

Small sample learning: An emergent trend in Practical AI



cinnamon

Châu Thành Đức, Ph.D
Technical Leader



CINNAMON AI LABS



- ❁ Tokyo-based AI Startup.
- ❁ Raised \$15M in Series B.
- ❁ Top 25 AI Startup to watch (Forbes2019)
- ❁ Miku (CEO): Woman of the year (NHK2018)
- ❁ Having: 80+ AI Researchers (15+ PhD)
- ❁ Main products:

FLAX SCANNER

Document Image Analysis

Rossa Voice

Speech Recognition

Introduction



Ph.D in Information Science (2014), Japan Advanced Institute of Science and Technology



Lecturer, Faculty of Information Technology
Ho Chi Minh University of Science



Technical Leader, Research Department
Cinnamon AI Labs



My mission: to bring AI research to real life



✿ My personal mission

To bring AI research to real life applications



✿ What is real life application?

- Real problem
- Feasibility
- Superiority

Only a few giant companies have big data.

→ Small data is the future of most of us!

The importance of small sample data



- ✿ **Small data** is data that is “small” enough for human comprehension. It is data in a volume and format that makes it accessible, informative and actionable (*)

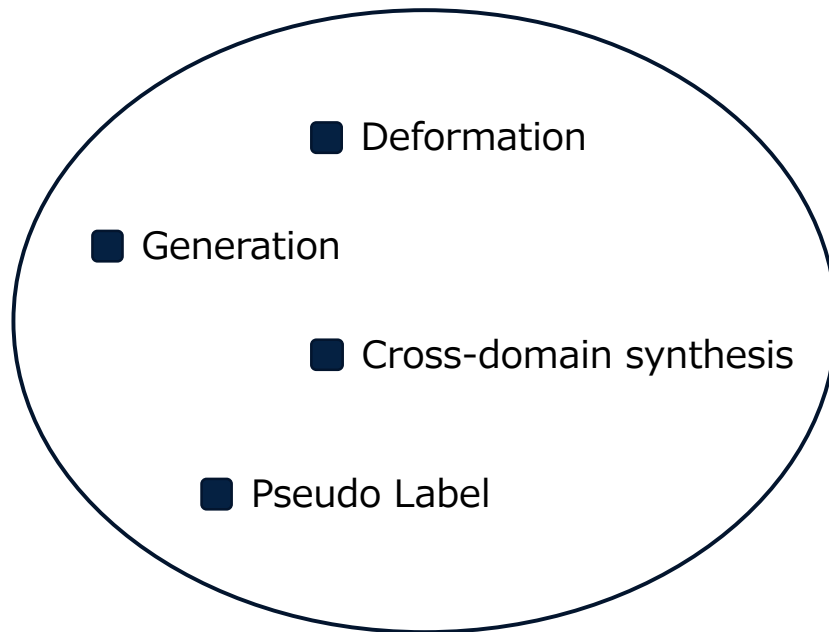
- ✿ Why small data?
 - Insufficient data for conventional big data approaches
 - Long-tail distribution existed extensively in big data
 - Lack of labels due to high cost of human annotations
 - Big data is not the way of human.

(*) Wikipedia

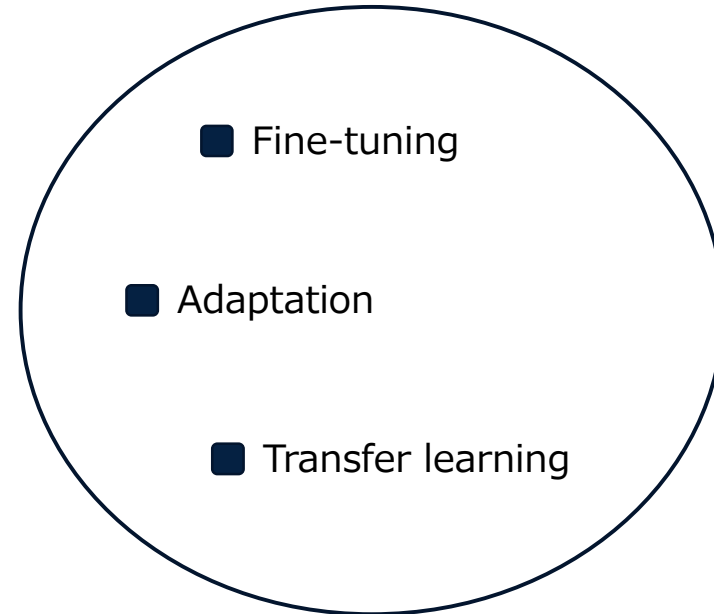
Two main approaches for small data: Data augmentation and Model rectification



Small sample learning (SSL)



Data Augmentation



Model Rectification



Deformation

(noise, mirror, scaling, pose, lighting)

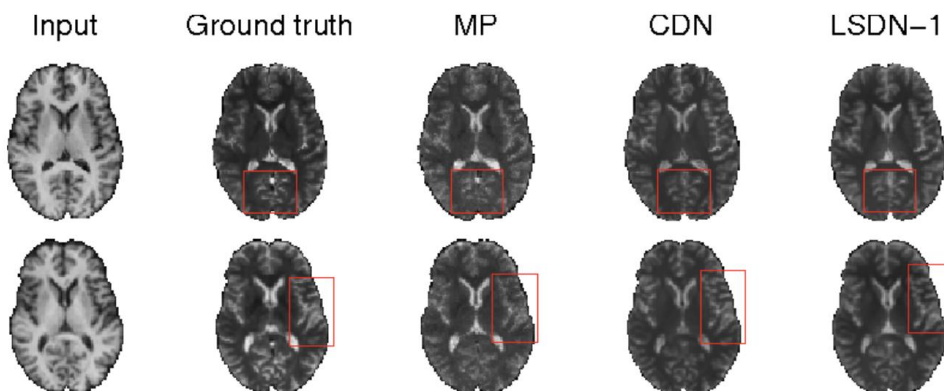


Generation

(VAE, GAN)

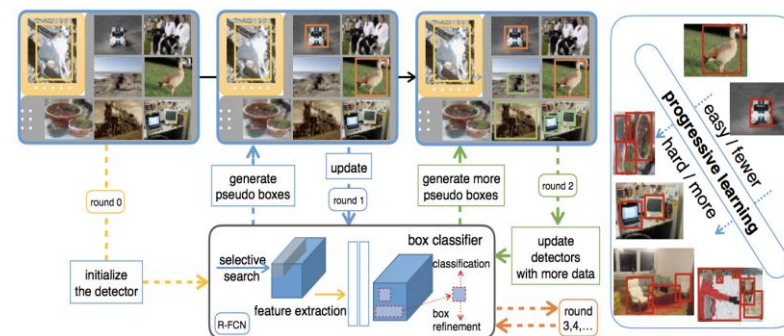


Cross-domain synthesis



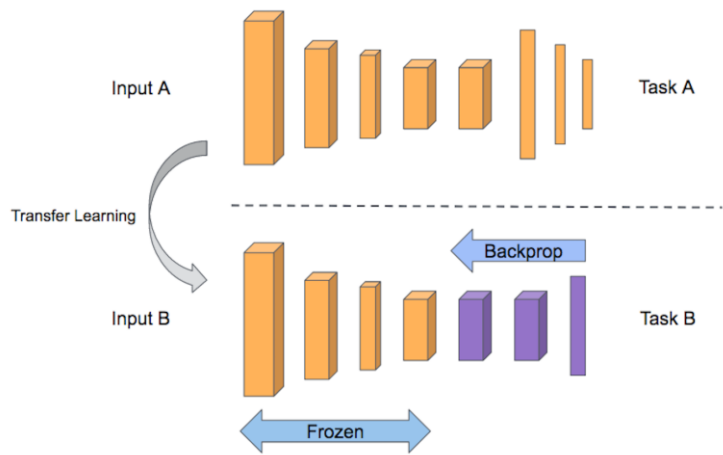
Pseudo label

(self-paced learning, dual learning)

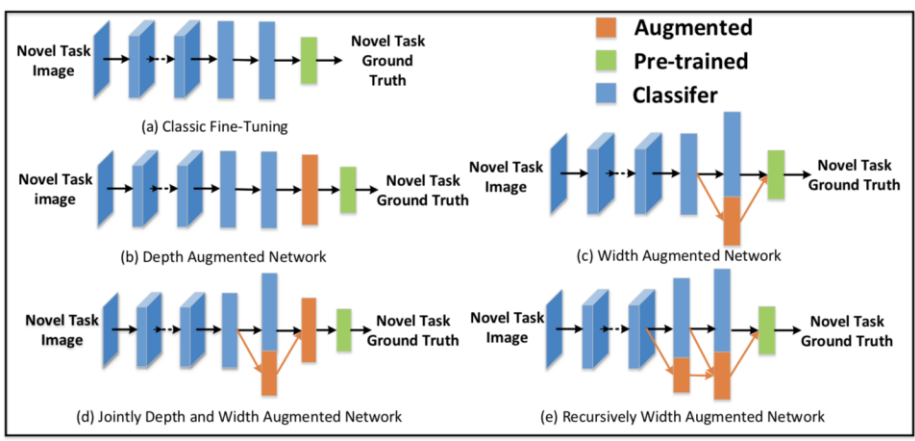




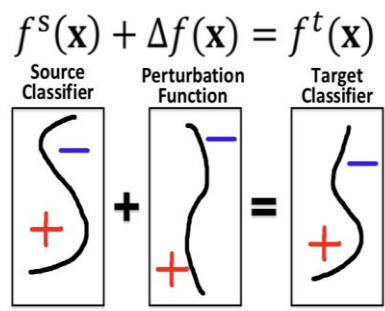
Transfer Learning



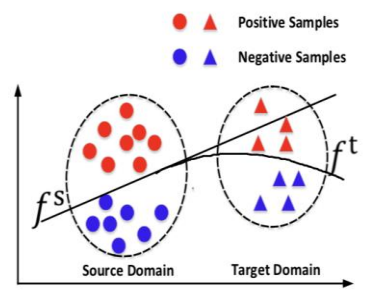
Fine-tuning



Domain adaptation



(a) Adapting existing model



(b) Diagram of model adaptation

Small data in Cinnamon (1): Document Image Analysis



❁ Document Image Analysis (DIA)

Unstructured Data

Unstructured Data,
Photo-scanned data,
PDF data, etc.

10sec/scan

Structured Data

Type	Company
Company	Nexus FrontierTech Ltd.
Address	1-3-1 Yuraku-Cho, Chiyoda, Tokyo
Date	2017-4-3
Name	Akio Tanaka
Marital Status	Married
DoB	1964-3-10
	...



Integrated to
Internal systems

Small data in Cinnamon (1): Small data techniques for DIA



Deformation (Layout analysis)

30年8月15日 社保本人		
医学管理等	在宅医療	検査
0点	0点	168
精神科専門療法	処置	手術
0点	0点	10,020
		保険
合計		102,610
負担額		30,780

30年8月15日 社保本人		
医学管理等	在宅医療	検査
0点	0点	168
精神科専門療法	処置	手術
0点	0点	10,020
		保険
合計		102,610
負担額		30,780

30年8月15日 社保本人		
医学管理等	在宅医療	検査
0点	0点	168
精神科専門療法	処置	手術
0点	0点	10,020
		保険
合計		102,610
負担額		30,780

30年8月15日 社保本人		
医学管理等	在宅医療	検査
0点	0点	168
精神科専門療法	処置	手術
0点	0点	10,020
		保険
合計		102,610
負担額		30,780

Synthesis and Generation (OCR)

嘉陽製作株式会社

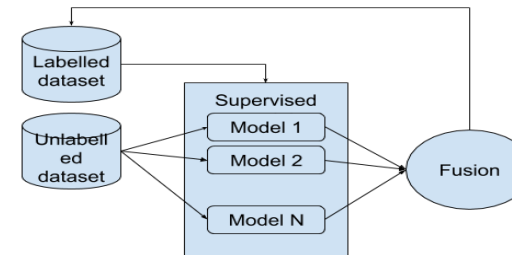
四谷4丁目29番地802

平衡濃度に近い値である

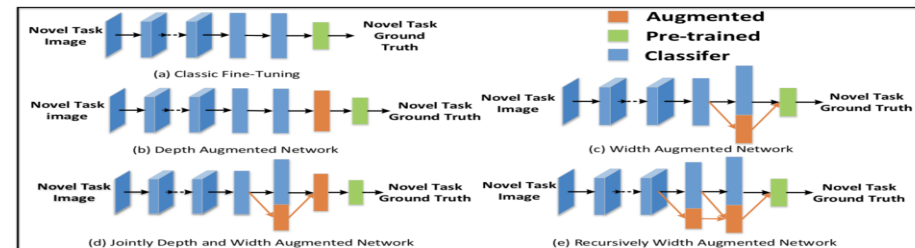
平衡濃度に近い値である

平衡濃度に近い値である

Pseudo label (OCR)



Fine-tuning



Small data in Cinnamon (1): Document Image Analysis

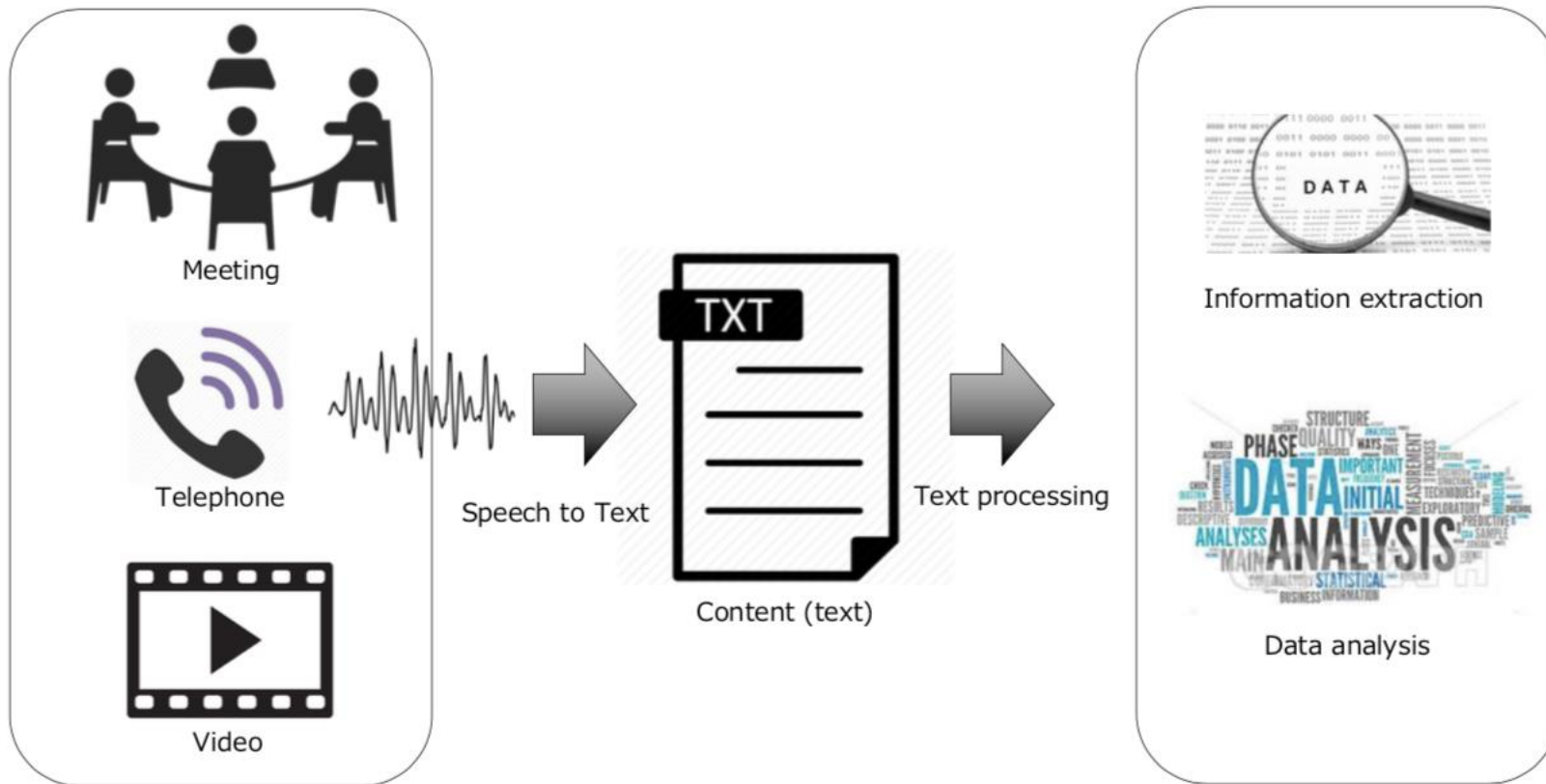


Document name	Comprehension type		Accuracy*	
	Form	Text	Image Recognition	+NLP
Insurance related A Company	Insurance policy	Nonstandard Typed	95% (75%)	98%- (93%-)
	Driving license	Nonstandard Typed	90% (84%)	97%- (96%-)
	Vehicle certificate	Standard Typed	95% (84%)	98%- (96%-)
Major insurance B Company	Application form	Standard Handwritten	89-95%	-
Major insurance C Company	Application form	Standard Handwritten	88-89%	95%-
Major printing D company	Account open application	Standard Handwritten	88-100%	98%-
	Residential certificate	Nonstandard Handwritten	-	90-100%

Small data in Cinnamon (2): Automatic Speech Recognition



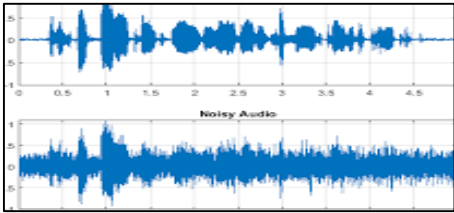
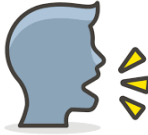
❁ Automatic Speech Recognition (ASR)



Small data in Cinnamon (2): Speech Recognition



✿ Small data techniques

Deformation		<ul style="list-style-type: none">- noise- reverberation- pitch- time stretch
Adaptation		<ul style="list-style-type: none">- speaker- gender

✿ Results:

CER (%)	Cinnamon ASR	Google ASR	Retrieva	NextGen
Zoom meeting	22.16	40.55	-	-
Telephone	28.26	-	34.67	31.44

Conclusion



- ✿ Small data is one of the future of AI
- ✿ Cinnamon is persistently to pursue small data
- ✿ Cinnamon is open for all who are interested in small data



cinnamon

Extend human potential by eliminating repetitive tasks