A Scalable Framework for Product Image Classification applied to Home Improvement E-commerce

Tianlong Xu
The Home Depot
Atlanta, Georgia, USA

Le Yu
The Home Depot
Atlanta, Georgia, USA

Yuanbo Wang
The Home Depot
Austin, Taxes, USA

Morgan White
The Home Depot
Atlanta, Georgia, USA

Estelle Afshar
The Home Depot
Atlanta, Georgia, USA

ABSTRACT

In this paper, we propose a highly scalable product image type classification framework that is designed to manage visual assets on home improvement retailer online platforms. With the continuous growth of online business and the expansion of catalog sizes, we find that the amount and diversity of product-related data continue to expand. In this research, we focus on one specific element of product data: images. The manual selection of the most relevant image to enable a specific customer experience (e.g. inspiration, advertisement, comparison, etc.) is impossible. Even when the product is known, we recognized that images could be described along different aspects such as their content (room scene, white background, line art, etc.), the product view (front-facing, angled views, etc.), the annotations, etc. Such descriptors are foundational elements in choosing the best images for many applications. Our solution consists of designing an ontology to define visual concepts and organizing their relationships. This ontology determines the series of classifiers we trained to predict multiple labels that define various image types. The classifiers are trained by leveraging both deep learning (fine-tuned convolutional neural networks with Siamese network triplet loss) and traditional computer vision (local pattern feature extraction) techniques. Besides, we further improve the prediction accuracies by using an active learning approach to select highly informative training data. Our latest models indicate accuracies of predicting the correct labels ranging from 84% to 98%. To automate the classification process, we developed a highly scalable production pipeline that predicts tens of millions of images, in parallel, in a matter of hours on a weekly basis. We also demonstrated the benefits of the proposed framework through three business applications where selecting the best images played a critical role in improving powering customer experiences.

DLP-KDD 2021, August 15, 2021, Singapore

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/1122445.1122456

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Neural networks**; **Active learning settings**; • **Information systems** → **Enterprise applications**.

KEYWORDS

image classification, ontology system, neural networks, active learning, multi-classifier pipeline, e-commerce, enterprise applications

ACM Reference Format:
Tianlong Xu, Le Yu, Yuanbo Wang, Morgan White, and Estelle Afshar. 2021. A Scalable Framework for Product Image Classification applied to Home Improvement E-commerce. In *Proceedings of DLP-KDD 2021.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Online retailers and more broadly e-commerce rely on content such as text description and specifications to properly present products. In addition, visual media (photos, videos, AR, etc.) is playing an increasing role in providing factual information when the appearance and aesthetics of the products are important factors of the customers selection, which is particularly relevant in fashion, home décor, art, automobile, real estate, etc. Websites usually provide multiple images of a product in order to create an enhanced experience and each type of images shows a different aspect (Figure 1). Customers gain a more comprehensive understanding of the product from these multiple perspectives, seeing a single product view ("silo", see appendix for more definitions), a close-up that highlight details, the product dimensions, or a lifestyle picturing the product in the context of a scene. Another benefit of knowing the type of image available for a product is in the selection of the best image as an input of an algorithm that powers applications such as visual search and recommendations. The definition of the best images varies in function of the application, for example the product dimensions are added to silo images , the marketing campaigns favor inspirational lifestyle images. This paper addresses the prerequisite needed to select the best images for each application, by providing a solution to accurately label all images with standardized descriptive concepts we call "image types".

Given the size of online catalogs, human labeling is certainly not conceivable. We initially thought about training a single Convolutional Neural Networks (CNN) classifier to automate the prediction

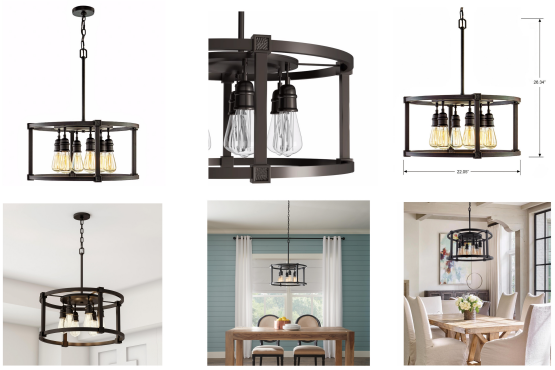

Figure 1: Product images of a chandelier. From left to right in the top row, image types are: silo-front; close-up; dimensions. Bottom row: lifestyle product, lifestyle limited scene, lifestyle full room.

of labels. However, due to the extreme diversity of product categories and image types, we faced the major challenge of generalization of CNN models to the most typical cases. Another challenge was the granularity of the labels: for some applications, particular concepts need to be further refined to distinguish nuances so we created sub-concepts (e.g. lifestyle is further described as product-only lifestyle, limited-scene lifestyle, and full-room lifestyle). These observations motivated us to devise an ontology of image types that define a set of concepts that can used across product categories. This ontology supervises and outlines the development of our framework. Also, we further improved the results of the initial CNN approach by developing a sequence of models combining Siamese network loss functions, CNN fine-tuning, traditional computer vision image features to reach higher model accuracy. Finally, we integrated an active learning framework to effectively select the most informative and representative images in the training data, to further boost model performance in each training round.

2 RELATED WORK

Product images are crucial assets for e-commerce platforms. Studies have shown that maintaining a high standard of product images has an extremely positive effect on keeping customer engagements [6] and product popularity [31]. In terms of leveraging visual assets, extensive applied data science research has primarily been focusing on visual search (e.g., Microsoft [2][11], Alibaba [33], Amazon [36], Facebook [27], Pinterest [23]) and visually similar product recommendation (Google [28], Alibaba [33], Ipkart [22], Amazon [36]). To power visual search and visually similar product recommendation, image classification on product category classes is always performed as a foundation (Facebook [13][1], Fashion classification [19][16]). While classifying the product category classes, some have documented the challenges caused by diverse image types. For instance, Bergamo [2] and Zhang [33] mentioned some misclassifications due to taking lifestyle images which contain a lot of background information that led to poorer results. This inconvenience can be alleviated by making the primary product salient (e.g., performing regional image cropping) as discussed by Li [15] who reported a five-class image type classification approach which served as an important image filter to support their "complete the look" project. In general, existing literature on e-commerce images have been dominated by the identification of products while solutions to tackle the description of the type of image remain very limited. The motivation of this work is to propose a systematic approach of describing the diverse type of e-commerce images in order to best support different applications.

With the rapid increase of computing power, image classification approaches have been transformed in the past decade by the evolution and success of Convolutional Neural Network (CNN). New CNN architectures kept being invented (i.e., AlexNet [14], GoogLeNet [26], VGG series [24], ResNet [10], Inception [25], Xception [4], and Squeeze & Excitation [12]), each outperforming the previous one and pushing the state-of-the-art accuracies to the next level. These CNN architectures contain abundant image features that are extremely useful for identifying the inherent image patterns and thus distinguishing different image types. Apart from these, additional insights can be generated by traditional computer vision approaches such as HOG [5] and SIFT [17] that effectively identify local image patterns. Studies have also shown that local features contain abundant latent information, which are supportive for image classification [32]. To achieve optimal outcomes of the classification tasks, we implemented an aggregated feature set that takes advantage of both CNN (via transfer learning) and traditional computer vision techniques.

With the rapid growth of CNN in image classification, active learning (AL) has become increasingly popular as it helps to obtain better informative and representative training samples to improve the performance of algorithms. Gal et al. [8] added a prior on the weights of neural networks, sample the weights from the dropout distribution at test time, and calculate the informativeness score of an unlabeled image based on the variation ratio of prediction results as a way to select the top-ranked-score images for manual labeling before adding them into the training data for the next training round. The main drawback of this method is that similar images could be selected in real system or it might favor bringing more images from certain categories. Elhamifar et al. [7] took advantage of convex programming to select informative and diversity examples with quadratic complexity. Future more, Sener et al. [21] applied the core set and a greedy approximation to find the most distant image from the training set. In [34] k-means clustering method was used as a proxy to select a subset of diversity examples from top-ranking examples with low margin scores. Zhu et al. [35] proposed an active learning based framework that leverages domain expertise and training data annotation to power product type classification for e-commerce, which boosts significant search efficiency. Our active learning approach is based on submodular optimization[29], and more importantly it is scalable to large datasets by applying feature-based selection method and narrowing down the selection pool via pre-filtering samples.

3 FRAMEWORK OF THE IMAGE TYPE CLASSIFIER PROJECT (ITC)

The framework built for the Image Type Classifier project is outlined in Figure 2. It is composed of an offline training, monitoring

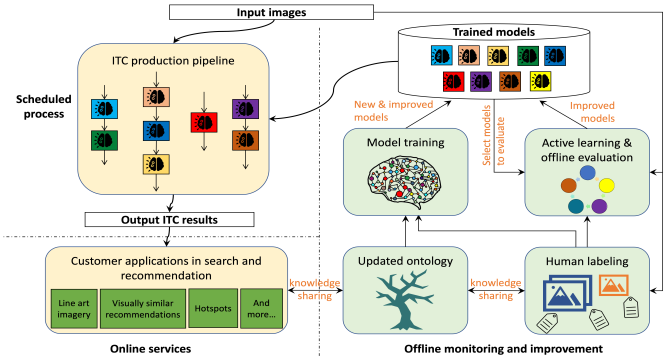

Figure 2: Illustration of the ITC project framework

and model improvement, and a production pipeline which is scheduled to run on a regular basis, for example daily or weekly, to serve the online services that enable customer experiences. Each component will be presented in detail in the next few sections. We are summarizing below the major contributions of our system:

- We have developed a highly scalable framework to solve a very challenging business problem related to the diverse image types. To properly characterize an image and organize the image types, we have created an expandable **Image Type Ontology**, which outlines the hierarchy of the image type concepts and relationships.
- Instead of relying on a large scale all-in-one classifier that might have relatively poor performances, we created a novel framework that leverages **multiple classifiers** and organizes them separately in specified order (launching or detaching any classifier won't affect others in service).
- To perform offline model evaluation and model improvement, we have integrated an **active learning** component. By applying several advanced sampling methods on both existing and newly generated label sets during model retraining, active learning helps boost the model accuracy significantly.
- To use ITC models in **production**, we built a highly scalable pipeline that automatically loads pretrained models and processes in parallel tens of millions of images in hours. Different sets of models can be assigned to different categories, which tailors to different business needs.
- Online services of ITC support several **e-commerce applications** such as alternative and complementary recommendations, visual search and generation of enhanced image asset by displaying product dimensions. ITC predictions are critical as input of customer experience and algorithms. Feedback from the owners of these algorithms also loops back in the offline processes and they are very valuable for growing ontology graphs and model training.

4 IMAGE TYPE ONTOLOGY

At the core of offline modules, the role of the expandable Image Type Ontology is to define the concepts that are used to characterize an image and organize their relationships. As the major challenge of building a highly representative ITC system, the complexity of image types not only come from the size and diversity of the catalog, but also from the inherent relationships of different image types. As we explore the image database, we continuously discover new and unique image type and expand the concepts of the ontology. So far, we have identified about 30 unique concepts, or image types. Some of these concepts are mutually exclusive, while others are sub-concepts which inherent from their parents and add additional granularity. While the majority of concepts are concrete and translate into a label, some concepts are only intermediate abstracts that do not translate into an actual label (e.g. angles). Based on the applications (online home improvement retailers) we are supporting, the Image Type Ontology we have created has three mains branches: Content, Annotation (Fig. 3) and View (Fig. 4). Ultimately, an image can be described by multiple concepts, for example the content or product view, and under each concept, we developed a series of class names that we want to classify product images into (illustrated in Figure 3).

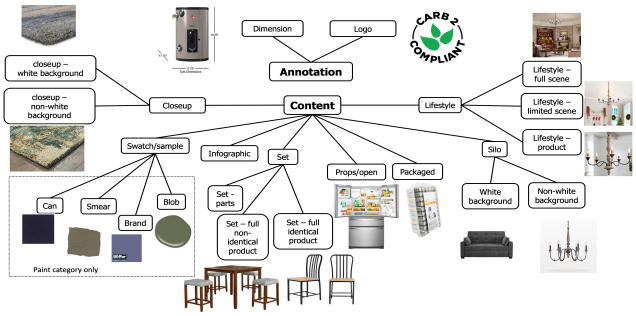

Figure 3: Representation and illustration of the concepts of image "Content" and "Annotation". Under these concepts, images can be classified into close-up, silo, set, lifestyle, dimensions, and logo, etc. based on the diverse possible composition of the image (some concepts evolve to higher degrees of granularity). See Appendix for detailed definitions of all image types.

The most represented concept in the image catalog we have used is "silo" (almost 50%) and because of the multiple application that this type of image drive, we further split the concepts into subconcepts, sometimes with several level of granularity (Fig. 4). As the application grows more diverse to satisfy new business needs, the ontology evolves to incorporate new concepts or to develop image descriptors of higher-level granularities (e.g., distinguish the angled-right from angled-right view of some non-symmetrical products). The concepts and relationships of the ontology serve as a means of efficient media that connects human-defined image type taxonomy with machine learning labeling system.

5 ITC MODEL TRAINING AND EVALUATION

5.1 Model training and comparison

Based on the ontology described above, we can expect an online catalog to have a non-uniform distribution of image types across all product categories. Furthermore, some concepts only exist for

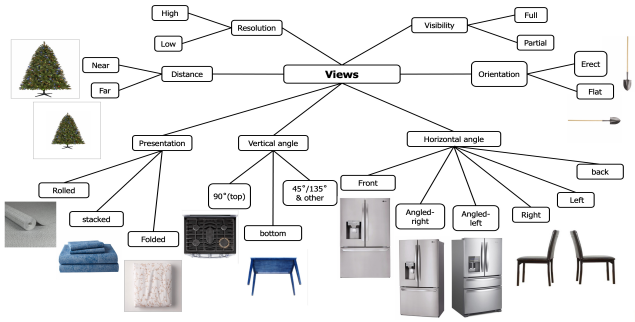

Figure 4: Representation and illustration of the concept of "View". Under this concept, an image can be described by several sub-concepts depending on the various stationary views the product is displayed in.

specific products (e.g. the "swatch/sample" concept is only relevant for covering surfaces such as paint, flooring, wall paper, etc.), and some concept do not apply to specific products (e.g., for symmetric products such as the Christmas trees, we do not need to distinguish the horizontal views). However, these following concepts are relatively universal and can be found in (almost) all product categories: lifestyle, closeup, dimensions, and silo.

To handle this situation, we have developed two types of ITC models: the category-specific models and the generic models. The category-specific models are trained on small human-labeled datasets sampled (usually 200 instances per label class) from categories of products that share similar shapes (e.g. dinning chair, patio chair, office chair), and are designed to obtain higher accuracies for granular concepts (e.g. chair views such as front, angled-right, angled-left, back, top, and bottom, etc.). The generic models are trained on a large human-labeled dataset (e.g. 50k images) that are stratified samples of images across all product categories so the trained models are robust to the variation of product types. The generic models cover about half the image types that apply to almost all product categories. The category-specific models aim to classify images into more detailed patterns and much higher granularity, while the generic models aim to capture some high-level characteristics of image types across all categories.

The ITC model training consists of two stages. The first stage is fine-tuning the pretrained state-of-the-art models (e.g., VGG16, VGG19, ResNet, and GoogLeNet etc.), which has already proven to carry rich features. Guided by [18], we implemented the fine-tuning by altering some of these pretrained CNNs (e.g., VGG16, VGG19, and ResNet usually win our lab tests). Specifically, we replace the last dense layer with a new dense layer and then retrain the new network as shown in Figure 5. During the training process, Siamese network and triplet loss minimization [9][20] were employed for their robustness in learning semantic similarity and distinguishing between subtle nuances. The loss function is detailed as below.

$$Loss = \sum_{i=1}^{N} [\|f_i^a - f_i^p\|_2^2 - \|f_i^a - f_i^n\|_2^2 + \alpha] \quad (1)$$

where $\|f_i^a - f_i^p\|^2$ is the cosine similarity distance between anchor and positive input; $\|f_i^a - f_i^n\|^2$ is the cosine similarity distance between anchor and negative input; α is a constant. During the training, two same-structure CNNs sharing the same weights are configured. For each iteration of the training process, every input image in the same batch has a chance to be selected as an anchor (the chairs with yellow border in Figure 5), which is randomly assigned a positive image (image of the same classification class, e.g., the blue-bordered chair of Scenario 1 in Figure 5) and a negative image (image of a different classification class, e.g., the blue-bordered chair of Scenario 2 in Figure 5). All three images are passed through the networks and the outputs of the top pooling layer are used to compute the cosine similarity distances. By minimizing the averaged distance difference, the network learns to adjust the CNN weights towards a direction to differentiate different classes. After this process, we have a transferred CNN (the yellow or blue neural network in Figure 5) that predicts image features as output vectors.

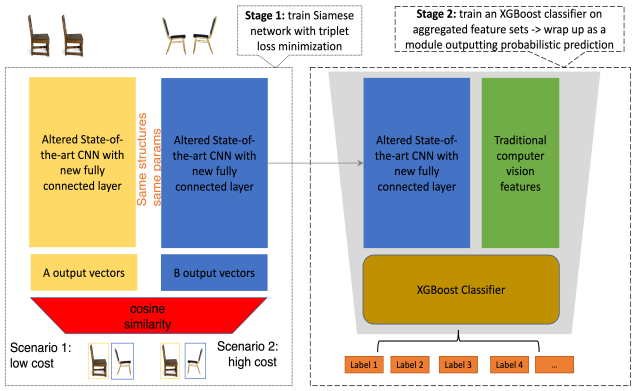

Figure 5: Key components in ITC model training. Major steps here are 1) train a finetuned CNN that output features representing latent image type characteristics; 2) train a classifier (e.g., XGBoost) that ultimately predicts image types with the CNN features and some additional computer vision features.

To further enrich the feature set, we also examine some additional image features based on traditional computer vision and concatenate these features with those extracted from the fine-tuned models (Stage 2). For instance, knowing whether a product intersects the image border will potentially help distinguish silo images (no intersection) from lifestyle and closeup images (where intersection is very common). Number of disconnected objects in the image is another important factor to predict whether or not an image belongs to the "set" class. Other features such as the dominant color pixels and white space area percentage are also very strong predictors in identifying some classes (e.g., silo). These merged feature sets are fed into classification models to build classifiers that output probabilistic predictions. One of the most frequently used classifiers is XGBoost [3] due to its enormous advantages (i.e., scalable, portable, and highly distributed nature, see [30]). Note that the reason of applying aggregated feature set is to perform an exhaustive exploration of all possible predictive features that might contribute, despite it being plausible that the traditional computer vision features are already covered by CNN latent features.

Table 1: Coverage and tested accuracy of generic models.

Generic models	Covered ontology classes	Overall test accuracy: baseline model	Overall test accuracy: Siamese model
Generic model#1	Closeup, lifestyle, silo	85%	93%
Generic model#2	All silo views (except resolution, distance, and orientation classes)	79%	84%

Prior to adopting this approach, we have tried building baseline models, finetuning single pretrained CNNs with new dense (same as that in the Siamese network) and softmax layers. We obtained some good results (see Table 1) but by applying the Siamese approach, we keep elevating our benchmarks as models trained with this approach outperform those from the single finetuned networks.

During every model training and testing tasks, we randomly split every human-labeled dataset into testing set (20%), and training-validation set (80%). We perform Gridsearch cross validation on the training-validation set to train classification models and search the best hyperparameters combinations for both the transferred CNN (i.e., learning rate, number of epochs, size of batch, dropout percentage, etc.) and XGBoost (e.g., size of the tree, regularization coefficients, sampling strategies, etc.). Then we use the testing set (which data has never been seen by any model) to report the generalized accuracy score. Our latest test accuracies of the current generic models range from 84.3% to 93.0% (Table 1), while those for the category-specific models are from 86.0% to 98.0%, varying by category (Table 2). Due to fact that generic models have to learn the inherent data features of a mixture of diverse categories, they always perform not as accurate as the category-specific ones which tackle cases in only one single category. We are currently leveraging both models but the question of how to achieve an optimal balance between accuracy and coverage is still open and worth further exploration.

For some important categories, we have performed human auditing on the ITC results prior to utilizing ITC results for business applications (e.g., digital asset management). Business users are particularly interested in how often they can trust an ITC predicted result (relevant evaluation metric: precision). Table 3 presents the precision results based on 400 examples randomly selected for each class from three categories. It demonstrates that for these a few categories, ITC predicted results are trustable with very high confidence levels (most precision scores are above 85%).

5.2 Active Learning integrated model evaluation and improvement

Although Deep Learning models have shown unprecedented success in many areas of computer vision and pattern recognition, one big challenge is obtaining a large amount of labeled data to train the parameters, or to fine-tune a pretrained model. Labeling a dataset can be expensive and time consuming. Active learning

Table 2: Coverage, test accuracy of category-specific models.

Model category	Covered ontology classes	Overall test accuracy
Area Rugs	Closeup non-white background, closeup white background, lifestyle full room, lifestyle limited scene, lifestyle product, silo white background, sketch	87%
Chairs/Tables	Silo-horizontal angle views, silo-vertical angle views, open, infographic	93%
Luggage	silo-vertical angle views, silo-horizontal angle views, open, graphic, sketch, packaged, set	86%
Décor Paint	Swatch-can, swatch-smear, swatch-brand, swatch-blob	100%
Small Electrics	Silo-front, silo-side, top, open	88%
Cabinet Knobs	silo-vertical angle views, silo-horizontal angle views	98%

Table 3: Human audited model performance results (Generic model#1).

Category name	Name of class	Precision of class	Percentage of images	Overall precision
Faucet	Closeup	83%	8.8%	91%
	Lifestyle	84%	23.9%	
	Silo	95%	61.4%	
Vanity	Closeup	97%	7.6%	88%
	Lifestyle	79%	30.6%	
	Silo	92%	52.0%	
Interior Furniture	Closeup	97%	12.5%	95%
	Lifestyle	88%	25.5%	
	Silo	99%	54.9%	

is a framework to help in reducing the amount of labeled data required to train deep learning models while maintaining the same performance or achieving even better performance. As shown in Figure 6, every iteration of the active learning process we developed starts with a small amount of cleaned data and then one or more samples are selected from a larger unlabeled pool using certain sampling methods. The next step is to perform human labeling on the selected samples, the results of which will be added into the original training data for model training at the next iteration until ending criterion (budget or performance) is met.

In our active Learning application system, we would like to address the following key points.

Batch aware: Traditional active learning algorithms select one best sample to be annotated at a time. After the label is obtained, the model retrains and then selects the next sample. However, on one hand, deep neural networks are computationally heavy, and it may take hours to train the model. Sampling one example at each round may not be reasonable for most practical systems. On the other hand, comparing to classical machine learning algorithms (e.g., SVM), the confidence score output from the final "softmax" layer of neural network model tends to be overconfident and one

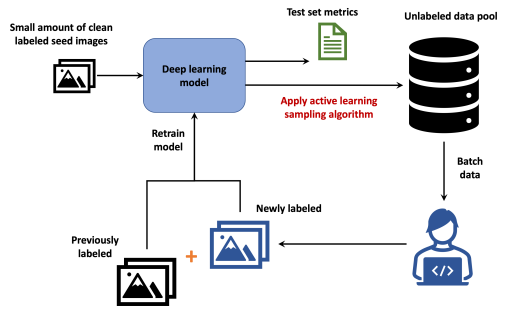

Figure 6: Illustration of the active learning pipeline.

example may not give enough information. Instead, selecting a batch of samples is recommended for neural network models.

Informative and representative: The goal of active learning is to select informative and representative samples to retain as many features of the larger dataset without knowing the "ground truth" as possible. Margin-based sampling method is a commonly used to measure the informativeness of unlabeled data:

$$f_i = P(\hat{y_1}|i) - P(\hat{y_2}|i) \tag{2}$$

where f_i represents the informativeness of sample i, $P(\hat{y_1}|i)$ and $P(\hat{y_2}|i)$) are the confidence scores of the highest possible class and second highest possible class predicted by the model for sample i, respectively. Samples with smaller difference between the top two label probabilities are more likely to be selected. In other words, instances with small margin are more ambiguous. However, one drawback of this method is that if we just select the top-ranking samples via margin-based sampling, we may obtain many similar examples or majority of the selected samples are from a certain category. Especially, image datasets of e-commerce are quite big and highly skewed with many duplicate images used by different products from the same supplier.

Considering the large size of unlabeled data pool in reality, we extend the feature-based submodular function provided in [29] for selecting a subset of diverse images instead of speech data. The general form of a feature-based submodular function is:

$$f(X) = \sum_{d=1}^{D} g(\sum_{i=1}^{N} X_{i,d}) \tag{3}$$

Where $g(*)$ is concave function operating on a subset X that has N selected examples and D dimension of features. Maximizing the objective function naturally encourages the diversity and coverage of the features within the selected dataset. Our new method includes extending the feature-based submodular function for image subset selection by leveraging the features discussed in Figure 5. To ensure informativeness of samples, a pre-filtered candidate pool is selected via margin-based sampling, and such strategy is named margin-based submodular sampling.

5.2.1 ***Case Study.*** We evaluate the margin-based submodular sampling as well as 3 other sampling methods, namely random sampling, pure margin-based sampling and margin-based K-Means sampling[34] on dataset (size: 12k) with three selected high-frequency classes: lifestyle, silo, closeup. 600 images are selected as validation dataset for evaluating the performance. At the first round, a VGG19 model is fine-tuned based on 600 images. Then, for each sampling method, at each iteration, a batch of 600 images are selected and added into the original training data to retrain the VGG19 model. Each experiment is bootstrapped 10 times independently to get the average performance and confidence interval. For margin-based submodular and K-Means sampling methods, pre-filtered top 600*k (k>1) images are selected firstly by margin-based sampling, and then 600 images are selected via submodular function or clustering method from the pre-filtered pool. In our case study, k equals to 10 and active learning loop includes 7 iterations until the total training size reaches 4,200 examples. Square root function is used as the concave function. The vector outputted from the pre-final layer of VGG19 model is used as the feature vector in K-means clustering.

Figure 7 is the validation accuracy of each sampling method. The line shows the average accuracy over 10 runs, and the shade shows the 90% confidence interval band around the mean. We can see that all active learning methods significantly outperform random sampling method. Methods combining informative and representative sampling outperform the method with only informative sampling. Our proposed method, margin-based submodule sampling, is better than margin-based K-Means clustering method. The validation accuracy is about 1% higher at each iteration after the initial round. Performance of model trained on 1,600 examples selected via margin-based submodule method is better than that of model trained on 4,200 examples selected randomly.

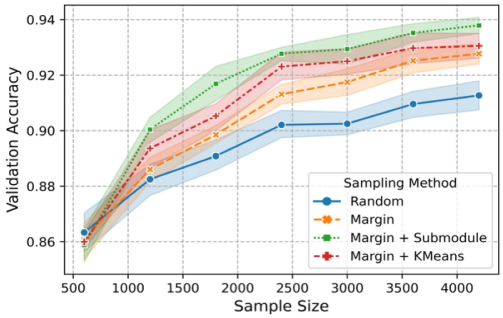

Figure 7: Validation accuracy on a generic labeled dataset.

We also compare the average sample selection time of margin-based submodule and margin-based K-Means at each iteration (except the first round) on CPU. From Table 4, we can see that margin-based submodule selection is faster than margin-based K-Means due to the quadratic complexity of K-Means clustering.

Table 4: Average selection running time comparison between margin-based submodule and margin-based K-Means at each iteration.

Model	Sampling size	Average time(s)
Margin + Submodule	600 over 6000	23.85
Margin + K-Means	600 over 6000	30.75

6 CONFIGURABLE ITC PIPELINE TO OVERCOME THE HETEROGENEITY OF IMAGE TYPES AMONG CATEGORIES

To automate the ITC models, we have built a highly scalable production that automatically loads pretrained models that are created with specific formats. Each model takes image URLs as input and outputs a prediction result decoded from its label set. To properly organize these models while productizing them, we developed configuration scripts to specify the positions of specific models in the production pipeline. Therefore, every image is routed to feed into a sequence of pretrained ITC models depending on the product category it belongs to, which is controlled by the configuration scripts. The pipeline (shown in Figure 8) is highly scalable in several aspects.

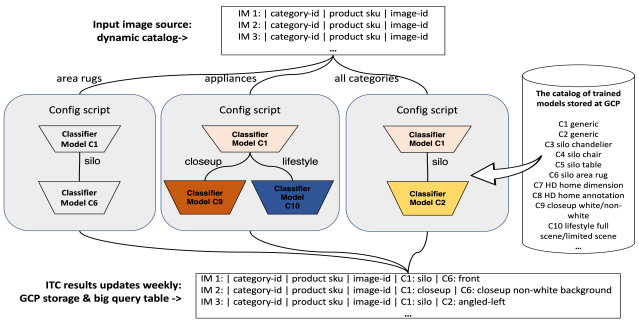

Figure 8: ITC framework and pipeline. The input source is dynamic since the online platform keeps adding new products and removing discontinued products. The pipeline directs every image to follow a specific prediction route, specified by the configuration scripts, and distribute the prediction results on weekly frequency.

- The pipeline allows old models to be replaced by new ones, so we keep publishing best-to-date models as we obtain more data or as we identify more advanced training strategies.
- The pipeline allows user-specified prediction routes, which can be managed by the configuration scripts. This will make possible same image being classified based on different ontology concepts and structures. When we update an ontology structure, the outputs can update accordingly as we manipulate the configuration scripts.
- The pipeline manages multiple classifiers efficiently and these classifier models are independent of each other since they are configured to category-wise prediction routes.
- The development of new models can be “on-demand” and tailor to specific business needs.
- The pipeline distributes the computation workload into multiple instances (implemented with Apache Beam) so it can routinely process massive amount of data (tens of millions) in an acceptable period of time (a few hours).
- As an option, the pipeline can run in "delta mode", in which it processes only the images that are newly introduced to the catalog and avoids significant redundant computing workload as it classifies only a small portion of the image catalog (i.e., the "delta).

7 APPLICATIONS OF ITC

7.1 Case 1: Line art imagery (LAI)

One of our online data science algorithms that relies heavily on ITC is "line art imagery" which takes the silo-front or silo-angled images (depending on product catagory) filtered by ITC as inputs and subsequently attach product dimension information onto the product image (Figure 9). With such line art images displayed on e-commerce product information page, customers can quickly see key product measurements, visually connect them with their respective edges, and build confidence to make a buying decision. Customer survey has shown that about 5% returned items are directly related to misunderstandings of the product sizes. This feature is especially important for some high-visibility categories such as lighting, area rugs, and vanities. Business analytics have shown that since the launch of this project, it has been generating 2% incremental revenue and 9 bps incremental conversion in these categories (presenting LAI to customers vs otherwise).

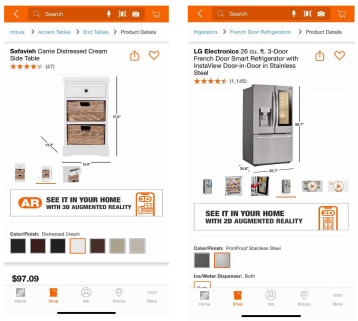

Figure 9: An Example of the LAI experience in two categories: end tables and refrigerators, where three dimensions (height, width, depth) of the products are displayed directly line art images.

As a critical backbone of LAI, ITC results are used to select as many silo-front and silo-angled images as possible to maximize the coverage. Lab experiment metrics suggest for a given category where such images are available, by 89% chances ITC generic models can successfully recommend the right image, while same metric for ITC category-specific models (if there is one for a given category) is 97%.

7.2 Case 2: more visually similar options (MVS)

One of our recommendation algorithms that takes ITC results as input is “more visually similar options” (e.g., see below for a real customer experience example). As a recommendation strategy, the MVS algorithm leverages visual information to offer alternative buying options to customers that are visually similar to the product being viewed. One critical step in the MVS pipeline is to select a "best image" for every product, which will be used to compute the pairwise visually similar score. The pairwise visually similar scores (cosine similarity scores of CNN image embeddings) are then ranked and for every product the top 5 alternative products that have similar visual appearances are recommended to customers.

Therefore, how often the right best image are selected is an important factor that affects the ultimate performance of MVS. In general, the MVS prefers silo-front or silo-angled images to be selected as best images since such images contain only the full product view without any additional information that might possibly trigger noise image embeddings. Figure 10 illustrates a sharp contrast in terms of the MVS performances when different types of images are selected as best images.

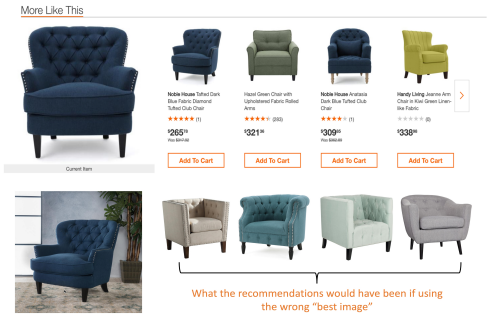

Figure 10: The first row is an example MVS customer experience. The image in the leftmost is the product being viewed by a customer, and the others represent the MVS recommendations. The second row is an example of a poorly performed MVS recommendations. The wrongly chosen best image in this case is a lifestyle image which, unlike silo-front or silo-angled images, contains too much background visual information so it results in irrelevant recommendations.

Prior to the development of ITC, the selection of best images is a rule-based approach (original baseline solution) that computes the white area percentage of every image of a product and the image with the highest value is picked as the best image. After leveraging ITC, significant improvement has taken place. Table 5 details a comparison of the best image selection quality between the original approach and the ITC involved approach (test category: interior furniture; sample size: 656 products). Therefore, ITC plays a critical role in supplying the correct images to e-commerce recommendation algorithms like MVS.

Table 5: Comparison of best image selection quality, before and after ITC get involved

	Baseline	ITC
Number of correct best images selected	551	638
Frequency of correct best images selected	83.9%	97.3%

7.3 Case 3: Hotspots

Another one of our recommendation algorithms currently in development that heavily utilizes ITC results is Hotspots, which aims to identify products within our product catalog based on information retrieved from lifestyle images (selected from ITC) (Figure 11).

Major steps of this algorithm are: 1) feed the lifestyle image into an object detection model which locates and categorizes the products of interest and crop the detected products of interest from the lifestyle images (preferably lifestyle full room where more categories of products can be detected); 2) retrieve candidate product images based on product categorization from the object detection model and the image classification from ITC; and 3) A Siamese network (also inspired from the ITC training) is then trained using the triplet loss function to generate similar embeddings from the cropped image and the matched product image and dissimilar embeddings from the cropped image and the remaining candidate product images. Consequently, when a cropped image containing a product of interest is searched against the product catalog, the product with the most similar embedding as the cropped image is retrieved from the list of potential candidates. This product will most likely be the best match for the product of interest in the cropped image.

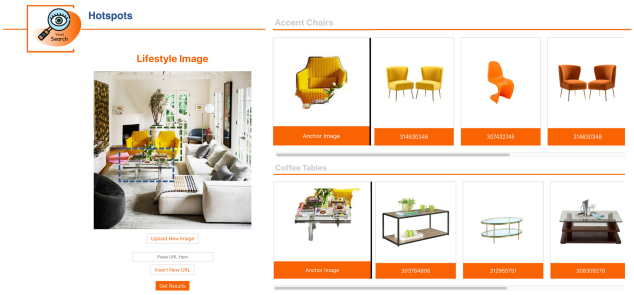

Figure 11: An Example of the Hotspots experience. In this example, detected accent chair and coffee table (anchors) are used to match against the products in our catalog and relevant results are recommended to customers.

ITC plays a critical role in the the Hotspots pipeline. Besides filtering lifestyle full room images during Step 1, it also helps identify the best product images to match with the customer's query (Step 2). Since each product has several types of images, it is important to identify the ones that do not contain unnecessary noise like background objects (i.e., silo, set, or closeup). Thus, the visual embeddings generated from these silo images will contain latent representations relevant to the key features of the product only. Additionally, ITC is crucial to creating the training sets for the Siamese Networks, since the training triplets require matching silo and lifestyle images as well as mismatches. Without the automatic results of ITC, we would not know which images belonging to a product are silo, and our only alternative would be to manually labeling images. In other words, ITC serves as an absolute prerequisite for Hotspots.

8 CONCLUSION AND FUTURE WORKS

In this paper, we have presented a highly scalable and sustainable image type classification framework that distinguishes diverse image types from a wide-range of products from a home improvement retailer online platform. As a comprehensive analysis of the mutual exclusive relationships of various concepts used to describe

the content and views of images, we have designed an extendable ontology to define and organize the variety of image characteristics. We keep growing the ontology graphs as we find more special cases and collect more feedback from business users. Guided by the ontology, we built robust ITC models by leveraging multiple techniques including deep learning, traditional computer vision approaches, as well as Siamese network triplet loss minimization. Trained models are evaluated offline and further improved by active learning approaches. To automate the classification process, we have developed a pipeline that process in parallel tens of millions of images routinely. The predicted labels are being used in production to enable customer experiences by rendering abundant product details and supplying sufficient best input images for recommendations and search. Lab experimentations have shown that ITC have strong potential to significantly boost the performances of many business applications such as LAI, MVS, and Hotspots.

As next steps of the continued development of ITC, we might consider the following a few areas of opportunities:

- **Personalization:** for different categories, leverage multi-armed bandit to automatically select the best image accordingly. For different customers, generate personalized types of images to display on the pages they are browsing.
- **Customer generated images:** apply ITC to customer generated images which can potentially help organize these user-contributed resources, which can be used to replenish missing types of images from the product catalog.

ACKNOWLEDGMENTS

We would like to express our special thanks to Xiquan Cui, Nian Yan, Mingming Guo, Jiaqi Wang, Haozheng Tian, and Lijiang Long for their insightful and constructive suggestions.

REFERENCES

[1] Sean Bell, Yiqun Liu, Sami Alsheikh, Yina Tang, Edward Pizzi, M. Henning, Karun Singh, Omkar Parkhi, and Fedor Borisyuk. 2020. GrokNet: Unified Computer Vision Model Trunk and Embeddings For Commerce. 2608–2616. https://doi.org/10.1145/3394486.3403311

[2] Alessandro Bergamo and Lorenzo Torresani. 2010. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. *Advances in neural information processing systems* 23 (2010), 181–189.

[3] Tianqi Chen and Carlos Guestrin. 2016. XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Aug 2016). https://doi.org/10.1145/2939672.2939785

[4] François Chollet. 2017. Xception: Deep Learning with Depthwise Separable Convolutions. arXiv:1610.02357 [cs.CV]

[5] N. Dalal and B. Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*.

[6] Wei Di, Neel Sundaresan, Robinson Piramuthu, and Anurag Bhardwaj. 2014. Is a Picture Really Worth a Thousand Words? - On the Role of Images in e-Commerce.

[7] Ehsan Elhamifar, Guillermo Sapiro, Allen Yang, and S Shankar Sasrty. 2013. A convex optimization framework for active learning. In *Proceedings of the IEEE International Conference on Computer Vision*. 209–216.

[8] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. *arXiv preprint arXiv:1703.02910* (2017).

[9] Rohith Gandhi. 2018. Siamese Network Triplet Loss. https://towardsdatascience.com/siamese-network-triplet-loss-b4ca82c1aec8

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV]

[11] Houdong Hu, Yan Wang, Linjun Yang, Pavel Komlev, Li Huang, Xi Chen, Jiapei Huang, Ye Wu, Meenaz Merchant, and Arun Sacheti. 2018. Web-Scale Responsive Visual Search at Bing. arXiv:1802.04914 [cs.CV]

[12] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. 2019. Squeeze-and-Excitation Networks. arXiv:1709.01507 [cs.CV]

[13] Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. 2015. Learning Visual Features from Large Weakly Supervised Data. arXiv:1511.02251 [cs.CV]

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.), Vol. 25. Curran Associates, Inc., 1097–1105. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[15] Eileen Li, Eric Kim, Andrew Zhai, Josh Beal, and Kunlong Gu. 2020. Bootstrapping Complete The Look at Pinterest. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining* (Virtual Event, CA, USA) *(KDD '20)*. Association for Computing Machinery, New York, NY, USA, 3299–3307. https://doi.org/10.1145/3394486.3403382

[16] Fengzi Li, Shashi Kant, Shunichi Araki, Sumer Bangera, and Swapna Samir Shukla. 2020. Neural Networks for Fashion Image Classification and Visual Search. arXiv:2005.08170 [cs.CV]

[17] D.G. Lowe. 2004. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60, 91–110 (2004). https://doi.org/10.1023/B:VISI.0000029664.99615.94. *International Journal of Computer Vision* 60 (2004), 91–110.

[18] A. Rosebrock. 2017. *Deep Learning for Computer Vision with Python: Starter Bundle*. PyImageSearch. https://books.google.com/books?id=9Ul-tgEACAAJ

[19] Alexander Schindler, Thomas Lidy, Stephan Karner, and Matthias Hecker. 2018. Fashion and Apparel Classification using Convolutional Neural Networks. arXiv:1811.04374 [cs.CV]

[20] Matthew Schultz and Thorsten Joachims. 2004. Learning a Distance Metric from Relative Comparisons. In *Advances in Neural Information Processing Systems*, S. Thrun, L. Saul, and B. Schölkopf (Eds.), Vol. 16. MIT Press, 41–48. https://proceedings.neurips.cc/paper/2003/file/d3b1fb02964aa64e257f9f26a31f72cf-Paper.pdf

[21] Ozan Sener and Silvio Savarese. 2017. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489* (2017).

[22] Devashish Shankar, Sujay Narumanchi, HA Ananya, Pramod Kompalli, and Krishnendu Chaudhury. 2017. Deep learning based large scale visual recommendation and search for e-commerce. *arXiv preprint arXiv:1703.02344* (2017).

[23] Raymond Shiau, Hao-Yu Wu, Eric Kim, Yue Du, Anqi Guo, Zhiyuan Zhang, Eileen Li, Kunlong Gu, Charles Rosenberg, and Andrew Zhai. 2020. Shop The Look: Building a Large Scale Visual Shopping System at Pinterest. 3203–3212. https://doi.org/10.1145/3394486.3403372

[24] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs.CV]

[25] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. 2016. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. arXiv:1602.07261 [cs.CV]

[26] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2014. Going Deeper with Convolutions. arXiv:1409.4842 [cs.CV]

[27] Yina Tang, Fedor Borisyuk, Siddarth Malreddy, Yixuan Li, Yiqun Liu, and Sergey Kirshner. 2019. MSURU: Large Scale E-commerce Image Classification with Weakly Supervised Search Data. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2518–2526.

[28] Jiang Wang, Yang song, Thomas Leung, Chuck Rosenberg, Jinbin Wang, James Philbin, Bo Chen, and Ying Wu. 2014. Learning Fine-grained Image Similarity with Deep Ranking. arXiv:1404.4661 [cs.CV]

[29] Kai Wei, Yuzong Liu, Katrin Kirchhoff, Chris Bartels, and Jeff Bilmes. 2014. Submodular subset selection for large-scale speech training data. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3311–3315.

[30] Wikipedia.org. 2020. XGBoost. https://en.wikipedia.org/wiki/XGBoost

[31] Stephen Zakrewsky, Kamelia Aryafar, and Ali Shokoufandeh. 2016. Item Popularity Prediction in E-commerce Using Image Quality Feature Vectors. arXiv:1605.03663 [cs.CV]

[32] J. Zhang, S. Lazebnik, and C. Schmid. 2007. Local features and kernels for classification of texture and object categories: a comprehensive study. *International Journal of Computer Vision* 73 (2007), 2007.

[33] Yanhao Zhang, Pan Pan, Yun Zheng, Kang Zhao, Yingya Zhang, Xiaofeng Ren, and Rong Jin. 2018. Visual Search at Alibaba.

[34] Fedor Zhdanov. 2019. Diverse mini-batch Active Learning. *arXiv preprint arXiv:1901.05954* (2019).

[35] Yun Zhu, Sayyed M. Zahiri, Jiaqi Wang, Han-Yu Chen, and Faizan Javed. 2020. Active Learning for Product Type Ontology Enhancement in E-commerce. arXiv:2009.09143 [cs.LG]

[36] Zhen Zuo, L. Wang, Michinari Momma, W. Wang, Yikai Ni, Jianfeng Lin, and Y. Sun. 2020. A flexible large-scale similar product identification system in e-commerce.

A ONTOLOGY CONCEPT DEFINITION

Table 6: Definition of class under concept group

Concept group	Class name	Definition
Annotation	Logo	Any image that has a logo or other marketing materials with it
Annotation	Dimensions	Image that contains dimension information adjacent to product
Content	Closeup (with white or non-white background)	Zoomed in product image as a purpose to highlight partial details
Content	Sketch	Image that contains some sketch or drawings of a product
Content	Infographic	Any image that has text descriptions, specifications, or other marketing materials with it
Content	Swatch/sample	Pure color/homogenous color in the whole image
Content	Swatch-Can	Image contains a product can in a corner and pure/homogenous color in the rest of the image
Content	Swatch-Smear	Smear textured pure/homogeneous color in the whole image
Content	Swatch-Brand	Images contain a brand logo in a corner and pure/homogeneous color in the rest of the image
Content	Swatch-Blob	Blob shaped pure/homogeneous color in the whole image
Content	Silo	Image contains a single full product with a white/non-white background
Content	Set	Image contains multiple products with a white/non-white background (can be full identical product set, full non-identical product set or part set)
Content	Open	Silo image that contains a full product that is open (having inside view)
Content	Props	Silo image contains a full product that come with several decorations around
Content	Packaged	Silo image that displays packaged product (s)
Content	Lifestyle (limited scene/full scene)	Images contain a single or multiple products taken in a rich scene that has abundant other contents
Views	Dist. far view	Silo image contains a small-scale product that it takes up small fraction of the whole image area
Views	Dist. near view	Image contains a large-scale product that it takes up majority of the whole image area
Views	Front/back	Silo image that is front facing or back facing
Views	Vert. angled 90/270	Silo image that is vertically rotated 90/270 degrees, relative to the front facing view
Views	Vert. angled 45/135	Silo image that is vertically rotated 45/135 degrees, relative to the front facing view
Views	Vert. angled other	Silo image that is vertically rotated by a random degree, relative to the front facing view
Views	Side – left/right	Silo image that is horizontally rotated 90/270 degrees (left/right), relative to the front facing view
Views	Horiz. angled – angled left/right	Silo image that is horizontally rotated towards the left/right, relative to the front facing view
Views	Orientation flat	Image contains slender product that is aligned flat
Views	Orientation erect	Image contains slender product that is aligned erectly
Views	Surface presentations	Silo image of surface shaped products that are folded, stacked or rolled