

조동빈 DONGBIN CHO

Birth 1993.01.05

Email jarvis08.cdb@gmail.com

2020.03.01 ~ 2022.02.25

한양대학교 컴퓨터소프트웨어학과 [MS]
강원대학교 산업공학과 [BS]

2019.12.26 ~ 2020.02.29

카카오 추천팀

추천 알고리즘 개발

2018.11.05 ~ 2019.04.22

한빛소프트 AR팀 인공지능파트

텍스트/음성 챗봇 개발

PROJECTS

| | | |
|--|---|--------|
| 한양대학교 대학원 Machine Learning System Lab. | [1] DNN Model Quantization | 1년 8개월 |
| | [2] Network Embedding Generation | 1년 8개월 |
| | [3] Artificial Intelligence Assistant | 2 개월 |
| | [4] DNN Model Quantization | 3 개월 |
| 카카오 | [1] Automobile Video Recommendation | 2 개월 |
| | [2] Comics Recommendation | 2 주 |
| 한빛소프트 | [1] (EN) Text Chatbot | 2 개월 |
| | [2] (KR) Multi-speaker Speech Synthesis Model | 4 개월 |



한양대학교 대학원 · Machine Learning System Lab.

- [1] DNN Model Quantization
- [2] Network Embedding Generation
- [3] Artificial Intelligence Assistant
- [4] DNN Model Quantization

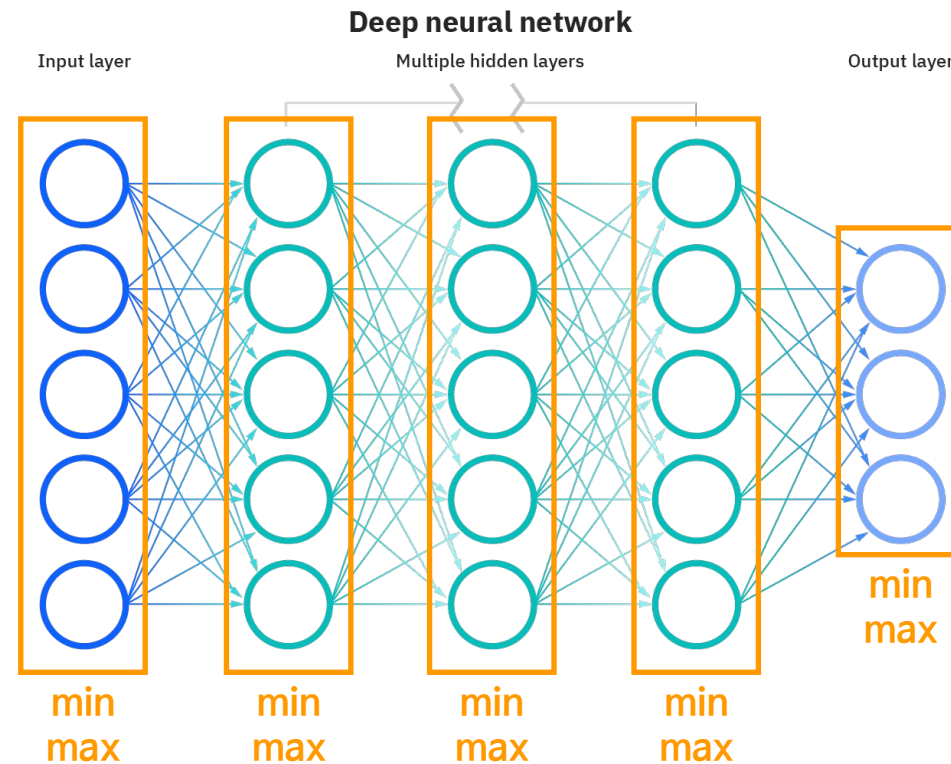


DNN Model Quantization

| | | |
|------------|----------|--|
| Definition | What is | General DNN models use Float32 type variables |
| | | Quantized models use low-bit INT types at inference |
| | What for | <ul style="list-style-type: none">• Model storage• In memory load• Matrix multiplication with Float32 type cause bottleneck/unusability in low performance H/W |

[1] DNN Model Quantization

| | | |
|---------|--------------------------|--|
| Problem | Problem | Quantized models' performance (e.g., accuracy) drops catastrophically when using sub-8bit INT type |
| | Why legacy tech. suffers | Too generalized Quantization parameters <ul style="list-style-type: none">Quantization parameters require: Per layer avg-ed min/max range of intermediate outputs across datasets |
| | | Averaged min/max values include outliers |



Solution

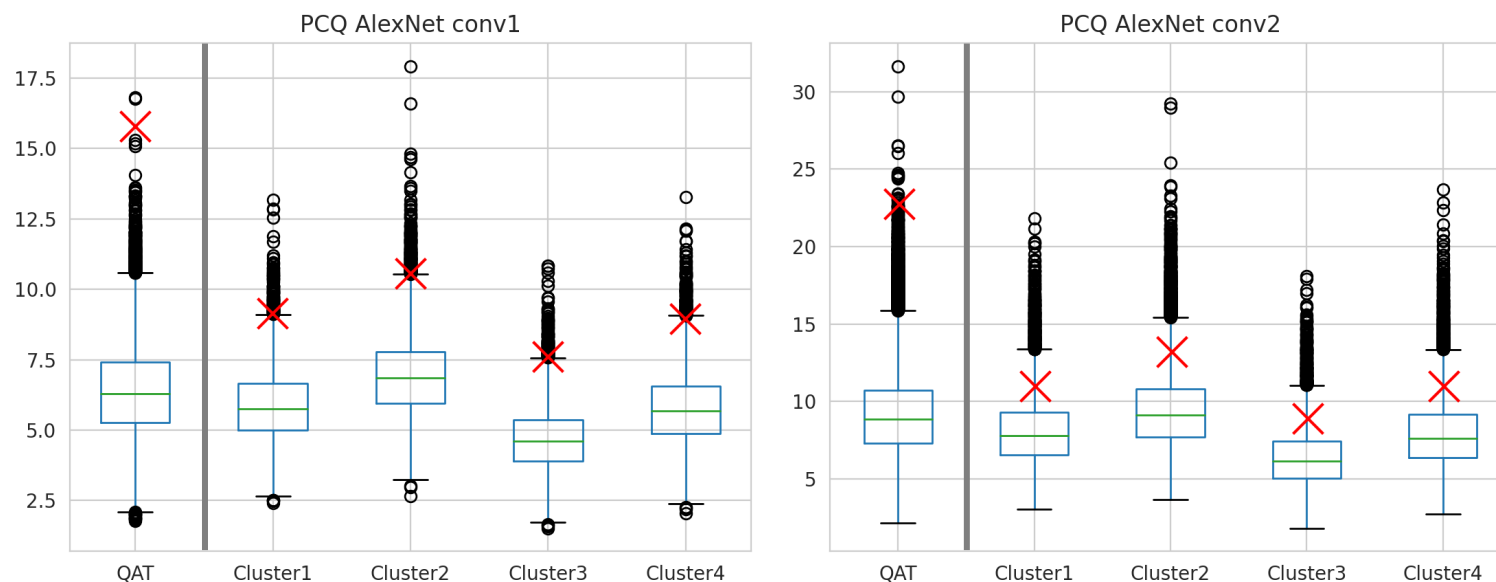
Granular Exponential Moving Average (Granular EMA)

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

Network Embedding Generation

* Currently submitted to BIB (Briefings in Bioinformatics)

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

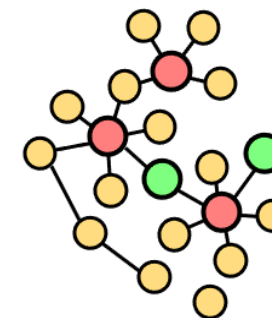
Extremely unbalanced dataset

- 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

● Cell line ● Drug ● Protein



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

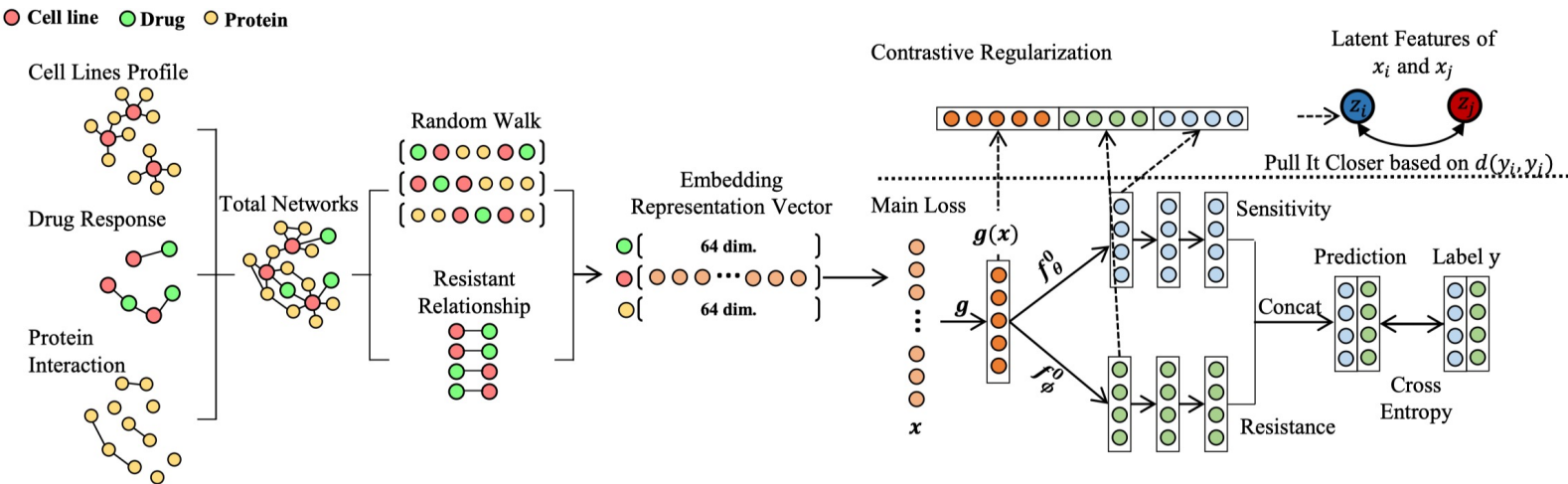


Fig. 1. The framework of RAMP. RAMP consists of two main stages. First, representation vectors are extracted from heterogeneous networks with RA-NS. Second, the multitask architecture of a Bayesian neural network is trained by representation vectors with contrastive regularization.

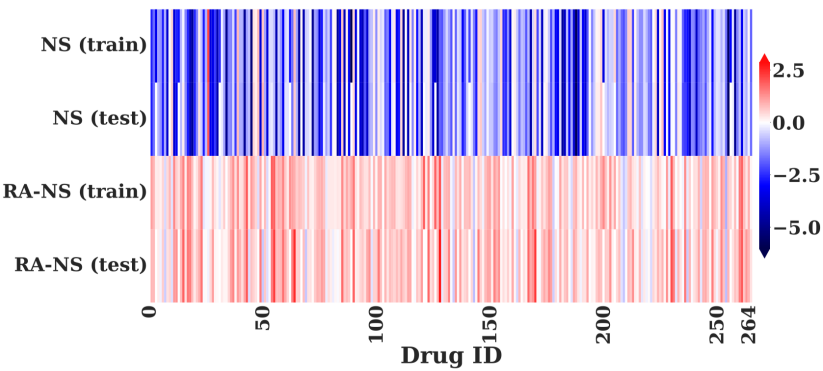


Fig. 2. Embedding similarities among drug and cellines. We subtract the similarity of a drug and its resistant cell lines from the similarity of the drug and its responsive cell lines. The results are normalized and plotted in a heatmap format. The higher (or redder) the value is, the better the embedding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

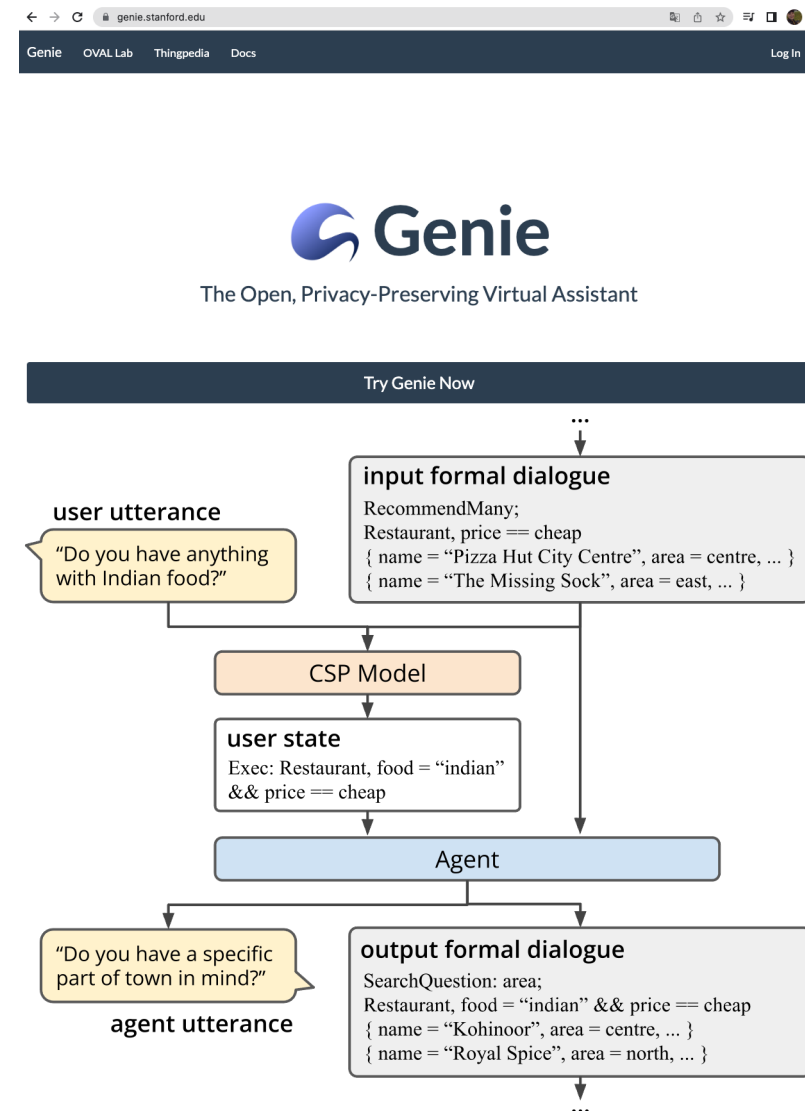
Artificial Intelligence Assistant

- AI Assistant App, Almond

- Currently, the service name has been modified to [Genie](#)
- Developed by Stanford OVAL Lab

- Training Korean Seq2SQL Model

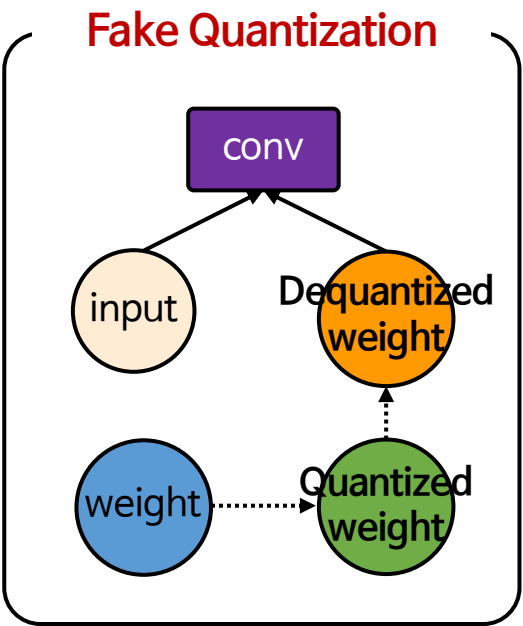
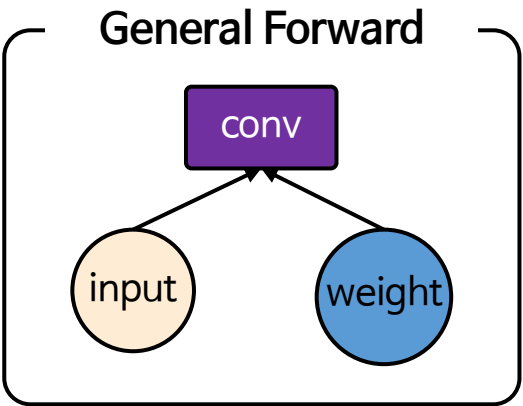
- Dataset preparation
 - Web Crawling
 - Construct templates of sentences (example of sentences)
 - Augment sentences based on templates
- Train & serve model



DNN Model Quantization

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

| | |
|----------|--|
| Problem | <div>Quantization Aware Training (Google)</div> <ul style="list-style-type: none">• Fake-quantize all of the weight matrices with a single low-bit type• Too much quantization errors occur and the trained model gets ruined <div>QuantNoise (Facebook)</div> <ul style="list-style-type: none">• Fake-quantize probabilistically selected subsets of matrices (a subset per matrix)• Trained models under-prepared for Quantization |
| Solution | <div>Fake Single Precision Training (FST)</div> <ul style="list-style-type: none">• 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 |





카카오

[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 <ul style="list-style-type: none">• 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 <ul style="list-style-type: none">• 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 | <div>[Exp-1] High matrix sparsity</div> <div>[Exp-2] Considering time bias enhanced by low traffic</div> |

Target Item (Bandit)




제네시스보다 저렴한 5천만원대 전기차 BMW i4 edrive 40

02:38 / 14:59

제네시스보다 저렴한 5천만원대 전기차 BMW i4 edrive 40

Reco. Result (Selected Arms)




테슬라도? 저라면 이거 삽니다

29:15

BMW i4 시승기, 날마다 비싸지는 테슬라 보단 이 전기차를 사겠습니다


(Arm #13)



우리나라 소비자들에게 최고의 전기차 물었더니 보인 반응

05:00


최고의 전기차는?



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


09:17

(Arm #7)




6천만원에 모하비 풀옵션 선택? 모하비가 달라졌다? 2023 모하비

10:02



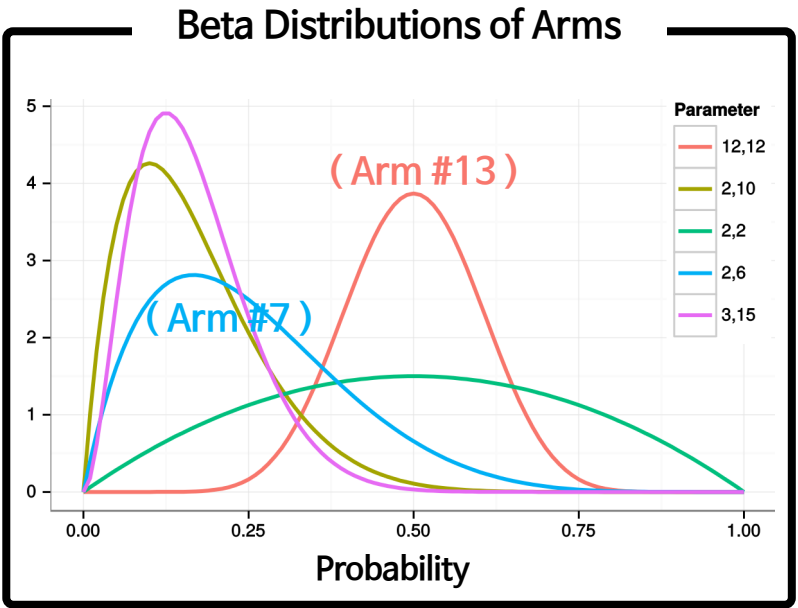
아우디 Q7 45 TDI, 이 차가 답답하면 성격이 급하신 겁니다?

22:27



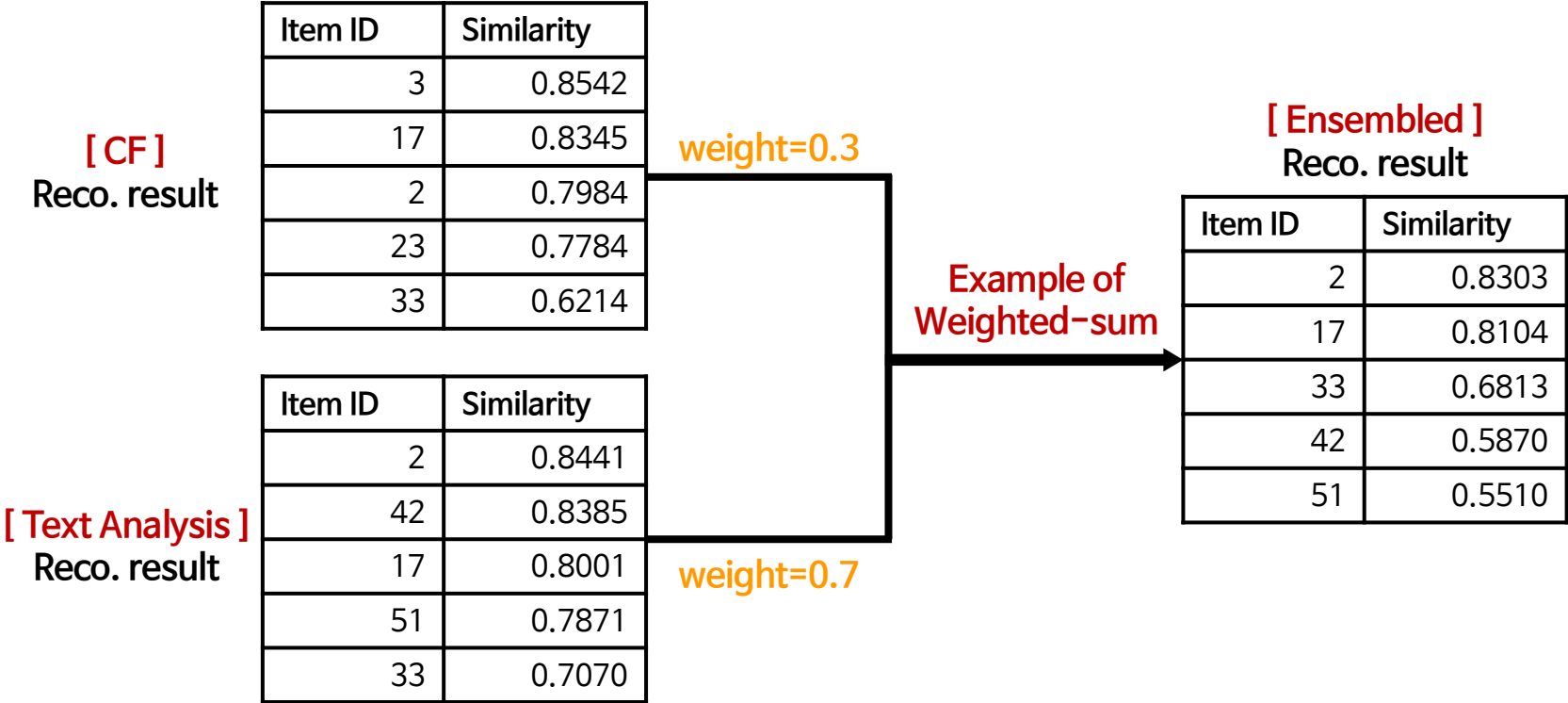
싼타페 쏘렌토 저격가능? 신형 QM6 미리보기? 르노 오스트랄 완전공개!

05:45



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 <ul style="list-style-type: none">Therefore, assumed that the composition of used models are good enough |



Automobile Video Recommendation

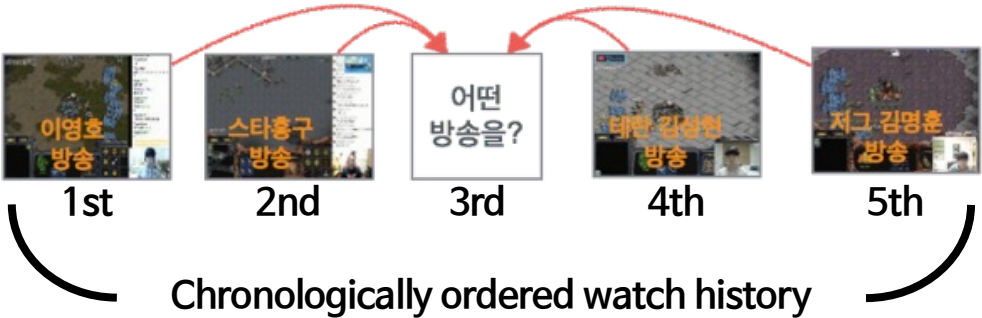
| | | | |
|----------|---|---------|--|
| Exp 5 | Item2Vec instead of Matrix Factorization | Purpose | Overcome Matrix Factorization model's limitation |
| | | Reason | Needed to generate reco. results within limited item list <ul style="list-style-type: none">The limited items rated 30~40th on avg., 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 〉

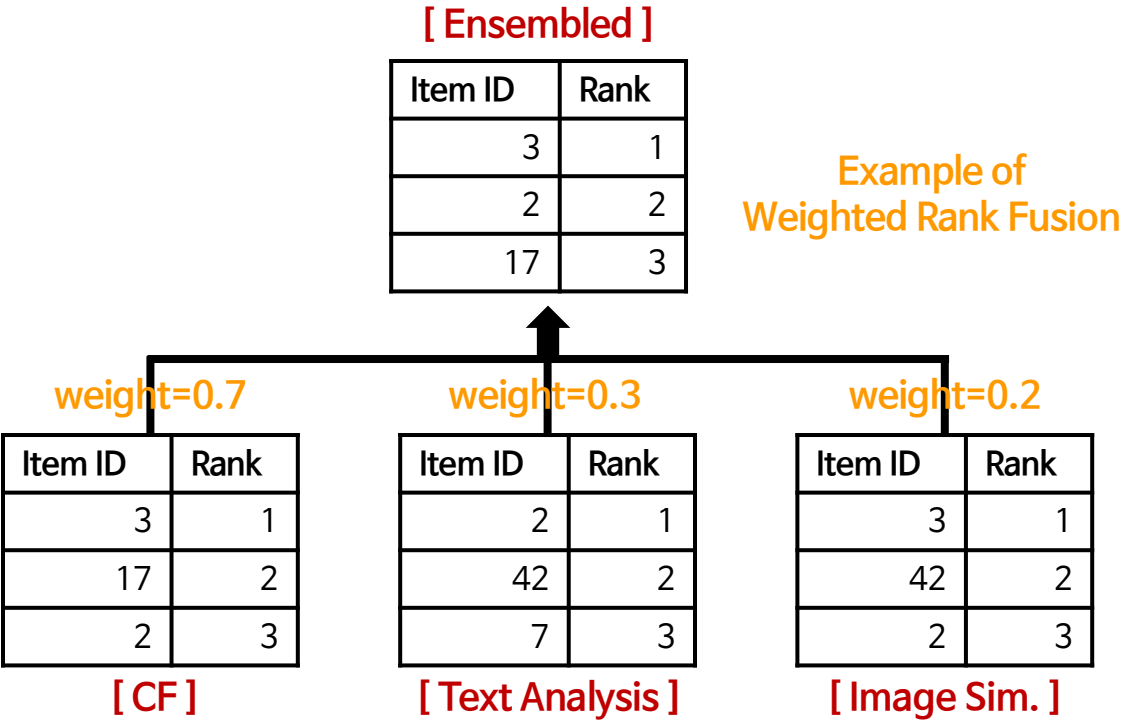


Comics Recommendation

| | | | |
|----------|--|---------|--|
| 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 |



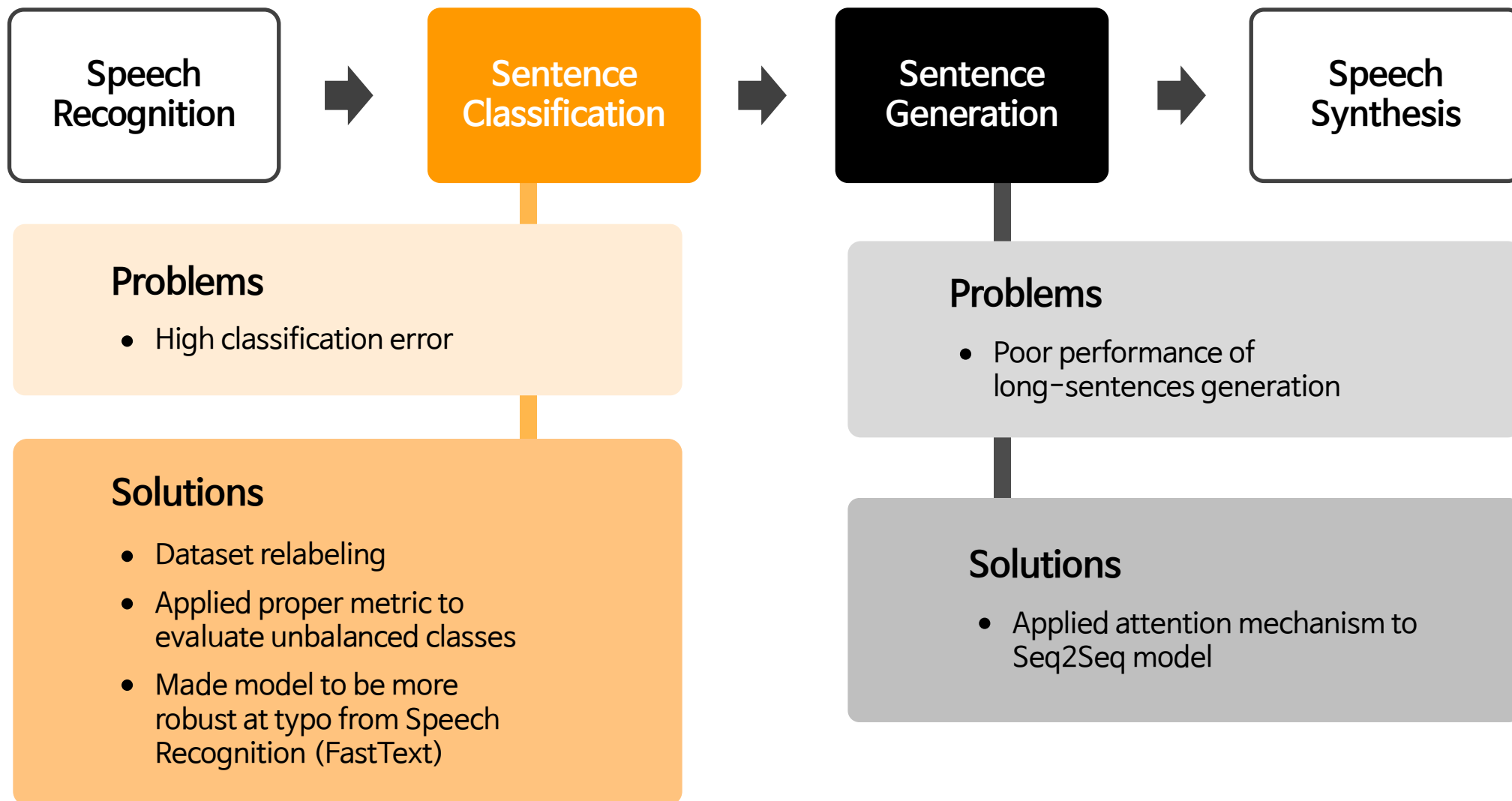


한빛소프트

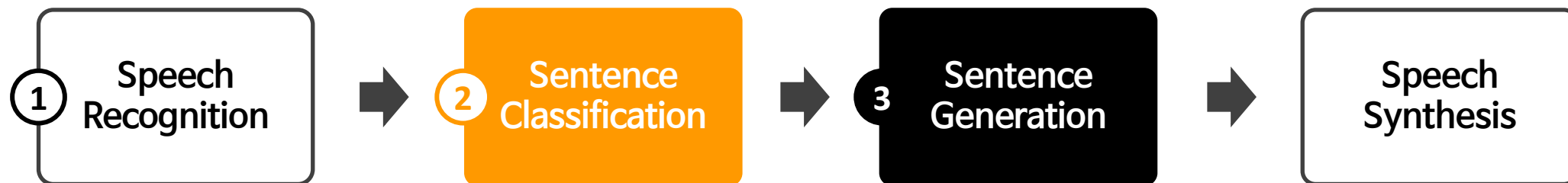
[1] (EN) Text/Audio Chatbot

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

(EN) Text/Audio Chatbot



(EN) Text/Audio Chatbot



```
>>>>> 2 OhEnglish Conversation - Domain [ MainTopic 1 : 일반 생활 ] <<<<<<<<<<
----- Say something! -----
----- HBS STT ( his table was dirty can you clean it. ) -----
1 User >> this table is dirty can you clean it
3 OE_Bot >> Sure, sorry about the mess.
<< TTS(Request)
```

(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