INFERRING THE IMPORTANCE OF PRODUCT APPEARANCE: A STEP TOWARDS THE SCREENLESS REVOLUTION

A PREPRINT

Yongshun Gong
JD AI Research and
University of Technology, Sydney
gongyongshun@jd.com

Jinfeng Yi
JD AI Research
yijinfeng@jd.com

Dong-Dong Chen
JD AI Research and
National Key Laboratory for
Novel Software Technology,
Nanjing University
chendd@lamda.nju.edu.cn

Jian Zhang
University of Technology, Sydney
jian.zhang@uts.edu.au

Jiayu Zhou
Michigan State University
jiayu@cse.msu.edu

Zhi-Hua Zhou
National Key Laboratory for
Novel Software Technology,
Nanjing University
zhouzh@lamda.nju.edu.cn

May 10, 2019

ABSTRACT

Nowadays, almost all the online orders were placed through screened devices such as mobile phones, tablets, and computers. With the rapid development of the Internet of Things (IoT) and smart appliances, more and more screenless smart devices, e.g., smart speaker and smart refrigerator, appear in our daily lives. They open up new means of interaction and may provide an excellent opportunity to reach new customers and increase sales. However, not all the items are suitable for screenless shopping, since some items’ appearance play an important role in consumer decision making. Typical examples include clothes, dolls, bags, and shoes. In this paper, we aim to infer the significance of every item’s appearance in consumer decision making and identify the group of items that are suitable for screenless shopping. Specifically, we formulate the problem as a classification task that predicts if an item’s appearance has a significant impact on people’s purchase behavior. To solve this problem, we extract features from three different views, namely items’ intrinsic properties, items’ images, and users’ comments, and collect a set of necessary labels via crowdsourcing. We then propose an iterative semi-supervised learning framework with three carefully designed loss functions. We conduct extensive experiments on a real-world transaction dataset collected from the online retail giant JD.com. Experimental results verify the effectiveness of the proposed method.

Keywords Screenless Retail · Multi-view Learning · Semi-supervised Learning

1 Introduction

As China’s largest online retailer and the third largest internet company globally in terms of revenue and market capitalization, JD.com serves over 300 million active customers and fulfills billions of orders per year. Despite these remarkable numbers, almost all the orders were placed through devices with an integrated screen, such as mobile phones, tablets, and computers. Similar situations also occur in other e-commerce giants such as Amazon and Alibaba.

https://www.jd.com/
https://www.amazon.com/
https://www.alibaba.com/
Gold Pendants

Running shoes

Mobile phone shell

Services

Books

Figure 1: Some typical examples of strong appearance-related (SA) items and weak appearance-related (WA) items. The WA items are suitable for screenless shopping but SA items are not.

Figure 1 demonstrates that the growth of the existing e-commerce model is strictly limited by the number of screen devices worldwide.

Today, we are witnessing the dawn of a new era of the Internet of Things (IoT). The statistic shows that the number of connected IoT devices worldwide will jump about 12 percent on average annually, from nearly 27 billion in 2017 to 125 billion in 2030[^4]. With the rapid development of IoT and smart appliances, more and more screenless smart devices appear in our daily lives, such as the smart speaker, smart refrigerator, and smart washer. They provide new means of interaction and may offer an opportunity to find new customers and increase sales [1, 2]. For example, a smart refrigerator can warn users or even place an order for them when running out of milk. Likewise, the smart speaker can learn users’ preference through voice conversations and make recommendations accordingly. However, not all the goods are suitable for screenless shopping, as their visual aspects are almost completely missing in such a scenario. For example, items like clothes, dolls, bags, and shoes should not be recommended in a screenless environment, since their appearances play an essential role in consumer decision making. Without seeing the item, users cannot firmly decide whether they like it or not. On the other hand, items like food, books, and virtual goods are more appropriate choices for screenless shopping since their products appearance only plays a minor role in consumer choice. Therefore, a key step towards screenless retail is to infer the significance of every item’s appearance in consumer decision making, which is the main focus of this paper. Specifically, we would like to distinguish between two classes of items: (i) items whose appearance plays an important role in consumer decision making, thus are inappropriate for screenless shopping, and (ii) items whose appearance has only negligible influence on people’s purchase behavior, thus are suitable for screenless shopping. We refer to them as strong appearance-related items (SA items for short), and weak appearance-related items (WA items for short), respectively. Figure 1 shows some examples of strong and weak appearance-related items.

To the best of our knowledge, none of the existing techniques concerns on identifying SA and WA items yet. It is a novel problem and a real commercial demand that faces several major challenges:

- **Limited Data**: In the real-world scenarios, almost all the orders were placed on screened devices. Therefore, how people behave in a screenless environment is largely unknown.

[^4]: [https://cdn.ihs.com/www/pdf/IoT_ebook.pdf](https://cdn.ihs.com/www/pdf/IoT_ebook.pdf)
Figure 2: Some examples of fine-grained categories studied in this paper.

- **Unlabeled Samples:** To be as the first work on studying the problem of distinguishing strong and weak appearance-related items in the e-commerce industry, there is thus no labeled information for any items.

- **Scalability:** JD.com has over 30 million sold items. This requires the proposed method to be highly efficient and scalable.

- **Cold Start:** At JD.com, a large number of new products are continually added. These new items do not have historical comment data which leads to the incomplete view problem, thus are more challenging to predict [3, 4, 5].

In this paper, we cast the problem as a two-class classification problem by utilizing three different views from items’ intrinsic features, items’ images, and users’ comments. To solve this problem, we propose an iterative semi-supervised learning framework with three carefully designed loss functions. Among them, the first one is the widely-used softmax loss [6] that penalizes the classification error on the labeled data. The second loss, named view alignment loss, not only aims to map the comment and image views into a similar feature space, but also handles the cold-start problem when the incomplete view occurs. The last one is a modified triplet loss function that utilizes the pseudo-labeled set to improve the performance of semi-supervised learning task. To effectively combine these loss functions to achieve the best overall performance, we propose a well-designed training process that organically integrates them. Finally, we note that this study is not only crucial in finding new customers and increase sales, but also providing excellent value for a product design process. After all, if the appearance of a product only plays a minor role in consumer choice, then its manufactory should make more efforts on improving the quality of products and services, rather than its appearance design. In contrast, if the appearance attributes of a product have an enormous influence on people’s purchase behavior, then the manufactory should invest more in product designs for increasing its overall attractiveness.

The rest of this paper is organized as follows. In Section 2 we present the multi-view pre-processing module. The proposed semi-supervised classification method is detailed in Section 3. The experimental results are presented in Section 4. Section 5 discusses the related work. Section 6 concludes the paper with future work.

### 2 Label Annotation and Feature Extraction

To the best of our knowledge, no prior work has attempted to tackle the problem of differentiating strong and weak appearance-related items, so we need to create a way to collect necessary labels and extract relevant features. We first collect manual annotations via crowdsourcing to capture the human perception of the importance of items’ appearance. Since items in the same category may have similar importance in the appearance, we label item categories instead of items for improving efficiency and reducing cost. To meet this requirement, we need to select the granularity of categories properly. For example, the category “Smart TV” is a better choice than the category “Electrical Equipment”, since the latter category covers a broad range of items with different appearance importance. To this end, we choose...
Table 1: Some typical examples of categories’ information

| Fine-grained category | Item ID        | Higher-level category | Basic features | Comments                                                                                           | Image for each item |
|-----------------------|----------------|-----------------------|----------------|---------------------------------------------------------------------------------------------------|---------------------|
| Electric cooker       | X990XX;        | Domestic appliance    | length, width  | (1) Buy it to my parents, the soup tastes so good; (2) The shipping is very fast, and packaging is also perfect, but unfortunately the lid is not made from glass; (3) Good service and nice soup; |                     |
|                       | X921XX;        |                       |                |                                                                                                   |                     |
|                       | ...            |                       |                |                                                                                                   |                     |
| Sports pants – man    | X544XX;        | Outdoor sport         | length, width  | (1) Fine workmanship, soft fabrics, comfortable to wear, and match everything; (2) The actual color looks more graded than the picture, but with a little bit of smell. Overall, it’s OK; |                     |
|                       | X523XX;        |                       |                |                                                                                                   |                     |
|                       | ...            |                       |                |                                                                                                   |                     |
| Children puree        | X982XX;        | Mother & baby         | length, width  | (1) Baby loves it, tastes very good. We always buy them in JD.com; (2) Not bad. Strawberry tastes better, but it is a little expensive; |                     |
|                       | X931XX;        |                       |                |                                                                                                   |                     |
|                       | ...            |                       |                |                                                                                                   |                     |

the lowest level (most fine-grained) category provided by JD.com as each item’s category, and assume that the items within the same category belong to the same appearance-related class. In the rest of this paper, we will abbreviate the term ‘the lowest level category’ as ‘the category’. Among all these categories, we randomly select 2, 707 categories that contain hundreds of thousands of items for crowd labeling. Figure 2 reveals some examples of such fine-grained categories, and a piece of data is shown in Figure 1. To collect the labels of SA and WA categories, 67 JD employees were requested to serve as the crowd workers. For each worker, we randomly selected 200 categories, and asked the worker to label their classes. On average, each category was labeled by five different workers, and its final label was decided by majority voting. In our work, only a small subset of labels are used in the training procedure, and all the other labels are used for evaluation.

To accurately classify strong and weak appearance-related categories, it is important to consider as much information as we can. Generally speaking, three aspects of categories’ information can be collected in the real-world scenario, including (i) categories’ intrinsic features such as their higher level categories and other basic attributes; (ii) categories’ images that contain their visual information, and (iii) comments information that expresses users’ opinions for the categories. Some typical examples of categories’ information are shown in Table 1. In the following, we introduce how to process the information and extract the features of the three views:

- **Intrinsic Feature View** This view contains the information about items’ intrinsic properties, such as their higher level categories and other basic attributes such as the average sizes of the items in each category. The higher level categorical information is one-hot encoded, while the other attributes are normalized between 0 and 1, as shown in Figure 3(a).

- **Comment Feature View** Users’ comments can usually provide valuable information on predicting the importance of product appearance. For example, a comment like “this shirt looks perfect, and wears comfortable at a rock-bottom price” suggests that the shirt is attractive since it “looks perfect”, thus it should be a strong appearance-related item and not suitable for screenless shopping.

For each category, we extract two kinds of features from users’ comments, and concatenate them into a longer one, as shown in Figure 3(b). The first kind uses a simple sum of all the comments words’ embedding learned via Word2Vec method. To extract the second kind of features, we build a keywords dictionary that contains 329 highly correlated words to product appearance. We then create a 329-dimensional normalized keyword matching vector for each category. Specifically, the normalized matching keywords vector of the $i$-th category, denoted as $vec_i$, is computed by

$$vec_i = \frac{TF_i}{Sum_i},$$

(1)

---

5 We select a subset of categories due to two reasons: (i) the total number of JD’s lowest level categories is confidential, and (ii) manually labeling all the categories are very difficult, if not impossible.

6 We have also tried other embedding methods such as Doc2Vec and Paragraph2Vec, and Word2Vec yields the overall best performance.
(a) The Intrinsic Feature View. It contains the higher level category information in one-hot encoding, and concatenate with the normalized basic attributes.

(b) The Comment Feature View. The left part is embedded by word2vec, and the right part is generated by matching the keywords dictionary.

Figure 3: Pre-Processing of the intrinsic feature view and the comment feature view.

where $TF_i$ indicates the $i$-th category’s keywords frequency vector, and $Sum_i$ is the total number of comment words of the $i$-th category.

- **Image Feature View** This view aims to extract a unified image embedding for all the fine-grained categories. For all the images in the same category, we first learn their image embeddings via ResNet-50 [8], and then use the mean of these embeddings as the features extracted from the image view.

3 Methodology

3.1 Problem Definition and Framework

Before presenting our method, we first introduce some basic notations. Given the three views, the $l$-th labeled category is denoted as $x^l = (x^l_1, x^l_2, x^l_3) \in \mathbb{R}^{m_1+m_2+m_3}$, where $x^l_v \in \mathbb{R}^{m_v}$ represents the $v$-th view ($v = 1, 2, 3$). The entire labeled categories are denoted as $\mathcal{X}^l = \{ x^l | l = 1, 2, \ldots, L \}$, corresponding with their labels $y^l = \{ y^l_l | l = 1, 2, \ldots, L \}$ where $y^l_l = 0$ denotes the SA category and $y^l_l = 1$ denotes the WA category. Besides, $\mathcal{X}^l$ can be segmented into three subsets $\mathcal{X}^l_v$, where $v$ stands for the $v$-th view of $\mathcal{X}^l$. For unlabeled data, $\mathcal{X}^u = \{ x^u_u | u = 1, 2, \ldots, U \}$ denotes $U$
unlabeled categories (U ≫ L), where \( x^u = (x^u_1, x^u_2, x^u_3) \in \mathbb{R}^{m_1 + m_2 + m_3} \) denotes an unlabeled category which is also combined by three views \( x^u_v \). The \( v \)-th view of \( X^u \) is denoted as \( X^u_v \). In this paper, we utilize view 1, view 2 and view 3 to represent the intrinsic feature view, the comment view, and the image view respectively. Our goal is to learn a model from the training set \( \{ X_l, Y_l \} \cup X^u \) to refer the SA/WA categories.

Here we briefly introduce an overview of our training procedure, with the details presented in the next subsection. Our training process in each epoch is divided into two phases. In the first phase, all labeled categories are trained by two loss functions: a softmax loss, denoted by \( L_{SM} \), and a view alignment loss \([9, 10, 11, 12]\), denoted by \( L_{VA} \). We use the view alignment loss in our task to map the comment and the image feature views into a close enough feature space. Besides, this loss function can help to solve the cold start problem with missing views. In this phase, we train the model by minimizing \( L_{phase1} \) as follows:

\[
L_{phase1} = L_{SM}(X_l, Y_l) + \alpha L_{VA}(X^u_2, X^u_3) \tag{2}
\]

where parameter \( \alpha > 0 \) is introduced to balance the two losses.

At the beginning of the second phase, the model trained based on all the labeled categories is used to provide pseudo-labels for all the unlabeled categories \( X^u \). Since pseudo-labels may not be accurate, especially at the beginning of the training procedure, we use the most reliable predictions with the highest confidence scores for creating pseudo-labeled triplet constraint \( I^u = < x^u, x^u(p), x^u(n) > \). This constraint means that \( x^u \) and \( x^u(p) \) are likely to be in the same category, while \( x^u \) and \( x^u(n) \) are not. Then, we introduce a pseudo-labeled triplet loss \( L_{PT} \) to narrow the distances between \( < g_W(x^u), g_W(x^u(p)) > \) while enlarge the distances between \( < g_W(x^u), g_W(x^u(n)) > \), where \( g_W(\cdot) \) is a non-linear mapping function that needs to be learned.

Note that the comment and the image feature views of the unlabeled categories also need to be aligned, the second phase simultaneously optimizes the following function:

\[
L_{phase2} = \alpha L_{VA}(X^u_2, X^u_3) + \beta L_{PT}(X^u) \tag{3}
\]

We inherit the parameter \( \alpha \) used in the first phase and fine-tune the weight \( \beta \) of \( L_{PT} \). A detailed parameter analysis is listed at Section 4.7. The proposed network architecture and the algorithm of the training process are shown in Figure 4 and Algorithm 1, respectively.

3.2 The Loss Functions

- **Softmax Loss** The standard softmax loss is used for the labeled set in the first phase of the training process. It is formulated as

\[
L_{SM}(X_l, Y_l) = \frac{1}{L} \left( \sum_{l=1}^{L} -[y^l \log p^l + (1 - y^l) \log(1 - p^l)] \right), \tag{4}
\]
Algorithm 1: The proposed network

Multi-view pre-processing to build fine-grained categories;

Input:
The multi-view labeled set \( \{ \mathcal{X}^l, \mathcal{Y}^l \} \) and unlabeled set \( \mathcal{X}^u \);
Balance parameters \( \alpha, \beta, \) the view alignment margin \( \lambda \), and the triplet margin \( \gamma \);
The learning rate \( \eta \);

Output:
The learned model;

1. **Training:**
   1. for \( t = 1 \rightarrow T \) do
   2. for labeled set \( \{ \mathcal{X}^l, \mathcal{Y}^l \} \) do
   3. Optimize the model by minimizing Equation 2
   4. end
   5. Provide pseudo-labels by the strategy in Section 3.2
   6. for unlabeled set \( \mathcal{X}^u \) do
   7. Optimize the model by minimizing Equation 3
   8. end
   9. end

where \( y^l \) is the annotated label, and \( p^l \) is the predicted probability of \( x^l \).

- **View Alignment Loss** The initial idea of the view alignment loss is to project different views into a unified latent subspace. Some relevant studies focus on seeking hierarchical nonlinear mappings in deep learning process as follows:

\[
f(x) = h^{(M)} = s(W^{(M)}h^{(M-1)} + b^{(M)}),
\]

where \( W^{(M)} \) is a projection matrix to be learned in the previous layer \( h^{(M-1)} \); \( b^{(M)} \) is a bias vector at the \( M \)-th layer; \( s(\cdot) \) denotes a nonlinear activation function, e.g., the \text{relu} or \text{tanh} function; and \( f(\cdot) \) is the non-linear mapping function determined by the parameters \( W^{(m)} \) and \( b^{(m)} \) \((m = 1, 2, 3, ..., M)\) [10].

In our task, we note that the comment and image feature views may not be able to perfectly aligned in a same subspace, since people’s comments are usually subjective and may contain the information not covered by the images. In this case, we instead assume that the learned subspaces of the two views are close enough. This leads to the view alignment loss for the labeled data:

\[
L_{VA}(X^l_2, X^l_3) = \frac{1}{L} \sum_{i=1}^{L} \max \{ d^2_j(x^i_2, x^i_3) - \lambda, 0 \},
\]

where

\[
d^2_j(x^i_2, x^i_3) = ||f(x^i_2) - f(x^i_3)||_2^2
\]

and

\[
f(x^i_v) = h_v^{(M)} = s(W_v^{(M)}h_v^{(M-1)} + b_v^{(M)}).
\]

Likewise, the view alignment loss for the unlabeled data can be formulated as:

\[
L_{VA}(X^u_2, X^u_3) = \frac{1}{U} \sum_{i=1}^{U} \max \{ d^2_j(x^i_2, x^i_3) - \lambda, 0 \}
\]

Another advantage of introducing the view alignment loss is to address the cold-start problem with missing views. Since the view alignment loss can project the comment and image features to close enough subspaces, the lack of one view can be approximately replaced by the other view. Figure 6 illustrates the framework of the view alignment loss, and its performance is reported in Section 4.

- **Pseudo-labeled Triplet Loss** To utilize the information of unlabeled data, we introduce a metric learning-based loss named pseudo-labeled triplet loss. After the first training phase, we use the model trained by the labeled data to annotate the unlabeled categories, and denote these annotations as pseudo-labels. The pseudo-labels may exist some errors, thus we only use the most reliable pseudo-labels with the highest confidence scores to assist the training procedure. Given a anchor category \( x^u \), we search the most confident pseudo-labels within a same mini-batch, and generate the triplet constraint using the most reliable positive and negative categories, denoted as \( x^{u(p)} \) and \( x^{u(n)} \), respectively. Figure 7
This shirt looks perfect, and wears comfortable at a rock-bottom price.

Figure 6: The flowchart of the view alignment loss

shows one such example, where the second and the fifth categories are marked as the positive and negative samples, respectively, since their probabilities are the highest. Notably, we use triplet constraints here but not the pairwise constraints, since the triplet constraints only enforce positive examples to be closer than negative examples, while the pairwise constraint devotes to gather all positive examples as close together as possible [13]. Therefore, the triplet constraint enjoys the flexibility to adapt to different levels of intra-class variance for different classes and is more robust to some inaccurate pseudo-labels.

Given a triplet constraint $I^u = \langle x^u, x^{u(p)}, x^{u(n)} \rangle$, we define the score of this triplet as

$$d_g(I^u) = ||g(x^u) - g(x^{u(p)})||^2_2 - ||g(x^u) - g(x^{u(n)})||^2_2 + \gamma,$$

where $g(\cdot)$ is a non-linear mapping function that needs to be learned, and $\gamma$ is a positive margin parameter.

By considering all the triplet constraints, the pseudo-labeled triplet loss function is defined as:

$$L_{PT}(X^u) = \frac{1}{U} \sum_{u=1}^{U} (\max \{ d_g(I^u), 0 \})$$

(11)
4 Experiments

In this section, we report the results of extensive experiments conducted on a large-scale, real-world dataset provided by JD.com.

4.1 Dataset

The dataset contains hundreds of thousands of items that belong to 2,707 fine-grained categories. Each item has one image, and the dataset contains over 1 million user comments in total. After the feature extraction step, we generate three feature representations from the intrinsic, comment, and image views, with their dimensions equaling to 412, 629, and 1000, respectively. The 2,707 category labels were annotated via crowdsourcing. Among them, 70% shuffled categories are used as the training set, while the others are used for evaluation. We randomly pick 30% data of training set as the initial labeled dataset \( \{X_l, Y_l\} \). We evaluate all the performances by Accuracy (%) since the collected data are balanced in label distribution.

4.2 Experimental Setup & Baselines

We design a network with seven fully connected layers as shown in Figure 4. Batch normalization (BN) is applied after each layer before the Leaky-ReLU (0.1) activation. The network is trained using Adam with the exponential decay rates for the first moment estimates 0.9, the second moment estimates 0.999, and an initial learning rate of 0.005 which is divided by 5 after epochs 150, 300, and 450 (600 epochs in total). The mini-batch size is 64. We also introduce dropout \((p = 0.5)\) after the concatenated layer. The margin parameters \( \lambda \) and \( \gamma \) of \( L_{VA} \) and \( L_{PT} \) are set to 0.01 and 0.5 through the experiments. The \( \alpha \) and \( \beta \) weights in \( L_{VA} \) and \( L_{PT} \) are set to 10 and 0.1, respectively.

**Baselines.** Since our proposed method is, to the best of our knowledge, the first algorithm that can infer the importance of product appearance and distinguish SA/WA items, there is no direct baseline for comparison. In this case, we compare our method with two baseline algorithms from clustering and semi-supervised learning perspectives. All parameters of the first two baselines are optimized by the grid search method. Furthermore, two variants of the proposed method are also involved for a comparison. These baseline algorithms are:

- **MLAN_C:** Multi-view learning with adaptive neighbors for clustering, a state-of-the-art unsupervised method to solve the multi-view problem [14]. It utilizes a local structure with adaptive neighbors to learn a graph-based representation, and iteratively updates the clustering latent space by the similarity graph.
Table 2: Test accuracies (%) compared with the first three baselines. Mean and standard deviation of 5 random runs are reported and the best result is in bold.

| Methods | MLAN C | MLAN SC | SWA(LSM) | Proposed |
|---------|--------|---------|----------|----------|
| ACC     | 72.920 ± 0.098 | 75.997 ± 0.325 | 72.435 ± 0.060 | 84.916 ± 0.238 |

Table 3: Comparison between SWA (Joint) and the proposed method. This test aims to discover the affects with the different training structure and various loss combinations. Accuracies (%) of 5 random runs are reported.

| Methods           | SWA (Joint) | Proposed |
|-------------------|-------------|----------|
| $L_{SM}+L_{VA}$   | 68.831 ± 0.724 | 82.927 ± 0.407 |
| $L_{SM}+L_{PT}$   | 60.074 ± 0.125 | 63.715 ± 0.913 |
| $L_{VA}+L_{PT}$   | 54.256 ± 0.049 | 54.073 ± 0.297 |
| $L_{SM}+L_{VA}+L_{PT}$ | 79.459 ± 1.227 | 84.916 ± 0.238 |

- **MLAN_SC**: Multi-view learning with adaptive neighbors for semi-supervised classification, a semi-supervised learning method that was also proposed in [14].
- **SWA(LSM)**: This is a simple method that solves the SA/WA classification problem by using the basic framework of our model, but only uses the softmax loss in the training process.
- **SWA (Joint)**: This approach jointly optimizes the three losses used in this paper, and its final loss function is $L_{SM} + \alpha L_{VA} + \beta L_{PT}$. This approach does not use the well-designed training process discussed in Section 3.1, but instead trains all the labeled and unlabeled categories together.

4.3 Classification Performance

In this experiment, we first compare the performance of our model and the first three baseline methods. We then evaluate the last baseline algorithm SWA (Joint) with various loss combinations. The classification performance with 30% of labeled categories is reported in Table 2. Table 2 clearly shows that the proposed algorithm yields a much higher classification accuracy than the first three baselines. The clustering method MLAN_C yields a very poor performance since it is an unsupervised learning method that does not utilize any label information. The semi-supervised learning method MLAN_SC achieves a slightly better performance, but is still much worse than the proposed method. This may be due to the reason that the linear transformation used in MLAN_SC cannot project the multi-view features into a good enough subspace. Furthermore, SWA(LSM) only trains the model using the labeled information, thus yields the overall worst performance.

Table 3 summarizes the classification performance using different training structure with various loss combinations. Still, the proposed algorithm yields the best performance among all the combinations. Notably, the performance significantly drops if the three loss functions are naively combined. This is because the pseudo-labeled triplet loss cannot be learned well if most of the positive and negative categories are mislabeled.

4.4 Evaluation with the Cold-start Problem

We then evaluate the performance of our model when facing with the cold-start problem. Here we assume that the comment view of all the test data is missing. According to the results shown in Table 4, both SWA (Joint) and the proposed method outperform SWA (LSM) by a large margin. This observation can be explained by the analysis in Section 3.2. The view alignment loss in SWA (Joint) and the proposed method can project the comment and image features to close enough subspaces, so the lack of one view can be approximately replaced by the other view. With the well-designed training process, the proposed model also achieves a better performance than SWA (Joint).

Table 4: The performance conducted on the cold-start items. Accuracies (%) of 5 random runs are reported with the best result in bold.

| Methods | SWA (LSM) | SWA (Joint) | Proposed |
|---------|-----------|-------------|----------|
| ACC     | 61.936 ± 0.084 | 74.319 ± 3.964 | 82.389 ± 0.262 |
4.5 Results of the Single View

To evaluate the effect of each view, we further conduct the experiment to test the contribution of every single view. At first, we modify the structure of the proposed network to fix the single view input. Then, the view alignment loss is removed because there is no longer pair of comment and image views.

Table 5 indicates that the intrinsic view contains more useful information than others. Since the real-world data exist massive noise, both comment and image feature views yield worse performance across all compared methods. Moreover, MLAN_C and MLAN_SC cannot capture the knowledge of intrinsic view effectively because the SA/WA classification problem is hard to be learned by the linear transformation strategy due to the complexity of the original distributions. For the latter three methods, there are no significant gaps among their results which illustrate that both SWA (L\text{SM}), SWA (Joint) and the proposed model can handle the single view problem.

4.6 Comparison with the Variation of Labeled Ratios

In this experiment, we evaluate how performance will change with varied sizes of labeled data. Figure 8 shows that with the number of labeled categories increases, the performance of almost all the methods improves at first, and then gradually stabilize. In all the cases, the proposed method still yields the best performance.

4.7 The Sensitivity of Balance Parameters

In this section, we evaluate the performance of our method by varying the balance parameters $\alpha$ and $\beta$. Figure 8 shows that the performance of almost all the methods improves at first, and then gradually stabilize. In all the cases, the proposed method still yields the best performance.

In particular, our model achieves the best result when $\alpha = 10$, and it is not sensitive to $\alpha$ in the range from 4 to 15.
Figure 9 reveals the impacts of varying $\beta$. The performance tends to stay stable at $0.01 \leq \beta \leq 0.2$, and $\beta = 0.1$ yields the best result for our model.

4.8 Case Study

Finally, we display some interesting observations in this section.

As shown in Figure 10, two red boxes reveal the classification results with a very high confidence (87%-99%). In the left red box, items like shirt, wallet, knitwear, and suitcase should not be shopped via screenless devices since their appearances play important roles in consumer decision making. However, SA categories in the blue box have lower confidence scores, which suggests that they may have opportunities to be sold in the screenless environment. In our model, semantic information extracted from three views can adjust the final inference result to make it more reasonable. For example, the laptop should be identified as the SA category with a high confidence. However, its confidence score decreases when the ‘second-hand’ tag occurs. This is not surprising since most items with the ‘second-hand’ tag are likely to have less importance on their appearance. For this reason, the ‘second-hand’ laptop lost its external attractiveness to some degree as shown in the left blue box of Figure 10. Similar conclusions can be also found in the WA category.

5 Related Work

In this section, we briefly review the existing works of multi-view learning and semi-supervised learning.

5.1 Multi-view Learning

Multi-view learning aims to merge the knowledge from multiple different views to improve the learning performance, and a range of approaches along this line of research have been proposed $[15,16]$. These approaches can be roughly divided into two categories: 1) View alignment learning methods, and 2) View agreement learning methods. The main idea of the first category is to align features learned from each view to obtain the common latent subspace. Canonical Correlation Analysis (CCA) $[17,18,19]$, Partial Least Squares (PLS) $[20,21]$ and similarity learning $[22,23]$ are utilized to align the inter-view features. On par with these view alignment learning methods, view agreement learning methods devote to require that the models learned from each view have agreed with the output. The representative methods include co-training $[24]$, co-EM $[25]$, co-regularization $[9]$, and co-regression $[26]$. Besides, the incomplete view problem has also been studied $[3,4,5,27]$.

5.2 Semi-supervised Learning

Semi-supervised learning seeks methods for utilizing unlabeled data in addition to labeled data to improve learning performance, and many efforts have been made for this research line $[28,29,30]$. Roughly speaking, these methods can be divided into four categories. The first category is generative methods $[31,32,33]$, which extend the supervised generative models by estimating the labels of the unlabeled data using the expectation-maximization (EM) algorithm. The second category is S3VM (Semi-Supervised Support Vector Machine) methods $[34,35]$, which utilize the unlabeled data to adjust the decision boundary to make it go through the less dense region while keeping the labeled data
being classified correctly. The Third category is graph-based methods [36,37,38], which construct a graph by both labeled data and unlabeled data and employ label propagation on the graph. The last category is disagreement-based methods [30,24,26,59,40,41,42,43], which train multiple learners and exploit the disagreements among the learners by labeling the unlabeled data.

6 Conclusions and Future Work

In this paper, we propose a novel framework for inferring the importance of product appearance in consumer decision making. To this end, we first collect category labels via crowdsourcing, and extract item features from three different views. We then propose a semi-supervised framework with three carefully designed losses to classify SA and WA items. Extensive experiments verify the effectiveness of the proposed method.

In this study, we focus on the items’ inherent properties and assume that the importance of product appearance is user-independent. However, we also realize that whether or not accepting screenless shopping may be a subjective behavior in some cases. For example, some people may accept purchasing an umbrella via screenless shopping since they care most about its functionality (protecting against rain). However, some other people may think that the design and appearance of the umbrella are equally important, thus will not purchase umbrellas via screenless devices. We will study the problem of learning personalized opinions about screenless shopping in our future work.

References

[1] Bin Zhou, Wentao Li, Ka Wing Chan, Yijia Cao, Yonghong Kuang, Xi Liu, and Xiong Wang. Smart home energy management systems: Concept, configurations, and scheduling strategies. Renewable and Sustainable Energy Reviews, 61:30–40, 2016.
[2] Biljana L Risteska Stojkoska and Kire V Trivodaliev. A review of internet of things for smart home: Challenges and solutions. *Journal of Cleaner Production*, 140:1454–1464, 2017.

[3] Yang Yang, De-Chuan Zhan, Xiang-Rong Sheng, and Yuan Jiang. Semi-supervised multi-modal learning with incomplete modalities. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 2998–3004, 2018.

[4] Shao-Yuan Li, Yuan Jiang, and Zhi-Hua Zhou. Partial multi-view clustering. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, pages 1968–1974, 2014.

[5] Xinwang Liu, Miaomiao Li, Lei Wang, Yong Dou, Jianping Yin, and En Zhu. Multiple kernel k-means with incomplete kernels. In *Proceedings of the 31th AAAI Conference on Artificial Intelligence*, pages 2259–2265, 2017.

[6] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67, 2005.

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pages 3111–3119, 2013.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Xiaonan Ding, Partha Niyogi, and Mikhail Belkin. A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of ICML workshop on learning with multiple views*, volume 2005, pages 74–79. Citeseer, 2005.

[10] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Discriminative deep metric learning for face verification in the wild. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1875–1882, 2014.

[11] Jingchao Ni, Shiyu Chang, Xiao Liu, Wei Cheng, Haifeng Chen, Dongkuan Xu, and Xiang Zhang. Co-regularized deep multi-network embedding. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 469–478, 2018.

[12] Vahid Noroozi, Sara Babaei, Lei Zheng, Sihong Xie, Weixiang Shao, and Philip S Yu. Semi-supervised deep representation learning for multi-view problems. *arXiv preprint arXiv:1811.04480*, 2018.

[13] R. Manmatha, Chao-Yuan Wu, Alexander J. Smola, and Philipp Krähenbühl. Sampling matters in deep embedding learning. In *Proceedings of the 2017 IEEE International Conference on Computer Vision*, pages 2859–2867, 2017.

[14] Feiping Nie, Guohao Cai, and Xuelong Li. Multi-view clustering and semi-supervised classification with adaptive neighbours. In *Proceedings of the 31th AAAI Conference on Artificial Intelligence*, pages 2408–2414, 2017.

[15] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *CoRR*, abs/1304.5634, 2013.

[16] Shiliang Sun. A survey of multi-view machine learning. *Neural Computing and Applications*, 23(7-8):2031–2038, 2013.

[17] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[18] Shotaro Akaho. A kernel method for canonical correlation analysis. In *International Meeting on Psychometric Society*, 2001, 2001.

[19] Galen Andrew, Raman Arora, Jeff A. Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1247–1255, 2013.

[20] Herman Wold. Soft modeling: the basic design and some extensions. *Systems under indirect observation*, 2:343, 1982.

[21] William Robson Schwartz, Aniruddha Kembhavi, David Harwood, and Larry S. Davis. Human detection using partial least squares analysis. In *Proceedings of the 2009 IEEE International Conference on Computer Vision*, pages 24–31, 2009.

[22] Bing Bai, Jason Weston, David Grangier, Ronan Collobert, Kunihiko Sadamasa, Yanjun Qi, Olivier Chapelle, and Kilian Q. Weinberger. Learning to rank with (a lot of) word features. *Information Retrieval*, 13(3):291–314, 2010.

[23] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc’Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pages 2121–2129, 2013.

[24] Avrim Blum and Tom M. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, 1998.
[25] Kamal Nigam and Rayid Ghani. Analyzing the effectiveness and applicability of co-training. In Proceedings of the 9th International Conference on Information and Knowledge Management, pages 86–93, 2000.

[26] Zhi-Hua Zhou and Ming Li. Semi-supervised regression with co-training. In Proceedings of the 19th International Joint Conference on Artificial Intelligence, pages 908–916, 2005.

[27] Chang Xu, Dacheng Tao, and Chao Xu. Multi-view learning with incomplete views. IEEE Transactions on Image Processing, 24(12):5812–5825, 2015.

[28] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. Semi-supervised learning. MIT Press, 2006.

[29] Xiaoqin Zhu. Semi-supervised learning literature survey. Technical Report 1530, University of Wisconsin-Madison, 2007.

[30] Zhi-Hua Zhou and Ming Li. Semi-supervised learning by disagreement. Knowledge and Information Systems, 24(3):415–439, 2010.

[31] Kamal Nigam, Andrew McCallum, Sebastian Thrun, and Tom M. Mitchell. Text classification from labeled and unlabeled documents using EM. Machine Learning, 39(2/3):103–134, 2000.

[32] David J. Miller and Hasan S. Uyar. A mixture of experts classifier with learning based on both labelled and unlabelled data. In Proceedings of the 9th International Conference on Neural Information Processing Systems, pages 571–577, 1996.

[33] Behzad M. Shahshahani and David A. Landgrebe. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon. IEEE Transactions on Geoscience and Remote Sensing, 32(5):1087–1095, 1994.

[34] Thorsten Joachims. Transductive inference for text classification using support vector machines. In Proceedings of the 16th International Conference on Machine Learning, pages 200–209, 1999.

[35] Olivier Chapelle and Alexander Zien. Semi-supervised classification by low density separation. In Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics, 2005.

[36] Avrim Blum and Shuchi Chawla. Learning from labeled and unlabeled data using graph mincuts. In Proceedings of the 18th International Conference on Machine Learning, pages 19–26, 2001.

[37] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In Proceedings of the 16th International Conference on Neural Information Processing Systems, pages 321–328, 2003.

[38] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research, 7:2399–2434, 2006.

[39] Zhi-Hua Zhou and Ming Li. Tri-training: Exploiting unlabeled data using three classifiers. IEEE Transactions on Knowledge and Data Engineering, 17(11):1529–1541, 2005.

[40] Zhi-Hua Zhou, Ke-Jia Chen, and Hong-Bin Dai. Enhancing relevance feedback in image retrieval using unlabeled data. ACM Transactions on Information Systems, 24(2):219–244, 2006.

[41] Ming Li and Zhi-Hua Zhou. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. IEEE Transactions on Systems, Man, and Cybernetics: Systems, Part A, 37(6):1088–1098, 2007.

[42] Zhi-Hua Zhou, De-Chuan Zhan, and Qiang Yang. Semi-supervised learning with very few labeled training examples. In Proceedings of the 22th AAAI Conference on Artificial Intelligence, pages 675–680, 2007.

[43] Dong-Dong Chen, Wei Wang, Wei Gao, and Zhi-Hua Zhou. Tri-net for semi-supervised deep learning. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, pages 2014–2020, 2018.