TÉCNICO LISBOA

Institute for Systems and Robotics Lisboa

FARFETCH

Carnegie Mellon University

ICCV 2019 Seoul, Korea

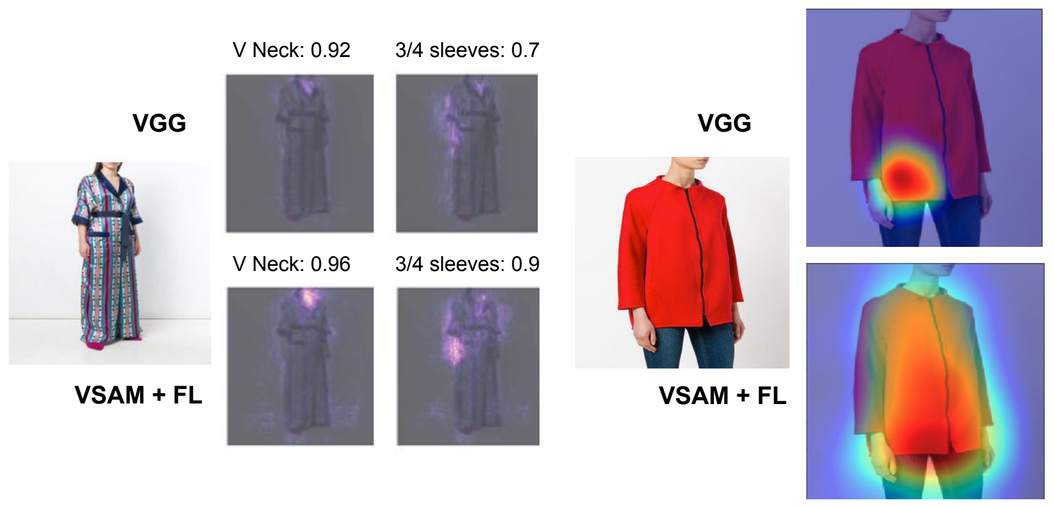
2nd Workshop on Computer Vision for Fashion, Art and Design

Pose Guided Attention for Multi-label Fashion Image Classification

Beatriz Quintino Ferreira[1], João P. Costeira[1], Ricardo G. Sousa[2], Liang-Yang Gui[3], João P. Gomes[1]

[1]ISR-IST, Universidade de Lisboa, [2]Farfetch, [3]Carnegie Mellon University

Abstract: We propose a compact framework with guided attention for **multi-label** classification in the fashion domain. Our **visual semantic attention model (VSAM)** is **supervised by automatic pose extraction** creating a discriminative feature space. VSAM **outperforms the state of the art for an in-house dataset** and **performs on par with previous works on the *DeepFashion*** dataset, **even without using any landmark annotations**. Additionally, we show that our semantic attention module **brings robustness to** large quantities of **wrong annotations** and **provides more interpretable results**.



The effect of pose guided attention. Saliency and class activation maps highlight most relevant regions

Introduction

- **Fashion attributes are associated with specific locations** (e.g. sleeve, neckline)

Purely data driven approaches disregard this fact and are less robust

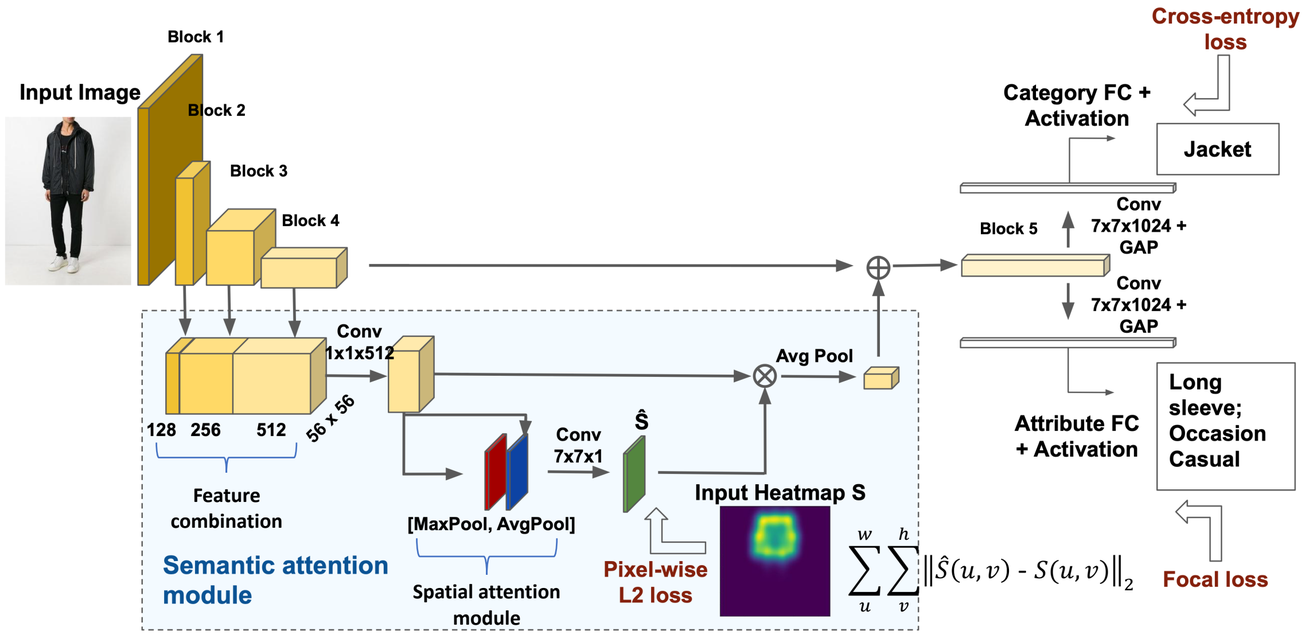
Other approaches require an expensive annotation process, impractical at large scale

- We exploit the **relation between attribute localization** and **visual appearance**, by **embedding a semantic attention** module **guided by body pose**

Proposed model - VSAM

- **Task:** predict category C (multi-class) and attribute vector A (multi-label) for each image
- The attention mechanism acts on a feature combination scheme and is supervised by the pose of the model wearing the clothing item

Examples of pose detections (by *OpenPose*) and respective heatmaps, used during training



Proposed Visual Semantic Attention Model (VSAM) architecture with the semantic attention mechanism regularization

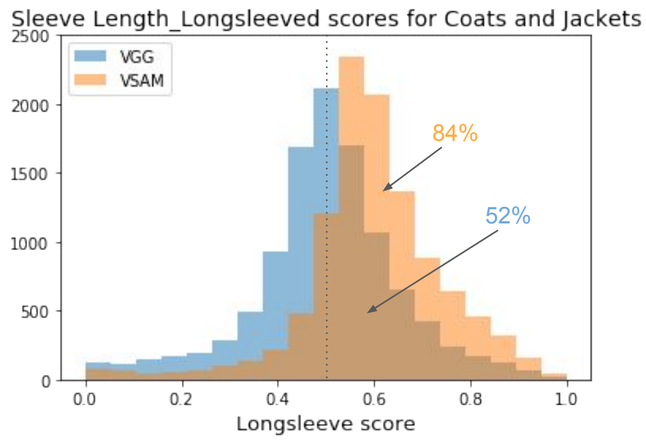
Experiments and Results

Datasets:

- "In-house" dataset, with approx. 245k images with 17 categories and 53 attributes, and avg. 1.2 attributes annotations per product
- *DeepFashion* dataset, with approx. 289k images with 50 categories and 1000 attributes

Method/Metric	P@k	R@k	F1@k	AP
Model from [1]	73.02	72.56	73.33	69.17
VSAM + FL	**80.70**	**81.56**	**80.63**	**75.44**

Quantitative results for multi-label classification for the "in-house" dataset



VSAM increases model robustness to large quantities of noisy annotations

Method/ Metric	Texture		Fabric		Shape		Part	
	Top-3	Top-5	Top-3	Top-5	Top-3	Top-5	Top-3	Top-5
Model from [1]	44.39	53.91	31.82	41.70	39.88	50.51	31.11	40.76
Fashion Net [2]	37.46	49.52	39.30	49.84	39.47	48.59	44.13	54.02
Liu et al. [3]	56.30	65.82	43.05	53.64	58.75	67.80	46.47	57.39
VSAM + FL	56.28	65.45	41.73	52.01	55.69	65.40	43.20	53.95

Quantitative results for multi-label classification for the *DeepFashion* dataset. VSAM does not use any landmark annotations

Contributions:

- Creating **complete** and **consistent fully labeled fashion datasets** requires tremendous **effort** and is **extremely expensive**
- We take advantage of ***a priori* cues** to look for details in images
- In spite of its **much lower complexity**, **VSAM outperformed** by a large margin previous work on **multi-label [1]**, and **performs on par with state-of-the-art** methods for **DeepFashion** even **without using landmark annotations**
- **VSAM** was **robust to wrong annotations** and provides **more meaningful visualizations and interpretability**

*This work was partially funded by FCT via grant [PD/BD/114430/2016] and project [UID/EEA/50009/2019]

References: [1] B. Quintino Ferreira, L. Baía, J. Faria, and R. Sousa. "A Unified Model with Structured Output for Fashion Images Classification". In *KDD'18 Workshop on AI for Fashion*, 2018.
[2] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. "DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations". In *CVPR*, 2016.
[3] J. Liu and H. Lu. "Deep fashion analysis with feature map up-sampling and landmark-driven attention". *ECCV Workshops on Computer Vision for Fashion, Art and Design, 2018*.