RnR: Extraction of Visual Attributes from Large-Scale Fashion Dataset

Sungjae Lee*
Department of Computer Science
Kookmin University
Seoul, South Korea
odobenus@kookmin.ac.kr

Yeonji Lee*
Department of Computer Science
Kookmin University
Seoul, South Korea
nelumbotus@kookmin.ac.kr

Junho Kim
Department of Computer Science
Kookmin University
Seoul, South Korea
junho@kookmin.ac.kr

Kyungyong Lee
Department of Computer Science
Kookmin University
Seoul, South Korea
leeky@kookmin.ac.kr

***Abstract*—Supervised visual perception models require a large number of annotated inputs so that a model can be trained and validated. However, manual annotation of large-scale datasets can be prohibitive in terms of time and expense. We present an automatic fashion image tagging algorithm that uses detailed item descriptions and carefully selected category information produced by a service provider. Compared to previous work, which used a relatively simple frequency-based heuristic, a set of image attributes extracted by our proposed algorithm could be used to create a model (ResNet-18 architecture) with exhibited 8% higher accuracy than a model created with attributes suggested by DeepFashion. This in-progress work identifies opportunities that carefully chosen visually recognizable attributes can be used to help build a more general and descriptive model.**

***Index Terms*—auto tagging, fashion dataset, visual attribute, image annotation**

I. Introduction

Supervised image recognition algorithms rely largely on annotated tagging of data to provide ground truths for image classification. ImageNet [1] is almost a de-facto standard for many image classification algorithms and contains over 14 million annotated images, manually tagged by human experts. Due to the nature of deep learning algorithms, the provision of more annotated images in the training phase is likely to improve the generalizability and accuracy of a model. However, even with the existence of a public market for human labor, such as the Amazon Mechanical Turk, it becomes prohibitive to manually annotate a large number of images with correct tags, especially using experts. The use of publicly-available text information paired with corresponding images can provide an alternative solution for mitigating the workload of manual image annotation when valid visual information can be extracted from a narrative description. Joulin and Maaten et. al. [2] proposed to use the information from users' descriptions of pictures in *Flikr*. DeepFashion [3] is a collection of fashion-related images with annotated attributes, landmark information, and fashion synthesis. The DeepFashion team crawled public online fashion websites (*Forever 21* and *mogujie*) to collect images and corresponding descriptions.

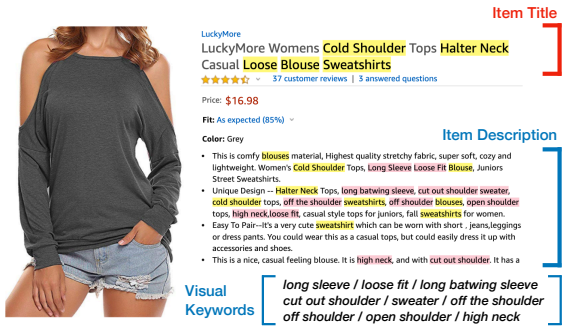

Fig. 1. Annotating visual keywords from image description and visual keyword set which is not included in title text but description text

In the process of extracting valuable information from descriptive texts, prior work [2], [3] relied on a simple mechanism involving counting n-gram keywords over a limited region, such as a title, and marking keywords with high frequency as valid attributes. Figure 1 shows a fashion item example downloaded from Amazon. It contains an image, title, and feature description expressed using bullet points. We highlight a few visually-recognizable features from the image. Although the Title section contains valuable visually annotatable information, it omits more detailed information that exists only in the Item Description section. For example, the title section misses *long sleeves, loose fit, cut out shoulder* attributes that exist only in the Item Description section.

As shown in Figure 1, the Item Description section contains a large amount of visually recognizable information and can be a valuable input into annotated images, helping to improve the quality of supervised models. In this work, we aim to extract visually-recognizable features from a detailed narrative image description, so that they could work as high-quality

*equal contribution

input datasets with which to build robust and accurate supervised visual perception models for fashion-related images. To achieve this goal, we developed a novel algorithm, called *RnR*, to select only the visual attributes from an item description. The algorithm calculates the importance of an n-gram keyword in a description, by comparing the frequency of a keyword for products in the same fashion category (***R**elevant*) with that of products in different fashion categories (***NonR**elevant*). To quantitatively evaluate the quality of attributes extracted by the algorithm, we applied a deep binary classifier using transfer learning. We found that a plain ResNet-18 [4] architecture binary classifier model built with attributes suggested by our proposed algorithm shows 8% higher accuracy than a model built with attributes suggested by DeepFashion [3].

The main contributions of this paper are as follows.

- a novel algorithm to pick visually recognizable keywords from an image and its detailed description
- a new method to quantitatively evaluate the accuracy of selected visual features
- making over 200K image URLs and item descriptions crawled from Amazon fashion publicly available[1]

The rest of this paper is organized in the following way. Section II presents an algorithm to extract visually noticeable information from an item description. Section III evaluates the quality of the proposed approach by using a well known plain deep learning model. Section IV discusses related work. Section V summarizes the findings in the paper and discusses future directions.

II. Image Attribute Auto Tagging

On the Web, there are a huge number of images and corresponding descriptions on sites such as image-tags pairs in *Instagram*, and item image-description pairs in many e-commerce sites, and the number keeps growing. In general, every image-description pair on the Web is generated by humans, and can accurately describe abstracted characteristics of an image. Such information is useful for the generation of annotated image sets that can be used as inputs for training and validating supervised visual perception models. To build high-quality datasets for such purposes, it is important to discriminate whether an n-gram keyword in an image description accurately represents the visual characteristics of an input image. For example, for an image in Figure 1, n-gram keywords such as [*halter neck, long sleeve, and open shoulder*] accurately represent visual features of the given image. However, keywords such as [*highest quality stretch fabric, unique design, easy to pair, and suitable for fall, winter*] are difficult to identify, even using a state-of-the-art image classification model, because these are not really visual features.

[1]We will make the dataset publicly accessible by the presentation date.

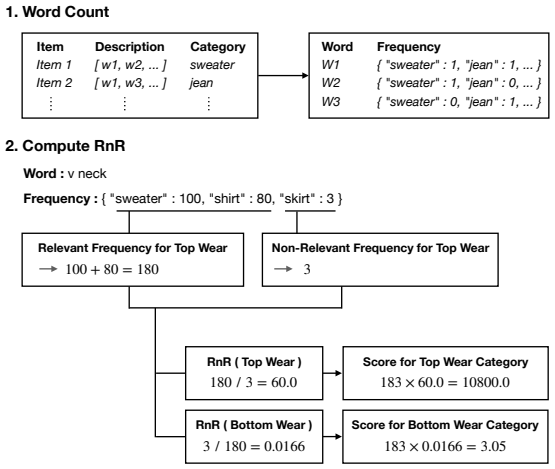

Fig. 2. An example of *RnR* algorithm after categorizing attributes into *Top-wear* and *Bottom-wear*. An attribute *v neck* appears more frequently in the *Top-wear* category and it becomes a visual attribute for the category.

A. Attribute Extraction in DeepFashion

DeepFashion [3] is one of the richest fashion databases. It contains fashion-related images crawled from public e-commerce websites and associated information which is vital to cutting-edge research. Among many benchmarks provided in the database, the attribute prediction benchmark is relevant to the work reported in this paper. In this benchmark, the curators provide approximately 300K annotated images with bounding boxes. To provide visual attributes for each image, the authors used information in the product title. After separating titles into words by spaces, they used the last word as a category. After removing stop-words, they created n-grams, the n being up to three. Then, they counted the number of occurrences of each n-gram word and took the most frequent thousand words as final attributes. This frequency-based attribute selection is simple, and Joulin and Maaten et. al. [2] used a similar method to extract attributes from a data set of Flikr images and descriptions.

However, this simple heuristic has some shortcomings. Referencing words only in the title may miss information in the description section, as shown in Figure 1. Words which occur frequently in the title section are not necessarily visually recognizable attributes, even after removing stop-words. Lastly, the authors did not provide a quantitative analysis as to how well the suggested attributes represent the visual characteristics of the images.

B. RnR*: Category-Aware Automatic Attribute Selection*

To improve user's shopping experience, many e-commerce sites provide a range of information about an item, such as title, detailed description, and product category. The *RnR* (Relevant-NonRelevant) algorithm attempts to address limitations in the prior approaches and aims to extract visual attributes from detailed descriptions of fashion items using carefully pre-configured category information.

The algorithm consists n-gram word counting, followed by weighting of the count values by the number of occurrences in the categories. Initially, we perform n-gram word counting after removing stop words. The n is configurable and an n of three seems to be enough to express visual features of fashion images. Amazon provides *Featured Categories* for each fashion-related item. When counting the number of occurrences of n-gram words, we aggregate the count by the *Featured Categories* to which an item belonged. In the *RnR* step, we allocate to each n-gram a weight calculated from the ratio of the number of occurrences in the *Relevant* and *NonRelevant* category, the *RnR* score. The algorithm then multiplies the number of occurrences by the *RnR* score, and n-gram words with higher scores becomes visual features for each category. Formally speaking, an attribute, i, gets a score of S_{ic}, for a category of c as shown in Equation 1. N_i indicates the number of occurrences for an attribute i across all categories. N_{ic} represents the number of occurrences of attribute i in a specific category c (the *Relevant* score). The $N_i - N_{ic}$ is the number of occurrences of attribute i except the category c (the *NonRelevant* score). The intuition behind the algorithm is that attributes that appear frequently in specific categories will represent characteristics of items in the category. An example of using the *RnR* algorithm is shown in Figure 2

$$S_{ic} = N_i \times \frac{N_{ic}}{N_i - N_{ic}} \tag{1}$$

For example, if we evaluate a keyword of *v neck* with the configured categories of *Top-wear* and *Bottom-wear*, we count the number of occurrences of the *v neck* attribute in the description section of the crawled items grouped by two different categories. Then, we multiply the occurrences of *v neck* in the entire description to the *RnR* score of *v neck* that is calculated by word frequency in *Top-wear* categories and *Bottom-wear* categories. From the calculated *RnR* score, we can find out that *v neck* attribute is *Top-wear* related word, because the score of *Top-wear* should be much larger than the other one due to the number of appearances in each category. Finally, we can designate the *v neck* as an attribute of *Top-wear* category, and if the total score of the word is higher than others, it could be one of meaningful visual attribute in fashion dataset.

III. Evaluations

After extracting visual features from an image description, it is important to evaluate how well the chosen attributes represent the visual characteristics of original images. To focus on the evaluation of the extracted attributes validation, we used an evaluator that uses a CNN-based binary classifier, as shown in Figure 3. With a set of attributes extracted using different algorithms, we prepare training datasets that had a binary output for each attribute in the ratio of 1 : 1. The *true* datasets were randomly selected from images that contained a target attribute, while the *false* datasets were randomly selected from images without the target attribute.

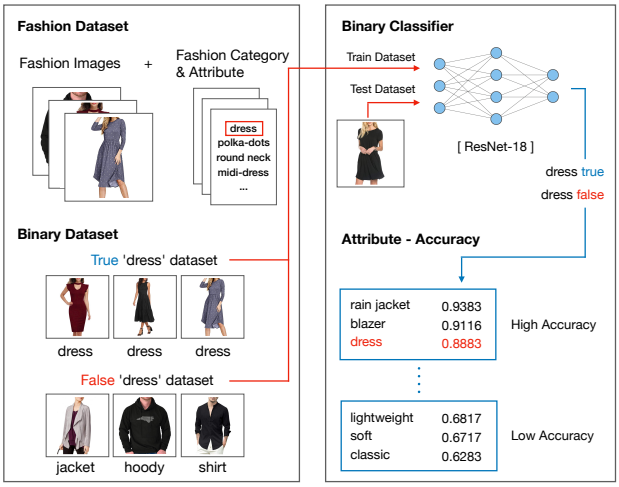

Fig. 3. Extracted image attribute evaluator sequence using a CNN-based binary classifier

The input images had various dimensions, and we padded images to make them square. Then, we transformed images to 224×224 and performed a horizontal flip with a probability of 0.5. Using the transformed dataset, we applied a transfer learning method using a pre-built ResNet-18 model [4] with a learning rate of 0.001 for 25 epochs. Using the trained CNN model, we measured the prediction accuracy on a validation dataset built in the same manner as the training dataset. The evaluation method focuses on the quality of the extracted attributes, without any tuning of the underlying model. The intuition behind the evaluation method is that with a vanilla model the accuracy will largely depend on the quality of the training and validation datasets. Thus, if an attribute is visually recognizable, the trait should be identified by the underlying model and will result in high accuracy. Otherwise, the accuracy of a model built by the training datasets will be low because the model cannot capture visual clues from the training dataset.

Figure 4 shows the binary classification accuracy of models built using three different training datasets. The 20 left-most plain blue bars show the top-20 accuracy of attributes extracted using our *RnR* algorithm, using the Amazon fashion dataset (*AmazonFashion RnR*). The circle-marked 20 bars in the middle represent the top-20 accuracy of attributes extracted from the Amazon fashion dataset using only the title section (*AmazonFashion Title*). We use only the number of occurrences of each attribute in the title section. The right-most 20 bars with right upward diagonal bars show the top-20 attribute accuracy of DeepFashion (*DeepFashion Title*). We use the attributes and the dataset provided by the DeepFashion [3] database. When choosing visual attributes, the *AmazonFashion Title* and *DeepFashion Title* use the same mechanism, but different datasets. The features selected by the three different algorithms represent the visual attributes quite well. Amongst the three algorithms, we could also observe that the proposed *RnR* algorithm contains unique attributes compared to other methods. We suspect that this is because the *RnR* algorithm

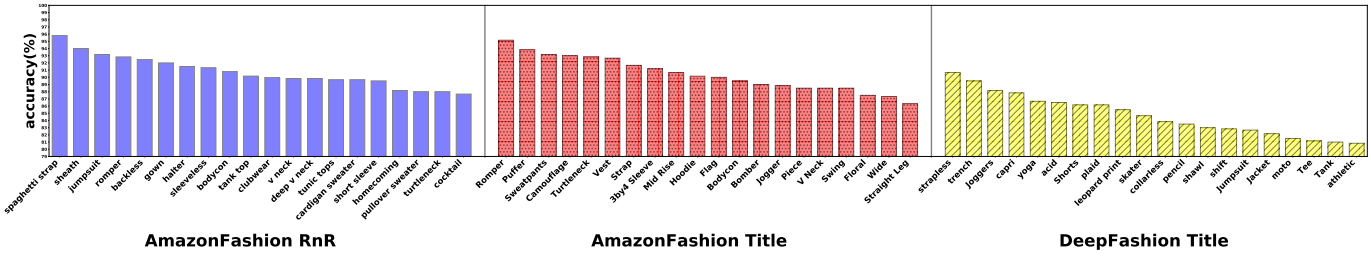

Fig. 4. Top-20 accuracy attributes of different attribute selection algorithms

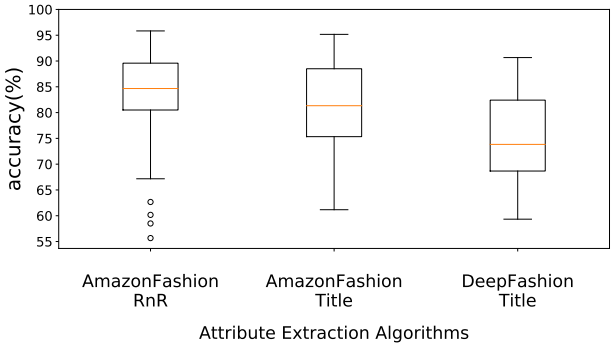

Fig. 5. The box-whisker plot of binary image classifier accuracy built with different attributes

references the item description section, in which more detailed information can appear.

Figure 5 shows the binary classification accuracy of models built with different attributes. Accuracy is expressed using box-and-whisker plots to reflect the accuracy of models trained using different attributes and different algorithms. Overall, the ResNet-18 model trained with the attributes suggested by our *RnR* algorithm showed the highest accuracy. The average accuracy was 8% higher than that of using DeepFashion attributes. The evaluation results shown in Figures 4 and 5 demonstrate that careful selection of attributes from a narrative item description enriches the attributes of the image items by uncovering visually recognizable characters in a verbose item description. We are confident that the enriched set of attributes and images can provide a foundation on which to build a supervised model achieving high accuracy.

IV. Related Work

In the field of computer vision, large scale of fine grained image-tag pair dataset is very important to build an accurate image classification model. To reduce the labor of image labeling, there were various attempts in algorithm and dataset for auto-tagging. For example, DeepFashion [3] used simple and powerful heuristic algorithm for extract attributes from e-commerce fashion dataset. Extracting attributes from title section was meaningful method when we start to build new kinds of fashion dataset, because it provides a method to make fashion image-attribute pair baseline from raw dataset. Gutierrez et. al. [5] applied Convolutional Neural Networks (CNN) to fashion-related images from Amazon.com to extract category and attribute information for classification.

Active learning is another method of building annotated datasets from unlabeled large-scale image dataset. *Active learning* [6] shows that reduction of human effort in image-category labeling, and it could be extended to image-attribute labeling. Similar to active learning, weakly supervised learning [7] provides a mechanism to generate annotated dataset from unlabeled dataset using a very small number of already annotated dataset. Those auto-tagging approaches decreases humans' involvement in the process of dataset annotation, but they still require human labor to make manual decision for items that are close to the decision boundary or building very accurate labeled dataset for starting point. Different from the previous work, the proposed work in this paper can perform image annotation tasks by using already existing detailed information of item description and category information without humans' intervension.

V. Discussion for Limitations and Future Work

In this paper, we describe an *RnR* algorithm that extracts visually recognizable attributes from an item description of a fashion-related image from e-commerce websites. The novelty of this algorithm lies in referencing the frequencies of items across different categories, a strategy which helps to detect features relevant to a target category. The preliminary evaluation results shed light on the possibility of automatic image annotation from publicly-available item descriptions. However, this in-progress work still has considerable rooms for improvement and there is a need to justify the validity of the visual features which have been extracted.

Building upon this work, we foresee many opportunities for improvements. 1) In the current version, we focused primarily on the fashion categories of *Top-wear* and *Bottom-wear*, so multiple other categories could be added to improve the robustness and generality of the attributes. 2) Applying more advanced natural language processing techniques, such as Part-of-Speech (POS) tagging [8], to the visual attributes selection from a detailed description. 3) Suggested visual attributes from the *RnR* algorithm shows higher accuracy than other methods. However, few non-visually recognizable attributes selected by the algorithm show very low accuracy in Figure 5. The attributes with lower accuracy include [*ro, ky, aus, oh*], and the algorithm should be improved to exclude such attributes without losing generality. 4) The evaluation method could

detect visually recognizable attributes. We can use the binary classification accuracy as an input when choosing appropriate visual attributes. 5) To use the binary classifier output for attribute suggestion, significant compute capacity is necessary, because there are many candidate attributes that could be used for training. For such purposes, cloud computing could be used. Building a work process pipeline in a cloud environment would make the proposed system generally applicable.

REFERENCES

[1] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[2] Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 67–84, Cham, 2016. Springer International Publishing.

[3] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[5] Patricia Gutierrez, Pierre-Antoine Sondag, Petar Butkovic, Mauro Lacy, Jordi Berges, Felipe Bertrand, and Arne Knudson. Deep learning for automated tagging of fashion images. In Laura Leal-Taixé and Stefan Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 3–11, Cham, 2019. Springer International Publishing.

[6] Brendan Collins, Jia Deng, Kai Li, and Li Fei-Fei. Towards scalable dataset construction: An active learning approach. In David Forsyth, Philip Torr, and Andrew Zisserman, editors, *Computer Vision – ECCV 2008*, pages 86–98, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[7] Paroma Varma and Christopher Ré. Snuba: Automating weak supervision to label training data. *Proc. VLDB Endow.*, 12(3):223–236, November 2018.

[8] Lluís Màrquez and Horacio Rodríguez. Part-of-speech tagging using decision trees. In Claire Nédellec and Céline Rouveirol, editors, *Machine Learning: ECML-98*, pages 25–36, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg.