주제분석

너의 목소리가 보여

1 딥러닝팀 이승우 안세현 이수정 전효림 홍지우

01

주제 소개

01 주제 소개

딥러닝을 활용한 화자 프로파일링

주제 선정 배경

사람의 음성을 바탕으로 성별, 나이, 출신 지역, 최종 학력 등을 유추할 수 있는 모델 구현

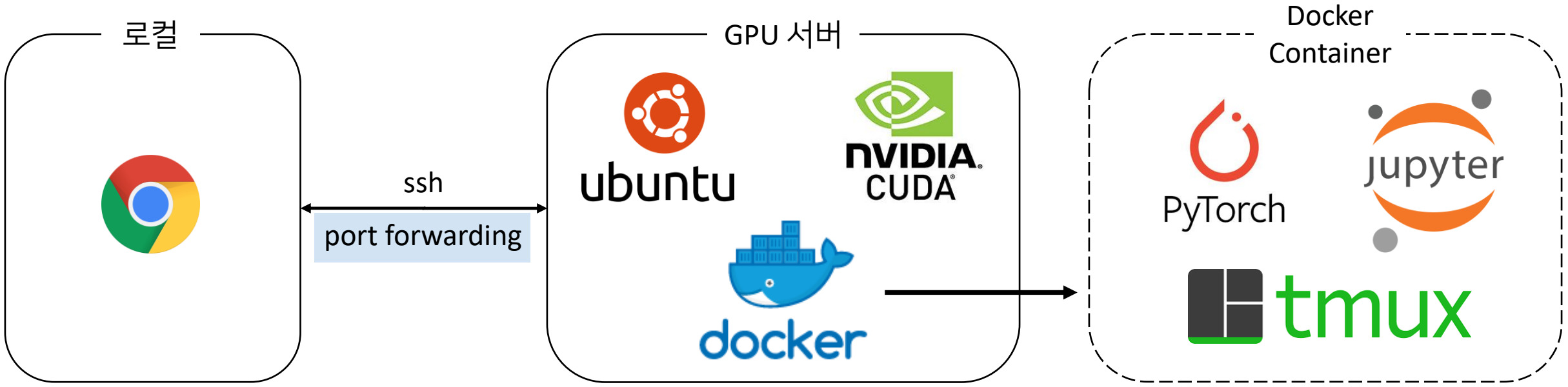
02

개발 환경

02 개발 환경

개발 환경

현재 개발 환경



서버에서 주피터 노트북을 실행해 서버를 연 후

포트 포워딩을 통해 로컬에서 직접 접속

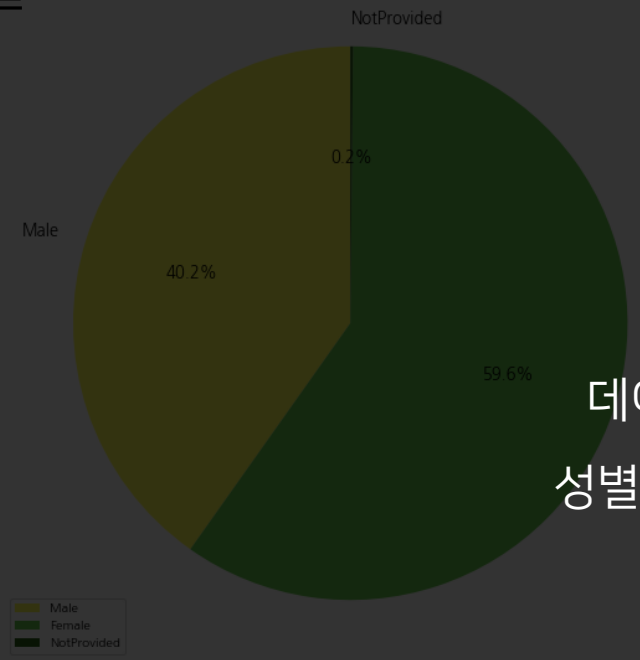
03

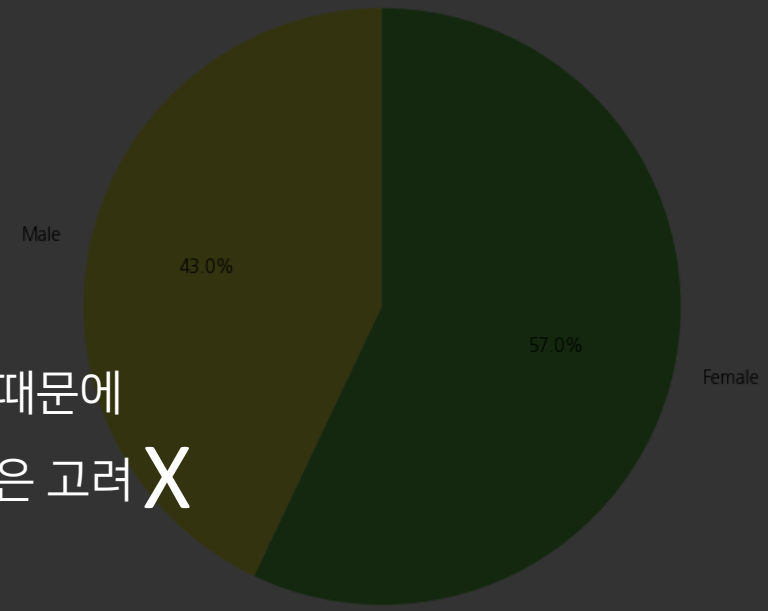
음성 데이터

03 음성 데이터

EDA

성별





데이터의 수가 아주 많기 때문에

성별에 대한 데이터 불균형은 고려 X

명령어 음성	자유 대화 음성
여성이 20%p 정도 더 많음	여성이 12%p 정도 더 많음

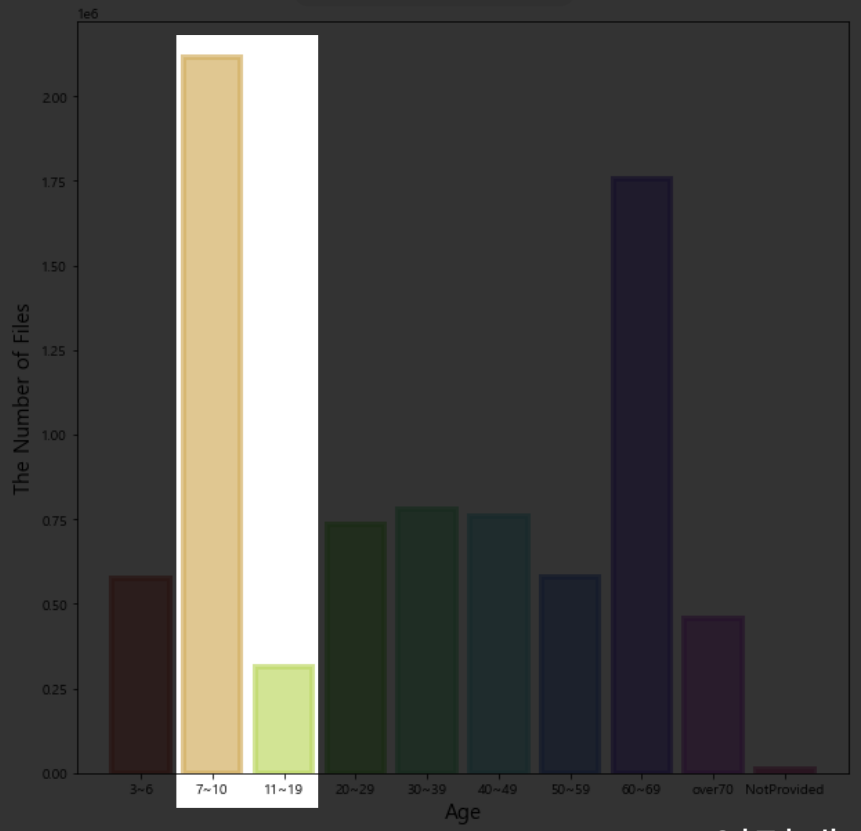
DeepVoice

03 음성 데이터

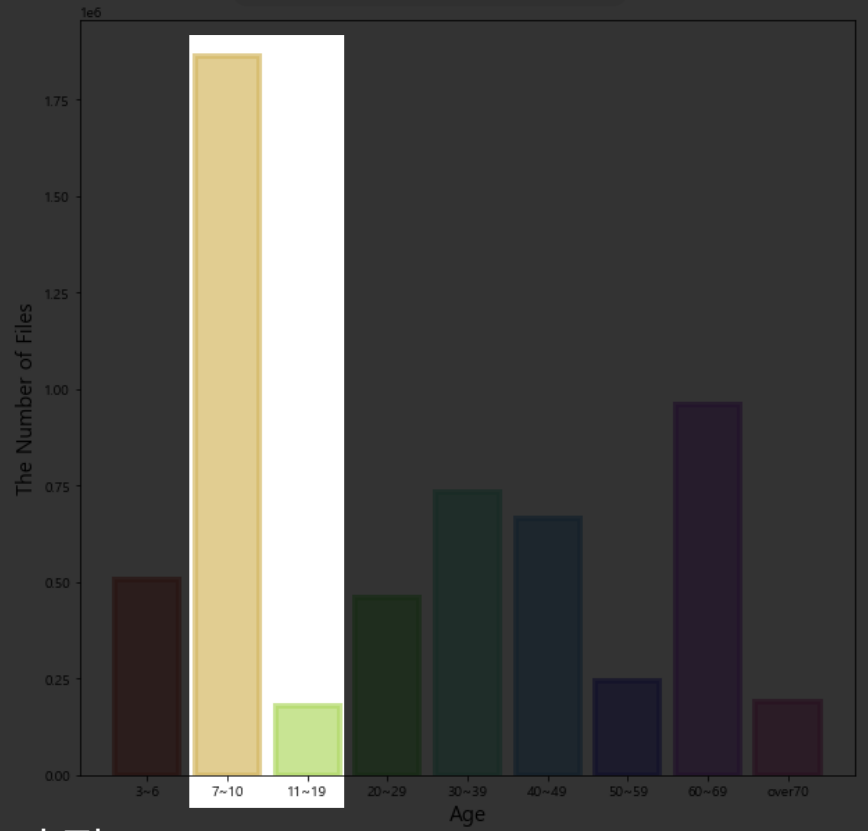
EDA

연령

명령어 음성



자유 대화 음성



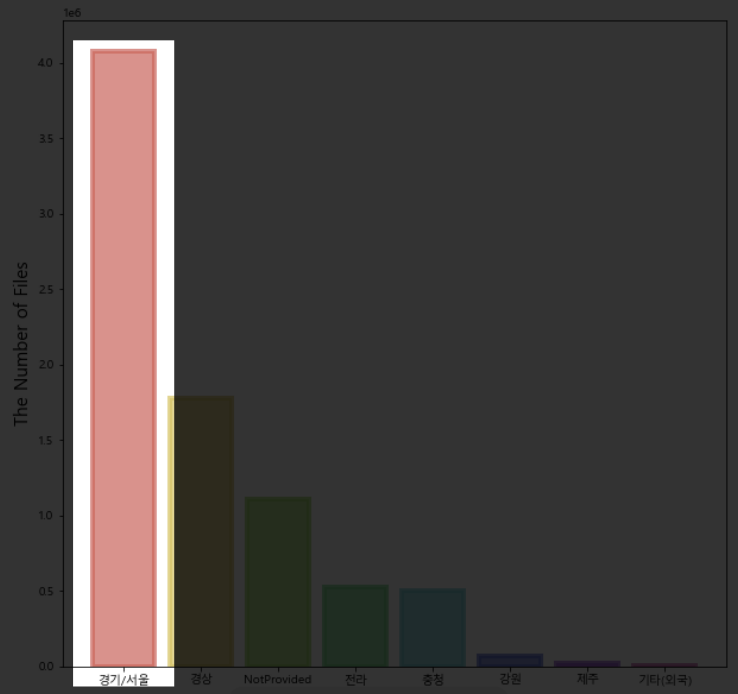
연령에 대한 데이터 불균형 고려 필요

DeepVoice

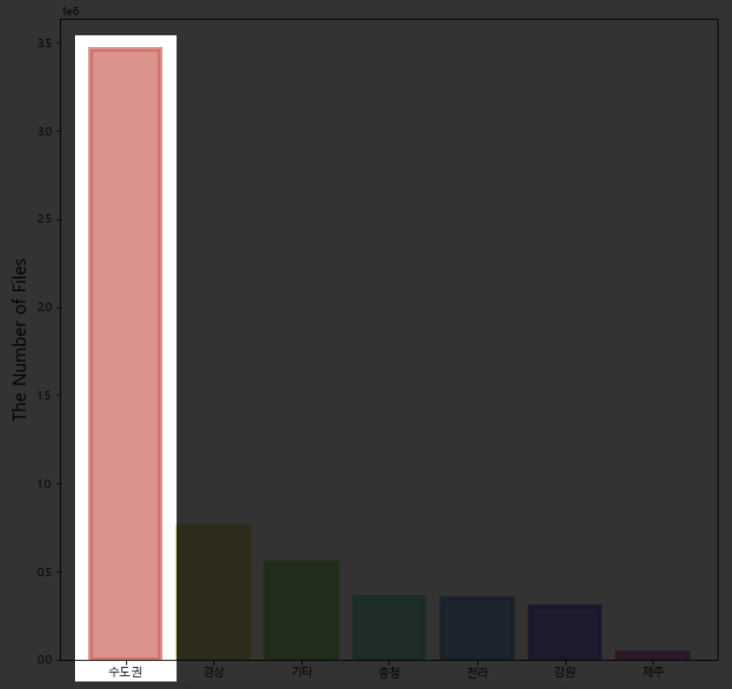
03 음성 데이터

EDA

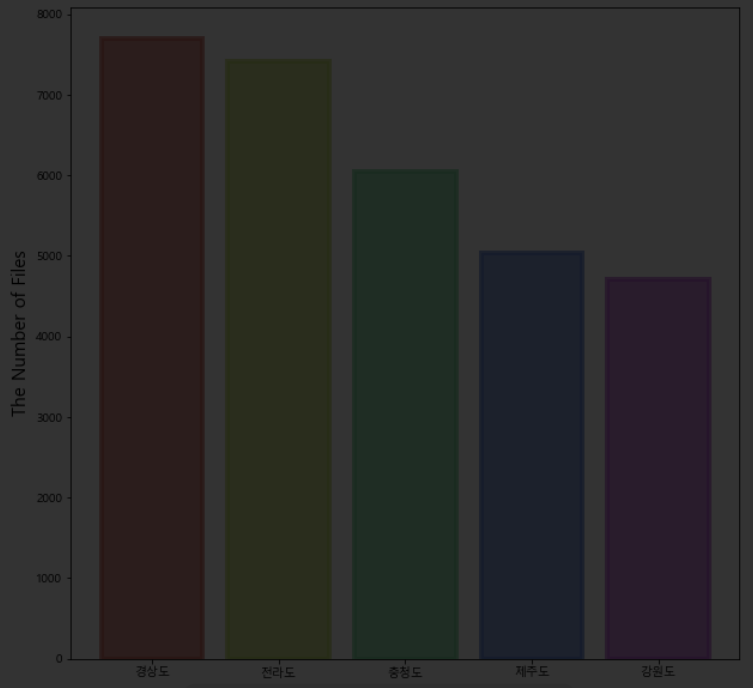
방언



명령어 음성



자유 대화 음성



한국어 방언 발화

방언에 대한 데이터 불균형 고려 필요

DeepVoice

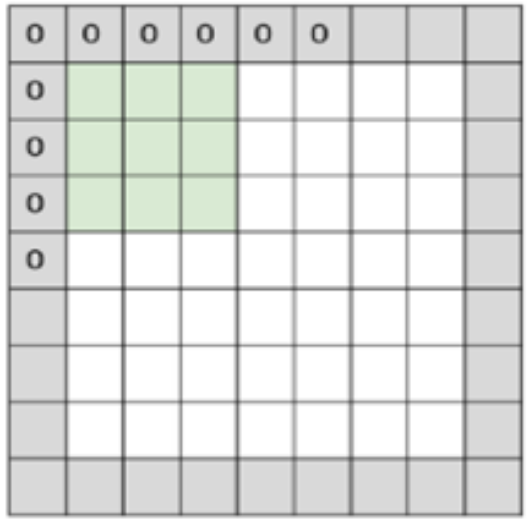
04

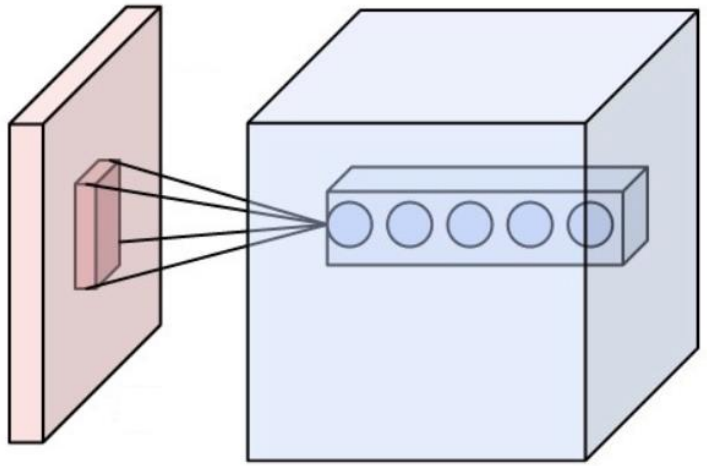
분석 과정

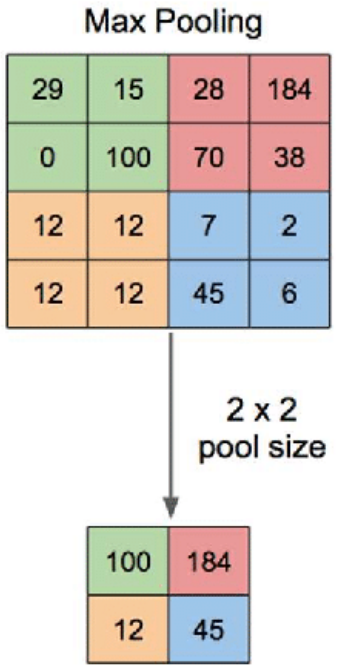
04 분석 과정

모델링

MFCC를 활용한 모델 - CNN







Input 위, 아래
2칸씩 Zero Padding

Convolutional Layer

Max Pooling

DeepVoice

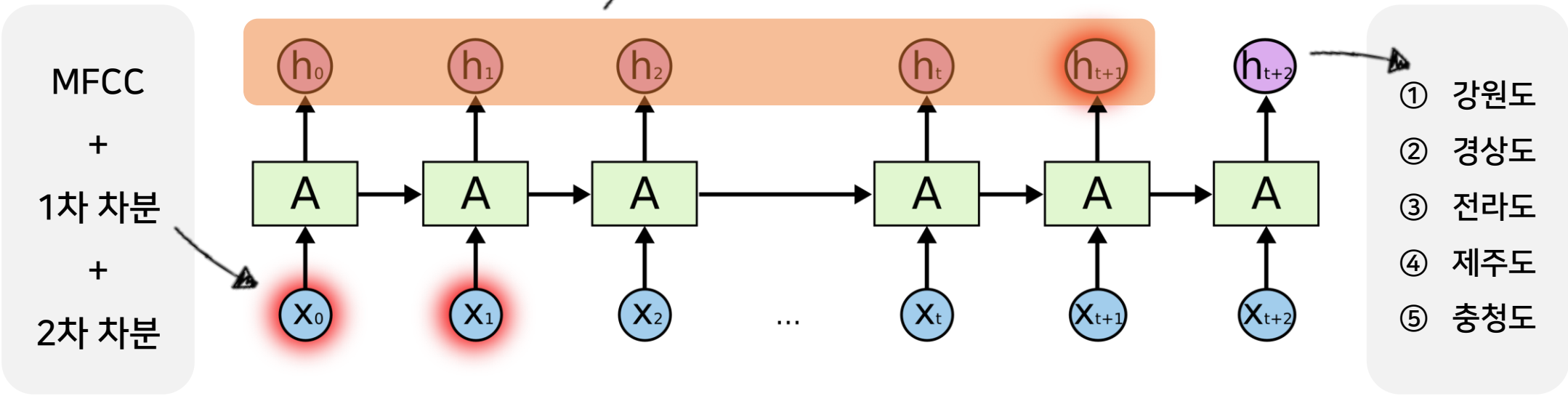
04 분석 과정

모델링

MFCC를 활용한 모델 - RNN

분류 모델이므로 Many-to-One의 형태

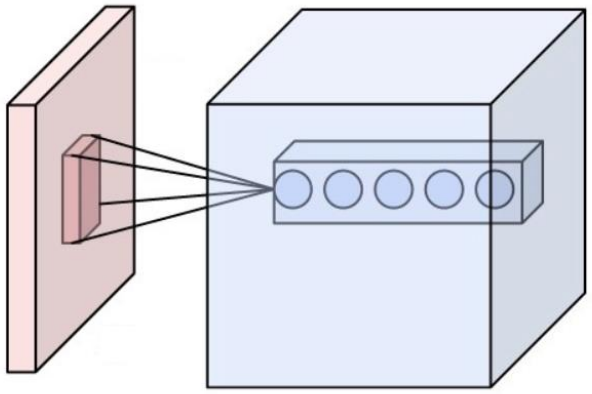
∴ 최종 Hidden Layer의 Output만 사용!



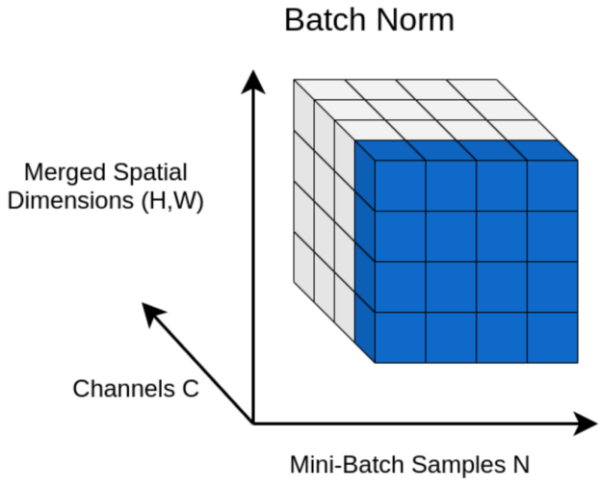
04 분석 과정

모델링

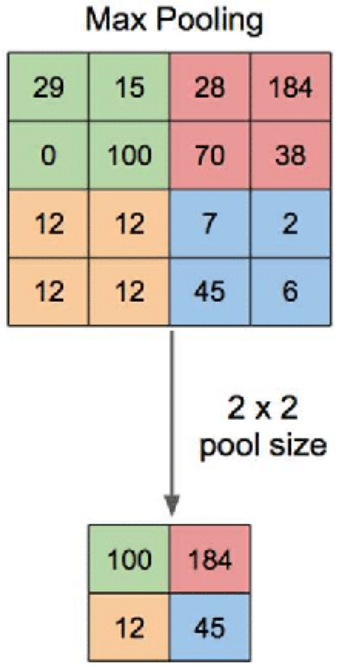
Mel Spectrogram을 활용한 모델 - CNN



Convolutional Layer



Batch Normalization



Max Pooling

DeepVoice

04 분석 과정

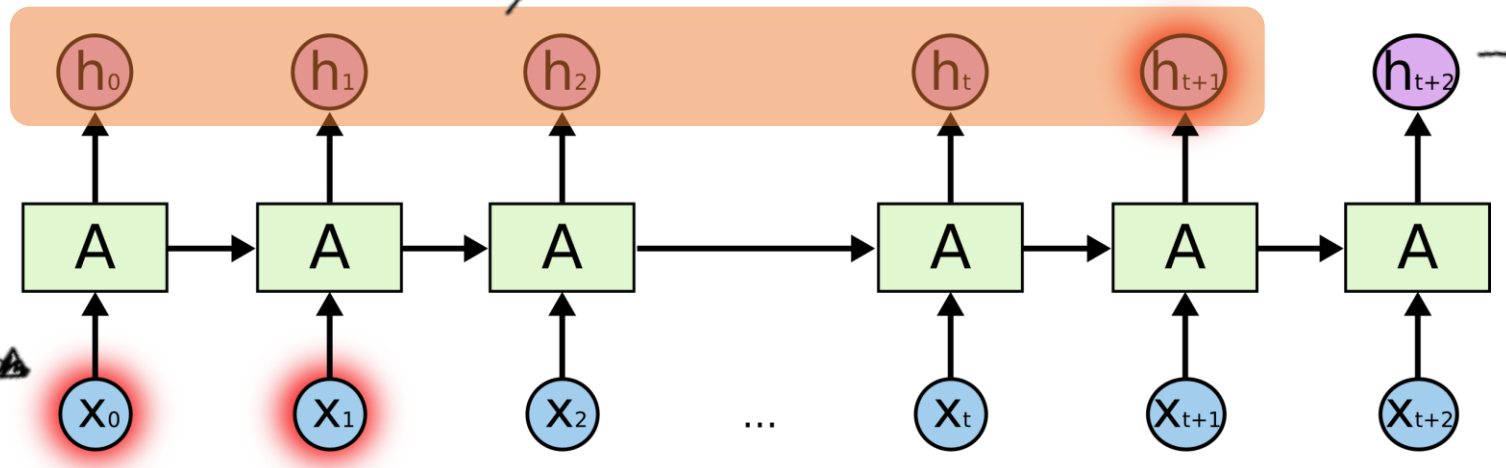
모델링

Mel Spectrogram을 활용한 모델 - RNN

분류 모델이므로 Many-to-One의 형태

∴ 최종 Hidden Layer의 Output만 사용!

Mel Spectrogram + 1차 차분



① 강원도
② 경상도
③ 전라도
④ 제주도
⑤ 충청도

04 분석 과정

모델링

모델 개요

모델 유형	Input 유형	Input Size	모델 구성
CNN (Convolutional NN)	MFCC	torch.size([3, 13, 126000])	[Conv * Pool] Layer 2개 + FC Layer 2개
Vanilla RNN (Recurrent NN)	MFCC	torch.size([126000, 39])	RNN Hidden Layer 2개 + Sequence 길이 126000
CNN	Mel Spectrogram	torch.size([2, 2000, 2000])	[Conv * BN * Pool] Layer 4개 + FC Layer 2개
Vanilla RNN	Mel Spectrogram	torch.size([5000, 256])	RNN Hidden Layer 2개 + Sequence 길이 5000

04 분석 과정

모델링

모델 성능 평가

모델의 유형과 무관하게 학습이 제대로 이루어지지 않음

모델 유형	모델 성능 평가	문제점
CNN (Convolutional NN)	① Loss (Cross Entropy Loss) ▪ Learning Rate: 5e-6~1e-9 적용 ▪ Batch Size: 10~25 적용 ▪ Epochs: 1~20 적용 ▪ 1.5000 ~ 1.7000 사이 진동 ② Train Accuracy ▪ Sample: 100개~1000개 사용 ▪ 20% ~ 40% 사이 진동	① Loss 수렴하지 않음 ② Train Set에 대한 정확도 낮음 데이터 전처리 방법 근본적인 개선 필요!
Vanilla RNN (Recurrent NN)		
CNN		
Vanilla RNN		

DeepVoice

감사합니다

주제분석

너의 목소리가 보여

1 딥러닝팀 이승우 안세현 이수정 전효림 홍지우

01

복습 및 정리

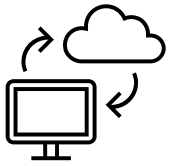
01 복습 및 정리

1주차 복습

딥러닝팀의 1주차

음성을 바탕으로 성별, 나이, 지역 유추하는 모델 구현

개발환경



서버에서 주피터 노트북을 실행해 서버를

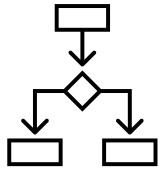
연 후 포트 포워딩을 통해 로컬에서 직접 접속

데이터

여러 개의 JSON 파일 하나의 CSV로 변환

전처리를 거쳐 Feature engineering

모델링



MFCC / Mel Spectrogram을 이용한

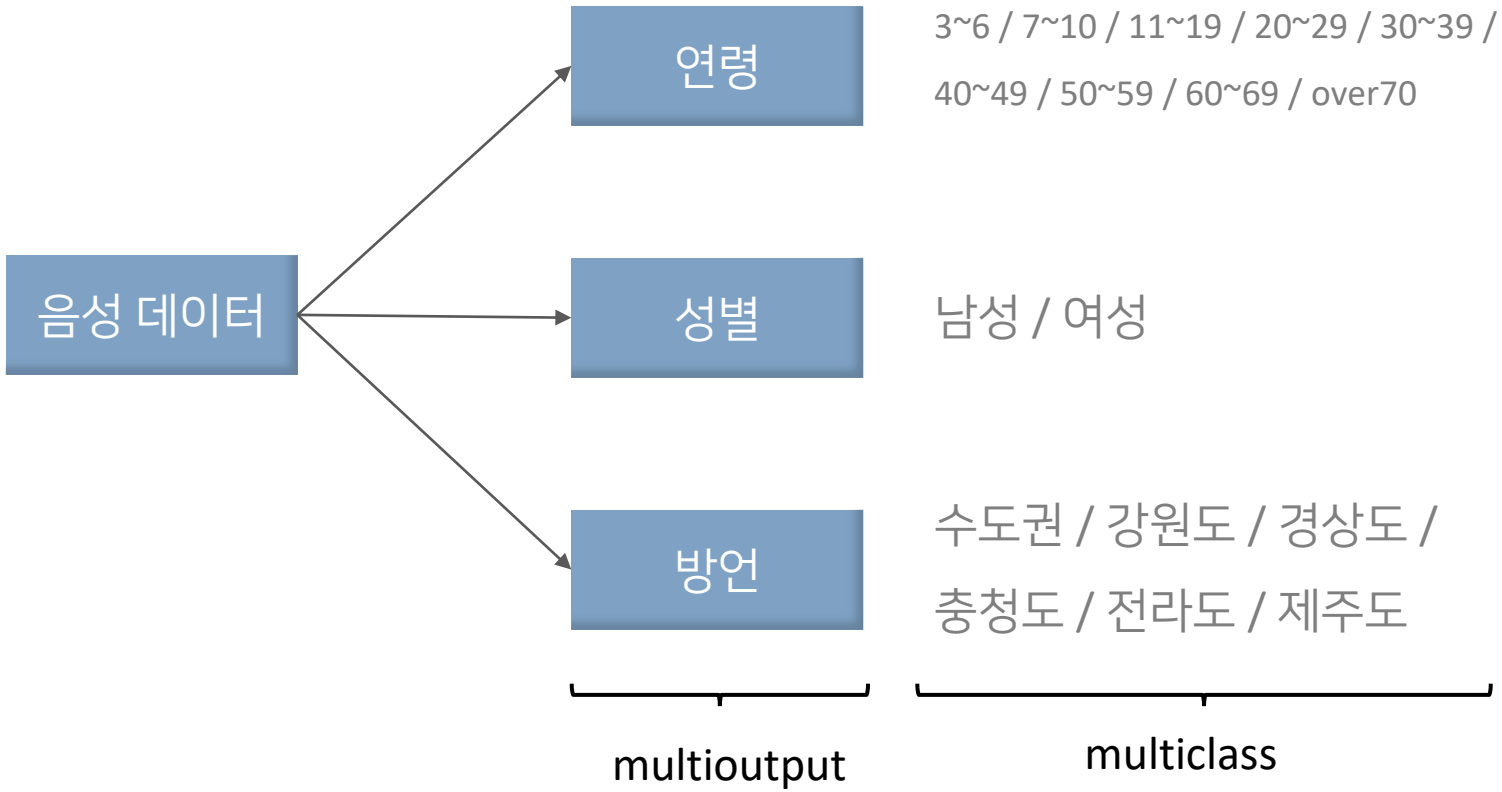
CNN 모델과 RNN 모델

DeepVoice

01 복습 및 정리

정리

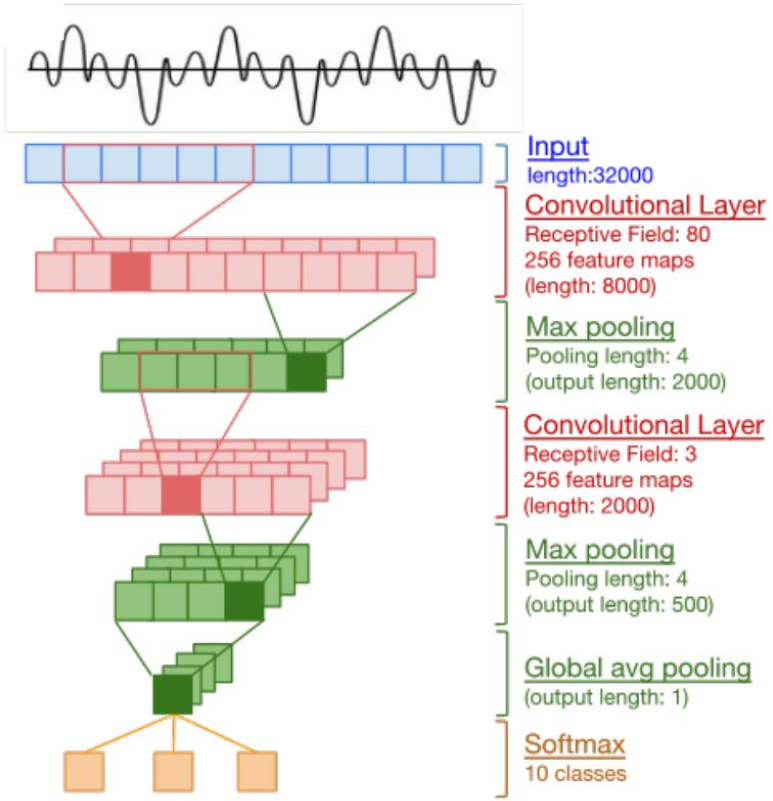
multiclass-multioutput classification



01 복습 및 정리

모델 추가

CNN for Raw Waveform



Feature Engineering 없이
입력으로 raw-audio를 사용하는
CNN 모델

VERY DEEP CONVOLUTIONAL NEURAL NETWORKS FOR RAW WAVEFORMS
- Wei Dai, et. al (2016)

02

시행착오 및 최종 input

02 시행착오 및 최종 input

시행착오

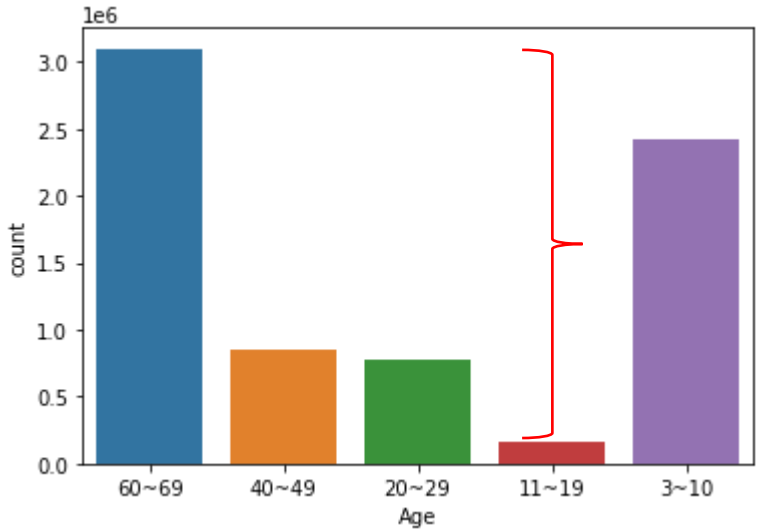
모델 성능 문제

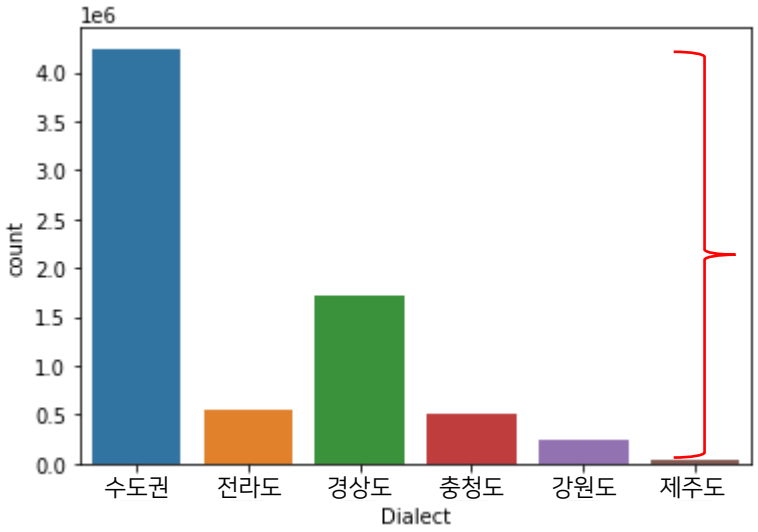
모델 유형	연령	성별	방언
Vanilla RNN (Recurrent NN)	train acc : 0.3 val acc : 0.28	train acc : 0.56 val acc : 0.55	train acc : 0.25 val acc : 0.16
CNN	train acc : 0.9 val acc : 0.55	train acc : 0.99 val acc : 0.82	train acc : 0.99 val acc : 0.28
CNN + LSTM	train acc : 0.52 val acc : 0.28	train acc : 0.79 val acc : 0.84	train acc : 0.45 Val acc : 0.2
CNN for raw waveform	train acc : 0.99 val acc : 0.5	train acc : 0.99 val acc : 0.86	train acc : 0.99 val acc : 0.25

02 시행착오 및 최종 input

시행착오

훈련 데이터 균등 샘플링





연령, 방언 label에 불균형이 심해 모델의 학습에 지장이 있을 것이라 판단

→ 데이터를 각 label마다 균등하게 샘플링하여 훈련 데이터로 이용

02 시행착오 및 최종 input

시행착오

Noise 제거 : RNNoise

validation accuracy

	연령 분류 딥러닝	성별 분류 딥러닝	방언 분류 딥러닝	Logistic Regression
노이즈 미제거	40.72	81.07	50.34	0.85
노이즈 제거	43.24	82.13	50.27	0.65

약간의 성능 향상

큰 성능 하락

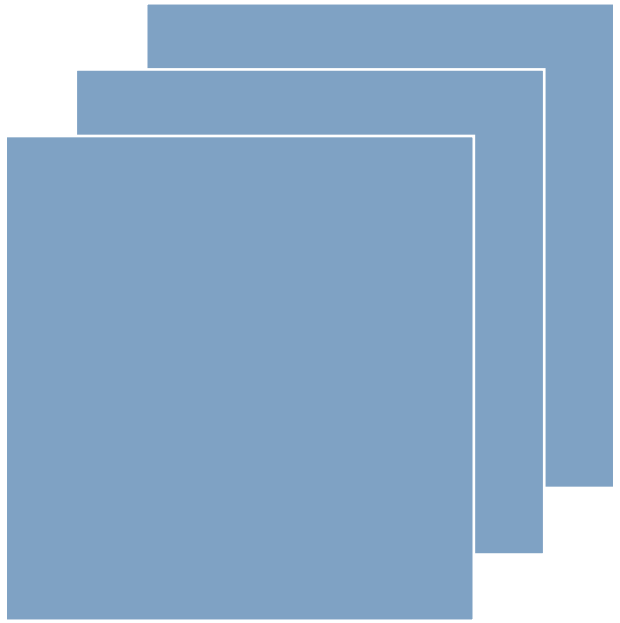
노이즈 제거 + 딥러닝 모델을 사용

성별 분류에 유용한 F0 주파수를 딥러닝 모델의 입력에 추가

02 시행착오 및 최종 input

최종 input

최종 Input



[3, 14, 400]

MFCC + F0 / MFCC + F0의 1차 차분 / MFCC + F0의 2차 차분

03

최종 모델링

03 최종 모델링

모델 연결

필요성

우리가 예측하려 하는 label이

다른 label의 예측에 도움을 줄 수 있지 않을까?

음성 외에 아무 정보도 주어지지 않은 상태의 방언 분류

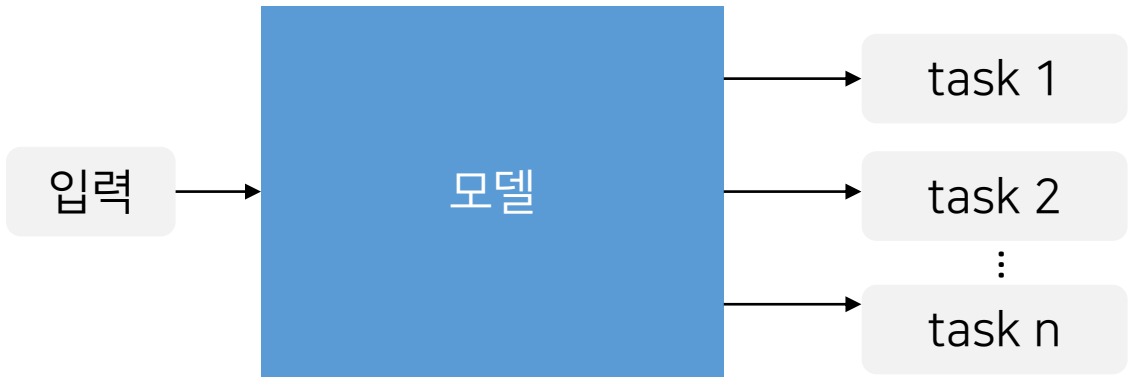
vs

음성과 함께 60대 남성이라는 정보가 주어진 상태의 방언 분류

03 최종 모델링

모델 연결

multitask learning



장점

- Knowledge Transfer

task 1을 학습하며 얻은 정보가

다른 task를 해결하는 데 도움

- Overfitting 감소

여러 task를 동시에 해결해야 하기 때문에

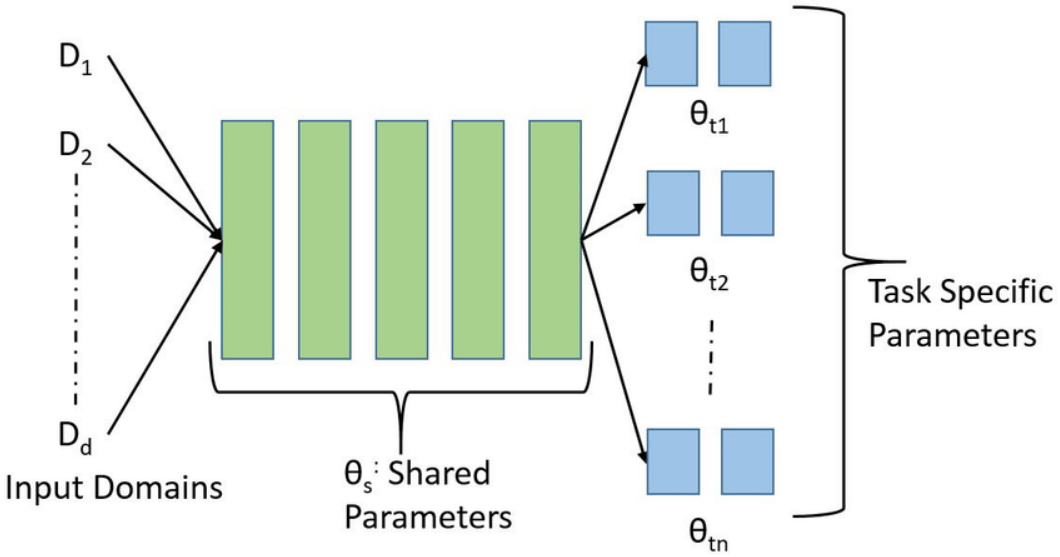
보다 일반화 된 feature를 추출하도록 학습

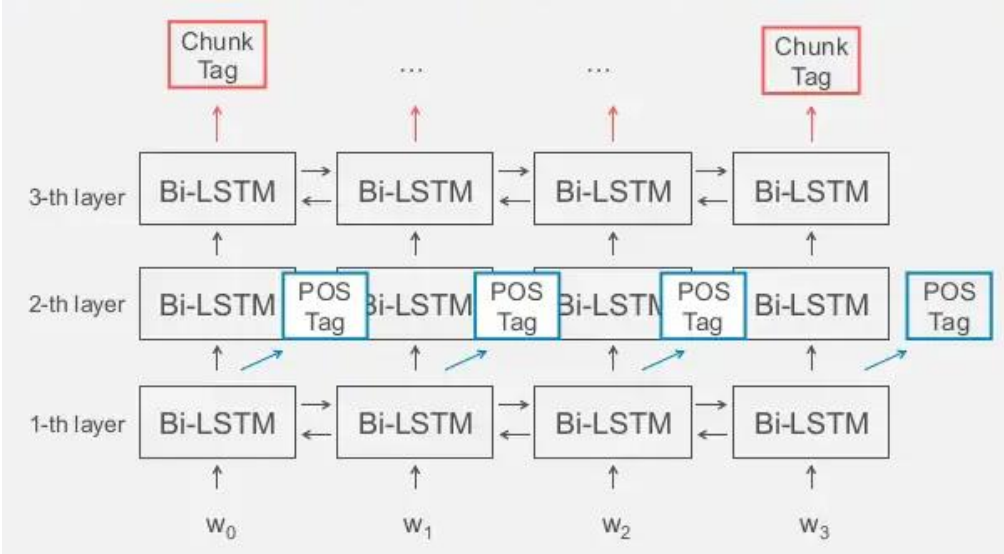
03 최종 모델링

모델 연결

개요

Hard Parameter Sharing	LSTM Base MTL





DeepVoice

04

결과 및 의의

04 결과 및 의의

최종 결과

최종 결과 정리

모델 유형	연령	성별	방언
MFCC + F0 Hard Parameter Sharing	train acc : 0.57 val acc : 0.55	train acc : 0.9 val acc : 0.84	train acc : 0.3 val acc : 0.29
CNN For raw-waveform Hard Parameter Sharing	train acc : 0.99 val acc : 0.5	train acc : 0.99 val acc : 0.85	train acc : 0.99 val acc : 0.28
CLSTM	train acc : 0.51 val acc : 0.45	train acc : 0.9 val acc : 0.85	train acc : 0.23 val acc : 0.35

감사합니다