AST: Audio Spectrogram Transformer

건국대학교
비전 랩실

김동환

2024.10.29(화)

핵심

Is CNN necessary in audio classification?

처음으로 CNN layer를 제거하고

순수 attention-based model를 도입한 모델

cross-modality transfer learning

서로 다른 유형의 데이터(예: 이미지와 오디오) 간에

학습된 지식을 전이하는 것

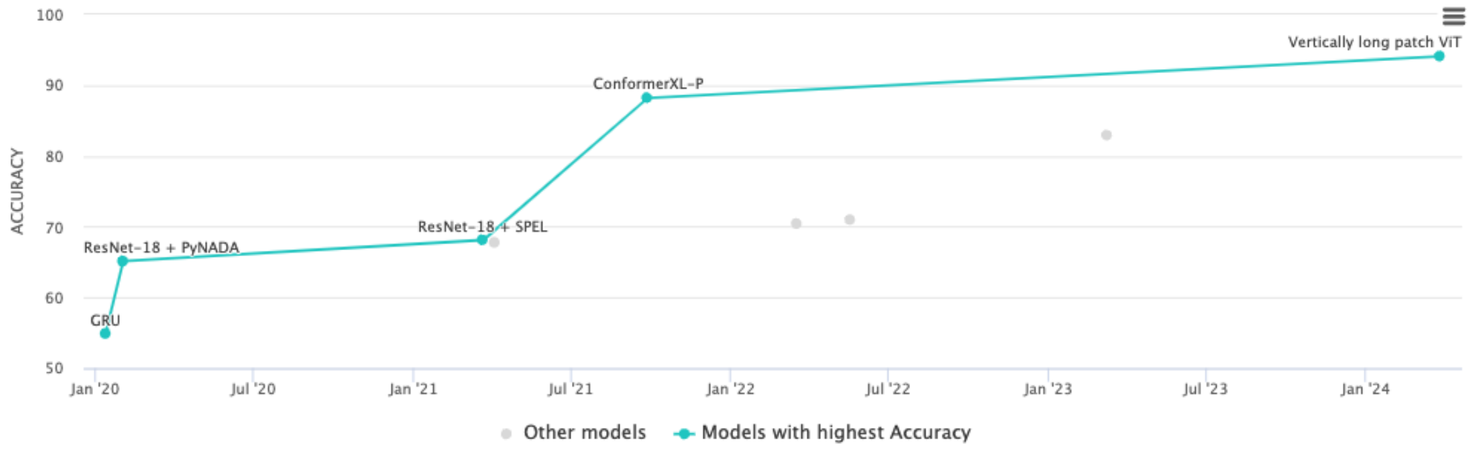
Audio domain 모델을 Image 데이터로

사전학습 시킴으로써 성능 향상

AST: Audio Spectrogram Transformer

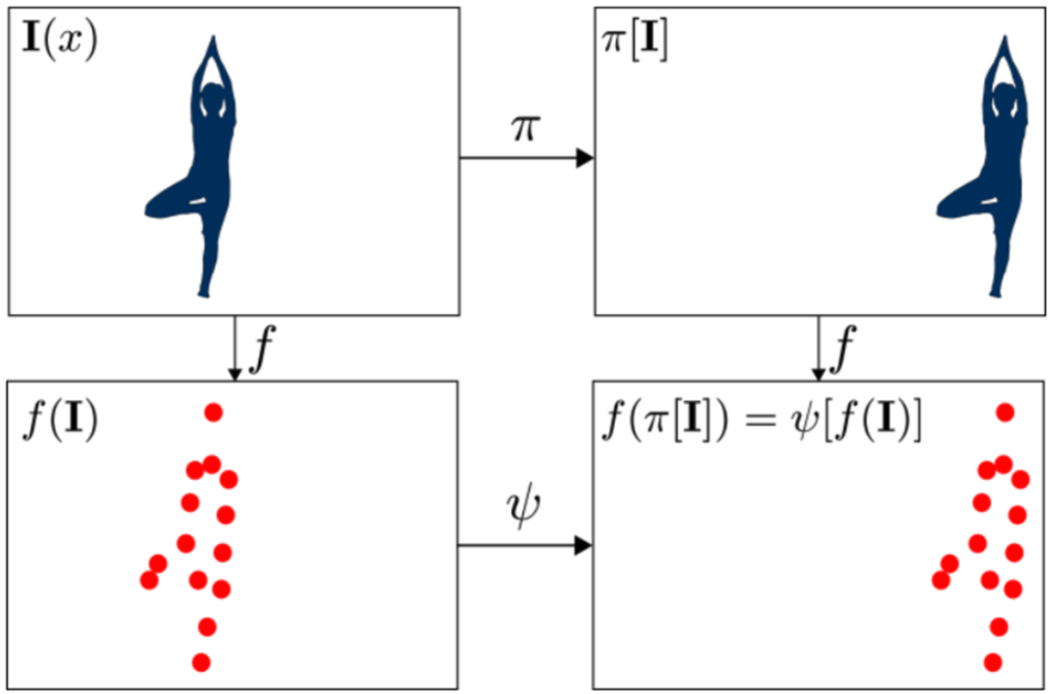
Introduction

음성 도메인에서도 CNN이 주류였던 당시 상황



AST: Audio Spectrogram Transformer

Introduction



Translation Equivariance and Pattern representation Invariance

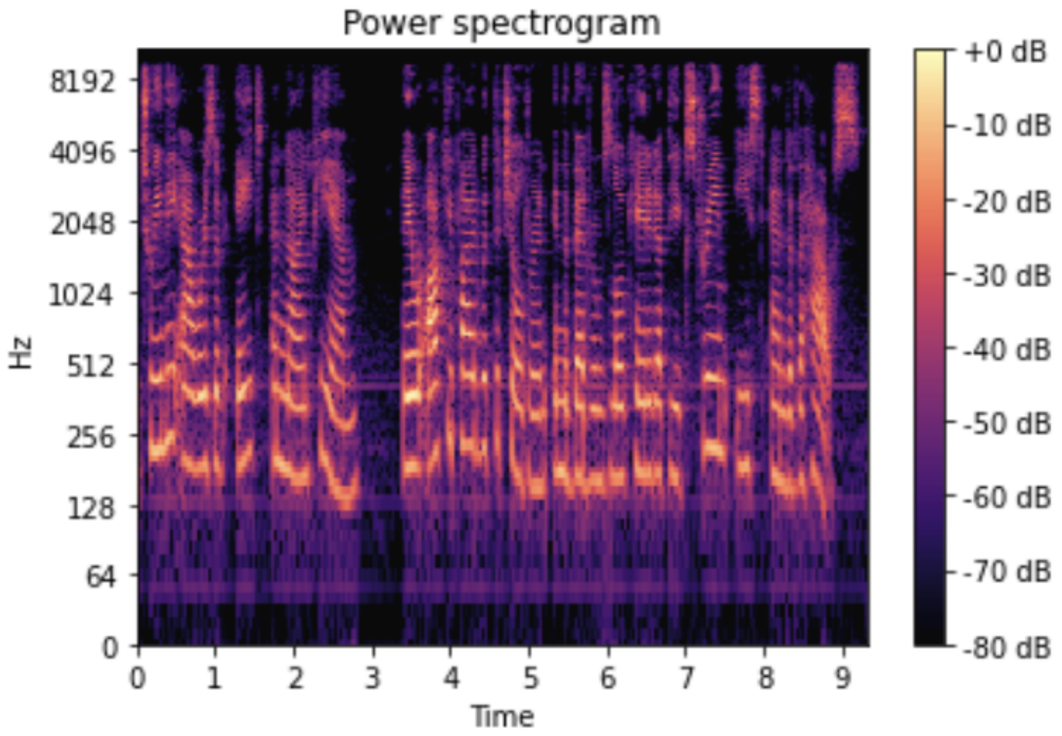
CNN's Inductive bias

Spatial locality : 인접 픽셀은 서로 관련

Translation equivariance : 사물의 위치가 변해도 출력 동일

AST: Audio Spectrogram Transformer

Introduction



CNN's Inductive bias

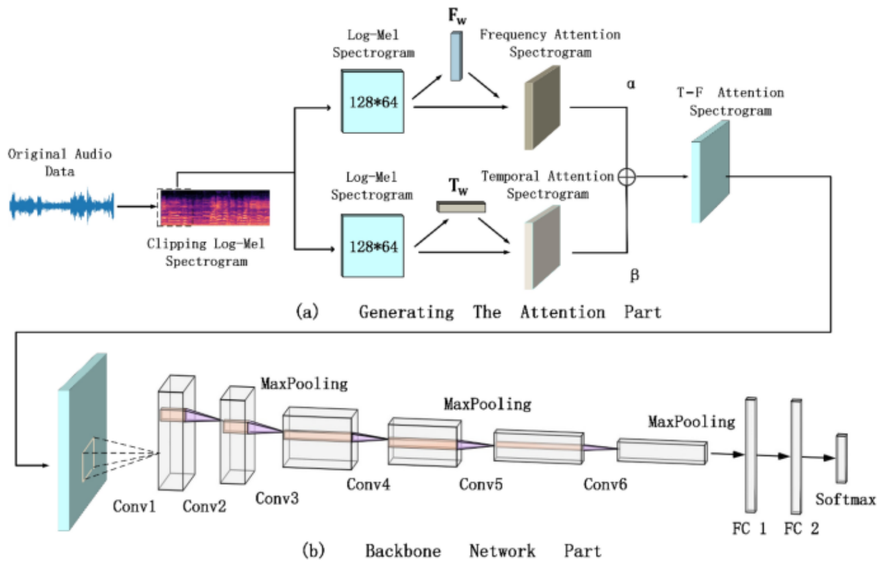
Spatial locality : 인접 픽셀은 서로 관련

Translation equivariance : 사물의 위치가 변해도 출력 동일

AST: Audio Spectrogram Transformer

Related Work

CNN Based model



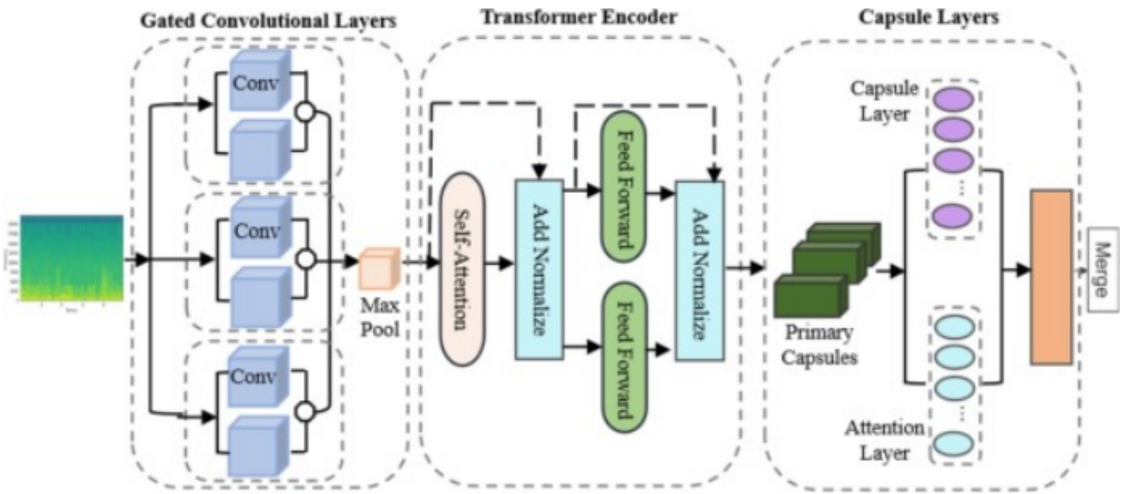
음성데이터가 CNN의 inductive bias에 잘 맞음

- Spatial locality : 인접 픽셀은 서로 관련
- Translation equivariance : 사물의 위치가 변해도 출력 동일

AST: Audio Spectrogram Transformer

Related Work

CNN-Transformer hybrid



weakly labelled data with CNN-transformer

CNN - Transformer hybrid

Long Range Dependency를

Transformer를 도입함으로서 해결

AST: Audio Spectrogram Transformer

Related Work

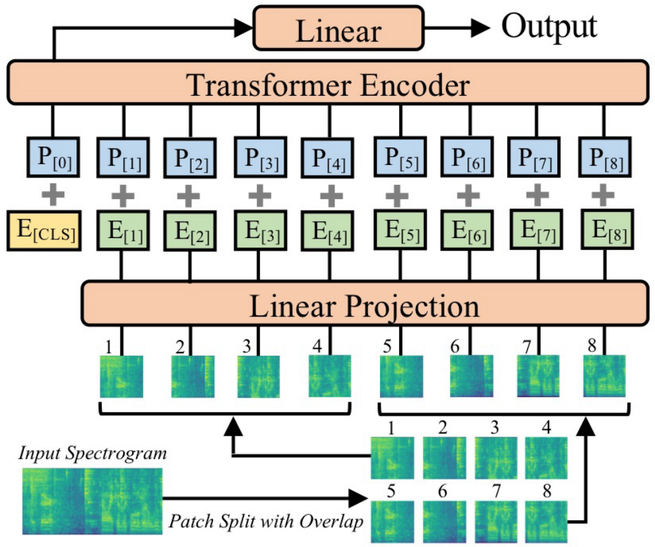


Figure 1: *The proposed audio spectrogram transformer (AST) architecture. The 2D audio spectrogram is split into a sequence of 16×16 patches with overlap, and then linearly projected to a sequence of 1-D patch embeddings. Each patch embedding is added with a learnable positional embedding. An additional classification token is prepended to the sequence. The output embedding is input to a Transformer, and the output of the classification token is used for classification with a linear layer.*

Is CNN necessary in audio classification?

처음으로 CNN layer를 제거하고

순수 attention-based model를 도입한 모델

AST: Audio Spectrogram Transformer

AST (Audio Spectogram Transformer)

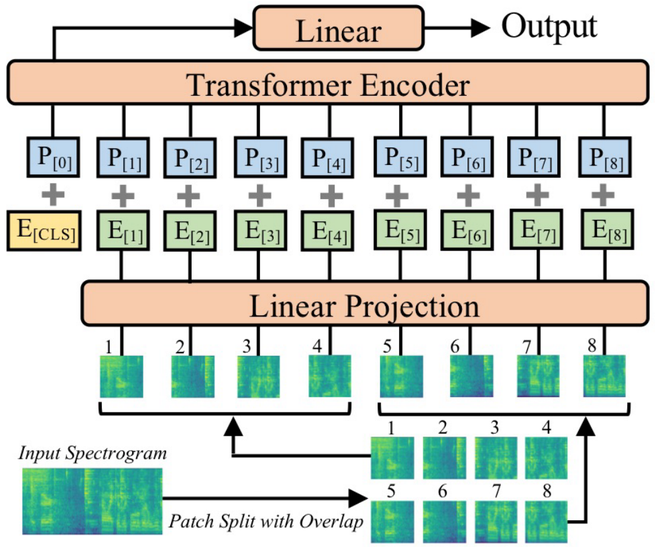


Figure 1: *The proposed audio spectrogram transformer (AST) architecture. The 2D audio spectrogram is split into a sequence of 16×16 patches with overlap, and then linearly projected to a sequence of 1-D patch embeddings. Each patch embedding is added with a learnable positional embedding. An additional classification token is prepended to the sequence. The output embedding is input to a Transformer, and the output of the classification token is used for classification with a linear layer.*

AST의 장점

1. 성능 : SOTA 달성
2. 범용성 : 가변 길이 입력 가능
3. 학습 속도 : CNN-attention hybrid model 보다 빠른 수렴

AST: Audio Spectrogram Transformer

AST : Model Architecture

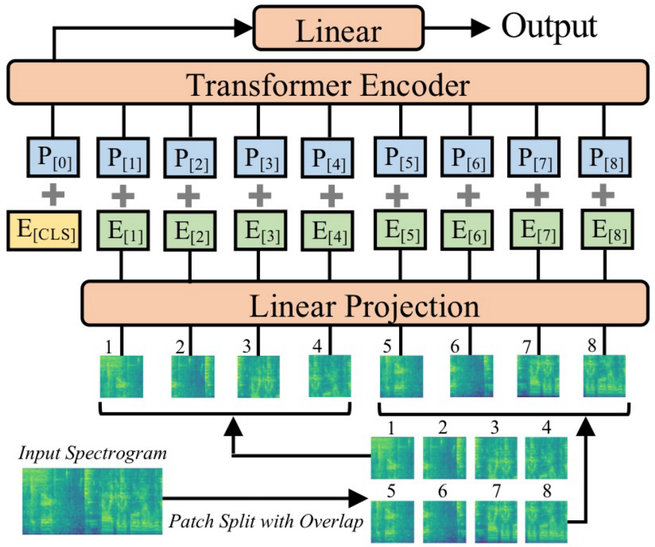
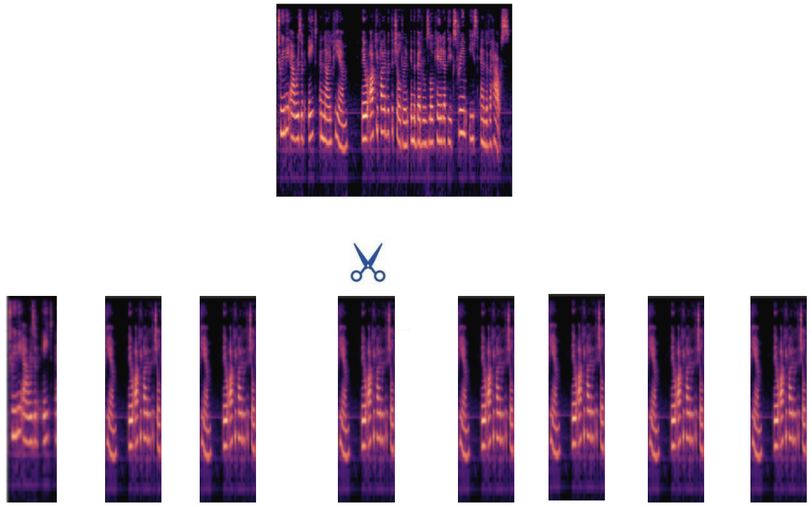


Figure 1: *The proposed audio spectrogram transformer (AST) architecture. The 2D audio spectrogram is split into a sequence of 16×16 patches with overlap, and then linearly projected to a sequence of 1-D patch embeddings. Each patch embedding is added with a learnable positional embedding. An additional classification token is prepended to the sequence. The output embedding is input to a Transformer, and the output of the classification token is used for classification with a linear layer.*

1. 자르기 (중복하여)



AST: Audio Spectrogram Transformer

AST : Model Architecture

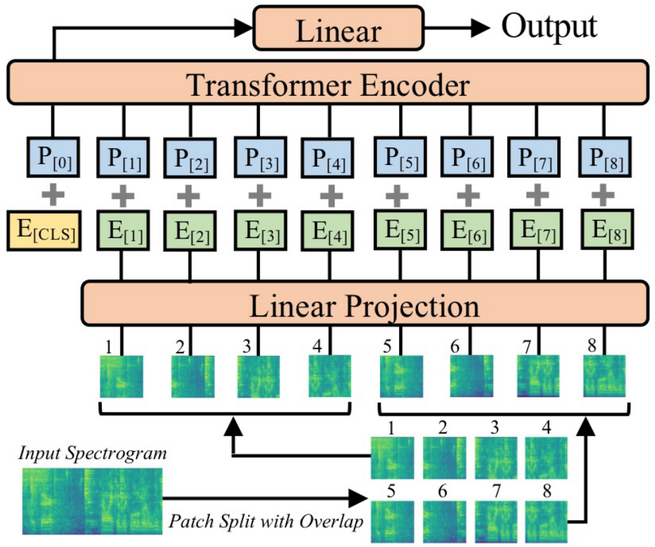
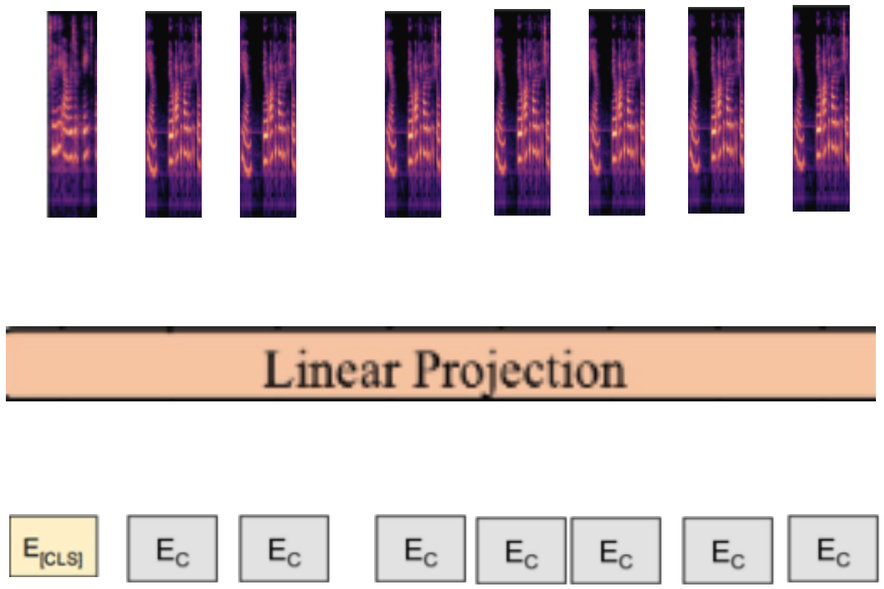


Figure 1: *The proposed audio spectrogram transformer (AST) architecture. The 2D audio spectrogram is split into a sequence of 16×16 patches with overlap, and then linearly projected to a sequence of 1-D patch embeddings. Each patch embedding is added with a learnable positional embedding. An additional classification token is prepended to the sequence. The output embedding is input to a Transformer, and the output of the classification token is used for classification with a linear layer.*

2. 선형 변환



AST: Audio Spectrogram Transformer

AST : Model Architecture

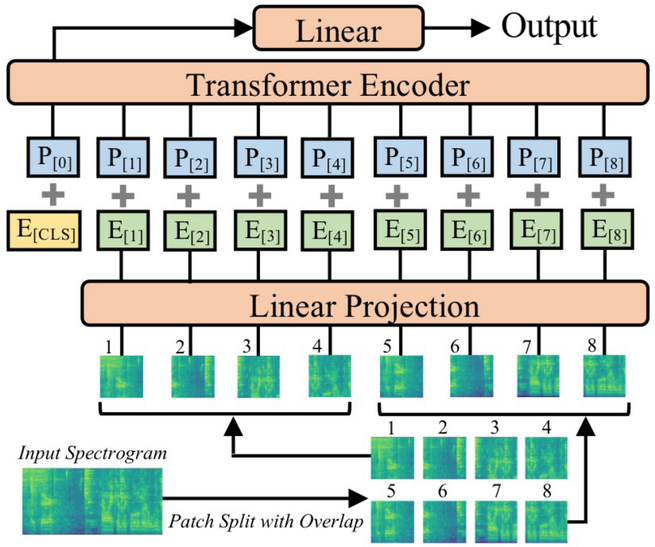
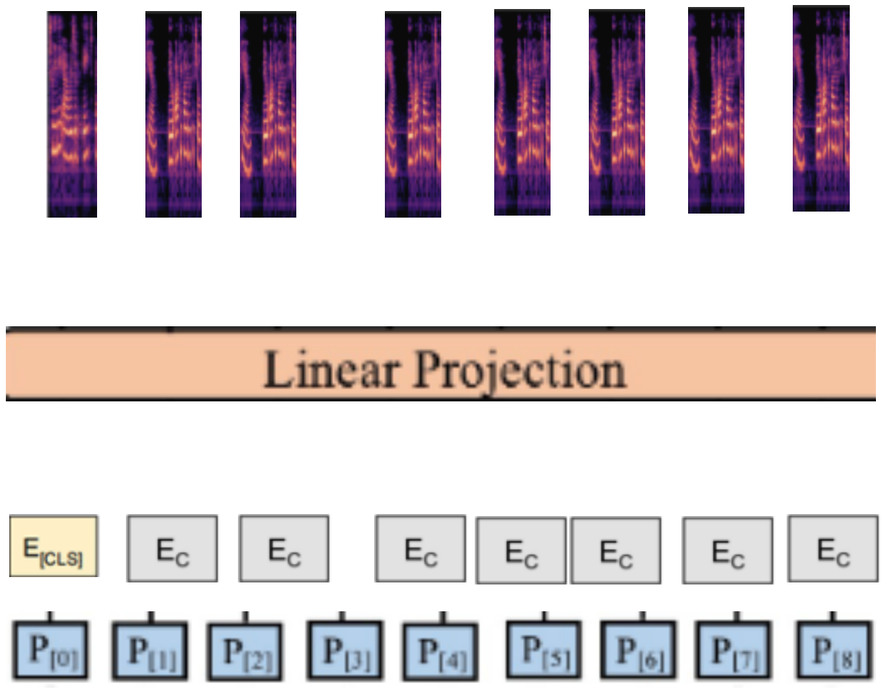


Figure 1: *The proposed audio spectrogram transformer (AST) architecture. The 2D audio spectrogram is split into a sequence of 16×16 patches with overlap, and then linearly projected to a sequence of 1-D patch embeddings. Each patch embedding is added with a learnable positional embedding. An additional classification token is prepended to the sequence. The output embedding is input to a Transformer, and the output of the classification token is used for classification with a linear layer.*

3. Positional Encoding



AST: Audio Spectrogram Transformer

AST : Model Architecture

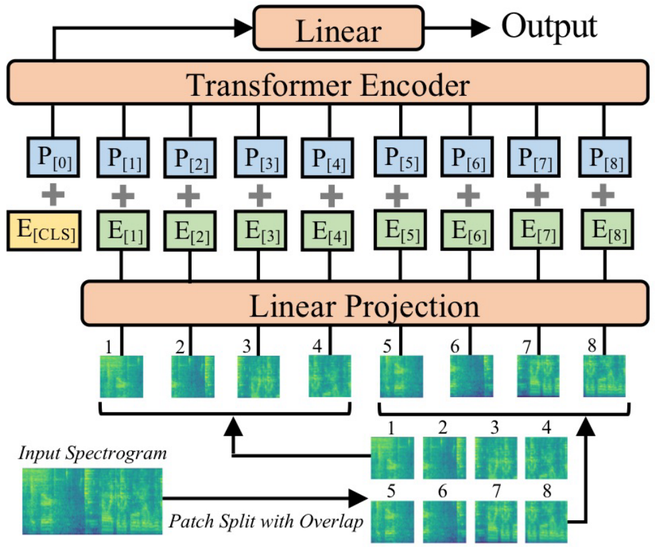
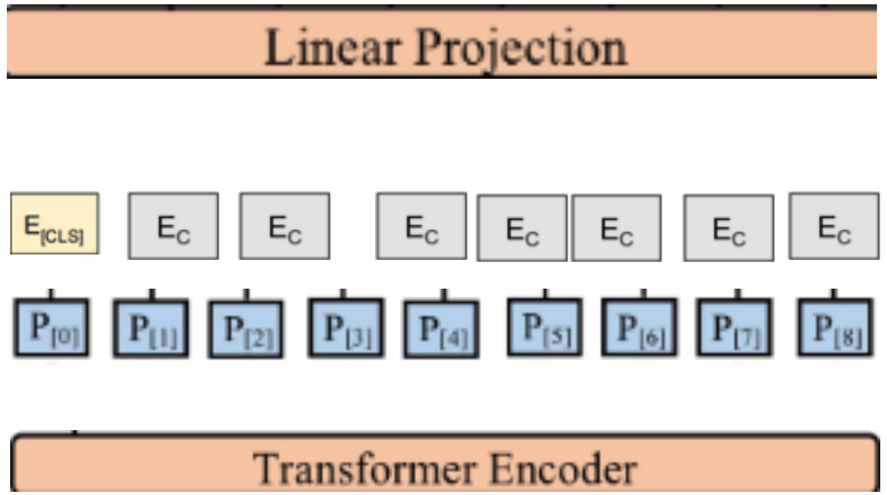


Figure 1: *The proposed audio spectrogram transformer (AST) architecture. The 2D audio spectrogram is split into a sequence of 16×16 patches with overlap, and then linearly projected to a sequence of 1-D patch embeddings. Each patch embedding is added with a learnable positional embedding. An additional classification token is prepended to the sequence. The output embedding is input to a Transformer, and the output of the classification token is used for classification with a linear layer.*

4.Transfomer Encoder에 입력



AST: Audio Spectrogram Transformer

AST : Model Architecture

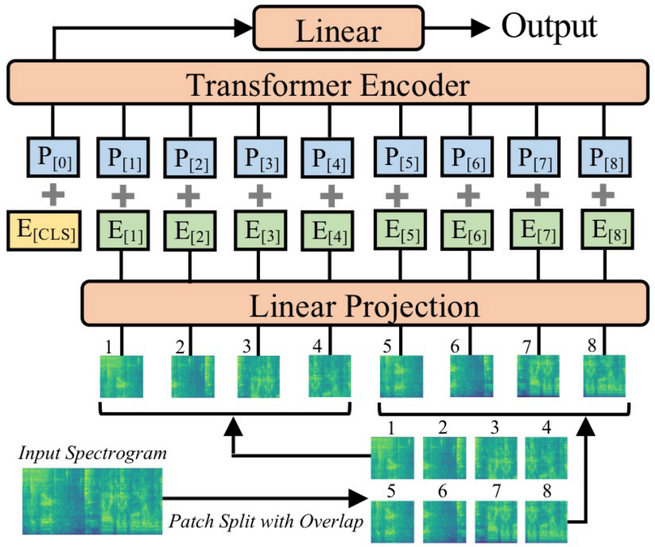
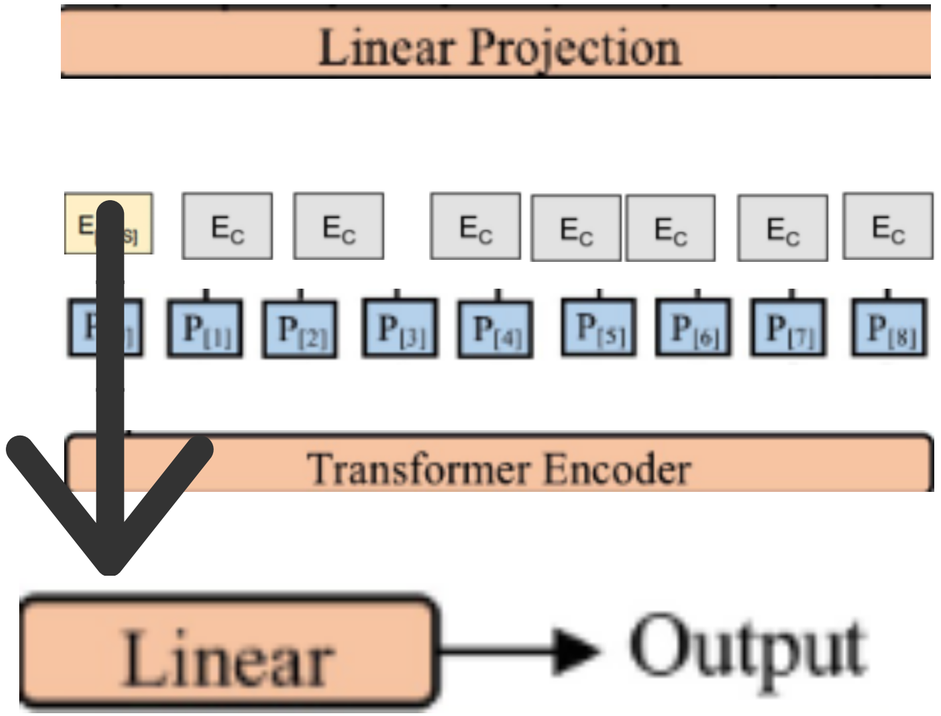


Figure 1: *The proposed audio spectrogram transformer (AST) architecture. The 2D audio spectrogram is split into a sequence of 16×16 patches with overlap, and then linearly projected to a sequence of 1-D patch embeddings. Each patch embedding is added with a learnable positional embedding. An additional classification token is prepended to the sequence. The output embedding is input to a Transformer, and the output of the classification token is used for classification with a linear layer.*

5.CLS 토큰의 잠재 벡터를 선형 변환 -> 결과값 출력



AST: Audio Spectrogram Transformer

AST : ImageNet Pretraining

Transformer is worse than CNN in small dataset

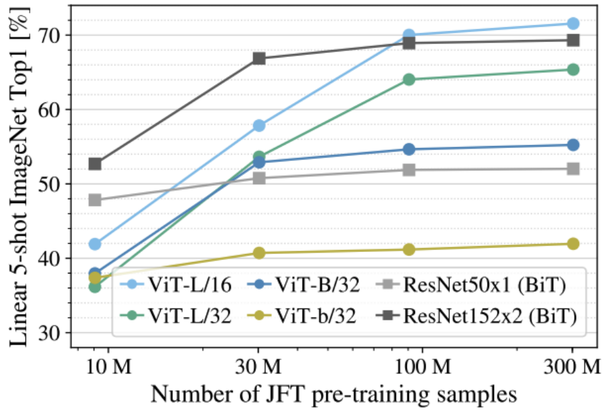


Figure 4: Linear few-shot evaluation on ImageNet versus pre-training size. ResNets perform better with smaller pre-training datasets but plateau sooner than ViT, which performs better with larger pre-training. ViT-b is ViT-B with all hidden dimensions halved.

Vision Transformer

Inductive bias를 줄이는 댓가

모델의 일반성을 얻기 위해
Inductive bias를 희생해야 한다.
=> 더 많은 데이터셋이 필요

문제점 : Audio Dataset 이 그리 많지 않다

AST: Audio Spectrogram Transformer

AST : ImageNet Pretraining

Transformer is worse than CNN in small dataset

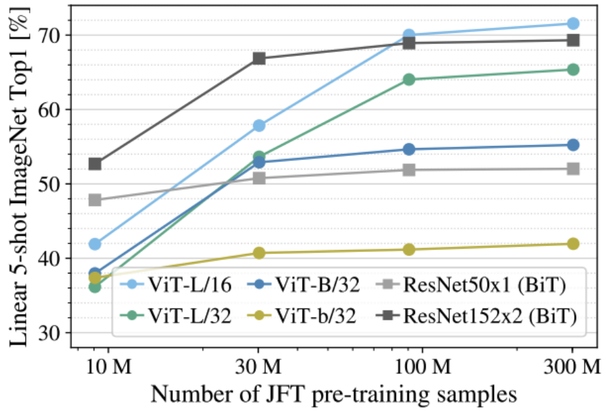


Figure 4: Linear few-shot evaluation on ImageNet versus pre-training size. ResNets perform better with smaller pre-training datasets but plateau sooner than ViT, which performs better with larger pre-training. ViT-b is ViT-B with all hidden dimensions halved.

Vision Transformer

Inductive bias를 줄이는 댓가

모델의 일반성을 얻기 위해
Inductive bias를 희생해야 한다.
=> 더 많은 데이터셋이 필요

문제점 : Audio Dataset 이 그리 많지 않다

해결책 : "cross-modality transfer learning !!"

AST: Audio Spectrogram Transformer

AST : ImageNet Pretraining

Transformer is worse than CNN in small dataset

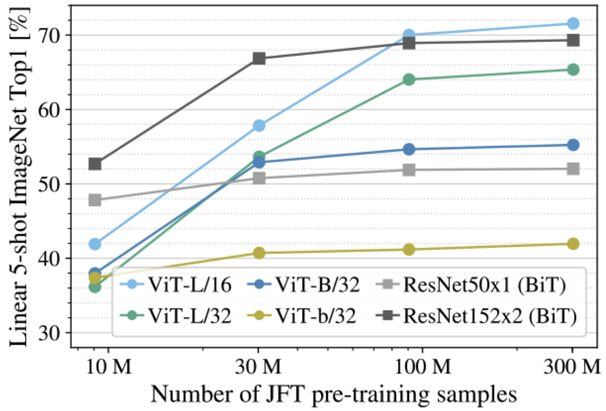


Figure 4: Linear few-shot evaluation on ImageNet versus pre-training size. ResNets perform better with smaller pre-training datasets but plateau sooner than ViT, which performs better with larger pre-training. ViT-b is ViT-B with all hidden dimensions halved.

Vision Transformer

Inductive bias를 줄이는 댓가

모델의 일반성을 얻기 위해
Inductive bias를 희생해야 한다.
=> 더 많은 데이터셋이 필요

문제점 : Audio Dataset 이 그리 많지 않다

해결책 : "cross-modality transfer learning !!"

결론 : "그냥 ViT 가져다 쓰자"

AST: Audio Spectrogram Transformer

AST : ImageNet Pretraining

Transformer is worse than CNN in small dataset

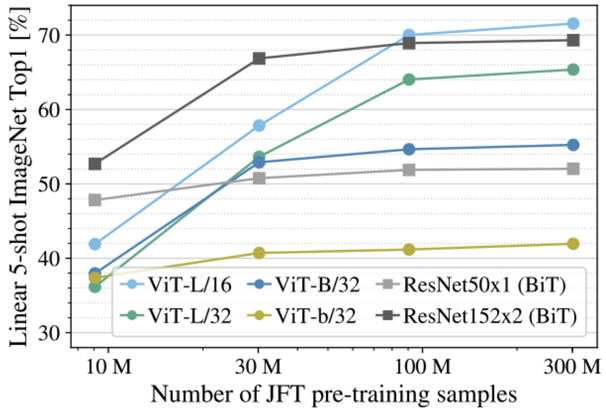


Figure 4: Linear few-shot evaluation on ImageNet versus pre-training size. ResNets perform better with smaller pre-training datasets but plateau sooner than ViT, which performs better with larger pre-training. ViT-b is ViT-B with all hidden dimensions halved.

Vision Transformer

Inductive bias를 줄이는 댓가

모델의 일반성을 얻기 위해
Inductive bias를 희생해야 한다.
=> 더 많은 데이터셋이 필요

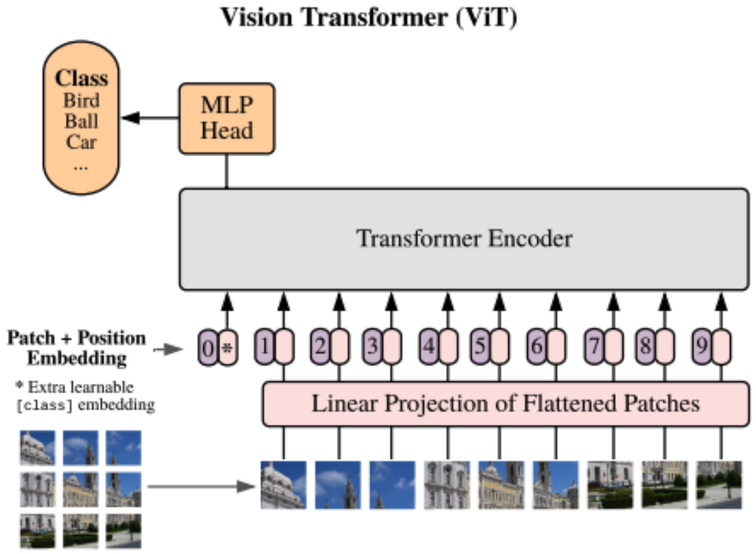
문제점 : Audio Dataset 이 그리 많지 않다

해결책 : "cross-modality transfer learning !!"

결론 : "그냥 ViT 가져다 쓰자"

AST: Audio Spectrogram Transformer

AST : ImageNet Pretraining

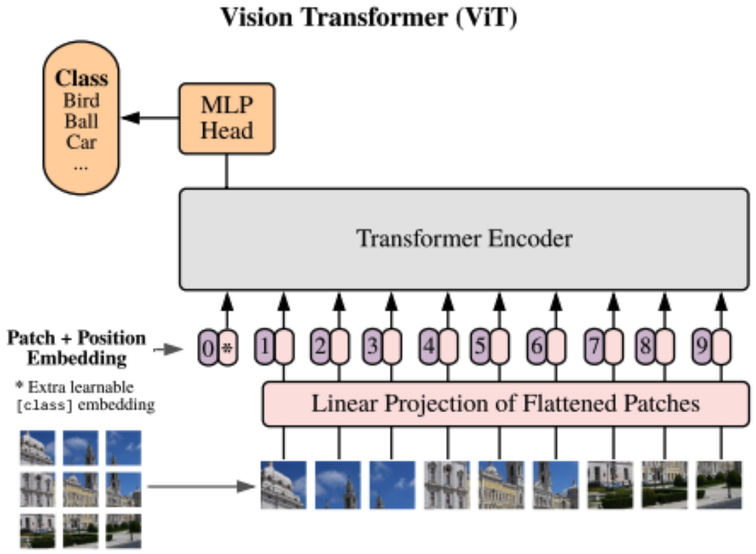


ViT를 약간만 조정 (오디오에 맞게)

1. 채널 수 조정
2. positional encoding 조정
3. 마지막 linear layer 조정

AST: Audio Spectrogram Transformer

AST : ImageNet Pretraining



ViT를 약간만 조정 (오디오에 맞게)

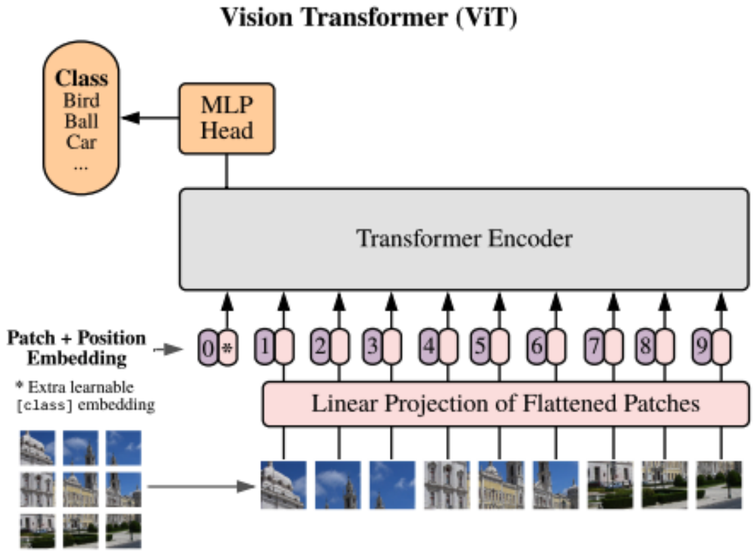
채널 수 조정

ViT의 입력인 이미지는 RGB 3차원이고
AST의 입력인 음성 데이터는 1차원이므로

패치 임베딩을 평균내어 사용

AST: Audio Spectrogram Transformer

AST : ImageNet Pretraining



ViT를 약간만 조정 (오디오에 맞게)

positional encoding 조정

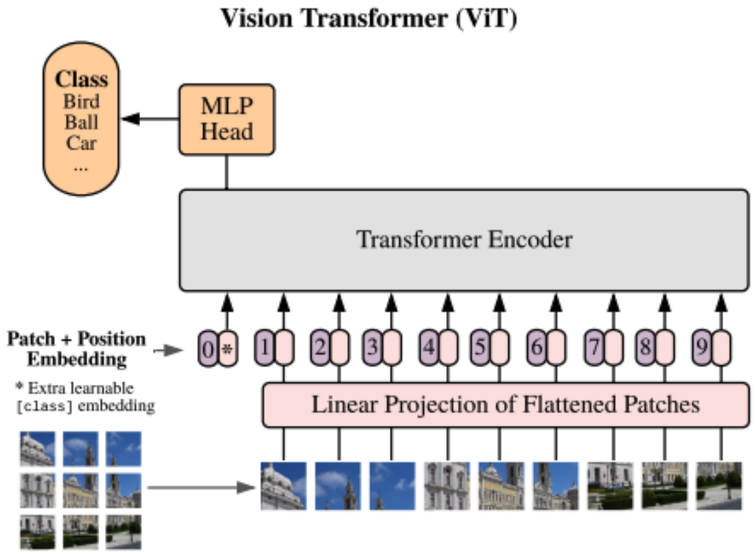
ViT 패치들에 맞는 위치 임베딩은 24×24 형태

AST는 10초 길이의 오디오 데이터를 받아 12×100 개의 패치로 나눔.
따라서 AST의 위치 임베딩 크기는 12×100이 필요

이를 위해 ViT의 24×24 위치 임베딩 중 첫 번째 차원을 12로 자르고, 두 번째 차원을 bi-linear interpolation하여 100으로 변환

AST: Audio Spectrogram Transformer

AST : ImageNet Pretraining



마지막 linear layer 조정

마지막 linear layer 조정

ViT 마지막 출력층은 "이미지 분류"에 사용
AST 마지막 출력층은 "음성 분류"에 사용될 것이므로

마지막 출력층을 새롭게 초기화후 학습

Experiment

"AudioSet" dataset & training detail

유튜브 비디오 10초 클립을 500여개 라벨로 분류

https://research.google.com/audioset/dataset/cat.html

배치 사이즈 : 12, Adam optimizer, BCE 사용, 5e-5 학습률,

매 25에폭마다 절반으로 학습률 감소시킴

"AudioSet" Results

서로 다른 유형의 데이터(예: 이미지와 오디오) 간에

학습된 지식을 전이하는 것

Audio domain 모델을 Image 데이터로

사전학습 시킴으로써 성능 향상

AST: Audio Spectrogram Transformer

Experiment

"AudioSet" dataset

유튜브 비디오 10초 클립을 500여개 라벨로 분류

https://research.google.com/audioset/dataset/cat.html

배치 사이즈 : 12, Adam optimizer, BCE 사용, 5e-5 학습률,

매 25에폭마다 절반으로 학습률 감소시킴

"AudioSet" Results

서로 다른 유형의 데이터(예: 이미지와 오디오) 간에

학습된 지식을 전이하는 것

Audio domain 모델을 Image 데이터로

사전학습 시킴으로써 성능 향상

AST: Audio Spectrogram Transformer

Experiment

"AudioSet" Results

유튜브 비디오 10초 클립을 500여개 라벨로 분류

https://research.google.com/audioset/dataset/cat.html

배치 사이즈 : 12, Adam optimizer, BCE 사용, 5e-5 학습률,

매 25에폭마다 절반으로 학습률 감소시킴

"AudioSet" Results

	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**

AST: Audio Spectrogram Transformer

Ablation Study

Subject

1. Impact of ImageNet Pretraining
2. Impact of Positional Embedding Adapation
3. Impact of Patch Split Overlap
4. Impact of Patch Shape and Size

Impact of ImageNet Pretraining

Table 2: *Performance impact due to ImageNet pretraining. "Used" denotes the setting used by our optimal AST model.*

	Balanced Set	Full Set
No Pretrain	0.148	0.366
ImageNet Pretrain (Used)	0.347	0.459

Ablation Study

Subject

1. Impact of ImageNet Pretraining
2. Impact of Positional Embedding Adapation
3. Impact of Patch Split Overlap
4. Impact of Patch Shape and Size

Impact of Positional Embedding Adapation

Table 4: *Performance impact due to various positional embedding adaptation settings.*

	Balanced Set
Reinitialize	0.305
Nearest Neighbor Interpolation	0.346
Bilinear Interpolation (Used)	0.347

Ablation Study

Subject

1. Impact of ImageNet Pretraining
2. Impact of Positional Embedding Adapation
3. Impact of Patch Split Overlap
4. Impact of Patch Shape and Size

Impact of Patch Split Overlap

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

Ablation Study

Subject

1. Impact of ImageNet Pretraining
2. Impact of Positional Embedding Adapation
3. Impact of Patch Split Overlap
4. Impact of Patch Shape and Size

Impact of Patch Shape and Size

Table 6: *Performance impact due to various patch shape and size. All models are trained with no patch split overlap.*

	# Patches	w/o Pretrain	w/ Pretrain
128×2	512	0.154	-
16×16 (Used)	512	0.143	0.336
32×32	128	0.139	-

Other Dataset

ESC, Speech Commands

	ESC-50	Speech Commands V2 (35 classes)
SOTA-S	86.5 [33]	97.4 [34]
SOTA-P	94.7 [7]	97.7 [35]
AST-S	88.7±0.7	**98.11±0.05**
AST-P	**95.6±0.4**	97.88±0.03

-S : ImageNet 사전 학습 + audio data로 학습 X
-P : ImageNet 사전 학습 + audio data로 학습 O

결론 : AST는 다른 데이터셋에서도 SOTA

AST: Audio Spectrogram Transformer

느낀점

cross modelity transfer learning -> 우리 연구에도 학습 가능

실험 설정 확인 : ex) 매 25에폭마다 절반으로 학습률 감소시킴

레퍼 찾아보는 것

AST: Audio Spectrogram Transformer