BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 03/06 pp 29–40
Volume 3, Number 1, March 2020
DOI: 10.26599/BDMA.2019.9020017

A Semi-Supervised Attention Model for Identifying Authentic Sneakers

Yang Yang, Nengjun Zhu, Yifeng Wu, Jian Cao, Dechuan Zhan*, and Hui Xiong*

Abstract: To protect consumers and those who manufacture and sell the products they enjoy, it is important to develop convenient tools to help consumers distinguish an authentic product from a counterfeit one. The advancement of deep learning techniques for fine-grained object recognition creates new possibilities for genuine product identification. In this paper, we develop a Semi-Supervised Attention (SSA) model to work in conjunction with a large-scale multiple-source dataset named YSneaker, which consists of sneakers from various brands and their authentication results, to identify authentic sneakers. Specifically, the SSA model has a self-attention structure for different images of a labeled sneaker and a novel prototypical loss is designed to exploit unlabeled data within the data structure. The model draws on the weighted average of the output feature representations, where the weights are determined by an additional shallow neural network. This allows the SSA model to focus on the most important images of a sneaker for use in identification. A unique feature of the SSA model is its ability to take advantage of unlabeled data, which can help to further minimize the intra-class variation for more discriminative feature embedding. To validate the model, we collect a large number of labeled and unlabeled sneaker images and perform extensive experimental studies. The results show that YSneaker together with the proposed SSA architecture can identify authentic sneakers with a high accuracy rate.

Key words: sneaker identification; fine-grained classification; multi-instance learning; attention mechanism

1 Introduction

The popularity of online retailing increases the importance of distinguishing between authentic and counterfeit goods. European Union statistics show that counterfeit shoes, clothes, and accessories accounted for about 10% of online shopping transaction volume in 2017, resulting in a loss of 43.3 billion Euros. Romania is losting as much as 403 million Euros from the sale of counterfeit products each year, which affects 27 000 jobs. Furthermore, research has shown that after a consumer purchases a counterfeit product, his spending on the online shopping platform will fall by a factor of almost 4. Counterfeiting is a particularly acute problem in the sneaker market, because of high profit margins and relatively easy replicability. Therefore, there are a large number of authentication requirements, which also generate large-scale data: one authentication platform generated nearly 4 TB of images in only 7 months. As shown in Fig. 1, the differences between authentic and counterfeit sneakers are often subtle and it can be extremely difficult to distinguish them. Current authentication methods are inefficient and expensive as they require specialized organizations or domain experts. Because

• Yang Yang and Dechuan Zhan are with National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China. E-mail: {yangy, zhandc}@lamda.nju.edu.cn.
• Nengjun Zhu and Jian Cao are with Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China. E-mail: {zhu_nj, cao-jian}@sjtu.edu.cn.
• Yifeng Wu is with Alibaba Company, Hangzhou 310000, China. E-mail: yixin.wyf@alibaba-inc.com.
• Hui Xiong is with Rutgers University, New York, NJ 07102, USA. E-mail: hxiong@rutgers.edu.
∗ To whom correspondence should be addressed.
Manuscript received: 2019-05-21; accepted: 2019-09-25

Fig. 1 Sampled examples in YSneaker for same class sneaker. Can you distinguish them? Answer: (a) counterfeit sneakers; (b) authentic sneakers.

of the difficulty of the task, there remain no effective semi-automatically methods to assist in authentication. Deep Convolutional Neural Networks (DCNNs) have achieved state-of-the-art results on a number of benchmarks[1–3] and are being applied to a wide range of practical applications: biomedical detection[4], recommender systems[5], intelligent agents[6], talent management[7], etc. Advancements have been made in unsupervised and semi-supervised DCNNs, with the development of variational autoencoders[8], adversarial networks[9], and semi-supervised generative adversarial networks[10]. With these demonstrations of the power of DCNN, much attention is now being directed toward challenging classification datasets, such as Fine-Grained object Recognition (FGR) tasks, with DCNN architecture being exploited to better learn discriminative image part localizations and part-based features[11, 12]. The task of sneaker authentication is similar in the need to recognize fine-grained categories, which requires the ability to handle objects with subtle inter-class difference and large intra-class variations. However, existing fine-grained labeled data is usually insufficient; for example, CUB-200-2011[13] only contains about 30 training images for each class. Furthermore, the FGR datasets mainly consist of the animal breeds (e.g., bird species[14] and dog species[15]) or objects (e.g., car models[16]) that are not easily transferable to a product identification task. More importantly, the existing fine-grained datasets only have a single source image for each instance, but in real applications it is extremely difficult to make an accurate product identification from single-source data, as Fig. 1 demonstrates, previous FGR methods also require additional information in the form of a bounding box. To the best of our knowledge, there are no public datasets especially designed for product identification tasks.

Aiming at the practical application of the sneaker authentication task, we introduce a large-scale dataset named YSneaker, which consists of different brands of authentic and counterfeit sneakers. For each sneaker extracts, there are labels at two levels, a coarse label (i.e., the sneaker category) and a fine-grained label (i.e., authentic/counterfeit), multiple image representations and a contextual description. The dataset includes 14 coarse sneaker categories, covering almost all common brands, and contains 240 625 unlabeled sneakers and 568 769 labeled sneakers from the authentication results of 7 domain experts. There are nearly 7 million images in total, which makes YSneaker much larger than previous fine-grained datasets and gives it greater practical application value.

YSneaker can be exploited for many machine learning and data mining problems: basic classification problems, including counterfeit detection and classification into categories; semi-supervised learning with unlabeled data; crowdsourcing problems that need to consider different expert capabilities and instance difficulties; online learning issues; meta-learning; transfer learning; active learning; novel class detection; anomaly detection; issues of data and computational efficiency; large-scale distribution optimization; opportunities for logical reasoning[17], etc. In this paper, to verify the application value of YSneaker, we mainly consider the identification task for authentication purposes. We implement a Semi-Supervised Attention architecture (SSA), including a self-attention structure for different images of a sneaker, and a novel prototypical loss approach for the data structure containing unlabeled data. This architecture is ideally suited to YSneaker because of the multi-source data. Specifically, different source images have different levels of importance for identification, which means that the attention mechanism will be confused if all of the multi-source images are treated equally. To solve this problem, the SSA uses a weighted average of the output feature representations, where the weights are determined by an additional shallow neural network. While the previous methods are always supervised, SSA takes advantage of unlabeled data, which can further minimize the intra-class variations for more discriminative feature detection.

In summary, the main contributions of this paper are as follows.

• We present the YSneaker dataset, which has practical applications and prospects for large-scale research projects. As well as the full dataset YSneaker is made available in a reduced size (YSneaker-small)

for model exploration;

- We propose SSA as an architecture for semi-supervised attention-based fine-grained classification tasks, which is well-suited to YSneaker;
- We comprehensively evaluate SSA against other methods and find that it achieves superior performance in real-world applications.

2 YSneaker Dataset

2.1 Details of YSneaker

The data in YSneaker is collected from HuPu, which is one of the largest and most authoritative sporting goods websites, and has professional experts to provide the specialized anthentication services. In the identification module of the website, there are 7 professional experts involved in our dataset collection. To become a professional expert needs to pass a strict examination, which includs identifying 50 pairs of sneakers from different brands correctly and supplying reasons for the judgement of authenticity. After becoming an expert, a long internship takes place, during which time the platform will randomly inspect authentication results. After the internship, site users also can challenge an expert's judgements on the platform. Consequently, it is difficult to become an expert and the identification results can generally be trusted.

Considering the fact that authentic and counterfeit sneakers are often very similar, even domain experts cannot normally identify authenticity from just a single source image (i.e., the way the sneakers are shown in Fig. 1). Therefore, in general, sneaker identification usually takes different local information of the sneakers into account, such as appearance, tag, midsole, insole, box logo, stamp, and addisional images (optional) shown in Fig. 2. To proceed with an authentication, a user first selects one of the 7 experts, then submits the above six images of the sneaker taken from specific

Fig. 2 An illustration of the YSneaker with multiple source images. Each instance includes: (a) appearance; (b) tag; (c) midsole; (d) insole; (e) box logo; (f) stamp; (g) extra images (not indispensable).

angles. Optionally, additional photos can be submitted to provide extra information. After examining the multiple source images, the domain expert will give one of three identification results: authentic, counterfeit, or uncertain. Uncertainty can occur for a number of reasons: the expert may be unfamiliar with the brand; the data provided may be incomplete; the images could be noisy, etc. Of these, the first reason is the most common. We consider sneakers given a result of uncertain to be unlabeled. Users are charged 5 RMB for each authentication, and can submit the same sneaker to different experts. When the same sneaker is submitted by the same user to a different expert, that expert cannot see the results of previous authentications.

We collected all sneakers identified by the 7 domain identification experts over a period of 7 months, which can be represented as $\{E_i\}_{i=1}^{7}$. Because of the different ability levels of the experts, each identified a different number of sneakers in this period: E_1 identified 116 606 sneakers, E_2 identified 119 449, E_3 identified 60 021, E_4 identified 137 100, E_5 identified 106 160, E_6 identified 132 680, and E_7 identified 137 378. Each sneaker example includes 6 localized source images (i.e., Figs. 2a–2f), with any additional images defined as the seventh source (i.e., Fig. 2h). A contextual description of the sneaker given by the user is also collected. Two ground truth labels are created: the "authentic/counterfeit" label as given by the experts, and the sneaker brand tagged by users from 14 options (Nike, Adidas, Jordan, Converse, Puma, Li Ning, Anta, Reebok, Under Armour, Skechers, New Balance, Kappa, Asics, or other). The raw pictures are uploaded in different sizes, so all images are resized to 448×448 for convenience. Table 1 gives a quantitative summary of the dataset.

Compared to the most widely used representative fine-grained datasets (i.e., CUB[13], Dogs[15], and Stanford Cars[16]), YSneaker covers a novel domain, and is highly associated with practical applications. It therefore presents a broad range of research prospects. YSneaker is a large-scale dataset with multiple source images, and is made available in two benchmark forms: one is of reduced size with the data from a single expert for model exploration (YSneaker-small); the other comprises of the complete authentication results from all experts (i.e., $\{E_i\}_{i=1}^{7}$). Our model is trained on the YSneaker-small dataset, on which we split the labeled set into 60% training, 10% validation, and 30% testing, with all unlabeled instances used for training.

Table 1 Dataset description. "Instance" is the number of sneakers; "Image" is the number of total images considering all sources.

Expert	Authentic		Counterfeit		Unlabel		All	
	Instance	Image	Instance	Image	Instance	Image	Instance	Image
E_1	43 892	394 926	47 514	362 477	25 200	237 259	116 606	994 662
E_2	45 647	417 633	36 488	274 141	37 314	328 691	119 449	1 020 465
E_3	33 989	296 626	10 720	80 216	15 312	129 895	60 021	506 737
E_4	43 680	378 321	48 220	373 212	45 200	390 038	137 100	1 141 571
E_5	50 240	414 378	25 000	186 829	30 920	263 286	106 160	864 493
E_6	41 440	378 149	51 120	394 387	40 120	369 165	132 680	1 141 701
E_7	45 640	445 997	45 179	352 835	46 559	411 096	137 378	1 209 928
All	304 528	2 726 030	264 241	2 024 097	240 625	2 129 430	809 394	6 879 557

2.2 Issues with the data

There are three main problems with the collected data and labels: anomalies, disorder, and overlapping examples. In regard to the first, there are a small number of anomalous items that are not sneakers (i.e., clothes, sport pants, bags, etc.), as shown in Fig. 3a. In regard to the second, the multi-source data may be incomplete or in an incorrect order, as shown in Fig. 3b. In regard to the third, a sneaker may be submitted multiple times for anthentication by the same user, thus creating duplication or overlap.

To minimize these problems, we exclude problematic items based on several rules, including the presence of certain keywords in the contextual description (i.e., clothes, bags, sport pants, etc.) and the inclusion of less than six source images. A sneaker with images uploaded in a disordered sequence will be treated as anomalous and its images denoted as duplicates. The problem of duplicate sneakers submitted by the same user is dealt with by excluding cases of repeated combinations of User ID and contextual description. This is done on the basis that users normally enter the same information into the description field when uploading the same sneaker. The number of items excluded by this preliminary data cleaning is limited. For example, E_1 still has 115 412 valid examples after these rules are applied.

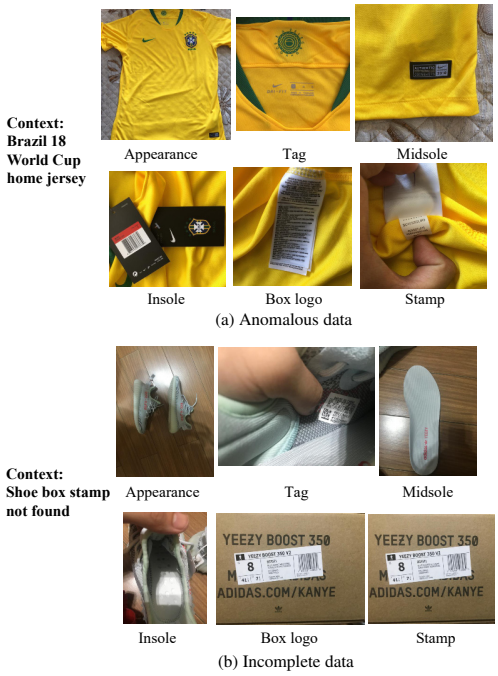


Fig. 3 Noise data: (a) Anomalous data, which can be identified as clothing from the context and (b) incomplete data, missing the stamp image, which causes the order confusion of different source images.

However, there are still some issues with the remaining data that could benefit from further research. Some outliers remain due to miscellaneous products that are not removed by the detection of keywords in the contextual information, and some images are of poor quality due to noisy backgrounds, lack of sharpness, etc. There is also the possibly of errors in the fine-grained labeling. Different experts have different identification capabilities, and even the ability level of the same expert will change over time. For the purposes of this paper, considering the strict expert assessment mechanism, user feedback mechanism, and platform supervision mechanism, we assume that the majority of labeled examples are correct.

3 Proposed Method

3.1 Notations

In this paper, without any loss of generality, it is supposed that there are N instances, including N_l examples labeled as "authenticity or counterfeit", denoted as $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{N_l}$, and N_u unlabeled instances, denoted as $\{\boldsymbol{x}_i\}_{i=N_l+1}^{N_l+N_u}$. Meanwhile, each instance has

7 different source images, where the seventh source is optional and may contain a number of images for each instance. Thus, a sneaker can be denoted by $\boldsymbol{x}_i = \{\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{iK}\}$, with a various number of instances.

3.2 Semi-supervised attention model

Accompanying the dataset, we propose SSA, an end-to-end deep network to conduct the identification task, as illustrated in Fig. 4. The motivation of SSA is to exploit the importance of different instances in the bag (sneaker), and utilize the unlabeled data to better maintain the manifold structure, which can further verify the application value of YSneaker.

Specifically, identification through the multiple-source images can be regarded as a case of Multiple Instance Learning (MIL), where the label is assigned at bag level. We propose an attention mechanism-based invariant aggregation operator, which learns the bag label probability using the neural networks to solve the MIL problem by learning the Bernoulli distribution of the bag label[18]. Meanwhile, we utilize the unlabeled data to minimize the intra-class variation, which is used to dynamically update the class center to maintain the global structure information. In conclusion, the SSA undertakes two tasks: (1) learning fine-grained feature representation with the attention-based deep network; and (2) calculating the loss with the labeled and unlabeled data.

Previous multi-instance learning pooling operators (max or mean pooling) have the clear disadvantage of being pre-defined and difficult to train. Therefore, an adaptive multi-instance pooling could potentially achieve better results by adjusting the importance of different instances. Thus, we operate the attention mechanism, as shown in Fig. 4, with a weighted average of instances, where the weights are determined by an additional neural network. $\boldsymbol{x}_i = \{\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{iK}\}$ are the K input images for the i-th sneaker, $\boldsymbol{x}_{i1}^{l_p}$ is the feature embedding learnt from the l_p-th layer by the deep network for $\boldsymbol{x}_{i1}$ (i.e., the 2048 dimension feature output for ResNet18). Thus, $\boldsymbol{x}_i^{l_p} = \{\boldsymbol{x}_{i1}^{l_p}, \boldsymbol{x}_{i2}^{l_p}, \ldots, \boldsymbol{x}_{iK}^{l_p}\}$ is the bag of the K instances with feature embedding. A weight α_k is determined for each embedding $\boldsymbol{x}_{ik}^{l_p}$ by an additional shallow neural network that is fully connected to our framework. The weights must sum to 1, which is invariant to the size of a bag. The weighted average fulfills the requirements of the fundamental theorem of symmetric functions[19]. Consequently, the attention mechanism and fine-grained classification are conducted in a unified end-to-end framework. The attention-based pooling can be represented as

$$\boldsymbol{x}_{iA}^{l_p} = \sum_{k=1}^{K} \alpha_k \boldsymbol{x}_{ik}^{l_p} \tag{1}$$

where $\boldsymbol{x}_{iA}^{l_p}$ is the bag representation, and $\alpha_k = h(\boldsymbol{x}_{ik}^{l_p}) / \sum_{j=1}^{K} h(\boldsymbol{x}_{ij}^{l_p})$, $h(\cdot)$ is the neural network for calculating the weight for each instance-level embedding, and can discover the relationships between instances.

The large amount of unlabeled data must still be dealt with, so the scenario must be extended to a semi-supervised process. The unlabeled data is handled by calculating the class center adaptively. Without any loss of generality, for Eq. (1) the semi-supervised multi-instance fine-grained loss (L) is determined as follows,

$$L = \frac{1}{N_l} \sum_{i=1}^{N_l} y_i \log(f(\boldsymbol{x}_{iA}^{l_p})) + \frac{\lambda}{N_l + N_u} \sum_{i=N_l+1}^{N_l+N_u} \hat{\ell}(\boldsymbol{x}_{iA}^{l_p}, p) \tag{2}$$

where the first term is the loss for supervised "authentic/counterfeit" identification, $f(\boldsymbol{x}_{iA}^{l_p})$ is the output prediction with weighted bag embedding. The second term is the semi-supervised prototypical loss for minimizing the intra-class variation, p is the class center representation, and $\hat{\ell}$ can be represented as

$$\hat{\ell}(\boldsymbol{x}_{iA}^{l_p}, p) = \sum_{c=1}^{2} \log \left(\frac{\exp(-\|\boldsymbol{x}_{iA}^{l_p} z_{c_i} - \hat{p}_c\|_2^2)}{\sum_j \exp(-\|\boldsymbol{x}_{jA}^{l_p} z_{c_i} - \hat{p}_c\|_2^2)} \right) \tag{3}$$

where

$$\hat{p}_c = \frac{\sum_{i=1}^{N_l} \boldsymbol{x}_{iA}^{l_p} z_{c_i} + \sum_{k=N_l+1}^{N_l+N_u} \boldsymbol{x}_{kA}^{l_p} \hat{z}_{c_k}}{\sum_{i=1}^{N_l} z_{c_i} + \sum_{k=N_l+1}^{N_l+N_u} \hat{z}_{c_k}},$$

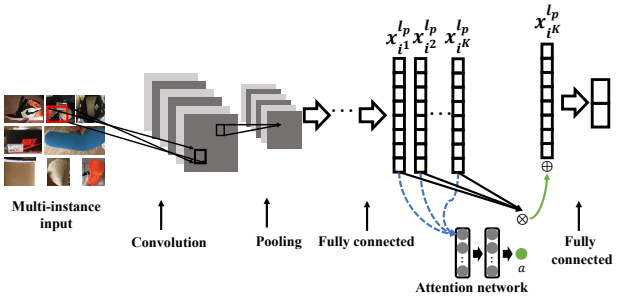


Fig. 4 Illustration of the proposed SSA. Specifically, sneaker is denoted as a bag with various number of instances. Then, SSA calculats the instance-level representations with the deep network, and utilizes additional attention-based network to get the final bag-level representation, which is used for semi-supervised fine-grained identification.

$$\hat{z}_{c_k} = \max\left(\frac{\exp(-\|\boldsymbol{x}_{k_A}^{l_p} - p_c\|_2^2)}{\sum_c \exp(-\|\boldsymbol{x}_{k_A}^{l_p} - p_c\|_2^2)}\right),$$

$$p_c = \frac{\sum_{i=1}^{N_l} \boldsymbol{x}_{i_A}^{l_p} z_{c_i}}{\sum_{i=1}^{N_l} z_{c_i}},$$

where z_{c_i} is 1 if the i-th instance belongs to the c-th identification class, otherwise it is 0, and $\hat{z}_{c_k}$ is the weight for the k-th unlabeled instance of the class with the highest probability. FGR differs from previous recognition tasks in its subtle inter-class differences and large intra-class variations, which cause difficulties for classification. In relation to these difficulties, the prototypical loss designed into SSA addresses two issues. The first term loss maximizes the inter-class distance for better classification, while the second term minimizes the intra-class differences by clustering the same class examples, noting that we utilize the unlabeled data to calculate the class center more accurately. Furthermore, SSA can be intuitively and conveniently integrated into a bilinear CNN[11], where the bilinear pooling aggregates the two-stream DCNN features for fine-grained visual recognition. In fact, $\boldsymbol{x}^{l_p}$ could utilize the bilinear CNN model in SSA, but the bilinear pooling operation cost is extremely high and makes the training phase very slow and demanding on memory resources. Surprisingly, although SSA only uses ResNet18 in the fine-grained network, the identification performance is also superior to a bilinear CNN, which validates the effectiveness of our model. In the future, we could further integrate the bilinear CNN model constrainted by low rank.

4 Related Work

Fine-grained classification: The classification task undertaken by proposed SSA is similar in spirit to FGR[11, 14, 20], which can generally be classified into two dimensions: fine-grained feature learning and discriminative part localization. To better model subtle visual differences in FGR, Ref. [11] proposed a bilinear structure to compute the pairwise feature interactions using two independent CNN, while Ref. [21] proposed to unify a CNN with a spatially-weighted representation using a Fisher vector[22]. A large number of works have proposed to leverage annotations in the form of bounding boxes and parts to localize significant regions. However, the heavy involvement of human effort makes this impractical for real-world large-scale applications. Thus, Ref. [12] proposed a novel Multi-Attention Convolutional Neural Network (MA-CNN), in which part generation and feature learning are mutually reinforcing, and Ref. [23] proposed a novel Recurrent Attention Convolutional Neural Network (RA-CNN), which recursively learns discriminative region attention and region-based feature representation at multiple scales in a mutually reinforcing manner. However, these methods come with high training demands. Meanwhile, the existing datasets for FGR are for limited tasks and have insufficient quantities of labeled data, making them impossible to transfer to the identification task or to carry out large-scale model validation. Besides, to the best of our knowledge, the state-of-the-art methods cannot effectively make use of unlabeled data.

Multi-instance learning: The previous multi-instance approaches have all utilized the mean pooling or max pooling[24–26]. These operators are non-trainable which limits their applicability. On the other hand, attention mechanisms are widely used in deep learning for computer vision[27], natural language processing[28], etc. Recent proposals have been put forward to adopt adaptively trainable multi-instance pooling. For instance, Ref. [29] proposed an attention-based multi-instance method, in which the attention weights are trained as the parameters of an auxiliary linear regression model, while Ref. [18] proposed the use of a two-layered neural network to learn the MIL, and experiments showed this approach outperforms commonly used multi-instance pooling operators. However, these methods are supervised and thus disregard the unlabeled data.

5 Experiments and Discussion

5.1 Datasets and configurations

YSneaker-small contains 14 brand categories, of which most data belong to 8 categories (Nike, Adidas, Jordan, Converse, Puma, Li Ning, New Balance, Under Armour) with other brands only adding up to 1600 instances. We therefore examine these 8 categories in our authentication experiments. DCNN is considered the state-of-the-art method for learning discriminative features. Therefore, to comprehensively evaluate YSneaker, we adopt the representative DCNN architecture, ResNet18[30], to set

baselines. The images are randomly flipped before being passed into the network, but no other data augmentation is utilized. The base learning rate is set to 0.001 and optimized with Adam. Finally, Accuracy, Precision, Recall, and F1-Measure are taken as the four criteria to measure identification performance. The parameter λ in the training phase is tuned within $\{10^{-6}, 10^{-5}, \ldots, 10^{5}, 10^{6}\}$. When the variation between the objective value of Eq. (2) is less than 10^{-5} in an iteration, we consider the SSA to be convergent. We run the following experiments on an environment implemented on NVIDIA K80 GPUs.

5.2 Alternative methods for comparison

Based on the data characteristics, we first set three baselines: CNN built with all images, MS-CNN, and MS-Bilinear. Considering that SSA is related to MIL and fine-grained classification, several multi-instance methods are used for comparison (DeepMIML[25], MI-CNN[26], MIL-Att[18], and MI-Net[31]) and two state-of-the-art fine-grained methods (Bilinear[11] and MA-CNN[12]) are also compared. These methods are all supervised, so for the purpose of comparison we train all of the networks as a regression problem, setting the confidence of unlabeled data as 0.5, and test with the labeled authentic and counterfeit instances. In detail, the compared methods are as follows.

- **CNN:** Trains a single CNN network with all source images, with the results being the ensemble of separate tests.
- **MS-CNN:** Trains a CNN network for each source image independently, with the results being the ensemble of separate tests.
- **MS-Bilinear:** Trains a bilinear network for each source image, with the results being the ensemble of separate tests.
- **DeepMIML:** Exploits a deep neural network to generate instance representation for MIML with max pooling.
- **MI-CNN:** Trains YSneaker as a multi-instance network considering label correlation with max pooling.
- **MI-Net:** A neural network that aims at solving MIL problems with 4 extensions (mi-Net, MI- Net, MI-Net-DS, and MI-Net-RC).
- **MIL-Att:** A deep attention-based multi-instance network with attention pooling.
- **Bilinear:** Models local pairwise feature interactions in a translational invariant manner, which is particularly useful for fine-grained categorization.
- **MA-CNN:** Reinforces part generation and feature learning with each other, based on convolution, channel grouping, and part classification sub-networks.

5.3 YSneaker identification results

The single source experimental results are shown in Table 2. These results validate the effectiveness

Table 2 Baselines on YSneaker-small. All results are evaluated on the test set and reported with Accuracy, Precision, Recall, and F1-Measure. The results of best performance are bolded.

(%)

Method	Accuracy							
	Source_1	Source_2	Source_3	Source_4	Source_5	Source_6	Source_7	Ensemble
CNN	72.13	78.45	78.12	77.18	82.08	73.87	73.31	**82.59**
MS-CNN	74.05	79.10	78.76	76.82	82.04	73.84	71.26	**82.36**
MS-Bilinear	74.15	80.24	79.84	78.05	**84.14**	75.06	71.46	83.57
Method	Precision							
	Source_1	Source_2	Source_3	Source_4	Source_5	Source_6	Source_7	Ensemble
CNN	70.10	76.24	76.21	74.92	79.07	71.08	75.04	**82.48**
MS-CNN	73.06	79.78	79.15	74.15	81.57	72.67	72.28	**82.98**
MS-Bilinear	73.71	79.62	79.74	76.04	**83.74**	74.23	72.51	83.33
Method	Recall							
	Source_1	Source_2	Source_3	Source_4	Source_5	Source_6	Source_7	Ensemble
CNN	79.30	83.69	82.82	77.52	84.29	75.13	**88.22**	81.13
MS-CNN	77.97	78.84	79.05	78.01	82.90	71.38	**90.39**	83.57
MS-Bilinear	76.81	82.16	80.90	78.04	82.35	72.11	**90.27**	83.27
Method	F1-Measure							
	Source_1	Source_2	Source_3	Source_4	Source_5	Source_6	Source_7	Ensemble
CNN	74.42	79.79	79.38	76.20	81.59	73.05	81.10	**81.80**
MS-CNN	75.43	79.31	79.10	76.03	82.39	72.19	80.33	**82.76**
MS-Bilinear	75.23	80.87	80.32	77.03	83.04	73.16	80.42	**83.30**

of a multi-source ensemble on the main criteria (Accuracy and F1-Measure), ensemble performance is mostly superior to single source performance. The results of comparisons with multi-instance methods are shown in Table 3. All of the methods achieve reasonable performance, which validates the reliability of YSneaker-small. These results demonstrate that several state-of-the-art multi-instance methods (MI-CNN, DeepMIML, and MIL-Att) are superior to the ensemble methods presented in Table 2, which indicates the existence of a certain degree of crossed sources (single source images may contain images from other sources). The attention-based multi-instance method (MIL-Att) is superior to the traditional untrainable methods. However, considering the presence of unlabeled data, it is still necessary to develop more advanced models for identification. To explore the effect of fine-grained methods, we conduct further experiments. We utilize the state-of-the-art fine-grained architectures (Bilinear and MA-CNN) modified as multi-instance methods with unlabeled data for comparison. SSA-Max and SSA-Mean denote SSA with max pooling or mean pooling. The results are recorded in Table 3, and reveal that SSA achieves superior performance to other fine-grained methods, and is better than max/mean pooling on most criteria except Recall, which indicates that the use of unlabeled data will further improve the performance. The details are released with the code.

5.4 Investigation of attention

Alongside the test of accurancy, we conduct a recognition performance analysis on the correctly classified sneakers of Bilinear and SSA (considering that Bilinear has the second highest performance). The attention localization is marked as the same as that in Ref. [12]. The analysis results are presented in Fig. 5. We find that SSA pays attention to more discriminative fine-grained localized features (tags and midsole) in the specified images.

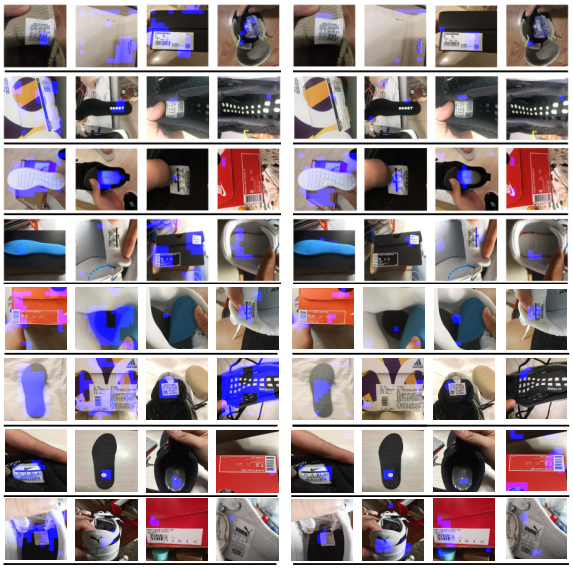
Table 3 Comparison on YSneaker-small. All results are evaluated on the test set and reported with Accuracy, Precision, Recall, and F1-Measure. The results of best performance are bolded.

(%)

Method	Accuracy	Precision	Recall	F1-measure
DeepMIML	86.95	81.40	**93.51**	87.12
MI-CNN	85.52	82.12	88.89	85.37
mi-Net	71.98	81.24	52.69	63.92
MI-Net	76.91	75.64	75.23	75.43
MI-Net-DS	74.52	79.08	62.44	69.79
MI-Net-RC	74.28	71.71	75.02	73.32
MIL-Att	87.00	83.63	91.34	87.31
Bilinear	87.01	86.94	87.48	87.21
MA-CNN	86.33	83.19	91.50	87.14
SSA-Mean	86.82	80.20	92.78	86.03
SSA-Max	**88.87**	87.41	90.12	**88.73**
SSA	88.75	**90.00**	86.64	88.29

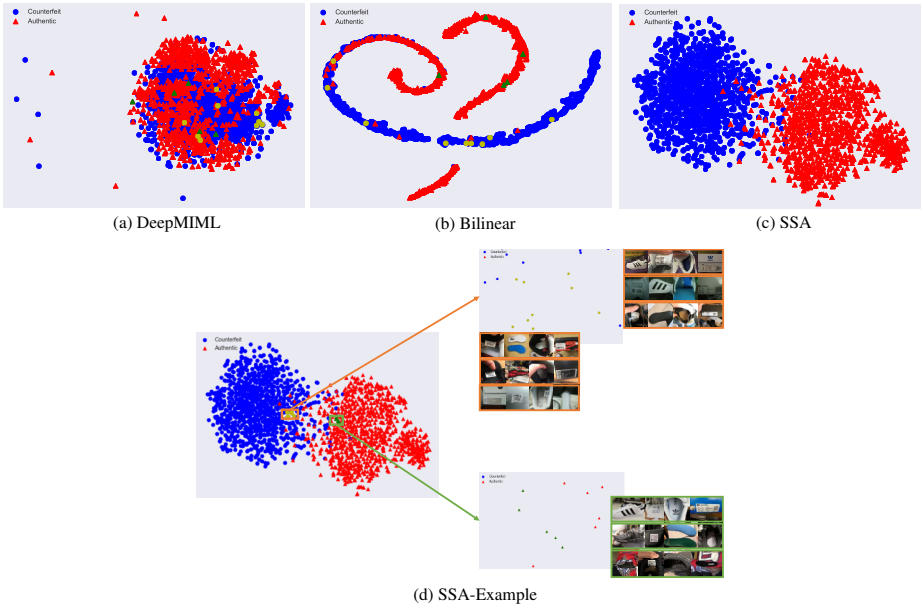
5.5 Investigation of embedding

To explore learned feature representations, we randomly select 1000 examples for each class and use t-SNE[32] to visualize the embedded feature. Figure 6 shows the projected feature maps of DeepMIML, Bilinear, and SSA for the sampled data. From Fig. 6c, it appears that instances are effectively clustered by SSA, which validates the assumption that semi-supervised information would be particularly valuable for reducing intra-class variation and enlarging inter-class differences. Furthermore, for the projected feature maps of SSA, we zoomed into two dense cluster margins of authentic and counterfeit examples, and sampled 5×5 patches to show the raw images (i.e., green points for counterfeit examples and yellow points for authentic examples) as shown in Fig. 6d. This reveals that SSA reduces the intra-class variation, as demonstrated by its capacity to cluster the same class of counterfeit sneakers, while enlarging the inter-class difference, as demonstrated by its capacity to correctly distinguish authentic and counterfeit sneakers even



(a) Bilinear (b) SSA

Fig. 5 An illustration of the local attention learning. The attention localization is marked with blue shadow.



(a) DeepMIML (b) Bilinear (c) SSA

(d) SSA-Example

Fig. 6 t-SNE visualisation of the sampled data for (a) DeepMIML, (b) Bilinear, (c) SSA, and (d) SSA-Example, where each point in the patch corresponds to a sneaker. Each instance (sneaker) is with 2048-dimension vector, then they are projected by t-SNE to two dimensions. For SSA, we have zoomed into two dense cluster margins of authentic (marked with yellow) and counterfeit (marked with green) examples, and sampled 5×5 patches to show the raw images. The same examples are also displayed for DeepMIML and Bilinear.

of the same class. However, Figs. 6a and 6b reveal that DeepMIML and Bilinear have more difficulty to distinguish the same examples.

5.6 Investigation of convergence

To investigate the convergence of SSA iterations empirically, the objective function value (i.e., the value of Eq. (2)) and the classification performance of SSA in each iteration are recorded in Fig. 7. This clearly reveals that the objective function value decreases as the iterations increase, and the classification performance is stable after several iterations. Moreover, these additional experimental results indicate that SSA converges very fast after only 10 epochs.

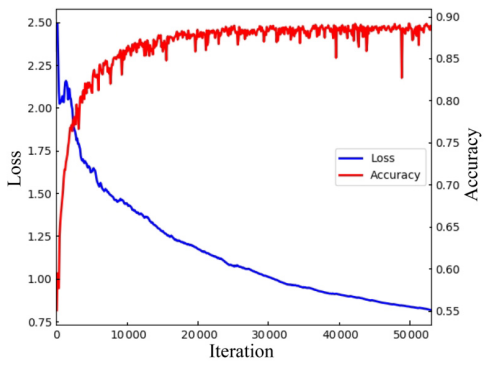


Fig. 7 Objective function value convergence and corresponding classification accuracy vs. number of iterations of SSA.

5.7 Parameter stability

To explore the influence of parameter λ, we tune λ in the range of $\{10^{-6}, 10^{-5}, \ldots, 10^{6}\}$, and record the average performance in Fig. 8. The results show that SSA achieves increasing performance as λ decreases, which reflects the trend of using unlabeled data. Unlabeled data presents more difficulties for classification, so increasing λ will decrease performance.

6 Conclusion

We introduce the large-scale YSneaker dataset in the hope of encouraging machine learning research on a practical, difficult, and important dataset.

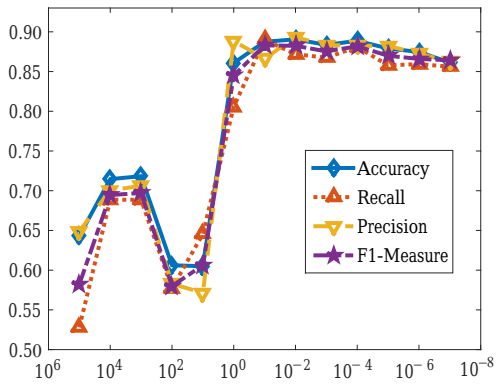


Fig. 8 Influence of the parameters λ.

Accompanying this dataset, we provide a benchmark method (SSA) for comparison. SSA adopts a self-attention-based semi-supervised multi-instance CNN architecture for product identification, which can exploit the instance correlation and both the inter-class differences and intra-class variations simultaneously. The experimental results indicate the promise of SSA, as sneaker authentication shifts from manual identification to semi-automatic identification. Only a fraction of the data is used in this paper, and further interpretation and a more robust architecture could be expected to result in better performance and improved understanding of the data. YSneaker is made available to encourage further advanced research into many other machine learning topics.

Acknowledgment

This research was supported by the National Key R&D Program of China (No. 2018YFB1004300), the National Natural Science Foundation of China (Nos. 61773198, 61632004, and 61751306), the National Natural Science Foundation of China-Korea Research Foundation Joint Research Project (No. 61861146001), Collaborative Innovation Center of Novel Software Technology and Industrialization, and Postgraduate Research & Practice Innovation Program of Jiangsu Province (No. KYCX18-0045).

References

[1] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, Densely connected convolutional networks, in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 2261–2269.

[2] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollar, Focal loss for dense object detection, in *Proceedings of the International Conference on Computer Vision*, Venice, Italy, 2017, pp. 2999–3007.

[3] K. He, G. Gkioxari, P. Dollar, and R. B. Girshick, Mask R-CNN, in *Proceedings of the International Conference on Computer Vision*, Venice, Italy, 2017, pp. 2980–2988.

[4] O. Ronneberger, P. Fischer, and T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *Proc. Medical Image Computing and Computer-Assisted Intervention*, Munich, Germany, 2015, pp. 234–241.

[5] J. Lian, X. Zhou, F. Zhang, Z. Chen, X. Xie, and G. Sun, xdeepfm: Combining explicit and implicit feature interactions for recommender systems, in *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, London, UK, 2018, pp. 1754–1763.

[6] S. Wang, L. He, B. Cao, C. Lu, P. S. Yu, and A. B. Ragin, Structural deep brain network mining, in *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, Halifax, Canada, 2017, pp. 475–484.

[7] H. Xu, Z. Yu, J. Yang, H. Xiong, and H. Zhu, Dynamic talent flow analysis with deep sequence prediction modeling, *Transactions on Knowledge and Data Engineering*, vol. 31, no. 10, pp. 1926–1939, 2019.

[8] D. P. Kingma and M. Welling, Auto-encoding variational bayes, in *Proceedings of the International Conference on Learning Representations*, Banff, Canada, 2014, pp. 34–42.

[9] Y. Li and J. Ye, Learning adversarial networks for semi-supervised text classification via policy gradient, in *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, London, UK, 2018, pp. 1715–1723.

[10] K. G. Dizaji, X. Wang, and H. Huang, Semi-supervised generative adversarial network for gene expression inference, in *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, London, UK, 2018, pp. 1435–1444.

[11] T. Lin, A. Roy Chowdhury, and S. Maji, Bilinear CNN models for fine-grained visual recognition, in *Proceedings of the International Conference on Computer Vision*, Santiago, Chile, 2015, pp. 1449–1457.

[12] H. Zheng, J. Fu, T. Mei, and J. Luo, Learning multi-attention convolutional neural network for fine-grained image recognition, in *Proceedings of the International Conference on Computer Vision*, Venice, Italy, 2017, pp. 5219–5227.

[13] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, The Caltech-UCSD Birds-200-2011 Dataset, Report, California Institute of Technology, CA, USA, 2011.

[14] X. Zhang, H. Xiong, W. Zhou, W. Lin, and Q. Tian, Picking deep filter responses for fine-grained image recognition, in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 1134–1142.

[15] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, Novel dataset for fine-grained image categorization: Stanford dogs, in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, 2011, p. 1.

[16] J. Krause, H. Jin, J. Yang, and F. Li, Fine-grained recognition without part annotations, in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 5546–5555.

[17] Z.-H. Zhou, Abductive learning: Towards bridging machine learning and logical reasoning, *Science China Information Sciences*, vol. 62, no. 7, pp. 76 101:1–76 101:3, 2019.

[18] M. Ilse, J. M. Tomczak, and M. Welling, Attention-based deep multiple instance learning, in *Proceedings of the International Conference on Machine Learning*, Stockholm, Sweden, 2018, pp. 2132–2141.

[19] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola, Deep sets, in *Proc. of Advances in Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 3394–3404.

[20] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, Compact bilinear pooling, in *Proceedings of the International Conference on Computer Vision*, Las Vegas, NV, USA, 2016, pp. 317–326.
[21] N. Zhang, J. Donahue, R. B. Girshick, and T. Darrell, Part-based R-CNNs for fine-grained category detection, in *Proceedings of the European Conference on Computer Vision*, Zurich, Switzerland, 2014, pp. 834–849.
[22] F. Perronnin and D. Larlus, Fisher vectors meet neural networks: A hybrid classification architecture, in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 3743–3752.
[23] J. Fu, H. Zheng, and T. Mei, Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition, in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 4476–4484.
[24] P. H. O. Pinheiro and R. Collobert, From image-level to pixel-level labeling with convolutional networks, in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 1713–1721.
[25] J. Feng and Z. Zhou, Deep MIML network, in *Proceedings of the AAAI Conference on Artificial Intelligence*, San Francisco, CA, USA, 2017, pp. 1884–1890.
[26] Y. Yang, Y. Wu, D. Zhan, Z. Liu, and Y. Jiang, Complex object classification: A multi-modal multi-instance multi-label deep network with optimal transport, in *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, London, UK, 2018, pp. 2594–2603.
[27] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in *Proceedings of the International Conference on Machine Learning*, Lille, France, 2015, pp. 2048–2057.
[28] H. Li, M. R. Min, Y. Ge, and A. Kadav, A context-aware attention network for interactive question answering, in *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, Halifax, Canada, 2017, pp. 927–935.
[29] N. Pappas and A. Popescu-Belis, Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, 2014, pp. 455–466.
[30] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 770–778.
[31] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, Revisiting multiple instance neural networks, *Pattern Recognition*, vol. 74, pp. 15–24, 2018.
[32] L. v. d. Maaten and G. Hinton, Visualizing data using t-SNE, *Journal of Machine Learning Research*, vol. 9, no. 11, pp. 2579–2605, 2008.

Yang Yang is working towards the PhD degree with the National Key Lab for Novel Software Technology, the Department of Computer Science & Technology in Nanjing University, China. His research interests lie primarily in machine learning and data mining, including heterogeneous learning, model reuse, and incremental mining.

Jian Cao received the PhD degree from Nanjing University of Science and Technology, Nanjing University. He is currently a professor with Department of Computer Science and Engineering at Shanghai Jiao Tong University (SJTU), China, and the deputy head of the Department. He is the director of the SJTU & Morgan Stanley Joint Research Center on Financial Service Innovation. He is also the leader of the Lab for Collaborative Intelligent Technology. Dr. Cao's research interests include network computing, service computing, and data analytics. He leads the research group of collaborative information system. He has authored or co-authored over 180 journal and conference papers in the above areas.

Nengjun Zhu is currently a PhD candidate in the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University (SJTU), Shanghai, China. His research interests include recommender system and data mining.

Yifeng Wu received the MS degree from Nanjing University, Nanjing, China, in 2018. He is working in Alibaba Company, Hangzhou, China. His research interests lie primarily in machine learning and data mining, including multi-modal learning.

Dechuan Zhan received the PhD degree from Nanjing University, China, in 2010. At the same year, he became a faculty member of the Department of Computer Science and Technology, Nanjing University, China. He is currently an associate professor with the Department of Computer Science and Technology at Nanjing University. His research interests are mainly in machine

learning, data mining, and mobile intelligence. He has published over 20 papers in leading international journals/conferences. He serves as an editorial board member of IDA and IJAPR, and serves as SPC/PC in leading conferences such as IJCAI, AAAI, ICML, NIPS, etc.

Hui Xiong received the PhD from the University of Minnesota-Twin Cities, MN, USA, in 2005. He is currently a full professor at Rutgers, the State University of New Jersey, where he received the ICDM-2011 Best Research Paper Award, and the 2017 IEEE ICDM Outstanding Service Award. His general area of research is data and knowledge engineering, with a focus on developing effective and efficient data analysis techniques for emerging data intensive applications. He has published prolifically in refereed journals and conference proceedings (4 books, more than 80 journal papers, and more than 100 conference papers). He is a co-editor-in-chief of *Encyclopedia of GIS*, an associate editor of the *IEEE Transactions on Knowledge and Data Engineering*, the *IEEE Transactions on Big Data*, the *ACM Transactions on Knowledge Discovery from Data*, and the *ACM Transactions on Management Information Systems*. He has served regularly on the organization and program committees of numerous conferences, including as a program cochair of the Industrial and Government Track for KDD-2012, a program co-chair for ICDM-2013, a general co-chair for ICDM-2015, and a program co-chair of the Research Track for KDD-2018. For his outstanding contributions to data mining and mobile computing, he was elected an ACM distinguished scientist in 2014. He is a fellow of the IEEE.