딥러닝 기반 비디오 스토리 학습 기술

허민오 · 김경민 · 장병탁*

1. 서 론

최근 딥러닝으로 대표되는 기계학습 기술의 발전에 힘입어 이미지 및 비디오 상의 물체 인식, 물체 검출 기술이 크게 발전하였다[1,2,3]. 학계의 많은 연구자들은 이러한 기술을 적극적으로 활용하여 영상 컨텐츠를 이해하는 기계를 구현하기 위한 연구를 지속하고 있다. 즉, 단순히 물체의 존재만이 아니라 사진이 찍힌 장소는 어디인지[4], 누구의 얼굴이 나왔는지[5,6], 어떤 행동이 이루어지고 있는지를 다루고[7,8,9], 여기에서 한 단계 더 나아가 사건의 전후 관계를 다루는 스토리 학습[10,11,12]까지도 관심을 보이고 있다.

이와 같이 영상 컨텐츠 이해 기술이 발전하게 된 세 가지 핵심 요인으로 딥러닝 기술 외에도 GPU 등의 하드웨어로 인한 계산 속도의 향상, 학습에 사용할 수 있는 빅데이터의 등장을 들 수 있다. 즉, 빅데이터를 재료로 딥러닝 기술을 이용하여 정보의 다양한 표현 방식을 학습하고 영상의 다양한 요소를 인식하는 지능 기술의 비약적인 발전이 최근 계속되고 있다.

※ 교신저자(Corresponding Author) : 장병탁, 주소 : 서울특별시 관악구 관악로 1 서울대학교 신공학관 302동 323호, 전화: 02-880-1847, FAX: 02-875-2240, E-mail : btzhang@bi.snu.ac.kr
* 서울대학교 공과대학 컴퓨터공학부
(E-mail : {moheo,kmkim}@bi.snu.ac.kr)

이에, 본 고에서는 딥러닝 기술을 통해 이미지와 비디오로부터 영상의 이해를 도모하는 최근까지의 연구와 기법들을 소개하고자 한다. 특히, 앞에서 언급한 핵심요소를 고려하여 어떠한 딥러닝 기술을 통하여, 어떠한 데이터로 무엇을 가르치고자 했는지를 중심으로 살펴볼 것이다.

이후의 구성은 다음과 같다. 2 절에서 학습의 주요 모듈이 되는 기반 딥러닝 기술을 소개하고, 3 절에서 영상 인식과 이해를 위한 학습 기술을 정리한다. 4 절에서는 질의응답(QA)을 중심으로 한 비디오 스토리 학습 기술을 살펴보고, 5 절에서 비디오 질의응답(VQA) 기술을 로봇에 적용한 예인 뽀로로봇(Pororobot)을 소개한다. 끝으로 6 절에서 결론을 맺는다.

2. 딥러닝 기술

본 절에서는 이미지와 비디오 학습에 쓰이는 주요 딥러닝 모델과 기술을 소개하여 후술할 기술에 대한 이해를 높이고자 한다.

2.1 영상처리를 위한 딥러닝 모델

2.1.1 컨볼루션 신경망

먼저 영상처리에 가장 널리 사용되는 컨볼루션

신경망(Convolutional Neural Network, 이하 CNN)을 소개한다. 이미지와 비디오는 기본적으로 픽셀마다 색상(RGB 또는 gray) 값을 가지는 매우 큰 차원의 벡터로 표현된다. 해상도가 낮은 이미지, 예를 들면 40 × 40픽셀의 작은 이미지 정보를 표현하는 데에도 1600차원의 벡터가 필요하기 때문에 처리와 학습에 필요한 계산량은 매우 크다. 이러한 어려움을 회피하기 위해 CNN은 데이터 차원 간 기하적 연관성을 이용하도록 제안되었다[13].

CNN은 모든 데이터 차원을 동시에 다루기보다는 공간적으로 가까이 있는 국부 영역의 패턴만을 모델링하도록 하여 모델의 복잡도를 크게 줄이고, 이러한 국부 영역 패턴을 필터(filter) 또는 커널(kernel)과 같이 사용하여 전 영역에서 공통적으로 모델링이 되도록 하였다. 이미지 데이터의 경우 이러한 특징에 매우 적합하다. 왜냐하면 이미지는 국부적 패턴이 모여 전체적 패턴을 구성하는 계층적 특성이 있으며, 이 국부적 패턴은 위치에 무관하게 전 영역에서 동일하게 사용될 수 있다. 만약, 이미지 안에 선으로 그린 사과가 있다고 할 때, 사과를 표현하는 윤곽선은 사과의 좌우측 테두리에 상관없이 동일한 패턴을 쓸 수 있다. 또한, 이 윤곽선(또는 사과 전체)는 이미지 내에서 어디에 있더라도 위치만 달라질 뿐 동일한 필터가 반응을 보이도록 하는 것이 효율적이다. 이렇게 커널을 전체 이미지 영역을 움직이면서 겹치는 정도를 계산하는 과정을 컨볼루션(convolution) 연산이라고 하며, 컨볼루션 신경망 모델의 가장 주요한 요소이다. 이를 ReLU (Rectified Linear Unit) 연산을 통해 음수인 경우 0으로 바꾸어 패턴 검출의 역할을 더한다. 이를 최대 풀링 연산을 통해 특정 영역 내의 최대값으로 대체해 출력함으로써 해상도를 낮추어 상위 계층에서 추상화된 정보를 다룰 수 있도록 한다(그림 2 참조). 이와 같은 연산의 조합을 갖는 여러 계층을 두어, 픽셀의 조합에서 윤곽선을 검출하고, 상위 층으로 갈수록 점차 추상화 정도가 높은 도형, 물체를 단계적으로 인식할 수 있다[2].

여기에 문제 해결에 적합한 계층의 수와 필터 크기 및 개수를 찾고자 시도하면서, 일부 은닉노드를 확률적으로 꺼버리는 방법인 dropout[14],

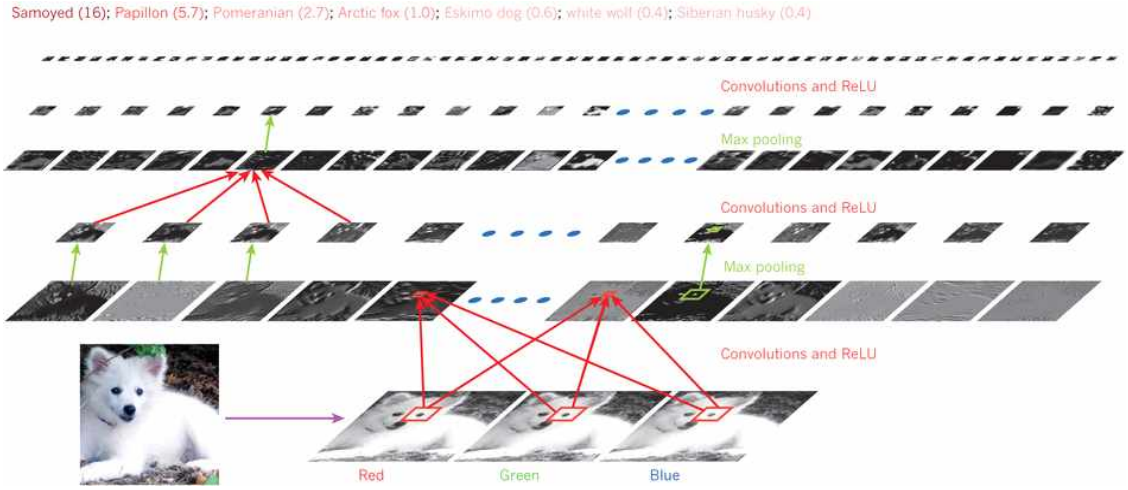


그림 1. 컬러 이미지를 위한 딥 컨볼루션 신경망(Deep CNN)의 작동방식 [1]. 계층별로 주어진 다수의 필터에 대해 컨볼루션과 ReLU 연산, 풀링 연산을 반복하면서 계층을 따라 올라간 후, 최종 계층에서 물체 인식의 결과를 확률적으로 나타낸다.

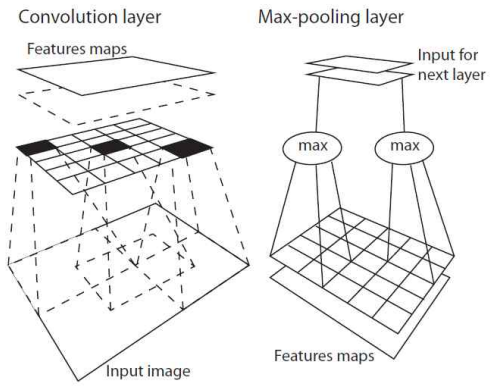


그림 2. 컨볼루션 연산과 최대 풀링 연산[18]

내부 네트워크를 두는 방법, 계층을 넘어 연결을 부여하는 방법 등을 조합한 다양한 변형 모델이 연구되었으며, 그 결과로 22 계층을 가지는 GoogLeNet[15], 152 계층을 갖는 Deep Residual Network[16]까지 발표되었다. 이 과정에서 학습된 딥러닝 모델은 과거에 영상처리에 자주 쓰였던 HOG, SIFT, SURF, 해리스 코너와 같은 여러 필터들보다 여러 응용에서 더 좋은 성능을 보이면서 영상 표현 인자로서 영상처리의 다양한 문제에 널리 적용되고 있다[17].

물체검출 문제에도 컨볼루션 연산을 이용한 딥러닝기법인 R-CNN[3]이 제안되었다. 단순히 이미지 안에 어떤 물체가 있는지만 다루지 않고 무엇이 어디에 있는지 검출해낸다. 이미지를 격자형 영역으로 나누고 각 영역이 검출영역에 속하는지 여부를 다루도록 한 영역제안망(region proposal network)을 모델 안에 포함한 것이 특징이다.

2.1.2 영상 재생성을 위한 신경망

CNN은 데이터 집합에 정답이 표지(label)로서 함께 주어지는 감독학습(supervised learning)의 틀 안에서 제안되었다. 하지만, 데이터에 적합한 표지가 있는 경우는 흔하지 않고, 이를 만드는 작업은 많은 비용이 필요하다. 반면, 인터넷과 웹에 널려 있는 데이터의 양은 막대하므로 무감독학습(unsupervised learning)이 가능한 딥러닝 기법도 다수 제안되었다. 무감독학습은 표지가 없기 때문에 데이터의 전체적 특성을 학습하게 되므로, 입력데이터를 재생성하는 생성적 학습을 다루게 된다. 즉, deep generative model을 통한 오토인코더(auto-encoder)에 대한 연구를 주로 다룬다. 대표적인 예로 Deep Belief Network (DBN)[19], Deep Boltzmann Machine (DBM)[20], Variational Auto-encoder (VAE)[21], Generative Adversarial Network (GAN)[22] 등이 있다.

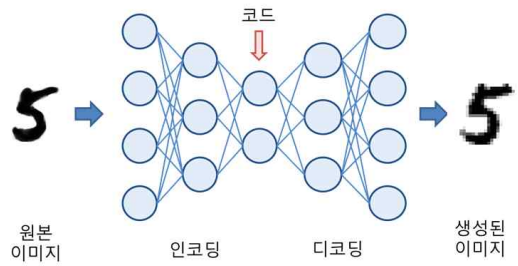


그림 3, 오토 인코더 개념도

이 중에 VAE와 GAN의 변형 모델들이 이미지 생성에 좋은 품질을 보여 최근에 많은 관심을 받고 있다. 특히, GAN은 DBN, DBM과는 달리 디코더 역할을 하는 생성모델의 사전 분포가 단순하게 주어질 수 있다는 특징이 있다. 즉, 일상을 담은 이미지들이 어떤 구조적 특성에 따라 인자 공간 안의 임베딩으로써 다양체(manifold)를 구성하고 있다면, 자연스럽게 비선형 차원축소를 하는 효과를 가지게 되는 장점이 생긴다. VAE는 생성모델 외에 사전 분포의 파라미터를 다루는 신경망을 가진다. 학습 데이터를 통해 이 신경망을 조절하여 생성모델이 학습데이터와 유사한 분포의 데이터를 생성하도록 사전분포를 조절한다. 반면, GAN은 생성모델이 생성하는 데이터가 본래 학

습데이터에 속하는 것인지 진위 여부를 분류하는 신경망을 가진다. 학습 데이터를 통해 이 분류기가 더 이상 구별하지 못하도록 생성모델을 학습시킨다.

2.2 언어처리를 위한 딥러닝 모델

딥러닝 기술의 도입으로 획기적인 발전이 계속되는 또다른 응용분야는 언어처리이다.

딥러닝이 크게 기여한 첫 번째 언어처리 기술은 단어 임베딩(embedding)[23]이다. 과거에는 단어를 표현하기 위해 사전의 단어 수만큼 크기를 가지는 벡터를 이용하여 해당 단어의 존재 여부 또는 빈도를 표시하는 방법을 사용하였다. 단어 임베딩기술에서는 단어를 고차원 연속 공간에 할당하되, 의미가 유사하면 거리도 가깝도록 임베딩된다. 단어 의미의 유사성은 문장 안에서 인접한 단어의 분포가 얼마나 유사한지를 기준으로 판단하며 기본적인 신경망 구조로 학습이 가능하다.

두 번째 중요기술은 딥러닝 기반 시퀀스 모델인 순환신경망(recurrent neural network, 이하 RNN)이다. RNN은 강력한 동적 시스템으로서, 입출력 사이에 은닉 계층이 있고, 이 안에 연속 벡터로 표현되는 상태값을 갖는다. RNN의 파라미터는 그림 4의 좌측과 같이 입력단 행렬 U, 상태 전이행렬 W, 출력단 행렬 V 만으로 구성되지

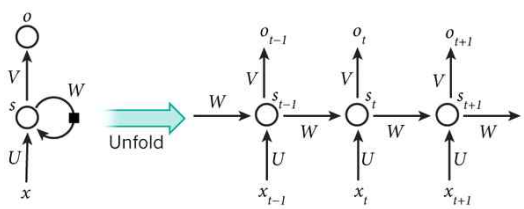


그림 4. 순환신경망 (recurrent neural network) [1]. 모델 파라미터는 U, W, V 뿐이지만, 추론 단계에는 오른쪽과 같이 시퀀스 길이만큼 펴지면서 심층 구조가 나타남.

만, 추론을 수행할 경우 우측과 같이 시퀀스의 길이만큼 펼쳐지면서 심층 구조가 생성된다. 이로 인해 학습에 있어 오류 역전파 과정에서 오류의 미분값이 지나치게 커지거나 사라지는 현상이 나타났으나, ReLU의 등장으로 개선되었다. 그림 4의 관계를 식으로 표현하면 아래와 같다.

$$s_t = g(U \bullet x_t + W \bullet s_{t-1})$$

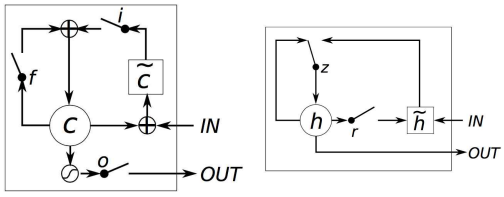
단, g는 smooth한 유계 함수이다. 예를 들면, 시그모이드 함수, 하이퍼볼릭 탄젠트 함수(tanh), ReLU 함수가 이에 해당된다.

이러한 구조만으로는 긴 시간 간격 간의 연관성을 다루기 어려우므로, 그림 5의 Long Short-term Memory (LSTM)[24], Gated Recurrent Unit (GRU)[25]과 같이 은닉 계층의 상태값과 입력에 따라 입출력, 전이 정보의 흐름을 조절할 수 있는 다소 복잡한 모델들이 제안되었다. 그림 5의 LSTM에서 볼 수 있듯이 입력(input), 망각(forget), 출력(output) 세 개의 게이트를 통해 입출력과 기억을 조절한다. GRU는 LSTM을 간략화한 모델이며 갱신(z), 리셋(r)의 두 개 게이트를 통해 정보의 흐름이 조절된다. 두 모델은 기능과 성능 면에서 거의 동일하지만, GRU의 파라미터 수가 좀 더 적다.

구체적으로는 j번째 LSTM 유닛은 t 시점에 메모리 c_t^j를 가지고, RNN 은닉상태값에 해당되는 s_t^j는 LSTM 유닛의 출력이며 다음과 같다.

$$s_t^j = \sigma_t^j \bullet \tanh(c_t^j)$$

여기서 σ_t^j는 출력 정도를 조절하는 출력 게이트 값이며, 다음과 같이 정해진다.



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

그림 5. LSTM과 GRU의 구조[26]. 입출력, 은닉상태값 간에 게이트가 추가됨

$$\sigma_t^j = \sigma(U_o \bullet x_t + W_o \bullet h_{t-1} + V_o \bullet c_t)^j$$

여기서 σ는 시그모이드 함수이고, V_o는 대각 행렬이다. 메모리셀 c_t^j는 다음과 같이 망각, 입력 게이트와 새로운 입력을 반영하여 갱신된다.

$$c_t^j = f_t^j \bullet c_{t-1}^j + i_t^j \bullet \tilde{c_t^j}$$
$$\tilde{c_t^j} = \tanh(U_c \bullet x_t + W_c \bullet h_{t-1})^j$$

윗 식의 입력, 망각 게이트 값은 아래와 같다.

$$f_t^j = \sigma(U_f \bullet x_t + W_f \bullet h_{t-1} + V_f \bullet c_{t-1})^j$$
$$i_t^j = \sigma(U_i \bullet x_t + W_i \bullet h_{t-1} + V_i \bullet c_{t-1})^j$$

단, V_i와 V_f는 대각 행렬이다.
반면, GRU에 대한 부분은 아래와 같다.

$$h_t^j = (1 - z_t^j)h_{t-1}^j + z_t^j \tilde{h_t^j}$$
$$z_t^j = \sigma(U_z \bullet x_t + W_z \bullet h_{t-1})^j$$
$$\tilde{h_t^j} = \tanh(U \bullet x_t + W \bullet (r_t \odot h_{t-1}))^j$$
$$r_t^j = \sigma(U_r \bullet x_r + W_r \bullet h_{t-1})^j$$

단, $\odot$ 연산은 요소별 곱을 의미한다.

이러한 순환신경망 모델을 이용하여 텍스트를 생성하는 흥미로운 응용 연구가 다수 소개되었다. 기계가 중국어로 된 시를 생성하거나[27], 낱글자 단위로 학습한 신경망으로 셰익스피어의 글, 위키피디아의 구조화된 글, 리눅스 소스코드를 생성하기도 하였으며[28], 영화 시나리오를 스스로 작성하기도 하였다[29]. 실제로 이 시나리오를 기반으로 'sunspring'이라는 제목의 단편 영화가 촬영되기도 하였다[1].

순환신경망 모델은 이미지 생성에도 적용되었는데, 대략적인 아이디어는 다음과 같다. 픽셀 생성기 또는 작은 영역을 위한 이미지 생성기를 별도로 두고, RNN으로 캔버스(이미지 전체) 상의 이미 생성된 결과와 생성기와의 관계를 표현하도록 하여 고품질의 이미지 생성을 도모한다[30,31].

2.3 질의응답(QA)을 위한 딥러닝 모델

전통적인 QA 시스템은 정보검색(information retrieval)과 긴밀히 연결되어 있다. 즉, 기존에는 질문에 대한 답을 내기 위해 주로 키워드 기반 검색을 수행하였다. 그러나, QA 문제에서도 딥러닝 방식의 새로운 도구가 개발되었으며, 대표적인 모델은 메모리망(memory network)이다[32,33]. 메모리망을 학습하기 위한 기본 재료는 주어진 지문과 지문의 내용을 바탕으로 한 질문·답 쌍이며, 기존의 신경망 구조에 명시적으로 메모리를 추가한 것이 특징이다. 메모리망은 크게 4 가지 모듈로 구성되어 있다. 질의 입력을 인자로 변형하는 모듈 (I), 새 입력을 반영하여 메모리를 업데이트 하는 모듈 (G), 메모리 내용에 따라 새로운 출력을 만드는 모듈 (O), 출력을 이용하여 답을

1) https://www.youtube.com/watch?v=LY7x2Ihqjmc

생성하는 모듈 (R)이다. 그림 6에서 정보의 흐름에 따른 정보의 처리 과정이 도식화되어 있다. 좌측에 지문(sentences)이, 아래쪽에 질의(question) q가 주어지면, 질의를 고려한 문장에서 주의(attention) p가 계산되고, 이를 가중합한 출력 o와 질의 q를 통합한 결과인 u를 거쳐 답 a를 얻는다. 필요에 따라 메모리를 여러 번 거치면서 u를 보정하고 답을 낼 수도 있다.

메모리망 안에서 쓰이는 문장 자체를 RNN으로 표현하는 동적 메모리망(dynamic memory network)[34]도 발표된 바가 있으며, 문장 외에도 비디오의 이미지들을 입력으로 다룬 Deep Embedded 메모리망[12]도 제안되었다.

3. 영상 이해와 Visual QA

3.1 영상 이해를 위한 학습 데이터의 대두

기계학습을 위해서는 학습 데이터가 반드시 필요하다. 물체를 촬영한 이미지 상에 어떠한 물체가 있는지 표지가 주어진 경우, 감독학습을 통해 학습을 수행할 수 있다. 만약, 이미지 상에 여러 물체

그림 7. 이미지 데이터에 부여된 다양한 보조정보의 예. ImageNet은 (a)와 (b) 형식, MS COCO는 (d)형식도 제공. VQA는 (e) 형식, Visual Genome은 (f)도 제공한다[36,37,38].

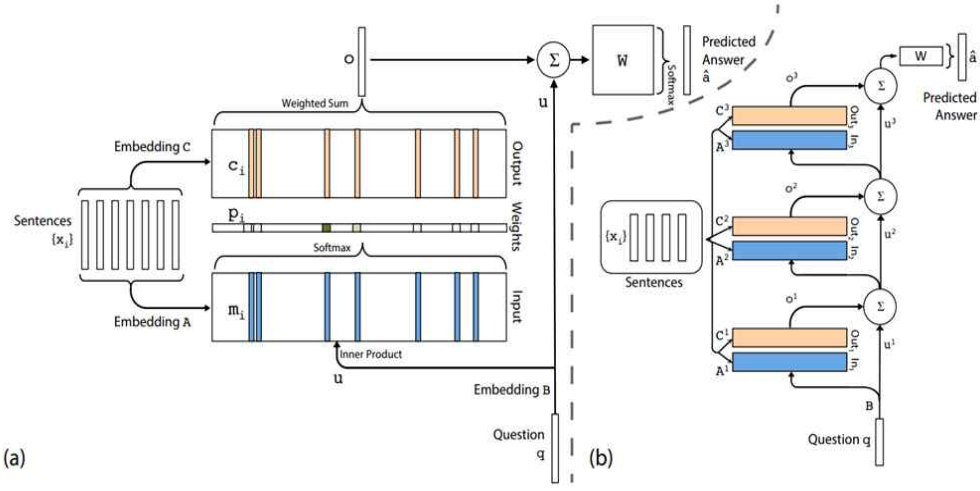


그림 6. 질의부터 응답까지 딥네트워크 구조로 연결된 end-to-end 메모리망[33]. (a)는 질의와 응답 사이의 계층을 1개만 둔 경우, (b)는 3개의 계층을 둔 경우이다.

가 있는 경우 테두리를 쳐서 어디에 있는지도 알려준다면, 물체를 찾는 문제도 다룰 수 있다. 만약, 이미지 상에 픽셀 단위로 영역을 잡아 물체를 알려준다면, 물체 단위로 영상을 쪼개는 문제인 의미적 영상 분할(semantic segmentation)을 다룰 수 있다(그림 7의(c)). 영상의 상황을 설명할 수 있는 정보가 풍부할수록 보다 다양한 흥미로운 기계지능을 구현할 수 있다.

과거에는 이러한 데이터 수집이 아주 어려운 일이었지만, 인터넷이 보편화되며 수십, 수백 만 단위의 영상이 공유되었다. 이러한 대규모 데이터에 그림 7과 같이 크라우드소싱을 통해 구축된 보조 정보가 제공되는 다양한 데이터 집합이 공개되고 있다. 대표적인 예로는 ImageNet[35], MS COCO[36], MS COCO의 일부에 QA를 덧붙인 Visual QA 데이터집합[37], 이미지에 묘사글과 QA, 객체-행동 지식표현을 덧붙인 Visual Genome[38]이 있다.

3.2 영상 이해를 위한 딥러닝 기술

위와 같이 이미지에 여러 연관 정보를 덧붙인 빅데이터가 나타남에 따라 기계가 이미지의 패턴에 따른 연관된 정보를 배우고 영상을 이해하는 능력

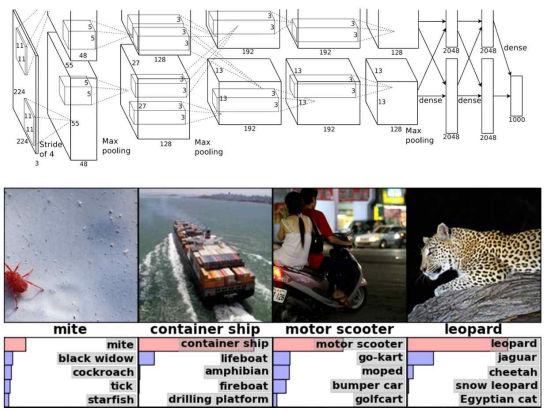


그림 8. AlexNet의 구조도와 실제 분류 사례[2]

을 갖출 수 있게 되었다. 이미지 속에 나타난 물체는 촬영 조건에 따라 워낙 변화가 다양하므로 이를 극복할 만큼의 충분한 데이터가 필요하다. 이미지 안에 나타나는 다양하고 복잡한 패턴 중에 문제 해결에 적절한 표현을 찾는 과정은 기존에는 컴퓨터 비전 연구자들의 고된 엔지니어링을 거쳐야 하였으나, 딥러닝 기술이 등장하여 학습을(representation learning) 통해 보다 질 좋은 표상을 자동으로 얻을 수 있게 되었다.

분류 문제: 딥러닝이 성공적으로 기여한 분야 중 하나는 이미지로 제공되는 물체가 어떤 표지에 속하는지 분류하는 문제이다. 2012년에 발표된 AlexNet[2]은 이러한 문제의 해법을 딥러닝으로 대체한 대표적인 성공사례이다. 학습데이터로 쓰인 ImageNet은 약 120만 장의 컬러사진에 표지가 1000개인 까다로운 데이터집합이며, 2010년부터 매년 대회의 재료로 제공되었다. 그림 8의 AlexNet은 2 등과 큰 차이를 보이며 우승하였고 이후 대회에서는 대부분의 참가팀이 딥러닝 기술을 필수로 사용되게 되었다.

2015 년에는 Deep Residual Network[16]이 ILSVRC, COCO 챌린지를 모두 우승하였으며 데이터만 놓고 보았을 때 사람의 분류성능을 능가하였다[39]. 이 모델은 계층 수가 152에 달하는 거대한 딥네트워크이다. 이와 같은 딥네트워크들은 상위 계층까지 지속적으로 일어나는 추상화 덕분에 상위 계층의 출력이 이미지를 나타내는 좋은 인자로 사용할 수 있음이 밝혀졌고 다양한 용도로 널리 쓰이고 있다.

묘사글 생성 문제: 사진 안의 물체를 넘어서 사진 안의 상황을 묘사하는 기술은 RNN의 기술적 문제해결과 학습데이터의 등장으로 시작되었다. MS COCO 데이터는 개별 사진별 묘사글이 5개씩 제공되었고, 구글에서 이 데이터를 이용하

여 이미지를 위한 캡션 생성기를 연구하여 발표하였다. 구조는 그림 9와 같이 CNN으로 이미지 인자를 인코딩하고 이 값을 디코더 역할을 하는 LSTM에 은닉상태값으로 전달하여 문장을 생성한다[40]. 여기에 어떤 단어가 이미지의 어느 부분에 크게 반응하는지 학습하는 주의(attention) 기제를 반영하여 개선된 성능을 얻었다[41]. 묘사글이 대개는 이미지 안의 물체와 연관된 경우가 많으므로, 물체검출 결과와 묘사글 간의 관계를 학습을 먼저 하여, 묘사글 생성학습의 효율을 높이기도 하였다[42,43,44].

묘사글 검색 문제: 상기 접근 방법은 이미지를 인코딩한 결과를 바탕으로 문장을 디코딩하는 접근인데, 이미지와 문장을 모두 인코딩하는 접근도 생각해 볼 수 있다. 즉, 입력 이미지와 문장을 동일한 임베딩 공간에 할당하고, 대응되는 이미지와 문장이 가까워지도록 학습을 시키면, 유사한 이미지와 문장들이 가까운 영역에 모이게 된다. 참고로, 이 과정은 zero-shot 학습[45]과 유사하다. 이 경우, 생성하기보다는 검색하는 문제가 되며, 평가방식에 따라 오류가 다소 나더라도 실제적으로는 의미가 통하는 결과를 더욱 효율적으로 찾을 수 있다[46].

위의 연구들은 대개 MS COCO 데이터에 기반을 두고 있으나, 새로운 데이터에 눈을 돌린 사례도 있다. 위키피디아에는 개념설명을 위해 사진이 함께 포함된 경우가 많은데, 이를 학습에 활용한 것이다[47]. 접근 방법은 [45]와 개념적으로 유사하다. 먼저 위키피디아 글의 인자와 이미지에서 CNN을 통과시킨 인자를 동일 임베딩 공간에 할당한다. 그 후, 대응되는 관계는 가깝게, 그 외의 것은 멀어지도록 학습을 수행한다.

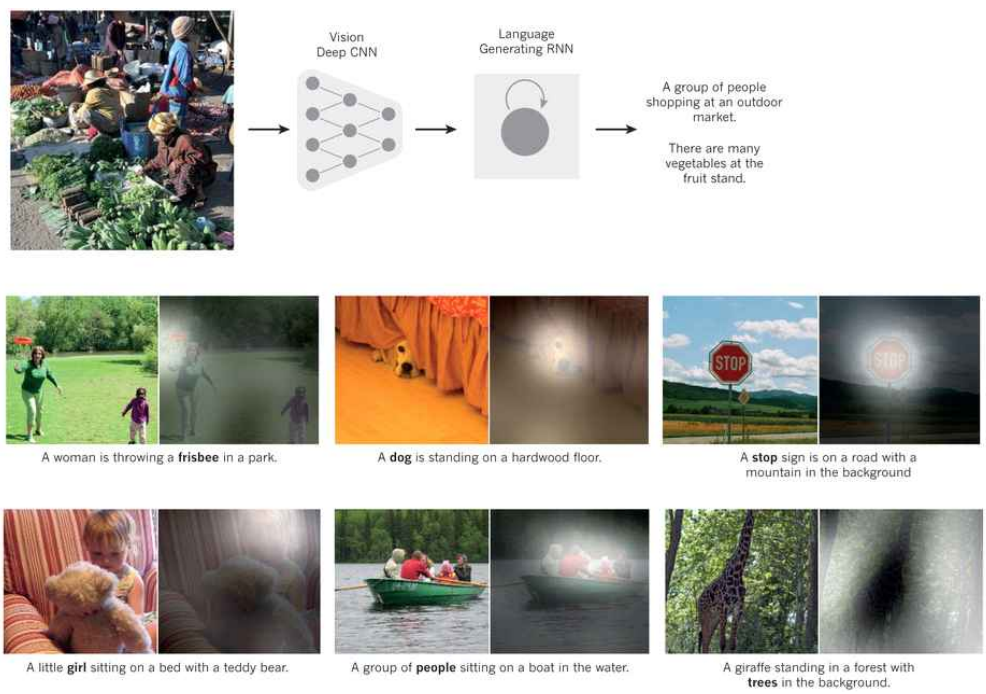


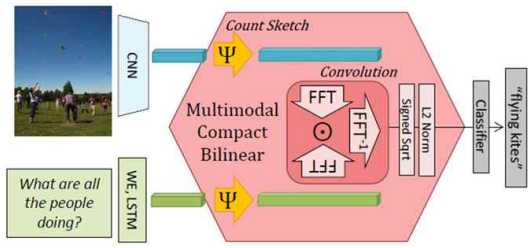
그림 9. 입력 이미지에 대한 순환망에 의해 자동 생성된 캡션〔1,40〕. 이미지를 CNN이 처리한 후, RNN을 통해 캡션 문장을 생성. 입력 이미지에 따라 생성된 문장 내 단어와 연관된 주의(attention)가 달라짐〔41〕.

이미지 생성 문제: 이와는 반대 방향의 접근으로 묘사글이 주어지면 이미지를 생성하는 문제도 생각해볼 수 있다. 글자 정보에 비해 이미지 정보는 매우 복잡하기 때문에 이미지 설명 생성 결과에 비하면 결과물의 수준은 그다지 만족스럽지 않다. 몇 가지 사례를 보면, [48]에서는 문장 표현에 양방향 (bidirectional) LSTM에 주의 모델을 반영한 모델을 사용하여 핵심 키워드를 인코딩하도록 하였다. 이를 바탕으로 이미지 생성을 위해 GAN의 변형과 RNN을 합쳐서 만든 이미지 모델(DRAW[30]의 확장)을 사용하기도 하였다. 또 다른 접근 방식은 이미지 생성 품질이 좋은 것으로 평을 받는 GAN을 기본 모델로 두고, 생성기에 문장정보를 입력받게 하여 학습하는 것이다[49].

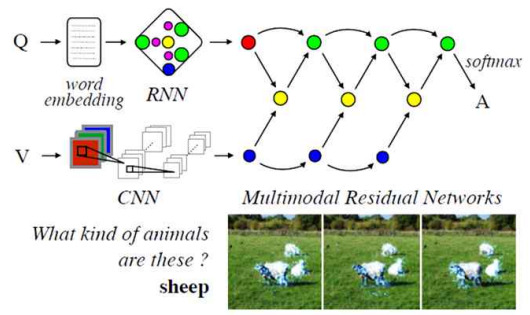
3.3 Visual QA

이 절에서는 지금까지 살펴본 응용 사례에서 한 걸음 더 나아가, 이미지에 대한 QA 데이터 집합을 구축하고, 딥러닝 모델로 이미지에 대한 질의응답을 수행하는 연구를 소개한다. 사용자 입장에서 기계가 이미지를 이해하는지 확인하는 가장 자연스러운 방법은 이미지에 대한 질문에 적절한 답을 하는가일 것이다. 버지니아텍 연구진은 2015년에 Visual QA 데이터 집합을 구축하고 공개하며 Visual QA 챌린지를 개최하였다[37]. 하나의 이미지에 대해서도 사람에 따라 다양한 질문과 답을 할 수 있는 점을 고려할 때, 단순하고 고정적이기보다는 다소 유연한 학습기준이 필요하다. 이 챌린지에서는 그런 이유로 이미지 당 세 개의 질문을 만들고, 질문 하나마다 10 명의 사람에게 답을 하게 하였다. 이 중에 세 명 이상이 공통된 답을 한 것은 모두 맞는 답으로 간주하였다.

QA에서 질의는 문장형식, 응답은 단어 또는 문장 형식을 가지므로, 대부분의 챌린지 참여자들



(a) Multimodal Compact Bilinear with Attention



(b) Multimodal Residual Networks

그림 10 Visual QA를 위한 멀티모달 모델들의 개요 [50, 51]

은 언어 처리에 RNN 계열의 모델을 주로 사용하고, 이미지 처리에는 CNN을 주로 사용하였다. 성능의 차이를 가져온 핵심적인 차이는 질의에서 얻은 인자와 이미지에서 얻은 인자, 답 사이의 연관 관계를 어떻게 구조화 하는가였다. 질의와 이미지 입력에 대해 compact bilinear pooling을 통해 이들 인자 간의 연관관계를 조합적으로 늘려 표현한 팀이 1위를 하였다[50]. 반면 인자 간 연관관계를 질의와 이미지 입력에 대해 요소별 곱으로 표현하고 그림 10과 같이 계층을 뛰어넘는 residual 연결을 추가한 멀티모달 잔차망을 제안한 팀도 좋은 성능을 얻었다[51]. 이 연구는 요소별 곱에서 비롯된 차이 정보를 CNN으로 역전파하는 방식의 새로운 주의 시각화에 대한 제안을 함께 하여 많은 관심을 끌었다. 최근 Visual QA 관련 서베이 논문이 arXiv에 공개되었다[52,53].

4. 비디오 학습과 스토리 QA

대용량 이미지 데이터에 기반한 딥러닝 기술이 성공을 거두면서 비디오 데이터로의 확장에 관심이 옮겨가기 시작했다. 과거에 연구된 비디오 이해 문제는 주로 수동적 특징 추출 방법에 의존한 행동 인식, 사건 검출, 비정규성 검출이었고, 데이터 집합의 크기도 소규모였다. 그러나 최근에는 대용량 비디오 데이터가 공개되면서 이미지의 경우와 유사하게 자동 묘사글 생성, QA, 비디오 분류 등의 보다 복잡한 문제가 다루어지게 되었다. 무엇보다도 비디오는 순서가 있는 데이터이므로 사건의 전후 관계를 다루게 되면서 스토리와 연관된 연구가 진행되고 있다. 4 장에서는 최근 공개되고 있는 대용량 비디오 데이터와 비디오 이해를 추구하는 딥러닝 기술, 비디오를 위한 QA 학습 기술을 소개한다.

4.1 비디오 스토리 학습을 위한 데이터집합

단순히 긴 비디오만 찾는다면 CCTV나 차량 블랙박스로 촬영한 동영상을 활용할 수도 있으나, 효과적인 학습을 위해서는 보조정보가 있는 데이터가 필요하다. 비디오 학습에 주로 다루어지는 데이터 중 하나는 상업용으로 제작된 비디오 컨텐츠에서 추출된 것이다. 시장 확대를 위해 부가 정보가 제공되는 경우가 있기 때문이다. 일례로 묘사글이 함께 제공되는 비디오 데이터인 MPII-MD (Max Planck Institute for Informatics - Movie Description)를 들 수 있다[54]. 여기에는 94 개 영화 속의 68K 개 비디오 클립-묘사글 쌍을 포함되어 있으며, 묘사글은 시각 장애인의 영화 시청을 돕기 위해 기존에 서비스 되고 있던 DVS

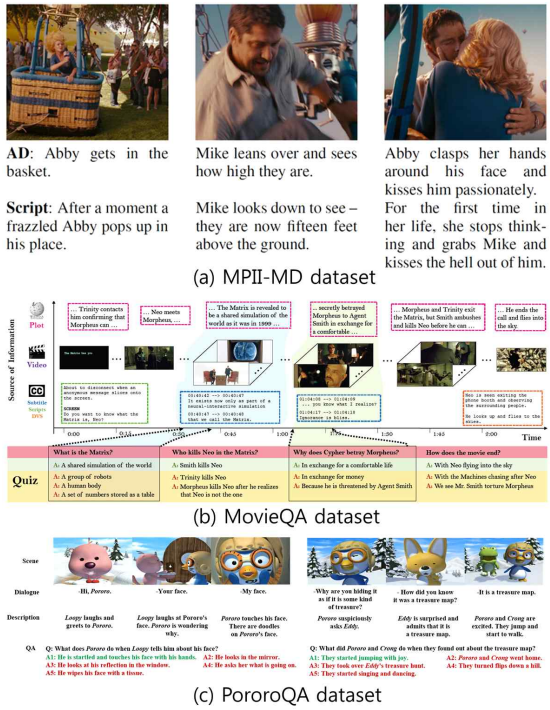


그림 11 상업용 컨텐츠 기반 비디오 스토리 학습용 데이터 집합의 예

(Descriptive Video Service)로부터 제작되었다. 비디오에 QA가 추가된 데이터도 최근 생성되고 있으며, 대표적인 예로 MovieQA[55]와 PororoQA[12]가 있다. MovieQA 데이터집합은 140 개 영화의 비디오 정보(비디오 클립, 위키피디아로부터 추출한 줄거리 정보, 자막, 스크립트, DVS)와 7천여 개의 영화 내용 관련 QA 쌍을 제공한다. PororoQA 데이터집합은 유아용 3D 애니메이션인 '뽀롱뽀롱 뽀로로'의 177 개 에피소드로부터 생성된 16K 개 비디오 클립-자막 쌍, 27K 개 비디오 묘사글과 9K 개 내용 관련 QA 쌍을 포함한다. 2016년 9월말에 공개된 구글의 Youtube-8M은 비디오 분류 문제를 위한 데이터 집합이다. 데이터 집합은 총 500K 시간 길이의 8M 개 유튜브 비디오 4800 개의 표지를 제공한다.

표지의 수가 큰 것은 ImageNet에서와 같이 일반화 된 인자를 생성하는 측면에서 잠재적으로 유리하다. 그림 12에 비디오 이해 기술 연구에 사용할 수 있는 데이터 집합과 통계를 정리하였다.

4.2 이미지 시퀀스 및 비디오를 위한 딥러닝 기술

여행 블로그글 생성 문제: 시간 순서가 있는 이미지 집합과 관련된 글이 있는 경우가 있다. 이러한 데이터는 주로 인터넷에 블로그 형식으로 공개되곤 하는데 이를 이용하여 이미지 스트림에 대한 글을 생성하는 연구가 진행되었다. [10]에서는 여행에서 촬영한 사진이 주어지면 자동으로 여행 블로그 글을 생성하는 문제를 다루었으며, 이를 위한 모델로 CRCN(Coherence Recurrent Convolutional Networks)가 제안되었다. CRCN은 이미지 인자를 위한 CNN, 문장처리를 위한 양방향 RNN 외에 유연한 문장 생성을 위한 coherence 모델이 합쳐져 구성된다. CRCN은 입력된 이미지 시퀀스에 적절한 문장들을 학습 데이터로부터 가져온 뒤 이미지와 문장 관계, 그리고 문장과 문장 관계를 평가하여 가장 적절한 문장들을 출력한다.

일상 묘사글 생성 문제: 내러티브 클립과 같은 라이프로깅 웨어러블 카메라를 이용해 일상을 촬영한 데이터는 좋은 학습데이터가 될 수 있다. [56]에서는 3.2절에서 소개한 이미지용 자동 묘사글 생성 모델을 이용하여 미세조정 학습을 시도하고, 유사 이미지 그룹화, 이미지-묘사글 간의 정렬 기술[42] 등을 종합하여 일상을 정리하는 글 생성을 시도하였다.

비디오 묘사글 생성 문제: Venugopalan 등은 기존 이미지 묘사글 생성 모델을 확장하여 비디오 묘사글 생성 모델을 제안했다[57]. 이러한 방식은 여러 이미지 프레임을 하나의 이미지와 같은 표현으로 바꾸는 과정이 필수적인데, 이를 위해 비디오 클립의 이미지 프레임들을 CNN을 거쳐 인자를 추출하되 이들의 평균값을 계산하여 비디오 벡터를 구했다. 그리고 비디오 벡터를 매 시간단계마다 입력으로 받아 LSTM으로 비디오와 관련한 문장을 생성하였다.

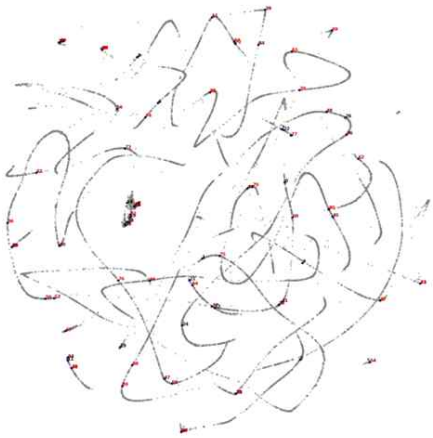
장면에 대한 정보 표현으로 물체, 행동, 장소에 대한 분류기를 쓸 수 있다면 도움이 될 수 있다. [58]에서는 MPII-MD 데이터를 이용하여 이러한 분류기들의 출력을 비디오 인코더로 사용하고 디코더로서 LSTM으로 묘사글 생성을 시도하였다.

비디오 묘사글 검색 문제: 책으로 나왔던 원작을 영화로 다시 제작하는 경우가 있다. 이러한 경우 영화 속 비디오와 책 속의 글이 대응될 수 있으므로 이 관계를 학습하여 임의의 비디오 장면과 글에 대한 상호 생성을 시도할 수 있다. 실제로 [11]에서 MovieBook 데이터 집합을 구축하고 학습하여 영화 일부에 대한 묘사를 위해 적절한 책의 글을 찾아 대신하고, 책의 글과 연관된 영상들을 찾는 예를 보여주었다. 이를 위해 비디오 영화 클립의 이미지 처리에 CNN의 일종인

Dataset	YouCook (Das et al. 2013)	TACos (Regneri et al. 2013)	TACos M.L (Rohrbach et al. 2014)	MSVD (Chen et al. 2011)	MPII-MD (Rohrbach et al. 2015)	M-VAD (Torabi et al. 2015)	MovieQA (Tapaswi et al. 2016)	PororoQA (ours)
# videos	88	127	185	N/A	94	92	140	177
# clips (scenes)	88	7K	14K	2K	68K	49K	7K	16K
# hours	N/A	N/A	N/A	5	74	85	N/A	20.5
# sentences	3K	18K	52K	70K	68K	55K	N/A	59K
# QA	-	-	-	-	-	-	6K	9K
Domain	cooking	cooking	cooking	open	movie	movie	movie	cartoon

그림 12. 비디오 학습을 위한 데이터 집합과 통계[12]

(a) 장면에 대한 묘사글 생성 예



(b) 이벤트 순서에 대해 임베딩하여 도출된 궤적 모양의 이벤트 코드들

그림 13 비디오 스토리 이해를 위해 이벤트 순서정보를 통한 궤적 모양의 임베딩 공간을 도출하려는 연구

GoogLeNet과 hybrid-CNN을 적용하고 DVS 정보와 동일 공간에 임베딩하여 이미지-묘사글 사이의 관계를 학습하였다. 그리고, 영화 속 자막과 책속의 문장/문단과의 관계는 BLEU, TF-IDF, skip-thought 벡터[59]를 통해 유사도를 결정하였다. 이렇게 얻은 이미지-묘사글, 대사-책속의 글 사이의 관계 정보를 바탕으로 시퀀스 정보처리를 위해 CRF(conditional random field)를 사용하고 CNN으로 장면-문단 간 유사도를 함수를 근사하였다.

비디오 내 이벤트 순서 학습 문제: 한편 본 저자 그룹은 다른 각도에서 응용을 시도하고 있다. 추구하는 바는 연속된 비디오 속 이벤트를 궤적으로 바꾸어 표현하여 추론을 쉽게 하고자 하는 것이다. 비디오를 직접 다루기 어렵다면 문장으로 바꾸어 처리하는 것을 고려할 수 있다. 즉, 장면을 묘사글로 변환하고, 대사도 글의 형태이므로 두 정보 모두 skip-thought 벡터를 통해 실수 벡터로 바꿀 수 있다. 이 둘을 연결하여 하나의 벡터로 합쳐(이벤트 벡터로 명명) 입력으로 사용한다. 이를 바탕으로 이벤트 간의 시간 거리가 가까운 것이 임베딩 공간에서 가까워지도록 학습하였다. 그림 13은 PororoQA 데이터[12]를 이용해 이벤트 표현에 사용된 장면에 대한 묘사글의 예와 도출된 궤적 모양의 임베딩 결과를 t-SNE[60]를 이용해 시각화한 것이다.

4.3 비디오 QA

3.3절에서 소개한 Visual QA 문제는 학습 데이터를 생성하는 과정의 주요 과정을 사람이 수행해야만 했다. 만약, 비디오 장면 묘사글에 대해 특정 단어를 지우고 괄호로 대체한 후 이를 맞추는 QA를 고려한다면, QA 데이터를 자동적으로 대량으로 만들 수 있다. 예를 들어, 비디오 클립의 주석 문장이 'Kids are playing basketball'이었다면, Kids are playing ( )이 질의가 되고, basketball이 답이 되며, QA 모델은 네 개의 답안 보기 중에 하나를 선택해야 한다. [61]에서 Zhu et al.은 기존 공개된 비디오 데이터의 주석 문장을 "괄호 채우기" 형식의 QA 문제로 바꾸고, CNN으로 얻은 비디오 클립 벡터와 정답 답안이 포함된 주석 문장을 동일 임베딩 공간 상에서 가까워지도록 학습시켰다. 이 방식으로 비디오를 바탕으로 과거, 현재, 미래에 대한 QA 문제를 다루었다.

4.1에서 소개한 MovieQA 데이터에 대해 [55]

에서는 video2text 모델과 메모리망을 사용하여 기준 결과를 제시하였다. 비디오의 이미지는 묘사글로 바꾸고, 자막과 함께 두 개의 문장 시퀀스로 바꾸었다. 그리고 각 시퀀스와 QA 데이터를 사용하여 두 개의 서로 다른 메모리 네트워크를 학습시킨 뒤 QA 결과를 앙상블 조합하였다. 향후 이 데이터를 통해 챌린지가 열릴 예정이다.

4.1에서 소개했던 PororoQA 데이터에 대한 학습 방법으로서, video2text 모델과 메모리망, 주의 모델을 함께 사용한 deep embedded 메모리망을 제안했다[12]. video2text를 이용하여 이미지에 대한 묘사글을 만들고 여기에 해당 이미지에서 사용된 대사를 합쳐 메모리망을 위한 주어진 글의 일부로 사용하였다. [55]에서 이미지 스토리 모델과 텍스트 스토리 모델의 앙상블 조합의 성능이 저하될 수 있었다는 점을 고려하여, 비디오를 번역한 문장과 기존 자막을 합쳐 하나의 융합된 스토리 문장 시퀀스를 만들고, 메모리망을 학습시킨 것이다. 그림 14는 deep embedded 메모리망의 모습을, 그림 15는 deep embedded 메모리망이 뽀로로 QA를 푼 예시를 보여준다.

5. 뽀로로봇(Pororobot)

Pororobot은 비디오 스토리 QA기술을 로봇 플랫폼에 설치하여 음성인식, 음성합성 기술을 통해 사용자와 상호작용을 할 수 있도록 제작한 로봇이다. PororoQA 데이터를 통해 학습한

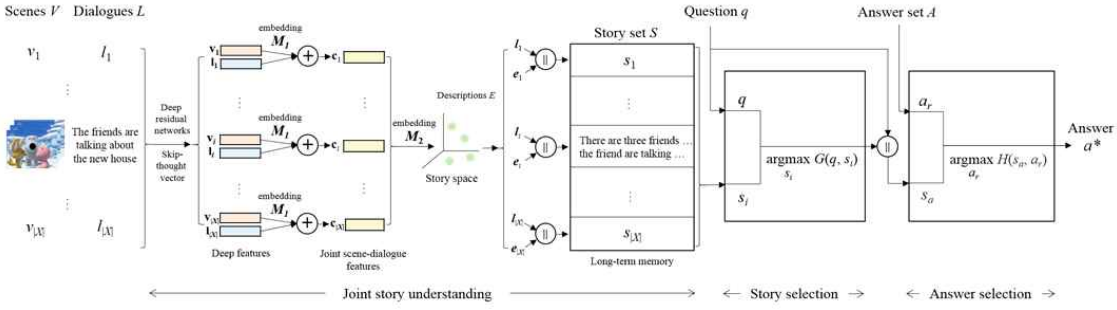


그림 14. Pororo QA 데이터를 위한 Deep Embedded 메모리망의 구조도[12]

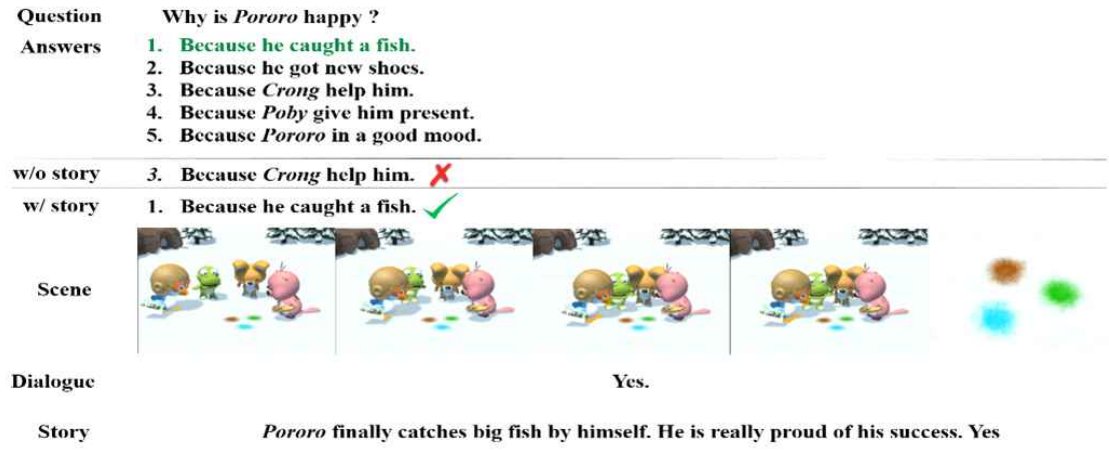


그림 15. Deep Embedded 메모리망이 Pororo QA를 푼 예시[12]

deep embedded 메모리망이 설치되어 있다. 이를 통해 사용자와 뽀로로 비디오를 보면서 함께 질의응답을 수행한다. 이는 교육용 시나리오로 쉽게 적용할 수 있다. 예를 들어, 어린이와 로봇은 함께 교육용 만화 비디오를 시청한다. 로봇은 시청한 비디오를 학습하여 이에 관련된 질문을 생성하고 어린이에게 질문한다. 로봇의 질문에 어린이가 대답하였을 때, 어린이의 대답이 올바르다면 로봇은 대답에 동의하고 다음 질문을 묻는다. 반대로 어린이의 대답이 틀렸다면, 로봇은 대답에 동의하지 않고 어린이에게 정답을 알려준다. 로봇은 어린이의 응답의 정답여부를 판별하여 올바른 피드백을 줄 수 있다.

그림 16은 시행 시나리오의 예시이다.

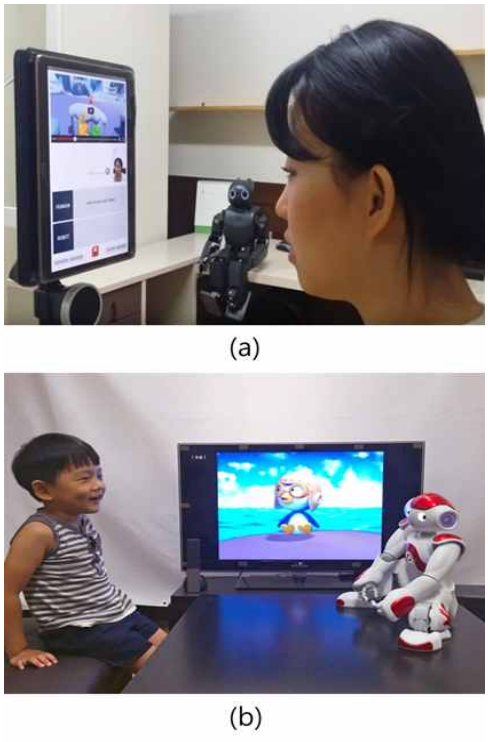


그림 16. 음성인식과 음성합성 기술을 통해 질의응답 상호작용을 하는 뽀로로봇 에이전트. (a) 태블릿 타입, (b) 로봇 타입

6. 결 론

본 고에서는 급격히 발전하고 있는 딥러닝 기술을 바탕으로 한 이미지 및 비디오 스토리 학습 기술을 살펴하였다. 이를 위해 영상 매체를 위한 딥러닝 기술과 언어 처리 및 질의응답을 위해 성공적으로 적용된 다양한 최신 딥러닝 기술을 살펴보았다. 이미지 및 비디오의 스토리 학습이 가능하도록 학습 데이터에 부가되는 보조데이터가 점차 풍부해지면서 다양한 응용 문제의 해법과 그 파급효과에 대한 기대도 커지고 있다. 실세계에는 로봇과 같은 사용자를 위한 인터페이스가 필수적인데 생산자와 소비자 모두 큰 관심을 가지고 이 분야를 바라보고 있다.

참 고 문 헌

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature vol. 521, pp. 436-444, 2015.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," In Advances in neural information processing systems (NIPS), 2012.

[3] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp. 580-587, 2014.

[4] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," In Advances in neural information processing systems (NIPS), pp. 487-495, 2014.

[5] F. Schroff, D. Kalenichenko, and J. Philbin. "Facenet: A unified embedding for face recognition and clustering." Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815-823, 2015.

[6] Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1701-1708).

[7] A. Toshev, and C. Szegedy. "Deeppose: Human pose estimation via deep neural networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1653-1660, 2014.

[8] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," In Advances in neural information processing systems (NIPS), pp. 1799-1807, 2014.

[9] S.-W. Lee, C.-Y. Lee, D. Kwak, J. Kim, J. Kim, and B.-T. Zhang, "Dual-memory deep learning architectures for lifelong learning of everyday human behaviors," International Joint Conference on Artificial Intelligence (IJCAI 2016), pp. 1669-1675, 2016.

[10] C. Park, and G. Kim, "Expressing an Image Stream with a Sequence of Natural Sentences," In Advances in neural information processing systems (NIPS), 2015.

[11] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 19-27, 2015.

[12] K.-M. Kim, C.-J. Nan, M.-O. Heo, S.-H. Choi, B.-T. Zhang, "DeepStory: video story qa by deep embedded memory networks," AAAI 2017 (submitted)

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol.86(11), 2278-2324, 1998.

[14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," J. Machine Learning Res. vol.15, pp. 1929 - 1958, 2014.

[15] C. Szegedy, W. Liu, W., Y. Jia, P. Sermanet, S. Reed, D. Anguelov, and A. Rabinovich, "Going deeper with convolutions," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1-9, 2015.

[16] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[17] Y. Bengio, A. Courville, and P. Vincent. "Representation learning: A review and new perspectives." IEEE transactions on pattern analysis and machine intelligence, vol.35.8 , pp. 1798-1828, 2013.

[18] W. W. Zhu, A. Berndsen, E. C. Madsen, M. Tan, I. H. Stairs, A. Brazier, P. Lazarus, R. Lynch, P. Scholz, K. Stovall, et al. "Searching for pulsars using image pattern recognition," The Astrophysical Journal, vol.781(2):117, 2014.

[19] G. Hinton, and R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." Science 313.5786, pp. 504-507, 2006.

[20] R. Salakhutdinov, and G. Hinton, "Deep Boltzmann machines," In Proc. International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 448 - 455, 2009.

[21] D. P. Kingma, M. Welling, "Auto-Encoding Variational Bayes," International Conference on Learning Representations (ICLR), 2014.

[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, and Y. Bengio,

"Generative adversarial nets," In Advances in Neural Information Processing Systems (NIPS), pp. 2672-2680, 2014.

[23] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," In Proc. Advances in Neural Information Processing Systems (NIPS), pp. 3111 - 3119, 2013.

[24] S. Hochreiter, and J. Schmidhuber, "Long short-term memory," Neural Comput. vol. 9, pp. 1735 - 1780, 1997.

[25] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.

[26] J. Chung, C. Gulcehre, K.H. Cho, and Y. Bengio, Empirical Evalutation of Gated Recurrent Neural Networks on Sequence Modeling, arXiv:1412.3555, 2014.

[27] W. Zhang and M. Lapata, "Chinese Poetry Generation with Recurrent Neural Networks," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.

[28] A. Karpathy, "The Unreasonalbe Effectiveness of Recurrent Neural Networks," http://karpathy.github.io/2015/05/21/rnn-effectiveness/

[29] http://benjamin.wtf

[30] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, D. Wierstra, "DRAW: a recurrent neural network for image generation," International Conference on Machine Learning (ICML), 2015.

[31] A. van den Oord, N. Kalchbrenner, K. Kavukcuoglu, "Pixel recurrent neural networks," International Conference on Machine Learning (ICML), 2016.

[32] J. Weston, S. Chopra, and A. Bordes, "Memory networks," International Conference on Learning Representation (ICLR), 2015.

[33] S. Sukhbaatar, J. Weston, and R. Fergus. "End-to-end memory networks." Advances in neural information processing systems (NIPS), 2015.

[34] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, and R. Socher, "Ask Me Anything: Dynamic Memory Networks for Natural Language Processing," International Conference on Machine Learning (ICML), 2016.

[35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imageneet: A large-scale hierarchical image database," In Computer Vision and Pattern Recognition (CVPR), pp. 248-255, 2009.

[36] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft coco: Common objects in context," In Computer Vision-ECCV 2014, pp. 740-755, 2014.

[37] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick and D. Parikh, "VQA: Visual Question Answering," In International Conference on Computer Vision (ICCV), 2015.

[38] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei, "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations," https://arxiv.org/abs/1602.07332, 2016.

[39] J. Markoff, "A Learning Advance in Artificial Intelligence Rivals Human Abilities". The New York Times, 2015-12-10.

[40] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," In Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3156-3164, 2015.

[41] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention ," International Conference on Machine Learning (ICML), 2015.

[42] A. Karpathy, and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[43] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig, "From Captions to Visual Concepts and Back," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[44] X. Chen, and C. L. Zitnick, "Mind's Eye: A Recurrent Visual Representation for Image Caption Generation," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015

[45] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "zero-shot learning through cross-modal transfer," In Advances in neural information processing systems (NIPS), pp. 935-943, 2013.

[46] R. Kiros, R. Salakhutdinov, and R. Zemel,."Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models," Transactions of the Association for Computational Linguistics, (To appear).

[47] L. Ba, K. Swersky, and S. Fidler. "Predicting deep zero-shot convolutional neural networks using textual descriptions," Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015.

[48] E. Mansimov, E. Parisotto, J. Ba, and R. Salakhutdinov, "Generating Images from Captions with Attention," International Conference on Learning Representation (ICLR), 2016.

[49] S. Reed, Z. Akata, X. Yan, L. Logeswaran, Bernt Schiele, and H. Lee, "Generative Adversarial Text to Image Synthesis," International Conference on Machine Learning (ICML), 2016.

[50] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, M. Rohrbach, "Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding," EMNLP 2016 (accepted).

[51] J.-H. Kim, S.-W. Lee, D.-H. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, B.-T. Zhang, "Multimodal Residual Learning for Visual QA, " Advances in neural information processing systems (NIPS) 2016 (accepted).

[52] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel, "Visual Question Answering: A Survey of Methods and Datasets," arXiv:1607.05910, 2016.

[53] K. Kafle, and C. Kanan, "Visual Question Answering: Datasets, Algorithms, and Future Challenges", arXiv:1610.01465, 2016.

[54] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, "A dataset for movie description," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[55] M. Tapaswi, Y Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, "Movieqa: Understanding stories in movies through question-answering," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[56] C. Fan, and D. J. Crandall, "DeepDiary: Automatic Caption Generation for Lifelogging Image Streams," arXiv:1608.03819v1, 2016.

[57] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, " Translating videos to natural language using deep recurrent neural networks," the 2015

Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT), 2015.

[58] A. Rohrbach, M. Rohrbach, and B. Schiele, "The long-short story of movie description," German Conference on Pattern Recognition, Springer International Publishing, 2015.

[59] R. Kiros, Y. Zhu, R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," In Advances in neural information processing systems (NIPS), pp. 3294-3302, 2015.

[60] L.J.P. van der Maaten and G.E. Hinton. "Visualizing High-Dimensional Data Using t-SNE," Journal of Machine Learning Research 9(Nov):2579-2605, 2008.

[61] L. Zhu, Z. Xu, Y. Yang, and A. Hauptmann, "Uncovering Temporal Context for Video Question and Answering," arXiv preprint arXiv:1511.04670, 2015.

허 민 오

- 2001~2004 ㈜인티 연구원
- 2005 고려대 전기전자전파공학부 학사
- 2005~현재 서울대 컴퓨터공학부 석박사 통합과정
- 2010.07~08 뮌헨공과대학 방문 연구원
- 관심분야: 인지기계학습, 딥러닝, 계산학적 비디오 내러티브 지능, 베이지안 모델링 방법론

김 경 민

- 2013 홍익대 컴퓨터공학부 학사
- 2013~현재 서울대 컴퓨터공학부 석박사 통합과정
- 관심분야: 멀티미디어 마이닝, 딥러닝, 인지과학, 기계학습

장 병 탁

- 1986 서울대 컴퓨터공학과 학사
- 1988 서울대 컴퓨터공학과 석사
- 1992 독일 Bonn 대학교 컴퓨터과학 박사
- 1992~1995 독일국립정보기술연구소 (GMD, 현 Fraunhofer Institutes) 연구원
- 1997~현재 서울대 컴퓨터공학부 교수 및 인지과학, 뇌과학, 생물정보학 협동과정 겸임교수
- 2003~2004 MIT 인공지능연구소(CSAIL) 및 뇌인지과학과(BCS) 객원교수
- 2007~2008 삼성종합기술연구원(SAIT) 객원교수
- 현 재 서울대 인지과학연구소 소장
- Applied Intelligence, BioSystems, Journal of Cognitive Science 등 국제저널 편집위원
- 관심분야: 바이오지능, 인지기계학습, 분자진화 컴퓨팅 기반 뇌인지 정보처리 모델링