# 비디오 행동인식을 위한 효과적인 딥러닝 알고리즘

*차상국 **한종기

세종대학교

*sgspet29@naver.com, **hjk@sejong.edu

# Efficient Deep-learning Algorithm for Action Recognition in Video

*Cha, Sangguk **Han, Jong-Ki

Sejong University

## 요약

본 논문은 비디오기반 행동인식을 연구하였으며, 기존의 구조를 참조하여 더 높은 인식률을 위한 새로운 구조를 제안한다. 딥러닝의 기본인 CNN과 RNN을 베이스로 한 구조이며 UCF-101 이라는 Data Set를 사용하였다.

## 1. 서론[1)]

2016년 3월 인공지능 프로그램 알파고가 프로기사 이세돌 9단을 이긴 '알파고 쇼크' 이후 사람들은 인공지능의 위력에 충격을 금치 못했다. 이 대국 이후 인공지능에 대한 사람들의 관심은 높아졌고, 다양한 분야에서 인공지능을 활용한 연구는 활발해 졌다.

딥러닝을 공부하기 위한 Data Set들 또한 만들어 졌으며, 그 종류와 양은 어마어마 하다. 본 논문에서 사용한 UCF-101 Data Set은 행동인식을 위한 비디오이다. 행동인식은 범죄 예방이나 로봇에 활용할 수 있으며, 현재 다양한 분야에서 사용되고있는 주요한 기술이다.

과거에는 SIFT, SURF, HoG 등 Hand-crafted 방식의 필터를 사용하여 특징을 추출한 뒤 분류하던 방식에서 현재는 특징 추출 및 분류를 딥러닝 구조 내에서 한번에 해결하고 있다. 본 논문 또한 딥러닝을 사용하여 비디오의 행동을 인식하였으며, 그 구조를 소개하고자 한다.

## 2. 기존의 딥러닝 구조

UCF-101 Data Set은 13320개의 짧은 비디오가 101개의 클래스로 구성되어 있으며, 딥러닝의 가장 기초구조인 Deep Neural Net은 연산의 숫자가 상상을 초월할 정도로 많기 때문에 막대한 양의 데이터를 학습하기에는 무리가 있다. 이러한 문제점을 해결하기 위한 구조로 CNN 과 RNN이 있다.

### 2.1. CNN

CNN이란 Convolution Neural Network의 줄인말로 데이터 처리를 위한 연산을 Convolution으로 한다. Convolution 과 Pooling을 기본으로한 구조와 기존의 Neural Net을 아주 작은 층을 사용하여 Full-connected Layer로 사용하며 그림 1. 과 같이 나타난다.

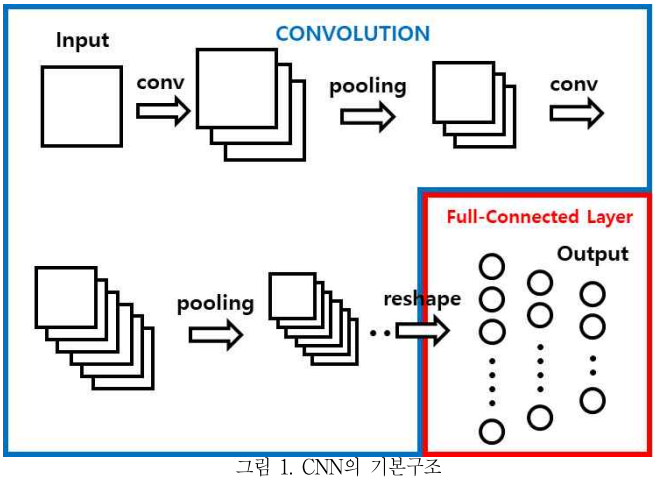


그림 1. CNN의 기본구조

입력데이터가 convolution 과 pooling 과정을 거친 뒤 간단한 Neural Net에 값이 전달되어 출력데이터를 만들어 내는 과정이다.

### 2.2. LSTM

LSTM은 RNN의 변형으로 시퀀스 데이터를 학습하기 위한 구조이다. RNN 구조로 학습을 할 때 관련정보와 그 정보를 사용하는 지점 사이의 거리가 멀어질 경우 초기 입력값을 기억하지 못하는 Vanishing gradient problem 이라는 문제점을 가지고 있기 때문에 긴 시퀀스를 가지는 데이터에서는 이러한 문제점을 해결하기 위해 RNN의 Hidden state에 cell-state를 추가한 구조인 LSTM을 사용한다. LSTM은 다음과 같은 식과 그림 2.와 같은 구조를 가진다

$$\text{input gate} : i_t = \sigma((x_t + s_{t-1})W^i + b_i \quad (1)$$

$$\text{forget gate} : f_t = \sigma((x_t + s_{t-1})W^f + b_{fi} \quad (2)$$

1) 연락저자: 한종기

Acknowledgement: 본 연구는 과학기술정보통신부의 재원으로 한국연구재단의 지원(NRF-2018R1A2A2A05023117)과 정보통신기술진흥센터의 지원(No. 2017-0-00486)을 받아 수행된 연구임

output gate : $o_t = \sigma((x_t + s_{t-1}) W^o + b_o$ (3)

frame state : $s_t = \tanh(c_t) \bullet o_t$ (4)

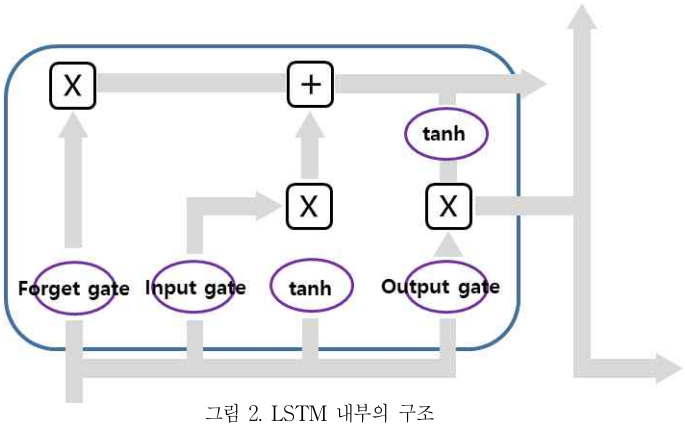


그림 2. LSTM 내부의 구조

cell state가 일종의 컨베이어 벨트 역할을 하여 학습이 깊어지더라도 초기 입력 값의 전파가 비교적 잘 되는 것이 LSTM의 특징이다.

### 2.3. CNN + LSTM

행동인식을 위해서는 다음 그림 3. 과 같은 CNN과 LSTM을 기본으로 한 구조를 가진다. 이미지 처리를 하는 CNN 부분과 시퀀스 데이터학습을 위한 LSTM을 사용하며, 본 논문에서 제안하는 구조도 각 프레임이 CNN을 거친 뒤 LSTM을 통해 시퀀스의 연관성을 학습하는 것을 목표로 한다.

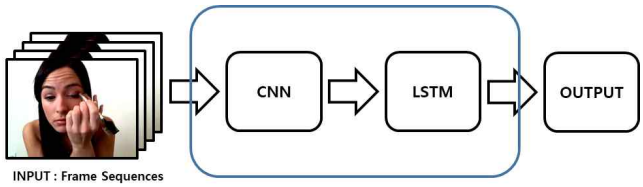


그림 3. 행동인식을 위한 기본적인 딥러닝 구조

## 3. 제안하는 방법

본 논문에서 제안하는 구조는 Deep Bidirectional - LSTM (DB - LSTM) 구조에서 착안하였다. DB-LSTM에서 제안된 구조는 CNN구조와 Foward Layer 와 Backward Layer로 이루어진 2개의 LSTM을 사용하는 구조이다. 하지만 이 구조의 CNN 부분은 AlexNet에서 착안한 것으로 층이 5개로 매우 얇다는 단점을 가지고 있다. 이는 언제든 이 부분의 학습을 강화시키면 학습률이 올라간다는 뜻을 가지고 있다. 나는 이 부분을 조금 더 깊게 짜는 것부터 시작 하였지만 일정 수준 이상으로 깊어지면 더 이상 학습률이 증가하지 않고 일정수준을 유지한다는 것을 알 수 있었다. 이러한 값의 변화를 통해 가장 적절한 층의 개수를 정하였다. 또한 GoogleNet에서 제안된 것과 같이 다양한 종류의 필터와 Pooling을 사용하여 성능향상을 시키려고 했다.

그리고 학습을 위해서는 입력 데이터의 수가 같아야 했다. 하지만 해당 Data Set의 영상은 길이가 각각 다르기 때문에 입력 데이터의 길이를 일정하게 셋팅 해야했다. 이 문제를 해결하기 위한 기존의 방법으로는 모든 영상에서 동일한 프레임 수의 영상을 추출하거나, 가장 짧은 동영상의 길이에 모든 영상을 맞추는 방식 등이 있었다. 하지만 이러한 방식들은 많은 데이터가 소실되는 단점이 있었다. 이를 해결하기 위한 방법으로 동영상을 1초 단위로 학습하는 방법을 사용하였다.

## 4. 실험

직접 구현해본 DB-LSTM은 제시된 것 만큼의 정확도는 나오지 않았다. 이미지의 전처리 과정이나 구현하는 방식이 다르기 때문에 똑같이 구현할 수는 없었다. 나는 직접 재현한 DB-LSTM구조의 성능을 향상시키는 방법을 찾아 실험하였으며 결과는 표 1.과 같다.

표 1. DB-LSTM 구조에 따른 정확도 단위 : %

| Method | 정확도 |
|---|---|
| DB-LSTM | 91.21 |
| 직접 재현한 DB-LSTM | 67.2 |
| 새롭게 제안한 DB-LSTM | 73.1 |

직접 재현한 DB-LSTM 구조는 기존의 구조만큼의 정확도는 나오지 않았으나 CNN 구조를 변형하여 새롭게 제안한 구조은 직접 재현한 DB-LSTM 구조보다 약 6%의 정확도가 증가하였다. 단순한 CNN 구조도 학습을 하는데 지장은 없었지만, 조금 더 깊은 층의 CNN과 다양한 종류의 필터와 pooling을 도입한 구조는 행동인식 학습률 향상에 도움을 주었다.

## 5. 결론

본 논문에서는 비디오에서의 행동인식을 위한 구조인 DB-LSTM의 단점을 보완한 새로운 구조를 제시하였다. 새롭게 제시된 구조는 각 영상을 초 단위로 학습하며 단순한 CNN 구조를 더 효과적인 구조로 개선하려고 하였다. 여기서 새롭게 제안된 CNN 구조는 기존의 DB-LSTM의 성능 향상에 도움이 될 것으로 보인다. 이러한 행동인식 기술의 발전은 CCTV 분석을 통한 실시간 범죄현장에 도움을 줄 수 있을 것이고, 더 나아가 사람의 행동에 반응하는 인공지능을 만들기 위한 데이터로써 큰 역할을 할 것이다. 우리는 여기서 만족하지 않고 더 좋은 성능을 위해 Dense-Net 이나 ResNet 등을 참고로 하여 새로운 구조를 제작할 예정이다

## 참고문헌

[1] Amin Ullah, Jamil Ahmad, Khan Muhammad, Muhammad Sajja, Sung Wook Baik, Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features, IEEE Journals & Magazines, Volume: 6, p.1155-1166, 2018

[2] Huafeng Chen, Jun Chen, Ruimin H,; Chen Chen, Zhongyuan Wang, Action recognition with temporal scale-invariant deep learning framework, IEEE Journals & Magazines, Volume: 14, p.163-172, 2017