TV 드라마 비디오 스토리 분석 딥러닝 기술

Deep Learning Technologies for Analysis of TV Drama Video Stories

□ 남장군, 김진화, 김병희, 장병탁 / 서울대학교

요약

비디오 정보를 자동으로 학습하고 관련 문제를 해결하기 위해서는, 비디오의 기본 구성요소인 영상, 음성, 언어 정보의 학습을 기반으로 고차원의 추상적 개념을 파악하는 기술이 필수적이다. 최근 딥러닝이 실용적인 수준으로 이러한 기술을 가능하게 함에 따라, 보다 도전적인 비디오 스토리 분석과 이해 문제 해결을 시도할 수 있게 되었다. 본 고에서는 비디오의 요소별 분석에 적용 가능한 최신 딥러닝 기술을 소개하고, 딥러닝 기술을 핵심으로 한 TV 드라마의 스토리 분석 사례를 살펴본다.

Ⅰ. 서 론

최근 딥러닝 기술이 크게 발전하면서 인공지능의 대표적 목표인 음성지능, 시각지능, 언어지능의 구현이 실용적인 단계로 올라섰다. 무엇보다 각 분야별 기술이 개별적으로 연구되던 기존 트렌드가 딥러닝이라는 공통의 기술 하에서 다중 지능 구현으로 융합되는 획기적인 변화가 이어지고 있다. 이러한 변화에 힘입어 비디오 스토리를 이해하는 수준의 지능 구현 연구를 본격적으로 시작할 수 있게 되었다.

본 고에서는 이러한 변화를 견인한 대표적인 딥러닝 기술을 정리하고, TV 드라마 비디오에서의 스토리 학습 사례를 소개한다.

이후 구성은 다음과 같다. Ⅱ장과 Ⅲ장에서는 비디오의 요소 분석과 스토리 학습 문제 및 관련 딥러닝 기술을 정리한다. Ⅳ장에서는 여러 딥러닝 기술을 기반으로 TV 드라마 비디오에서 스토리를 분석한 응용사례를 소개한다. 마지막으로 Ⅴ장에서는 결론을 맺는다.

Ⅱ. 비디오 정보의 추출 및 분석

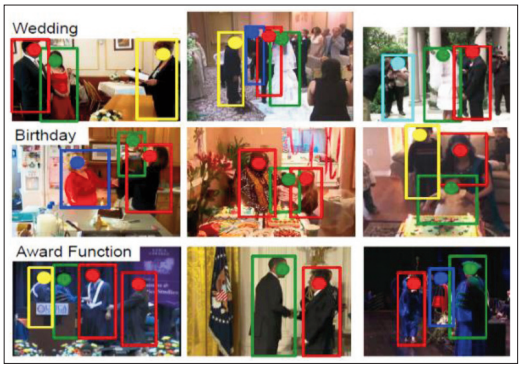
비디오를 자동으로 분석하는 문제는 인공지능에

서 오랜 기간 다룬 문제이며, 영상의 시각정보 추출부터 스토리 구성에 따른 비디오 분류 문제까지 다양한 세부 문제를 해결해야 한다. 최근 딥러닝 기술을 이용한 컴퓨터 영상 분석 기술에 힘입어 비디오 분류, 이벤트 인식, 비디오 자동 주석 등의 분야에서 큰 진전이 이어지고 있다. 몇 가지 사례를 살펴보자면, [1]에서는 다중 해상도 컨볼루션 신경망(Multiresolution CNN) 구조를 제안하여 대규모 비디오 데이터의 자동 주석 및 분류문제에 적용하였다. [2]는 다중모달 딥러닝을 사용하여 비디오의 시각과 음성정보의 매핑 성능을 보여주었다. [3]에서는 시공간 특성 간섭(Spatio-temporal feature coherence)을 통해 대용량 비디오에서 자주 나타나는 개념들을 구별하였다.

이와 같은 사례는 대용량 비디오 데이터의 자동 주석과 분류, 특정 이벤트 인식 등 구체적이지만 제한된 정보를 분석하는데 공통점이 있다. 그러나 실제 비디오의 스토리 자동 분석을 위해서는 보다 추상적인 수준의 내용 인식과 흐름에 대한 모델이 필요하다. 관련 사례로서, [4]에서는 TV 드라마의 등장인물 중심의 정보를 학습하여 등장인물 간 소셜 네트워크를 분석하였다. [5]는 Deep Embedded 메모리망을 이용하여 유아 애니메이션 뽀로로의 내용 관련 질문에 자동으로 응답하는 인공지능 시스템을 선보였다.

Ⅲ. 딥러닝 기반 비디오 스토리 분석 연구

비디오 스토리를 이해하는 딥러닝 모델을 만들기 위해서는 비디오를 구성하는 다양한 구성 요소가 종합된 대규모 데이터가 필요하다. Ⅲ장에서는 최근에 공개된 여러 대용량 비디오 데이터셋을 소개하고, 비디오 구성 요소 중 영상과 언어 학습에 필수적인 대표적인 딥러닝 기술을 소개한다.



〈그림 1〉 비디오 소셜 이벤트 검출 예시

1. 비디오 스토리 학습 데이터셋

이미지 인식을 비롯한 다양한 컴퓨터 비전 문제의 해법이 크게 개선된 가장 큰 계기 중의 하나가 바로 ImageNet(2009년)과 같은 대용량 공개 데이터셋의 출시이다. 비디오는 이미지에 비해 데이터의 복잡도가 훨씬 높기 때문에, 특정 문제 해결에 특화된 벤치마크용 데이터셋 위주로 공개되고 있으며, ImageNet 데이터셋 정도의 대규모 비디오 데이터셋은 많지 않다.

대표적인 데이터셋으로 독일의 MPI에서 공개한 MPII-MD가 있다. MPII-MD 데이터셋은 94개 영화의 68K개 비디오 클립-묘사글의 쌍을 포함하였고 LSMDC2016(Large Scale Movie Description and Understanding Challenge) 대회에서 비디오 묘사글 생성 연구의 벤치마크 데이터셋으로 사용되었다. 비디오 분류 문제를 풀기 위해 최근 구글에서는 ImageNet에 대응되는 비디오 분류 데이터셋 YouTube-8M를 공개하였다. YouTube-8M는

(a) PororoQA 데이터셋 예시 (b) MPII-MD 데이터셋 예시

(c) Movie QA 데이터셋 예시 (d) YouTube 8M 데이터셋 예시

〈그림 2〉 비디오 스토리 학습 데이터셋

8백만 개(총 5백만 분 분량)의 비디오 URL과 비디오 단위의 표지(Video-level labels) 데이터가 포함된다. 이는 기존의 스포츠 동영상 분류를 위한 Sports-1M 데이터셋보다도 큰 규모의 데이터셋이다. 또한 비디오에 관한 질의응답 문제를 풀기 위해 구축한 MovieQA과 PororoQA 데이터셋이 있다. MovieQA는 140개 영화의 비디오 클립과 영화 소개, 그리고 자막, 묘사글이 포함되고 영화 스토리에 관한 질의응답 데이터가 약 7천 개 포함되어 있다. PororoQA는 유아용 애니메이션 'Pororo'의 177개 에피소드에서 추출한 16K개의 비디오 클립과 자막, 27K개 비디오 묘사글과 9K개의 스토리 질의응답 데이터가 포함된다. 〈그림 2〉에서 각 데이터셋의 예시를 볼 수 있다.

2. 비디오 분석을 위한 딥러닝 기술

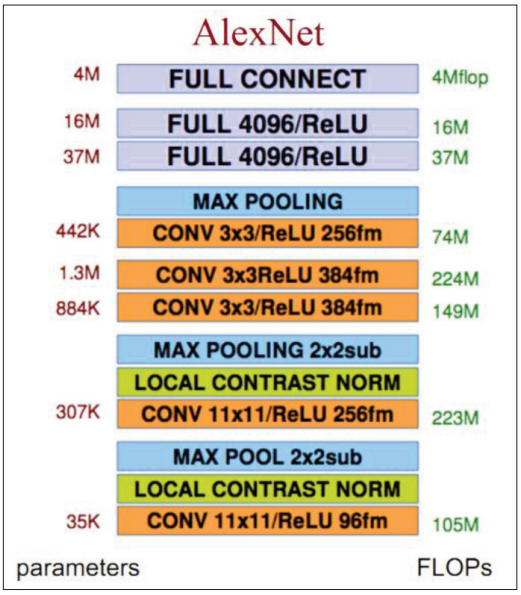
이 절에서는 비디오 영상과 언어 분석을 위한 대표적인 딥러닝 기술을 소개한다.

1) 영상처리 딥러닝 기술

이미지 분류 문제: 전통적인 영상처리에서는 SIFT, HOG와 같은 특징점 추출 방법을 사용하여 문제에 접근하였다. 이러한 방법은 전문가의 지식이 필요하고 각 문제마다 수작업으로 디자인을 해야하는 문제점이 있다. 반면 딥러닝 기술은 데이터에서 분류에 필요한 특징점을 자동으로 학습한다. 영상처리에서 대표적 딥러닝 기술인 컨볼루션 신경망(Convolutional neural network, CNN)은 깊은 층의 네트워크를 통해 다양한 단계의 특징점 조합을 학습하여 성능을 크게 개선하였다.

대표적인 CNN 모델로 딥러닝 기반 영상처리를 촉발한 AlexNet[6]이 있다. AlexNet(〈그림 3〉)을 통해 CNN의 대표적 요소 기술을 살펴본다.

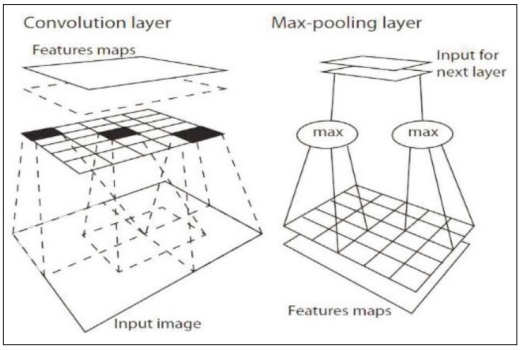
컨볼루션(Convolution): 이 층에서는 입력 이미지에 학습 가능한 필터를 적용한 컨볼루션 연산을 수행한다(식 (1)). 각 필터별로 2차원 이미지를 훑은 결과로 2차원 활성화맵을 출력한다.



〈그림 3〉 AlexNet 모델의 구조도

$$y = \sum_{k=1}^{K} h_k^{n-1} * h_{kj}^n \quad (1)$$

풀링(Pooling): 풀링 과정은 지정영역의 대표 값을 계산하는 과정을 통해 모델의 복잡도를 줄이고 영상의 정보를 추상화한다.



〈그림 4〉 컨볼루션과 최대 풀링 연산

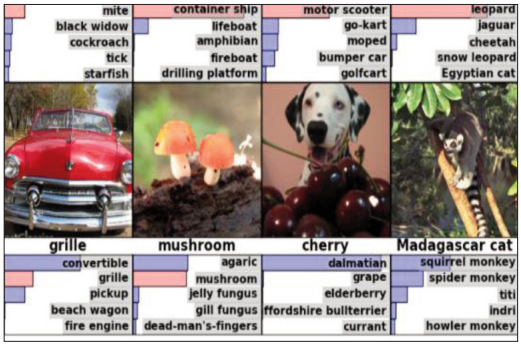
활성화와 표준화(Activation and normalization): 층사이의 값의 전달 과정에 활성화 함수 ReLU (식 (2))를 사용하였고 표준화 함수(식 (3))는 인접된 특징값 사이의 영향을 최소화한다.

$$F(x) = \max(0, x) \quad (2)$$

$$b^i{}_{x,y} = a^i{}_{x,y} / (k + \alpha \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} (a^j{}_{x,y})^2)^\beta \quad (3)$$

완전 연결층(Fully connected layer): 완전연결층은 일반 인공신경망 구조와 같으며 모델의 출력단에 연결되어 softmax를 통해 레이블별 예측 확률을 출력한다.

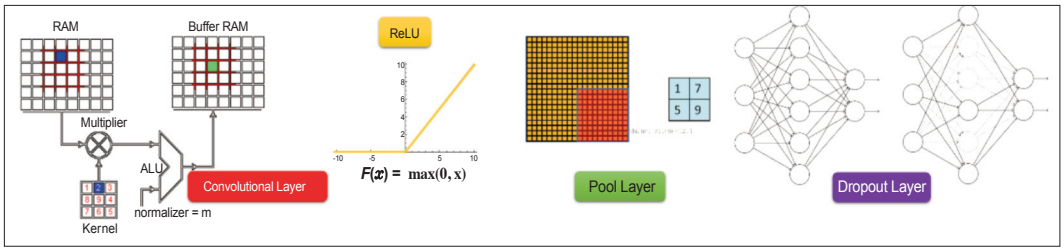
Dropout: 학습 단계에서 일부 은닉 노드를 확률적으로 제외하여, 노드에 연결된 부분의 학습을 일시적으로 중단시킨다. 과적합을 방지하고 성능 향상에 기여한다.



〈그림 5〉 ImageNet 이미지 분류 결과 예시

AlexNet은 2012년에 ImageNet 이미지 분류대회에서 압도적 성능으로 우승하였다. 이후 대회에서는 CNN의 다양한 변형 모델이 발표되어 성능을 향상시켰다. 〈표 1〉에 AlexNet, VGG-Net[7], GoogLeNet[8]과 ResNet[9] 모델을 정리하였다.

최근 이미지 분류와 인식 문제에서 최고의 성능



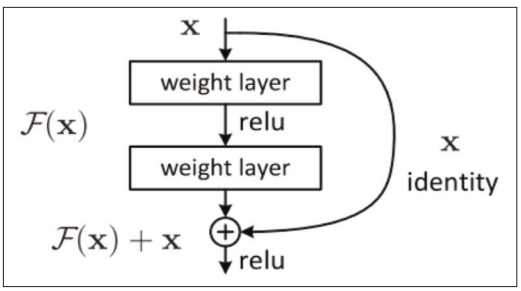
〈그림 6〉 AlexNet의 구성 요소(컨볼루션 , 활성화 함수 , 풀링, Dropout의 구조도)

〈표 1〉 대표적인 CNN모델 및 ImageNet 분류 성능 비교

Model	year	layer	Top-5 error	DA	Conv. layer	Kernel size	FC layer	FC layer size	Dropout	LRN
AlexNet	2012	8	16.4%	+	5	11, 5, 3	3	4096, 4096, 1000	+	+
VGGNet	2014	19	7.3%	+	16	3	3	4096, 4096, 1000	+	-
GoogLeNet	2014	22	6.7%	+	12	7,1,3,5	1	1000	+	+
ResNet	2015	152	3.57%	+	151	7,1,3,5	1	1000	+	-

DA: Data Augmentation; FC: Full Connection; LRN: Local Response Normalization;

을 보이는 딥러닝 모델은 ResNet(deep residual network, 딥 잔차망)이다. ResNet은 신경망의 층이 층의 입력값 x를 중심으로 항등함수와(지름길 연결로 구현) 비선형 학습이 필요한 잔차 F(x)를 구분하여(여러 신경망 층으로 구현) x+F(x) 형태의 매핑을 학습하도록 하였다. 〈그림 7〉은 완전연결층을 중심으로 구성한 ResNet 층의 개념도이며, 이미지 처리 문제의 경우 컨볼루션층을 적용한다. 그 결과, 기존 모델에서 층을 깊게 쌓을 때 성능이 하락하는 문제를 해결하고, 필요에 따라 충분히 깊은 층을 가진 모델을 학습하여 성능을 향상하는 것이 가능하게 되었다. 〈표 2〉는 ResNet을 이용하여 ImageNet 데이터셋에서 이미지 분류 문제를 푼 실험 결과이다. ResNet은 2015년의 이미지넷과 Microsoft COCO 대회의 탐지와 분리 분야에서 우승하였다.

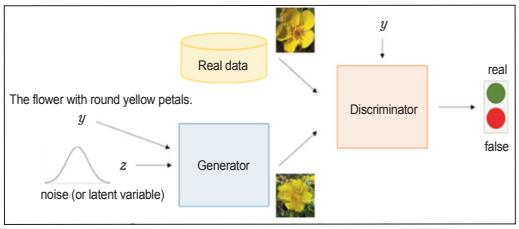


〈그림 7〉 ResNet의 잔차 학습 단위 도식화

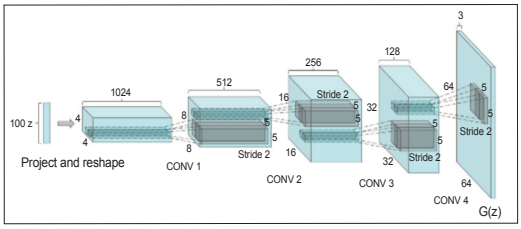
〈표 2〉 2015년 ImageNet 분류 실험 결과

Method	Top-5 err. (test)
VGG(ILSVRC'14)	7.32
GoogLeNet(ILSVRC'14)	6.66
VGG(v5)	6.8
PReLu-net	4.94
BN-inception	4.82
ResNet(ILSVRC'15)	**3.57**

이미지 생성 문제: 이미지 분류 문제 외에 CNN 구조의 또다른 성공적 응용 분야는 이미지 생성이다. 최근 이미지 생성에서 각광을 받고 있는 대표적인 모델은 GAN(Generative Adversarial Network,



〈그림 8〉 GAN 모델의 모식도



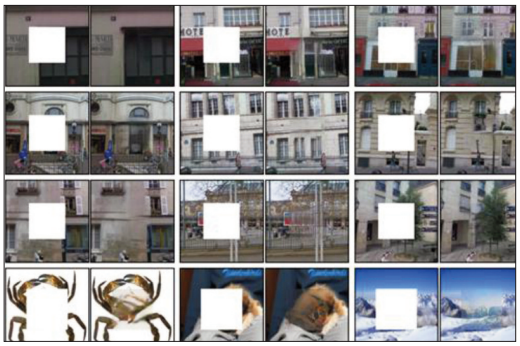
〈그림 9〉 DCGAN의 생성 모듈 구조도

생성 대립넷)[10]이다. GAN은 생성 모듈과 분류 모듈이 학습 과정에서 서로 적대적으로 대결을 한 결과 생성 모듈의 성능을 극대화한다. 생성 모듈은 사전 분포로부터 임의로 표집된 z로부터 데이터 $x=G(z)$를 생성한다. 분류 모듈은 생성 모듈이 생성한 데이터와 실제 데이터를 구분하려 한다. 반대로 생성 모듈은 분류 모듈을 속일 수 있는 실제와 같은 데이터를 생성하려 한다.

실제 데이터의 확률 분포를 P_{data}라 하고 생성 모듈이 학습한 확률 분포를 P_{model}이라고 할 때, 생성 대립넷은 학습 과정에서 최적의 $v(G,D)$를 갖는 G와 D를 찾는다(수식 (4)).

$$v(G,D) = \min_{G}\max_{D} E_{x\sim P_{data}}[logD(x)] + E_{x\sim P_{model}}[\log(1-D(x))] \quad (4)$$

이미지 생성 모델로서의 GAN의 뛰어난 가능성은 곧 다양한 후속 모델 개발로 이어진다. 대표적 사례로, 강력한 영상처리 성능을 보이는 CNN모델을 결합한 DCGAN(Deep Convolutional GAN)[11]은 실제 사진 수준의 이미지 생성도 가능하다. 〈그림 9〉와 같이 DCGAN모델은 네 개 이상의 컨볼루션 층으로 분류 모듈을 구성하고, 비슷한 수의 디컨볼루션 층으로 생성 모듈을 구성하였다. 디컨볼루션은 컨볼루션의 흐름이 거울에 반사된 것과 같은 정반대 형태이고 이에 따라 필터의 크기가 역으로 커진다. 이 모델에 사용되는 컨볼루션은 필터 추출 간격을 2로 하고 풀링을 생략하여 과대적합화 문제를 완화하였다. 배치 정규화(batch normalization)를 적용하여 학습의 속도를 높이고 아담(adam) 최적화 기법이 응용되었다. DCGAN을 비롯한 이미지 생성 모델이 발전함에 따라, 관련한 다양한 문제에도 적용되었다. 이미지의 가려진 영역을 재생하는 사례를 〈그림 10〉에서 볼 수 있다.

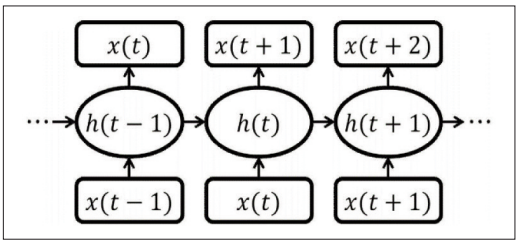


〈그림 10〉 DCGAN으로 가려진 공간을 채우는 예

2) 언어처리 딥러닝 기술

비디오의 연속적 이미지뿐만 아니라 음성, 자막 등과 같은 순서 정보를 학습하기 위해서는 다른 방식의 모델이 필요하다.

순환신경망(Recurrent Neural Network,



〈그림 11〉 순환신경망 구조도

RNN)[12]은 순서를 고려한 동적 신경망 모델로서 특히 언어처리 응용분야에 획기적인 발전을 가져왔다. RNN의 가장 큰 특징은 과거에 대한 기억을 가지고 있다는 것이다. 과거의 입력된 단어의 순서를 고려하여 새로운 단어를 예측하는 언어 모델링에서는 RNN이 기존의 모델을 대체하였다. 〈그림 11〉은 RNN의 구조도이다.

$$h(t)=\sigma\left(W_{in}x(t)+W_r h(t-1)+b_h\right),\\ y(t)=\phi\left(W_{out}h(t)+b_y\right) \quad (5)$$

RNN의 학습 과정에서도 다른 신경망 모델과 마찬가지로 일부 데이터를 기준으로 오차를 줄이는 방향으로 연결 가중치를 조절하는 SGD(Stochastic Gradient Descent) 및 오류 역전파 방법을 사용하며, 특히 시간축 방향에도 동일한 방식을 적용하는 BPTT(Back Propagation Through Time) 알고리즘이 사용된다.

기본 RNN은 긴 시간 간격 간의 연관성을 학습하는 과정에서 안정적인 학습이 어려운 문제가 있으며, 이를 해결한 확장 모델로서 LSTM(Long Short-Term Memory)[13]과 GRU(Gated Recurrent Unit)[14]가 기본 구성 요소로 많이 사용된다. 이들은 모델에 입출력과 기억 정보를 선별적으로 조절하는 게이트(gate)를 두어 RNN에 비해 긴 문장의 생성과 학습의 성능을 높였다.

RNN을 기반으로 한 다양한 딥러닝 모델이 문장 생성, 기계번역 등과 같은 대표적 언어처리 문제에서 획기적 성능 향상 결과를 보이고 있으며, 영화 시나리오 작성[15], 이미지 묘사글 생성[16] 등에도 적용되고 있다.

Ⅳ. 딥하이퍼넷 기반 TV 드라마 분석

이 장에서는 TV 드라마 비디오 스토리 분석의 직접적 사례로서 딥하이퍼넷을 이용하여 TV 드라마로부터 인물관계를 자동으로 분석하는 연구를 소개한다. 딥하이퍼넷은 계층구조를 통해 데이터로부터 자동으로 지식을 학습한다. 기존의 고정된 신경망 모델의 구조와는 달리 구조는 유동적으로 변할 수 있어 동적인 정보를 다루기에 적합하다.

1. 딥하이퍼넷

이 절에서는 딥하이퍼넷의 기술적인 부분을 살펴본다. 〈그림 12〉는 딥하이퍼넷의 구조도이다[17]. 모델 자체는 다층 구조로 구성되었고 이미지-자막 쌍을 구성하여 Monte Carlo Sampling 방법을 통해 H층의 하이퍼에지를 구성한다. C^1층 노드는 H층 하이퍼에지의 부분 집합을 클러스터링한 조합이고 노드의 갯수는 학습에 따라 변하게 된다.

$$Sim(\mathbf{h}^m) = Dist(\mathbf{h}^m)/|\mathbf{h}^m| \quad (6)$$

h^m는 C^1층의 m번째 노드에 연결된 하이퍼에지(hyperedge)의 집합이고 함수 $Dist$는 에지 사이들의 유클리드 거리이다. $Sim(\mathbf{h}^m)$가 임계 값을 넘을 때 노드는 두개로 갈라지게 된다. 그중 임계 값

은 *Sim*들의 평균과 표준편차에 의해 정한다. C^2 층의 노드는 등장인물에 대응되며 C^1층과의 연결은 등장인물들이 나타나는 장면(Scene)에 의해 결정된다.

본 연구에서는 비디오의 매개 이미지-자막 쌍을 입력하여 딥하이퍼넷의 학습과정에서 순차적으로 시각적 언어 개념망을 만드는 동시에 에피소드를 관찰하면서 순차적 베이지안 추론에 의해 개념망의 이미지-자막 쌍을 업데이트한다.

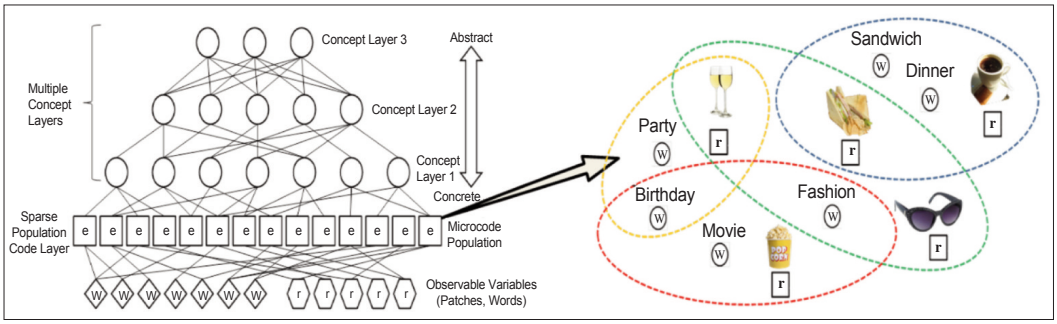
$$P_t(\mathbf{e},\boldsymbol{\alpha},\mathbf{c^1}|\mathbf{r},\mathbf{w},\mathbf{c^2}) = \frac{P(\mathbf{r},\mathbf{w}|\mathbf{e},\boldsymbol{\alpha},\mathbf{c^1},\mathbf{c^2})P(\mathbf{c^2}|\mathbf{e},\boldsymbol{\alpha},\mathbf{c^1})P_{t-1}(\mathbf{e},\boldsymbol{\alpha},\mathbf{c^1})}{P(\mathbf{r},\mathbf{w},\mathbf{c^2})} \quad (7)$$

이진 벡터 r, w는 이미지 조각과 단어의 특징 벡터이고 c^1, c^2는 노드의 존재여부를 판단한다. e는 하이퍼에지들의 집합이고 α는 에지들의 가중치이다. 파라미터 θ(e,α)와 c^1, c^2가 주어졌을 때 (7)의 수식으로 학습이 진행된다. P_t는 t번째 에피소드에 대한 매개변수의 확률 분포이다. t번째 에피소드를 관찰하였을 때 사전 확률 분포 P_{t-1}(θ)는 우도와 표준 값을 계산함으로써 사후 확률 분포를 업데이트한다. 자세한 학습 과정은 [17]에서 소개하였다.

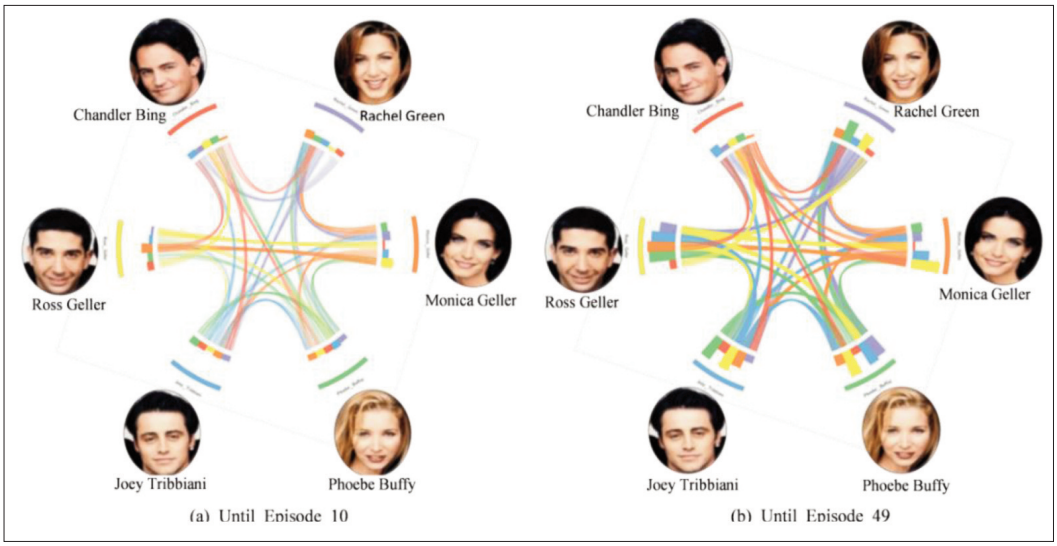
2. 비디오 정보 추출

등장인물 인식 방법: 이 절에서는 딥러닝 기술을 활용하여 드라마 영상에서 등장인물을 인식한 결과를 소개한다. 먼저 TV 드라마속의 얼굴을 검출기를 이용하여 장면 속 얼굴 영역을 검출한다. 검출된 얼굴을 인식하기 위하여 본 연구에서는 2012년 R. Socher 등이 제안한 컨볼루션-재귀 신경망(Convolutional - recursive neural network, CNN-RNN)[18]을 적용하였다. 미국 TV 드라마 'Friends'의 등장인물 총 6명의 얼굴 이미지 총 6000장을 수집하여 학습한 결과 89%의 인식률을 보였다.

장소 분류 방법: 본 연구에서는 장소를 분류하기 위해 Bag of features(Bof)모델을 사용한다. Bof모델은 각 장소 이미지를 고유 특징 벡터(Eigen-vectors)의 집합으로 정의하고 특징 벡터들의 분포를 학습하여 분류하는 기술이다. TV 드라마에서 주로 등장하는 7개 장소 이미지에서 각각 200개의 학습 데이터와 100개의 테스트 데이터를 선정하였다. 장소 분류 실험을 한 결과 77.0%의 인식률을



〈그림 12〉 딥하이퍼넷 학습 모델 구조도



〈그림 13〉 TV 드라마 'Friends' 등장인물 사이의 연관성 분석 결과

보였다.

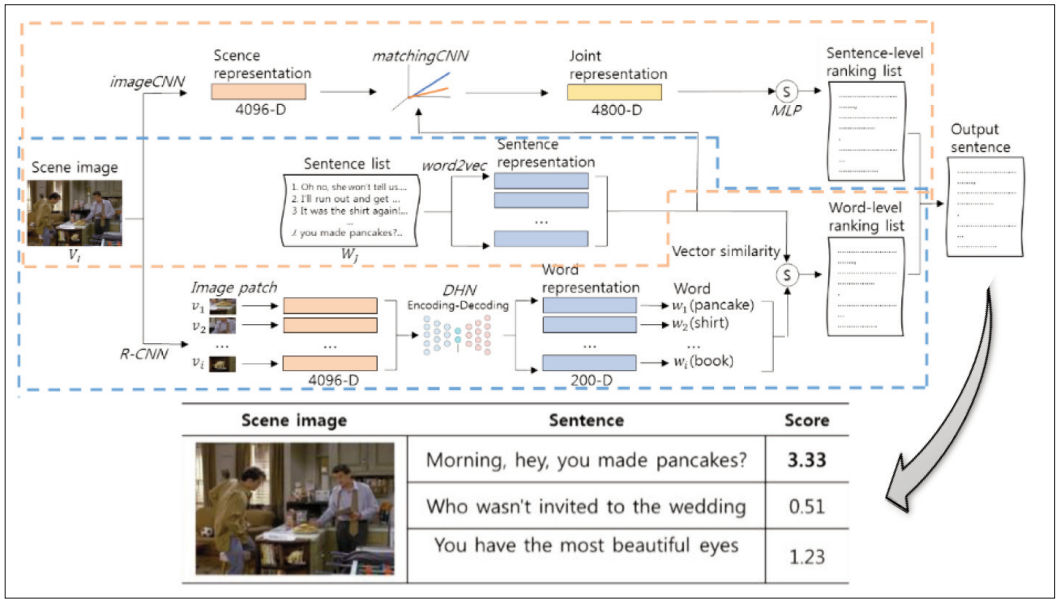
3. 데이터 전처리 과정

실험 대상으로서 TV 드라마 'Friends'의 183편 에피소드, 총 4400분 분량의 비디오 데이터를 사용하였다. 전체 비디오는 자막의 출현시간을 기준으로 이미지 프레임을 추출하여 약 6M개의 이미지-자막 쌍의 데이터로 변환하였다. 등장인물 인식과 장소 분류 방법을 통해 시각 정보를 추출하였고 의자, 램프, 컵 등 기타 물체 인식은 R-CNN모델[19]을 적용하여 이미지 조각을 생성하였다. 이미지 조각은 4096차원의 컨볼루션 신경망 특징 벡터로 표현한다. 출현된 자막은 단어(Word) 단위로 Word2vec[20]을 적용하여 200차원의 실수 벡터로 변환하였다.

4. TV 드라마 스토리 분석

이 절에서는 딥하이퍼넷의 학습을 통해 구축된 지식망을 이용하여 TV 드라마속 등장인물의 관계를 분석한 실험 결과를 소개한다. 〈그림 13〉은 등장인물의 연관성을 분석한 실험 결과이다. 그래프에서 두 인물 사이에 연결된 선의 개수는 그들이 공유하는 하이퍼에지의 개수를 표시하고 각 등장인물에 표시된 히스토그램은 기타 인물과 공유하는 하이퍼에지들의 가중치의 합이다. 즉 통계값이 높을수록 연관성이 높음을 의미한다. 그래프를 보면 드라마 10편의 인물관계에 비해 49편까지 학습한 등장인물사이의 연관성은 상대적으로 높아졌음을 확인할 수 있다. 이런 정보는 드라마 속의 인물 등장 비율, 대본의 양, 인물 중요성 등 기타 관련 정보를 추측하는데 정량적인 근거가 될 수 있다.

또한 학습한 지식을 이용하여 비디오 묘사글 검색 문제에 적용할 수 있다. 〈그림 14〉는 비디오 스토리 묘사글 검색 과정과 실험 결과 예시이다.



〈그림 14〉 비디오 스토리에 대한 묘사글 검색 문제 예시

V. 요약 및 결론

본 고에서는 비디오 분석 연구를 위한 대용량 비디오 데이터셋과 대표적인 딥러닝 기술을 살펴 보았다. 비디오와 같은 다중모달 데이터를 다루기 위한 영상처리 기법과 언어처리 기법의 최신 연구동향을 정리하고 본 연구진의 TV 드라마 분석 연구를 통해 실제 응용 사례를 소개하였다. 딥러닝 기술의 발전과 컴퓨팅 능력의 향상은 비디오와 같은 대용량 데이터를 분석하는데 기술적인 배경이 되었다. 이러한 동향을 바탕으로 비디오 스토리의 학습에 필요한 데이터 또한 점차 풍부해질 것으로 기대되며, 앞으로도 보다 혁신적인 후속 연구가 계속 나올 것으로 기대된다.

참 고 문 헌

[1] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In Proceedings of *IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1725-1732. (2014)

[2] I.-H. Jhuo and D.T. Lee. Video event detection via multi-modality deep Learning. In Proceedings of *International Conference on Pattern Recognition*. pp. 666-671. (2014)

[3] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri. C3D: Generic features for video analysis. arXiv preprint arXiv:1412.0767. (2014)

[4] C.-J. Nan, K.-M. Kim and B.-T. Zhang. Social network analysis of TV drama characters via deep concept hierarchies. In *Proceedings of International Conference on Advances in Social Networks Analysis and Mining*. pp. 831-836. (2015)

[5] K. Kim, C. Nan, M.-O. Heo, S.-H. Choi and B.-T. Zhang. PororoQA: Cartoon video series dataset for story understanding. In *Proceedings of NIPS 2016 Workshop on Large Scale Computer Vision System*. (2016)

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

[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. (2014)

[8] 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*. pp. 1-9. (2015)

[9] K. He, X. Zhang, S. Ren and J. Sun. Deep residual learning for image recognition. In Proceedings of *the IEEE Conference on Computer Vision and Pattern Recognition*. (2016)

[10] I. Goodfellow, J. Pouget-Abadie et al. Generative adversarial nets. In Proceedings of *Advances in Neural Information Processing Systems*. pp.2672-2680. (2014)

[11] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In Proceedings of *International Conference on Learning Representations*. (2015)

[12] A. Graves, A. Mohamed, G. Hinton. Speech recognition with deep recurrent neural networks. In Proceedings of *2013 IEEE international conference on acoustics, speech and signal processing*. pp. 6645-6649. (2013)

[13] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput. vol. 9*. pp. 1735-1780. (1997)

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

[15] http://benjamin.wtf

[16] 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*. pp. 3156-3164. (2015)

[17] J.-W. Ha, K.-M. Kim and B.-T. Zhang. Automated construction of visual-linguistic knowledge via concept learning from cartoon videos. In Proceedings of *the Twenty-Ninth AAAI Conference on Artificial Intelligence*. pp. 522-528. (2015)

[18] R. Socher, B. Huval, B. Bath, C. D. Manning and A. Y. Ng. Convolutional-recursive deep learning for 3D object classification. In Proceedings of *Advances in Neural Information Processing Systems*. pp. 665-673. (2012)

[19] R. Girshick, J. Donahue, T. Darrell and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of *International Conference on Pattern Recognition*. pp. 580-587. (2014)

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

필 자 소 개

남 장 군

– 2014년 : Harbin Engineering University 전자정보공학부 학사
– 2014년 ~ 현재 : 서울대학교 컴퓨터공학부 석사과정
– 주관심분야 : 기계학습, 컴퓨터 비전, 인지과학

김 진 화

– 2011년 : 광운대학교 컴퓨터소프트웨어학과 학사
– 2015년 : 서울대학교 협동과정 인지과학전공 석사
– 2015년 ~ 현재 : 서울대학교 협동과정 인지과학전공 박사과정
– 주관심분야 : 딥러닝, 주의기반 인지시스템

김 병 희

– 2003년 : 서울대학교 컴퓨터공학부 학사
– 2006년 : 서울대학교 컴퓨터공학부 박사과정 수료
– 2006년 : 독일 베를린공대 방문연구원
– 2006년 ~ 현재 : 서울대학교 컴퓨터공학부 연구원
– 주관심분야 : 기계학습 기반 인공지능, 딥러닝, 순서 정보 학습 및 생성

장 병 탁

– 1986년 : 서울대학교 컴퓨터공학과 학사
– 1988년 : 서울대학교 컴퓨터공학과 석사
– 1992년 : 독일 Bonn 대학교 컴퓨터과학 박사
– 1992년 ~ 1995년 : 독일국립정보기술 연구소 연구원
– 1997년 ~ 현재 : 서울대학교 컴퓨터공학부 교수 및 인지과학, 뇌과학, 생물정보학 협동과정 겸임교수
– 2003년 ~ 2004년 : MIT 인공지능연구소(CSAIL) 및 뇌인지과학과(BCS) 객원교수
– 2007년 ~ 2008년 : 삼성종합기술연구원(SAIT) 객원교수
– 현재 : 서울대학교 인지과학연구소 소장, Applied Intelligence, BioSystems, Journal of Cognitive Science 등 국제 저널 편집위원
– 주관심분야 : 바이오지능, 인지기계학습, 분자진화 컴퓨팅기반 뇌인지 정보처리 모델링