

Generalization 강화와 Overfitting 해결을 통한 Fake Voice Detection

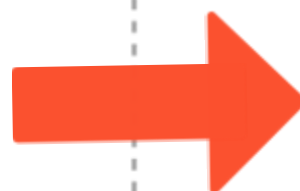
SW중심대학 디지털 경진대회_SW와 생성AI의 만남 : AI 부문

2024.08.06

데이터 분석

① 초기 Data 통계

- Train Data
 - 55,438개, 평균 3.1초
- [Real] Label Data
 - 27,818개, 평균 4.53초
- [Fake] Label Data
 - 27,620개, 평균 1.68초



② 데이터 시간 조정

모든 Audio Data → 5초로 정규화

- 모델의 시간 민감도 문제를 해결하기 위함
- Model이 Audio Data 길이에 민감하지 않도록 하여, 더 일관된 학습 성능을 기대할 수 있게 되었음

전처리 방법

1

MixUp과 Noise Data 생성을 통한 label 추가

: [1, 1] label & [0, 0] label 추가

2

음성길이 조정 후 Data Smampling

: 음성 Data를 모두 5초로 처리 후 label별 Data를 각각 2만개씩 추출

3

CleanUNet를 이용한 Data 처리

: 사람 목소리를 제외한 음성 제거

Noise Data 생성 이유

Test Data Set에 음성이 안 들어가고, 환경 소음만 들어간 Data가 존재



- (fake:0, real:0)인 label을 갖는 Data 학습 필요
→ But, 외부 Data 사용 불가로 인해 "Noise 데이터 생성"하기로 결정

Noise Data 생성 이유

▶ Noise Data 생성에 사용된 기술

- ㉠ 다양한 Noise 생성
ex) white, pink, brown noise 등
- ㉡ 진폭 변조
- ㉢ Chirp(처프) 신호 추가
- ㉣ Mask 적용



총 1만개의 Noise Data 생성

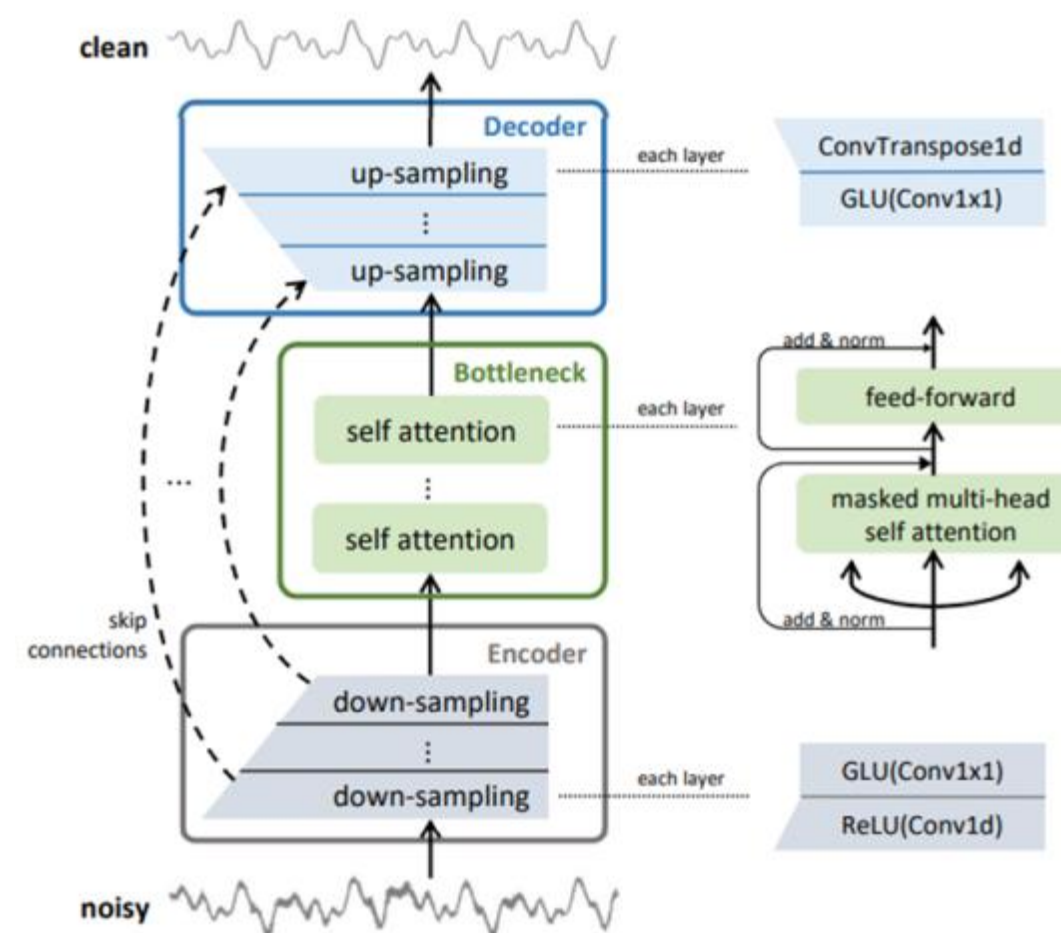
[추가 작업] CleanUNet

① CleanUNet

: 음성을 제외한 Noise, 환경 소음을 제거하기 위한 모델

② 원리

: 원래 해상도를 유지하며,
중요한 특징을 추출하고 다시 복원하는 원리를 사용함
→ 중요한 음성 특징만 남기고, Noise와 환경 소음 제거



출처: Kong, Zhifeng, et al. "Speech Denoising in the Waveform Domain with Self-Attention," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2022): 7867-7871.

03. Model Selection

모델 선택

Model 1

RawNet2

음성 파일에서 CNN-LSTM을 사용하여
feature 추출 후
MLP로 분류 작업을 하는 Model

Model 2

AASIST

RawNet2 기반 인코더에 Graph
Attention Network를 더한 Model



Model 3

Wav2vec 2.0

Transformer 아키텍처를 이용한
음성 임베딩 Model

03. Model Selection

모델 선택

Model 3

Wav2vec 2.0

(단점 1)

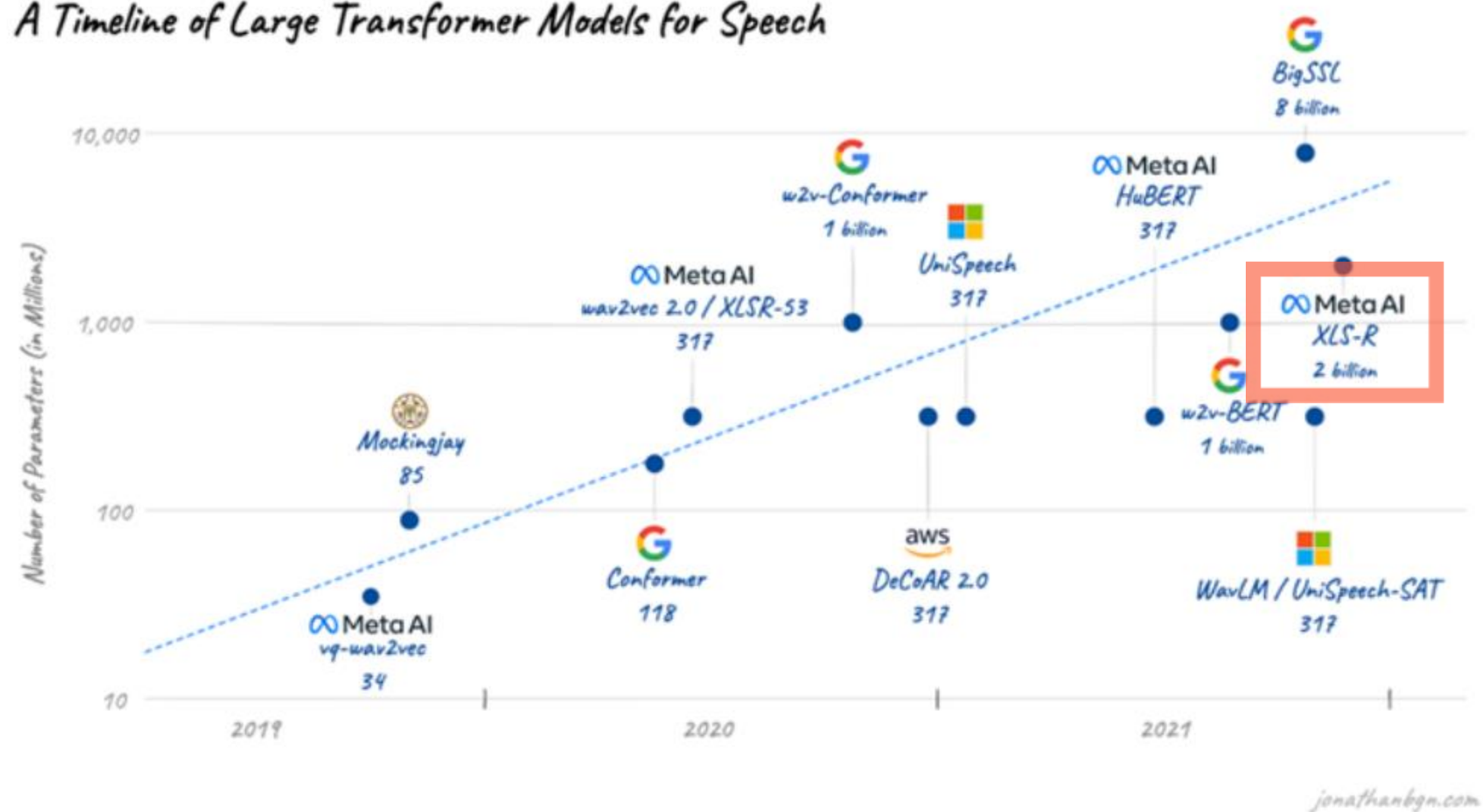
: 성능이 좋을 것으로 예상되나,
모델이 너무 큼

(단점 2)

: 작은 모델의 경우 성능이 좋지 않았음



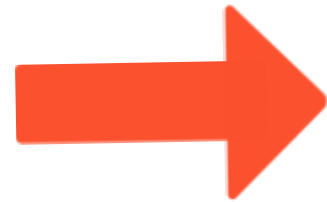
A Timeline of Large Transformer Models for Speech



03. Model Selection

모델 선택

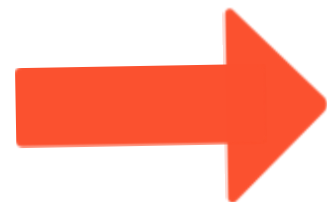
Model 1
RawNet2



1026179 rawnet2_aug_submit1.csv
rawnet2_aug10_2ep edit

성능: 0.3526803212

Model 2
AASIST



1026768 aasist_w_augS7_submit7.csv
aasist_w_augS7_8ep_1e-5 edit

성능: 0.2731644615

성능 향상
약 22.55%



모델 최종 선택

① AASIST란?

: RawNet2 기반 인코더에 Graph Attention Network를 더한 모델
→ 사전 학습 모델 사용

② AASIST 특징

: 주파수 및 시간 특징을 동시에 포착

㉠ 주파수 Feature

: 특정 주파수에서 이상한 패턴을 포착

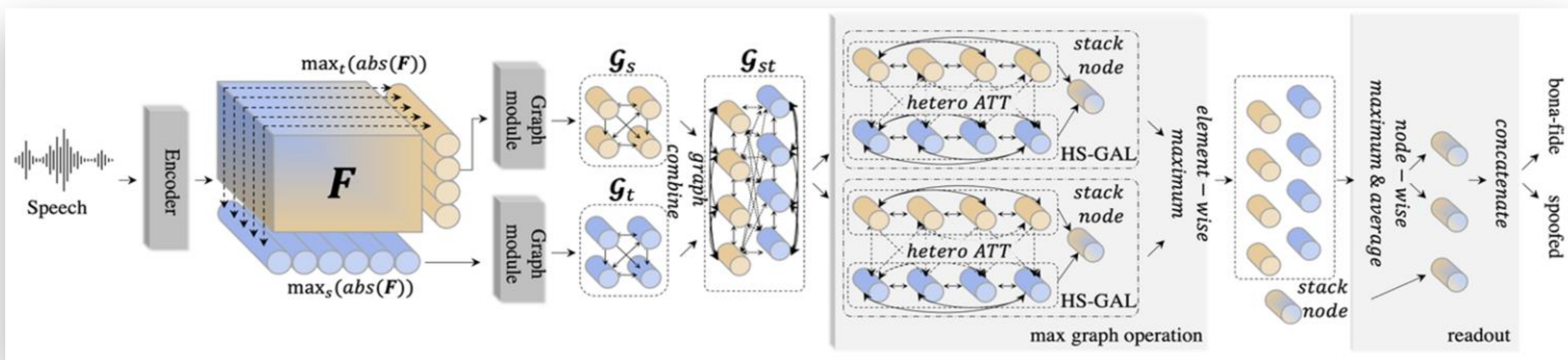
㉡ 시간 Feature

: 시간에 따라 이상한 변화를 포착

03. Model Selection

모델 최종 선택

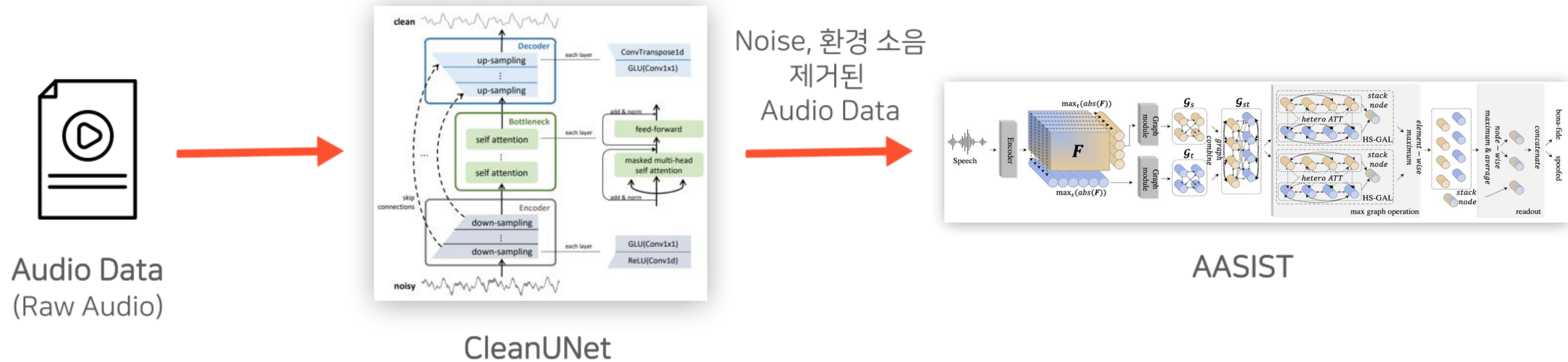
③ AASIST Model Pipeline



출처: Kong, Zhifeng, et al. "Speech Denoising in the Waveform Domain with Self-Attention."
ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2022): 7867-7871.

04. Model Flow

모델 흐름도



05. Calibration

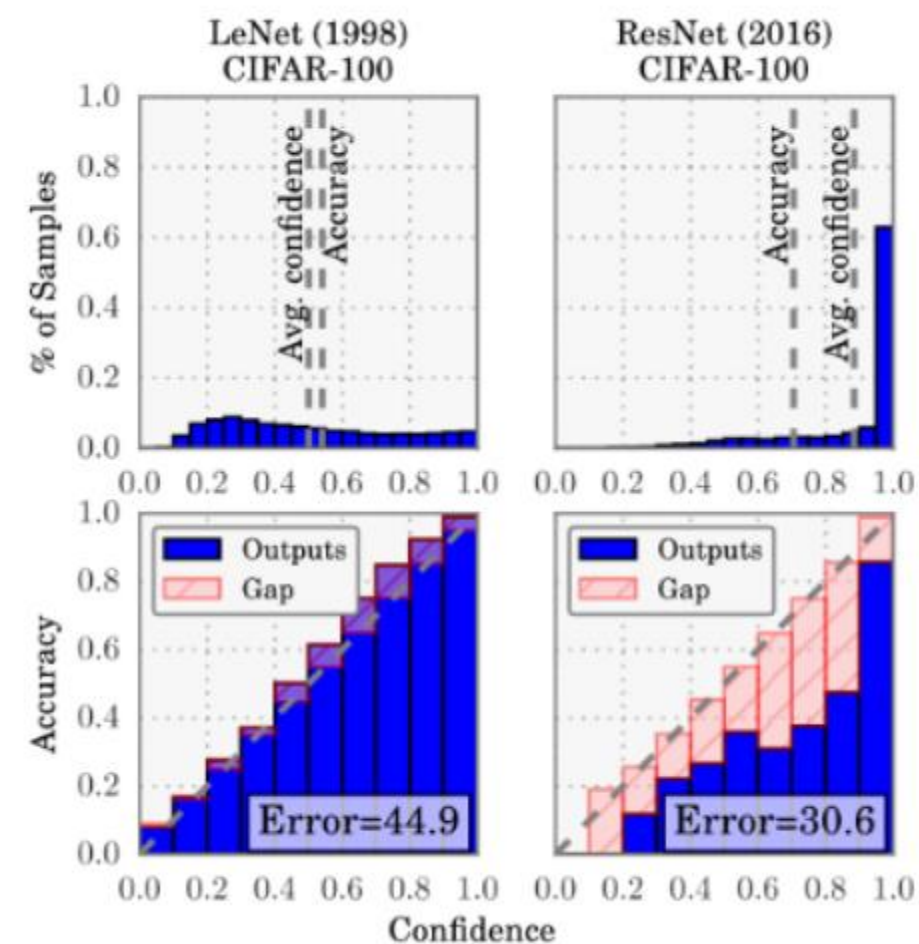
Label Smoothing

① Label Smoothing이란?

: Model이 과도하게 확신하는 것을 방지하기 위해
정답 label을 약간 부드럽게 만드는 기법

② 사용 목적

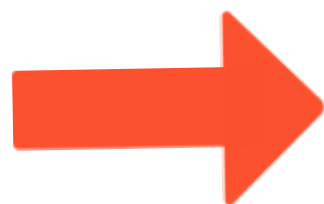
: 훈련 과정에서 Hard encoding이 아닌 Soft encoding을 통해서
overconfidence를 해결하기 위해서 사용함



출처: On Calibration of Modern Neural Networks
Chuan Guo, Geoff Pleiss, Yu Sun, Kilian Q. Weinberger

Label Smoothing

적용 전



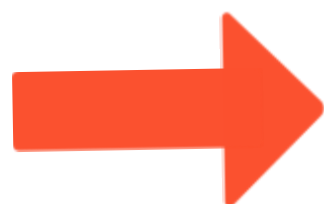
1029370 RawBoost_aasist_w_f1_augSS7_submit6.csv
edit

성능: 0.2106135711

[Alpha = 0.2] 기준



적용 후



1030833 Double_Smoothed_with_Epsilon_0_2_Label_Data_13.csv
edit

성능: 0.2038972239



성능 향상
약 3.19%

06. Applicability

적용 가능성 [모델 파라미터 수]

≐ 85.4 %

Wav2vec에 비해 CleanUNet + AAsist의
Parameter 수가 약 85.4% 더 적음

Model 1

AASIST

0.3M

Model 2

CleanUNet

46M

Model 3

AASIST + CleanUNet

46.3M

Model 4

Wav2vec 2.0

Small: 317M

Large: 1B

제출 점수

① Public Score

PUBLIC

PRIVATE

순위기준




● WINNER

● 1%

● 4%

● 10%

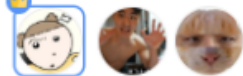
전체 랭킹 >

#	팀	팀 멤버	점수	제출수	등록일
11	VoiceWizards	  	0.19854	65	8시간 전

07. Score

제출 점수

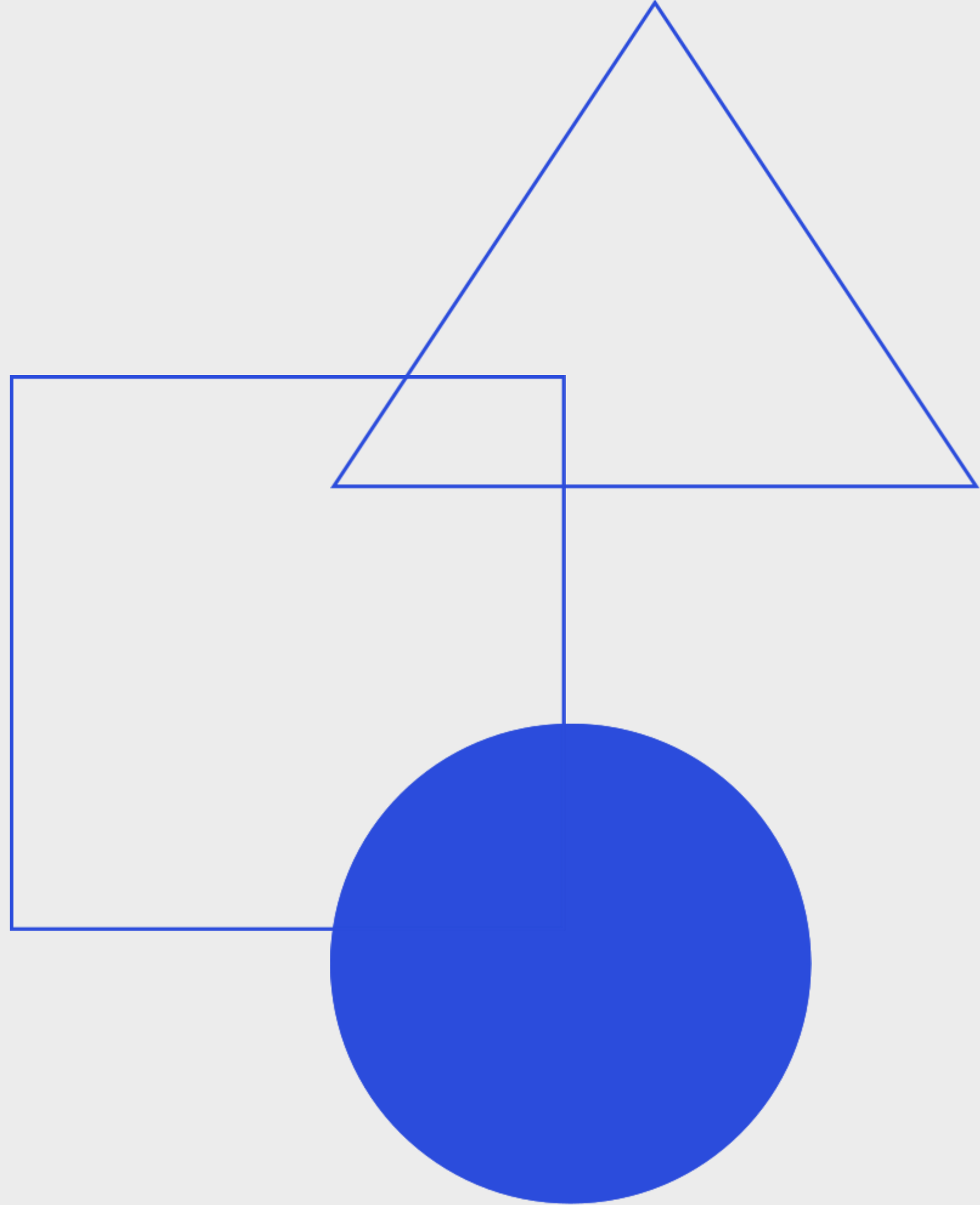
② Private Score

PUBLIC		PRIVATE		순위기준		
● WINNER		● 1%		● 4%		● 10%
#	팀	팀 멤버	최종점수	제출수	등록일	
11	VoiceWizards		0.19839	65	8시간 전	

➡ Public Score에 비해 Private Score가 좋게 나옴 (일반화가 잘 됨)

➡ 12팀 가운데 유일하게 Public Score보다 Private Score가 더 좋게 나옴

감사합니다



Data Sampling

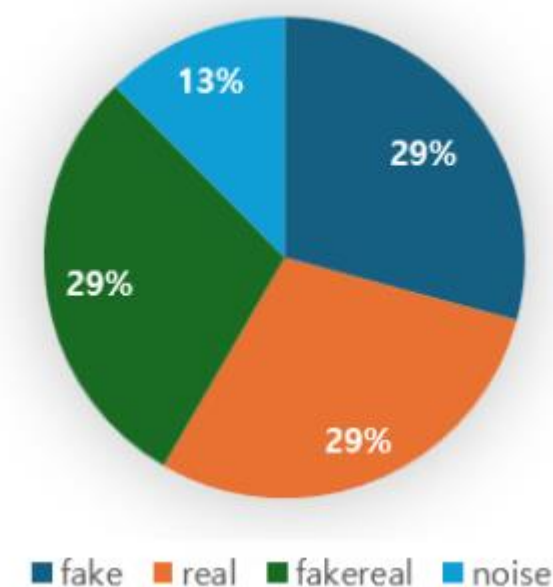
① Train Data set

: 각 클래스를 비슷한 비율로 샘플링

→ 특정 Class를 덜 뽑거나, 더 뽑으면 과소적합 혹은 과적합이 발생할 수 있음

- Fake: 2만개
 - 화자 1명: 1만개, 화자 2명 1만개
- Real: 2만개
 - 화자 1명: 1만개 & 화자 2명: 1만개
- FakeReal: 2만개
 - Fake & Real 화자 1명씩
- Noise: 8천개

Label 개수



Data Sampling

② Validation Data set

- Fake, Real, FakeReal: 1만개
- Noise: 2천개

MixUp

① MixUp이란?

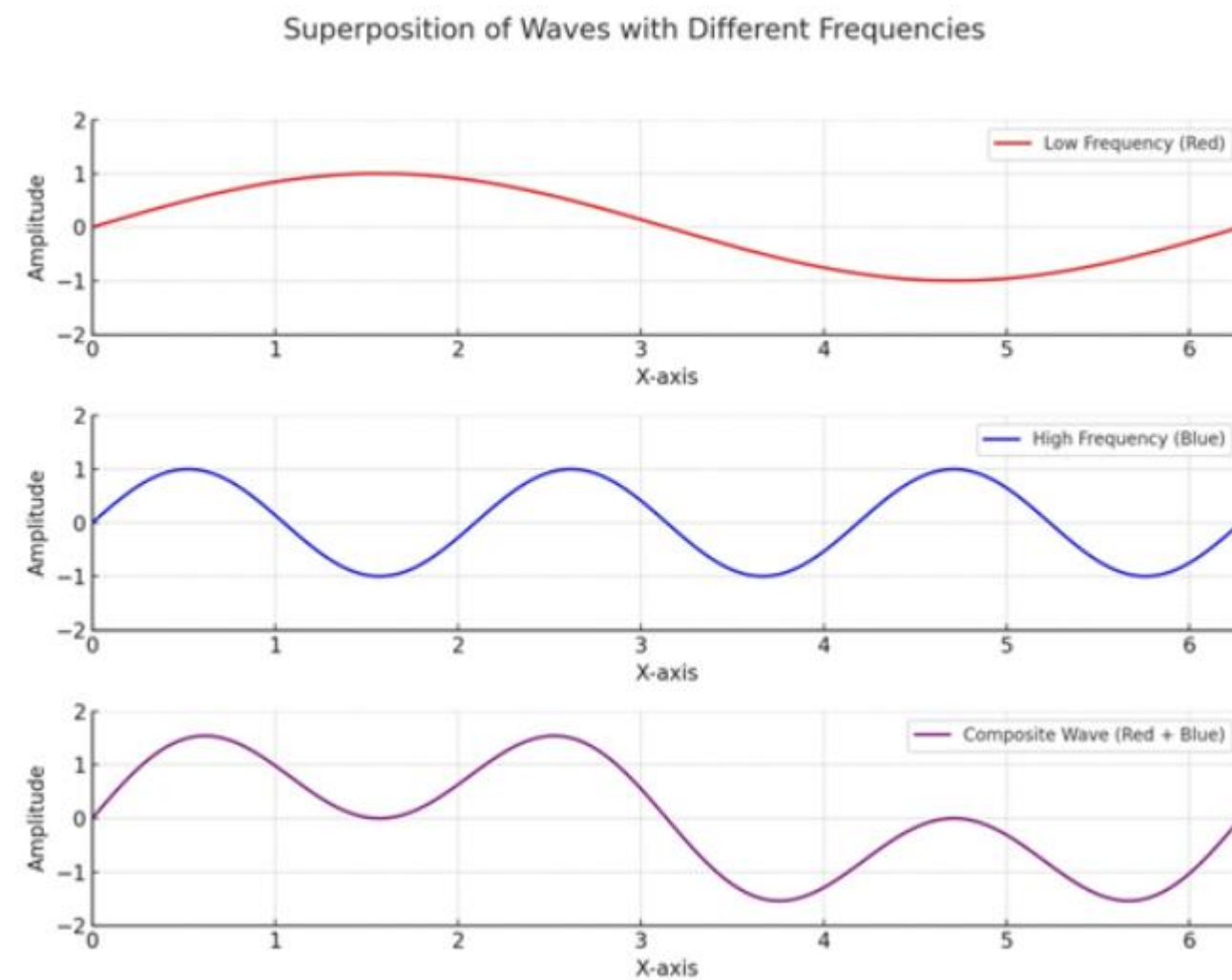
: 화자 두 명인 상태를 만들기 위한 기법

② 원리

: Train Data에 랜덤한 Train Data를 하나 더 합성하여 생성함
→ Multi-Label classification이 가능함

③ 활용

: 총 Train Data와 비슷한 크기로 Mixup Data 생성



[자세히] Label Smoothing

① Label Smoothing

: Hard label(One-hot encoded vector로 정답 인덱스는 1, 나머지는 0으로 구성)을 Soft label(라벨이 0과 1 사이의 값으로 구성)로 스무딩하는 것을 뜻함

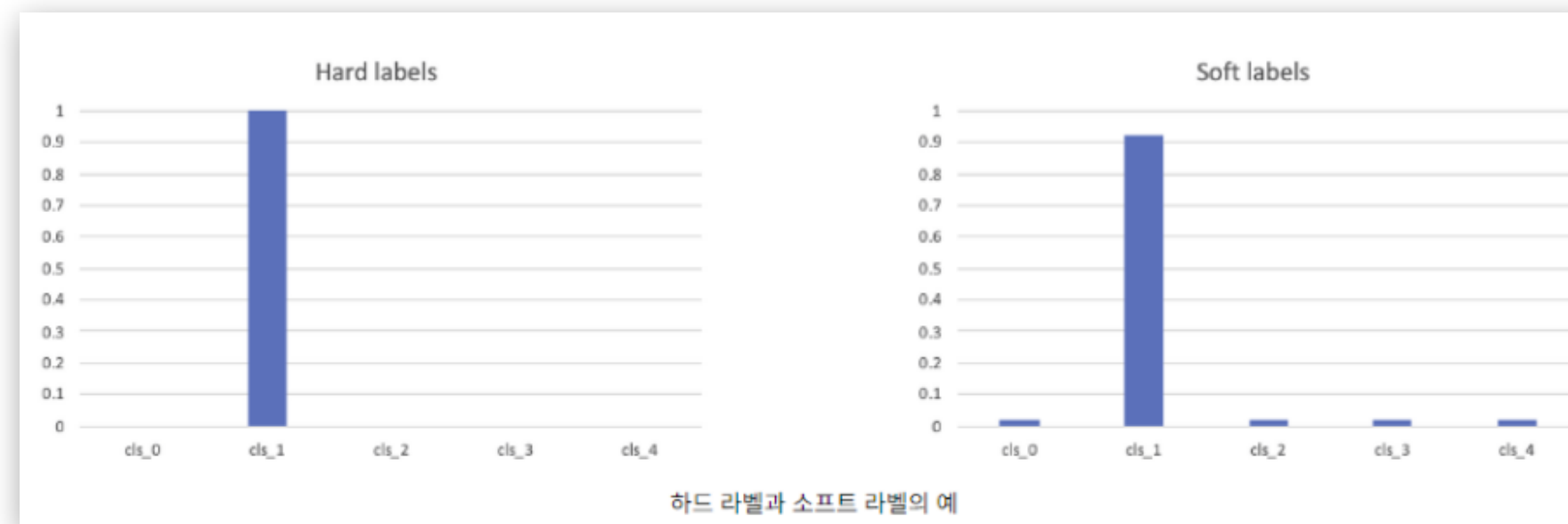
* K개의 클래스에 대해서, 스무딩 파라미터(Smoothing parameter)를 α 라고 할 때, k번째 클래스에 대한 Smoothing 식은 아래와 같음

$$y_k^{LS} = y_k(1 - \alpha) + \alpha/K$$

[자세히] Label Smoothing

② Hard Label vs Soft Label

```
# Class number K = 5, 2nd class is ground-truth.  
# y2_hard_label = [0, 1, 0, 0, 0]  
# y2_soft_label = [0.02, 0.92, 0.02, 0.02, 0.02]
```



: One-hot encoded vector에서 0인 값은 모두 0보다 큰 값인 0.02로,
정답을 나타내는 1은 1보다 작은 0.92로 라벨 스무딩을 통해 벡터값을 변환함
→ 이러한 변환을 다른 말로 label-smoothing regularization(LSR)이라고 부름

실행 시간

① with Denoise

- T4
 - 50분 41초
- L4
 - 21분 17초
- A100
 - 21분 47초

② without Denoise

- T4
 - 16분 53초
- L4
 - 6분 16초
- A100
 - 6분 03초