뽀로로Q&A: 뽀로로 비디오 스토리 질의응답

남장군°, 김경민, 장병탁
서울대학교 컴퓨터공학부
{cjnan, kmkim, btzhang}@bi.snu.ac.kr

Pororo Q&A: Story-based Question Answering from Pororo Videos

Chang-Jun Nan°, Kyung-Min Kim and Byoung-Tak Zhang
School of Computer Science and Engineering, Seoul National University

요 약

딥러닝 기술에 GPU를 활용한 병렬처리 기법이 접목되면서 이미지 이해, 음성인식, 자연언어처리 등의 분야에서 놀라운 연구 성과들이 보고되고 있다. 최근 연구자들은 기술 접목 분야를 넓혀 비디오 스토리 질의응답 문제에 도전을 하고 있지만 아직까지 인간 수준의 성능을 얻지 못하고 있고 풀어야할 문제들은 여전히 남아있다. 특히 비디오의 멀티모달한 지식 표현 특성과 질의응답의 정량적 평가방법은 벤치마크로 사용될 대용량의 비디오 스토리 질의응답 데이터가 부족한 상황에서는 고안되기 어렵다. 본 논문에서는 어린이 만화 비디오 "뽀롱뽀롱 뽀로로"에 대한 스토리 기반 질의응답 데이터와 베이스라인 성능을 제시한다. 데이터 수집을 위해 사람 제작자들이 "뽀롱뽀롱 뽀로로" 비디오 183개 에피소드를 본 뒤 스토리에 대한 설명-질문-답 쌍을 자유로운 형식으로 작성하였다. 수집된 질의응답 쌍 2500여개로 베이스 라인 테스트를 한 결과 12.93%의 정확도를 보였다.

1. 서 론

최근 딥러닝 기술에 GPU를 활용한 병렬처리 기법이 접목되면서 인공지능의 놀라운 연구 성과들이 보고되고 있다. 특히 이미지 이해, 음성인식, 자연언어처리 등의 분야에서 인간 수준에 가까운 정확성을 보여주었다.[1] 최근 연구자들은 기술 접목 분야를 넓혀 비디오 스토리 질의응답 문제에 도전을 하고 있다. 최신 연구성과들을 보면 아직까지 인간 수준의 성능을 얻지 못하고 있고 풀어야할 문제들은 여전히 남아있다.[2-4] 특히 비디오의 멀티모달하고 동적인 지식 표현 특성과 질의응답의 정량적 평가방법은 벤치마크로 사용될 대용량의 비디오 스토리 질의응답 데이터가 부족한 상황에서는 고안되기 어렵다.

만화 비디오 데이터는 동적이고 멀티모달 하면서도 비전 처리가 상대적으로 쉽고 스토리 내용이 용이한 장점이 있어 실세계에서 인터랙션을 하는 어린이 교육용 로봇을 구현하기에 가장 적합한 테스트베드이다. 이러한 특성을 이용해 [5]에서는 하이퍼네트워크 모델을 이용하여 만화 비디오로부터 질의응답을 학습하는 프레임워크를 제안하였고 [6]에서는 이를 어린이 교육용 로봇에 적용하여 만화 비디오로부터 질의응답을 추론하는 아키텍처를 구현하였다. 상기 연구결과들은 하이퍼네트워크 모델이 질의응답 데이터를 효율적으로 학습할 수 있다는 특성을 입증하였으나 데이터가 부족하는 문제가 있었다.

본 논문에서는 어린이 만화 비디오에 대한 스토리 기반 질의응답 데이터의 수집 방안을 제안하고 최근에 수집된 데이터를 베이스 라인 테스트한 성능을 제시한다. 데이터 수집을 위해 사람 제작자들이 데이터 수집 사이트[7]를 통해 "뽀롱뽀롱 뽀로로" 비디오 183개 에피소드를 본 뒤 스토리에 대한 설명-질문-답 쌍을 자유로운

Video Stream	
Description	Crong missing his toycar.
Question	What are they finding?
Answer	Toycar.
Video Stream	
Description	Pororo and his friends are playing hide-and-sick.
Question	What are Pororo and his friends playing?
Answer	Hide-and-sick.

그림 1 뽀로로Q&A 데이터 형식

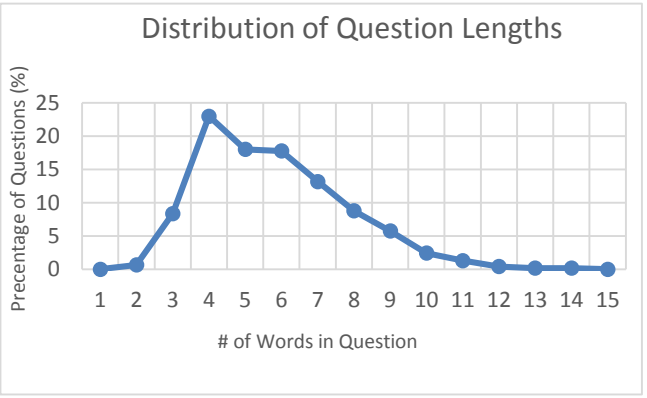

그림 2 뽀로로Q&A 데이터 질문의 단어 길이 분포

형식으로 작성하였다. 수집된 질의응답 쌍 2500여개로 베이스라인 테스트를 한 결과 12.93%의 정확도를 보였다.

본 논문의 구성은 다음과 같다. 2절에서는 뽀로로 질의응답 데이터의 구성을 소개하고, 3절에서는 수집 데이터를 분석한다. 4절에서는 비디오 스토리 질의응답 데이터를 적용한 실험을 소개하고 베이스라인 성능을 제시한다. 5절에서는 본 논문에 대한 결론을 맺는다.

2. 뽀로로Q&A 데이터 구성

뽀로로Q&A는 만화 비디오 "뽀롱뽀롱 뽀로로"의 스토리에 대한 자유형식(Free-form) 질의응답 데이터셋이다. 그림 1에서 수집된 데이터의 형식을 보여준 바와 같이 뽀로로Q&A 데이터는 일정 시간 단위의 비디오 클립과 비디오 클립의 스토리에 상응한 설명-질문-답(Description-Question-Answer, DQA)의 텍스트 쌍 형식으로 구성된다.

2.1 뽀로로 만화 비디오

뽀로로Q&A 데이터는 어린이 만화 비디오 "뽀롱뽀롱 뽀로로"의 183편 에피소드, 총 1232분 분량을 대상으로 수집하였다. 비디오 데이터는 자막이 출현하는 시점에서 앞뒤 3초를 기준으로 비디오 클립 형식으로 저장하였고 사람 제작자들은 각 비디오 클립의 스토리에 대해 질의응답 데이터를 작성하였다.(그림 1)

2.2 설명-질문-답

데이터의 범용성을 고려하여 설명-질문-답 쌍은 전부 영어로 작성하였다. 설명문은 스토리에 기반하여 자유로운 형식으로 작성하였고 질문은 "What is...", "What did...", "Who is...", "Where are..."와 같은 여러가지 종류의 질문들이 포함되여 있다. 질의응답의 평가기준을 고려하여 답은 하나의 단어로 작성하였고 여러 개의 단어로 표현해야 하는 경우 "Hide-and-sick"와 같은 방식으로 하나의 단어로 결합하였다.

3. 질의응답 데이터 분석

3.1 질문 데이터 분석

수집된 뽀로로Q&A 데이터를 분석한 결과 질문들을 보면 주로 "What is...", "What did...", "Who is...", "Where are..."와 같은 유형들이 있었고 표 1에서 이들의 출현하는 비율을 보여주었다. 질문들은 여러 개의 단어들로 조합되였고 문장의 단어 개수의 분포를 나타내고 그림 2과 같다. 결과를 보면 25%이상의 질문은 4개의 단어로 구성되였다는 것을 알수가 있다.

Question Type		Answer Type	
What is...	11.65%	Pororo	4.30%
What did...	11.55%	Eddy	3.80%
Who is...	7.21%	Crong	3.72%
What are...	4.37%	Poby	3.48%
Where did...	4.25%	Yes	2.63%
Where is...	3.52%	Loopy	2.51%
What was...	3.20%	House	2.47%
Who did...	2.63%	harry	1.94%

표 1 질문의 유형과 출현하는 답들의 분석 결과

3.2 답 데이터 분석

표 1에서 가장 많이 나타나는 답들을 순서대로 보여주었다. 결과를 보면 "Yes/No"문제 외에 "House"와 같은 오브젝트가 있다. 하지만 등장인물의 이름으로 답이 가장 많은 것을 알 수가 있다. 특히 "Pororo"와 같은 주인공은 제일 많이 나타났다.

4. 베이스 라인 성능 테스트

베이스 라인 성능 테스트를 위해 183편의 "뽀롱뽀롱 뽀로로"가 데이터로 사용되였다. 또한 수집된 데이터 약 2500개의 설명-질문-답 쌍을 준비하였다. 실험을 위하여 학습/테스트 데이터를 4:1 비율로 랜덤으로 나누었다.

본 논문에서는 먼저 R-CNN모델을 사용하여 이미지 조각을 벡터로 표현하고 설명-질문-답 쌍의 문장들을 BoW(Bag of Words)와 TF-IDF(Term frequency-inverse document frequency) 벡터 두가지로 표현하였다. BoW는 단순히 문장에서 단어의 출현여부를 나타내는 벡터이다. 이러한 표현 방법은 "is", "are", "be" 와 같은 가치가 높지않은 단어들이 많이 출현하는 문제점을 고려하지 않았다. TF-IDF는 단어의 중요성을 고려하여 단어들을 실수 가중치로 표현할 수가 있다.

정량적 평가를 위하여 본 논문에서는 데이터 마이닝 기법을 참고한다. 비디오 클립과 설명-질문 쌍이 주어졌을 때 이의 특징벡터와 학습 데이터에 있는 특징벡터들을 비교하여 가장 유사한 데이터의 답을 출력으로 한다. 특징벡터의 비교 기준인 Jaccard similarity 다음과 같다.

$$Sim(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (1)$$

출력된 답과 실제 테스트셋의 답과 비교하여 정확도를 알수가 있다. 표 2는 총 30회의 실험을 통해 얻은 베이스 라인 성능을 보여주며 결과로부터 비디오 질의응답 태스크의 난이도를 알수가 있다.

	BoW Score	TF-IDF Score
Jaccard similarity	12.93%(avg)	10.49%(avg)
	15.59%(max)	11.94%(max)

표 2 베이스 라인 성능 테스트 결과

5. 결론 및 논의

본 논문에서는 어린이 만화 비디오 "뽀롱뽀롱 뽀로로"에 대한 스토리 기반 질의응답 데이터와 베이스라인 성능을 제시했다. 데이터의 수집 방안과 데이터 분석결과를 통해 데이터의 구조를 살펴보았고 베이스라인 성능 테스트의 결과로부터 비디오 스토리 기반 질의응답 태스크의 난이도를 확인할 수 있었다. 제안된 뽀로로Q&A 데이터는 비디오 스토리 기반 질의응답 시스템 학습에 사용할 수 있으며 정량적 성능평가 기준이 될 것으로 기대한다. 이러한 응용을 위해서는 현재까지 수집된 데이터에 비해 더 많은 데이터가 필요함으로 보아 "Amazon"과 같은 사이트를 통해 작업을 진행해 볼 수가 있다.

감사의 글

이 논문은 2016년도 정부(미래창조과학부, 국방부)의 재원으로 정보통신기술진흥센터(R0126-16-1072-SW스타랩), 한국산업기술평가관리원(10044009-HRI.MESSI, 10060086-RISF), 국방과학연구소(UD130070ID-BMRR)의 지원을 받았음.

참고문헌

[1] H. Fang, S. Gupta, F. N. Iandola, R. Srivastava, L. Deng, P. Doll ár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From Captions to Visual Concepts and Back. *In Computer Vision and Pattern Recognition*. 2015.

[2] X. Chen and C. L. Zitnick. Mind's Eye: A Recurrent Visual Representation for Image Caption Generation. *In Computer Vision and Pattern Recognition*. 2015.

[3] M. Hodosh, P. Young, and J. Hockenmaier. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *In Journal of Artificial Intelligence Research*. 2013. pp. 853-899.

[4] S. Antol, A. Agrawal, J. Lu, et al. VQA: Visual question answering. *In Proceedings of the IEEE International Conference on Computer Vision*. 2015. pp. 2425-2433.

[5] 김경민, 남장군, 하정우, 허유정, 장병탁. 비디오 질의응답을 위한 듀얼 심층 메모리. *2015 한국정보과학회 동계학술발표회 논문집*. 2015. pp. 654-656.

[6] 허유정, 김경민, 장병탁. 뽀로로봇: 딥러닝 기반의 질의응답 로봇. *2015 한국정보과학회 동계학술발표회 논문집*. 2015. pp. 645-647.

[7] http://ai.snu.ac.kr:3000