비디오 다중 형식 데이터 기반의 질의 응답을 위한 동적 메모리 네트워크 학습 기법

황보선, 온경운, 김은솔, 장병탁
서울대학교 컴퓨터공학부
{bshwang, kwon, eskim, btzhang}@bi.snu.ac.kr

Multimodal Dynamic Memory Networks Learning For Video Question Answering

Bosun Hwang, Kyoung Woon On, Eun-Sol Kim, Byoung-Tak Zhang
School of Computer Science and Engineering, Seoul National University

요 약

본 논문은 고전 동화 내용에 기반을 둔 아동 비디오의 다중 형식(multimodal) 데이터를 입력으로 하여, 해당 비디오 스토리의 질의에 대한 응답을 동적 메모리 네트워크를 사용하여 출력하는 시스템을 제안한다. 제안된 시스템은 순차적인 비디오 이미지와 텍스트 정보를 모두 사용하여 기존의 이미지 질의/응답 문제에서 확장되어, 비디오 스토리를 학습하여 문장 형태의 응답을 한다는 차별성을 가진다. 실험을 위하여 4가지 동화 비디오로부터 이미지와 자막 정보를 추출하였고, 해당 비디오의 스토리에 대해 묻고 답하는 질의 응답 데이터 셋을 Amazon Mechanical Turk를 이용 수집하여 학습에 사용하였다. 실험 결과, 제안된 스토리 기반 학습 시스템이 에피소드 기반의 질의/응답에 문제에 적용 가능함을 확인할 수 있었으며, BLEU(Bilingual Evaluation Understudy) 스코어 기준 video + text 결합 형태의 질의/응답 성능(평균 BLEU score: 0.276)이 text만을 이용한 DMN 구조의 성능(평균 BLEU score: 0.272)보다 좋은 성능 결과를 나타내는 것을 확인할 수 있었다.

1. 서론.

현재 인공지능 분야의 자연어 처리 관련 부문의 새로운 패러다임은 질의/응답(Question Answering: QA) 문제로 집중되고 있다. 전 세계적으로 이를 확장한 VQA(Visual Question Answering) 콘테스트가 개최되는 등 많은 인공지능 연구자들이 연구를 진행하고 있는 상황이다.

본 논문에서는 이러한 QA 문제를 스토리 기반의 QA 문제로 확장하였다. VQA 문제가 주어진 이미지 안에서의 질문에 대한 적절한 대답을 도출하는 문제라면, 본 연구는 에피소드 기반의 스토리에 대한 질문과 이에 대한 대답을 하는 시스템의 개발을 주 목적으로 한다. 이를 위해서 기본적인 framework은 QA 문제에서 좋은 성능을 보여주고 있는 동적 메모리 네트워크(Dynamic Memory Networks: 이하 DMN)를 사용하였으며, text 기반의 스토리 QA 문제를 확장하여, 동영상 정보까지 결합하는 framework을 구현하였다.

2. 동적 메모리 네트워크

동적 메모리 네트워크는 Kumar et al. 그룹이 2015년 ICML에서 발표한 신경망의 일종으로서, QA 문제를 풀기 위한 일반적인 모델이라고 할 수 있다. [2]

기본적으로 Input Module, Question Module, Episodic Memory Module, Answer Module의 네 가지 모듈로 구성이 되어 있으며, 각각은 데이터의 흐름에 따라 각 모듈의 역할에 맞는 데이터 표현(representation)을 만들어 내게 된다. 구체적으로 Input Module은 GRU를 이용하여 입력 데이터의 hidden representation을 만들어 내며, Question Module 역시 GRU를 이용하여 질의에 대한 vector representation을 만들어 낸다. Episodic Memory Module은 주어진 질의에 대한 적절한 대답을 추출할 수 있도록 memory network를 구성하는 역할을 한다. 이러한 Episodic Memory Module은 크게 attention mechanism과 memory update mechanism 두 가지 요소로 구성이 된다. 마지막으로 Answer Module은 question vector와 memory network의 상태를 전달받아 예상되는 대답을 하게 된다. [1][2]

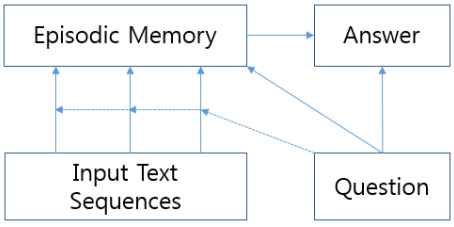

그림 1) Dynamic Memory Networks Module 구조 [1]

3. 아동 비디오 데이터

본 논문에서 사용된 데이터는 Pink Pong이라는 아동을

대상으로 제작된 고전 동화의 내용에 속하는 5분 분량의 짧은 비디오 데이터를 사용하였다. 4편의 비디오 데이터는 자막이 나타나는 시점마다 하나의 instance로 데이터를 생성하였으며, 이에 따라 자막이 나타나는 시간을 기준으로 자막과 함께 비디오 이미지 프레임을 추출하여 데이터 쌍을 생성하였다.
이와 함께 비디오 내용에 따라 관련된 질문과 응답 쌍을 만들었으며, 총 400개의 training question set과 200개의 test question set을 만들어 사용하였다.

Video titles	상영시간	자막 개수
The snow white and The seven dwarves	4m 35s	52
The little mermaid	7m 23s	61
The three little pigs	5m 09s	48
The wolf and the seven sheep	5m 26s	56

표 1) Data descriptions

4. 데이터 전처리 과정

Input Module의 입력으로서 text sequence는 기존 DMN 구조에서의 과정을 따르게 되지만, video image의 경우는 text sequence와의 결합을 위해 적절히 추상화된 representation을 만들어 줘야 한다. 이를 위해 variatoinal CNN autoencoder를 사용하였다. Variational CNN autoencoder는 그림 2)와 같은 구조로 네트워크의 입력과 출력을 동일 이미지로 교사 학습시킴으로써 원본 이미지를 복원하게 되며, 이 때의 mid-layer의 추상화된 vector를 video image의 feature vector로 사용하게 된다. Variational CNN autoencoder는 데이터 자체 값이 아닌 데이터의 분포를 학습함으로써 일반적인 autoencoder보다 좋은 data representation을 얻어낼 수 있다고 알려져 있다. [3]

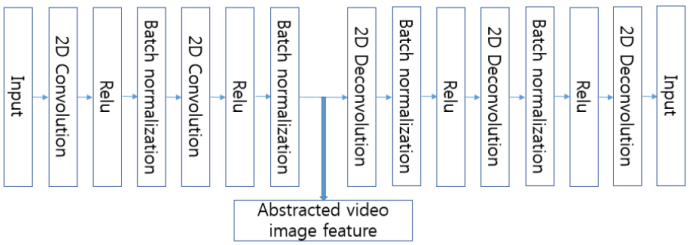

그림 2) Variational Autoencoder 구조

그림 3) 원본 비디오 이미지

그림 4) 복원된 비디오 이미지

5. 데이터 결합 과정

Variational autoencoder를 통해 추출된 Image feature는 자막과 결합된 형태로 전체 시스템의 입력으로 사용되게 되는데, 이종간의 데이터 결합을 위해 Multi-layer perceptron을 이용해 데이터를 결합하게 된다. 이 때 수정된 DMN의 구조는 그림 5)와 같다.

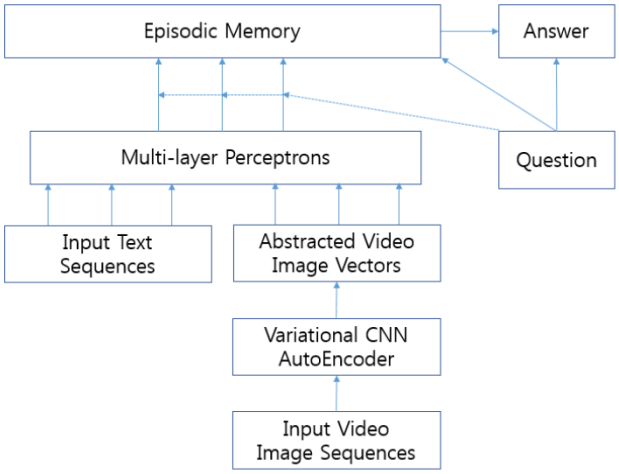

그림 5) Multi-modal pooling DMN module 구조

6. 실험 결과

의뢰를 통해 제작된 질문과 대답 쌍을 이용해 전체 DMN을 학습하여 주어진 질문에 대해 적절한 답이 이루어지는지 실험하였다. 자막만을 입력으로 한 DMN 구조와 비디오 이미지를 결합한 DMN 구조의 두 가지 framework에 대해 결과를 비교하였다. VQA와 달리 대답의 형태가 문장의 형태로 주어지기 때문에, 기계 번역의 성능 척도로 사용되는 BLEU(Bilingual Evaluation Understudy) score를 평가 지표로 사용하였다. 표 2)는 동화 비디오 별 BLEU score값을 나타내며, 그림 6)는 QA 예를 보여주고 있다.

Video titles	자막	자막 + 이미지
The snow white and The seven dwarves	0.257	0.256
The little mermaid	0.270	0.271
The three little pigs	0.244	0.253
The wolf and the seven sheep	0.316	0.322
Averaged BLEU scores	***0.272***	***0.276***

표 2) BLEU score results

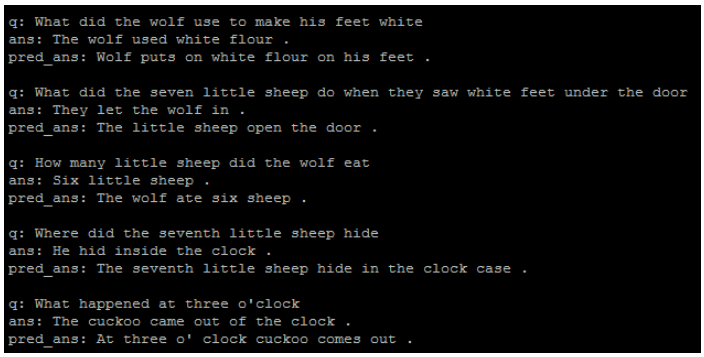

그림 6) QA examples

7. 결론 및 향후 연구 방향

본 논문에서는 DMN framework을 기반으로 기존의 QA나 VQA 문제에서 나아가 에피소드 기반의 QA 문제를 DMN 구조를 활용하여 풀어갈 수 있음을 보였다. 또한, 텍스트 QA와 더불어 비디오 이미지 데이터의 결합된 형태가 시스템 전체 성능에 도움을 줄 수 있음을 확인할 수 있었다.

향 후 연구에서는 QA 데이터 생성 시, 전체 내용에 대한 질문과 답뿐만이 아닌, 비디오 이미지상에서 대답을 구할 수 있는 QA set을 만들어 실험을 진행할 예정이며, 이종 데이터간의 추가적인 결합 방법에 대한 연구도 진행될 예정이다.

8. 참 고 문 헌

[1] Ankit Kumar et al., “Ask Me Anything: Dynamic Memory Networks for Natural Laguage Processing”, ICML 2015

[2] Caiming Xiong, Stephen Merity, Richard Socher, “Dynamic Memory Networks for Visual and Textual Question Answering, ICML 2016

[3] Diederik P Kingma, Max Welling, “Auto-Encoding Variational Bayes”, Stat ML, January 2014

[4] Akira Fukui et al., “Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding”, CVPR 2016

[5] Sainbayar Sukhbaatar et al., “End-To-End Memory Networks”, NIPS 2015

[6] Eun-Sol Kim, Kyoung-Woon On and Byoung-Tak Zhang, “DeepSchema: Automatic Schema Acquisition from Wearable Sensor Data in Restaurant Situations”, IJCAI 2016