carry out an experiment of 50 categories with 5 samples in each category. Moreover, with an increase in the number of categories up to 250, the extracted features are barely satisfactory. We then resort to a more powerful tool, deep learning. In the past few years, deep learning has achieved unprecedented success in various domains and ranked top on many tasks. It has been also applied on automated inspection of dual energy X-ray imagery [14]. The leading architecture in deep learning models, convolutional neural network, is proven to surpass the traditional machine learning algorithms in feature extractions.

As aforementioned, representative features from the same class are expected to be close to each other whereas a margin is forced between those from distinct classes. Inspired by the great work [15], we can seek for an embedding f to obtain our feature dataset by training a Triplet network to achieve the goal. With pairs of images from the same class and images from different classes as input, a deep convolutional network, in this paper we choose Resnet-v1-50 to act as feature extractors and learn the feature representations of images. Thereafter, all the features are mapped onto a unit sphere with L2 norm. The process can be formulated as

$$\begin{aligned} \|f(x_i) - f(x_j)\|_2^2 + \alpha < \|f(x_i) - f(x_k)\|_2^2, \\ \forall\big(f(x_i), f(x_j), f(x_k)\big) \in \mathcal{T}, \end{aligned} \quad (2)$$

where x_i(anchor) and x_j(positive) share the same label, x_k comes from other classes. Here α is the margin between positive and negative samples. And f is the embedding from input images to N dimensional unit sphere satisfying $\|f(\cdot)\|_{L^2} = 1$. The inequality above gives directly the triplet loss

$$L_T = \sum_{i,j,k}^{N} \Big[\|f(x_i) - f(x_j)\|_2^2 - \|f(x_i) - f(x_k)\|_2^2 + \alpha\Big]_+ . \quad (3)$$

And the total loss, which is combination of classification loss and triplet loss, is minimized by adjusting the weights in the network. Figure 1 illustrates the triplet network.

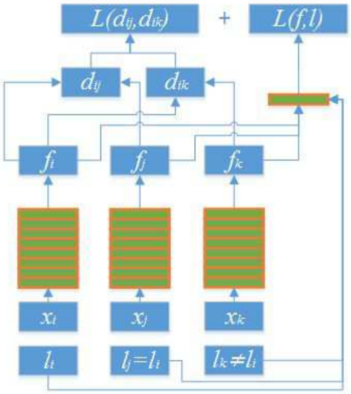


Figure 1. Triplet Network

3.1.2 Fast object detection algorithm

Following [3], each filter-sized window of a feature can be encoded as a high-dimensional sparse binary descriptor called WTA(Winner-Take-all) hash [21]. As a result, a detection process is thus transformed to be a search problem. Two preconditions of this algorithm are listed as follows
(1) Compared with a linear space, the ordinal space is more appropriate to qualitatively describe the differences between data
(2) The measure of similarity in the linear space is less insensitive than in ordinal space

It is observed that each WTA hash function defines an ordinal embedding and is suitable as a basis for locality-sensitive hashing [3], which implies that the first condition is satisfied. As to the second one, it suffices to prove that for any $\bar{x}, \bar{y}, \bar{z} \in \mathcal{R}^d$,

$$L_2(\bar{x}, \bar{y}) < L_2(\bar{x}, \bar{z}) \Leftrightarrow K(\bar{x}, \bar{y}) < K(\bar{x}, \bar{z}), \quad (4)$$

where $L_2(\cdot, \cdot)$ and $K(\cdot, \cdot)$ are defined as

$$\begin{aligned} L_2(\bar{x}, \bar{y}) &= \|\bar{x} - \bar{y}\|_2^2 \quad (5) \\ K(\bar{x}, \bar{y}) &= \bar{x} \cdot \bar{y}. \quad (6) \end{aligned}$$

It is easily verified that

$$L_2(\bar{x}, \bar{y}) = K(\bar{x}, \bar{x}) + K(\bar{y}, \bar{y}) - 2K(\bar{x}, \bar{y}) \quad (7)$$

which leads to the desired result (4) for normalized vectors $\|\bar{y}\|_2 = \|\bar{z}\|_2 = 1$.

This fast object detection algorithm allows us to implement feature matching process with highest response in $O(1)$ time independent of the quantity of features. Figure 2 demonstrates the computation process. In the experiment, we keep the first $k = 4$ indices for each of $N = 3000$ permutations and band size $W = 4$ to implement a WTA hash function.

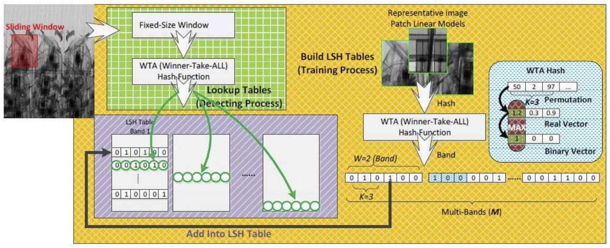


Figure 2. The training and testing process of fast object-detection algorithm. Training-Process: encode all fix-sized filter vectors from section 3.1.1 by WTA hashing function and decompose every binary descriptor into M bands with W spans of length K and store each band in its respective LSH(local sensitive hash) table. Testing-Process: slide window and compute the WTA hash and break into bands, look up each band in its corresponding LSH table and count how many times each filter occurs in all tables. The class label is obtained by those filters with more occurrences.

3.2. Data analysis model

It is required to submit cargo manifests prior to or upon the arrival of containers. And abundant information is carried by cargo manifests. By taking product scopes of companies into account, the range of predictions on cargo categories can be narrowed. In addition, port statistics including location, arrival time correlates cargo categories as well, which is summarized as follows

$$p(c|I_{3rd}) \propto p(c|\text{company})p(c|\text{port, time}), \tag{8}$$

where $p(c|\cdot)$ denotes the probability distribution of prediction with the given condition and I_{3rd} represents the information collected from cargo manifests. In the following we will discuss $p(c|\text{company})$ and $p(c|\text{port, time})$ in detail.

3.2.1 Modelling $\mathbf{p(c|company)}$

The main product scopes are collected from either yellow pages from internet or our training set collected from cargo manifests. Samples from yellow pages present in the following way

$$\text{AMERICAN TACK\&HARDWARE} \Leftrightarrow 28 : 1$$

With supplier-buyer brand names on the left side, 2-digit HS code and its corresponding transportation frequency are listed on the right side.

And our training dataset may offer a longer digits of HS code,

$$\text{AMERICAN TACK\&HARDWARE} \Leftrightarrow$$
$$28046190 : 1; 28049010 : 1.$$

We define

$$p(c|\text{company}) = \begin{cases} 0.8 * F * \alpha, & \text{c}\in \text{product scopes} \\ 0.2, & \text{others} \end{cases} \tag{9}$$

where F is the cumulative frequency of c, and α is given by

$$\alpha = \begin{cases} 0.4, & \text{product scopes with 2-digit HScode} \\ 0.6, & \text{product scopes with 4-digit HScode} \\ 0.8, & \text{product scopes with 6-digit HScode} \\ 1.0, & \text{product scopes with 8-digit HScode} \end{cases} \tag{10}$$

It is pointed out that if c belongs to some certain class, it falls into all its superclasses automatically. And the training dataset will update by incrementing all related frequencies by one each time.

3.2.2 Modelling $\mathbf{p(c|port, time)}$

Our port statistics are collected monthly. Each record is edited as

$$\text{port} - 01 \Leftrightarrow 03034300 : 2; 05074900 : 1 \cdots$$

$$\text{port} - 02 \Leftrightarrow 03045700 : 1; 05074900 : 2 \cdots$$

Here 'port-month' is located on the left side. HS codes and their corresponding frequencies are shown on the right side. Let us define

$$p(c|\text{port, time}) = \begin{cases} 0.8 * F, & \text{if c was recorded} \\ & \text{ at the port during} \\ & \text{the time interval} \\ 0.2, & \text{others} \end{cases} \tag{11}$$

And the corresponding frequencies will be refreshed by adding one afterwards.

3.3. User response model

By introducing human-in-the-loop, we leverage human and computer intelligence to improve the performance for real-life application. At the beginning t_0, the computer sends a request and waits for the user to input $u_0 = \{\tilde{p}, \tilde{x}_s\}$ with $\tilde{p}$ denoting the cargo arrangement mode and $\tilde{x}_s$ being a salient region. Subsequently, at each timestamp $t_i, i =$

$1, \cdots, n$, the similarity selection mode is activated. The computer will display ranking top N similar images from out feature dataset and users are required to pick up the most similar one with the index $\tilde{k}$ therein. The user response u_i is thus obtained with the index iteratively. Let $U_t = \{u_0, u_1, \cdots, u_t\}$ be the sequence of user responses. With c and x denoting the label and the image, the conditional probability $p(c|U_t, x)$ yields

$$p(c|U_t, x) \propto p(u_i, x|c) p(U_{t(t>0)}, c|x). \tag{12}$$

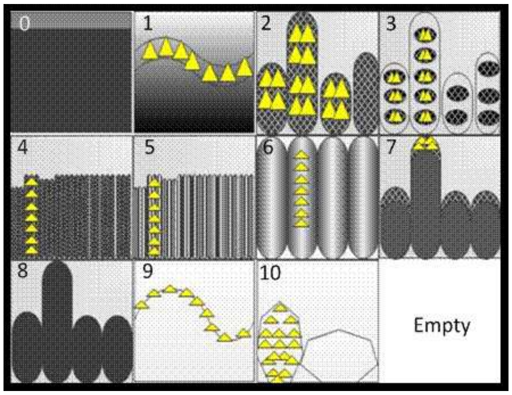


Figure 3. Cargo arrangement modes and their corresponding salient regions(yellow parts). For mode 0 and 8, no salient regions found.

3.3.1 Cargo arrangement mode analysis

It is noticed that $u_0 = \{\tilde{p}, \tilde{x}_s\}$. With the given image x, by applying chain rule we have

$$p(u_0, x|c) \propto p(\tilde{x}_s|c, \tilde{p}) p(\tilde{p}|c). \tag{13}$$

Figure 3 shows a total eleven cargo arrangement modes and their corresponding salient regions. For each cargo arrangement mode, with Texton features [17] extracted from its corresponding salient regions, we train a GMM (Gaussian Mixture Model) to characterize $p(\tilde{x}_s|c, \tilde{p})$. For the mode 0 and 8, we set $p(\tilde{x}_s|c, \tilde{p})$ to be one. Implementation details are given in figure 4.

3.3.2 Similarity selection process

Following [19] and [16], our probabilistic model is formulated as follows.

By integrating over all possible states z of image x in the perceptual space, we obtain

$$p(U_{t(t>0)}, c|x) = \int p(U_{t(t>0)}, c, z|x) dz, \tag{14}$$

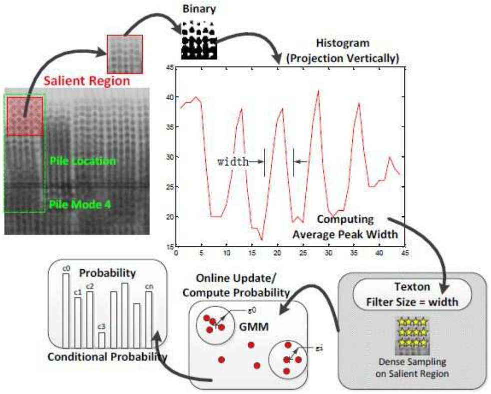


Figure 4. The process consists of feature extraction on salient regions, GMM update and the conditional probability computation. The arrangement mode and the salient region are determined by the user. Texton features extracted on salient regions are used to train one GMM model per class at training stages. These GMM models are used to compute the conditional probability at testing stages.

where $U_t = \{u_0, u_1, \cdots, u_t\}$. It is assumed that user responses depend only on the location in perceptual space, which implies that

$$\begin{aligned} p(U_{t(t>0)}, z, c|x) &= \prod_{t:t>0} p(u_t|c, z, x) p(c, z|x) \\ &= \prod_{t:t>0} p(u_t|z) p(c, z|x) \end{aligned} \tag{15}$$

by using chain rule. This leads to

$$p(U_{t+1}, z, c|x) = p(u_{t+1}|z) p(U_t, z, c|x). \tag{16}$$

On one hand, we suppose that the probability of the user choosing the most similar image $\tilde{k}$ within D candidates is proportional to its perceptual similarity,

$$p(u_{t+1}|z) = p(\tilde{k}|z) = \frac{s(z, z_{\tilde{k}})}{\sum_{i \in D} s(z, z_i)}, \tag{17}$$

where $s(\cdot, \cdot)$ is a similarity metric in perceptual space. On the other hand, by feeding the prediction obtained in the last round of iteration back into the algorithm,

$$p(c, z|x) = \frac{1}{N_c} p(c|x), \tag{18}$$

where N_c denotes the total number of training images in perceptual space with class label c. Consequently, $p(U_{t(t>0)}, c|x)$ is obtained by integrating (14)(15)(17) and (18). Further details may be found in [19] and [16]. Then the inaccurate predictions may be corrected by combining (1), (12), (13) and (14).

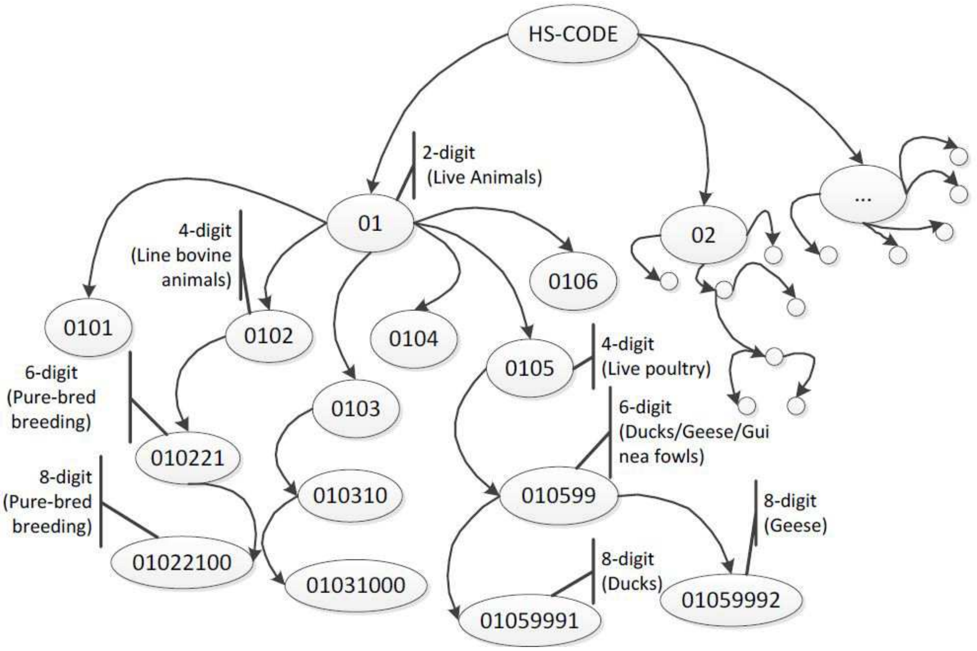


Figure 5. The HS code semantic tree describes the semantic correlations among different digit HS codes.

3.4. HS code semantic tree

Before we proceed directly, we would like to give a brief introduction to HS (The Harmonized System [1])code here. HS, an international nomenclature for the classification of traded products, comprises a hierarchical structure of commodity description, in which goods are designated from top to down. The HS comprises of 21 sections and 96 chapters. Each section groups several chapters which include a particular class of commodities. These chapters provide broad categories and are subdivided into about 5000 headings and subheadings so as to describe commodities in detail. The HS code consists of 6 digits and each country can modify by adding two or four digits to meet domestic needs with first 6 digits adopted universally. For each HS code, the first two digits designate the HS chapter while the latter two digits identify the position of heading in the chapter. The third two or more digits, if necessary, designate the subheadings. A HS code is determined by a list of factors including its material, form and function. Take whole potatoes for example, the classification differs depending on whether they are fresh or frozen. Frozen potatoes locate in 07101000 while fresh ones are classified in position 07019000.

Based on the hierarchical organization of HS, we build a HS semantic tree, as shown in Figure 5. The root node of the tree covers all HS codes and its child nodes include 2-digits HS codes. One level down the tree each time means two more digits added to the HS code. The process proceeds until every terminal node presents an 8-digit HS code. The semantic tree can be integrated seamlessly with our system. Once HS codes are predicted with low confidence, users can alternatively trace their parent nodes to improve the accuracy.

Let V denote the set of all different level HS codes including 2 digits, 4 digits, 6 digits and 8 digits. And Y is a subset of V and contains all 8-digit HS code. For any image $x \in X$, the classifier $f' : X \rightarrow V$

$$f'_\lambda(x) = \mathrm{argmax_{v \in V}(r_v + \lambda)P(v|x)} \tag{19}$$

is learned. Here

$$p(v|x) = \sum_{v_l \in \text{child}(v)} p(v_l|x) \tag{20}$$

with $p(v_l|x)$ obtained by our ensemble model and the information gain r_v is given by

$$r_v = \log_2 |Y| - \log_2 \sum_{y \in Y} I_{\{v \in \pi(y)\}} \tag{21}$$

with the characteristic function $I_{\{\cdot\}}$ and $\pi(y)$ representing all the ancestors of y.

By choosing different λ, we can flexibly balance between specificity and accuracy. A larger λ implies that a more abstract level rather than a more concrete level in the semantic tree is preferable. Further details may be found in [4].

4. Experiment

We perform the experiments on real customs data including approximately 10000 X-ray images from 632 categories and their corresponding cargo manifests including HS codes and other detailed information, which were collected from NUCTECH inspection systems.

4.1. Image categorization

The Triplet Network is trained in our experiment. We use sliding window of $64 * 64$ to loop over each training images and a packet of patches with size $64 * 64$ are thus obtained. The patches, in which the area of air is larger than some certain threshold, are thrown away. And the other patches are ready for training after standard image processing. It is worth mentioning that the number of patches per image must be the same in training process in order to be consistent with the input layer of Triplet Network. By initializing from ResNet-v1-50 pretrained on the ImageNet dataset, We use ResNet-v1-50 as our backbone architecture by replacing fully connected layers with 1×1 convolution layer to reduce feature dimensions. As a matter of fact, ResNet-v1-50 will suffice for the task. The choice of the networks is up to the developers. We use 0.0001 weight decay and 0.9 momentum and set an initial learning rate to be 0.01, which is annealed by an exponential decay every 5 epochs.

We evaluate our performance on our test images. A stack of patches, which are obtained by sliding window with the same fixed size on test images, are input into well-trained Triplet Network. Then we compute the average intra-class distances and average inter-class distances, which are shown in the blue part and orange part in Figure 6 respectively. Compared with the histogram by Fisher kernel framework [13] in Figure 7, it is demonstrated that the Triplet Network surpasses the traditional method and is more powerful to extract features.

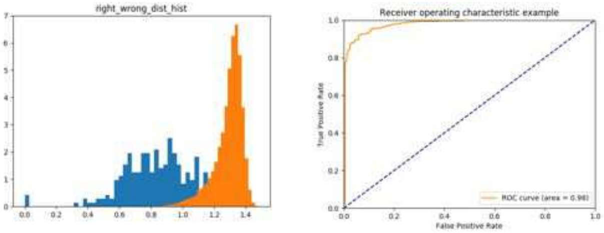


Figure 6. Histogram of average distances for features in Triplet network and its ROC curve

4.2. Data analysis

The abundant information in the cargo manifests is helpful for our HS classification task. The probability model filters out uncorrelated predictions effectively so that the accuracy increases by 10%.

4.3. Human-in-the-loop

If users could pick up the most similar image correctly in the process, the performance would rise up to more then 80%. However, users may not respond perfectly. The cargo arrangement mode 3 might easily be confused with the mode 4 and 5. Moreover, user responses are affected by subjective differences. These two reasons limit the performance of the system. Table 1 exhibits the accuracies and their average query times.

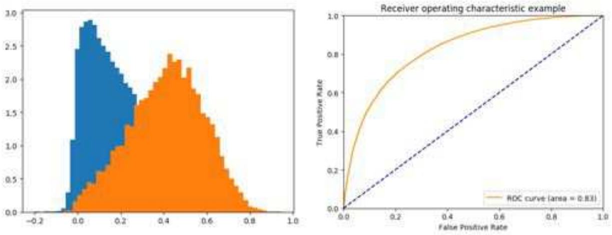


Figure 7. Histogram of average distances for features in Fisher kernel framework [13] and its ROC curve

Classification Accurary	Average Query Times
45%	3.2
50%	5.8
55%	15.3
60%	30.3

Table 1. Relations between accuracy and average query times.

4.4. HS code semantic tree

We set $\lambda = 5$ in our experiment and the depths of predictions in the semantic tree is about 3.1 averagely, which implies that the predicted results mostly represent 8-digit HS codes and only a few results with 2-digit and 4-digit HS codes are obtained. Table 2 shows that the relations among λ, average predicted depth and the accuracy.

Parameter λ	Average Level	Accuracy
$\lambda = 0.0$	4	30%
$\lambda = 0.5$	3.7	33%
$\lambda = 1.5$	3.1	60%

Table 2. Relations among λ, average predicted depth and the accuracy.

5. Conclusion

In this paper, we propose a comprehensive system to verify commodity descriptions in the containers. Supported by deep learning, we build a feature dataset for fine-grained image categorization. Moreover, by investigating cargo manifests, main product scopes of each company can be obtained and used to correct inaccurate predictions. And the human expertise is incorporated to empower our model by analyzing user responses in the human-in-the-loop. Therefore, we can improve the overall performance so that the inspection system is applicable. Currently, deep learning serves as a feature extractor in the first image categorization model of our system. In the future, we will look into ways of building an end-to-end deep learning framework for HS classification.

6. Acknowledgement

This research was supported by NUCTECH company limited. We would like to thank Jian Zhang for his insightful suggestions to this work. Also we are thankful to our colleagues and especially grateful to Qiang Li and Gang Fu for their discussions.

References

[1] https://en.wikipedia.org/wiki/Harmonized_System.

[2] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *European Conference on Computer Vision*, pages 438–451. Springer, 2010.

[3] T. Dean, M. A. Ruzon, M. Segal, J. Shlens, S. Vijayanarasimhan, and J. Yagnik. Fast, accurate detection of 100,000 object classes on a single machine. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1814–1821. IEEE, 2013.

[4] J. Deng, J. Krause, A. C. Berg, and L. Fei-Fei. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3450–3457. IEEE, 2012.

[5] B. Fernando, E. Fromont, and T. Tuytelaars. Effective use of frequent itemset mining for image classification. In *European conference on computer vision*, pages 214–227. Springer, 2012.

[6] S. Gao, I. W.-H. Tsang, and Y. Ma. Learning category-specific dictionary and shared dictionary for fine-grained image categorization. *IEEE Transactions on Image Processing*, 23(2):623–634, 2014.

[7] X. He and Y. Peng. Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification. In *AAAI*, pages 4075–4081, 2017.

[8] X. He, Y. Peng, and J. Zhao. Fast fine-grained image classification via weakly supervised discriminative localization. *arXiv preprint arXiv:1710.01168*, 2017.

[9] S. Lazebnik, C. Schmid, and J. Ponce. Semi-local affine parts for object recognition. In *British Machine Vision Conference (BMVC'04)*, pages 779–788. The British Machine Vision Association (BMVA), 2004.

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[11] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[12] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Computer Vision, Graphics & Image Processing, 2008. ICVGIP'08. Sixth Indian Conference on*, pages 722–729. IEEE, 2008.

[13] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[14] T. W. Rogers, N. Jaccard, and L. D. Griffin. A deep learning framework for the automated inspection of complex dual-energy x-ray cargo imagery. In *Anomaly Detection and Imaging with X-Rays (ADIX) II*, volume 10187, page 101870L. International Society for Optics and Photonics, 2017.

[15] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[16] L. Van Der Maaten and K. Weinberger. Stochastic triplet embedding. In *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, pages 1–6. IEEE, 2012.

[17] M. Varma and A. Zisserman. A statistical approach to texture classification from single images. *International journal of computer vision*, 62(1-2):61–81, 2005.

[18] C. Wah, S. Branson, P. Perona, and S. Belongie. Multiclass recognition and part localization with humans in the loop. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2524–2531. IEEE, 2011.

[19] C. Wah, G. Van Horn, S. Branson, S. Maji, P. Perona, and S. Belongie. Similarity comparisons for interactive fine-grained categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 859–866, 2014.

[20] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-ucsd birds 200. 2010.

[21] J. Yagnik, D. Strelow, D. A. Ross, and R.-s. Lin. The power of comparative reasoning. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2431–2438. IEEE, 2011.

[22] J. Zhang, L. Zhang, Z. Zhao, Y. Liu, J. Gu, Q. Li, and D. Zhang. Joint shape and texture based x-ray cargo image classification. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pages 266–273. IEEE, 2014.

[23] N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Part-based r-cnns for fine-grained category detection. In *European conference on computer vision*, pages 834–849. Springer, 2014.

[24] X. Zhang, H. Xiong, W. Zhou, W. Lin, and Q. Tian. Picking deep filter responses for fine-grained image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1134–1142, 2016.

[25] Y. Zhang, X.-S. Wei, J. Wu, J. Cai, J. Lu, V.-A. Nguyen, and M. N. Do. Weakly supervised fine-grained categorization with part-based image representation. *IEEE Transactions on Image Processing*, 25(4):1713–1725, 2016.