A Semi-supervised Deep Learning Approach Based on Attentive Visual Features for Fashion Clothing Classification

Shajini Majuran

ABSTRACT

Fashion clothing analysis encompasses spotting and identifying items and other characteristics of clothes in an image or a video. Earlier fashion analysis models mostly relied on hand-crafted features to search for powerful clothing representations such as graph models, contextual information, general object proposals, human parts, bounding boxes, and semantic masks. At present, an evolving technique for clothing related tasks is the deep neural networks that simultaneously learns features and classifies clothing-related images. In this study, the contributions are four-fold based on convolutional neural network (CNN) approaches: (i) Landmark-driven model, (ii) landmark-free model, (iii) semi-supervised learning architecture, and (iv) clothing detector for fashion clothing classification. The proposed landmark-driven technique aims to achieve clothing classification by producing regularised landmark layouts incorporating two attention pipelines. The proposed *landmark-driven network* extracts features accompanying a coarse-to-fine view in scale to effectively localise clothing landmarks and the spatial-channel attention pipeline prioritises the prominent features by identifying and giving peak to the feature values wherever they exist in an input image. The proposed *landmark-free network* facilitates multiscale contextual feature supervision through combining lower and higher levels of extracted features from the network. Generally, CNN embodies the semantic gap between lower and higher levels of architecture, where the clothing classification requires both spatial and semantic representation. Therefore, a multistage network is introduced to combine different stages of extracted features along with a semantic feature-based attention to strengthen the feature representation for multitask learning. In addition, a *teacher-student collaborative* (i.e., semi-supervised) *learning architecture* is proposed to make use of labelled and unlabelled samples together. The annotation of large-sized clothing imageset requires more time and man power, thus it is an expensive process. The teacher model learns the mapping from labelled data and the student model learns from both the mapping and the pseudo labelled data for weighted minimisation. Moreover, locating a clothing item in an image for further analysis becomes a complex task, hence a *single-stage attention-based network* is proposed for fashion clothing detection and classification. The proposed detector benefits by adopting multilevel contextual features through multistage architecture, so that the semantic gap between the lower and upper-level features from different levels of feature representation is better addressed. Furthermore, the classification and detection branches maintain less number of trainable parameters, thus the model shows efficiency and have faster inference time. The proposed methods are evaluated on DeepFashion-C, Fashion Landmark Detection (FLD), and DeepFashion2 benchmark datasets. In summary, this study suggests that a landmark-free model with semi-supervised learning setup is highly plausible for fashion clothing classification and attribute prediction. The results show 93.31 and 97.15 top-3 and top-5 accuracies for classification and 57.29 and 66.14 top-3 and top-5 recall rate for attribute prediction, respectively. This study further highlights the generalisation and scalability of fashion models for fashion analysis to enforce discriminatory aspect in feature representation with consistent performance.