**< 캡스톤디자인 최종보고서 >**

프로젝트 명: 사용자와 교감하는 대화형 인공지능을 위한 백채널 예측 모델 개발

팀원(참여학생): 이재형(소프트웨어융합학과), 최진욱(소프트웨어융합학과)

**요 약**

백채널(Backchannel)이란 대화 중 청자의 짧은 반응을 의미하며, 이는 대화의 자연스러움과 품질을 향상시키는 핵심 요소이다. 인간처럼 백채널 응답을 생성할 수 있는 인공지능은 사용자에게 더욱 풍부한 대화 경험을 제공한다. 본 연구에서는 인공지능과의 더 자연스러운 대화를 위해 기존 백채널 예측 모델을 검토하고, 트랜스포머 아키텍처와 focal loss, 오버샘플링 등을 도입한 개선된 백채널 예측 모델을 제안한다.

# 서론

## 연구배경

인공지능(AI) 기술의 발전은 사회의 다양한 분야에서 혁신적인 변화를 가져왔다. 특히, 대화형 인공지능은 우리 일상에 깊숙이 자리 잡으며, 사람과 기계 간의 소통 방식을 새롭게 정의하고 있다. 이러한 대화형 인공지능은 사용자의 질문에 대한 응답, 정보 제공, 그리고 특정 작업 수행 능력 등을 통해 일상 생활의 편리함을 제공하고 있다. 그러나 현존하는 많은 인공지능 시스템들은 아직 인간과 같은 수준의 백채널 반응을 생성하는 데에는 한계를 보이고 있다.

-------------------

## 연구 목표

본 연구는 인공지능 시스템이 대화 중 적절한 백채널 반응을 예측하는 능력을 향상시키는 것을 목표로 한다. ------

## 기대 효과

본 연구는 인공지능 시스템이 인간과의 상호작용에서 보다 섬세하고 자연스러운 대화 파트너로서의 역할을 수행할 수 있도록 하는 데 기여할 것이다. 또한, 이를 통해 ……

In summary, our contribution is a learned alternative to the Motion Matching algorithm which replaces the three key stages of the algorithm with specialized neural networks, resulting in state-ofthe-art animation generation results in terms of animation quality, runtime performance, and memory usage.

# 배경 지식 및 관련 연구

Ortega의 모델은 세 가지 병렬로 작동하는 컴포넌트(Lexical, Acoustic, Listener Embedding)로 구성되어 있다. Lexical 컴포넌트는 텍스트 입력에서 벡터 표현을 생성하기 위한 합성곱 신경망(CNN)과 fully connected (FC) 레이어로 구성되어 있다. Acoustic 컴포넌트는 음성 입력에서 벡터 표현을 생성하기 위해 합성곱 신경망(CNN)과 FC 레이어를 사용한다. Listener Embedding 컴포넌트는 청자의 인덱스(고유 식별자)을 입력으로 받아 청자의 벡터 표현을 생성하는데 임베딩 레이어와 FC 레이어를 활용한다.

----------------

## 합성곱 신경망(Convolutional Neural Networks, CNN)

합성곱 신경망(Convolutional Neural Networks, CNN)은 딥 러닝의 핵심 구조 중 하나로, 주로 이미지 인식, 분류 및 처리 작업에 널리 사용된다. ……

## 완전 연결 계층(Fully Connected Layer, FC Layer)

완전 연결 계층(Fully Connected Layer, FC Layer)인공 신경망의 구성 요소로, 'FC Layer' 또는 'Dense Layer'라고도 불린다. 이 층의 각 뉴런은 입력 및 출력 층을 제외하곤 이전 및 이후 층의 모든 뉴런과 연결되어 있으며, 이 연결은 가중치로 표현되어 연결의 강도를 나타낸다. 이러한 구조는 데이터의 특성 간 복잡한 관계를 학습하는 데 목적이 있다.

------------------------

## 손실함수(Loss Function)

손실 함수(Loss Function), 또는 비용 함수(Cost Function)는 기계 학습에서 예측값과 실제 목표값 사이의 차이를 정량적으로 나타내는 수학적 함수이다. 이 함수는 알고리즘의 예측이 실제 데이터와 얼마나 유사한지 판단하는 기준으로 사용되며, 이러한 예측값과 실제값 간의 차이가 '손실'로 간주된다. 손실 함수의 주된 역할은 모델 학습 과정을 안내하고, 모델의 성능을 평가하는 것이다.

….

## 기존 연구 문제점 및 해결 방안

본 프로젝트에서는 베이스라인 모델인 BPM\_MT의 문제점을 파악하고, 이를 해결하기 위한 방안을 모색해 보았다. 첫 번째 문제는 실세계 발화 데이터에서 백채널 카테고리별 샘플 수의 불균형이다. 이러한 불균형은 모델의 학습 과정에서 특정 카테고리에 편향되게 만들 수 있으며, 이는 실제 환경에서의 모델 성능 저하로 이어질 수 있다. BPM\_MT는 이러한 불균형을 효과적으로 다루는 메커니즘이 부족하다. 따라서 우리는 데이터 불균형을 해결하기 위한 방법을 모색해야 한다.

두 번째 문제는 오디오 처리에 있어서 BPM\_MT가 LSTM을 사용한다는 것이다. LSTM은 상대적으로 오래된 모델으로, 최근에는 더욱 발전된 오디오 처리 기술이 존재한다. LSTM보다 더 효율적이고 정확한 오디오 처리 방법을 사용하는 모델을 사용 하는 것을 고려해야 한다. 해결 방안으로 첫 번째 문제점인 데이터 불균형을 해결하기 위해 우리는 다음과 같은 접근 방식을 채택하였다.

* 데이터 불균형 해결

손실 함수 개선: 모델이 더욱 효과적으로 다양한 데이터를 학습할 수 있도록 손실 함수를 조정한다. 이를 통해 모델이 소수의 샘플에 대해서도 더 잘 학습할 수 있게 한다.

* 데이터셋 입력 방식 개선

데이터 입력 방식을 개선하여 모델이 더 균형 잡힌 데이터를 받도록 한다. 이는 모델의 일반화 능력을 향상시키는 데에 중요하다.

* 오버샘플링

실제 입력 데이터 수를 증가시키기 위해 오버샘플링을 진행한다. 이는 특히 소수의 샘플이 존재하는 카테고리에 도움이 된다.

* 오디오 모델 교체

Wav2Vec 2.0[7]과 HuBert[8] 모델 적용: LSTM을 대신하여 최신 오디오 처리 모델인 Wav2Vec 2.0과 HuBert를 적용한다. 이 모델들은 LSTM보다 오디오 데이터의 특성을 더 잘 포착하고, 결과적으로 더 정확한 백채널 예측을 가능하게 할 것이라 예상된다

# ……………………

# 추진 내용

## 팀 구성 및 역할

|  |  |  |
| --- | --- | --- |
| 이름 | 역할 | 수행업무 |
| 이재형 | 팀장 |  |
| 최진욱 | 팀원 |  |

## 전체 시스템 구성

…….

## 사용자 시나리오

---

## 핵심 기능 개발

1. LSTM의 대체

기존의 오래된 LSTM 모델 대신, BPM\_MT에 최적화된 Wav2Vec 2.0과 HuBert 모델을 도입함으로써 성능을 향상시키려고 한다. LSTM은 그 구조상 긴 시퀀스 데이터에서의 의존성을 학습하기에 유용하지만, 연산 부하가 크고 최근에 개발된 모델들에 비해 성능이 상대적으로 떨어진다. Wav2Vec 2.0과 HuBert 모델을 BPM\_MT에 적용함으로써, 더 빠른 학습 속도와 높은 정확도를 달성할 수 있을 것으로 기대한다. 이러한 Transformer 모델들은 self-attention 메커니즘을 통해 시퀀스 내의 각 요소들 사이의 관계를 더 효과적으로 학습할 수 있어, 최신 모델의 도입은 BPM\_MT의 성능을 크게 향상시킬 수 있을 것으로 예상된다.

…..

## …..

## 이슈 및 대응

……………

# 결과

## GPT를 통한 Text Augmentation 진행 시 SOTA

------------

## GPT의 결과

-------

--------------

# 결론

본 연구는 Focal Loss와 Transformer 기반 모델(Wav2Vec2-Large-XLSR- Korean과 HuBERT)을 활용한 백채널 카테고리 예측에 대한 새로운 접근법을 제안하였다. 기존의 CE Loss 대신 Focal Loss를 적용함으로써, 클래스 불균형이 심한 데이터에서 발생하는 오분류 문제를 완화시켰다. 이는 Focal Loss가 각 클래스에 대한 모델의 예측 오류에 다른 가중치를 부여함으로써 보다 균형 잡힌 학습을 유도하기 때문이다. 더욱이, LSTM 대신 Transformer 기반 모델을 채택함으로써, LSTM이 가진 시퀀스 데이터의 장기 의존성 학습에 대한 한계를 극복하고, self- attention 메커니즘을 통해 모델의 성능을 향상시켰다. 이러한 모델 구조의 변화는 특히 오디오 데이터 처리에 있어서 중요한 의미를 가지며, 실험 결과에서도 Wav2Vec2-Large와 Focal Loss의 조합이 가장 우수한 성능을 나타내었다.

본 연구의 중요한 성과 중 하나는 GPT를 이용한 Text Augmentation이었다. 이 방법을 통해 소수 클래스의 데이터를 인위적으로 확장함으로써, 클래스 불균형 문제를 효과적으로 완화시키고, 모델의 일반화 능력을 크게 향상시켰다. GPT를 사용한 Text Augmentation은 EDA 기법보다 더 효과적인 Text Augmentation임을 시사한다. GPT는 새로운 문맥과 다양한 표현을 지닌 데이터를 생성함으로써, 모델이 실세계의 다양한 상황에 대응할 수 있도록 훈련시킨다. 이는 백채널 카테고리 예측의 정확도를 현저히 향상시켜, 자연스러운 의사소통 분석에 큰 도움이 될 것으로 기대된다.

이러한 결과는 오디오 모델과 손실 함수의 선택이 백채널 예측의 정확성에 결정적인 영향을 미친다는 것을 보여주며, 특히 GPT를 이용한 Text Augmentation의 효과는 이 분야에서의 중요한 발전으로 평가될 수 있다. 그러나 실험에서는 LSTM과 CE Loss가 특정 조건에서 여전히 우수한 성능을 보였으며, 이는 손실 함수와 모델 구조의 선택이 단순히 클래스 불균형 문제를 해결하는 것을 넘어 다양한 요소를 고려해야 함을 나타낸다. 향후 연구에서는 loss 함수를 더욱 개선시키고, ELECTRA와 같은 더 진보된 언어 모델을 탐색함으로써, 백채널 카테고리 예측의 정확도를 더욱 향상시킬 방안을 모색할 예정이다. 이는 백채널 모델의 발전에 크게 기여할 것으로 기대된다.

참고문헌

[1] D. Ortega, C.-Y. Li and N. T. Vu, "OH, JEEZ! or UH-HUH? A Listener-Aware Backchannel Predictor on ASR Transcriptions,” ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, pp. 8064-8068. 2020.

[2]