Journal of Korea Robotics Society (2019) 14(1):022-030
https://doi.org/10.7746/jkros.2019.14.1.022 ISSN: 1975-6291 / eISSN: 2287-3961

효과적인 인간-로봇 상호작용을 위한 딥러닝 기반 로봇 비전 자연어 설명문 생성 및 발화 기술

Robot Vision to Audio Description Based on Deep Learning for Effective Human-Robot Interaction

박 동 건[1] · 강 경 민[2] · 배 진 우[3] · 한 지 형†
Dongkeon Park[1], Kyeong-Min Kang[2], Jin-Woo Bae[3], Ji-Hyeong Han†

Abstract: For effective human-robot interaction, robots need to understand the current situation context well, but also the robots need to transfer its understanding to the human participant in efficient way. The most convenient way to deliver robot's understanding to the human participant is that the robot expresses its understanding using voice and natural language. Recently, the artificial intelligence for video understanding and natural language process has been developed very rapidly especially based on deep learning. Thus, this paper proposes robot vision to audio description method using deep learning. The applied deep learning model is a pipeline of two deep learning models for generating natural language sentence from robot vision and generating voice from the generated natural language sentence. Also, we conduct the real robot experiment to show the effectiveness of our method in human-robot interaction.

Keywords: Human-Robot Interaction, Video To Audio Description, Video Captioning, Speech Synthesis, Text To Speech

1. 서 론

현재 인공지능 기술은 deep neural network (DNN) 를 필두로 매우 빠르게 발전하고 있다. DNN 기반의 인공지능 기술은 convolutional neural network (CNN)를 이용한 이미지 관련 연구와 recurrent neural network (RNN)을 이용한 음성 및 자연어 처리 관련 연구에서 괄목할 만한 성능을 내고 있다[1-5]. 반면 현재 로봇 기술은 휴머노이드 로봇의 걷기, 휠 베이스 로봇의 제어, 로봇 팔의 제어 등 하드웨어 제어 기술 성숙도는 높은 반면, 아직 로봇 지능 기술의 성숙도는 낮은 편이다. 따라서 현재 발전하고 있는 이미지, 자연어, 음성 관련 인공지능 기술을 로봇에 접목하여 로봇 지능 기술을 발전시킬 필요성이 있다.

기존의 비디오 기반의 자연어 설명문을 생성하는 연구는 다양한 딥러닝 모델을 적용시키며 발전되고 있다[6,7]. 계층적으로 long short term memory (LSTM)을 사용하여 짧은 문장들로 문장간 종속석을 파악하여 문장을 결합하거나[6] attention을 사용하여 문장생성을 한다[7]. 그리고 음성 합성은 Hidden Markov Model (HMM)[8] 기반의 통계적 파라미터 방식으로부터 DNN 방식으로 발전되어 성능이 향상된 연구결과들이 나오고 있다[9-11]. 하지만 기존의 연구는 모두 독립적으로 진행되어 로봇 지능에 접목되지 못했다.

효과적인 인간-로봇 상호작용을 위해서는 로봇이 현재의 상황을 정확하게 인식하는 것뿐만 아니라, 로봇이 현재 상황을 어떻게 인식했는지 상호작용 상대방인 인간에게 효과적으

Received : Dec. 8. 2018; Revised : Jan. 15. 2019; Accepted : Jan.16 . 2019
※ This study was supported by the research fund for a new professor by the SeoulTech (Seoul National University of Science and Technology).
1. Undergraduate Student, Dept. of Computer Science and Engineering, Seoul National University of Science and Technology, Seoul, Korea (zsx160@ seoultech.ac.kr)
2. Undergraduate Student, Dept. of Computer Science and Engineering, Seoul National University of Science and Technology, Seoul, Korea (pinky4117@ seoultech.ac.kr)
3. Undergraduate Student, Dept. of Computer Science and Engineering, Seoul National University of Science and Technology, Seoul, Korea (sjg02122@ seoultech.ac.kr)
† Assistant Professor, Corresponding author: Dept. of Computer Science and Engineering, Seoul National University of Science and Technology, Seoul, Korea (jhhan@seoultech.ac.kr).

로 전달하는 것 역시 필요하다. 인간의 입장에서 가장 원활한 상호작용 방식은 자연어 설명문의 발화를 통한 정보 전달이다. 따라서, 본 논문은 효과적인 인간-로봇 상호작용을 위한 로봇 지능 개발의 한 부분으로서 딥러닝 모델 기반 로봇 비전 자연어 설명문 생성 및 발화 기술 개발을 목표로 한다. 로봇이 상황을 인식하는 주요 센서는 로봇 비전 센서로 효과적인 상호작용을 위해서는 로봇 비전의 단일 이미지 기반 상황 인식이 아닌 시간 흐름이 포함된 비디오 기반의 맥락 인식이 필수적이다. 또한, 로봇 비전의 동영상 맥락 인식 결과를 자연어 설명문으로 생성하고 음성으로 발화하여 상대방에게 전달하는 것이 상호작용 중인 인간 상대방에게 로봇이 인식한 상황을 가장 효과적으로 전달할 수 있는 방법이다.

따라서, 본 논문에서는 로봇 지능 개발을 위해 비디오 기반의 자연어 설명과 음성 합성을 연결하고자 한다. 로봇 비전 영상을 입력으로 받아 자연어 설명문을 생성하는 딥러닝 모델과 자연어 문장을 입력으로 받아 합성음을 생성하는 딥러닝 모델을 파이프라인으로 통합한 모델을 로봇에 적용한다. 로봇 비전 동영상, 자연어 설명문, 합성음은 모두 시간의 흐름을 포함한 sequence 이므로 RNN 기반의 딥러닝 모델을 적용한다. 본 논문에서 적용한 딥러닝 모델의 인간-로봇 상호작용을 위한 로봇 지능에의 효용성을 검증하기 위해 실시간 로봇 실험을 수행한다. 로봇 실험에서는 robot operation system (ROS) 기반의 로봇 모바일 플랫폼인 Turtlebot3를 이용하여 로봇 비전으로 인지된 사람의 행동 혹은 주변 상황을 자연어 설명문으로 생성하고, 합성음으로 발화하여 효과적인 인간-로봇 상호작용을 확인한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문에서 적용한 자연어 설명문과 합성음을 생성하는 딥러닝 모델에 대해 설명한다. 3장에서는 모델 학습에 사용한 데이터와 실험 환경, 실험 결과를 설명하고 그것에 대해 고찰한다. 마지막으로 4장에서는 본 논문의 결론을 맺는다.

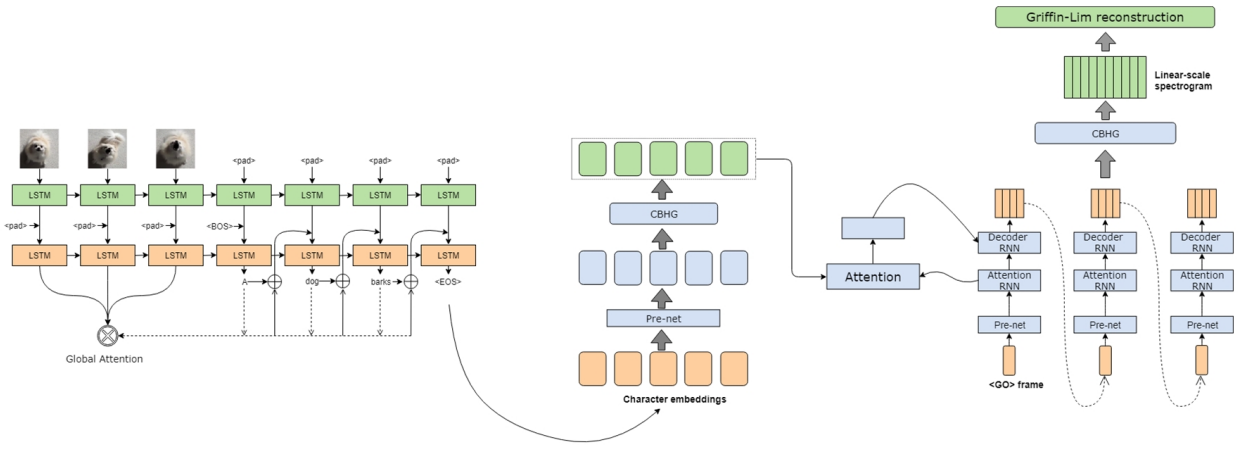
2. 자연어 설명문과 합성음을 생성하는 딥러닝 모델

2.1 전체 모델 구조

본 논문에서 적용한 딥러닝 모델은 [Fig. 1]과 같다. 로봇의 시야에서 상황을 인식해 자연어 문장을 생성하고 생성된 문장을 기반으로 상황을 설명하는 스펙트로그램을 합성한다. 문장과 음성을 생성하는 두 가지 모델 모두 content 기반의 attention 매커니즘을 기반으로 한 모델을 적용한다[12,13].

2.2 로봇 비전 비디오 자연어 설명문 생성 모델

본 논문에서는 모델의 입력인 로봇 비전 비디오와 출력인 자연어 설명문이 모두 시간 흐름을 포함한 sequence 임에 착안하여 Sequence to Sequence - Video to Text (S2VT) 모델을 적용한다[14]. S2VT 는 시간적인 순서에 따른 이미지 프레임의 흐름과 단어의 흐름이 대응되도록 학습하는 RNN 기반 모델이다. 모델의 입력인 영상 프레임의 흐름과 출력인 단어의 흐름은 가변적인 길이를 가지기 때문에 Sequence to Sequence (Seq2Seq) 모델을 사용한다[15,16]. 또한, 본 논문에서는 단어 생성에 이미지 프레임의 어느 부분이 더 중요한지를 반영하기 위해 S2VT 모델에 추가적으로 global attention을 접목하여 문장을 생성한다. Attention을 사용한 이유에는 두가지가 있다. 첫째는 사람이 비디오를 볼 때 모든 장면을 자세히 보는 것이 아닌 특정 부분을 중점을 두어 보기 때문에 영상 부분에 attention을 적용한다. 둘째는 생성될 단어의 앞 부분의 몇 단어들이 문장 생성에 중요하게 영향을 미칠 수 있기 때문에 자연어 부분에도 attention을 적용한다. Attention 모델은 출력과 관련 있는 비디오 정보인 context vector를 이용해 더 자연스러운



[Fig. 1] Overall structure which incorporates video description and speech synthesis model.

문장을 만들 수 있다. Seq2Seq 모델에서 attention을 적용하기 위해서는 인코딩 단계의 은닉 상태를 사용한다. 부분적인 프레임이 특정 위치에 있는 단어 생성에 영향을 미치기도 하지만 일반적으로 전체적인 영상의 은닉 상태를 고려한 attention이 전반적인 문장 생성에 더 많은 기여를 하기 때문에 local attention이 아닌 global attention을 적용한다.

본 논문에서 적용한 global attention을 적용한 S2VT 모델에 대한 자세한 설명은 다음과 같다. S2VT 모델은 비디오 입력 처리와 문장 생성 처리를 위해 두 개의 LSTM[17] 층으로 구성되어 있다. 두 개의 LSTM 층 중, 첫번째 LSTM 층의 은닉표현이 두번째 LSTM에 입력으로 주어지는데, 첫번째 LSTM층은 프레임의 시각적인 흐름을 모델링하는데 사용되고, 그 다음 층은 출력인 단어의 흐름을 모델링하는데 사용된다. 인코딩 단계에서 첫번째 LSTM 층은 CNN을 거친 이미지 프레임을 받아 시간의 흐름에 따라 처리한다. [Fig. 1]에서 볼 수 있듯이 디코딩 단계에서 두번째 LSTM 층은 첫번째 LSTM층의 출력과 인코딩 단계에서의 LSTM 층의 모든 은닉상태를 고려한 global attention으로 생성된 context vector를 동시에 입력으로 받아 단어 기준으로 문장을 생성한다. 각 시간 단계 t에 따라 현재 타겟의 은닉상태는 $\boldsymbol{h}_t$이고 모든 인코딩 단계와 타겟의 이전 단계의 source인 s의 은닉상태는 $\overline{\boldsymbol{h}_s}$이다. $\boldsymbol{h}_t$와 $\overline{\boldsymbol{h}_s}$을 기준으로 영상 입력의 길이와 같은 alignment vector $\mathbf{a}_t$을 추론한다. 최종적으로 global context vector는 $\mathbf{a}_t$의 가중합의 평균으로 계산된다.

$$\mathbf{a}_t(\mathrm{s}) = \mathrm{align}\left(\boldsymbol{h}_t, \overline{\boldsymbol{h}_s}\right) = \frac{\exp\left(\mathrm{score}\left(h_t, \overline{h_s}\right)\right)}{\Sigma_{s'}\exp\left(\mathrm{score}\left(h_t, \overline{h_{s'}}\right)\right)} \quad (1)$$

위 식에서 alignment vector를 구하기 위한 score는 식(2)와 같이 $\overline{\boldsymbol{h}_s}$, $\boldsymbol{h}_t^{\mathrm{T}}$에 학습되는 가중치인 $\boldsymbol{W_a}$을 곱해 content 기반의 general 형태로 계산한다[13].

$$\mathrm{score}(\boldsymbol{h}_t, \overline{\boldsymbol{h}_s}) = \boldsymbol{h}_t^T \boldsymbol{W_a} \overline{\boldsymbol{h}_s} \quad (2)$$

S2VT는 비디오의 시각적인 특징을 추출하여 가변적인 입력 프레임을 처리하고 자연스러운 문장을 생성하도록 학습한다. 기계번역과 유사하게 주어진 입력에 대한 조건부 확률로 단어를 추정하고, 인코딩과 디코딩 단계에서 단일 LSTM을 사용하여 파라미터를 공유한다.

두번째 LSTM층의 출력은 softmax 함수를 이용해 최대확률로써 단어를 선택하고, <EOS> 태그가 나타날 때까지 단어를 만들어 낸다. 예측된 문장과 정답 문장을 이용해 log-likelihood로 손실함수를 계산해 Adam Optimizer[18]를 사용해 전체 훈련 데이터셋에 대해 최적화한다.

2.3 로봇 비전 비디오 자연어 설명문 생성 모델의 입출력 데이터

모델의 입력인 각각의 영상 프레임은 256x256 크기로 조정되고 무작위로 227x227크기로 자른 후, CNN 의 입력으로 들어간다. CNN은 ImageNet dataset 중 ILSVRC-2012 물체 분류의 1.2M 이미지에 대해 미리 학습한 VGG-16 모델을 이용한다. VGG-16의 마지막 fully connected (FC) 분류 층을 제거하고 500 차원으로 대응되도록 새로운 임베딩 층을 추가한다. 500차원의 CNN출력을 S2VT의 첫번째 LSTM층의 입력으로 사용하고, 새롭게 추가된 임베딩 층의 가중치들은 S2VT와 동시에 학습된다.

모델의 출력인 단어는 one-hot 벡터 인코딩을 사용하여 표현된다. one-hot 벡터는 단어의 총 개수가 벡터의 크기이며, 해당되는 단어만 1이고 나머지는 0인 벡터를 말한다. 이는 위에서 언급한 영상 프레임과 동일하게 임베딩 층 학습을 통해 500차원으로 임베딩한다. 첫번째 LSTM의 출력과 임베딩된 단어 벡터가 합쳐져 두번째 LSTM의 입력으로 들어간다.

2.4 합성음 생성 모델

본 논문에서는 모델의 입력인 자연어 설명문과 출력인 음성이 모두 시간 흐름을 포함한 sequence 임에 착안하여 tacotron 모델을 적용한다[19]. Tacotron은 글을 음성으로 합성하는 end-to-end generative text-to-speech 모델로서 문자열을 입력으로 하고 일치하는 스펙트로그램을 출력으로 하는 통합적인 모델이다. Tacotron은 encoder와 attention 기반의 decoder의 Seq2Seq와 post processing 네트워크로 이루어져 있다[12,20].

인코더의 목적은 문장의 흐름을 파악하는 것이다. tacotron의 인코더는 S2VT에서 나온 문장을 토대로 문자열을 받는다. 인코더의 각 문자는 one-hot 벡터에서 continuous 벡터로 임베딩된다. 그리고 각 임베딩된 문자는 pre-net이라는 비선형변환을 하게 된다. pre-net은 FC 층과 rectified linear unit (ReLU) 활성화 함수, 드랍아웃[21]으로 구성되어 있다. 드랍아웃은 정규화 기법으로 훈련 중 무작위로 뉴런을 선택한다. 뉴런은 학습 중에 서로 연관되어 과대적합이 발생할 수 있기 때문에 드랍아웃을 통해 이를 막고 좋은 특징들이 잘 학습되게 한다. 두 번째 FC layer는 첫 번째 층보다 2배 적은 뉴런을 사용하기 때문에 학습의 수렴과 일반화를 향상시킨다. pre-net 출력은

CBHG[22] 모듈에 의해 최종적인 encoder 출력이 된다. 이때 CBHG의 의미는 1-D convolution bank (CB), Highway network (H)[23] and a bidirectional gated recurrent unit (G)[24]이다. 1-D convolution 필터의 K 층들은 convolution bank를 형성하는데 사용된다. K는 너비가 k인 C_k필터를 포함한다. 이 구조를 통해서 unigram, bigram, k-gram을 모델링할 수 있다. 이후 CNN에서 일반적으로 적용되는 max-pooling이 사용된다. max pooling은 locally invariant를 학습하는 강점이 있다. locally invariant는 local gram에서 변하지 않는 값들을 뜻하며 정성적으로 봤을 때 문맥이 달라져도 변하지 않는 문자들을 말한다. 그리고 시간 축을 유지하기 위해 stride를 1로 사용한다. CBHG의 모든 convolutional 층에서 입력 데이터의 최적 스케일과 평균을 학습하여 신경망의 안정성과 성능을 향상시키기 위해 batch normalization[25]을 사용한다. Max-pooling 층 이후에 두 개의 1-D convolutional 층이 사용된다. Residual 연결[26]은 처음의 입력과 두 번째 convolutional 층의 출력을 합치는 역할을 한다. Residual 연결을 통해 모델의 층이 깊을 때도 gradient가 잘 전달될 수 있다. Highway 네트워크 또한 층을 통과할 때 수행되어야 하는 선형 연산과 활성화 연산을 하지 않고 빠르게 지나가기만 하는 우회로를 만들어 빠른 학습을 도와준다. 이러한 CBHG 기반의 encoder는 과대적합을 줄여줄 뿐만 아니라 일반적인 다층 RNN보다 잘못 학습된 합성음이 발음되는 것을 적어지게 한다.

디코더에서는 content기반의 attention 매커니즘을 이용한다. Context 벡터와 attention RNN cell 출력을 합쳐서 디코더 RNN의 입력으로 넣는다. 학습의 속도를 높이기 위해 gated recurrent unit (GRU)를 vertical residual connection[27]에 사용한다. 디코더 타겟, 즉 Seq2Seq 타겟은 80-band mel-scale 스펙트로그램으로 한다. 샘플링된 스펙트로그램은 가공되지 않은 스펙트로그램에 비해 정보량이 적지만, 80-band mel-scale 스펙트로그램은 음성의 리듬과 규칙들의 정보를 충분히 포함한다. 디코더 타겟을 예측하기 위해서 FC 층을 사용한다. 여기서 중요한 점은 각 디코더 단계에서 여러 개의 겹치지 않는 프레임을 예측한다는 것이다. 한번에 r개의 프레임을 예측하면 디코더 단계가 r로 나눈 만큼 줄어들어 모델 크기, 훈련 및 추론 시간도 줄어든다[28]. 각 디코더 프레임은 옆에 있는 프레임과 연관 되어있고 각 문자들은 여러 개의 프레임에 포함되어 있으므로 attention 메커니즘을 사용할 시에 훨씬 빠르게 학습을 할 수 있다. 첫 번째 디코더 단계에서는 <GO> 프레임, 즉 0으로 채워진 프레임을 사용한다. 추론단계에서는 전 단계에서 r번째로 나온 디코더 출력을 디코더 입력으로 사용한다. 학습 단계에서는 항상 r번째의 실제 정답 값을 디코더로 넣어준다. 디코더 입력 프레임은 인코더와 마찬가지로 pre-net으로 전달된다.

후처리는 학습 타겟인 mel-스펙트로그램을 최종 타겟인 음성으로 변환하는 과정이다. CBHG 모듈은 decoder 단계 이후에도 사용되는데 bidirectional GRU로 인해 앞뒤 특징들을 추출할 뿐만 아니라 예측한 프레임의 오류를 없애는 역할도 한다. 최종적으로 CBHG 모듈을 통해 스펙트로그램이 예측된다. 스펙트로그램은 음성을 표현하기에 좋은 방법이지만 위상에 대한 정보가 부족하다. 때문에 스펙트로그램의 위상을 추정할 수 있는 신호처리 알고리즘인 Griffin-Lim[29,30]을 사용하여 음성 파형을 예측할 수 있다. STFT 크기와 생성된 스펙트로그램이 일치하는 파형을 찾을 수 있도록 반복적으로 수행한다.

3. 실 험

3.1 학습 데이터 셋, 하이퍼파라미터 및 로컬 서버 환경

S2VT 학습에는 Microsoft Video Description Corpus (MSVD)[31]를 사용했다. MSVD는 하나의 행동을 설명하는 짧은 유튜브 클립을 모아둔 데이터이다. MSVD의 원래 데이터 셋에는 여러 언어로 설명되어 있지만 본 논문에서는 영어 데이터만 사용한다. 또한 문장을 모두 소문자로 바꾸고 단어를 토큰화하고 구두점을 삭제하는 전처리 과정을 거친다. 프레임은 10프레임에 하나씩 샘플링 한다.

Tacotron 학습에는 LJ Speech 데이터셋을 사용했다. 이 데이터는 공개적으로 사용 가능하여 최근 text to speech (TTS) 모델에서 벤치마크로써 널리 사용되고 있다. 이 데이터는 한 명의 성우가 녹음한 13,000개의 짧은 오디오 클립으로 구성되어 있으며 클립의 길이는 1초에서 10초 사이로 총 길이는 약 24시간이다.

각 모델 학습에 사용된 hyper-parameter는 [Table 1]과 [Table 2]와 같다. 로컬 서버의 GPU는 NVIDIA GTX 1080Ti, CPU는 AMD Ryzen 7 1700X 그리고 RAM은 삼성 DDR4 16GB*2을 사용했다. 학습에 걸린 시간은 배치크기가 32일 때 한 스텝에 약 1.1초가 걸려 총 397K 스텝을 진행한 결과 약 121시간으로 약 5일동안 학습하였다.

3.2 동영상에 대한 음성 평가 및 상황 설명 평가

로봇 실험을 수행하기 앞서, 3.1절 학습 데이터 셋과 동일한 MSVD 동영상에 대한 자연어 설명문 생성 및 합성음 생성에

[Table 1] S2VT model hyperparameter

Image embedding dimension	4096
LSTM dimension	1000
The Number of Encoding Time Step	80
The Number of Decoding Time Step	20

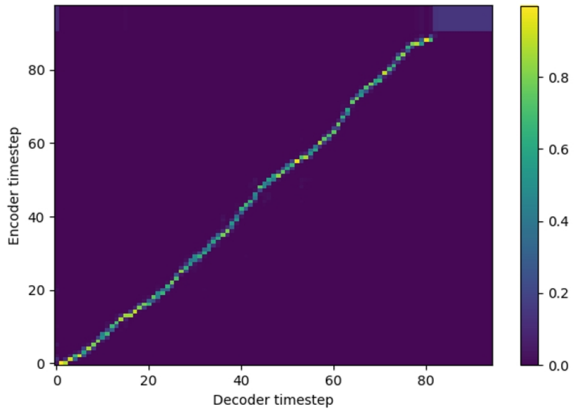
[Table 2] Tacotron model hyperparameter

Spectral analysis	Pre-emphasis:0.97, Frame length: 50ms, Overlap length: 12.5ms, Window: Hann
Character embedding dimension	256
Encoder CBHG	Conv1D bank : K=16, conv-k-128-ReLU Max pooling : stride=1, width=2 Conv1D projections : conv-3-128-ReLU → conv-3-128-Linear Highway network : 4 layers of FC-128-ReLU Bidirectional GRU : 128 cells
Encoder pre-net	FC-256-ReLU → Dropout (0.5) → FC-128-ReLU → Dropout (0.5)
Decoder pre-net	FC-256-ReLU → Dropout (0.5) → FC-128-ReLU → Dropout (0.5)
Decoder RNN	2-layer residual GRU (256 cells)
Attention RNN	1-layer GRU (256 cells)
Post-processing net CBHG	Conv1D bank : K=8, conv-k-128-ReLU Max pooling : stride=1, width=2 Conv1D projections : conv-3-256-ReLU → conv-3-80-Linear Highway network : 4 layers of FC-128-ReLU Bidirectional GRU : 128 cells
Reduction factor (r)	2

대한 모델 평가를 진행했다. 음성을 생성하는 모델이기 때문에 객관적인 지표를 통한 성능분석을 하기가 어렵다. 따라서 attention alignment를 통한 시각적인 비교와 주관적 음질 평가로 대신한다.

[Fig 2]에 나타나 있는 attention alignment은 학습에 적용된 attention이 인코더와 디코더 단계에서 서로 다른 중요도를 부여해 가장 적절한 인코더 스텝의 가중치를 높인 것을 시각화한 것으로 직선에 가까울수록 좋은 그래프이다.

주관적 음질 평가는 영상의 소리가 사람과 얼마나 비슷한



[Fig. 2] Attention alignment plot

[Table 3] Voice quality and situation explanation evaluation scores

Data Set	Subjective speech evaluation	Context explanation evaluation
MSVD & LJ Speech	3.07 ± 0.73	2.84 ± 0.67

[Table 4] Situation explanation evaluation by the NLP metric

METEOR	0.289
BLEU_1	0.678
BLEU_2	0.558
BLEU_3	0.462
BLEU_4	0.365
ROUGE_L	0.646
CIDEr	0.554

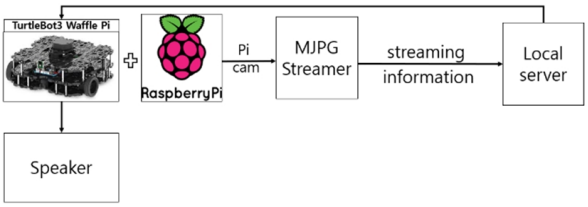
지 자연스러움을 1-5점으로 절대 평가하는 방법이다. 이 때, 사람과 비슷하다는 척도는 억양, 발음, 잡음 등을 고려하였다. 마찬가지로 상황 설명 평가는 생성된 합성음이 영상을 잘 설명하는지 1-5점으로 구분하여 평가하는 방법이다. 각 설문지는 MSVD 데이터셋 중 학습하지 않은 영상을 무작위로 10~15개를 선정하고 추론 결과인 합성음을 영상에 추가하여 만들어졌다. 설문조사는 청력이 정상인 성인 54명을 대상으로 이루어졌다. 평가결과는 [Table 3]과 같다.

또한 실험의 신뢰성을 보다 향상시키기 위해 MSVD 데이터의 10%를 테스트 셋으로 만들어 상황 설명 평가를 추가적으로 진행하였다. 실험은 주로 기계 번역에서 사용되는 자연어 평가 지표로 진행하였고 결과는 [Table 4]이다. 자연어 생성을 평가하기 위한 지표에는 여러가지가 있지만 가장 많이 사용되는 BLEU, METEOR, ROUGE_L 그리고 CIDEr[32-35] 점수를 사용하였다. BLEU[32] 점수는 각각의 생성된 문장을 정답 문장과 비교하여 계산하고 최종적으로는 전체 말뭉치에 평균을 낸 값이다. BLEU_n은 unigram부터 n-gram까지 점수를 구해 평균을 구한 것이고 1에 가까울수록 정답 문장과 비슷하다. BLUE는 예측한 문장의 정밀도에 가중치가 있는 평가지표라면 ROUGE-L[33]는 재현율에 가중치가 있다. METEOR[34]는 정밀도와 재현율의 조화평균으로 구성된 평가 척도이다. 정답 문장의 토큰과 생성된 문장의 토큰이 얼마만큼 일치하는 지와 문장 구조도 함께 고려하므로 BLEU보다 문장 단위에서 측정하는데 적합하다. CIDEr[35]을 구할 때는 어간을 기준으로 변형된 단어들을 기본형으로 만들어주는 과정을 거친 후, 점수 계산을 한다. 사진에 대한 자연어 설명을 목적으로 만들어진 지표이기 때문에, 여러 개의 이미지를 보고 문장을 생성하는 상황을 가정한다. TF-IDF방법처럼 특정 영상에서만 나온 단어는 긍정적인 점수를, 대부분의 영상의 문장이 공통적으로 가지고 있는 단어들은 부정적인 점수를 주는 방식이다.

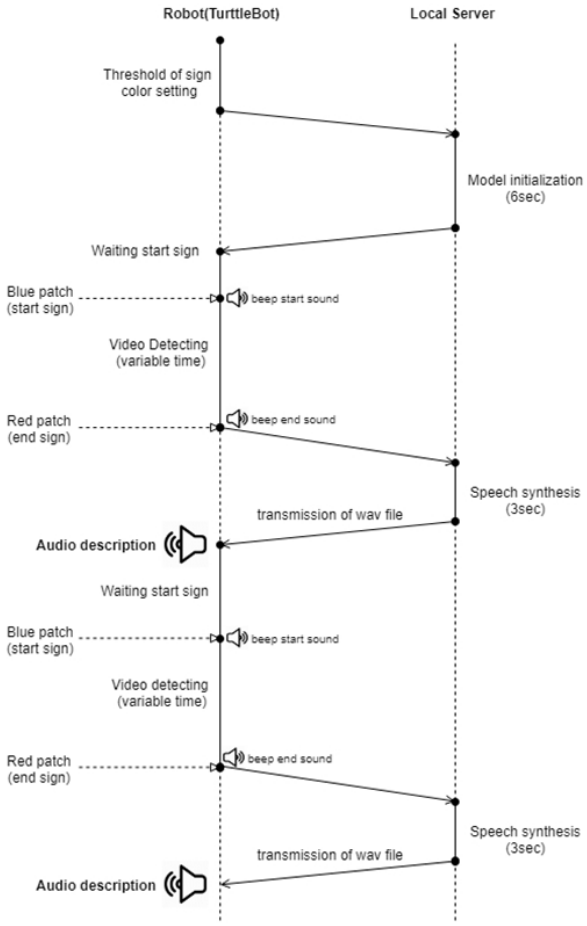
3.3 실시간 로봇 실험 환경

로봇을 이용한 실시간 자연어 설명문 생성 및 발화 실험을 위해 Turtlebot3 waffle Pi 모델에 OpenCR Board 와 Raspberry Pi3을 장착하여 사용했다. Raspberry Pi3 에는 Raspbian OS에 ROS Kinetic을 설치하여 ROBOTICS RC100을 이용하여 원격 제어 했다. [Fig. 3]처럼 OpenCV를 이용해 로봇 비전 동영상을 로컬 서버로 전송하고, 서버는 영상을 받아 딥러닝 모델 추론 후 결과를 Raspberry Pi3 로 보낸다.

로봇실험의 순서흐름도는 [Fig. 4]와 같다. 특정 색이 인식



[Fig. 3] Robot experiment environment and structure



[Fig. 4] Sequence flow diagram of robot experiment

되는 것으로 비디오의 시작과 끝을 판별하기 때문에 환경에 따라 색 임계 값을 조정해준다. 그 다음 딥러닝 모델의 그래프를 로컬서버에 올리는데 6초가 걸린다. 파란색 패치가 인식되면 받기 시작 소리와 함께 비디오 입력을 시작하고 빨간색 패치가 인식되면 정지 소리를 내고 비디오 입력을 멈추고 로컬 서버로 입력된 비디오를 넘겨준다. 로봇이 오랫동안 상황에 놓여 있다면 그만큼의 영상 입력 시간이 소요되지만, 비디오를 총 80프레임으로 샘플링하기 때문에 음성 합성에 필요한 시간은 매 실험이 동일하다. 로컬 서버가 비디오를 받아 음성 합성까지 3초가 소요되고 생성된 오디오를 로봇으로 전달해 재생한다.

3.4 실시간 로봇 실험 결과

실시간 로봇 실험 결과를 3.2 장의 설문조사와 같은 방식으로 32명의 사람을 대상으로 평가하였다. 평가결과는 [Table 5]와 같다. 설문조사는 문장이 잘 나온 상황과 그렇지 않은 상황을 섞어서 진행하였기 때문에 상황 설명평가 결과의 표준편차가 [Table 3] 보다 높은 것을 확인할 수 있다. 인간-로봇 상호작용 상황에서 상대방이 어떤 상황인지를 로봇이 인식하고, 이를 자연어 설명문을 이용한 음성으로 설명하는 것에 본 논문이 적용한 방식이 적절함을 보여준다.

[Fig. 5]는 실시간 로봇 실험의 대표 장면을 보인다. [Fig. 5]의 각 이미지는 실시간 로봇 실험의 촬영 장면과 로봇 비전 화면이고, 아래의 캡션은 딥러닝 모델이 생성한 자연어 설명문이다. 그림의 1행은 상황에 알맞게 생성

된 자연어 설명문이고 2행은 그렇지 못한 설명문이다.

첫번째 행에서는 로봇이 비전을 통해 인식한 상대방의 상황에 대해 적절한 자연어 설명문을 생성한 것을 알 수 있다. 하지만 두번째 행에서는 상황을 잘 인식하지 못하여 잘못된 문장을 생성하였다. 첫번째 영상은 주변에 노트북, 물통, 휴대폰 다양한 물체가 있고 attention이 휴대폰으로 잡혀서 잘못된 문장이 생성되었다. 두번째는 물체보다는 기타를 치는 행동에 초점이 맞춰져 빗자루를 들고 있지만 기타를 친다고 생성되었다. 마지막 비디오는 칠판에 행동하는 것은 맞췄지만 남자와 여자가 동시에 잡히면서 잘못된 문장을 만들어 냈다.

실시간 로봇 실험에서 로봇 시야에 여러 물체가 한번에 보

[Table 5] Voice quality and situation explanation evaluation scores on robot experiment

Subjective speech evaluation	Context explanation evaluation
3.37 ± 0.97	2.86 ± 1.43

[Fig. 5] Audio descriptions resulted in robot environment

이는 것보다 다양한 물체가 보이는 것보다 특정 물체를 직접적으로 보여주는 영상이 문장 생성에 더 좋은 결과를 가져왔다. 한 프레임에서 가장 많은 부분을 차지하고 있는 사물에 특징점을 두는 CNN에 기반을 두고 있기 때문이다. 한 화면에서 여러 물체를 감지하는 object detection을 이용해 작은 물체들에 대해서도 문장을 생성한다면 자연어 설명문 생성 성능을 개선시킬 수 있을 것이다.

4. 결 론

본 논문에서는 효과적인 인간-로봇 상호작용을 위해 딥러닝 기반 로봇 비전 자연어 설명문 생성 및 발화 모델을 학습시키고 로봇에 적용하였다. 로봇이 인식한 상황에 대하여 자연어를 이용해 발화를 함으로써 상호작용 상대방이 로봇의 인식 결과를 가장 효율적으로 알 수 있어, 효과적인 인간-로봇 상호작용이 가능함을 보였다. 추후 연구과제로써 본 논문에서 사용한 파이프라인 형태의 딥러닝 모델이 아닌 end-to-end 딥러닝 모델 연구가 필요하다. 또한, 상황설명을 위한 자연어 설명문 생성에 다중 물체 감지 딥러닝 모델을 활용하여 자연어 설명문 생성 성능을 발전시킬 것이다.

References

[1] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, “Long-term recurrent convolutional networks for visual recognition and description,” *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, DOI: 10.1109/CVPR.2015.7298878.

[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, DOI: 10.1109/CVPR.2015.7298935.

[3] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, “Describing videos by exploiting temporal structure,” *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, DOI: 10.1109/ICCV.2015.512.

[4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L, Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *arXiv:1409.0575 [cs.CV]*, 2014.

[5] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv:1409.1556 [cs.CV]*, 2014.

[6] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu., “Video Paragraph Captioning Using Hierarchical Recurrent Neural Networks,” *arXiv:1510.07712 [cs.CV]*, 2016

[7] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen., “Video captioning with attention-based LSTM and semantic consistency,” *IEEE Transactions on Multimedia*, vol. 9, no. 9, pp. 2045-2055, Sept., 2017.

[8] K. Tokuda, H. Zen, and A. Black, “An HMM-based speech synthesis system applied to English,” *2002 IEEE Workshop on Speech Synthesis*, Santa Monica, CA, USA, 2002.

[9] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu,

"WaveNet: A generative model for raw audio," *arXiv:1609.03499 [cs.SD],* 2016.

[10] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2Wav: End-to-end speech synthesis," *ICLR 2017*, 2017.

[11] S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, J. Raiman, S. Sengupta, and M. Shoeybi, "Deep voice: Real-time neural text-to-speech," *arXiv:1702.07825 [cs.CL],* 2017.

[12] O. Vinyals, Ł.Kaiser, T. Koo, S. Petrov, I. Sutskever, and G.Hinton, "Grammar as a foreign language," *arXiv:1412.7449 [cs.CL]*, 2015.

[13] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv:1508.04025 [cs.CL]*, 2015.

[14] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to Sequence -- Video to Text," *2015 IEEE International Conference on Computer Vision (ICCV),* 2015.

[15] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoderdecoder approaches," *arXiv:1409.1259 [cs.CL]*, 2014.

[16] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *arXiv:1409.3215 [cs.CL]*, 2014.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation,* vol. 9, no. 8, Nov., 1997.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980 [cs.LG],* 2014.

[19] Y. Wang, R.J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards End-to-End Speech Synthesis," *Interspeech 2017*, 2017, DOI: 10.21437/Interspeech.2017-1452.

[20] D. Bahdanau, K.H. Cho, and Y. Bengio "Neural machine translation by jointly learning to align and translate," *arXiv:1409.0473 [cs.CL],* 2014.

[21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*., vol. 15, no. 1, pp. 1929-1958, 2014.

[22] J. Lee, K. Cho, and T. Hofmann, "Fully character-level neural machine translation without explicit segmentation," *Transactions of the Association for Computational Linguistics,* vol. 5, pp. 365-378, 2017.

[23] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *arXiv:1505.00387 [cs.LG]*, 2015.

[24] J. Chung, C. Gulcehre, K.H. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv:1412.3555 [cs.NE]*, 2014.

[25] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv:1502.03167 [cs.LG],* 2015.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp.770-778, 2016.

[27] Y. Wu, M. Schuster, Z. Chen, Q. V Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Ł. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv:1609.08144 [cs.CL]*, 2016.

[28] H. Zen, Y. Agiomyrgiannakis, N. Egberts, F. Henderson, and P. Szczepaniak, "Fast, compact, and high-quality LSTM-RNN based statistical parametric speech synthesizers for mobile devices," *Interspeech,* 2016, DOI: 10.21437/Interspeech.2016-522.

[29] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *ICASSP '83. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236-243, 1984.

[30] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," *arXiv:1603.04467 [cs.DC],* 2016.

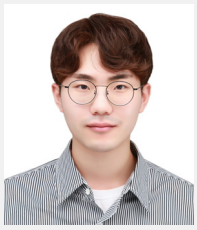
[31] D. L. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," *49th Annual Meeting of the Association for Computational Linguistics,* pp. 190-200, 2011.

[32] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," *40th Annual Meeting on Association for Computational Linguistics*, pp. 311-318, Philadelphia, PA, USA, 2002.

[33] C.-Y. Lin. "ROUGE: A Package for Automatic Evaluation of Summaries," *ACL-04 Workshop*, pp. 74-81, 2004.

[34] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," *Ninth Workshop on Statistical Machine Translation*, pp. 376-380, Baltimore, MD, USA, 2014.

[35] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus- based image description evaluation," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp.4566-4575, 2015.



박 동 건

2015~현재 서울과학기술대학교 컴퓨터공학과

관심분야: 물체 인식, 자연어 처리, text to speech

배 진 우

2014~현재 서울과학기술대학교 전기정보공학과

관심분야: 기계학습, 딥러닝

강 경 민

2016~현재 서울과학기술대학교 컴퓨터공학과

관심분야: 기계학습, 딥러닝

한 지 형

2008 한국과학기술원 전기 및 전자공학과 (학사)

2015 한국과학기술원 전기 및 전자공학과 (박사)

2015~2017 한국전자통신연구원 선임연구원

2017~현재 서울과학기술대학교 컴퓨터공학과 조교수

관심분야: 인간-로봇 상호작용, 인간 의도 파악, 지능형 로봇, 기계학습, 딥러닝