**< 캡스톤디자인 최종보고서 >**

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

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

**요 약**

MotionMatching이란 가상 캐릭터 애니메이션을 생성하는데 사용되고 있는 기술로써, Clavet와 Büttner에 의해 연구되어 발표되었습니다.[1] MotionMatching기술은 애니메이션 클립을 블렌드 트리(Blend Tree)나 상태 기계(State Machine)에 배열할 필요가 없이, 캐릭터의 각 상황에 맞는 애니메이션 클립을 대규모애니메이션 데이터베이스에서 찾아 매칭 시킨 후 시퀀스들 간의 블렌딩과 전환을 연속적으로 수행하여 보다 자연스럽고 유연한 애니메이션을 생성할 수 있습니다. 때문에 어플리케이션 실행 시간에서 주어진 상황이 계속적으로 바뀌는 게임 산업에서 많이 활용되고 있습니다.

본 보고서에서는 기존 MotionMatching에서 딥러닝 기술을 도입하여 발전된 Learned MotionMatching[2] 을 구현하고, 이를 바탕으로 “모션 액션 택배 배달” 게임 컨텐츠를 제작 하였습니다.

# 서론

## 연구배경

상호작용가능한 어플리케이션, 특히 콘솔게임이나 PC게임 등에서는 점점 기술이 발전함에 따라 이용자는 AAA게임과 같은 더 사실적이고 몰입감 있는 동적인 세계를 원하며, 그들의 요구의 맞추어서 동적인 세계 안에서 시각적으로 보이는 그래픽 기술들이 어떻게 하면 사실 적으로 표현 될지에 관해서 많은 연구와 개발을 통해 발전 되고 있습니다. 여기서 시각적으로 보이는 그래픽 기술은 비단 완성도가 높은 게임 배경이나, 많은 양의 파티클을 요구하는 화려한 이펙트 뿐만 아니라 캐릭터의 움직임이 얼마나 사실적인가 또한 포함하고 있습니다.

이미 많은 사람들이 사실적인 캐릭터 애니메이션에 관하여 고민을 해왔고, 다양한 기술이 고안이 되었으며, 그 중 최근에 가장 많이 쓰이는 기술이 “MotionMatching”이라는 기술입니다. 하지만 “MotionMatching”기술에는 여러 단점이 존재 합니다.

첫 번째 단점은 애니메이션 데이터의 규모입니다. 모션 매칭은 방대한 양의 사전 녹화된 애니메이션 데이터를 필요로 합니다. 데이터베이스에 저장되는 애니메이션 클립이 많아질수록 메모리 사용량도 증가합니다. 메모리 사용량은 데이터 양에 따라 선형적으로 증가하여, 데이터가 많아질수록 더 많은 메모리를 요구합니다. 이는 고품질 애니메이션을 위해 많은 데이터를 필요로 하는 현대의 AAA 게임에서 특히 문제가 될 수 있습니다.

두 번째 단점은 **실시간 검색 비용이 많이 듭니다.** 최근접 이웃 검색(nearest neighbor search)을 실시간으로 수행해야 하므로, 많은 애니메이션 데이터가 있을 경우 런타임 성능이 저하될 수 있습니다. 리얼 타임이 중요한 게임 어플리케이션에서는 치명적으로 다가올수 있습니다.

세 번째 단점은 데이터베이스 구축 비용입니다. 대규모 애니메이션 데이터베이스를 구축하고 유지하는 데 많은 시간과 노력이 필요합니다. 또한 “MotionMatching”에서는 애니메이션 품질에 크게 의존 됩니다.  
원본 데이터가 부자연스럽거나, 부정확하면 결과 애니메이션도 품질이 떨어 질 수 있습니다.  
 네 번째 단점은 일반 “MotionMatching”을 사용했을 때 캐릭터간의 상호작용 혹은 캐릭터와 오브젝트 사이의 복잡한 상호작용이나 환경과의 복잡한 상호작용을 처리하는 데 어려움이 있습니다. 이는 더 복잡한 알고리즘이나 추가적인 데이터 처리 기술을 필요로 할 수 있습니다.  
 이러한 단점들 때문에, 모션 매칭은 특정 상황에서는 매우 유용하지만, 모든 상황에서 최적의 솔루션은 아닙니다. 따라서 신경망 기반의 생성 모델을 활용하여 고안된 “Learned MotionMatching”을 구현하여 기존 “MotionMatching”에서의 단점을 해결하려 합니다.

## 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] Michael Büttner and Simon Clavet. 2015. Motion Matching - The Road to Next Gen Animation. In Proc. of Nucl.ai 2015. https://www.youtube.com/watch?v= z\_wpgHFSWss&t=658s

# [2] DANIEL HOLDEN, Ubisoft La Forge, Ubisoft, Canada SIGGRAPH 2020 – learned Motion Matching https://youtu.be/16CHDQK4W5k?si=CfZw2QgU5GXvKqGY