ICLAL: 인 컨텍스트 러닝 기반 오디오-언어 멀티 모달 딥러닝 모델

박준영[1], 여진영[2], 이고은[3], 최창환[4] 최상일[5]
[1]단국대학교 인공지능융합학과 석사과정
[2]연세대학교 인공지능학과 조교수
[3]단국대학교 컴퓨터학과 박사과정
[4]단국대학교 컴퓨터공학과 학사과정
[5]단국대학교 컴퓨터공학과 교수

j72220216@dankook.ac.kr, jinyeo@yonsei.ac.kr, ge971010@naver.com, ho03206@naver.com, choisi@naver.com

ICLAL: In-Context Learning-Based Audio-Language Multi-Modal Deep Learning Models

Jun Yeong Park[1], Jinyoung Yeo[2], Go-Eun Lee[3], Chang Hwan Choi[4], Sang-Il Choi[5]
[1]Dept. of AI-Based Convergence, Dankook University
[2]Dept. of Artificial Intelligence, Yonsei University
[3,4,5]Dept. of Computer Engineering, Dankook University

요 약

본 연구는 인 컨택스트 러닝 (In-Context Learning)을 오디오-언어 작업에 적용하기 위한 멀티모달 (Multi-Modal) 딥러닝 모델을 다룬다. 해당 모델을 통해 학습 단계에서 오디오와 텍스트의 소통 가능한 형태의 표현 (Representation)을 학습하고 여러가지 오디오-텍스트 작업을 수행할 수 있는 멀티모달 딥러닝 모델을 개발하는 것이 본 연구의 목적이다. 모델은 오디오 인코더와 언어 인코더가 연결된 구조를 가지고 있으며, 언어 모델은 6.7B, 30B 의 파라미터 수를 가진 자동회귀 (Autoregressive) 대형 언어 모델 (Large Language Model)을 사용한다 오디오 인코더는 자기지도학습 (Self-Supervised Learning)을 기반으로 사전학습 된 오디오 특징 추출 모델이다. 언어모델이 상대적으로 대용량이기 언어모델의 파라미터를 고정하고 오디오 인코더의 파라미터만 업데이트하는 프로즌 (Frozen) 방법으로 학습한다. 학습을 위한 과제는 음성인식 (Automatic Speech Recognition)과 요약 (Abstractive Summarization) 이다. 학습을 마친 후 질의응답 (Question Answering) 작업으로 테스트를 진행했다. 그 결과, 정답 문장을 생성하기 위해서는 추가적인 학습이 필요한 것으로 보였으나, 음성인식으로 사전학습 한 모델의 경우 정답과 유사한 키워드를 사용하는 문법적으로 올바른 문장을 생성함을 확인했다.

1. 서론

언어모델을 사용하여 텍스트로부터 언어적 맥락과 일반적 특징을 추출한 후 이를 다양한 분야에서 적용하는 자연어 기법[1, 2]은 전통적으로 높은 성능을 보장하던 방법이었다. GPT-3[3]의 등장은 모델의 크기를 175B 수준까지 증가시켰을 때 자동회귀 (Autoregressive) 언어모델은 ICL (In-Context Learning)으로 문제를 해결하는 능력이 생김을 시사했다. ICL 은 언어모델에게 특정한 과제(Task)를 별도로 지도학습(Supervised Learning)하지 않더라도 과제와 관련된 예시 (Demonstration)나 지시문 (Instruction)을 0 개~n 개까지 제시하면 주어진 예시에 따라 모델이 과제에 대한 해결방법을 문맥적으로 파악한다는 것이다. 이 때 주어진 과제 해결을 위한 예시의 개수에 따라 zero-shot (0 개), one-shot (1 개), few-shot (n 개)으로 구별한다.

대형 언어모델이 자연어 처리 과제에서 높은 성능을 보이자 이를 다른 모달리티 (Modality)에 적용하려는 움직임이 나타났다. Frozen[4]은 비전과 언어를 이해하고 연결하는 멀티모달 (Multi-Modal) 멀티태스크 (Multi-Task) 모델로, 텍스트와 비주얼 정보를 동시에 처리하고 이해할 수 있어 이미지에 대한 텍스트 설명을 생성하는 등의 작업을 수행한다. 이 때 대형 모델

을 항상 전부 미세조정 해야 한다면 컴퓨터 자원의 낭비와 높은 계산 시간 복잡도의 문제를 초래할 수 있으므로, 상대적으로 파라미터 양이 많은 언어모델의 파라미터는 고정하고 텍스트 외의 모달 정보를 처리하는 모듈만 파라미터를 업데이트한다. Frozen 은 6.7B 개의 파라미터를 가진 GPT 기반 언어모델과 25M 개의 파라미터를 가진 NF-ResNet[5, 6] 기반 비전 인코더를 사용하여 다양한 비전-언어 과제를 수행했다.

본 연구는 비전-언어 모델로 개발된 Frozen 을 오디오 (Audio) 모달로 이식하는 초기연구를 진행한다. Frozen 과 동일한 학습 방법을 적용하여 GPT 기반 모델의 파라미터를 고정하고 오디오 인코더의 파라미터만 업데이트한다. 오디오 인코더에서 입력으로 들어온 음성 신호를 사전학습 된 대형 언어 모델이 이해 가능한 형태로 인코딩 한 것을 *Audio Prefix* 라고 명명하고 텍스트 임베딩 (Text Embedding)과 함께 사용하여 다양한 멀티모달 과제를 수행하는 것이 본 연구의 목적이다.

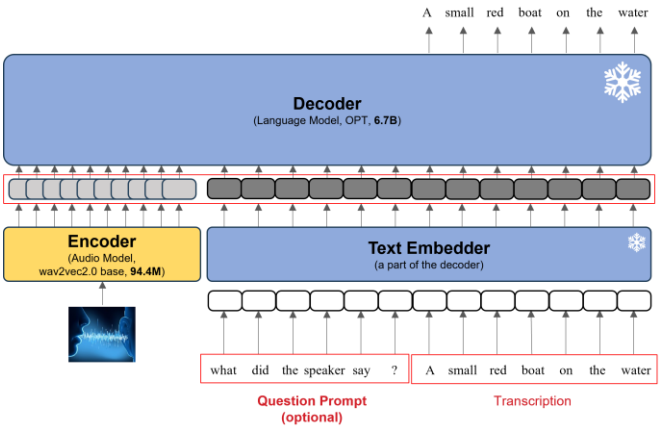
본 논문에서는 해당 구조로 멀티모달 분류 과제에 대해 연구한 WavPrompt[7]를 참고하여 모델의 학습 세부사항을 설정한 후 분류 뿐 만 아니라 생성 과제에서도 Frozen 방법이 적용 되는지 실험을 진행하고 결과를 분석했다.

2. 모델 구조

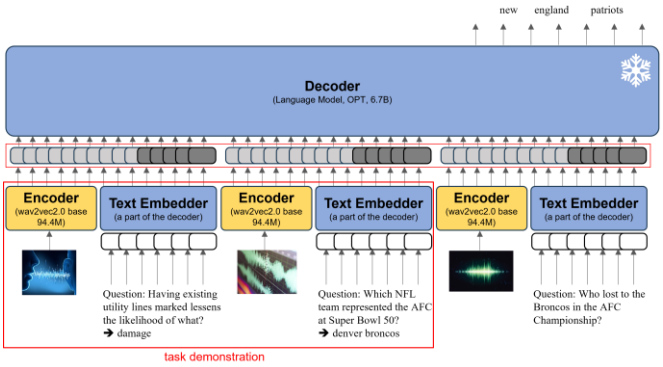
모델의 전체 구조는 두 가지 모듈로 이루어진다. 오디오 모달의 정보를 처리하는 오디오 인코더 (Audio Encoder)와 언어 정보를 처리하는 언어 디코더 (Language Decoder)이다.

오디오 인코더로 사용할 모듈은 Wav2Vec 2.0[8]의 base 모델을 사용했다. 94.4M 개의 가중치를 가지고 있다. Wav2Vec2.0 은 TCN (Temporal Convolution Network)[9]로 오디오의 내재적 특징을 추출하고 트랜스포머 (Transformer)[10]와 마스크 언어모델링 (MLM, Masked language Modeling)[2] 기법을 사용하여 특징을 학습하는 자기지도학습 (Self-Supervised Learning)을 적용했다. 사전학습을 위해 라벨링 되지 않은 53,000 시간의 오디오 데이터를 사용했다. 본 연구에서는 이와 같은 방식으로 사전학습이 완료된 Huggingface 에서 제공하는 'facebook/wav2vec2.0-base'를 사용해 실험을 진행했다.

언어 디코더로는 OPT[11] 6.7B 과 30B 모델을 사용했다. WavPrompt 에서는 GPT-2[12] Medium (355M) 모델을 사용했으나, Frozen 에서는 7B 모델을 사용할 뿐 만 아니라 In-Context Learning 으로 사용할 일반적인 지식 (General Knowledge)를 추출하기 위해서는 더 큰 언어모델이 필요하다. 따라서 모델의 크기가 비교적 다양하고 공개된 사전학습 파라미터가 존재하는 OPT 모델을 디코더로 사용했다.



(그림 1) 학습단계



(그림 2) 추론단계

3. 학습

모델 학습을 위해 ASR (Automatic Speech Recognition, 음성인식)을 위한 데이터셋과 Abstractive Summarization (요약)을 위한 데이터셋 두 가지를 사용했다. 각 데이터셋은 별도의 표기가 없다면 함께 사용되지 않았으며, 별개의 모델을 학습해 어떤 작업하는 것이 더 나은지에 대한 결과를 비교했다.

음성인식 데이터셋은 LibriSpeech[13]의 학습용 'train-clean-360'과 테스트용 4 종 데이터셋[표 1]을 사용했다. 학습 데이터는 총 360 시간의 노이즈가 적은 정제된 문장 읽기 음성과 대본 (Transcription)으로 구성돼 있다.

요약 데이터셋은 AMI[14] 데이터셋으로 회의 내용을 사람이 직접 요약한 것이다. 총 100 시간의 회의 음성과 사람이 직접 요약한 텍스트로 이루어졌다. 음성인식 데이터셋보다 오디오 신호의 길이가 더 긴 것이 특징이다.

학습 (그림 1)은 다음과 같은 절차로 이루어진다. 원본 waveform 오디오를 오디오 인코더에 넣고 결과로 *Audio Prefix* 를 받는다. 음성인식 데이터셋으로 사전학습 한다면 대본 (Transcription) 텍스트, 요약 데이터셋으로 사전학습 한다면 요약 텍스트를 언어모델의

임베딩 모듈에 넣어 텍스트 임베딩을 얻는다. *Audio Prefix* 와 텍스트 임베딩을 연결 (Concatenate)해 하나의 프롬프트로 만들고 디코더에 넣는다. 정답 (Groundtruth)으로는 각 모델의 입력 텍스트를 사용한다. 추론 때는 음성과 지시문 (Task Instruction)과 예시 (Demonstration)만 넣은 뒤 결과 텍스트를 생성한다. (그림 2)

4. 결과

	ICLAL	1	2	3	4
dev-clean	0.0521	0.0211	0.0296	0.0380	0.0292
test-clean	0.0345	0.0215	0.0311	0.0392	0.0324
dev-other	0.1968	0.0432	0.0720	0.1046	0.0491
test-other	0.1133	0.0451	0.0739	0.1077	0.0527

[표 1] 음성인식 테스트 결과
1. HuBERT-L/ 2.Wav2Vec2.0-L
3.Wav2Vec-base/ 4. Wav2Vec2.0-Conformerr-L

Context	before the foundation can be done contractors are typically required to verify and have existing utility lines mark either by the utilities themselves or through a company specializing in such services. this lessens the likelihood of damage to the existing electrical water sewage phone and cable facilities which could cause outages and potentially hazardous situations. … (중략)
Question	Having existing utility lines marked lessens the likelihood of what?
Target	damage
Predict	before the foundation can be done contractors are typically required to verify and have existing utility lines marked either by the utilities themselves or through a company specialising in such services. this lessens the likelihood of damage to the existing electrical water suidgee foam and cable facilities which could cause outages and potentially hazardous situations. … (중략) </s> I have noticed that more damage is done to the existing electrical water suidgee foam and cable facilities than an to the existing electrical wires themselves. </s> 1. A method to improve the flow of heat energy in a heat engine having a cooling jacket, in which heat energy generated bythe engine is removed from the cooling jacket, said method comprising the steps of :

[표 2] 질의응답 문장 생성 결과

음성인식 데이터셋으로 학습한 모델은 이후 LibriSpeech 의 dev, test 데이터셋으로 성능을 측정했다. 사용한 성능 지표는 WER (Word Error Rate)로 값이 작을수록 높은 성능을 의미한다. 비교 모델은 Huggingface 에서 사전학습이 완료된 상태로 제공되는 모델들을 사용했다. 비교 실험 결과 음성인식 데이터셋으로 학습한 모델은 음성인식 작업을 잘 수행하는 것으로 보인다. [표 1] 이 모델로 SpokenSQuAD[15] 데이터셋을 사용해 QA (Question Answering, 질의응답) 과제를 테스트 하면 음성인식 결과를 먼저 생성하고 이 후 추가적인 문장을 생성했는데, 정답과 키워드는 유사하나 질문의 답에 대한 내용은 아닌 문장들이었다. [표 2] 이러한 문제를 해결하기 위해 음성인식으로 사전학습한 모델에 AMI 데이터셋을 사용해 요약 작업으로 추가 학습을 진행했다. 이 후 동일한 QA 작업을 수행한 결과 음성인식 결과는 더 이상 출력되지 않았으나 생성된 문장이 정답이 아니었다. 그러나 문법적으로 올바른 문장을 생성하고 있으며 답변의 내용으로 미루어 보아 '질문에 응답하고 있음'을 인지하고 있음을 확인했다. 이에 비해 요약 데이터셋으로 사전학습한 모델은 음성인식으로 사전학습한 모델에 비해 문장 생성이 불안정함을 확인했다.

5. Future Work

후속 연구에서는 음성인식 데이터셋과 요약 데이터셋의 내재적 특징을 분석하고 오디오 인코더의 파인튜닝 (Finetuning)이나 지시문 튜닝 (Instruction Tuning) 기법 등을 사용해 오디오의 다양한 특성을 반영한 표현을 학습할 예정이다. 그리고 요약 작업으로 사전학습 할 경우 불완전한 문장이 생성되는 이유에 대해 분석하고 다양한 작업에서 모델의 문장 생성 결과를 비교할 것이다.

6. Acknowledgements

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 학석사연계 ICT 핵심인재양성사업의 연구결과로 수행되었음" (RS-2023-00259867)

참고문헌

[1] Peters, Matthew E. et al. "Deep Contextualized Word Representations." ArXiv abs/1802.05365 (2018): n. pag.
[2] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
[3] Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.
[4] Tsimpoukelli, Maria, et al. "Multimodal few-shot learning with frozen language models." Advances in Neural Information Processing Systems 34 (2021): 200-212.
[5] Brock, Andy, et al. "High-performance large-scale image recognition without normalization." International Conference on Machine Learning. PMLR, 2021.
[6] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[7] Gao, Heting, et al. "Wavprompt: Towards few-shot spoken language understanding with frozen language models." arXiv preprint arXiv:2203.15863 (2022).
[8] Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." Advances in neural information processing systems 33 (2020): 12449-12460.
[9] Lea, Colin, et al. "Temporal convolutional networks for action segmentation and detection." proceedings of the IEEE Conference on Computer Vision and Pattern

Recognition. 2017.
[10] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
[11] Zhang, Susan, et al. "Opt: Open pre-trained transformer language models." arXiv preprint arXiv:2205.01068 (2022).
[12] Radford, Alec, et al. "Language models are unsupervised multitask learners." OpenAI blog 1.8 (2019): 9.
[13] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 2015, pp. 5206-5210, doi: 10.1109/ICASSP.2015.7178964.
[14] Carletta, Jean, et al. "The AMI meeting corpus: A pre-announcement." International workshop on machine learning for multimodal interaction. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005.
[15] Li, Chia-Hsuan, et al. "Spoken SQuAD: A study of mitigating the impact of speech recognition errors on listening comprehension." arXiv preprint arXiv:1804.00320 (2018).