

# A Multimodal Recommender System for Large-scale Assortment Generation in E-commerce

Murium Iqbal  
Overstock  
Midvale, Utah  
miqbal@Overstock.com

Adair Kovac  
Overstock  
Midvale, Utah  
akovac@Overstock.com

Kamelia Aryafar  
Overstock  
Midvale, Utah  
karyafar@Overstock.com

## ABSTRACT

E-commerce platforms surface interesting products largely through product recommendations that capture users' styles and aesthetic preferences. Curating recommendations as a complete complementary set, or assortment, is critical for a successful e-commerce experience, especially for product categories such as furniture, where items are selected together with the overall theme, style or ambiance of a space in mind. In this paper, we propose two visually-aware recommender systems that can automatically curate an assortment of living room furniture around a couple of pre-selected seed pieces for the room. The first system aims to maximize the visual-based style compatibility of the entire selection by making use of transfer learning and topic modeling. The second system extends the first by incorporating text data and applying polylingual topic modeling to infer style over both modalities. We review the production pipeline for surfacing these visually-aware recommender systems and compare them through offline validations and large-scale online A/B tests on Overstock<sup>1</sup>. Our experimental results show that complementary style is best discovered over product sets when both visual and textual data are incorporated.

## CCS CONCEPTS

•Information systems → Recommender systems; Collaborative filtering; Presentation of retrieval results; •Computing methodologies → Neural networks;

## KEYWORDS

recommender system, visual document, topic modeling, set recommendation, quadratic knapsack problem, product recommendation

## ACM Reference format:

Murium Iqbal, Adair Kovac, and Kamelia Aryafar. 2018. A Multimodal Recommender System for Large-scale Assortment Generation in E-commerce. In *Proceedings of ACM SIGIR Workshop on eCommerce, Ann Arbor, Michigan, USA, July 2018 (SIGIR 2018 eCom)*, 9 pages.  
DOI: 10.1145/nnnnnnn.nnnnnnn

<sup>1</sup>[www.overstock.com](http://www.overstock.com)



Figure 1: An automatically generated assortment from the multimodal approach is shown.

## 1 INTRODUCTION

Overstock<sup>1</sup> is an e-commerce platform with the goal of creating *dream homes for all*. Users browse Overstock's catalog to select pieces that complement one another while matching the stylistic settings and color palettes of their rooms. As furniture is not a disposable product, furniture purchases are subject to careful scrutiny of aesthetics and strict budgeting. Brick and mortar stores often inspire consumers by creating furniture showrooms, in some cases pushing consumers to walk through their carefully selected furniture displays. These assorted showrooms alleviate the creative and stylistic pressure on the consumer, since the set is already on display with all the necessary pieces.

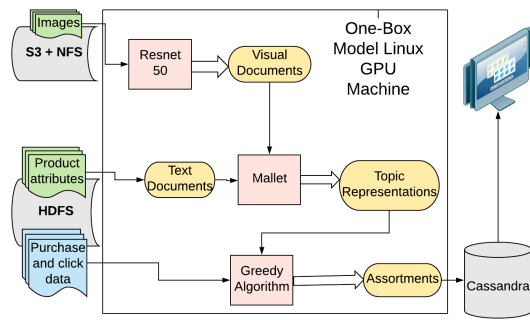
By crafting an appropriate recommender system, this experience can be recreated on an e-commerce platform. Figure 1 illustrates an example of the proposed system's automatically generated showroom, built around two seed products selected from Overstock<sup>1</sup>. The goal of this recommender system is to provide set recommendations that adhere to a general theme or cohesive visual style while accounting for essential item constraints<sup>2</sup>. Set recommendations have become more prominent with the rise of subscription box services in various domains such as fashion (e.g. StitchFix<sup>3</sup>), jewelry (e.g. Rocksbox<sup>4</sup>) and beauty products (e.g. Birchbox<sup>5</sup>). The criteria for selecting an assortment depends heavily on the product space which can inform the latent space of user preferences and product

<sup>2</sup>Essential item constraints refers to specific products which must be included in an assortment, e.g. a bed frame in a bedroom set or a vendor must-have in a subscription box.

<sup>3</sup>[www.stitchfix.com](http://www.stitchfix.com)

<sup>4</sup>[www.rocksbox.com](http://www.rocksbox.com)

<sup>5</sup>[www.birchbox.com](http://www.birchbox.com)



**Figure 2:** A system overview diagram is presented above. We use product data and user engagement data residing within our Hadoop cluster and our NFS/S3 store to train the model. Most of the training is done on a one box GPU server, which runs the Resnet-50, Mallet and our greedy algorithm. All results are served to Overstock<sup>1</sup> users from Cassandra.

representations. Personal care product recommendations, for example, rely on user preferences, such as skin type, which can be inferred from product text-based attributes. Fashion, jewelry and furniture shopping, on the other hand, are predominantly visual experiences. This motivates a *visually-aware representation* for products and style-based preferences for users.

We propose two assortment recommender systems. The first takes advantage of visually semantic features transferred from a deep convolutional neural network to learn style. The second uses both these visual features and product text-based attributes to learn style across the multimodal dataset. Our hypothesis is that while the visual style can help find similar products, use of text data in conjunction with images will result in more complimentary, and cohesive stylized assortments.

## 2 RELATED WORK

As the prevalence of recommendation systems grows, e-commerce platforms are looking to increase the influence on customer purchases by not just recommending one item at a time but several in a bundle together [25] [27] [1] [6]. This is true across industries, from travel agencies to clothing [10] [14] [15]. Specifically Zhu et al. [27] model the problem of bundle recommendations as the Quadratic Knapsack Problem (QKP), and find an approximation to it. They use only implicit feedback data, rather than content data to build their system, allowing for the bundles to be personalized. Using this approach, we also view our system as a QKP, but look to leverage primarily content data, incorporating implicit feedback as a minor signal. To capture a user’s preference, the system we propose allows a user to present a seed product, around which we can create the bundle. This allows us to circumvent the cold start problem, as any user, regardless of having a history on our site, can still build assortment recommendations. New products are also seamlessly incorporated, as long as images or text data is made available.

Recommendation systems is a well studied field which has two large over arching types: collaborative filtering methods [18] [12] [7], and content based methods [4] [21] [22] of which many rely on learning latent representation of the input data. Topic modeling,

specifically Latent Dirichlet Allocation (LDA) [3], has also been applied to create topic based recommendations both off of text input as well as from implicit feedback [11] [2] [26]. These systems work by scoring individual products against one another or against users to see which are most similar. Here we look to extend this methodology by applying topic modeling over both image and text data. Rather than simply concatenating image and text features and linearly combining them, PolyLDA enables us to learn two distinct but coupled latent style representations. This allows for a versatile interpretation of style [13]. We then use these learned styles to make bundle or assortment recommendations rather than the traditional single item recommendations.

To facilitate style discovery from the images, we process them via prevalent deep learning techniques. Deep residual networks have recently been shown as a powerful model to capture style based preferences for creating visual wardrobes from fashion images [10]. With the goal of learning visually coherent styles, we apply transfer learning, by using a Resnet-50<sup>6</sup> which was pre-trained on ImageNet<sup>7</sup> [8]. We explore interpreting the convolutional neural network by indexing the activations on channels within the convolutional layers [19]. We use a neural network to learn filters for our data and simply index their responses to images to create visual documents, similar to the older bag of word methods [24]. To discover visually-aware trends, we use these documents with Latent Dirichlet Allocation (LDA) [3].

While we use LDA for topic modeling on single-modality image data, we need an extension to interpret both visual semantic features and text-based attribute data as complementary modalities. Roller et al. [20] create a multimodal topic model that assumes words and corresponding visual features occur in pairs and should be captured as tuples within the topic distribution. This model would work well for descriptions of images, but the underlying assumption could prove to be too stringent to generalize to our application. Mimno et al. [17] offer a more flexible extension of LDA, Polylingual LDA (PolyLDA), which handles tuples of documents in different languages with assumed identical topic distributions. The documents within a tuple do not need to be direct translations of one another, and the topics themselves have distinct sets of words for each language. By treating our image data and our text data as two separate languages, we can use PolyLDA directly to create a more flexible multimodal topic model with which to infer style.

Hsiao and Grauman [10] also apply PolyLDA to infer style from images, but they use complementary clothing types (e.g. pants and blouse) as languages in a compatibility model instead of using PolyLDA to handle different modalities of data for the same item. The documents they use for their topic models are generated using a Resnet trained to automatically detect their pre-defined text attributes within images and apply the appropriate labels.

Much of the work performed on traditional recommendation systems can be applied here, but additional constraints must be taken into consideration. Products can’t simply be visually similar to be purchased together, but must be complimentary as well. For example, having two similar brightly colored upholstered chairs as a result in a traditional recommendation is perfectly acceptable, but if they

<sup>6</sup><https://github.com/facebook/fb.resnet.torch>

<sup>7</sup><http://www.image-net.org/>

are both taken into one assortment with the intention of being bought together, they may clash. Here we show that by incorporating text data, we are able to avoid assortments of incredibly visually similar, interchangeable products which are not at all complimentary. These nuances can be further addressed by incorporating purchase data to infer what types of products are not just similar, but also compatible, and likely to be bought together. To create the final assortment recommendations, we define a distance measure that combines stylistic similarity with implicit feedback from users, in the form of purchases, to develop a concept of affinity similar to approaches in composite recommendation systems such as [9, 10, 23, 27].

This rest of this paper is organized as follows: In Section 3, we present our production pipeline for assortment recommendations as deployed in Overstock<sup>1</sup> and explore our product embeddings as used with topic modeling techniques. We also review our formulation of assortment generation as a 0-1 quadratic knapsack problem and our greedy optimization for building assortments. Section 4 presents our offline and online experimental results. Finally, we conclude our paper in Section 5 and propose future directions.

### 3 METHODOLOGY

Here we provide an overview of the production system used to surface recommendations throughout Overstock<sup>1</sup> and describe our assortment recommender system in detail. We first explain our product embeddings as text-based bag-of-words and bag-of-visual-words (BoVW) documents. Then we explain the LDA approaches used to discover text-based and visual styles across our platform. We finally discuss our formulation of assortment generation as a 0-1 quadratic knapsack problem with budget constraints and our greedy assortment recommendations algorithm.

#### 3.1 System Overview

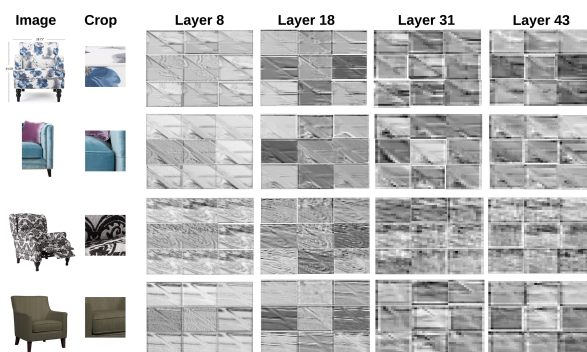
Product recommendations on Overstock<sup>1</sup> are surfaced through multiple carousels across the platform. User interactions with products and recommendations is recorded along side purchase data. Each product on the platform is represented by corresponding title, text attribute tags, text-based descriptions and images. The first step in generating recommendations is processing the large volume of product and user data which are hosted on Hadoop Servers.

After initial processing on Hadoop, the data is transferred to a one-box CUDA-enabled GPU server. Here product images are passed through a publicly available Resnet-50 to create image-based documents. Both the image and text documents are then fed into Mallet<sup>8</sup> to generate topic representations of products. We create two topic variants, a visual variant, that uses the image data with LDA and a multimodal variant, that uses both text and image data with PolyLDA.

Once the topic distributions are created, the products reside within their defined feature space. Hereafter, the methodology for both variants is identical. We construct a distance measure over the topic space, using implicit feedback from our users, to model compatibility. The assortments are then generated by finding products which minimize this distance to predefined seed product pairs.

The generated assortments are pushed to our Cassandra database cluster. Our website servers can query Cassandra directly to see

<sup>8</sup><http://mallet.cs.umass.edu/>



**Figure 3:** Here we depict activations from the layers used to create our visual documents. The leftmost column is the raw image. The next column to the right is the cropped patch which is passed through the network. In the next 4 columns, we see the response to these patches from 9 channels each of convolutional layers 8, 18, 31 and 43. Lighter colors mean strong responses, while darker colors mean weak responses. We threshold these responses to determine whether or not the "word" each channel represents is present within the image. This methodology seems to work particularly well for our upholstered furniture, showing distinct responses to the different patterns.

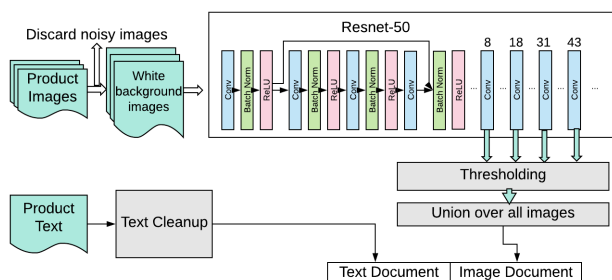
which products have an associated assortment to present to users. Assortment recommendations are featured on corresponding product pages, below the product description, providing an option for users to purchase an entire assortment, or just parts of it. A brief diagram of the systems involved is presented in Figure 2.

#### 3.2 Product Embeddings

Products on Overstock<sup>1</sup> have several images of the item for sale and different forms of associated text data including title, description and attribute tags. Product attributes are descriptive tags associated with a product. Some example attribute categories are color, size, style, composing materials, finish, and brand. In this paper, we utilize these attributes as the primary text-based information since they often provide a rich text representation of the item that compliments provided images. All product attribute data is processed to remove stop words and are concatenated with each product title to form a bag-of-words document.

We use the methodology described in [13] to create visual documents to be used with LDA via a Resnet. We then combine both visual and text data together via PolyLDA.

The convolutional layers within a Resnet are composed of a series of small learned filters. Each filter is convolved with the input in steps along its height and width. The result of this process is a 2-dimensional grid of activations that represents the response from the filter at corresponding locations of the input. The learned filters within a Resnet respond to specific patterns. By viewing these filters as words, and indexing them, Iqbal et al. [13] show that a BoVW document can be created which, when combined with LDA, is effective at uncovering style. To index a layer, we simply threshold the response from the filter to indicate whether or not it is sufficiently activated by the image to be included within the image's document.



**Figure 4: Each product’s corresponding text and images are collected. Our system then discards images with noisy scenes, keeping only those with white backgrounds and passing them through Resnet-50. Channel activations from layers 8,18,31, and 43 are thresholded and indexed to create visual documents. The union of these documents is taken as the visual representation of the product. The text data is stripped of stopwords and paired with the corresponding visual document to create a tuple for ingestion by PolyLDA**

Given this methodology for creating image documents, our visual word vocabulary is defined entirely by the channels in the layers that we choose to index. We tested creating documents using various layers through the network. We found that combining several middle layers gave us the best results. After experimenting we empirically chose to use Resnet-50 layers 8, 18, 31, and 43. The combination of these layers provided significantly better results, per visual inspection, than any single layer or other combinations. This yielded a total vocabulary size of 2816 visual words over our image documents. Since multiple images can be provided for each product, we take the union of all present visual words within the corresponding images as the visual document for the product itself. The process is depicted in Figure 4.

### 3.3 Topic Modeling

Once the visual documents are created, they can be used directly with LDA to create topics for our visual variant. LDA is an unsupervised generative model which infers probability distributions of words within topics and models documents as a mixture of those topics [3]. Given a set of documents and a number of target topics the model assumes the following generative process was used to create the documents.

- (1) For  $M$  documents, initialize the set of topic distributions,  $\theta_i \sim \text{Dir}(\alpha)$  for  $\{i = 1 \dots M\}$
- (2) For  $K$  topics, initialize the set of word distributions,  $\phi_k \sim \text{Dir}(\beta)$  for  $\{k = 1 \dots K\}$
- (3) For the  $k^{th}$  word in the  $i^{th}$  document select a topic  $z_i$  from  $\theta_i$  and a word,  $w_{i,k}$ , from  $\phi_{z_i}$

where Dir is the Dirichlet distribution.

The model initializes the parameters randomly then iteratively updates these parameters via Gibbs Sampling and variational Bayes inference. After many iterations the topic distributions converge to a stable state and the resulting topics can be used as a low dimensional feature space that captures the salient content of the original documents.

For the multimodal variant, each visual document is paired with its text equivalent in a tuple. These tuples are then used with Polylingual LDA (PolyLDA) [17]. Products that are missing either a text document or a visual document can still be used with PolyLDA, as it is robust to missing documents within tuples. A diagram of the system for multimodal document creation is provided in Figure 4. PolyLDA is an extension of LDA meant to handle loosely equivalent documents in different languages. It takes as input tuples of documents, each in a different language but with the same exact topic distribution. Each resulting latent topic from this method contains a distinct associated set of word distributions for each language. Documents aren’t required to be direct translations of one another, which allows for flexibility. A word in one language could be attributed to a given topic while its direct translation in another language may be attributed to a different topic. Words that only appear in one language can also be present in topics despite having no equivalent translations in the corpus. PolyLDA extends LDA by assuming the following generative procedure:

- (1) For a given tuple of documents,  $\{d_1 \dots d_L\}$ , initialize a single set of topic distributions,  $\theta \sim \text{Dir}(\alpha)$
- (2) For  $K$  topic sets with  $L$  languages, initialize the set of word distributions,  $\phi_{k,l} \sim \text{Dir}(\beta)$  for  $\{k = 1 \dots K\}$  and  $\{l = 1 \dots L\}$
- (3) For the  $k^{th}$  word in the  $l^{th}$  document in the  $i^{th}$  tuple select a topic  $z_{i,l}$  from  $\theta_i$  and a word,  $w_{i,l,k}$ , from  $\phi_{z_{i,l}}$

For our multimodal variant, we make use of the flexibility afforded by PolyLDA by applying it to our visual and text documents, representing each modality as a distinct language attempting to describe the same product. This method should allow for topics that afford better generality. Intuitively for our corpus, we suspect that a topic could capture relationships between certain visual features, like thick wood, with textual attributes, like "quality", that don’t have a direct visual representation but do share an underlying contextual relationship. By combining modalities the model also affords us the ability to infer a more complete style representation given incomplete data. For example, we can apply our multimodal topics to products with only text attributes available and still infer style that would normally be captured by the missing image data.

We provide visualizations of the resulting topics from both variants in Figure 5. To create these visualizations we selected the products that are maximally aligned with each topic. In our case, we have empirically seen that the topics relate well to various styles while reducing our feature space from a large number of attribute tags, title words, and image-based channel activations to a succinct  $k$ -dimensional space (where  $k$  is the number of topics). These styles are not necessarily complimentary though, especially in the case of the visual topics. We can see that the items within the topics seem more substitutable, e.g. in the cases of brightly colored furniture, similarly patterned upholstery, and mirrored surfaces.

### 3.4 Knapsack Problem

Once we have all our products residing within the topics’ feature space, we can begin to build out our assortments. The process for generating the assortments is the same for both the visual variant and the multimodal variant, with the only difference between the two



Figure 5: Representative items from each of six topics in the visual variant on the left (a) and multimodal variant on the right (b) are depicted above.

being the topic representations of products. To build these assortments, we first define the necessary components of an assortment as verticals. The verticals we define here enable us to build living room assortments, but we would like to emphasize that verticals can be defined for any e-commerce platform. If desired, the verticals can be left undefined, and the assortment can be built with no vertical constraints. For our specific application verticals are critical pieces of living room furniture, e.g. coffee table, chair, or accent table. Our definitions are manually applied to the products and are listed below. (Here a couch set can be either a sectional sofa or a sofa and loveseat combination.)

*Verticals:* {Couch Set, Coffee Table, Accent Table, Entertainment Center, Bookshelf, Ottoman, Chair}

We chose to build our assortments around seed pieces. For our current application, we generated seeds, but we can easily allow users to provide their own seeds around which we can curate an assortment. We selected seeds as preferred pairings of the most crucial verticals in any living room space, namely a couch set and a coffee table. Although all the other verticals can be seen as optional, these two verticals are defining members of a living room and therefore must be present in any assortment. We chose pairs of sofas and coffee tables which are most frequently co-clicked for our seeds.

We then assume a budget constraint for the entire assortment. As such we can formulate our assortment generation as a 0-1 quadratic knapsack problem (0-1 QKP) as defined by Gallo et al [5]. The 0-1 QKP assumes that a knapsack with limited space needs to be filled with a set of items. These items all have an intrinsic value. Items

also have an additional associated pair-wise profit when selected together for the knapsack.

For a given seed, we can find the optimal assortments by trying to find the products that are nearest the seed in the topic space. We can add the additional constraint that we want all products within the assortment to be nearest one another in the topic space as well. The total assortment must remain within the assumed budget constraint. Each product has an associated vertical, and we can add in some constraints on the minimum and maximum number of products each vertical can contribute to the final assortment.

To calculate proximity we use a Mahalanobis distance built on purchase data similar to [16]. The Mahalanobis distance allows us to leverage the implicit feedback of our users to understand how our topics distribute and relate to one another across multiple products within a purchase.

We assume that purchases of living room furniture made by the same user over a 3 month period are being used in the same room, and as such can be said to represent a compatible assortment. After trimming to include only purchases with 3-10 pieces of living room furniture, so as to avoid noise and erroneous signals from bulk purchasers, we create a dataset of roughly 7.5k purchased assortments. We take the  $L_2$  normalized sum of the topic distributions of the co-purchased products as a topic representation of the purchase itself, and use this dataset to calculate the covariance matrix used for our Mahalanobis distance.

The Mahalanobis distance is defined as  $d_M(x_i, x_j) = (x_i - x_j)M(x_i - x_j)^T$ , Where  $M$  is the covariance matrix for the topics learned from our purchase data. For each seed pair of coffee table and couch set, we greedily swap products in and out of our assortment until we

**Algorithm 1** Generate Assortments

---

```

 $\delta \leftarrow \epsilon + 1$ 
2:  $A_{i,prev} \leftarrow \{\} \forall i \in \{V-S\}$ 
    $A_{i,prev} \leftarrow \{p_i\} \forall i \in \{S\}$ 
4: while  $\delta \geq \epsilon$  do
    $\delta \leftarrow 0$ 
6:   for  $i \in \{V-S\}$  do
      $t_A \leftarrow \|(\sum t_p)\| \forall p_j \in A_{j,prev}, \forall j \in V \neq i$ 
8:     while  $\text{size}(A_i) < i_{size}$  do
        $A_i \leftarrow A_i \cup \text{argmin}(d_M(t_o, t_A))$ 
10:    $\delta \leftarrow \delta + d_M(t_{A_i}, t_{A_{i,prev}})$ 

```

---

converge to an ideal set of products. The only verticals which are not allowed to change are those in the seed.

To formulate the 0-1 QKP we first define some terminology. Each product  $o$  will have a vertical label  $a_o$ , an associated price  $c_o$ , and associated topic distribution  $t_o$ . All products will have a pairwise distance associated with them and every other product  $d_{i,j}$  which is the Mahalanobis distance between products  $o_i$  and  $o_j$  built on the purchases and topic distributions. Additionally the products all have a score  $q_o$  to represent their value to the seed, here we use the inverse of the distance from the product to the seed. Let  $M$  be the set of products comprising the assortment. Let  $B$  be the total available budget minus the cost of the seed. The 0-1 knapsack problem can then be formulated as:

$$\begin{aligned}
 & \text{maximize} \quad \sum_{i=1}^M q_i + \sum_{j=1}^M 1/d_{i,j}, j \neq i \\
 & \text{subject to:} \\
 & \quad \forall i : \sum_{i=1}^M c_i \leq B \\
 & \quad \forall k : \min_k \leq \text{count}(a_o = k) \leq \max_k
 \end{aligned} \tag{1}$$

A greedy approximation can now be formulated to follow the constraints of the 0-1 QKP. First a solution is initialized, then iteratively, items are swapped until the system converges to an optimal solution. To get the initial set, products are all scored by their total potential for compatibility (the sum of the score of this product with the sum total of all reciprocals of pairwise Mahalanobis distance) divided by their price. Products are then sorted by this ratio and the highest scored are added until vertical constraints or budget constraints are met. Then each product is considered for a swap with other products to improve the total compatibility of the assortment.

### 3.5 Greedy Method

For our online evaluation, we relax the budget constraints, so that we can display the assortment on product pages. This allows us to select the best assortment per product. We can easily reapply the budget constraint for user assortments. As such, we use the following greedy method to generate the assortments. The assortment is initialized as first only containing the seed items, all other verticals are empty. Each vertical is then iteratively considered for new products while holding all other verticals constant. New products are chosen from

the pool of products that are labeled as candidates for this vertical. Those which are closest to the total assortment (minus the current vertical) at the previous time-step are added in, until the size of the vertical is met. The sizes of the verticals in our experiment were set by us, but a user can dictate how many side tables, bookshelves, chairs, etc they wish to add to their assortment.

Let  $V$  be the set of all verticals.  $S$  is the set of verticals used in the seed.  $A_i$  is the set of products selected for the assortment in vertical  $i$  for the current timestep. Each vertical  $i$  has an associated size,  $i_{size}$ , which is the number of products we want for this vertical.  $A_{i,prev}$  is the set of products selected for the assortment in vertical  $i$  for the previous timestep.  $p_i$  is an element of  $A_{i,prev}$ .  $o_i$  is a candidate product for vertical  $i$ .  $t_o$  is the topic distribution of a product.  $d_M$  is the Mahalanobis distance as defined above.  $\delta$  is the change in the distance of the assortment from iteration to iteration. The greedy approach is described using this notation in Algorithm 1. Figure 6 illustrates the generated assortments using this method for both the visual and multimodal variants.

## 4 RESULTS AND DISCUSSION

This section describes a large-scale experiment on Overstock<sup>1</sup> to determine whether simultaneously learning style from text and images provides better results than learning style from only images. All offline validations and online A/B test assume 2 variants: a visual-only assortment recommender system, which is built on top of BoVW representation of product images, and a multimodal one, which is built on text-based attributes and visual words. Our site does offer manually curated collections, but there is little overlap on the products selected in the manual process and those selected for our automated system. The methodology used to create these collections also involves selecting products from the same vendor from the same product line, which can often be viewed as substitutable products as well. These would thus serve as a poor comparison to our model, which attempts to find stylistically complimentary products of different categories regardless of vendor, rather than similar/identical products. As such we do not use them as a baseline to compare our model against.

*Online Evaluation.* We run an A/B test on Overstock<sup>1</sup> with both the visual variant and multimodal variant assortments on product pages. The users are provided an option to add an entire assortment or individual items from the recommendations module to their cart. User engagement with recommendations are measured via click-through rate, CTR, on assortment recommendations is a classic measure of user engagement with product recommendations in e-commerce platforms. Our findings show that the multimodal variant outperforms the visual variant, showing a statistically significant lift of 10.9% relative to a visual only baseline.

*Offline Evaluations.* As we are recommending sets of products together, we can't use traditional offline evaluations, such as AUC, MAP and Recall. For scoring, we have taken the assortments and scored them with the average click-based Jaccard coefficient calculated pairwise over all products contained within the assortment. We make use of logs of product clicks over the last 2 months to build this score. This tells us how compatible the various products within an assortment are to one another.



**Figure 6:** Assortments from our visual variant (top) and multimodal variant (bottom) with the same seeds are depicted above. The visual assortments tend to provide a more-of-the-same assortment, while the multimodal is able to create a diverse set of products that still forms a cohesive style. When the visual assortment can't find products similar to either seed, it selects products which are similar to one another. In some cases it can match products to one seed, and ignores the other, which leaves the second seed item looking out of place. The multimodal variant, which can take advantage of text data as well, is able to find pieces to create a cohesive look incorporating elements of both seeds.

To calculate the assortment Jaccard, we look at  $S$ , the set of all user browsing sessions that included clicks on at least 2 items of living room furniture. A user browsing session only spans one visit, so we can assume that when items are co-clicked within the same session, the user still has the same intent, and is implicitly seeing both products as satisfying this intent. We take all user browsing sessions over a two month span for our dataset. We denote the subset of  $S$  that includes sessions that have clicked on product  $a$  as  $S_a$ .

We calculate the Jaccard coefficient for all pairs of products  $(a, b)$  within our corpus as follows:

$$J(a, b) = \frac{|S_a \cap S_b|}{|S_a \cup S_b|}$$

In contrast to the standard Jaccard formulation, if  $|S_a \cup S_b| = 0$ , we consider  $J(a, b) = 0$ . This modification is to prevent products that were never visited together from being counted as perfectly compatible with each other, since the standard formulation defaults these cases to 1.

For assortment  $A$  we calculate an assortment Jaccard score  $J_A$  as a simple average of the pairwise Jaccard coefficient for all items within the assortment, excluding the Jaccard coefficient between the two seeds. An example assortment with a good Jaccard score is shown in Figure 1.

The multimodal maintains a much higher average assortment Jaccard than the visual-only variant. The dramatic difference between

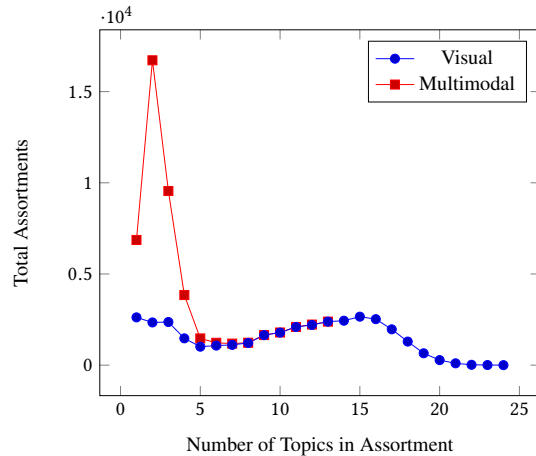
**Table 1:** Jaccard Score averages and max values for assortments of both variants.

Modality	Avg	Max
visual	0.0015	0.0220
multimodal	0.0027	0.0362

the multimodal variant and the visual variant may reflect an underlying bias in the click data: The text modality looks at the same data used in site search and navigation to surface groups of products to users, so co-click data may reflect better discoverability of items with similar textual features as opposed to similar visual features.

We also examine the distribution of topics from each assortment. The topic representation of products in our PolyLDA generated space is more diverse than that of the LDA. The products in the LDA space usually are strongly attributed to one or two products, while the PolyLDA does a better job of representing a product as a mixture of topics. This affects style as the products within the topics that are learned are very visually similar, lending themselves to substitutability rather than complimentary behavior.

Figure 7 shows how many topics compose each assortment. The distribution for the visual variant is heavily skewed, with the majority of assortments only being composed of a few topics. The multimodal



**Figure 7:** The above chart tabulates the number of topics that non-trivially contribute to an assortment. The vast majority of visual assortments have 1-4 topics associated with them, as depicted by the peak in the red line. On the other hand, the multimodal assortments can have anywhere from one to 20 topics associated with them, which allows for more diverse product representation. This is desirable as the uncovered topics show very similar items, not necessarily complimentary ones as depicted in Figure 5. A user would likely not want an entire set of bright red furniture. Mixing among the topics to create a complimentary set is more desirable than an entire set of the exact same visual features.

variant has much more variety, with assortments composed of only a few topics, to assortments composed of nearly half the topics being equally likely. This shows us that the multimodal variant offers more diverse assortments. This is preferred since, as depicted in 5, the topics themselves often contain substitutable rather than complimentary items. Properly blending the topics together to create a cohesive look results in better assortments, as depicted in Figure 6.

## 5 CONCLUSION

Online shopping is a visual experience. Visually-aware recommendations are crucial to online shopping and e-commerce platforms. Yet, systems which rely solely on images suffer from a lack of diversity. By incorporating both image and text data we are able to create cohesive styles.

In this paper we introduced a deep visually-aware large-scale assortments recommender system for Overstock<sup>1</sup>. Our assortment recommender system takes advantage of product images to create visually coherent trends from Overstock products. We introduced two variants: a visual-only variant and a multimodal variant. Our visual-only variant creates a bag-of-visual-words representation of product images by thresholding the activations from specific layers of a pre-trained deep residual neural network, Resnet-50. It then applies topic modeling (LDA) on product image representations to create visual trends among Overstock products. We then proposed a greedy approach (with and without budget constraints) to create assortment recommendations based on seed items that maximize the visual compatibility of the set. Our multimodal variant takes

advantage of text-based product attributes in addition to image representation. This variant utilizes Polylingual LDA (PolyLDA) to create trends that are based on two modalities, images and text.

We have featured multiple assortments generated from both models. We have evaluated our results through a set of offline validations and an online large-scale A/B test on Overstock. Our experimental results indicate that incorporating both image and text data provides more a more cohesive visual style than using only images and can enhance user engagement metrics with recommendations module. We also show that PolyLDA provides a meaningful way to simultaneously learn style across text and image data.

## REFERENCES

- [1] Moran Beladev, Lior Rokach, and Bracha Shapira. 2016. Recommender systems for product bundling. *Knowledge-Based Systems* 111 (2016), 193–206.
- [2] Sonia Bergamaschi, Laura Po, and Serena Sorrentino. 2014. Comparing Topic Models for a Movie Recommendation System. In *WEBIST (2)*. 172–183.
- [3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
- [4] Tiago Cunha, Carlos Soares, and André CPLF Carvalho. 2017. Metalearning for Context-aware Filtering: Selection of Tensor Factorization Algorithms. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 14–22.
- [5] Giorgio Gallo, Peter L Hammer, and Bruno Simeone. 1980. Quadratic knapsack problems. In *Combinatorial optimization*. Springer, 132–149.
- [6] Robert Garfinkel, Ram Gopal, Arvind Tripathi, and Fang Yin. 2006. Design of a shopbot and recommender system for bundle purchases. *Decision Support Systems* 42, 3 (2006), 1974–1986.
- [7] Prem Gopalan, Jake M Hofman, and David M Blei. 2015. Scalable Recommendation with Hierarchical Poisson Factorization. In *UAI*. 326–335.
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [9] Ruining He, Charles Packer, and Julian McAuley. 2016. Learning compatibility across categories for heterogeneous item recommendation. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, 937–942.
- [10] Wei-Lin Hsiao and Kristen Grauman. 2017. Creating Capsule Wardrobes from Fashion Images. *arXiv preprint arXiv:1712.02662* (2017).
- [11] Diane J Hu, Rob Hall, and Josh Attenberg. 2014. Style in the long tail: Discovering unique interests with latent variable models in large scale social e-commerce. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1640–1649.
- [12] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, 263–272.
- [13] Murium Iqbal, Adair Kovac, and Kamelia Aryafar. 2018. Discovering Style Trends through Deep Visually Aware Latent Item Embeddings. *CVPR Workshops* (2018).
- [14] Qi Liu, Yong Ge, Zhongmou Li, Enhong Chen, and Hui Xiong. 2011. Personalized travel package recommendation. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. IEEE, 407–416.
- [15] Si Liu, Jiashi Feng, Zheng Song, Tianzhu Zhang, Hanqing Lu, Changsheng Xu, and Shuicheng Yan. 2012. Hi, magic closet, tell me what to wear!. In *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 619–628.
- [16] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 43–52.
- [17] David Mimno, Hanna M Wallach, Jason Naradowsky, David A Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*. Association for Computational Linguistics, 880–889.
- [18] Andriy Mnih and Ruslan R Salakhutdinov. 2008. Probabilistic matrix factorization. In *Advances in neural information processing systems*. 1257–1264.
- [19] Ivet Rafeagas, Maria Vanrell, and Luis A Alexandre. 2017. Understanding trained CNNs by indexing neuron selectivity. *arXiv preprint arXiv:1702.00382* (2017).
- [20] Stephen Roller and Sabine Schulte Im Walde. 2013. A multimodal LDA model integrating textual, cognitive and visual modalities. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 1146–1157.
- [21] Aaron Van den Oord, Sander Dieleman, and Benjamin Schrauwen. 2013. Deep content-based music recommendation. In *Advances in neural information processing systems*. 2643–2651.



- [22] Robin Van Meteren and Maarten Van Someren. 2000. Using content-based filtering for recommendation. In *Proceedings of the Machine Learning in the New Information Age: MLnet/ECML2000 Workshop*. 47–56.
- [23] Andreas Veit, Balazs Kovacs, Sean Bell, Julian McAuley, Kavita Bala, and Serge Belongie. 2015. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 4642–4650.
- [24] Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. 2009. Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 1113–1120.
- [25] Min Xie, Laks VS Lakshmanan, and Peter T Wood. 2010. Breaking out of the box of recommendations: from items to packages. In *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 151–158.
- [26] Shengli Xie and Yifan Feng. 2015. A recommendation system combining LDA and collaborative filtering method for Scenic Spot. In *Information Science and Control Engineering (ICISCE), 2015 2nd International Conference on*. IEEE, 67–71.
- [27] Tao Zhu, Patrick Harrington, Junjun Li, and Lei Tang. 2014. Bundle recommendation in ecommerce. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 657–666.