FashionCapsNet: Clothing Classification with Capsule Networks

Araştırma Makalesi/Research Article

Furkan KINLI, Furkan KIRAÇ

Department of Computer Science, Özyeğin University, İstanbul, Turkey
furkan.kinli@ozyegin.edu.tr, furkan.kirac@ozyegin.edu.tr
(Geliş/Received:20.06.2019; Kabul/Accepted:26.01.2020)
DOI: 10.17671/gazibtd.580222

***Abstract*—** Convolutional Neural Networks (CNNs) are one of the most commonly used architectures for image-related deep learning studies. Despite its popularity, CNNs have some intrinsic limitations such as losing some of the spatial information and not being robust to affine transformations due to pooling operations. On the other hand, Capsule Networks are composed of groups of neurons, and with the help of its novel routing algorithms, they have the capability for learning high dimensional pose configuration of the objects as well. In this study, we investigate the performance of brand-new Capsule Networks using dynamic routing algorithm on the clothing classification task. To achieve this, we propose 4-layer stacked-convolutional Capsule Network architecture (FashionCapsNet), and train this model on DeepFashion dataset that contains 290k clothing images over 46 different categories. Thereafter, we compare the category classification results of our proposed design and the other state-of-the-art CNN-based methods trained on DeepFashion dataset. As a result of the experimental study, FashionCapsNet achieves 83.81% top-3 accuracy, and 89.83% top-5 accuracy on the clothing classification. Based upon these figures, FashionCapsNet clearly outperforms the earlier methods that neglect pose configuration, and has comparable performance to the baseline study that utilizes an additional landmark information to recover pose configuration. Finally, in the future, proposed FashionCapsNet may inherit extra performance boost on the clothing classification due to advances in the relatively new Capsule Network research.

***Keywords*—** deep learning, capsule networks, clothing classification, fashion analysis

FashionCapsNet: Kapsül Ağları ile Kıyafet Sınıflandırma

***Özet*—** Konvolüsyonel Sinir Ağları (KSA) görsel ilişkili derin öğrenme çalışmalarında en sık kullanılan mimarilerden biridir. Popülaritesine rağmen, KSA'lar ortaklama işlemi yüzünden konumsal bilgi kaybı ve afin dönüşümlerine dayanıklı olmama gibi bazı yerleşik sınırlamalara sahiptir. Öte yandan, gruplanmış nöronlardan oluşan Kapsül Ağları, özgün yönlendirme algoritmalarının yardımıyla, nesnenin yüksek boyutlu poz konfigürasyonunu da öğrenme kapasitesine sahiptir. Bu çalışmada dinamik yönlendirme algoritmasını kullanan Kapsül Ağları'nın kıyafet sınıflandırma performansını inceledik. Bu amaçla, arka arkaya yerleştirilmiş 4 Konvolüsyonel katmanlı bir Kapsül Ağ mimarisi (FashionCapsNet) önerdik, ve bu modeli 46 kategoriye ayrılmış 290 bin kıyafet resmi içeren DeepFashion adlı veri seti ile eğittik. Akabinde, modelimizin ve DeepFashion veri seti ile eğitilmiş CNN tabanlı en gelişmiş metotların kategori sınıflandırma sonuçlarını karşılaştırdık. Çalışmamızın sonucunda, FashionCapsNet, kıyafet sınıflandırma için %83,81'lik en yüksek-3 başarım oranı ve %89,83'lük en yüksek-5 başarım oranı sonuçlarını elde etmiştir. Bu rakamlara dayanarak, FashionCapsNet, poz konfigürasyonunu ihmal eden eski metotları açık bir şekilde geride bırakmıştır, ve poz konfigürasyonunu belirgin nokta bilgisinden faydalanarak telafi eden referans çalışmasıyla benzer bir performans göstermiştir. Son olarak, görece yeni olan Kapsül Ağları üzerine yapılacak araştırmalardaki gelişmeler sayesinde, önerdiğimiz bu modelin (FashionCapsNet) kıyafet sınıflandırma performansında ekstra bir artış gözlemlenebilir.

***Anahtar Kelimeler*—** derin öğrenme, kapsül ağları, kıyafet sınıflandırma, moda analizi

1. INTRODUCTION

Fashion is one of the most prominent industries in the world. In recent years, with the advent of the cutting-edge technologies in computer science and the emergence of e-commerce, clothing recommendation systems have drawn the attention of large fashion companies. Although these systems have great potential for most companies, it is essential to be come up with certain solutions for the problems such as clothing classification [1-3], attribute prediction [3-5], clothing segmentation [5], and style prediction [6,7]. The common solution approach in Computer Vision was to use feature descriptors like Histogram of Gradients [8] (HOG) or Scale-Invariant Feature Transform [9] (SIFT) to extract the features from the images and to deploy these features on traditional Machine Learning algorithms. Thereafter, the achievement of AlexNet [10] on ImageNet Challenge (ILSVRC) [11] and the prime motivating force of using deeper architectures in neural networks due to the improvements of GPU technologies lead up to significant accomplishments of fashion recommendation systems using deep learning methods [12-14].

Broadly speaking, Convolutional Neural Networks (CNNs) work well in most Computer Vision tasks. A CNN is basically composed of an input, an output and some hidden layers (e.g. convolutional, pooling, dense and normalization layers). Furthermore, CNNs are capable of learning some different representations directly from sample images by obtaining regional spatial information with local receptive fields [15]. Although CNNs have outstanding performance on image-related deep learning tasks, in real life, there are some intrinsic limitations of this architecture. First, pooling layer used for down-sampling the output of the previous layer ignores the spatial relationship between some parts of the image. For that matter, CNNs cannot gather the hierarchical information between important pieces that identify the object. Likewise, CNNs are not robust to affine transformations. Due to pooling operations, this architecture cannot employ pose information for recognizing the objects. An image with pose configuration that is not encountered during training could be misclassified by CNNs on testing phase. Therefore, the training data needs to include different kinds of transformations of the sample images to get better performances in CNN-based architectures.

Recently, an alternative Deep Learning approach called Capsule Networks (CapsNets), with a novel routing algorithm between capsules, has been proposed by Sabour and Hinton et al. [16]. In this design, it is supposed to learn the information about the object and the intrinsic spatial relationship between the parts of the object by harnessing the routing-by-agreement algorithm. Thus, CapsNets are able to recognize the objects regardless of the viewing angle and without needing different transformations of them during training.

In recent studies on clothing recognition, CNN-based methods neglecting pose configuration achieve arguably good performances [4,18,20]. However, it hinders further improvements on the recognition accuracy. On the contrary, most state-of-the-art CNN-based methods figure out to recover pose configuration of the objects by employing hand-crafted landmark information [3] or bag-of-words descriptors [21], or by adding attention mechanisms for the guidance of domain knowledge [22,23]. In this study, we indicate that Capsule Networks can achieve similar, even better performance on the clothing category classification without using any side information or extra module such as attention maps. To achieve this, we propose a novel Capsule Network architecture, which extracts the features from the images by a number of stacked-convolutional layers, and forwards these features to the fully-connected capsule layers. Our proposed design is trained on DeepFashion dataset [3] that contains 290k images with 46 fine-grained category labels.

The rest of the paper is organized as follows: Section 2 surveys the previous studies on the clothing category classification. Section 3 describes required mathematical background for CNNs and Capsule Networks. Section 4 gives detailed information about DeepFashion dataset. Proposed Capsule Network architecture and our methodology are introduced in Section 5. While Section 6 discusses the experimental results of our methodology, and finally, Section 7 concludes the paper.

2. RELATED WORKS

Clothing recognition is one of the starting points of visual fashion analysis. In earlier studies, fashion models are mostly relied on hand-crafted feature descriptors [8-9], and it is attempted to generate a powerful clothing representation with these features. In recent years, with the emergence of deep learning techniques, neural networks can have better performances on clothing recognition tasks as in the case of several different domains (e.g. fine-grained object recognition, face recognition) [10,24,27,30]. To survey the recent studies on clothing recognition, Kiapour *et al.* [4] proposes introduces an excessively challenging task, namely *Exact Street to Shop*, where the goal is to match street photos that are captured in uncontrolled settings to the same item in online shop photos that are captured by professionals. In this study, it is experimentally shown that learning the category-specific similarity between street and online shop photos has better performance on both category classification and image retrieval tasks than applying traditional machine learning techniques to the hand-crafted features.

Huang *et al.* [18] shows that incorporation of semantic attributes and visual similarity between cross-scenario images increases the expressive power of extracted features. To demonstrate this argument, Dual Attribute-aware Ranking Network (DARN) is introduced in this study. This network consists of two Network-in-Network (NiN) mechanisms [19] to learn both category and semantic attributes of the images in cross-domain scenario.

Thereafter, the outputs of two sub-networks are concatenated, and fed into triplet ranking loss to learn the visual similarity.

Liu *et al.* [3] introduces a large-scale clothing dataset, called DeepFashion, with comprehensive annotations. In this dataset, there are nearly 800k images with numerous attributes, landmarks and cross-domain image pairs. To show the usefulness of this dataset, Liu *et al.* proposes FashionNet [3] architecture that iteratively estimates the landmark information in the intermediate step to gate the learned features for the prediction of the category and attributes. Moreover, for an ablation study, FashionNet is represented by different building blocks where the model has different numbers of attributes (i.e. 100, 500 and 1000), or fashion landmarks are replaced with human joints or poselets to gate the features.

Lu *et al.* [20] proposes an automated learning framework that contains a multi-task deep network using a novel dynamic branching procedure. In this framework, the network model is initialized as a thin model, and is expanded with dynamic branching that is mainly responsible for grouping the shared features in each layer by considering the task and the complexity of the model.

Corbière *et al.* [21] demonstrates that it is possible to achieve promising results on the clothing category classification and image retrieval by integrating bag-of-words approach to weakly-supervised learning process. The proposed model encodes weakly-annotated noisy data using the bag-of-words descriptors to generate separable visual concepts, and to provide meaningful similarity between images. Along with that, the learned representations are suitable for both classification and retrieval tasks, and this study essentially addresses the issue of finding a large, rich-annotated and clean-labeled dataset for training of deep neural networks.

Wang *et al.* [22] introduces a novel knowledge-guided deep network that captures the kinematic and symmetric relations between the clothing landmarks to address the landmark localization and category classification problems. In addition, two attention modules (i.e. landmark-aware and category-driven) are employed for boosting the category classification performance. With the help of the attention modules, the network is enforced to focus on the functional parts of the clothes, and also to reinforce the expressive power of extracted features on classifying the clothing category.

Analogous with the approach in [22], Liu *et al.* [23] attacks to the landmark localization and category classification problems with the guidance of a single attention map. In this study, transpose-convolutional up-sampling is used for generating the feature maps in order to detect the landmarks in low-resolution images more accurately. Instead of using two isolated attention mechanisms as in the case of [22], Liu *et al.* employs a single attention mechanism generated by landmark heatmaps to the network, and it is reported that this approach improves the precision of fashion category classification and attribute prediction.

3. BACKGROUND INFORMATION

The main objective of this study is to classify the clothing images by category with Capsule Networks, brand-new deep learning architecture. Next, we compare the performances of Capsule Networks, our baseline study [3] and the other state-of-the-art methods using CNNs (with heuristic down-sampling layers) on DeepFashion [3] dataset. Before revealing the details of our methodology, we concisely review the basic structure of CNNs frequently used in many Computer Vision tasks, and the structure of Capsule Networks. At the end of this section, we investigate which limitations of CNNs are overcome by Capsule Networks, and how to achieve it.

3.1. Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) generally contain a certain number of convolutional layers combined with a non-linear activation function and down-sampling layers (e.g. max-pooling). The convolution is performed by sliding the kernel over the input data to form the receptive fields, as shown in Figure 1. Due to the consecutive convolutional layers in this architecture, low-level features such as corner, texture and edge are extracted from the input data by sharing the weights. Then, these features are combined in deeper layers to compose higher level features.

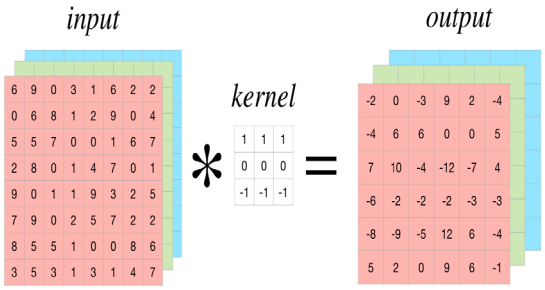


Figure 1. Performing convolution operation by sliding the kernel over the input data to form the receptive fields.

CNNs are applicable for a wide range of Computer Vision tasks such as image classification [10,24-26], object detection [27-29], object localization [24-26], synthetic image generation [30-32]. This popularity stems from the fact that they can automatically learn by deriving the identical properties of the data without needing any prior knowledge about the domain. There are also a number of studies emphasizing the importance of parameter changes in success rate of CNNs [39-41]. However, there are some intrinsic limitations on CNN architecture. First, CNNs classify an image by joining some components of the object in the image regardless of the spatial relationship between them. The main reason of this problem is that pooling operations are kinds of rudimentary routing methods, where the neurons are picked by a heuristic (e.g.

maximum, average or minimum) without considering the task. Therefore, CNNs can easily confuse an object with the fake one that contains some components of the object with improper alignment. Secondly, CNNs are not robust to affine transformations since the output of pooling operations completely throws away pose information that is important for correct recognition. In other words, CNNs are not able to capture different pose information of the images if this information is not seen on the training set. To address it, this design needs to be trained with several different transformations of the images to generalize the performance. However, for real world applications, it is impractical to have all possible transformations of the images in the training set. Based upon these reasons, a new deep learning architecture called *Capsule Networks* is proposed by Sabour and Hinton *et al.* [16]. In fact, the idea behind this approach [33] goes back decade ago, but it recently starts to work well after inventing dynamic routing algorithm.

3.2. Capsule Networks (CapsNets)

Basically, a capsule could be considered as a group of neurons who together pack a high dimensional information. This information refers to the existence of the entity and pose configuration describing the underlying behavior of the entity in a more refined way. The activation vector within an active capsule represents several features of a specific entity such as position, size, orientation, deformation and texture, while the overall length of the vector states the probability of the existence of that specific entity. Capsule output in a layer is routed to the capsules in the next layer by multiplying it with the weight matrix (as coupling coefficient). The magnitude of the coupling coefficient represents the strength of a parent capsule to be routed. In other words, this algorithm is kind of a top-down feedback mechanism where the predictions in lower levels determine which capsule in the higher level is activated. This is called "routing-by-agreement" [16]. This algorithm is a far more powerful routing algorithm than pooling variants that pick the neurons by a heuristic. The overall architecture can be seen in Figure 2.

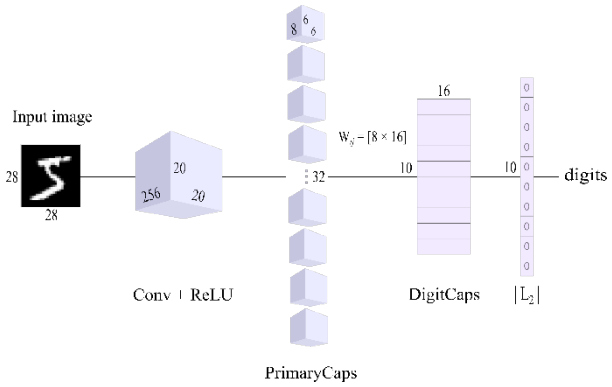


Figure 2. Capsule Network architecture proposed by Sabour and Hinton *et al.* [16].

Considering u_i as the output of capsule i, and W_{ij} as the weight matrix

$$\hat{u}_{j|i} = W_{ij} u_i \tag{1}$$

where $\hat{u}_{j|i}$ is the vector that predicts the output of the parent capsule j by capsule i. The relationship between capsules in the previous layer and the possible parent capsule is encoded to a coefficient c_{ij} as "routing soft-max" whose initial logits b_{ij} are the log prior probabilities of routing i^{th} capsule in the previous layer to j^{th} capsule in the next layer. The logits of all capsules in each layer are initialized to 0 at the beginning of the routing-by-agreement algorithm.

$$c_{ij} = \frac{e^{b_{ij}}}{\sum e^{b_{ij}}} \tag{2}$$

The input for the parent capsule j is calculated as weighted sum over all prediction vectors from the capsules in the previous layer.

$$s_j = \sum_i c_{ij} \hat{u}_{j|i} \tag{3}$$

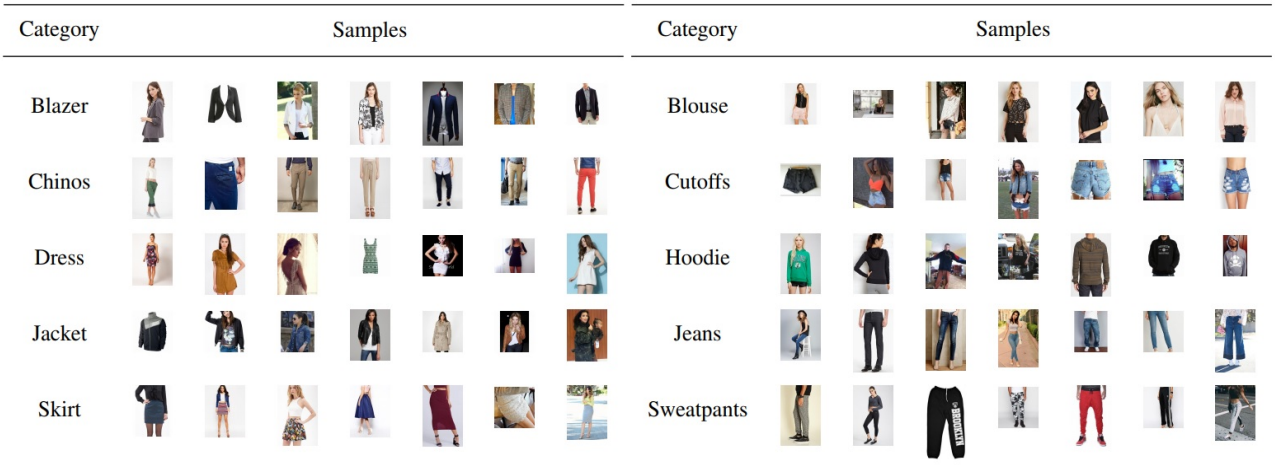


Figure 3. Example images from DeepFashion [3] dataset.

A non-linear function called *squashing* is applied to the input for the parent capsule j to ensure that the values in this vector are compressed in a range between zero and slightly below one. Note that epsilon value (10^{-7}) is added to the denominator of unit scaling of the input vector since we observed that the gradients vanish at the early stage of our experiments. The final version of squashing formula is calculated as follows.

$$v_j = \frac{\|s_j\|^2}{1+\|s_j\|^2} \frac{s_j}{\|s_j\|+\varepsilon} \quad (4)$$

Therefore, the magnitude of the inner product of v_j and $\hat{u}_{j|i}$ decides which capsule in the next layer is likely to route (agreement).

$$a_{ij} = v_j \cdot \hat{u}_{j|i} \quad (5)$$

For Capsule Networks, the loss is the sum of the losses of all category capsules that are calculated as separate margin loss L_k, for each category capsule k.

$$L_k = T_k max(0, m^+ - \|v_k\|)^2 + \lambda(1 - T_k) max(0, \|v_k\| - m^-)^2 \quad (6)$$

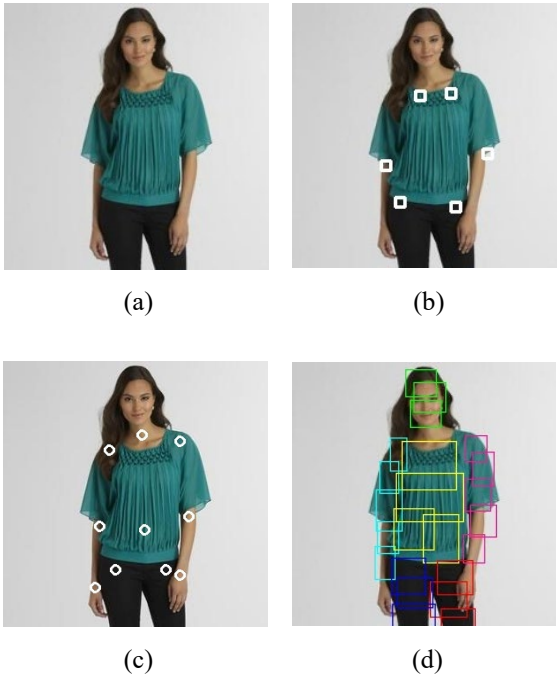
where T_k represents the existence of the instantiation in category capsule k; and m^+, m^- and λ hyper-parameters that control the loss value by the existence, and set to 0.9, 0.1 and 0.5 respectively as proposed in [16].

4. DATASET

DeepFashion [3] is a dataset of 800K high/mid-resolution images that belong to 50 fine-grained categories. The images in this dataset were collected by Multimedia Laboratory of The Chinese University of Hong Kong, from two representative online shopping Web pages and user-generated contents on blogs and forums by querying from Google Images. Sample images of DeepFashion dataset with category labels can be seen in Figure 3.

Table 1. Forming five main attribute groups, and the examples of the attributes in each group.

Groups	Attributes
Texture	*Floral, Stripe, Paisley, Distressed, Dot, Plaid, Panel, Raglan …*
Fabric	*Lace, Denim, Chiffon, Pleated, Woven, Leather, Cotton, Linen …*
Shape	*Crop, Maxi, Fit, Longline, Boxy, Mini, Skinny, Midi, Pencil, Sheath …*
Part	*Sleeveless, Pocket, V-Neck, Hooded, Racerback, Peplum, Strappy …*
Style	*Graphic, Muscle, Tribal, Peasant, Surplice, Polka, Retro, Yoga…*



(a) (b) (c) (d)

Figure 4. Example images from DeepFashion [3] dataset with (a) and without (b) landmarks employed; with (c) human joints and with (d) poselets, a part of pose.

DeepFashion is an extensively annotated clothing dataset that contains numerous attributes, localization parameters and correspondence of images shot under different scenarios ranging from well-posed online shopping photos to unstructured customer photos. For the clothing category classification task, there are 210k training images, 40k validation images and 40k test images with 46 different categories in this dataset. Moreover, the attributes form five groups: texture, fabric, shape, part, and style, and an image can have +8 landmark locations. The attributes and their group information can be seen in Table 1, while different extra information to the visual features in the dataset that can be included to the training is demonstrated in Figure 4. We specifically did not include these hand-crafted landmark or attribute information in the dataset to our training process since our proposed architecture, namely *FashionCapsNet*, has the capacity for learning pose information by itself.

5. EXPERIMENTAL STUDY

In this study, our main objective is to observe the category classification performance of Capsule Networks on realistic, diverse and large clothing images. To achieve this, we propose 4-layer stacked-convolutional Capsule Network architecture (FashionCapsNet) with dynamic routing algorithm. Next, we train our proposed model on 210k training images in DeepFashion [3] dataset. Finally, we examine the best top-3 and top-5 accuracy of this model, namely FashionCapsNet, on the category classification, and compare the results with both the baseline study (FashionNet) [3] and the other state-of-the-art methods trained on DeepFashion dataset.

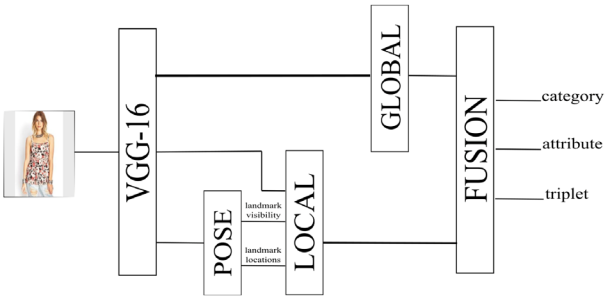


Figure 5. FashionNet [3] model architecture.

5.1. Baseline: FashionNet

We picked FashionNet proposed in Liu *et al.* [3] as the representative of CNN-based architectures. The limitations of traditional CNN architectures are addressed in FashionNet [3] by taking advantage of hand-crafted landmark information.

FashionNet is built on VGG-16 model [34] by ramifying the last layer into 3 different branches. The first branch in the intermediate layer, named as *pose branch*, is responsible for learning the location value and the visibility of the key-points on the structure of clothes from the images. Separately, *local branch* is the second branch that captures the local features by passing over the output of pose branch. The last branch in this design directly learns the global features from images. At the end, as shown in Figure 5, these three branches are concatenated into a single output layer in order to predict the clothing category and attributes, as well as to learn the pair-wise relationships between clothing images. In addition, there are several different variants of FashionNet utilizing different number of attributes (100, 500 or 1000) to create comprehensive profiles of different clothing variations. Liu *et al.* [3], first, pre-train FashionNet by using a large subset of DeepFashion as training and validation data. Thereafter, another (smaller) subset of DeepFashion is used for fine-tuning the pre-trained FashionNet model, where any item in the smaller subset overlaps with the larger one.

FashionNet is optimized by weighted sum of four different loss functions with iterative training strategy. In the first step of this strategy, the location and visibility information for landmarks is estimated with the help of local and global features of the images. Then, the clothing category is predicted by utilizing the estimated landmark locations to gate the local features. At this point, it clearly shows that FashionNet tries to cope with the lack of pose information in CNN-based models by supporting the model with the extra information extracted from the hand-crafted landmark annotations.

5.2. Our Proposed Architecture: FashionCapsNet

In this study, we propose FashionCapsNet, 4-layer stacked-convolutional Capsule Network design for clothing category classification. This design inherently has the capability of preserving the pose configuration and being robust to affine transformations without any extra information, while it learns low/mid-level features by deriving the identical characteristics of the objects without needing any prior knowledge about the domain.

FashionCapsNet is a hybrid architecture for clothing category classification which provides feasible training process for Capsule Networks on large-sized clothing images. To achieve this, instead of directly resizing the images before fitting to the model, several consecutive convolution operations with stride has been incorporated. Before primary capsule, a number of convolutional layers with different number of filters are stacked without any pooling operation between layers, as distinct from the default methodology proposed by Sabour and Hinton *et al.* [16] (i.e. 1 convolutional layer with 64 filters). This mechanism helps us to extract the features of the objects in the images, and these features are sent to the primary capsule layer as the input. At the same time, adding more convolutional layers before capsules reduces the number of trainable parameters of fully-connected capsule layers, so that it can provide feasible training of such a large dataset within limited computational resources.

We used leaky rectified linear unit (Leaky ReLU) [35] as activation function that allows for a small, non-zero gradient (when the unit is saturated and not active) and batch normalization [36] between stacked-convolutional layers for regularization purposes. As for the rest of our

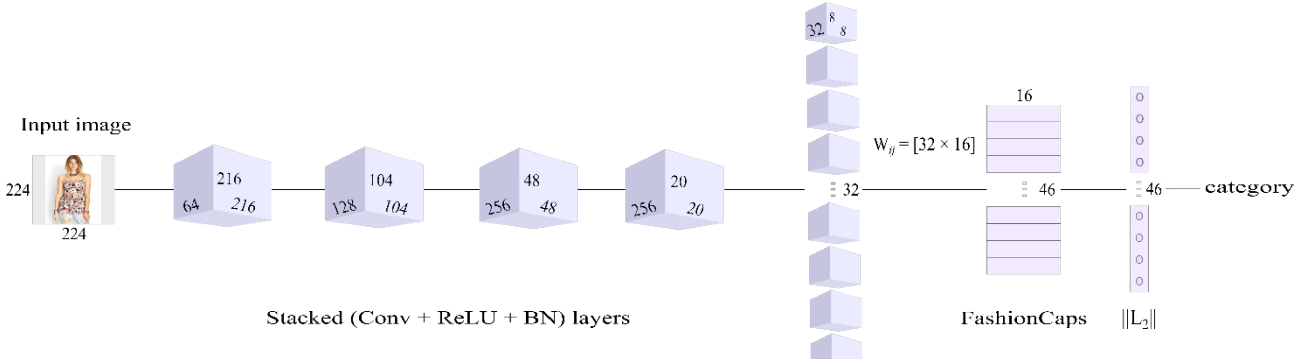


Figure 6. FashionCapsNet (our) model architecture.

proposed model, the primary capsule layer has 32 channels of 8-dimensional fully-connected capsules after stacked-convolutional layers. Capsules are activated by slightly different version (see Equation 4.) of squashing (activation) function proposed in Sabour and Hinton *et al.* [16]. At this point, dynamic routing mechanism is iterated 3 times to route the activation to 16-dimensional capsules in the final layer, namely *FashionCaps*. The length of the activation of each capsule in FashionCaps represents the presence of an instance for each category.

Table 2. Hyper-parameter settings of FashionCapsNet.

Hyper-parameters	
Optimizer	*Adam [37]*
Learning Rate	*0.001*
Decay Rate	*0.00005*
Batch Size	*32*
Routings	*3*
Reconstruction Weight (λ)	*0.0001*
Normalization	*Pixel-wise*

Furthermore, we masked the activity vector of the correct capsule, and used it for reconstructing the images. To achieve this, the output of FashionCapsNet is fed into a decoder network that contains 4 transpose-convolutional layers followed by Leaky ReLU activation function and batch normalization. The mean-squared difference between the original images and the reconstructed ones is added to the loss function in order to create a regularization effect on double margin loss [16]. Final loss is calculated as follows:

$$L = L_k + \frac{\lambda}{N}\sum_i^N (x_i - r_i) \qquad (7)$$

where x represents a set of the original images, r the reconstructed images, and λ is a coefficient that scales down the reconstruction loss, so that it does not dominate the margin loss.

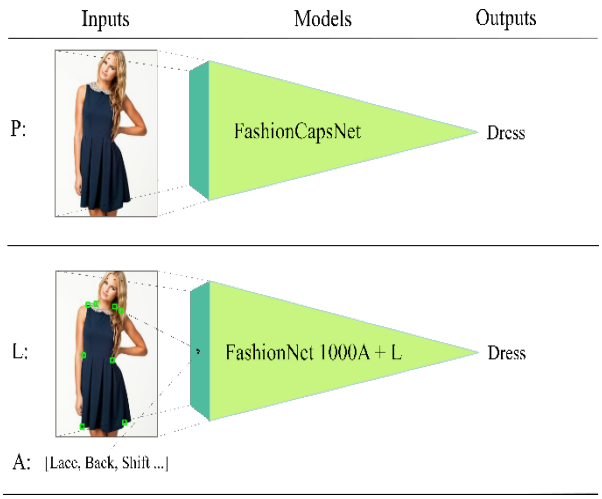


Figure 7. Our proposed architecture sees only images during training, while the baseline model [3] utilizes several attributes and landmark information besides to images in order to train the model.

During training, we picked Adam [37] to optimize the loss function with learning rate 10^{-3}, and decay rate 5×10^{-4}. Image batch for each gradient step contains 32 different samples, and dynamic routing algorithm is iterated 3 times.

The reconstruction coefficient λ is set to 10^{-4}. Pixel-wise normalization is applied to all samples in the dataset. All hyper-parameter settings are introduced in Table 2, and for the sake of clarity, the final version of our proposed architecture, FashionCapsNet is shown in Figure 6.

6. RESULTS & DISCUSSION

In this study, we mainly investigate the performance of Capsule Networks [16] on clothing classification. Therefore, we propose a hybrid Capsule Network architecture called FashionCapsNet, and train our proposed model on 210k training samples of DeepFashion [3] dataset with hyper-parameters presented in Table 2. This task is fine-grained category classification (46 classes), hence the performance is measured by top-K accuracy metric, where K equals to 3 or 5. We test our proposed model with 40k testing samples of DeepFashion dataset.

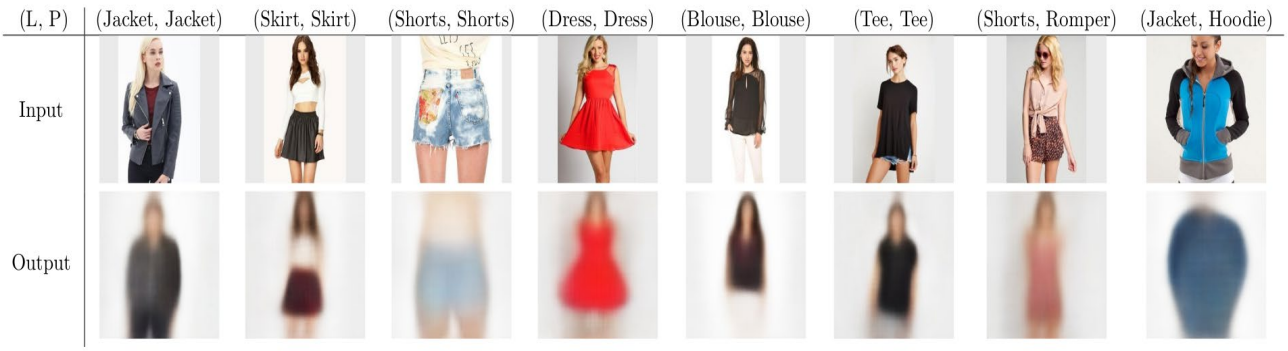


Figure 8. Examples for DeepFashion test reconstructions. (L): Labels, (P): Predictions

Although CNNs have a proven track record of accomplishments in Computer Vision tasks, as a matter of fact, CNNs neglect pose configuration of the object and are not robust to affine transformations. In the literature, several different approaches (see in Section 2) have been enforced to CNNs in order to mitigate the negative effects of these problems so far. However, Capsule Networks inherently learn pose information without needing any side information and extra modules. In this study, we compare the results of our proposed Capsule Network model with the results of the other state-of-the-art CNN-based models that employs any extra side information or module in the clothing category classification task. The overall idea is represented in Figure 7. Table 3 summarizes clothing category classification results of FashionCapsNet and FashionNet variants.

Moreover, in our design, we include the reconstruction loss that measure the difference between the original and reconstructed images in our final loss. The main reason behind this is to regularize the margin loss, and to prevent over-fitting. As illustrated in Figure 8, the reconstructions of the correct capsule consist of the structurally identical features, independent from their pose configuration. At this point, the most dominant pose configuration among the images (i.e. upper body/middle shot) occurs in most of the reconstructed images regardless of their categories.

In the course of the first experimental comparison, we discuss the results of our proposed Capsule Network architecture, and our baseline study FashionNet [3]. We report that our proposed FashionCapsNet achieves 83.81% top-3 accuracy and 89.83% top-5 accuracy on clothing category classification. As shown in Table 3, these figures demonstrate that FashionCapsNet outperforms most of FashionNet variants including building blocks that employ a smaller number of attributes, and use human joints or poselets instead of landmarks. In the meantime, FashionCapsNet has comparable performance to the best variant of FashionNet which utilizes 1000 attributes and the landmark information during training. Capsule Network model seeing only the images during training has close to, even better performance than, CNN-based architectures supported by hand-crafted landmarks and attributes. Consequently, CNNs may utilize various kinds of side information to deal with deformation and pose variation, but capsules can inherently learn pose information within the activation vectors flowing in the network.

Table 3. Top-K accuracy performance of the variants of the baseline study [3] and our proposed model. FashionNet has different building blocks where the model has different numbers of attributes (A) (i.e. 100, 500 and 1000), or fashion landmarks (L) are replaced with human joints (J) or poselets (P). FashionCapsNet does not use any extra side information during training.

Models & Side Information		**Top-3**	**Top-5**
FashionNet [3]	*+100 A + L*	*47.38%*	*70.57%*
	+500 A + L	*57.44%*	*77.39%*
	+J + L	*72.30%*	*81.52%*
	+P +L	*75.34%*	*84.87%*
	+1000 A +L	*82.58%*	***90.17%***
FashionCapsNet	***No SI/EM***	***83.18%***	*89.83%*

Table 4 summarizes the clothing category classification results of our proposed model and the state-of-the-art methods with the information of applied techniques and the number of parameters in the models. These figures indicate how successful FashionCapsNet is, and what are the performance limits it has when compared to the state-of-the-art CNN-based architectures. First, FashionCapsNet clearly outperforms WTBI [4] and DARN [18] which both use semantic attributes disparately to improve the classification performance, but neglect pose configuration during training. Moreover, as aforementioned before, FashionCapsNet and the best variant of FashionNet [3] (i.e. supported by 1000 attributes and +8 landmarks) have closely contested classification performances on DeepFashion [3] dataset. At this point, while FashionNet

Table 4. Experimental results on DeepFashion dataset for the clothing category classification.

Architectures	**Backbone**	**Side information (SI) / Extra Module (EM)**	**# Parameters**	**Top-3 (%)**	**Top-5 (%)**
WTBI [4]	*AlexNet*	*Category-specific Similarity (SI)*	*60*	*43.73*	*66.26*
DARN [18]	*Custom NiN*	*Visual Similarity (SI)*	*105*	*59.48*	*79.58*
FashionNet [3]	*VGG-16*	*Landmark Information (SI)*	*134*	*82.58*	*90.17*
FashionCapsNet	*CapsNet*	***No SI / EM Used***	*45*	*83.18*	*89.83*
Corbière *et al.* [21]	*ResNet50*	*Bag-of-words Descriptors (EM)*	*25*	*86.30*	*92.80*
Lu *et al.* [20]	*VGG-16*	*Dynamic Branching (EM)*	*134*	*86.72*	*92.51*
Wang *et al.* [22]	*VGG-16*	*Two Attention Modules (EM)*	*142*	*90.99*	*95.78*
Liu *et al.* [23]	*VGG-16*	*Single Attention Module (EM)*	*138*	*91.16*	*96.12*

employs landmark information to recover pose configuration thrown away due to pooling operations, FashionCapsNet can learn pose information by preserving the spatial relationship between pixels. However, our proposed architecture cannot achieve the performance of more advanced CNN-based architectures. The underlying reason for this is that encapsulating the neurons as a densely connected layer is completely different, but not a complex structure. On the other hand, state-of-the-art CNN-based methods referred in Table 4 adopt different techniques (e.g. bag-of-words descriptors, dynamic branching and attention mechanisms) to their models to improve the overall clothing classification performance. In the future, our proposed model, FashionCapsNet, may inherit extra performance boost on the clothing classification, due to advances in the relatively new Capsule Network research.

7. CONCLUSION

Digital transformation is taking over nearly all businesses, and fashion industry is one of the recent adopters of deep learning-based solutions. The initial task for fashion recommendation systems is *generally* to classify the clothing categories. Despite its popularity, the clothing category classification has never been easy to achieve when employed in real world applications. It was attacked with several different methodologies that utilizing feature extractors [8-9] in the past decade. In recent studies, CNN-based deep learning architectures [3-4,18] *generally* perform better when compared to the heuristic methods, and applying some different techniques such as bag-of-descriptors [20], dynamic branching [21] and attention modules [22-23] significantly improves the performance of clothing recognition models.

In this study, we investigate the performance of a custom Capsule Network architecture on the clothing category classification by using DeepFashion dataset [3] that contains the clothing images with 46 fine-grained category labels. Thereafter, we compare our results with the results of the other state-of-the-art CNN-based models. Our first observation is that our Capsule Network design trained only on images perform even better than CNN-based models [3] that is supported by extra information (e.g. hand-crafted landmarks and attributes) besides the images. Secondly, our proposed architecture outperforms the methods [4,18] that neglect pose configuration, while it needs some improvements to reach the performances of more advanced methods [20-23].

As the future work, using matrix capsule structure with EM routing, which has recently been introduced by Hinton *et al.* [38], may increase the classification accuracy of our proposed architecture. Furthermore, extracting the features by residual blocks instead of plain convolutional blocks may increase the representative power of the input of primary capsules. In addition, transfer learning for feature extraction, and adding attention mechanism to the capsule block of FashionCapsNet could be considered as the methods that may improve the overall performance of our model.

REFERENCES

[1] L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack, L. V. Gool, "Apparel Classification with Style", *Proceedings of the 11th Asian conference on Computer Vision (ACCV)*, 321-335, 2012.

[2] B. Willimon, I. Walker, S. Birchfield, "Classification of Clothing Using Midlevel Layers", *ISRN Robotics,* 2013.

[3] Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, "DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2016.

[4] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, T. L. Berg, "Where To Buy It: Matching Street Clothing Photos in Online Shops", *IEEE International Conference on Computer Vision (ICCV)*, 2015

[5] S. Zheng, F. Yang, M. H. Kiapour, R. Piramuthu, "ModaNet: A Large-Scale Street Fashion Dataset with Polygon Annotations", *Proceedings of ACM Multimedia conference (ACM Multimedia'18),* 2018.

[6] Z. Al-Halah, R. Stiefelhagen, K. Grauman, "Fashion Forward: Forecasting Visual Style in Fashion", *IEEE International Conference on Computer Vision (ICCV),* 2017.

[7] W. Di, C. Wah, A. Bhardwaj, R. Piramuthu, N. Sundaresan, "Style Finder: Fine-Grained Clothing Style Recognition and Retrieval", *IEEE International Workshop on Mobile Vision (IWMV)*, 2013.

[8] N. Dalal, B. Triggs, "Histograms of Oriented Gradients for Human Detection", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 1, 886–893, 2005.

[9] D. G. Lowe, "Distinctive Image Features from Scale-invariant Key-points", *International Journal of Computer Vision (IJCV)*, 60(2), 91–110, 2004.

[10] A. Krizhevsky, I. Sutskever, G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", *Neural Information Processing Systems (NIPS),* 1106–1114, 2012.

[11] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh et al., "ImageNet Large Scale Visual Recognition Challenge (ILSVRC)", *International Journal of Computer Vision (IJCV),* 2015.

[12] K. Zhao, X. Hu, J. Bu, C. Wang, "Deep Style Match for Complementary Recommendation", **Workshops at the Thirty-First AAAI Conference on Artificial Intelligence**, 2017.

[13] Yu, H. Zhang, X. He, X. Chen, L. Xiong, Z. Qin, "Aesthetic-based Clothing Recommendation", *WWW*, 649–658, 2018.

[14] H. Tuinhof, C. Pirker, M. Haltmeier, "Image-based Fashion Product Recommendation with Deep Learning", **IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, 2018.

[15] W. Luo, Y. Li, R. Urtasun, R. Zemel, "Understanding the Effective Receptive Field in Deep Convolutional Neural Networks", **Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS)**, 4905–4913, 2016.

[16] S. Sabour, N. Frosst, G. E. Hinton, "Dynamic Routing between Capsules", *Neural Information Processing Systems (NIPS),* 3859–3869*,* 2017.

[17] J. Lafferty, A. McCallum, F. C. N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", **Proceedings of the 18th International Conference on Machine Learning (ICML)***,* June 2001.

[18] J. Huang, R. S. Feris, Q. Chen, S. Yan, "Cross-domain Image Retrieval with A Dual Attribute-aware Ranking Network", **Proceedings of the IEEE International Conference on Computer Vision (ICCV)***,* 1062–1070, 2015.

[19] M. Lin, Q. Chen, S. Yan, "Network in Network", **International Conference on Learning Representations (ICLR)***,* 2014.

[20] Y. Lu, A. Kumar, S. Zhai, Y. Cheng, T. Javidi, R. S. Feris, "Fully-adaptive Feature Sharing in Multi-task Networks with Applications in Person Attribute Classification", *A Computing Research Repository (CoRR)*, abs/1611.05377, 2016.

[21] C. Corbiere, H. Ben-Younes, A. Rame, C. Ollion, "Leveraging Weakly Annotated Data for Fashion Image Retrieval and Label Prediction", **IEEE International Conference on Computer Vision (ICCV) workshop***,* 2017.

[22] W. Wang, Y. Xu, J. Shen, S. C. Zhu, "Attentive Fashion Grammar Network for Fashion Landmark Detection and Clothing Category Classification", **IEEE Conference on Computer Vision and Pattern Recognition (CVPR)***,* 4271–4280, 2018.

[23] J. Liu, H. Lu, "Deep Fashion Analysis with Feature Map Upsampling and Landmark-driven Attention", **IEEE European Conference on Computer Vision (ECCV) workshop***,* 2019.

[24] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition", **IEEE Conference on Computer Vision and Pattern Recognition (CVPR)***,* 770–778, 2016.

[25] C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, "Inception-v4, Inception-ResNet and The Impact of Residual Connections on Learning", *AAAI*, 4, 12, 2017.

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed et al., "Going Deeper With Convolutions", **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)***,* 1–9, 2015.

[27] J. Redmon, S. K. Divvala, R. B. Girshick, A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection", **IEEE Conference on Computer Vision and Pattern Recognition (CVPR)***,* 779–788*,* 2016.

[28] R. Girshick, "Fast R-CNN", **Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)***,* 1440–1448, 2015.

[29] S. Ren, K. He, R. Girshick, J. Sun, "Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks", **Proceedings of the Neural Information Processing Systems (NIPS)***,* 91–99, 2015.

[30] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley et al., "Generative adversarial nets", *Neural Information Processing Systems (NIPS),* 2014.

[31] A. Radford, L. Metz, S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks", **International Conference on Learning Representations (ICLR)**, 2016.

[32] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford et al., "Improved Techniques for Training GANs", *Neural Information Processing Systems (NIPS),* 2234–2242, 2016.

[33] G. E. Hinton, A. Krizhevsky, S. D. Wang, "Transforming auto-encoders.", **International Conference on Artificial Neural Networks (ICANN)***, Springer*, 44–51, 2011.

[34] K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", *arXiv preprint,* abs/1409.1, 1–10, 2014.

[35] A. L. Maas, A. Y. Hannun, A. Y. Ng. "Rectifier Nonlinearities Improve Neural Network Acoustic Models", **International Conference on Machine Learning (ICML)***,* 30, 2013.

[36] S. Ioffe, C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift", **Proceedings of International Conference on Machine Learning (ICML)***,* 37, 448–456, 2015.

[37] D. P. Kingma, J. Ba, "ADAM: A Method for Stochastic Optimization", **3rd International Conference for Learning Representations (ICLR)**, San Diego, 2014.

[38] G. Hinton, S. Sabour, N. Frosst, "Matrix Capsules with EM Routing", **International Conference on Learning Representations (ICLR)***,* 2018.

[39] Y. Hu, X. Li, N. Zhou, L. Yang, L. Peng, S. Xiao, "A Sample Update-Based Convolutional Neural Network Framework for Object Detection in Large-Area Remote Sensing Images", *IEEE Geoscience and Remote Sensing Letters*, 16(6), 947-951, 2019.

[40] M. M. Ozguven, K. Adem, "Automatic detection and classification of leaf spot disease in sugar beet using deep learning algorithms", *Physica A: Statistical Mechanics and its Applications*, 535, 122537, 2019.

[41] Y. Wei, X. Liu, "Dangerous goods detection based on transfer learning in X-ray images", *Neural Computing and Applications*, 1-14, 2019.