논문 20-45-07-05 The Journal of Korean Institute of Communications and Information Sciences '20-07 Vol.45 No.07
https://doi.org/10.7840/kics.2020.45.7.1174

딥러닝을 이용한 시각적 의류 분석 기술: 서베이

이 수 민◆, 오 성 찬*, 정 찬 호°, 김 창 익**

Visual Fashion Analysis Using Deep Learning: A Survey

Sumin Lee◆, Sungchan Oh*, Chanho Jung°, Changick Kim**

요 약

전자상거래가 증가하고 온라인 쇼핑 산업이 발달함에 따라 대량의 의류 이미지의 분석과 관리에 대한 필요성이 커지고 있다. 이러한 경향에 따라, 시각적 의류 분석 기술이 크게 주목을 받고 있다. 시각적 의류 분석에는 옷의 카테고리와 특성정보를 파악하는 의류 인식, 같은 의류를 찾는 의류 검색 등이 있다. 일반 물체 영상과 달리, non-rigid 특성을 가지는 의류는 영상 안에서 변형과 겹침이 빈번하게 발생한다. 알고리즘 적용을 까다롭게 만드는 이러한 특성을 극복하기 위해 위한 다양한 연구들이 활발하게 이루어지고 있다. 최근에는 대량 의류 데이터 셋의 등장과 딥러닝의 발달로, 딥러닝 기반의 의류 분석 기술들이 활발히 연구되고 있으며 이는 기존 방법들과 비교 했을 때 상당한 성능 향상을 이루었다. 이러한 연구 동향에 따라 본 논문에서는 최근 주목받고 있는 딥러닝 기반의 시각적 의류 분석 방법들을 중 의류 인식 기술과 의류 검색 기술에 대해서 소개하고자 한다.

Key Words : Visual Fashion Analysis, Deep-learning, Fashion Recognition, Fashion Retrieval

ABSTRACT

Due to the huge potential in the industry, understanding fashion images has driven a lot of attention. The visual fashion analysis techniques are used in various ways, such ais fashion recognition, fashion retrieval. Unlike general object images, clothes with non-rigid properties usually suffer from a deformation and occlusion in a image. Since the non-rigid characteristic makes hard to apply algorithms to fashion images, various research have been actively conducted to overcome this problem. Recently, with an advent of large-scale datasets and a development of deep learning, diverse fashion analysis methods based on deep-learning are introduced, which achieved huge performance improvement. In this paper, we introduce fashion recognition and fashion retrieval methods among prominent deep learning-based visual fashion analysis methods.

Ⅰ. 서 론

최근 전 세계적으로 온라인 쇼핑 산업이 발달함에 따라 의류 영상 분석을 위한 기술에 대한 관심이 높아지고 있다. 의류 영상 분석에는 이미지 속 옷에 대한 정보를 분석하는 의류 인식 (Fashion Recognition), 이미지로 원하는 옷을 찾을 수 있는 의류 검색 (Fashion Retrieval) 등이 있다. 이러한 의류 영상 분석 기술들은 사용자가 직접 촬영한 사진 속에 존재하는 옷과 같거나 유사한 옷을 찾아주거나, 어울리는 옷

◆ First Author : School of Electrical Engineering, Korea Advanced Institute of Science and Technology, suminlee94@kaist.ac.kr, 정회원
° Corresponding Author : Department of Electrical Engineering, Hanbat National University, peterjung@hanbat.ac.kr, 정회원
* Electronics and Telecommunications Research Institute, sungchan.oh@etri.re.kr, 선임연구원
** School of Electrical Engineering, Korea Advanced Institute of Science and Technology, changick@kaist.ac.kr
논문번호 : 202004-080-A-RN, Received April 7, 2020; Revised May 20, 2020; Accepted May 20, 2020

을 추천 해주는 서비스와 같이 실제 산업에서 다양하게 활용되고 있다. 예를 들어 LG전자에서는 사용자의 일정에 적합한 옷을 추천해주고 가상 피팅을 할 수 있는 제품을 개발하고 있으며, 네이버나 인터파크와 같은 온라인 쇼핑 사이트에서는 이미지로 의류 상품을 검색할 수 있는 서비스를 제공하고 있다.

일반적인 물체와 달리, 의류는 non-rigid 특성을 가지기 때문에 영상 안에서 변형(deformation)과 겹침(occlusion) 현상이 빈번하게 발생한다. Non-rigid 특성은 의류 영상에 대한 알고리즘 적용을 까다롭게 만들기 때문에 이를 극복하기 위한 연구들이 진행되고 있다. 연구 초반에는 SIFT, HOG와 같은 수제 특징(Handcrafted Feature)을 사용한 기술들[8,9]이 제안되었으나, 최근에는 대량 데이터셋의 등장과 딥러닝의 발달로 딥러닝 기반의 방법들이 활발하게 연구되고 있다. 딥러닝 기반으로 전환되면서 성능이 크게 향상되었을 뿐만 아니라, 기술에 대한 활용 분야가 넓어졌다. 이에 따라 본 논문에서는 최근 주목받고 있는 딥러닝 기반의 의류 분석 기술 중에서도 의류 인식과 의류 검색 기술에 대해서 소개한다.

본 문의 구성은 다음과 같다. 2장에서는 의류 인식 기술에 대해, 3장에서는 의류 검색 기술에 대해서 설명한다. 4장에서는 시각적 의류 분석 기술의 연구 방향과 전망을 살펴본다.

Ⅱ. 의류 인식

의류 인식은 주어진 영상 안에 존재하는 옷의 카테고리와 특성정보를 분석하는 것을 목표로 한다 (그림 1). 의류 카테고리에는 Shirt, Skirt, Dress, Pants 등이 있고, 특성정보는 V-neck, Jean, Lace, Print 등을 예로 들 수 있다. 대부분의 경우 데이터셋에서 주어지는 바운딩 박스(Bounding Box)를 이용하여 해당 옷만 잘라 내어 (Cropping) 의류 인식을 진행한다.

그림 1. 해당 의류의 카테고리와 특성 정보를 파악하는 의류 인식의 예시
Fig. 1. An example of fashion recognition, which aims to predict a category and attributes of given clothes

2.1 FashionNet[1]

FashionNet[1]은 처음으로 의류 인식에 컨볼루션 뉴럴 네트워크 (Convoltuion Neural Network)를 적용한 방법으로, 의류 인식의 벤치마크 (Benchmark)를 제안하였다. 특히 의류의 non-rigid 특성을 극복하기 위해서 의류 랜드마크 (Fashion Landmark)를 처음 도입하여 의류 non-rigid 특성을 극복하고자 하였다. 의류 랜드마크는 소매나 칼라, 허리 라인 등과 같이 옷의 구조적으로 중요한 부분들에 해당된다.

FashionNet은 VGG-16[4] 네트워크에 기반하며, VGG-16 네트워크의 마지막 컨볼루션 레이어를 세 개의 분기로 대체한 구조를 가진다. 의류 랜드마크를 국소화 하기 위한 1) 랜드마크 분기와, 전역 특징 (Global Feature) 추출을 위한 2) 전역 분기, 그리고 국소화된 랜드마크 정보를 활용하여 지역 특징 (Local Feature) 추출을 위한 3) 지역 분기가 그것이다. 전역 특징 추출 분기는 컨볼루션 레이어와 완전 연결레이어 (Fully Connected Layer)로 이루어져 있으며 입력 영상에 전체에 컨볼루션이 적용되어 전역 특징을 추출하게 된다. 랜드마크 국소화 분기에서는 회귀 (Regression) 방식으로 랜드마크 좌표를 계산하고, 해

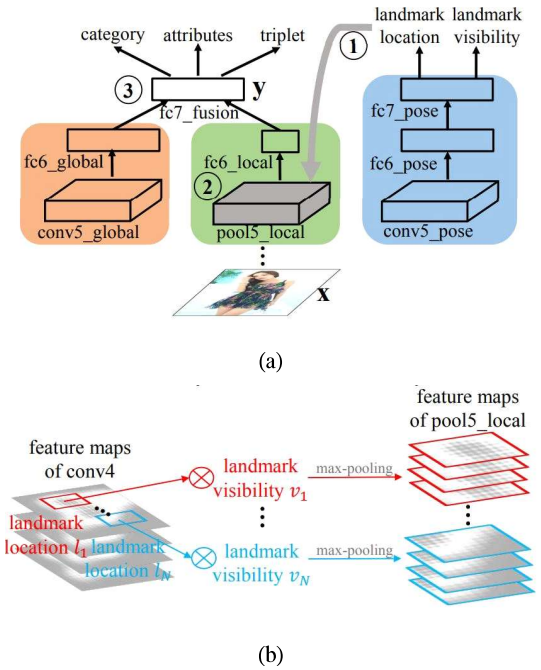

그림 2. (a) FashionNet의 구조 [1], (b) 랜드마크 풀링 (Landmark Pooling) 레이어[1]
Fig. 2. (a) The architecture of FashionNet[1], (b) Schematic illustration of landmark pooling layer[1]

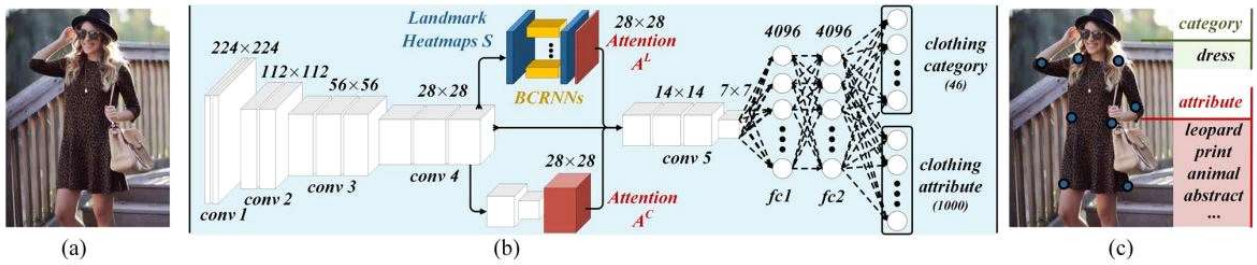

그림 3. Attentive Fashion Grammar Network의 도식화[3]
Fig. 3. Illustraction of the Attentive Fashion Grammar Network[3]

당 랜드마크의 가시성 (Visibility)를 측정한다. 지역 특징 추출 분기에서는 랜드마크 분기에서 측정된 랜드마크 좌표에 랜드마크 풀링 (Landmark pooling)을 적용하여 랜드마크 정보를 추출한다. 랜드마크 풀링은 가시성을 반영하여 랜드마크 좌표를 중심으로 최대 풀링 (Max Pooling)을 하는 방식이다. FashionNet은 마지막 합성 완전 연결 레이어에서 전역 특징 정보와 지역 특징 정보를 활용하여 카테고리와 특징 정보를 분류하게 된다.

카테고리 분류를 위해서는 일반적인 교차 엔트로피 손실함수 (Cross-Entropy loss) $L_{category}$를 사용하였으며, 다중 레이블인 특성정보 분류를 위해서는 가중 교차 엔트로피 (Weighted Cross-Entropy) 손실함수 L_{attr}를 사용하였다. 이 때, $\boldsymbol{x}_j$, $\boldsymbol{c}_j$, $\boldsymbol{a}_j$는 각각 j번째 의류 영상, 카테고리 레이블, 특성정보 레이블이다.

$$L_{category} = \sum_j \left(\begin{matrix} \boldsymbol{c}_j \log p(\boldsymbol{c}_j|\boldsymbol{x}_j) + \\ (1-\boldsymbol{c}_j)\log(1-p(\boldsymbol{c}_j|\boldsymbol{x}_j)) \end{matrix} \right) \quad (1)$$

$$L_{attr} = \sum_j \left(\begin{matrix} w_{pos}\boldsymbol{a}_j \log p(\boldsymbol{a}_j|\boldsymbol{x}_j) + \\ w_{neg}(1-\boldsymbol{a}_j)\log(1-p(\boldsymbol{a}_j|\boldsymbol{x}_j)) \end{matrix} \right) \quad (2)$$

랜드마크 국소화를 위해서는 랜드마크 좌표에 대한 $l2$ 손실함수 $L_{landmark}$가 사용되었다.

$$L_{landmark} = \sum_j \| \boldsymbol{v}_j \circ (\hat{l}_j - l_j) \|_2^2 \quad (3)$$

FashionNet은 의류 랜드마크와 컨볼루션 뉴럴 네트워크를 사용하여 기존 handcrafted feature를 사용하는 방법들[8,9]와 비교했을 때 성능 향상을 이루었지만, 랜드마크가 회기 방법으로 국소화 된 점과 랜드마크 구성의 선수 지식 (Prior Knoweldge)를 활용하지 못했다는 단점이 있다. 랜드마크를 회기 방법으로 국소화 할 경우 비선형 문제가 되기 때문에 정확한 학습이 어렵다[15].

2.2 Attentive Fashion Grammar Network[3]

Attentive Fashion Grammer Network[3]은 랜드마크 인지 어텐션 (Landmark-aware Attention)과 카테고리 중심 어텐션 (Category-driven Attention)을 활용한 의류 인식 방법이다. 이 역시 FashionNet과 마찬가지로 VGG-16 모델에 기반하고 있으며 Conv.4 특징맵 두 종류의 어텐션을 적용한 뒤 Conv.5 컨볼루션과 완전연결 계층을 통해 카테고리와 특성 정보를 추정하는 방식이다.

랜드마크 인지 어텐션 생성을 위해서 패션 문법 (Fashion Grammar)과 순환 신경망 (Recurrent Neural Net)으로 이루어진 Bidirectional Convoltuional Recurrent Neural Network (BCRNN)가 제안되었다. 비선형적이고 학습이 어려운 회기 방식 대신, BCRNN은 랜드마크 위치 분포의 신뢰도를 측정하여 heatmap을 생성한다. 랜드마크에 대한 4개의 동적 (Kinematic) 문법과 4개의 대칭 (Symmetry) 문법을 정의하였고 이를 순환 신경망을 이용한 메시지 패싱 (Message Passing)으로 구현하였다. 메시지 패싱이 포함된 BCRNN은 Conv.4 특징맵 F_c에 대해서 정밀한 랜드마크 인지 어텐션 A^L텐션은 예측한다.

카테고리 중심 어텐션 A^C은 Conv.4의 특징맵 F_c에 대하여 Bottom-up Top-down 방식[16]의 네트워크를 적용하여 생성된다. 각각의 방법으로 생성된 두 가지 어텐션은 식 (4)과 같이 Conv.4의 특징맵에 적용되어 특징맵이 의류에 특화된 풍부한 정보를 담을 수 있게 한다. 이 때, n_{F_C}는 F_c의 채널 수를 의미한다.

$$G_c = \left(1 + A^L + A_c^C\right) \circ F_c,\ c \in \{1, ..., n_{F_C}\} \quad (4)$$

카테고리와 특성정보 학습에 사용된 손실함수는 FashionNet과 동일하며, 랜드마크 학습을 위해서 랜

드마크 히트맵에 대한 $l2$ 손실함수가 사용되었다.

FashionNet[1]은 처음으로 의류 인식에 컨볼루션 뉴럴 네트워크 (Convoltuion Neural Network)를 적용한 방법으로, 의류 인식의 벤치마크 (Benchmark)를 제안하였다. 특히 의류의 non-rigid 특성을 극복하기 위해서 의류 랜드마크 (Fashion Landmark)를 처음 도입하여 의류 non-rigid 특성을 극복하고자 하였다. 의류 랜드마크는 소매나 칼라, 허리 라인 등과 같이 옷의 구조적으로 중요한 부분들에 해당된다.

2.3 Leveraging weakly annotated data[12]

[12]는 온라인 쇼핑몰 사이트에서 크롤링한 정보를 수동 레이블링을 거치지 않고 바로 학습에 사용하는 방법을 제안하였다. 온라인 사이트에서 크롤링한 정보는 완벽하지 않기 때문에 weakly annotated 영상을 가지고 학습을 시켰다.

잡음이 포함된 레이블을 활용하기 위해서 균일 샘플링을 사용하였다[17]. 학습 시, 단어 셋 (Vocabulary)에서 단어 w를 균일하게 샘플링한다. 다음 bag-of-words에 단어 w가 포함된 이미지 x를 무작위로 선택하여 모델을 학습시킨다. 영상 하나에서 단어 하나를 예측하도록 학습시키기 위해서 교차 엔트로피 손실 함수를 사용하였다. 이미지 임베딩을 위해서는 Resnet50[18]가 사용되었고 임베딩 된 단어 특징 벡터를 곱하여 의류 인식을 진행한다. 이 때, 단어 특징 벡터는 이미지 특징 벡터에 대해서 어텐션 기능을 하게 된다.

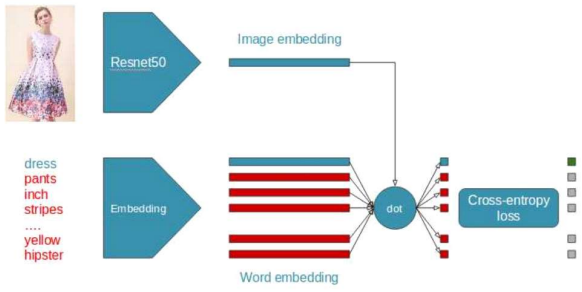

그림 4. [12]의 제안하는 방법
Fig. 4. The illustration of propose model at [12]

2.4 비교

2.4.1 데이터셋

의류 인식 연구에서는 Deepfashion-C[1]을 벤치마크 (Benchmark) 데이터셋으로 활용한다. Deepfashion-C 데이터셋은 카테고리와 특성정보 레이블, 바운딩 박스, 그리고 랜드마크 정보를 제공한다. 총 282,000개의 의류 영상은 202,000장의 학습 영상과, 40,000장의 확인 영상, 40,000장의 테스트 영상으로 구성되어 있다. 카테고리는 Tee, Shirt, Pants, Dress 등 총 50개가 있으며, 특성정보는 V-neck, Shiffon, Jean, Lace, Print 등 1000개가 존재한다. 각 영상은 하나의 의류 아이템에 대한 어노테이션 (Annotation)을 제공하며, 하나의 아이템은 여러 개의 특성정보를 가질 수 있다. 평균적으로 3.38개의 성정보를 가진다. 1000개의 특성정보는 Texture, Fabric, Shape, Part, 그리고 Style의 다섯 개의 소분류로 나뉜다.

Deepfashion-C에서 제공하는 랜드마크는 좌우 소매, 좌우 칼라, 좌우 허리 라인, 좌우 옷 끝, 총 8 종류가 있다. 옷 종류에 따른 랜드마크 개수는 표 1과 같다. 옷이 모양 변형과 겹침이 심하기 때문에, 각 랜드마크의 가시성에 대한 정보가 함께 제공된다.

표 1. Deepfashion-C 데이터셋의 옷 종류에 따른 의류 랜드마크 갯수
Table 1. The number of fashion landmark with different clothes type in Deepfashion-C

Clothes type	Upper	Lower	Full-body
# of fashion landmark	6	4	8

2.4.2 성능

카테고리 분류의 평가지표로는 top-1, top-3 정확도 (Accuracy)를 사용하며, 특성정보 예측 평가지표로는 top-1, top-3 회수율 (Recall Rate)을 사용한다.

특성정보는 각 소분류 클래스들에 대해서 성능을 측정하고 전체 클래스에 대해서 측정하였다.

표 2은 수제 특징 기반의 방법들[8,9]과 앞서 소개하였던 방법들의 Deepfashion-C 데이터셋에 대한 성능 비교 표이다. 대부분의 항목에서 Attentive Fashion Grammar Network[3]이 높은 성능을 나타내고 있으며, 특성정보의 'All' 성능이 크게 향상되었다.

Ⅲ. 의류 검색

이미지 회수 (Image Retrieval)은 쿼리 (Query) 이미지에 대해서 같은 ID를 가지는 이미지를 갤러리 (Gallery) 이미지셋에서 찾는 테스크이다. 의류 검색 또한 이미지 회수에 속하며, 소비자가 촬영한 이미지와 비교적 제한된 환경에서 촬영된 매장 이미지를 매칭하는 테스크이다 (그림 5). 소비자가 촬영한 이미지가 쿼리에 해당하며 매장 이미지는 갤러리에 해당한다. 소비자가 촬영한 이미지는 겹침과 잘림이 심할 뿐

표 2. Deepfashion-C 데이터셋에 대한 카테고리 분류와 특성정보 예측 성능. 가장 높은 값은 진하게 표시됨
Table 2. Quantitative results for category classification and attribute prediction on the Deepfashion-C dataset. The best score are marked in bold.

Method	Category		Texture		Fabric		Shape		Part		Style		All	
	top-3	top-5	top-3	top-5	top-3	top-5	top-3	top-5	top-3	top-5	top-3	top-5	top-3	top-5
WTBI[8]	43.73	66.26	24.21	32.65	25.38	36.06	23.39	31.26	26.31	33.24	49.85	58.68	27.46	35.37
DARN[9]	59.48	79.58	36.15	48.15	36.64	48.52	35.89	46.93	39.17	50.14	66.11	71.36	42.35	51.95
FashionNet[1]	82.58	90.17	37.46	49.52	39.90	**49.84**	39.47	48.59	**44.13**	54.02	66.43	73.16	45.52	54.61
Want et al.[3]	**90.99**	**95.78**	50.31	**65.48**	**40.31**	48.23	**53.32**	**61.05**	40.65	**56.32**	**68.70**	**74.25**	**51.53**	**60.95**
Corbiere et al. [12]	86.30	92.80	**53.60**	63.20	39.10	48.80	50.10	59.50	38.80	48.90	30.50	38.30	23.10	30.40

만 아니라 그 외 조명 등과 같은 촬영 환경에 의한 노이즈 (Noise)가 심하다. 반면 매장 이미지는 조명이나 배경이 제한된 환경에서 촬영하여 깨끗하고 선명한 영상이며, 소비자가 촬영한 영상보다 겹침이나 변형 현상이 적다.

그림 5. 의류 검색의 예시. 검은색 테두리의 이미지는 사용자가 촬영한 입력영상이고 오른쪽 이미지는 매장 영상의 검색 결과이며, 녹색 테두리는 입력 영상과 같은 의류의 이미지[1]
Fig. 5. Examples of fashion retrieval. Consumer-taken images with black lines are input images, and retrieved in-shop images are on the right. Images with green line are correct results[1]

3.1 GRNet[13]

GRNet(Graph Reasoning Network)[13]은 쿼리와 갤러리 이미지 사이의 유사성 (Similarity)을 여러 스케일 (Scale)의 전역 특징과 지역 특징을 사용하여 학습한 방법이다 (그림 6). 전역 특징이 지역 특징을 포함한다고 가정했던 이전 방법들과 달리, 지역 특징을 따로 분리하면서 지역 특징이 최적화가 되지 않을 수 있는 문제를 해결하였다. 또한, 여러 스케일의 지역 특징과 전역 특징을 그래프 추론 (Graph Reasoning)을 통해 유사성을 측정함으로써 의류의 디테일한 부분에 대한 정보를 활용할 수 있도록 하였다.

GRNet은 먼저 쿼리와 갤러리 입력 영상에 대하여 피라미드 공간 윈도우 (Pyramid Spatial Window)를 이용하여 여러 스케일의 피쳐를 추출한다. 쿼리와 갤러리의 피라미드 스케일 l의 i번째 지역 특징을 각각 $\boldsymbol{x}_l^i$, $\boldsymbol{y}_l^i$라고 했을 때, 스케일이 같은 특징들끼리 비교하여 다음과 같이 유사성 S_p을 측정한다.

$$S_p(\boldsymbol{x}_l^i, \boldsymbol{y}_l^i) = \frac{\boldsymbol{P}\left|\boldsymbol{x}_l^i - \boldsymbol{y}_l^i\right|^2}{\left\| \boldsymbol{P}\left|\boldsymbol{x}_l^i - \boldsymbol{y}_l^i\right|^2 \right\|_2} \tag{5}$$

그래프 노드 (Node)에 해당되는 각 스케일의 유사성 $\boldsymbol{s}_{l_1}^{ij}$과 $\boldsymbol{s}_{l_2}^{mn}$를 사용하여 두 노드 간 스칼라 엣지 $w_p^{l_1 ij,\, l_2 mn}$은 다음과 같이 측정한다.

$$w_p^{l_1 ij,\, l_2 mn} = \frac{\exp\left(\left(\boldsymbol{T}_{out}\boldsymbol{s}_{l_1}^{ij}\right)^{\top}\left(\boldsymbol{T}_{\in}\boldsymbol{s}_{l_2}^{mn}\right)\right)}{\sum_{l.p.q}\exp\left(\left(\boldsymbol{T}_{out}\boldsymbol{s}_{l_1}^{ij}\right)^{\top}\left(\boldsymbol{T}_{\in}\boldsymbol{s}_{l_2}^{pq}\right)\right)} \tag{6}$$

이때, $\boldsymbol{T} \in R^{D \times D}$는 선형 변형을 위한 매트릭스이다.

노드 값과 엣지 값을 이용하여 다음과 같이 유사성 추론 (Similarity Reasoning)을 수렴할 때까지 반복적으로 적용한다.

$$\hat{\boldsymbol{s}}_{l_1}^{ij} = \sum_{l_2, m, n} w_p^{l_1 ij,\, l_2 mn}\boldsymbol{s}_{l_2}^{mn} \tag{7}$$

GRNet은 전역 특징과 지역 특징을 그래프 추론을 활용하여 의류 검색 문제를 해결하여 성능 향상을 이끌었다. 하지만 비교적 작은 특징이 존재하기 어려운 외투 (Outerwear)에 대해서는 성능 저하가 일어나는

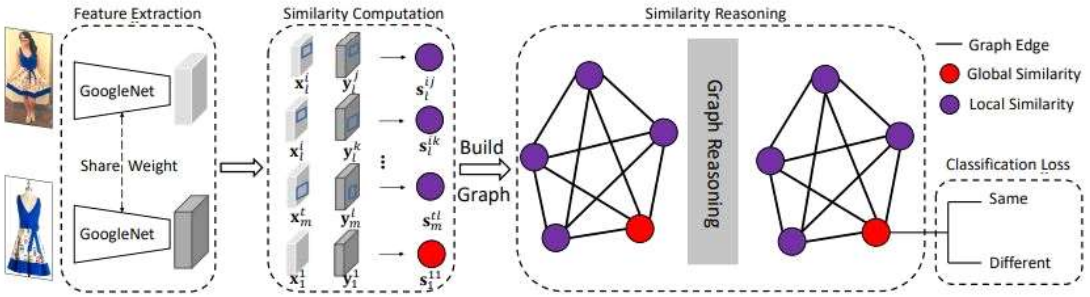

그림 6. GRNet의 도식화[13]
Fig. 6. Illustration of the GRNet[13]

등 지역 특징을 충분히 활용하기 힘든 경우에 대해서 한계점을 보인다.

3.2 FLAM[14]

FLAM(Feature-level Attribute Manipulation)[14]은 의류 이미지 검색 (Fashion Image Retrieval, FIR)과 의류 특성 조작 (Fashion Attribute Manipulation, FAM)을 동시에 학습시킴으로써 특징 벡터의 정보 수용량 (Capacity)를 증가시켜 성능 향상을 이끌어낸 방법이다 (그림 7).

FLAM는 임베딩 단계와 쿼리 특성 조작 단계로 나누어진다. 먼저, 임베딩 단계에서는 FIR로 충분히 학습된 특징 추출기에서 특징 x을 추출한다. 특성 정보 임베더 (Attribute-specific Embedder) ϕ_a로 특징 x를 특성 별 임베딩 공간 (Attribute-specific Embedding Space)에 임베딩한다. 이 때 x에 대한 특성 정보를 학습하기 위해서 특성 정보 사전 d_a (Attribute Dictionary)를 사용한다. 특성 정보 사전은 임베딩 공간이 Shape, Color, Pattern에 대한 정보를 나타낼 수 있게 하고, 임베딩 된 특징 벡터와 관련 있는 특성 정보와는 가까이, 관련이 없는 정보와는 멀리 존재할 수 있도록 한다. 이 특성 정보 사전은 학습 초기 랜덤하게 설정되며, 학습 가능한 파라미터이다. 임베딩 공간에서 특징 벡터의 거리 학습을 위해서는 Triplet 손실 함수 L_{trip}가 사용된다. 이 때, x와 다른 특성 정보를 가지는 특성 벡터를 x^-라고 한다.

$$L_{trip} = \max\{0, dist(x, x^+) - dist(x, x^-) + \mu\} \quad (8)$$

다음 쿼리 특성 조작 단계에서는 GAN (Generative Adversarial Network)이 사용된다. 특징 벡터 페어 (x, x^-)에 대해서 생성기 (Generator) G는 다음과 같이 x^-의 타겟 특징을 가지는 특징 벡터 $\tilde{x}$를 생성한다.

$$G(x, \phi_a(x^-)) \rightarrow \tilde{x} \quad (9)$$

특징 벡터의 합성 여부를 판단하는 판별기 D와 합성 특징 벡터 $\tilde{x}$에 대하여 다음과 같이 적대적 손실함수 (Adversarial Loss) L_{adv}가 적용된다.

$$L_{adv} = E_x[\log D(x)] + E_{x,x^-}[\log(1 - D(G(x, \phi_a(x^-))))] \quad (10)$$

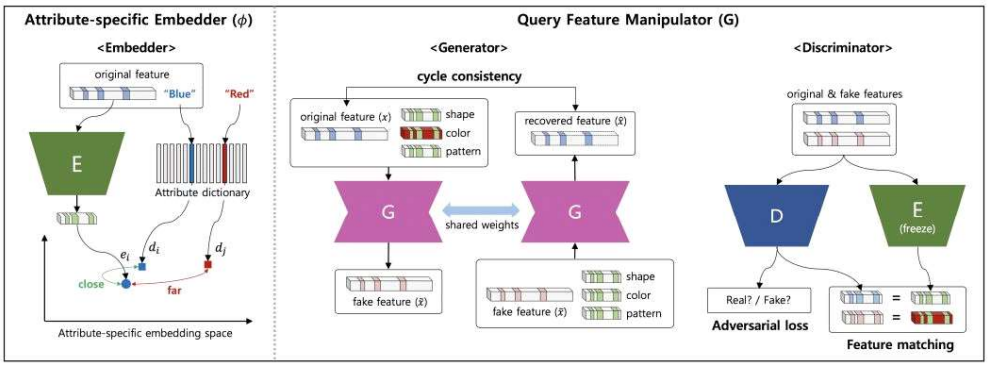

그림 7. FLAM 방법의 도식화[14]
Fig. 7. Illustration of the FLAM method[14]

또한 생성된 $\tilde{x}$가 $\phi_a(x)$로 다시 x로 변환 될 수 있도록 하는 순환 일관성 손실함수와, 특성 정보에 대한 매칭 손실함수가 사용된다.

특성 정보를 조작하여 합성 이미지를 생성하는 것은 ill-posed 문제가 발생하는데, FLAM은 이미지가 아닌 특징 벡터 단에서 조작을 수행하여 그러한 단점들을 극복하였다. 다만, FIR로 미리 학습된 특징 추출기를 사용해야하기 때문에 End-to-End 학습이 불가하다는 단점이 있다.

3.3 비교

3.3.1 데이터셋

의류 인식 연구에서는 Deepfashion-C2S[1]을 벤치마크 데이터셋으로 사용한다. Deepfashion-C2S 데이터셋은 의류 아이템의 고유 아이디와 바운딩 박스, 옷 종류 (상의, 하의, 전신), 그리고 소스 종류 (사용자 촬영, 매장 촬영) 정보를 제공한다. 총 33,881개의 아이템에 대해서 239,557개의 의류 영상을 제공한다.

3.3.2 성능

의류 검색의 평가지표로는 Top-k 회수율 (Recall Rate)을 사용하였다. Top-k 회수율은 주어진 쿼리 영상에 대해서 가장 유사도가 높은 k개의 갤러리 이미지들을 선별하고, 선별된 이미지 중에서 쿼리 영상과 ID가 일치하는 경우를 측정한다. 수제 특징을 사용한 WTBI[8]와 DARN[9]보다 GRNet과 FLAM 방법은 상당한 성능 향상을 얻었다. 특히 FLAM 방법이 가장 높은 성능을 달성하였다.

표 3. Deepfashion-C2S 데이터셋에 대한 의류 검색 성능 평가. 가장 높은 값은 진하게 표시됨
Table 3. Quantitative results for fashion retrieval on the Deepfashion-C2S dataset. The best score are marked in bold.

	R@1	R@20	R@50
WTBI[8]	0.24	0.63	0.87
DARN[9]	0.36	1.11	1.52
GRNet[13]	25.7	64.4	75.0
Shin et al. [14]	26.5	66.4	75.5

Ⅳ. 결 론

본 논문에서는 딥러닝을 이용한 시각적 의류 분석 기술들 중 가장 주목받고 있는 의류 인식과 의류 검색 기술들에 대해서 소개하였다. 딥러닝의 도입은 수제 특징을 사용하였던 이전 방법들에 비해서 모델의 의류 이해 능력을 향상시켜서 높은 성능을 달성하였고, 실제 산업에서 활용될 수 있도록 하였다.

이미지에서 의류 정보를 추출하는 의류 인식과 대량의 이미지 셋에서 동일한 옷을 가지는 이미지를 찾아내는 의류 검색 기술은 앞으로도 활발하게 성장할 온라인 쇼핑 시장에서 반드시 필요한 기술이라고 생각한다. 두 기술은 웹 상에 존재하는 거대한 의류 이미지 데이터 속에서 원하는 의류 이미지에 쉽게 접근할 수 있게 만들 것이다. 사용자가 단어가 아닌 이미지로 검색을 하고 원하는 옷을 찾을 수 있는 서비스를 상용화하기 위한 노력이 꾸준히 이루어지고 있는 만큼, 의류 인식과 의류 검출과 관련된 연구는 앞으로도 활발히 이루어질 것으로 예측된다.

References

[1] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proc. IEEE Conf. CVPR*, pp. 1096-1104, Jun. 2016.

[2] Y. Ge, R. Zhang, X. Wang, X. Tang, and P. Luo, "DeepFashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images," in *Proc. IEEE Conf. CVPR*, pp. 5337-5345, Jun. 2019.

[3] W. Wang, Y. Xu, J. Shen, and S. C. Zhu, "Attentive fashion grammar network for fashion landmark detection and clothing category classification," in *Proc. IEEE Conf. CVPR*, pp. 4271-4280, Jun. 2018.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, May 2015.

[5] X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, and L. S. Davis, "Automatic spatially-aware fashion concept discovery," in *Proc. IEEE ICCV*, pp. 1463-1471, Oct. 2017.

[6] X. Zou, X. Kong, W. Wong, C. Wang, Y. Liu, and Y. Cao, "FashionAI: A hierarchical dataset for fashion understanding," in *Proc.*

IEEE Conf. Comput. Vision and Pattern Recognition Workshops (ICCVW), Oct. 2019.

[7] M. Hadi Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg, "Where to buy it: Matching street clothing photos in online shops," in *Proc. IEEE ICCV*, pp. 3343-3351, Oct. 2015.

[8] H. Chen, A. Gallagher, and B. Girod, "Describing clothing by semantic attributes," in *Proc. ECCV*, pp. 609-623, Oct. 2012.

[9] J. Huang, R. S. Feris, Q. Chen, and S. Yan, "Cross-domain image retrieval with a dual attribute-aware ranking network," in *Proc. IEEE Int. Conf. Computer Vision*, pp. 1062-1070, Oct. 2015.

[10] Z. Liu, S. Yan, P. Luo, X. Wang, and X. Tang, "Fashion landmark detection in the wild," in *Proc. ECCV*, pp. 229-245, Oct. 2016.

[11] M. I. Vasileva, B. A. Plummer, K. Dusad, S. Rajpal, R. Kumar, and D. Forsyth, "Learning type-aware embeddings for fashion compatibility," in *Proc. ECCV*, pp. 390-405, Oct. 2018.

[12] C. Corbiere, H. Ben-Younes, A. Ramé, and C. Ollion, "Leveraging weakly annotated data for fashion image retrieval and label prediction," in *Proc. IEEE ICCV*, pp. 2268-2274, Oct. 2017.

[13] Z. Kuang, Y. Gao, G. Li, P. Luo, Y. Chen, L. Lin, and W. Zhang, "Fashion retrieval via graph reasoning networks on a similarity pyramid," in *Proc. IEEE ICCV*, pp. 3066-3075, Oct. 2019.

[14] M. Shin, S. Park, and T. Kim, "Semi-supervised feature-level attribute manipulation for fashion image retrieval," in *Proc. BMVC*, Sep. 2019.

[15] J. J. Tompson, A, Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Proc. Advances in NIPS*, pp. 1799-1807, Dec. 2014.

[16] J. Long, E. Shelhamer, and T. Darrell, "Fully-convolutional networks for semantic segmentation," in *Proc. IEEE Conf. CVPR*, pp. 3431-3440, Jun. 2019.

[17] A. Joulin, L. van der Maaten, A. Jabri, and N. Vasilache, "Learning visual features from large weakly supervised data," in *Proc. ECCV*, pp. 67-84, Oct. 2016.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. CVPR*, pp. 770-778, Jun. 2016.

이 수 민 (Sumin Lee)

2018년 2월 : 경북대학교 전자공학부 졸업
2020년 2월 : 한국과학기술원 전기및전자공학부 석사
2020년 3월~현재 : 한국과학기술원 전기 및 전자공학부 박사과정
<관심분야> 전자공학, 영상처리, 딥러닝
[ORCID:0000-0003-1490-1355]

오 성 찬 (Sungchan Oh)

2006년 2월 : 서강대학교 전자공학과 졸업
2008년 2월 : 서강대학교 전자공학과 석사
2014년 2월 : 서강대학교 전자공학과 박사
2014년 11월~2016년 10월 : LG 전자 CTO부문 선임연구원
2016년 11월~현재 : 한국전자통신연구원 선임연구원
<관심분야> 전자공학, 영상인식, 딥러닝
[ORCID:0000-0003-3700-6492]

정 찬 호 (Chanho Jung)

2004년 2월 : 서강대학교 전자공학과 졸업
2006년 2월 : 서강대학교 전자공학과 석사
2013년 2월 : 한국과학기술원 전기및전자공학부 박사
2016년 9월~현재 : 한밭대학교 전기공학과 부교수
<관심분야> 전자공학, 영상인식, 딥러닝
[ORCID:0000-0003-3145-6732]

김 창 익 (Changick Kim)

1989년 : 연세대학교 전기공학과 학사 졸업.
1991년 : 포항공과대학교 전기전자 공학과 석사 졸업.
2000년 : 워싱턴주립대학교 전기전자 공학과 박사 졸업.
1991년~1997년 : ㈜ KC 중앙연구소 선임연구원
2000년~2005년 : Epson Palo Alto Lab. 책임연구원
2005년~2009년 : 한국정보통신대학교 조교수, 부교수
2009년~2010년 : HP Labs, Palo Alto 방문연구원
2009년~2014년 : 한국과학기술원 부교수
2015년~2016년 : UC Berkeley 방문교수
2014년~현재 : 한국과학기술원 교수
<관심분야> 영상처리, 영상이해, 컴퓨터비전, 패턴인식
[ORCID:0000-0001-9323-8488]