뽀로로봇: 딥러닝 기반의 질의응답 로봇

허유정○, 김경민, 장병탁
서울대학교 컴퓨터공학부
{yjheo, kmkim, btzhang}@bi.snu.ac.kr

Pororobot: A Deep Learning Robot that Plays Video Q&A Games

Yu-Jung Heo○, Kyung-Min Kim, and Byoung-Tak Zhang
School of Computer Science and Engineering, Seoul National University

요 약

최근 기계 학습 기술의 발전으로 로봇 지능과 인간-로봇 상호작용에 대한 연구가 활발히 진행되고 있다. 특히 딥러닝 기법을 활용한 이미지 내 물체 인식, 이미지 기반 설명문 생성 등의 일부 연구는 인간과 유사한 수준의 결과를 보이고 있다. 하지만 불확실하며 고려해야 할 변수가 매우 많은 실세계 환경에서 동적으로 작용하는 로봇 지능을 구현하기 위해, 시간에 따른 개념의 변화에 유연하며 강건하게 반응할 수 있는 의사결정 등 해결해야할 문제가 여전히 남아있다. 본 논문에서는 어린이와 로봇('뽀로로봇')이 질의응답하며 교감하는 시나리오를 제시하고, 비디오 데이터로부터 질의응답을 추론하는 아키텍쳐을 제안한다. 제안된 아키텍쳐는 컨볼루션 신경망과 순환 신경망으로부터 각각 비디오의 이미지, 문장 특징을 추출하고 이를 기반으로 심층 개념구조모델을 구축한다. 아키텍쳐 평가를 위한 초기 실험으로 모델은 183개 에피소드 분량의 '뽀로로' 만화비디오와 1200개 질의응답 데이터를 학습했으며, 모델이 생성한 질문들을 평가한 결과, 비디오 질의응답 시스템으로서 활용 가능성을 확인할 수 있었다.

1. 서 론

최근 개인 교육 서비스를 제공하는 학습 보조 및 가정 교사 로봇에 대한 관심이 증가하고 있으며, 인간-로봇 상호작용(HRI)은 이러한 로봇의 개발에 필수적인 요소로 여겨진다. 기계학습 기술의 발전은 로봇 지능과 인간-로봇 상호작용에 대한 연구를 가속화하였고, 최근 로봇의 시각 장면 이해를 통한 설명문 생성 및 질의응답 생성을 주제로 다양한 연구가 이루어지고 있다[1]. 그러나 불확실하고 고려해야할 변수가 매우 많은 실제 환경에 동적으로 반응하는 인간-로봇 상호작용은 여전히 매우 어려운 문제로 여겨진다[2].

본 논문에서는 이러한 실제 환경에 반응하여 동작할 수 있는 시나리오를 적용하여, 비디오 학습을 통한 질의응답 아키텍쳐의 프로토타입을 제안한다. 제안된 시나리오에서 어린이와 로봇('뽀로로봇')은 실제 환경에서 비디오를 학습한 후 질의응답을 수행한다. 본 연구에서는 컨볼루션 신경망[3,4]과 순환 신경망[5]을 통해 추출한 특징을 기반으로 심층 개념구조모델[6]을 구성하여 영상을 학습하고, 학습된 모델에서 질의응답을 추론한다. 로봇 플랫폼으로는 나오 에볼루션 V5를 사용하였고, 해당 로봇 플랫폼에 제안된 아키텍쳐를 적용한 질의 생성을 통해 시스템의 유효함을 확인하였으며, 추후 연구의 방향성을 제시한다.

본 논문의 구성은 다음과 같다. 2절에서는 비디오 질의응답 로봇을 소개하고, 3절에서는 비디오 질의응답 아키텍쳐를 소개한다. 4절에서는 교육용 만화 비디오 데이터를 적용한 실험을 설계하고 실제 도출된 중간 결과를 소개하며, 마지막으로 5절에서는 본 논문에 대한 결론을 맺는다.

2. 비디오 질의응답 로봇

비디오 질의응답 게임 로봇은 두 가지 측면에서 어린이의 교육을 돕는다. 첫째, 어린이와 로봇은 질의응답을 통해 영상에 등장한 새로운 개념 또는 지식을 학습할 수 있다. 둘째, 어린이와 로봇은 함께 영상을 보고 상호작용하며 공통된 경험을 공유하고, 이는 어린이의 사회성과 의사소통능력을 향상시킨다.

본 연구에서 주어진 시나리오는 다음과 같다. 어린이와 로봇('뽀로로봇')은 함께 교육용 만화 비디오를 시청한다. 로봇은 시청한 비디오를 학습하여 어린이에게 질문한다. 어린이가 질문에 대답하였을 때, 어린이의 대답이 올바르다면 로봇은 대답에 동의한다. 어린이의 대답이 올바르지 않다면, 로봇은 어린이에게 정답을 알려준다. 로봇은 어린이의 응답의 정답 여부를 판별하여, 올바른 피드백을 줄 수 있다. 제안된 비디오 질의응답 로봇 시나리오의 실험환경은 그림 1과 같다[7].

그림 1 비디오 질의응답 로봇의 실험환경

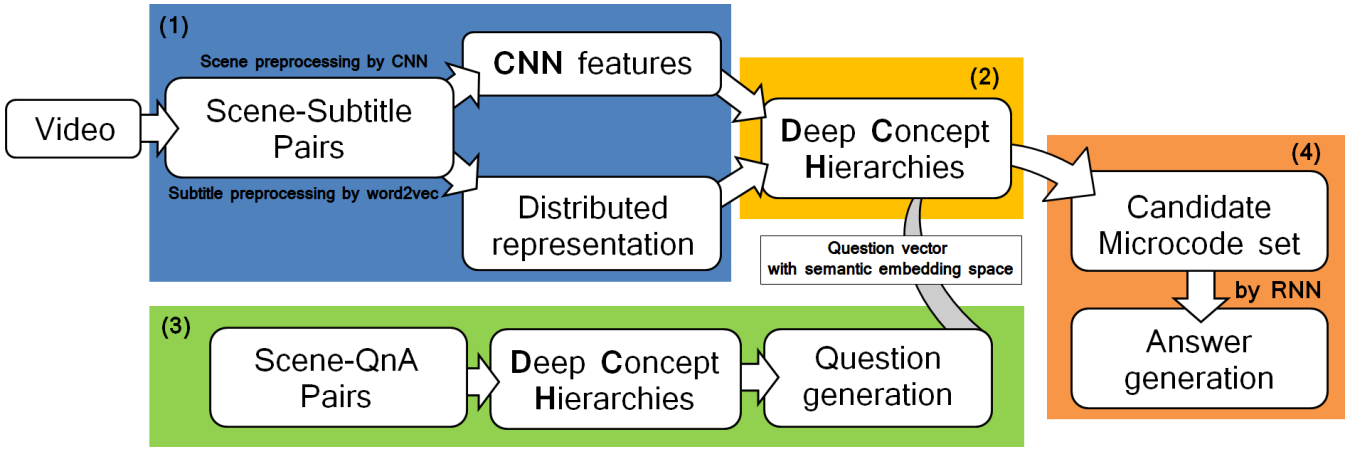

그림 2 제안된 질의응답 로봇 시스템의 흐름도

3. 비디오 질의응답 시스템

비디오 질의응답 시스템의 흐름도는 그림 2와 같으며, 4단계로 구성된다. 각 단계는 (1) 전처리 : 컨볼루션 신경망과 순환 신경망을 통해 영상에서 시각, 언어적 특징을 추출하는 단계 (2) 비디오의 개념 학습: 추출된 특징을 기반으로 시각-언어적 개념을 포함하는 심층 개념구조모델을 구성하는 단계 (3) 질문 생성 : 영상-질의응답 데이터를 학습한 심층 개념구조모델을 구성 및 주어진 이미지를 대상으로 질문을 생성하는 단계 (4) 답변 생성 : 순환 신경망을 통해 학습 모델의 마이크로 코드를 추출하여 질문에 대한 응답을 생성하는 단계이며, 자세한 설명은 다음과 같다.

3.1 전처리

시각, 언어적 특징을 추출하기 위해 해당 영상에서 자막이 나타날 때마다 장면과 자막의 쌍을 수집한다. 수집된 장면은 컨볼루션 신경망[3,4]을 통해 이미지 패치로 변환되고, 수집된 자막은 순환 신경망을 통해 실제값의 벡터로 변환된다. 자막 변환시, word2vec을 적용하여 단어의 문맥적 의미를 내포하게 표현한다[5].

3.2 비디오의 개념 학습

3.1에서 추출된 시각, 언어적 특징을 기반으로 심층 개념구조모델 D_1 을 구성한다. 구성된 개념구조모델은 그림 3과 같다. 심층 개념구조모델은 이미지 패치 r과 텍스트 벡터 w의 SPC(Sparse Population Coding)[8]로 인코딩된 마이크로 코드 h를 포함한다. 첫 번째 개념층 (C_1)은 마이크로코드의 집합을 의미하고 두 번째 개념층 (C_2)은 영상에서 관찰된 각각의 등장인물을 의미한다.

3.3 질문 생성

영상에 대한 질의응답 데이터를 이용하여 질의응답 패턴을 학습한 심층 개념구조 모델 D_2를 구성한다.

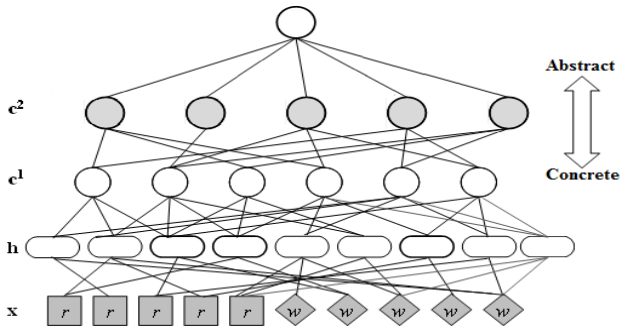

그림 3 심층 개념구조모델의 구조

구성된 모델의 매개변수가 θ 이고, 쿼리로 주어지는 영상에서 추출된 이미지 패치가 $\boldsymbol{r}$이며, 생성된 최적의 질문이 q^* 일때, 이미지-질의 변환의 공식은 식 (1)과 같고, 질의 생성에 사용된 알고리즘은 [9]과 같다.

$$q^* = \underset{q}{\operatorname{argmax}} P(\boldsymbol{q}|\boldsymbol{r},\theta) = \underset{q}{\operatorname{argmax}} P(\boldsymbol{r}|\boldsymbol{q},\theta)\,P(\boldsymbol{q},\theta) \quad (1)$$

3.4 답변 생성

3.3에서 추출된 질의 q^*에 word2vec을 적용하여 3.2의 심층 개념구조모델과 같은 의미 벡터 공간에 사상하고, 마이크로코드 $[E_1,\ldots,E_i]$의 코사인 유사도를 측정하는 함수 S_m을 통해, 후보 마이크로코드 m을 추출한다.

$$\mathrm{m} = \underset{i}{\operatorname{argmax}}\, S_m(q^*, E_i) \quad (2)$$

선택된 마이크로코드 m에 순환 신경망을 적용해 응답을 생성한다. 3.3에서 생성된 질문의 마이크로코드가 M_q이고, 응답을 생성하기 위한 유클리드 유사도 비교 함수가 S_a 일 때, 응답을 생성하는 공식은 다음과 같다.

$$a = \operatorname{argmax} S_a(q^*, [\mathrm{m}_1, \mathrm{m}_2, \ldots, \mathrm{m}_i], M_q) \quad (3)$$

답변 생성을 위해 사용된 알고리즘은 최근 이미지 질의응답 문제에 적용된 m-RNN 기술과 같다[1].

4. 실험 설계

4.1 로봇 플랫폼

본 연구에서는 초음파, 적외선 등의 센서를 갖추어 주변 환경을 스스로 인식할 수 있으며, 25개의 관절로 구성되어 전신 동작 제어가 가능한 알데바란 로보틱스사의 휴머노이드 로봇 나오 에볼루션 V5를 사용하였다.

4.2 실험 데이터 명세

본 실험에서는 183개의 에피소드, 1232분 분량의 교육용 만화 비디오 뽀로로를 실험 데이터로 사용하였다. 만화 비디오는 어린이가 학습할 수 있는 쉬운 줄거리로 구성되며, 간단한 언어 패턴을 보이고 이미지 처리 또한 실세계 영상에 비해 수월하다는 장점을 가진다. 영상에 대한 질의응답 생성을 위한 학습데이터로 만화 비디오를 보고 사용자가 생성한 1200개의 질의응답쌍을 사용하였다. 질의는 크게 두 가지 유형으로, 영상의 이미지 정보를 묻는 유형(예를 들어, "뽀로로와 친구들은 무엇을 타고있는가?")과 영상의 내용 정보를 묻는 유형(예를 들어, "왜 뽀로로는 요리를 하는가?") 으로 이루어진다.

4.3 질문 생성 결과

3.3의 심층 개념구조모델을 통해 생성된 질문의 적합성을 판단하기 위해, BLEU Score 평가 및 피험자에 의한 직접 평가를 수행하였다. BLEU Score는 주로 기계번역 문제에 사용되는 지표로 [0,1]의 범위를 가지며 값이 높을수록 적합한 번역을 의미한다. 해당 실험에서는 0.3513의 BLEU Score를 기록하였다. 사람에 의한 직접 평가는 7명의 피험자를 대상으로 학습모델이 생성한 80개의 질문에 대한 적합도를 평가하였으며, [0,1]의 범위로 척도화하여 0.5068을 기록하였다. 기록된 질문의 적합성 판단 수치는 표 1과 같으며, 영상에 따라 생성된 질문의 예는 그림 4와 같다.

5. 결론 및 향후 연구방향

본 논문에서는 딥러닝 기반의 질의응답 아키텍쳐를 제안하였다. 제안된 아키텍쳐는 시각-언어적 특징 추출, 시각-언어적 특징을 반영하는 심층 개념구조모델 구성, 질의응답 패턴을 반영하는 심층 개념구조모델의 구성 및 질문 생성, 마이크로코드 추출 및 순환 신경망을 통한 응답 생성의 4단계로 구성된다. 제안된 아키텍쳐에서 도출된 질의 생성을 통해 해당 시스템의 유효함을 확인할 수 있었다. 제안된 아키텍쳐는 목적이 분명한 에이전트로 확장될 수 있고, 점증적 학습이 가능하다. 이러한 구현을 위해, 보다 높은 차원의 데이터를 실시간으로 처리할 수 있는 대규모의 GPU 클러스터가 요구된다.

표 1 생성된 질문에 대한 평가 결과

Generated Questions From DCH	BLEU Score	Averaged Human Rated Score
	0.3513	0.5068

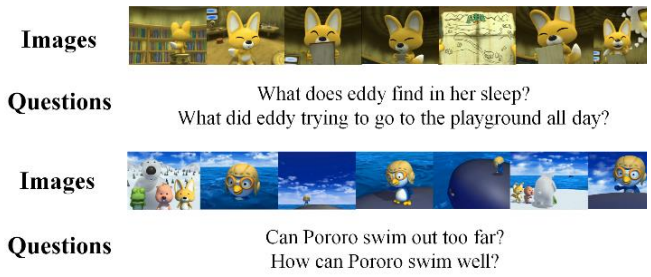

그림 4 영상에 대해 생성된 질문의 예

감사의 글

이 논문은 2015년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원(R0126-15-1072-SW스타랩, 10044009-HRI.MESSI)을 받아 수행된 연구임.

6. 참고 문헌

[1] Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., Xu, W. Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering. arXiv preprint arXiv:1505.05612.

[2] Zhang, B.-T. 2013. Information-Theoretic Objective Functions for Lifelong Learning. *AAAI 2013 Spring Symposium on Lifelong Machine Learning*. 62-69.

[3] Krizhevsky, A., Sutskever, I.., and Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Proceedings of Advances in Neural Information Processing Systems (NIPS)*. 1097-1105. 2012

[4] Girshick, R., Donahue, J., Darrell, T., Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 580-587. 2015

[5] Mikolov, T., sutskever I., Chen K., Corrado G., Dean J. Distributed Representation of Words and Phrases and their Compositionality. *Proceedings of Advanced in Neural Information Processing Systems (NIPS)*. 3111-3119. 2013

[6] Ha, J.-W., Kim, K.-M., Zhang, B.-T. Automated Visual linguistic Knowledge Construction via Concept Learning from Cartoon Videos. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*. 522-528. 2015

[7] Kim, K.-M., Nan, C.-J., Ha, J.-W., Heo, Y.-J., and Zhang, B.-T., Pororobot: A deep learning robot that plays video Q&A games, *AAAI 2015 Fall Symposium on AI for Human-Robot Interaction (AI-HRI 2015),* 2015

[8] Zhang, B.-T., Ha, J.-W., and Kang, M. 2012. Sparse Population Code Models of Word Learning in Concept Drift. *In Proceedings of the 34th Annual Conference of Cognitive Science Society(Cogsci 2012)*. 1221-1226, 2012

[9] Kim, K.-M., Ha, J. -W., Lee, B.-J., and Zhang, B.=T., Character-based Subtitle Generation by Learning of Multimodal Concept Hierarchy from Cartoon Videos, *in Journal of Korean Institute of Information Scientists and Engineers*, 2015