PROFILE

조동빈 DONGBIN CHO

Work Experience	2023.02.13 ~ 현재	SOOP (구 AfreecaTV)	데이터기술팀 / 팀원 ML Engineer
	2022.06.20 ~ 2022.12.31	Undefined	개발팀 / 팀원 Data Scientist
	2019.12.26 ~ 2020.02.29	Kakao	추천팀 / 인턴 Data Scientist
	2018.11.05 ~ 2019.04.22	한빛소프트	인공지능 파트 / 팀원 Data Scientist
Education	2020.03.01 ~ 2022.02.25	한양대학교 대학원 컴퓨터소프트웨어학과	ML System Lab. / 석사 DL Model Optimization
	2012.03.01 ~ 2018.08.31	강원대학교 산업공학과	Management of Tech. Lab. / 학부 Gamification
Publications	2022 BIB Journal (Briefings in Bioinformatics)		*RAMP: Response-Aware Multi-task Learning with Contrastive Regularization for Cancer Drug Response Prediction* (link)
	2022 ICEIC (International Conference on Electronics, Information, and Communication)		*Quantization training with two-level bit width* (link)

PROJECTS

SOOP (구 AfreecaTV)	[1] Global SOOP Reco. Pipeline	(진행중)
	[2] MLOps	1 개월
	[3] Clip & Short-form VOD Reco.	3 + 3 개월
	[4] Streamer Exploration	2 개월
	[5] Streamer Representative VOD	2 개월
	[6] VOD View Valuation	2 개월
	[7] Offline Simulation	2 개월
	[8] LIVE Broadcast Reco.	2 주

PROJECTS

Undefined	[1] Match Result Recorder	1 개월
	[2] Competition Rule Recommendation	2 개월
	[3] FAQ Chatbot	3 개월
ML System Lab.	[1] Network Embedding Generation	2 년 2 개월
	[2] DNN Model Quantization - 1	2 년
	[3] DNN Model Quantization - 2	3 개월
	[4] Artificial Intelligence Assistant	2 개월
Kakao	[1] Automobile Video Recommendation	2 개월
	[2] Comics Recommendation	2 주
HanbitSoft	[1] (KR) Multi-speaker Speech Synthesis Model	4 개월
	[2] (EN) Text Chatbot	2 개월

SOOP (구 AfreecaTV)

[1] Global SOOP Reco. Pipeline

[2] MLOps

[3] Clip & Short-form VOD Reco.

[4] Streamer Exploration

[5] Streamer Representative VOD

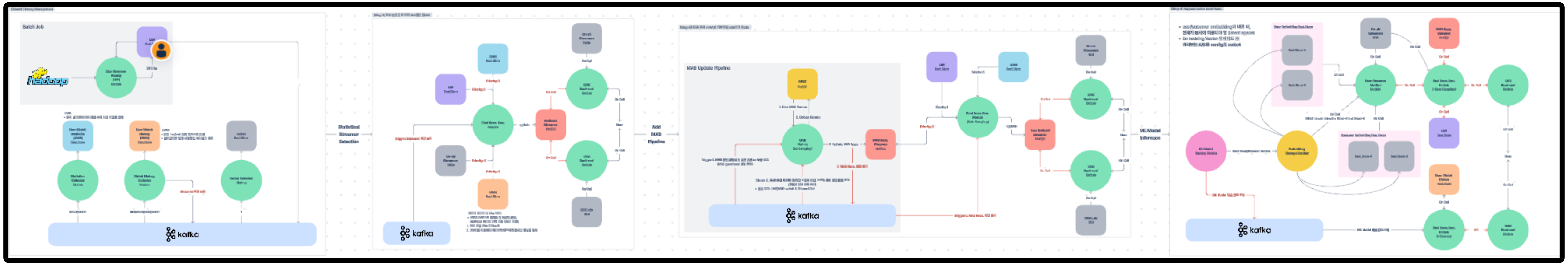
[6] VOD View Valuation

[7] Offline Simulation

[8] LIVE Broadcast Reco.

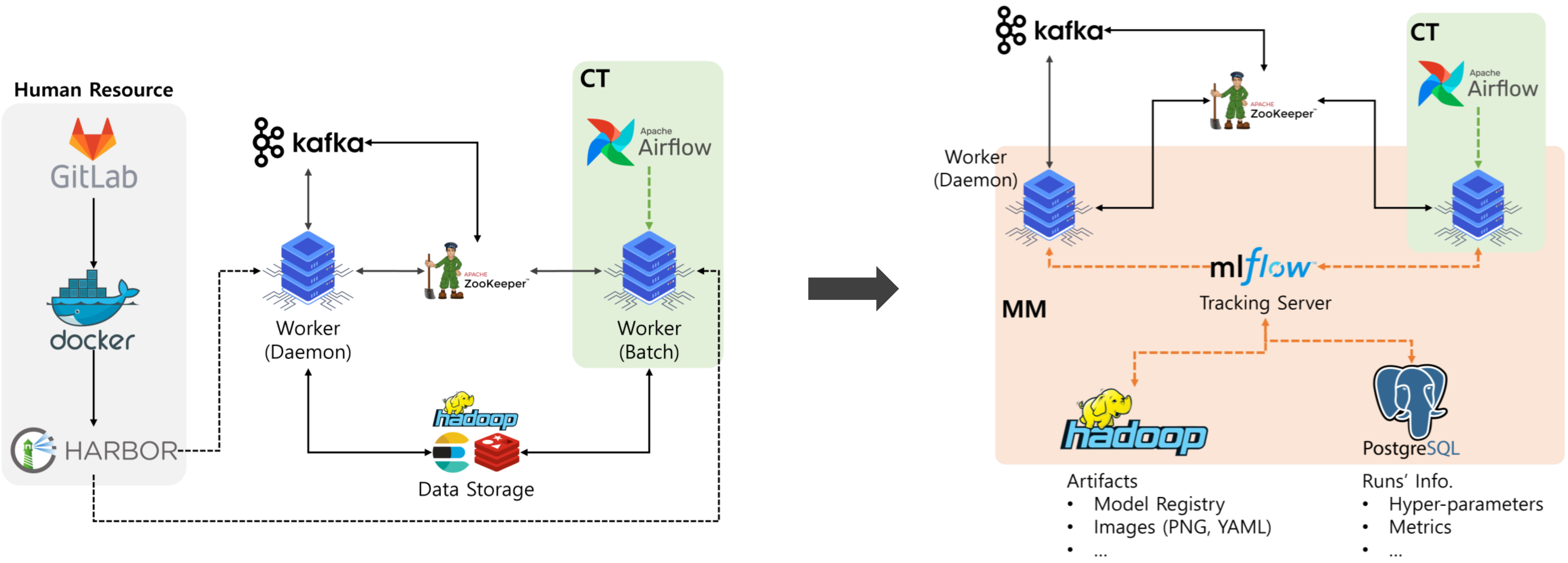
[신규 글로벌 서비스] 기반 기술 및 추천 파이프라인 개발

추천 파이프라인 개발 4 단계



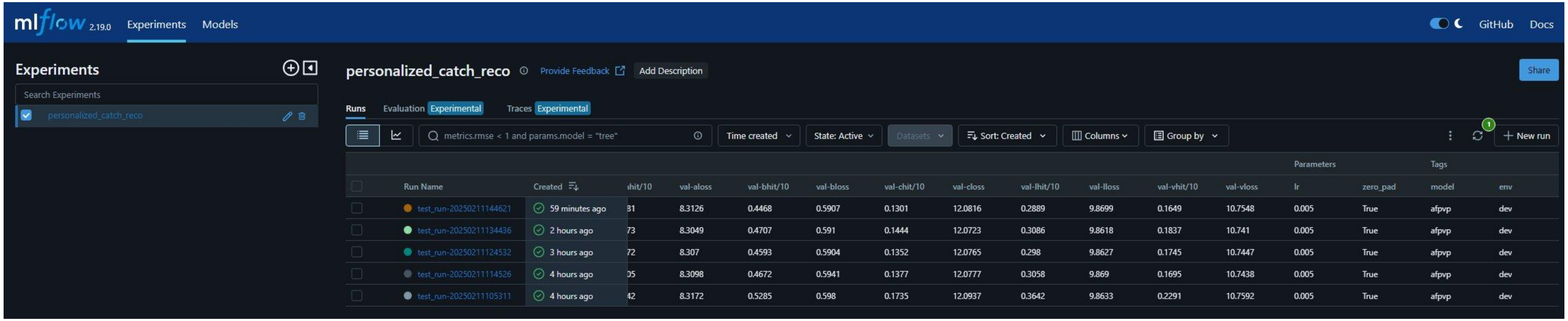
완료	- 신규 서비스 추천 파이프라인 구상 및 계획 수립 - 유저 시청 데이터 수집 및 \`선호\` 스트리머 선정 로직 개발 (Batch & Stream)
진행중	Feature Store 도입
예정	1. 선호 스트리머 기반, LIVE 방송 및 VOD 추천 2. MAB를 이용한 \`연관\` 스트리머 탐색 〉 LIVE 방송 추천 강화 3. DL 모델을 이용한 유저 별 모든 스트리머 및 VOD와의 관계 연산 및 추천
기술	AWS , k8s, Kafka, Airflow, Feast, Hive, Spark, MAB, PyTorch, ElasticSearch, Redis

MLflow 도입



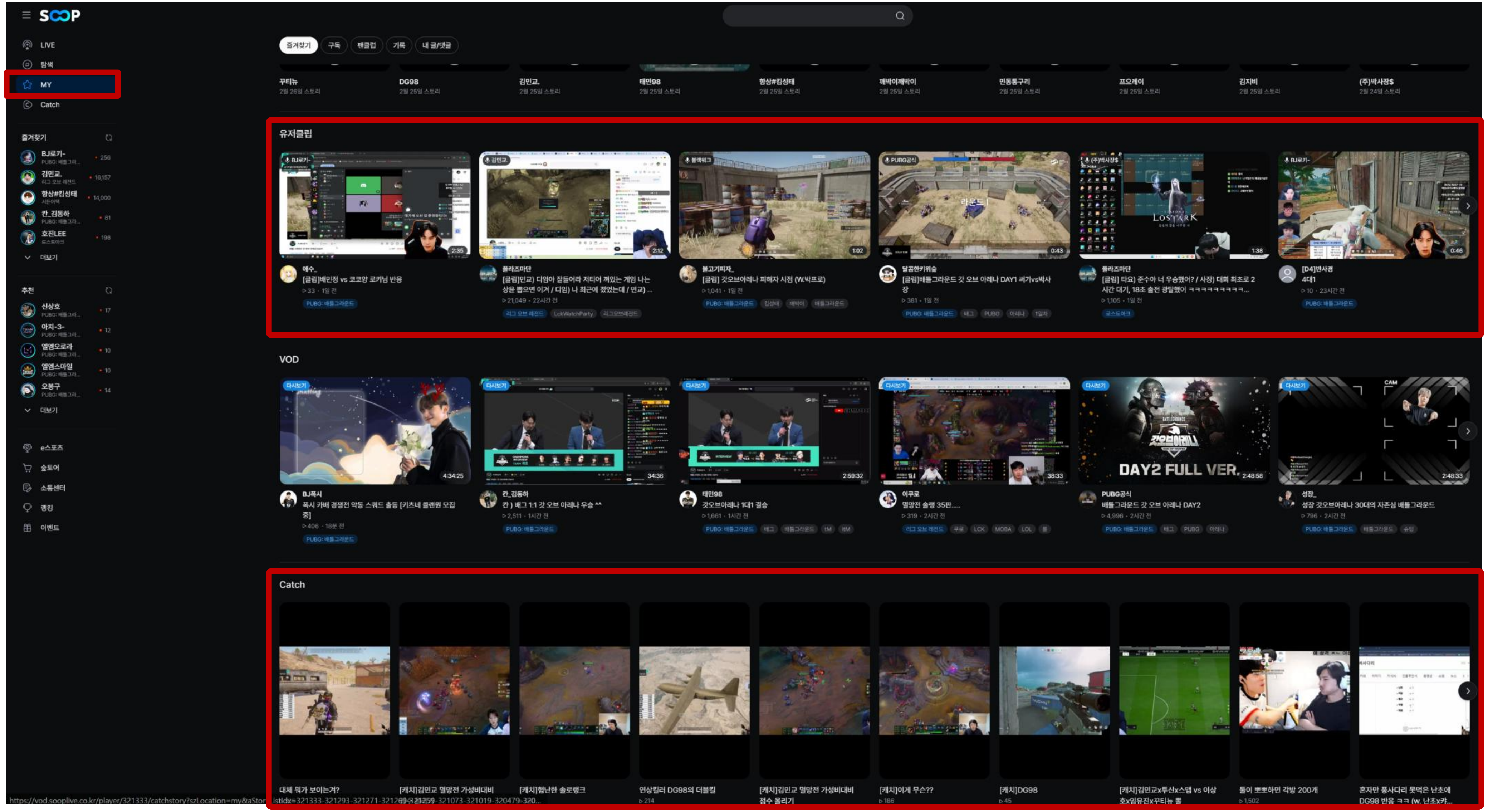
요약	MLflow 도입 및 이중화
근거	ML/DL 연구 내용 자산화 및 모델 관리 필요
기술	MLflow, nginx, HDFS, PostgreSQL, Python

MLflow 도입


mlflow 2.19.0
Experiments
Models
GitHub
Docs
Experiments
Search Experiments
personalized_catch_reco
personalized_catch_reco
Provide Feedback
Add Description
Share
Runs
Evaluation Experimental
Traces Experimental
metrics.rmse < 1 and params.model = "tree"
Time created
State: Active
Datasets
Sort: Created
Columns
Group by
New run
Parameters
Tags
Run Name
Created
val-aloss
val-bhit/10
val-bloss
val-chit/10
val-closs
val-lhit/10
val-lloss
val-vhit/10
val-vloss
lr
zero_pad
model
env
test_run-20250211144621 59 minutes ago 8.3126 0.4468 0.5907 0.1301 12.0816 0.2889 9.8699 0.1649 10.7548 0.005 True afpvp dev
test_run-20250211134436 2 hours ago 8.3049 0.4707 0.591 0.1444 12.0723 0.3086 9.8618 0.1837 10.741 0.005 True afpvp dev
test_run-20250211124532 3 hours ago 8.307 0.4593 0.5904 0.1352 12.0765 0.298 9.8627 0.1745 10.7447 0.005 True afpvp dev
test_run-20250211114526 4 hours ago 8.3098 0.4672 0.5941 0.1377 12.0777 0.3058 9.869 0.1695 10.7438 0.005 True afpvp dev
test_run-20250211105311 4 hours ago 8.3172 0.5285 0.598 0.1735 12.0937 0.3642 9.8633 0.2291 10.7592 0.005 True afpvp dev


mlflow 2.19.0
Experiments
Models
GitHub
Docs
personalized_catch_reco
test_run-20250211144621
Register model
Overview
Model metrics
System metrics
Artifacts
val
Refresh
val-aloss
val-bhit3/10
val-vloss
Step
test_run-20250211144621 (val-aloss)
test_run-20250211144621 (val-bhit3/10)
test_run-20250211144621 (val-vloss)
val-vhit/20
val-vhit/10
val-bloss

메인&즐겨찾기 페이지 Clip&Short-form 동영상 추천



메인&즐겨찾기 페이지 Clip&Short-form 동영상 추천

Model Architecture

Deep Neural Networks for YouTube Recommendations

Paul Covington, Jay Adams, Emre Sargin
Google
Mountain View, CA
{pcovington, jka, msargin}@google.com

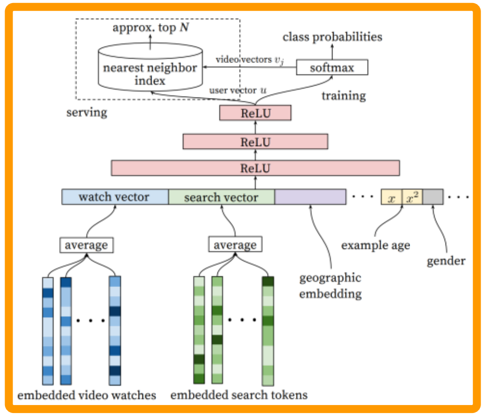
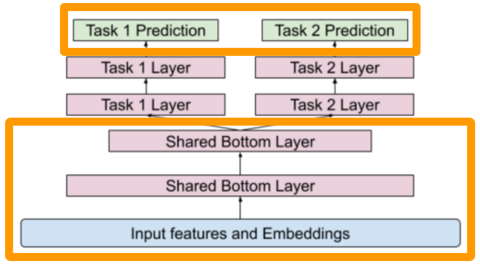


Figure 3: Deep candidate generation model architecture showing embedded sparse features concatenated with dense features. Embeddings are averaged before concatenation to transform variable sized bags of sparse IDs into fixed-width vectors suitable for input to the hidden layers. All hidden layers are fully connected. In training, a cross-entropy loss is minimized with gradient descent on the output of the sampled softmax. At serving, an approximate nearest neighbor lookup is performed to generate hundreds of candidate video recommendations.

Multi-task Loss (without task layers)

Recommending What Video to Watch Next: A Multitask Ranking System

Zhe Zhao, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sathiamoorthy, Xinyang Yi, Ed Chi
Google, Inc.
{zhezhao,lichan,liwei,jilinc,aniruddhnath,shawnandrews,aditeek,nlogn,xinyang,edchi}@google.com



(a) Shared-Bottom Model with shared bottom hidden layers and separate towers for two tasks.

Batch Softmax

Sampling-Bias-Corrected Neural Modeling for Large Corpus Item Recommendations

Xinyang Yi, Ji Yang, Lichan Hong, Derek Zhiyuan Cheng, Lukasz Heldt, Aditee Kumthekar, Zhe Zhao, Li Wei, Ed Chi
Google, Inc.
{xinyang,jiyangjy,lichan,zcheng,heldt,aditeek,zhezhao,liwei,edchi}@google.com

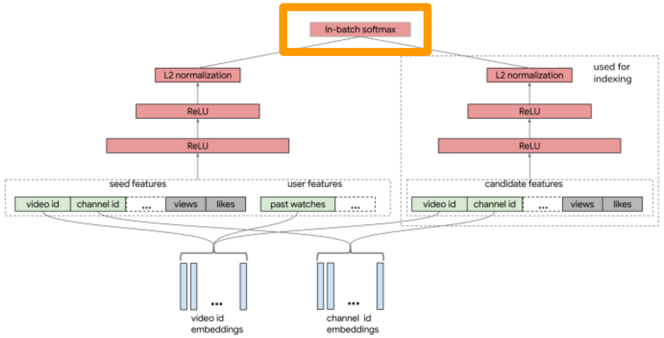
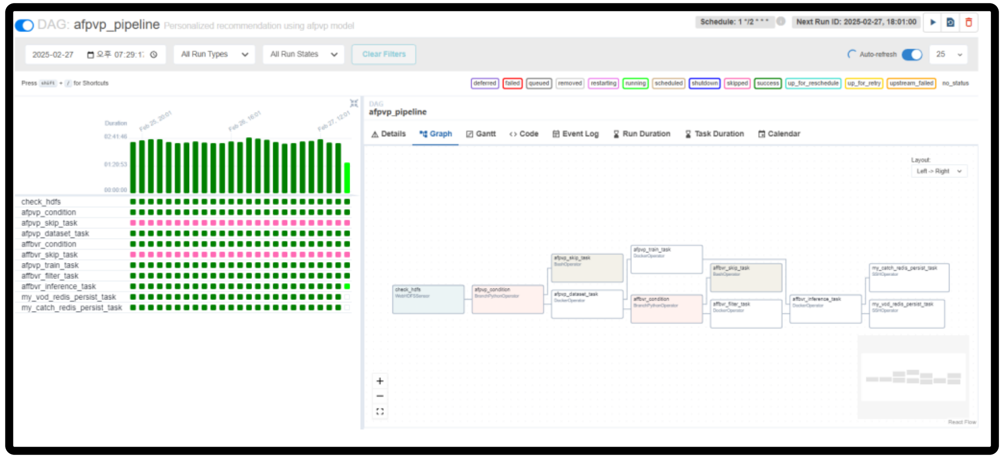


Figure 2: Illustration of the Neural Retrieval Model for YouTube.

Post-processing (preference ranking)

Streamer Selection
1. Short-term
2. Long-term

Airflow Dag Graph



요약	1. 서비스 메인 숏폼 동영상 추천 (3 개월) 〉 CTR 7% 상승 2. 즐겨찾기 페이지 클립&숏폼 동영상 추천 (3 개월) 〉 재생 20% 상승
근거	늘 보는 스트리머 위주로 시청하는 유저에게 추천 할 데이터 필요
기술	Airflow, PyTorch, Hive

스트리머 탐험 로직 개발

요약	숏폼 VOD 연속 재생 횟수 증대를 위해, 기존 개인화 추천 데이터에 유저가 시청할 만 한 다른 스트리머들의 최신/인기 VOD 추가
근거	다른 VOD 유형 대비, 숏폼을 시청할 때 비교적 다양한 스트리머의 VOD를 시청
방법	1. 네 가지 탐험 유형을 설계(e.g., 함께 시청되는 스트리머) 2. 로그에 탐험 데이터 유형을 기록하여 평가
기술	Airflow, PySpark, Hive

스트리머 별 대표 VOD 선정 로직 개발

요약	1. 스트리머와 관련된 모든 VOD들을 점수화 하여 2. 스트리머를 대표할 수 있는 n 개 VOD 선정
근거	스트리머 채널(개인 페이지) 상단에 노출 할 대표 VOD 선정 필요 (기존: 스트리머들의 방치 영역)
방법	VOD 평가를 위한 여러 통계 지표 생성 (최신성, 조회수, 양질성 등)
기술	Airflow, Hive, HBase, RDB

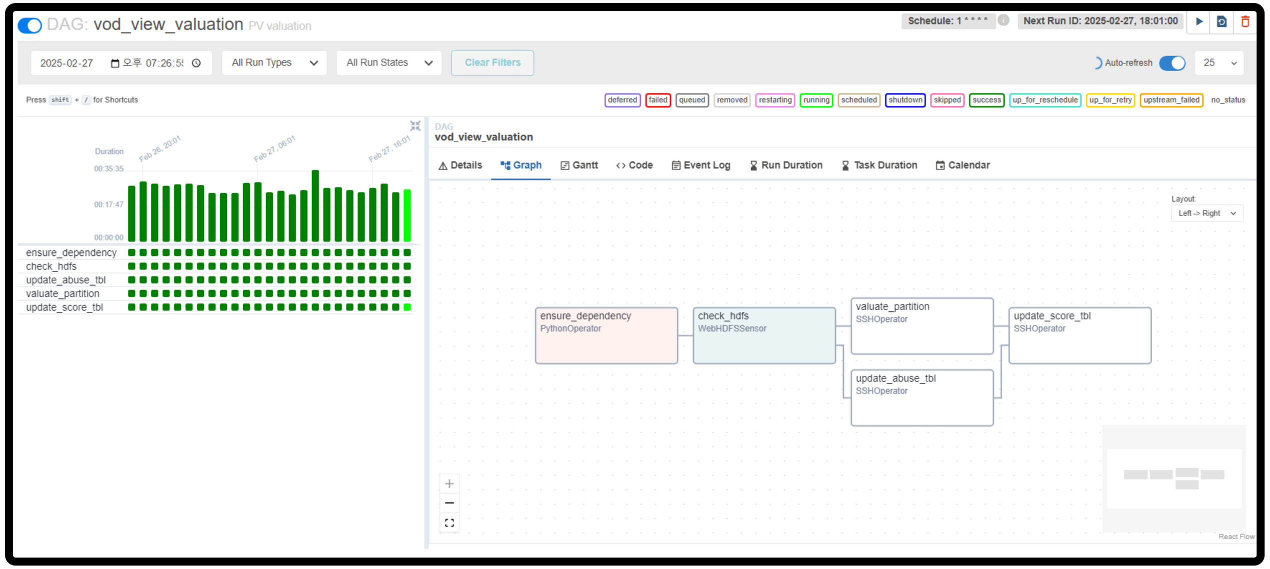
조회수 가치 평가

요약	시청 행위(초)의 가치를 절대적 수치가 아닌 상대적 비교를 통해 VOD의 품질을 분석
근거	1. 시청 행위는 implicit feedback 이므로, 시청 시간 만으로는 명확한 만족도 측정이 어려움 2. 다양한 어뷰징으로 인해 실질 지표 산출이 어려움

핵심	• SOOP의 VOD 특징: 쉽게 라이브 방송에서 클릭 몇 번으로 VOD를 생성할 수 있음
	  – VOD의 퀄리티 보장이 어려움
	• 다음을 활용하여 VOD의 품질을 예상 및 유저 만족도를 평가
	  – 유저 별 평균 시청 시간
	  – item 별 평균 시청 되는 시간

기술	Airflow, PySpark(SQL), Hive

Airflow Dag Graph

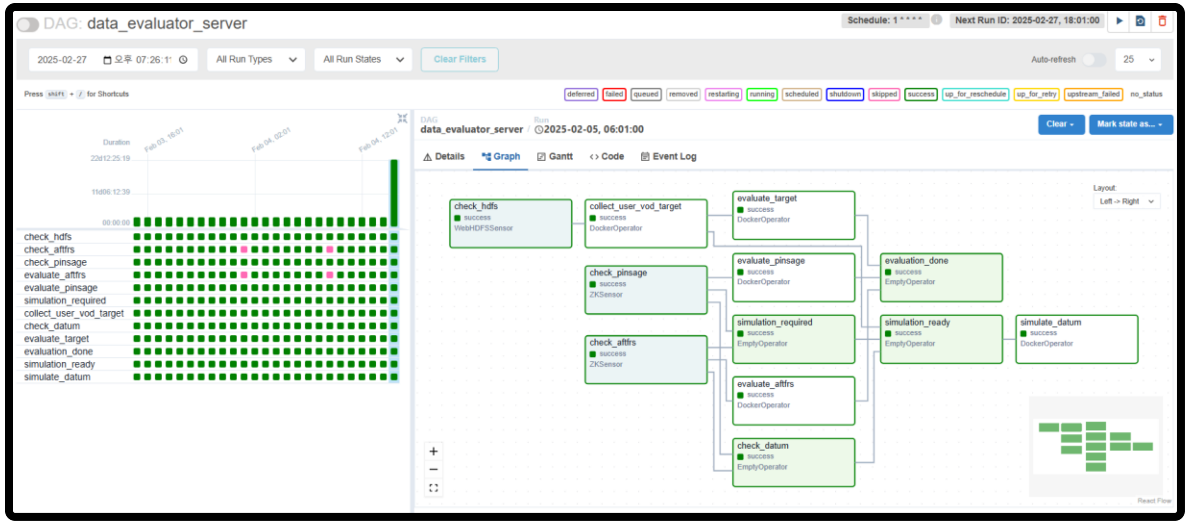


[SOOP-7] Offline Simulation

추천 데이터 오프라인 시뮬레이션 평가

요약	추천 모델들의 데이터를 offline 에서 평가 및 비교
방법	로직: 1. Batch 생성되는 같은 시간 대의 추천 모델 별 데이터 수집 2. 다음 batch의 시청 이력을 토대로 모델 별 데이터 평가 지표: - 정확성 관련 지표 (HitRate/MAP 등) - 데이터 특성 관련 지표 (Freshness/Coverage 등)
기술	Airflow, SQLite, Redash

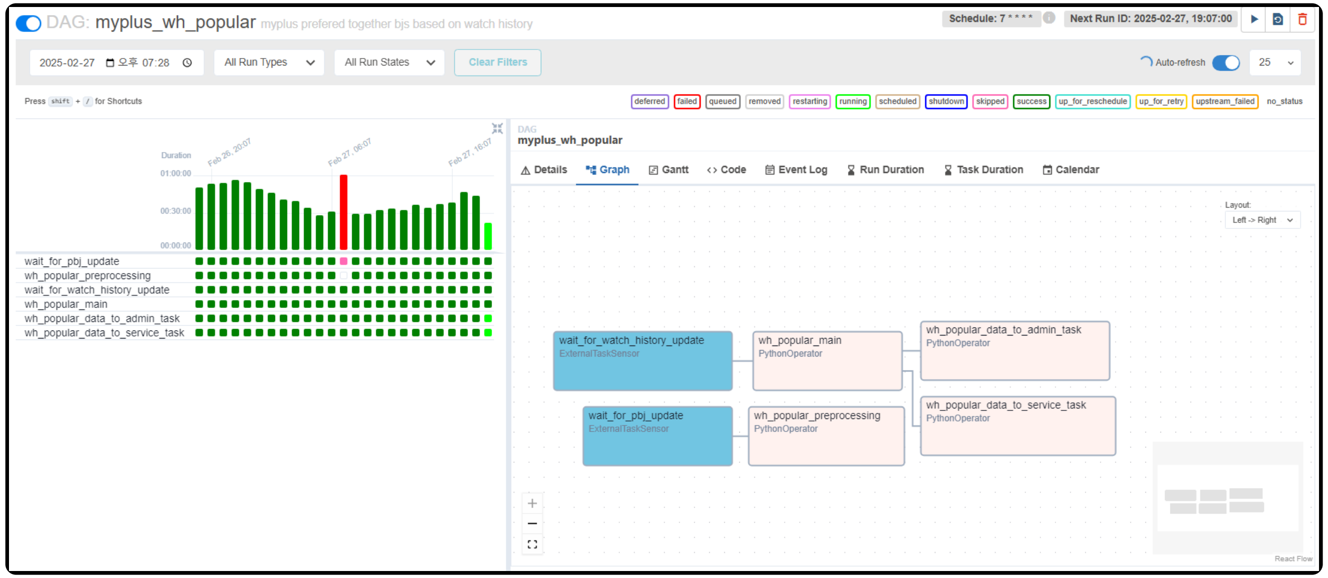
Airflow Dag Graph



라이브 방송 추천

요약	선호할 만 한 라이브 방송 추천 모듈 〉 CTR 3% 상승 및 현재 배포중
내용	- 타겟 유저: 방송 카테고리 가 아닌, 스트리머 위주의 성향을 띄는 유저 - 방송 선정: 타겟 유저가 선호하는 스트리머들을 선호하는, 다른 사람들의 선호 스트리머를 통계
기술	Airflow, HDFS

Airflow Dag Graph



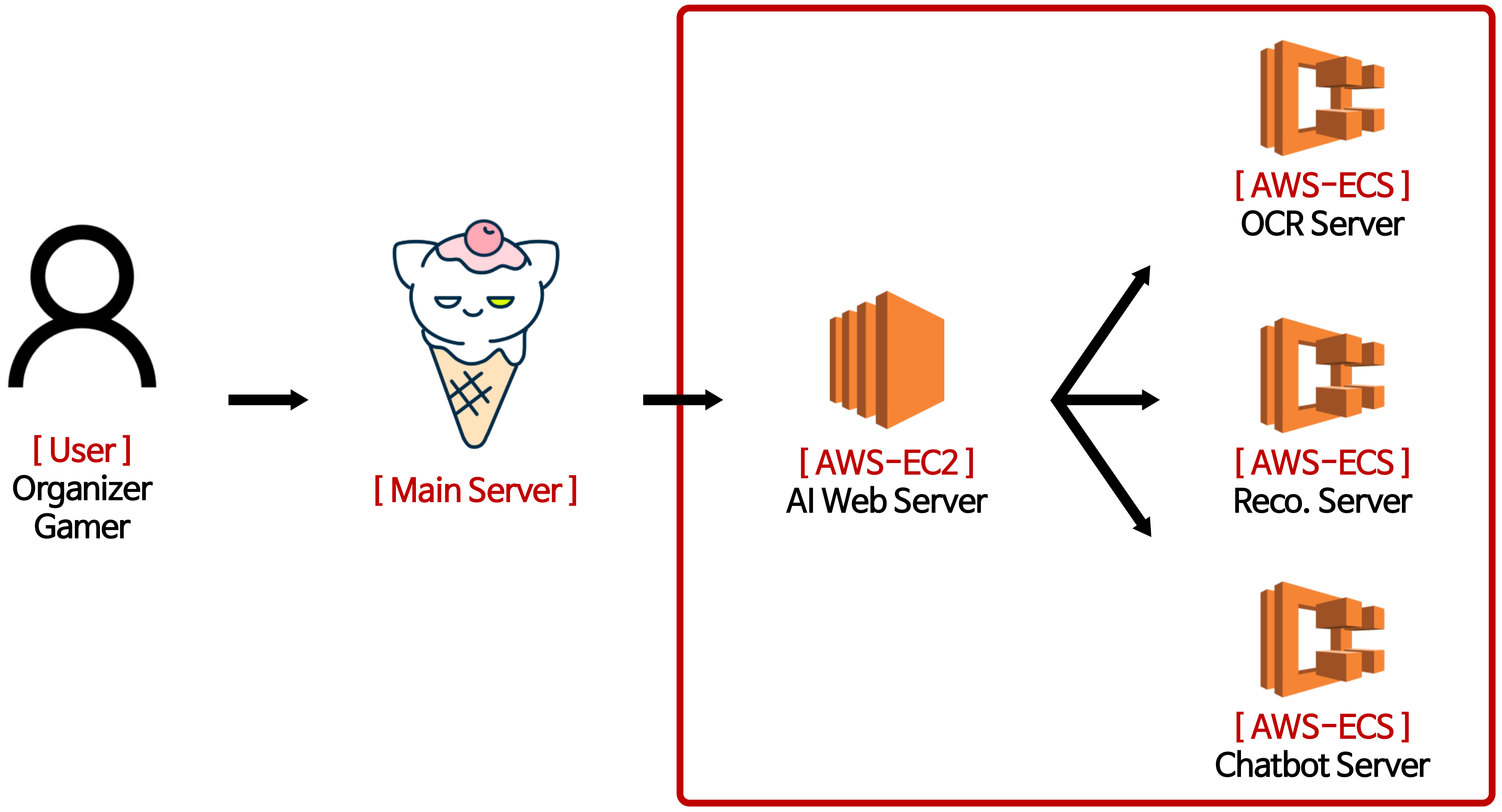
Undefined

[1] Match Result Recorder

[2] Competition Rule Recommendation

[3] FAQ Chatbot

AI Serveing Pipeline



Match Result Recorder (OCR)	Model	Tesseract (Google, LSTM-based)

	Works	• Define Problem
		• Define Pipeline
		    • Our Tesseract Model
		    • Cloud API (in case of poor confidence)
		    • Finetuning
		• Model Serving

Competition Rule Recommendation	Model	Matrix Factorization (Alternative Least Squares)
	Works	• Define Problem • EDA and Feature Selection (via Correlations) • Model Selection/Tuning • Model Optimization (remove operations) • Model Serving
Chatbot	Model	Multi-lingual BERT, StarSpace (Facebook)
	Works	• Dataset Preprocessing • Model Selection/Tuning • Model Serving

ML System Lab.

[1] Network Embedding Generation

[2] DNN Model Quantization – 1

[3] DNN Model Quantization – 2

[4] Artificial Intelligence Assistant

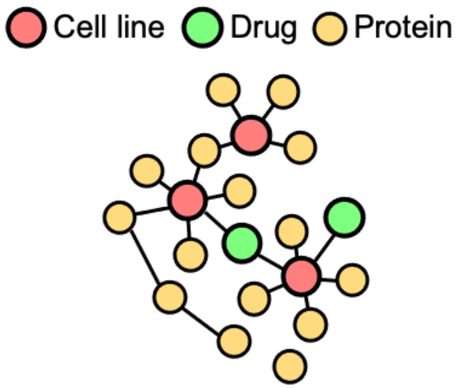
Network Embedding Generation

Paper link

Github Code link

* **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



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

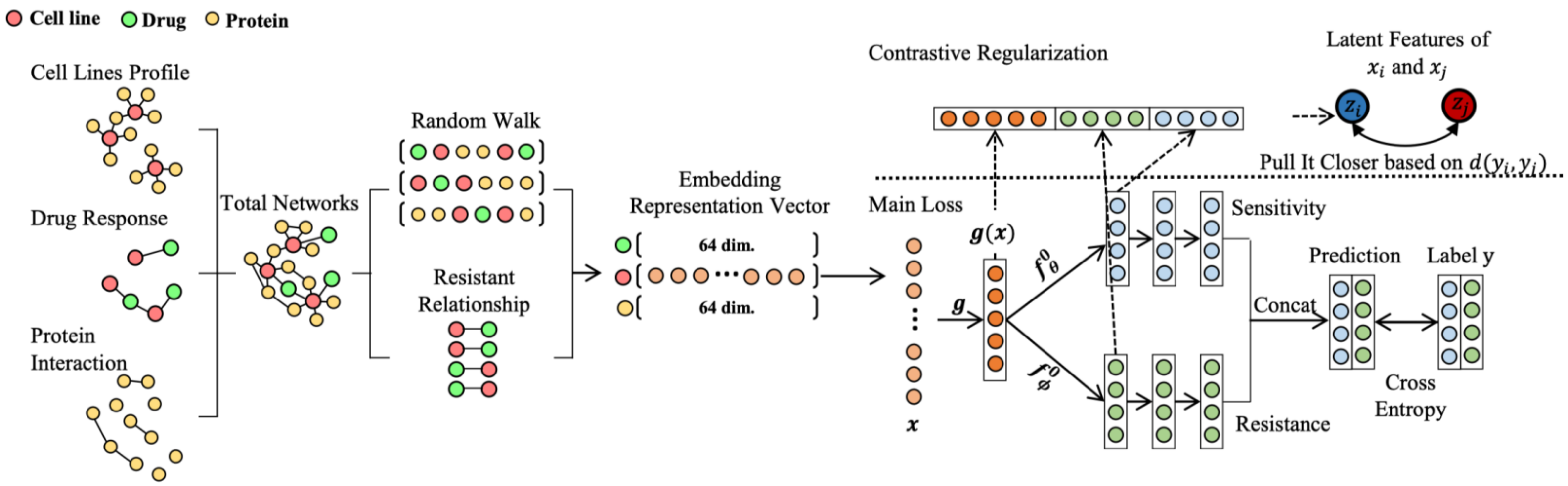


Fig. 1. The framework of RAMP. RAMP consists of two main stages. First, representation vectors are extracted from heterogeneous networks with RA-NS. Second, the multitask architecture of a Bayesian neural network is trained by representation vectors with contrastive regularization.

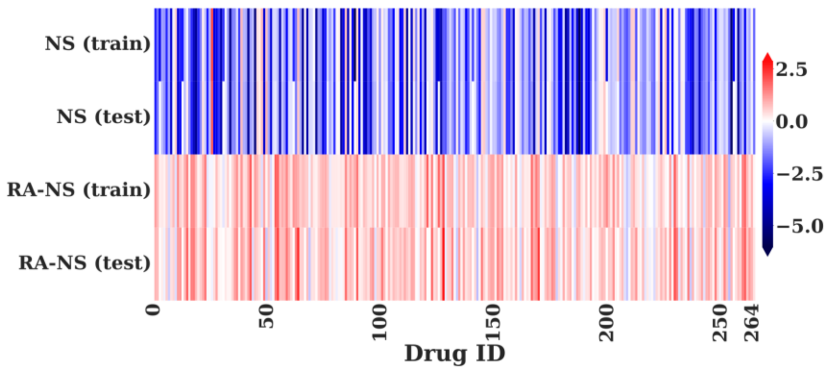


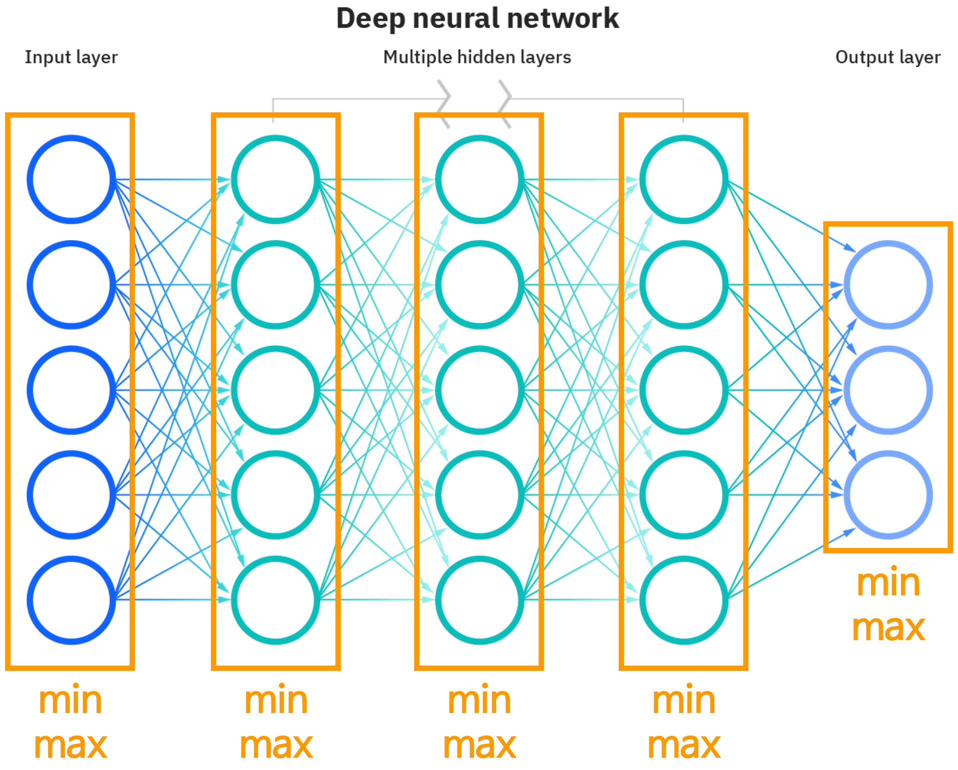
Fig. 2. Embedding similarities among drug and cellines. We subtract the similarity of a drug and its resistant cell lines from the similarity of the drug and its responsive cell lines. The results are normalized and plotted in a heatmap format. The higher (or redder) the value is, the better the embedding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

Github Code link

DNN Model Quantization – 1

Definition	What is Quantization	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	Poor Performance	Quantized models' **performance (e.g., accuracy) drops catastrophically when using sub-8bit INT type**
	Why	**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

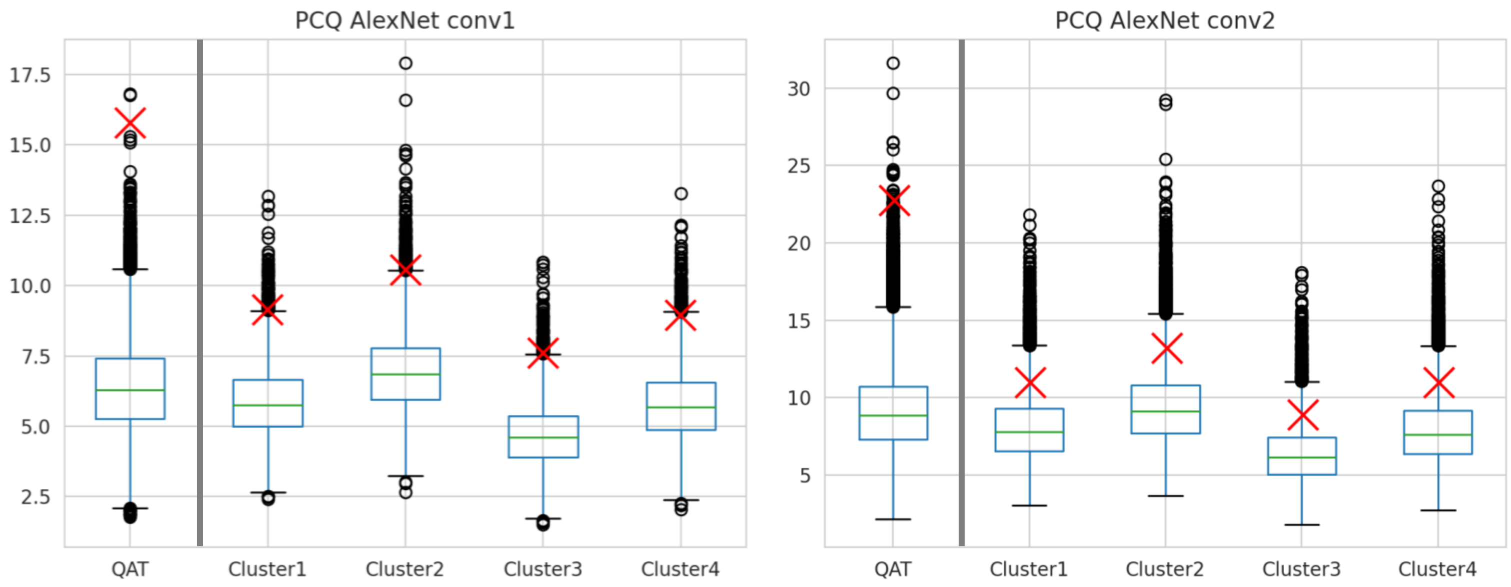
Granular Exponential Moving Average (Granular EMA)

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

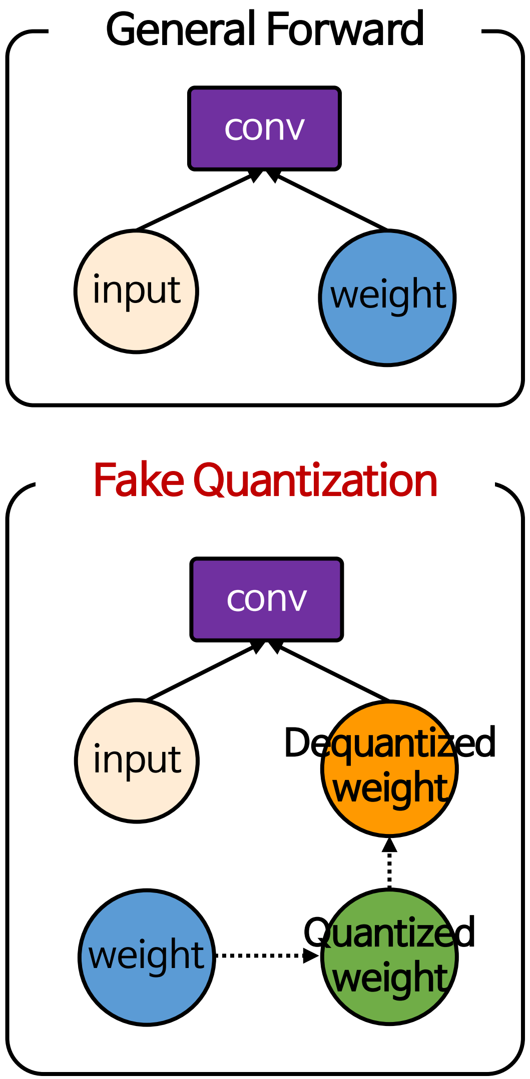
- Shows that our method
  - how efficiently exclude outliers
  - how to work with clusters
- QAT : Baseline (Google)
- Cluster* : Ours
- X : Trained maximum value
- Box-plots : Actual max values per image

Paper link

DNN Model Quantization – 2

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

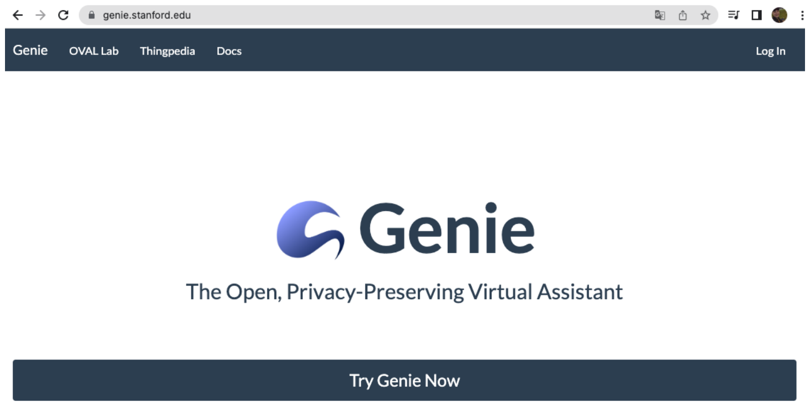
Problem	**Quantization Aware Training** (Google) • Fake-quantize all of the weight matrices with a single low-bit type • **Too much quantization errors** occur and the trained model gets ruined
	QuantNoise (Facebook) • Fake-quantize probabilistically selected subsets of matrices (a subset per matrix) • 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

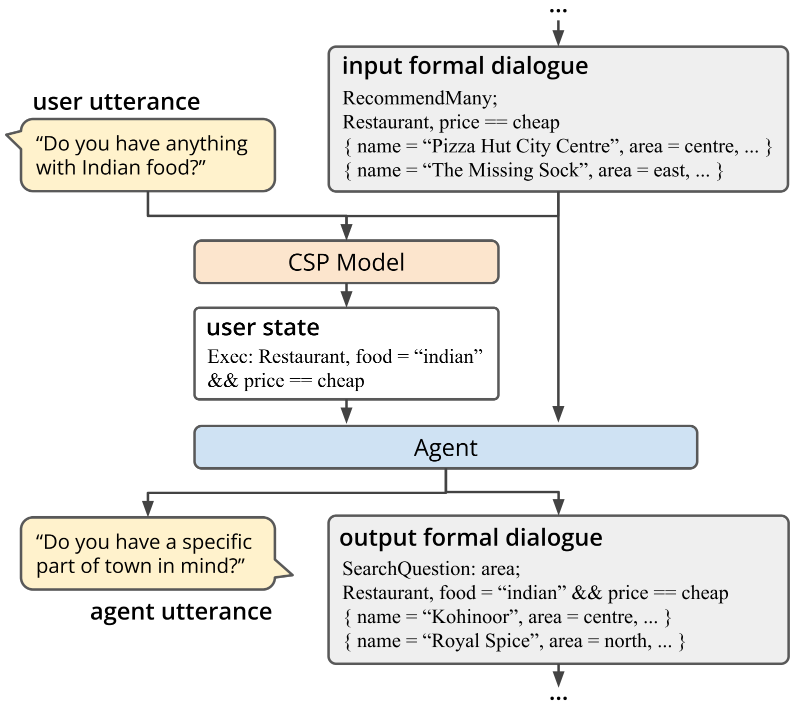


Artificial Intelligence Assistant

- 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





Kakao

[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** (RRF to Weighted-sum)	Purpose	Searching the key model among ensembled models
		Reason	Other well performing services had been used **similar model combination** • 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

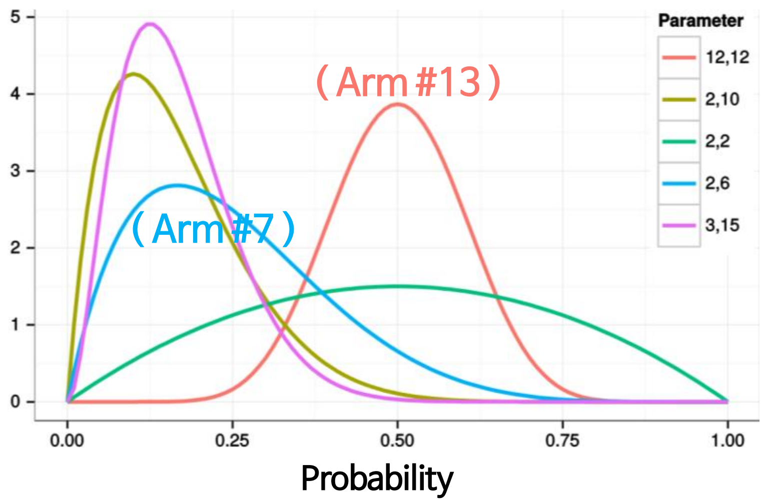
Exp 1, 2	Thomson Sampling h-params tuning	Purpose	Adjustment of trade-off between exploration & exploitation
		Reason	[Exp-1] High **matrix sparsity**
			[Exp-2] Considering **time bias** enhanced by low traffic

Target Item (Bandit)

제네시스보다 저렴한 5천만원대 전기차 BMW i4 edrive 40

Reco. Result (Selected Arms)

Beta Distributions of Arms



Automobile Video Recommendation

Exp 3, 4	Ranking Algorithm (RRF to Weighted-sum)		
		Purpose	Searching the key model among ensembled models
		Reason	Other well performing services had been used **similar model combination** • 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

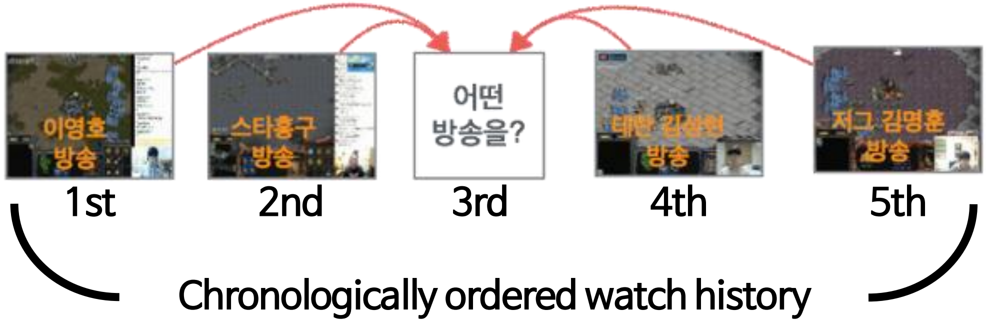
Item ID	Similarity
2	0.8303
17	0.8104
33	0.6813
42	0.5870
51	0.5510

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 〉

〈 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.]

HanbitSoft

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

[2] (EN) Text/Audio Chatbot

(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

(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