Effective Fashion Retrieval Based on Semantic Compositional Networks

Dan Valle and Nivio Ziviani
CS Dept., UFMG & Kunumi
Brazil
Email: {dan,nivio}@kunumi.com

Adriano Veloso
CS Dept., UFMG
Brazil
Email: adrianov@dcc.ufmg.br

***Abstract*—Typical approaches for fashion retrieval rank clothing images according to the similarity to a user-provided query image. Similarity is usually assessed by encoding images in terms of visual elements such as color, shape and texture. In this work, we proceed differently and consider that the semantics of an outfit is mainly comprised of environmental and cultural concepts such as occasion, style and season. Thus, instead of retrieving outfits using strict visual elements, we find semantically similar outfits that fall into similar clothing styles and are adequate for the same occasions and seasons. We propose a compositional approach for fashion retrieval by arguing that the semantics of an outfit can be recognised by their constituents (i.e., clothing items and accessories). Specifically, we present a semantic compositional network (*Comp-Net*) in which clothing items are detected from the image and the probability of each item is used to compose a vector representation for the outfit. *Comp-Net* employs a normalization layer so that weights are updated by taking into consideration the previously known co-occurrence patterns between clothing items. Further, *Comp-Net* minimizes a cost-sensitive loss function as errors have different costs depending on the clothing item that is misclassified. This results in a space in which semantically related outfits are placed next to each other, enabling to find semantically similar outfits that may not be visually similar. We designed an evaluation setup that takes into account the association between different styles, occasions and seasons, and show that our compositional approach significantly outperforms a variety of recently proposed baselines.**

***Index Terms*—Fashion Retrieval, Learning Representations**

I. Introduction

An extraordinarily large number of photos are posted every day on fashion blogs. As a consequence, users searching for fashion inspiration may navigate for hours with no guarantee of finding relevant or potentially inspiring outfits. A number of fashion retrieval approaches have being proposed recently, and they usually operate by retrieving outfits that are similar to a query image in terms of visual elements such as color, shape and texture [1]. While these approaches are effective in retrieving outfits with similar visual appearance, they do not take into consideration the semantics of the outfits, which is related to abstract concepts such as occasion, season and style, e.g., certain outfits may be appropriate for the same times and places, even if these outfits are not visually similar.

In this paper, we would like to answer questions like "which outfits are associated with similar styles, and are suitable for the same occasions and seasons?" We present semantic compositional networks to improve the retrieval of semantically similar outfits posted in fashion sites and blogs, that is, the user provides a query image showing her outfit and we find outfits that are semantically related to the query. The accuracy in effectively ranking the returned outfits may drastically reduce the time and frustration that occurs while searching for fashion inspiration.

Some of the challenges in this process are:

- Outfits are semantically similar if they are associated with the same style and are appropriate for the same occasions and seasons [2]. However, outfits have a large variation in style, occasion and season, and it is hard to directly compare outfits based on these abstract dimensions.
- The visual appearance and the semantics of an outfit may not be directly correlated. Outfits may be visually similar and still be associated with different styles or be appropriate to different occasions. Also, outfits may be visually different but be associated with the same styles, occasions and seasons. For example, boots are associated with several clothing styles, but the co-occurrence of boots and hat strongly suggests that the outfit is associated with the boho chic style. Similarly, shorts may be appropriate to several occasions, but the co-occurrence of shorts, slippers and sunglasses suggests that the outfit is appropriate to an event in the beach.

In view of these challenges, we propose a compositional approach for fashion retrieval. Compositionality allows us to learn feature vectors for accurately representing outfits based on the occurrences of clothing items, and this has a fundamental motivation since it is relatively easy to obtain images of outfits labeled with their constituent items such as hat, bag, pants, shoes, and so on. By contrast, there may be debate on whether an outfit should be associated with a style or other, or if the outfit is suitable or not to certain occasions. Further, while low-level visual features, such as color, shape and texture, do not carry enough information to find outfits that are semantically similar, compositionality allows us to match semantically close outfits that may not be visually similar.

Convolutional networks (CNNs) have long been applied to object recognition in images [3]. Still, recognizing clothing items in images is particularly hard, and recent works have proposed increasingly complex CNN models, with millions of parameters to learn [4], [5]. Clothing items are frequently subject to deformations and occlusion [6], and exhibit serious variations when they are taken under different scenarios [5]. Further, mutually exclusive clothing items (e.g., sandals and high-heels) may get misclassified by the network, and these specific errors may hurt retrieval performance seriously, as they are likely to result in representations associated with antagonic styles, occasions and seasons. For instance, sandals are associated with informal occasions and are typically used in the summer. High-heels, on the other hand, are more formal and elegant. Therefore, misclassifying mutually exclusive items may associate a wrong semantics to the outfit.

Fortunately, there are clear co-occurrence patterns between the various clothing items, and we may exploit these patterns in different ways in order to learn improved representations with reduced network capacity. Firstly, if the network gets confused between skirts and dresses, then the occurrence of tops can be used to increase the odds of skirts. To model these co-occurrence patterns, we introduce a normalization layer which updates the probabilities of clothing items by taking into account the co-occurrence information between them. Further, in order to help the network to better distinguish between mutually exclusive clothing items, we employ a cost-sensitive mimization which puts a higher cost when mutually exclusive items are misclassified. After some optimization iterations the learned representation is expect to converge to the most likely probabilities for each clothing constituent.

Differently from previous works, we consider that the semantics of an outfit is mainly comprised of environmental and cultural concepts such as occasion, style and season. Thus, we judge relevance based on how semantically similar are the query and the returned outfit. In this case, a perfect match occurs if both outfits are associated with the same style and are appropriate to the same occasions and seasons. On the other extreme, we consider that two outfits are completely irrelevant to each other if they are associated with different styles, occasions and seasons. Relevance varies between these extremes and we considered two possible scenarios to assess relevance. In a more strict scenario, we do not take into account any possible relationship between different styles, occasions and seasons. Thus, an outfit must be associated with only one style, and there is only one occasion and only one season for which it is appropriate. As a result, relevance vanishes if there is not an exact match between the semantics of the query and the semantics of the returned outfit. The second scenario takes into account the relationship between different styles, occasions and seasons. This means that an outfit that is suitable for a wedding may also be suitable (to some extent) for a graduation. We built a SkipGram model [7] in order to measure the extent to which different styles, occasions and seasons are related to each other, and relevance is assessed by considering the relationship between them.

Contributions and Findings: The main contribution of this paper is to present new effective fashion retrieval models based on semantic compositional networks. In practice, we claim the following benefits and contributions:

- We represent outfits in a semantic level, following a compositional approach in which dimensions correspond to the occurrence of clothing items. A semantic compositional network (which we call *Comp-Net*) computes the probability of each clothing item, following a cost-sensitive optimization that punishes more severely errors involving mutually-exclusive items. Cost-sensitive optimization enables our network to remain relatively simple when compared with deeper models for fashion retrieval.
- We introduce a conditional normalization layer that is used in conjunction with a softmax function to update the probabilities so that the interplay between clothing items are taken into consideration. The final result is that outfits are placed on a semantic space, enabling the search for outfits that are semantically related.
- We built a new dataset for fashion retrieval and conducted a rigorous set of experiments to show that the proposed model outperforms recent fashion retrieval models based on dense representations.

II. Related Work and State of the art

First we review clothing recognition methods and second we discuss the literature on machine learning for fashion retrieval.

A. Clothing Recognition and Parsing

Early work focused on investigating the capability of typical image descriptors for clothing attribute classification [8]. In [9], the authors formulate clothing parsing as a pixel labeling problem, where images are segmented into superpixels and then clothing labels for every segment are predicted using a Conditional Random Field (CRF) model. In [10], [11], the authors proposed a deep convolutional network model for detecting various clothing items a person in the image is wearing or carrying. In [12], the authors proposed a joint image segmentation and parsing approach for clothing parsing. Authors in [13] proposed a Dual Attribute-aware Ranking Network, called DARN, addressing the problem of cross-domain image retrieval, where there is large discrepancy between query images and images in the database.

Some works have already exploited the relationship between objects in the image in order to improve recognition. In [14], the authors used a multi-attribute CRF to capture the mutual dependencies between clothing attributes in order to make attribute predictions. The final output of the system is a list of nameable attributes that describe the clothing appearance. Similarly, in [15] the authors have taken into account the compatibility of clothing items in order to improve attribute predictions. They considered inter-object or inter-attribute compatibility in the recognition problem, and formulate a CRF that seeks a proper combination in the given outfit. Authors in [16] addressed the problem of automatically parsing fashion images with weak supervision from the user-generated color-category tags such as “red jeans” and “white T-shirt”. This problem is similar to ours in the sense that we also perform a weakly supervised task, but we focus on fashion retrieval, while [16] focused on clothing parsing.

B. *Fashion Retrieval*

Fashion retrieval encompasses the task of finding the most similar outfits to a query. In [17], the authors study the problem of personalized outfit recommendation, that is, to automatically suggest outfits to users that fit their personal fashion preferences. Unlike existing recommendation systems that usually recommend individual items, they suggest sets of items. In [18], the authors propose an approach to match an example of a clothing item to the same item in an online shop. They used a standard CNN model to represent outfits. In [19], the authors present a CNN model for rapid fashion retrieval in a recommendation system. In [20], the authors propose an approach to find similar styles from a large database of tagged fashion images. In [1], the authors propose an approach to find outfits with similar styles using typical image descriptors as features for learning to rank algorithms. In [21], the authors develop a fashion retrieval system considering local similarity, where users can retrieve clothes which are globally similar to an image and partially similar to another image. In [22], the authors introduce a new problem in which the goal is to predict how fashionable a person looks on a photograph and to suggest subtle improvements she could make to improve her appeal.

Authors in [23] proposed a Siamese CNN architecture that learns a visual notion of compatibility across fashion categories. Compatibility is modeled based on co-occurrence of clothing items in co-purchase data from Amazon.com. Authors in [24] proposed models for the One-Class Collaborative Filtering setting, where the goal is to estimate users' fashion-aware personalized ranking functions based on their past feedback. In [25], the authors propose an approach that matches a user specified occasion (e.g., wedding or shopping), with suitable clothes from online shops. In [4], the authors propose the *StyleNet-1.0* model which employs weak labels, such as tags associated with the outfits, in order to learn representations for outfits. Such representations are used to place outfits into a fashion style space. Instead of considering attributes such as styles and occasions independently as in [25] and [4], our proposed model places outfits in a multi-attribute semantic space where styles, occasions and seasons are taken into account jointly, so that we can assess and compare their appropriateness to specific styles, occasions and seasons. Authors in [5] proposed the *FashionNet* model which learns clothing features by jointly predicting clothing attributes and landmarks. We consider [4] and [5] the closest works to ours and consequently we include *StyleNet-1.0* and *FashionNet* as baselines in order to evaluate our proposed *Comp-Net* model.

Our Work: In order to achieve superior ranking performance, fashion retrieval models are becoming increasingly complex and hard to train [4], [5]. In contrast, in this paper we claim that fashion retrieval models can be greatly simplified by following a cost-sensitive minimization and by normalizing the output probabilities according to the observed interplay between different clothing items. Our proposed network is able to achieve state-of-the-art performance while being relatively simple when compared with state-of-the-art representatives.

III. *Comp-Net*: SEMANTIC COMPOSITIONAL NETWORK

We tackle the fashion retrieval task by using a retrieval based approach − the user provides a query image showing her outfit, and we find outfits that are semantically related to the query in a large database of outfits posted in blogs and fashion communities. Our approach for fashion retrieval, which we call *Comp-Net*, is divided into two main steps: (i) learning a semantic space in which outfits are effectively represented, and (ii) ranking outfits according to their similarity to a given query. We assume that information of occurrence of clothing items is abundant and available in the form of "weak labels" [4]. This information is necessary for learning compositional feature vectors.

A. *Learning the Semantic Space*

Building an effective feature set to represent clothing images is of paramount importance for improved fashion retrieval. In particular, we want features to be robust to background changes and to focus entirely on the outfit. Further, features should be meaningful to fashion attributes such as styles, occasions and seasons. Thus, we exploit the composition of outfits, so that outfits are represented by observing how likely are the possible constituents. Convolutional neural networks (CNNs) are renowned for their high recognition performance and are thus one of the must-try algorithms [26]. In particular, we used 2×2 kernels for the convolutional filters to keep the number of weights down for the network and allow increasing the number of layers [27]. A preliminary analysis showed that dropout in the convolutional layers was not beneficial, and thus dropout is used only in the fully-connected layer to prevent overfitting throughout the architecture. The network output is given as a vector of probabilities associated with k clothing items (i.e., blazer, shirt, skirt, dress etc.), that is, a compositional feature vector. A full overview of the architecture can be seen in Table I.

1) Conditional Normalization: We observed that certain clothing items are often related to other items. Essentially, mutual dependencies between various clothing items capture the rules of style [14]. To model these rules, a normalization layer is employed on top of the CNN predictions for each individual clothing item, so that softmax probabilities are updated by taking into account co-occurrence information. We assume that co-occurrence information is given in terms of conditional probabilities $p(x_j|x_i)$ which gives the probability that clothing item x_j occurs given that item x_i also a constituent of the outfit. Conditional probabilities are easily obtained directly from the training set. The conditional normalization layer essentially applies a normalization factor to the probability of each clothing items as shown in:

$$P'(x_i) = P(x_i) \times \nabla_{x_i} \tag{1}$$

where $P(x_i)$ is the initial probability estimate for clothing item x_i, and $P'(x_i)$ is the normalized estimate which will be given

TABLE I
COMPOSITION LEARNING NETWORK ARCHITECTURE.

type	kernel size	output size	#params
convolution	2×2	$31 \times 31 \times 32$	416
convolution	2×2	$30 \times 30 \times 32$	4,128
max pooling	2×2	$15 \times 15 \times 32$	
convolution	2×2	$14 \times 14 \times 64$	8,256
convolution	2×2	$13 \times 13 \times 64$	16,448
max pooling	2×2	$6 \times 6 \times 64$	
convolution	2×2	$5 \times 5 \times 128$	32,896
convolution	2×2	$4 \times 4 \times 128$	65,664
max pooling	2×2	$2 \times 2 \times 128$	
fully-connected		1,024	525,312
dropout (0.5)		1,024	
fully-connected		1,024	1,049,600
dropout (0.5)		1,024	
fully-connected		k	1,024×k
cond-normalization		k	
softmax		k	
Total			1,702,700 + 1,024×k

as input to a softmax layer. Further, ∇_{x_i} is the normalization factor, as shown in:

$$\nabla_{x_i} = \frac{\sum_{j}^{k} P(x_j) \times P(x_j|x_i)}{k} \tag{2}$$

The basic intuition is to normalize $P(x_i)$ according to the observed conditional probabilities $P(x_j|x_i)$. For instance, if item x_i is likely to occur, and it is highly associated with other item x_j, then the probability of x_j will increase accordingly. Similarly, if item x_i is unlikely to occur, then the probability of x_j will decrease accordingly.

2) Cost-Sensitive Minimization: Prediction errors are not equally harmful. Misclassifications involving mutually exclusive clothing items may shift the outfit representation, making it closer to wrong styles, occasions and seasons. We used stochastic gradient descent [28] to minimize the binary cross-entropy of the training set, but we put a higher cost when mutually exclusive items are misclassified. We used conditional probabilities $P(x_j|x_i)$ to build a cost matrix C, as shown in:

$$C_{i,j} = \alpha - (1 - \alpha) * P(x_j|x_i) \tag{3}$$

where α is the highest cost imposed to a misclassification. Thus, costs in C vary from 1 (i.e., all misclassifications have the same weight) to α. Cost-sensitive minimization may not produce fewer errors, but it is likely to result in networks that perform better in practical terms because it has a built-in bias in favor of less expensive errors.

3) Training: Learning rate was set to 0.01 and k was set to 20. We used Rectifier Linear Units (Relu) as non linear activations and a dropout probability of 0.5. The mini-batch size is fixed to 16 and training was stopped after 50 epochs with no improvement. We perform a grid search for these hyper-parameters, tuning on the validation set, with early stopping. The best model was chosen according to the smallest loss on the validation set.

B. Ranking Outfits using the Semantic Space

Once outfits are properly represented by compositional feature vectors, we can compare them using the semantic space. We hypothesize that the nearest neighbors in the semantic space will more similarly match a query image. The semantic space contains:

- A set of images (i.e., outfits) $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$. Each outfit $x_i \in \mathcal{X}$ is associated with a triple $f(x_i) =$<*style, occasion, season*> that denotes the semantics of x_i.
- Vector representations for outfits, i.e., $\forall\ x_i \in \mathcal{X}$ there exists a vector $v(x_i) \in \mathcal{R}^k$.
- A measure for similarity $s : \mathcal{R}^k \times \mathcal{R}^k \rightarrow \mathcal{R}$, such that $0 < s(v(x_1), v(x_2)) \leq 1$.

Assumptions: With the above definitions, we have the following assumptions, which serve as motivation for our fashion retrieval model:

- The semantic space $\mathcal{X}$ can be viewed as an arrangement of semantic regions. Each region refers to specific styles, occasions and seasons.
- The semantics of an outfit is embedded in its similarity with other outfits. That is, the similarity $s(v(x_i), v(x_j))$ is compatible with $|f(x_i) - f(x_j)|$.

Outfits can be efficiently compared using the Euclidean distance $||\ .\ ||_2$. In this case, feature pairs are first normalized to have unitary norm. This process is efficient and does not require additional steps devoted to learn ranking functions.

IV. EXPERIMENTAL RESULTS

In this section, we present the data and baselines used to evaluate fashion retrieval models. Then we discuss our evaluation procedure and report our results. In particular, our experiments aim to answer the following research questions:

RQ1: How effective is conditional normalization?
RQ2: How meaningful are the semantic spaces created from constituent clothing items?
RQ3: How effective is cost-sensitive minimization? What is the effect of α in Equation (3)?
RQ4: Does *Comp-Net* improve state-of-the-art effectiveness in semantic fashion retrieval?

A. Dataset

Chictopia (www.chictopia.com) is a website designed for fashion enthusiasts and bloggers to create profiles, post outfits, and socialize with others interested in fashion. The site currently has over 300,000 users. Each post is associated with an image (i.e., an outfit) and several tags indicating the occurrence of certain clothing items in the outfit. A post also contains tags indicating the fashion style of the outfit, as well as the occasion and season for which the outfit is most appropriate. To evaluate the effectiveness of *Comp-Net* we collected a subset of Chictopia, which comprises 67,715 images accompanied with tags indicating the constituent clothing items and also tags indicating the style, occasion and season, enabling us to compare outfits in a semantic level.

B. Relevance

Relevance $r(x_i, x_j)$ is given as some difference between the semantics of query x_i and a returned outfit x_j. A perfect match occurs when both outfits share the style and are appropriate for the same season and occasion. By contrast, a totally irrelevant outfit does not share any of these attributes with the query. However, we interpret the difference between $f(x_i)$ and $f(x_j)$ in two different ways:

- We consider a strict scenario where styles, occasions and seasons are not semantically related. In this case, relevance is simply given as the intersection-union ratio involving styles, occasions and seasons associated with query and a returned image:
$$r_1(x_i, x_j) = \frac{|f(x_i) \cap f(x_j)|}{|f(x_i) \cup f(x_j)|}$$
- We also considered a more realistic scenario where styles, occasions and seasons are semantically related. In this case, relevance is given as:
$$\begin{aligned} r_2(x_i, x_j) = \text{avg}(\cos(f_{style}(x_i), f_{style}(x_j)) \quad &+ \\ \cos(f_{occasion}(x_i), f_{occasion}(x_j)) \quad &+ \\ \cos(f_{season}(x_i), f_{season}(x_j))) \end{aligned}$$
where:
  - $f_{style}(x_i)$ is a $k-$dimensional SkipGram vector representation associated with the style tag in $f(x_i)$.
  - $f_{occasion}(x_i)$ is a $k-$dimensional SkipGram vector representation associated with the occasion tag in $f(x_i)$.
  - $f_{season}(x_i)$ is a $k-$dimensional SkipGram vector representation associated with the season tag in $f(x_i)$.

  Basically, we learn SkipGram representations for the tags by considering the triple <*style, occasion, season*> as the context. Thus, similar vector representations are assigned to style tags that frequently appear together with the same occasion and season tags. In the same way, vector representations are assigned to occasion and season tags.

C. Training and Evaluation Procedure

We extracted the 67,715 images to be used as training and validation sets. Clothing tags associated with these images are used as labels. We conducted five-fold cross-validation, and at each run, four folds are used as training set, and the remaining fold is used as validation set. Once the best hyper-parameters are found, we finally trained each network using its respective images and produced their representations for unseen data in their test sets.

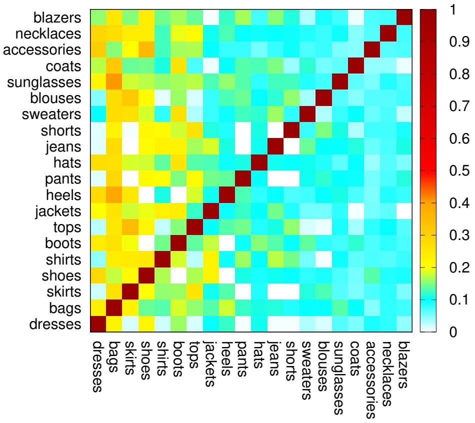

Fig. 1. (Color online) Conditional probability matrix for different items.

1) Conditional Probabilities: We also used the training images to calculate the conditional probabilities for each permutation pair of clothing items, as shown in Figure 1 (notice that conditional probability is asymmetric, and thus the full matrix is shown in the figure). The figure considers only the 20 most frequent clothing items in order to avoid clutter. Some expected relationships are evident, such as $P(\text{dresses}|\text{shorts}) = 0$ (i.e., these items are mutually exclusive, and thus do not occur together in any outfit), or $P(\text{shoes}|\text{dresses})$ is high. This conditional probability matrix is used in Equations (2) and (3).

2) Ranking: The transformed test images are used as queries, and for each query a ranking containing 1,000 images/outfits (sampled from the same set) is returned. To evaluate the retrieval performance we used the standard NDCG (normalized discount cumulative gain) measure [29]. Notice that NDCG numbers will vary greatly depending on whether we are using $r_1(x_1, x_2)$ or $r_2(x_i, x_j)$ to measure relevance.

D. Baselines

We considered the following methods in order to provide baseline comparison:

- *StyleNet-1.0* [4]: A feature extraction network that minimizes a ranking loss and a classification network that minimizes the cross-entropy loss are trained jointly. The input for the network is composed of "weak labels", which is similar to the input of our models.
- *FashionNet* [5]: A deep network devised to jointly predict clothing items and landmarks. The estimated landmarks are then employed to pool/gate the learned features.

E. Results

We divided our analysis into qualitative and quantitative discussions in order to answer the research questions presented at the beginning of Section IV.

1) Conditional Normalization: The first experiment is devoted to answer RQ1. For this, we show the benefits of taking into account the relationships between clothing items using conditional normalization while learning *Comp-Net*.

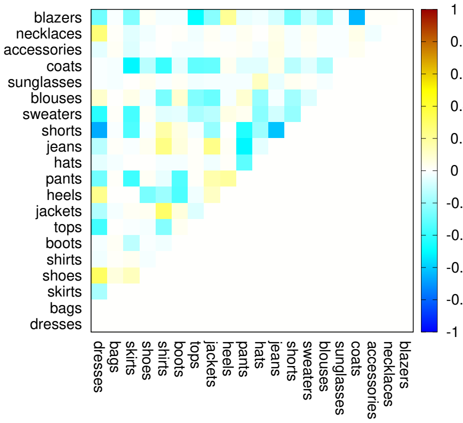

Fig. 2. (Color online) The difference in correlation between $\mathcal{M}$ and $\mathcal{M}'$, showing pairs of items for which the correlation increased or decreased.

Specifically, we evaluate the extent to which the conditional normalization layer modifies item probabilities and, consequently, the compositional representations. For this, we first ran a version of *Comp-Net* without the conditional normalization layer, which we call *Comp-Net$'$*. Then we took all the compositional vectors and since each dimension of these vectors correspond to a specific clothing item, we calculated the Pearson correlation between every pair of items, resulting in a correlation matrix which we call $\mathcal{M}'$. Then, we also ran *Comp-Net* with the conditional normalization layer, and calculated the corresponding correlation matrix, which we call $\mathcal{M}$. Finally, we calculated $\mathcal{M} - \mathcal{M}'$, resulting in the matrix shown in Figure 2. Most of the differences are low, ranging from -0.1 to 0.1. In some cases, however, it can be seem that the correlation between some clothing items (such as shoes and dresses, or jackets and shirts) has greatly increased. There are also pairs of clothing items for which the correlation has greatly decreased (e.g., shorts and dresses, or blazers and coats). We conclude that conditional normalization is effective in capturing the interplay between clothing items.

2) Semantic Fashion Space: The second experiment is devoted to answer RQ2. Since our goal is to match styles, occasions, and seasons, we considered a reference model that directly learns to predict this information. Specifically, we used SkipGram representations for each possible tag as the reference model, so that we can show how close the proposed compositional representation is to a representation directly learned from tags/labels related to style, occasion, and season.

Figure 3 (Left) shows the reference model. Clearly, similar representations are assigned to tags that frequently appear together in the same context. For instance, "rocker" is a style tag that is close to "music concert" or "clubbing" occasion tags. Similarly, "chic" and "elegant" are two related style tags.

Figure 3 (Right) shows the average vector considering all outfits associated with specific season, occasion and style tags. More specifically, a season, occasion or style tags are placed at the centroid of the vectors associated with the corresponding tags. This results in another semantic fashion space. Qualitatively, the inspection of the semantic fashion space enables us to grasp the potential of our compositional representations. The "romantic" and "50s" styles are close to each other, as are the "vintage" and "60s" styles. The "retro" style is placed somewhere in between these styles. The same trend is observed for occasions: "museum" and "theatre" are close to each other, as are "pool party" and "beach". Inter-attribute relationships are also evident. For instance, the "sexy" style is close to occasions such as "girls night out" and dating occasions, or "runway" style is close to "fashion show".

Figure 4 shows the observed associations between clothing items and styles, seasons and occasions, and help us

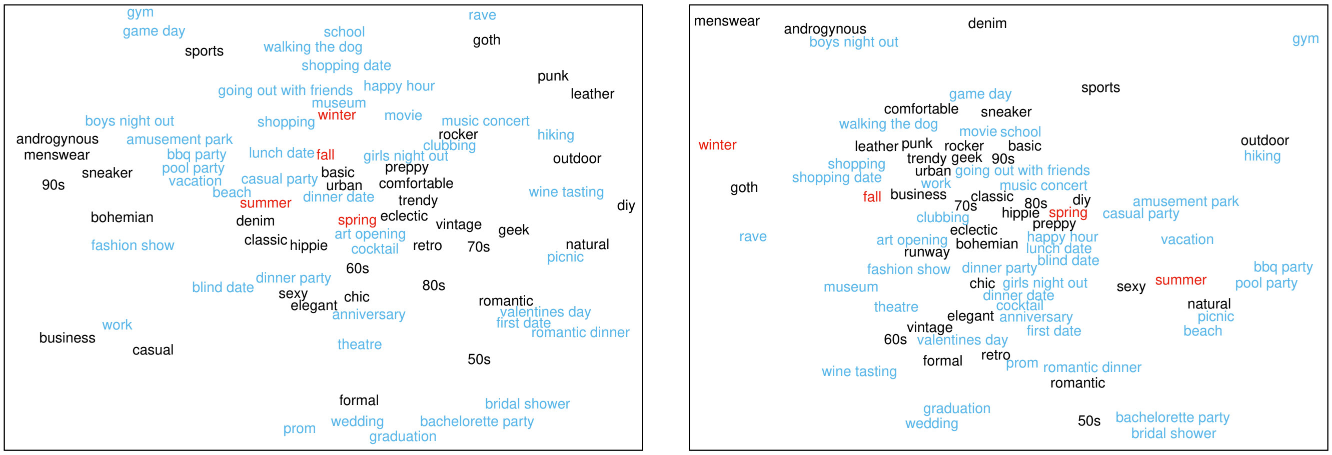

Fig. 3. (Color online) Visualization of the semantic fashion space using t-SNE [30]. Minor adjustments were made in order to avoid clutter, and the relative distances between the tags were kept the same. Left – Each style (black), occasion (blue) and season (red) tag is placed according to its SkipGram vector. Right – Each style (black), occasion (blue) and season (red) tag is placed at the centroid of the corresponding outfit representations learned by *Comp-Net* ($\alpha = 1.2$). For instance, the position of "winter" corresponds to the centroid considering the vector representations of all outfits that are related to winter.

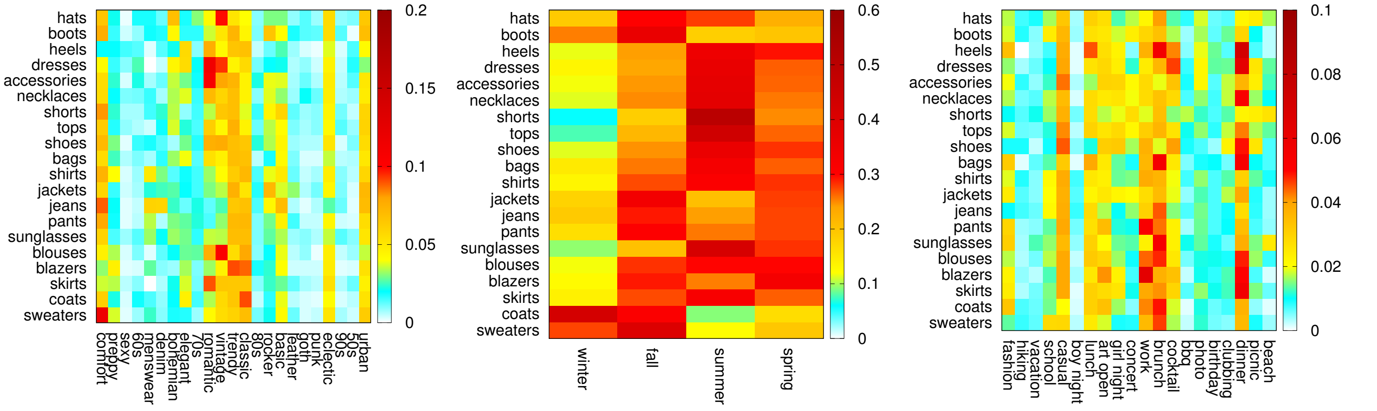

Fig. 4. (Color online) Left – Observed association (i.e., $P(x|y)$) between clothing items and styles. Center – Observed association between clothing items and seasons. Right – Observed association between clothing items and occasions. The figure considers only the 20 most frequent clothing items, and only few styles and occasions in order to avoid clutter

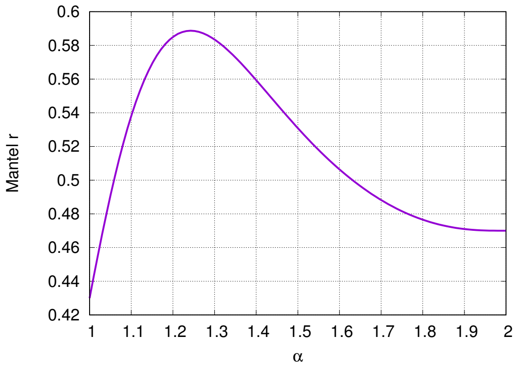

Fig. 5. Mantel r correlation coefficients for varying α values.

to understand why the semantic space learned by *Comp-Net* is meaningful. Clearly, some clothing items are largely associated with certain styles, seasons and occasions. “Coats” is much more associated with “winter” than with “summer”, or “heels” is not associated with “comfortable”. *Comp-Net* learns representations based solely on clothing items, but the association that exists between specific clothing items and the various possibilities of style, season, and occasions, makes the semantic space meaningful.

3) Cost-Sensitive Minimization: The third experiment is devoted to answer RQ3, and thus we show the benefits of cost-sensitive minimization. Specifically, we show how different cost factors α affect the semantic space learned by *Comp-Net*. We varied α from 1 to 2, that is, α varies from a scenario in which misclassifications have equal costs to another scenario in which misclassifications involving mutually exclusive items are highly penalized.

To evaluate the impact of α, we compared the semantic space learned by *Comp-Net* (using a specific α value) with the semantic space obtained with the reference SkipGram model. First, we built a distance matrix for each semantic space, where the dimensions are style, occasion and season tags. Then, we employed the Mantel r coefficient [31], which gives the spatial auto-correlation between two distance matrices. It is non-parametric and computes the significance of the correlation through permutations of the rows and columns of one of the distance matrices. The test statistic is the Pearson product-moment correlation coefficient r, which falls in the range of -1 to $+1$, where being close to -1 indicates strong negative correlation and $+1$ indicates strong positive correlation. An r value of 0 indicates no correlation. Figure 5 shows Mantel r coefficients for different α values. An r value greater than 0.40 indicates that there is relatively strong positive correlation between the two semantic spaces. Clearly, α has great impact in the resulting semantic space, and the best semantic space learned by *Comp-Net* was obtained with $\alpha = 1.2$.

4) Baseline Comparison: The last set of experiments is devoted to answer RQ4. For this, we show NDCG numbers for the execution of *Comp-Net* ($\alpha = 1.2$), *StyleNet-1.0*, and *FashionNet*. We employed both $r_1(x_i, x_j)$ and $r_2(x_i, x_j)$ in order to assess relevance. We considered $r_1(x_i, x_j)$ to be a more strict evaluation scenario, and $r_2(x_i, x_j)$ to be a more realistic evaluation scenario, as the semantic difference between predicted and actual tags are taken into account while assessing relevance.

Table II shows the retrieval effectiveness of *Comp-Net*, as well as the retrieval effectiveness of the baselines. *Comp-Net* outperforms *FashionNet* and *StyleNet-1.0*, specially in the topmost positions. Their performances tend to approach as the rank size increases. Still, *Comp-Net* numbers are statistically significant at 0.05. Finally, as expected, NDCG numbers are much higher when the different models are evaluated using $r_2(x_i, x_j)$, as in this case the relationship between different styles, occasions and seasons is taken into account.

V. Conclusions and Future Work

We have presented a novel approach for fashion retrieval. The proposed approach, *Comp-Net*, learns compositional feature vectors by estimating the probabilities of the clothing

TABLE II
RANKING EFFECTIVENESS OF THE DIFFERENT NETWORKS.

	$r_1(x_i, x_j)$			$r_2(x_i, x_j)$		
	@1	@5	@10	@1	@5	@10
FashionNet	0.394	0.505	0.698	0.746	0.826	0.916
StyleNet-1.0	0.398	0.512	0.710	0.758	0.839	0.925
Comp-Net	0.408	0.521	0.724	0.776	0.856	0.939

items that compose an outfit. Observed co-occurrence patterns between different clothing items are taken into account in order to avoid misclassifications involving mutually exclusive clothing items, leading to improved probability estimates without increasing network capacity. We employ two ways to assess relevance. The first one is more strict, and simply makes a binary matching decision. We also introduce a novel, more realistic way of assessing relevance by taking into account the relationship between different styles, occasions and seasons. Our experiments show that *Comp-Net* achieves a retrieval performance which is superior than the performance of state-of-the-art representatives. Further, the experiments also showed that the semantic space learned by *Comp-Net* is significantly similar to a reference semantic space obtained using SkipGram directly over style, occasion and season tags. Thus, qualitatively, *Comp-Net* learns compositional representations that result in a meaningful semantic fashion space that uncovers rules about what matches, what is appropriate to certain seasons, what to wear in certain situations and so on.

ACKNOWLEDGEMENTS

This work was partially funded by projects InWeb (grant MCT/CNPq 573871/2008-6) and MASWeb (grant FAPEMIG/PRONEX APQ-01400-14), and by the authors individual grants from CNPq, FAPEMIG and Kunumi.

REFERENCES

[1] M. Moreira, J. dos Santos, and A. Veloso, "Learning to rank similar apparel styles with economically-efficient rule-based active learning," in *ACM ICMR International Conference on Multimedia Retrieval*, 2014, pp. 361–371.

[2] A. Lurie, *The Language of Clothes*. New York, Random House, 1981.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE CVPR Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[4] E. Simo-Serra and H. Ishikawa, "Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction," in *IEEE CVPR Conference on Computer Vision and Pattern Recognition*, 2016, pp. 298–307.

[5] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 1096–1104.

[6] N. Wang and H. Ai, "Who blocks who: Simultaneous clothing segmentation for grouping images," in *IEEE ICCV International Conference on Computer Vision*, 2011, pp. 1535–1542.

[7] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS 27th Annual Conference on Neural Information Processing Systems*, 2013, pp. 3111–3119.

[8] J. Lorenzo-Navarro, M. C. Santana, E. Ramón-Balmaseda, and D. Freire, "Evaluation of LBP and HOG descriptors for clothing attribute description," in *International Workshop on Video Analytics for Audience Measurement*, 2014, pp. 53–65.

[9] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *IEEE CVPR International Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3570–3577.

[10] K. Nogueira, A. Veloso, and J. Santos, "Pointwise and pairwise clothing annotation: Combining features from social media," *Multimedia Tools Appl.*, vol. 7, no. 75, pp. 4083–4113, 2016.

[11] A. Veloso, J. Santos, and K. Nogueira, "Learning to annotate clothes in everyday photos: Multi-modal, multi-label, multi-instance approach," in *SIBGRAPI*, 2014, pp. 327–334.

[12] X. Liang, L. Lin, W. Yang, P. Luo, J. Huang, and S. Yan, "Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1175–1186, 2016.

[13] J. Huang, R. S. Feris, Q. Chen, and S. Yan, "Cross-domain image retrieval with a dual attribute-aware ranking network," in *IEEE ICCV International Conference on Computer Vision*, 2015, pp. 1062–1070.

[14] H. Chen, A. C. Gallagher, and B. Girod, "Describing clothing by semantic attributes," in *ECCV European Conference on Computer Vision*, 2012, pp. 609–623.

[15] K. Yamaguchi, T. Okatani, K. Sudo, K. Murasaki, and Y. Taniguchi, "Mix and match: Joint model for clothing and attribute recognition," in *BMVC British Machine Vision Conference*, 2015, pp. 1–12.

[16] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, and S. Yan, "Fashion parsing with weak color-category labels," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 253–265, 2014.

[17] Y. Hu, X. Yi, and L. S. Davis, "Collaborative fashion recommendation: A functional tensor factorization approach," in *ACM MM Multimedia Conference*, 2015, pp. 129–138.

[18] 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 *IEEE ICCV International Conference on Computer Vision*, 2015, pp. 3343–3351.

[19] K. Lin, H. Yang, K. Liu, J. Hsiao, and C. Chen, "Rapid clothing retrieval via deep learning of binary codes and hierarchical search," in *ACM ICMR Intl Conference on Multimedia Retrieval*, 2015, pp. 499–502.

[20] K. Yamaguchi, M. H. Kiapour, and T. L. Berg, "Paper doll parsing: Retrieving similar styles to parse clothing items," in *IEEE ICCV International Conference on Computer Vision*, 2013, pp. 3519–3526.

[21] M. Mizuochi, A. Kanezaki, and T. Harada, "Clothing retrieval based on local similarity with multiple images," in *ACM MM Multimedia Conference*, 2014, pp. 1165–1168.

[22] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun, "Neuroaesthetics in fashion: Modeling the perception of fashionability," in *IEEE CVPR Conference on Computer Vision and Pattern Recognition*, 2015, pp. 869–877.

[23] A. Veit, B. Kovacs, S. Bell, J. McAuley, K. Bala, and S. Belongie, "Learning visual clothing style with heterogeneous dyadic co-occurrences," in *IEEE ICCV International Conference on Computer Vision*, 2015, pp. 4642–4650.

[24] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *WWW International Conference on World Wide Web*, 2016, pp. 507–517.

[25] S. Liu, J. Feng, Z. Song, T. Zhang, H. Lu, C. Xu, and S. Yan, "Hi, magic closet, tell me what to wear!" in *ACM MM Multimedia Conference*, 2012, pp. 619–628.

[26] A. Marczewski, A. Veloso, and N. Ziviani, "Learning transferable features for speech emotion recognition," in *ACM Multimedia*, 2017, pp. 529–536.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR International Conference on Learning Representations*, 2015.

[28] G. Hinton, "A practical guide to training restricted boltzmann machines," in *Neural Networks: Tricks of the Trade - Second Edition*. Springer, 2012, pp. 599–619.

[29] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, 2002.

[30] L. V. der Maaten and G. Hinton, "Visualizing high-dimensional data using t-SNE," *J. of Machine Learning Res.*, vol. 9, pp. 2579–2605, 2008.

[31] N. Mantel, "The detection of disease clustering and a generalized regression approach," *Cancer Res.*, vol. 27, no. 2, pp. 209–220, 1967.