뻐끔뻐끔

Video-LipReading-to-Script

대구 1기

박혜령 이창수 김선아 성은지 이동섭

Contents

Contents

04

Experiments

1. 모델 학습 결과 비교
2. Train 시각화
3. Test 결과

05

Demo

1. 데모 실행
2. 데모 분석

06

Future Works

1. 기대효과 및 활용방안
2. 향후 발전 가능성
3. 앞으로의 개발 일정
4. 서비스 구현 프로세스

Our Project

Video-LipReading-to-Script

01. 팀원 소개

뻐끔뻐끔

박혜령	이창수	김선아	성은지	이동섭
- 팀장 - 논문 리뷰 - **모델링** - 총괄	- 팀원 - 논문 리뷰 - 모델링 - **개발 환경 구축**	- 팀원 - **논문 리뷰** - 모델링 - 발표	- 팀원 - 논문 리뷰 - **데이터셋 구축**	- 팀원 - 논문 리뷰 - 데이터셋 구축 보조 - **도메인 조사**

02. 프로젝트 소개

붕어립(Bung-eo-lip) 프로젝트

1. 영상의 **입 모양 모션을 인식**하는 모델 구현
2. 영상에 **한국어 자막**을 제공하는 서비스 개발

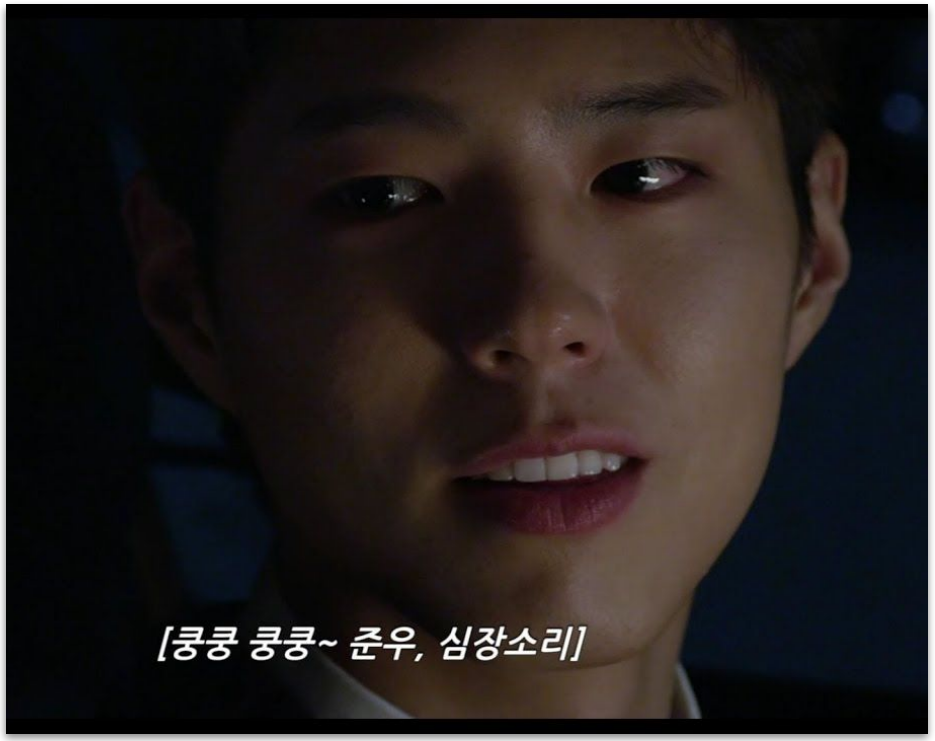
03. 개발 배경

🧐 배리어 프리란?

장애인을 포함한 모든 사회의 구성원이 살기 좋은 사회를
만들기 위해 물리적·제도적·심리적 장벽을 허물자는 운동이다.

한국의 독순술 연구와 배리어프리 서비스 미흡

해외 독순술 연구 사례



한국 최초의 배리어프리 영화
박보검 주연의 '반짝반짝 두근두근'(2015)

04. 전체 프로세스

도메인 조사 → 논문 리뷰 → 개발 환경 구축 → 데이터셋 구축 → 모델링

05. 개발 진행 상황

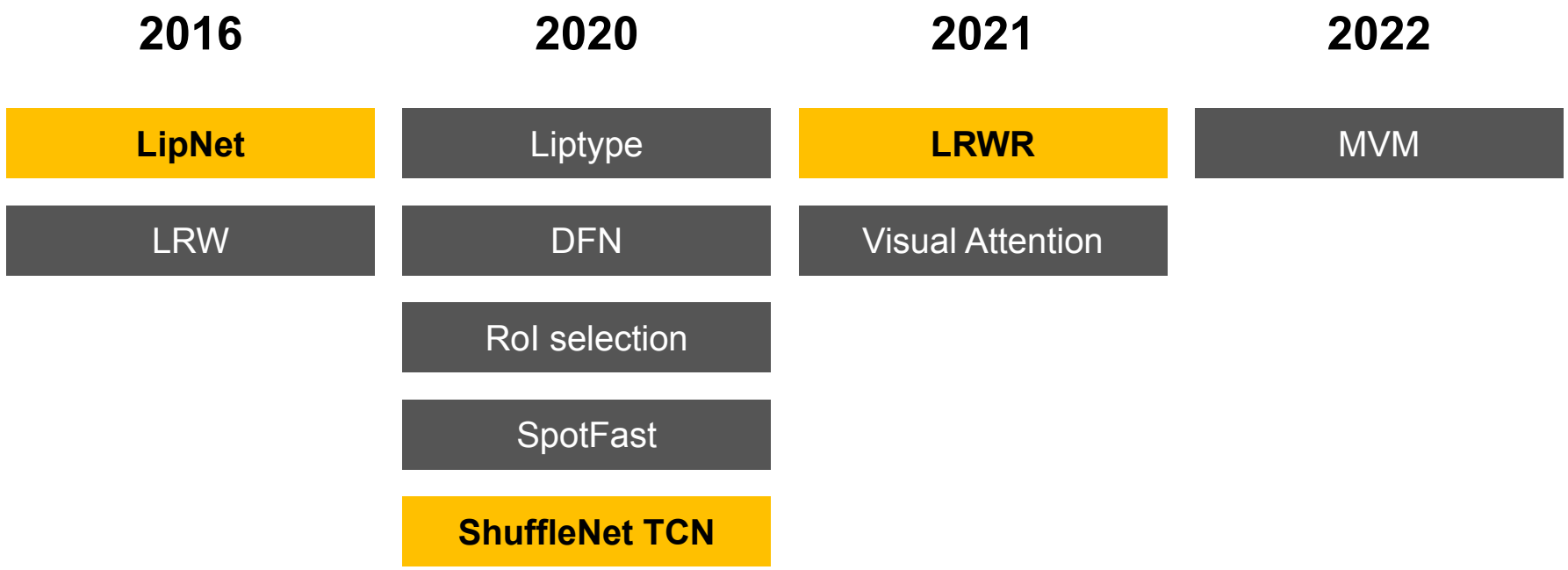
내용 \ 주		M1	M2	H1	H2	H3	H4	H5	H6
1단계	데이터셋 이해/제작								
	관련 논문 리뷰								
	모델 설정								
	검수								
2단계	서비스 개발								
	검수								
3단계	전체 유지보수								

M 미니 해커톤 / H 해커톤

Literature Review

Video-LipReading-to-Script

01. 논문 도표


2016
2020
2021
2022
LipNet
LRW
Liptype
DFN
RoI selection
SpotFast
ShuffleNet TCN
LRWR
Visual Attention
MVM

01. 논문 도표

2016

LipNet

- **최초 문장 단위 립리딩 모델**
- VGG, STCNN, Bi-Gru, CTC Loss

2020

ShuffleNet-TCN

- **모델 경량화 (계산 비용 절감)**
- MobileNet 에서 좀 더 변형된 SuffleNet
- TCN

2021

LRWR

- **러시아어 데이터셋 구축**
- D3D

02. Introduction - 립리딩 모델 구조

- 딥러닝 적용 방식 : **2단계 접근법**을 따름
  - **Frontend** : 3D-CNN(3D conv layer + deep 2D conv)이 최근 연구에서 많이 쓰임
  - **Backend** : LSTM, Attention mechanisms, self-attention modules, Temporal Convolutional Networks (TCN)

LipNet

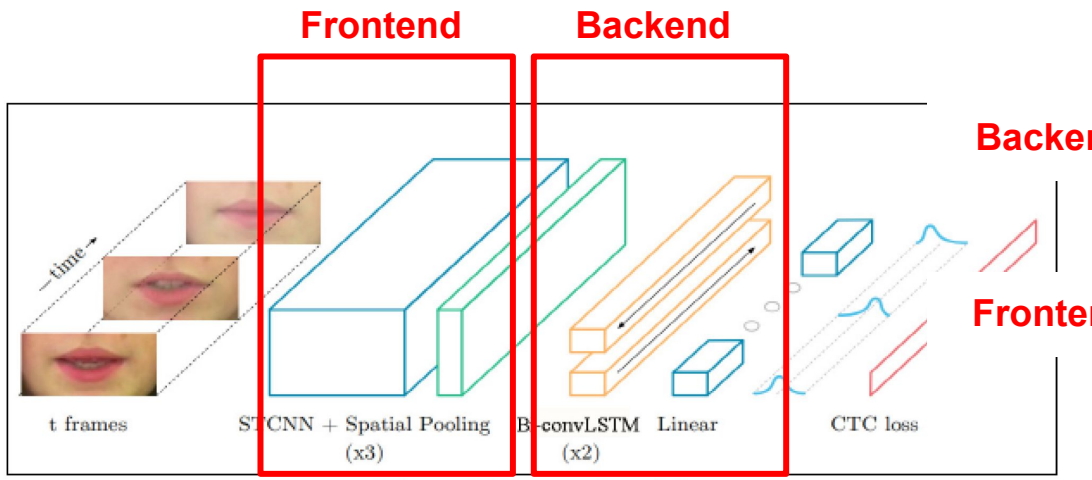
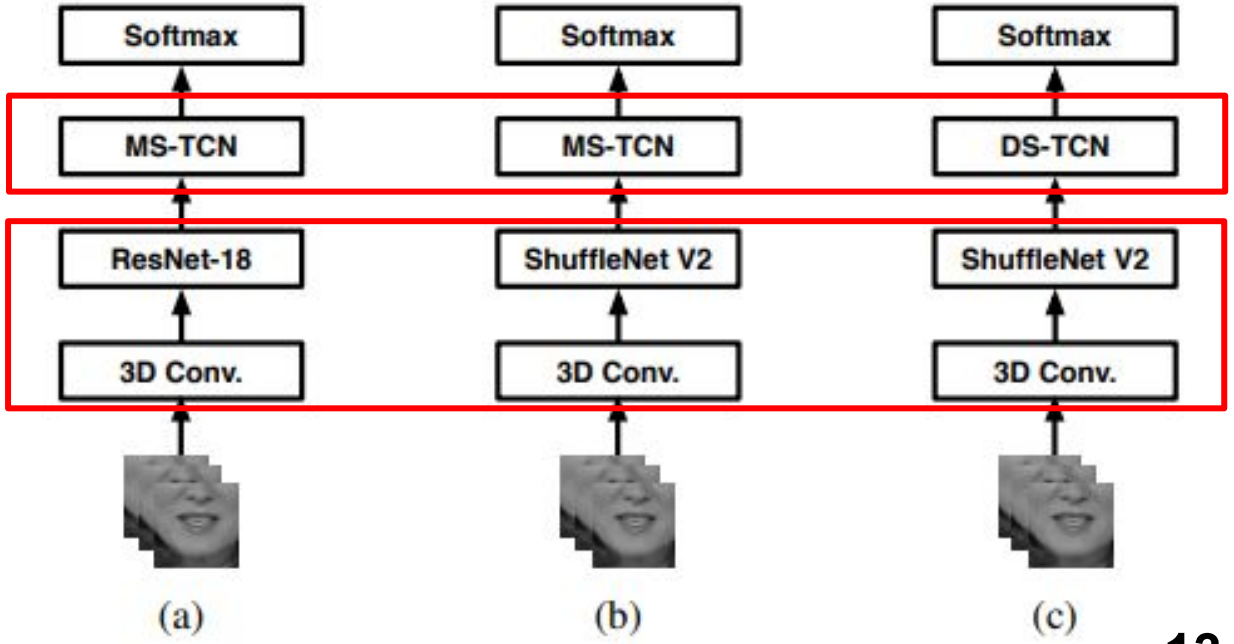


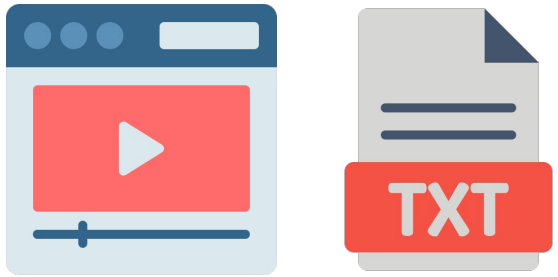
Figure 5. Improved model

ShuffleNet-TCN



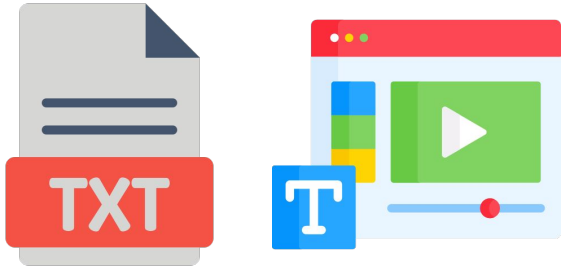
02. Introduction - 립리딩 모델 I/O

I/O



Input

- **Video** → Crop → Lip Image Frames
- **Text Align** → Label



Output

- **Text**
- Subtitle이 생성된 영상

*Assael, Yannis M., et al. "Lipnet: End-to-end sentence-level lipreading." *arXiv preprint arXiv:1611.01599* (2016)

03. LipNet Summary

Key Contributions

1. 최초의 end-to-end **문장** 단위의 모델
2. GRID Corpus dataset

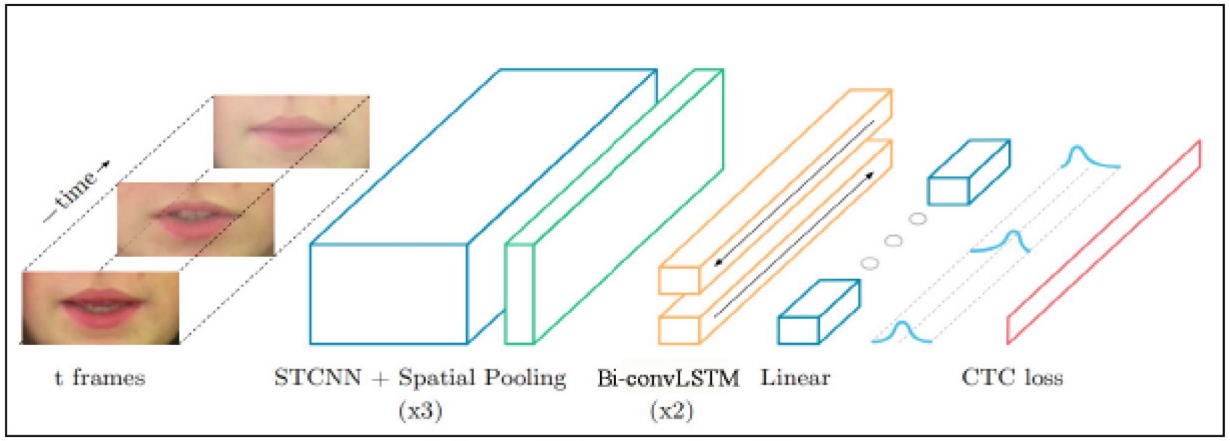
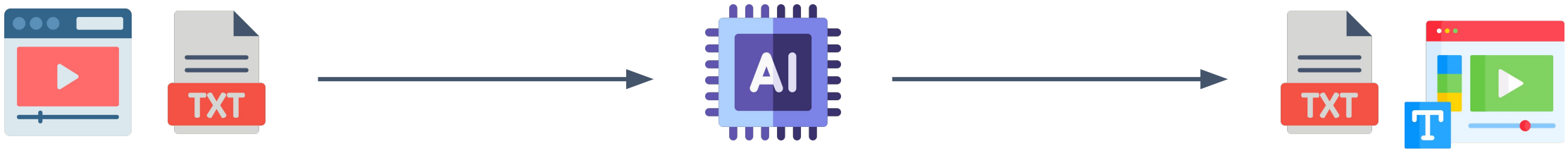


Figure 5. Improved model



Process

- **STCNN**: 논문에 따르면 video에서 시간의 흐름과 공간의 차원을 모두 convolution
- **bi-LSTM**: STCNN의 output sequence 정보를 전파하기 위해 bi-LSTM사용, 정보 흐름을 제어 학습
- **CTC Loss**: Target sequence와 output sequence의 길이가 다를때 사용
- Label → UNICODE (encoding) → 한국어 (decoding)

*Assael, Yannis M., et al. "Lipnet: End-to-end sentence-level lipreading." *arXiv preprint arXiv:1611.01599* (2016)

04. ShuffleNet-TCN - 선정 이유

1. LipNet의 연산량 이슈

a. LipNet은 연산량이 매우 많음

tensorflow 버전	3 epoch 진입 시, 메모리 부족으로 GCP 꺼짐
pytorch 버전	10 epoch 798min 소요 (default : 1만 epoch로 설정되어 있었음)

2. 경량 모델 사용

a. Backbone 교체

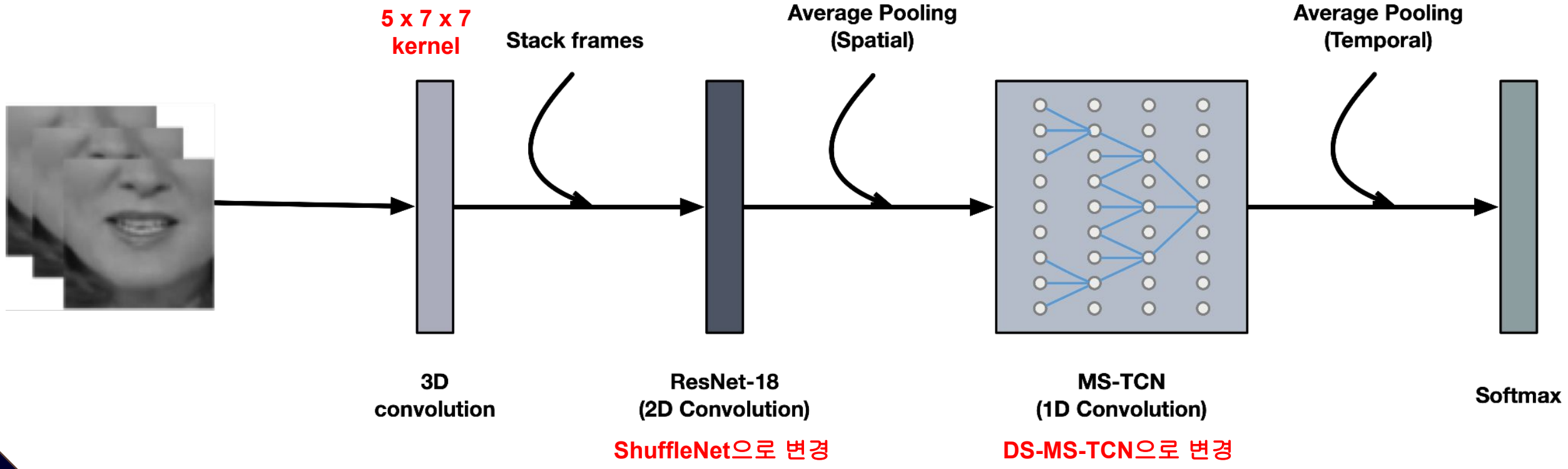
b. DS-MS-TCN 사용

c. Knowledge Distillation

*Pingchuan Ma., et al. "Towards Practical Lipreading with Distilled and Efficient Models." *arXiv preprint arXiv: 2007.06504* (2020)

04. ShuffleNet-TCN - Summary

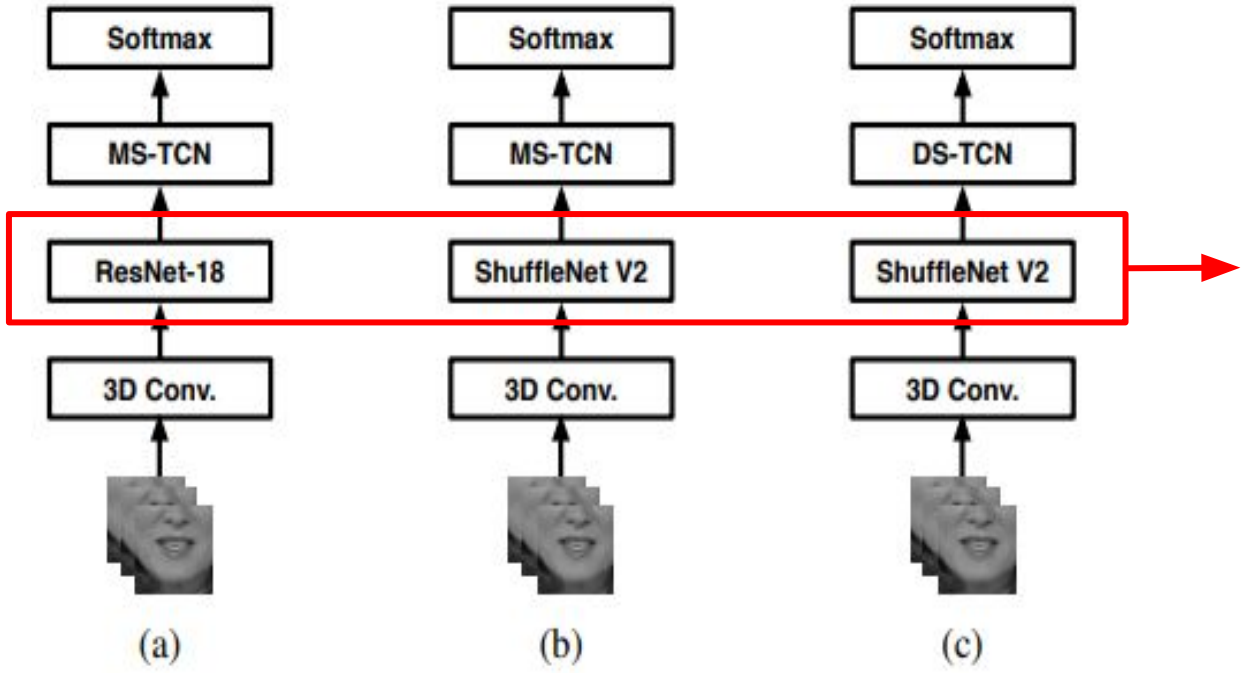
Towards Practical Lipreading with Distilled and Efficient Models



*Pingchuan Ma., et al. "Towards Practical Lipreading with Distilled and Efficient Models." *arXiv preprint arXiv: 2007.06504* (2020)

04. ShuffleNet-TCN - Key Contribution

1-1. Backbone 교체 → ShuffleNetV2



- backbone을 ResNet-18이 아닌, 경량화된 ShuffleNet V2로 변경한다.
- ResNet-18보다 파라미터 수가 5배 적고, FLOPs가 12배 적다.

*Pingchuan Ma., et al. "Towards Practical Lipreading with Distilled and Efficient Models." *arXiv preprint arXiv: 2007.06504* (2020)

04. ShuffleNet-TCN - Key Contribution

1-2. Shuffle Grouped Convolution

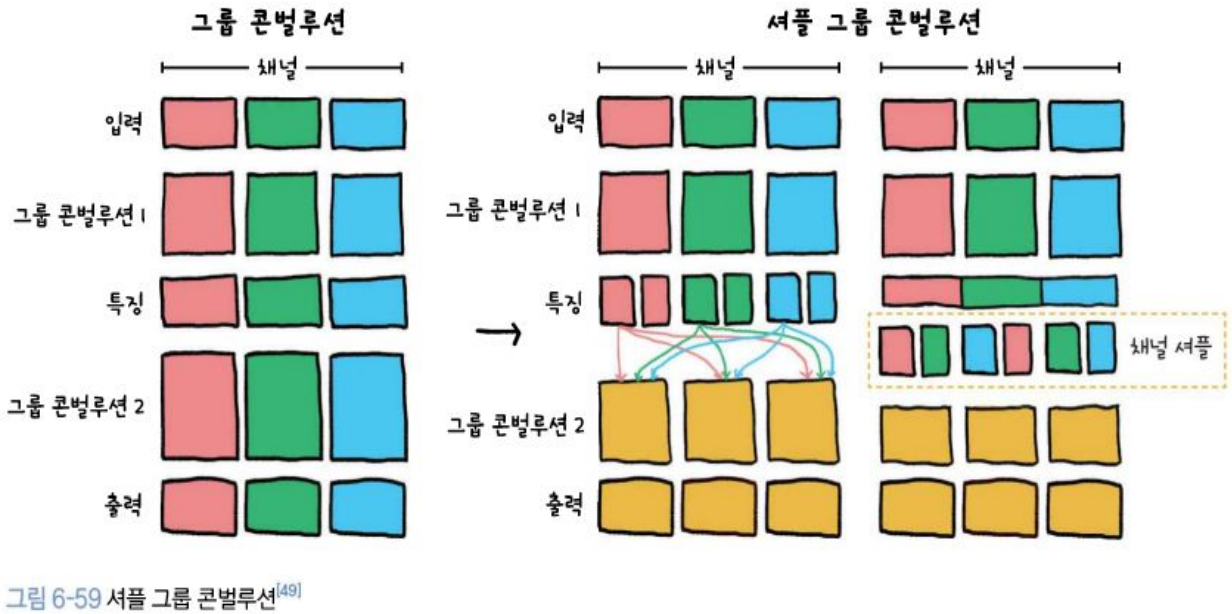


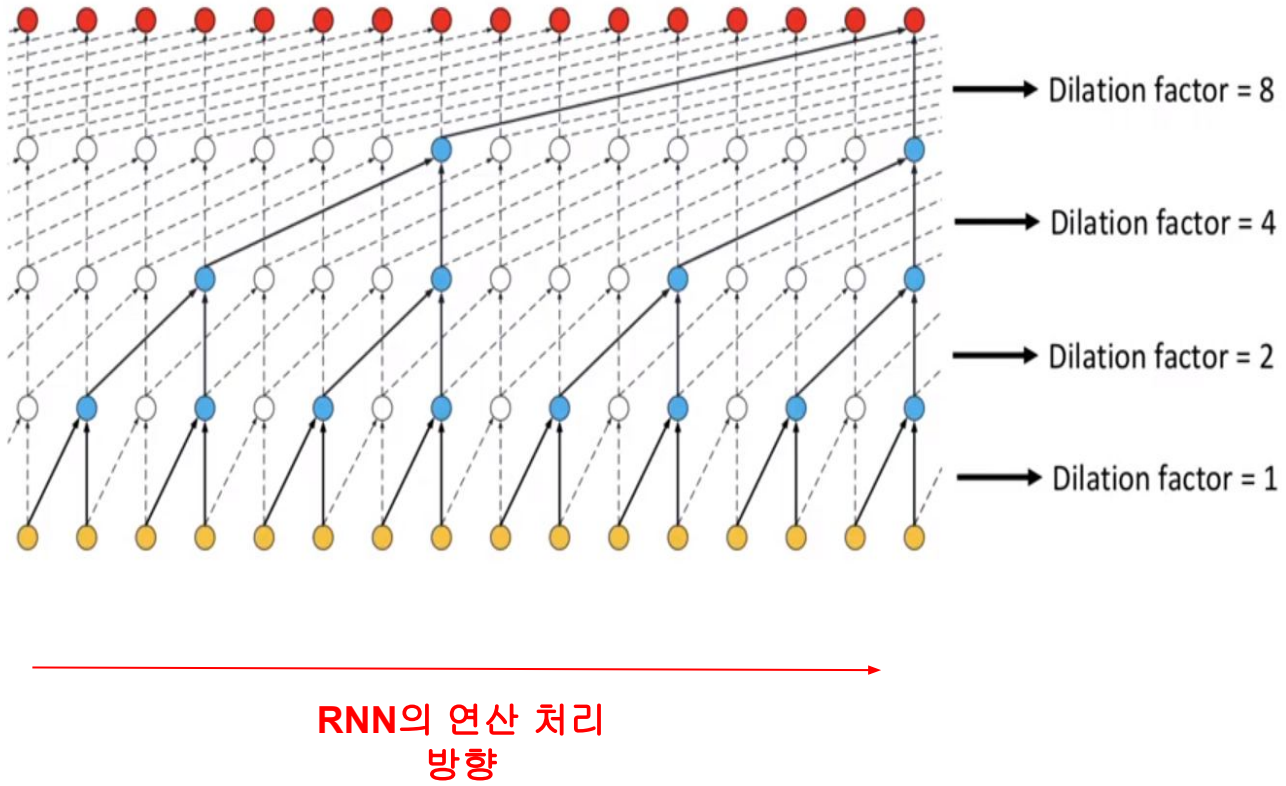
그림 6-59 셔플 그룹 콘벌루션[49]

- ShuffleNet에서 제안된 **Shuffle Grouped Convolution**
- 일반적인 Group Conv는 같은 채널그룹 안에서만 정보가 흐르고, 그룹 간에 서로 정보 교환 X
- 채널 그룹 간에 정보를 교환하면 표현이 강화될 수 있다는 아이디어
- 주기적으로 그룹 간에 채널을 섞어서 정보가 교환되도록 만든 Group Conv 방식

*Pingchuan Ma., et al. "Towards Practical Lipreading with Distilled and Efficient Models." *arXiv preprint arXiv: 2007.06504* (2020)

04. ShuffleNet-TCN - Key Contribution

2-1. TCN (Temporal Convolution Network)



- **1D conv**을 Sequence 데이터에 적용하는 방식
- TCN은 conv을 사용하므로 같은 파라미터에 대해 **병렬적으로 연산**→ 직렬 RNN 보다 빠름
- 모델의 깊이와 dilation으로 **receptive field 크기를 조절 가능**
- TCN은 하나의 layer에 대하여 같은 파라미터가 공유 → **메모리 소요가 적음**

*Pingchuan Ma., et al. "Towards Practical Lipreading with Distilled and Efficient Models." *arXiv preprint arXiv: 2007.06504* (2020)

04. ShuffleNet-TCN - Key Contribution

2-2. DS-TCN
(Depthwise Separable Temporal Convolution Network)

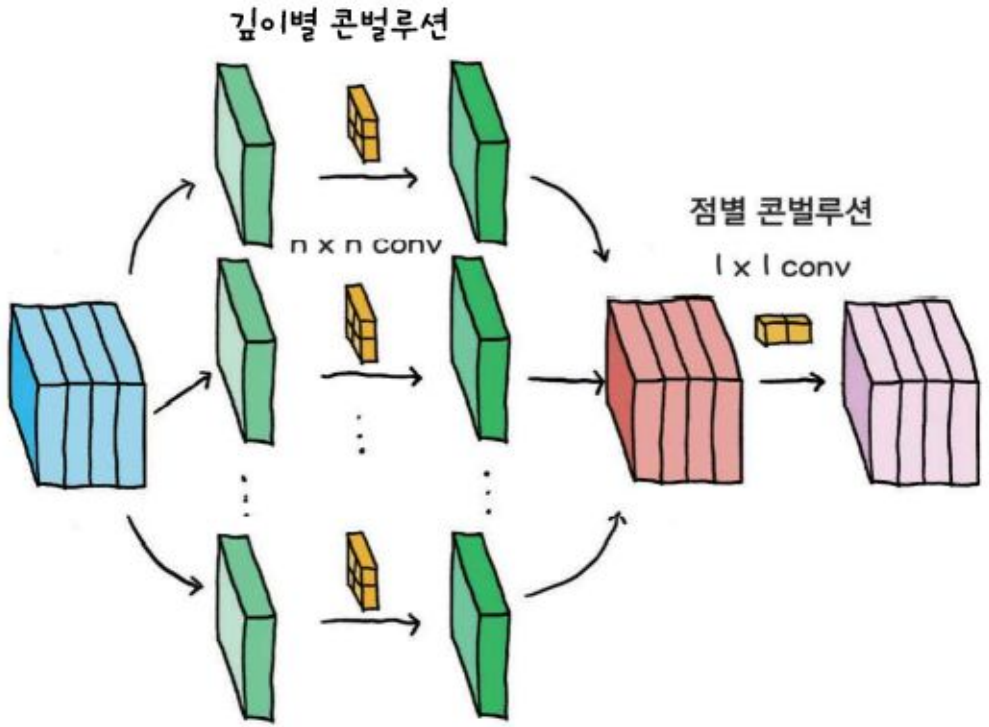


그림 6-57 깊이별 분리 콘벌루션

- **Depthwise Conv** : 공간적 특징 추출
- **Pointwise Conv** : 채널간 특징 추출

- 표준 컨볼루션보다 **연산량이 8~9배 줄어든다.**

- 기존에 사용하던 **MS-TCN**의 헤드부분에 추가하여 **DS-MS-TCN** 을 최종적으로 사용

*Pingchuan Ma., et al. "Towards Practical Lipreading with Distilled and Efficient Models." *arXiv preprint arXiv: 2007.06504* (2020)

04. ShuffleNet-TCN - Key Contribution

3. Knowledge Distillation

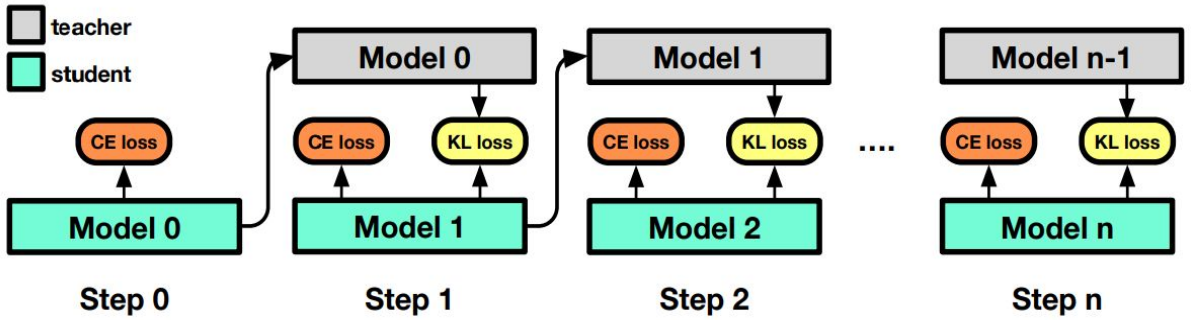


Fig. 1: The pipeline of knowledge distillation in generations

Knowledge Distillation :
큰 모델(Teacher Network)로부터 증류한 지식
→ 작은 모델(Student Network)로 transfer하는 과정

transfer learning : 서로 다른 도메인에서 지식을 전달하는 방식

knowledge distillation : 같은 도메인의 B모델에게 A모델이 가진 지식을 전달하는 방식(Model Compression 효과)

self-distillation 과정은 더 이상의 개선이 관찰되지 않을 때까지 반복된다.

*Pingchuan Ma., et al. "Towards Practical Lipreading with Distilled and Efficient Models." *arXiv preprint arXiv: 2007.06504* (2020)

05. LRWR

LRWR: Large-Scale Benchmark for Lip Reading in Russian language

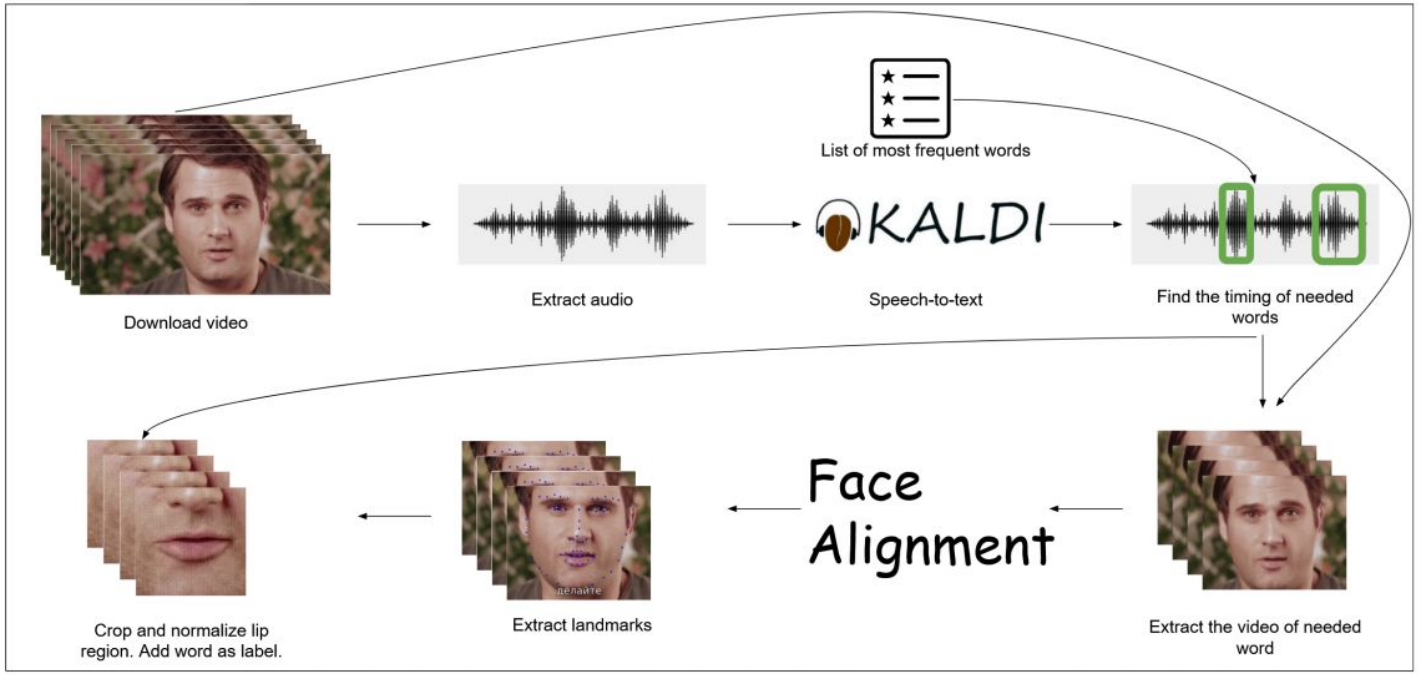


Figure 1: *Data collection pipeline*

*Evgeniy Egorov., et al. "LRWR: Large-Scale Benchmark for Lip Reading in Russian language." *arXiv preprint arXiv: 2109.06692* (2021)

Datasets

Video-LipReading-to-Script

01. 한국어 데이터셋 구축 - 계기

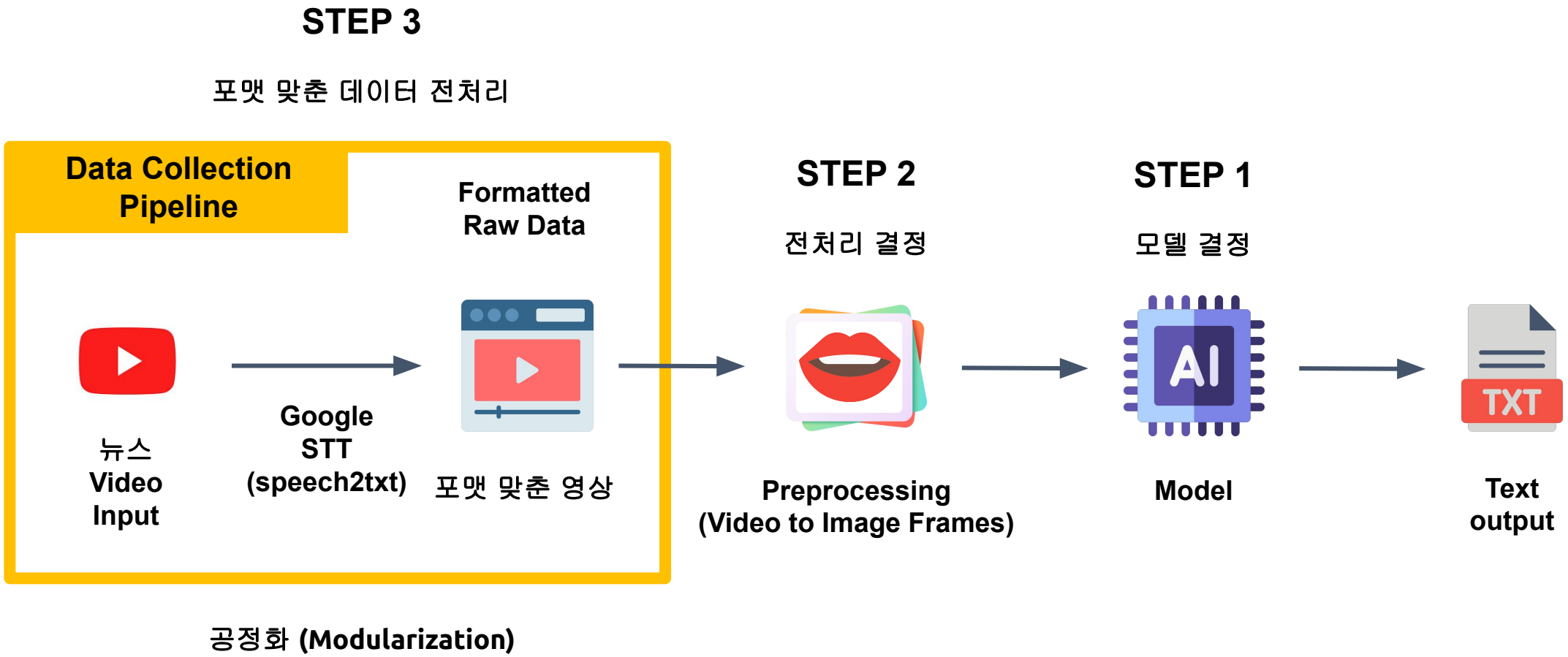
ENGLISH	MIRACL-VC1	**Grid Corpus**	**LRW**
CHINESE	CMLR	LRW-1000	
GERMAN	GLips		
KOREAN	신체 말단 움직임 영상		

01. 한국어 데이터셋 구축 - 계기

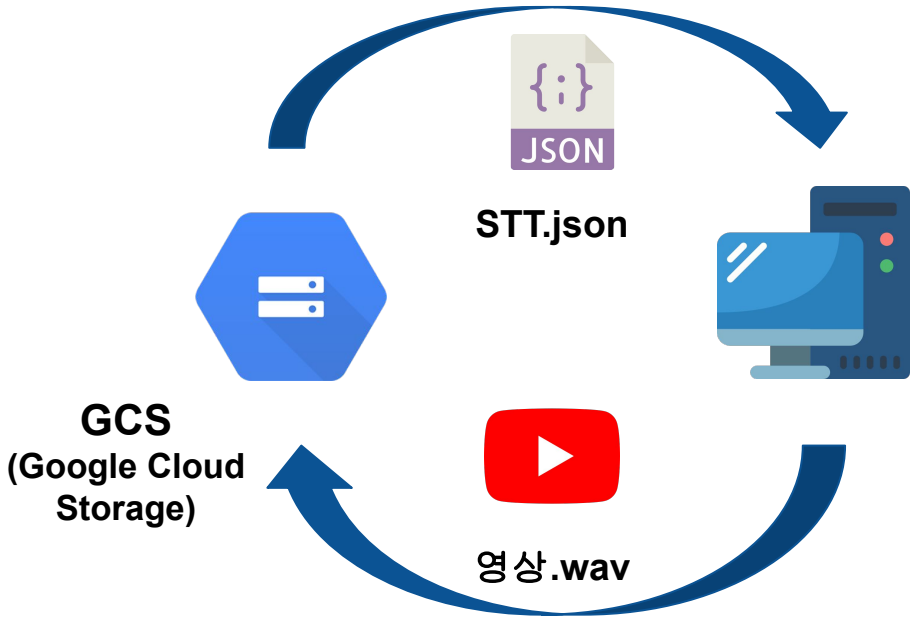
	신체 말단 움직임 영상	Grid Corpus	LRW
video			

align	{
	"annotations": [
	{
	"id": 230615,
	"image_name": "0001_M003_C_0000000.jpg",
	"image_id": 246541,
	"video_id": 42261,
	"bbox": [...],
	"actor_id": "M003",
	"word": "그",
	"word_id": 1,
	"num_keypoints": 24,
	"2D keypoints": [...]
	]
	}
		0 11000 sil
		11000 17000 bin
		17000 25250 blue
		25250 32750 at
		32750 36500 e
		36500 47750 one
		47750 60250 now
		60250 74500 sil
			Disk reference: 6221443953311207281
			Channel: BBC One HD
			Program start: 2015-11-26 13:00:00 +0000
			Clip start: 1428.12 seconds
			Duration: 0.45 seconds

01. 한국어 데이터셋 구축 - 프로세스



01. 한국어 데이터셋 구축 - Google STT 자동화



STEP 1

1. 영상 다운받고 **GCS** 로 옮기기
2. **STT(Speech-To-Text)** 받아오기

STEP 2

STT 타임스탬프로
단어별로 잘라내기

그리고_00036.avi 에 해당하는
그리고_00036.txt 파일

```
Disk reference: 1
Channel: Sejong
Program start: 2022-05-30 18:40:00 +0000
Clip start: 63.3 seconds
Duration: 1.2 seconds
```

STEP 3

alignment 포맷 txt

01. 한국어 데이터셋 구축 - 영상 선정

명확한 발음

정면

화면 전환 X

영상 효과

옆면

오프닝

아나운서
[뉴스 데이터]

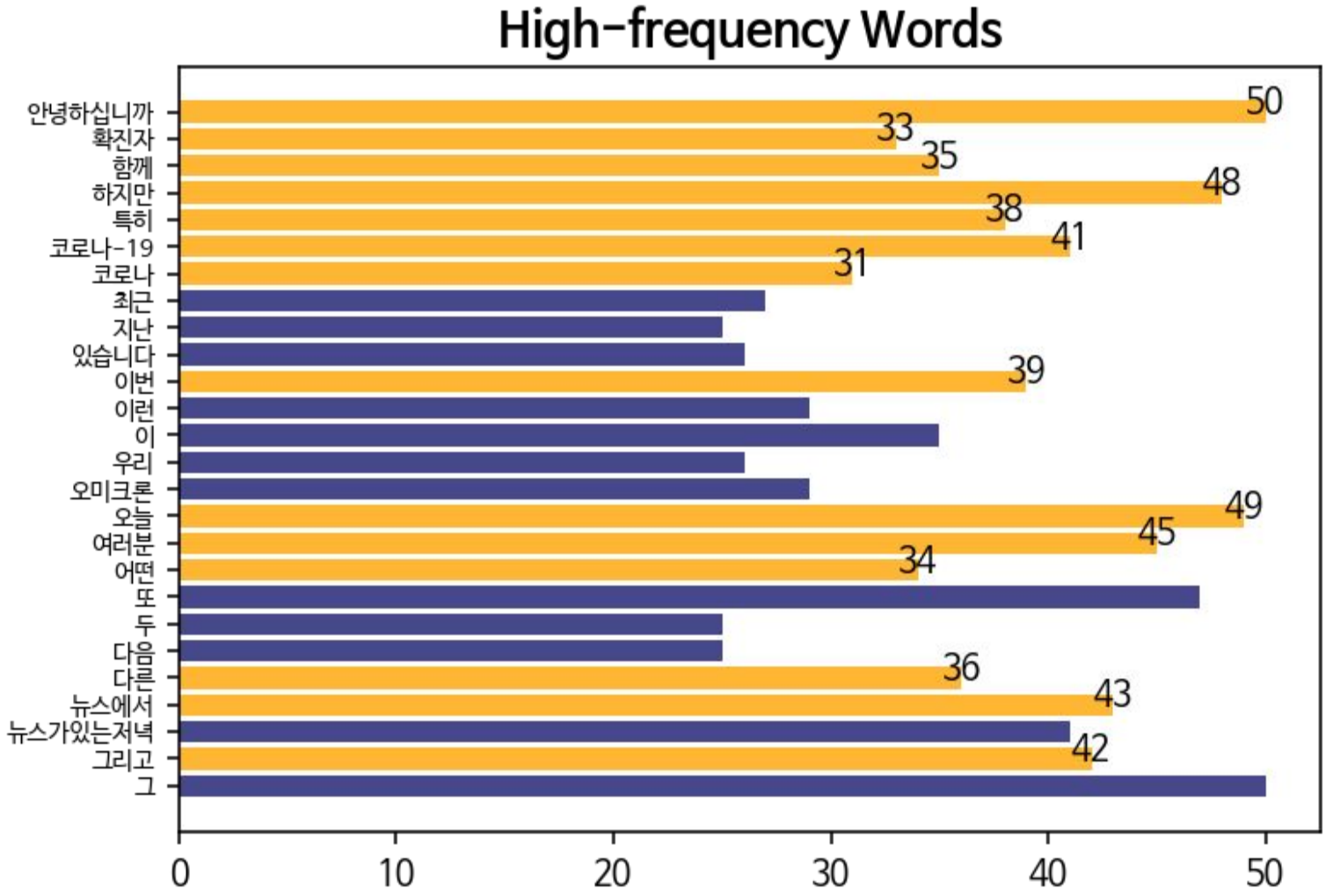
02. Preprocessing - 총 데이터

02. Preprocessing - 데이터 선정

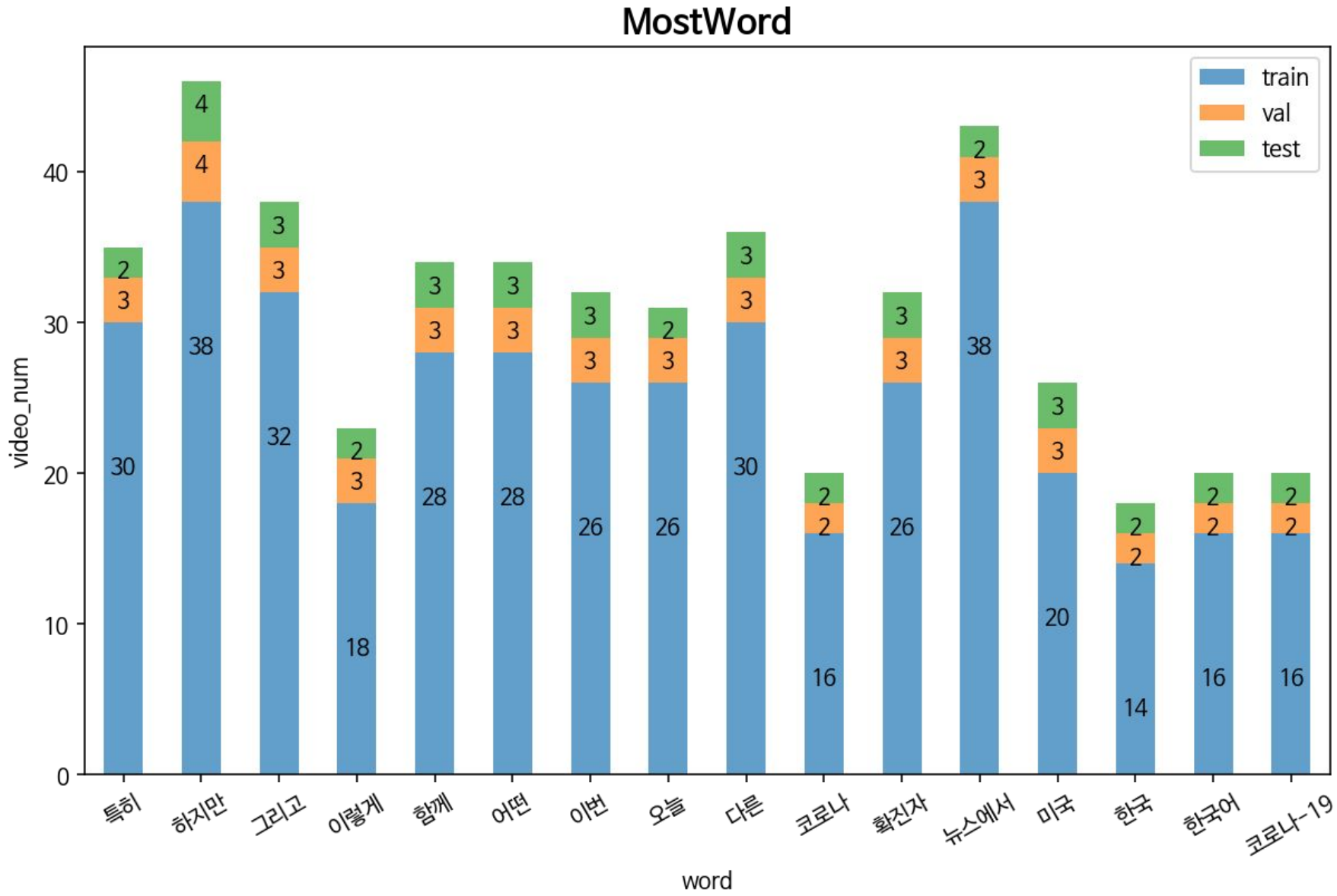
단어 꾸러미에 담은 것: 16개
<face_recognition 영상 고르기 전>

"안녕하십니까 : 50"
"확진자 : 33"
"함께 : 35"
"하지만 : 48"
"특히 : 38"
"코로나 : 31"
"코로나-19 : 41"
"이번 : 39"
"오늘 : 49"
"여러분 : 45"
"어떤 : 34"
"다른 : 36"
"뉴스에서 : 43"
"그리고 : 42"

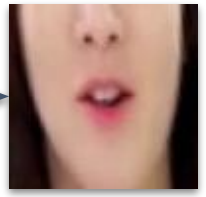
(+이렇게, 한국, 한국어, 미국, 여러분)



02. Preprocessing - 사용 데이터



02. Preprocessing - Face Landmark → 입술 Crop



STEP 1

Video
→ Image(프레임 추출)

STEP 2

Grayscale 변환
→ 프레임 29개로 맞추기
→ Face Landmark 찾기

STEP 3

입술 crop (96,96) 크기
→ numpy 변환
→ .npz 파일 저장

Experiments

Video-LipReading-to-Script

01. 모델 학습 결과 비교

1. LipNet

```
Epoch 0: Curriculum(train: False, sentence_length: -1, flip_probability: 0.5, jitter_probabili
.05)
Epoch 0: Curriculum(train: False, sentence_length: -1, flip_probability: 0.5, jitter_probabili
.05)
Epoch 0: Curriculum(train: False, sentence_length: -1, flip_probability: 0.5, jitter_probabili
.05)
/root/.pyenv/versions/3.6.9/lib/python3.6/site-packages/nltk/translate/bleu_score.py:472: User
ng:
Corpus/Sentence contains 0 counts of 3-gram overlaps.
BLEU scores might be undesirable; use SmoothingFunction().
  warnings.warn(_msg)


[Epoch 0] Out of 256 samples: [CER: 10.000 - 0.833] [WER: 4.000 - 1.000] [BLEU: 0.351 - 0.351]

['게 께', '께 께']
```

Input 단어 ‘함께’ 라벨 -> **Output** 단어 ‘게 께’ 텍스트

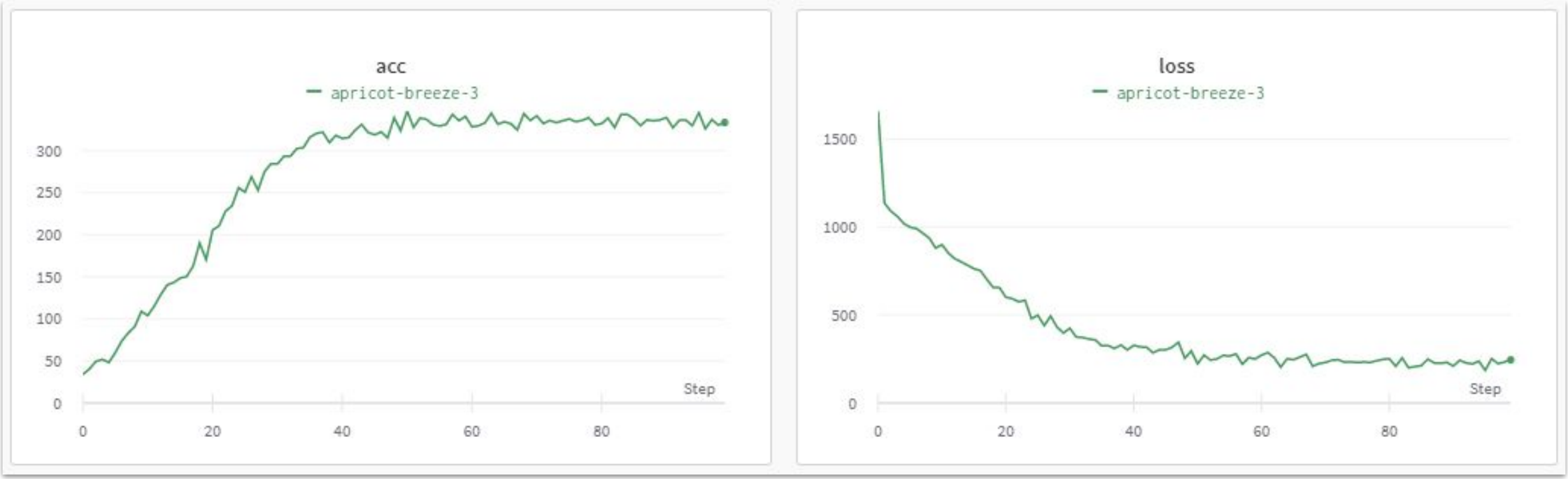
2. ShuffleNet-TCN

```
Model and log being saved in: ./train_logs/tcn/2022-06-04T16:30:01
2-norm of the neural network: 48.1286
Partition train loaded
Partition val loaded
Partition test loaded
Model has been successfully loaded from ./train_logs/tcn_backup/2022-06-04T12:13:35/ckpt.best.pth.tar
  0%|
                                                                                | 0/2 [00:00<?, ?it/s]/home
eading_using_TCN_running/Lipreading_using_Temporal_Convolutional_Networks-master/lipreading/dataset.py
rWarning: Creating a tensor from a list of numpy.ndarrays is extremely slow. Please consider convertin
t to a single numpy.ndarray with numpy.array() before converting to a tensor. (Triggered internally at
h/csrc/utils/tensor_new.cpp:210.)
  data = torch.FloatTensor(data)

Prediction: 오늘
Confidence: 1.0

 50%|
                                                                                | 1/2 [00:01<00:01,  1.13s/it]
Prediction: 오늘
Confidence: 1.0

100%|
                                                                                | 2/2 [00:01<00:00,  1.10it/s]
Test Dataset 5 In Total          CR: 1.0
Test-time performance on partition test: Loss: 0.0000   Acc:1.0000
```

Input 단어 ‘오늘’ 영상 -> **Output** 단어 ‘오늘’ 텍스트

02. Train 시각화

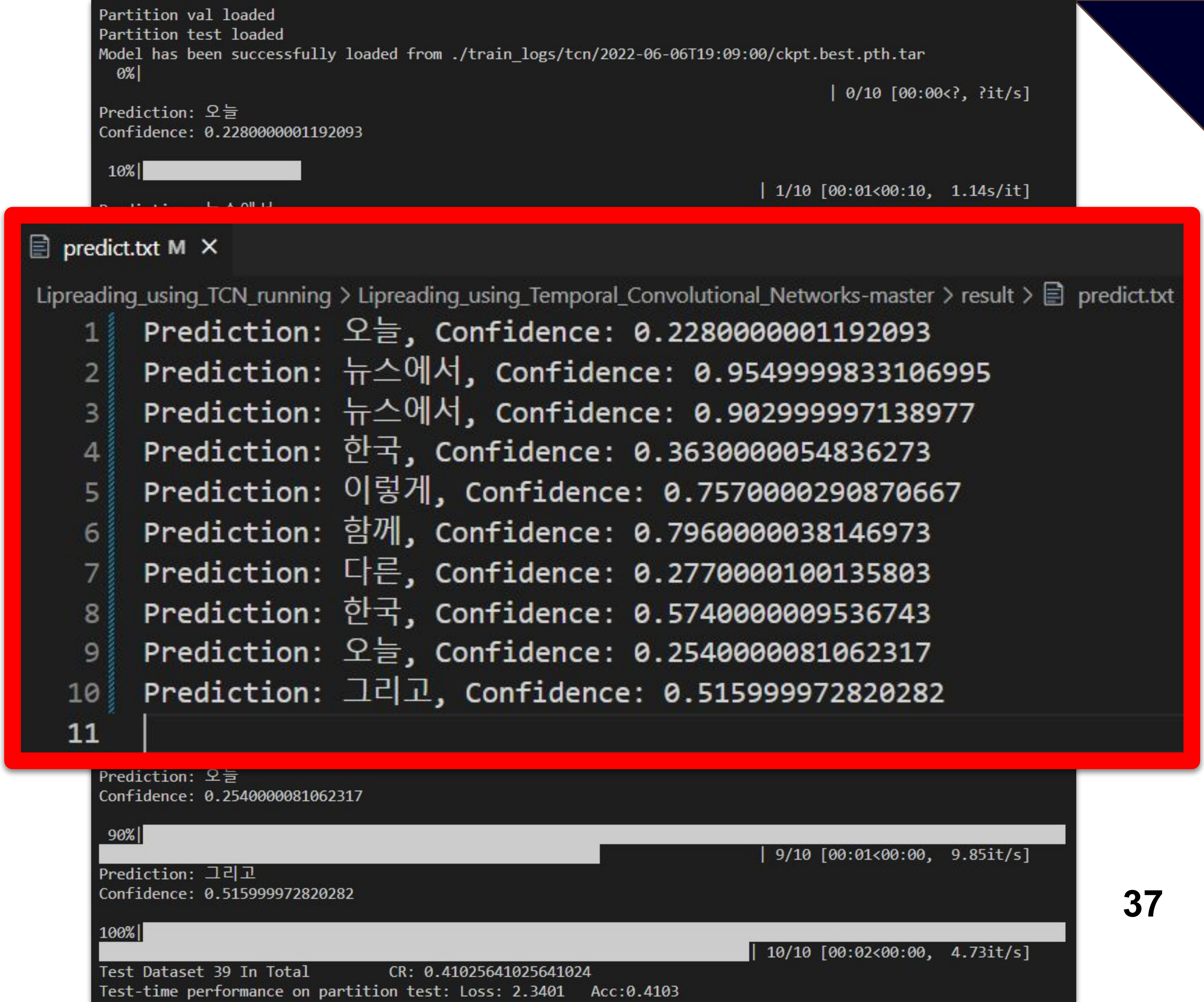
Using Wandb Tool



- **100 epochs**
- **Metrix: Acc, Loss**
- 학습이 진행될 수록 **Acc** 는 높아지고 **Loss** 는 낮아지는 것을 확인함

03. Test 결과

- **100 epochs**
- **Acc Avg 0.4** 로 낮은 수치를 보임
- 이후에 파라미터나 학습 횟수 변경을 통해 개선이 필요함



Demo

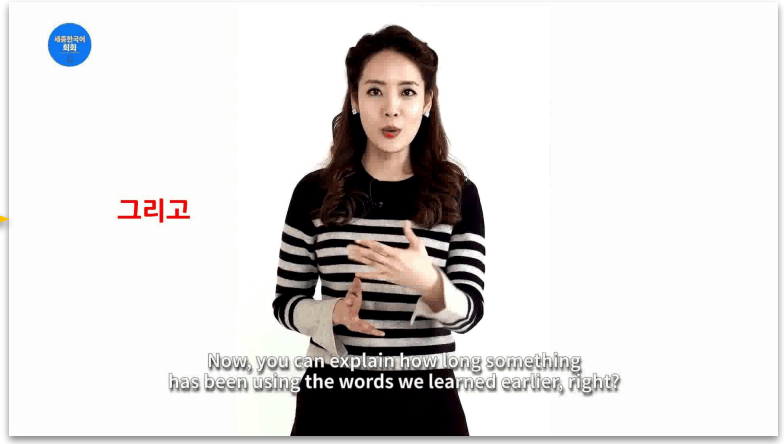
Video-LipReading-to-Script

01. 데모 실행

Input: ‘함께’ 단어 영상

```
------------ START ------------

------------ FRAME 0 ------------
------------ FRAME 1 ------------
------------ FRAME 2 ------------
------------ FRAME 3 ------------
------------ FRAME 4 ------------
------------ FRAME 5 ------------
------------ FRAME 6 ------------
------------ FRAME 7 ------------
------------ FRAME 8 ------------
------------ FRAME 9 ------------
------------ FRAME 10 ------------
------------ FRAME 11 ------------
------------ FRAME 12 ------------
------------ FRAME 13 ------------
------------ FRAME 14 ------------
------------ FRAME 15 ------------
------------ FRAME 16 ------------
------------ FRAME 17 ------------
------------ FRAME 18 ------------
------------ FRAME 19 ------------
------------ FRAME 20 ------------
------------ FRAME 21 ------------
------------ FRAME 22 ------------
------------ FRAME 23 ------------
------------ FRAME 24 ------------
------------ FRAME 25 ------------
------------ FRAME 26 ------------
------------ FRAME 27 ------------
------------ FRAME 28 ------------

------------ PREDICT ------------

Prediction: 그리고
Confidence: 0.7929999828338623

------------ END ------------


------------ GIF OUTPUT ------------


------------ GIF DONE ------------
```

1초 영상 처리 시간 → 15초



Output: ‘그리고’ 자막, 자막붙은 영상

02. 데모 분석 - Bad Case Analysis

결과 원인 분석

1. 학습에 사용한 데이터가 적어서 데모 output 이 별로인 것으로 판단됨
2. Google STT API 가 불완전하기 때문에 구축한 영상 데이터의 Loss 존재
3. 하이퍼 파라미터의 최적값을 찾아내지 못함

개선 방법

1. 학습에 사용할 데이터셋을 더 많이 늘려서 train, test 시도
2. 음성 데이터와 대조하여 영상 데이터의 Loss 를 줄여서 구축
3. 모델 학습에 사용한 하이퍼 파라미터 변경 및 학습 횟수 변경을 통해 개선

Future Works

Video-LipReading-to-Script

01. 기대효과 및 활용방안

외국 상용 서비스

- 외국 독순술 딥러닝 서비스 X
- 독순술은 대부분 사람이 투입됨
- Deep learning Framework
  AV-HuBERT 공개되어 있음

한국어 서비스 기대효과

- 소리가 겹치는 상황에서 도움됨
- 인건비 절감
- 한국어 독순술 컨텐츠 실현
  - 음성인식 기술과의 시너지
  - 한국어 발음 교육 사업
- 장애인들의 불편함 해소

02. 향후 발전 가능성

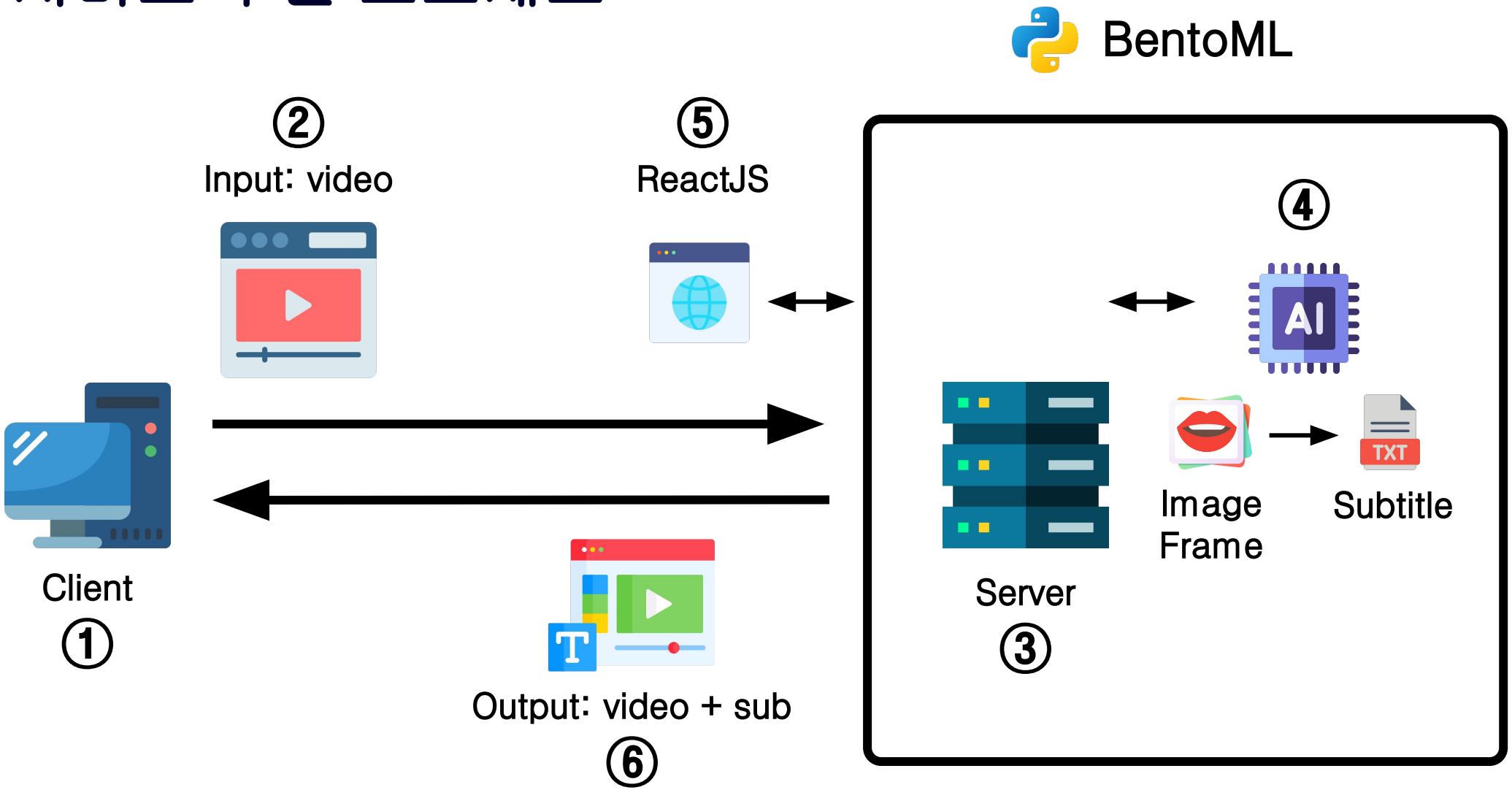
1. BentoML을 통한 Web Service, 앱 개발을 통한 App Service 제공
2. 실시간 자막 생성 서비스도 기대해 볼 수 있음
3. 음성인식 자막 서비스와 결합
4. 소리 없는 영상에 대한 자막 출력
5. Face Detection Module을 변경하여 다양한 각도에서의 Detection 성능 개선

03. 앞으로의 개발 일정

내용 \ 주		M1	M2	H1	H2	H3	H4	H5	H6
1단계	데이터셋 이해/제작								
	관련 논문 리뷰								
	모델 설정								
	검수								
2단계	서비스 개발								
	검수								
3단계	전체 유지보수								

M 미니 해커톤 / H 해커톤

04. 서비스 구현 프로세스



Our Links

1. Our Figma
2. Our Notion
3. Our GitHub
4. Our PPT

References

1. LipNet: End-to-End Sentence-level Lipreading [Paper] [GitHub_1] [GitHub_2]
2. Towards Practical Lipreading with Distilled and Efficient Models [Paper] [GitHub]
3. LRWR: Large-Scale Benchmark for Lip Reading in Russian language [Paper]

Q&A

Thank you

빠끔빠끔

부록

02. ShuffleNet-TCN - Key Contribution

3. Knowledge Distillation

Method	Top-1 Acc. (%)
3D-CNN [23]	61.1
Seq-to-Seq [9]	76.2
ResNet34 + BLSTM [6]	83.0
ResNet34 + BGRU [24]	83.4
2-stream 3D-CNN + BLSTM [25]	84.1
ResNet-18 + BLSTM [26]	84.3
ResNet-18 + BGRU + Cutout [27]	85.0
ResNet-18 + MS-TCN [21]	85.3
ResNet-18 + MS-TCN - Teacher	85.3
ResNet-18 + MS-TCN - Student 1	87.4
ResNet-18 + MS-TCN - Student 2	87.8
ResNet-18 + MS-TCN - Student 3	**87.9**
ResNet-18 + MS-TCN - Student 4	87.7
Ensemble	**88.5**

Born-Again Distillation

동일 모델을 사용한 self-distillation 방식

Student Backbone (Width mult.)	Student Back-end (Width mult.)	Distillation	Top-1 Acc.	Params $\times 10^6$	FLOPs $\times 10^9$
ResNet-18 [21]	MS-TCN(3×)	-	41.4	36.7	15.78
3D DenseNet [20]	BGRU (256)	-	34.8	15.0	30.32
ShuffleNet v2 (1×)	TCN (1×)	✗	40.7	3.9	1.73
	TCN (1×)	✓	41.4	3.9	1.73
ShuffleNet v2 (1×)	DS-TCN (1×)	✗	39.1	2.5	1.68
	DS-TCN (1×)	✓	40.4	2.5	1.68
ShuffleNet v2 (0.5×)	TCN (1×)	✗	40.5	3.0	0.89
	TCN (1×)	✓	41.1	3.0	0.89
ShuffleNet v2 (0.5×)	DS-TCN (1×)	✗	39.1	1.6	0.84
	DS-TCN (1×)	✓	40.2	1.6	0.84

Sequential Distillation

Ablation Study를 통해 성능이 높은 모델을 차용 하기 위해, Standard distillation을 진행한다.

ResNet-18 → ShuffleNet V2 → ShuffleNet V2

(MS-TCN) (MS-TCN) (DS-MS-TCN)

*Pingchuan Ma., et al. "Towards Practical Lipreading with Distilled and Efficient Models." *arXiv preprint arXiv: 2007.06504* (2020)