Improving Continuous Japanese Fingerspelling Recognition with Transformers: A Comparative Study against CNN-LSTM Hybrids

○Akihisa Shitara[†*], Yuhki Shiraishi[*]

[†]Graduate School of Library, Information, and Media Studies, University of Tsukuba, Japan
[*]Faculty of Industrial Technology, Tsukuba University of Technology, Japan

Email: theta-akihisa@digitalnature.slis.tsukuba.ac.jp

DigitalWorld 2025 Congress
The Eighteenth International Conference on Advances in Computer-Human Interactions
ACHI 2025

May 18, 2025 to May 22, 2025 - Nice, France

Presenter: Akihisa Shitara

- d/Deaf
- Ph.D Student
  - Graduate School of Library, Information, and Media Studies, University of Tsukuba
- Part-time Researcher
  - Faculty of Industrial Technology, Tsukuba University of Technology
- Research Topics
  - Accessibility
  - Human Computer Interaction
  - Deaf Physicality

Contents

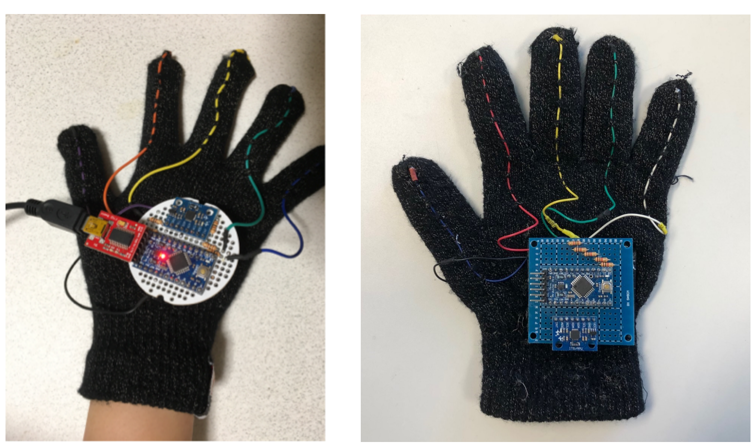
Contents

Introduction Part1

To achieve smooth communication between
d/Deaf and hard of hearing and hearing people

Development of
Continuous Japanese Fingerspelling Recognition System
using **Sensor Gloves** and **Deep Learning**[1-2]

Sensor Gloves



1st Proto 2nd Proto[1-2]

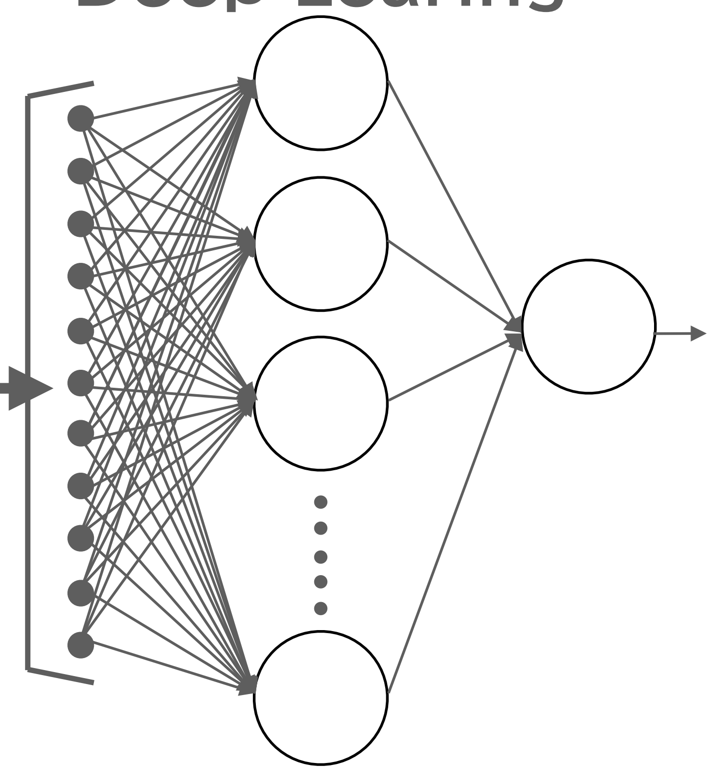
Conductive Fiber Weaving Techniques

Finger movement
(5 dims)

IMU Sensors

Acceleration / Angular
(3 dims / 3 dims)

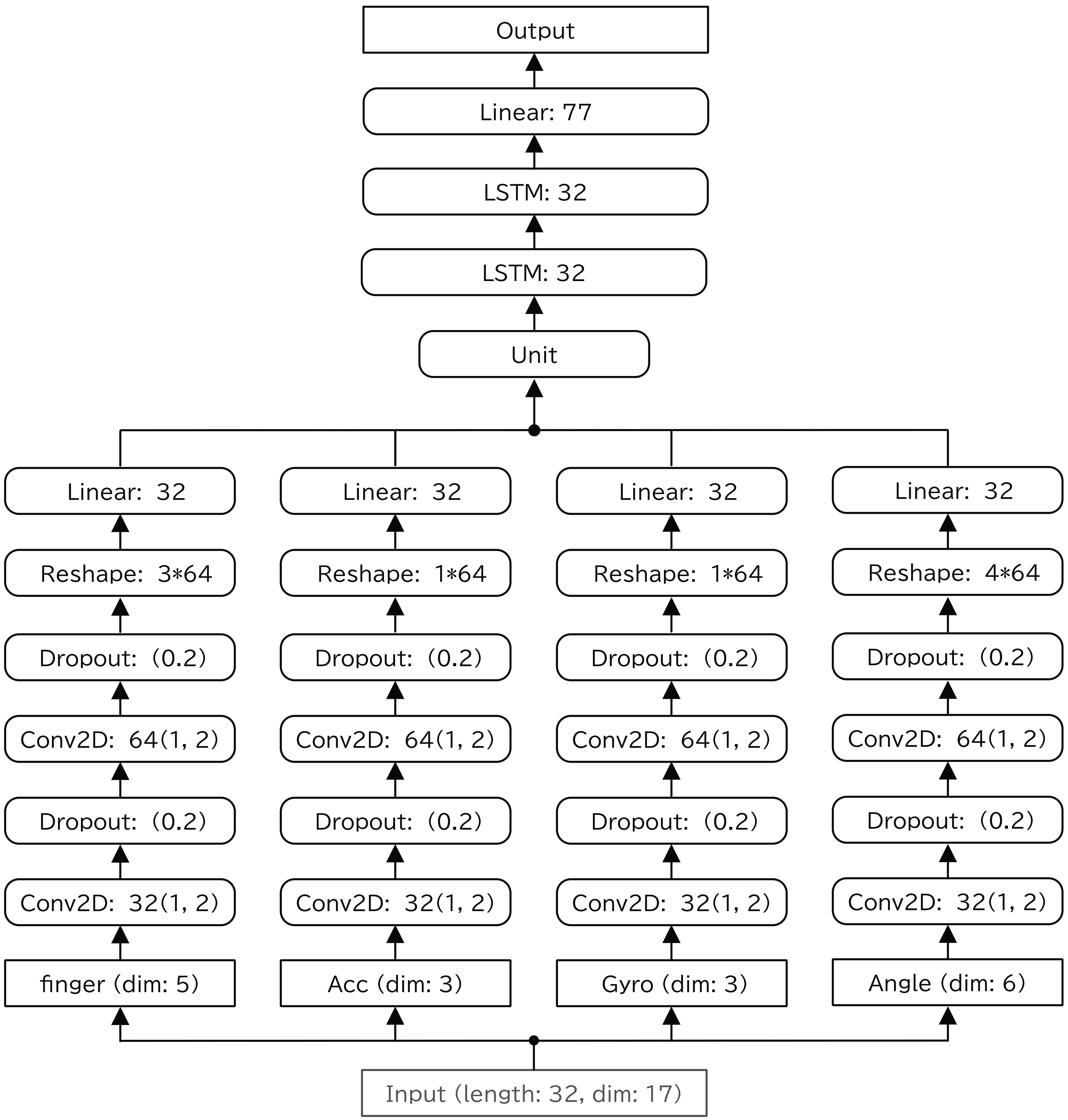
Deep Learing



Output Label

[1] Tomohiko Tsuchiya, Akihisa Shitara, Fumio Yoneyama, Nobuko Kato, Yuhki Shiraishi, Sensor Glove Approach for Japanese Fingerspelling Recognition System Using Convolutional Neural Networks, Proceedings of The Thirteenth International Conference on Advances in Computer-Human Interactions (ACHI 2020), pp.152-157, 2020
[2] Yuhki Shiraishi, Akihisa Shitara, Fumio Yoneyama, Nobuko Kato, Sensor Glove Approach for Continuous Recognition of Japanese Fingerspelling in Daily Life, International Journal on Advances in Life Sciences, Vol.14, No. 3&4, pp.53-70, 2022

Introduction Part2



CNN-LSTM combined machine learning model:
2CNN-2LSTM

Features

- Branch of Input dims
  1. Finger movement (5 dims)
  2. Acceleration (3 dims)
  3. Angular (3 dims)
  4. Angle (sin, cos. 6 dims)
- Branch of Input dims for CNN
- Combin of Output dims by CNN
- Combin of Output dims for LSTM

Accuracy

- Avaerage micro F-measure
  92.1%
- Avaerage macro F-measure
  64.7%

Contents

Related Work Part1

Recognition Method	Features	Points
Image Recognition	Hand → Camera → PC	- Occlusion factors - Environmental factors
Sensor glove Recognition	Hand → Sensor glove  → PC	- Sensors can be attached - Easy to perform

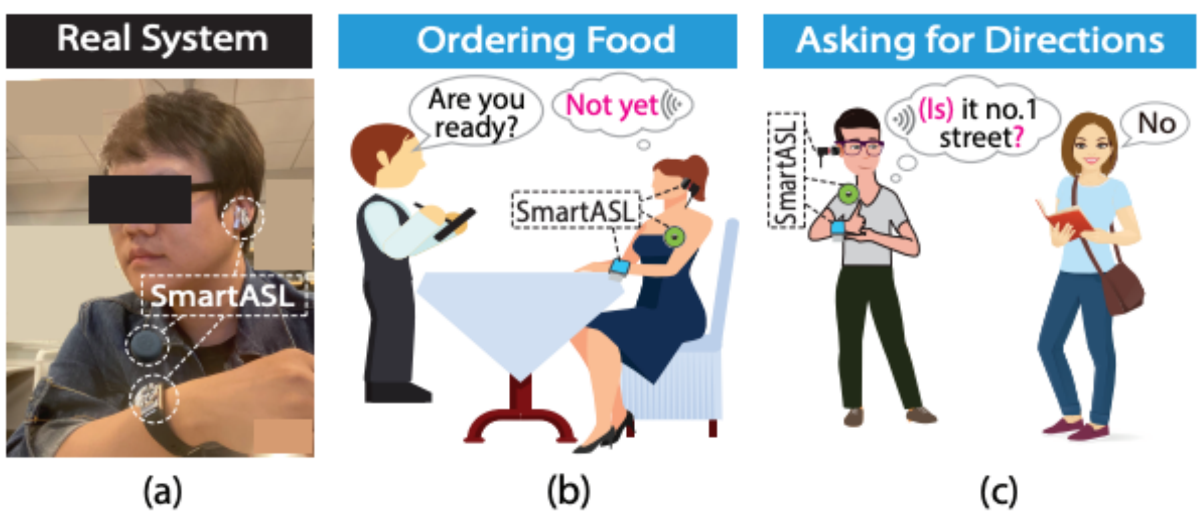
Image Recognition

1. Survey of related research on Sign Language Recognition[14]
   - CNN, LSTM, and **Transformer** are used in many research
2. Reported Survey of SOTA in Sign Language Recognition[15]
   - **Transformer** is used in many cases
3. Emergence of machine learning models based on **Transformer**
   - Spatial Temporal Graph Convolutional Networks (ST-GCN)[18] + Transformer, Conformer[19], Signformer[21]

[14] Patel, C. C. and Patel, P.: A Comparative Analysis of Sign Language Recognition Approaches Across Varied Sign Languages, Universal Threats in Expert Applications and Solutions, Singapore, Springer Nature Singapore, pp. 355-372 (2024)
[15] De Coster, M., Shterionov, D., Van Herreweghe, M. and Dambre, J.: Machine translation from signed to spo ken languages: state of the art and challenges, Universal Access in the Information Society, Vol. 23, Singapore, Springer Nature Singapore, pp. 1305-1331 (2024)
[18] Yan, S., Xiong, Y. and Lin, D.: Spatial Temporal Graph Convolutional Networks for Skeleton-Based Ac- tion Recognition, Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, No. 1 (online), DOI: 10.1609/aaai.v32i1.12328 (2018)
[19] Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y. and Pang, R.: Conformer: Convolution-augmented Transformer for Speech Recognition, Interspeech 2020, pp. 5036-5040 (online), DOI: 10.21437/Interspeech.2020-3015 (2020).
[21] Yang, E.: Signformer is all you need: Towards Edge AI for Sign Language (2024)

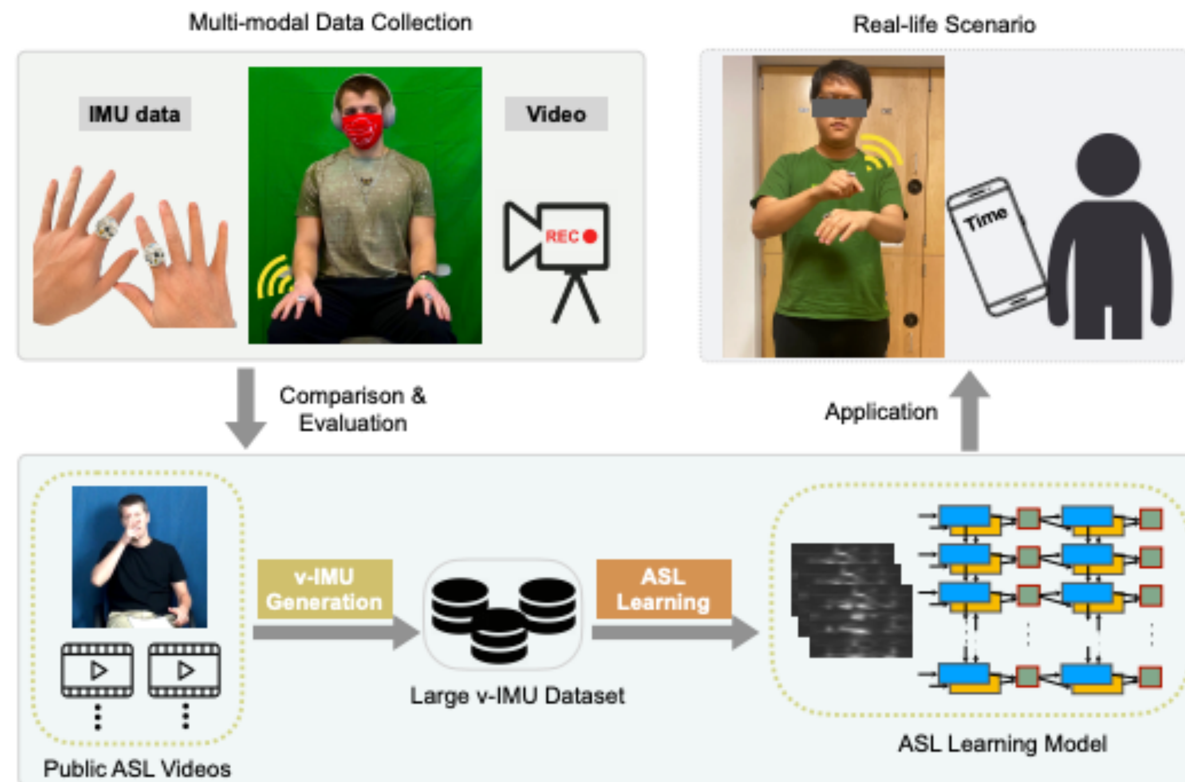
Related Work Part2

Sensor glove Recognition



SmartASL[30]

Earphone, Smartwatch



SignRing[32]

Ring-shaped devices

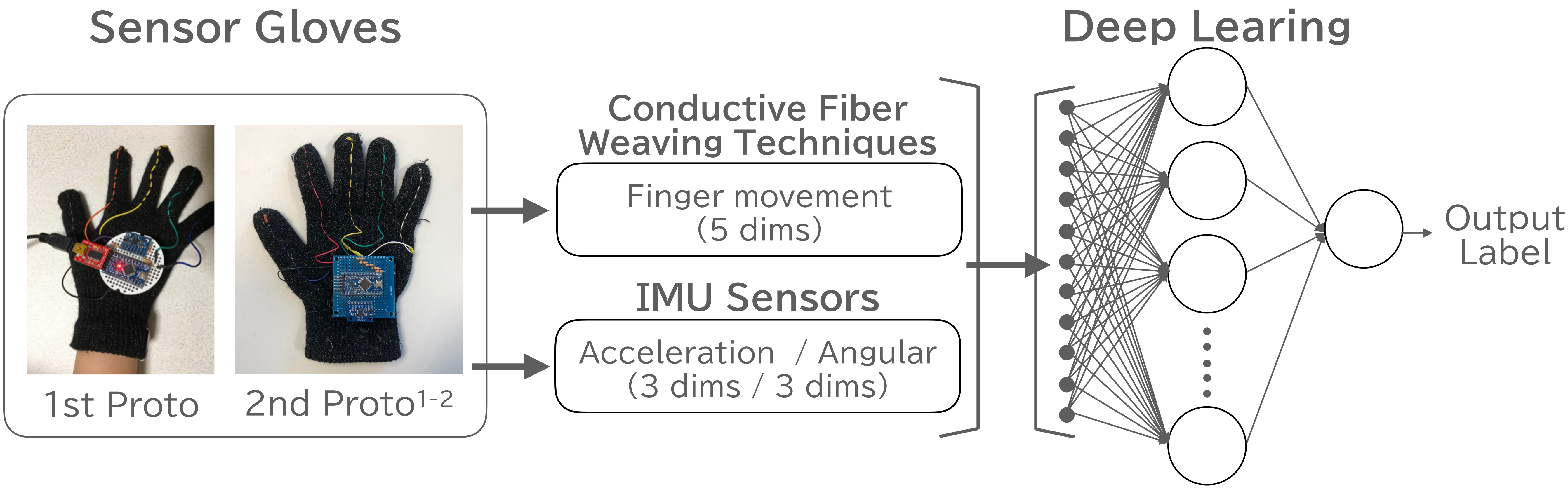
<u>IMU Sensor</u> and **<u>Transformer</u>** is used

[30] Jin, Y., Zhang, S., Gao, Y., Xu, X., Choi, S., Li, Z., Adler, H. J. and Jin, Z.: SmartASL: "Point-of-Care" Comprehensive ASL Interpreter Using Wearables, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., Vol. 7, No. 2 (online), DOI: 10.1145/3596255 (2023)

[32] Li, J., Huang, L., Shah, S., Jones, S. J., Jin, Y., Wang, D., Russell, A., Choi, S., Gao, Y., Yuan, J. and Jin, Z.: SignRing: Continuous American Sign Language Recog- nition Using IMU Rings and Virtual IMU Data, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., Vol. 7, No. 3 (online), DOI: 10.1145/3610881 (2023).

Purpose and Hypothetical

Development of
Continuous Japanese Fingerspelling Recognition System
using **Sensor Gloves** and **Deep Learning**[1-2]



Transformer Encoder could:

1. Improving **Recognition rates**?
2. Handling **individual differences among signers expressing**?

Contents

First Comparative Analysis Part1

Development Environment

- Ubuntu 22.04
- NVIDIA CUDA 12.3.0
- Python 3.10
- PyTorch 2.2.0a0+6a974bec

Compare Model

- 2LSTM (batchsize: 64)
- branch-2CNN-unit-2LSTM (batchsize: 64)
- Transformer Encoder (batchsize: 16)
- branch-CNN-unit-Transformer Encoder (batchsize: 32)
- branch-2CNN-unit-Transformer Encoder (batchsize: 32)

Dataset is same. (described later)

First Comparative Analysis Part2

Sensor Gloves

1st Proto

2nd Proto[1-2]

Conductive Fiber Weaving Techniques

Finger movement
(5 dims)

IMU Sensors

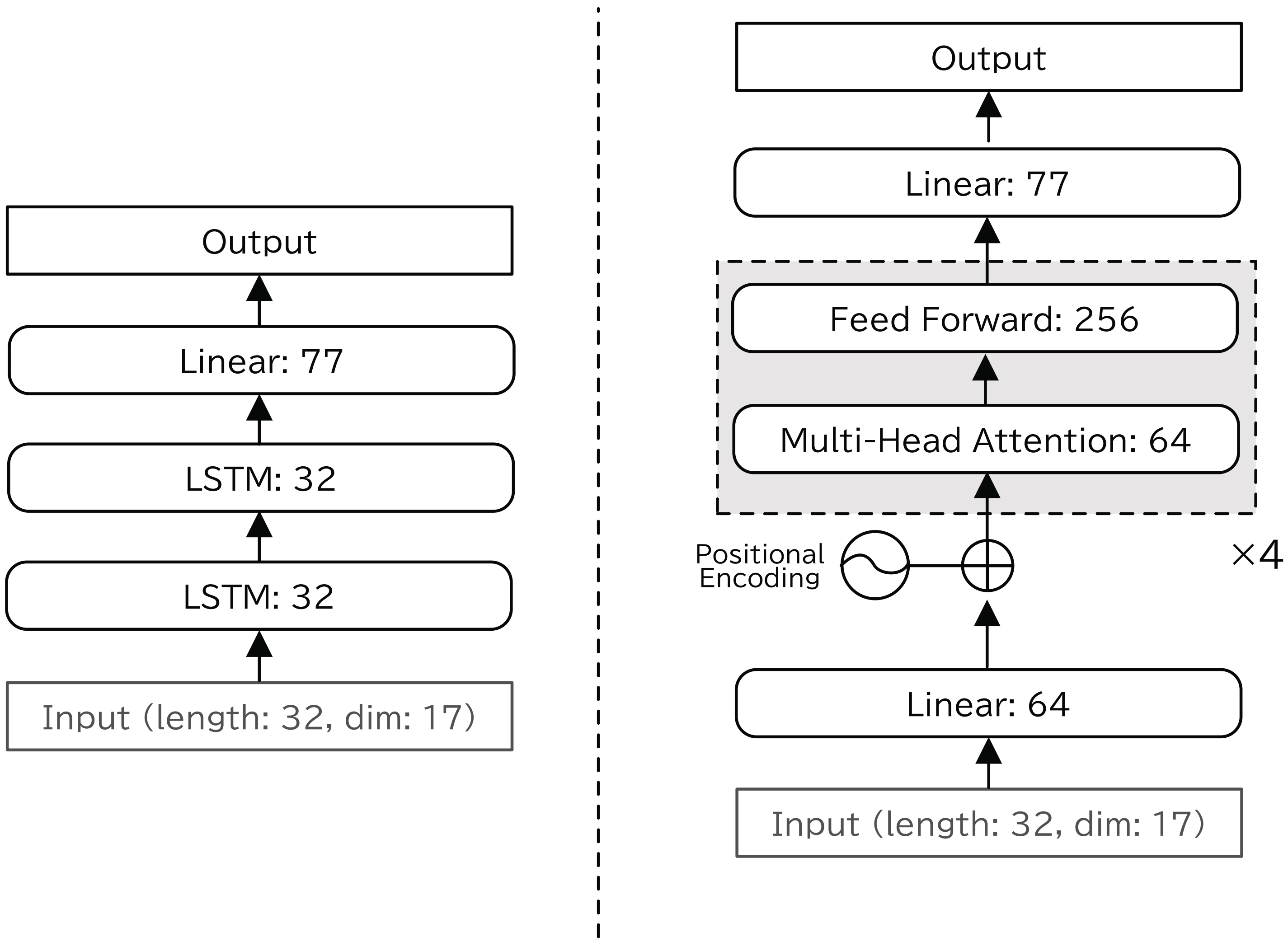
Acceleration / Angular
(3 dims / 3 dims)

Continuous Japanese fingerspelling dataset

- 33 Participants
- 64 words × 5 times (each person)
  - 1 word: 3~5 fingerspelling characters
- Input dims: 17 dims
  - add angles (calculated usng angular, 6 dims)
- Using moving average calculations: 32 samples each word (4sps × 8sec)
  - 960 samples each word (120sps × 8sec)

First Comparative Analysis Part3

1. Improving compared Transformer Encoder than LSTM?



(a) (b)

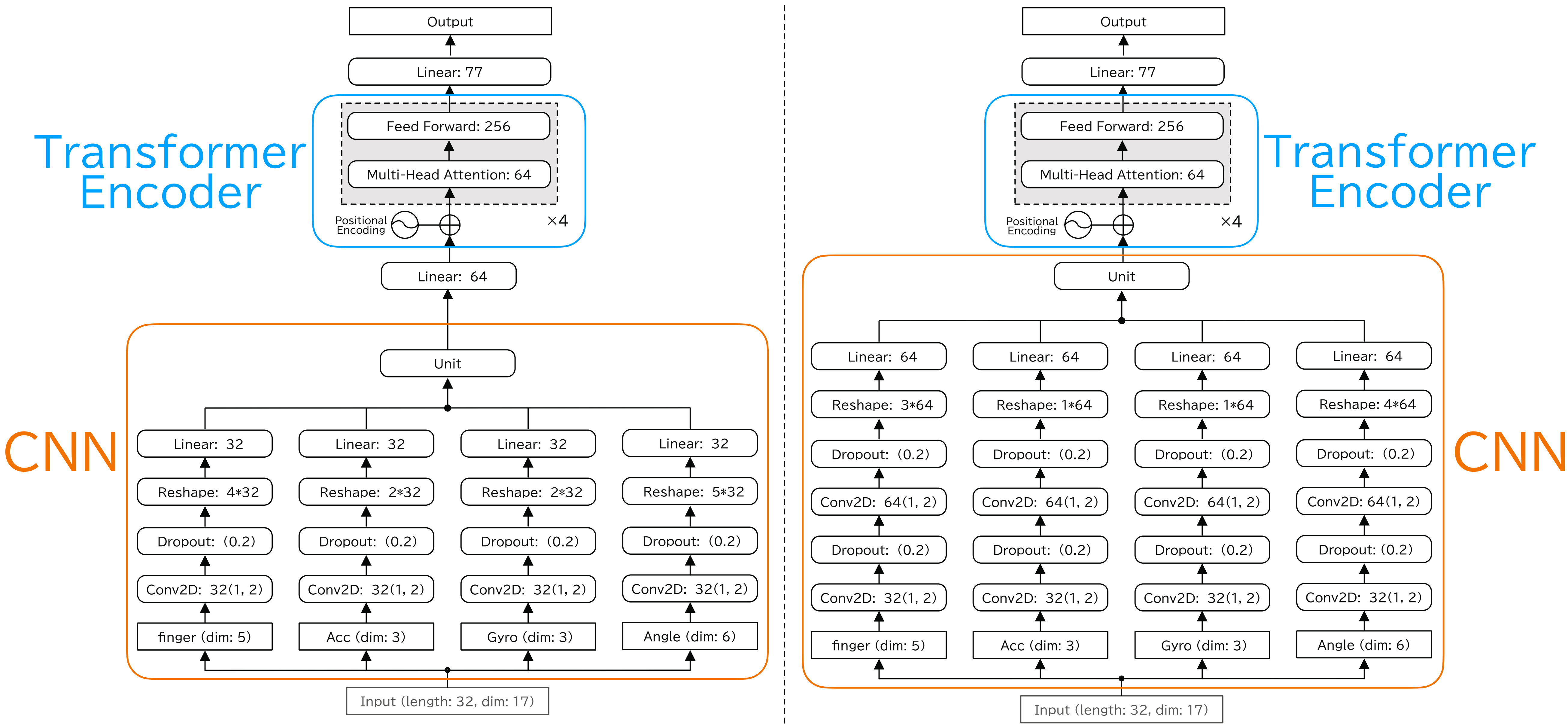
LSTM (a)

- Layer number : 2
- Hidden size : 32

Transformer Encoder (b)

- Layer number : 4
- Input (d_model) : 64
- Output (Feed Forwad) : 256

2. Improving Transformer Encoder combined with CNN?



(a) (b)

Branch → CNN → Combin (Same as previous study)

First Comparative Analysis Results Part1

k-fold CV (CrossValidation) Evaluation

- 5-fold CV, 10-fold CV
  - Input data was shuffled
  - Input data was divided into Training and Test data
  - Average and Standard Deviation of the 5 and 10 runs
  - Values at the epoch when the Validation Loss was minimized

model	k=5, F-measure [%]		k=10, F-measure [%]	
	micro	macro	micro	macro
2LSTM	89.9 (0.2)	50.5 (1.2)	90.2 (0.3)	51.6 (1.6)
branch-2CNN-unit-2LSTM	91.9 (0.1)	64.6 (0.7)	92.1 (0.2)	65.4 (1.1)
Transformer Encoder	92.8 (0.3)	72.5 (1.3)	93.3 (0.3)	74.9 (1.8)
branch-CNN-unit-Transformer Encoder	**93.4 (0.1)**	**75.8 (0.6)**	**93.6 (0.2)**	**76.5 (0.8)**
branch-2CNN-unit-Transformer Encoder	93.3 (0.2)	74.8 (0.9)	93.6 (0.3)	76.6 (1.0)

Transformer Encoder could:

1. **Improving Recognition rates**

First Comparative Analysis Results Part2

33-fold CV Evaluation

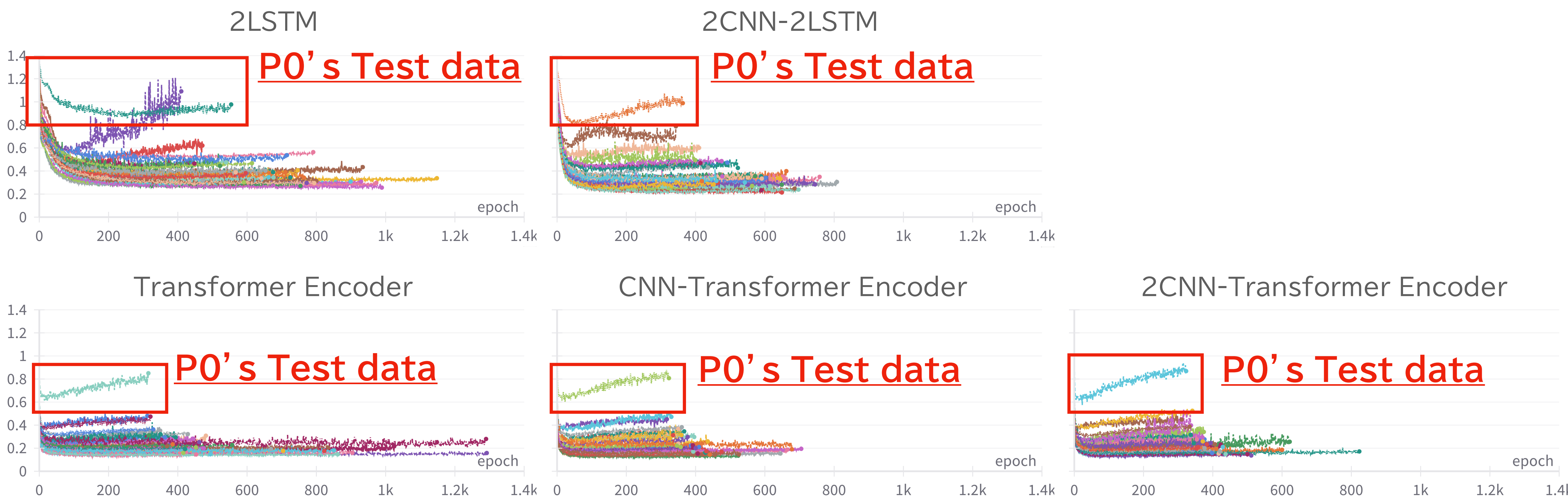
- **random**
  - Input data was shuffled
  - Input data was divided into Training and Test data
  - Average and Standard Deviation of 33 runs
  - Values at epoch when Validation Loss was minimized

- **person**
  - Input data for one person used as Test data
  - Input data for the remaining 32 people used Training data
  - Average and Standard Deviation of Test data each person
  - Values at epoch when Validation Loss was minimized

model	k=33, F-measure [%]			
	random		person	
	micro	macro	micro	macro
2LSTM	90.4 (0.4)	52.1 (2.1)	88.8 (2.2)	40.5 (9.6)
branch-2CNN-unit-2LSTM	92.2 (0.4)	66.0 (2.5)	90.0 (0.3)	50.7 (13.1)
Transformer Encoder	93.5 (0.5)	75.8 (2.4)	92.2 (2.8)	69.1 (10.3)
branch-CNN-unit-Transformer Encoder	93.8 (0.5)	77.7 (2.2)	92.4 (2.9)	70.8 (10.1)
branch-2CNN-unit-Transformer Encoder	93.8 (0.4)	77.1 (2.2)	92.4 (2.7)	70.8 (9.1)

First Comparative Analysis Results Part3

33-fold CV Evaluation



Found that significant Validation Loss for the same person's Test data

First Comparative Analysis Results Part4

2CNN-2LSTM

k=33		P0		P1	
chi	41.3	ho	0.0	me	0.0
ho	44.2	bo	0.0	hu	0.0
pe	46.6	pa	0.0	du	0.0
du	48.9	hu	0.0	re	21.1
te	49.9	zi	0.0	xya	25.0
pu	53.4	ha	0.0	ni	27.3
hu	53.7	no	0.0	ne	28.6
he	54.5	ze	0.0	pe	28.6
di	54.6	ni	0.0	di	37.5
so	54.6	ro	0.0	yu	40.0
:	:	:	:	:	:
	66.0		22.7		63.1

CNN-TransformerEncoder

k=33		P0		P1	
du	55.3	nu	0.0	nu	0.0
di	59.7	da	0.0	yo	40.0
chi	63.9	di	0.0	du	40.0
nu	65.2	ze	0.0	yu	50.0
pe	67.1	ke	0.0	re	53.3
ho	68.0	ta	11.8	di	54.5
pi	71.4	pi	13.3	ra	54.5
tsu	71.7	ko	15.4	chi	54.5
yu	71.9	ha	17.4	ta	57.1
na	72.6	du	17.4	xyo	60.0
:	:	:	:	:	:
	66.0		22.7		63.1

2CNN-TransformerEncoder

k=33		P0		P1	
du	53.9	na	0.0	di	33.3
di	58.6	so	0.0	nu	50.0
chi	63.8	ke	0.0	re	52.6
pe	64.3	di	0.0	ma	53.8
ho	68.6	pu	14.3	ho	54.5
bo	70.0	ho	16.7	te	54.5
so	70.7	ko	20.0	chi	54.5
he	70.8	a	22.2	xtsu	59.5
tsu	70.9	ze	23.5	ra	60.0
nu	71.8	chi	24.4	du	62.5
:	:	:	:	:	:
	77.1		41.7		75.6

A notable improvement was observed
in the recognition of challenging characters

Contents

Second Comparative Analysis Part1

Dataset is same. (described later)

Compare Model

6 Combinations: 2(1, 2) × 3(A, B, C)

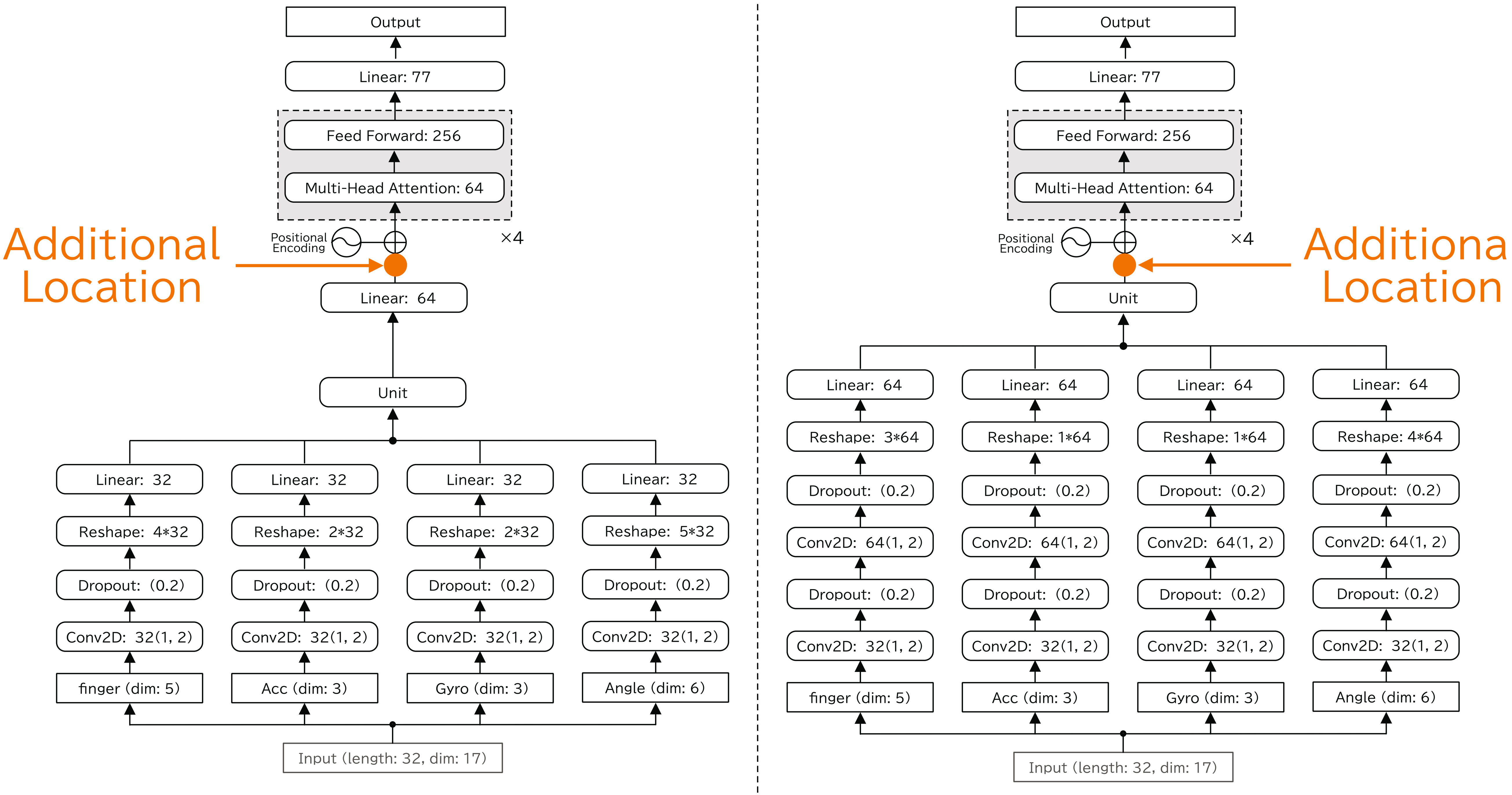
1. branch-CNN-unit-Transformer Encoder (batchsize: 32)
2. branch-2CNN-unit-Transformer Encoder (batchsize: 32)

A. Removing P0’s data
B. Add BatchNorm-1D and ReLU
C. Both A and B

※ Others are same as in First Comparative Analysis

Second Comparative Analysis Part2

Add BatchNorm-1D and ReLU



(a) (b)

Second Comparative Analysis Results Part1

k-fold CV (CrossValidation) Evaluation

- 10-fold CV
  - Input data was shuffled
  - Input data was divided into Training and Test data
  - Average and Standard Deviation of the 5 and 10 runs
  - Values at the epoch when the Validation Loss was minimized

model	k=10, F-measure [%]	
	micro	macro
branch-CNN-unit-Transformer Encoder	93.6 (0.2)	76.5 (0.8)
(Removing P0's data)	93.9 (0.2)	77.6 (0.7)
(Add BatchNorm-1D and ReLU)	93.8 (0.2)	77.4 (1.0)
(Removing P0's data & Add BatchNorm-1D and ReLU)	**94.1 (0.3)**	**78.4 (1.0)**
branch-2CNN-unit-Transformer Encoder	93.6 (0.3)	76.6 (1.0)
(Removing P0's data)	93.9 (0.2)	77.3 (1.0)
(Add BatchNorm-1D and ReLU)	93.5 (0.3)	76.1 (1.4)
(Removing P0's data & Add BatchNorm-1D and ReLU)	93.9 (0.2)	77.7 (0.6)

Contents

Discussion Part1

Transformer Encoder could:

1. Improving **Recognition rates**?
   —> "Yes"

2. Handling **individual differences among signers expressing**?
   —> Cannot make strong claims that "Yes"
   - The results for P0 raise two potential explanations:
     1) Represents an "extreme individual difference"
     2) Contains "noise" that transcends typical individual variations

Future data collection efforts should include?:

Comprehensive participant attribute information
(Experience with Japanese Sign Language (JSL) or Signed Japanese, etc.)

Discussion Part2

Limitation

1. Not yet explored
   - Full parameter space for CNN combinations
   - Relationship with preprocessing parameters
     - moving averages calculations
2. Not designed for real-time processing
3. Cannot evaluate sentence-level
   - Including also NM expressions and grammatical omission
     - Not include sentence-level data
4. Only utilizes the Transformer Encoder
   - Not conducted comparative analyses
     - Involving Transformer Decoder or CTC
     - Conformer
     - Spatial Temporal Graph Convolutional Networks
     - Signgformer

Contents

Conclusion and Future Work

Conclusion

Transformer Encoder could improving Recognition rates:

For 76 Japanese fingerspelling characters

- Average micro F-measures: 93.8% (0.2)
- Average macro F-measures: 77.4% (1.0)

Future Works

1. Comparative analyses
   - Involving Transformer Decoder or CTC
   - Conformer
   - Spatial Temporal Graph Convolutional Networks
   - Singformer
2. Comparative analysis with one-handed sign language
   - examine the spatial characteristics differentiating
     - fingerspelling from sign language