커뮤니케이션 이론
2022년 18권 1호 111~154쪽
https://doi.org/10.20879/ct.2022.18.1.111

미디어 인공지능
: 컴퓨터 비전 관련 딥러닝 모델의 미디어 동영상 분야 적용 가능성에 관한 연구*

박대민 선문대학교 미디어커뮤니케이션학부 조교수**

미디어 동영상 분야는 컴퓨터 비전 관련 딥러닝 모델을 활용해 연구 차원에서는 동영상의 자동화된 내용분석을 수행하고 실무 차원에서는 미디어 분야의 디지털 전환을 통해 서비스를 개선할 여지가 큰 영역이다. 이 논문에서는 미디어 동영상의 분석과 생성에 활용도가 높은 비전 관련 딥러닝 기반 모델을 검토했다. 우선 다양한 모델의 기축이 되는 알고리즘으로서 분류 모델로 널리 사용되는 합성곱 신경망(CNN)과 순환 신경망(RNN), 생성 모델로 사용되는 적대적 생성 신경망(GAN)과 오토인코더(AE), 사전 훈련 모델을 활용하는 전이학습을 살펴보았다. 다음으로 미디어 동영상 영역에서 활용도가 높은 과업을 객체탐지, 행동인식, 사건탐지, 동영상 요약, 동영상 분류 등 5개 대분류와 객체탐지, 안면인식, 표정인식, 랜드마크 인식, 상품인식, 행동인식, 자세추정, 이상탐지, 상황인식, 동영상 요약, 동영상 분류 등 11개 소분류로 제시했다. 이어 각 과업별 SOTA(state-of-the-art)와 벤치마크 데이터셋을 소개했다. 끝으로 이러한 모델의 학문적, 실무적 활용 가능성을 제시해보았다. 본 논문은 수식이나 프로그래밍에 대한 지식이 없이 미디어 연구자나 미디어 서비스 기획자가 비전 분야 딥러닝의 큰 흐름을 파악하고 관련 모델을 직접 활용하거나 컴퓨터공학 분야의 연구자 또는 개발자와 협업할 때 배경지식을 제공할 것으로 기대한다. 또한 비전 관련 딥러닝이 발전함에 따라 미디어 인공지능 기반 동영상 빅데이터 분석 시스템의 개발 가능성도 높아질 것이다.

---

* 이 연구는 2022학년도 선문대학교 교내학술연구비 지원에 의하여 이루어졌음.

** dmpark@sunmoon.ac.kr

KEYWORDS 딥러닝, 컴퓨터 비전, 합성곱 신경망, 순환 신경망, 적대적 신경망, 미디어 인공지능, 동영상 내용분석

1. 문제제기

2012년 합성곱 신경망(convolutional nueral network, CNN)을 사용한 AlexNet이 등장한 이후 딥러닝(deep learning)은 컴퓨터 비전(Computer Vision, CV) 전반에 혁신적인 발전을 가져왔다. 이에 따라 미디어 동영상 서비스에서도 인공지능(Artificial Intelligence, AI) 기반 서비스를 제공할 토대가 마련되고 있다. 최근에는 미디어 연구에서도 텍스트를 넘어서 방송 동영상 등의 분석에 딥러닝 기반의 동영상 내용분석(Video Analytics, VA)을 적용하려는 시도가 나오고 있다(윤호영, 2021; 최윤정 등, 2020).

그러나 미디어 연구자가 딥러닝을 활용해 영상을 분석하기가 쉽지는 않다. 특히 동영상의 경우, 과업(task)에 따라서는 사진 간의 시계열 정보를 추가로 학습하고, 분석 결괏값을 준 실시간에 가깝게 내놓아야 한다는 점에서 난이도가 높다. 게다가 비전 분야 딥러닝 연구는 매우 방대할 뿐만 아니라 하루가 다르게 최고 성능의 최신 모델(state-of-the-art, SOTA)[1]이 갱신되고 있다[2].

다행히 비전 분야의 연구는 큰 줄기가 되는 알고리즘의 계열이 존재한다. SOTA는 이러한 계열들을 조합하거나 응용한 결과물인 경우가 많다. 따라서 미디어 연구자는 우선 토대가 되는 알고리즘 계열을 이해한 뒤, SOTA 중에 본인 연구에 맞는 최적의 모델을 검토하는 식으로 비전 분야의 딥러닝을 활용할 수 있다.

---

1) 벤치마크 데이터셋으로 학습하고 평가하여 F1 점수(F1 score), mAP(mean average precision) 등 성능 지표 점수를 기준으로 가장 높은 점수를 갖는 모델을 뜻한다.

2) 예컨대 해외의 AI 연구 포털 서비스인 페이퍼스위드코드닷컴(paperswithcode.com)에 소개된 비전 분야 연구만 해도 939개 과업, 2,630개 벤치마크 데이터셋(benchmark dataset), 22,962개의 논문에 달한다. 또한 비전 분야의 최상위권 학술대회인 CVPRC(onference on Computer Vision and Pattern Recognition)에서는 2021년 1,600개가 넘는 연구 성과를 공개하고 있다.

이 연구는 컴퓨터 비전 분야의 딥러닝 기술 중 미디어 인공지능(media artificial intelligence), 즉 미디어 연구 분야에서 특히 동영상 관련 내용분석 연구, 또는 방송 등에서 AI 기반 미디어 동영상 서비스 개발에 활용도가 높은 모델들을 검토하고자 한다. 이를 위해 우선 2장에서 비전과 딥러닝의 기본 개념을 간략히 살펴본다. 이어 3장에서 비전 분야 딥러닝의 주요 계열을 소개한다. 이 연구에서 우선 검토할 비전 분야 딥러닝의 큰 줄기가 되는 알고리즘은 크게 네 가지다. 첫째, 객체 탐지에 널리 활용되는 CNN이다. 이 논문에서는 CNN의 기본 원리와 AlexNet에서 VGG, ResNet, R-CNN, YOLO 등 CNN의 발전을 훑어본다. 둘째, 시계열 데이터 분석에 사용되는 순환 신경망(Recurrent Neural Network, RNN)이 있다. RNN은 LSTM과 GRU로 발전한다. 셋째, 영상 생성에 널리 활용되는 적대적 생성 신경망(Generative Adversarial Network, GAN)과 오토인코더(Autoencoder, AE)를 살펴본다. 넷째, 사전 훈련 모델(pretrained model)을 활용해 모델을 효율적으로 개선하는 전이학습(transfer learning)을 다룬다[3]. 이 논문에서는 특히 계열별 알고리즘이 어떤 내적 한계들을 극복하면서 발전해왔는지를 최대한 유기적으로 기술하고자 한다. 이를 통해 미디어 연구자가 자신의 연구 목적에 맞는 적정 모델을 선별하는데 도움을 주고자 한다. 다음으로 4장에서는 미디어 동영상 분야에서 활용도가 높은 주요 과업을 정해, 과업별 SOTA를 살펴본다. 그러한 과업으로 객체와 관련된 객체탐지(object detection), 안면인식(face recognition), 표정인식(facial expression recognition), 랜드마크 인식(landmark recognition), 상품인식(product recognition), 행동과 관련된 행동인식(action recognition), 자세추정(pose estimation), 사건(event)과

3) 일반적으로 CNN과 RNN은 분류 모델, GAN과 AE는 생성 모델로 간주되지만, GAN을 분류에 사용하거나 RNN을 생성에 활용할 수도 있다.

관련된 상황인식(context recognition), 이상탐지(anomaly detection), 그리고 동영상 요약(video summarization), 동영상 분류(video classification) 등 11개 과업을 선정했다. 이들 과업은 미디어 분야와 관련성이 높으면서, 동시에 한국지능정보사회진흥원이 구축한 AI 통합 플랫폼인 AI허브(aihub.or.kr)에 관련 과업의 AI학습데이터가 공개된 경우를 우선 선정했다. 이렇게 함으로써 본 연구에서 검토한 모델의 경우 일반적으로 활용되는 벤치마크 데이터셋에 더하여 국내 미디어 분야에 맞는 AI학습데이터도 추가로 얻을 수 있어 그 활용도가 더욱 높을 것으로 기대된다. 과업별 SOTA의 기본은 CNN, RNN, GAN, AE 등 주요 알고리즘을 조합하고 응용해 최적 성능을 내도록 개선한 것이다. 따라서 주요 알고리즘을 이해하고 있다면 대부분의 SOTA의 원리도 파악할 수 있다. 다만 과업별 SOTA는 수시로 갱신되므로 연구자는 실제 연구 시점에서 SOTA를 검토할 필요가 있다. 이 연구에서는 충분히 뛰어나면서도 미디어 서비스에서 활용 가능할 정도로 널리 사용되어 검증된 모델을 우선 소개한다. 즉 페이퍼스위드코드에서 높은 활용도와 관심도를 보여주거나, AI허브에서 AI학습데이터셋의 검증용 모델로 활용된 경우를 중심으로 간략히 살펴본다. 이어 미디어 연구에서 활용 방안을 모색한 뒤 딥러닝 기반 모델을 적용할 수 있는 미디어 서비스의 예를 제시한다. 대표적으로는 동영상 내용분석, 영상검색, 영상추천, 영상편집, AI 인간의 활용 등을 들 수 있다.

2. 컴퓨터 비전과 딥러닝의 기본 개념

이 장에서는 비전 분야 딥러닝의 발전을 살펴보기에 앞서, 컴퓨터 비전의 기본 개념과 적용 범위, 그리고 딥러닝의 기본 개념과 딥러닝의 토대가 되는 퍼셉트론의 발전에 대해 살펴본다.

1) 컴퓨터 비전의 기본 개념

사진이나 동영상을 다양한 연산을 통해 목적에 맞게 컴퓨터로 처리하는 것을 디지털 영상처리(digital image processing) 또는 컴퓨터 비전이라고 한다. 디지털 영상처리가 가능한 것은 영상을 디지털화하여 함수로 표현할 수 있기 때문이다. 디지털 영상을 함수로 표현함으로써 딥러닝에 필요한 다양한 연산이 가능해진다. 디지털 영상의 함수 표현과 연산에 대해 좀 더 자세히 살펴보자. 영상은 시간에 따라 진폭이 변화하는 파장 형태의 가시광선으로 이뤄진다. 디지털 영상(digital image)은 아날로그 영상(analogue image)을 디지털 변환[4] 한 것이다. 아날로그 영상은 가시광선의 파장 변화를 시간에 따라 연속적으로 변화하는 실수 형태로 기록한다. 반면 디지털 영상은 보통 RGB, 즉 적녹청의 색상 채널을 사용하며 각 채널의 데이터 값을 정숫값으로 표현한다. 예를 들어 24비트 트루 컬러(true color)의 경우 색상 채널당 0에서 255 사이의 정숫값을 갖는다. 디지털 영상의 기본 단위는 화소(pixel)이다. 디지털 영상은 결국 영상을 화소 단위로 쪼갠 뒤 각 화소에 색상 채널별로 정숫값을 할당한 데이터이다.

영상을 단순한 정수로 표현할 수 있게 된 덕분에, 영상의 함수 표현이 가능해진다. 디지털 영상의 함수 표현은 디지털 영상을 2차원 평면상에서 입력값과 출력값의 함수로 처리할 수 있게 표현하는 것을 뜻한다. 영상이 표현된 2차원 평면은 가로 x 축, 세로 y축으로 나타낼 수 있다. 입력값을 위치 (x, y)로 주면, 해당 위치에 있는 화소의 RGB 값이 출력

---

4) 디지털 변환 과정은 표본화(sampling), 양자화(quantization), 부호화(coding)로 이뤄진다. 표본화는 연속적인 아날로그 영상에서 표본 값을 추출하는 것이다. 표본값은 일정한 시간의 표본화 주기에 따라 추출한다. 표본화된 영상 신호의 시간대별 진폭 크기는 아직 실숫값으로 표현된다. 양자화는 이 실숫값을 근사한 정숫값으로 표현하는 것이다. 끝으로 부호화는 정숫값을 다시 2진법으로 표현하는 과정을 뜻한다. 표본화는 시간 축에 따른 데이터를, 양자화는 진폭 축에 따른 데이터를 연속형에서 이산형으로 바꾸는 과정이다. 부호화는 이산형으로 표현된 데이터를 이진 데이터로 바꾸는 과정이다.

값으로 나온다. 영상의 함수 표현으로 다양한 영상의 연산이 가능해진다. 예컨대 검정의 RGB 값은 (0, 0, 0)이며, 흰색의 RGB 값은 (255, 255, 255)이다(황선규, 2019). 어떤 영상의 화소값 전체에 양의 정수를 더하면 영상이 전체적으로 흰색에 가까워지는 효과, 즉 밝아지는 효과를 줄 수 있다. 뺄셈의 경우 영상이 어두워지며, 곱셈은 밝기 차이가 더욱 벌어지면서 영상이 뚜렷해지고, 나눗셈은 영상이 희미해진다(이희석, 2016). 비단 사칙연산만이 아니라 딥러닝에서 널리 활용되는 행렬 연산, 미적분 등도 가능하다.

디지털 영상처리는 영상 변환(image transformation), 영상 분석(image analysis), 영상 인식(image recognition)으로 유형화된다. 영상을 합성하는 등 영상을 만들어 내는 생성(generating)도 넓은 의미의 디지털 영상 처리에 포함될 수 있다. 영상 변환은 화소의 RGB 값이나 화소의 위치를 변경하여 입력 영상의 특성을 변형하는 것을 뜻한다. 영상 변환은 광학적 변환과 기하학적 변환으로 나눌 수 있다. 광학적 변환은 화소의 RGB 값을 변경하는 방식이다. 밝기나 명암비 조절, 영상의 잡음 제거, 영상의 보정이나 강조 등에 활용된다. 기하학적 변환은 화소의 위치를 변경하는 방식으로 영상을 변경한다. 영상분석은 입력 영상의 물리적 정보를 추출하는 과정이다. 경계선 검출(edge detection)이나 외곽선 검출(counter detection), 직선 및 원 검출, 코너나 크기 불면점 등을 찾는 특징 추출(feature extraction), 방향 등 공간 정보 파악, 서로 다른 두 영상의 같은 지점 등을 찾는 정합(registration) 등이 영상분석의 주요 과업이다(이희석, 2016).

영상 인식은 입력 영상의 의미 정보를 추출하는 과업이다. 영상 인식은 영상이 담고 있는 객체(object)가 어떤 유형, 즉 어떤 클래스(class)인지 파악하는 분류(classification), 위치를 파악하는 로케이션(location), 여러 객체의 종류와 위치를 구별해 파악하는 탐지(detection), 객체를 배경과 구분해 파악하는 분할(segmentation)로

나눠진다. 분할은 다시 서로 다른 객체라도 배경만 아니면 하나로 묶어주는 의미론적 분할(semantic segmentation)과 서로 다른 객체를 구분하는 인스턴스 분할(instance segmentation)로 세분된다(Khan & Al-Habsi, 2020). 영상 인식의 주요 객체로는 인간, 동물, 인공물(artifact), 랜드마크(landscape) 등이 있다. 영상 인식의 대상으로는 객체 외에 행동(action)이 있다. 수화(sign language)와 같은 몸짓(gesture)의 인식은 행동 인식의 특별한 사례이다. 교통사고나 화재, 범죄 등 사건 탐지(event detection)는 객체와 행동의 인식을 통해 수행한다. 상황(context)은 행동 외에도 계절이나 밤낮 등 시간과 장소, 발화 내용 등 의미를 활용해 파악한다.

2) 딥러닝의 기본 개념

딥러닝은 다수의 은닉층(hidden layer)이 있는 인공신경망을 활용하는 기계학습 방법이다(LeCun, Bengio, & Hinton, 2015). 기계학습은 데이터, 즉 입력값과 출력값(정답)을 제공하면 컴퓨터가 스스로 입력값과 출력값 간의 함수 관계를 찾아내는 방법을 뜻한다. 기계학습은 기본적으로 예측(prediction) 문제를 다룬다. 기계학습은 크게 지도학습(supervised learning)과 비지도학습(unsupervised learning)으로 나눠진다. 지도학습은 객체 등의 클래스가 사전에 주어진, 즉 라벨링(labeling)된 학습데이터를 바탕으로 학습한다. 지도학습은 실수형 변수에 대한 예측인 회귀(regression)나 이산형 변수에 대한 예측인 분류에 사용된다. 비지도학습은 라벨링되지 않은 데이터로 학습하며 클래스도 사전에 주어지지 않는다. 비지도학습은 기본적으로 군집화(clustering)에 활용된다. 강화학습(reinforcement learning)은 넓게 보면 비지도학습의 일종으로 역시 정답이 사전에 주어지지 않는다. 대신 게임의 규칙 등 주어진 환경에서 보상을 극대화하는 데이터를 수집하는 과정을 포함한다(조태호, 2020).

딥러닝 이전의 가장 기본적인 인공신경망은 입력층(input layer)과 출력층(output layer)으로만 이루어진 단층 퍼셉트론(perceptron)이다(Rosenblatt, 1958). 단층 퍼셉트론을 활용한 학습은 1차 방정식을 푸는 과정으로 볼 수 있다. 즉 여러 입력값 x과 출력값 y가 주어질 때 가장 많은 문제를 근사적으로 풀 수 있는 1차 방정식의 기울기와 절편을 구하는 것이다. 이때 기울기에 해당하는 가중치와 절편에 해당하는 바이어스(bias)가 학습을 통해 찾아야 할 파라미터(parameter)이다. 딥러닝의 문제는 예측의 오차를 최소화하는 파라미터를 찾는 최적화(optimization) 문제다[5].

단층 퍼셉트론은 인간 신경망 1개를 모방한 구조를 갖고 있다(Rosenblatt, 1958). 단층 퍼셉트론의 정보 전달 방식은 인간이 자극을 전달하는 과정과 유사하다. 입력값에 가중치를 곱하고 바이어스를 더한 가중합이 임계치보다 크면 활성화하고 임계치보다 작으면 비활성화한다. 인간 신경망이 역치를 넘는 자극만 다음 신경세포에 전달하는 것과 비슷하다. 활성화 여부를 결정하는 함수가 활성함수(activation function)다. 활성함수는 보통 활성화되면 1의 값을, 비활성화되면 0의 값을 출력한다. 초기에는 활성함수로 시그모이드(sigmod) 함수[6]가 사용됐다.

---

5) 가장 간단한 최적화 방법으로는 최소제곱법(least squares method)이 있다. 데이터의 x값에서 주어진 y값과 회귀식의 y값 간의 차이를 제곱의 합인 평균 제곱 오차(mean square error, MSE)를 최소화하는 파라미터를 찾는 방법이다. 다만 최소제곱법은 입력값이 한 종류인 단순 선형 회귀에만 적용된다. 입력값이 여럿인 다중 선형 회귀에서는 경사하강법(gradient descent)을 사용한다. 그 절차는 다음과 같다. 우선 임의의 가중치를 대입하고 학습데이터에서 주어진 x, y값들을 입력해 MSE를 구한 뒤 다른 가중치를 대입하고 MSE를 구하고, MSE가 작은 쪽의 가중치를 선택한다. 이때 대입하는 파라미터 간의 거리는 학습률(learning rate)이라고 한다. 이러한 과정을 반복해 MSE가 최솟값으로 수렴하는 가중치를 구한다. MSE가 최소가 되는 지점에서 가중치의 미분 값은 0이 된다. 여기서 반복 횟수는 에폭(epoch)이라고 한다. MSE 외의 다른 손실함수(loss function)를 비용함수(cost function)로 사용할 수도 있다(조태호, 2020).

6) 시그모이드 함수는 S자 곡선 모양으로 함숫값이 0과 1 사이에 모여 있으며 0과 1 값 근처의 기울기는 완만하고 그 외의 부분은 경사가 급한 형태다.

그러나 단층 퍼셉트론은 XOR 문제와 같은 간단한 문제도 풀지 못하는 한계를 갖고 있다(Minsky & Papert, 1972). XOR 연산이란 2개의 입력값 중 하나만 참일 때 참의 값을 출력하는 연산을 뜻한다[7]. XOR 문제는 다층 퍼셉트론으로 해결했다. 다층 퍼셉트론은 단층 퍼셉트론과 달리 입력층과 출력층 사이에 은닉층이 추가돼 있다. 이는 x축과 y축만으로 이루어진 2차원 평면을 z축이 추가된 3차원 공간으로 만든 뒤, 선이 아닌 면을 이용해 입력값에 따른 출력값을 분류하는 것과 같다.

하지만 다층 퍼셉트론 역시 순전파만으로는 은닉층의 파라미터를 찾기 어렵다는 문제를 갖고 있었다. 이 문제는 오차 역전파(error backpropagation)로 풀었다(Rumelhart, Hinton, & Williams, 1986). 우선 순전파를 통해 입력값을 초기 가중치에 따라 계산해 출력값을 생성하고 이를 실제값, 즉 정답과 비교해 오차를 산출한다. 이어 역전파 과정을 통해 결괏값에 영향을 주는 출력층과 은닉층의 가중치들을 오차를 줄이는 방향으로 업데이트한다. 이 과정을 일정 횟수 반복하면서 오차가 최소화되면 학습이 완료된다.

역전파 또한 기울기 소실(gradient vanishing) 또는 폭주(gradient exploding) 문제가 있었다(Hochreiter, 1998). 기울기 소실이란 활성함수를 시그모이드 함수를 쓴 채로 역전파를 반복하면 입력층에 가까울수록 가중치의 기울기가 0에 수렴해 학습이 중단되는 것을 의미한다[8]. 폭주는 기울기 소실과 반대로 기울기가 폭증하고 발산하는 문제다. 이러한 문제를 해결하기 위해 초기에는 활성함수로 하이퍼볼릭 탄젠트 (hyperbolic tangent) 함수[9]를 도입했다. 현재는 이보다 간단한

---

7) 단층 퍼셉트론으로 2개의 입력값이 모두 참(1)일 때 참을 출력하는 AND 게이트와 2개의 입력값 중 하나라도 참일 때 참을 출력하는 OR 게이트의 연산은 가능하다.

8) 미분 값이 0~0.25에만 있으므로 경사하강법 적용 시 세 번 이상 미분하면 0에 가까워진다.

9) 함숫값은 -1~1 사이에 있고 S자형이면서 미분 값은 0~1 사이에 있다.

ReLU(Rectified Linear Unit) 함수[10]가 널리 사용된다(Nair & Hinton, 2010). 오차역전파의 또 다른 문제는 경사하강법을 실행할 때 너무 시간이 많이 소모되고, 그렇다고 속도를 높이면 정확도가 떨어진다는 점이었다. 이 문제를 해결하기 위해 모멘텀(SGD with momentum), RMSProp, Adam 등의 고급 경사 하강법을 사용한다[11]. 역전파 문제가 해결되면서 딥러닝은 본격적인 발전의 토대를 마련했다.

3. 컴퓨터 비전에서 딥러닝의 적용과 발전

3장에서는 다양한 SOTA의 기본이 되는 CNN, RNN, GAN, AE, 전이학습에 대해 살펴본다. 기본적으로 CNN과 RNN은 분류 모델로, GAN과 AE는 생성 모델로 볼 수 있지만 분류 모델이 생성에, 생성 모델이 분류에 사용되기도 한다. 전이학습은 사전 훈련 모델을 활용해 모델의 성능을 효율적으로 개선한다.

1) 합성곱 신경망

다층 퍼셉트론은 완전연결층(fully-connected layer, FCL), 즉 한 계층의 모든 노드가 다음 계층의 모든 노드와 1대 1로 완전히 연결된 형태를 띤다. 이렇게 되면 망의 크기가 너무 커지고, 구해야 할 파라미터 수가 너무 많아지며, 그 결과 학습 시간이 너무 길어진다. 주어진 학습데이터에서만 고성능을 발휘하고 조금만 다른 상황이 되면 성능이 저하되는 과적합(overfitting)도 쉽게 발생한다. 특히 영상은 근접한 화소 간의 영향력이 크지만, 멀리 있는 화소 간의 영향력은 적다. 이는 컴퓨터 비전에

10) 함숫값은 0 이상이고 미분 값은 0 또는 1이다.

11) 현재는 모멘텀과 RMSProp을 결합한 Adam이 가장 널리 사용된다.

서 영상을 전역(global)으로 처리하는 것보다는, 처음에는 국소(local)적으로 처리하고 이를 종합하는 것이 더 효율적임을 시사한다. 예를 들어 새를 탐지하려면 배경은 덜 중요하지만, 부리가 가장 중요한 정보일 수 있다[12].

CNN은 영상처리에 합성곱(convolution) 연산을 활용하는 인공신경망 모델이다. CNN은 영상을 여러 개로 분할해 처리한다. 부분의 특성에 따라 영상을 분류하므로 영상의 일부가 달라지더라도 같은 클래스로 분류할 수 있어서 사진이나 동영상에 널리 활용된다.

CNN의 구조는 학습과 분류의 두 단계로 구성된다. 특징 학습(feature learning) 단계에서는 합성곱 연산과 활성함수로 구성된 필터링(filtering) 과정과 연산량을 줄이기 위한 풀링(pooling) 과정이 하나의 합성곱 층을 이룬다. 이러한 합성곱 층이 많을수록 더 깊은 신경망이다. 필터링 과정은 입력 영상(input image)을 작은 정방행렬 형태의 필터(filter)를 사용해 보폭(stride)에 맞춰 이동하면서 내적(inner product)을 구한다. 이 과정을 영상의 모든 영역에 대해 수행해 특징 지도(feature map)를 완성한다. 이때 특징 지도의 크기가 필터링을 할 때마다 점점 작아지는 데이터 손실 문제가 발생한다[13]. 이 문제를 해결하기 위해 필터링에 앞서 입력 영상의 둘레에 일정 폭의 값을 추가하는 패딩(padding) 과정을 거친다[14]. 단계별 필터링 과정에서 영상의 특징을 저수준(low level)의 특징 지도부터 순차적으로 겹쳐가면서 고수준

---

12) 인간이나 고양이와 같은 동물도 이러한 방식으로 시각 정보를 처리한다. 시각 피질의 뉴런은 각자 국부 수용 영역(local receptive field)을 갖는다. 시각 피질 내 뉴런들은 각각 고유한 특정 시각 자극에만 반응한다. 즉 어떤 뉴런은 저수준 패턴 반응(예: 수직, 각도)에만, 다른 뉴런은 고수준의 복잡한 패턴(예: 무늬, 객체)에만 반응한다. 고수준의 반응은 저수준의 반응에 기반한다. 뉴런의 수용 영역들은 겹칠 수 있으며 겹친 수용 영역들이 전체 시야를 이룬다(Hubel & Wiesel, 1959).

13) 필터링을 한 번 거치면 n*n의 행렬이 (n-1)*(n-1)의 행렬로 작아진다.

14) 보통 간단하게 '0'의 값을 갖는 1폭짜리 제로 패딩(zero padding)이 사용된다.

(high level)의 특징 지도를 산출한다. 문제는 필터링을 한 번 거칠 때마다 하나의 특징 지도가 산출되기 때문에, 하나의 입력 영상에 필터링 횟수만큼의 특징 지도가 생겨난다는 점이다. 이렇게 되면 학습해야 할 파라미터의 수가 너무 많아진다. 따라서 차원 축소(dimension reduction)를 위해 정방행렬 모양의 필터를 사용해 특징 지도의 크기를 줄이는 풀링 과정을 거친다[15]. 보통 필터링하는 영역에서 간단히 가장 큰 값을 선택하는 맥스 풀링(max pooling)이 널리 사용된다. 풀링 층은 단순히 특징 지도의 크기만 줄여주기 때문에 학습할 파라미터는 없다. 풀링 층은 필요에 따라 생략할 수 있다. 최종적으로 분류 단계에서는 특징 학습의 결괏값을 1차원 벡터로 평탄화(flatten)한 뒤 완전연결층을 거쳐 분류 결과를 출력한다(Géron, 2019/2020).

CNN은 1989년 르쿤 연구팀이 숫자 필기체 인식에 적용하면서 처음 공개됐다(LeCun et al., 1989). 이후 2012년 8개의 은닉층을 쓰는 AlexNet이 ILSVRC에서 객체 인식에서 압도적인 성능으로 우승하며 CNN이 각광 받기 시작했다(Krizhevsky, Sutskever, & Hinton, 2012). 2014년 ILSVRC에서는 22개의 은닉층을 사용한 GooLeNet과 19개의 은닉층을 사용한 VGG19(Very Deep Convolutional Networks)[16]가 근소한 차이로 1위와 2위를 차지했다(Simonyan & Zisserman, 2014; Szegedy et al., 2015). GooLeNet은 1*1 합성곱 필터가 포함된 인셉션 모듈(inception module)을 사용했다. 이를 통해 같은 깊이의 신경망이라고 해도 구해야 할 파라미터 값은 크게 줄일 수 있었다.

기존 신경망보다 많은 은닉층을 사용한 VGG와 GoogLeNet은 더

15) 2*2 크기의 필터를 사용하면 특징 지도의 크기는 반으로 줄어든다.

16) VGG 뒤의 숫자는 은닉층의 개수를 나타낸다. VGG19 외에 VGG16이 널리 사용된다.

깊은 층을 사용하면 모델 성능이 높아질 수 있다는 점을 보여줬다. 그러나 실제로는 층을 너무 많이 쌓게 되면 오히려 정확도(accuracy)가 떨어지는 문제(degradation problem)가 발생한다. 2015년 등장한 ResNet은 우선 층을 무작정 쌓는 것보다 적절한 깊이로 유지할 때 성능이 극대화된다는 것을 보여줬다. 디그레데이션 문제는 잔차 학습(residual learning)으로 해결한다. 즉 이전 은닉층에서 학습한 결과 x를 다음 은닉층으로 그대로 보내어 출력값 H(x)로 내놓는 대신, x를 다음 은닉층의 잔차 학습 결과 f(x)와 합쳐준다. 일종의 지름길 역할을 하는 건너뛰기 연결(skip connection)을 이용하는 것이다. 이렇게 하면 모델은 함수 전체를 업데이트하는 것이 아니라 잔차만 추가 학습하게 되며, 결국 학습이 빨리 완료될 수 있게 된다. ResNet은 이러한 방법으로 깊이를 152개 층까지 늘렸다. ResNet은 ILSVRC에서 인간의 객체인식 오류율(5.1%)보다 낮은 3.6%를 기록하면서 SOTA를 갱신했다(He, Zhang, Ren, & Sun, 2016).

R-CNN(Regions with CNN features)은 분류를 위한 CNN과 위치 파악을 위한 RPN(region proposal network)을 결합한 모델이다. 이를 통해 한 영상에서 여러 종류의 객체를 동시에 인식하는 객체탐지가 가능해졌다(Girshick, Donahue, Darrell,& Malik, 2014). R-CNN은 속도를 개선한 Fast R-CNN, Faster R-CNN, 그리고 R-CNN를 분할에 적용한 Mask R-CNN 등으로 발전했다(He, Gkioxari, Dollár, & Girshick, 2017; Girchick, 2015; Ren, He, Girshick, & Sun, 2015).

R-CNN과 같은 2단계 방식(two-stage methods)은 정확도는 높지만, 속도를 높이는 데에는 한계가 있다. 이 때문에 미디어에서 널리 사용되는 30fps(frame per second) 이상 동영상의 객체탐지에는 적합하지 않다. YOLO(You Only Look Once)와 같은 1단계 방식(one-stage methods)은 실시간 동영상의 객체탐지에 적합하다.

YOLO는 합성곱 층을 하나만 사용한 단일 신경망(single neural network)으로 위치를 나타내는 바운딩박스(bounding box) 구획과 객체 분류를 동시에 수행함으로써 객체탐지 속도를 대폭 개선했다(Redmon, Divvala, Girshick, & Farhadi, 2016). YOLO 모델은 최초 개발자가 ResNet을 응용해 더 깊은 층을 쌓은 YOLOv3까지 업데이트했다(Redmon & Farhadi, 2018). 이후 다른 연구진에 의해 속도 및 정확도를 개선한 YOLOv4와 이전 버전에 비해 용량을 줄이고 속도를 개선한 YOLOv5[17] 등이 공개됐다(Bochkovskiy, Wang, & Liao, 2020).

YOLO는 영상을 화소 7*7 크기의 그리드 셀(grid cell)로 나누어 바운딩박스를 예측한다. 따라서 7*7 크기 이하의 작은 객체탐지가 어렵다. 또한 고도로 추상화된 최종 단계의 특징 지도에서만 객체탐지를 수행하기 때문에 객체탐지 정확도가 다소 낮다. 이를 개선하기 위해 중간 단계의 특징 지도에서도 객체탐지를 진행하는 SSD(Single Shot Multibox Detector)가 제안됐다(Liu et al., 2016). 이어 상위층의 특징들도 함께 이용하는 FPN(Feature Pyramid Network), FPN을 개선한 BiFPN(weighted bi-directional FPN)을 적용한 EfficientDet이 공개됐다(Lin et al., 2017; Tan, Pang, & Le, 2020).

동영상에서는 객체탐지 외에도 프레임이 이동할 때 동일한 객체를 정합시킬 수 있어야 한다. 이를 다중객체추적(Multi-Object Tracking, MOT)이라고 한다. YOLO만으로도 MOT가 가능할 수 있지만, 할당 모델이 없다면 한 객체가 가려졌다가 나타날 때 프레임 간 동일 객체를 제대로 인식하지 못할 수 있다. MOT는 객체탐지 모델(object detection model)과 할당 모델(association model) 등 두 단계로 이뤄진다. 전자는 YOLO와 같이 일반적인 객체탐지를 하는 부분이다. 후자는 객체탐지

17) https://github.com/ultralytics/yolov5

에서 얻은 바운딩박스에 ID를 부여하고 이를 다른 바운딩박스에 재할당(Re-Identification, Re-ID)을 시켜주는 부분이다(Zhang, Wang, Wang, Zeng, & Liu, 2020).

MOT는 처음에는 탐지와 할당을 독립된 단계로 처리했다(Bewley, Ge, Ott, Ramos, & Upcroft, 2016). 이는 2단계 모델(2-stage model)로 불렸다. MOT를 2단계로 처리하는 것은 속도 문제를 초래한다. 이를 해결하기 위해 하나의 신경망에서 탐지와 할당을 한꺼번에 수행하는 JDE(Jointly learns the Detector and Embedding)이 제시됐다(Wang et al., 2020). 그런데 JDE는 객체가 겹쳐 있는 경우 하나의 객체에 여러 개의 바운딩박스가 겹쳐진 경우 재할당 과정이 모호해질 수 있다. 이를 개선한 것이 FairMOT이다. FairMOT에서는 JDE와 달리 탐지와 재할당에 저수준 특징(low level feature)을 이용함으로써 탐지 시 과적합도 줄이고 Re-ID의 정확도도 높이며 학습도 쉬워졌다(Zhang, Wang, Wang, Zeng, & Liu, 2021).

2) 순환신경망

RNN은 이전 은닉층의 출력을 다음 은닉층에서 입력으로 사용함으로써 시계열 데이터를 다룰 수 있게 한다(Rumelhart, Hinton, & Williams, 1986). RNN은 일련의 데이터가 주어졌을 때, 그다음 데이터값을 예측하는 데 적합하다. 이런 RNN을 활용해 데이터를 분류하거나 생성하는데 활용될 수 있다. RNN에 적합한 대표적인 시계열 데이터로는 문장이나 음성을 들 수 있다. 그러나 프레임별 순서가 있는 동영상은 물론 상하좌우로 화소가 배열된 사진도 시계열 데이터로 이해할 수 있다. RNN은 동영상 분류(video classification) 등에 활용된다(Ng et al., 2015). RNN의 과정을 간단히 설명하면 첫 단계(t=0)와 다음 단계(t=1)의 순전파와 역전파의 흐름은 아래와 같이 나타낼 수 있다.

t=0: 일반적인 순전파와 역전파

순전파: 입력층(t=0) → 은닉층(t=0) → 출력층(t=0)

역전파: 출력층 (t=0) → 은닉층(t=0) → 입력층(t=0)

t=1: 이전 단계의 은닉층이 포함됨

순전파: 입력층(t=1) + 은닉층(t=0) → 은닉층(t=1) → 출력층(t=1)

역전파: 출력층(t=1) → 은닉층(t=1) → 입력층(t=1) + 은닉층(t=0)

첫 단계(t=0)에서는 일반적인 순전파 '입력층→은닉층→출력층'과 역전파 '출력층→은닉층→입력층'의 단계를 거친다. 그러나 다음 단계(t=1)에서 순전파의 은닉층에 전 단계 은닉층의 출력과 현 단계 입력층의 입력이 함께 입력된다. 역전파에서는 현 단계의 출력층이 은닉층에 입력되면 해당 은닉층의 출력값이 현 단계의 입력층에 입력될 뿐만 아니라 이전 단계의 은닉층에 입력되어 이전 단계의 가중치를 업데이트한다.

RNN은 기울기 소실 문제로 은닉층 초기의 장기 기억은 사라지고 단기 기억만 남는 장기의존성(problem of long-term dependency) 문제가 심하다. 이를 해결한 것이 LSTM(long short-term memory)이다(Hochreiter & Schmidhuber, 1997). LSTM은 은닉층에 기억을 담아두는 저수지 역할을 하는 기억 셀(memory cell)과 기억에 남길 비율을 정하는 수문 역할을 하는 게이트를 갖는다. 데이터는 과거 기억의 비율을 조정하는 망각 게이트(forget gate), 추가할 새 기억의 비율을 조정하는 입력 게이트(input gate)를 거쳐 기억 셀에 입력된다. 출력 게이트(output gate)는 기억 셀의 데이터를 출력층에 반영하는 비율을 조정한다. 즉 LSTM에서는 게이트 3개와 기억 셀 1개 등 총 4개의 가중치 파라미터 집합을 학습한다(我妻幸長, 2020/2020).

LSTM은 다소 복잡한 구조여서 학습 시간이 길다는 문제가 있다. 이를 해결한 것이 GRU(gated recurrent unit)이다(Cho et al.,

2014). GRU는 업데이트 게이트(update gate)와 리셋 게이트(reset gate)를 갖는 것이 특징이다. 업데이트 게이트는 입력 게이트와 망각 게이트를 합친 형태로, 이전 출력층의 데이터가 현재의 은닉층과 출력층에 영향을 미치는 비율을 조정한다. 리셋 게이트는 기억 셀과 출력 게이트를 합한 구조로, 새로 입력된 데이터가 이전 출력층과 현재 은닉층에 영향을 미치는 비율을 결정한다.

3) 적대적 생성 신경망, 오토인코더

GAN은 가짜 영상을 만드는 생성자(generator)와 생성자가 만든 영상의 진위를 판별하는 식별자(discriminator)가 서로 경쟁하면서 점점 더 진짜 같은 가짜 영상을 생성하는 모델이다. GAN은 강화학습의 일종으로 별도로 라벨링된 학습데이터가 필요하지 않다. GAN의 학습은 식별자와 생성자에서 이뤄진다. 식별자는 진짜 영상을 정답으로, 생성자가 만든 가짜 영상을 오답으로 학습한다. 식별자의 식별 결과는 0과 1 사이의 값을 가지며 정답에 가까울수록 1에 가깝게 제시된다. 만일 생성자가 만든 영상이 진짜에 가깝다면 식별자도 그 영상을 1에 가까운 값을 결괏값으로 내놓을 것이다. 생성자의 학습은 일단 가중치를 랜덤 노이즈로 시작해 영상을 생성한다. 그다음 생성한 영상이 오차를 줄이도록, 즉 식별자의 식별 결과가 정답인 1이 나오도록 가중치를 업데이트한다(Goodfellow et al., 2014).

GAN 기반 생성은 합성(synthesis), 증강(augmentation), 각색(adaptation), 변환(translation) 등으로 세분화할 수 있다. 합성의 대표적인 사례로는 DCGAN(deep convolutional generative adversarial networks)을 들 수 있다. 예를 들어 미소 짓는 여성의 얼굴에서 무표정한 여성을 빼고 무표정한 남성을 더하면 미소 짓는 남성의 얼굴을 합성할 수 있다(Radford, Metz, & Chintala, 2015). GAN을 활용해 적은 데이터로 가상의 학습데이터를 만드는 것은 증강의 예이다(Tanaka & Aranha,

2019). 사진을 반 고흐의 화풍으로 그림을 바꿔주거나, 그림을 대신 완성해주는 것은 각색의 예이다(Gatys, Ecker, & Bethge, 2015). 변환 기법으로는 이미지를 1대 1 또는 그룹 대 그룹으로 쌍으로 학습하여 바꿔주는 pix2pix 또는 Cycle GAN을 들 수 있다(Chen & Hays, 2018). 한편 GAN은 생성 외에 분류에 활용되기도 한다.

GAN과 함께 널리 활용되는 생성 모형으로는 AE가 있다. AE는 기본적으로 입력 영상의 특징을 더 낮은 차원으로 압축하는 차원 축소(dimensionality reduction) 과정과 차원 축소된 특징에서 영상을 재생해 출력으로 내보내는 정보 추출(information retrieval) 과정으로 이뤄진다. 영상 생성에서는 입력 영상을 인코더(encoder)로 차원 축소하고, 디코더(decoder)로 복원하는 식으로 작동한다. 일반적인 신경망과 달리 AE는 은닉층이 입력층보다 작고, 입력층과 출력층의 크기는 같다(Goodfellow, Bengio, & Courville, 2016).

AE 중에서 현재 널리 쓰이는 것이 VAE(variational autoencoder)이다. VAE의 경우 인코더에서는 평균 벡터 뮤(μ)와 분산 벡터 시그마(σ)로 입력 영상의 특징을 찾은 뒤 잠재변수(latent variable) 제타(ξ)를 확률 샘플링한다. 즉 인코더에는 평균, 분산, 잠재변수에 각각 신경층이 할당된다. 디코더에서는 잠재변수를 입력으로 받아 이를 원래 데이터로 재현한다. VAE에서는 제타를 조금씩 조정하면 연속적으로 조금씩 변화하는 다양한 이미지들을 쉽게 생성할 수 있다(Kingma & Welling, 2013).

4) 전이학습

전이학습은 인간이 기존 지식을 다른 영역에 적용하는 것과 유사한 방식으로 학습을 진행한다. 전이학습은 소스 과제를 통해 얻은 소스 모델의 영역 지식(domain knowledge)을 유사한 표적 과제의 대상 모델 학습에 적용하는 방법이다(Caruana et al., 1995; Pan & Yang, 2009). 예를 들어 ImageNet으로 사전 훈련한 모델을 이용해 각종 의료 영상을

인식하는 AI 모델을 만들 수 있다(Shie, Chuang, Chou, Wu, & Chang, 2015). 기존 학습 방식이 특정 객체탐지 등 특정 과업에 특화된 학습데이터를 활용해 특정 과업에 특화된 가중치 값을 학습한 모델을 만드는 것이라면, 전이학습은 다른 영역의 학습데이터로 학습한 사전 훈련 모델을 사용하고 최종층과 일부 은닉층만 학습한다.

전이학습이 필요한 이유는 크게 두 가지로 정리할 수 있다. 첫째, 학습데이터가 부족한 경우에 전이학습이 유용하다. 어떤 클래스의 학습데이터가 저작권 문제로 비용이 너무 많이 들거나, 의료 사진과 같이 개인정보 문제로 확보하기 힘들거나, 천연기념물처럼 그 자체가 희소한 경우 대규모 학습데이터 구축이 어렵다. 둘째, 전이학습은 "바퀴를 두 번 발명하지 않게(Don't reinvent the wheel)" 해준다. 즉 이미 누군가가 충분히 많은 학습데이터를 쌓아 놓고 막대한 컴퓨팅 자원을 활용해 잘 만들어진 사전 훈련 모델을 이미 만들어 놓았다면 이를 활용하는 것이 효과적이다.

전이학습은 사전 훈련 모델 개발, 특성 추출(feature extraction), 미세 조정(fine tuning)으로 구성된다. 사전 훈련 모델은 ImageNet 등 규모가 큰 데이터셋을 활용해 깊은 신경망을 사용해 만든다(Huh, Agrawal, & Efros, 2016). 특성 추출은 표적 모델의 은닉층에 대한 지식, 즉 파라미터 값을 사전 훈련 모델에서 대거 가져오는 방식으로 이뤄진다. 미세 조정은 최종층 이전의 은닉층 일부를 새로운 층으로 대체 입력한 뒤 타겟 과제에 특화된 소수의 학습데이터로 교체된 은닉층과 최종층의 학습을 진행하는 과정이다(Sarkar, Bali, & Ghosh, 2018).

전이학습은 성공적일 경우, 더 나은 초기 성능, 더 나은 최종 성능, 더 빠른 학습이 가능하다는 장점을 갖는다(Torrey & Shavlik, 2010). 다만 소스 모델의 지식이 타겟 모델에 전이될 때 지식 개선이 이뤄지지 않고 성능이 저하되는 부정적 전이가 나타날 수도 있다(Rosenstein, Marx, Kaelbling, & Dietterich, 2005). 이 경우, 목표 과업이 전이학습을 적용하는데 적정한지(전이 시기), 전이학습을 적절한 부분에서

수행했는지(전이 항목), 전이학습의 방법이 적절한지(전이 방법) 등을 재검토해야 한다(Sarkar, Bali, & Ghosh, 2018).

전이학습은 분류, 생성 등 모든 종류의 과업에 적용할 수 있다. 최근에는 트랜스포머(transformer)에 전이학습을 접목하는 방식이 각광을 받고 있다. 대표적으로 자연어처리(natural language processing, NLP) 분야의 BERT(Bidirectional Encoder Representations from Transformers)와 GPT-3(Generative Pre-trained Transformer 3)가 있다. 최근에는 트랜스포머를 비전 분야에 적용하려는 시도도 나오고 있다(Khan et al., 2021). 다만 텍스트에 비해 동영상의 데이터 크기가 훨씬 더 크고[18] 이에 따라 사전 학습에 필요한 데이터 규모도 훨씬 더 커질 수밖에 없기에, 텍스트 분야의 트랜스포머나 GPT-3에 준하는 모델을 동영상 분야에서 구축하는 것은 난이도가 훨씬 더 높을 것으로 예상된다.

4. 미디어 동영상 영역의 비전 분야 딥러닝 적용 가능성

이 장에서는 미디어 동영상 분야에 적용할 수 있는 주요 과업을 크게 5개, 세부적으로 11개로 정리하고 과업별로 사용 가능한 SOTA와 벤치마크 데이터셋을 간단히 소개한다.

1) 미디어 동영상 분야의 딥러닝 적용 가능 과업과 모델

(1) 객체탐지

우선 다양한 객체탐지 모델로는 앞서 소개한 YOLOv5를 활용할 수 있

18) 예를 들어 기사 1개당 원고지 10매로 보면 1개 기사의 텍스트 파일 용량은 4KB이다. 이것이 90초 분량의 FHD급 동영상 파일에 대응한다고 가정하면 1개 동영상 클립은 대략 25MB 안팎에 달한다.

다. 객체 클래스는 MS COCO(Common Objects in Context)를 참고한다. 해당 데이터셋은 80개 객체 클래스로 분류된 33만 장의 사진을 제공한다(Lin et al., 2014). YOLOv5에서는 MS COCO로 사전 훈련한 모델을 제공한다.

객체탐지의 하위 과업으로 미디어 동영상에서 특히 중요한 얼굴, 장소, 상품을 살펴보자. 얼굴 관련 주요 과업으로는 영상 속 등장인물을 파악할 수 있는 얼굴 인식과 감성 인식에 활용되는 표정인식이 있다. 얼굴 인식은 얼굴과 배경을 구분하는 얼굴 탐지(face detection)와 얼굴을 정면을 바라보도록 변환하는 정규화를 거쳐 서로 다른 얼굴을 구분하는 과정을 거친다. 얼굴 인식 분야의 SOTA로는 RetinaFace와 ArcFace를 꼽을 수 있다(Deng, Guo, Xue, & Zafeiriou, 2019; Deng et al., 2019). 관련 데이터셋으로는 LFW(labeled faces in the Wild)가 대표적이다[19]. 표정인식에는 VGG와 같은 CNN 모델을 사용할 수 있다. 관련 벤치마크 데이터셋인 FER2013은 화남(angry), 역겨움(disgust), 두려움(fear), 행복(happy), 슬픔(sad), 놀람(surprise), 무표정(neutral) 등 7개 표정을 분류한다(Goodfellow et al., 2013). 또 다른 벤치마크 셋인 AffectNet은 추가로 불만(contempt)을 포함한다(Mollahosseini, Hasani, & Mahoor, 2017).

유명 장소를 인식하는 랜드마크 인식의 SOTA로는 ResNet 기반의 DELF(DEep Local Feature)를 꼽을 수 있다(Noh, Araujo, Sim, Weyand, & Han, 2017). 벤치마크 셋으로는 구글 랜드마크 셋(Google Landmarks Dataset)이 있다(Weyand, Araujo, Cao, & Sim, 2020). 상품인식의 경우 작은 객체를 인지할 수 있는 EfficientDet이 적합하다. 미디어 동영상에서는 광고를 제외하면 상품이 작게 나오는 경우가 많기 때문이다. 상품인식 관련 데이터셋으로는 스탠

19) http://vis-www.cs.umass.edu/lfw/

포드 온라인 상품 셋(Stanford Online Products)[20] 등이 있다.

(2) 행동인식, 자세추정

행동인식은 I3D(Inflated 3D Convolutional Neural Networks)를 활용할 수 있다. I3D는 객체탐지를 위한 합성곱 연결망(ConvNet)과 시계열 데이터 학습을 위한 LSTM을 결합하는 방법, 2D ConvNet(2D Convolutional Neural Networks)을 시간 축으로 확장한 3D ConvNet, 2개의 ConvNet을 사용하고 각각 개별 프레임의 RGB 정보와 프레임 간 객체의 픽셀이 이동한 방향과 거리의 분포를 담은 광학 흐름(optical flow)을 입력하는 2-스트림(two-stream) 방식을 거쳐 2-스트림에 사용하는 ConvNet을 3D ConvNet으로 바꾼 것이다. 행동 관련 벤치마크 데이터셋으로는 I3D와 함께 공개된 딥마인드의 데이터셋인 Kinetics 400을 활용할 수 있다(Carreira & Zisserman, 2017).

자세추정은 행동을 바운딩박스가 아닌 관절 단위의 키 포인트(key point)로 학습한다. 자세추정은 스포츠나 댄스의 동작을 분류할 때 활용할 수 있다. 적용 가능한 관련 모델로는 고해상도의 고수준 특징을 함께 처리하는 HRNet(High-Resolution Net)가 있다(Wang et al., 2020).

(3) 상황 분류와 이상탐지를 이용한 사건 탐지

사건 탐지는 상황 분류와 이상탐지를 활용할 수 있다. 상황 분류 모델로는 SlowFast가 있다. SlowFast 모델은 2-스트림 방식을 이용한다. 즉 동영상에서 초당 적은 프레임을 추출한 경로(slow pathway)에서 공간 정보를, 초당 많은 프레임을 추출한 경로(fast pathway)에서 시간 정보를 학습하여 동영상의 상황을 시공간 특성에 따라 분류한다(Feichtenhofer,

---

20) https://cvgl.stanford.edu/projects/lifted_struct/

Fan, Malik, & He, 2019).

이상탐지는 3D ConvNet을 이용해 이상행동을 탐지하는 방법, MOT를 활용해 사람이나 자동차 등의 궤적이 정상이 아닌 경우를 탐지하는 방법, 인스턴스 분할이 가능한 YOLACT 등을 활용해 화재 등을 탐지하는 방법 등이 있다(대전교통공사, 2021.6; 미디어그룹사람과숲, 2021.6; Sultani, Chen, & Shah, 2018).

(4) 동영상 요약

미디어 분야에서 동영상 요약은 가편집, 예고편 생성 등에 활용될 수 있다. 동영상 요약은 주요 정보의 손실을 최소화하면서 대용량 동영상을 적은 수의 키 프레임(key frame) 또는 키 쇼트(key shot)로 구성된 동영상으로 축약하는 기술이다(Fajt, Sokeh, Argyriou, Monekosso, & Remgnino, 2018). 동영상 요약은 움직임이나 색 등 특징 기반(feature based) 방식, 유사도에 따른 군집화 기반(cluster based) 방식, 주요 사건을 정의하고 해당 사건을 탐지해 요약하는 사건 기반(event based) 방식, 시공간 정보나 돌출 객체를 활용하는 궤적 기반(trajectory based) 방식, 숏 선택 기반(shot selection based) 방식 등이 있다. 또한 높은 시청률 등 사용자 로그 정보나, 음성이나 대본 등을 자연어처리 기법으로 요약하고 해당 동영상 부분을 추출하는 멀티 모달 접근이 활용됐다.

비전 분야 딥러닝 모델이 발전함에 따라 다양한 범주, 다양한 길이의 동영상을 비전 기반으로 요약할 수 있을 것으로 보인다. 예를 들어 AI허브의 요약영상 AI 학습데이터는 유튜브의 20개 카테고리별로 20분 이상과 20분 이하의 동영상 클립에서 1분 간격으로 3초짜리 1개 숏을 추출한다. 이 학습데이터를 학습한 AI 모델은 10분짜리 동영상을 30초 영상으로 요약해줄 수 있다(한국과학기술원, 2021.2). 더 나아가 복수의 요약영상을 하나의 축약된 클립으로 생성할 수도 있다. 요약 관련 모델로는

CNN 기반 DeSumNet, LSTM을 활용한 ddpLSTM, LSTM에 GAN을 접목한 Adversarial LSTM Networks 등을 거쳐 현재 어텐션(Attention) 기반 VASNet이 SOTA를 달성했다(홍순기, 2021; Fajtl et al., 2018). 관련 데이터셋으로는 TvSum 등이 있다.

(5) 동영상 분류

동영상 분류는 동영상을 프레임 단위가 아니라 전체 동영상 단위 또는 클립 단위로 분류하는 것을 목표로 한다. 동영상 분류 방법으로 가장 기본적인 방식은 프레임 단위 분류를 바탕으로 동영상 분류를 수행하는 것이다. 그러나 개별 프레임의 정보가 전체 동영상의 분류와 무관할 수 있으므로 좀 더 복잡한 기법이 필요하다. DCGN(deep convolutional graph neural network)은 우선 CNN으로 추출한 영상의 특징을 바탕으로 프레임을 노드로 하고 유사한 프레임을 서로 연결한 프레임 수준의 그래프(graph)를 도출한다. 이어 이를 샷 수준, 사건 수준, 전체 동영상 수준 순으로 그래프를 도출하고 이를 분류하는 방식으로 학습한다(Mao, Wu, Xue, & Zhang, 2018). 동영상 분류 셋으로는 22개 장르 35만 시간 분량의 610만 개 동영상 클립에 3,862개 클래스에 대한 시각 정보를 라벨링한 YouTube-8M이 있다.

동영상은 행동인식에 따라 분류할 수도 있다. 이 밖에도 영상 정보 외에 장르 정보 등 메타데이터, 대본이 없는 경우 STT(Speech to text)를 활용해 대본을 생성하고, 대본이 있는 경우 대본을 자연어처리하여 의미론적으로 분류하는 멀티모달(multi-modal) 접근이 병행될 수 있다. 즉 미디어에서는 대본에는 없지만 영상에는 포함된 배경과 같은 영상 특성과, 정치 뉴스와 같은 메타데이터 또는 논평 대상이 되는 인물 등 대본에는 있지만 영상에는 없는 멀티모달 특성이 결합돼야 활용도 높은 동영상 분류가 가능하다.

이상으로 이 논문에서 검토한 과업별 SOTA 모델의 예시를 정리하

면 〈표 1〉과 같다.

표 1. 미디어 동영상 분야 적용 가능 과업별 주요 모델 및 벤치마크 데이터셋 예시

과업		모델	벤치마크 데이터셋
객체	객체탐지	YOLOv5	MS COCO
	안면인식	ArcFace	LFW
	표정인식	VGG	FER2013, AffectNet
	랜드마크 인식	DELF	Google Landmarks Dataset
	상품인식	EfficientDet	Stanford Online Products
행동	행동인식	I3D	Kinetics 400
	자세추정	HRNet	MS COCO
사건	이상탐지	FairMOT	MOT20
		YOLACT	MS COCO
	상황인식	SlowFast	Kinetics 600
요약	동영상 요약	VASNet	TvSum
분류	동영상 분류	DCGN	YouTube-8M

2) 미디어 동영상 영역의 비전 분야 딥러닝 적용 방안

이상 검토한 모델은 무엇보다 동영상의 내용분석, 즉 VA에 활용될 수 있다. 국내 미디어 연구에서도 방송 동영상 등의 내용분석에 딥러닝 기반 VA를 적용한 경우가 있다. 그 방식을 살펴보면 우선 클라우드 컴퓨팅 서비스에서 제공하는 API(application programming interface)를 활용하는 방식이 있다. 예컨대 마이크로소프트의 애저(Microsoft Azure)가 제공하는 API를 사용해 방송사의 대선 보도에 나타난 대선 후보의 감정을 얼굴 표정을 통해 분석한 사례가 있다. 이 연구는 규칙 기반으로 키프레임을 추출하는 방식으로 분석 대상 프레임 수를 4,990개로 줄였다(김윤환·길우영·이종혁, 2017). 또 다른 연구에서는 구글이 제공하는 텐서플로우(TensorFlow) 기반 AI 프레임워크인 미디어파이프(Mediapipe)[21]를 활용해 걸그룹 브레이브걸스의 자세추정을 클라우드 상의 분석 환경인 구

21) https://mediapipe.dev/

글 코랩(Colab)[22]에서 분석한다(윤호영, 2021). 최근에는 API나 AI 프레임워크를 사용하지 않고, 직접 학습데이터를 구축하고 CNN 계열 모델이자 표정인식 분야의 SOTA인 VGG16와 전이학습을 사용해 대선 후보별 신원을 확인하는 수준의 안면인식과 표정인식을 수행한 연구가 있다. 이 연구는 키 프레임 추출 없이 125만 장의 프레임을 분석했다(최윤정 등, 2020).

딥러닝을 활용한 동영상 분석 절차는 다음과 같다. 첫째, 분석 영상을 수집한다. 둘째, 과업을 정하고 해당 과업의 SOTA를 선별한다. 셋째, 과업과 가장 관련성이 높은 벤치마크 데이터셋을 활용해 모델을 학습시킨다. 이런 모델링 과정에서는 필터 개수, 필터 크기, 보폭의 값, 패딩 방식, 풀링 방식, 풀의 크기, 풀의 보폭 값, 최종층에서 완전연결층의 넓이, 활성함수, 손실함수, 경사하강법 등 사람이 직접 설정하는 변수인 하이퍼 파라미터(hyper parameter)를 조정하는 과정이 병행된다. 학습 및 분석 환경은 고성능 GPU를 장착한 개인용 컴퓨터를 활용할 수도 있지만, 구글 코랩과 같은 클라우드상의 분석 환경을 사용하는 것도 효과적이다. 목표 과업이 소스 과업과 다를 경우, 벤치마크 데이터셋과 다른 기 구축된 학습데이터를 확보하거나 직접 구축한 학습데이터를 활용한 전이학습을 수행한다. 넷째, 모델의 성능을 평가하고 필요시 모델을 개선한다. 모델이 충분한 성능이 나오지 않을 경우, 하이퍼 파라미터를 조정하거나, 학습데이터를 추가하거나, 모델을 바꿔 비교해본다. 한편 전이학습 시에는 미세 조정을 다시 하거나, 부정적 전이가 일어났다면 이를 수정하는 등의 작업이 필요할 수 있다(Sarkar et al., 2018). 최종적으로 만족할만한 성능의 모델을 확보했다면, 수집한 동영상을 전처리한 뒤 해당 모델을 통해 VA를 수행한다. 만일 목표 과업에 맞는 모델을 제공하는 공개된 API가 있으며 분석할 데이터가 비교적 많지 않다면, 앞의 단계를

22) https://colab.research.google.com/

건너뛰고 마지막 단계만 API를 활용해 수행하면 된다. 어떤 경우든 VA는 인간이 수작업으로 하기 힘든 방대한 양의 데이터를 수집해 반복적으로 분석할 때 그 가치가 극대화된다.

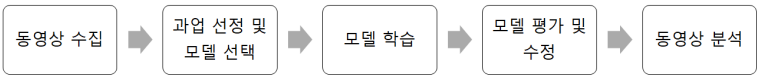

그림 1. 딥러닝 기반 동영상 분석 절차

다른 한편 딥러닝 기반 VA를 활용해 영상검색이나 영상요약 등 다양한 미디어 동영상 서비스를 AI 기반으로 고도화할 수 있다. 지금의 동영상 검색은 대부분 텍스트에 의존한다. 즉 동영상의 태그나 제목, 설명, 대본 등을 활용해 검색을 수행한다. 그러나 VA를 적용하면 수작업으로 입력한 태그나 제목, 대본에 포함돼 있지 않더라도 동영상 기반으로 정확한 검색이 가능하다. 예를 들어 '피아노'가 텍스트에 쓰여 있지 않더라도 영상 속 피아노를 사전에 오토 태깅하고 검색에 쓸 수 있다.

영상을 사전 분류하거나, 관련 동영상을 미리 군집화해둘 수도 있다. 텍스트 분석에 의존한 영상검색은 클립 단위 검색만 가능하지만, VA에 의존한 영상검색은 프레임 단위 검색의 정밀한 검색도 가능하다. 동영상의 모든 프레임에 수작업으로 태깅하는 것은 비용 문제로 사실상 불가능하다.

영상 정보 기반의 고도화된 영상요약도 가능해진다. 현재의 영상 요약은 방송의 시청률이나 게임의 득점 장면 등과 같은 영상 외적 정보, 영상 내 문자 인식이나 텍스트 요약 등과 결합한 멀티 모달 방식의 영상 요약 기법이 많다(금준경, 2018.10.25.; 이영현, 2019.7.9.). 그러나 VA를 활용하면 키 샷의 영상적 특징을 학습해 동영상을 요약하는 것도 가능해질 것이다.

영상검색이나 영상요약을 활용해 영상편집을 효율화할 수도 있다.

우선 영상요약 기술을 통해 촬영본의 1차 편집에 활용할 수 있다. 영상요약이 고도화될 경우 영화나 방송 프로그램의 예고편을 다양하게 만들 수도 있다. 아카이브로 구축한 동영상의 검색도 훨씬 풍부해질 수 있다. VA 속도가 개선되면, 최근 촬영본 역시 여러 버전을 동시에 빠르게 검색하고 편집할 수 있을 것이다. 영상을 설명하는 자막을 1차적으로 자동 부착하는 것도 가능하다.

제작이나 편집 외에도 콘텐츠 유통 측면에서도 도움을 받을 수 있다. 우선 동영상 플랫폼에서 정교한 추천도 가능하다. 기존의 추천에 활용되는 동영상의 정보는 주로 장르, 국적, 출연자, 제작진 등 동영상 관련 메타 정보와 수작업으로 부착한 태그 정보 수준이었다. 그러나 VA를 통해 사전에 동영상의 태그 정보를 장면 단위로 부착하고 있다면, 내용 기반이든 행동 기반이든 사용자에게 더욱 정확한 추천을 특정 프레임 단위로 추천해줄 수도 있을 것이다. 다만 이러한 영상 기반 준 실시간 추천은 실제 서비스 수준에 적용할 정도의 확장 가능성(scalability)이 아직은 낮다. 기술적으로 안정적인 서비스 운영을 뒷받침하기 어려울 뿐만 아니라, 투자 대비 수익(return on investment, ROI)가 낮기 때문이다.

VA 외에도 영상 생성 기술을 활용할 수도 있다. 이미 인스타그램이나 틱톡 등 많은 영상 플랫폼들은 다양한 AI 기반 도구를 사용해 아바타 생성이나 필터 효과 등을 제공함으로써 사용자 관여도를 높이고 있다. 또한 AI 아나운서 등을 활용해 로봇 저널리즘을 구현할 수도 있다. 가상 인플루언서를 이용한 광고나 관객과 24시간 상호작용하는 예능 프로그램을 만들 수도 있을 것이다.

5. 요약 및 제언

이 논문에서는 미디어 동영상의 분석과 생성에 활용도가 높은 비전 관련

딥러닝 기반 모델을 검토했다. 우선 다양한 모델의 기축이 되는 알고리즘으로서 분류 모델로 널리 사용되는 CNN과 RNN, 생성 모델로 사용되는 GAN과 AE, 사전 훈련 모델을 활용하는 전이학습을 살펴보았다. 다만 RNN과 같은 분류 모델이 생성에 사용되거나 GAN과 같은 생성 모델이 분류에 사용되는 경우도 있다. 다음으로 미디어 동영상 영역에서 활용도가 높으면서 AI허브에 학습데이터가 공개된 과업을 중심으로 주요 과업을 객체탐지, 행동인식, 사건탐지, 동영상 요약, 동영상 분류 등 5개 대분류와 객체탐지, 안면인식, 표정인식, 랜드마크 인식, 상품인식, 행동인식, 자세추정, 이상탐지, 상황인식, 동영상 요약, 동영상 분류 등 11개 소분류로 제시했다. 이어 각 과업별 SOTA와 벤치마크 데이터셋을 소개했다. 끝으로 이러한 모델의 학문적, 실무적 활용 가능성을 제시해보았다.

딥러닝 분야는 컴퓨터공학 전공자라고 해도 딥러닝 연구자가 아니면 최신 연구를 따라가기 힘들 정도로 빠르게 발전 중이다. 게다가 관련 논문은 물론 개론서조차 낯선 전문용어와 약어, 수학적 설명, 도안, 소스코드로 가득하다. 이 때문에 딥러닝 활용에 필요한 공학적 지식을 미디어 연구 연구자들이 습득하기가 쉽지 않다. 본 논문은 수식이나 프로그래밍에 대한 지식이 없이 미디어 연구자가 CNN, RNN, GAN, AE, 전이학습 등 비전 분야 딥러닝의 큰 줄기를 파악하는데 도움을 줄 것으로 기대한다. 또한 본 논문에서는 미디어 동영상 분야에 적용 가능성이 높은 주요 과업과 과업별 SOTA 및 SOTA의 학습과 평가에 활용되는 벤치마크 데이터셋을 소개했다. 이를 통해 미디어 연구자나 미디어 서비스 기획자는 관련 모델을 직접 활용하거나 컴퓨터공학 분야의 연구자 또는 개발자와 협업할 때 필요한 배경지식을 얻을 수 있을 것으로 생각한다.

미디어 동영상 분야에 딥러닝 모델의 활용도가 높아진다면, 설명 가능한 AI(Explainable Artificial Intelligence, XAI) 등에 대한 미디어 연구의 필요성도 그만큼 커질 수 있다. 이에 따라 비판적 미디어 연구자와 딥러닝 관련 컴퓨터공학자들과의 협업이 요청될 수 있다. 이상적인

경우 비판적 미디어 연구자가 영역 지식을 바탕으로 미디어 동영상 관련 비전 분야의 XAI 연구에 기여할 수도 있다. 이때 미디어 연구자 역시 CNN에서 각각의 합성곱층이 어떤 특징을 학습하는지 확인함으로써 해석가능한 CNN(interpretable CNN)을 구현한다거나 학습 시 설명까지 함께 학습하는 등의 기술적 시도를 이해한다면 더욱 깊이 있는 협업 참여가 가능할 것이다(김시원·윤성로·김용대, 2021).

미디어 연구자가 클라우드 서비스에서 제공하는 API를 사용하지 않고 직접 AI를 모델링을 할 경우, 과업 및 모델 선정, 학습데이터 구축, 하이퍼 파라미터 설정, 모델 학습, 미세 조정, 분석 등을 직접 수행해야 한다. 사실상 딥러닝 분야의 컴퓨터공학자에 버금가는 수준의 코딩 역량이 필요하거나, 최소한 컴퓨터공학자와의 협업이 요청될 수 있다. API를 사용할 경우, 연구 목적에 부합하는 API가 제공되고 있어야 한다. 연구자 역시 API를 활용할 수 있는 코딩 역량은 갖춰야 한다. 여기서 유의할 점은 제공되는 API는 물론 널리 통용되는 SOTA 역시, 사실 미디어 연구자가 그대로 쓰기에는 여전히 다소 단순한 수준의 내용분석을 제공한다는 점이다. 즉 현재의 AI는 한정된 환경의 동영상에서 객체, 행동, 장소 등을 탐지하며, 상황은 이러한 정보를 조합하는 형태로 제시하고, 그 역시 제한된 클래스 내에서만 적정 성능을 보여준다. 따라서 편향성이나 선정성과 같은 고차원적인 분석에서는 이러한 간단한 내용분석 결과를 사회과학의 이론적 전통 위에서 타당성 있게 개념화하고 조작적으로 신중하게 정의하는 과정이 선행돼야 한다. 또한 분석 결과는 데이터 수집이나 모델의 성능에 최대한 영향을 덜 받도록 강건하게 해석해야 한다. 엄밀한 분석이 필요할 경우, AI의 분석 결과를 인간이 추가로 검증하고 재작업해야 할 필요도 있다.

이러한 제한점에도 불구하고 미디어 동영상 분석에 딥러닝을 적용하는 것은 가치가 있다. 2010년대 이후 미디어 연구에서는 텍스트를 중심으로 자연어처리와 의미연결망 분석(semantic network analysis) 기

법을 활용한 컴퓨터 기반 내용분석(computerized content analysis)이 확산되어 왔다(예를 들어 박대민, 2016). 비전 분야 딥러닝 기법은 흔히 미디어 연구에서 빅데이터 분석(big data analysis)이라고 불리는 이러한 자동화된 내용분석 방법을 텍스트에서 사진과 동영상까지 확대해 줄 것으로 기대한다. 분석할 데이터를 대규모로 수집하고, 학습데이터를 구축하고, AI 모델링을 하는 작업은 처음에는 수작업으로 분석하는 것 못지않은 노력이 들어갈 수도 있다. 그러나 일단 시스템을 구축하고 난 뒤에는, 반복적인 분석 작업이 훨씬 수월하게 진행될 수 있다. 분석 시스템에서 적절한 사용자 인터페이스가 제공된다면, 코딩을 전혀 모르는 연구자도 방대한 양의 동영상을 최적화된 높은 성능의 딥러닝 모델을 활용해 분석할 수도 있을 것이다. 2021년 12월 현재 AI허브에 무료로 공개된 방송 동영상 기반 학습데이터는 그 시작으로 볼 수 있다. AI허브, 특히 비전 분야[23]에서는 방송 동영상을 원천데이터로 수집하여, 이를 수작업으로 라벨링한 약 3,700시간의 학습데이터가 무료로 공개되어 있다. 또한 AI허브는 방송 동영상은 아니지만 유튜브 동영상이나 스포츠, 댄스 등 방송과 관련성이 높은 동영상을 활용한 학습데이터도 구축해놓았다.

궁극적으로는 개체명 인식과 같은 자연어처리에 딥러닝 모델인 BERT를 활용하는 한국언론진흥재단의 뉴스 빅데이터 분석 시스템 〈빅카인즈〉[24]와 같은 시스템을 미디어 동영상 분야에 도입할 필요가 있다. AI 기반의 고도화된 대중문화 아카이브 구축에도 활용할 수 있다(박대민·오세욱, 2016; 박대민·유재연, 2021.10.). 비전 분야 딥러닝의 발전은 동영상 분야에 미디어 인공지능을 적용할 토대를 마련했다. 미디어 동영상 빅데이터 분석 시스템은 방송은 물론 유튜브 등 다양한 소스에서 제공되는 미디어 동영상을 실시간에 가깝게 수집하고, 딥러닝 기반으

23) https://aihub.or.kr/aihub-data/vision/about

24) https://www.bigkinds.or.kr/v2/intro/service.do

로 상당한 수준의 영상분석, 영상검색, 영상편집, 분류, 요약 등의 서비스를 준 실시간으로 제공하는 형태로 개발될 수 있다. 이러한 분석 시스템이 있다면 미디어 연구자들은 수집, 분석, 모델링과 관련된 자원 부족과 기술적 어려움을 극복하고 이론화에 집중할 수 있을 것이다. 미디어 서비스 제공자 역시 딥러닝 기반 디지털 전환(digital transformation)을 실현하고 더 나은 사용자 경험을 고객에게 제공할 수도 있을 것이다. 또한 이러한 과정에서 미디어 연구자들이 영역 지식과 이 논문에서 소개한 딥러닝의 배경지식을 바탕으로 미디어 분야의 사회적 수요에 맞는 고차원적인 시스템 기획 및 개발에 기여하기를 기대해본다.

본 논문은 미디어 동영상 영역에 적용 가능한 비전 분야 딥러닝 모델을 체계적으로 소개하는 것에 목적을 두어 딥러닝을 활용한 실제 분석 사례를 구체적으로 제시하지 못한 한계를 갖는다. 향후 미디어 연구에서 미디어 동영상 분석에 딥러닝 활용도를 높이기 위해 다양한 세부 주제의 동영상을 수집하고, 딥러닝을 활용해 분석하고, 분석 결과를 이론화한 연구 사례가 많아지기를 고대해본다.

참고문헌

금준경 (2018. 10. 25). SBS가 인공지능으로 클립 영상을 만들기 시작했다. 〈미디어오늘〉. URL: http://www.mediatoday.co.kr/news/articleView.html? idxno=145137

김시원·윤성로·김용대 (2021). 인공지능의 작동원리와 설명가능성에 대하여: 고급 인공지능 학습 및 해석/설명 방법론. 〈인공지능 원론: 설명가능성을 중심으로〉. 서울: 박영사.

김윤환·길우영·이종혁 (2017). 미디어 정서의 의제설정과 점화 효과 연구. 〈사이버커뮤니케이션학보〉, 34권 4호, 53-98.

대전교통공사 (2021.6). 〈인공지능 데이터 구축•활용 가이드라인: 지하철 역사 내 CCTV 이상행동 영상 소개〉. 서울: 한국지능정보사회진흥원.

미디어그룹사람과숲 (2021.6). 〈인공지능 데이터 구축•활용 가이드라인: 화재 발생 예측 영상 소개〉. 서울: 한국지능정보사회진흥원.

박대민 (2016). 장기 시계열 내용 분석을 위한 뉴스 빅데이터 분석의 활용 가능성: 100만 건 기사의 정보원과 주제로 본 신문 26년. 〈한국언론학보〉, 60권 5호, 353-407.

박대민·오세욱 (2016). 방송 뉴스 빅데이터 분석의 가능성. 〈방송문화연구〉, 28권 1호, 7-51.

박대민·유재연 (2021.10.). 한국대중문화아카이브연구원 구축 전략. 한국 대중문화아카이브 구축: 기술, 제도, 추진 전략. 서울대학교, 시흥.

윤호영 (2021). 사람에서 컴퓨터 자동화로의 연결을 위한 탐색 : 객체 인식(Object Detection) 딥러닝 알고리즘 YOLO4, 자세 인식(Pose Detection) 프레임워크 MediaPipe를 활용한 음악 프로그램의 여성 신체 대상화, 선정적 화면 검출 연구. 〈한국언론학보〉, 65권 6호, 452-481.

이영현 (2019.7.9). 게임과 AI #10 이미지 인식 기술을 이용한 야구 하이라이트 클립 생성. 〈엔씨소프트 블로그〉. URL: https://blog.ncsoft.com/게임과-ai-10-이미지-인식-기술을-이용한-야구-하이라이/

이희석 (2016). 〈영상 처리 프로그래밍: 기초 알고리즘부터 라이브러리 활용까지 영상 처리 프로그래밍〉. 부천: 프리렉.

조태호 (2020). 〈모두의 딥러닝〉. 서울: 길벗.

최윤정·정유진·윤호영·김민정·김나영·첸루·신주연·이주희·김나영·여은·강제원 (2020). 딥러닝(Deep learning)기반 동영상 처리 알고리즘을 통한 19대 대선 TV토론 영상분석: 후보자들의 등장빈도, 표정, 응시방향에 대한 분석. 〈한국언론학보〉, 64권 5호, 319-372.

한국과학기술원(2021.2). 인공지능 데이터 구축·활용 가이드라인: 감정인식 및 요약영상 AI 데이터. URL: https://aihub.or.kr/aidata/27717

홍순기 (2021). AI 기반 자동 편집 기술 동향. 〈방송과 미디어〉, 26권 1호, 76-96.

황선규 (2019). 〈OpenCV 4로 배우는 컴퓨터 비전과 머신러닝〉. 서울: 길벗.

我妻幸長 (2020). はじめてのディープラーニング 2: Pythonで実装する再帰型ニューラルネットワークとVAE´ GAN. SBクリエイティブ. 최재원·장건희 (역) (2020). 〈핵심 딥러닝 입문: RNN, LSTM, GRU, VAE, GAN 구현〉. 안양: 책만.

Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (2016, September). *Simple online and realtime tracking.* Paper presented at 2016 IEEE international conference on image processing (ICIP).

Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934.*

Carreira, J., & Zisserman, A. (2017). *Quo vadis, action recognition? A new model and the kinetics dataset.* Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition.

Chen, W., & Hays, J. (2018). *Sketchygan: Towards diverse and*

realistic sketch to image synthesis. Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). *Arcface: Additive angular margin loss for deep face recognition*. Paper presented at the IEEE/CVF conference on computer vision and pattern recognition.

Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., & Zafeiriou, S. (2019). Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*.

Fajtl, J., Sokeh, H. S., Argyriou, V., Monekosso, D., & Remagnino, P. (2018, December). *Summarizing videos with attention*. Paper presented at the Asian conference on computer vision.

Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). *Slowfast networks for video recognition*. Paper presented at the IEEE/CVF international conference on computer vision.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2015). A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*.

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. 박해선 (역). (2020). 〈핸즈온 머신러닝 사이킷런, 케라스, 텐서플로 2를 활용한 머신러닝, 딥러닝 완벽 실무〉. 서울: 한빛미디어.

Girshick, R. (2015). *Fast R-CNN*. Paper presented at the IEEE international conference on computer vision.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). *Rich feature hierarchies for accurate object detection and semantic segmentation.* Paper presented at the IEEE conference on computer vision and pattern recognition.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, Massachusetts: The MIT Press.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D. H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., Ionescu, R., Popescu, M., Grozea, C. Bergstra, J., Xie, J., Romaszko, L., Xu, B., Chuang, Z., & Bengio, Y. (2013, November). *Challenges in representation learning: A report on three machine learning contests.* Paper presented at the international conference on neural information processing, Berlin, Heidelberg.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27.*

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). *Mask R-CNN.* Paper presented at the IEEE international conference on computer vision.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition.* Paper presented at the IEEE conference on computer vision and pattern recognition.

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *6*(02), 107-116.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term

memory. *Neural Computation*, *9*(8), 1735-1780.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*(3), 574-591.

Huh, M., Agrawal, P., & Efros, A. A. (2016). What makes ImageNet good for transfer learning?. *arXiv preprint arXiv:1608.08614.*

Katsov, I. (2017). *Introduction to algorithmic marketing: Artificial intelligence for marketing operations*. Sunnyvale, CA: Ilya Katosv.

Khan, A. I., & Al-Habsi, S. (2020). Machine learning in computer vision. *Procedia Computer Science*, *167*, 1444-1451.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2021). Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169.*

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097-1105.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541-551.

Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). *Feature pyramid networks for object detection*. Paper presented at the IEEE conference on

computer vision and pattern recognition.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., & Dollár, P. (2014, September). *Microsoft COCO: Common objects in context.* Paper presented at the European conference on computer vision.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). *SSD: Single shot multibox detector.* Paper presented at the European conference on computer vision.

Mao, F., Wu, X., Xue, H., & Zhang, R. (2018). *Hierarchical video frame sequence representation with deep convolutional graph network.* Paper presented at the European conference on computer vision workshops.

Minsky, M. & Papert, S. (1969). *Perceptrons.* Cambridge, Massachusetts: The MIT Press.

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, *10*(1), 18-31.

Nair, V., & Hinton, G. E. (2010, January). *Rectified linear units improve restricted boltzmann machines.* Paper presented at the Icml.

Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. (2015). *Beyond short snippets: Deep networks for video classification.* Paper presented at the IEEE conference on computer vision and pattern recognition.

Noh, H., Araujo, A., Sim, J., Weyand, T., & Han, B. (2017). *Large-scale image retrieval with attentive deep local features.* Paper presented at the IEEE international conference on computer vision.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE transactions on knowledge and data engineering*, *22*(10), 1345-1359.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434.*

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). *You only look once: Unified, real-time object detection.* Paper presented at the IEEE conference on computer vision and pattern recognition.

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, *28*, 91-99.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review, 65*(6), 386.

Rosenstein, M. T., Marx, Z., Kaelbling, L. P., & Dietterich, T. G. (2005, December). *To transfer or not to transfer*. Paper presented at NIPS 2005 workshop on transfer learning.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J.. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533-536.

Sarkar, D., Bali, R., & Ghosh, T. (2018). *Hands-on transfer learning with python: Implement advanced deep learning and neural network models using TensorFlow and Keras*. Mumbai: Packt Publishing Ltd.

Shie, C. K., Chuang, C. H., Chou, C. N., Wu, M. H., & Chang, E.

Y. (2015, August). *Transfer representation learning for medical image analysis.* Paper presented at the 2015 37th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556.*

Sultani, W., Chen, C., & Shah, M. (2018). *Real-world anomaly detection in surveillance videos.* Paper presented at the IEEE conference on computer vision and pattern recognition.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke,V., & Rabinovich, A. (2015). *Going deeper with convolutions.* Paper presented at the IEEE conference on computer vision and pattern recognition.

Tan, M., Pang, R., & Le, Q. V. (2020). *EfficientDet: Scalable and efficient object detection.* Paper presented at the IEEE/CVF conference on computer vision and pattern recognition.

Tanaka, F. H. K. D. S., & Aranha, C. (2019). Data augmentation using GANs. *arXiv preprint arXiv:1904.09135.*

Torrey, L., & Shavlik, J. (2010). Transfer learning. In E. S. Olivas, J. D. M. Guerrero, M. Martinez-Sober, J. R. Magdalena-Benedito, & L. Serrano (Eds.), *Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques,* Hershey, PA: IGI global.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W. & Xiao, B. (2020). *Deep high-resolution representation learning for visual recognition.* Paper presented at the IEEE transactions on pattern analysis and machine intelligence.

Wang, Z., Zheng, L., Liu, Y., Li, Y., & Wang, S. (2020, August). *Towards real-time multi-object tracking.* Paper presented at the computer vision-ECCV 2020: 16th European conference, Glasgow, UK.

Weyand, T., Araujo, A., Cao, B., & Sim, J. (2020). *Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval.* Paper presented at the IEEE/CVF conference on computer vision and pattern recognition.

Zenati, H., Foo, C. S., Lecouat, B., Manek, G., & Chandrasekhar, V. R. (2018). Efficient gAADn-based anomaly detection. *arXiv preprint arXiv:1802.06222.*

Zhang, Y., Wang, C., Wang, X., Zeng, W., & Liu, W. (2020). A simple baseline for multi-object tracking. *arXiv preprint arXiv:2004.01888*, *3*(4), 6.

Zhang, Y., Wang, C., Wang, X., Zeng, W., & Liu, W. (2021). FairMOT: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, *129*(11), 3069-3087.

투 고 일 자: 2022년 01월 05일

심 사 일 자: 2022년 02월 02일

게재확정일자: 2022년 02월 24일

Abstract

A Study on the Applicability of Media Videos of Deep Learning Models Related to Computer Vision

Daemin Park
Sunmoon University

Recently, media researchers employ deep learning models related to computer vision to perform automated content analysis of videos. Understanding deep learning models is also essential to AI(artificial intelligence) driven digital transformation in the media industry. In this paper, we reviewed computer vision-related deep learning models that are widely used for video analytics and generation. First, we looked at convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which are widely used for classification, as well as generative adversarial network (GANs) and autoencoders (AEs) as generation models, and transfer learning using pre-training models. Following that, we proposed tasks in five major categories for which AI could be highly useful: object detection, action recognition, event detection, video summarization, and video classification. Then, for subtasks such as object detection, face recognition, facial expression recognition, landmark recognition, product recognition, pose estimation, anomaly detection, context recognition, video summarization, and video classification, we introduced state-of-the-arts (SOTAs) and benchmark datasets. Finally, the potential academic and practical applications of these models were discussed. We anticipate that media researchers or media service providers will understand the major trend of deep learning in computer vision and will be able to

get knowledge when using deep learning models independently or collaborating with programmers.

KEYWORDS deep learning, computer vision, convolutional neural networks(CNN), recurrent neural networks (RNNs), generative adversarial network (GANs), media AI(media artificial intelligence), video analytics