Fashion Image Retrieval with Capsule Networks

Furkan Kınlı[1] Barış Özcan[2] Furkan Kıraç[3]
Department of Computer Science, Özyeğin University, İstanbul, Türkiye
{[1]furkan.kinli, [3]furkan.kirac}@ozyegin.edu.tr, [2]baris.ozcan.10097@ozu.edu.tr

Abstract

In this study, we investigate in-shop clothing retrieval performance of densely-connected Capsule Networks with dynamic routing. To achieve this, we propose Triplet-based design of Capsule Network architecture with two different feature extraction methods. In our design, Stacked-convolutional (SC) and Residual-connected (RC) blocks are used to form the input of capsule layers. Experimental results show that both of our designs outperform all variants of the baseline study, namely FashionNet, without relying on the landmark information. Moreover, when compared to the SOTA architectures on clothing retrieval, our proposed Triplet Capsule Networks achieve comparable recall rates only with half of parameters used in the SOTA architectures.

Capsule Networks

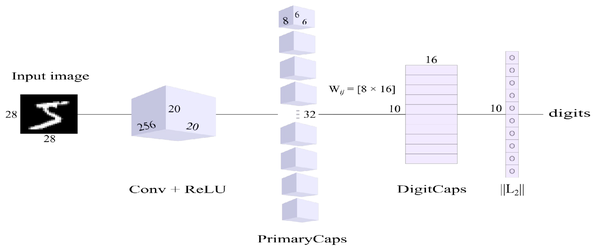


Figure 1. Capsule Network architecture proposed by Sabour and Hinton *et al.* [1].

Capsules are groups of neurons that convey higher dimensional information throughout the network in more refined way. This information is interpreted as the pose configuration and the existence probability of an instance. Each capsule in a higher level is formed by the routing of incoming votes from the capsules in lower level. At this point, these votes are calculated by the linear transformation of the pose configuration. During *dynamic routing* [1], the linear combination of incoming votes weighted by their coefficients (i.e. *coupling coefficients*) forms the non-activated outputs in higher level capsules. For each iteration, the weights of these votes are updated with respect to the dot product of the incoming votes and the outputs in higher level capsules. This is called *agreement between capsules*. Finally, the output of each capsule in lower level is determined by *squashing* function as proposed in [1].

Our Architectures

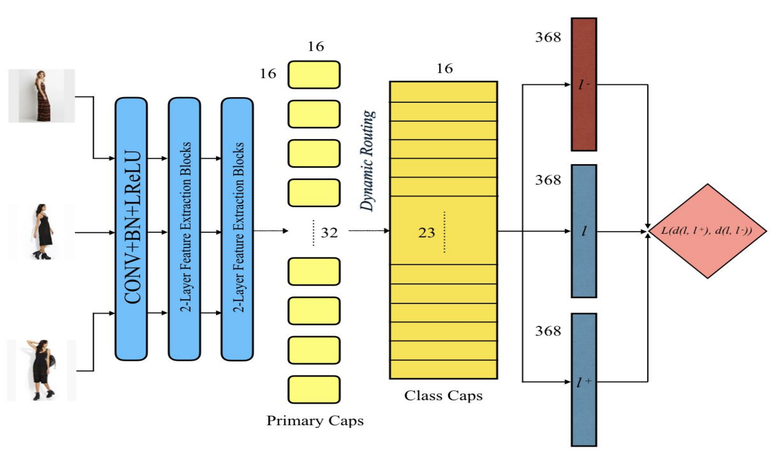


Figure 2. Illustration of our proposed architectures containing different feature extraction blocks.

DeepFashion



Figure 3. Example of a triplet containing an anchor image (a), a positive image (b) and a negative image (c) from in-shop partition of DeepFashion [2] data set.

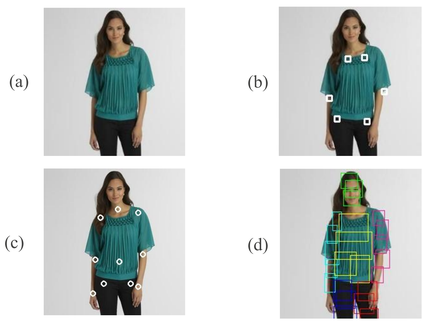


Figure 4. Example images from DeepFashion [2] data set with (a) and without (b) landmarks employed; with (c) human joints and with (d) poselets, a part of pose.

Triplet Relation

$$d(l, l^+) = \|f(x) - f(x^+)\|_2^2$$

$$d(l, l^-) = \|f(x) - f(x^-)\|_2^2$$

$$d(l, l^+) + \alpha < d(l, l^-)$$

Objective Function

$$L = \sum_i [d(l_i, l_i^+) - d(l_i, l_i^-) + \alpha]$$

Experimental Settings

Table 1. Hyper-parameter settings.

Optimizer	Adam [3]
Learning Rate	0.001
Decay Rate	0.0005
Batch Size	32
Routings	3
Margin	2.0
Normalization	Pixel-wise

Results

Table 2. Recall@K performance of the variants of the baseline study [2] 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). SCCapsNet and RCCapsNet do not use any extra side information during training.

Models		Top-20 (%)	Top-50 (%)
FashionNet +	(100 A + L)	57.3	62.5
	(500 A + L)	64.6	69.5
	(1000 A + J)	68.0	73.5
	(1000 A + P)	70.0	75.0
	(1000 A + L)	76.4	80.0
SCCapsNet	Not Used	81.8	90.9
RCCapsNet	Not Used	**84.6**	**92.6**

Table 3. Experimental results of in-shop clothing retrieval task on DeepFashion data set. "-": not reported.

Architectures	#P	Top-1 (%)	Top-10 (%)	Top-20 (%)	Top-50 (%)
WTBI [4]	60	35.0	47.0	50.6	54.5
DARN [5]	105	38.0	56.0	67.5	72.5
FashionNet	134	53.2	72.5	76.4	80.0
Corbière *et al.* [6]	25	39.0	71.8	78.1	85.6
SCCapsNet	2.5	32.1	72.4	81.8	90.9
RCCapsNet	4.5	33.9	75.2	84.6	92.6
HDC	5	62.1	84.9	89.0	93.1
VAM	6	66.6	88.7	92.3	-
BIER	5	76.9	92.8	95.2	97.1
HTL	5	80.9	94.3	95.8	97.8
A-BIER	5	83.1	95.1	96.9	98.0
ABE	10	87.3	96.7	97.9	98.7

References

[1] S. Sabour, N. Frosst, G. E. Hinton, "Dynamic Routing between Capsules", *Neural Information Processing Systems (NIPS)*, 3859–3869, 2017.
[2] 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.
[3] D. P. Kingma, J. Ba, "ADAM: A Method for Stochastic Optimization", *3rd International Conference for Learning Representations* (ICLR), San Diego, 2014.
[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] 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.
[6] 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.
[7] Y. Yuan, K. Yang, and C. Zhang. Hard-Aware Deeply Cascaded Embedding. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 814–823, Oct 2017.
[8] Z. Wang, Y. Gu, Y. Zhang, J. Zhou, and X. Gu. Clothing retrieval with visual attention model. 2017 VCIP, pages 1–4, 2017.
[9] M. Opitz, G. Waltner, H. Possegger, and H. Bischof. BIER: Boosting Independent Embeddings Robustly. 2017 IEEE International Conference on Computer Vision (ICCV), pages 5199–5208, 2017.
[10] W. Ge. Deep Metric Learning with Hierarchical Triplet Loss. In The European Conference on Computer Vision (ECCV), September 2018.
[11] J.W. Kim, B. Goyal, K. Chawla, J. Lee, and K. Kwon. Attention-based ensemble for deep metric learning. In The European Conference on Computer Vision (ECCV), September 2018.