法政大学
HOSEI University

Introduction of our research activities on machine learning

Intelligent Information Processing Lab. (IIPL)

Hitoshi Iyatomi

Applied Informatics, Hosei University

iyatomi@hosei.ac.jp

Recent topics @ IIPL

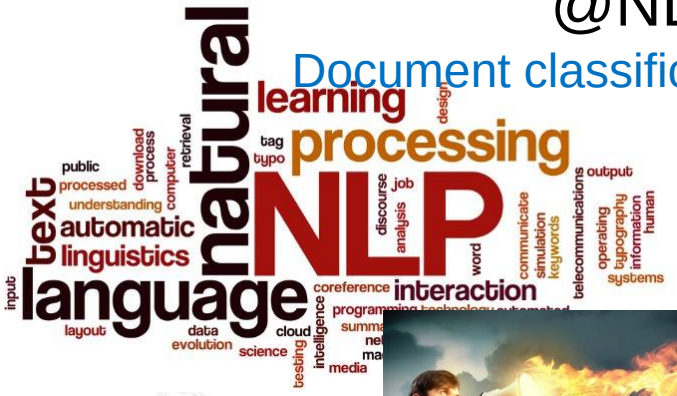
@recognition

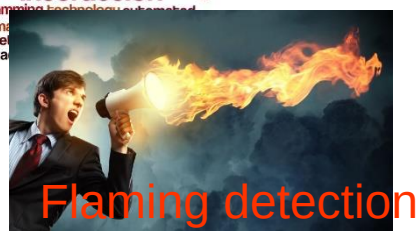
Movie analysis

basic

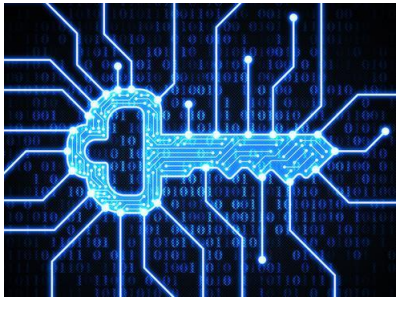
@NLP

Document classification

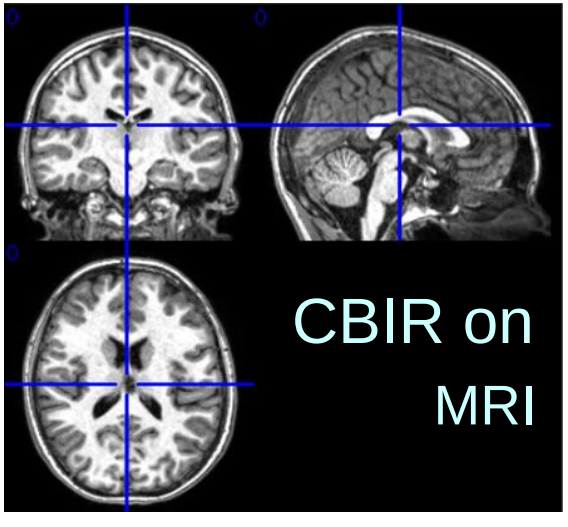




Flaming detection

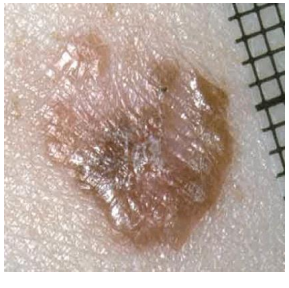


@security



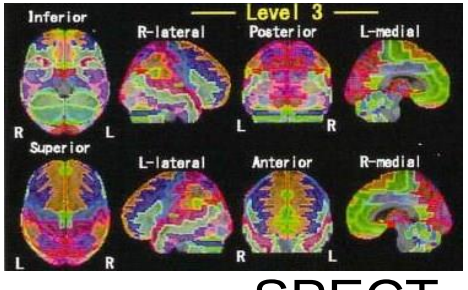
@diagnosis

plant

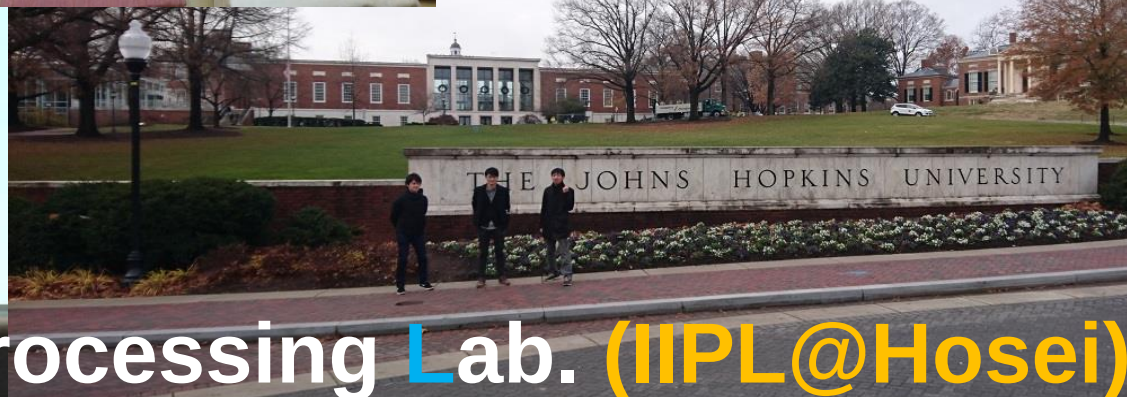


cancer

@medical app



SPECT



Intelligent Information Processing Lab. (IIPL@Hosei)

Paradigm change of image recognition / analysis

Before deep learning

Input image

neural net, SVM… etc.

features

Feature extraction → classification → results

- Handmade features
- Troublesome pre-processing
  （e.g. segmentation）

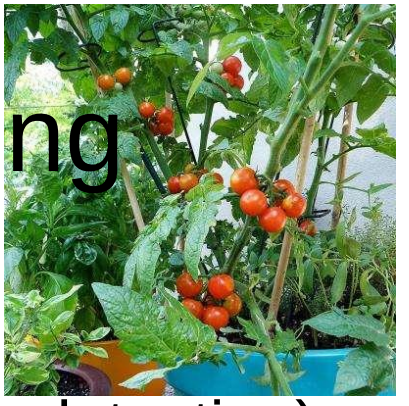
Deep learning

Automatically extract essential features

Feature extraction + classification → results

+ fully automatic
- Huge training data

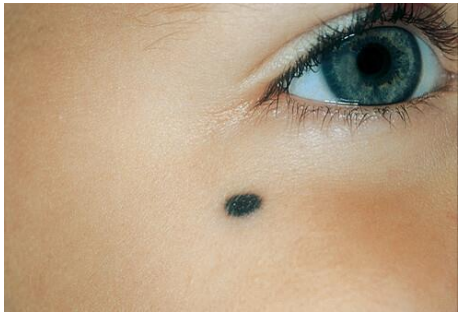
Very brief introduction what we do with deep learning



1. Automated plant diagnosis

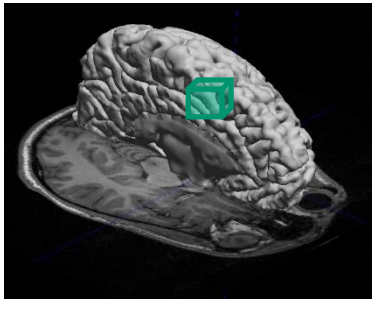
2. Security applications (web application firewall, malware detection)

3. Medical application
   - Melanoma diagnosis
   - Significant dimension reduction
     for content-based image retrieval (CBIR)



4. Natural language processing (NLP) research.

twitter



at IIPL@Hosei

Automated plant disease diagnosis

農林水産省
Ministry of Agriculture, Forestry and Fisheries

Background

Monetary loss by plant disease

est. **220B** USD [Agrious., 2005]

Early detection and diagnosis are essential

- visual inspection by expert
- geometrical exam if needed

time consuming, expensive, limited availability

Automated plant diagnosis system is desired.

農研機構 NARO 国立研究開発法人 農業・食品産業技術総合研究機構

HOSEI University

農林水産省 Ministry of Agriculture, Forestry and Fisheries

NTT DATA

with 24 prefectures

5 years national project

Basic investigation on a robust and practical Plant diagnosis system

*[1]Erika Fujita, *[1]Yusuke Kawasaki, *[2]Hiroyuki Uga, *[1]Satoshi Kagiwada, *[1]Hitoshi Iyatomi

*[1]Hosei University, Japan, *[2]Saitama Agricultural Technology Research Center, Japan

Goal of first stage:

Easy, fast, accurate (and low cost) automated plant diagnosis system

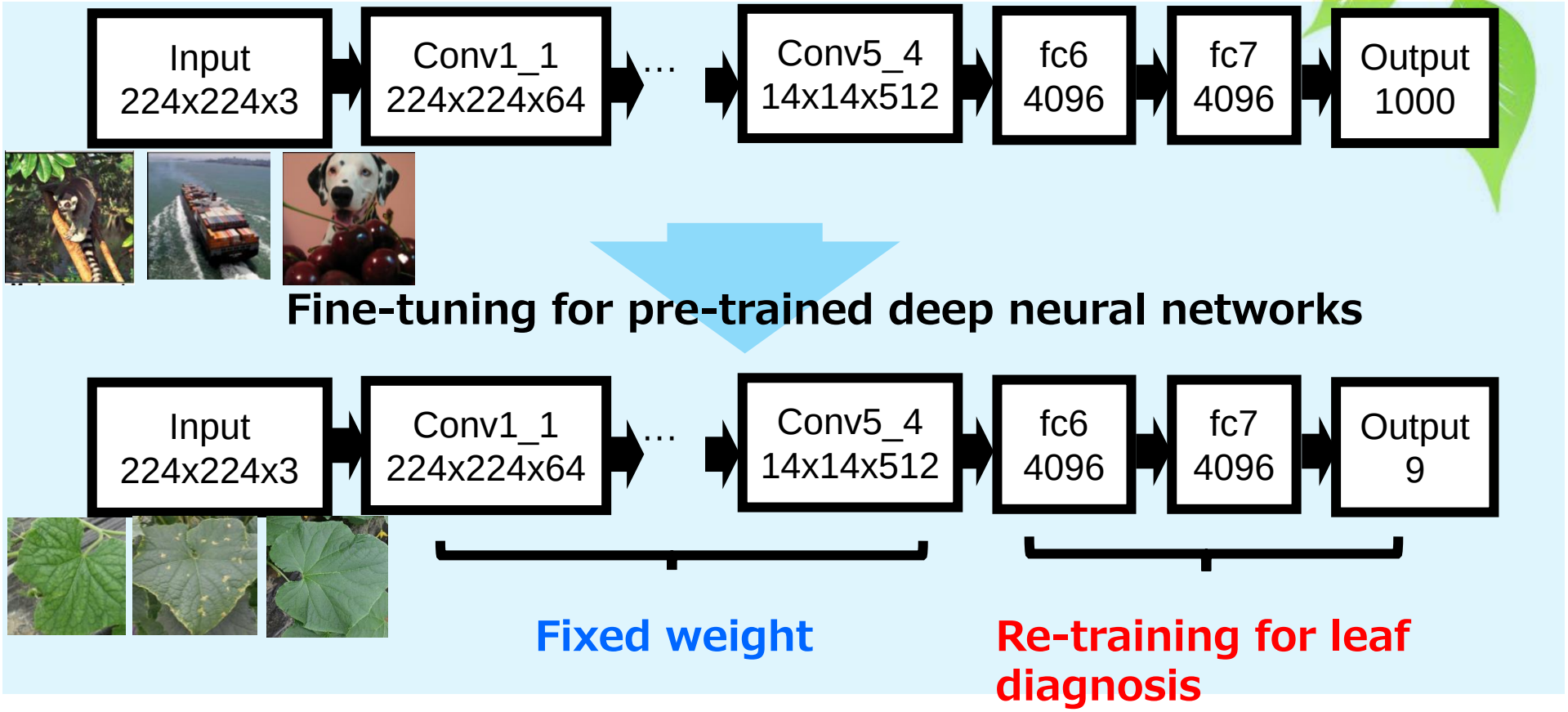
Current target (ex: cucumber)

Fig. 1. Terminal symptom of each diseases

Fig. 2. Initial symptom of each diseases

Effects of light, shadow, white-balance etc…

Classification – Fine-tuned CNNs

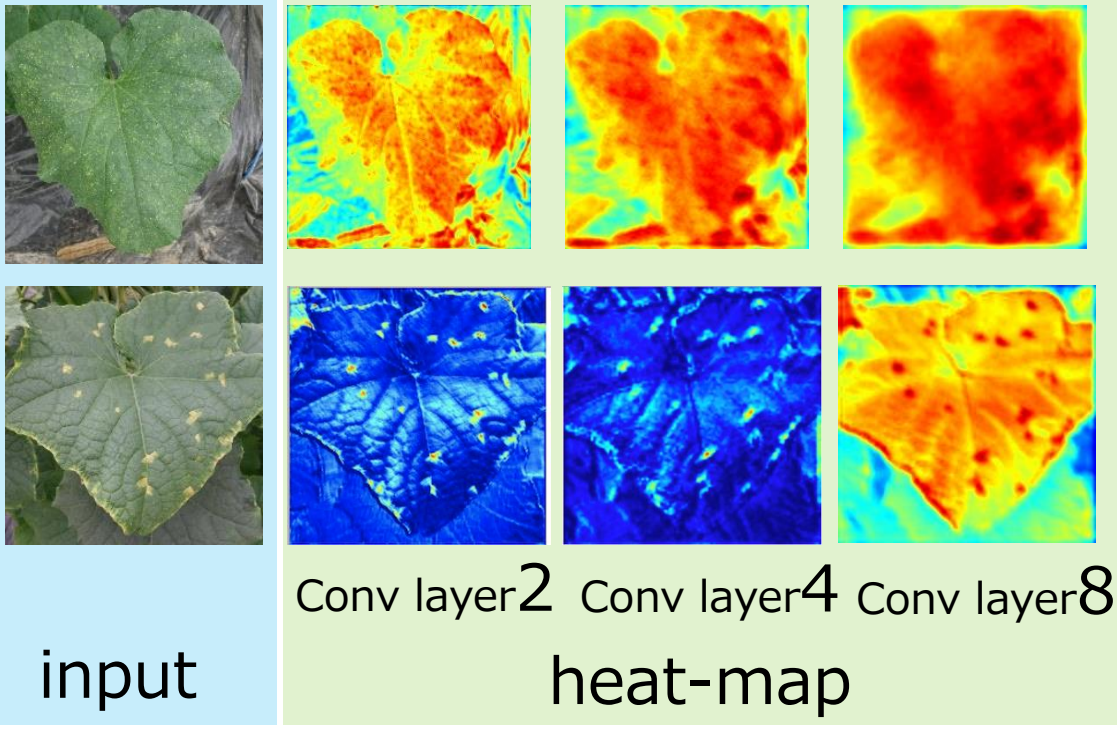


8 diseases＋healthy classification

Trained (1000x9) x 72 images　　10 fold cross-validation

93.6% mean accuracy　(min 90.8% - max 99.6%, SP=94.5%)

Visualization of “clue” for diagnosis with GradCAM*

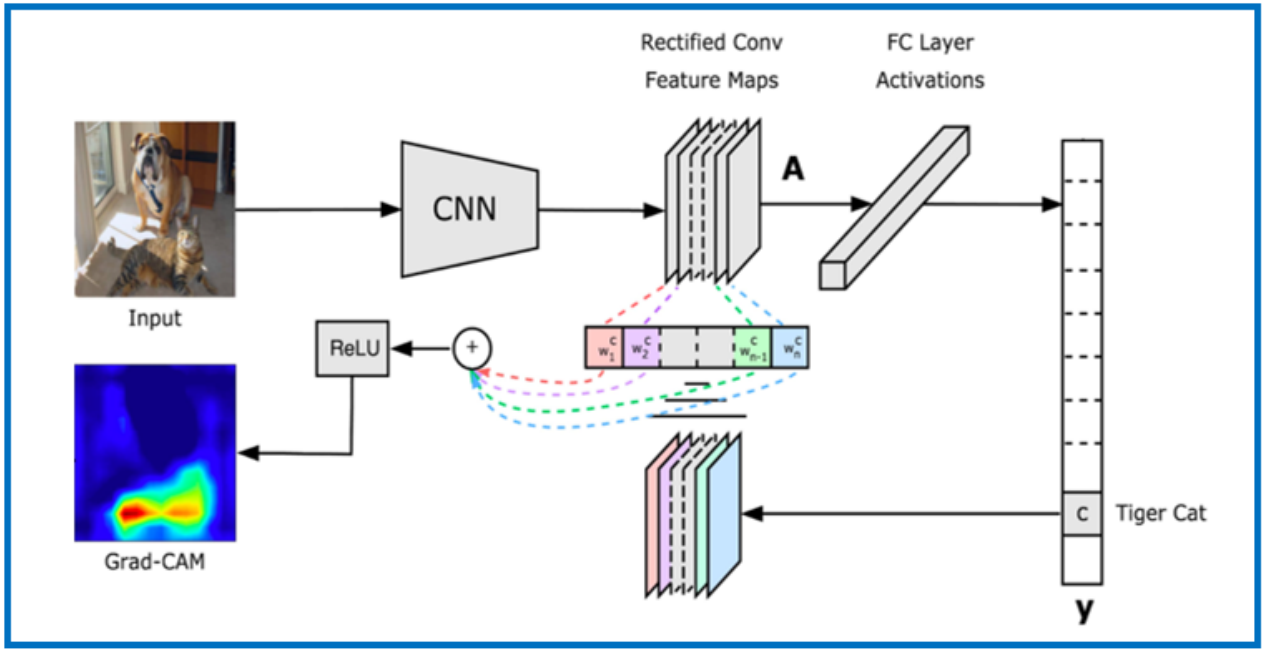


Heat-map representation

Visualize the areas
with large impact for the output

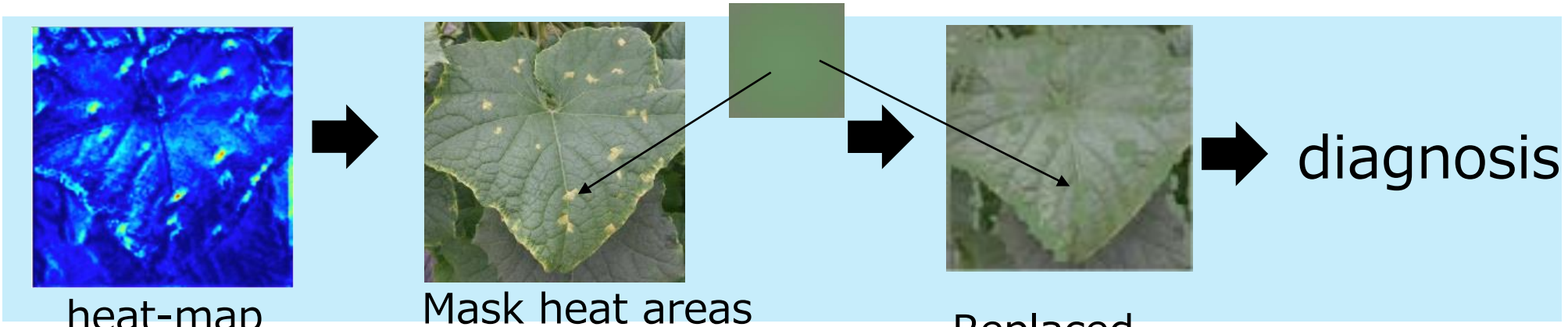
Visually reasonable.

How to evaluate it
quantitatively?



GradCAM [Selvaraju+2017]

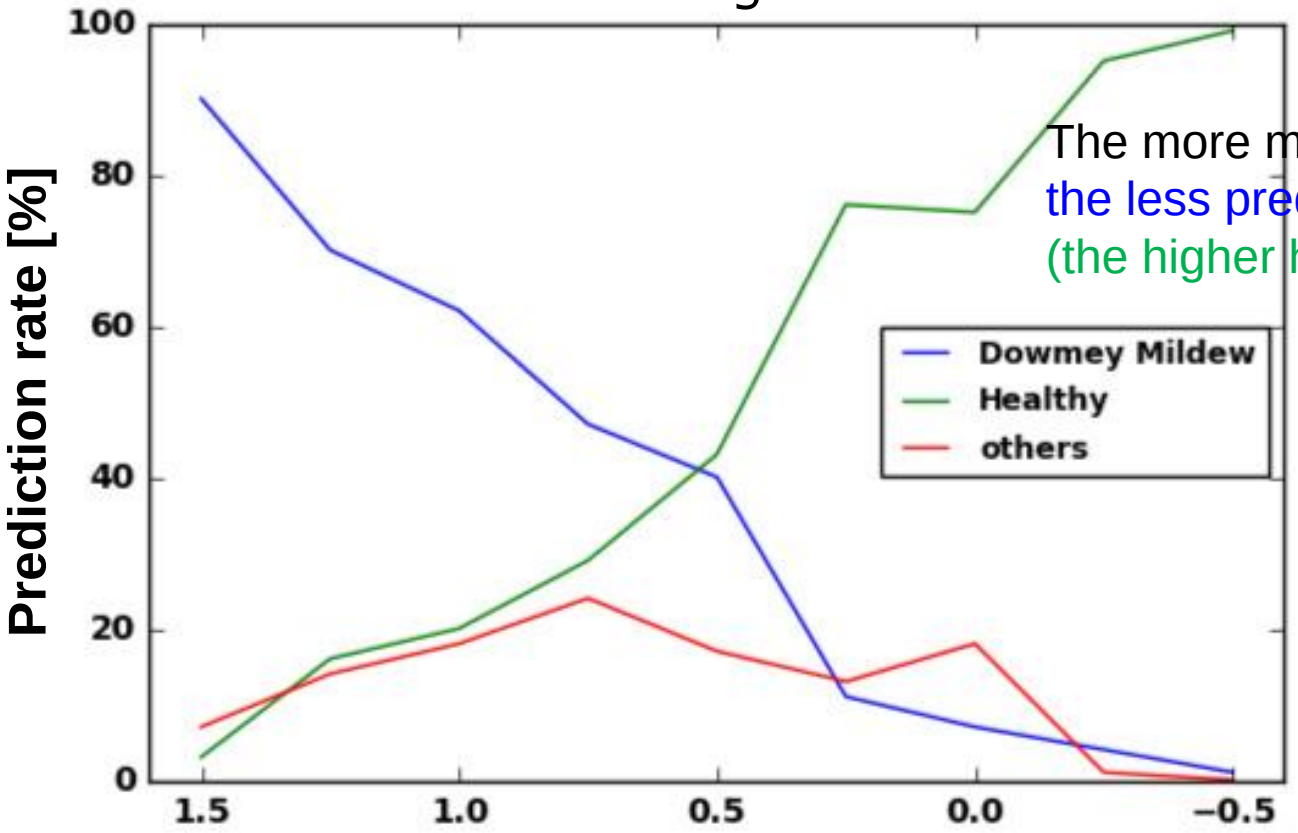
Evaluation of “clue” for diagnosis



heat-map

Mask heat areas
with average leaf color

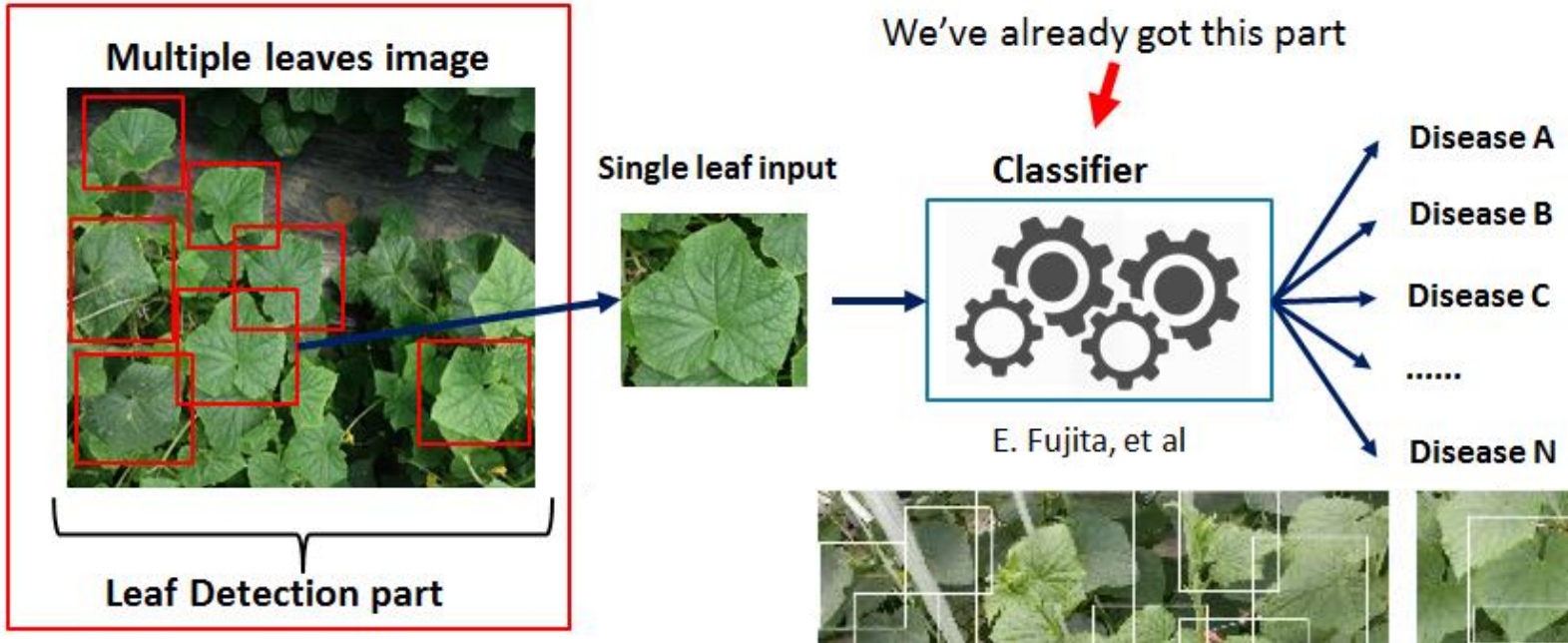
Replaced



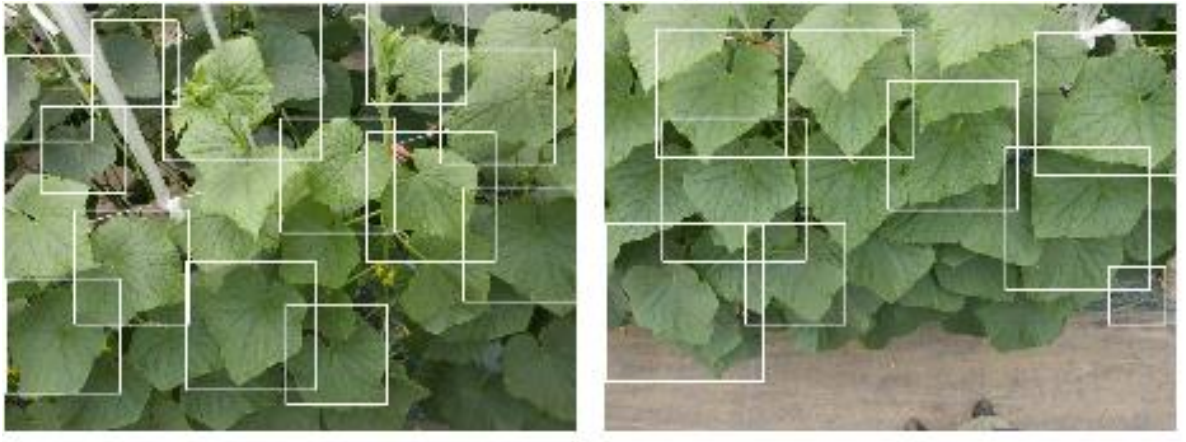
The more mask processing performed,
the less prediction rate for Dowmey mildew.
(the higher healthy prediction rate is observed.)

Confirmed the validity of observed heat-map.

One-shot diagnosis from wide-angled image



trained 1.44M image patch



Fast and accurate
leaf detection with
deep learning approach

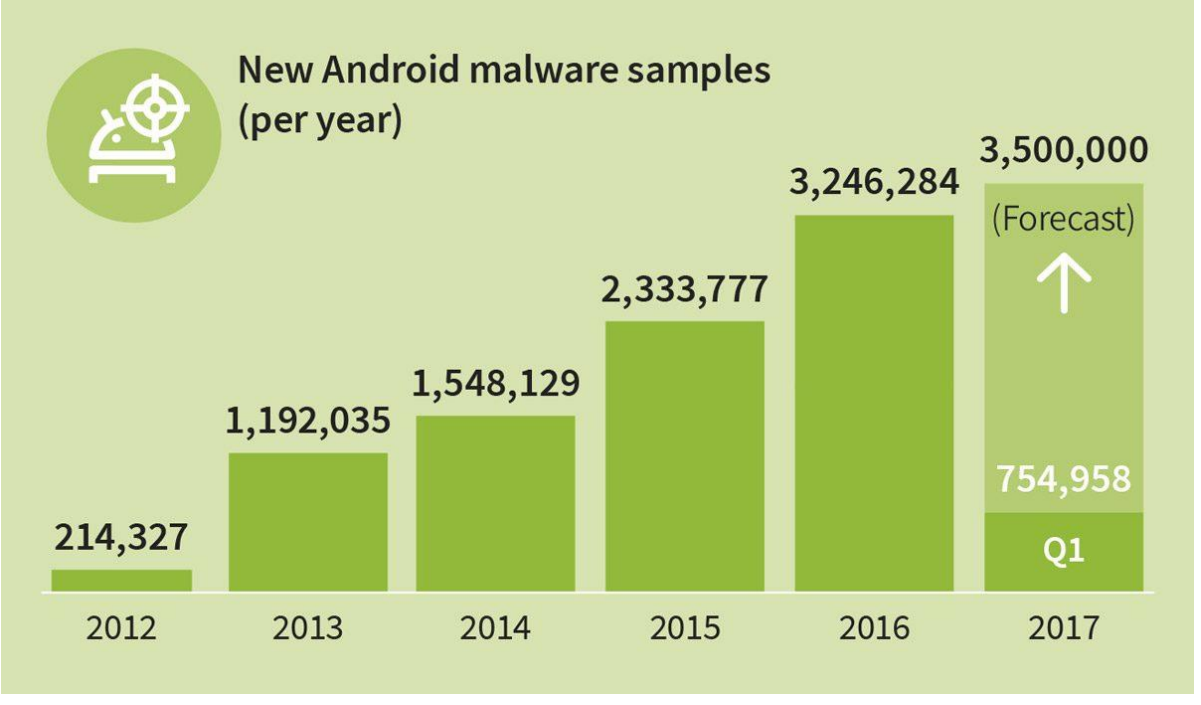
78.0% in F-measure
@ 2.0 fps

Second and third stages:

Automated diagnosis with stationary camera device etc.

Detection of ROI (each leaf etc.) and diagnose it with CNN

Security + Machine learning

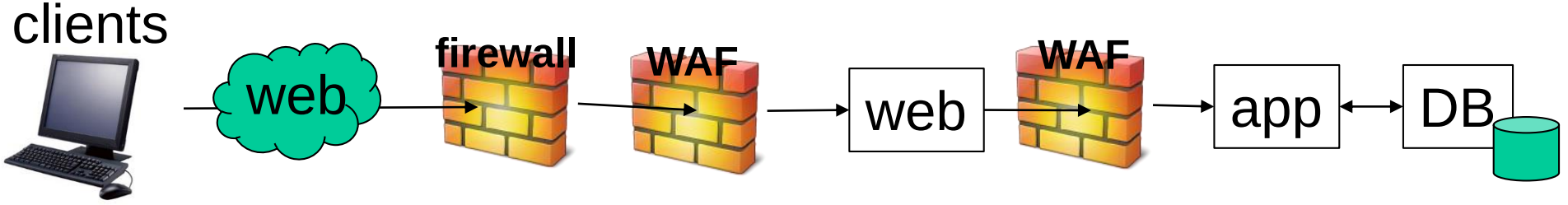


will be presented IEEE CSPA 2018 (Malaysia in 2 weeks)

Best Paper Award at IEEE CSPA2018

Deep learning-based Web Application Firewall (WAF)

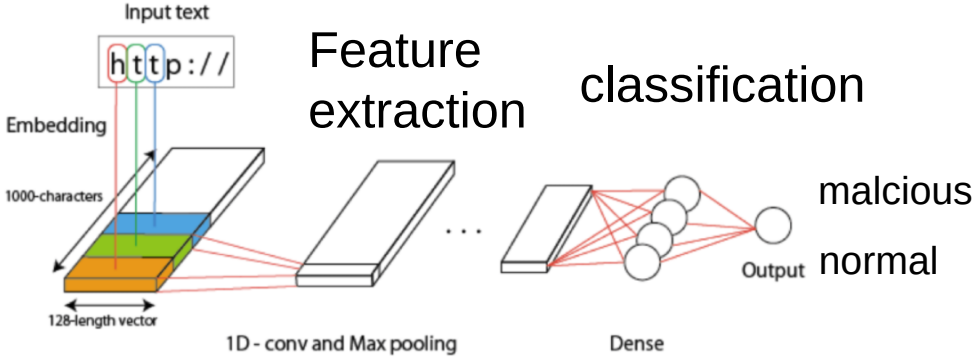
Conventional firewall is difficult to block new or subspecies attacks.



motivation: Low-cost and accurate WAF to prevent malicious connection

```
GET http://localhost:8080/tienda1/publico/anadir.jsp?id=2&nombre=Jam%F3
n+Ib%E9rico&precio=85&cantidad=%27%3B+DROP+TABLE+usuarios%3B+
SELECT+*+FROM+datos+WHERE+nombre+LIKE+%27%25&B1=A%F1adir+al+carrito
User-Agent: Mozilla/5.0 (compatible; Konqueror/3.5; Linux) KHTML/3.5.8 (like Gecko)
Pragma: no-cache
Cache-control: no-cache
```

Question: How to determine and implement for “attack” ?



Character-level CNN

HTTP DATASET CSIC 2010 Dataset (normal 36K + attack 25K = 61K)

Detection accuracy: 98.8% @ 2.35ms 10-fold cross validation

Beat current state-of-the-art : 82.0% with Naiive Bayes

Deep learning-based Android Malware detection

New application is released one after another everyday.

Detection of malware is essential.

Question:
What is the clue of malware?
How to define and implement them?

→ **Character-level CNN bring solutions for above questions.**

Our method investigates only 1024 bytes of Android APK file.
attained average 93-96% accuracy.

10-fold cross validation

- AMD dataset（malware 5,000)
- Drebin dataset (malware 5,000)
- normal applications passed 64 security checks (5,000)

- State-of-the-art performance
- Much faster than conventional

Automated skin cancer diagnosis

Automated melanoma diagnosis

Malignant melanoma -- **The worst skin cancer**

× difficult to cure (metastasize easily)

× high mortality rate

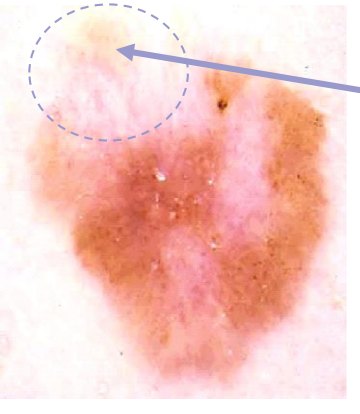
Early stage melanoma ○ 5-year survival rate : 93%

Early detection and resection is highly required

But diagnosis by naked eye is extremely difficult

tumor image

dermoscopy



dermoscopy image

Diagnosis with dermoscopy × Still difficult and subjective

(diagnostic accuracy: 75-80% by expert dermatologists)

Internet-based melanoma diagnostic system

- automated diagnosis **24hours-365days**
  → **Find early stage patient**, screening
- clinical database
  → Data standardization

http://dermoscopy.k.hosei.ac.jp

Dermoscopy image

Diagnostic results
（malignancy score: 0-100）

Extraction of tumor area

Calculation of parameters

Select adequate classifier

Classifier for non-acral lesion

Classifier for acral lesion

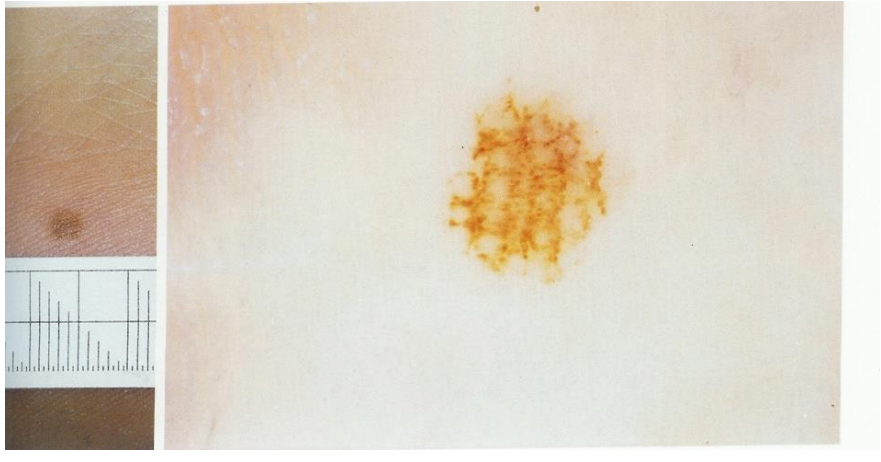
Registration to database

Asian-specific melanomas in acral area

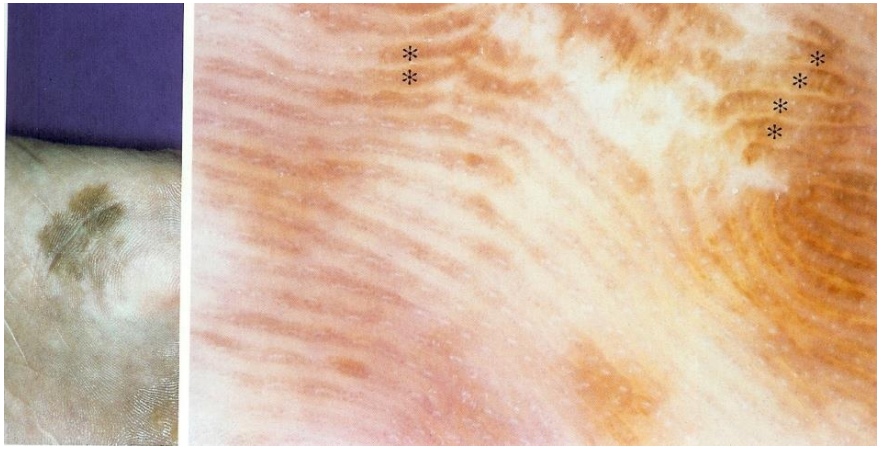
Acral (palm and sole) lesions have completely different appearance.

Half of Asian melanomas are from acral area.

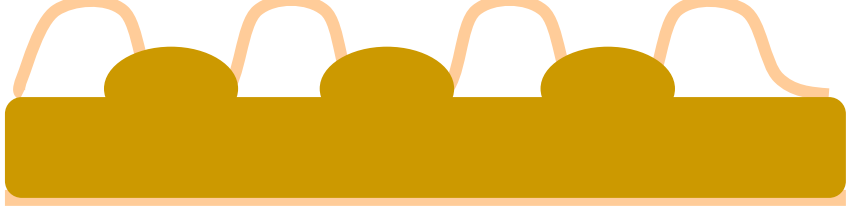
But no research for automated analysis have been made



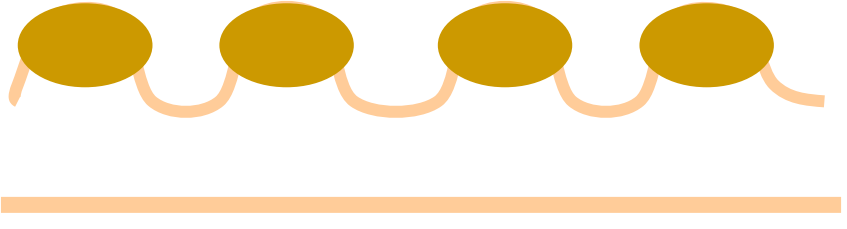
nevus (benign)



melanoma



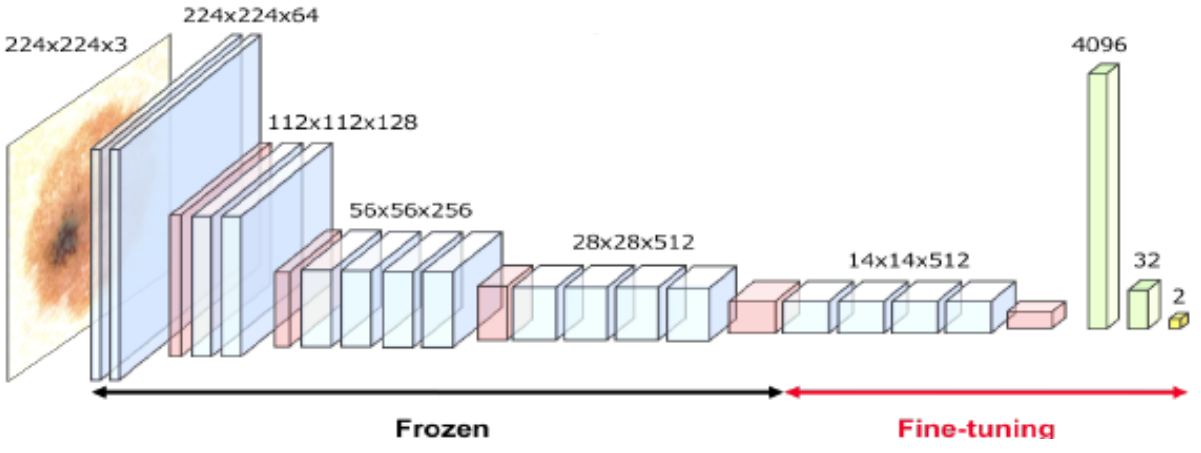
Furrow areas are selectively pigmented.

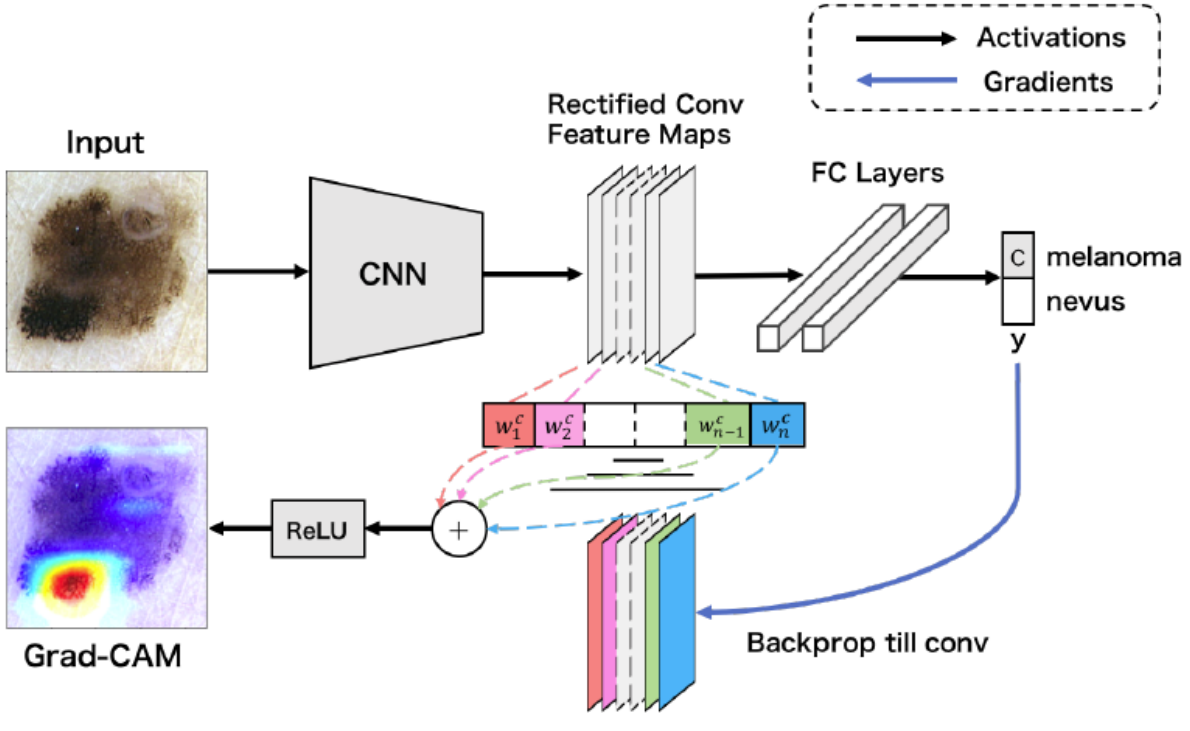


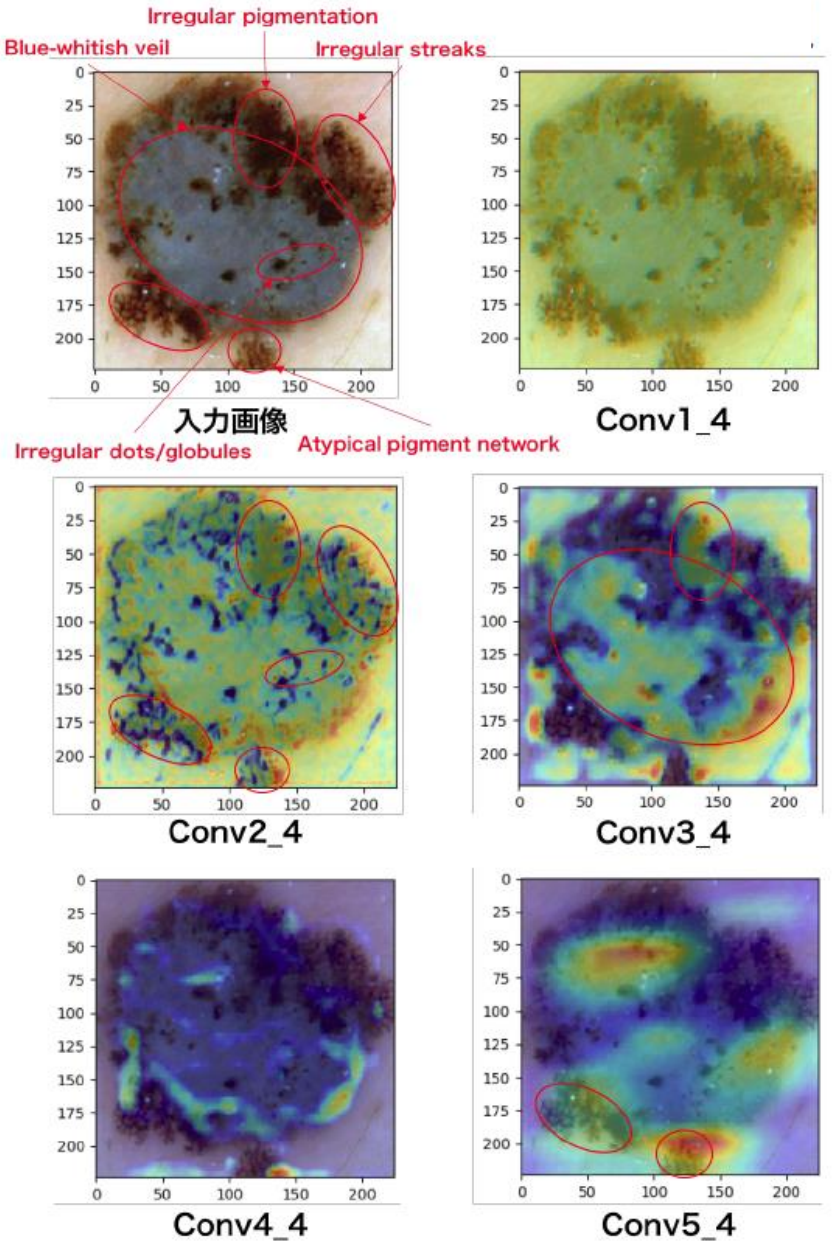
Ridge areas are selectively pigmented.

Automated melanoma diagnosis with transfer-learning

& visualization of clue of diagnosis







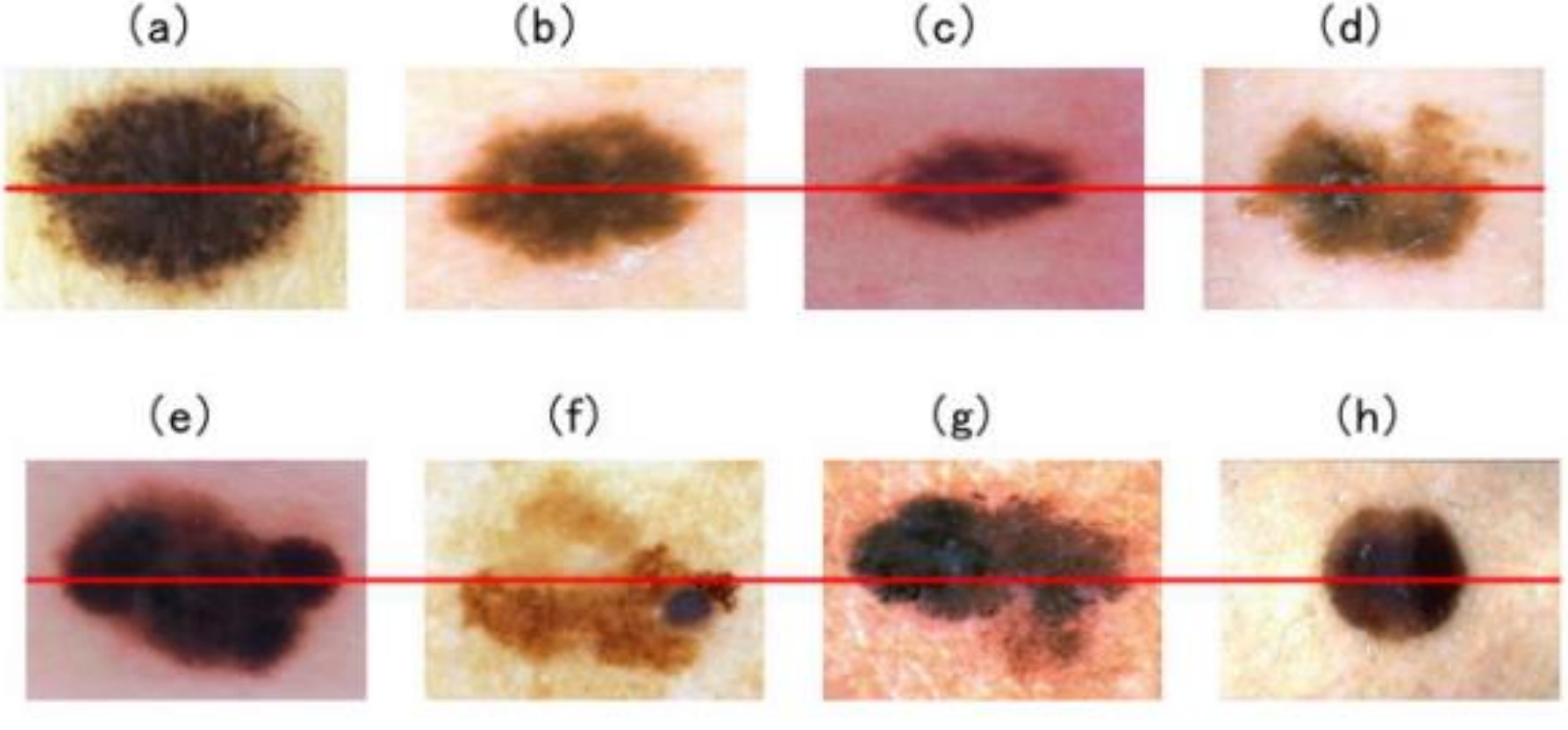
We attained diagnostic accuracy of around 90% and visualize “ground for diagnosis.”

7-point checklist

Major Criteria
1. Atypical pigment network(不規則な網構造)
2. Blue-whitish veil(青白い領域)
3. Atypical vascular pattern(不規則な血管パターン)
Minor Criteria
4. Irregular streaks(不規則な枝状構造)
5. Irregular pigmentation(不規則な色素沈着)
6. Irregular dots/globules(不均一な点，粒状構造)
7. Regression structures(色素抜け構造)

We focused on the biological characteristics of pigmented skin lesions

Quite simple approach – aligned major axis of tumor



This pre-process is effective more than x10 data augmentation.

3D Content-based image retrieval (CBIR)

using 3D convolutional autoencoders

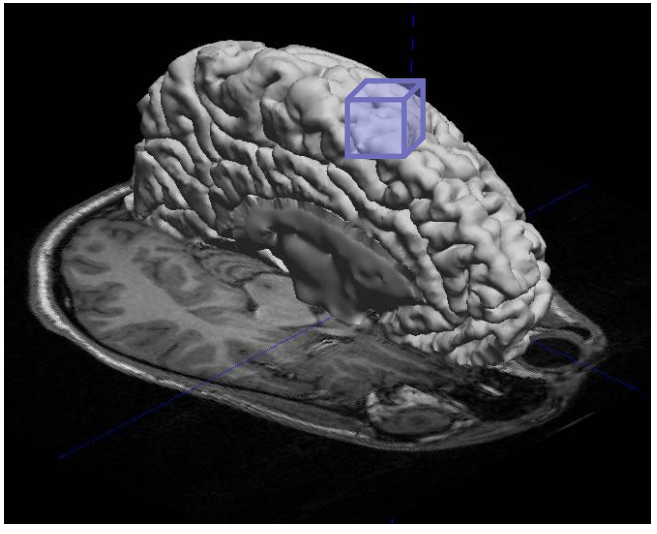
collaborate with

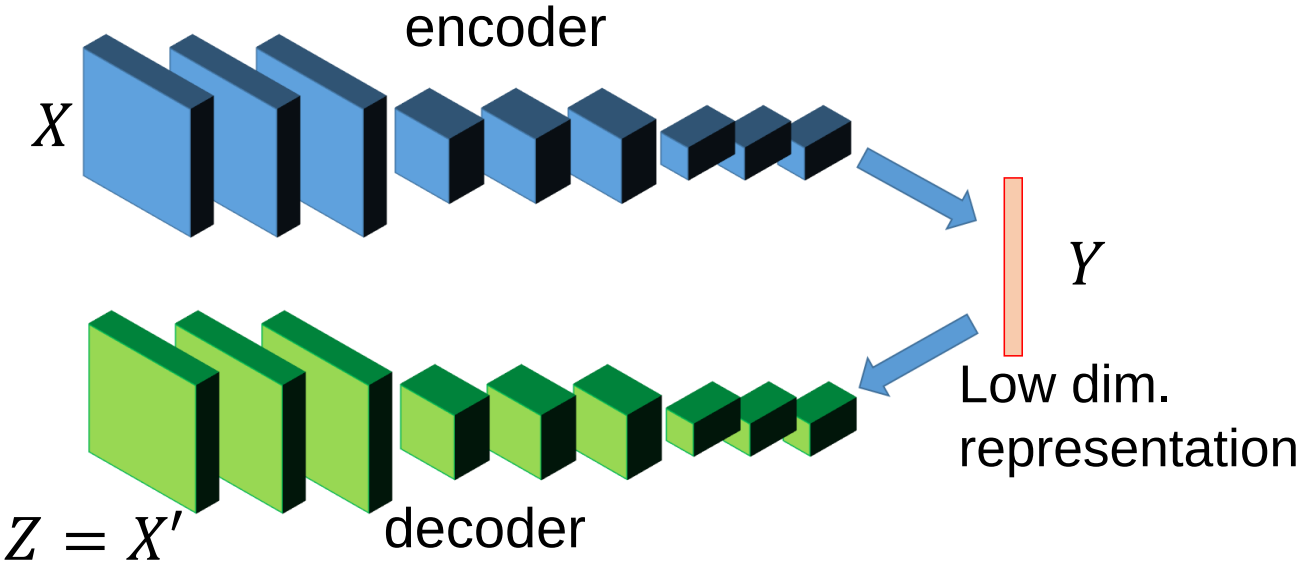
Significant dimension reduction for "Deep" Content-based image retrieval (CBIR)

- Retrieving past clinical cases are important for medical practice.
- Text-based search is currently used, but has a limitation
- CBIR technique is required, but not easy (ultra high dimension).

Dimensional reduction is essential for machine learning.

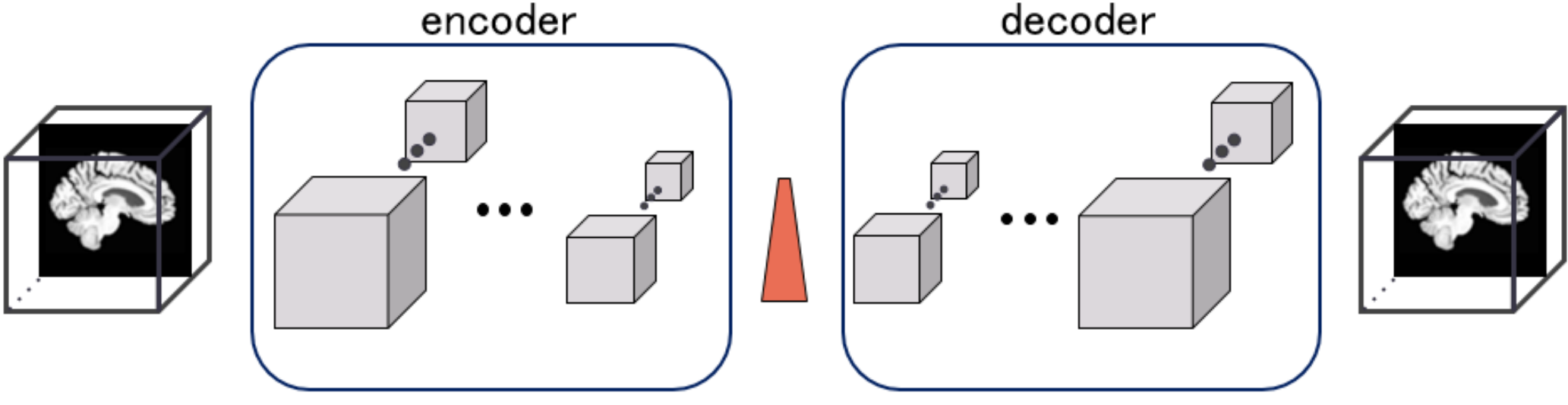
Dimension reduction with 3D convolutional autoencoders (3D-CAE)





$X \in \mathcal{R}^D \quad Z = X' \sim X \in \mathcal{R}^D \quad Y \in \mathcal{R}^d \quad D \gg d$

Develop original 3D convolutional auto encoders (3D-CAE)
with pseudo tied-weight



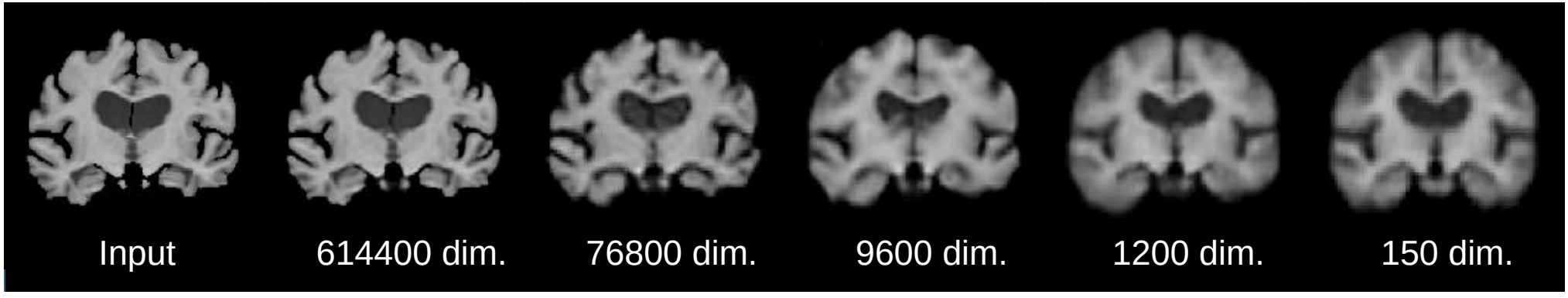
Input
$X \in \mathcal{R}^D$
160x192x160
≒5M

conv + pooling
5 layers

representation
$Y \in \mathcal{R}^d$
150, 1200, …

deconv + unpooling
5 layers

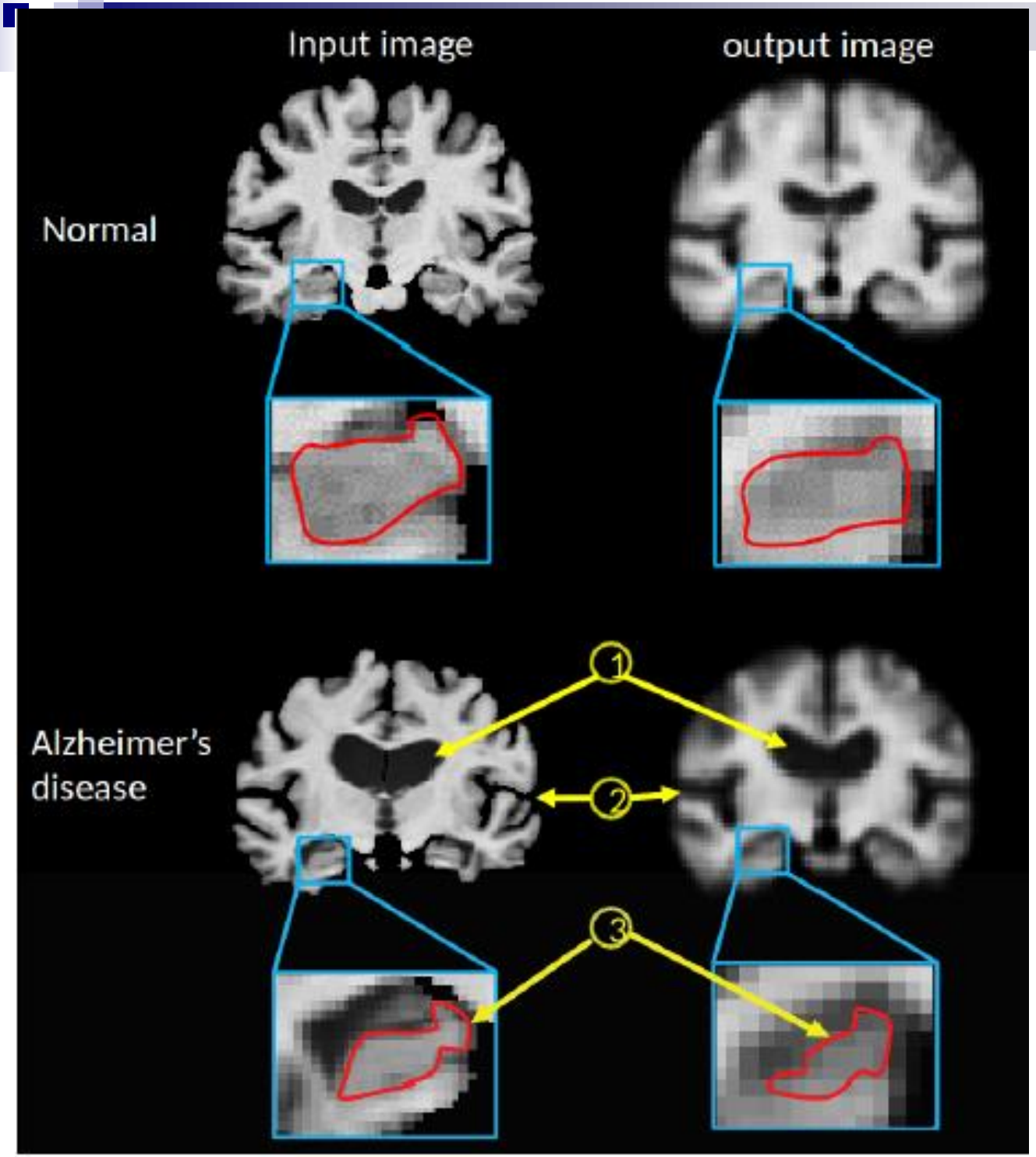
Output
$Z = X' \sim X \in \mathcal{R}^D$
160x192x160



Represent 5M dim 3D brain MRI image with only 150 dim.

Only with 150 dim,

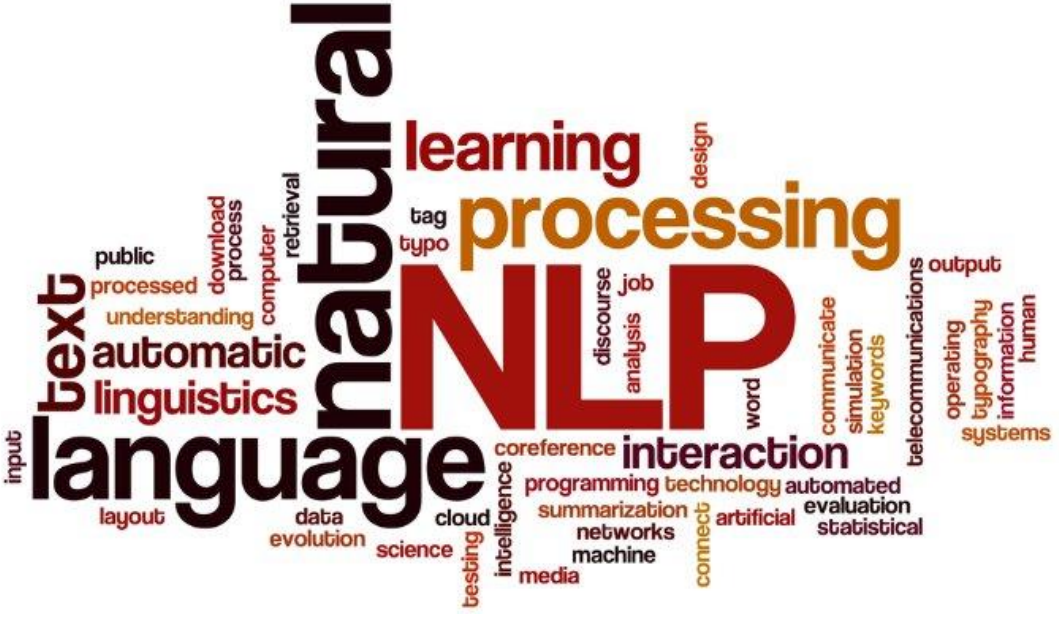
RMSE = 8.4% on not trained 3D MRI reconstruction



Biomarker areas for diagnosing Alzheimer's disease

Only 150 dim representation preserves involved areas

Deep learning for NLP application



Introduction

- Difficulty of processing Japanese / Chinese text
  - **No word boundary**
    - Word segmentation preprocess
    - Hard to segment words include coinages and slang words
  - **Large number of alphabets**
    - More than 2,000 alphabets for daily use (Japanese)

Common NLP methodologies (designed for English) cannot be applied.

SankeiBiz

リアルタイムの

官房長官、対北朝鮮「さらに厳しい対応とる」 ＮＳＣ開催

日本経済新聞 - 59 分前

北朝鮮による弾道ミサイルの発射を受け、日本政府は６日午前、首相官邸で国家安全保障会議（ＮＳＣ）を開いた。菅義偉官房長官は記者会見で、北朝鮮への制裁強化について「度重なる制裁が科されている中でも、全く無視する形で挑発行動を続けている。国連としてさらなる …

SankeiBiz

リアルタイムの

官房長官、対北朝鮮「さらに厳しい対応とる」 ＮＳＣ開催

日本経済新聞 - 59 分前

北朝鮮による弾道ミサイルの発射を受け、日本政府は６日午前、首相官邸で国家安全保障会議（ＮＳＣ）を開いた。菅義偉官房長官は記者会見で、北朝鮮への制裁強化について「度重なる制裁が科されている中でも、全く無視する形で挑発行動を続けている。国連としてさらなる …

Head news from Google news Japan

March 5. 2017

すももももももものうち

すもも　も　もも　も　もも　のうち

Sour peach and peaches are kinds of peach

Introduction

- Difficulty of processing Japanese / Chinese text
  - **No word boundary**
    - Word segmentation preprocess
    - Hard to segment words include coinages and slang words
  - **Large number of alphabets**
    - More than 2,000 alphabets for daily use (Japanese)

Common NLP methodologies (designed for English) cannot be applied.

Kanji

1st grade

Hiragana

Katakana

2nd grade

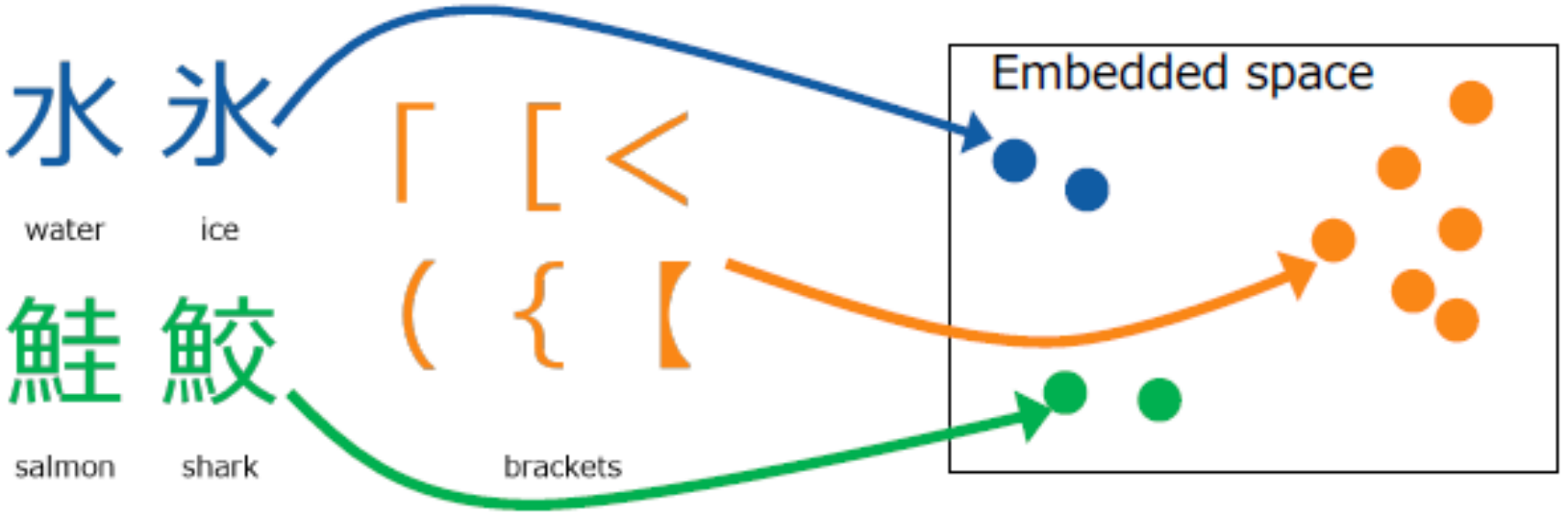
All Japanese learn 1,000 Kanji characters in their elementary school.

Large variety of “alphabet” causes over-fitting of the model.

Introduction

- Two New Document Classification Techniques for CLCNN

i. Image-based Character Embedding



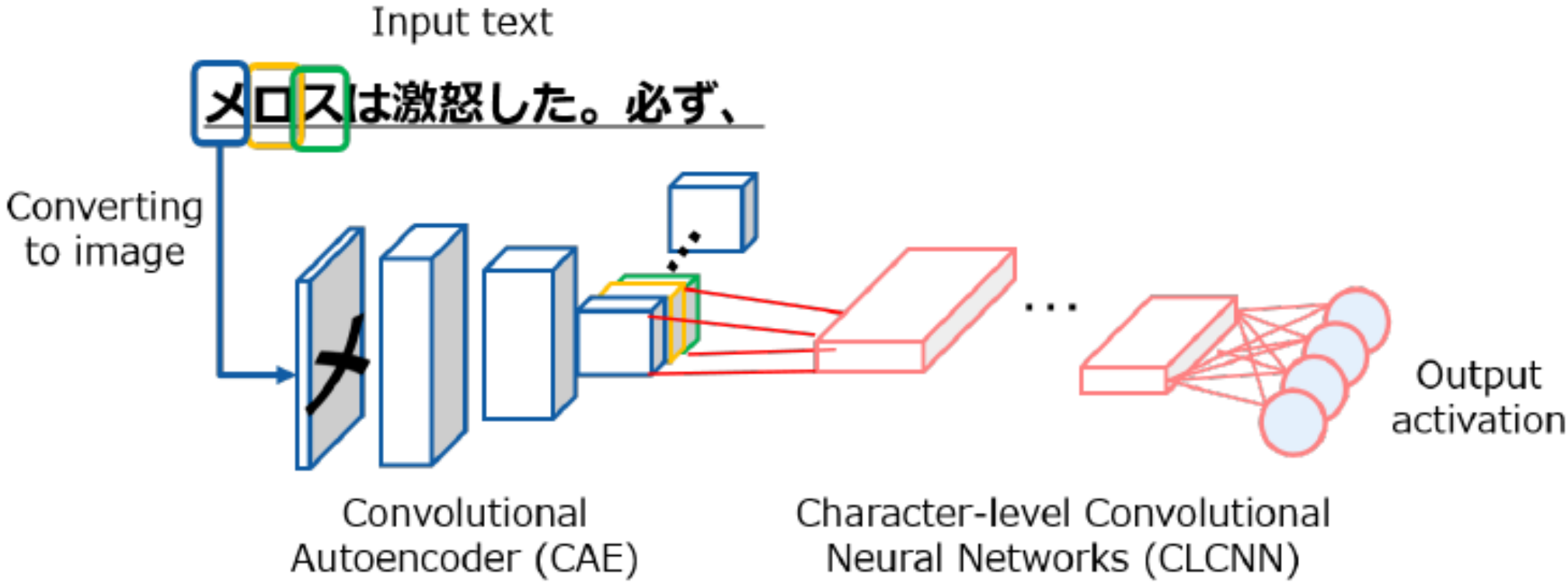
ii. Data augmentation without word segmentation, "wildcard training"

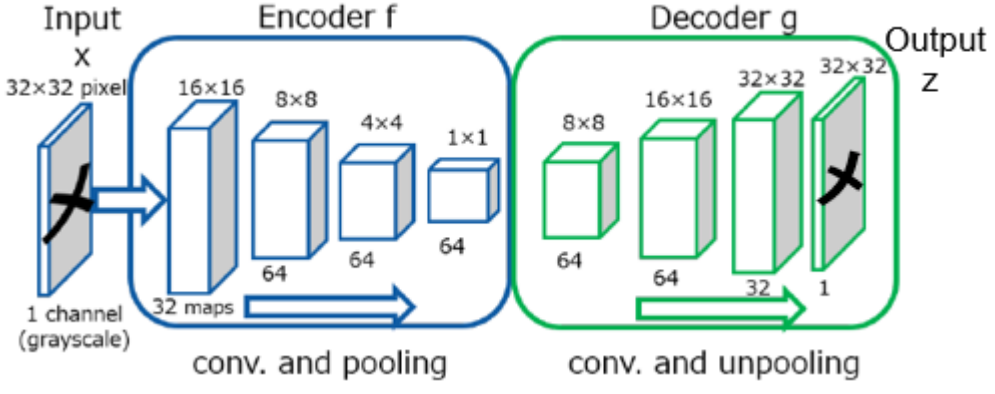
メロスは激怒した。 ⟶ メロス＊激＊した。

The Proposed Method

a. Image-based Character Embedding (CAE)

b. Character-level Classifier with Wildcard Training (CLCNN)





Original image

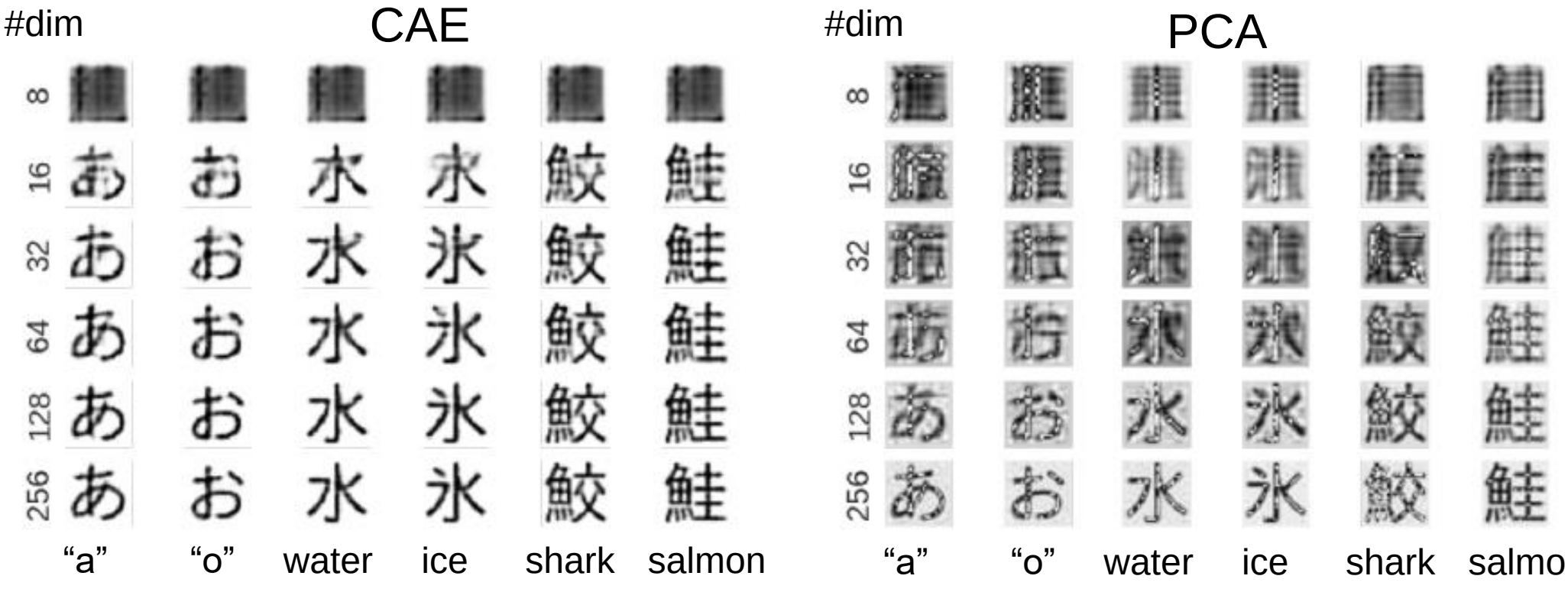
Encoded images

by CAE

by PCA

Each character (32x32 dim) can be represented by 64 dim vector with convolutional auto encoder (CAE)

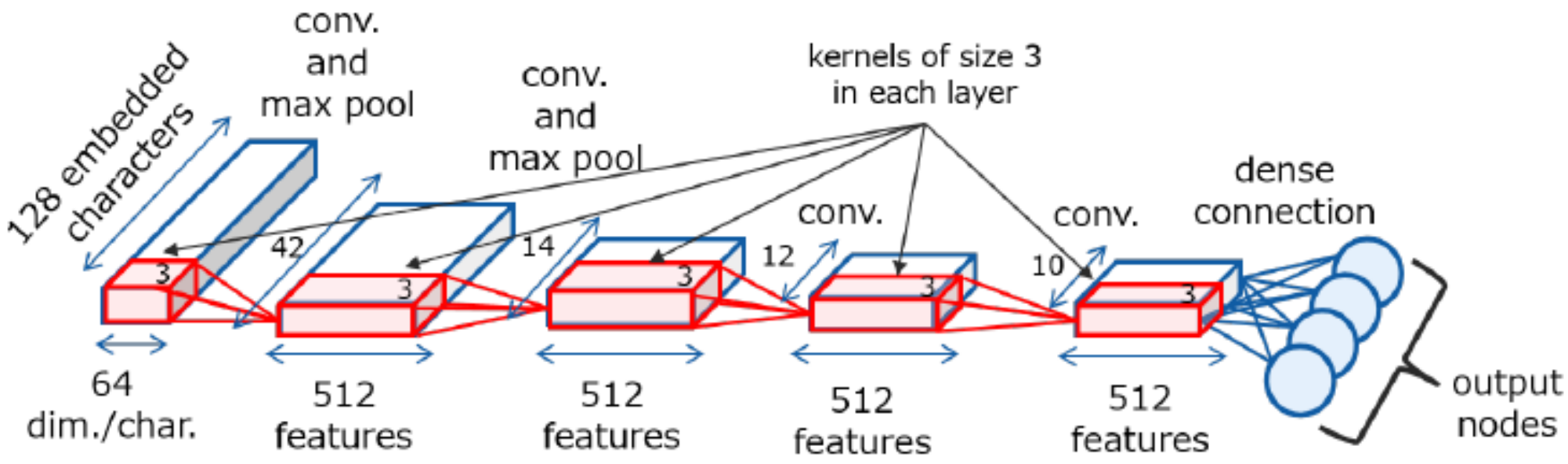
32x32 -> 64 dim

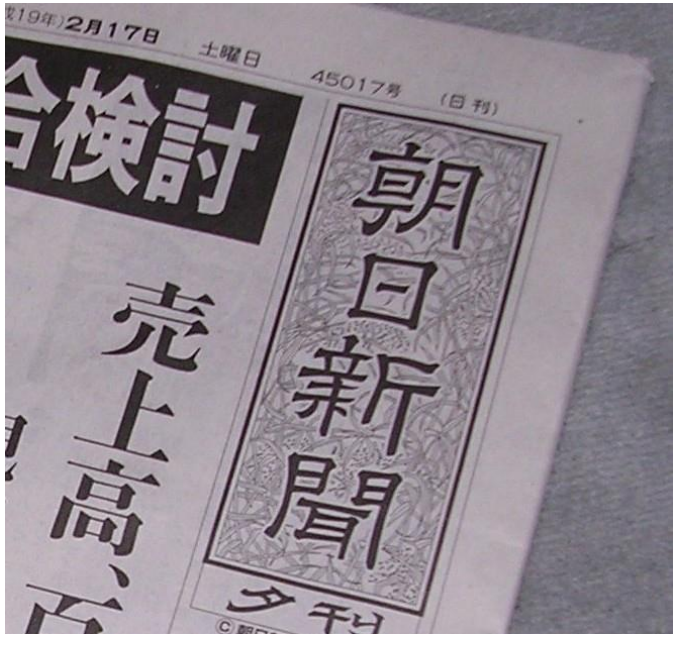


Comparison of image-based character embedding

b. Character-level Convolutional Neural Networks (CLCNN)

- CLCNN performs hierarchical feature extraction and classification.
- It takes image-based embedded characters as input.
- It's trained with ***wildcard training (WT)***, dropping some characters randomly.
  - Wildcard training augments the combinations of characters.





Publisher estimation of
Japanese major four newspapers
from single article

Experiments and Results

(2) Publisher Estimation from Japanese Newspaper Articles

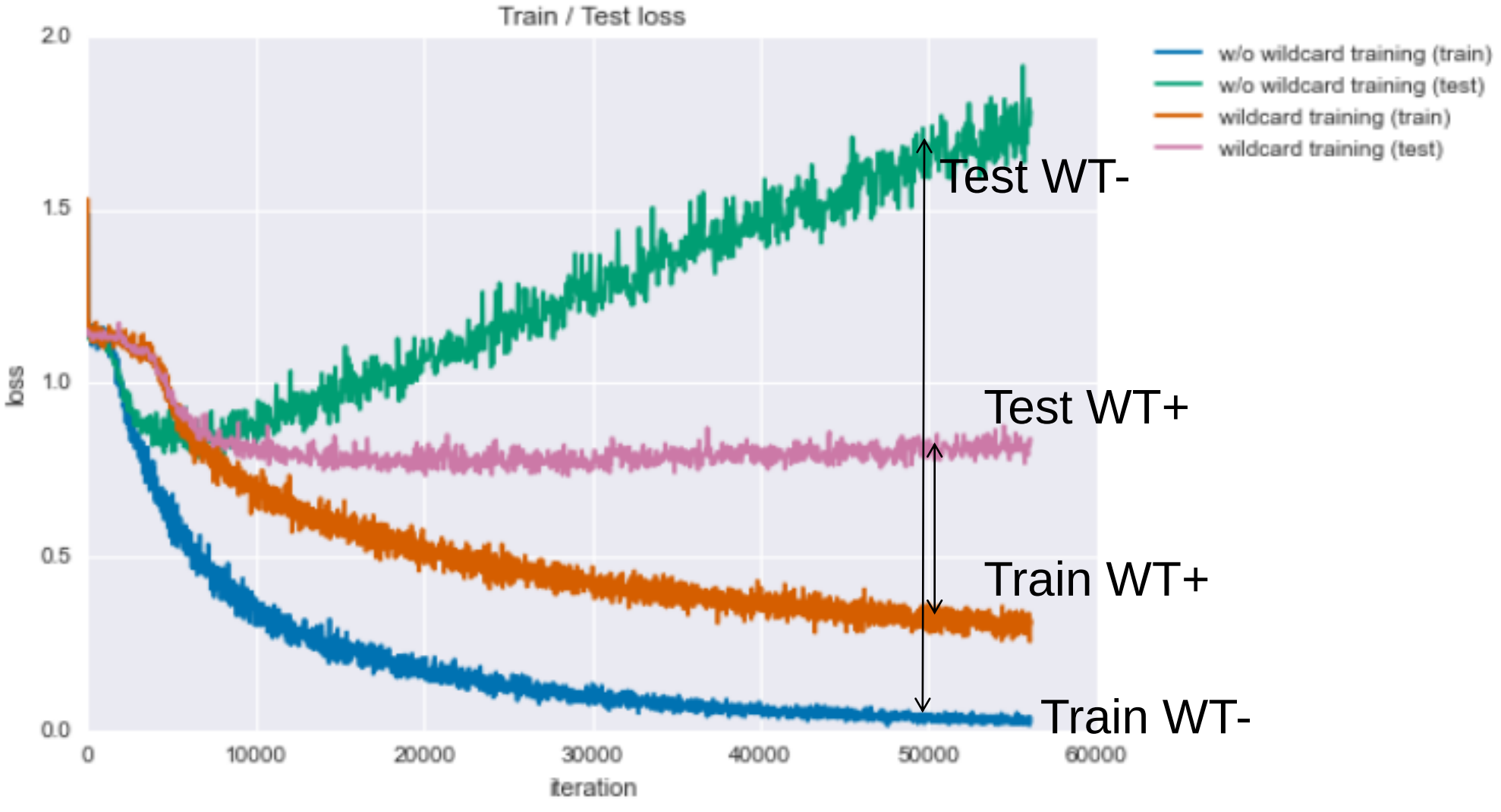
Methods	Accuracy [%]
(proposed) CAE + CLCNN + WT	**87.81**
(proposed) CAE + CLCNN w/o WT	80.95
(proposed) Lookup Table + CLCNN + WT	79.66
Lookup Table + CLCNN w/o WT	73.13
3-gram* + TF-IDF	84.27
Word segmentation** + TF-IDF	67.22
LSI (# topics = 2,000)	84.00
LDA (# topics = 70)	56.10

Latent semantic indexing

* 3-gram approach uses top-30,000 most frequently tokens.
** Word segmentation approach uses all of morphemes in training data.

- Our methods shows the best score in this task.
- Other character-level methods also shows higher score.
  - ◆ Newspaper text include many coinages, proper words and so on.

Appendix | Loss Curve of CLCNN Training (Publisher Estimation)

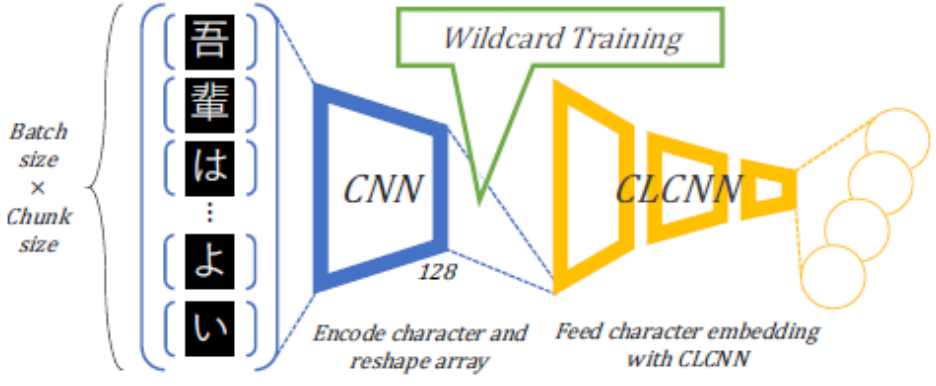


Wildcard training (WT) significantly reduces “over-fitting”

Further extension..

Character Encoder Character-level CNN (CE-CLCNN)

- Enable end-to-end training
- Applicable image-based data augmentation



Method	Accuracy[%]
(Proposed) RE + CE-CLCNN + WT	**58.4**
(Proposed) RE + CE-CLCNN	58.0
(Proposed) CE-CLCNN	54.4
CLCNN + WT [4]	54.7
CLCNN [4]	36.2
VISUAL model [6]	47.8
LOOKUP model [6]	49.1
Ensemble (VISUAL + LOOKUP) [6]	50.3

Wikipedia title category estimation task
(12 class, 206K data)

Our method attained 8% better performance than state-of-the-art.
(Liu et al. ACL2017)

Toward objective bias detection of news media

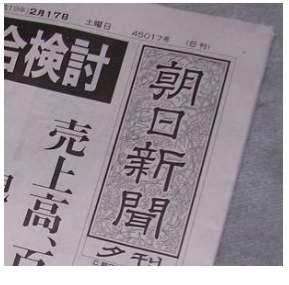
		prediction(件)			
		読売新聞	朝日新聞	毎日新聞	産経新聞
actual(件)	読売新聞	**582**	63	8	65
	朝日新聞	41	**962**	12	139
	毎日新聞	4	15	**1569**	28
	産経新聞	31	61	19	**1237**

Classification performance on major four newspaper : 90.0%

Visualization of “fingerprint” of news.

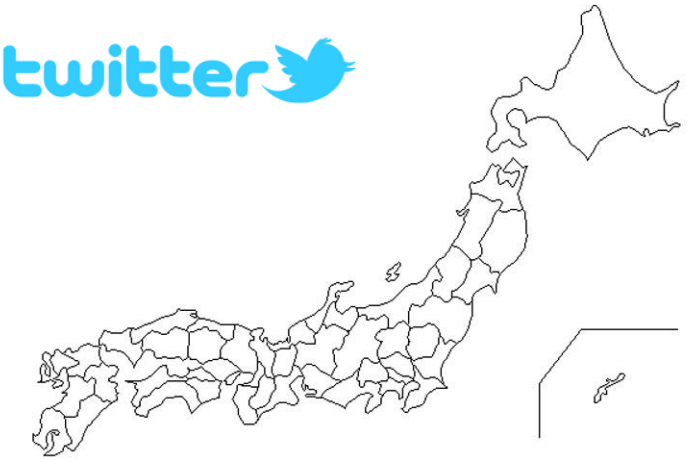
産経新聞社とＦＮＮ（フジニュースネットワーク）が９、１０両日に実施した合同世論調査で、安倍晋三内閣の支持率は５１・８％となり、一昨年１２月の第２次安倍内閣発足以降、最低だった前回調査（７月１９、２０日）より６・２ポイント回復した。不支持率は３６・３％だった。集団的自衛権を限定的に容認する閣議決定や、滋賀県知事選における与党推薦候補の敗北が影響した前回調査より持ち直した。朝日新聞が慰安婦問題をめぐり「強制連行した」との証言に基づく記事を取り消し、自社の過去の報道を検証する記事を掲載したことについては、「検証は十分だと思わない」とする回答が７０・７％を占め、「十分だと思う」（１１・９％）を大きく上回った。女性はどの年代も「十分だ」とする回答が１割に届かず、男性よりも厳しかった。安倍首相が、９月第１週に行う予定の内閣改造・自民党役員人事で女性を積極登用する姿勢を示していることについては、７５・１％が「評価する」とした。首相が新設する方針の安全保障法制や地方創生の各担当相に関し「期待する」と答えたのはそれぞれ５５・４％、５９・２％だった。冷え込んだ日中、日韓関係の改善を求める声も多く、「首脳会

安倍晋三首相は２２日、豪州のアボット首相と電話で会談し、イスラム過激派組織「イスラム国」に日本人２人が拘束された事件をめぐり、情報収集などで協力を求めた。アボット氏は「豪州としても、国際社会と共に出来る限りの協力をしたい」と応じた。電話はアボット首相側からあり、約１５分間行われた。外務省などによると、安倍首相は「『イスラム国』により、邦人の殺害予告動画が配信された。人命を盾にとって脅迫することは許し難い行為で、強い憤りを覚える」と「イスラム国」を非難した。その上で「テロに屈することなく、国際社会によるテロとの戦いに貢献していく。事実関係に関する情報収集、邦人の早期解放に向けた協力などでご支援をいただきたい」と要請した。アボット氏はこれに応じる考えを示し「日本政府及び国民は、この困難を乗り越えるものと確信している」と語った。両首相は、アボット首相が年内に訪日できるよう調整を進めることでも一致した。安倍首相は２２日夜、英国のキャメロン首相とも電話で会談し、人質事件で協力を求めた。キャメロン首相は「日本が困難な時期にある中で、自分は日本と共にあり、情報協力を含め、できる支援はすべて行う考えだ



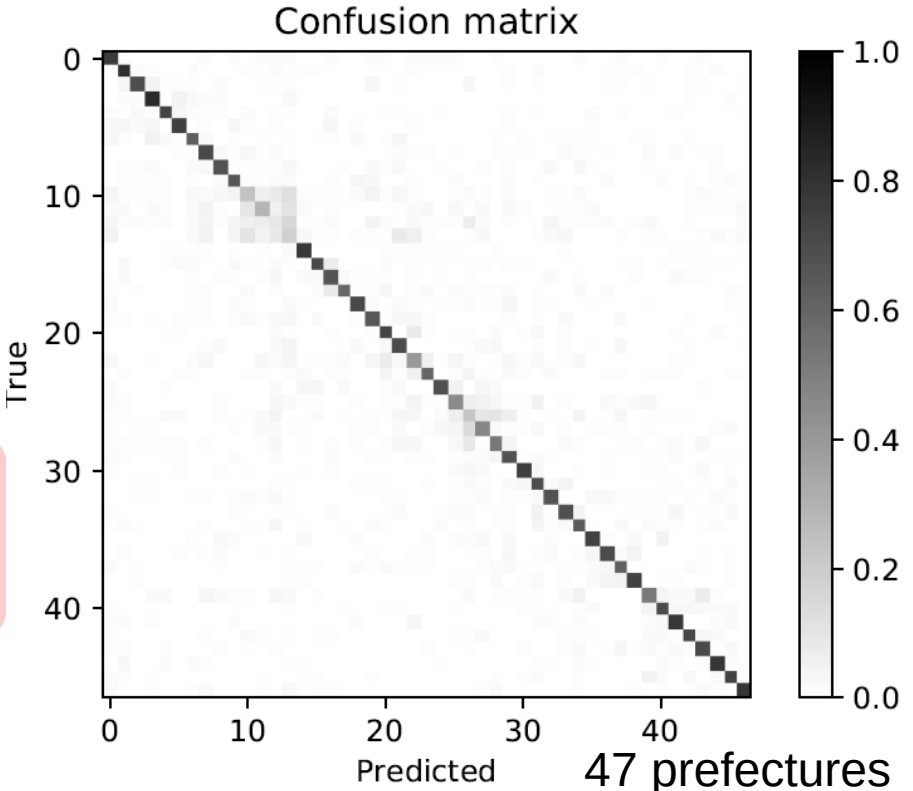
Estimation of user location on Twitter

Information of “user location” is increasing.
- exact and timely ad., job offer, etc.



Collect
2000 users x 500 tweets x 47 pref.
analyze 47M tweets in total

- conventional topic model
- several cutting-edge deep models



63% accuracy in prefecture-level estimation (47 class classification)

Text data augmentation considering the meaning

A Quick Brown Fox Jumps Over The Lazy Dog 0123456789
A Quick Brown Fox Jumps Over The Lazy Dog 0123456789
A Quick Brown Fox Jumps Over The Lazy Dog 0123456789
A Quick Brown Fox Jumps Over The Lazy Dog 0123456789
A Quick Brown Fox Jumps Over The Lazy Dog 0123456789
A Quick Brown Fox Jumps Over The Lazy Dog 0123456789
A Quick Brown Fox Jumps Over The Lazy Dog 0123456789
A Quick Brown Fox Jumps Over The Lazy Dog 0123456789
A Quick Brown Fox Jumps Over The Lazy Dog 0123456789
A Quick Brown Fox Jumps Over The Lazy Dog 0123456789
A Quick Brown Fox Jumps Over The Lazy Dog 0123456789

deep learning needs large amount of training data

- For images: Easy augment is available -- rotation, shift, crop, enlarge
- For text: Not easy : Need to analyze the meaning of the target

Replace word with similar other one in the sentence using
sparse composite document vector（SCDV）[*1] + fastText[*2]

Table 4.1: 各手法と推定結果の関係

手法名	Accuracy	Precision	Recall
SCDV-fastText(M-2)	**63.3**	**69.7**	**63.3**
SCDV-fastText(M-1)	58.9	63.5	58.9
SCDV-GloVe(M-2)	54.4	64.7	53.4
GloVe(M-2)	54.4	63.9	54.4
Thesaurus Random	53.5	60.8	53.3
Nearest fastText	53.2	57.6	63.4
Nearest GloVe	51.3	54.2	60.3
GloVe(M-1)	51.1	55.1	61.1
SCDV-Word2Vec(M-2)	50.8	56.4	51.1
SCDV-GloVe(M-1)	50.0	64.7	44.4
Nearest Word2Vec	49.9	55.4	50.1
Word2Vec(M-2)	48.3	56.2	45.6
Word2Vec(M-1)	48.1	56.8	48.8
SCDV-Word2Vec(M-1)	47.8	55.4	43.1
fastText(M-1)	42.2	54.5	42.2
fastText(M-2)	41.1	51.0	42.1
No Augmentation	**31.1**	**41.8**	**34.1**

原文

－ ...今回 の 討論会 は かなり の 収穫 だ た に 違い ない ...

置き換え後（拡張）

－ ...この度 の 討論会 は かなり の 収穫 だ た に 違い ない ...

－ ...今回 の ディベート は かなり の 収穫 だ た に 違い ない ...

－ ...今回 の 討論会 は 豊富 の 収穫 だ た に 違い ない ...

－ ...今回 の 討論会 は かなり の 取れ高 だ た に 違い ない ...

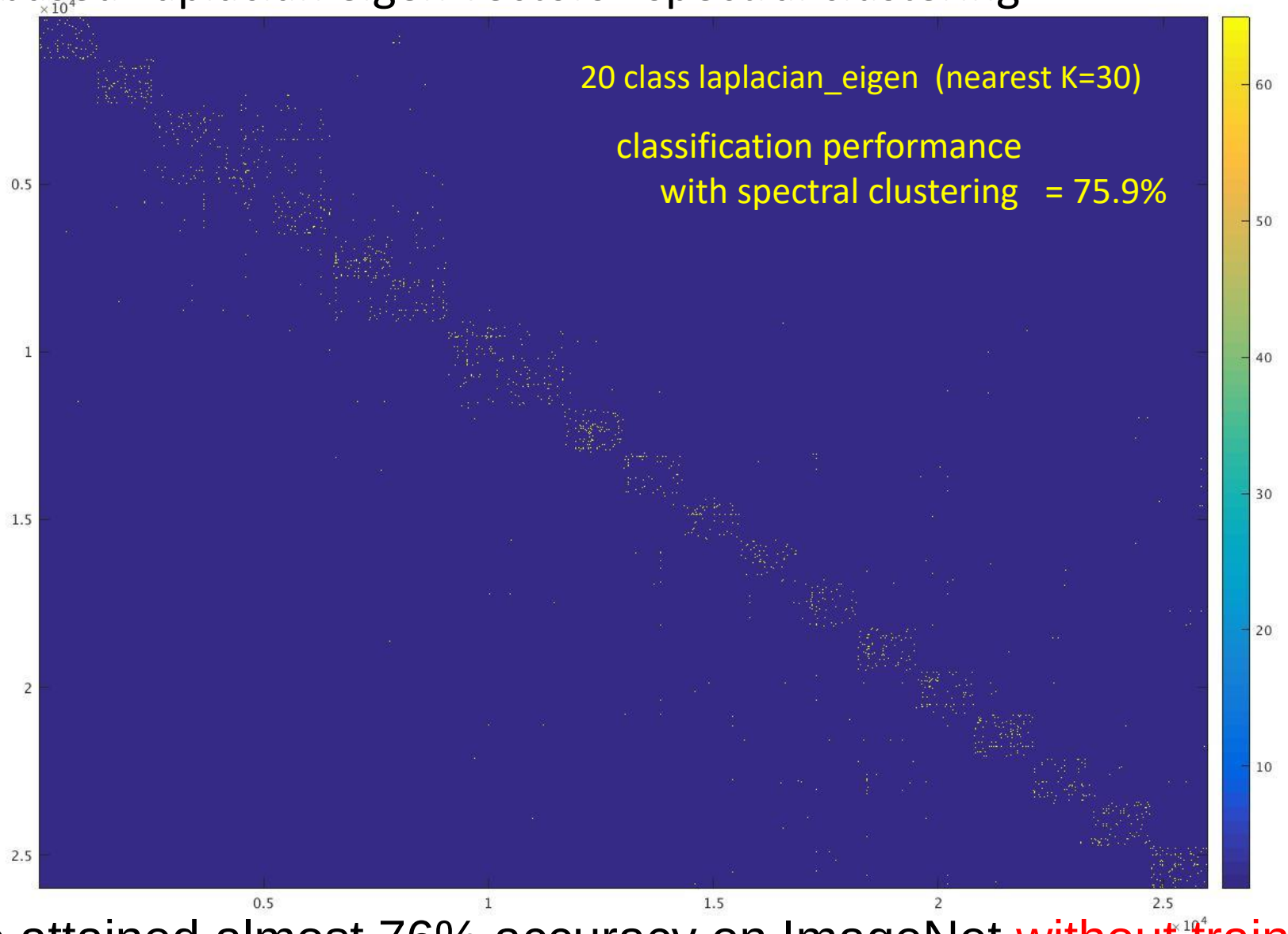
With this augmentation,
30+% gain for 9-class topic estimation task

*1 Dheeraj Mekala, Vivek Gupta, Bhargavi Paranjape, Harish Karnick, "SCDV : Sparse Composite Document Vectors using soft clustering over distributional representations," arXiv:1612.06778, 2017

*2 Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov, "Enriching Word Vectors with Subword Information," arXiv:1607.04606, 2016

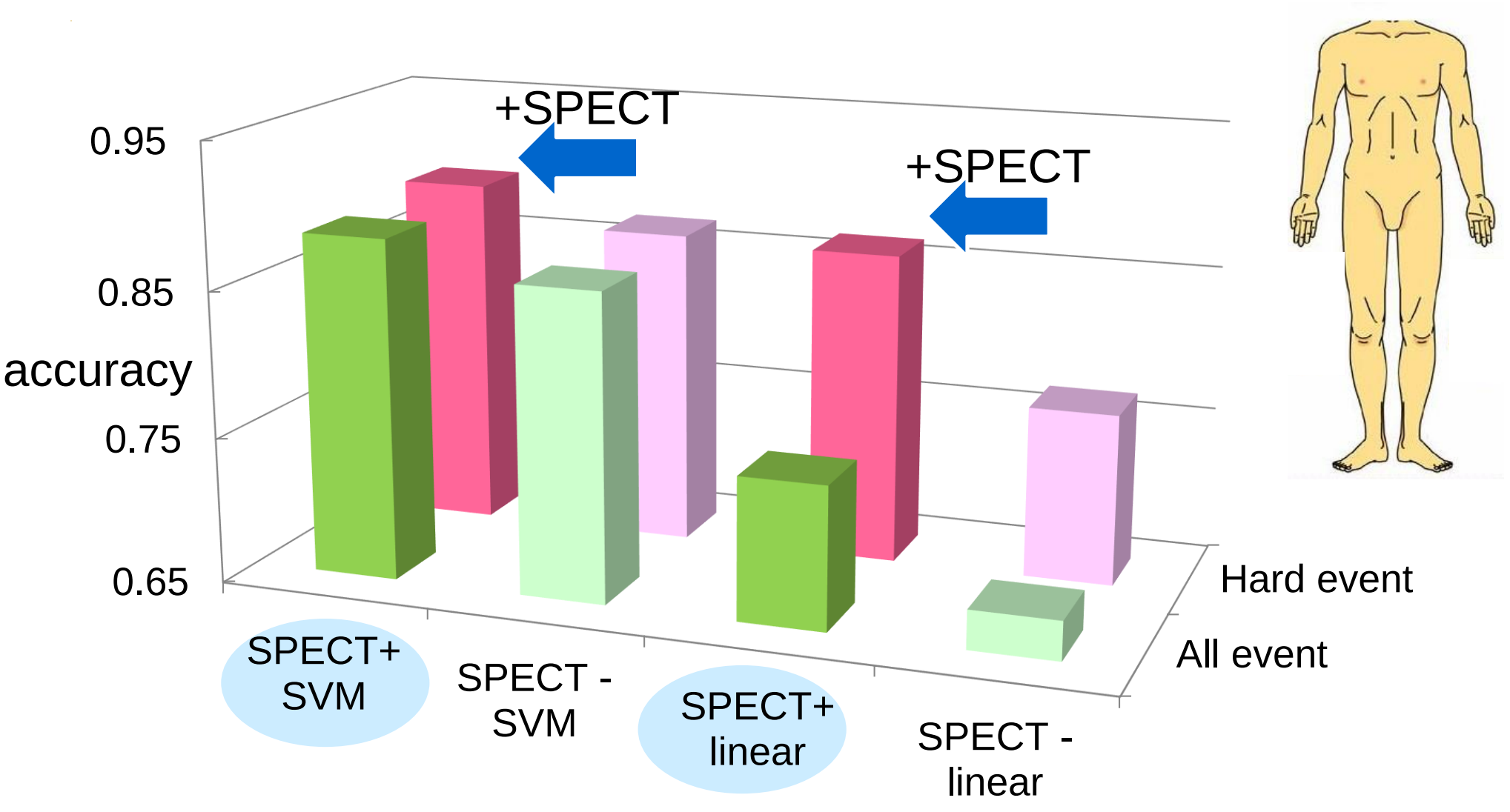
Unsupervised training - clustering performance on ImageNet 20 class

- VGG-19 pre-trained network
- K-NN based Laplacian eigen vectors+ spectral clustering



We attained almost 76% accuracy on ImageNet without training!

Risk estimation of heart events during surgery

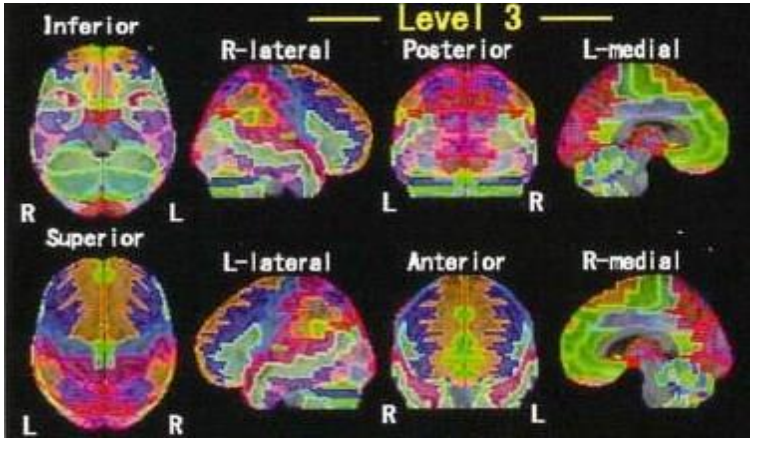


We predicted critical heart accident “hard event” during surgery with 90% accuracy.

Prediction of progression of mild cognitive impairment (MCI) to dementia

With new quantitative index from SPECT

	SE(%)	SP(%)
within 1 yr.	87.0	75.9
3 yr.	89.0	79.6
5 yr.	90.4	88.5



vbSEE: voxel-based analysis stereotactic extraction estimation

→ Our method predict the progression accurately.













Intelligent Information Processing Lab (IIPL@Hosei)

Our computational resources

As of Mar. 2018

one computer / person +

GPUs

nVidia GTX1080Ti ×40

（+ 20-30 each / year）

nVidia Tesla P100

3x Xeon 24cores with

1TB, 512GB, 256GB RAM

（+2-3 each /year）

60TB RAID-6 Disk array

(upto 144TB)

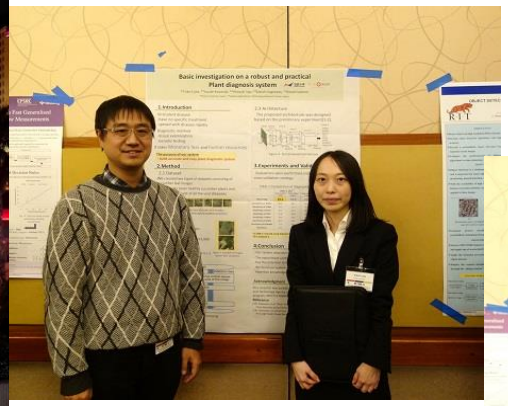
IEEE BigData 2016,
(Washington D.C, USA:
2016/12)

IEEE TALE 2014,
(Wellington, New Zealand: 2014/11)

IEEE CSPA 2018,
(Penang, Malaysia:
2018/3)
Best Paper Award

SPIE Medical Imaging 2014
(San Diego, US: 2014/2)

ISVC 2015
（Las Vegas, USA, 2015/12）



IEEE ICMLA 2016,
(Anaheim, USA:
2016/12)

Intelligent Information Processing Lab (IIPL@Hosei)

We welcome you!