SHIFT15M: Multiobjective Large-Scale Fashion Dataset with Distributional Shifts

Masanari Kimura
ZOZO Research
Department of Statistical Science,
School of Multidisciplinary Sciences,
The Graduate University for Advanced Studies (SOKENDAI)
masanari.kimura@zozo.com

Takuma Nakamura
ZOZO Research
takuma.nakamura@zozo.com

Yuki Saito
ZOZO Research
yuki.saito@zozo.com

Abstract

Many machine learning algorithms assume that the training data and the test data follow the same distribution. However, such assumptions are often violated in real-world machine learning problems. In this paper, we propose SHIFT15M, a dataset that can be used to properly evaluate models in situations where the distribution of data changes between training and testing. The SHIFT15M dataset has several good properties: (i) Multiobjective. Each instance in the dataset has several numerical values that can be used as target variables. (ii) Large-scale. The SHIFT15M dataset consists of 15 million fashion images. (iii) Coverage of types of dataset shifts. SHIFT15M contains multiple dataset shift problem settings (e.g., covariate shift or target shift). SHIFT15M also enables the performance evaluation of the model under various magnitudes of dataset shifts by switching the magnitude. In addition, we provide software to handle SHIFT15M in a very simple way: https://github.com/st-tech/zozo-shift15m

Figure 1: Overview of SHIFT15M dataset.

Preprint. Under review.
1 Introduction and motivation

Many machine learning algorithms assume that the training data and the test data are generated from the same distribution. However, in real-world problem settings, this assumption is often violated. This situation is called dataset shift or distribution shift and is a very important problem setting for machine learning \cite{33,39}. It is well known that obtaining robust prediction models under dataset shifts is still a difficult task \cite{33,39}. As a result, a number of strategies have been proposed to evaluate robustness to dataset shifts. Common examples include noise corruptions \cite{13,14}, spatial transformations \cite{7,8}, and adversarial examples \cite{2,31}. Encouragingly, the past few years have seen substantial progress in robustness to these synthetic distribution shifts, e.g., see \cite{5,12,17,18,24,34,40,42,44,46} among many others. Although synthetic distribution shifts are a good starting point for experiments, a recent study have reported that there is no correlation between robustness to artificial dataset shifts and robustness to natural dataset shifts \cite{32}. The main motivation of this paper is to provide a novel dataset with natural dataset shifts.

Recall that empirical risk minimization (ERM) is supported by the i.i.d. assumption \cite{37,38,3}:

\[
\hat{h} = \arg\min_{h \in H} \bar{R}(h) = \arg\min_{h \in H} \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \ell(y_i, h(x_i)).
\]

(1)

This is because if the distributions of the training data and the test data are identical, the expected value of the empirical risk strictly matches the expected risk due to the law of large numbers:

\[
\mathbb{E}_{p_{tr}}[\bar{R}(h)] = \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \mathbb{E}_{p_{tr}}[\ell(y_i, h(x_i))] = \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} R(h) = R(h).
\]

(2)

Although ERM is a very powerful framework, it assumes distributional identities, as mentioned above. In other words, when the distributions followed by the training data and the test data are different, ERM cannot select an appropriate hypothesis. Many researches working on dataset shift have the ultimate goal of solving this situation.

However, since most existing public datasets do not explicitly consider distribution shifts, it is difficult for researchers to evaluate their models under dataset shifts. In this paper, we propose a novel dataset, namely SHIFT15M, which can be used to properly evaluate models under the dataset shift situations. The SHIFT15M dataset has several good properties: (i) Multiobjective. Each instance in the dataset has several numerical values that can be used as target variables. (ii) Large-scale. The SHIFT15M dataset consists of 15million fashion images. (iii) Coverage of types of dataset shifts. SHIFT15M contains multiple dataset shift problem settings (e.g., covariate shift or target shift). SHIFT15M also enables the performance evaluation of the model under various magnitudes of dataset shifts by switching the magnitude. In addition, we provide software to handle SHIFT15M in a very simple way.\footnote{https://github.com/st-tech/zozo-shift15m} We also provide a datasheet for SHIFT15M, which follows the template \cite{11} (see Appendix).

In the following, we first introduce the notations used throughout this paper and the famous taxonomy of dataset shifts.

1.1 Problem formulation

First, we formulate the problem of supervised learning. We denote by $\mathcal{X} \subset \mathbb{R}^d$ the input space. The output space is denoted by $\mathcal{Y} \subset \mathbb{R}$ (regression) or $\mathcal{Y} \subset \{1, \ldots, K\}$ ($K$-class classification). We assume that training examples $\{(x_i^{tr}, y_i^{tr})\}_{i=1}^{n_{tr}}$ are independently and identically distributed (i.i.d.) according to some fixed but unknown distribution $p_{tr}(x, y)$, which can be decomposed into the marginal distribution and the conditional probability distribution, i.e., $p_{tr}(x, y) = p_{tr}(x)p_{tr}(y|x)$. We also denote the test examples by $\{(x_i^{te}, y_i^{te})\}_{i=1}^{n_{te}}$ drawn from a test distribution $p_{te}(x, y) = p_{te}(x)p_{te}(y|x)$.

Let $H$ be a hypothesis class. The goal of supervised learning is to obtain a hypothesis $h: \mathcal{X} \rightarrow \mathbb{R} (h \in H)$ with the training examples that minimizes the expected loss over the test distribution:

\[
R(h) := \mathbb{E}_{(x^{te}, y^{te}) \sim p_{te}(x, y)}[\ell(h(x^{te}), y^{te})],
\]

(3)

where $\ell: \mathbb{R} \times \mathcal{Y} \rightarrow \mathbb{R}$ is the loss function that measures the discrepancy between the true output value $y$ and the predicted value $\hat{y} := h(x)$.
1.2 Taxonomy of dataset shifts

SHIFT15M contains several types of dataset shifts. Dataset shift problems can be categorized in terms of which variables and conditional probabilities change, as follows.

Definition 1.1. (Covariate shift [28]) We consider that the two distributions $p_{tr}(x, y)$ and $p_{te}(x, y)$ satisfy the covariate shift assumption if the following three conditions hold:

$$
p_{tr}(x) \neq p_{te}(x),
\quad p(y|x) = p_{tr}(y|x) = p_{te}(y|x).
$$

Definition 1.2. (Target shift [45]) We consider that the two distributions $p_{tr}(x, y)$ and $p_{te}(x, y)$ satisfy the target shift assumption if the following three conditions hold:

$$
p_{tr}(y) \neq p_{te}(y),
\quad p(x|y) = p_{tr}(x|y) = p_{te}(x|y).
$$

Definition 1.3. (Concept drift [35]) We consider that the two distributions $p_{tr}(x, y)$ and $p_{te}(x, y)$ satisfy the concept drift assumption if the following three conditions hold:

$$
p_{tr}(y|x) \neq p_{te}(y|x) \quad \text{or} \quad p_{tr}(x|y) \neq p_{te}(x|y).
$$

Definition 1.4. (General dataset shift [30]) Let $Z \subseteq \{Z, Y\}$ be a set of immutable variables whose marginal distribution should remain fixed, $W \subseteq \{X, Y\} \setminus Z$ be a set of mutable variables whose distribution can be shifted, and $V = \{X, Y\} \setminus (W \cup Z)$ be the remaining dependent variables. This partition of the variables defines a factorization of $p_{tr}$ into

$$
p_{tr}(v|w, z)p_{tr}(w|z)p_{tr}(z),
$$

where $z \in Z$, $w \in W$ and $v \in V$. We consider that the two distributions $p_{tr}(x, y)$ and $p_{te}(x, y)$ satisfy the general dataset shift assumption if the following hold:

$$
p_{tr}(w|z) \neq p_{te}(w|z).
$$

Notably, this formulation generalizes other dataset shifts. For example, if we let $Z = \emptyset$ and $W = X$, then this corresponds to a covariate shift.
1.3 Contributions

Our contributions are summarized as follows:

- We propose SHIFT15M, a dataset that can be used to properly evaluate models in situations where the distribution of data changes between training and testing. The SHIFT15M dataset has several good properties: (i) Multiobjective. Each instance in the dataset has several numerical values that can be used as target variables. (ii) Large-scale. The SHIFT15M dataset consists of 15 million fashion images. (iii) Coverage of types of dataset shifts. SHIFT15M contains multiple dataset shift problem settings (e.g., covariate shift or target shift). SHIFT15M also enables the performance evaluation of the model under various magnitudes of dataset shifts by switching the magnitude;
- We provide open-source software to handle the SHIFT15M dataset in a very simple way. Figure 3 shows the minimum sample code of our software;
- We propose a novel set-to-set matching method under dataset shift. From the experimental results, we can see that our proposed method outperforms existing methods.

```python
from shift15.datasets import NumLikesRegression

dataset = NumLikesRegression(root="./data", download=True)
(x_train, y_train), (x_test, y_test) = dataset.load_dataset(
    target_shift=True,
    train_size=train_sample_size,
    test_size=test_sample_size,
    test_mu=test_mu,
    test_sigma=test_sigma,
    train_mu=train_mu,
    train_sigma=train_sigma,
    random_seed=seed,
)
```

Figure 3: Minimum sample code using SHIFT15M data loader.

2 Dataset overview

In this section, we describe the overview of the SHIFT15M dataset. SHIFT15M consists of about 15 million image features of clothes and about two million instances that are combinations of them. Figure 1 shows an overview of each instance of SHIFT15M. From this figure, we can see that each instance of SHIFT15M has several clothing image IDs and several numerical values. Table 1 also shows some statistics of SHIFT15M. From this table, we can see that the numerical values of each instance are different for each year. This is due to the fact that SHIFT15M is constructed on IQON, a web service launched in 2010 as shown in Figure 2, and the number of users is gradually increasing. We also provide a datasheet for SHIFT15M, which follows the template [11] (see Appendix).

| Property                  | Total | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 |
|---------------------------|-------|------|------|------|------|------|------|------|------|------|------|------|
| Number of Instances       | 2,555,147 | 1,423 | 4,813 | 131,611 | 466,583 | 730,443 | 617,844 | 299,502 | 137,510 | 92,944 | 59,412 | 13,062 |
| Number of Items           | 15,218,721 | 8,327 | 29,140 | 756,532 | 2,644,564 | 4,293,802 | 3,731,864 | 1,853,647 | 855,036 | 576,022 | 375,549 | 84,238 |
| Mean Number of Likes      | 26.98   | 0.94  | 2.00  | 15.74  | 16.84  | 23.24  | 37.37  | 35.67  | 32.41  | 24.89  | 21.34  | 16.01 |
| Median Number of Likes    | 9.00    | 0.00  | 1.00  | 8.00   | 6.00   | 6.00   | 13.00  | 18.00  | 23.00  | 19.00  | 17.00  | 12.00 |

The numerical values and image features in the dataset are affected by distributional shifts from year to year. Figure 4 shows the results of a simple experiment to observe the distributional shift of image features. This experiment is a binary classification that identifies the corresponding year of the image features. The right panel of this figure shows the results of an experiment to predict the year of the image features, where one set is the year 2010, and the other is the years 2011 to 2020. The results show that it is difficult to discriminate between image features that are close to each other in terms of
the year, while discrimination becomes easier as the years are farther apart. This fact suggests that the image features shift over the years, which can be attributed to the gradual change in fashion trends.

2.1 Possible tasks

Here, we list candidate tasks for which SHIFT15M can be applied as follows:

- regression (e.g., number of likes or sum of prices);
- classification (e.g., category ids or publish years);
- set-to-set matching.

In particular, the last task of set-to-set matching is unique and will be described in detail below. Note that the tasks to which SHIFT15M can be applied are not limited to those listed above. We strongly encourage users to contribute new ideas for tasks to which SHIFT15M can be applied, and we provide a task proposal guide in the GitHub repository.

2.1.1 Set-to-set matching

We describe the problem setup for set-to-set matching [26], which is one of the unique tasks to which SHIFT15M can be applied. We consider fashion set matching, where provided examples of the outfits are used as correct combinations of items (clothes). Using a large number of examples of the outfits in the form of images, we aim to match the correct pair of defined sets using the SHIFT15M. Let \( v_n, w_m \in \mathcal{X} \) be feature vectors representing the features of each individual item. Let \( \mathcal{V} = \{v_1, \ldots, v_N\} \) and \( \mathcal{W} = \{w_1, \ldots, w_M\} \) be sets of these feature vectors, where \( \mathcal{V}, \mathcal{W} \in 2^X \).

The function \( f : 2^X \times 2^X \rightarrow \mathbb{R} \) calculates a matching score between the two sets \( \mathcal{V} \) and \( \mathcal{W} \). The set-to-set matching is a task where the matching function \( f \) is used per pair of sets to select a correct matching \([47, 26]\). Given candidate pairs of sets \((\mathcal{V}, \mathcal{W}^{(k)})\), where \( \mathcal{V}, \mathcal{W}^{(k)} \in 2^X \) and \( k \in \{1, \ldots, K\} \), we choose \( \mathcal{W}^{(k^*)} \) as a correct one so that \( f(\mathcal{V}, \mathcal{W}^{(k^*)}) \) achieves the maximum score from amongst the \( K \) candidates.

Since the image features are shifted, we have to take into account the covariate shift assumption. In the following, we propose a powerful benchmark method for set-to-set matching that addresses covariate shifts.

3 Importance weighted set-to-set matching

To solve set-to-set matching under dataset shift, which is one of the unique tasks that SHIFT15M can be used for, we propose a new benchmark method. First of all, we introduce the importance of weighted ERM, which is the most famous strategy of dataset shift adaptation.
3.1 Importance weighted ERM

For shifts in the marginal distribution (such as covariate shift), we can expect that the strategy of weighting each instance will be effective, as follows:

\[
\hat{h} = \arg \min_{h \in \mathcal{H}} \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} w(x_i) \ell(y_i, h(x_i)),
\]

where \( w(\cdot) : \mathcal{X} \rightarrow \mathbb{R}_+ \) is the weighting function. This strategy can be regarded as equivalent to minimizing the empirical error by increasing the weights of the instances that are likely to be generated by the test distribution.

**Definition 3.1.** (Importance weighted ERM \([28]\)) IWERM uses the density ratio \( p_{te}(x)/p_{tr}(x) \) as the weighting function:

\[
\hat{h} = \arg \min_{h \in \mathcal{H}} \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \frac{p_{te}(x_i^{tr})}{p_{tr}(x_i^{tr})} \ell(h(x_i^{tr}), y_i^{tr}).
\]

Adopting the density ratio as the weighting function, as in Definition 3.1, leads to the following statistically important property.

**Theorem 3.1.** (Consistency of IWERM\([28]\)) If we set \( w(x) = p_{te}(x)/p_{tr}(x) \) as the weighting function, the empirical error computed by the weighted ERM is consistent estimator of the expected error in the test distribution.

Proof.

\[
\mathbb{E}_{p_{tr}} \left[ w(X) \ell(Y, h(X)) \right] = \int_{\mathcal{X} \times \mathcal{Y}} \frac{p_{te}(x)}{p_{tr}(x)} \ell(y, h(x)) p_{tr}(x) p(y|x) dx dy
\]

\[
= \int_{\mathcal{X} \times \mathcal{Y}} \ell(y, h(x)) p_{te}(x) p(y|x) dx dy = \mathbb{E}_{p_{te}} \left[ \ell(y, h(x)) \right] = \mathcal{R}^\ell(h).
\]

Since the calculation of density ratio is often unstable in practical use, several variants exist as follows.

**Definition 3.2.** (Adaptive importance weighted ERM \([28]\)) AIWERM uses \( (p_{te}(x)/p_{tr}(x))^\alpha \) for \( \alpha \in [0, 1] \) as the weighting function:

\[
\hat{h} = \arg \min_{h \in \mathcal{H}} \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \left( \frac{p_{te}(x_i^{tr})}{p_{tr}(x_i^{tr})} \right)^\alpha \ell(h(x_i^{tr}), y_i^{tr}).
\]

**Definition 3.3.** (Relative importance weighted ERM \([41]\)) RIWERM uses \( p_{te}(x)/(1-\alpha)p_{tr}(x) + \alpha p_{te}(x) \) for \( \alpha \in [0, 1] \) as the weighting function:

\[
\hat{h} = \arg \min_{h \in \mathcal{H}} \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \frac{p_{te}(x_i^{tr})}{(1-\alpha)p_{tr}(x_i^{tr}) + \alpha p_{te}(x)} \ell(h(x_i^{tr}), y_i^{tr}).
\]

3.2 Importance weighted set-to-set matching

Using the above ideas, we propose a novel covariate shift adaptation method for set-to-set matching. Let \( \mathcal{L}(\mathcal{V}, \mathcal{W}; f) \) be the \( K \)-pair-set loss \([26]\) function for the set matching, which is defined as follows:

\[
\mathcal{L}(\mathcal{V}, \mathcal{W}, f) = - \frac{1}{K} \sum_{i=1}^{K} \sum_{j=1}^{K} \delta_{ij} \log \frac{\exp(f(V_i, W_j))}{\sum_{k=1}^{K} \exp(f(V_i, W_k))},
\]

where \( \delta \) is Kronecker’s delta. Then, we can modify \( \mathcal{L}(\mathcal{V}, \mathcal{W}, f) \) as follows:

\[
\mathcal{L}_w(\mathcal{V}, \mathcal{W}, f) = - \frac{1}{K} \sum_{i=1}^{K} \sum_{j=1}^{K} \delta_{ij} \exp(p\text{test}|V_i \cup W_j)) \log \frac{\exp(f(V_i, W_j))}{\sum_{k=1}^{K} \exp(f(V_i, W_k))}.
\]
Then, we have
\[ w(x) = \frac{p(x)}{p(x)} \]. Let \( g(x) \) be the optimal source discriminator which identifies whether \( x \) is generated \( p_{tr} \) or \( p_{te} \). Then, we can write as \( g(x) = p(s = 1|x) = \frac{1}{1 + w(x)} \). Suppose that the density ratio \( p_{te}(x)/p_{tr}(x) \) is bounded by \( \beta > 0 \), we have \( \frac{1}{1 + \beta} \leq g(x) \leq 1 \) for all \( x \). From the unlabeled data generated from \( p_{tr} \) and \( p_{te} \), we can learn the estimator \( \hat{g} \) of \( g \). Then, we can write

\[ \text{max-IW} : p(\text{test}|V_i \cup W_j) = \max_{x \in V_i \cup W_j} w(x), \]

\[ \text{mean-IW} : p(\text{test}|V_i \cup W_j) = \frac{1}{|V_i \cup W_j|} \sum_{x \in V_i \cup W_j} w(x). \]

Here, \( w(x) \) is the weighting function.

Next, we approximate \( w(x) \) by using unlabeled data from both \( p_{tr} \) and \( p_{te} \). In IWERM, the squared error can be decomposed as follows:

\[
\mathbb{E}_{p_{tr}} \left[ \|f(x) - y\|^2 \right] = \mathbb{E}_{p_{tr}} \left[ w(x)\|f(x) - y\|^2 \right]
= \mathbb{E}_{p_{tr}} \left[ \hat{w}(x)\|f(x) - y\|^2 \right] + \mathbb{E}_{p_{tr}} \left[ (w(x) - \hat{w}(x))\|f(x) - y\|^2 \right],
\]

where \( \hat{w}(x) \) is the approximator of \( w(x) \). Second term of Eq. (12) is bounded as

\[
\mathbb{E}_{p_{tr}} \left[ (w(x) - \hat{w}(x))\|f(x) - y\|^2 \right] \leq \frac{1}{2} \mathbb{E}_{p_{tr}} \left[ \|f(x) - y\|^2 \right] + \mathbb{E}_{p_{tr}} \left[ (w(x) - \hat{w}(x))^2 \right].
\]

Let \( s \) is the indicator of the distributions, where \( s = 1 \) corresponds to the train distribution and \( s = 0 \) corresponds to the test distribution, and we assume that \( p(s) = 0.5 \). Then, we also assume that

\[
p(x|s) = \begin{cases} p_{tr}(x) & s = 1, \\ p_{te}(x) & s = 0. \end{cases}
\]

Then, we have \( w(x) = \frac{p(x|s=0)}{p(x|s=1)} \). Let \( g(x) \) be the optimal source discriminator which identifies whether \( x \) is generated \( p_{tr} \) or \( p_{te} \). Then, we can write as \( g(x) = p(s = 1|x) = \frac{1}{1 + w(x)} \). Suppose that the density ratio \( p_{te}(x)/p_{tr}(x) \) is bounded by \( \beta > 0 \), we have \( \frac{1}{1 + \beta} \leq g(x) \leq 1 \) for all \( x \). From the unlabeled data generated from \( p_{tr} \) and \( p_{te} \), we can learn the estimator \( \hat{g} \) of \( g \). Then, we can write

| Models          | \( W = 0 \) | \( W = 10 \) | \( W = 20 \) | \( W = 30 \) | \( W = 40 \) | \( W = 50 \) |
|-----------------|------------|------------|------------|------------|------------|------------|
| ERM             | 9.36(±0.02) | 10.44(±0.04) | 17.10(±0.06) | 28.80(±0.05) | 39.56(±0.05) | 48.84(±0.05) |
| IWERM (optimal) | 9.36(±0.02) | 25.67(±0.12) | 32.58(±0.12) | 26.83(±0.11) | 20.19(±0.10) | 14.52(±0.10) |
| RIWERM (\( \alpha = 0.25 \)) | 9.36(±0.02) | 9.34(±0.04) | 9.53(±0.03) | 11.37(±0.04) | 14.89(±0.09) | 17.00(±0.14) |
| RIWERM (\( \alpha = 0.50 \)) | 9.36(±0.02) | 9.73(±0.04) | 9.57(±0.03) | 9.69(±0.04) | 12.37(±0.10) | 14.68(±0.15) |
| RIWERM (\( \alpha = 0.75 \)) | 9.36(±0.02) | 11.50(±0.05) | 11.35(±0.05) | 9.30(±0.03) | 10.46(±0.09) | 12.60(±0.14) |

Figure 5: Experimental results of the regression problem for the number of likes.

Table 2: Experimental results for the regression problem.
the weight estimation term as
\[ E_{p_{tr}} \left[ (w(x) - \hat{w}(x))^2 \right] = E \left[ \left( \frac{g(x) - \hat{g}(x)}{g(x)\hat{g}(x)} \right)^2 \right] \leq (1 + \beta)^4 E_{p_{tr}} \left[ (g(x) - \hat{g}(x))^2 \right] \]
\[ = (1 + \beta)^4 E_p \left[ (g(x) - \hat{g}(x))^2 \frac{p_{tr}(x)}{p(x)} \right] \leq 2(1 + \beta)^4 E_{p} \left[ (g(x) - \hat{g}(x))^2 \right] \]
\[ = 2(1 + \beta)^4 \left\{ E_p \left[ (s - g(x))^2 \right] - E_p \left[ (g(x) - \hat{g}(x))^2 \right] \right\}. \tag{15} \]

This indicates that the weighting function is approximated by the function \( g(x) \).

4 Benchmarks

In this section, we introduce three types of benchmark experiments using SHIFT15M. However, the tasks to which SHIFT15M can be applied are not limited to these three but can be applied to many more. For more benchmarks and details of each experimental setting, see our GitHub repository.

4.1 Regression problem under the target shift assumption

First, we present benchmark results for a regression problem with the target shift. The target variable is the number of likes that each instance possesses, and the input variables are the user ID and the prices of the items. In this experiment, we evaluate the robustness of the model for different shift magnitudes, and the magnitudes of the shift are measured by the Wasserstein distance. We use the simple linear regression as the ordinal ERM, and we compare this with two other covariate shift adaptation methods, IWERM and AIWERM. Figure 5 and Table 2 show the experimental results of the regression task. These results show that the performance of ERM decreases with increasing shift magnitude, while IWERM and its variants are able to gain robustness to target shifts.

4.2 Category classification problem under the covariate shift assumption

Next, we introduce benchmark results for a category classification task with the covariate shift. The target variable is the category ID, and the input variable is the item image feature extracted by VGG16 [29]. In this experiment, we sample the training and test data from different years to evaluate the performance. Figure 6 shows the experimental results of the classification task.

4.3 Set-to-set matching problem under the covariate shift assumption

Finally, we introduce benchmark results for a set-to-set matching under the covariate shift. The model architecture is the same as the previous work [26], which is based on the architecture of Transformer [38, 16, 21].
Very recently, the WILDS dataset [15] has been proposed, which is a valuable dataset that has a natural dataset shift. This data set is very useful and many experimental results have been reported [43, 27, 39, 1]. In addition to the good properties of the WILDS dataset, our SHIFT15M dataset is novel in two aspects: it has multiple shift types, and the models can be evaluated under multiple magnitudes of shifts.

### 5 Related works

There are a number of studies addressing dataset shift [20, 30, 9]. However, most existing studies report the results of numerical experiments using datasets with artificially induced distribution shifts to demonstrate the usefulness of their proposed method [4, 10, 14]. Although such an experimental setup seems reasonable, a recent study have reported that there is no correlation between robustness to artificial dataset shifts and robustness to natural dataset shifts [32].

Very recently, the WILDS dataset [15] has been proposed, which is a valuable dataset that has a natural dataset shift. This data set is very useful and many experimental results have been reported [43, 27, 39, 1]. In addition to the good properties of the WILDS dataset, our SHIFT15M dataset is novel in two aspects: it has multiple shift types, and the models can be evaluated under multiple magnitudes of shifts.

### 6 Limitations and conclusion

In this paper, we proposed the SHIFT15M dataset and reported benchmark results for several possible tasks using SHIFT15M. This dataset is unique and valuable in two aspects: (i) it contains multiple types of shifts, and (ii) it can manipulate multiple shift magnitudes. Due to legal restrictions, image features are provided instead of raw images, and users cannot view the raw images. However, we believe that this limitation is not critical, as we have been able to achieve satisfactory benchmark performance even from image features.

#### 6.1 Future works

Our future works are as follows:

- We will add benchmarks for new tasks other than those presented in this paper. We also plan to actively encourage users to submit ideas for benchmarks, and we are preparing and providing the task proposal guide.

- We will continue to develop OSS to handle SHIFT15M. We provide developers’ guides to make it easy for anyone to get involved in development.

---

Table 3: Experimental results of the Fill-In-The-N-Blank with four candidates.

| Models     | 2013         | 2014         | 2015         | 2016         | 2017         |
|------------|--------------|--------------|--------------|--------------|--------------|
| ERM        | 0.924(±0.005)| 0.907(±0.006)| 0.880(±0.009)| 0.865(±0.006)| 0.855(±0.003)|
| ERM + mean-IW | 0.924(±0.005)| 0.917(±0.002)| 0.880(±0.003)| 0.866(±0.003)| 0.860(±0.002)|
| ERM + max-IW | 0.924(±0.005)| 0.921(±0.002)| 0.896(±0.006)| 0.871(±0.001)| 0.865(±0.005)|

Table 4: Experimental results of the Fill-In-The-N-Blank with eight candidates.

| Models     | 2013         | 2014         | 2015         | 2016         | 2017         |
|------------|--------------|--------------|--------------|--------------|--------------|
| ERM        | 0.845(±0.000)| 0.822(±0.001)| 0.791(±0.005)| 0.762(±0.008)| 0.741(±0.004)|
| ERM + mean-IW | 0.845(±0.000)| 0.831(±0.008)| 0.792(±0.002)| 0.766(±0.004)| 0.749(±0.002)|
| ERM + max-IW | 0.845(±0.000)| 0.842(±0.004)| 0.807(±0.003)| 0.769(±0.005)| 0.753(±0.005)|

Our task can be considered an extended version of a standard task, Fill-In-The-Blank [6], which requires us to select an item that best extends an outfit from among four candidates. Because selecting a set corresponds to filling multiple blanks, we consider the set matching problem as Fill-In-The-N-Blank [26]. To construct the correct pair of sets to be matched, we randomly halve the given outfit $O$ into two non-empty proper subsets $V$ and $W$, as follows: $O \rightarrow \{V, W\}$, where $V \cap W = \emptyset$.

Tables 3 and 4 show the experimental results of the Fill-In-The-N-Blank with four and eight candidates. From these results, we can see that our covariate shift adaptive set-to-set matching methods can achieve better performances than the ordinal ERM.
References

[1] Anders Andreassen, Yasaman Bahri, Behnam Neyshabur, and Rebecca Roelofs. The evolution of out-of-distribution robustness throughout fine-tuning. *arXiv preprint arXiv:2106.15831*, 2021.

[2] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

[3] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Summer School on Machine Learning*, pages 169–207. Springer, 2003.

[4] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

[5] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019.

[6] Guillem Cucurull, Perouz Taslakian, and David Vazquez. Context-aware visual compatibility prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12617–12626, 2019.

[7] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning*, pages 1802–1811. PMLR, 2019.

[8] Alhussein Fawzi and Pascal Frossard. Manitest: Are classifiers really invariant? *arXiv preprint arXiv:1507.06535*, 2015.

[9] Jacob Fiksel, Abhirup Datta, Agbessi Amouzou, and Scott Zeger. Generalized bayes quantification learning under dataset shift. *Journal of the American Statistical Association*, pages 1–19, 2021.

[10] Angus Galloway, Thomas Tanay, and Graham W Taylor. Adversarial training versus weight decay. *arXiv preprint arXiv:1804.03308*, 2018.

[11] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

[12] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[13] Robert Geirhos, Carlos R Medina Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *arXiv preprint arXiv:1808.08750*, 2018.

[14] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[15] Pang Wei Koh, Shiori Sagawa, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanus Phillips, Irena Gao, Tony Lee, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[16] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pages 3744–3753. PMLR, 2019.

[17] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[18] Pratyush Maini, Eric Wong, and Zico Kolter. Adversarial robustness against the union of multiple perturbation models. In *International Conference on Machine Learning*, pages 6640–6650. PMLR, 2020.
[19] Jose G Moreno-Torres, Troy Raeder, Rocio Alaiz-Rodriguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. Pattern recognition, 45(1):521–530, 2012.

[20] Jose García Moreno-Torres, José A Sáez, and Francisco Herrera. Study on the impact of partition-induced dataset shift on k-fold cross-validation. IEEE Transactions on Neural Networks and Learning Systems, 23(8):1304–1312, 2012.

[21] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In European conference on computer vision, pages 483–499. Springer, 2016.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.

[23] Joaquin Quiñonero-Candela, Masashi Sugiyama, Neil D Lawrence, and Anton Schwaighofer. Dataset shift in machine learning. Mit Press, 2009.

[24] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. arXiv preprint arXiv:1801.09344, 2018.

[25] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In International Conference on Machine Learning, pages 5389–5400. PMLR, 2019.

[26] Yuki Saito, Takuma Nakamura, Hirotaka Hachiya, and Kenji Fukumizu. Exchangeable deep neural networks for set-to-set matching and learning. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16, pages 626–646. Springer, 2020.

[27] Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. arXiv preprint arXiv:2104.09937, 2021.

[28] Hidetoshi Shimomotai. Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of statistical planning and inference, 90(2):227–244, 2000.

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations, 2015.

[30] Adarsh Subbaswamy, Roy Adams, and Suchi Saria. Evaluating model robustness and stability to dataset shift. In International Conference on Artificial Intelligence and Statistics, pages 2611–2619. PMLR, 2021.

[31] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013.

[32] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In Advances in Neural Information Processing Systems (NeurIPS), 2020.

[33] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In CVPR 2011, pages 1521–1528. IEEE, 2011.

[34] Florian Tramer and Dan Boneh. Adversarial training and robustness for multiple perturbations. arXiv preprint arXiv:1904.13000, 2019.

[35] Alexey Tsybalkin. The problem of concept drift: definitions and related work. Computer Science Department, Trinity College Dublin, 106(2):58, 2004.

[36] Vladimir Vapnik. The nature of statistical learning theory. Springer science & business media, 2013.

[37] Vladimir N Vapnik. An overview of statistical learning theory. IEEE transactions on neural networks, 10(5):988–999, 1999.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.
[39] Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On calibration and out-of-domain generalization. arXiv preprint arXiv:2102.10395, 2021.

[40] Eric Wong, Frank Schmidt, and Zico Kolter. Wasserstein adversarial examples via projected sinkhorn iterations. In International Conference on Machine Learning, pages 6808–6817. PMLR, 2019.

[41] Makoto Yamada, Taiji Suzuki, Takaafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. Neural computation, 25(5):1324–1370, 2013.

[42] Fanny Yang, Zuowen Wang, and Christina Heinze-Deml. Invariance-inducing regularization using worst-case transformations suffices to boost accuracy and spatial robustness. arXiv preprint arXiv:1906.11235, 2019.

[43] Haoran Zhang, Natalie Dullerud, Laleh Seyyed-Kalantari, Quaid Morris, Shalmali Joshi, and Marzyeh Ghassemi. An empirical framework for domain generalization in clinical settings. In Proceedings of the Conference on Health, Inference, and Learning, pages 279–290, 2021.

[44] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In International Conference on Machine Learning, pages 7472–7482. PMLR, 2019.

[45] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In International Conference on Machine Learning, pages 819–827. PMLR, 2013.

[46] Richard Zhang. Making convolutional networks shift-invariant again. In International conference on machine learning, pages 7324–7334. PMLR, 2019.

[47] Pengfei Zhu, Lei Zhang, Wangmeng Zuo, and David Zhang. From point to set: Extend the learning of distance metrics. In Proceedings of the IEEE international conference on computer vision, pages 2664–2671, 2013.
A Details of the experimental setup

A.1 Regression problem under the target shift assumption

In this experiment, we specify the means and variances of the target variables in the training and test data, and sample accordingly to induce the target shift problem. All hyperparameters are default settings of the scikit-learn [22]. All experiments are conducted on Dual Intel Xeon Platinium 2.7 GHz, 24-cores CPU. The source code can be found at the following url: https://github.com/st-tech/zozo-shift15m/tree/main/benchmarks. See the above url for more details.

A.2 Category classification problem under the covariate shift assumption

In this experiment, we use the following hyperparameters: batchsize=32, epochs=10, learning rate=0.005, optimizer=SGD. The network architecture can be found at https://github.com/st-tech/zozo-shift15m/tree/main/benchmarks/item_category_prediction. This experiment is conducted on GCP n1-standard-8 (8 vCPU, 30 GB RAM, NVIDIA Tesla K80). The datasets are generated from SHIFT15M by random sampling each year, and we split it into train/validation/test as 3500/500/500.

A.3 Set-to-set matching problem under the covariate shift assumption

Experimental setup is same as the previous work [26]. In this experiment, we use the following hyperparameters: batchsize=32, epochs=32, learning rate=0.001, optimizer=SGD. The source code can be found at the following url: https://github.com/st-tech/zozo-shift15m/tree/main/benchmarks/set_matching. This experiment is conducted on AWS SageMaker ml.p2.xlarge.
Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

Many machine learning algorithms assume that the training data and the test data are generated from the same distribution. In the real world, however, this assumption is most often violated. Many robust algorithms against such dataset shifts have been studied, but they often experiment with artificially induced dataset shifts on originally i.i.d. datasets. Although such experiments seem reasonable, recent studies have reported that there is no correlation between robustness to artificial dataset shifts and robustness to natural dataset shifts. The main motivation of the SHIFT15M project is to provide a dataset that contains natural dataset shifts collected from a web service that was actually in operation for several years. In addition, the SHIFT15M dataset has several types of dataset shifts, allowing us to evaluate the robustness of the model to different types of shifts (e.g., covariate shift and target shift).

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The initial version of the dataset was created by Masanari Kimura, Yuki Saito, Kazuya Morishita, Ryosuke Goto, and Takuma Nakamura most of whom were researchers at the ZOZO Research.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Not applicable.

Any other comments?

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The SHIFT15M dataset is a collection of outfits posted to the fashion website IQON (which is no longer providing this service). A record represents the posted outfit, the user who posted it, and some meta-information, it has 5 fields.

- set_id: An ID that identifies the outfit that was posted.
- items: Provides information about the items that comprise the posted outfit and consists of 4 subfields.
- category_id1: An ID indicating the item category (e.g., outerwear, tops, ...).
- category_id2: An ID indicating the item subcategory (e.g., T-shirts, blouses, ...).
- price: Price of the item (Japanese yen).
- user: Provides information about the user who posted the outfit and consists of 2 subfields. An ID that identifies the user who posted the outfit. A list of brands that users have voted for as their favorites. The number is an ID that identifies the brand.
- like_num: the number of times this outfit has been favorited by other users.
- publish_date: The date the outfit was posted.

How many instances are there in total (of each type, if appropriate)?

The dataset consists of 15,218,721 item images and 2,555,147 outfits which created by users of IQON.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

We collected outfits posted on a Japanese fashion website "IQON". This website has about 2M users, almost all are Japanese women. Most of them are in their 20s and 30s. The collection period was from 01/01/2010 to 04/06/2020. An outfit is a set of multiple items, and each item has a corresponding category. In SHIFT15M, outfits that contain 4 or more items belonging to the main categories (outerwear, tops, bottoms, shoes, bags, hats, and accessories) were collected.

What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each item consists of 4096-dimensional features extracted via the VGG16 model trained using the ILSVRC2012 dataset.

Is there a label or target associated with each instance? If so, please provide a description.

Yes. Each instance has several numerical values (category ID, number of likes). We can switch between several tasks by choosing one of these as the target variable.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is...
missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Items that do not belong to the main categories, such that underwear and background images for collage, are missing. The items field consists only of items that belong to the main categories, but the original outfit may contain items other than these.

**Are relationships between individual instances made explicit? (e.g., users’ movie ratings, social network links)?**

If so, please describe how these relationships are made explicit.

Each instance is assigned the ID of the user who submitted the outfit.

**Are there recommended data splits (e.g., training, development/validation, testing)?**

If so, please provide a description of these splits, explaining the rationale behind them.

SHIFT15M is a dataset with multiple dataset shifts observed in the real world. We provide software that makes it easy to experiment with different types and sizes of shifts. SHIFT15M was collected between 2010-2020, and our software allows automatic train/val/test splitting by specifying the shift type and magnitude.

**Are there any errors, sources of noise, or redundancies in the dataset?**

If so, please provide a description.

No.

**Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?**

If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained.

**Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)?**

If so, please provide a description.

No.

**Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**

If so, please describe why.

No.

**Are the dataset relate to people?**

Yes. Each instance is a combination of outfits created by an individual and preferred by that individual.

**Does the dataset identify any subpopulations (e.g., by age, gender)?**

If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

**Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?**

If so, please describe how.

It is impossible to identify individuals from the dataset.

**Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?**

If so, please provide a description.

No.

**Any other comments?**

---

**Collection Process**

**How was the data associated with each instance acquired?**

Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Except for the item attributes, the data was generated by users. Item attributes (category and price) were collected from e-commerce sites that sell the item. All data was viewable on the website.

**What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?**

Users were able to create and publish their outfits using an editor provided by the website. The items selected in the editor are registered as an outfit, and this function was tested based on general software development procedures.

**If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**
We collected a complete dataset without sampling to create our dataset, except for data deleted by the user.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? The employees in ZOZO Technologies, Inc. or VASILY, Inc. (merged into ZOZO Technologies, Inc.) were involved in collecting data.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The dataset was collected in the period of 2010-2020. Each outfit includes a timestamp that describes when the outfit created.

Were any ethical review processes conducted (e.g., by an institutional review board)? No.

Does the dataset relate to people? Yes. Each instance is a combination of outfits created by an individual and preferred by that individual.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? Collected directly through the website.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to the exact language of the notification itself.

Notified in the Terms of Service.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to the exact language to which the individuals consented.

The use of the service was deemed as consent.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

It is possible to contact the company that provided the service (see https://tech.zozo.com/privacy/info/).

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

No, there had been no potential impact analysis conducted.

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

We extracted the CNN features from images and treated them as input data in our image-based tasks. As a result, our dataset contains the features but does not include raw photos, making them anonymized. The CNN we used is an official pre-trained VGG16, and we adopted the outputs of the 'fc6' layer before applying ReLU as the feature. We used the Chainer implementation for extracting CNN features. For more information on the Chainer implementation, please refer to the reference page: https://docs.chainer.org/en/v7.8.0/reference/generated/chainer.links.VGG16Layers.html. We exclude the outfits that contain less than four items. Other than that, we did not remove any instances in creating our dataset. However, we excluded some data in each independent task. In detail, please refer to each task description.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

No.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

All software are provided on the SHIFT15M repository.

Any other comments?
Uses

Has the dataset been used for any tasks already? If so, please provide a description.

Benchmarks using this dataset and the specified evaluation protocol are listed in https://github.com/st-tech/zozo-shift15m/tree/main/benchmarks.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

All benchmarks that use this dataset will be available at https://github.com/st-tech/zozo-shift15m/tree/main/benchmarks.

What (other) tasks could the dataset be used for?

Here, we list candidate tasks for which SHIFT15M can be applied as follows:

- regression (e.g., number of likes or sum of prices);
- classification (e.g., category ids or publish years);
- set-to-set matching.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

No.

Are there tasks for which the dataset should not be used? If so, please provide a description.

This dataset is distributed in a way that excluding raw images and anonymizing the users/brands. Therefore, it requires the dataset users not to reconstruct raw images from the image features or restore the anonymized parts in a future task.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

ZOZO Research is supporting/maintaining the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

All changes to the dataset will be announced through the GitHub Releases.

Is there an erratum? If so, please provide a link or other access point.

All changes to the dataset will be announced through the GitHub Releases. Errata are listed under the “Errata” section of SHIFT15M repository.
Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

All changes to the dataset will be announced through the GitHub Releases.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

No.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

They will continue to be supported with all information on SHIFT15M repository. We also provide the contribution guides for software that supports the dataset.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Others may do so and should contact the original authors about incorporating fixes/extensions.

Any other comments?