DMQA open seminar

Dive into Audio Transformer

2022. 04. 01

Data Mining & Quality Analytics Lab.

발표자: 고은지

ejkoh21@korea.ac.kr

KOREA UNIVERSITY Data Mining Quality Analytics

Contents

Introduction

❖ 발표자 소개

- 고은지
- 고려대학교 산업경영공학과
- Data Mining & Quality Analytics Lab. (김성범 교수님)
- M.S student (2021.03 ~)

✓ 관심 연구 분야

- Deep Learning for multichannel signal analysis
- Semi-supervised learning on graphs

Introduction

❖ Introduction

- 다양한 오디오 데이터에 인공지능 기술을 적용하는 사례를 쉽게 접할 수 있음
- 오디오를 분석하기 위한 다양한 연구가 활발히 진행되고 있음

오디오 인식

인공지능 스피커

인공지능 기술을 기반으로 사람의 말을
인식하여 적절한 기능 등을 수행

오디오 분류

차량 소음을 통한 차량 상태 진단

인공지능 기술을 적용하여 차량 소음으로
고장 여부를 판별하고,
고장의 원인이 되는 부품을 진단

오디오 생성

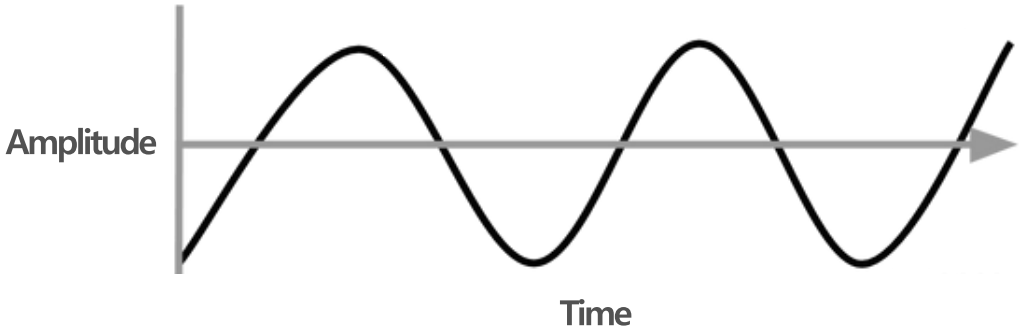
광고를 담당하는 인공지능 성우

인공지능 기술을 사용하여 브랜드별
특성이나 나라별 억양에 맞는
광고 멘트 생성

Audio data

❖ Audio

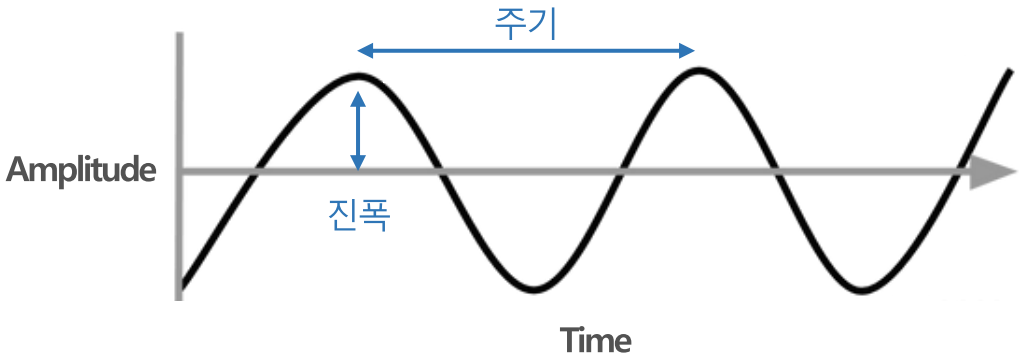
- 오디오는 물체가 진동함에 따라 발생
  - ➢ 사람의 목소리는 공기 분자가 진동을 하며 발생
- 오디오의 특징을 표현하는 중요한 요소로 **소리의 크기**, **높낮이**, 음색이 있음



Audio data

❖ Audio

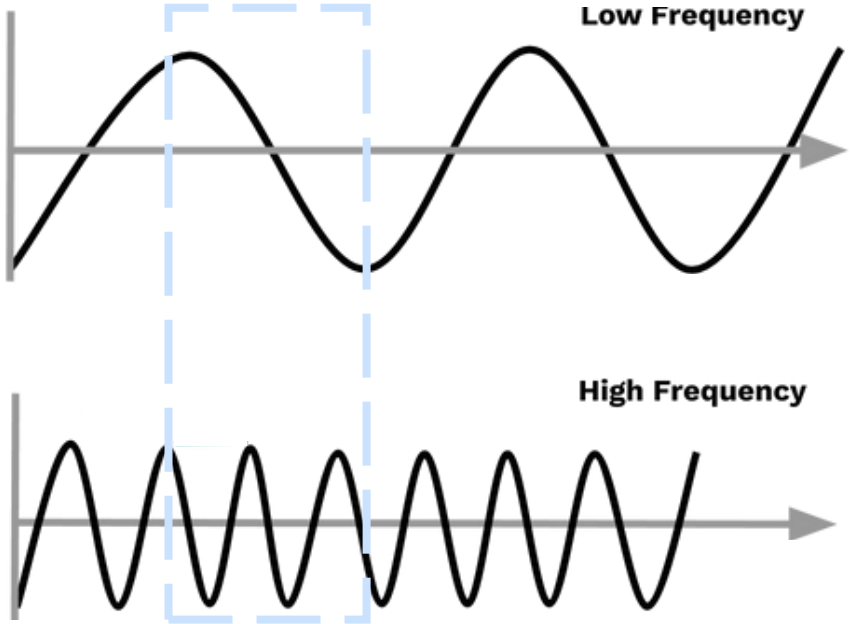
- 오디오의 **크기(amplitude)**: waveform의 높이에 해당함
  - ➢ 일반적으로 진폭이 높으면 큰 소리를 의미
- 오디오의 **높낮이(frequency)**: 1초동안의 진동 횟수로 표현됨
  - ➢ 같은 시간동안 상대적으로 많이 진동하면(주기가 짧으면) 높은 소리를 의미



Audio data

❖ Audio

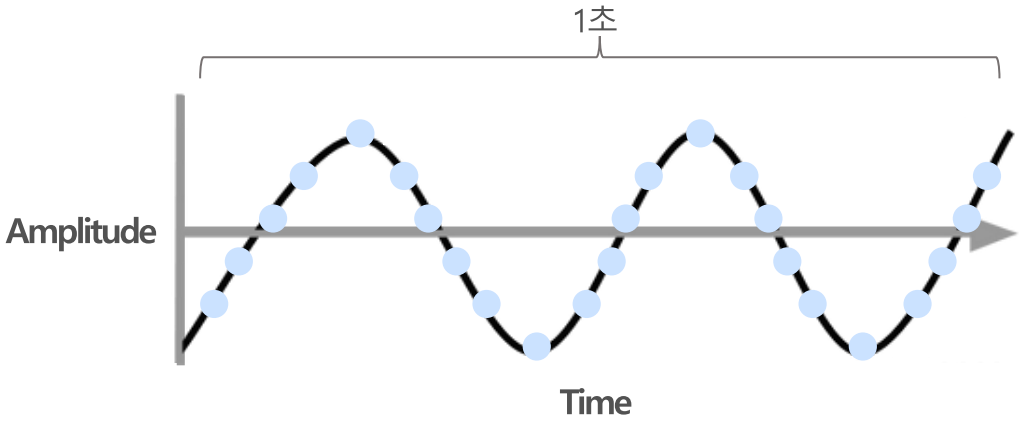
- 오디오의 **크기(amplitude)**: waveform의 높이에 해당함
  - ➢ 일반적으로 진폭이 높으면 큰 소리를 의미
- 오디오의 **높낮이(frequency)**: 1초동안의 진동 횟수로 표현됨
  - ➢ 같은 시간동안 상대적으로 많이 진동하면(주기가 짧으면) 높은 소리를 의미



Audio data

❖ Audio data

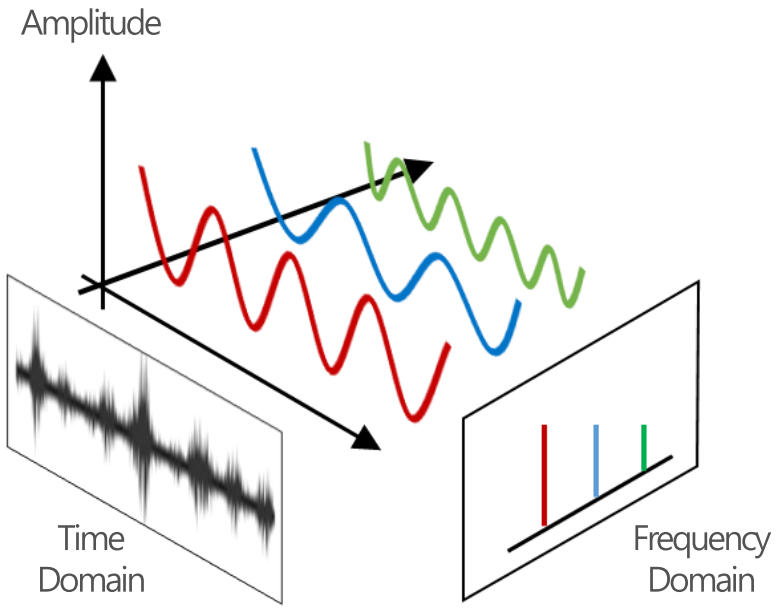
- 컴퓨터가 오디오를 이해하기 위해서는 오디오를 숫자로 표현해야 함
- Sampling 과정을 통해 아날로그 신호(오디오)를 디지털(오디오 데이터)로 변환
  - ➢ Sampling rate는 오디오에서 초당 sampling 하는 값의 개수를 의미
  - ➢ 예를 들어 sampling rate가 44,100인 경우, 오디오로부터 1초에 44,100개의 값을 추출한 것을 의미



Audio data

❖ Audio data

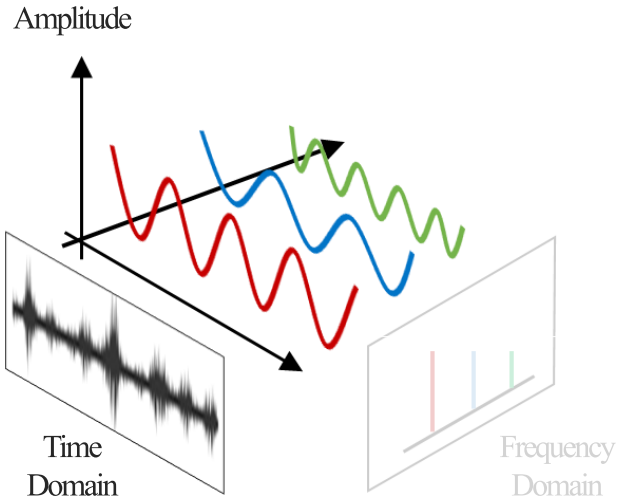
- 오디오 데이터는 time domain과 frequency domain에서 표현 가능
- Time domain은 시간에 따른 오디오의 특징 표현에 집중
- Frequency domain은 오디오를 구성하는 여러 주파수의 관점에서 특징을 표현

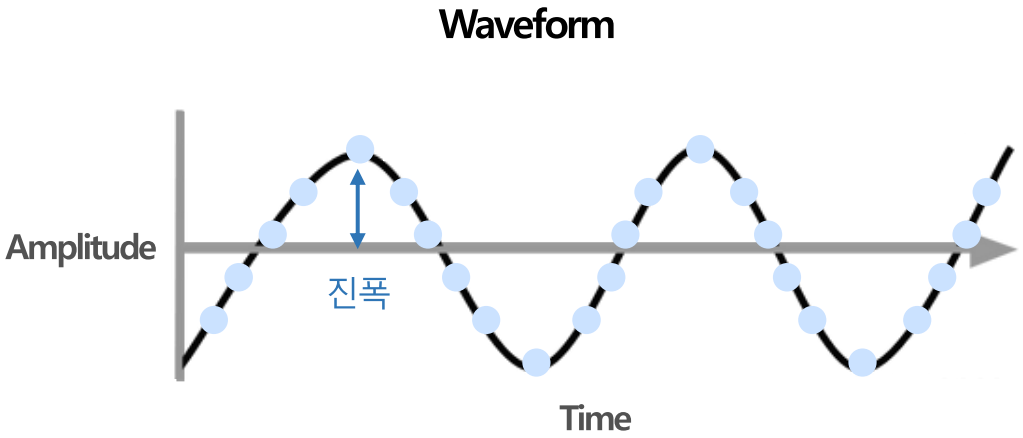


Audio data

❖ Time domain

- Waveform은 시간에 따른 진폭의 변화를 표현함
  - ➢ 소리의 크기 정보에 집중하여 표현하는 방식
- 오디오를 구성하는 여러 frequency별 특징을 표현하기 어려움

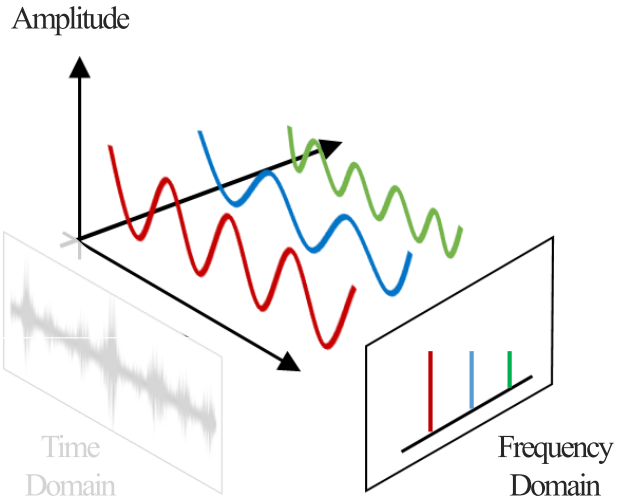




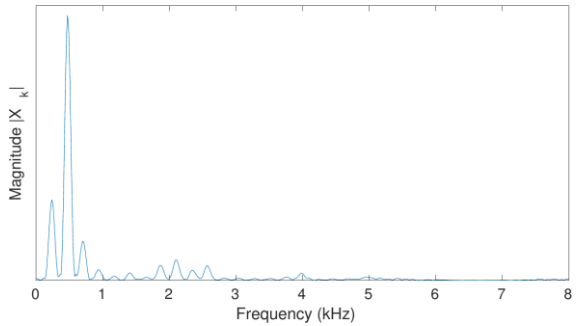
Audio data

❖ Frequency domain

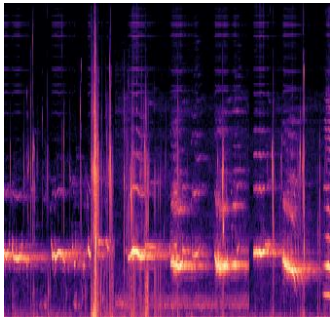
- Frequency domain에서는 오디오를 구성하는 주파수들의 정보를 중심으로 표현함
- 오디오를 구성하는 여러 주파수 정보를 알기 위해서 푸리에 변환을 통한 특징 추출이 필요
  - ➢ 푸리에 변환은 오디오를 서로 다른 frequency의 합으로 표현
- 푸리에 변환을 통한 특징 추출을 통해 스펙트럼, 스펙트로그램, MFCC 등을 얻을 수 있음



Spectrum



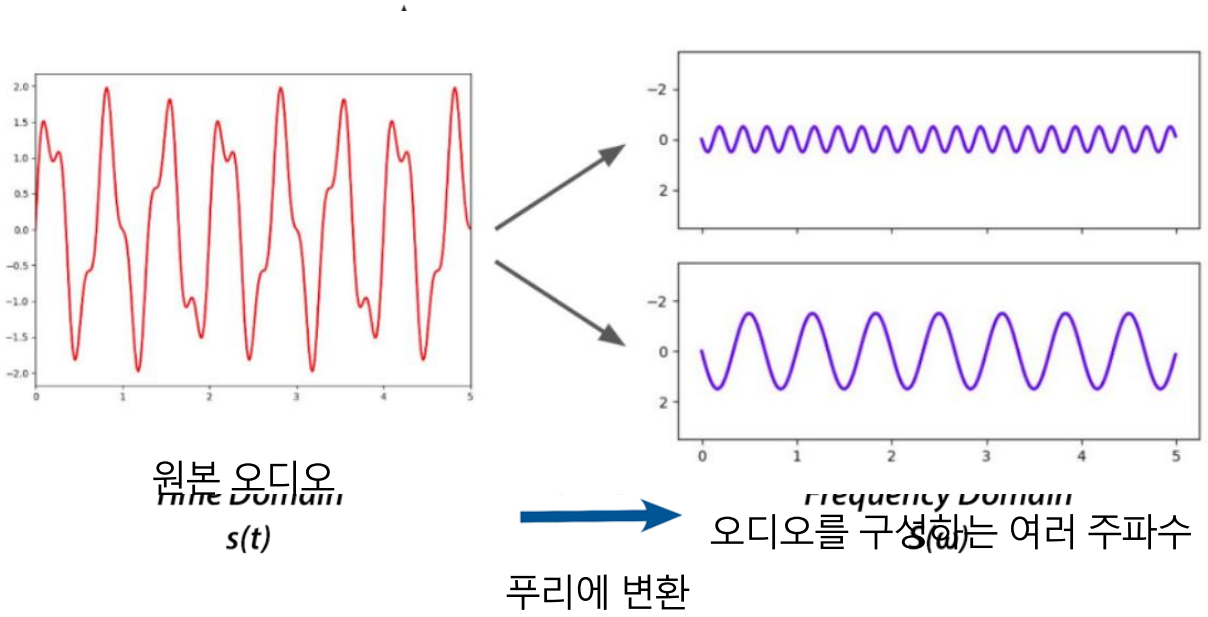
Spectrogram



Feature engineering for audio data

❖ 푸리에 변환

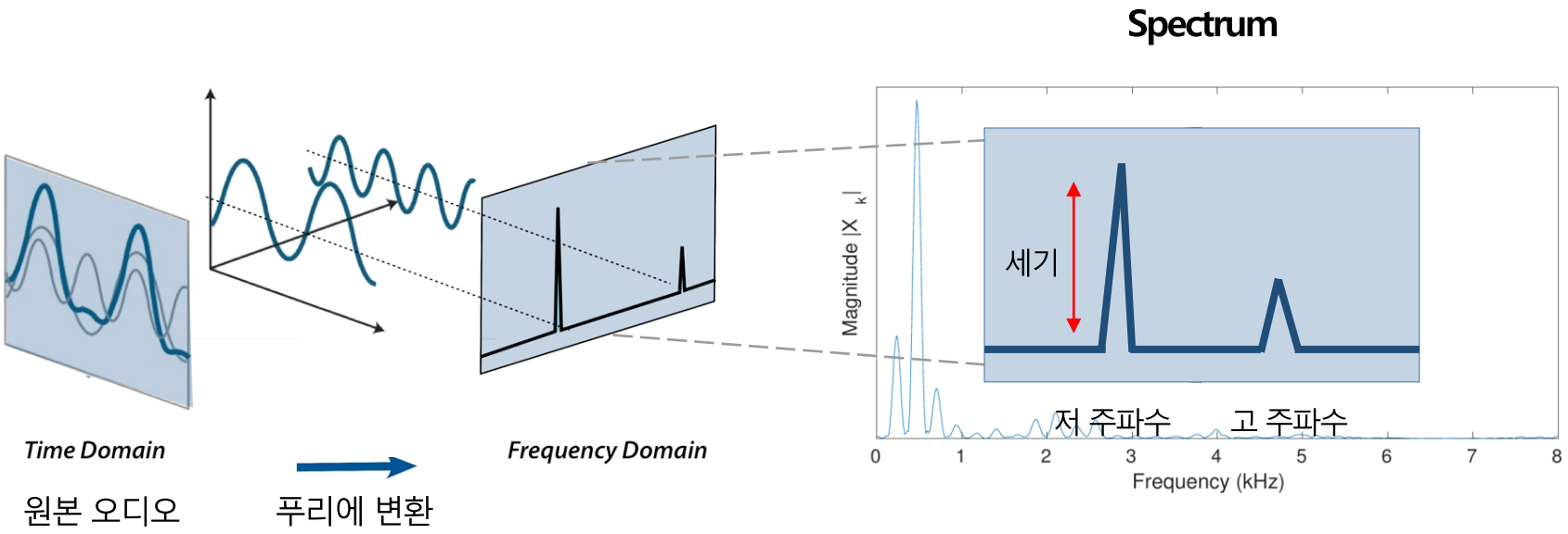
- 푸리에 변환은 오디오 신호를 다양한 주파수들의 합으로 표현
  - ➢ 각각의 주파수에 대한 해석을 가능케 함
- 푸리에 변환을 통해 스펙트럼 추출 가능



Feature engineering for audio data

❖ 스펙트럼

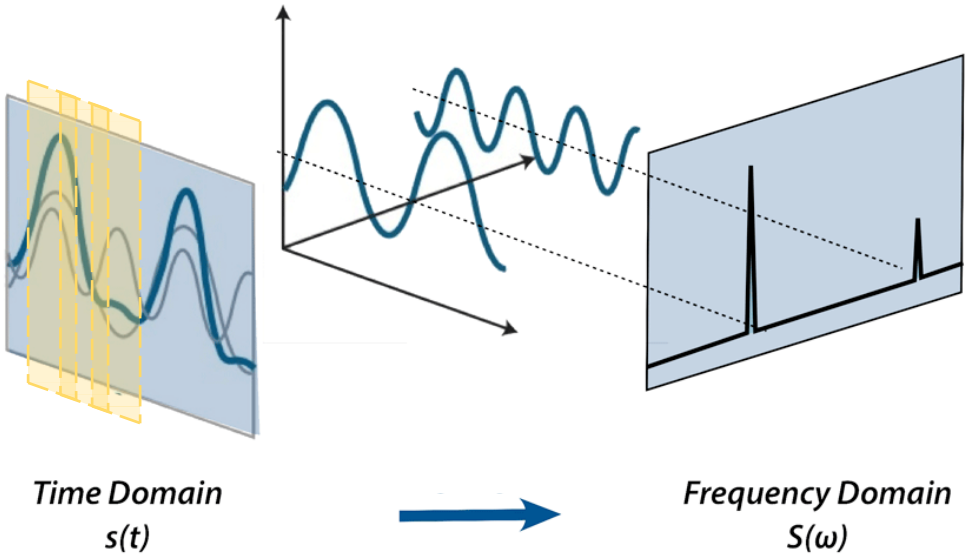
- 오디오를 구성하는 여러 frequency별 강도로 표현
- X축은 주파수(frequency), Y축은 세기(magnitude)를 나타냄
- 오디오를 time domain에서 frequency domain으로 변환한 특징 추출 결과
- 오디오의 시간 정보는 표현 불가하다는 한계가 있음



Feature engineering for audio data

❖ 단시간 푸리에 변환(STFT)

- 오디오에 푸리에 변환을 적용하면 시간 정보가 사라진다는 단점이 있음
- 시간 정보를 보존하기 위해서 단시간 푸리에 변환(short time fourier transform; STFT) 사용
- STFT는 오디오의 일정 시간 마다 푸리에 변환을 취해 시간 순으로 나열하는 방법
- STFT를 통해 스펙트로그램 추출 가능

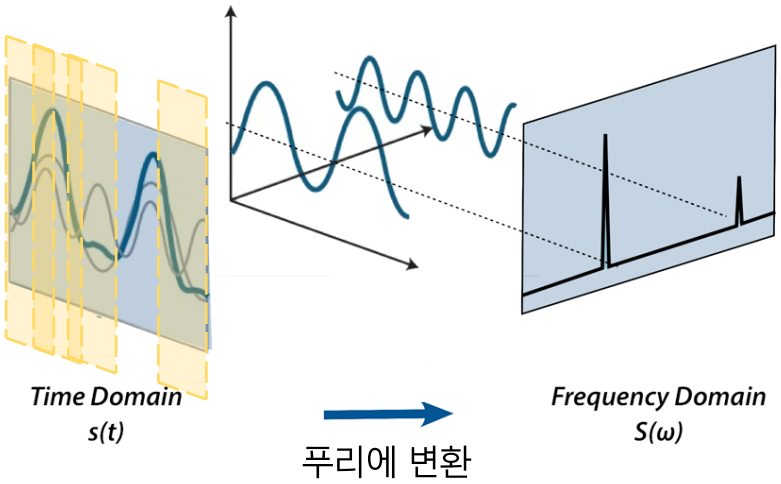


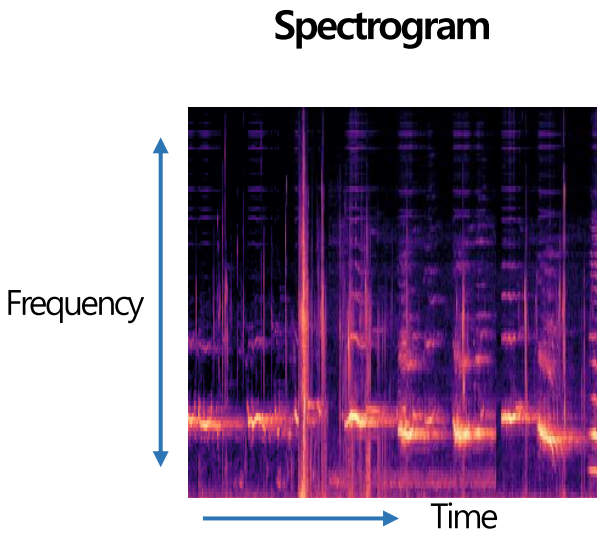
푸리에 변환

Feature engineering for audio data

❖ 스펙트로그램

- STFT를 통해 시간에 따른 주파수별 세기를 나타내는 스펙트로그램 추출
- X축은 시간(time), Y축은 주파수(frequency), 색상은 세기(magnitude)를 나타냄
- 오디오의 시간 정보와 주파수 특징을 모두 표현할 수 있음

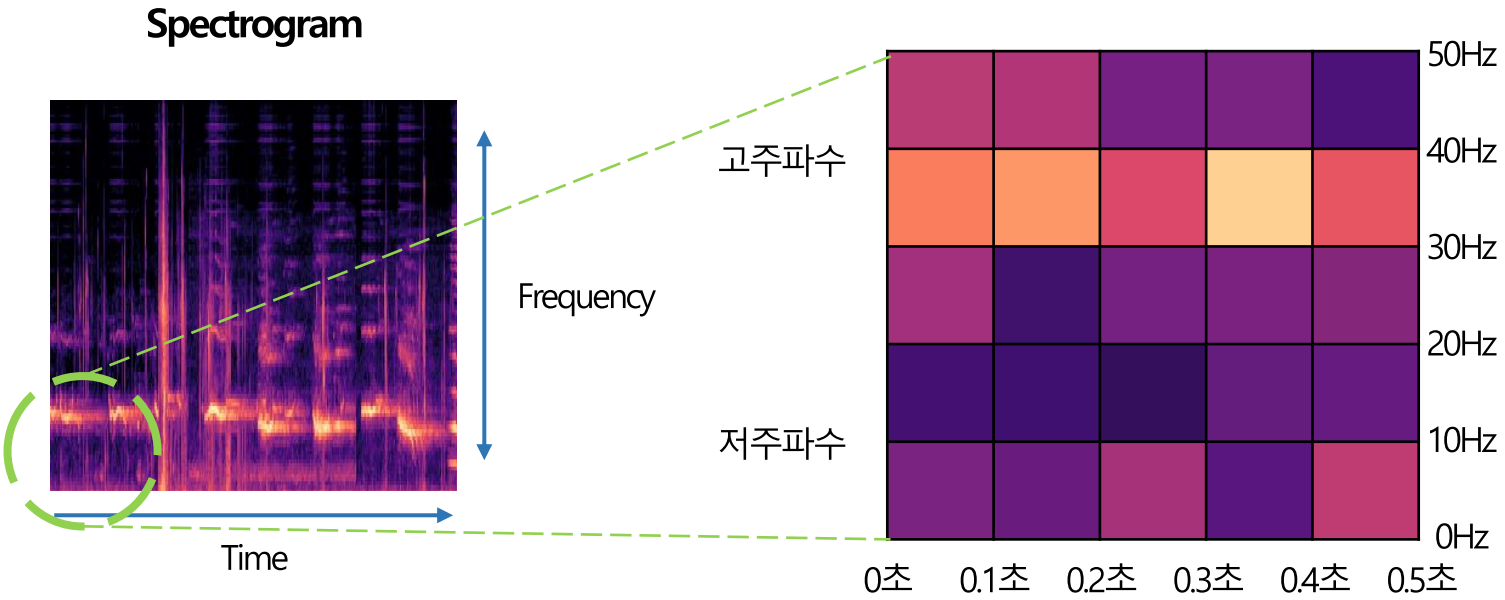




Feature engineering for audio data

❖ 스펙트로그램

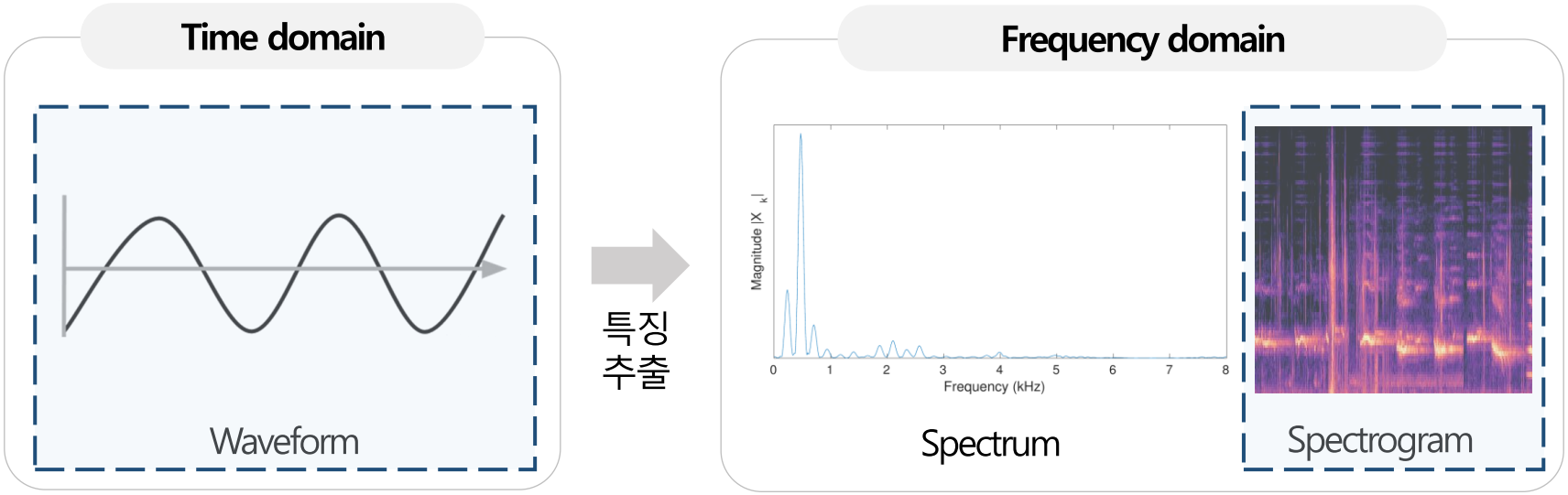
- STFT를 통해 시간에 따른 주파수별 세기를 나타내는 스펙트로그램 추출
- X축은 시간(time), Y축은 주파수(frequency), 색상은 세기(magnitude)를 나타냄
- 오디오의 시간 정보와 주파수 특징을 모두 표현할 수 있음



Feature engineering for audio data

❖ 특징 추출

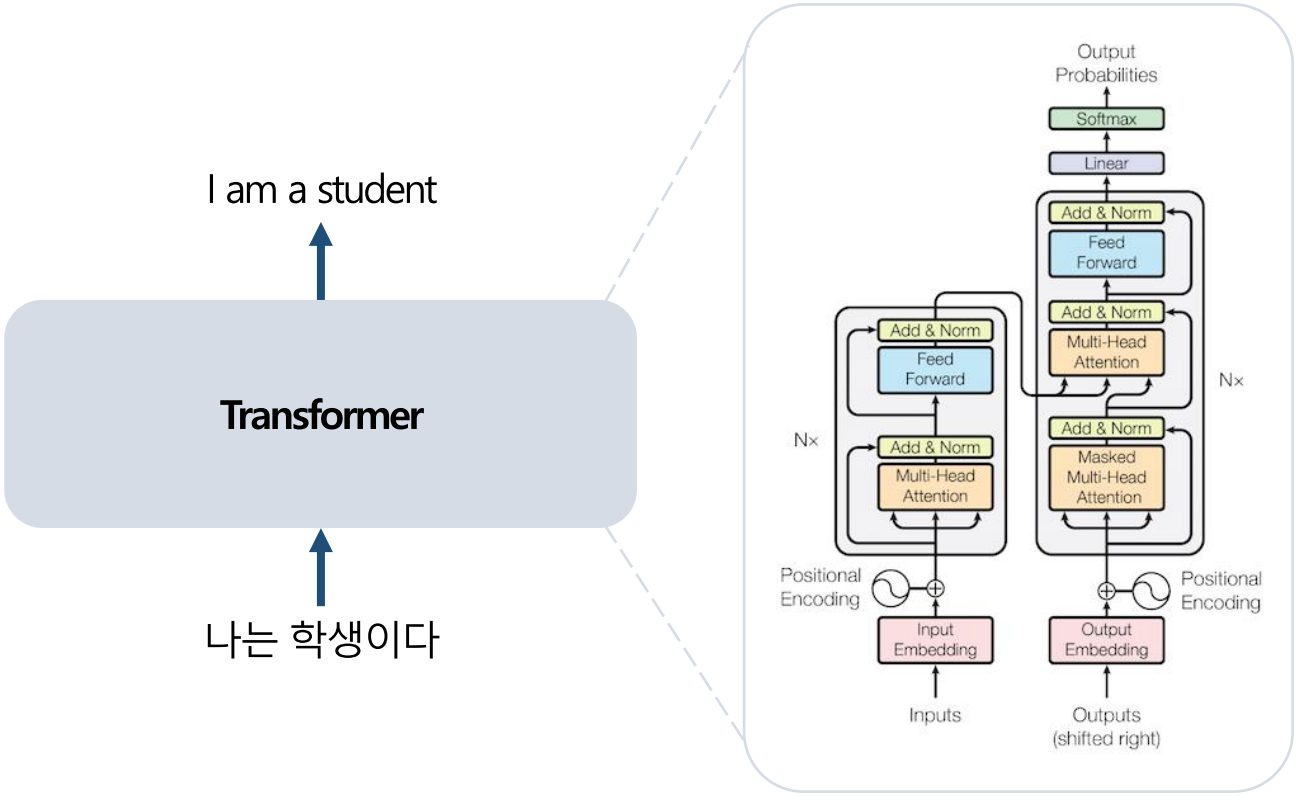
- Time domain인 waveform으로부터 푸리에 변환을 통해 frequency domain의 스펙트럼, 스펙트로그램 추출
  - ➢ Waveform: 오디오를 구성하는 주파수 정보 표현 불가
  - ➢ 스펙트럼: 시간 정보 표현 불가
  - ➢ 스펙트로그램: 시간에 따른 주파수 정보 표현 가능



Transformer for Audio data

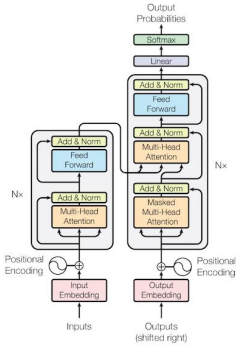
❖ Transformer

- Sequence 데이터를 순차적으로 처리함으로써 발생하는 높은 계산 복잡도와 연산 시간 문제를 해결
  - ➢ Sequence 데이터를 병렬 처리하여 계산 복잡도와 연산 시간을 줄이고 입력 간의 dependency를 모델링
- NLP와 vision 분야에서 RNN 및 CNN 모델을 대체하며 성능 향상을 이루고 우수성을 입증

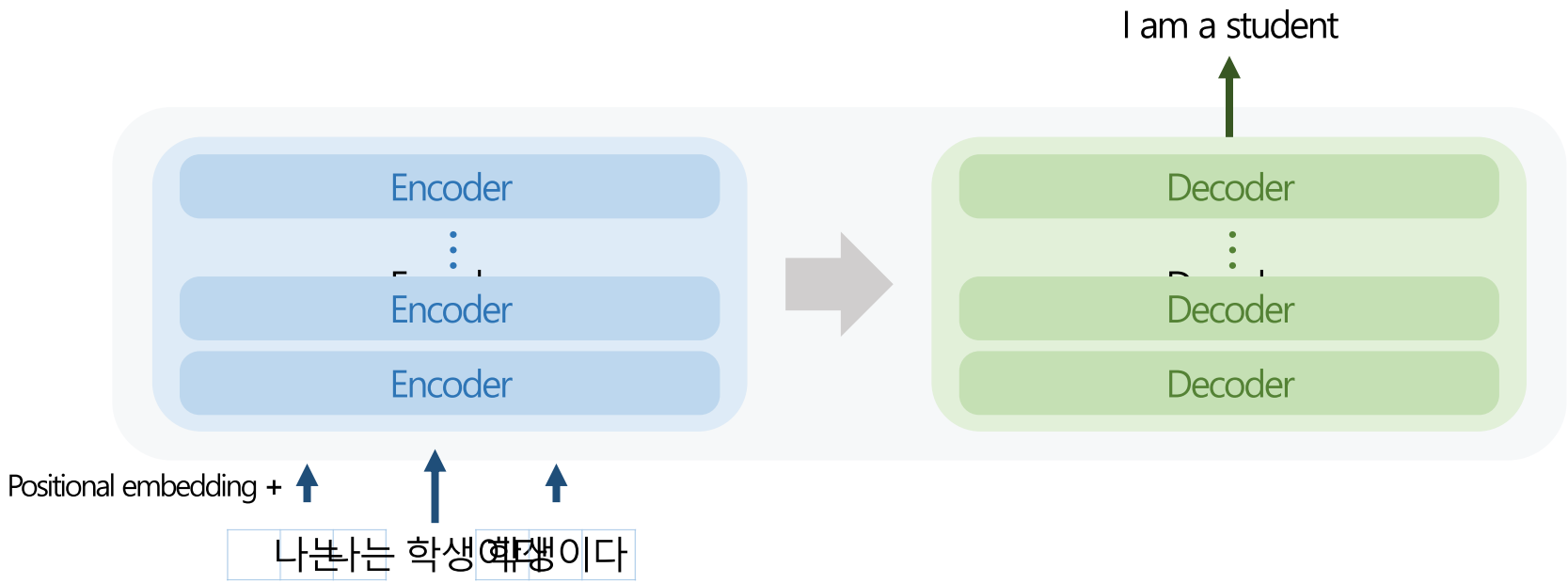


Transformer for Audio data

❖ Transformer

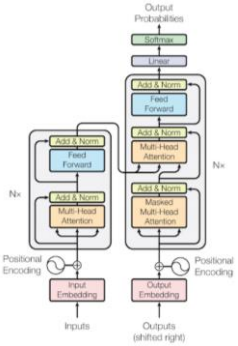


- Transformer는 기존의 seq2seq과 동일한 encoder-decoder 구조
  - Encoder에서 sequence 데이터를 입력 받고 decoder에서 타겟 sequence를 출력
  - Encoder와 decoder 각각 6개의 encoder, decoder로 구성
- Sequence 데이터가 입력되면 단어(토큰)들을 임베딩하여 encoder에 전달
  - 임베딩 과정에서 문장 내 단어의 위치정보를 제공하는 positional 임베딩을 더함

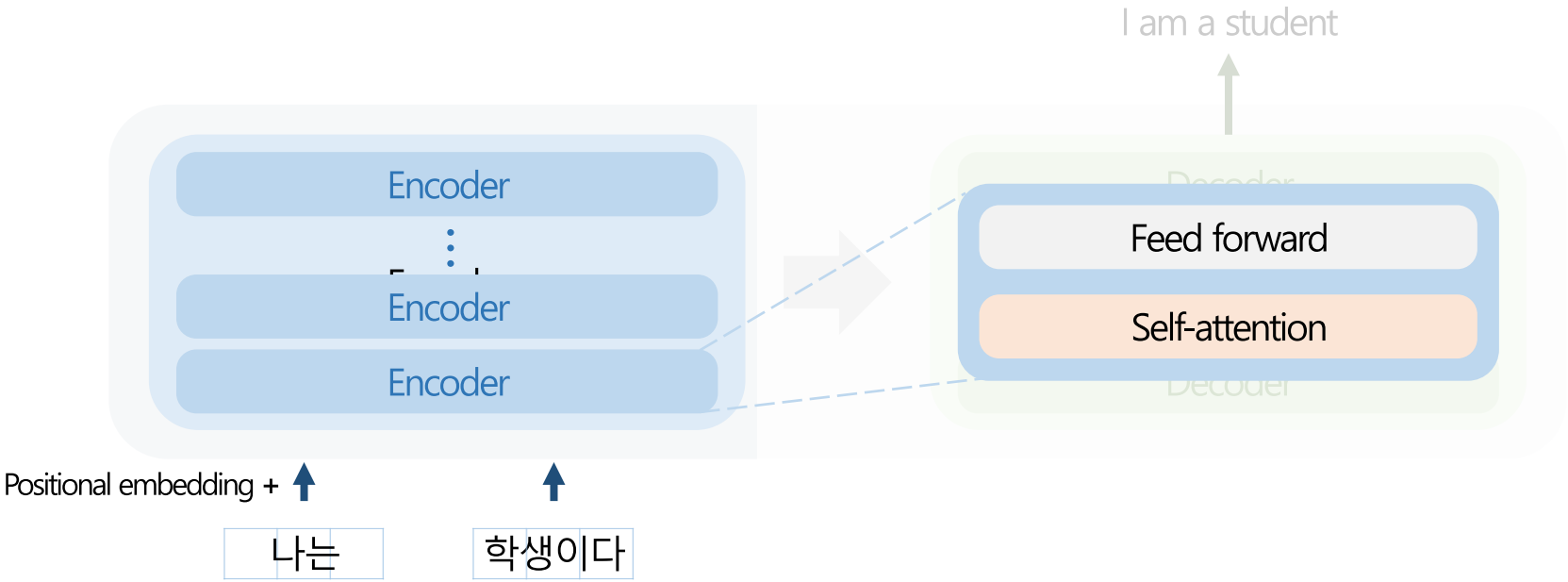


Transformer for Audio data

❖ Transformer



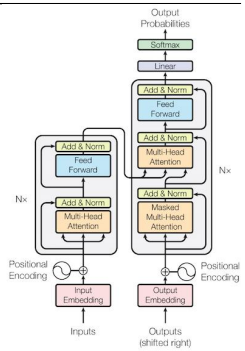
- Encoder는 self-attention과 feed forward neural network로 구성
  - ➢ Self-attention은 개별 단어에 대한 정보 요약 및 단어 간의 관계 파악
  - ➢ Multi-head attention을 사용함으로써 여러 개의 head를 통해 단어 간 다양한 관계 학습
  - ➢ Encoder는 attention 벡터를 출력하여 decoder에 전달
- Encoder 내의 residual connection, layer normalization 과정 포함
  - ➢ Self-attention과 feed forward를 각각 통과할 때마다 normalization 수행

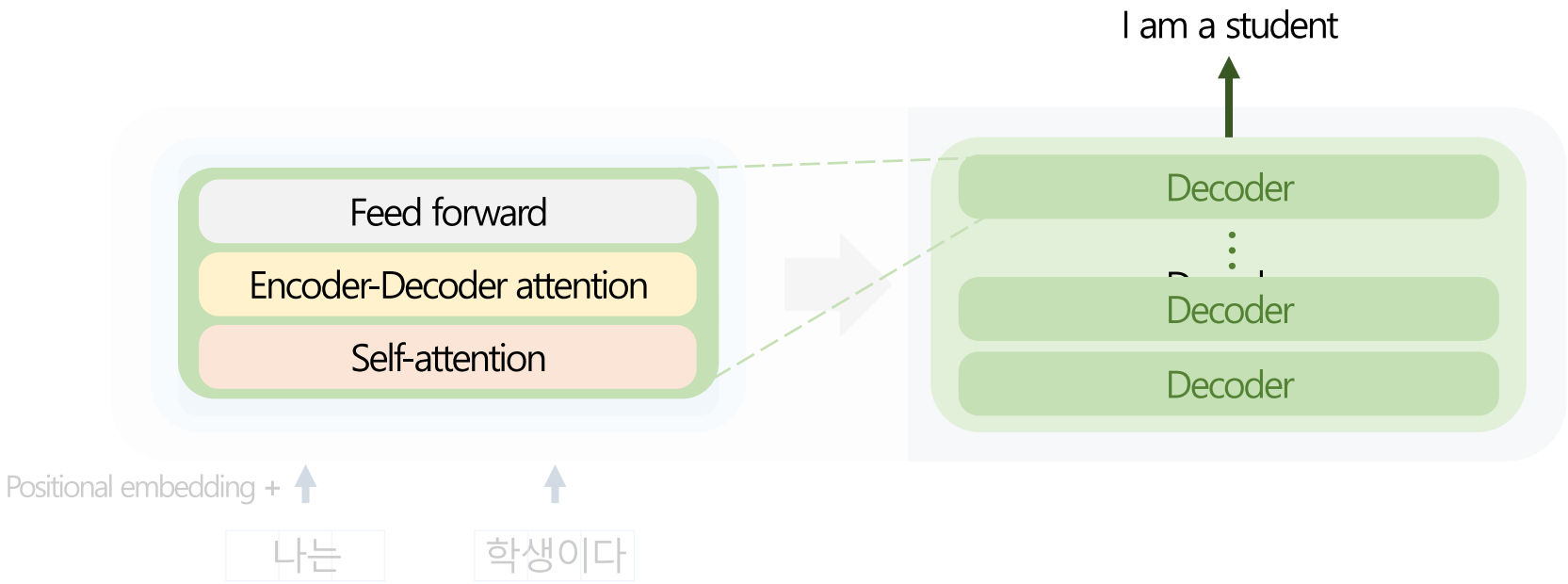


Transformer for Audio data

❖ Transformer

- Decoder는 self-attention, encoder-decoder attention, feed forward neural network로 구성
- Self-attention 수행 시 미래 시점의 정보를 반영하지 않도록 masking
- Encoder-decoder attention은 encoder에서 출력된 attention 벡터를 사용하여 self-attention 수행 시 decoder가 입력 sequence의 적절한 위치에 집중하도록 함

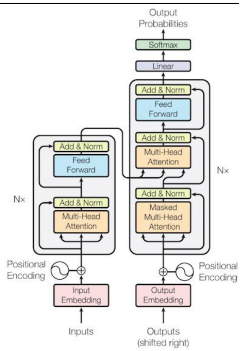


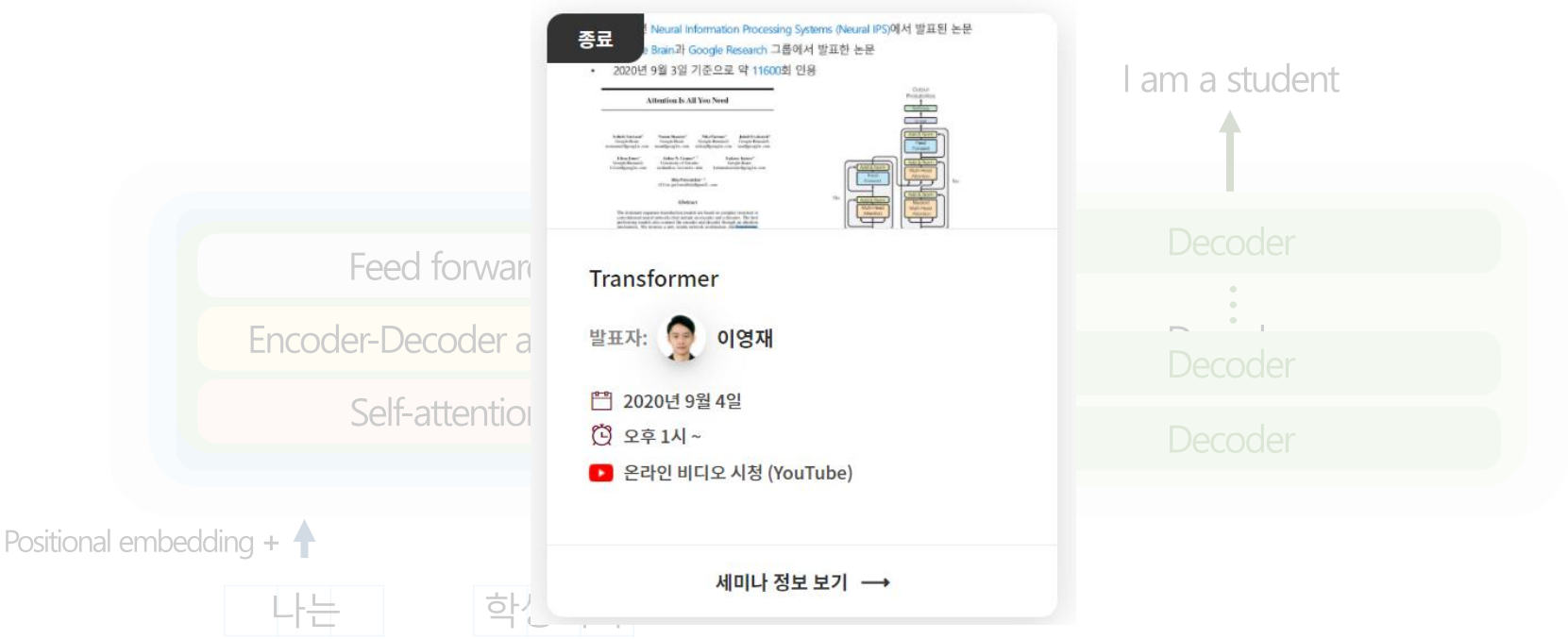


Transformer for Audio data

❖ Transformer

- Decoder는 self-attention, encoder-decoder attention, feed forward neural network로 구성
- Self-attention 수행 시 미래 시점의 정보를 반영하지 않도록 masking
- Encoder-decoder attention은 encoder에서 출력된 attention 벡터를 사용하여 self-attention 수행 시 decoder가 입력 sequence의 적절한 위치에 집중하도록 함





Transformer for Audio data

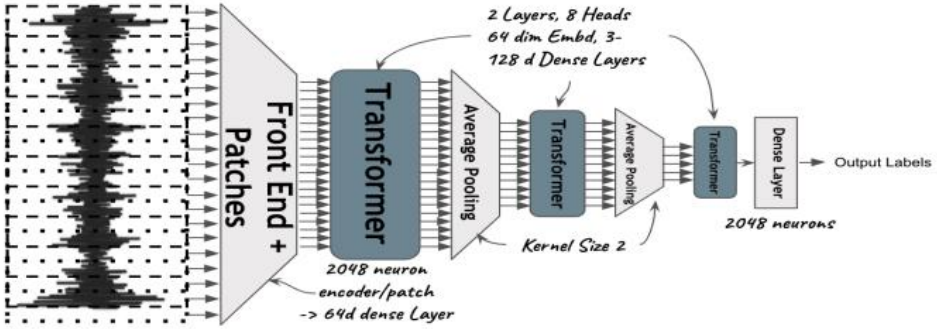
❖ Audio Transformers: transformer architectures for large scale audio understanding.

- Stanford 대학에서 연구하였으며(arXiv, 2021), 2022년 3월 31일 기준 11회 인용되었음
- 오디오 분류를 위해 waveform에 적합한 transformer 구조를 제안
- CNN에서 효과적으로 사용되는 pooling layer를 접목한 것이 특징

AUDIO TRANSFORMERS:
TRANSFORMER ARCHITECTURES FOR LARGE SCALE AUDIO UNDERSTANDING.
ADIEU CONVOLUTIONS*

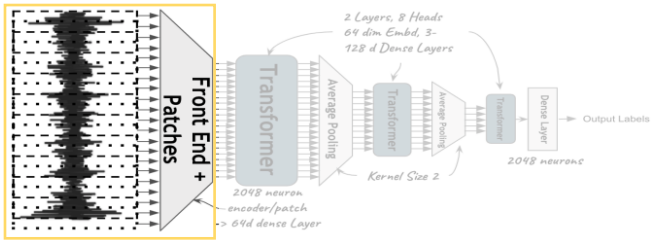
Prateek Verma and Jonathan Berger

Stanford University
450 Jane Stanford Way, Stanford CA, 94305,
prateekv, brg@stanford.edu



Transformer for Audio data

❖ Audio Transformer



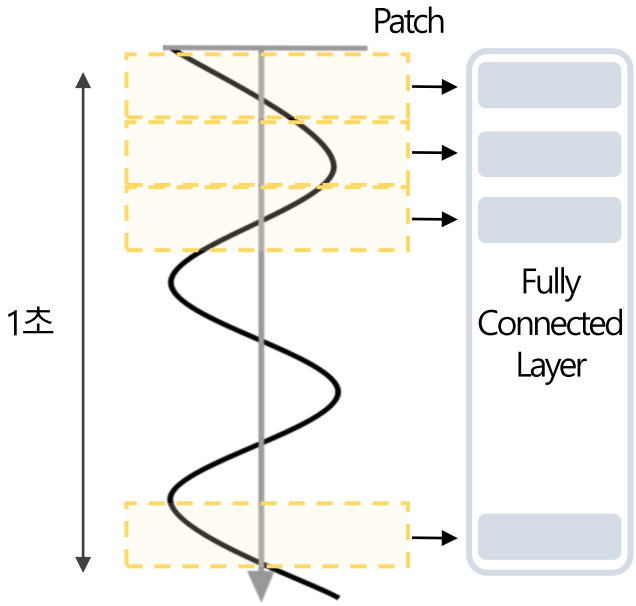
- Audio transformer 동작 구조

1) Waveform 임베딩

- ➢ 1초 길이의 waveform을 겹치지 않는 윈도우로 25ms 길이의 패치 생성
- ➢ Frequency 특징 반영을 위해 filter bank를 학습하는 fully connected layer를 사용하여 패치 임베딩

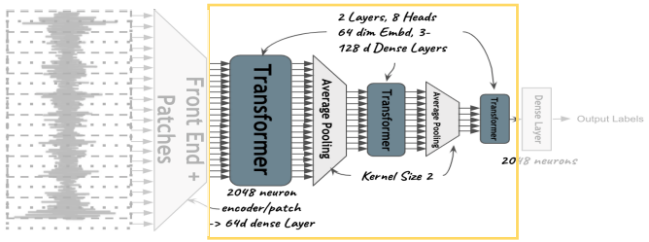
2) Transformer & Average pooling

3) Dense layer

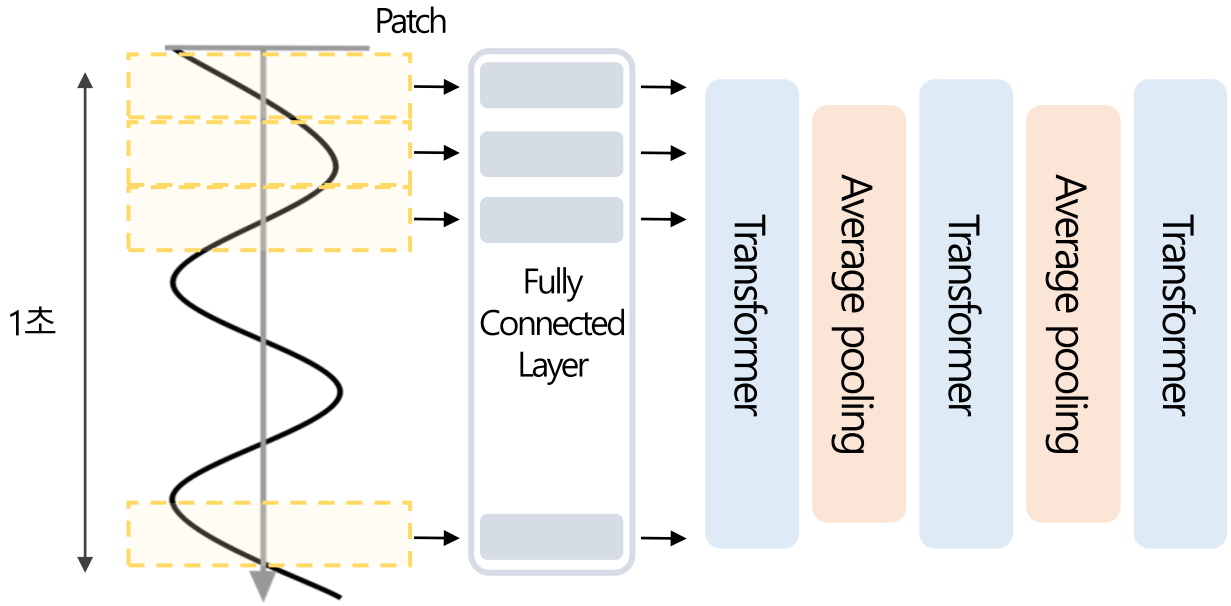


Transformer for Audio data

❖ Audio Transformer

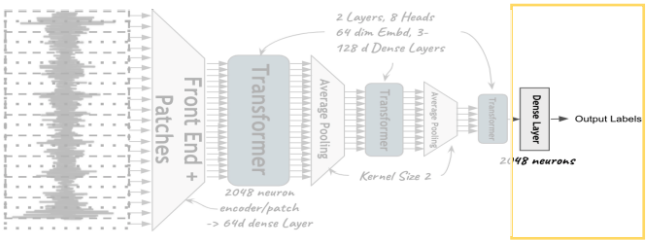


- Audio transformer 동작 구조
  1) Waveform 임베딩
  2) Transformer & Average pooling
     - ➢ 3번의 transformer layer를 통과하며 각 layer 사이에 average pooling을 적용
     - ➢ 각 transformer encoder는 self-attention과 feed forward로만 구성
  3) Dense layer

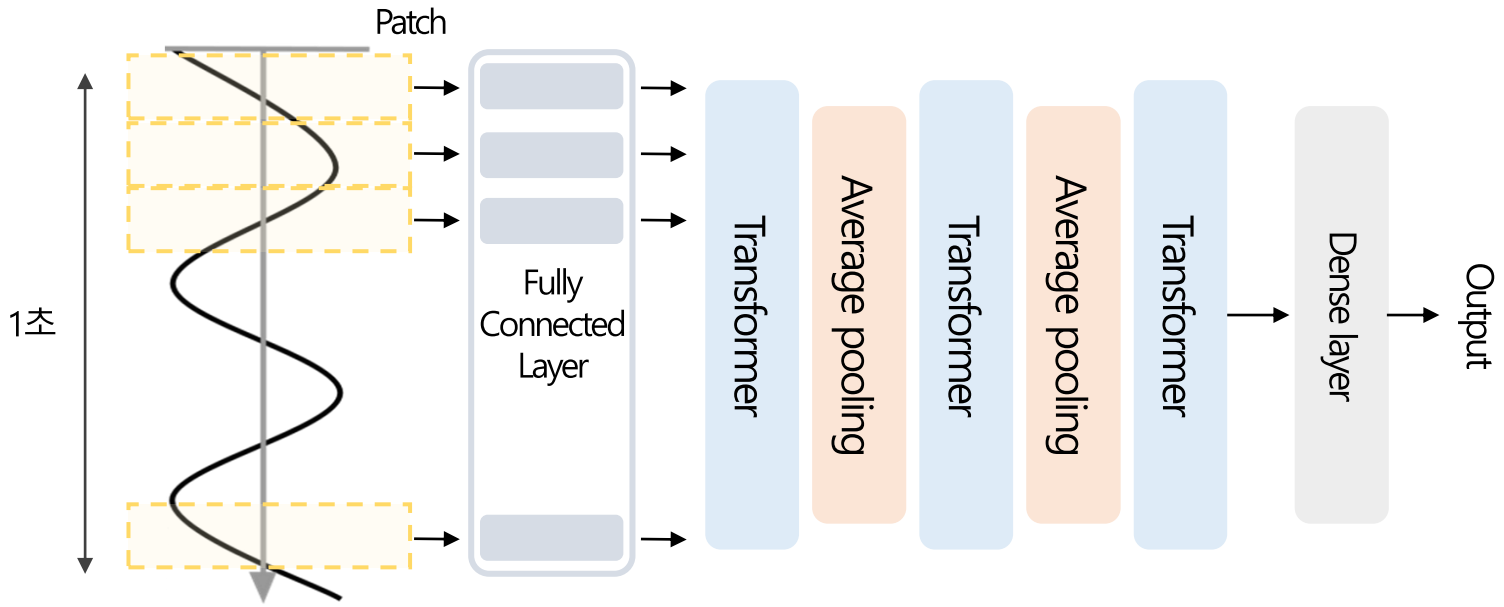


Transformer for Audio data

❖ Audio Transformer



- Audio transformer 동작 구조
  1) Waveform 임베딩
  2) Transformer & Average pooling
  3) Dense layer
     - ➢ Fully connected layer를 통과하여 최종적인 출력 값 산출



Transformer for Audio data

❖ Audio Transformer

- 오디오 클립을 1초 단위로 파싱한 데이터셋에 대해 200개의 class를 분류하는 task를 수행
- CNN에서 좋은 성능을 보였던 pooling layer를 추가함으로써 transformer 학습을 도움
  - ➢ 입력의 크기를 줄여 계산 효율성 증대
  - ➢ 넓은 receptive field를 갖게 되어 hierarchical한 특징을 학습하는데 유리함
- 실험을 통해 pooling을 적용한 transformer가 타 모델에 비해 좋은 성능을 내는 것을 보임

Table 1: *Comparison of various proposed architecture as shown in the table below for mean average precision (mAP) metric. We see how even baseline Transformer architectures without using any convolutional layers can outperform widely used CNN architectures for acoustic scene understanding by significant margins. [21]*

Neural Model Architecture	mAP	# Param
CRNN [21]	0.417	0.96M
VGG-like [21]	0.434	0.27M
ResNet-18 [21]	0.373	11.3M
DenseNet-121 [21]	0.425	12.5M
Small Transformer	0.469	0.9M
Large 6- Layer Transformer	**0.525**	2.3M
Large 6- Layer Transformer with Pooling	**0.537**	2.3M

Pooling에 따른 성능 차이 존재

Transformer for Audio data

❖ AST: Audio Spectrogram Transformer

- MIT에서 연구하였으며(arXiv, 2021), 2022년 3월 31일 기준 46회 인용되었음
- 오디오 분류를 위해 스펙트로그램을 위한 transformer 구조를 제안

AST: Audio Spectrogram Transformer

Yuan Gong, Yu-An Chung, James Glass

MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139, USA

{yuangong, andyyuan, glass}@mit.edu

Abstract

In the past decade, convolutional neural networks (CNNs) have been widely adopted as the main building block for end-to-end audio classification models, which aim to learn a direct mapping from audio spectrograms to corresponding labels. To better capture long-range global context, a recent trend is to add a self-attention mechanism on top of the CNN, forming a CNN-attention hybrid model. However, it is unclear whether the reliance on a CNN is necessary, and if neural networks purely based on attention are sufficient to obtain good performance in audio classification. In this paper, we answer the question by introducing the *Audio Spectrogram Transformer* (AST), the first convolution-free, purely attention-based model for audio classification. We evaluate AST on various audio classification benchmarks, where it achieves new state-of-the-art results of 0.485 mAP on AudioSet, 95.6% accuracy on ESC-50, and 98.1% accuracy on Speech Commands V2.

Index Terms: audio classification, self-attention, Transformer

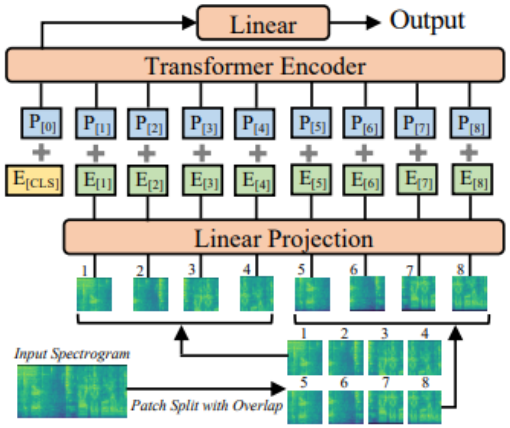
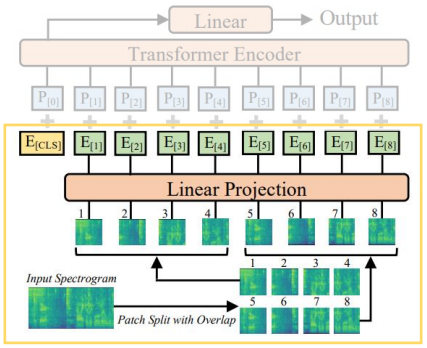


Figure 1: *The proposed audio spectrogram transformer (AST) architecture. The 2D audio spectrogram is split into a sequence*

Transformer for Audio data

❖ Audio Spectrogram Transformer(AST)



- Audio spectrogram transformer 동작 구조

1) 스펙트로그램 임베딩

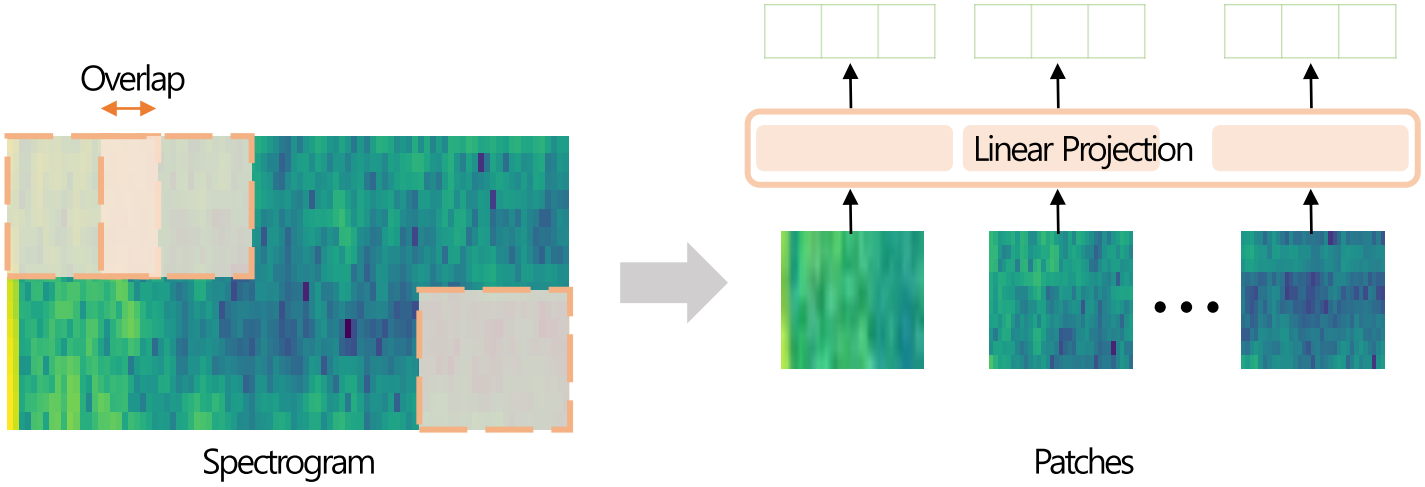
   ➢ 스펙트로그램을 패치로 분할하고 linear를 사용하여 임베딩

   ➢ 16x16 크기의 패치로 6만큼 overlap하여 분할

2) Positional 임베딩

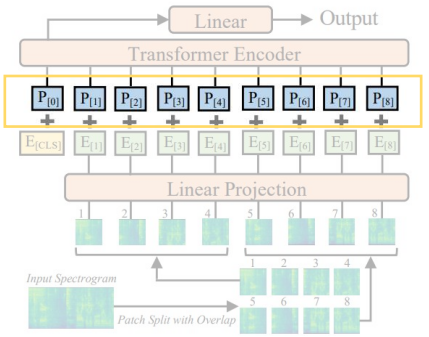
3) Transformer encoder

4) MLP block

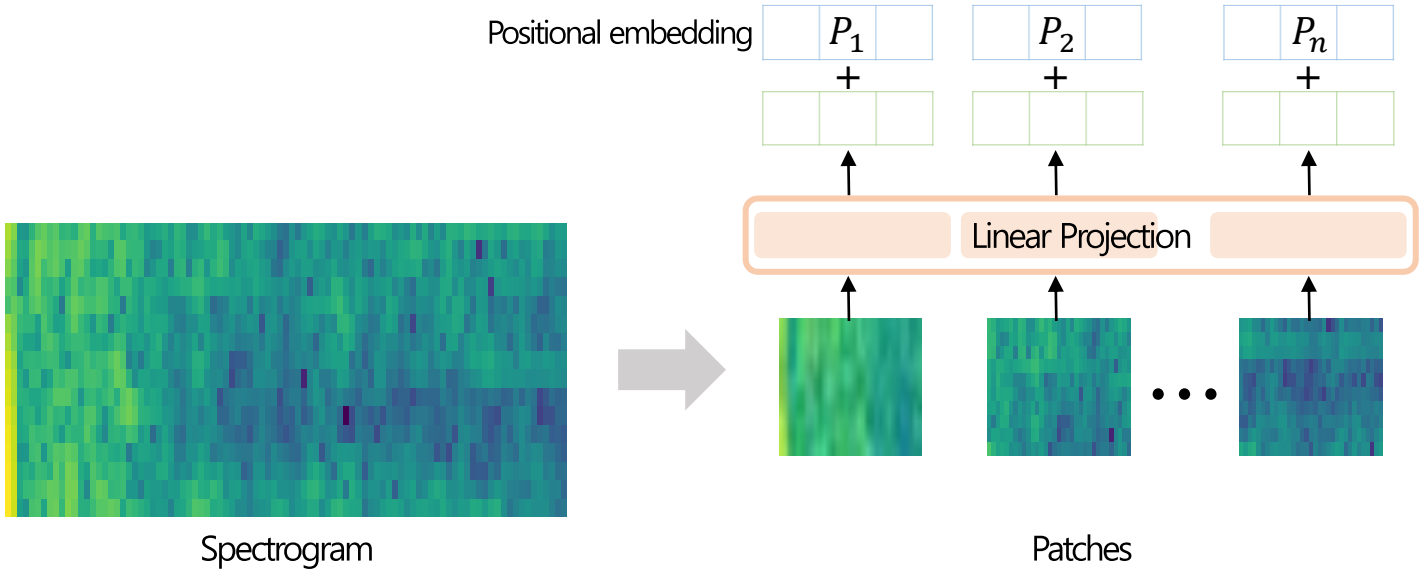


Transformer for Audio data

❖ Audio Spectrogram Transformer(AST)

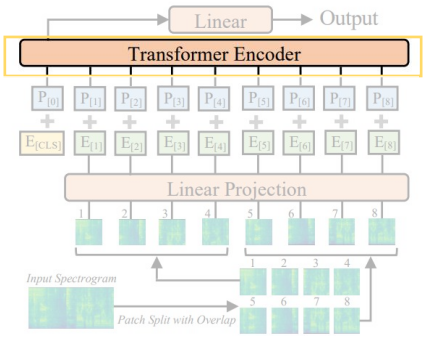


- Audio spectrogram transformer 동작 구조
  1) 스펙트로그램 임베딩
  2) Positional 임베딩
     - ➢ 스펙트로그램에서의 패치 위치 정보를 제공하기 위함
  3) Transformer encoder
  4) MLP block

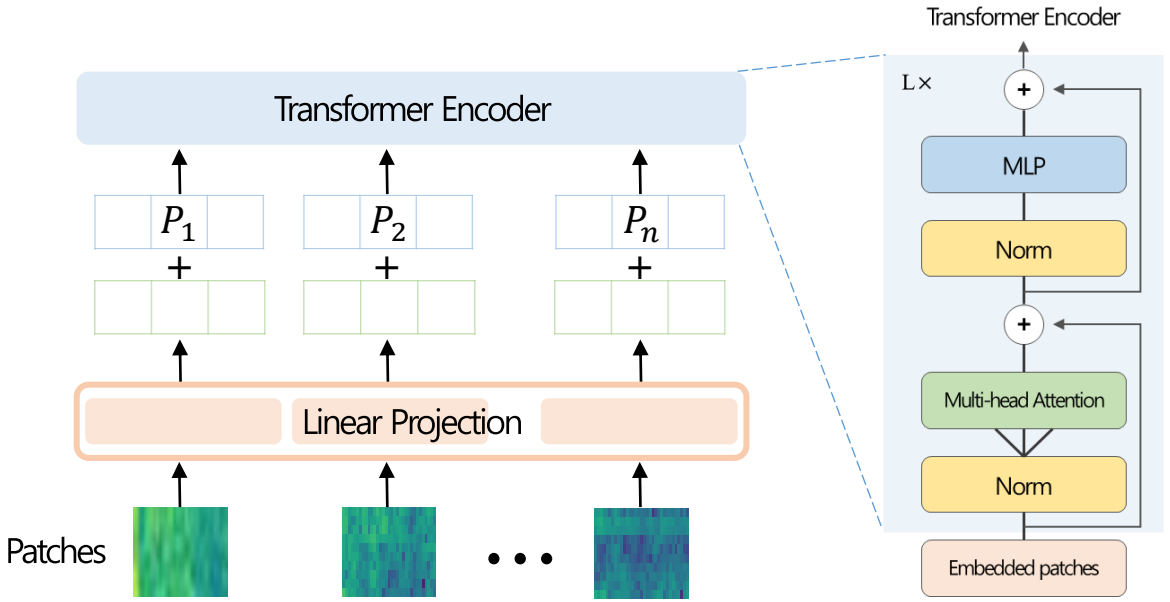


Transformer for Audio data

❖ Audio Spectrogram Transformer(AST)

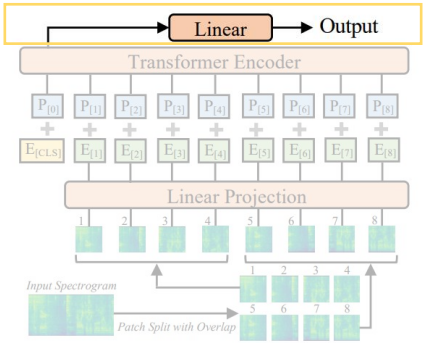


- Audio spectrogram transformer 동작 구조
  1) 스펙트로그램 임베딩
  2) Positional 임베딩
  3) Transformer encoder
     - ➢ Multi-head로 구성된 self-attention 메커니즘 적용
     - ➢ NLP task에서 주로 사용하는 transformer encoder와 normalization 순서에서 차이가 있음
  4) MLP block

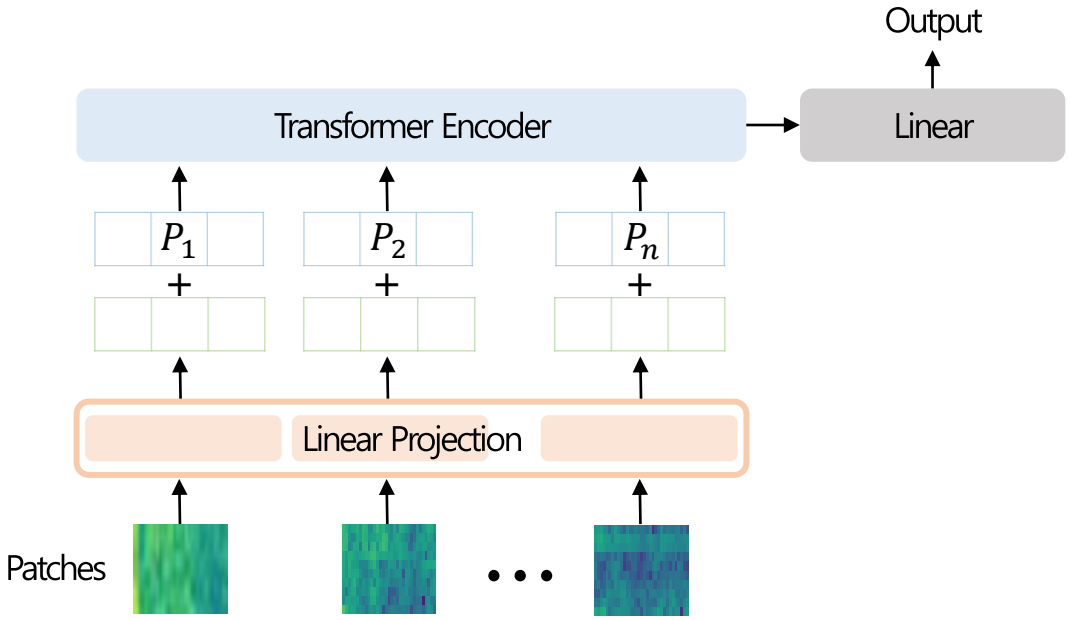


Transformer for Audio data

❖ Audio Spectrogram Transformer(AST)



- Audio spectrogram transformer 동작 구조
  1) 스펙트로그램 임베딩
  2) Positional 임베딩
  3) Transformer encoder
  4) MLP block
     - ➢ MLP block을 거쳐 최종적인 출력 값 산출



Transformer for Audio data

❖ Audio Spectrogram Transformer(AST)

- 10초 길이의 오디오 데이터셋에 대해 527개의 class를 분류하는 task에서 비교 방법론보다 좋은 성능을 보임
- 실험적으로 6만큼 overlap하여 패치를 생성하는 것이 가장 좋은 성능을 냄
- 이 외에도, 패치 크기와 positional 임베딩 사용 여부, 사전 학습 여부에 따른 성능 차이를 보여줌

Table 1: *Performance comparison of AST and previous methods on AudioSet.*

	Model Architecture	Balanced mAP	Full mAP
Baseline [15]	CNN+MLP	-	0.314
PANNs [7]	CNN+Attention	0.278	0.439
PSLA [8] (Single)	CNN+Attention	0.319	0.444
PSLA (Ensemble-S)	CNN+Attention	0.345	0.464
PSLA (Ensemble-M)	CNN+Attention	0.362	0.474
AST (Single)	Pure Attention	0.347 ± 0.001	0.459 ± 0.000
AST (Ensemble-S)	Pure Attention	0.363	0.475
AST (Ensemble-M)	Pure Attention	**0.378**	**0.485**

Table 5: *Performance impact due to various patch overlap size.*

	# Patches	Balanced Set	Full Set
No Overlap	512	0.336	0.451
Overlap-2	657	0.342	0.456
Overlap-4	850	0.344	0.455
Overlap-6 (Used)	1212	0.347	0.459

Transformer for Audio data

❖ Efficient Training of Audio Transformers with Patchout(PaSST)

- Johannes Kepler 대학에서 연구하였으며, 2022년 3월 31일 기준 2회 인용됨
- Audio Transformer의 계산 복잡성과 메모리 문제를 CNN 수준으로 감소시키고자 함

Efficient Training of Audio Transformers with Patchout

Khaled Koutini[1,2], *Jan Schlüter*[1], *Hamid Eghbal-zadeh*[1,2], *Gerhard Widmer*[1,2]

Institute of Computational Perception[1] & LIT AI Lab[2], Johannes Kepler University Linz, Austria

first.last@jku.at

Abstract

The great success of transformer-based models in natural language processing (NLP) has led to various attempts at adapting these architectures to other domains such as vision and audio. Recent work has shown that transformers can outperform Convolutional Neural Networks (CNNs) on vision and audio tasks. However, one of the main shortcomings of transformer models, compared to the well-established CNNs, is the computational complexity. In transformers, the compute and memory complexity is known to grow *quadratically* with the input length. Therefore, there has been extensive work on optimizing transformers, but often at the cost of degrading predictive performance. In this work, we propose a novel method to optimize and regularize transformers on audio spectrograms. Our proposed models achieve a new state-of-the-art performance on Audioset and can be trained on a single consumer-grade GPU. Furthermore, we propose a transformer model that outperforms CNNs in terms of both performance and training speed.[1]

Index Terms: transformers, audio-tagging, attention models, audio classification

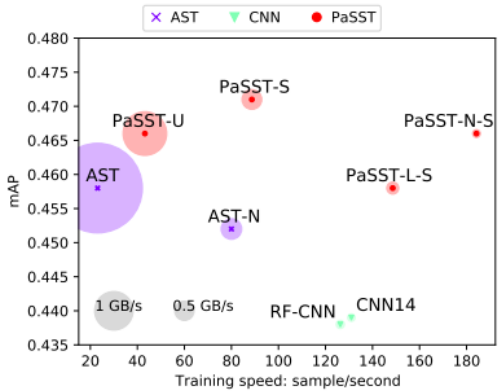
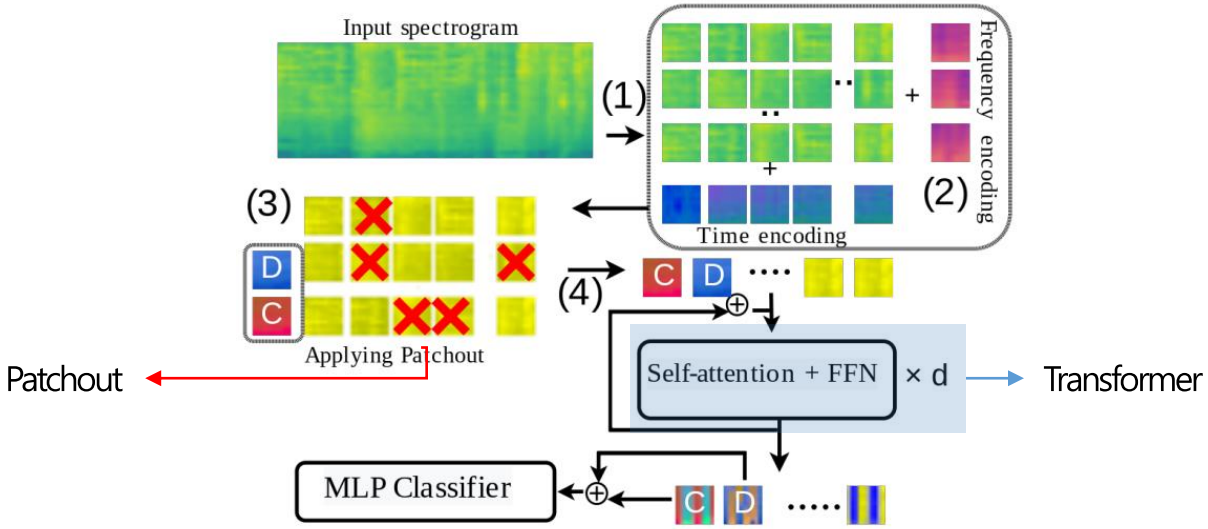


Figure 1: *Performance vs training speed on Audioset. The radius of the circle indicates the required GPU memory per sam-*

Transformer for Audio data

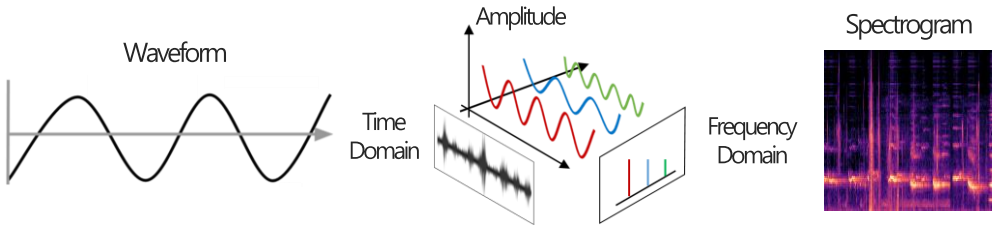
❖ Efficient Training of Audio Transformers with Patchout(PaSST)

- 스펙트로그램에서 패치 생성 시 overlap으로 인한 패치 총 수 증가는 계산 복잡성과 필요한 메모리 증가로 이어짐
- Patchout은 생성된 패치 중 일부를 drop하고 frequency 임베딩과 time 임베딩을 추가하여 보
- Patchout을 통해 전체 패치의 수가 줄고 효율적인 학습 및 정규화가 가능

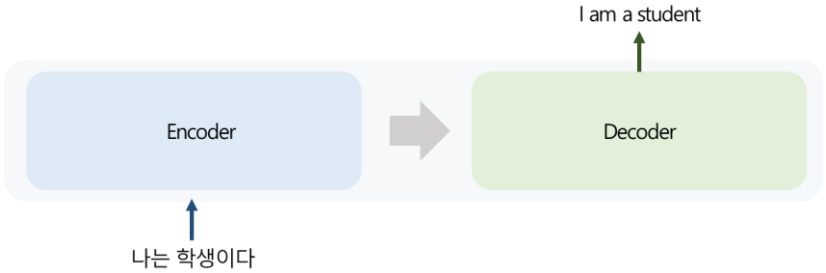


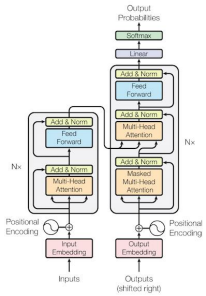
Summary

- **Time domain과 Frequency domain에서의 오디오 데이터**
  - Waveform에서 푸리에 변환을 통해 스펙트로그램 추출

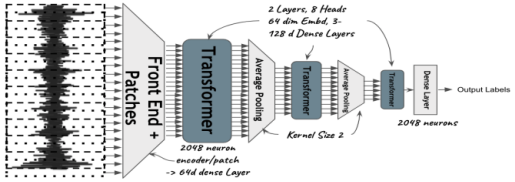


- **Transformer networks 구조**
  - Encoder-Decoder 구조로, sequence의 효율적인 병렬 처리가 가능

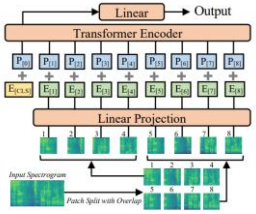




- **Audio Transformer & Audio Spectrogram Transformer**



Waveform을 위한 transformer



Spectrogram을 위한 transformer

Conclusion

- 최근 오디오 데이터를 분석하여 새로운 인사이트를 얻기 위한 다양한 연구가 활발히 수행되고 있음
- 오디오 데이터의 특징 추출을 위해 다양한 기법을 사용함
- 일반적으로 자주 사용하는 CNN과 RNN 뿐만 아니라 다른 도메인에서 좋은 결과를 내는 새로운 알고리즘을 적용하려는 시도가 많음

Reference

[1] Verma, P., & Berger, J. (2021). Audio transformers: Transformer architectures for large scale audio understanding. adieu convolutions. arXiv preprint arXiv:2105.00335.

[2] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[3] Gong, Y., Chung, Y. A., & Glass, J. (2021). Ast: Audio spectrogram transformer. arXiv preprint arXiv:2104.01778.

[4] Koutini, K., Schlüter, J., Eghbal-zadeh, H., & Widmer, G. (2021). Efficient Training of Audio Transformers with Patchout. arXiv preprint arXiv:2110.05069.

감사합니다