PROJECTS

회사1	[1] (EN) Text Chatbot	2 months
	[2] (KR) Multi-speaker Speech Synthesis Model	4 months
회사2	[1] Automobile Video Recommendation	2 months
	[2] Comics Recommendation	2 weeks
대학원 연구실	[1] Network Embedding Generation	2y 6m
	[2] DNN Model Quantization - 1	1y 10m
	[3] DNN Model Quantization - 2	3 months
	[4] Artificial Intelligence Assistant	2 months
회사3	[1] FAQ Chatbot	3 months
	[2] Competition Rule Recommendation	2 months
	[3] Match Result Recorder	1 month

회사1

[1] (EN) Text/Audio Chatbot

[2] (KR) Multi-speaker Speech Synthesis Model

(EN) Text/Audio Chatbot

Speech Recognition → Sentence Classification → Sentence Generation → Speech Synthesis

Sentence Classification

Problems

- High classification error

Solutions

- Dataset relabeling
- Applied proper metric to evaluate unbalanced classes
- Made model to be more robust at typo from Speech Recognition (FastText)

Sentence Generation

Problems

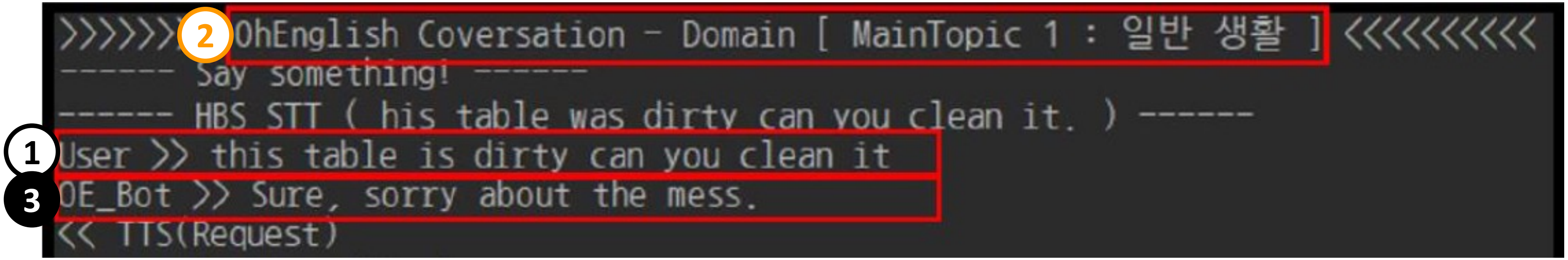
- Poor performance of long-sentences generation

Solutions

- Applied attention mechanism to Seq2Seq model

(EN) Text/Audio Chatbot

1 Speech Recognition → 2 Sentence Classification → 3 Sentence Generation → Speech Synthesis



(KR) Multi-speaker Speech Synthesis Model

- **Dataset preparation**

Web Crawling	Audio files
	Script files
Preprocessing	Cut audio files into files of sentences
	Cut script files into sentences (by comparing STT results)

- **H-params optimization**

- **Demo** https://jarvis08.github.io/pjt_hbs_multi.html

회사2

[1] Automobile Video Recommendation

[2] Comics Recommendation

Automobile Video Recommendation

Exp 1, 2	**Thomson Sampling** h-params tuning	Purpose	Adjustment of trade-off between exploration & exploitation
		Reason	[Exp-1] High **matrix sparsity**
			[Exp-2] Considering **time bias** enhanced by low traffic
Exp 3, 4	**Ranking algorithm's** h-params tuning	Purpose	Searching the key model among ensembled models
		Reason	Other well performing services had been used **similar pipelines** • Therefore, assumed that the composition of used models are good enough
Exp 5	**Item2Vec** instead of Matrix Factorization	Purpose	Overcome Matrix Factorization model's limitation
		Reason	Needed to generate reco. results **within limited item list** • The limited items rated 30~40th on avg., if we force the limitation off
			Needed some models which **capture information** which MF can't

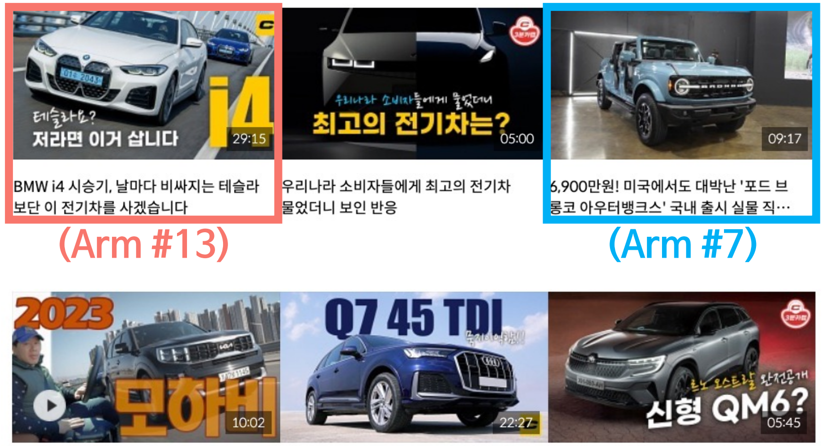
Automobile Video Recommendation

Exp 1, 2	**Thomson Sampling** h-params tuning	Purpose	Adjustment of trade-off between exploration & exploitation
		Reason	[Exp-1] High **matrix sparsity**
			[Exp-2] Considering **time bias** enhanced by low traffic

Target Item (Bandit)

제네시스보다 저렴한 5천만원대 전기차 BMW i4 edrive 40

Reco. Result (Selected Arms)



BMW i4 시승기, 날마다 비싸지는 테슬라 보단 이 전기차를 사겠습니다

우리나라 소비자들에게 최고의 전기차 물었더니 보인 반응

6,900만원! 미국에서도 대박난 '포드 브롱코 아우터뱅크스' 국내 출시 실물 직…

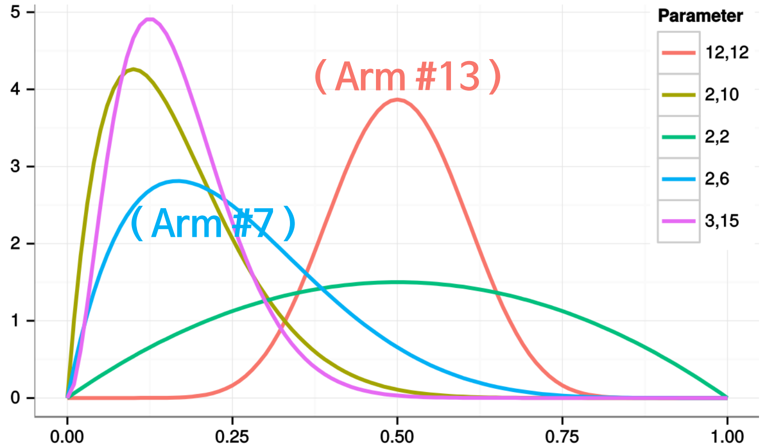
(Arm #13) (Arm #7)

6천만원에 모하비 풀옵션 선택? 모하비가 달라졌다? 2023 모하비

아우디 Q7 45 TDI, 이 차가 답답하면 성격이 급하신 겁니다?

싼타페 쏘렌토 저격가능? 신형 QM6 미리보기? 르노 오스트랄 완전공개!

Beta Distributions of Arms



Automobile Video Recommendation

Exp 3, 4	**Ranking algorithm's** h-params tuning	Purpose	Searching the key model among ensembled models
		Reason	Other well performing services had been used **similar pipelines** • Therefore, assumed that the composition of used models are good enough

[CF]
Reco. result

Item ID	Similarity
3	0.8542
17	0.8345
2	0.7984
23	0.7784
33	0.6214

weight=0.3

[Text Analysis]
Reco. result

Item ID	Similarity
2	0.8441
42	0.8385
17	0.8001
51	0.7871
33	0.7070

weight=0.7

Example of Weighted-sum

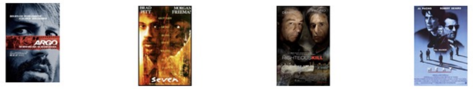
[Ensembled]
Reco. result

Item ID	Similarity
2	0.8303
17	0.8104
33	0.6813
42	0.5870
51	0.5510

Automobile Video Recommendation

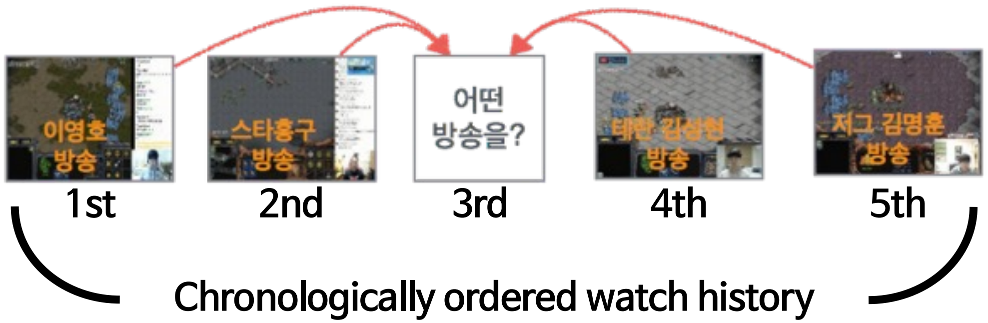
Exp 5	**Item2Vec** instead of Matrix Factorization	Purpose	Overcome Matrix Factorization model's limitation
		Reason	Needed to generate reco. results **within limited item list** • The limited items rated 30~40th on avg., if we force the limitation off
			Needed some models which **capture information** which MF can't

〈 MF Model's Reward Matrix 〉



John	5	1	3	5
Tom	?	?	?	2
Alice	4	?	3	?

〈 Item2Vec Model's Input Sequence 〉



Comics Recommendation

Exp 6	**Word2Vec input dataset** reconstruction	Purpose	Better reflection of Japanese characteristics
		Reason	Previously, model used **nouns** and **pronouns** only
			According to past researches, **verbs** and **adjectives** are also important for JP
Exp 7	Modified **ranking algorithm (RRF to WRF)**	Purpose	Strengthen the key model
		Reason	By previous experiment logs, the only MF used reco. pipeline without ensemble method outperformed ensembled pipeline
			But the ranking algorithm the system was using weakened MF's power

Comics Recommendation

Exp 7	Modified ranking algorithm to Weighted Rank Fusion	Purpose	Strengthen the key by giving weight to rank values
		Reason	By previous experiment logs, the only MF used reco. pipeline without ensemble method outperformed ensembled pipeline
			But the the Weighted-sum Ranking Algorithm weakened MF's power

[Ensembled]

Item ID	Rank
3	1
2	2
17	3

Example of
Weighted Rank Fusion

weight=0.7

Item ID	Rank
3	1
17	2
2	3

[CF]

weight=0.3

Item ID	Rank
2	1
42	2
7	3

[Text Analysis]

weight=0.2

Item ID	Rank
3	1
42	2
2	3

[Image Sim.]

대학원 연구실

[1] Network Embedding Generation

[2] DNN Model Quantization – 1

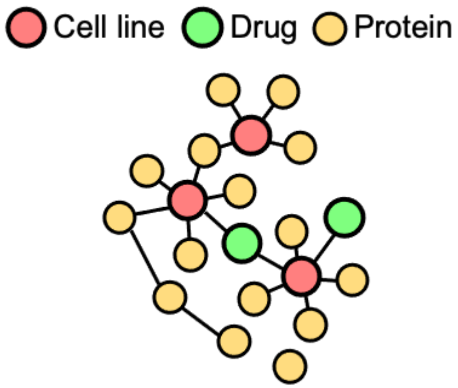
[3] DNN Model Quantization – 2

[4] Artificial Intelligence Assistant

Network Embedding Generation

* **Published in 2022 BIB** (Briefings in Bioinformatics) **Journal**

Project description	[Human Cell lines – Cancer Drugs] Response Prediction
	Network (graph) dataset consist of • **Cell line** nodes • **Drug** nodes • Protein nodes (connected to Cell lines)
	My Task: **Train embedding vectors of Cell lines and Drugs**
Problem	Extremely unbalanced dataset • About **20,000 Protein nodes** • About 900 Cell line nodes • About 300 Drug nodes
	Fails to reflect the relationships between Cell lines & Drugs
	As a result, we got poor response prediction performance



[1] Network Embedding Generation

Solution	Make training process to **focus on relationships between Cell lines & Drugs**
	Response-aware Negative Sampling (RA-NS) • **Cell line & Drug** nodes use **resistant Drug & Cell line** nodes as their negative samples

* Tested Models: **Node2Vec**, Graph Convolutional Network, Graph Transformer Network

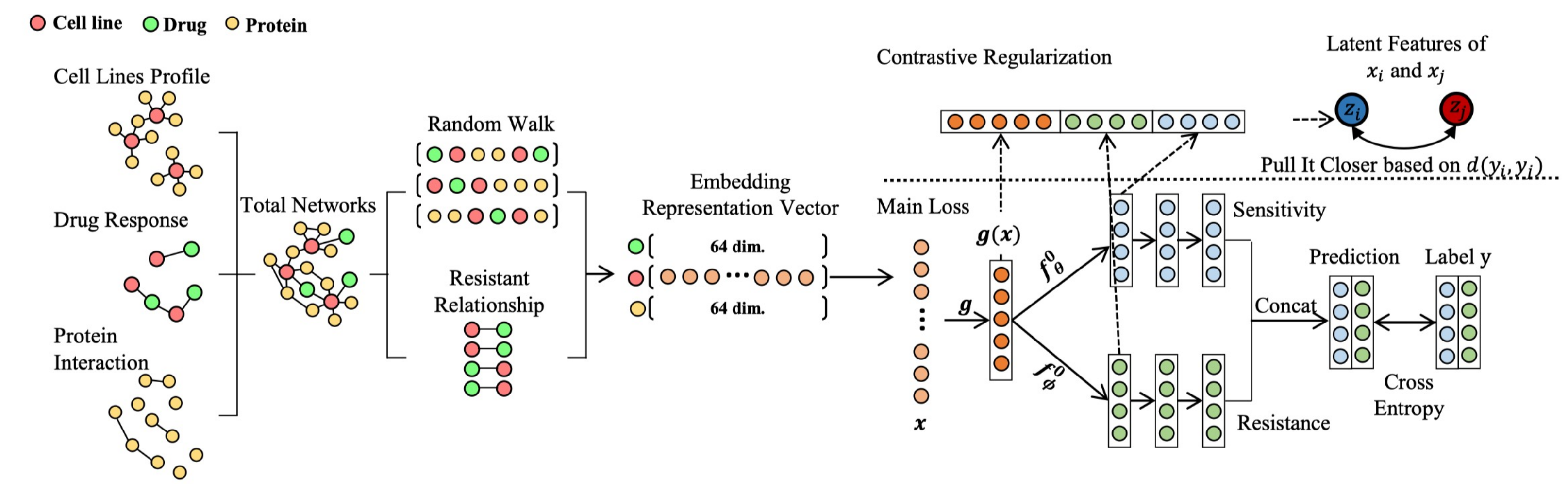


Fig. 1. The framework of RAMP. RAMP consists of two main stages. First, representation vectors are extracted from heterogeneous networks with RA-NS. Second, the multitask architecture of a Bayesian neural network is trained by representation vectors with contrastive regularization.

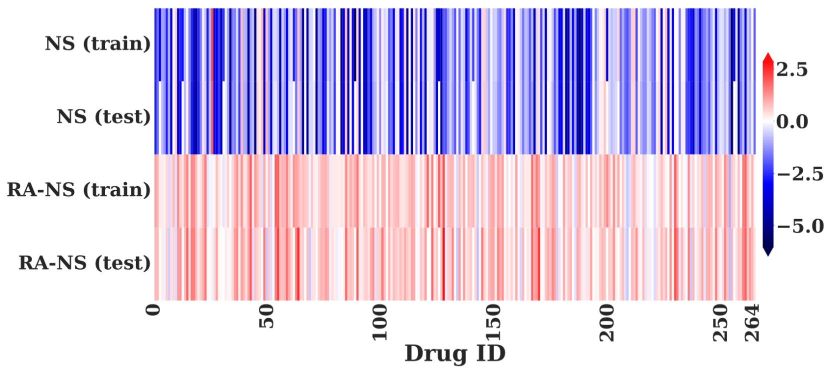
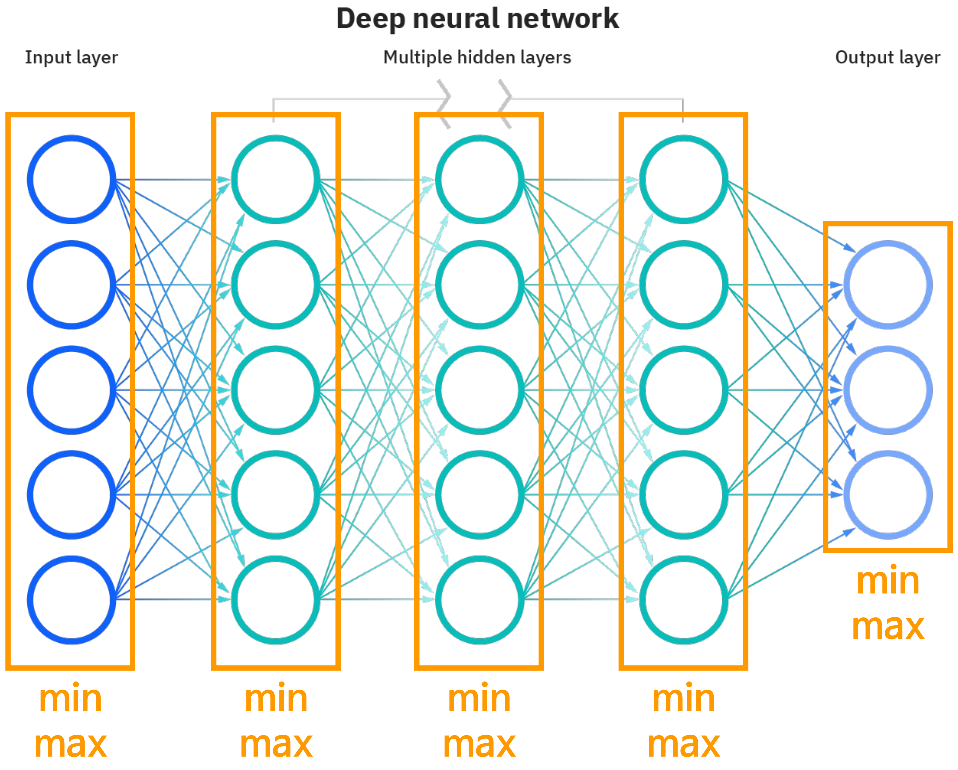


Fig. 2. Embedding similarities among drug and cellines. We subtract the similarity of a drug and its resistant cell lines from the similarity of the drug and its responsive cell lines. The results are normalized and plotted in a heatmap format. The higher (or redder) the value is, the better the embedding reflects the network structure.

DNN Model Quantization – 1

Definition	What is	General DNN models use Float32 type variables
		Quantized models use low-bit INT types at inference
	What for	• Model storage • In memory load • Matrix multiplication with Float32 type cause bottleneck/unusability in low performance H/W

Problem	Problem	Quantized models' **performance (e.g., accuracy) drops catastrophically when using sub-8bit INT type**
	Why legacy tech. suffers	**Too generalized Quantization parameters** • Quantization parameters require: Per layer avg-ed min/max range of intermediate outputs across datasets
		Averaged min/max values include outliers



Solution

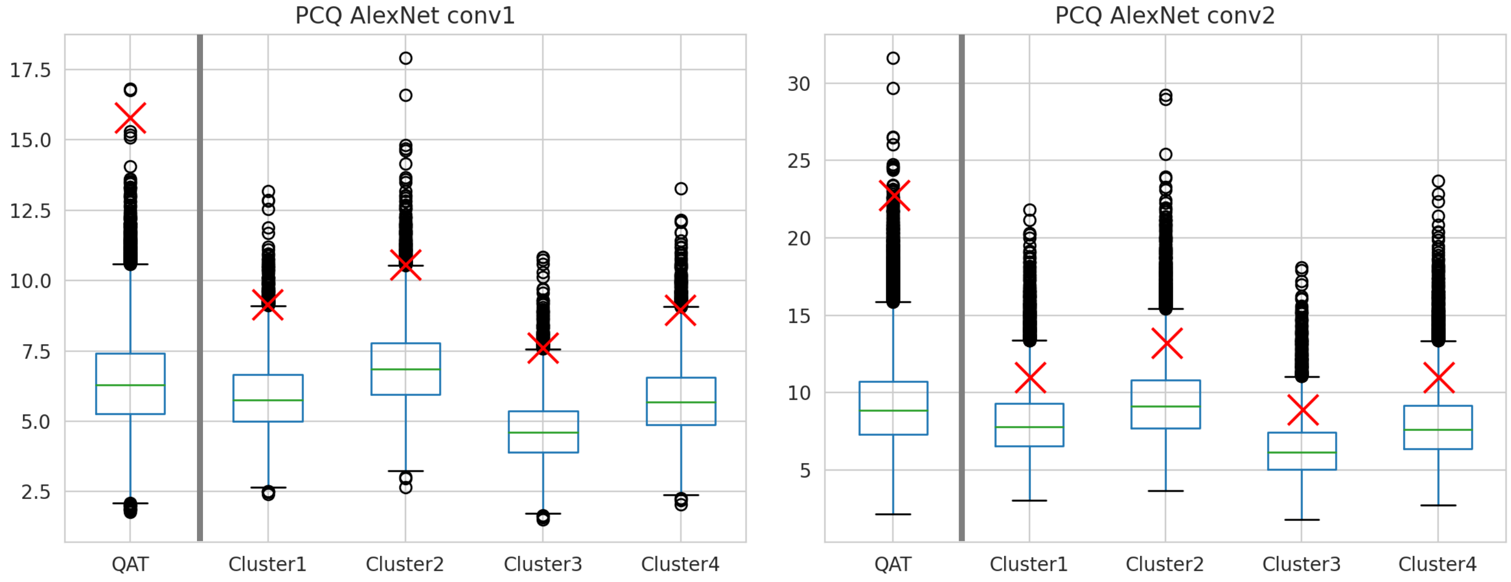
Granular Exponential Moving Average (Granular EMA)

Train Quantization Parameters while **excluding outliers**

Neural Network Aware Clustering (NNAC)

Train Quantization Parameters separately **across clusters of input images**

- Some data might need **shorter min/max range**
- Shorter range means **less information loss**



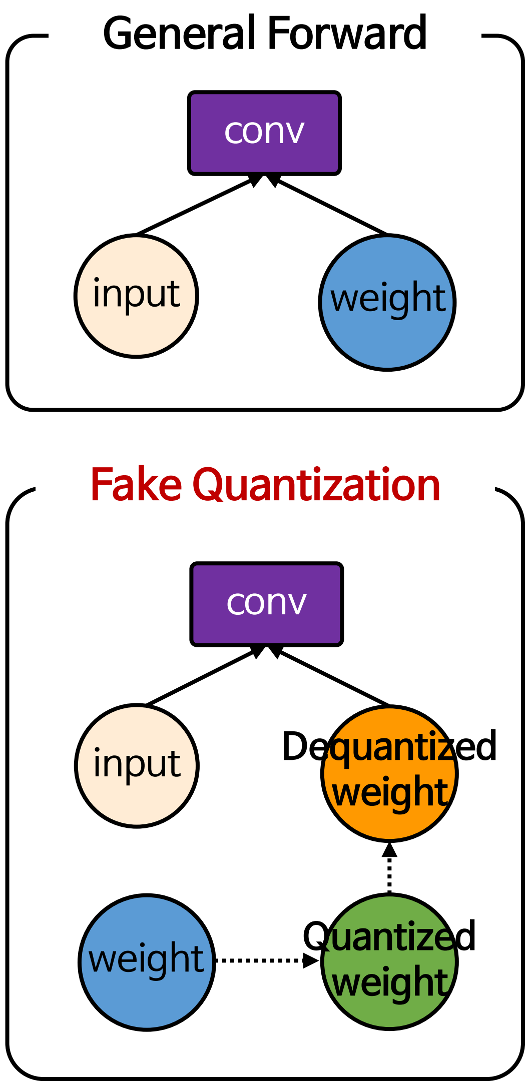
Figures' Description

- Shows that our method
  - how efficiently exclude outliers
  - how to work with clusters
- QAT : Baseline (Google)
- Cluster* : Ours
- X : Trained maximum value
- Box-plots : Actual max values per image

DNN Model Quantization – 2

* **Published in 2022 ICEIC** (International Conference on Electronics, Information, and Communication)

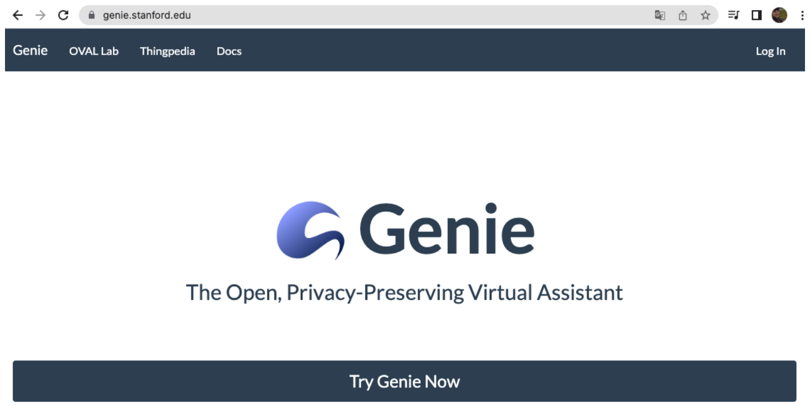
Problem	**Quantization Aware Training** (Google) • Fake-quantize all of the weight matrices with a single low-bit type • **Too much quantization errors** occur and the trained model gets ruined
	QuantNoise (Facebook) • Fake-quantize probabilistically selected subsets of matrices (a subset per matrix) • Trained models **under-prepared** for Quantization
Solution	**Fake Single Precision Training (FST)** • Probabilistically select subsets of weight matrices as QuantNoise • Fake-quantize **selected subsets** with **low-bit type** • Fake-quantize **the rests** with **higher bit type** than the selected

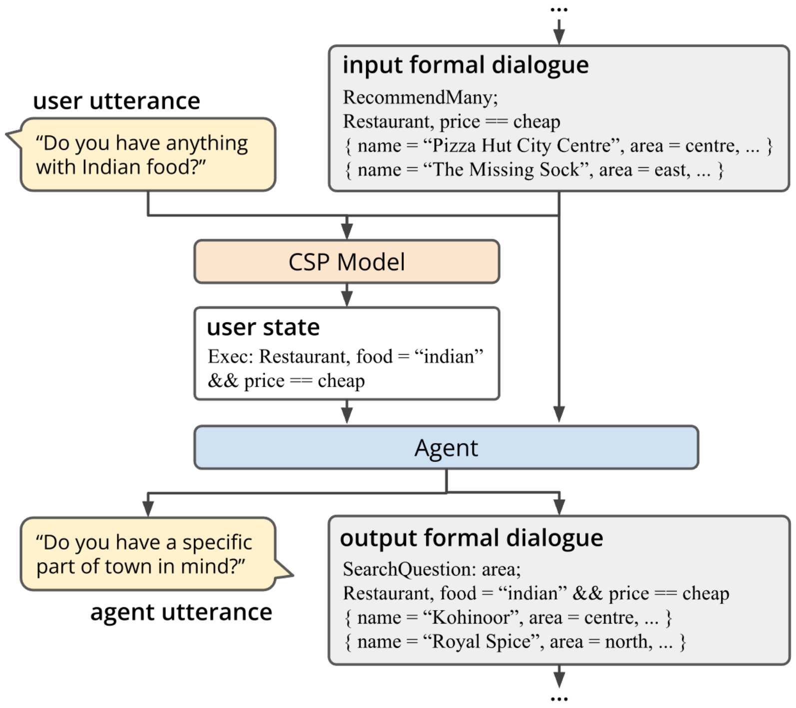


Artificial Intelligence Assistant

- AI Assistant App, **Almond**
  - Currently, the service name has been modified to Genie
  - Developed by Stanford OVAL Lab

- **Training Korean Seq2SQL Model**
  - Dataset preparation
    - Web Crawling
    - Construct templates of sentences (example of sentences)
    - Augment sentences based on templates
  - Train & serve model





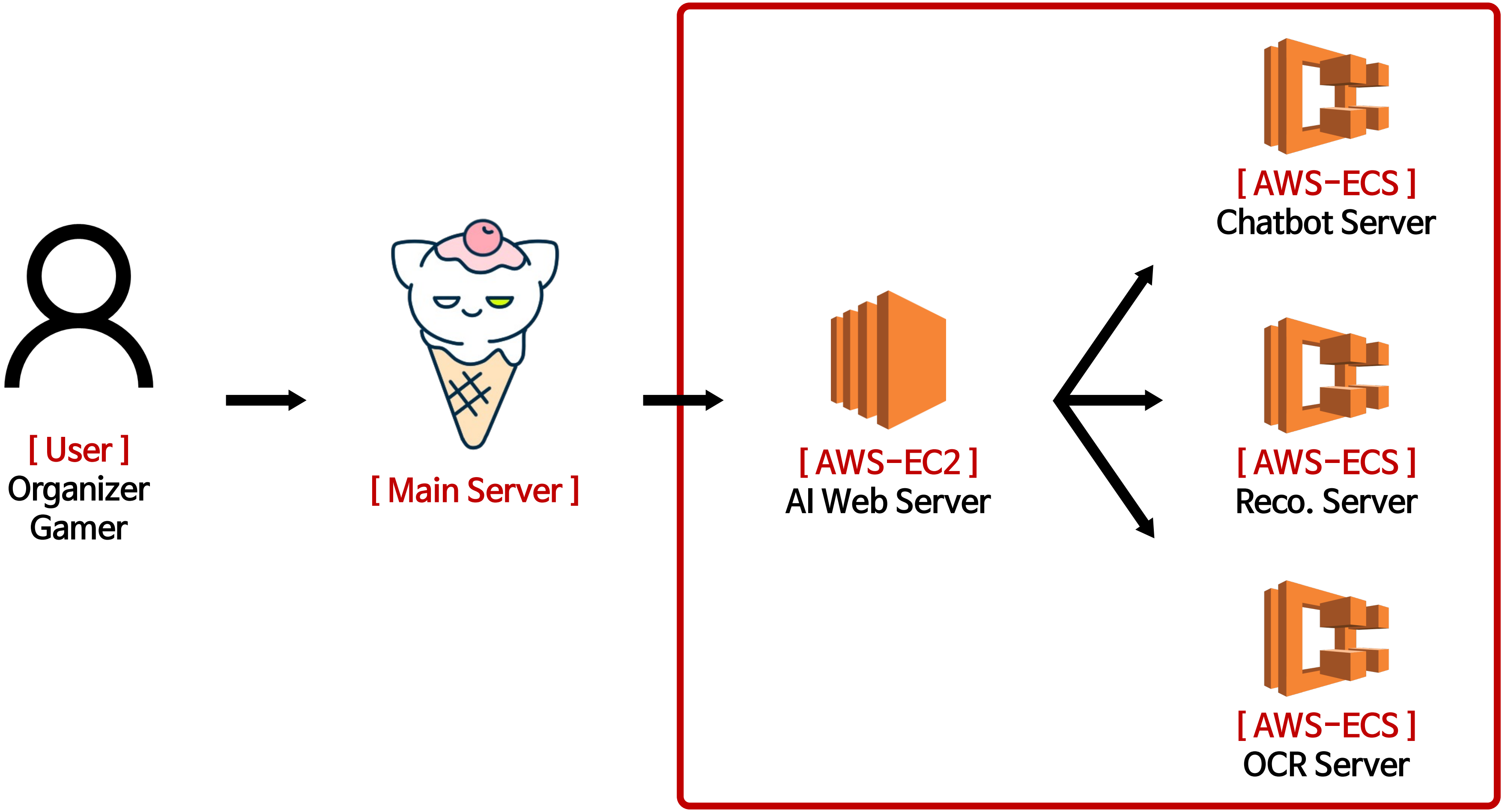
회사3

[1] FAQ Chatbot

[2] Competition Rule Recommendation

[3] Match Result Recorder

AI Server Pipeline



Chatbot	Open Source	RASA
	Utilized	Multi-lingual BERT, StarSpace
	What I've Done	• Dataset Preprocessing • Model Selection • Model Tuning • Model Serving
Competition Rule Recommendation	Open Source	LibRecommender (Alternative Least Square)
	What I've Done	• Define Problem • Dataset Preprocessing • Feature Selection (via Correlations) • Model Selection • Model Tuning • Model Optimization (removed operations)
Match Result Recorder (OCR)	Open Source	Tesseract, Google Vision API

	What I've Done	• Define Problem
		• Define Pipeline
		    • Our Tesseract Model
		    • Cloud API (in case of poor confidence)
		    • Serving
		    • Finetuning