CRAFT: Complementary Recommendations Using Adversarial Feature Transformer

Cong Phuoc Huynh   Arridhana Ciptadi   Ambrish Tyagi   Amit Agrawal

Amazon.com

{conghuyn, ambrisht, aaagrawa}@amazon.com

Abstract

Traditional approaches for complementary product recommendations rely on behavioral and non-visual data such as customer co-views or co-buys. However, certain domains such as fashion are primarily visual. We propose a framework that harnesses visual cues in an unsupervised manner to learn the distribution of co-occurring complementary items in real world images. Our model learns a non-linear transformation between the two manifolds of source and target complementary item categories (e.g., tops and bottoms in outfits). Given a large dataset of images containing instances of co-occurring object categories, we train a generative transformer network directly on the feature representation space by casting it as an adversarial optimization problem. Such a conditional generative model can produce multiple novel samples of complementary items (in the feature space) for a given query item. The final recommendations are selected from the closest real world examples to the synthesized complementary features. We apply our framework to the task of recommending complementary tops for a given bottom clothing item. The recommendations made by our system are diverse, and are favored by human experts over the baseline approaches.

1. Introduction

Recommendation algorithms are central to many commercial applications, particularly for online shopping. In domains such as fashion, customers are looking for clothing recommendations that visually complement their current outfits, styles, and wardrobe. Traditional content-based and collaborative recommendation algorithms [1, 17] do not make use of the visual cues to suggest complementary items. Among these, collaborative filtering [15, 24] is a commonly used approach, which primarily relies on behavioral and historical data such as co-purchases, co-views, and past purchases to suggest new items to customers. In contrast to these approaches, this work addresses the problem

Figure 1. Recommending tops for a given query bottom. Tops that are visually similar to the *actual* top worn with the query item are acceptable options, but lack diversity. Our approach generates both complementary and diverse recommendations that are also preferred by the fashion specialists. Please see the supplementary material for more examples.

of providing complementary item recommendations for a given query item based on visual cues.

We develop an unsupervised learning approach for Complementary Recommendation using Adversarial Feature Transform (CRAFT), by learning the co-occurrence of item pairs in real images. The assumption here is that the co-occurrence frequency of item pairs is a strong indicator of the likelihood of their complementary relationship. We define an adversarial process to train a conditional generative transformer network that can learn the joint distribution of item pairs by observing samples from the real distribution, *i.e.* image features of co-occurring items.

In contrast to traditional generative adversarial approaches that aim to directly synthesize images, our generative transformer network is trained on and generates samples in the *feature* space. To generate a recommendation
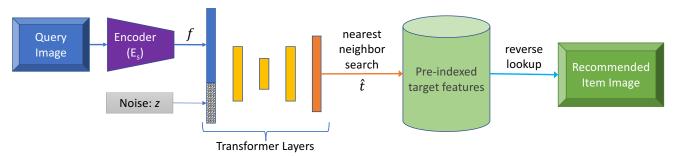
Figure 2. Generating recommendations using the proposed CRAFT network.

for a given query image, we first extract its features via a pre-trained encoder network, E_s (Figure 2). The query feature, f, along with a sampled noise vector, z, is fed to the transformer network to generate the feature vector for a complementary item. We then use the generated feature, $\hat{t}$, to retrieve the recommended target image by performing a nearest neighbor search on a pre-indexed candidate subspace. Our proposed approach is general (modeling visual co-occurrences) and can be applied to different domains such as fashion, home design, *etc*.

Our approach is a novel and unique way of utilizing generative adversarial training with several advantages over traditional Generative Adversarial Network (GAN) [7] based image generation. While the quality of visual image generation using GANs has improved significantly (especially for faces [12]), it still lacks the realism required for many real-world applications, such as in the fashion/apparel industry. More importantly, the goal of a recommendation system in such applications is often not to generate synthetic images, but rather to recommend real images from a catalog of items. An approach that generates synthetic images will thus still need to perform a search, which is typically done in the feature space, to find the most visually similar images in the catalog. CRAFT directly generates features of the recommended items, thereby bypassing the need to generate synthetic images and enabling a simpler and more efficient algorithm. By working in the feature space, we can use a simpler network architecture, which improves stability during training and avoids common pitfalls such as mode collapse [3].

We evaluate our algorithm (CRAFT) on a real-world problem of recommending complementary top clothing items for a given query bottom. For this task, a recommendation algorithm is expected to produce both complementary and diverse set of results. Figure 1 shows an example of the recommendation results from CRAFT and from a naïve approach of recommending nearest neighbors of the top. It is clear that generating recommendations by simply finding visually similar tops lacks the capability of producing *diverse* set of results.

Given that the assessment of fashion recommendations is a subjective problem, we conducted rigorous studies with fashion specialists to evaluate the quality of our algorithm. Through these studies, we demonstrate the effectiveness of our approach as compared to several baseline approaches (Section 4.3).

2. Related Work

Generative Adversarial Networks: GANs [7] have recently emerged as a powerful framework for learning generative models of complex data distributions. They have shown impressive results for various tasks including image generation [12, 21], image-to-image translation [11, 34], domain adaptation [2, 26, 28], *etc*. In a GAN framework, a generator is trained to synthesize samples from a latent distribution and a discriminator network is used to distinguish between synthetic and real samples. The generator's goal is to fool the discriminator by producing samples that are as close to real data as possible. A recent work by Zhu *et al*. [35] used the GAN framework to generate new clothing on a wearer. Our approach differs from these methods since we do not aim to generate an image of the complementary item. Instead, we use the adversarial training framework to learn the joint distribution between the source and target *features* in an *unsupervised* manner. We train a transformer that takes as input a random noise vector as well as the features of a query image and generates a feature vector representation of a complementary item.

The GAN paradigm has also found applications in the areas of image manipulation and image transformation [2, 11, 26, 34]. For example, Srivastava *et al*. [26] add realism to synthetic data by training an adversarial network that transforms a synthetic image into a real image. While such an approach can be applied to transform a given image into that of a complementary item, it only provides a fixed mapping. In contrast, ours is a generative approach that can provide multiple complementary items by learning the joint distribution in the feature space. Contrary to methods such as CycleGAN [34] that perform image-to-image translation using raw pixels, our approach works directly in the feature space. Feature-based domain adaptation approaches such as [28] attempt to directly learn a visual encoder for the target domain by minimizing an adversarial loss defined

GAN Input	Generative (w/ random seed)	Output	Example
N/A	Yes	Image	Image Generation [7]
Image	No	Image	Image-to-Image Translation [34]
Image + Attribute	No	Image	Image Manipulation [16]
Synthetic Image	No	Image	Adding Realism [26]
Synthetic Image	Yes	Image	Adding Realism [2]
Image	No	Features	Domain Adaptation [28]
Features	Yes	Features	Ours

Table 1. Similarities and differences between our approach and those that use adversarial loss for training.

on the source and target features. In contrast, we train a generative transformer network that operates in the feature space. Table 2 shows similarities and differences between our approach and those that use adversarial loss for training.

Unsupervised Learning: Recent applications of unsupervised learning for visual tasks include object discovery in videos [4]. In addition, there have been demonstrations of self-supervised learning [6] for the tasks of image colorization, image in-painting, hole filling, jigsaw puzzle solving from image patches, future frame prediction using video data [31], *etc*. In the fashion domain, annotated data are typically used for predicting fashion-related attributes and matching street-to-catalog images [13, 20]. These approaches involve visual search to find *similar* looking items, whereas our approach is focused on finding complementary items. Furthermore, our approach is unsupervised: we only take as input a set of images to learn the feature transformation between complementary objects.

Recommendation: There is a rich body of literature on using behavioral customer data such as browsing and purchasing history to develop recommender systems [15, 32]. Specific to the fashion domain, McAuley *et al*. [23] employed convolution neural network (CNN) features and non-visual data to build a personalized model of user's preference. In [9], the authors proposed a mixture of (non-metric) embeddings to recommend visually compatible items in several categories. A related approach is to learn a common embedding across categories and use a metric function in the embedding subspace as a measure of visual compatibility [30]. In [8], the authors proposed to learn a bi-directional Long Short Term Memory (LSTM) model in a supervised manner, to suggest items that complement each other in an entire outfit.

The aforementioned recommendation approaches use customer's behavioral data as training labels. Behavioral signals do not necessarily reflect that items viewed or purchased together are visually complementary. In contrast, our unsupervised approach learns item co-occurrences from only visual data. In multiple methods [8, 23, 30], the recommendation model is non-generative in the sense that it can only evaluate the compatibility between two given items. In others [9], the diversity of recommendation is limited by the (fixed) number of embeddings employed. In contrast, our

generative model is not subject to such a constraint, thanks to its ability to sample an infinite amount of noise vectors.

3. Generative Feature Transformer Network

This section describes our generative approach for complementary recommendation based on the co-occurrence of item pairs in real-world images. This could include, for example, combinations of top and bottom clothing items that people wear as part of their outfits or pairs of furniture items such as sofa and chairs that are present in most household scenes. We hypothesize that learning the joint distribution of such pairs can be useful for recommending new items that complement a given query. We adopt an adversarial learning paradigm, where our transformer network learns to generate features of the complementary items conditioned on the query item.

3.1. Network Architecture

In contrast to traditional GANs, our approach synthesizes visual features rather than images, which offers several advantages. It is more tractable to sample the feature space rather than the image space. The bulk of computation in the convolution and transpose convolution layers, which are usually required for feature learning and image synthesis, can be avoided, leading to a simpler architecture and stable training.

We first select an appropriate visual representations for the source and target images. The fixed feature representations (encodings) are generally derived from pre-trained CNNs. Typically, it is advisable to use application-specific feature representations, *e.g*., apparel feature embeddings for clothing recommendations, but a general representation such as one trained on ImageNet [5] or MS-COCO [19] offer robust alternatives. Figure 3 depicts the overall architecture of the CRAFT network. The source and the target feature encoders, E_s and E_t, respectively, are fixed and are used to generate feature vectors for training and inference.

Our architecture resembles traditional GAN designs with two main components: a conditional feature transformer and a discriminator. The role of the feature transformer is to transform the source feature s_q into a complementary target feature $\hat{t}_q$. The input to the transformer also consists of a random noise vector z sampled uniformly from
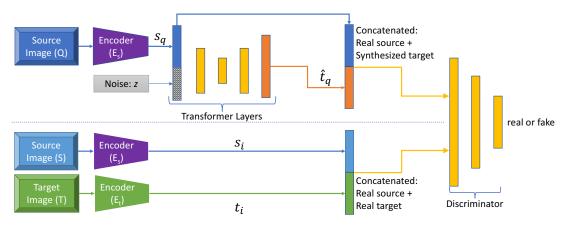
Figure 3. Architecture for the CRAFT framework. The transformer is trained using the adversarial loss to generate the target features conditioned on the source features and a sampled noise vector.

a unit sphere in a d_z-dimensional space. By design, the transformer is generative since it is able to sample various features in the target domain.

As discussed, since our approach works in the feature space, we can adopt a simple architecture for the feature transformer and discriminator. The transformer consists of several fully-connected layers, each followed by batch normalization [10] and leaky ReLU [22] activation layers. The discriminator is commensurate to the transformer in capacity, consisting of the same number of layers. This helps balance the power between the transformer and the discriminator in the two-player game, leading to stable training and convergence.

3.2. Training

Our training data consists of N co-occurring feature pairs $\mathcal{C} = \{(s_i, t_i), i = 1, \ldots, N\}$, where $s_i \in \mathbb{R}^{d_s}$ and $t_i \in \mathbb{R}^{d_t}$ denote the features corresponding to the source and the target images, respectively. Given a sample s_q from the source space, the complementary recommendation task is to generate target features $\{\hat{t}_q\}$ that maximizes the likelihood that the pair $(s_q, \hat{t}_q)$ belongs to the joint distribution $p_{\mathcal{C}}$ represented by the training data. To this end, we model the composition of layers in the feature transformer and the discriminator as two functions $T_\phi(s, z) : (s, z) \mapsto \hat{t}$ and $D_\theta(s, t) : (s, t) \mapsto [0, 1]$, respectively. Here, ϕ and θ are the learnable parameters of the two players, transformer and discriminator, respectively, and (s, t) is a pair of source and target feature vectors, and z is a random noise vector.

The training process emulates an adversarial game between the feature transformer and the discriminator, where the discriminator aims to classify feature pairs as real (co-occurring) or synthetic. On the other hand, the feature transformer synthesizes target features $\{\hat{t}_q\}$ conditioned on a given source feature s_q. Its objective is to fool the discriminator into the belief that $\hat{t}_q$ co-occurs with s_q. The feedback from the discriminator encourages the transformer to pro-

duce a target feature $\hat{t}_q$ so as to maximize the co-occurrence probability of the synthetic pair.

The adversarial game can be formulated as a mini-max optimization problem. The optimization approach can be implemented by alternating the training of the discriminator and the feature transformer. In the discriminator step (D-step), the discriminator's goal is to assign a binary label, i.e. 0 to the synthesized feature pair $(s_q, \hat{t}_q)$, where $\hat{t}_q = T_\phi(s_q, z)$, and 1 to an actual pair (s_i, t_i). The discriminator maximizes the cross entropy loss in Equation 1.

$$
\begin{aligned}
\mathcal{L}_D \triangleq\ & \mathbb{E}_{(s_i, t_i) \sim p_{\mathcal{C}}} \log D_\theta(s_i, t_i) \\
& + \mathbb{E}_{z \sim p_z, s_q \sim p_s} \log(1 - D_\theta(s_q, T_\phi(s_q, z)))
\end{aligned}
\tag{1}
$$

where p_z and p_s are the probability distribution function (pdf) of the random noise and the source feature.

The feature transformer maximizes the likelihood that the discriminator recognizes synthetic pairs as belonging to the data-generating (joint) distribution $\mathcal{C}$, i.e. it assigns a label 1 to such pairs. Therefore, the transformer step (T-step) aims to minimize the loss in Equation 2.

$$
\mathcal{L}_T = \mathbb{E}_{z \sim p_z, s_q \sim p_s} \log(1 - D_\theta(s_q, T_\phi(s_q, z)))
\tag{2}
$$

The overall objective function of the adversarial training process is formulated in Equation 3.

$$
\begin{aligned}
\min_\phi \max_\theta \mathcal{L} \triangleq\ & \mathbb{E}_{(s_i, t_i) \sim p_{\mathcal{C}}} \log D_\theta(s_i, t_i) \\
& + \mathbb{E}_{z \sim p_z, s_q \sim p_s} \log(1 - D_\theta(s_q, T_\phi(s_q, z)))
\end{aligned}
\tag{3}
$$

While our approach is completely unsupervised and does not require labels for complementary relationships, it can be easily extended to train a semi-supervised transformer that can benefit from additional complementary annotations. For example, if we know that certain complementary combinations are better or are ranked higher than others, then we can employ an additional discriminator that can take concatenated source and target features from labeled data and predict the ranking of these pairs.

3.3. Generating Recommendations

The recommendation work-flow is depicted in Figure 2. Here, we retain only the transformer's layers shown in Figure 3 for recommendation. From a query image, the query feature f is extracted by the source encoder, E_s, and multiple samples of transformed features $\{\hat{t}_i\}$ are generated by sampling random vectors $\{z_i\}$. This allows us to generate a diverse set of complementary recommendations by sampling the underlying conditional probability distribution function. Subsequently, we perform a nearest neighbor search within a set of pre-indexed target features extracted using the same target encoder, E_t, used during training. Actual recommendation images are retrieved by a reverse lookup that maps the selected features to the original target images.

4. Experimental Setup

We demonstrate the efficacy of our approach by applying it to the problem of complementary apparel recommendations. Specifically, we train the generative transformer network to synthesize features for top clothing items that are visually compatible to a given query bottom item. The generated features are used to retrieve the images of the nearest neighbors in a pre-indexed catalog of candidate top items as the complementary recommendations.

4.1. Datasets

We trained the proposed CRAFT network from scratch on unlabeled images, without the need for any human annotation to define complementary relationships. Our training data consisted of 473k full-length outfit images, each containing a top and a bottom clothing part, under the assumption that their outfits are highly complementary.

Each image was preprocessed to extract the regions of interest (ROIs) corresponding to the top and bottom clothing items. To extract clothing part ROIs, we trained a semantic segmentation network [33] on the "Human Parsing in the Wild" dataset [18]. We consolidate the original labels in the dataset into 15 labels, where top clothing items correspond to the label "upper-clothes" and bottom ones correspond to "pants" and "skirt". Using this segmentation network, we parse the training images into clothing parts, based on which tight bounding boxes around the segments corresponding to top and bottom regions are selected. In this manner, the training pairs are obtained automatically from images without the need for manual complementary annotation. Only the pre-processing stage requires supervised training. Having obtained the input feature from this stage, the feature transformer was trained in an unsupervised manner.

As discussed earlier, we use a fixed visual feature encoding for both the training and the recommendation process.

To this end, we extract the global averaging pooling feature encoded by the Inception-v4 model [27]. Rather than working in the original feature space with 1536 dimensions, we further reduce the dimensionality to 128 by Principal Component Analysis. This helps address the curse of dimensionality and further reduces the computational load.

4.2. Training and Network Parameters

We use the Adam optimizer [14] with starting learning rate of 0.0002 for both the discriminator and the transformer networks. To improve training stability, we use one-sided label noise [25]. Each minibatch for training the discriminator consists of an equal proportion of synthetic and real feature pairs. Our transformer network is composed of 3 fully connected layers with 256 channels in the first two layers and 128 channels in the third. Our discriminator is composed of 3 fully connected layers with 256, 256, and 1 channel(s), respectively. The noise vector z is uniformly sampled from the unit sphere in $\mathbb{R}^{128}$. We use leaky ReLU ($\alpha = 0.2$) and batch normalization for the first two layers of both the transformer and the discriminator.

4.3. Baseline Algorithms

To demonstrate the effectiveness of CRAFT, we compare it with the following baseline algorithms:

Random recommendations: A trivial baseline generates random recommendations from a given set of candidate options, referred to as *Random*. A random selection can offer a diverse set of target items, but they may not necessarily be complementary to the query item.

Nearest neighbors of source items: A strong and useful baseline method is to find items similar to the query in the source space, and recommend their *corresponding* target items. For example, if the query item is blue jeans, this approach will recommend various tops that other people have worn with similar blue jeans. Note that these can be retrieved by performing a visual search for the nearest neighbors of the bottom item. We refer to this method as *NN-Source*.

Incompatible recommendations: Additionally, we illustrate that CRAFT has not only learned to recommend complementary apparel items, but also learned the concept of visual incompatibility. We design a method for generating *Incompatible* recommendations by suggesting tops that are assigned low discriminator scores by the CRAFT network.

5. Results

Now we present an in-depth analysis of our algorithm and comparisons with baseline methods. In Section 5.1, we visualize how the learned transformer network *dynamically* reacts to given queries in terms of assigning compatibility scores for candidate tops. We then show qualitative
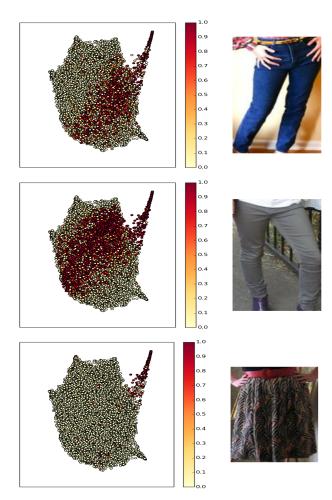
Figure 4. Each row shows a 2D t-SNE embedding of all the candidate tops (left) with the corresponding query image (right). The colors represent the discriminator score for tops conditioned on the query (red: high score, yellow: low score). Note that the discriminator in CRAFT is able to learn that common bottoms such as blue jeans and gray pants are compatible with a wide range of tops as compared to rarer query items such as the patterned skirt shown in the last row.

results of recommendations produced by various algorithms in Section 5.2. Finally, we describe our study with domain experts in Section 5.3 and analyze the results in Section 5.4.

5.1. Visualization of The Discriminator Output

In order to visualize the space of candidate top items, we projected them to a two-dimensional (2D) subspace using t-SNE [29]. The discriminator output can be seen as a proxy for the compatibility score between any top and a given query item. Each row in Figure 4 shows 2D embedding of all the tops in the dataset, color coded by the discriminator score for each top, given the bottom query item (shown on the right). Note that the compatibility scores for various candidate tops change depending on the query bottom. The yellow colors in the t-SNE plot denote low compatibil-

ity, while shades of orange to red denote high compatibility (see color bar). It is interesting to note how universal items such as blue jeans or gray pants are compatible with a large set of candidate tops, while rare bottoms like the richly textured pattern skirt shown on the bottom row are compatible with only a handful of tops. This illustrates that CRAFT is able to model the distribution of real item pairs.

5.2. Qualitative Results

Figure 5 shows qualitative results of the different recommendation methods for two query items. For this experiment, we generated 8 top recommendations from each algorithm and asked a fashion specialist to identify the top items that complement the given bottom query. While all of the approaches produce visually diverse recommendations, not all of them are compatible with the query. For a common bottom outfit such as dark jeans (Figure 5(a)), NN-Source perform as well as CRAFT, while for a less common bottom such as bright pink skirt (Figure 5(b)) they perform worse (see Section 5.4 for a more thorough analysis). This is aligned with our intuition that the quality of NN-Source recommendation highly depends on the proximity of the neighbors of the query. Interestingly, the results of the 'incompatible' algorithm demonstrate that our discriminator is able to learn not only the concept of visual compatibility, but also *incompatibility*: it often produces unusual outfit recommendation (e.g., the fur top as the third item in Figure 5(a)) that is not likely to complement the given bottom query.

5.3. User Study Design

When recommendations are provided from an open ended set, they are difficult to evaluate in absolute terms. Typically, different recommendation approaches are compared via A/B testing. Furthermore, for subjective domains such as fashion, it is preferable to obtain input from domain experts who are familiar with nuances involved in making style-appropriate recommendations. We adopt A/B testing as the main methodology to compare our proposed approach to various baselines described in Section 4.3. We evaluate the relevance of recommendations generated by each algorithm by measuring their *acceptance* by domain experts.

We approached a panel of four fashion specialists (FS) to provide feedback on recommendations generated by various algorithms. Each FS was presented with 17 recommendations for a given query (bottom) item, for each of the four algorithms. Among these recommendations, the FS were asked to select those that they judge to be complementary to the query. We used a total of 64 different query bottoms in this study, ranging for popular bottoms such as blue jeans to less common bottoms such as richly patterned skirts (see the Supplementary material for the full list). The images

(a) Complementary recommendation for a common query item (dark jeans)

(b) Complementary recommendation for a less common query item (pink skirt)
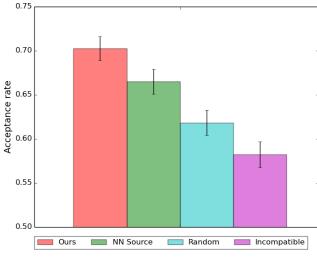
Figure 5. Comparison of the results from different recommendation algorithms. Highlighted in green are the items that have been marked as complementary to the query input by a fashion specialist. The CRAFT approach generates better (see Section 5.4 for quantitative analysis) and diverse recommendations.

were presented to FS in a random order to eliminate any bias for the algorithm or query items.

Since some FS are in general more selective than others, we need to normalize for their individual bias. To achieve this, we add the *actual* top worn by the user in the query outfit to the set of 17 recommendations at a random location. We normalize the FS acceptance scores by their likelihood of selecting the actual top as an acceptable recommendation. Note that we only perform analysis on the newly recommended tops, and exclude the original top from our results.

5.4. Analysis

Figure 6(a) shows the average rate of acceptance of generated recommendations for all FS for the four algorithms. As discussed, acceptance rates were normalized by the probability of each FS accepting the actual top for the given query bottom. The error bar denotes the 95% confidence interval for each of the results. Non-overlapping error bars indicate that the differences between the two results are statistically significant. The NN-Source algorithm has the overall acceptance score of 66.5 ± 1.4 and works bet-

(a) Overall acceptance rates
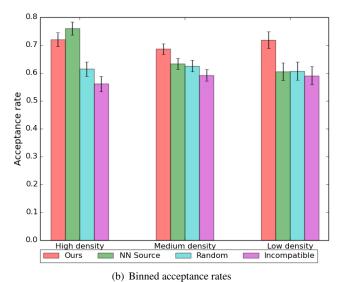

(b) Binned acceptance rates

Figure 6. Mean acceptance rate of recommendations for the different algorithms, as evaluated by fashion specialists (error bars indicate 95% confidence intervals). (a) Overall acceptance rating for each algorithm. (b) Acceptance ratings binned according to the density (high, medium, low) of query items in the feature space.

ter than the *Random* and *Incompatible* baseline algorithms as expected. The CRAFT approach generates recommendations with the highest FS acceptance score (70.3 ± 1.4).

Stratification by Feature Space Density: It is even more interesting to break down the analysis of the results in terms of the density of the query items in the feature space. Intuitively, the NN-Source algorithm should perform well in providing diverse complementary recommendations for high density regions in bottom feature space (*i.e.*, popular bottoms such as blue jeans). However, a nearest neighbor based approach would not perform well in low density regions, corresponding to rare bottom examples (e.g., striped purple skirt).

To validate this hypothesis, we approximate the density of each query point by taking the average distance to $K = 25$ nearest neighbors and bin the queries into low, medium, and high density regions, respectively. Figure 6(b) shows the average recommendation acceptance rate provided by FS for each algorithm in each density region. Again, the error bars denote the 95% confidence interval for each result. For queries that fall in the high density regions, the difference between our proposed approach and the NN-Source algorithm is statistically insignificant (error bars overlap). This is expected since nearest neighbor search is a good estimator of the conditional distribution of tops given a bottom for high density regions, where a large number of bottoms are available. However, the NN-Source algorithm starts to degrade at the medium density level, and eventually degenerates to similar performance as the *Random* and the *Incompatible* recommendation algorithms for low density regions. In contrast, the performance of CRAFT is consistent across all regions and is better than baseline algorithms for mid and low density regime. Thus, the proposed conditional transformer is able to generalize well irrespective of the density of the neighborhood surrounding the query item.

6. Conclusion and Future Work

We presented CRAFT, an approach to visual complementary recommendation by learning the *joint* distribution of co-occurring visual objects in an unsupervised manner. Our approach does not require annotations or labels to indicate complementary relationships. The feature transformer in CRAFT samples from a *conditional* distribution to generate diverse and relevant item recommendations for a given query. The recommendations generated by CRAFT are preferred by the domain experts over those produced by competing approaches.

By modeling the feature level distributions, our framework can potentially enable a host of applications, ranging from domain adaptation to one- or few-shot learning. The current work could be extended to incorporate the end-to-end learning of domain-related encoders as part of the generative framework.

References

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowledge and Data Engineering*, 17(6):734–749, June 2005. 1

[2] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR*, pages 95–104, 2017. 2, 3

[3] T. Che, Y. Li, A. P. Jacob, Y. Bengio, and W. Li. Mode regularized generative adversarial networks. In *ICLR*, 2017. 2

[4] I. Croitoru, S.-V. Bogolin, and M. Leordeanu. Unsupervised learning from video to detect foreground objects in single images. In *ICCV*, pages 4335–4343, 2017. 3

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 3

[6] C. Doersch and A. Zisserman. Multi-task self-supervised visual learning. In *ICCV*, 2017. 3

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, and S. Ozair. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014. 2, 3

[8] X. Han, Z. Wu, Y. Jiang, and L. S. Davis. Learning fashion compatibility with bidirectional lstms. In *ACM Multimedia*, 2017. 3

[9] R. He, C. Packer, and J. McAuley. Learning compatibility across categories for heterogeneous item recommendation. In *IEEE 16th International Conference on Data Mining, ICDM*, pages 937–942, 2016. 3

[10] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456, 2015. 4

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial nets. In *CVPR*, 2017. 2

[12] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *CoRR*, abs/1710.10196, 2017. 2

[13] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg. Where to buy it: Matching street clothing photos in online shops. In *ICCV*, 2015. 3

[14] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 5

[15] Y. Koren and R. Bell. Advances in collaborative filtering. In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, *Recommender Systems Handbook*, pages 145–186. 2011. 1, 3

[16] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, and M. Ranzato. Fader networks: Manipulating images by sliding attributes. *CoRR*, abs/1706.00409, 2017. 3

[17] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1):1–19, Feb. 2006. 1

[18] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan. Deep human parsing with active template regression. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(12):2402–2414, Dec. 2015. 5

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, T. Pajdla, B. Schiele, and T. Tuytelaars. *Microsoft COCO: Common Objects in Context*. Springer International Publishing, 2014. 3

[20] S. Liu, Z. Song, M. Wang, C. Xu, H. Lu, and S. Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. *ACM Multimedia*, pages 1335–1336, 2012. 3

[21] Y. Lu, Y. Tai, and C. Tang. Conditional CycleGAN for Attribute Guided Face Image Generation. *CoRR*, abs/1705.09966, 2017. 2

[22] Maas, Andrew L and Hannun, Awni Y and Ng, Andrew Y. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, volume 30, 2013. 4

[23] J. J. McAuley, C. Targett, Q. Shi, and A. van den Hengel. Image-based recommendations on styles and substitutes. In *SIGIR*, 2015. 3

[24] P. Melville, R. J. Mooney, and R. Nagarajan. Content-boosted collaborative filtering for improved recommendations. In *Eighteenth National Conference on Artificial Intelligence*, pages 187–192, 2002. 1

[25] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In *NIPS*, pages 2234–2242, 2016. 5

[26] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *CVPR*, 2017. 2, 3

[27] C. Szegedy, S. Ioffe, and V. Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *AAAI*, abs/1602.07261, 2017. 5

[28] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Adversarial discriminative domain adaptation. In *CVPR*, 2017. 2, 3

[29] L. van der Maaten and G. Hinton. Visualizing high-dimensional data using t-SNE. *JMLR*, 9:2579–2605, Nov 2008. 6

[30] A. Veit, B. Kovacs, S. Bell, J. McAuley, K. Bala, and S. Belongie. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *International Conference on Computer Vision (ICCV)*, 2015. 3

[31] C. Vondrick and A. Torralba. Generating the future with adversarial transformers. In *CVPR*, 2017. 3

[32] S. Zhang, L. Yao, and A. Sun. Deep learning based recommender system: A survey and new perspectives. *CoRR*, abs/1707.07435, 2017. 3

[33] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, 2017. 5

[34] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2, 3

[35] S. Zhu, S. Fidler, and R. Urtasun. Be your own prada: Fashion synthesis with structural coherence. In *ICCV*, 2017. 2