KIPS Tr. Software and Data Eng.
Vol.5, No.8 pp.393~400 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2016.5.8.393

A Safety Score Prediction Model in Urban Environment Using Convolutional Neural Network

Hyeon-Woo Kang† · Hang-Bong Kang††

ABSTRACT

Recently, there have been various researches on efficient and automatic analysis on urban environment methods that utilize the computer vision and machine learning technology. Among many new analyses, urban safety analysis has received a major attention. In order to predict more accurately on safety score and reflect the human visual perception, it is necessary to consider the generic and local information that are most important to human perception. In this paper, we use Double-column Convolutional Neural network consisting of generic and local columns for the prediction of urban safety. The input of generic and local column used re-sized and random cropped images from original images, respectively. In addition, a new learning method is proposed to solve the problem of over-fitting in a particular column in the learning process. For the performance comparison of our Double-column Convolutional Neural Network, we compare two Support Vector Regression and three Convolutional Neural Network models using Root Mean Square Error and correlation analysis. Our experimental results demonstrate that our Double-column Convolutional Neural Network model show the best performance with Root Mean Square Error of 0.7432 and Pearson/Spearman correlation coefficient of 0.853/0.840.

Keywords : Urban Safety, Convolutional Neural Network, Crime Prediction, Visual Perception

컨볼루션 신경망을 이용한 도시 환경에서의 안전도 점수 예측 모델 연구

강 현 우† · 강 행 봉††

요 약

최근, 컴퓨터 비전과 기계 학습 기술의 도움을 받아 효율적이고 자동적인 도시 환경에 대한 분석 방법의 개발에 대한 연구가 이루어지고 있다. 많은 분석들 중에서도 도시의 안전도 분석은 지역 사회의 많은 관심을 받고 있다. 더욱 정확한 안전도 점수 예측과 인간의 시각적 인지를 반영하기 위해서, 인간의 시각적 인지에서 가장 중요한 전역 정보와 지역 정보의 고려가 필요하다. 이를 위해 우리는 전역 칼럼과 지역 칼럼으로 구성된 Double-column Convolutional Neural Network를 사용한다. 전역 칼럼과 지역 칼럼 각각은 입력은 크기가 변환된 원 영상과 원 영상에서 무작위로 크로핑을 사용한다. 또한, 학습 과정에서 특정 칼럼에 오버피팅되는 문제를 해결하기 위한 새로운 학습방법을 제안한다. 우리의 DCNN 모델의 성능 비교를 위해 2개의 SVR 모델과 3개의 CNN 모델의 평균 제곱근 오차와 상관관계 분석을 측정하였다. 성능 비교 실험 결과 우리의 모델이 0.7432의 평균 제곱근 오차와 0.853/0.840 피어슨/스피어맨 상관 계수로 가장 좋은 성능을 보여주었다.

키워드 : 도시 환경 분석, 컨볼루션 신경망, 안전도 예측, 시각적 인지

1. 서 론

최근 도시 환경에 대한 광범위한 분석에 대한 관심이 증가하고 있고 이에 대한 다양한 연구들이 이루어지고 있다[1-9]. 이러한 연구들은 “도시에서 소득이 높고/낮은 지역은 어디인가?”, “도시에서 범죄가 많이 발생하는 지역은 어디인가?” 같은 경제, 인구 그리고 안전 등에 대한 분석이 많이 이루어지고 있다. 과거 이러한 분석은 그 규모가 매우 크기 때문에 많은 정보 수집과 비용이 필요했다. 하지만, 최근 컴퓨터 비전과 기계 학습 기술의 도움을 받아 영상으로부터 도시 환경을 예측 할 수 있는 새로운 방법들이 개발되고 있다. 특히, 범죄 혹은 안전도 예측은 프레드폴(PredPol), 히타치(Hitachi) 등에서 연구를 시작하면서 사회적으로 큰 관심을 받고 있다.

※ 이 논문은 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2015R1A2A1A10056304).
† 준 회 원 : 가톨릭대학교 디지털미디어과 석사과정
†† 종신회원 : 가톨릭대학교 디지털미디어학부 교수
Manuscript Received : March 4, 2016
First Revision : April 15, 2016
Accepted : April 26, 2016
* Corresponding Author : Hang-Bong Kang(hbkang@catholic.ac.kr)

도시의 안전도를 예측하는 기술들 중 하나는 Wilson과 Kelling의 깨진 창문 이론[10]을 기반으로 하는 방법이 있다. 깨진 창문 이론이란 사회에 존재하는 사소한 무질서를 방치하게 되면 사회적으로 큰 혼란으로 이어진다는 것이다. 예를 들어, 깨진 유리창, 낙서와 쓰레기 등과 같은 시각적으로 인지되는 무질서의 방치는 그 지점을 중심으로 범죄의 발생 및 확산의 원인이 된다는 것이다. 다시 말해 사람들이 도시에서 시각적으로 무질서하거나 혹은 위험하게 인지되는 장소는 실제로 낮은 안전도를 가진다는 것이다. 이 이론을 기반으로 인간의 시각적 인지(human visual perception)를 이용하여 도시 영상에서 인지되는 안전도를 평가하고 이를 기계학습 알고리즘을 통해 학습 및 예측하는 연구들이 이루어졌다[2-6].

대부분의 기존 연구들은 세 단계를 통해서 도시 영상의 안전도 점수 예측 모델을 생성한다. 우선, 사람들의 인지적인 평가를 이용한 실험을 통해 도시 영상의 안전도 점수를 구한다. 실험은 주로 두 장의 영상을 비교하여 상대적인 평가를 실시하는 쌍대비교(pairwise comparison)실험을 사용한다. 이어서 학습을 위해 영상으로부터 특징을 추출하고, 마지막으로 실험을 통해 얻어진 안전도 점수와 추출된 특징을 학습하여 안전도 점수 예측 모델을 생성한다. 이런 방법들은 도시 영상으로부터 특징을 추출하고 학습 알고리즘을 사용하여 안전도 점수 예측 모델을 생성하기 때문에 사용하는 특징 추출 방법과 학습 알고리즘에 따라 성능이 결정된다. 기존의 방법들은 영상으로부터 특징을 추출하기 위해서 전역 특징 추출 방법을 사용해왔다. 이러한 방법들은 영상으로부터 색상, 텍스쳐, 그라디언트 등의 정보를 인코딩하는 저수준 특징(low level feature)을 사용하는 방법이다. 때문에 인간의 시각적 인지를 기반으로 하는 영상의 안전도 점수와 같은 추상적인 정보의 예측에는 한계가 존재한다. 이를 극복하기 위해 우리는 학습을 통해 영상으로부터 추상화된 고수준 특징(high level feature)을 추출하는 능력이 뛰어난 컨볼루션 신경망(CNN: Convolution Neural Network) 알고리즘을 통한 안전도 점수 예측 방법을 연구한다. CNN은 얼굴 인식, 영상 분류 및 사물 검출 등 영상 처리 및 컴퓨터 비전 분야에서 놀라운 성능의 향상을 보여주었다. CNN은 2012년 Imagenet Large Scale Visual Recognition Challenge (ILSVRC)에서 기존의 다른 알고리즘에 비해 압도적인 성능을 보여주며 1위를 차지하였고 이를 계기로 유명해졌다[11]. CNN은 기존의 특징 추출과 예측 모델로 두 단계를 통해 예측하는 방법과는 다르게 특징 추출과 예측 모델을 한 번에 학습하는 end-to-end 학습을 실시한다. 또한, 입력 영상을 추상화하고 추상적인 정보를 추출하는 능력이 뛰어난 모델로 다양한 분야에서 우수한 성능을 보여주고 있다.

본 논문에서는 쌍대비교실험을 통해 안전도 점수를 구한 Place Pulse 1.0 데이터 셋[2]을 사용한다. Place Pulse 1.0 데이터 셋으로부터 수집한 영상과 안전도 점수를 이용해서 CNN의 학습을 실시한다. CNN을 통해서 고수준의 추상적인 특징 정보를 추출하여 학습을 할 수 있지만 인간의 시각적 인지를 기반으로 평가되는 안전도 점수를 정확하게 예측하기에는 부족함이 있다. 왜냐하면 일반적으로 CNN은 전역 정보를 추출하여 사용하며 CNN의 적용과정에서 영상의 크기와 해상도의 변형이 필요하고 이 과정에서 영상의 세부적인 정보는 손실된다. 하지만, 깨진 창문 이론과 인간의 시각적 인지를 좀 더 정확하게 반영하기 위해서는 전역 정보로부터 얻을 수 있는 컨텍스트 같은 시각적 단서 정보뿐만 아니라 세부적인 지역 정보로부터 얻을 수 있는 오브젝트 같은 시각적 단서 정보의 반영이 필요하다. 때문에 우리는 전역 정보와 지역 정보를 추출하는 두 개의 칼럼을 가진 Double-column CNN 구조(DCNN)[13, 14]를 통해 기존의 예측 방법에서 고려하지 못한 깨진 창문 이론과 인간의 시각적 인지를 좀 더 정확하게 반영하는 도시의 안전도 점수 예측 모델을 제안하고자 한다. DCNN에서 전역 정보와 지역 정보는 각각 Alexnet과 동일한 구조를 가지고 독립적으로 동작하는 두 개의 칼럼의 컨볼루션 층을 통해 추출하였고, 각 정보는 완전 연결 층에서 융합되어 도시의 안전도 예측에 사용된다. 또한, 우리는 DCNN의 학습을 위해 일반적인 학습 방법인 공동 훈련(joint training) 방식이 하나의 칼럼에 오버피팅(overfitting) 될 수 있다는 문제로 인해 새로운 학습 방법을 사용해 학습을 실시한다.

본 논문의 구성은 다음과 같다. 2절에서 본 논문의 관련 연구에 대해서 기술한 다음, 3절에서 Place Pulse 1.0 데이터 셋에 대해서 설명한다. 그리고 4절에서 본 논문에서 사용하는 Alexnet[11]과 우리의 DCNN의 구조와 학습 방법에 대한 내용을 설명한다. 마지막으로 5절에서는 제안한 모델의 실험 방법 및 평균 제곱근 오차(RMSE: Root Mean Square Error)와 상관관계 분석을 통한 성능 평가 결과를 보여주고 6절에서 결론을 맺는다.

2. 관련 연구

2.1 안전도 예측

안전도 예측 혹은 범죄 예측에 관련된 다양한 연구들이 존재한다. Salesses 등[2]은 깨진 창문 이론[10]을 기반으로 인간의 시각적 인지를 이용한 쌍대비교실험을 통해 도시 영상의 부유함, 독특함과 안전도 점수를 측정하였고 이를 활용하여 도시 환경의 분석 및 예측 방법을 실시했다. Ordonez 등[3]은 Salesses 등[2]의 데이터 셋을 뉴욕, 보스턴, 볼티모어 그리고 시카고에서 추가 수집한 데이터를 이용하여 확장했다. 또한, 도시 영상의 부유함, 독특함과 안전도 점수를 예측하기 위해서 GIST[18], Fisher Vector[19], 그리고 DeCAF [20] (Deep Learning 특징) 특징을 추출하고 서포트 벡터 머신(SVM: Support Vector Machine)과 SVR을 사용해 학습하였다. Khosla 등[4]은 도시의 영상의 장소로부터 원하는 목적지를 찾아가거나, 특정 가게와 가까이 있는지 그리고 안전한 장소인지라는 매우 새로운 연구를 실시하였다. 다양한 특징(GIST, 텍스쳐, 색상 등)과 SVR을 사용해서 8개 도시에서 사람과 컴퓨터의 성능을 비교했다. Naik 등[5]은 Salesses 등[2]의 데이터 셋의 영상을 이용해서 마이크로소

프트 트루스킬 알고리즘을 적용하여 안전도 점수를 새로 구했다. 또한, HoG[17], GIST 등 11개의 특징을 이용해서 SVR을 학습시켰다. 하지만, 복잡한 낙서, 다리 밑 그리고 현대식 건물 사진에서는 안전도 점수 예측에서 낮은 정확도를 보여주었다. Kang and Kang[6]은 context를 고려한 도시 안전도 예측 방법을 제안하였다. 그들은 컨텍스트 정보를 안전도 점수에 관해서 긍정과 부정으로 분류하였고, 컨텍스트에 대한 쌍대비교 실험을 실시하였다. 그들은 실험을 통해 영상에서 나타나는 컨텐스트에 따른 안전도 점수의 변화를 측정하였다. 구해진 컨텍스트 정보는 SVR 모델을 통해 예측된 안전도 점수를 보정해주는 사후처리에 사용하였다. 또한, 영상과 사람의 시각적 인지를 이용한 평가를 통한 안전도 예측 방법 외에도 범죄 발생 기록, 트위터 등을 활용한 방법도 존재한다. Mohler 등[7]은 범죄가 지진-여진과 같이 최초 범죄 발생 지역을 중심으로 퍼져나간다는 패턴을 발견했다. 이를 기반으로 지진-여진 알고리즘을 활용하여 과거 범죄 발생 기록을 사용해 범죄 발생을 모델링했다. Gerber 등[8]과 Chen 등[9]은 과거 범죄 발생 기록과 범죄 발생 지역 주변에서는 트위터에서 부정적인 감정이 많이 발생한다는 점을 이용해서 범죄 발생을 예측했다.

2.2 CNN

CNN은 다른 딥 러닝 알고리즘과는 다르게 컨볼루션 층을 가지고 있고 때문에 컴퓨터 비전의 다양한 분야에서 뛰어난 성능을 보여주고 있다. Krizhevsky 등[11]은 CNN에 Rectified Linear Unit (ReLU), Local Response Normalization (LRN), Dropout 등을 적용하여 ILSVRC 2012에서 종전의 기록을 압도적인 성능으로 뛰어넘어 우승하였고 약 10% 정도의 성능 향상을 보여주면서 CNN을 알리게 되었다. Simonyan과 Zisserman[12]은 매우 작은 3×3 크기의 커널을 사용하여 층의 깊이를 깊게 쌓는 구조(16-19층)를 통해 ILSVRC에서 Krizehvsky 등의 방법보다 약 10% 높은 성능을 보여주었다. Lu 등[13]은 영상의 미학 평가를 위해 CNN을 통해 영상의 전역 정보와 세부 정보를 융합하여 사용했다. 이를 위해 두 개의 칼럼을 가지는 CNN 구조를 적용하였다. 각 칼럼은 컨볼루션 층을 통해 전역 정보와 세부 정보를 추출하고, 추출된 정보는 완전 연결 층에서 융합되어 학습을 실시한다. Jung 등[14]은 얼굴 표정 인식을 위해서 얼굴 영상과 얼굴 특징점을 사용하여 신경망을 학습했다. 하나의 칼럼이 영상이 아닌 얼굴 특징 점을 입력으로 사용하기 때문에 Lu 등의 방식과는 다르게 얼굴 영상을 입력으로 사용하는 CNN과 얼굴 특징 점을 입력으로 사용하는 신경망을 통해 더블 칼럼 신경망을 구성했다.

3. Plcae Pulse 1.0

본 논문에서는 공개된 데이터 셋인 Salesses 등이 수집한 Place Pulse 1.0을 사용한다[2]. 데이터 셋의 각 영상은 구글 스트리트 뷰를 통해 뉴욕, 보스턴, 린츠와 잘츠부르크 4개의 도시에서 수집되었으며 총 4,136개의 위치 정보(위도, 경도)와 카메라 회전 정보(heading, pitch)를 가지고 있다. 또한, 각 데이터 i는 부유함(wealth), 독특함(uniqueness) 그리고 안전도(safety) 점수를 제공하고 각 점수는 인간의 시각적 인지를 이용해 점수를 구하는 쌍대비교 실험을 통해 구해졌다. 실험은 웹상에서 크라우드소싱을 통해 실시했고 총 7,872명의 참가자가 실험에 참여하였으며 208,738번의 쌍대비교 실험이 이루어졌다. 우리는 Place Pulse 1.0 데이터 셋에서 제공하는 좌표를 이용해서 구글 스트리트뷰 API를 통해 영상을 수집하였다. 구글 스트리트뷰 API를 통해 수집된 영상 중 영상이 존재하지 않는 데이터는 모두 제거하여 뉴욕 1,461장, 보스턴 1,047장, 린츠 253장과 잘츠부르크 212장 총 2,973장의 영상과 각 영상의 안전도 점수를 사용한다.

Fig. 1A는 Place Pulse 1.0의 쌍대 비교 실험의 예시를 보여준다. 실험은 도시 영상 데이터 셋에서 무작위로 선택된 두 장의 영상을 보여주고 다음과 같은 질문을 한다. “두 영상 중 더 안전해 보이는 영상은 무엇인가?” 그러면 실험 참가자는 질문에 따라 “왼쪽 영상, 오른쪽 영상 혹은 동등하다” 세 가지 중 하나를 선택한다. 실험의 결과는 Equation (1)을 통해서 최종 안전도 점수가 구해진다[15]. Equation (2)와 (3)은 각각 영상 i가 선택 받은 확률과 선택받지 못한 확률을 나타낸다.

Fig. 1. A: Example of pairwise comparison experiment of Placle Pulse 1.0. The users select an answer according to the following question: “Which place looks safer?”, B: Example of high and low safety score images of Place Pulse 1.0. The upper row shows the high safety score images. The bottom row shows low safety score images.

$$s_i = \frac{10}{3}\left(W_i + \frac{1}{w_i}\sum_{j_1=1}^{w_i} W_{j_1} - \frac{1}{l_i}\sum_{j_2=1}^{l_i} L_{j_2} + 1\right) \quad (1)$$

$$W_i = \frac{w_i}{w_i + l_i + t_i} \quad (2)$$

$$L_i = \frac{l_i}{w_i + l_i + t_i} \quad (3)$$

w_i, l_i 그리고 t_i는 각각 실험 중 영상 i가 다른 영상과 비교하여서 선택 받거나, 선택받지 못하거나 혹은 동등하다고 선택한 횟수를 나타낸다. Equation (1)의 두 번째와 세 번째 항은 각각 영상 i가 선택받았을 경우 선택받지 못한 영상들의 W_i의 평균과 선택받지 못했을 경우 선택받은 영상들의 L_i의 평균을 나타낸다. Equation (1)의 네 번째 항 1과 상수 항 $\frac{10}{3}$은 안전도 점수 s_i를 0-10 점으로 정규화 시켜주는 역할을 한다. Fig. 1B는 Place Pulse 1.0에서 쌍대비교 실험을 통해 구해진 안전도 점수 중 높은 안전도 점수를 가지는 영상과 낮은 안전도 점수를 가지는 영상들의 예시를 보여준다.

4. 예측 모델

4.1 Alexnet

본 논문에서는 Krizhevsky 등의 ILSVRC 2012에서 우승한 Alexnet[11] 구조를 사용한다. Fig. 2는 Alexnet의 구조를 보여준다. Alexnet은 5개의 컨볼루션 층과 3개의 완전 연결 층으로 구성되어 있으며 ReLU, max-pooling, LRN과 dropout 기법을 사용한다. Alexnet은 고정된 크기의 영상($227 \times 227 \times 3$)을 입력으로 사용하며 첫 번째 컨볼루션 층은 $11 \times 11 \times 3$ 크기의 96개의 커널을 가지고 있다. 두 번째 컨볼루션 층은 첫 번째 컨볼루션 층의 결과인 $55 \times 55 \times 96$ 크기를 입력으로 사용하며 $5 \times 5 \times 96$ 크기의 256개의 커널을 가지고 있다. 세 번째 컨볼루션 층은 $3 \times 3 \times 256$ 크기의 38개의 커널을 가지고 있으며 네 번째와 다섯 번째 컨볼루션 층은 $3 \times 3 \times 384$ 크기의 384개와 256개의 커널을 각각 가지고 있다. 각 컨볼루션 층의 결과는 ReLU를 적용하며 첫 번째, 두 번째 그리고 다섯 번째 컨볼루션 층은 3×3 크기의 max pooling을 적용한다. 또한, 첫 번째와 두 번째 컨볼루션 층은 max pooling의 결과에 LRN을 적용한다. 컨볼루션 층에 이어지는 여섯 번째, 일곱 번째 그리고 여덟 번째 3개의 완전 연결 층은 각각 4096개, 4096개 그리고 1개의 뉴런을 가진다. 여덟 번째 완전 연결 층은 기존의 Alexnet의 경우 1000종류의 영상 인식 때문에 1000개의 뉴런을 사용했지만 본 논문에서는 안전도 점수 예측이 목표이기 때문에 여덟 번째 완전 연결 층은 안전도 점수를 나타내는 1개의 뉴런을 가지도록 변경하여서 사용했고, 손실 층(loss layer)은 유클리디안 손실(euclidean loss)을 사용했다.

4.2 DCNN

1절에서 설명하였듯이 우리는 깨진 창문 이론과 인간의 시각적 인지를 반영하기 위해서 전역 정보로부터 얻어지는 시각적 단서 정보(e.g. 배경, 컨텍스트 등)와 지역 정보로부터 얻어지는 시각적 단서 정보(e.g. 오브젝트 등)를 반영하는 CNN 모델을 학습하고자 한다. 하지만, 일반적으로 CNN은 고정된 크기의 영상(예를 들어, Alexnet의 경우 $227 \times 227 \times 3$) 하나를 입력으로 사용한다. 때문에 데이터 셋의 영상을 고정된 크기로 변환시키는 과정이 필수적이다. 하지만, 영상의 크기를 변환시키는 과정 때문에 전역 정보와 지역 정보를 모두 사용하기 어렵다는 문제점이 있다. 영상을 변환시켜서 입력으로 사용하는 경우 영상의 전역 정보는 얻을 수 있지만 세밀한 지역 정보들은 크기 변환 과정에서 변형 및 손실이 발생하게 된다. 때문에 우리는 두 개의 칼럼을 가지는 DCNN 구조를 사용한다[13].

Fig. 3은 DCNN의 구조를 보여준다. DCNN은 두 개의 칼럼으로 구성되어 각 칼럼마다 하나의 영상을 입력으로 사용하여 총 2개의 영상을 입력으로 사용하는 CNN의 구조이다. 각 칼럼의 컨볼루션 층은 독립적으로 작동하며 서로 다른 입력 영상을 사용하여 컨볼루션 과정을 거쳐서 각각 다른 정보를 추출하게 되고, 각 칼럼에서 컨볼루션 층을 거쳐 추출된 결과는 완전 연결 층에서 하나의 특징 벡터로 합쳐진다. 즉, DCNN은 두 개의 정보를 융합하여 사용하는 CNN 구조라고 볼 수 있다. 본 논문에서는 DCNN의 두 개의 칼럼을 전역 정보를 다루는 전역 칼럼과 지역 정보를 다루는 지역 칼럼으로 구성한다.

전역 칼럼은 영상의 전체를 입력으로 사용하는 칼럼으로

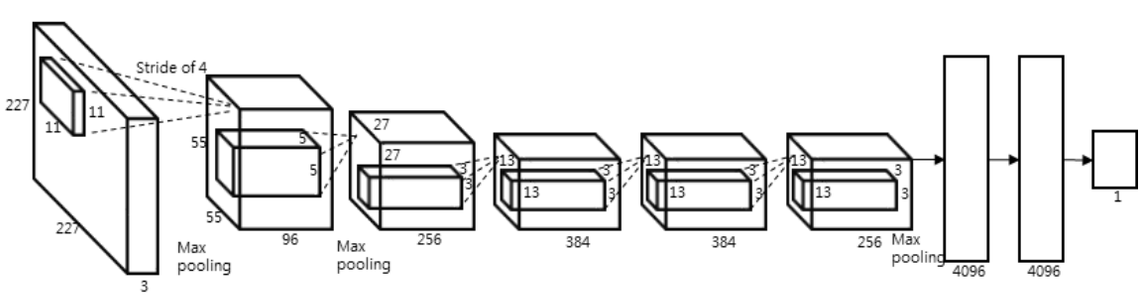

Fig. 2. Alexnet structure. It consist of five convolutional layers and three fully connected layers. We changed size of last fully connected layer 1000 to 1 because our CNN model predict safety score of urban image.

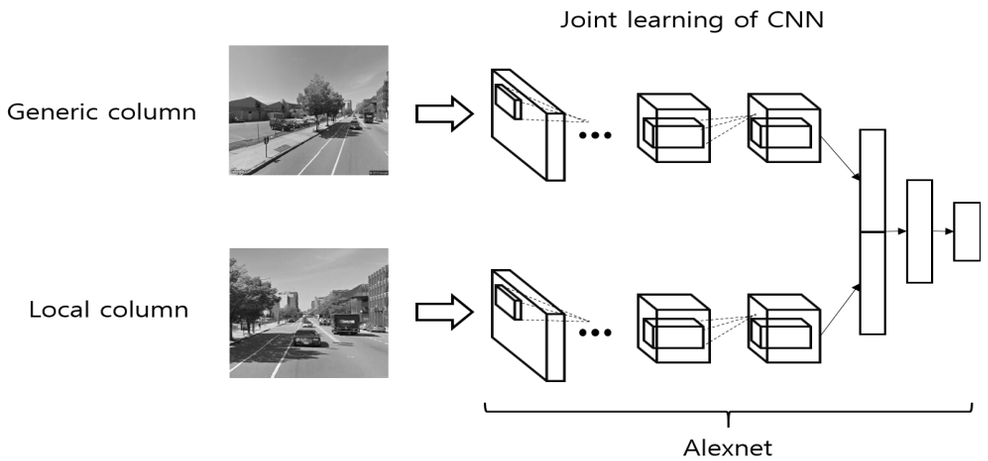

Fig. 3. Structure of DCNN. It have two column. One is generic column which use the resized full image as input, the other is local column which use the random cropping of original image as input

원 영상을 $227 \times 227 \times 3$의 크기로 변환하여 입력으로 사용한다. 영상의 크기를 조정하는 과정에서 원 영상과는 다른 해상도와 비율로 영상을 변형 하게 되고 이로 인해 영상의 세밀한 정보를 손실하거나 왜곡되는 문제가 발생한다. 지역 칼럼은 원 영상으로부터 $227 \times 227 \times 3$ 크기의 영상을 무작위 크로핑하여 입력으로 사용한다. 이는 영상의 크기나 해상도를 변형하지 않은 원 영상을 그대로 사용하기 때문에 변형으로 인해 손실 되거나 왜곡되는 문제가 발생하지 않는다. Fig. 4는 본 논문에서 사용한 DCNN의 전역 칼럼과 지역 칼럼의 입력 영상의 예시를 보여준다.

Fig. 4. Example of input images of generic column and local column of our dataset. The upper row shows the input of generic images. The bottom row shows the input of local images.

4.3 학습

DCNN은 두 개의 칼럼으로 구성되어 있기 때문에 일반적으로 두 개의 칼럼을 동시에 학습하는 공동 학습 방법을 사용한다. 공동 학습 방법의 문제점은 두 개의 칼럼이 한 번에 학습이 진행되기 때문에 DCNN이 각 칼럼에서 추출되는 정보를 잘 융합하여 예측을 하는 것이 아닌 특정 칼럼에 대해서 오버피팅되는 문제가 발생 할 수 있다는 것이다. 예를 들어 두 개의 칼럼 c_i와 c_j가 있을 때, 학습 과정 중 c_i 칼럼의 정보가 먼저 최적화 되는 경우 손실 층에서 발생하는 손실 값이 급격하게 감소하게 된다. 때문에 c_j 칼럼의 최적화는 어려워지게 되고 DCNN의 완전 연결 층은 두 칼럼의 정보를 융합하여 예측을 하는 것이 아닌 c_i 칼럼의 정보에 더욱 의존적이게 예측하도록 학습이 될 것이다. 특히, 본 논문에서는 하나의 칼럼이 무작위로 크로핑한 영상을 사용하는 경우 하나의 칼럼에 오버피팅되는 문제는 더욱 치명적이다.

공동 학습의 문제점을 해결하기 위해서 우리는 새로운 학습 방법을 사용해서 DCNN을 학습한다. Fig. 5는 본 논문에서 적용한 학습 방법의 절차를 보여준다. 우선, 각 칼럼을 독립적인 CNN으로 학습을 진행한다. 독립적으로 학습이 완료되면 두 칼럼의 컨볼루션 층의 가중치를 사용하여 새로운 완전 연결 층을 가지는 DCNN으로 구성한다. 그리고 DCNN을 학습 데이터 셋을 사용해서 다시 한 번 학습을 실시한다. 학습 과정에서 컨볼루션 층의 가중치는 변화시키지 않고 오직 완전 연결 층의 가중치만을 학습한다. 각 칼럼의 특징을 추출하는 컨볼루션 층은 독립적으로 학습된 CNN의 가중치를 변화하지 않고 사용하기 때문에 특정 칼럼에 오버피팅되는 문제점이 발생하지 않는다. 또한, 컨볼루션 층의 가중치가 고정되기 때문에 완전 연결 층은 두 칼럼에서 추출되는 정보를 잘 융합하여 예측하도록 학습이 된다.

5. 실험 방법 및 결과

5.1 실험 방법

우리는 도시 영상의 안전도 점수 예측을 위해서 전역 정보와 지역 정보를 융합하는 CNN 구조를 사용한다. 이를 위해 기본적인 DCNN 구조 외에 전역 정보와 지역 정보를 융합하는 다양한 학습 방법과 네트워크를 구성하여 학습을 실시한다. 본 논문에서는 기존 연구에서 주로 사용된 SVR 모델 2개와 다양한 CNN 모델의 성능을 평균 제곱근 오차와 상관관계 분석을 통해 성능 비교 실험을 실시한다. SVR 모델은 전역 영상에서 HoG 특징을 추출하여 학습한 모델(SVR + HoG)과 전역 영상과 지역 영상에서 HoG 특징을

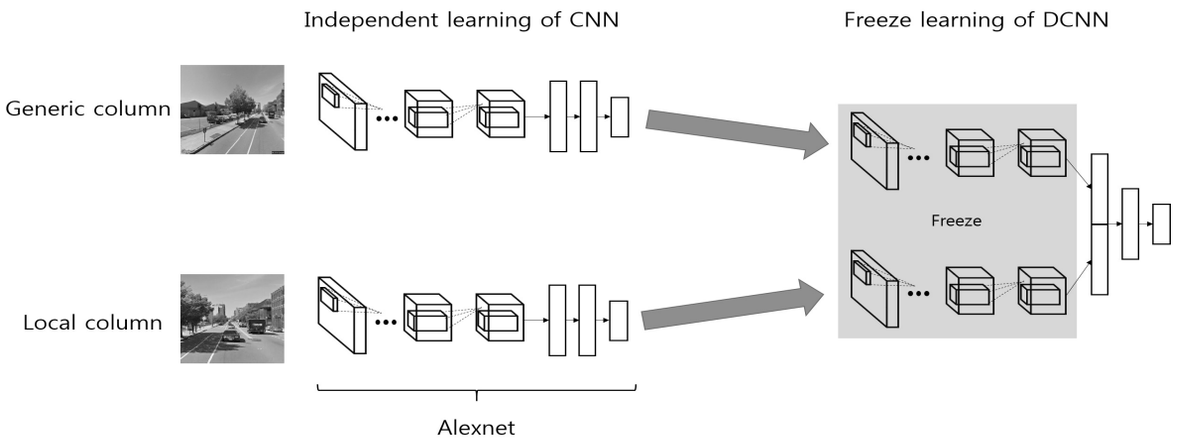

Fig. 5. Procedure of our learning method. In the first phase, we train the two independent CNN. In the second phase, we freeze weights of convolution layers during learning.

추출하여 하나의 특징 벡터로 연결하여 학습한 모델(SVR + joint)을 사용한다. CNN 모델은 Alexnet, 공동 학습을 실시한 DCNN 그리고 특징 공유(feature sharing) 구조를 사용한다. 특징 공유 구조는 전역 영상과 지역 영상을 동일한의 CNN의 입력으로 사용하여 특징을 추출하는 구조이다.

본 논문에서는 Place Pulse 1.0 데이터 셋으로부터 사용하는 영상은 총 2,973장으로 이 중 학습에 2,500장, 테스트에 473장을 사용하였다. SVR은 L2 손실(L2 loss)을 사용하는 L2-regularized SVR 모델을 사용하고, 파라미터 p와 c를 변경해가면서 학습을 실시하였다. 각 CNN 모델들은 초기 학습율은 0.001로 설정하였고 dropout 비율은 0.5로 설정하였다. CNN과 SVR을 학습하기 위해서 딥러닝 프레임워크 caffe [16]와 기계 학습 라이브러리 LIBLINEAR[21]를 사용하였고, Geforce GTX TITAN X 그래픽 카드를 사용하여 학습을 실시했다. 우리의 DCNN 모델을 학습하는데 11시간이 걸렸으면, 473장을 테스트하는데 약 3.5초의 시간이 걸린다.

5.2 실험 결과

본 논문에서는 모델들 사이의 성능 비교를 실시하기 위해서 평균 제곱근 오차와 상관관계 분석을 실시하였다. 평균 제곱근 오차는 원 안전도 점수와 예측된 안전도 점수가 얼마만큼의 차이를 가지는지를 측정하는 방법이다. Table 1은 평균 제곱근 오차의 성능 비교 결과를 보여준다. 성능 비교 결과 우리의 DCNN 모델이 0.7432의 평균 제곱근 오차로 가장 좋은 성능을 보여주었다. 즉, 우리의 모델의 예측 결과는 가장 적은 차이를 가짐을 알 수 있다.

우리의 DCNN 모델의 추가적인 성능 비교를 위해서 원 안전도 점수와 예측된 안전도 점수 사이의 상관관계 분석을 실시하였다. 상관관계 분석은 두 변수간의 관련성을 구하는 피어슨 상관계수(Pearson correlation coefficient)와 순위를 이용하는 스피어맨 상관계수(Spearman correlation coefficient)를 구해서 성능 비교를 실시했다. 통계 분석은 SPSS ver 18을 사용하여 실시하였다. Table 2는 상관관계 분석 결과를 보여주며, Fig. 6은 원 안점도 점수와 우리의 DCNN 모델의

Table 1. Results of RMSE performance comparison

Model	RMSE	Remarks
SVR + HoG	0.9893	best p - 1 best c - 0.001
SVR + joint	0.8238	best p - 1 best c - 0.001
Alexnet	0.8607	
Sharing	0.7606	
DCNN	0.7727	
Our DCNN model	0.7432	

Table 2. Results of Pearson and Spearman correlation coefficient

Model	Pearson correlation coefficient (p-value)	Spearman correlation coefficient (p-value)
SVR + HoG	0.739 (< 0.01)	0.741 (< 0.01)
SVR + joint	0.804 (< 0.01)	0.804 (< 0.01)
Alexnet	0.768 (< 0.01)	0.761 (< 0.01)
Sharing	0.845 (< 0.01)	0.837 (< 0.01)
DCNN	0.849 (< 0.01)	0.832 (< 0.01)
Our DCNN model	0.853 (< 0.01)	0.840 (< 0.01)

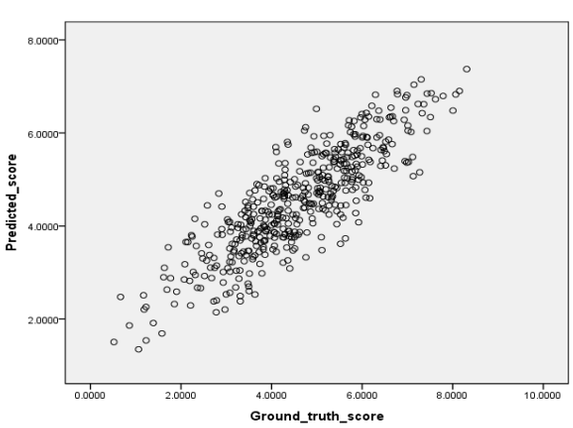

Fig. 6. Scatter plot of between predicted score of our DCNN model and ground truth score

Fig. 7. Example of prediction results using our DCNN model. The upper row shows the success results. The bottom row shows the failure results. s_g and s_p denote ground truth safety score and predicted safety score, respectively

산점도를 보여준다. 모든 모델의 상관관계 분석 결과 유의확률이 0.01보다 작은 값을 가지므로 상관계수는 유의수준 0.01에서 통계적으로 유의미한 결과를 보여주었다. 우리의 DCNN 모델은 0.853/0.840의 피어슨/스피어맨 상관계수로 원 안전도 점수와 예측된 안전도 점수가 가장 높은 양의 상관관계를 가짐을 보여주었고, 이를 통해 우리의 DCNN 모델이 가장 정확하게 안전도 점수를 예측한 것을 알 수 있다. Fig. 7은 우리의 DCNN 모델을 사용해 예측한 안전도 점수의 성공적인 결과와 실패한 결과의 예시를 보여준다.

실험 결과에서 평균 제곱근 오차와 상관관계 분석에서 전역 정보와 지역 정보를 모두 사용하는 SVR 모델과 CNN 모델이 전역 정보만을 사용하는 SVR 모델과 CNN 모델보다 좋은 성능을 보여주었다. 이를 통해 안전도 점수 예측에 전역 정보만 사용하는 것보다 지역 정보를 추가하여 예측을 실시하는 것이 더욱 정확한 안전도 점수 예측이 가능함을 알 수 있다. 또한, 공동 학습을 사용한 DCNN 모델보다 본 논문에서 제안하는 학습 방법을 사용한 DCNN 모델이 전역 정보와 지역 정보를 잘 융합하고 안전도 점수 예측 성능이 더 뛰어남을 알 수 있었다.

6. 결 론

본 논문에서는 전역 정보와 지역 정보를 융합하는 DCNN 구조를 사용하여 도시의 영상으로부터 안전도 점수를 예측하는 방법을 보여주었다. 우리는 DCNN의 공동 학습의 문제점을 극복하기 위해 새로운 학습 방법을 통해 DCNN을 학습하였다. 우리의 모델의 성능 비교를 위해 2개의 SVR과 3개의 CNN 모델을 학습하여 평균 제곱근 오차와 상관관계 분석을 통해 성능 비교를 실시하였다. 그 결과 0.7432의 평균 제곱근 오차로 가장 좋은 성능을 보였고 원 안전도 점수와 예측된 점수 사이의 상관관계 분석 결과는 0.853/0.840의 피어슨/스피어맨 상관계수로 가장 높은 양의 상관관계를 보였다.

우리의 연구는 거대한 도시에서 효율적으로 안전도 예측을 가능하게 하며 나아가 도시에서 발생하는 범죄 예측에도 활용이 가능하다고 생각된다. 또한, 실제 높은 안전도 점수를 보이는 영상과 낮은 안전도 점수를 가지는 영상을 분석해본 결과 높은 안전도 점수를 가지는 영상은 가로수, 깨끗한 거리 혹은 건물이 등장하는 영상이고, 반대로 낮은 안전도 점수를 보이는 영상은 자동차나 사람이 없는 텅 빈 거리, 지저분한 거리 혹은 건물이 등장하는 영상이 주로 존재한다는 패턴을 알 수 있었다.

차후에는 안전도 영상 데이터 셋을 서울과 북경 등과 같은 아시아권 도시로 확장하여 추가 수집을 할 계획이다. 또한, 현재는 전역 정보와 지역 정보를 단순하게 하나의 특징 벡터로 만들어 사용했지만 이를 안전도 점수 예측의 성능을 높일 수 있는 융합 방법의 개발이 필요하다. 이외에도 더 정확한 안전도 점수 예측을 위해 영상 정보 외에 추가적인 데이터를 수집하여 다중 데이터 융합을 통한 예측 모델의 개발을 할 계획이다.

References

[1] K. Lynch, “The image of the city,” MIT press, 1960.

[2] P. Salesses, S. Katja, and C. A. Hidalgo, “The collaborative image of the city: mapping the inequality of urban perception,” *PloS one*, Vol.8, No.7, 2013.

[3] V. Ordonez and T. L. Berg, “Learning high-level judgments of urban perception,” in *Proceedings of the European Conference on Computer Vision*, pp.494-510, 2014.

[4] A. Khosla, B. An, J. J. Lim, and A. Torralba, “Looking beyond the visible scene,” in *Proceedings of the Computer Vision and Pattern Recognition*, pp.3710-3717, 2014.

[5] N. Naik, J. Philipoom, R. Raskar, and C. Hidalgo,

"Streetscore--Predicting the Perceived Safety of One Million Streetscapes," in *Proceedings of the Computer Vision and Pattern Recognition Workshops*, pp.793-799, 2014.

[6] H. W. Kang and H. B. Kang, "A new context-aware computing method for urban safety," in *International Conference on Image Analysis and Processing Workshops*, pp.298-305, 2015.

[7] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita, "Self-exciting point process modeling of crime," *Journal of the American Statistical Association*, Vol. 106, pp.100-108, 2012.

[8] M. S. Gerber, "Predicting crime using Twitter and kernel density estimation," *Decision Support Systems*, Vol.61, pp.115-125, 2014.

[9] X. Chen, Y. Cho, and S. Jang, "Crime prediction using Twitter sentiment and weather," in *Proceedings of the Systems and Information Engineering Design Symposium*, pp.63-68, 2015.

[10] J. Q. Wilson and G. L. Kelling, "Broken windows," *Atlantic monthly*, Vol.249, No.3, pp.29-38, 1982.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the Advances in Neural Information Processing Systems*, pp.1097-1105, 2012.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *arXiv preprint arXiv:1409.1556*, 2014.

[13] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rapid: Rating pictorial aesthetics using deep learning," in *Proceedings of the ACM International Conference on Multimedia*, pp.457-466, 2014.

[14] H. Jung, S. Lee, J. Yim, S Park, and J. Kim, "Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition," in *Proceedings of the International Conference on Computer Vision*, pp.2983-2991, 2015.

[15] J. Park and M. E. Newman, "A network-based ranking system for us college football," *Journal of Statistical Mechanics: Theory and Experiment*, Vol.10, 2005.

[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the ACM International Conference on Multimedia*, pp.675-678, 2014.

[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of Computer Vision and Pattern Recognition*, pp.886-893, 2005.

[18] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision,* Vol.42, No.3, pp.145-175, 2001.

[19] F. Perronnin, J. S´anchez, and T. Mensink, "Improving the fisher kernel for largescale image classification," in *Proceedings of the European Conference on Computer Vision*, pp.143-156, 2010.

[20] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition," in *arXiv preprint arXiv:1310.1531*, 2013.

[21] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin, "LIBLINEAR: A library for large linear classification," *The Journal of Machine Learning Research*, Vol.9, pp.1871-1874, 2008.

강 현 우

e-mail : znxlwm@catholic.ac.kr
2015년 가톨릭대학교 디지털미디어학부 (학사)
2015년~현 재 가톨릭대학교 디지털미디어과 석사과정
관심분야 : Computer Vision, Artificial Intelligence, Machine Learning, Big Data

강 행 봉

e-mail : hbkang@catholic.ac.kr
1980년 한양대학교 전자공학과(학사)
1986년 한양대학교 전자공학과(석사)
1989년 Ohio State Univ. 컴퓨터공학(석사)
1993년 Rensselaer Polytechnic Institute 컴퓨터공학(박사)
1993년~1997년 삼성종합기술원 수석연구원
1997년~현 재 가톨릭대학교 디지털미디어학부 교수
관심분야 : Computer Vision, Machine Learning, HCI, Artificial Intelligence, Computer Graphics, Big Data