MetaShift: A Dataset of Datasets for Evaluating Contextual Distribution Shifts

Anonymous Authors[1]

Abstract

Understanding the performance of machine learning models across diverse data distributions is critically important for reliable applications. Motivated by this, there is a growing focus on curating benchmark datasets that capture distribution shifts. In this work, we present MetaShift—a collection of 12,868 sets of natural images across 410 classes—to address this challenge. We leverage the natural heterogeneity of Visual Genome and its annotations to construct MetaShift. The key construction idea is to cluster images using its metadata, which provides context for each image (e.g. *cats with cars* or *cats in bathroom*) that represent distinct data distributions. MetaShift has two important benefits: first, it contains orders of magnitude more natural data shifts than previously available. Second, it provides explicit explanations of what is unique about each of its data sets and a distance score that measures the amount of distribution shift between any two of its data sets. Importantly, to support evaluating ImageNet trained models on MetaShift, we match MetaShift with ImageNet hierarchy. The matched version covers 867 out of 1,000 classes in ImageNet-1k. Each class in the ImageNet-matched Metashift contains 19.3 subsets capturing images in different contexts.

1. Introduction

A major challenge in machine learning (ML) is that a model can have very different performances and behaviors when it's applied to different types of natural data. For example, if the user data have different contexts compared to the model's training data (e.g. users have outdoor dog photos and the model's training was mostly on indoor images), then the model's accuracy can greatly suffer. A model can have disparate performances even within different subsets within its training and evaluation data. In order to assess the reliability and fairness of a model, we therefore need to evaluate its performance and training behavior across heterogeneous types of data. However, the lack of well-structured datasets representing diverse data distributions makes systematic evaluation difficult.

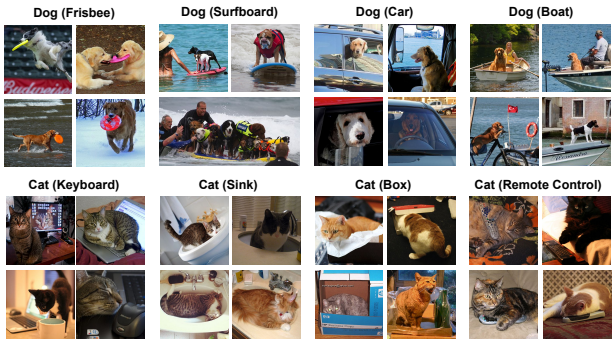

Figure 1: **Example subsets of natural images from MetaShift.** MetaShift leverages the natural heterogeneity within each class (e.g., "cat", "dog") to provide many subsets of images. Each subset corresponds to images in a similar context (the context is stated in parenthesis) and represents a coherent real-world data distribution. Here, we only show 2 out of 410 classes and 8 out of 12,868 subsets of images from MetaShift.

In this work, we present MetaShift to tackle this challenge. MetaShift is a collection of 12,868 sets of natural images from 410 classes. Each set corresponds to images in a similar context and represents a coherent real-world data distribution, as shown in Figure 1. Different from and complementary to other efforts to curate benchmarks for data shifts, MetaShift pulls together data across different experiments or sources. It leverages heterogeneity within the large sets of images from the Visual Genome project by clustering the images using metadata that describes the context of each image.

Different from ImageNet, images from Visual Genome usually contains much more than one objects, which already poses a distribution shift.

MetaShift also implements a score that measures the distance between any two subsets, which could study ML models' behavior under different carefully modulated amounts of distribution shift.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Importantly, to support evaluating ImageNet trained models on MetaShift, we match MetaShift classes with ImageNet hierarchy using WordNet synsets. We match the meta-data tags in MetaShift to ImageNet-1k and selected 5,040 sets of images from 261 classes. This ensures that all the labels of the regenerated dataset are a subset of the ImageNet-1k which means it could be a benchmark of evaluating the robustness of ImageNet model across distribution shifts.

Our contributions: We present MetaShift as an important dataset with heterogeneous contexts. We match the labels in MetaShift to ImageNet-2012, constructing a new labeled dataset where the labels are a subset of the 1000 labels of ImageNet-2012. We also demonstrate that the labels of this dataset can cover 867 of the 1,000 labels of ImageNet-2012, which means it could be used as a benchmark to evaluate the robustness of ImageNet model across distribution shifts. Enumerable binary classification tasks can be constructed over MetaShift to evaluate the performance of ImageNet Model across diverse distribution shift.

2. The MetaShift Construction Methodology

The MetaShift is a collection of subsets of data together with an annotation graph that explains the similarity/distance between two subsets (edge weight) as well as what is unique about each subset (node metadata). For each class, say "cat", we have many subsets of cats, and we can think of each subset as a node in the graph, as shown in Figure 2. Each subset corresponds to "cat" in a different context: e.g. "cat with sink" or "cat with fence". The context of each subset is the node metadata. The "cat with sink" subset is more similar to "cat with faucet" subset because there are many images that contain both sink and faucet. This similarity is the weight of the edge; a higher weight means the contexts of the two nodes tend to co-occur in the same data.

Base Dataset: Visual Genome We leverage the natural heterogeneity of Visual Genome and its annotations to construct MetaShift. Visual Genome contains over 100k images across 1,702 object classes. For each image, Visual Genome annotates the class labels of all objects that occur in the image. Formally, for each image $x^{(i)}$, we have a list of meta-data tags $m^{(i)} = \{t_1^{(i)}, t_2^{(i)}, \dots, t_{n_m}^{(i)}\}$, each indicating the presence of an object in the context. We denote the vocabulary of the meta-data tags as $\mathbb{M} = \{m_0, \dots, m_{|\mathbb{M}|}\}$. MetaShift is constructed on a class-by-class basis: For each class, say "cat", we pull out all cat images and proceed with the following steps.

Step 1: Generate Candidate Subsets We first generate candidate subsets by enumerating all possible meta-data tags. We construct $|\mathbb{M}|$ candidate subsets where the ith subset contains all images of the class of interest (i.e., "cat") that has a meta-tag m_i. We then remove subsets whose sizes are less than a threshold (e.g., 25).

Step 2: Construct Meta-graphs Since the meta-data are not necessarily disentangled, the candidate subsets might contain significant overlaps (e.g., "cat with sink" and "cat with faucet"). To capture this phenomenon, we construct a meta-graph to model the relationships among all subsets of each class. Specifically, for each class $j \in \mathbb{Y}$, we construct meta-graph, a weighted undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where each node $v \in \mathcal{V}$ denotes a candidate subset, and the weight of each edge is the overlap coefficient between two subsets:

$$\text{overlap}(X, Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)}, \tag{1}$$

We remove the edges whose weights are less than a threshold (e.g., 0.1) to sparsify the graph. As shown in Figure 2, the meta-graph $\mathbb{G}$ captures meaningful semantics of the multi-modal data distribution of the class of interest.

Step 3: Quantify Distances of Distribution Shifts The geometry of meta-graphs provides a natural and systematic way to quantify the distances of shifts across different data distributions: Intuitively, if two subsets are far away from each other in the MetaGraph, then the shift between them tend to be large. Following this intuition, we leverage *spectral embeddings* to assign an embedding for each node based on the graph geometry. Spectral embedding minimizes the expected square distance between nodes that are connected:

$$\min_{X: X^T 1 = 0, X^T X = I_K} \sum_{i,j \in V} A_{ij} \left\| X_i - X_j \right\|^2 \tag{2}$$

where X_i is the embedding for node $i \in \mathcal{V}$ and K is the dimension of the embedding, and A is the adjacency matrix. We denote by X the matrix of dimension $n \times K$ whose i-th row X_i corresponds to the embedding of node i. The constraint $X^T 1 = 0$ forces the embedding to be centered and $X^T X = I_K$ ensures that we do not get trivial solution like all node embeddings located at the origin (i.e., X = 0). Denoting by $L = D - A$ the Laplacian matrix of the graph, we have:

$$\text{tr}(X^T L X) = \frac{1}{2} \sum_{i,j \in V} A_{ij} \left\| X_i - X_j \right\|^2 \tag{3}$$

$$\min_{X: X^T 1 = 0, X^T X = I_K} \text{tr}(X^T L X) = \sum_{k=2}^{K+1} \lambda_k \tag{4}$$

The minimum is reached for X equal to the matrix of eigenvectors of the Laplacian matrix associated with the eigenvalues $\lambda_2, ..., \lambda_{K+1}$. After calculating the spectral embeddings, we use the euclidean distance between the embeddings of two nodes as their distance.

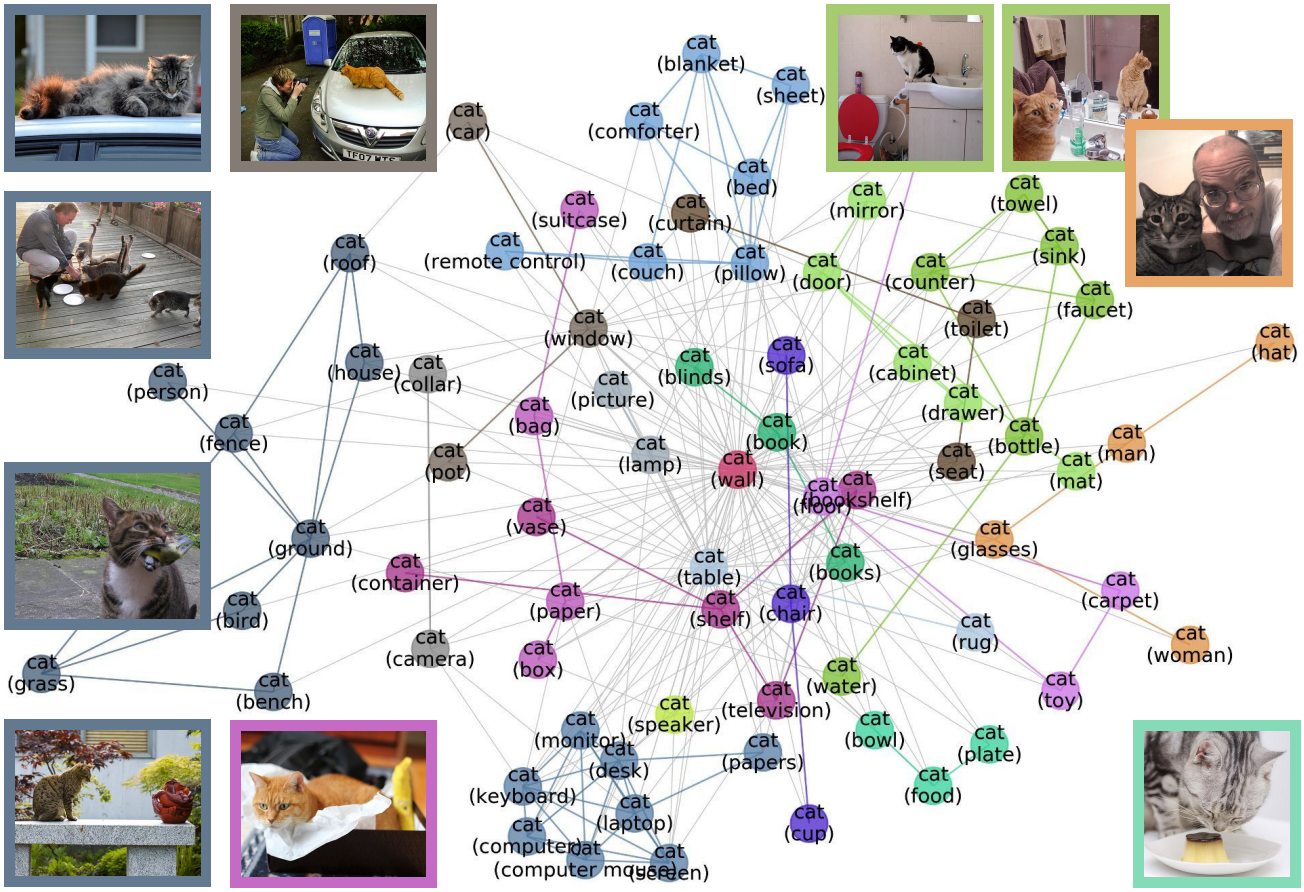

Figure 2: **Meta-graph—visualizing the diverse data distributions within the "cat" class.** Each node represents one subset of the cat images. Each subset corresponds to "cat" in a different context: e.g. "cat with sink" or "cat with fence". Each edge indicates the similarity between the two connecting subsets. Node colors indicate the communities automatically detected by graph-based algorithms. Inter-community edges are colored and intra-community edges are grayed out for better visualization. The border color of each example image indicates its community in the meta-graph. We have one such meta-graph for each of the 410 classes in the MetaShift. Beyond visualization, meta-graph also provides a natural and systematic way to quantify the distance between any two subsets (i.e., nodes), which is not available in previous benchmarks of natural data.

The subsets are characterized by a diverse vocabulary of 1,853 distinct contexts. Beyond 1,702 contexts defined by object presence, MetaShift also leverages the 37 distinct general contexts and 114 object attributes from Visual Genome. The general contexts typically describe locations (i.e., indoor, outdoor), weather (e.g., rainy, cloudless), and places (e.g., bathroom, ocean). The object attributes include color (e.g., white, red), material (e.g., wood, plastic), shape (e.g., round, square), and other object-specific properties (e.g., empty/full for plates). See Appendix A for examples and more information.

3. Matching MetaShift with ImageNet

Given that MetaShift is a flexible framework to generate a large number of real-world distribution shifts that are well-annotated and controlled, we can use it to construct a new dataset of specific classes and subpopulations.

ImageNet is an image database organized according to the WordNet hierarchy, in which each node of the hierarchy is depicted by hundreds and thousands of images. The full ImageNet contains 60,942 nodes, while the 1,000 ImageNet classes contains only 2,155 nodes. The leaf nodes represent the 1,000 classes, and the relationship between the non-leaf nodes and their children can be somewhat considered a "breed" relationship. For example, "English springer" is the child of "springer spaniel", and it can be viewed as a breed of the "springer spaniel".

To generate the dataset which fits in ImageNet, we need to match the labels in MetaShift with ImageNet-2012. For each meta-data tag of the classes and the subsets of the context as well as the attributes, we search in ImageNet hierarchy to find if it has the label with the same wordnet id. The meta-data tag in MetaShift may represent a greater domain than the leaf nodes of the ImageNet hierarchy, for example, MetaShift has only one general "cat" class, while

the ImageNet has "domestic cat" and "wildcat" under the "cat" hierarchy, and each kind of cat also has several different breeds. When doing the matching, if we can't find the matching labels in the current hierarchy, we then trace back to the parent nodes to do the further matching. That is to say, the matching should go through both leaf nodes and non-leaf nodes to ensure correctness.

Originally, the MetaShift contains a collection of 12,868 sets of natural images across 410 classes. After matching, we selected 5,040 sets of images across 261 classes, where each tag of it can be found in the ImageNet-2012 dataset. To verify the coverage of the generated dataset over the ImageNet-2012, we count in the following methodology: for each meta-data of the regenerated dataset where the labels are a subset of the 1,000 labels of the ImageNet-2012, we locate the tag in the ImageNet hierarchy. If it is a non-leaf node, then mark all of its leaf nodes, otherwise mark the leaf node itself. After matching all tags, we count the number of the marked leaf nodes, which represent the class labels of the ImageNet. The 261 classes match to 120 leaf nodes to the ImageNet directly, and the remainint 141 classes remains to be the non-leaf nodes, which will cover a larger scale of leaf nodes. Totally, we get 867 leaf nodes marked, which means most of the 1,000 labels of the ImageNet are included in the dataset we generate. The unmatched portions of our datasets can be potentially used for OOD (out-of-distribution) detection, and we will delay it in future work.

4. Task Construction under the MetaShift

The 261 classes over 5,040 sets of images provide enumerable options for task construction. We can select two classes of the dataset and do a binary classification task on it. Here we represent a method to construct the tasks with the MetaShift: We first filter the classes whose subsets are less than a threshold. For the selected classes, we find their parent nodes in the ImageNet hierarchy. The highest hierarchy of the common parents in the ImageNet of two classes can be used to evaluate their "breed" similarity. The higher the hierarchy is, that is, the more closed the node is to the root, the less similarity there is between the two classes. The construction method is simple, and we can also change the magnitude of the "breed" similarity to make the classification tasks more challenging.

We present a specific example of constructing the tasks. We first filter the classes with subsets less than 5. Considering the 20 objects under the root in the second hierarchy in the ImageNet, which include *menu, carbonara, bubble, vegetable* and other categories, we intuitively construct tasks under these 20 categories (demand the common parents of the two classes to be either the 20 objects or their children), and get 19,024 binary classification tasks with the MetaShift.

In each task, we can evaluate both (1) domain generalization and (2) subpopulation shifts in a well-annotated (explicit annotation of what drives the shift) and well-controlled (easy control of the amount of distribution shift) fashion.

- In *domain generalization*, the train and test distributions comprise data from related but distinct domains. This arises in many real-world scenarios since it is often infeasible to construct a comprehensive training set that spans all domains. To simulate this setting, we can sample two distinct collections of subsets as the train domains and the test domains respectively (e.g. bathroom vs. outdoor contexts). To adjust the magnitude of the shift, we can fix the test domains and change the train domains with different distances. For example, if we use cats-in-living-room as the test set, then this is a smaller distribution shift.
- In *subpopulation shifts*, the train and test distributions are mixtures of the same domains, but the mixture weights change between train and test. It is also an important setting since ML models are often reported to perform poorly on under-represented groups. To simulate this setting, we can sample the training set and test set from the same subsets but with different mixture weights. To adjust the magnitude of the shift, we can use different mixture weights for the training set while keeping the test set unchanged.

Given the enumerable tasks and the diverse distribution shifts across each task, we can use MetaShift as a benchmark to evaluate the performance of ImageNet models across tasks and distribution shifts.

5. Conclusion

We present MetaShift—a collection of 12,868 sets of natural images from 410 classes—as an important dataset with heterogeneous contexts. MetaShift contains diverse natural data shifts and provides explicit explanations of what is unique about each of its data sets and a distance score that measures the amount of distribution shift between any two of its data sets. Matching the meta-data tags in MetaShift to ImageNet, we can construct a new dataset with 5,040 sets of natural images across 261, where the labels are a subset of the 1,000 labels of ImageNet-2012. We present methods to construct binary classification tasks over MetaShift and propose that it can evaluate the robustness of ImageNet model across distribution shifts.

References

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Guo, Y., Codella, N. C., Karlinsky, L., Codella, J. V., Smith, J. R., Saenko, K., Rosing, T., and Feris, R. A broader study of cross-domain few-shot learning. In *European Conference on Computer Vision*, pp. 124–141. Springer, 2020.

He, Y., Shen, Z., and Cui, P. Towards non-IID image classification: A dataset and baselines. *Pattern Recognition*, 110, 2020.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Beery, S., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. WILDS: A benchmark of in-the-wild distribution shifts. *CoRR*, abs/2012.07421, 2020.

Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J. B., Larochelle, H., and Zemel, R. S. Meta-learning for semi-supervised few-shot classification. In *International Conference on Learning Representations*, 2018.

Saenko, K., Kulis, B., Fritz, M., and Darrell, T. Adapting visual category models to new domains. In *ECCV (4)*, volume 6314 of *Lecture Notes in Computer Science*, pp. 213–226. Springer, 2010.

Santurkar, S., Tsipras, D., and Madry, A. Breeds: Benchmarks for subpopulation shift. In *International Conference on Learning Representations*, 2020.

Triantafillou, E., Zhu, T., Dumoulin, V., Lamblin, P., Evci, U., Xu, K., Goroshin, R., Gelada, C., Swersky, K., Manzagol, P.-A., et al. Meta-dataset: A dataset of datasets for learning to learn from few examples. *arXiv preprint arXiv:1903.03096*, 2019.

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pp. 5385–5394. IEEE Computer Society, 2017.

Worrall, D. E., Garbin, S. J., Turmukhambetov, D., and Brostow, G. J. Harmonic networks: Deep translation and rotation equivariance. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 5028–5037, 2017.

A. Additional Dataset Information

For each image class (e.g. *Dogs*), the MetaShift dataset contains different sets of dogs under different contexts to represent diverse data distributions. The contexts include presence/absence of other objects (e.g. *dog with frisbee*). Contexts can also reflect attributes (e.g. *black dogs*) and general settings (e.g. *dogs in sunny weather*). These concepts thus capture diverse and real-world distribution shifts. We list the attribute and general location contexts below.

A.1. General location and attribute contexts

A.1.1. General location contexts

Figure 3: **Example subsets based on general contexts** (the global context is stated in parenthesis). MetaShift covers global contexts including location (e.g., indoor, outdoor) and weather (e.g., sunny, rainy).

```
GENERAL_CONTEXT_ONTOLOGY = {
    'indoor/outdoor': ['indoors', 'outdoors'],
    'weather': ['clear', 'overcast', 'cloudless', 'cloudy', 'sunny', 'foggy', 'rainy'],
    'room': ['bedroom', 'kitchen', 'bathroom', 'living room'],
    'place': ['road', 'sidewalk', 'field', 'beach', 'park', 'grass',
              'farm', 'ocean', 'pavement',
              'lake', 'street', 'train station', 'hotel room',
              'church', 'restaurant', 'forest', 'path',
              'display', 'store', 'river', 'yard',
              'snow', 'airport', 'parking lot']
}
```

Figure 4: **The general contexts and their ontology in MetaShift.** MetaShift covers 37 general contexts including location (e.g., indoor, outdoor, ocean, snow) and weather (e.g., couldy, sunny, rainy).

A.1.2. ATTRIBUTE CONTEXTS

Figure 5: **Example Subsets based on object attribute contexts** (the attribute is stated in parenthesis). MetaShift covers attributes including activity (e.g., sitting, jumping), color (e.g., orange, white), material (e.g., wooden, metallic), shape (e.g., round, square), and so on.

```
ATTRIBUTE_CONTEXT_ONTOLOGY = {
 'darkness': ['dark', 'bright'], 'dryness': ['wet', 'dry'],
 'colorful': ['colorful', 'shiny'], 'leaf': ['leafy', 'bare'],
 'emotion': ['happy', 'calm'], 'sports': ['baseball', 'tennis'],
 'flatness': ['flat', 'curved'], 'lightness': ['light', 'heavy'],
 'gender': ['male', 'female'], 'width': ['wide', 'narrow'],
 'depth': ['deep', 'shallow'], 'hardness': ['hard', 'soft'],
 'cleanliness': ['clean', 'dirty'], 'switch': ['on', 'off'],
 'thickness': ['thin', 'thick'], 'openness': ['open', 'closed'],
 'height': ['tall', 'short'], 'length': ['long', 'short'],
 'fullness': ['full', 'empty'], 'age': ['young', 'old'],
 'size': ['large', 'small'], 'pattern': ['checkered', 'striped', 'dress', 'dotted'],
 'shape': ['round', 'rectangular', 'triangular', 'square'],
 'activity': ['waiting', 'staring', 'drinking', 'playing', 'eating', 'cooking', 'resting',
              'sleeping', 'posing', 'talking', 'looking down', 'looking up', 'driving',
              'reading', 'brushing teeth', 'flying', 'surfing', 'skiing', 'hanging'],
 'pose': ['walking', 'standing', 'lying', 'sitting', 'running', 'jumping', 'crouching',
           'bending', 'smiling', 'grazing'],
 'material': ['wood', 'plastic', 'metal', 'glass', 'leather', 'leather', 'porcelain',
           'concrete', 'paper', 'stone', 'brick'],
 'color': ['white', 'red', 'black', 'green', 'silver', 'gold', 'khaki', 'gray',
           'dark', 'pink', 'dark blue', 'dark brown',
           'blue', 'yellow', 'tan', 'brown', 'orange', 'purple', 'beige', 'blond',
           'brunette', 'maroon', 'light blue', 'light brown']
}
```

Figure 6: **The attributes and their ontology in MetaShift.** MetaShift covers over 100 attributes including activity (e.g., sitting, jumping), color (e.g., orange, white), material (e.g., wooden, metallic), shape (e.g., round, square) and so on.

B. Related Work

Existing Benchmarks for Distribution Shift Distribution shifts have been a longstanding challenge in machine learning. Early benchmarks focus on distribution shifts induced by synthetic pixel transformations. Examples include rotated and translated versions of MNIST and CIFAR (Worrall et al., 2017); surface variations such as texture, color, and corruptions like blur in Colored MNIST (Gulrajani & Lopez-Paz, 2020), ImageNet-C (Hendrycks & Dietterich, 2019). Although the synthetic pixel transformations are well-defined, they generally do not represent realistic shifts in real-world images that we capture in MetaShift.

Other benchmarks do not rely on transformations but instead pull together data across different experiments or sources. Office-31 (Saenko et al., 2010) and Office-home (Venkateswara et al., 2017) contain images collected from different domains

like Amazon, clipart. These benchmarks typically have only a handful of data distributions. The benchmarks collected in WILDS (Koh et al., 2020) combine data from different sources (e.g., medical images from different hospitals, animal images from different camera traps). Similarly, some meta-learning benchmarks (Triantafillou et al., 2019; Guo et al., 2020) focuses on dataset-level shift by combining different existing datasets like ImageNet, Omniglot. While valuable, they lack systematic annotation about what is different across different shifts. (Santurkar et al., 2020; Ren et al., 2018) utilize the hierarchical structure of ImageNet to construct training and test sets with disjoint subclasses. For example, the "tableware" class uses "beer glass" and "plate" for training and testing respectively. Different from their work, we study the shifts where the core object remains the same while the context changes. NICO (He et al., 2020) query different manually-curated phrases on search engines to collect images of objects in different contexts. A key difference is the scale of MetaShift: NICO contains 190 sets of images across 19 classes while MetaShift has 12,868 sets of natural images across 410 classes.

To sum up, the advantages of our MetaShift are:

- Existing benchmark datasets for distribution shifts typically have only a handful of data distributions. In contrast, our MetaShift has over 12,868 data distributions, thus enabling a much more comprehensive assessment of distribution shifts.
- Distribution shifts in existing benchmarks are not annotated (i.e. we don't know what drives the shift) and are not well-controlled (i.e. we can't easily adjust the magnitude of the shift). The MetaShift provides explicit annotations of the differences between any two sub-datasets, and it quantifies the distance of the shift.