Tiered Deep Similarity Search for Fashion

Dipu Manandhar, Muhammet Bastan, and Kim-Hui Yap

Nanyang Technological University, Singapore 639798
{dipu002,muhammetbastan,ekhyap}@ntu.edu.sg

Abstract. *How similar are two fashion clothing?* Fashion apparels demonstrate diverse visual concepts with their designs, styles and brands. Hence, there exist a hierarchy of similarities between fashion clothing, ranging from exact instance or brand to similar attributes, styles. An effective search method, thus, should be able to represent the tiers of similarities. In this paper, we present a deep learning based fashion search framework for learning the tiers of similarity. We propose a new attribute-guided metric learning (AGML) with multitask CNNs that jointly learns fashion attributes and image embeddings while taking category and brand information into account. The two tasks in the framework are linked with a guiding signal. The guiding signal, first, helps in mining informative training samples. Secondly, it helps in treating training samples by their importance to capture the tiers of similarity. We conduct experiments in a new BrandFashion dataset which is richly annotated at different granularities. Experimental results demonstrate that the proposed method is very effective in capturing a tiered similarity search space and outperforms the state-of-the-art fashion search methods.

Keywords: Fashion Search, Deep Metric Learning, Multitask Learning

1 Introduction

Fashion contributes a significant portion in rapidly growing online shopping and social media [9, 2]. With such a growth, visual fashion analysis has received a huge research attention [19, 1, 24, 16, 15] and has been successfully deployed in large e-commerce companies and websites [29, 23, 13] such as *eBay, Amazon, Pinterest, Flipkart,* etc. One of the most important aspects in visual fashion analysis is fashion search. This paper presents a new deep learning based fashion search framework with an interesting fusion of multitask and metric learning.

With the recent advances in deep learning, end-to-end metric learning methods for visual similarity measure have been proposed [22, 3]. The main task here is to learn a discriminative feature space for image representation. Particularly for fashion search, the feature space should incorporate various elements of fashion. As fashion domain demonstrates a huge diversity in visual concepts with their designs, styles, brands, there exist tiers of similarity for fashion clothing. Visual fashion similarity can be defined based on various concepts such as categories (e.g. *dress, hoodie*), brands (e.g. *Adidas, Nike*), attributes (e.g. *color, pattern*) or design (*cropped, zippered*). Fig. 1 illustrates tiers of similarity for

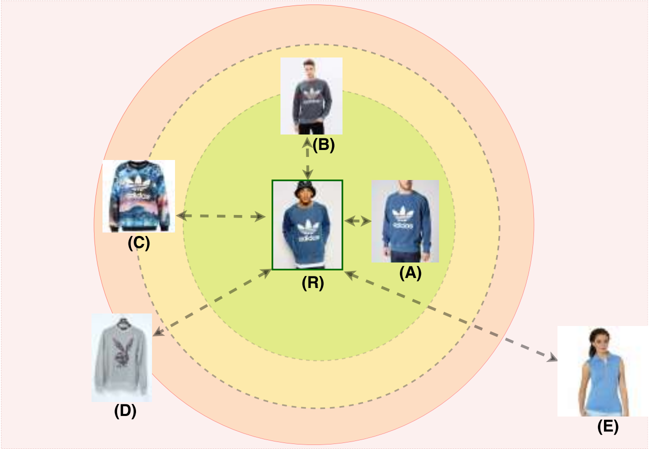

Fig. 1. Example of tiers of similarity in feature space for fashion clothing. Different tiers of similarity are denoted by the dotted concentric circles. Distances between the reference clothing (R) and clothing (A)-(E) represent the degrees of visual similarity.

clothing images. Clothing (A) and the reference clothing (R) are the exact same (brand, model, categories, color etc.) clothing and hence lies closest within the inner circle. Clothing (B) shares the same model as clothing (R) with a different color and hence it is second nearest to (R). Similarly, clothing items (C),(D) & (E) lie farther away. We aim to learn such a tiered feature space, as this provides the desired retrieval outcome for practical fashion search applications.

Deep metric learning has demonstrated huge success in learning visual similarity [22, 3, 8, 28, 11]. Siamese networks [8, 28, 5] and triplet networks [27, 22] are the most popular models for metric learning, the latter being reported to be better [22, 10]. Although successful, the existing triplet based methods [22, 3, 10] have few limitations. First, they require exact instance/ID level annotations, and do not perform well with weak label annotations *e.g.* , category labels (shown in Sec. 3). Secondly, these methods employ hard binary decisions during the triplet selection and treat the selected triplets with equal importance, which creates a restriction to learn tiers of similarity.

To learn discriminative feature space, researchers have combined metric learning with auxiliary information using multitask networks which have achieved better performance for face identification and recognition [6, 30, 22], person re-identification [18, 14], clothing search [12]. Particularly for fashion representation, multitask learning with attribute information is used in [12, 24]. *Where-To-Buy-It* (WTBI) [15] used pre-trained features and learned a similarity network using Siamese networks. Recently, *FashionNet* in [19] proposed to jointly optimize classification, attribute prediction and triplet loss for fashion search. However, they do not explore the possible interaction between the tasks and hence do not effectively learn a tiered similarity space required for fashion search.

In view of this, we propose a new *attribute guided metric learning* (AGML) framework using multitask learning for fashion search. The proposed framework utilizes the interactions between the attribute prediction and triplet network, by jointly training them. This has two major advantages over the existing methods. First, it helps in mining informative triplets especially when exact anchor-positive pair annotations are not available. Second, training samples are treated based on their importance in a soft manner which helps in capturing multiple tiers of similarity required for fashion search. We demonstrate its effectiveness for fashion search using a new BrandFashion dataset. Compared to the existing fashion datasets [7, 4, 19], this dataset is richly annotated with essential elements of fashion including clothing categories, attributes, and brand information which capture different tiers of information in fashion.

2 Proposed Method

The architecture of the proposed framework is shown in Fig. 2. It consists of three identical CNNs with shared parameters θ and accepts image triplets $\{x^a, x^p, x^n\}$ *i.e.* an anchor image (x^a), a positive image (x^p) from the same class as the anchor, and a negative image (x^n) from a different class. The last fully connected layer has two branches for learning the feature embedding $f(x)$, and the attribute vector $\mathbf{v}$. The guiding signal links two tasks and helps triplet sampling based on the importance of the samples. The network is trained end-to-end using the loss,

$$L_{total}(\theta) = L^G_{tri}(\theta) + \lambda L_{attr}(\theta) \tag{1}$$

where $L^G_{tri}(\theta)$ & $L_{attr}(\theta)$ represent the attribute-guided triplet loss & attribute loss respectively, and λ balances the contribution of the two losses.

2.1 Attribute Prediction Network

We use K semantic attributes to describe the image appearance, denoted $\mathbf{a} = [a_1, a_2, \ldots, a_K]$, where each element $a_i \in \{0, 1\}$ indicates the presence or absence of the i^{th} attribute. The problem of attribute prediction is treated as multilabel classification. We pass the first branch from last fully connected layers into a sigmoid layer to squash the output to $[0, 1]$ and output $\mathbf{v}$. The attribute prediction is optimized using binary-cross entropy loss $L_{attr}(\theta) = -\sum_{i=1}^{K} [a_i \log(v_i) + (1 - a_i) \log(1 - v_i)]$, where a_i is binary target attribute labels for image x, and v_i is a component of $\mathbf{v} = [v_1, v_2, \ldots, v_K]$, which is the predicted attribute distribution.

2.2 Attribute Guided Triplet Training

We use the predicted attribute vectors to guide both triplet mining and triplet loss training.

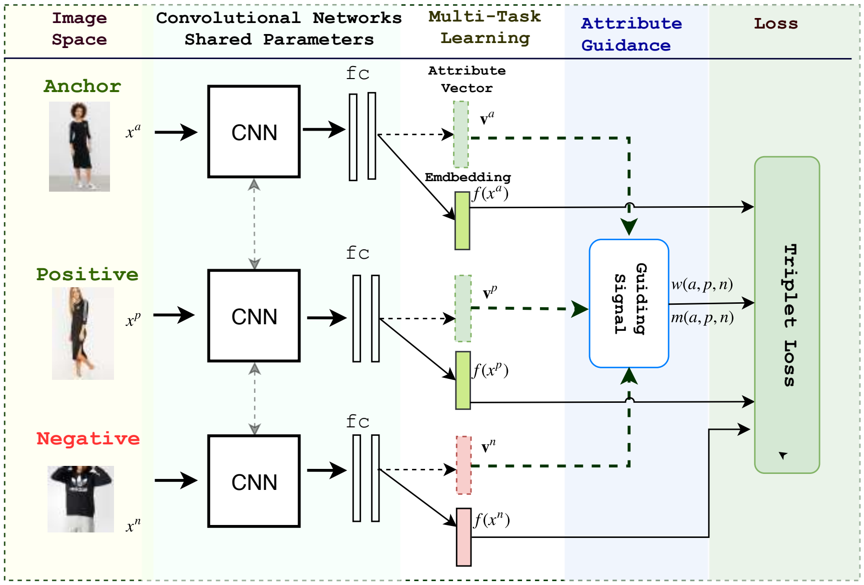

Fig. 2. Architecture of the proposed attribute-guided triplet network.

A. Triplet Mining

Random sampling based on class/ID labels for triplet does not assure the selection of the most informative examples for training. This is especially critical when only category information is available. For effective training, anchor-positive pairs should be reliable. Therefore, we propose to leverage cosine similarity $\langle \mathbf{x}, \mathbf{y} \rangle = \frac{\mathbf{x}^\intercal \mathbf{y}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2}$, between the anchor-positive attribute vectors (outputs of the attribute prediction network) to sample better triplets. In particular, we use a threshold (Φ) such that only the triplets with $\langle \mathbf{v}^a, \mathbf{v}^p \rangle > \Phi$ are selected for the training. This ensures that the anchor-positive pairs are similar in attribute space and hence are reliable.

B. Attribute Guided Triplet Training

We propose two ways to guide the triplet metric learning network. The first weights the whole triplet loss while the second operates on the margin parameter of the loss function. Let $\{x^a, x^p, x^n\}$ be an input triplet sample and $\{f(x^a), f(x^p), f(x^n)\}$ be the corresponding embedding. The proposed attribute-guided triplet loss given by,

$$L_{tri}^{G}(\theta) = w(a,p,n) \left[\|f(x^a) - f(x^p)\|_2^2 - \|f(x^a) - f(x^n)\|_2^2 + m(a,p,n) \right]_+ , \quad (2)$$

where $w(a,p,n)$ and $m(a,p,n)$ are the loss weighting factor and margin factor, which are functions of attribute distributions $\{\mathbf{v}^a, \mathbf{v}^p, \mathbf{v}^n\}$, as explained below.

B1. Soft-Weighted (SW) Triplet Loss
The SW triplet loss operates on the overall loss using the weight factor $w(a,p,n)$. We use $w(a,p,n) = \langle \mathbf{v}^a, \mathbf{v}^p \rangle \langle \mathbf{v}^a, \mathbf{v}^n \rangle$, the product of similarities between attribute vectors of anchor-positive and anchor-negative pairs. The above function adaptively alters the magnitude of the triplet loss. When the anchor-positive pair is similar in attribute space (*i.e.* $\langle \mathbf{v}^a, \mathbf{v}^p \rangle$ is high), the sample is more confident and reliable. Likewise, when the anchor-negative pair is similar in attribute space (*i.e.* $\langle \mathbf{v}^a, \mathbf{v}^n \rangle$ is high), it forms a *hard negative example i.e.* high information. Hence, the triplet is given higher priority and more attention during the network update. This is analogous to hard negative mining [22], but we handle them in a soft manner.

B2. Soft-Margin (SM) Triplet Loss
The SM triplet loss operates on the margin parameter using $m(a,p,n)$. The naive triplet loss uses a constant margin m, which treats all triplets equally and restricts learning desired tiered similarity. The soft margin is an adaptive margin $m(a,p,n) = m_0 \log(1 + \langle \mathbf{v}^a, \mathbf{v}^p \rangle \langle \mathbf{v}^a, \mathbf{v}^n \rangle)$, and promotes a tiered similarity space. Similar to SW triplet loss, when both $\langle \mathbf{v}^a, \mathbf{v}^p \rangle$ and $\langle \mathbf{v}^a, \mathbf{v}^n \rangle$ are high, the triplet is more reliable and informative (hard negative), and hence the effective margin becomes larger. In other words, when the negative image and anchor image are very similar in the attribute space, a reliable margin is used to learn the subtle difference and avoid the confusion. Hence, both SW and SM triplet loss explore the importance of the triplets, which helps in learning a tiered similarity space.

3 Experiments

We collected a new BrandFashion dataset with about $10K$ clothing images with distinctive logos from 15 brands. The images are categorized into 16 clothing categories and annotated with 32 semantic attributes. The goal is to demonstrate the tiered similarity space using the category, brand and attribute annotations. There are 50 query images in the dataset. We evaluated the performance for instance search using mean average precision (mAP) and the performance of tiered similarity search using normalized discounted cumulative gain (NDCG) *i.e.* $NDCG@k = \frac{1}{Z}\sum_i^k \frac{2^{r(i)}-1}{\log_2(1+i)}$. The relevance score of i^{th} ranked image is calculated based on similarity match considering three levels of information, namely category, brand and attribute *i.e.* $r(i) = r_i^{cat} + r_i^{brand} + r_i^{attr}$, where $r_i^{cat} \in \{0,1\}$, $r_i^{brand} \in \{0,1\}$. The attribute match r_i^{attr} is computed by taking the ratio of the number of matched attributes to the total number of query attributes [12]. Overall, the relevance score summarizes the tiered similarity search performance.

We used VGG16 [25] as the base CNN network, which is trained using the loss defined in Eq. (1), with SGD momentum of 0.5 & learning rate of 0.001. We set λ to 1. The value of margin m_0 is experimentally set to 0.5. For the SM triplet loss, the value of m_0 set to 0.8 such that the effective margin swings around the original value. We set the threshold (ϕ) to 0.7 and observe that the performance is fairly stable on $\phi \in [0.5, 0.9]$.

Table 1. Comparison of the proposed method with state-of-the-art-methods

Method	mAP (%)	NDCG@20(%)
Triplet Loss [22]	33.38	69.92
Multitask Network (Triplet+Attribute)	56.41	76.71
R-MAC [26]	30.03	70.70
Rapid-Clothing [17]	33.17	70.78
Visual-Search@Pinterest [13]	37.11	72.11
VisNet [23]	34.01	63.76
WTBI [15]	41.73	56.43
FashionNet [19]	50.14	80.63
Proposed AGML (SW)	**63.79**	**85.12**
Proposed AGML (SM)	63.71	83.66
Proposed AGML (SW) with Re-ranking	**71.25**	96.16
Proposed AGML (SM) with Re-ranking	71.05	**96.24**

Items from the same category and brand are sampled for the anchor-positive pairs, and items from different categories or brands constitute the negative samples. L_2-normalized feature from the last fully connected layer is used as the feature vector. We used PyTorch [20] for the implementation. Similar to [19, 15, 23], we crop out the clothing region prior to feature extraction. We used Faster-RCNN [21] to jointly detect the brand logo and clothing items in the images.

Table 1 compares performance of different methods in terms of mAP and NDCG@20. In terms of mAP, the naive triplet loss achieves 33.8%, while the multitask network (triplet+attribute) achieves 56.4%. This shows that there is a clear benefit of using auxiliary information using multitask metric learning. The proposed method additionally guides the triplet loss using the predicted attributes. The proposed AGML-SW and AGML-SM achieve mAPs of 63.79% and 63.71%. This demonstrates the advantage of attribute guided triplet loss. The proposed method clearly outperforms the deep feature encoding based methods [26, 17, 13], and state-of-the art metric learning methods [23, 15, 19].

Similar trend in results can be observed for NDCG in Table 1. The proposed attribute-supervised SW and SM triplet network achieve NDCG@20 of 83.66% and 85.12% respectively. Our method clearly outperforms other state-of-the-art methods which demonstrate the advantage of learning a tiered similarity space. We further take advantage of logo detection to re-rank the retrieval results. The proposed method achieves mAP $\approx$ 71% and NDCG@20 $\approx$ 96% with re-ranking based on detected brand logo information. Figure 3 shows example search results obtained using WTBI[15], FashionNet[19] and the proposed method which further demonstrates the advantage of the proposed method.

4 Conclusions

We presented a new deep attribute-guided triplet network which explores the importance of training samples and learns a tiered similarity space. The method

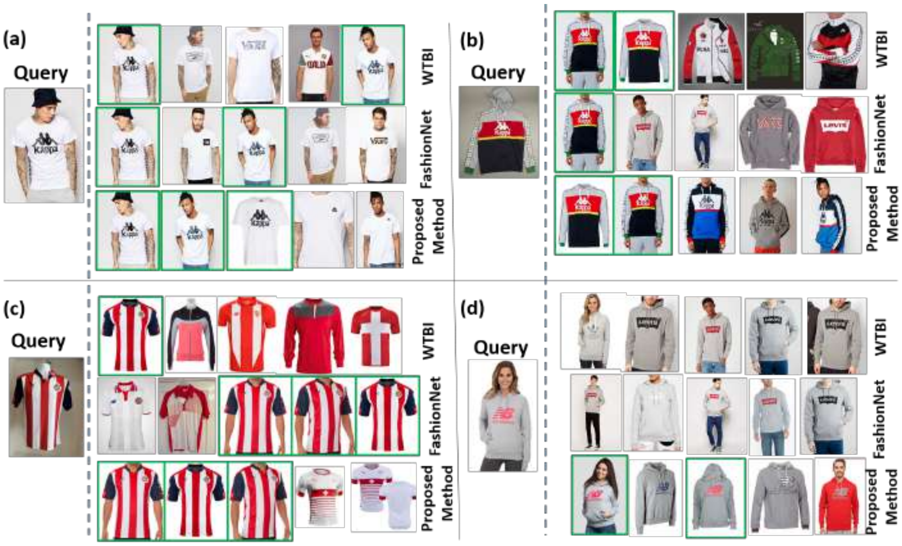

Fig. 3. Sample search results with query images and top-5 retrieved images. Exact same instance matches are highlighted with green borders. Best viewed in color.

uses multitask CNN which shares the mutual information among the tasks for better tuning the loss. Using the predicted attributes, the proposed method first mines informative triplets, and then uses them to train the triplet loss in a soft-manner, which helps in capturing the tiered similarity desirable for fashion search. We believe that the tiered similarity search will be appreciated by fashion companies, online retailers as well as customers.

Acknowledgment

This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the Infocomm Media Development Authority, Singapore. We gratefully acknowledge the support of NVIDIA AI Technology Center for their donation of GPUs used for our research.

References

1. Al-Halah, Z., Stiefelhagen, R., Grauman, K.: Fashion forward: Forecasting visual style in fashion. In: IEEE International Conference on Computer Vision (ICCV). pp. 388–397. IEEE (2017)
2. Baldwin, C.: Online spending continues to increase thanks to fashion sector (2014), https://www.computerweekly.com/news/2240225386/Spend-online-continues-to-increase-thanks-to-fashion-sector

3. Bell, S., Bala, K.: Learning visual similarity for product design with convolutional neural networks. ACM Transactions on Graphics (TOG) **34**(4), 98 (2015)
4. Bossard, L., Dantone, M., Leistner, C., Wengert, C., Quack, T., Van Gool, L.: Apparel classification with style. In: Asian Conference on Computer Vision. pp. 321–335. Springer (2012)
5. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a siamese time delay neural network. In: Advances in Neural Information Processing Systems. pp. 737–744 (1994)
6. Chechik, G., Sharma, V., Shalit, U., Bengio, S.: Large scale online learning of image similarity through ranking. Journal of Machine Learning Research **11**(Mar), 1109–1135 (2010)
7. Chen, H., Gallagher, A., Girod, B.: Describing clothing by semantic attributes. In: European Conference on Computer Vision. pp. 609–623. Springer (2012)
8. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol. 1, pp. 539–546 (2005)
9. Financial Times: Online retail sales continue to soar (2018), https://www.ft.com/content/a8f5c780-f46d-11e7-a4c9-bbdefa4f210b
10. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv:1703.07737 (2017)
11. Hu, J., Lu, J., Tan, Y.P.: Discriminative deep metric learning for face verification in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1875–1882 (2014)
12. Huang, J., Feris, R.S., Chen, Q., Yan, S.: Cross-domain image retrieval with a dual attribute-aware ranking network. In: International Conference on Computer Vision. pp. 1062–1070 (2015)
13. Jing, Y., Liu, D., Kislyuk, D., Zhai, A., Xu, J., Donahue, J., Tavel, S.: Visual search at pinterest. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1889–1898. ACM (2015)
14. Khamis, S., Kuo, C.H., Singh, V.K., Shet, V.D., Davis, L.S.: Joint learning for attribute-consistent person re-identification. In: European Conference on Computer Vision. pp. 134–146. Springer (2014)
15. Kiapour, M.H., Han, X., Lazebnik, S., Berg, A.C., Berg, T.L.: Where to buy it: Matching street clothing photos in online shops. In: IEEE International Conference on Computer Vision. pp. 3343–3351 (2015)
16. Kiapour, M.H., Yamaguchi, K., Berg, A.C., Berg, T.L.: Hipster wars: Discovering elements of fashion styles. In: European Conference on Computer Vision. pp. 472–488. Springer (2014)
17. Lin, K., Yang, H.F., Liu, K.H., Hsiao, J.H., Chen, C.S.: Rapid clothing retrieval via deep learning of binary codes and hierarchical search. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. pp. 499–502. ACM (2015)
18. Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Yang, Y.: Improving person re-identification by attribute and identity learning. arXiv:1703.07220 (2017)
19. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1096–1104 (2016)
20. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: PyTorch. http://pytorch.org

21. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (2015)
22. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNnet: A unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 815–823 (2015)
23. Shankar, D., Narumanchi, S., Ananya, H., Kompalli, P., Chaudhury, K.: Deep learning based large scale visual recommendation and search for e-commerce. arXiv:1703.02344 (2017)
24. Simo-Serra, E., Ishikawa, H.: Fashion style in 128 floats: joint ranking and classification using weak data for feature extraction. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 298–307 (2016)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
26. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of cnn activations. arXiv:1511.05879 (2015)
27. Wang, J., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., Wu, Y., et al.: Learning fine-grained image similarity with deep ranking. arXiv:1404.4661 (2014)
28. Wang, X., Sun, Z., Zhang, W., Zhou, Y., Jiang, Y.G.: Matching user photos to online products with robust deep features. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval. pp. 7–14. ACM (2016)
29. Yang, F., Kale, A., Bubnov, Y., Stein, L., Wang, Q., Kiapour, H., Piramuthu, R.: Visual Search at eBay. arXiv:1706.03154 (2017)
30. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv:1411.7923 (2014)