딥러닝 기술 동향

- CNN과 RNN을 중심으로 -

곽노준 박성헌* 김대식*
서울대학교 교수
서울대학교 박사과정*

본 고에서는 딥러닝의 여러 가지 분야 중 최근 영상인식 분야에서 기존 방법들보다 월등한 성능을 보이고 있는 컨볼루션 신경망(Convolutional Neural Networks: CNN)과 음성인식이나 자연어처리 등에 적용되어 뛰어난 성능을 보이는 순환 신경망(Recurrent Neural Networks: RNN)의 최근 동향을 소개하도록 한다.

Ⅰ. 서론

최근 전국적으로 알파고 열풍이 대단하다. 구글의 자회사 딥마인드에서 개발한 인공지능 엔진인 알파고가 바둑 초고수인 이세돌을 4:1로 이기고 나서 사람들은 "이제 드디어 영화에서나 보던 인공지능 로봇이 멀지 않았다", "미래에는 인공지능이 인간을 지배할 것이다"라는 등 기술의 발전을 놀라운 눈으로 바라보고 있다.

하지만 일반인들이 전율에 떨며 일면 공포심과 경외심을 가지고 바라보는 알파고 기술의 발전상황은 이 분야를 연구하는 연구자의 입장에서는 어느 정도 당연한 결과이다. 이미 1997년 IBM의 Deep Blue가 체스계의 이세돌인 카스파로프를 이겼으며, 2011년에는 같은 회사의 Watson이 미국의 유명 퀴즈쇼인 Jeopardy에서 역대 우승자들을 상대로 큰 점수차로 이기는 성과를 거두기도 했다. 이러한 이벤트 외에도 이미 인공지능 기술은 사람들의 생각보다 우리 생활에 깊숙이 파고들어 우리 삶의 모습을 하나씩 바꾸어 가고 있다. 일례로 주차장을 드나들 때 자동으로 번호판을 인식하는 기술이라든지, 사진을 찍을 때 사람 얼굴을 찾아서 네모 박스를 쳐 주는 기술, 인터넷 쇼핑이나 음악 감상할 때 나와 성향이 비슷한 사람들이 좋아하는 아이템이라고 추천을 해 주는 시스템, 외국어로 되어 있는 웹 사이트를 자동으로 한글로 번역해 주는 자동번역시스템 등에서부터 주식 매매, 날씨 예측, 자율주행 등 이루 헤아릴 수 없는 많은 곳에서

* 본 내용은 곽노준 교수(☎ 031-888-9166, nojunk@snu.ac.kr)에게 문의하시기 바랍니다.
** 본 내용은 필자의 주관적인 의견이며 IITP의 공식적인 입장이 아님을 밝힙니다.

이미 초보적인 수준에서부터 보다 차원이 높은 다양한 인공지능 기술이 활용되고 있다.

그럼 도대체 인공지능이란 무엇인가? 인공지능이란 말 그대로 컴퓨터와 같이 인간이 아닌 것이 인간의 지능을 모방하여 인간의 판단이 필요한 분야에서 독자적인 가치 판단 및 결정을 내릴 수 있도록 하는 기술을 통틀어서 일컫는 용어이다. 이 중 특히 기계학습은 백지 상태에 있던 갓난아이가 1년도 안 되어 말을 알아듣고, 의사표현을 하고, 걷고, 뛰는 것을 배우는 과정처럼 컴퓨터가 학습을 통해 새로운 분야의 전문가가 될 수 있게 하는 연구 분야로, 1940년대부터 생물의 신경망을 논리회로로 모델링하고자 하는 연구를 그 시초로 볼 수 있다. 그 후 1957년 인공신경망을 이용하여 영상을 인식하고자 하는 목적으로 퍼셉트론(perceptron)이라는 알고리즘이 개발되었고, 1986년 Rumelhart 등이 퍼셉트론을 여러 층으로 쌓아서 입출력간의 관계를 학습할 수 있도록 한 다층 퍼셉트론(multi-layer perceptron)이라는 구조와 이를 학습하는 역전파(back-propagation) 알고리즘을 개발함으로써 인공신경망의 첫 번째 붐을 이끌었다. 이러한 기반 위에 현재 2006년 이후로 최근 많은 사람들의 입에 오르내리며, 알파고의 성공을 이끌어낸 한 축이기도 한, 신경망 층을 매우 깊게 쌓아서 학습을 수행하는 딥러닝 기술이 엄청난 속도로 발전하고 있다.

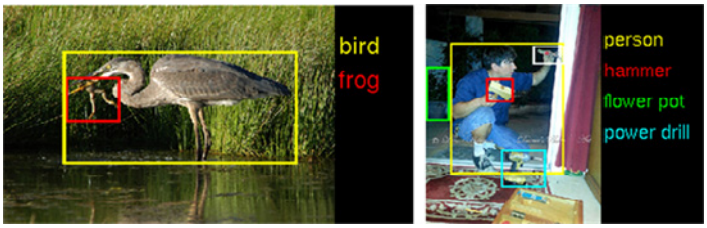
본 고에서는 딥러닝의 여러 가지 분야 중 최근 영상인식 분야에서 기존 방법들보다 월등한 성능을 보이고 있는 컨볼루션 신경망과 음성인식이나 자연어처리 등에 적용되어 뛰어난 성능을 보이는 순환 신경망의 최근 동향을 소개하도록 한다.

Ⅱ. 컨볼루션 신경망

1. 개요

컨볼루션 신경망(CNN)은 영상에 적용이 용이하도록 만들어진 인공 신경망의 한 종류이다. CNN은 Lecun et al.[1]이 1998년 처음 제안하였으며 일반적인 다층 퍼셉트론에서 사용되는 구조와 다르게 컨볼루션 레이어와 풀링 레이어로 이루어져 있다.

CNN은 처음 제안된 이후 성능 면에서 다른 알고리즘에 비해 뛰어나지 못했기 때문에 큰 주목을 받지 못하고 있었다. 하지만 이후 2012년 ImageNet Challenge[2] 의 영상 분류 문제에서 CNN 기반의 알고리즘이 2위를 큰 폭으로 누르고 우승하여 이후 CNN 연구에 불을 지피는 계기가 되었다. ImageNet Challenge는 영상 분류와 객체 검출 분야 경진대회로 이전에 존재하지 않



<자료> Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge", International Journal of Computer Vision 115.3 (2015): 211-252.

[그림 1] ImageNet 영상 및 ground truth 예제

았던 대용량 데이터베이스를 구축하여 대용량 영상 분류 및 객체 검출 분야의 연구를 활성화시키기 위해 2010 년 시작된 워크샵이다. 이 중 영상 분류의 경우 1,000 개의 서로 다른 클래스 영상을 학습하고 테스트 영상이 들어왔을 때 이를 알맞게 분류하는 문제이다. 영상 분류 데이터셋은 100 만 장 이상의 학습 데이터, 5 만 장의 밸리데이션 데이터, 10 만 장의 테스트 데이터로 구성되어 있다. [그림 1]은 이미지넷(ImageNet)의 영상과 ground truth(정답)의 예를 보여준다.

CNN 이 최근 들어 다른 알고리즘에 비해 영상 분류 및 객체 검출에 우수한 성능을 보이고 있는 이유는 크게 세 가지를 들 수 있다. 첫 번째는 Rectified Linear Unit(ReLU)[3]이라는 활성화 함수(activation function)의 도입으로 이전 sigmoid, tanh 등의 활성화 함수에서 나타나던 문제인 그레이디언트 베니싱(gradient vanishing) 문제가 없어진 것이다. Gradient vanishing 은 신경회로망을 학습하는 대표적인 알고리즘인 backpropation 알고리즘에서 낮은 층으로 갈수록 전파되는 에러의 양이 적어짐으로 인해 그레이디언트 변화가 거의 없어져 학습이 일어나지 않는 현상이다. 이 문제로 인해 깊은 인공 신경망의 학습이 어려웠는데 ReLU 의 도입으로 이 문제를 해결하여 깊은 인공 신경망에서도 낮은 층까지 학습이 가능해졌다.

두 번째 이유는 이미지넷과 같은 대용량 데이터베이스의 출현이다. 하드웨어의 발달로 인해 대용량 저장장치가 보편화되었고 Amazon Mechanical Turk[4] 등을 이용한 크라우드소싱이 가능해지면서 대용량 학습 데이터의 정답을 수작업으로 레이블링하는 일이 가능해졌다. 이러한 100 만 장 이상의 대용량 영상 데이터베이스를 바탕으로 여러 층으로 이루어진 CNN 을 학습함으로써 과적합(overfitting) 문제를 해결할 수 있었다. 일반적인 인공 신경망의 경우 학습해야 하는 변수의 개수가 매우 많기 때문에 적은 양의 학습 데이터로는 과적합이 쉽게 일어나게 되는데 대용량 데이터베이스의 출현으로 깊은 인공 신경망을 과적합 없이 학습할 수 있게 된 것이다. 마지막 이유는 Dropout[5]을 활용한 regularization 을 들 수 있다. Dropout 은 인공 신경망의 과적합

을 방지하기 위해 학습 알고리즘 상에서 특정 비율의 뉴런을 무작위로 작동하지 않게 만든 채 학습을 수행하게 된다. 매 iteration 마다 작동하지 않는 뉴런을 다르게 뽑아서 학습을 시켜 각각의 뉴런이 같은 정보를 학습하거나 아무런 정보도 학습하지 않는 것을 방지하였다.

위와 같은 이유로 컨볼루션 신경망은 대용량의 영상 데이터가 존재할 때 영상 분류 및 객체 검출을 효과적으로 수행하며 현존하는 알고리즘 중 가장 좋은 성능을 보이는 것으로 보고되고 있다. 다음 절에서 영상 분류에 사용되는 다양한 CNN 의 구조에 대해 알아보도록 하겠다.

2. CNN 을 이용한 영상 분류 동향

이미지넷 챌린지(ImageNet Challenge)의 영상 분류 성능은 Top-5 에러로 측정이 된다. 테스트 영상에서 가장 확률이 높은 5 개의 부류(class)를 알고리즘을 통해 선택한 뒤 5 개 중 그라운드 트루즈에 해당하는 부류가 있을 경우 정답을 맞춘 것으로 처리해 정답을 맞추지 못한 이미지의 비율을 Top-5 에러로 측정한다. 2012 년 이미지넷 챌린지에서 우승한 AlexNet[6]의 경우 16.4%의 Top-5 에러를 보였다. 2012 년 2 위에 오른 방법이 26.2%의 에러를 보인 것과 비교하면 이는 매우 큰 격차이다. 이 결과로 인해 많은 사람들이 CNN 의 연구에 뛰어드는 계기가 되었다. AlexNet 은 5 개의 컨볼루션 층과 2 개의 fully-connected 층(일반적인 다층 퍼셉트론과 같이 위와 아래의 모든 뉴런이 연결된 구조)으로 이루어져 있다.

이후 2015 년까지 이미지넷 챌린지에 참여한 팀들은 대부분 CNN 기반 알고리즘을 사용하고 있으며 주로 CNN 의 구조를 변형하여 학습의 효율성을 높이고 성능을 향상시킨 방법들이 많이 등장하였다. 2013 년 우승팀인 Clarify 사의 ZFNet[7]은 AlexNet 의 컨볼루션 필터 크기를 줄여 Top-5 에러를 11.7%까지 줄였다. 2014 년 등장한 VGGNet[8]과 GoogleNet[9]은 층의 개수를 늘려 네트워크를 깊게 만들고 필터 크기를 줄이는 것이 네트워크의 표현력을 높이는데 유리하다는 것을 증명하였다. VGGNet 의 경우 모든 컨볼루션 층의 필터 크기를 3×3 또는 1×1 로 고정하고 17 개의 층을 사용해 매우 깊은 CNN 구조를 만들었다. 1×1 컨볼루션의 경우 주위 정보를 포함하지 않음에도 불구하고 차원 축소와 비슷한 효과로 표현력을 높일 수 있음이 밝혀졌다. GoogleNet 은 1×1, 3×3, 5×5, 3×3 pooling 으로 이루어진 인셉션 구조 바탕으로 네트워크를 구축하였다. VGGNet 은 7.33%, GoogleNet 은 6.67%의 Top-5 에러를 기록하여 2013 년에 비해 많은 성능 향상을 보였다.

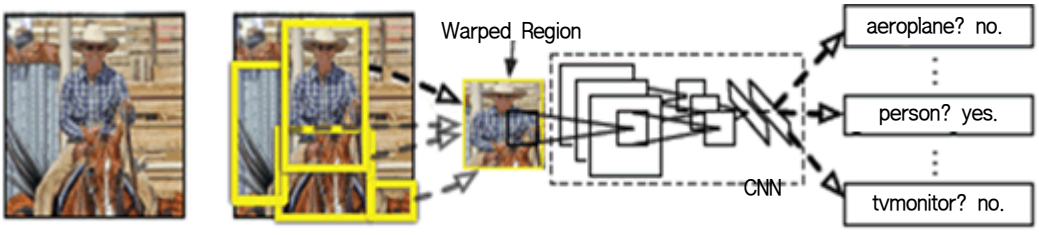
2015 년 영상 분류에서 가장 좋은 성능을 보인 Microsoft research Asia 의 ResNet[10]은 150

개 이상의 층으로 이루어진 네트워크를 효과적으로 학습이 가능하도록 하는 알고리즘을 개발하여 3.57%의 Top-5 에러를 기록하였다. 사람이 이미지넷 데이터를 분류했을 때 5.1% 정도의 Top-5 에러를 가진다고 보고된 것을 참고하면[11], 이는 CNN 이 영상을 분류하는 능력이 사람보다 뛰어남을 의미한다. 이외에도 Batch Normalization[12], Parametric ReLU[13] 등 CNN 의 학습 속도와 성능을 향상시킬 수 있는 방법이 많이 등장하여 이미지넷 영상 분류 문제는 거의 정복된 문제로 여겨지는 분위기이다.

3. CNN 을 이용한 객체 검출 동향

객체 검출 문제는 영상 분류 문제와 달리 영상 한 장의 부류를 분류하는 것이 아니라 영상에서 객체에 해당하는 부분을 찾아서 객체의 외곽상자(bounding box)를 결과로 출력해야 한다. 따라서 한 장의 영상에서도 여러 종류의 객체가 검출될 수 있어 영상 분류보다 어려운 문제로 다루어진다. 객체 검출에 사용되는 CNN 은 주로 영상 분류에서 사용된 CNN 과 같은 구조가 사용되며 CNN 의 성능이 객체 검출의 성능에 큰 영향을 미친다. 다만 외곽상자를 추정하기 위해 객체의 외곽상자에 해당하는 후보군을 먼저 검출하는 작업이 필요하다. 초기 CNN 을 이용한 객체 검출은 후보군 검출을 위한 방법으로 비지도 학습 기반의 방법인 selective search[14], edgeBox[15] 등이 주로 사용되었다. 이렇게 검출된 후보군 부분을 영상에서 잘라내 CNN 을 통해 분류하여 최종적으로 객체를 검출하게 된다. 이러한 프레임워크를 사용한 대표적인 방법이 R-CNN[16]이다([그림 2] 참조).

이후 후보군 또한 CNN 으로 함께 학습시켜 사용하는 방법이 더 좋은 성능을 보이게 되었다. 기존의 후보군 검출 방법은 이미지 한 장에 후보군이 1,000 개 이상 검출되었기 때문에 이를 모



<자료> Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[그림 2] R-CNN 구조도 개요

두 분류하는데 시간이 오래 걸렸다. 반면에 CNN 학습 시에 객체 위치를 함께 이용하여 학습함으로써 객체 검출 시간 및 분류 시간을 크게 단축하였다. 특히, Microsoft research Asia 의 Faster R-CNN[17]은 앞서 설명한 ResNet 과 결합하여 사용해 2015 년 이미지넷 객체 검출 분야에서 2 위와 큰 격차로 우승을 차지하였다.

4. CNN 의 발전 방향 – Generative 모델

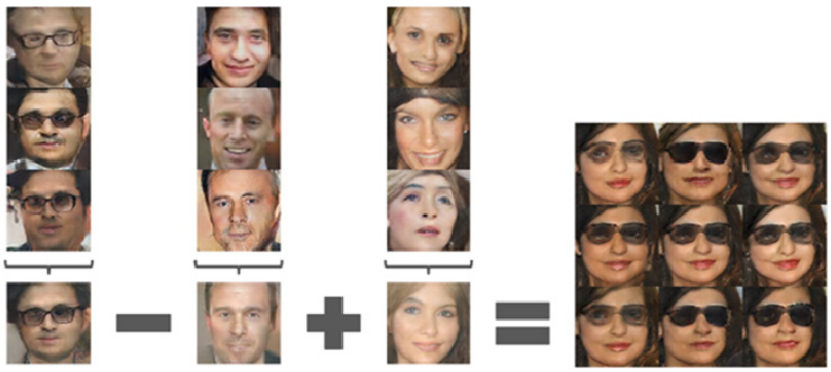
CNN 은 영상 분류에 가장 먼저 활용되면서 인기를 끌기 시작하였으나 컴퓨터 비전 분야 전반에 걸쳐 활용될 수 있는 여지가 많다. 특히, 하위 층에서 간단한 특징을 학습하고 위로 갈수록 고차원적인 특징을 학습하는 성질 때문에 CNN 의 학습은 새로운 레프리젠테이션을 학습하는 과정으로도 볼 수 있다. 최근 CNN 에서 화두가 되고 있는 문제는 generative model 이다. 기존 영상 분류나 다른 여러 분야에서 사용된 CNN 은 주로 discriminative model 을 학습하는 용도로 사용되어 왔다. 이는 데이터 x 가 주어졌을 때 각 부류 C 마다 p(C¦x)를 학습하는 모델이다. 반면에 generative model 은 데이터 x 의 확률분포 θ 의 추정을 통해 p(x¦θ)를 학습하는 모델이다. Generative model 보다 discriminative model 이 분류 등 기본적인 문제에 있어서 우수한 성능을 보인다고 알려져 있지만 Generative model 의 장점은 학습된 모델로부터 새로운 데이터를 생성할 수 있다는 점이다. 따라서 성능이 discriminative model 에 비해 좋지 않음에도 새로 생성된 데이터의 활용 방안이 많기 때문에 중요히 여겨지고 활발히 연구되고 있는 분야이다. CNN 과 같은 인공 신경망을 generative model 에 어떻게 접목시킬 것인가에 대해 많은 연구가 진행되어 왔다. 딥러닝의 시대를 처음으로 열었던 Hinton et al.의 Deep belief network(DBN)[18]의 경우 generative model 로써 층별로 학습시키는 모델이다. 이후 CNN 에서는 discriminative model 의 연구가 주로 이어지다가 최근 variational auto-encoder(VAE)[19]와 generative adversarial network (GAN)[20]의 등장으로 인해 generative model 이 다시 인기를 얻고 있다. VAE 의 경우, 베리에이션(variational) 인퍼런스 문제를 단순화시켜 이를 gradient descent 방법을 이용하여 학습할 수 있도록 만들었다. GAN 의 경우, 데이터의 확률 분포를 학습하는 네트워크와 실제 데이터와 생성된 데이터를 분류하는 네트워크의 미니맥스 게임을 통해 데이터의 확률 분포가 네트워크를 통해 학습이 되게 하였다.

DRAW[21] 네트워크에서는 다음 장에서 좀 더 자세히 설명될 RNN 을 VAE 와 결합하여 이미지를 생성하는 알고리즘을 개발하였다. 또한, 최근 Deep convolutional GAN(DCGAN)[22]에서는

주) DCGAN[21]을 이용하여 생성한 침실의 영상. 실제 침실의 사진이라 할 수 있을만큼 사실적인 영상을 보여준다.
<자료> Radford, Alec et al. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," arXiv preprint arXiv:1511.06434, 2015.

[그림 3] DCGAN[22]를 이용한 침실 영상 생성 예

GAN을 이용하여 더 크고 자세한 이미지를 생성하는 방법을 연구하였다. 그 결과 생성된 얼굴, 침실 등의 영상은 꽤 자연스러운 모습을 보여주고 있다([그림 3, 4] 참조). 특히, 얼굴 데이터로 학습한 DCGAN에서 생성에 입력으로 사용한 확률분포들 간의 벡터 연산을 통해 생성된 새로운 이미지를 생성하는 실험을 통해서 입력으로 사용한 벡터 공간이 고차원적인 의미를 표현하는



주) DCGAN[21]을 이용하여 생성한 얼굴 영상과 입력 벡터 간의 영상을 통해 새로 생성한 영상. 의미적으로 더하기, 빼기 연산을 수행한 결과를 영상으로 생성해 주는 것을 알 수 있다.
<자료> Radford, Alec et al. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," arXiv preprint arXiv:1511.06434, 2015.

[그림 4] DCGAN[22]를 이용한 얼굴 영상 생성 예

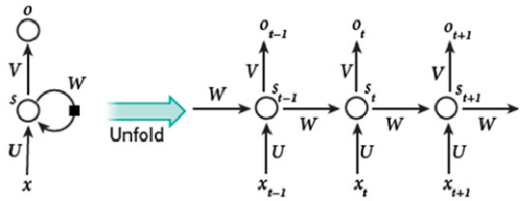
공간이 되도록 학습이 수행된다는 것을 볼 수 있다([그림 4] 참조).

앞으로 이러한 generative model을 활용한 새로운 영상 또는 동영상의 생성 문제가 CNN의 주요 연구 분야가 될 것으로 보여진다.

Ⅲ. 순환 신경망(RNN)의 소개

1. 개요

최근 음성이나 언어 등 연속된 입력 데이터를 다루는 연구자들이 딥러닝 모델 중 순환 신경망(Recurrent Neural Network: RNN)에 주목하고 있다. 순환 신경망은 [그림 5]에 표현한 것과 같이 연속된 데이터 상에서 이전 순서의 히든 노드(hidden node)의 값을 저장한다. 이후 다음 순서의 입력데이터로 학습할 때 이전에 저장해 놓은 값을 이용하게 된다. 결국 학습이 진행되어도 과거 학습의 정보를 잃지 않고 연속적인 정보의 흐름을 학습에 반영할 수 있다. 즉 순환 신경망은 강력한 동적 시스템의 역할을 한다. 학습 방법은 인공 신경망의 역전파 방법을 따르게 되지만 시간 방향의 학습이 추가 되어 backpropagation through time(BPTT)[24] 이라는 변형된 학습방법을 따르게 된다. 이러한 BPTT 학습 방법은 역전파의 거리가 늘어나면서 gradient 값이 폭증 하거나 사라지는 현상이 발생하는 문제점이 있으며 이로 인해 데이터의 길이가 길어질수록 학습은 힘들어진다[25].

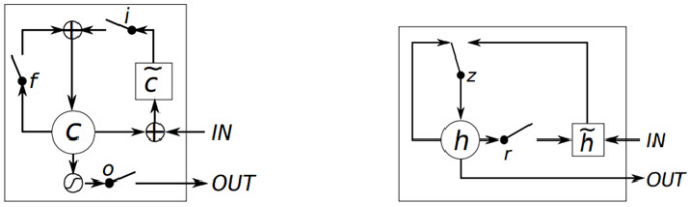


<자료> LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." Nature 521.7553 (2015): 436-444.

[그림 5] RNN의 구조

2. LSTM 및 GRU

기본 인공신경망 구조를 바탕으로 한 순환 신경망 학습의 문제점을 극복하기 위해 Long short-term memory(LSTM) 구조가 제안되었다[26]. 인공신경망의 히든 노드 대신에 LSTM cell을 사용하는 아이디어인데 구간이 길어지더라도 정보를 지속하는데 효과적이다. LSTM cell은 여러 게이트들과 상태값들이 조합된 구조로 이루어져 있다. 게이트는 입력(i), 출력(o), 망각(f)의 세가



(a) Long Short-Term Memory (b) Gated Recurrent Unit

<자료> Chung, Junyoung, et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling." arXiv preprint arXiv:1412.3555, 2014.

[그림 6] LSTM 의 구조와 GRU 의 구조

지 종류가 있으며, 시그모이드(sigmoid) 함수를 통해 0 과 1 사이의 값으로 제한시켜 통과 비율을 정해주게 된다. 입력 게이트는 입력값의 사용 비율을 결정하며 망각 게이트는 이전 단계에서 넘어온 히든 노드 값의 사용 비율을 결정한다. 출력 게이트는 최종 아웃풋의 사용 비율을 정해준다. 즉, 학습을 통해 게이트들은 얼마만큼 정보를 통과시켜줄지 결정할 수 있게 되고 이로 인해 장기간 기억을 보존하여 사용할 수 있게 된다. [그림 6 (a)]는 LSTM 의 구조를 간략히 보여준다.

LSTM 의 기본 구조 이외에도 다양한 변형구조가 연구되고 있다. 그 중 Gated Recurrent Unit (GRU)의 경우 보다 간단한 구조임에도 LSTM 못지 않은 성능을 보여 최근 연구에 많이 사용되고 있다[28]. GRU 의 경우는 게이트의 수가 [그림 6 (b)]와 같이 2 개로 LSTM 보다 적으며, GRU 에서는 리셋 게이트(r)와 업데이트 게이트(z)가 LSTM 의 3 가지 게이트의 역할을 나누어 수행한다. 리셋 게이트는 입력값과 이전단계의 히든 노드의 값을 어떻게 섞을지 그 비율을 정해주어 업데이트 게이트는 리셋게이트를 통과한 후보값과 이전 단계의 히든 노드 값의 조합 비율을 결정하여 최종 결과값을 결정한다. GRU 는 구조가 간단하고 변수의 수가 적어 LSTM 보다 학습시간이 짧게 걸리고 과적합이 덜 일어나는 장점이 있다. 하지만 데이터의 수와 복잡도에 따라 LSTM 이 좀 더 나은 성능을 보이는 경우도 있어 어느 구조가 낫다고 보기 힘들다[27].

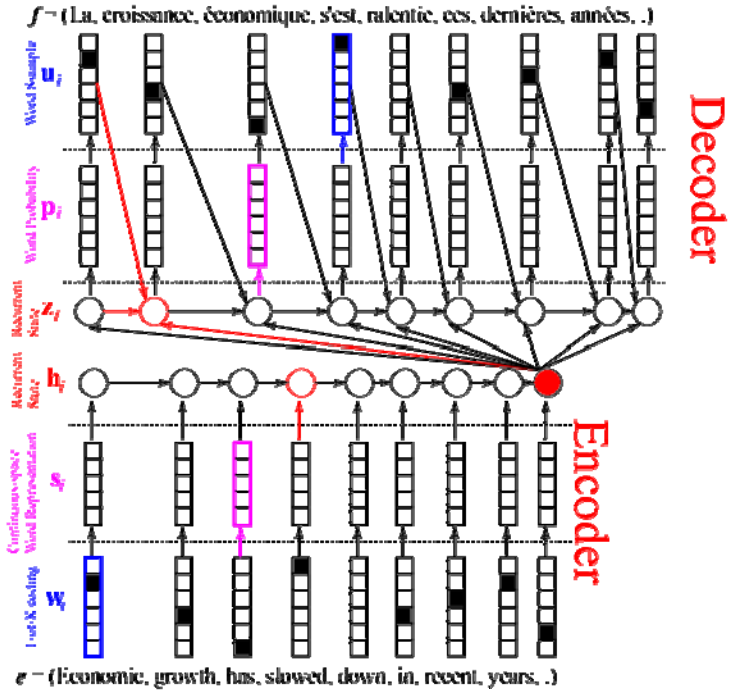
3. 순환 신경망의 응용

가. 자연어처리

순환 신경망은 최근 학습방법 및 구조 면에서 많은 연구가 진행되고 있다. 그와 더불어 음성인식, 자연어처리 등 연속 데이터를 처리하는 분야에 다양하게 응용되고 있다. 그 중 순환신

경망이 자연어로 문장을 이루는 패턴을 잘 익히면서 자연어처리 분야에서 큰 두각을 나타내고 있다. 예를 들어, 영어 자연어 문장을 독일어 및 프랑스 언어로 번역할 때 기존 번역 연구와 달리 순환 신경망을 이용할 수 있다. 최근 뉴럴 기계 번역(Neural Machine Translation: NMT) 모델은 순환 신경망을 기반으로 하여 언어 간의 특징을 스스로 학습하여 좋은 성능을 보여준다[29],[30].

NMT 모델은 [그림 7]과 같이 인코더 네트워크와 디코더 네트워크로 구성되며 두 네트워크 모두 순환 신경망 구조로 이루어져 있다. 앞서 소개한 LSTM 이나 GRU 등의 메모리 셀이 히든 노드를 이루게 된다. 인코더 네트워크의 입력은 문장을 이루는 단어들을 벡터로 변환하여 처리한다. 연속된 단어 벡터의 입력 값은 입력 문장을 의미하는 하나의 표현 벡터를 생성하여 디코더 네트워크로 넘겨지게 되며 디코더 네트워크는 인코더 네트워크에서 넘어온 벡터를 기반으로 다른 언어의 자연어 문장을 생성한다. 디코더 네트워크는 확률 모델로서 후보 단어 중 가장 높은 확률의 단어를 연속적으로 생성한다. 첫 번째 단어를 생성한 후 그 단어가 다시 입력값이 되어 두 번째 단어의 생성에 영향을 주는 방식으로 자연스러운 문장을 만드는 패턴을 따르게



<자료> "Introduction to Neural Machine Translation with GPUs," <https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-gpus-part-3/>

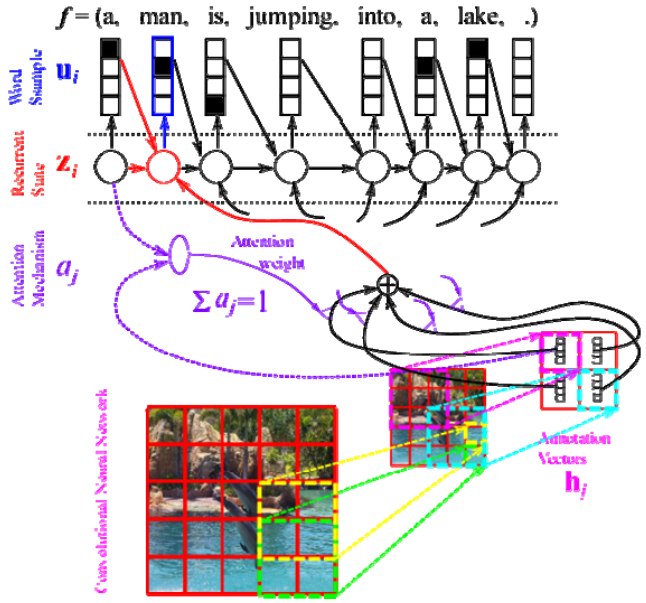
[그림 7] 뉴럴 기계 번역 모델의 구조

된다. 이를 통해 서로 다른 언어 간에도 자연어 입력과 출력이 가능하게 되는 것이다[30].

나. 이미지 설명 생성

순환 신경망을 이용한 모델은 단지 언어에만 국한되지 않는다. 순환 신경망은 앞서 소개한 CNN 과 결합하여 이미지와 언어 간의 번역 모델을 구축할 수 있다[31]. [그림 8]과 같이 이미지 설명 생성 모델을 자세히 살펴보면 CNN 을 인코더 네트워크로 대체하여 입력 이미지의 특징을 담은 표현 벡터를 디코더 네트워크에 넘기게 되고, 이후 디코더 네트워크는 이미지의 내용에 따라 자연어 문장을 생성하게 된다. 본 모델과 마찬가지로 연속적으로 출력값이 다음 순서의 입력값이 되는 방식으로 자연스러운 문장 패턴을 생성하게 된다. 더 많은 수의 이미지와 설명 데이터를 이용하여 학습할수록 이미지의 내용을 자연스럽게 설명해주는 모델을 만들 수 있게 된다.

이 모델에서는 NMT 모델에서의 인코더 네트워크를 CNN 으로 대체하여 다른 종류의 입력값을 사용하였다고 볼 수 있고 딥러닝 모델의 큰 장점으로 생각할 수 있다. 인공신경망이 이미지, 언어, 소리, 비디오 등 다양한 종류의 데이터를 표현 벡터로 변환하고 이를 학습하는데 적합하기 때문이다. 또한, 사용자가 입력 데이터를 모델에 입력만 하게 되면 모델 스스로 학습하고 특



<자료> "Introduction to Neural Machine Translation with GPUs," <https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-gpus-part-3/>

[그림 8] 이미지 설명 생성 모델의 구조

징을 찾아 결과를 도출하는 엔드 투 엔드(end-to-end) 모델링이 가능하다는 점이 딥러닝 모델의 또 다른 장점이다.

Ⅳ. 결론

본 고에서는 최근 이미지넷, 알파고 등의 성공에 따른 인공지능의 붐을 이끈 가장 큰 견인차 역할을 하고 있는 딥러닝 기술에 관해 소개하였다. 딥러닝 기술 중 특히 영상인식 및 검출 분야에서 활발히 사용되고 있는 CNN과 이의 최신 동향인 generative model들에 대해 살펴보았으며, 자연어처리 및 이미지 설명 자동 생성에서 사용되고 있는 RNN에 대해 살펴보았다. 딥러닝 기술은 최근 몇 년간 빅데이터 기술과 GPU 처리 속도의 향상 등에 힘입어 빠르게 발전하고 있으며, 많은 연구자들이 활발히 활동하고 있는 분야로 최근의 연구 동향에 비추어 보았을 때 영상인식 및 자연어처리 등의 분야 외에도 다양한 분야에 응용될 수 있을 것으로 보인다.

[참고문헌]

[1] LeCun, Yann, et al. "Gradient-based learning applied to document recognition," Proceedings of the IEEE 86.11, 1998, 2278-2324.

[2] Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge," International Journal of Computer Vision 115.3, 2015, 211-252.

[3] Nair, Vinod, and Geoffrey E. Hinton. "Rectified linear units improve restricted boltzmann machines," Proceedings of the 27th International Conference on Machine Learning(ICML-10). 2010.

[4] https://www.mturk.com/

[5] Srivastava, Nitish, et al. "Dropout: A simple way to prevent neural networks from overfitting," The Journal of Machine Learning Research 15.1, 2014, 1929-1958.

[6] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems. 2012.

[7] Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks," Computer vision–ECCV 2014. Springer International Publishing, 2014. 818-833.

[8] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[9] Szegedy, Christian, et al. "Going deeper with convolutions," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[10] He, Kaiming, et al. "Deep Residual Learning for Image Recognition," arXiv preprint arXiv:1512.03385, 2015.

[11] http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/

[12] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv preprint arXiv:1502.03167, 2015.
[13] He, Kaiming, et al. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," Proceedings of the IEEE International Conference on Computer Vision. 2015.
[14] Uijlings, Jasper RR, et al. "Selective search for object recognition," International journal of computer vision 104.2, 2013, 154-171.
[15] Zitnick, C. Lawrence, and Piotr Dollár. "Edge boxes: Locating object proposals from edges," Computer Vision–ECCV 2014. Springer International Publishing, 2014. 391-405.
[16] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation," Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.
[17] Ren, Shaoqing, et al. "Faster R-CNN: Towards real-time object detection with region proposal networks," Advances in Neural Information Processing Systems. 2015.
[18] Hinton, Geoffrey E., "Deep belief networks," Scholarpedia 4.5, 2009, 5947.
[19] Kingma, Diederik P., and Max Welling., "Auto-encoding variational bayes," arXiv preprint arXiv:1312.6114, 2013.
[20] Goodfellow, Ian, et al. "Generative adversarial nets," Advances in Neural Information Processing Systems. 2014.
[21] Gregor, Karol, et al. "DRAW: A recurrent neural network for image generation," arXiv preprint arXiv:1502.04623, 2015.
[22] Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," arXiv preprint arXiv:1511.06434, 2015.
[23] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton., "Deep learning," Nature 521.7553, 2015, 436-444.
[24] Williams, Ronald J., and David Zipser., "Gradient-based learning algorithms for recurrent networks and their computational complexity," Back-propagation: Theory, architectures and applications, 1995, 433-486.
[25] Bengio, Yoshua, Patrice Simard, and Paolo Frasconi. "Learning long-term dependencies with gradient descent is difficult," Neural Networks, IEEE Transactions on 5.2, 1994, 157-166.
[26] Hochreiter, Sepp, and Jürgen Schmidhuber., "Long short-term memory," Neural computation 9.8, 1997, 1735-1780.
[27] Chung, Junyoung, et al., "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint arXiv:1412.3555, 2014.
[28] Cho, Kyunghyun, et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014.
[29] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio., "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473 2014.
[30] "Introduction to Neural Machine Translation with GPUs," <https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-gpus-part-3/>
[31] Xu, Kelvin, et al., "Show, attend and tell: Neural image caption generation with visual attention," arXiv preprint arXiv:1502.03044 2015.