Diversity in Fashion Recommendation Using Semantic Parsing

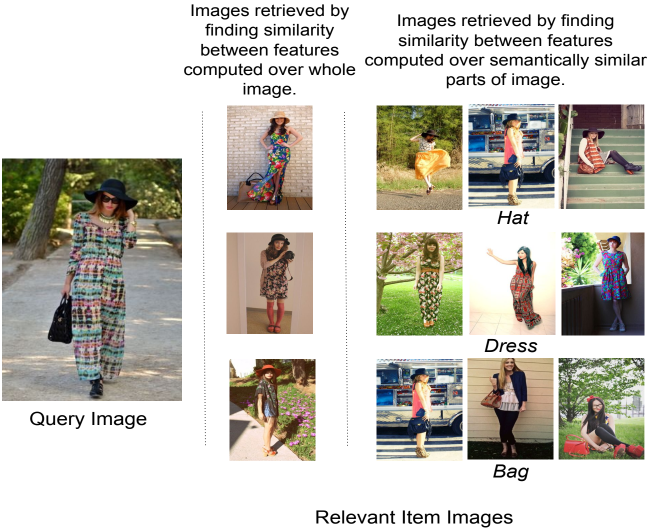
Sagar Verma[1], Sukhad Anand[1], Chetan Arora[1], Atul Rai[2]

[1]Department of Computer Science and Engineering
Indraprastha Institute of Information Technology, Delhi.

[2]Staqu Technologies

ICIP, 2018

Recommendation based on contextual similarity



Problem Statement

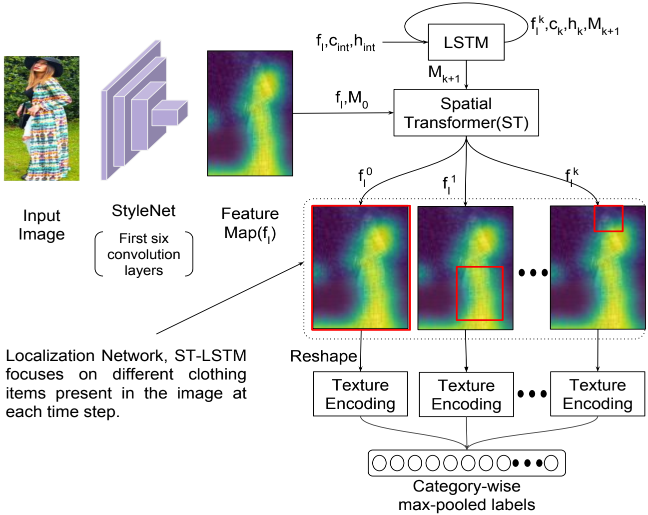
- Developing recommendation system for fashion images is challenging due to the inherent ambiguity associated with what criterion a user is looking at.
- Suggesting multiple images where each output image is similar to the query image on the basis of a different feature or part is one way to mitigate the problem.

Contributions

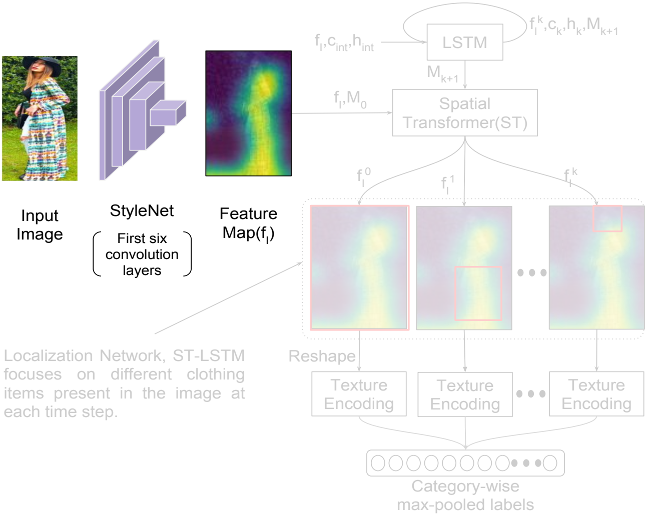
1. Given the ambiguous nature of user intent, we propose a new recommendation system to produce diverse recommendations on the basis of similarity of different parts in the query image.
2. To generate semantically meaningful parts in the fashion image, we propose to use attention based deep neural network which learns to attend to different parts using the weakly labeled data available in the benchmark dataset.
3. Instead of features from standard pre-trained neural networks, we suggest using texture-based features which, as we show in our experiments, are better suited for finding clothing similarity.
4. Apart from diversity in fashion recommendation, our experiments and evaluations on multiple datasets demonstrate the superiority of the proposed model in attribute classification and cross-scene image retrieval tasks.

Related Work

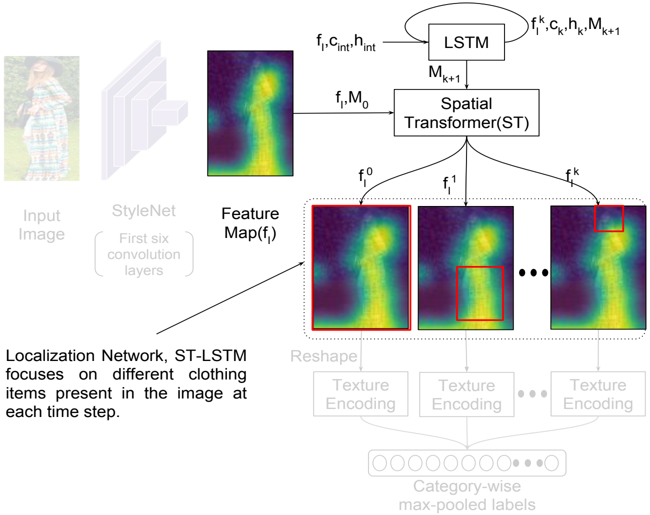
Proposed Architecture



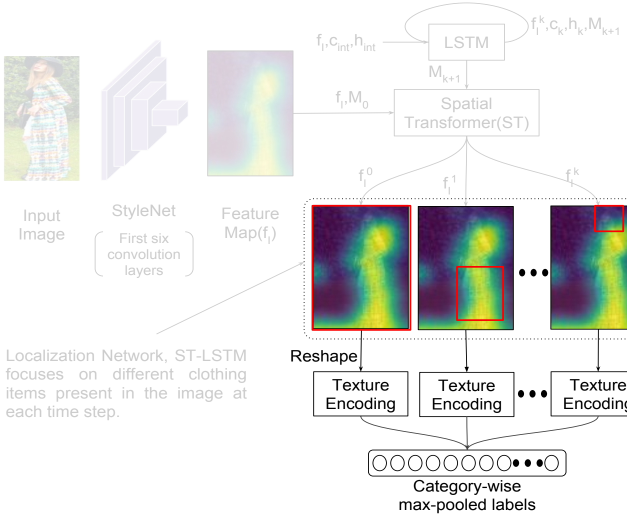
CNN for Global Image Features



Visual Attention Module



Texture Encoding Layer



Loss

Multi-label classification loss

$$\mathcal{L}_{cls} = \frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} (p_i^c - \hat{p}_i^c)^2 \tag{1}$$

where, N is training sample, C is total number of classes, $\hat{p}_i$ is ground truth label vector of sample i and p_i is predicted label vector of sample i.

Diversity loss

Diversity loss is the correlation between adjacent attention maps,

$$\mathcal{L}_{div} = \frac{1}{K-1} \sum_{k=2}^{K} \sum_{i=1}^{HxW} l_{k-1,i} . l_{k,i} \tag{2}$$

where, K is the total steps of recurrent attention, HxW is the height and width of attention maps, l_k is the k^{th} attention map.

Localisation loss

Localisation loss, $\mathcal{L}_{loc}$ from [] is used to remove redundant locations and force localization network to look at small clothing parts.

Combined Loss

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{div} + \lambda_2 \mathcal{L}_{loc} \tag{3}$$

where λ_1 and λ_2 are multiplicative factors. We use 0.01 for all our experiments.

Datasets

- **Fashion144K**
  - 90,000 images with multilabel annotation.
  - 128 classes.
  - Image resolution is 384x256.
- **Fashion550K**
  - 66 classes.
- **DeepFashion**
  - 800,000 images
  - Similarity pairs is available for consumer-to-shop and in-shop retrieval tasks

Experiments

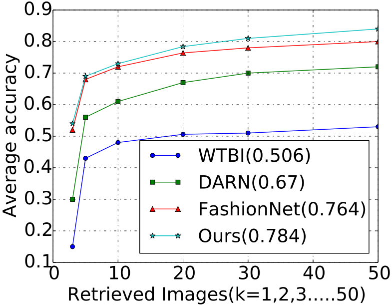
- Model is trained on Fashion144K dataset with 59 item labels, color labels were excluded.
- Evaluated item recognition task on Fashion144K and Fashion550K dataset.
- Consumer-to-shop and in-shop retrieval tasks are evaluated on DeepFashion dataset.

Results

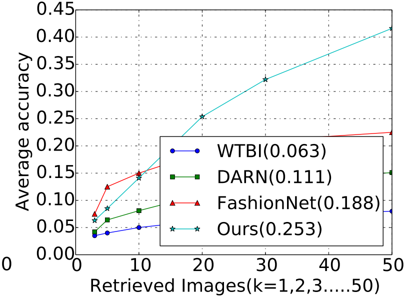
Dataset	Fashion144k [1]		Fashion550k [2]	
Model	AP_{all}	mAP	AP_{all}	mAP
StyleNet [1]	65.6	48.34	69.53	53.24
Baseline [2]	62.23	49.66	69.18	58.68
Viet et al. [3]	NA	NA	78.92	63.08
Inoue et al. [2]	NA	NA	79.87	**64.62**
Ours	**82.78**	**68.38**	**82.81**	57.93

Multi-label classification on Fashion144k [1] and Fashion550k [2]

Results



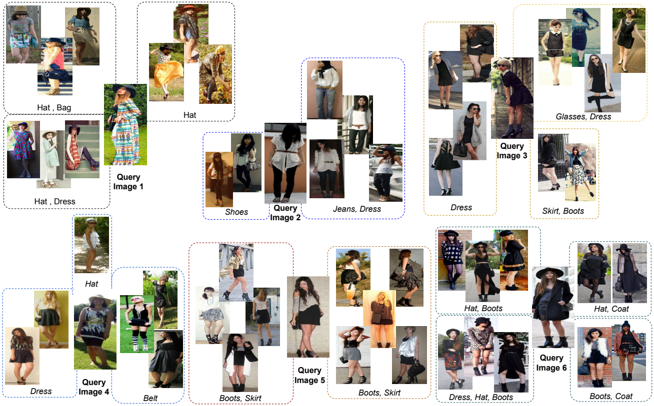
(a) In-Shop retrieval



(b) Consumer-to-shop retrieval

Retrieval results for In-shop and Consumer-to-shop retrieval tasks on DeepFashion dataset [4].

Results



Semantically similar results for some of the query images from Fashion144k dataset [1] using our method.

Conclusion

- Using clothing items for recommendation gives much variability in the recommendation results.
- Attention can be used to learn discriminative features from weak labels.
- Texture cues are good for learning different parts.

References

E. Simo-Serra and H. Ishikawa,
"Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction,"
in *CVPR*, 2016, pp. 298–307.

Naoto Inoue, Edgar Simo-Serra, Toshihiko Yamasaki, and Hiroshi Ishikawa,
"Multi-label fashion image classification with minimal human supervision,"
in *ICCVW*, 2017, pp. 2261–2267.

Andreas Veit et al.,
"Learning from noisy large-scale datasets with minimal supervision,"
in *CVPR*, 2017.

Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang,
"Deepfashion: Powering robust clothes recognition and retrieval with rich annotations,"
in *CVPR*, 2016, pp. 1096–1104.

Thank you
Questions?