DEVIEW 2021

Look-alike Modeling and Serving:
비슷한 사람을 찾아주세요

정찬주, 이소현 NAVER Biz AI Dev

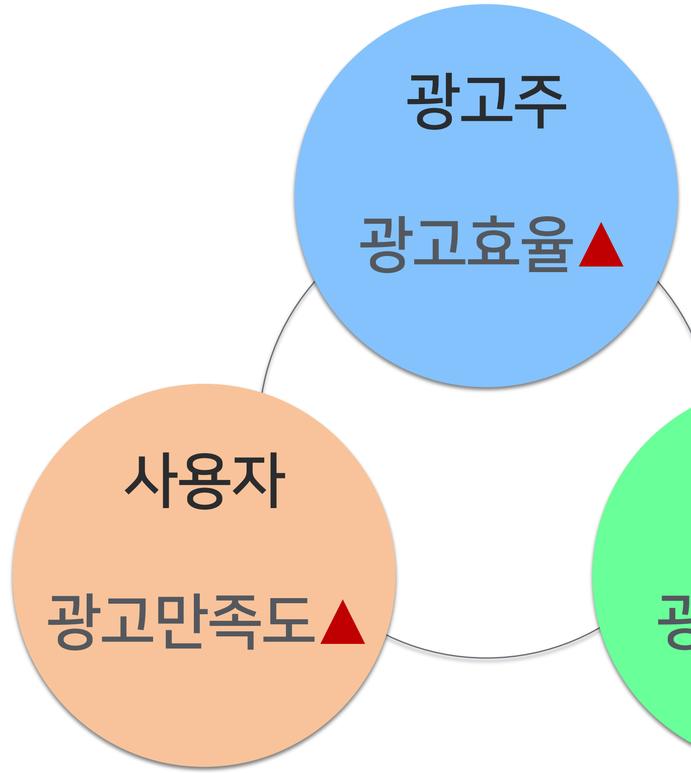
CONTENTS

1.Look-alike

1.1. 광고주의 요구

광고주는 광고에 맞는 유저를 원한다

광고주, 사용자, 플랫폼 모두의 요구에 부합



해당 광고에 관심있고
반응할 사용자에게만 송출

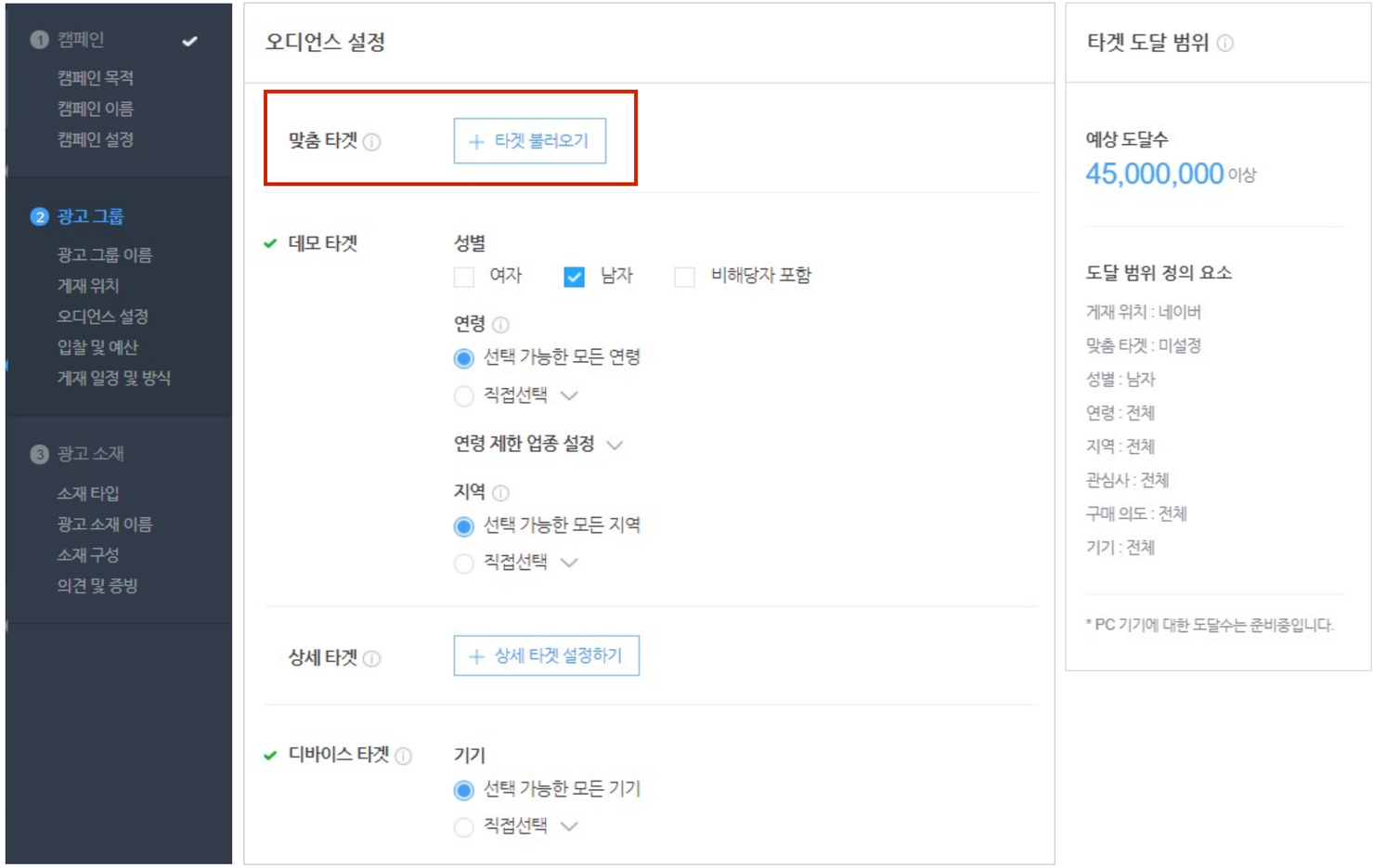
관심 없는 광고 미 송출

불필요 광고 송출 축소
추가 광고 슬롯 확보

1.2. 오디언스 기능

광고주센터 > 광고 그룹

- 오디언스 설정 제공



1.3. 유사 타겟(Look-alike)

DEVIEW 2021

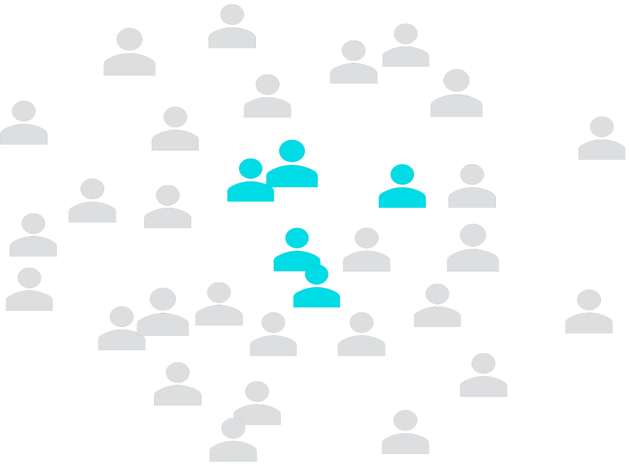
자신에게 더 맞는 오디언스를 원하는 광고주를 위해
유사 타겟 기능을 제공

예) 광고주 사이트 방문자와 비슷한 유저
이전 광고 반응자와 비슷한 유저

1.3. 유사 타겟(Look-alike)

1. 광고주 소유 오디언스(seed)를 이용

2. 이와 유사한 사람들을 찾고

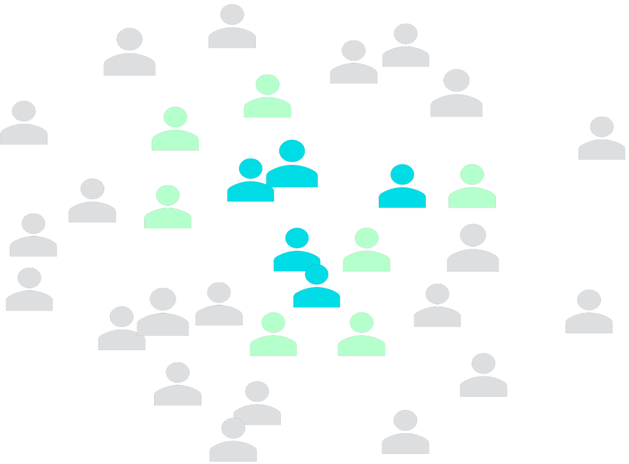


3. 이 대상으로 광고를 수행

1.3. 유사 타겟(Look-alike)

1. 광고주 소유 오디언스(seed)를 이용

2. 이와 유사한 사람들을 찾고



3. 이 대상으로 광고를 수행

1.3. 유사 타겟(Look-alike)

요구사항 분석

단순 광고주 오디언스와 유사한 사람 ×

광고 효율 ▲ (전환)

CVR ▲

하지만 CTR, 노출 수도 높았으면

노출 (IMP): 광고가 사용자에게 보여지는 행위

클릭: 광고 플랫폼의 주 매출 수단 (클릭 당 과금 모델)

전환: 구매, 사이트 가입, 장바구니 담기 등 사용자의 의미 있는 행위

CTR: Click through rate (클릭 수 / 노출 수)

CVR: conversion rate (전환 수 / 클릭 수)

CTCVR: CTR * CVR (전환 수 / 노출 수)

1.3. 유사 타겟(Look-alike)

현실적인 문제

Seed는 광고주가 소유

기계학습에서 가장 중요한 것은 정답 셋
하지만 이 seed는 우리에게 없다

개발 시 모델 수행 결과를 측정할 방법 필요

우리가 만든 seed에서 이전 목표들을 달성하면 좋은 모델로 인정

몰(스마트 스토어) 방문자 seed, 상세 상품 페이지 조회자 seed 사용

- 오프라인 테스트

1.3. 유사 타겟(Look-alike)

목표

성능

요청 사용자(seed)와 비슷한 새로운 사용자 셋 추출
더불어 CTCVR, CVR, CTR이 높아야 한다

처리량

1억 이상의 유저(브라우저 별 익명화 된) 에 대해 유사도 계산을 10분 이내

CTCVR: CTR * CVR (전환 수 / 노출 수)
CVR: conversion rate (전환 수 / 클릭 수)
CTR: Click through rate (클릭 수 / 노출 수)

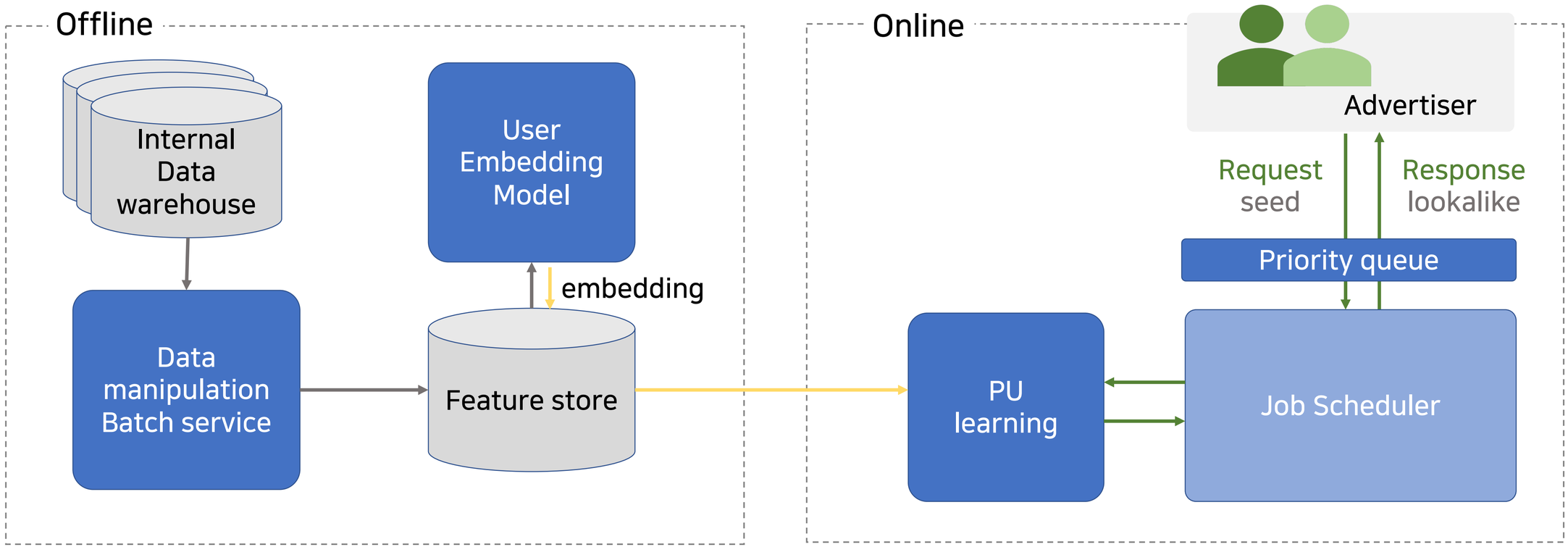
2.Model Architecture

2.1. 전체 구조

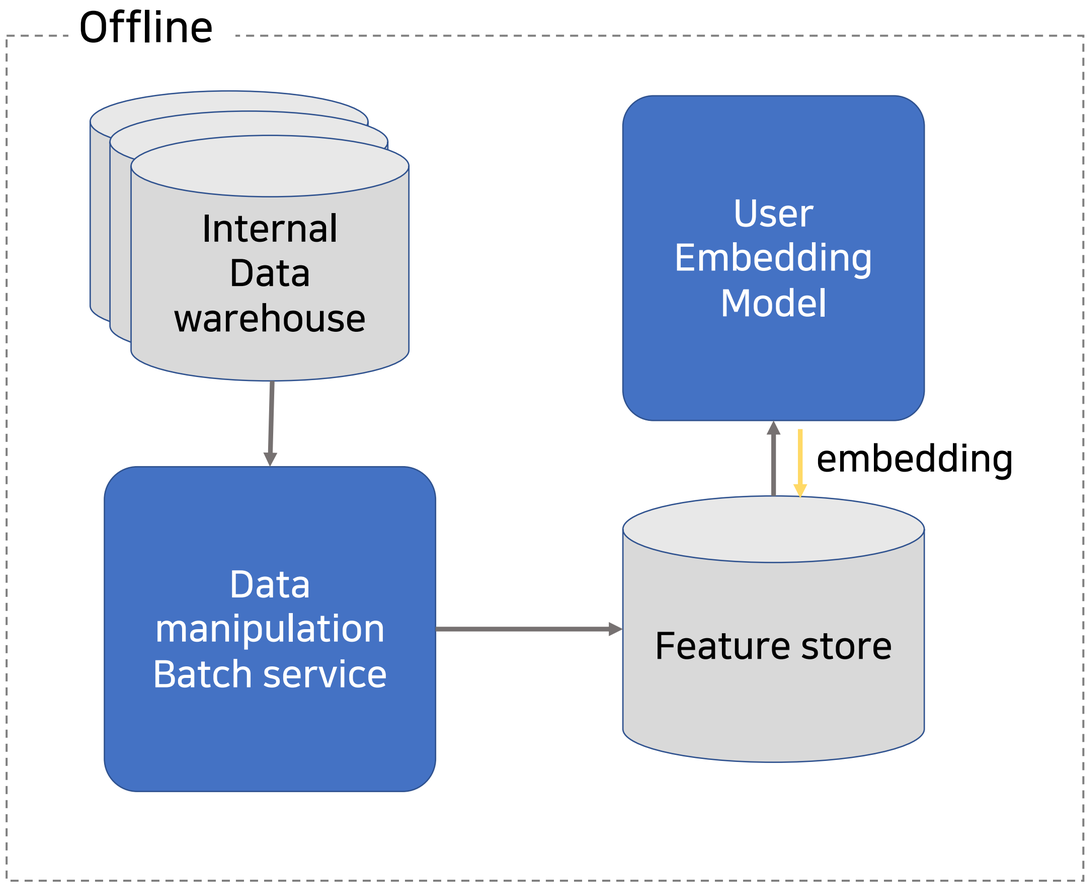
파트 분리

실시간 계산량을 줄여 빠른 생성을 대응하기 위해

Real-time Attention Based Look-alike Model for Recommender System[1]의 아이디어 차용



2.1.1. Offline part



사용자 행태 학습 후 embedding 저장

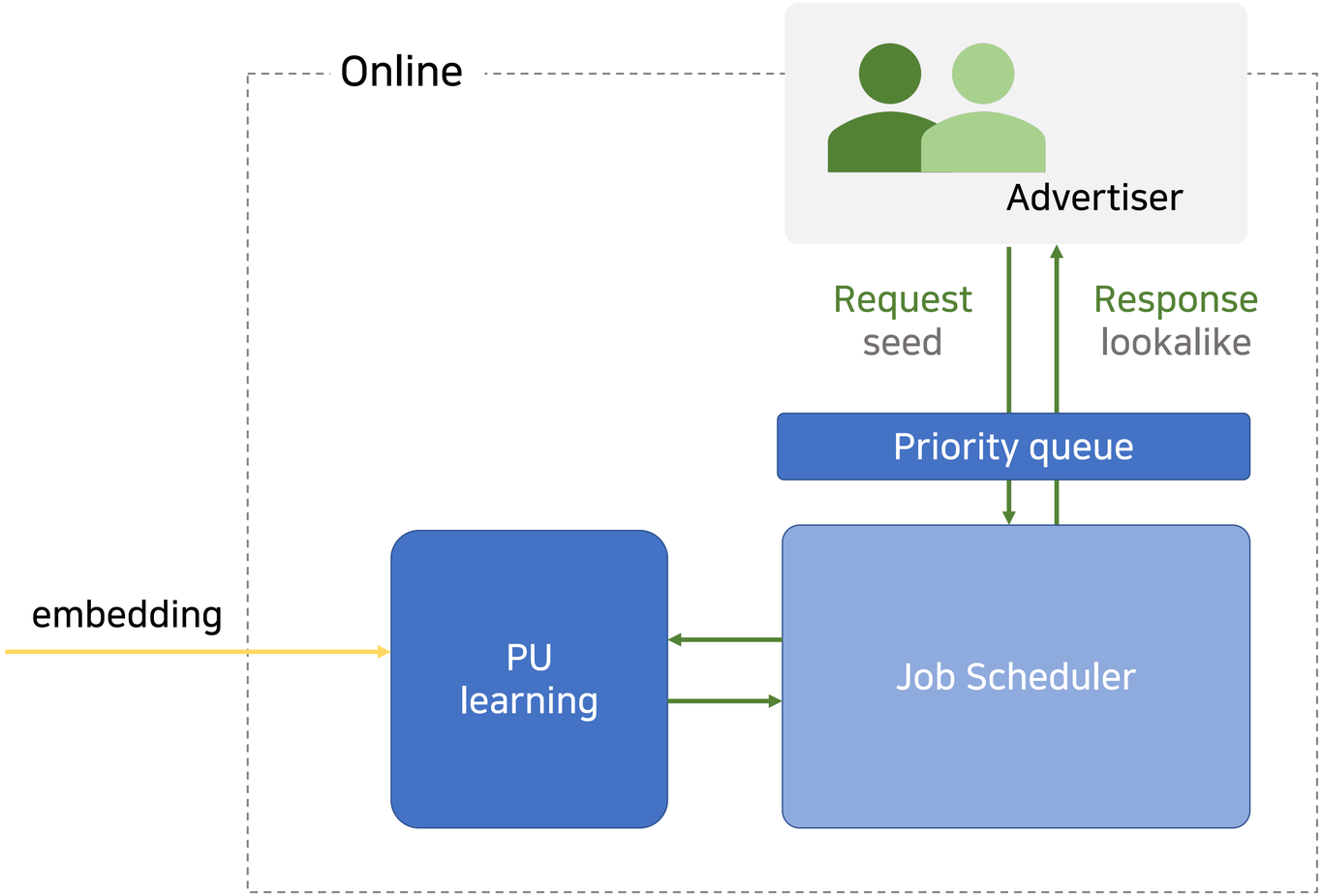
충분한 계산 시간을 사용

2.1.2. Online part

저장된 embedding 이용

실시간 요청에

Look-alike learning ▸ 응답



3.Offline Part:

User Embedding

3.1. Offline Part 목표

DEVIEW 2021

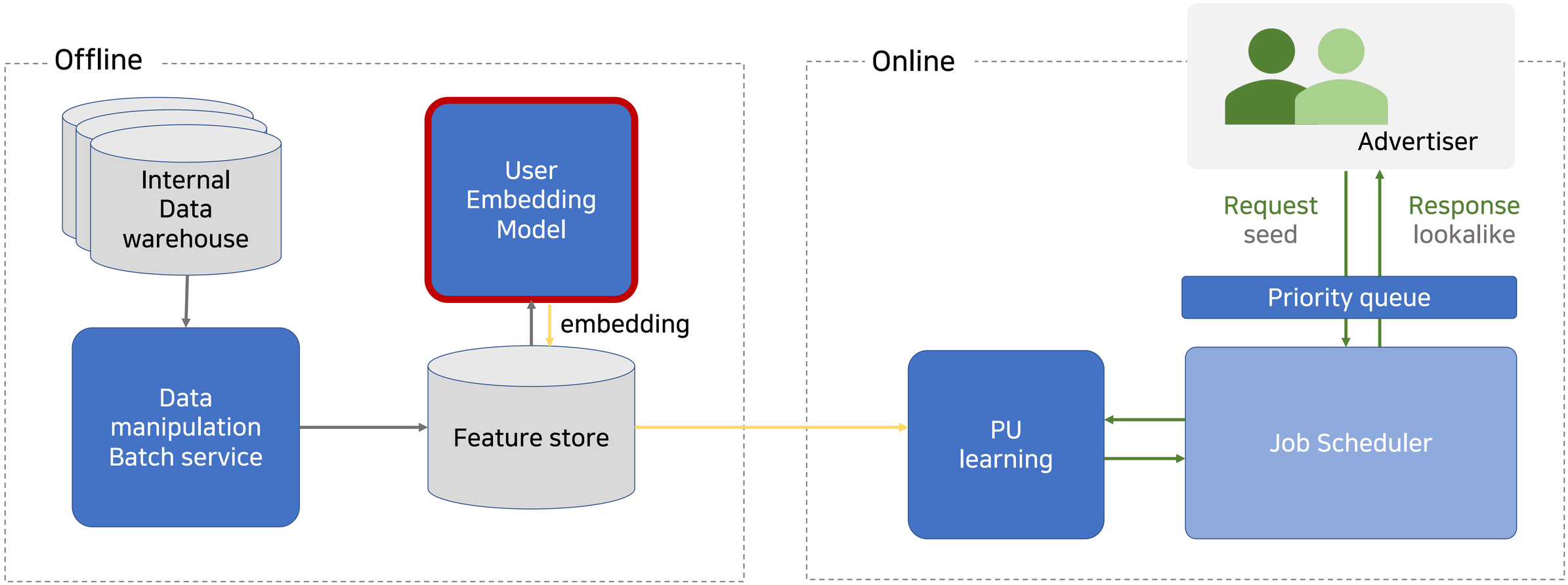
User embedding vector 생성

어떤 **특징**을 추출해야 할까?

광고주 의도에 따라 주요 목표(노출 수, CTR, CVR, ROAS 등)가 달라지게 되나,
Look-alike 사용 광고주의 목표: CTCVR, CVR

따라서 **전환 행태**에 초점

3.2. User Embedding Model



3.2. User Embedding Model

DEVIEW 2021

Multi-class classification task (참고 논문: [1])

Input X: 익명화 된 사용자의 행동 데이터

Label Y: 상품 카테고리 구매 여부

카테고리 체계: IAB audience taxonomy - purchase intent [2]

카테고리 수: 약 600개

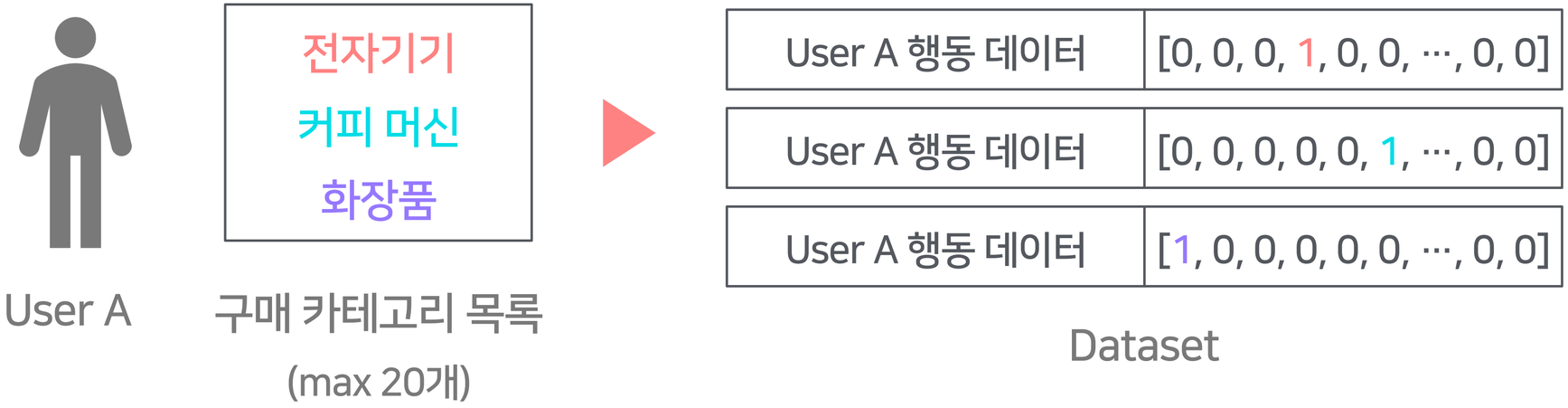
iab. TECH LAB

IAB (Interactive Advertising Bureau)

온라인 광고 산업의 표준 개발, 연구 및 법률 지원을 제공하는 광고 전문기구

3.2. User Embedding Model

Dataset 생성



3.2. User Embedding Model

Dataset 생성: Negative sampling

Input X	Label Y (약 600개)
User A 행동 데이터	[0, 1, 0, 0, 0, 0, ⋯, 0, 0]
User A 행동 데이터	[0, 0, 0, 1, 0, 0, ⋯, 0, 0]
User A 행동 데이터	[1, 0, 0, 0, 0, 0, ⋯, 0, 0]

▶

Input X	Sampled Label Y (1+10개)
User A 행동 데이터	[0, 1, 0, 0, ⋯, 0]
User A 행동 데이터	[0, 0, 1, 0, ⋯, 0]
User A 행동 데이터	[1, 0, 0, 0, ⋯, 0]

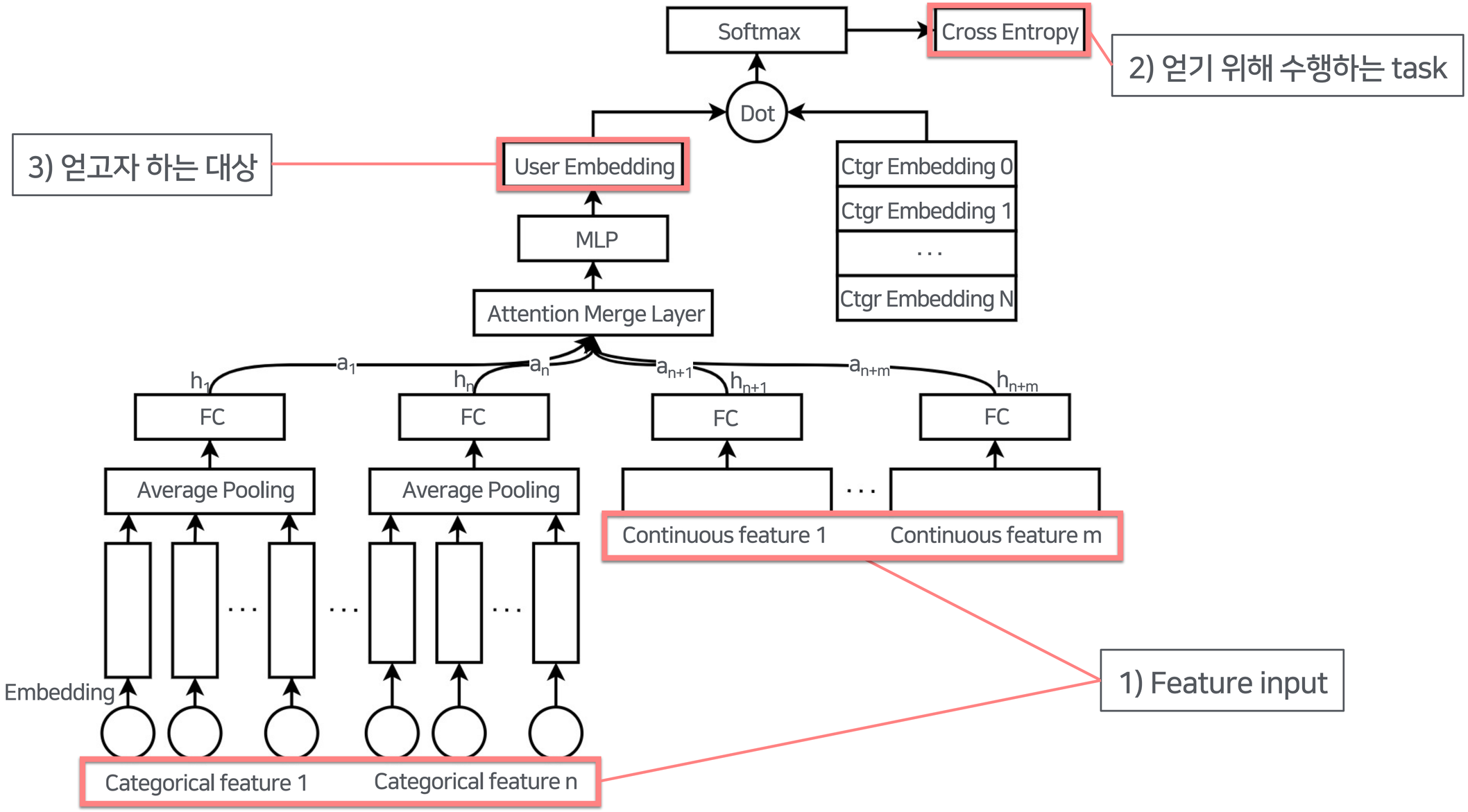
- 1:10 negative sampling
- Sampling distribution: approximately log-uniform distribution

$$p(x_i) = \frac{\log(k+2) - \log(k+1)}{\log(D+1)}$$

(k: 카테고리 x_i 등장 빈도 순위, D: 전체 카테고리 수)

3.2. User Embedding Model

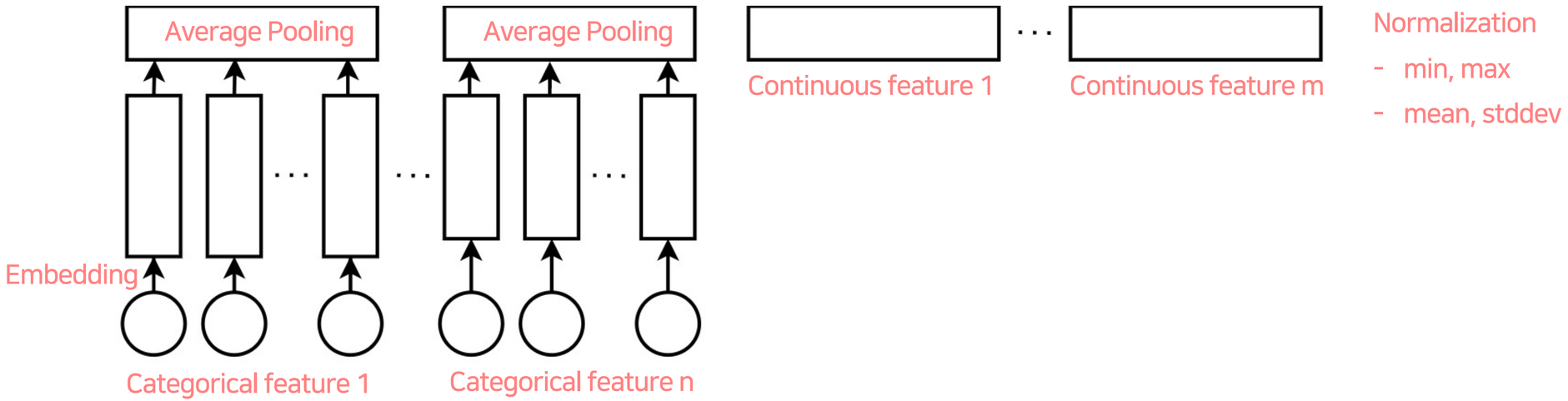
Model



3.2. User Embedding Model

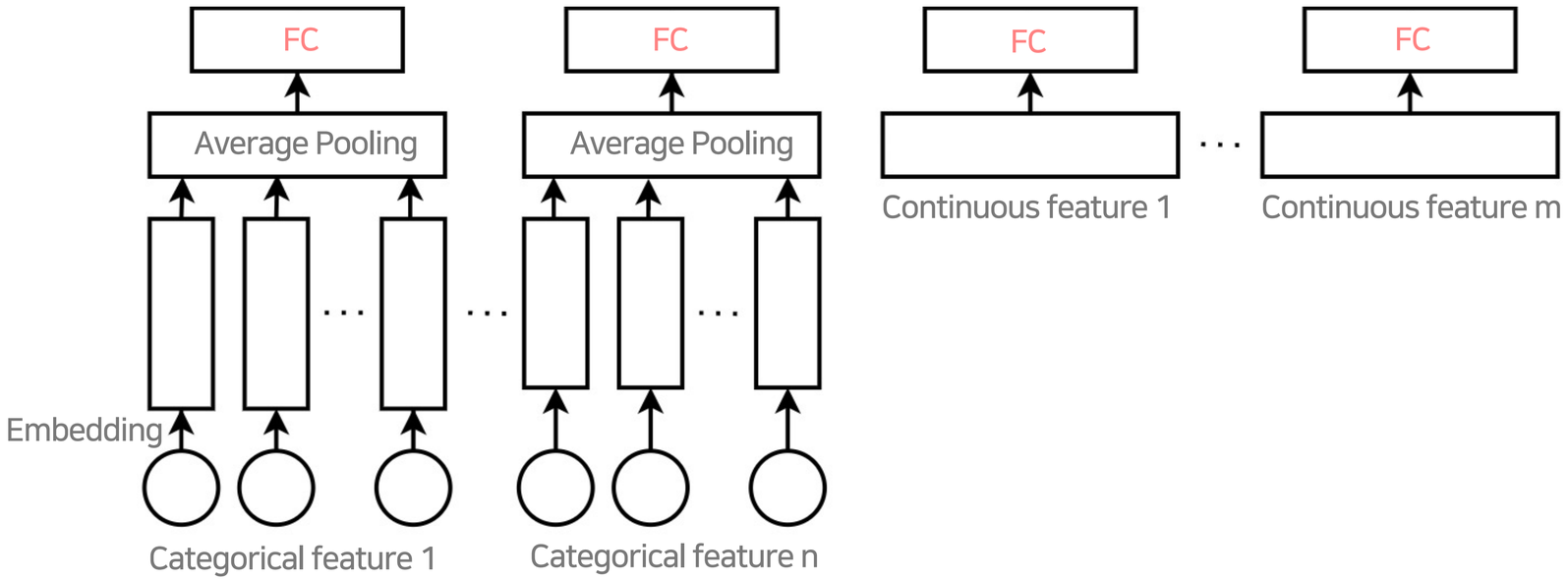
DEVIEW 2021

Model



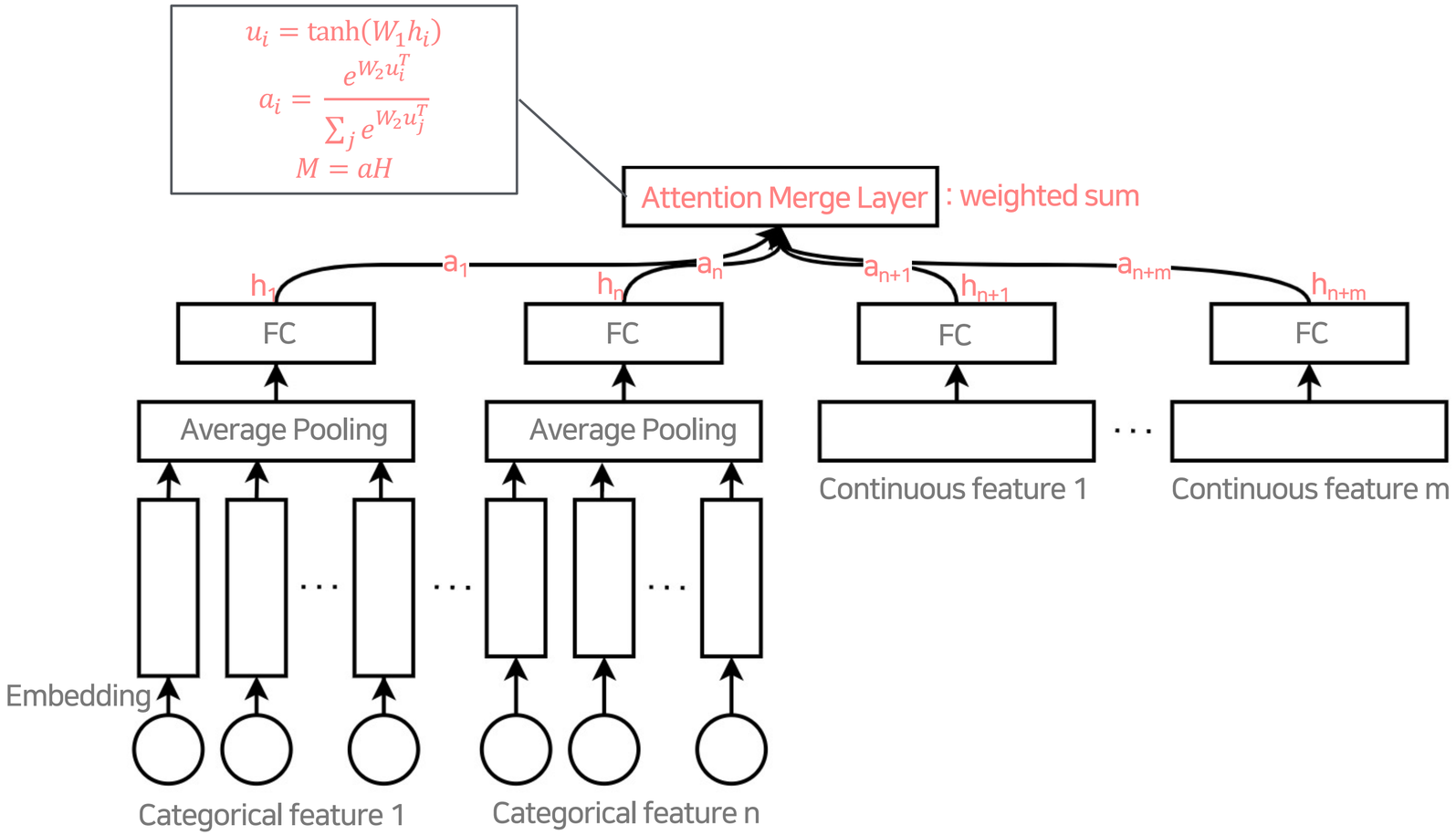
3.2. User Embedding Model

Model



3.2. User Embedding Model

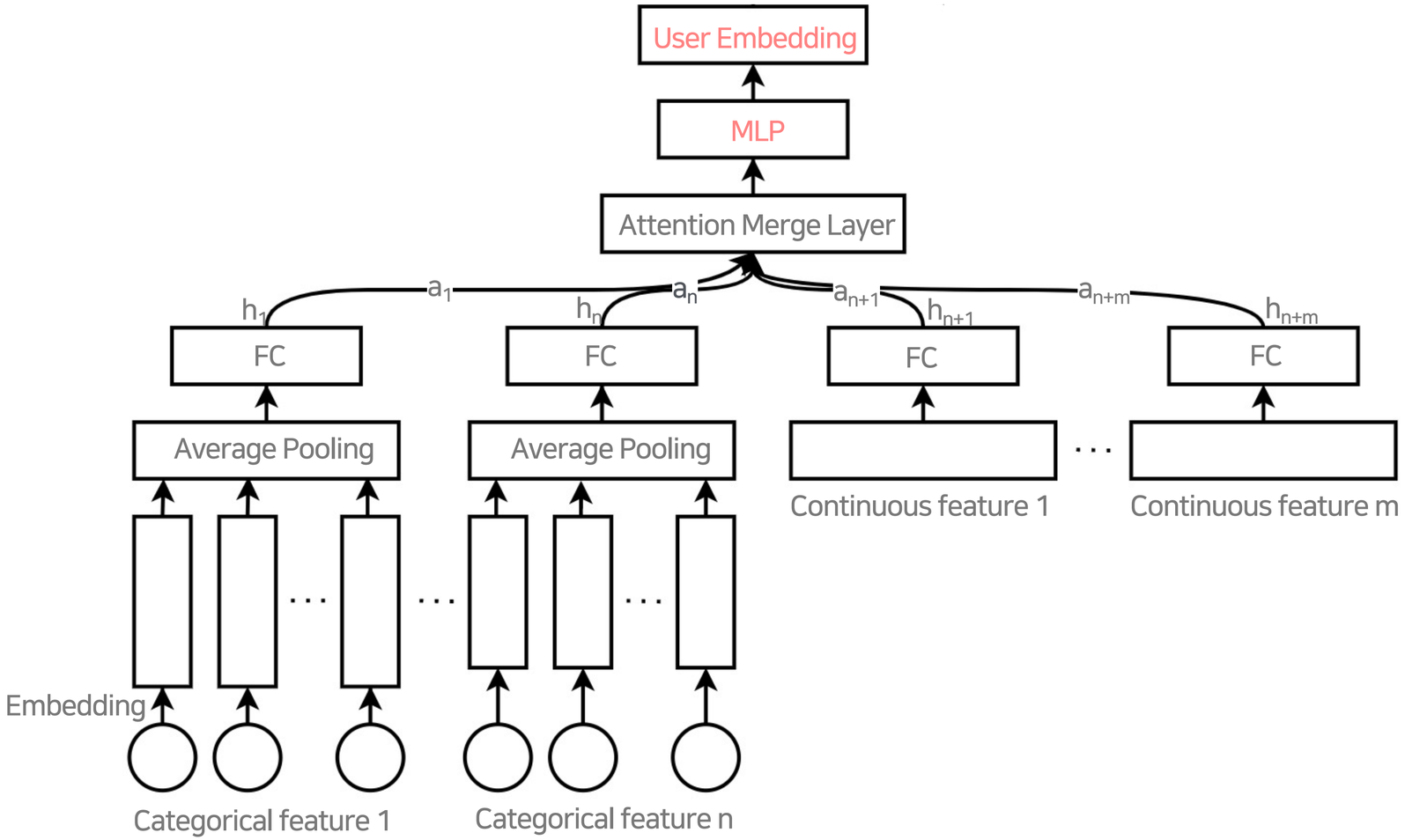
Model



3.2. User Embedding Model

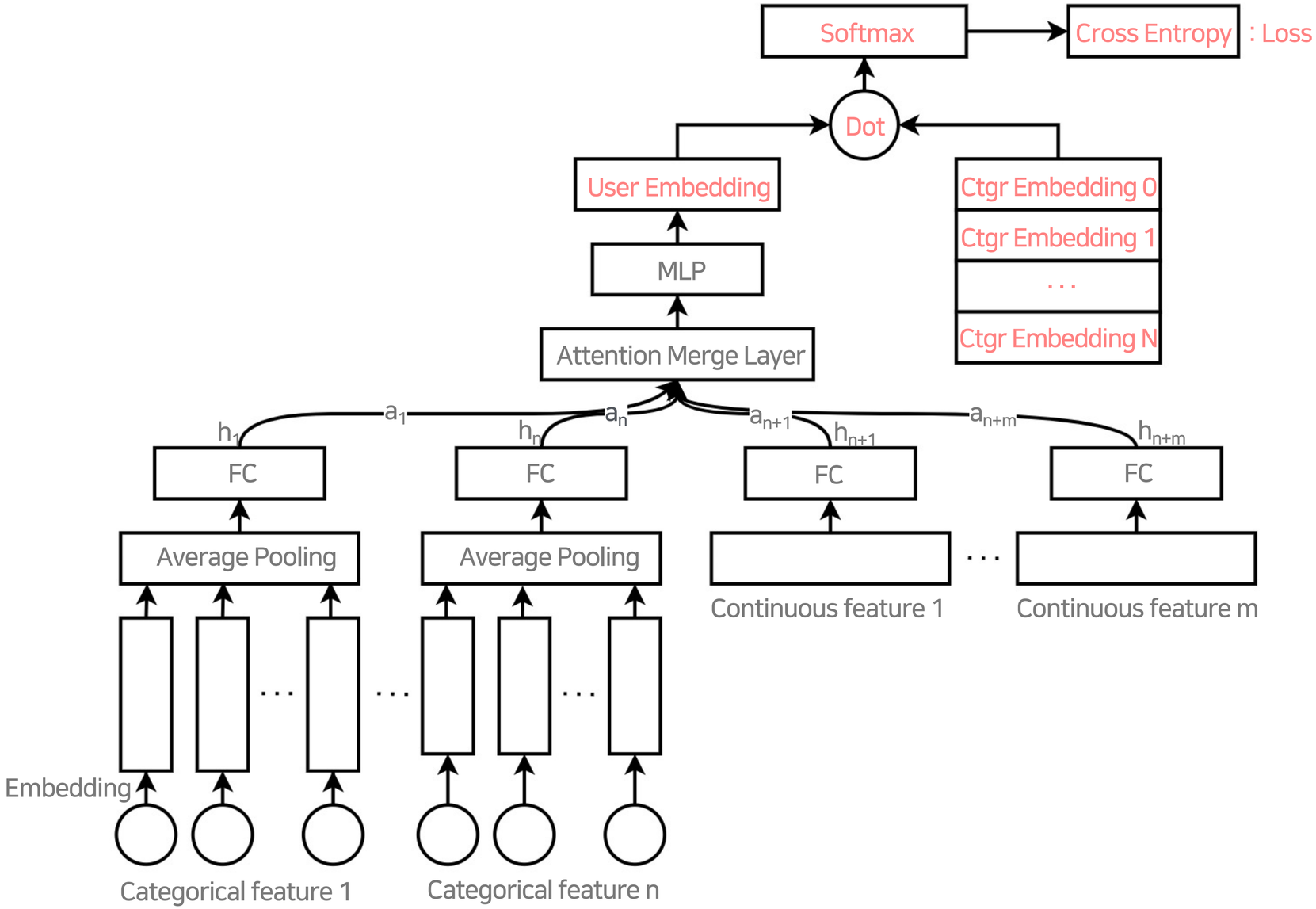
DEVIEW 2021

Model



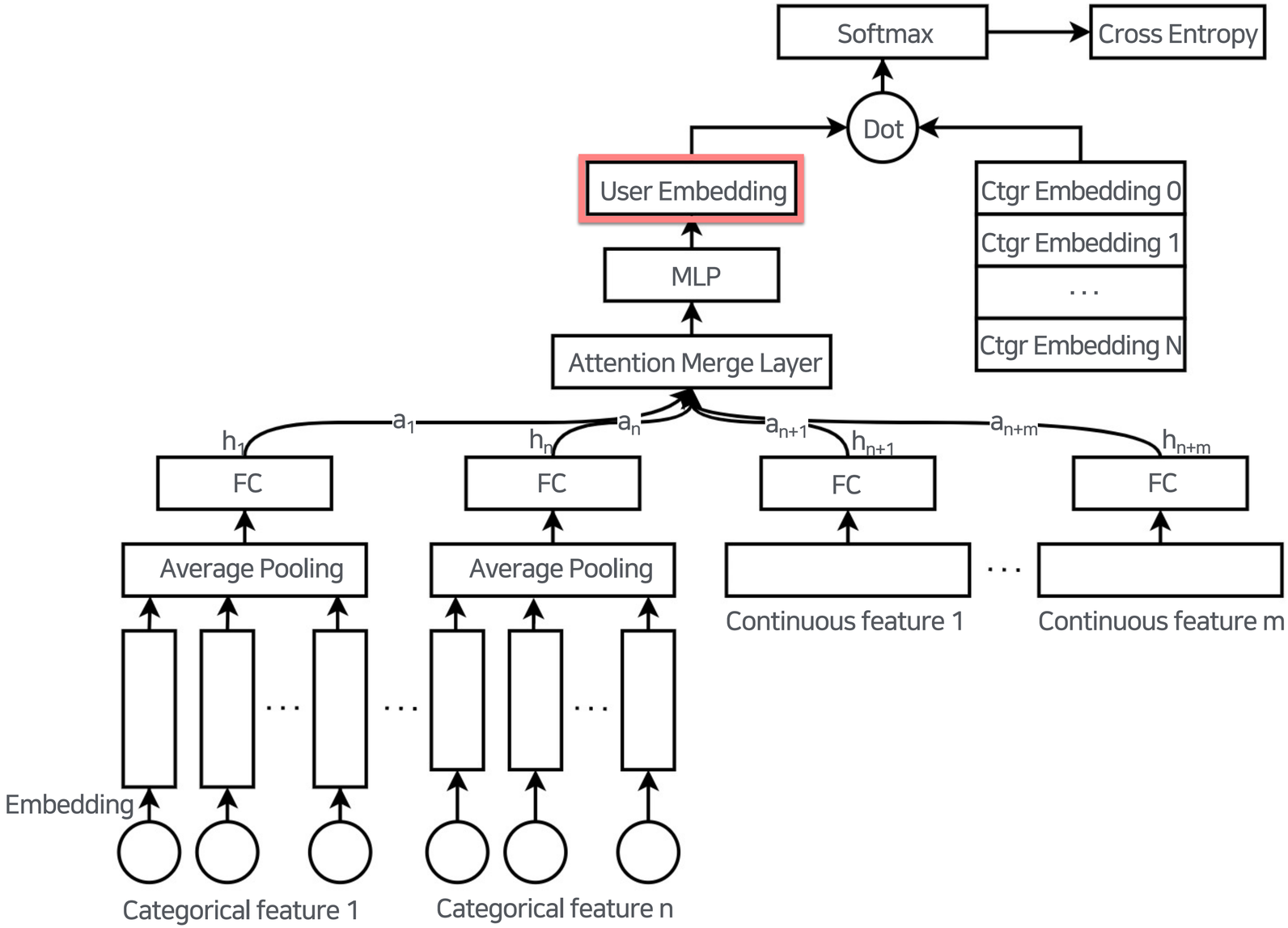
3.2. User Embedding Model

Model

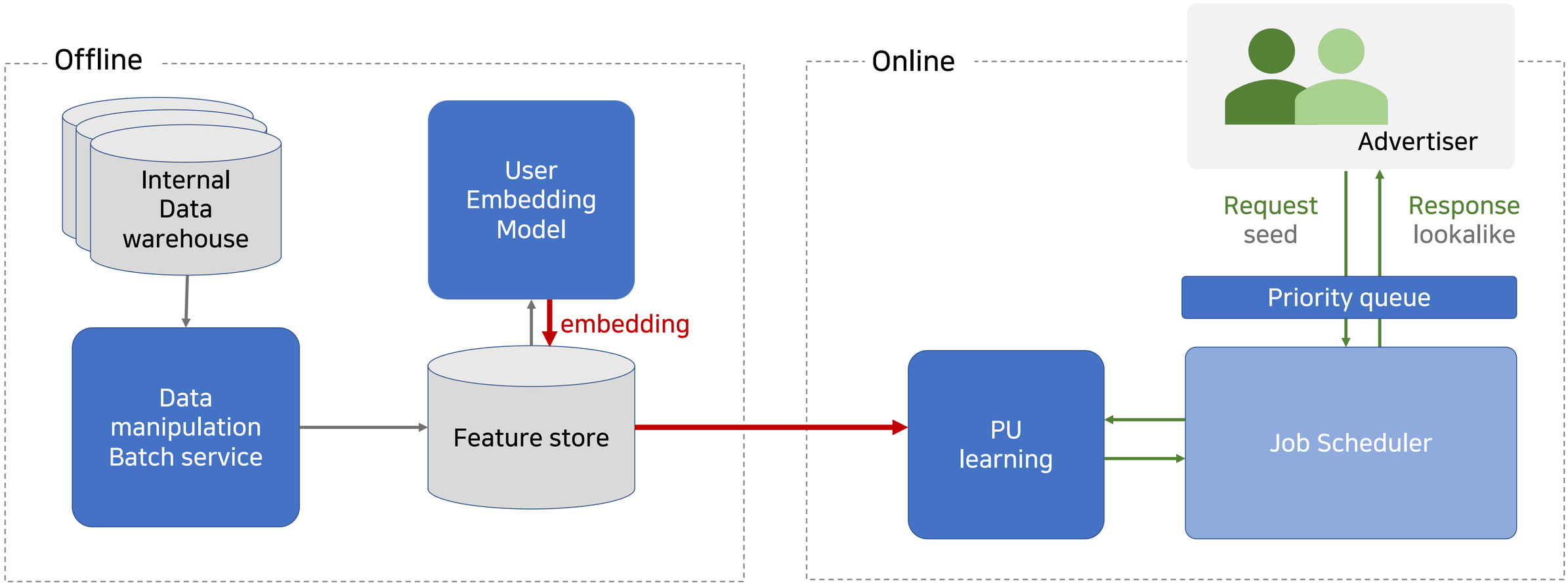


3.2. User Embedding Model

Model



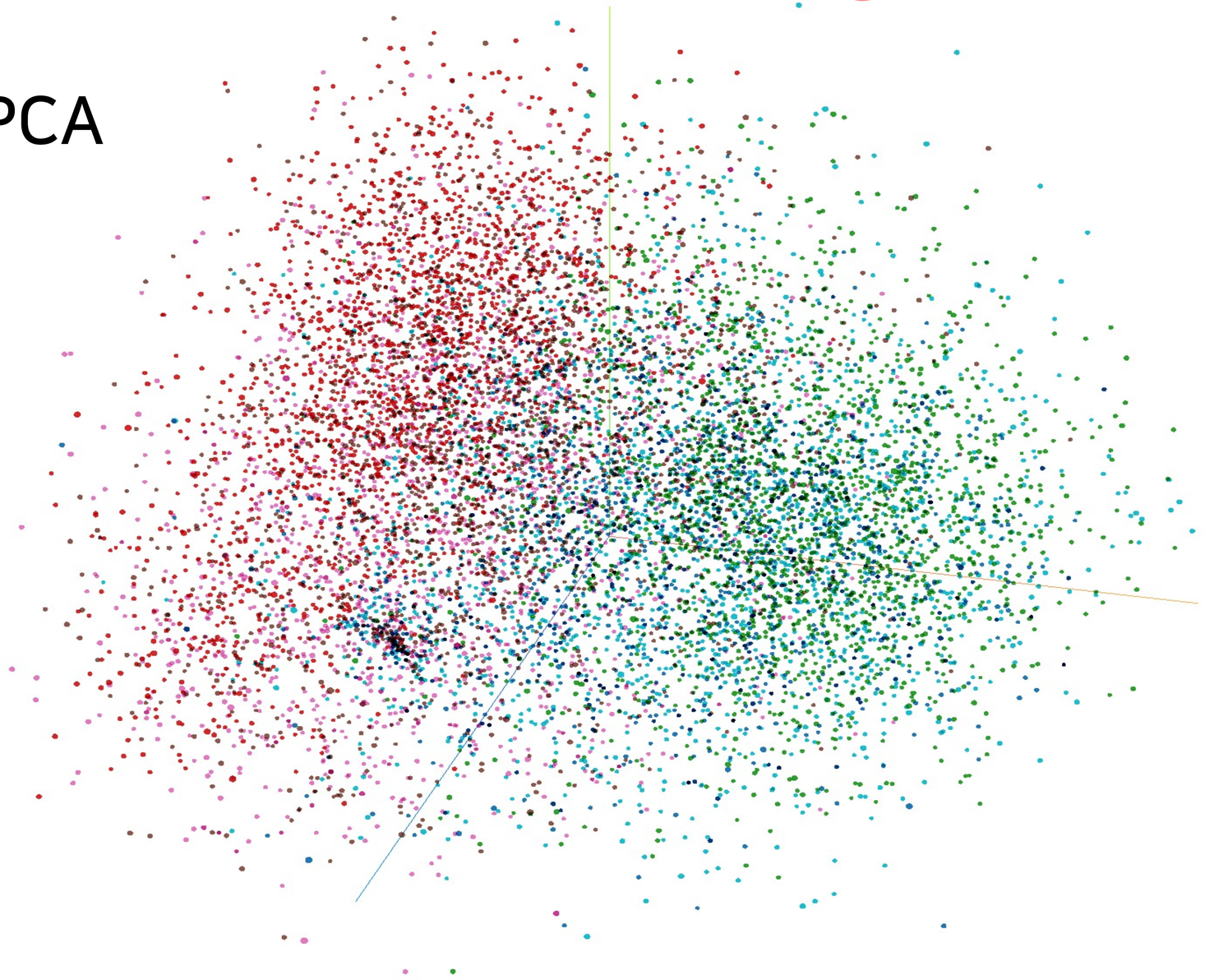
3.3. User Embedding 저장



3.4. User Embedding 시각화

DEVIEW 2021

PCA



화장품 B사

Seed
추천
구매

스포츠 U사

Seed
추천
구매

Seed: 해당 브랜드 스마트스토어 방문자
- B사: 2020.12.01 ~ 2020.12.10
- U사: 2020.11.26 ~ 2020.12.10

추천: Seed 기반 look-alike 모델 추천 유저
(2020.12.10 데이터 기준)

구매: 해당 브랜드 구매자 (2020.12.11)

4.Online Part:

Look-alike Learning

4.1. Online Part 목표

Seed와 user embedding 정보를 사용, 가까운 사용자 셋 반환

Seed: 1,000 ~ 3M (1만 ~ 10만 권장) 사용자 셋

Embedding 사용자 정보: Offline part에서 생성

처리 속도: 10분 이내 (ASAP) 완료

4.2. PU Learning

Positive example과 unlabeled example만 존재

Positive example (P): Seed

Unlabeled example (U): 전체 유저 – Seed

U를 positive와 negative로 분리하는 classifier를 만드는 것이 목표

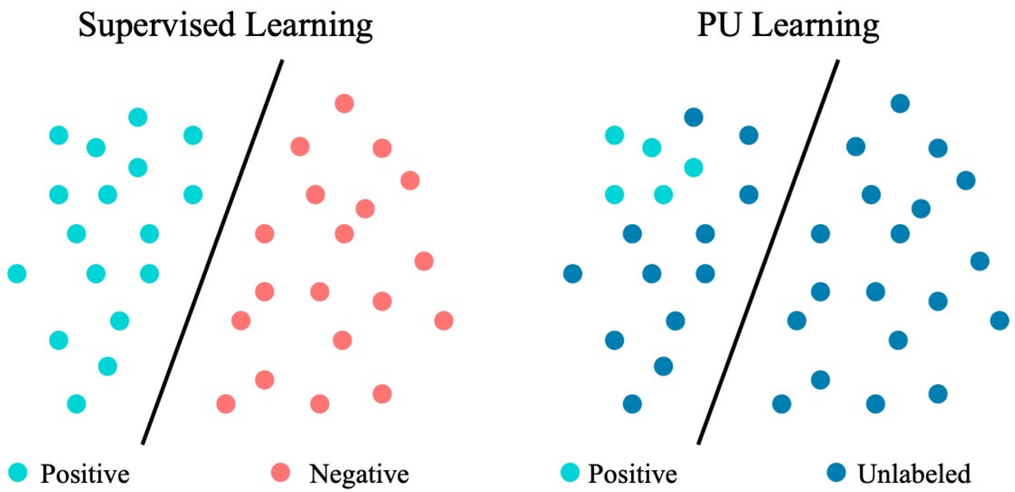


Fig. Illustration of PU learning[3]

4.2. PU Learning

DEVIEW 2021

2-stage learning

Stage 1

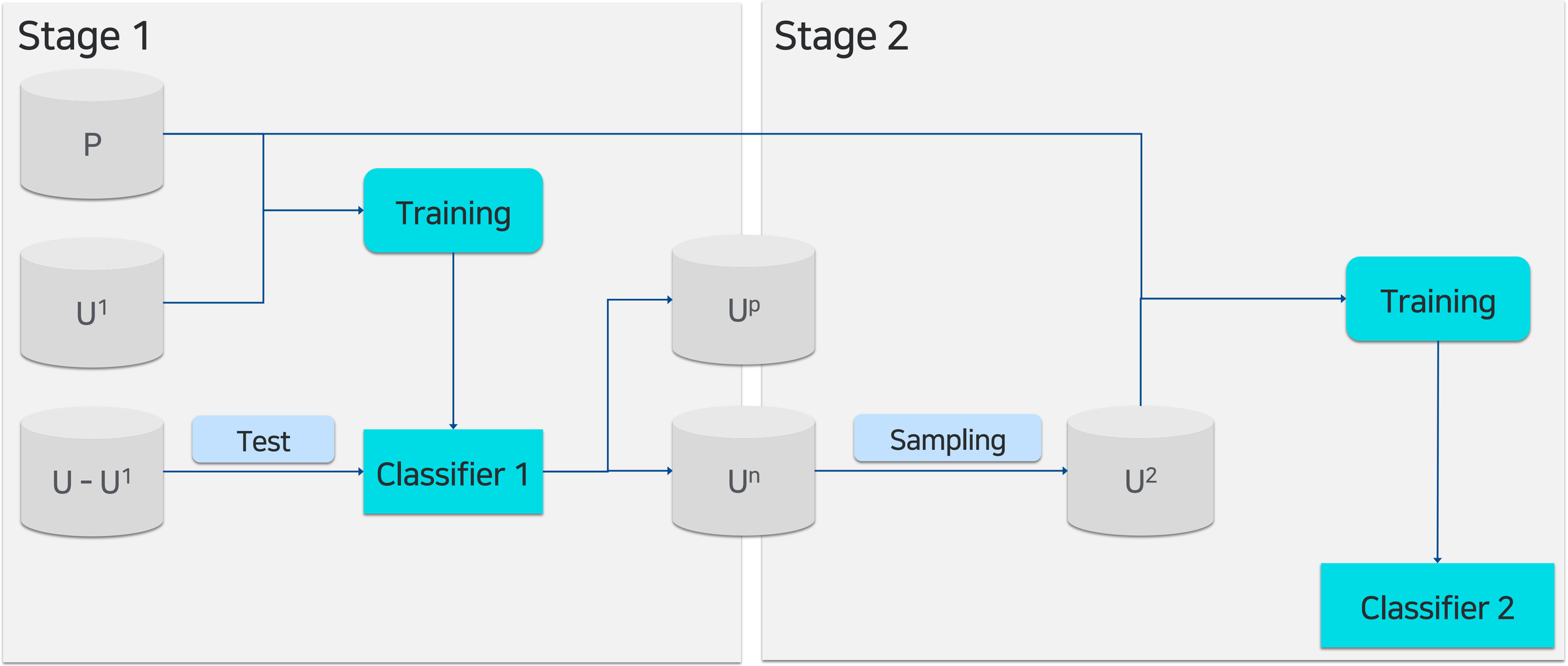
믿을 만한(reliable) negative를 찾는다

Stage 2

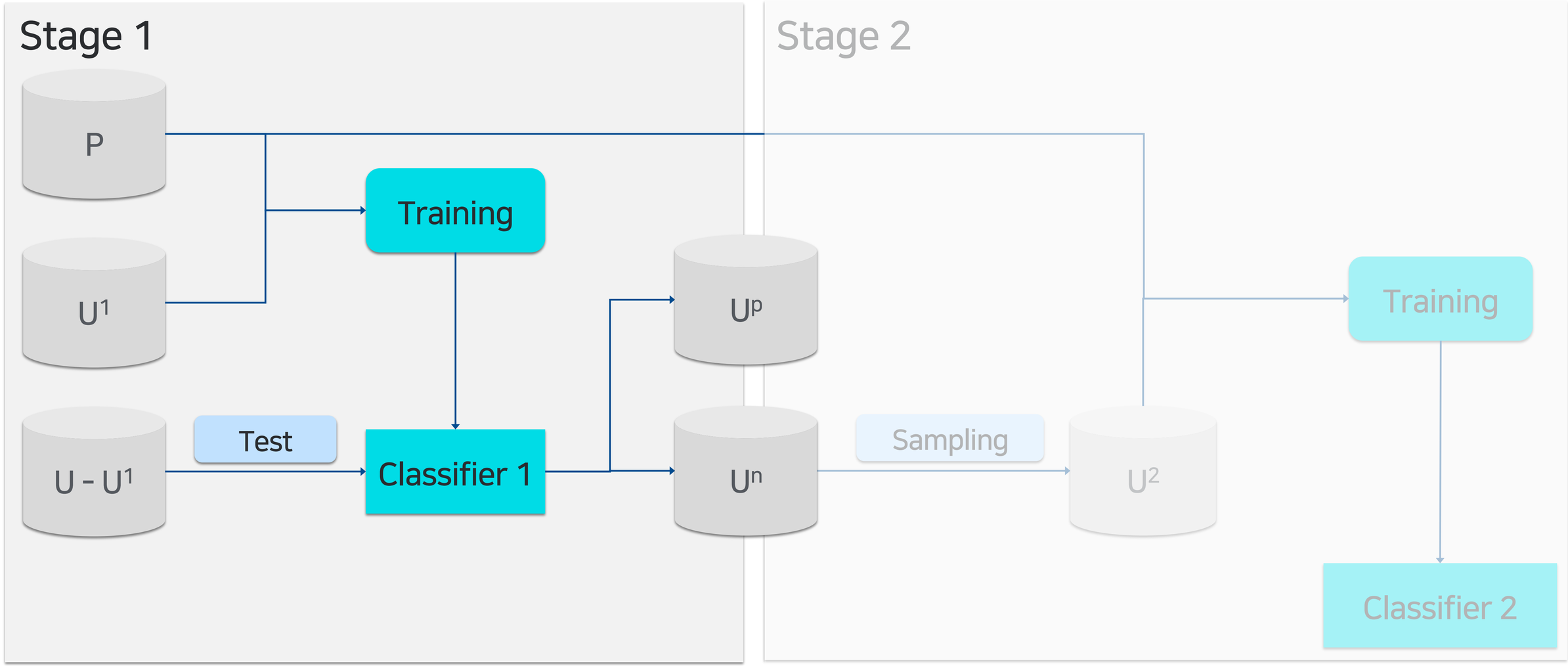
Positive와 stage 1의 reliable negative로 학습

Classifier로 U를 positive와 negative로 분리

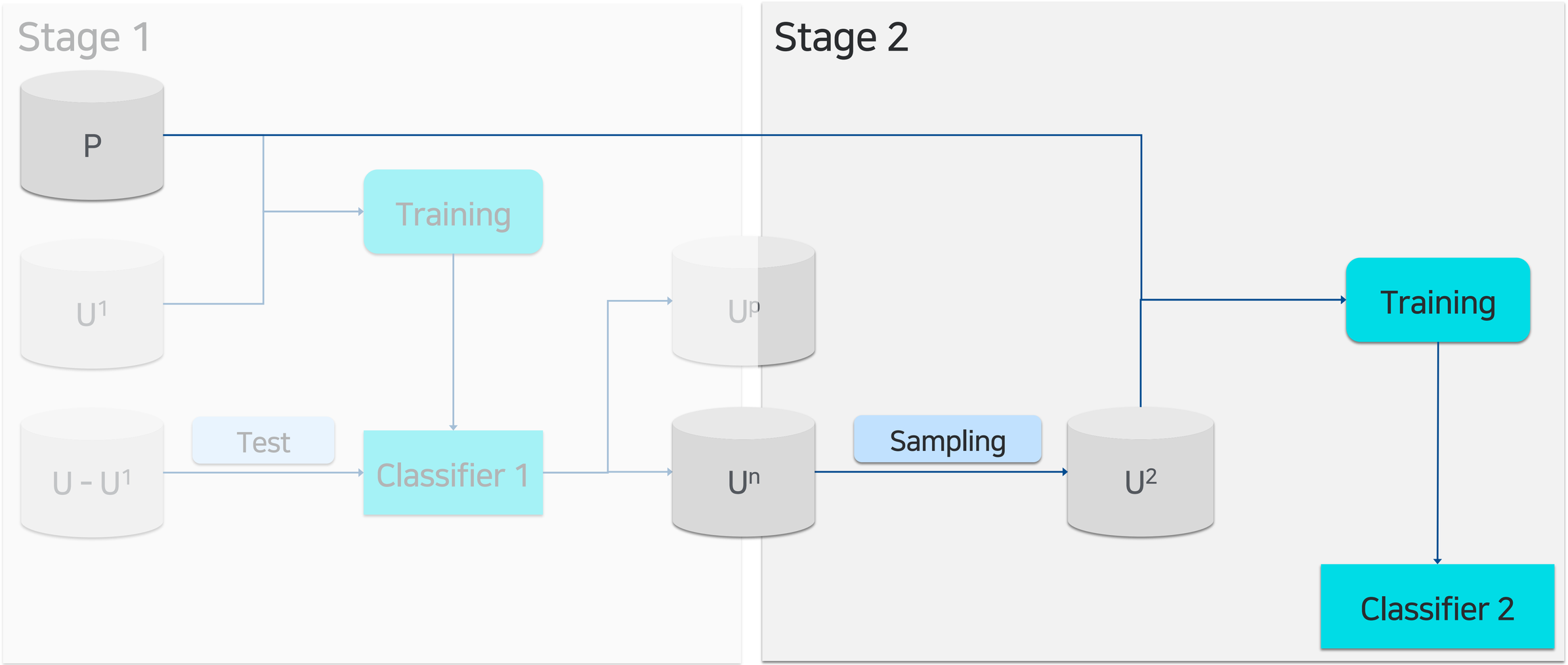
4.2. PU Learning



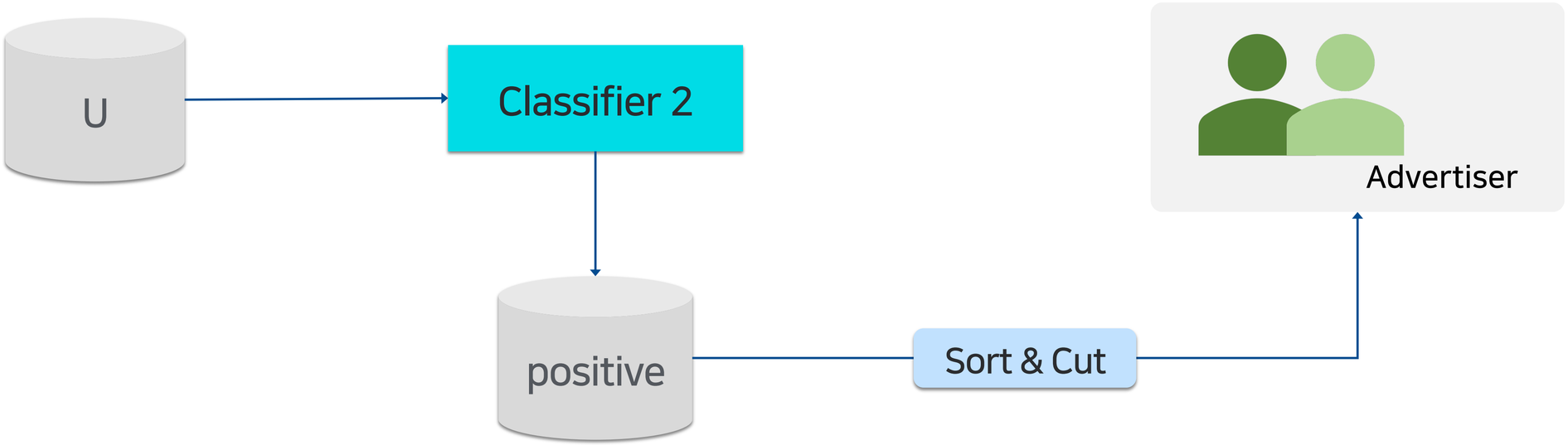
4.2. PU Learning



4.2. PU Learning



4.3. 추천 유저 선별



DEVIEW 2021

5.Service Part

5.1. 서비스 시스템 구성

Feature manipulation

Offline learning

Online learning / Service

Batch

Data Storage

5.2. 최적화

오프라인 학습, 추론 속도 향상

- TF graph 최적화
  - CPU 병목 제거, GPU Utilization 향상
    - non-TF python 코드 최소화, TF primitives 사용 등
- Data I/O 최적화
  - tf.data api 사용으로 대용량 데이터 hdfs를 통해 학습 및 추론

실시간 학습, 추론 속도 향상

- Spark Physical Plan 분석을 통한 최적화

5.3. 최적화 결과

DEVIEW 2021

오프라인 학습

학습 속도 (최적화 전 대비) 600% 향상

온라인 학습

온라인 학습, 추론 (최적화 전 대비) 200% 향상

6.Experiments & Results

6.1. 오프라인 테스트

DEVIEW 2021

목표 및 방법

모델 개발 중 모델 성능에 대한 피드백을 받기 위해 존재

과거 배너광고(Display Ad) data를 활용

대조군 1: Non-target (≒ 전체 유저 대상)

대조군 2 - 구모델: 기존 서비스 Look-alike 기법 (간략화된 거리 기반 계산 방법)

실험군 - 신모델: 현 Look-alike 기법

평가 지표: CTCVR, CVR, CTR (중요도 순)

* CTR은 이번 개선의 주요 목표는 아니나 기존 서비스와의 성능 비교를 위해

6.1. 오프라인 테스트

DEVIEW 2021

목표 및 방법

대상 광고: 가전, 패션잡화, 화장품 등 다양한 카테고리 별 광고

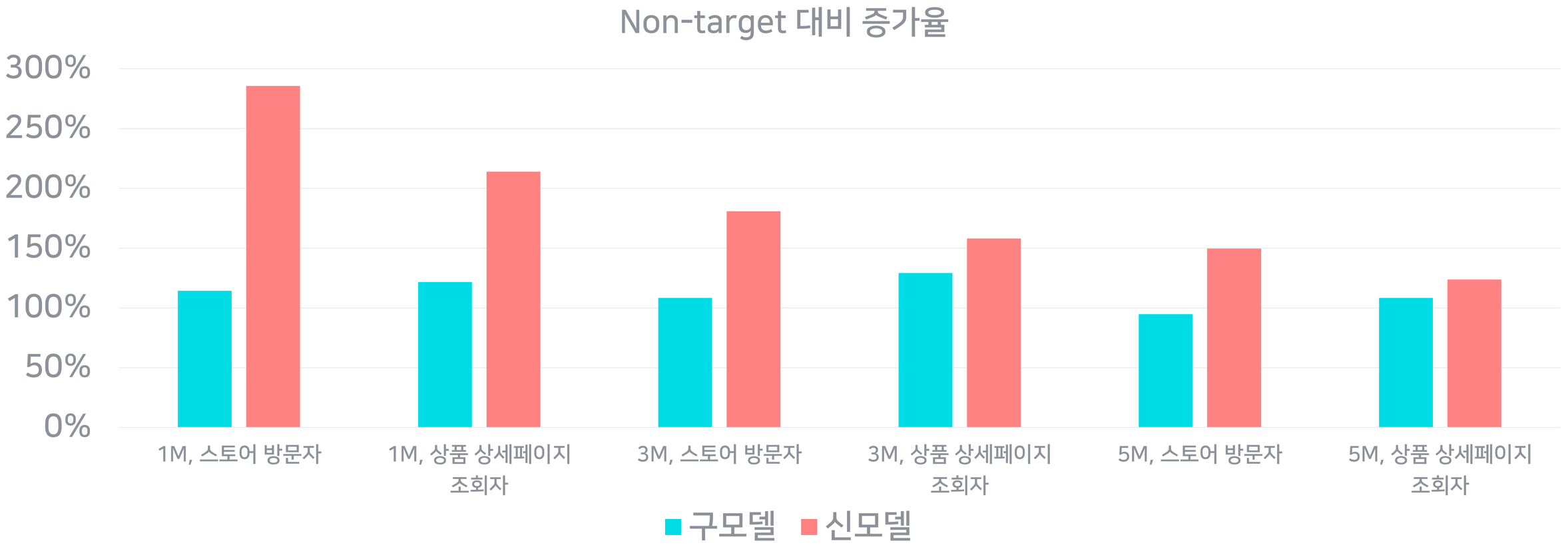
Seed 기준

- 광고주 스토어 방문자
- 광고주 스토어 상품 상세 페이지 조회자

확대 규모: 1M, 3M, 5M

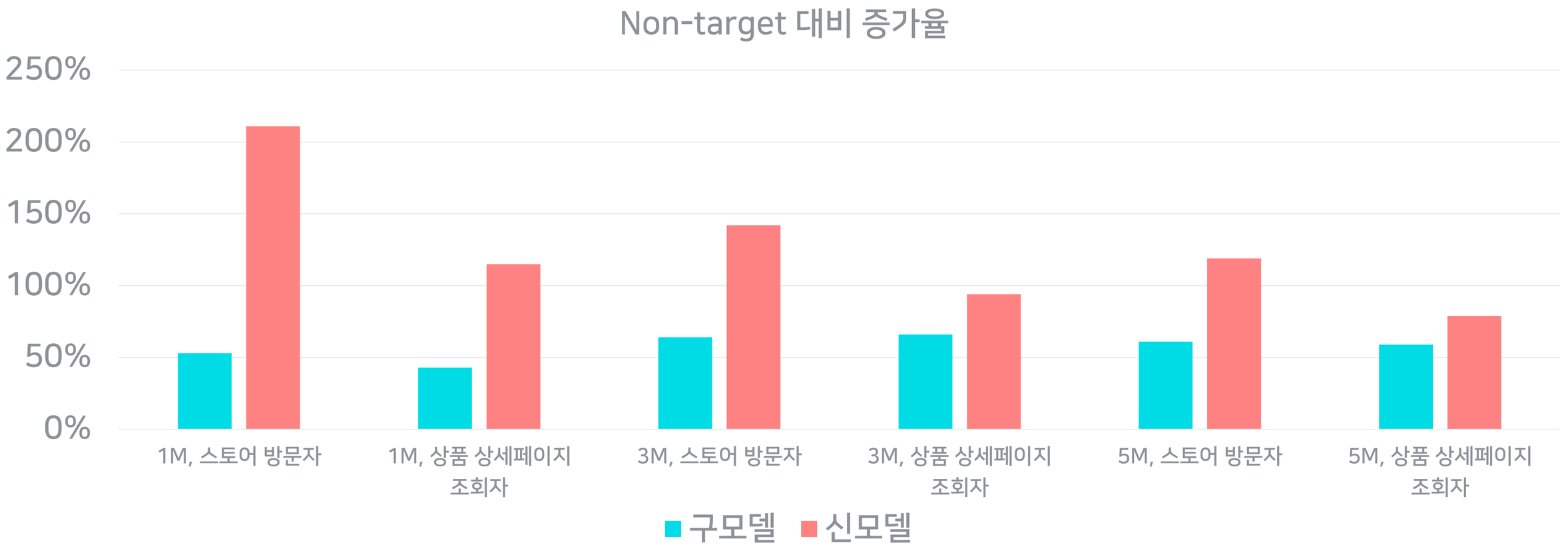
6.1. 오프라인 테스트

CTCVR 결과



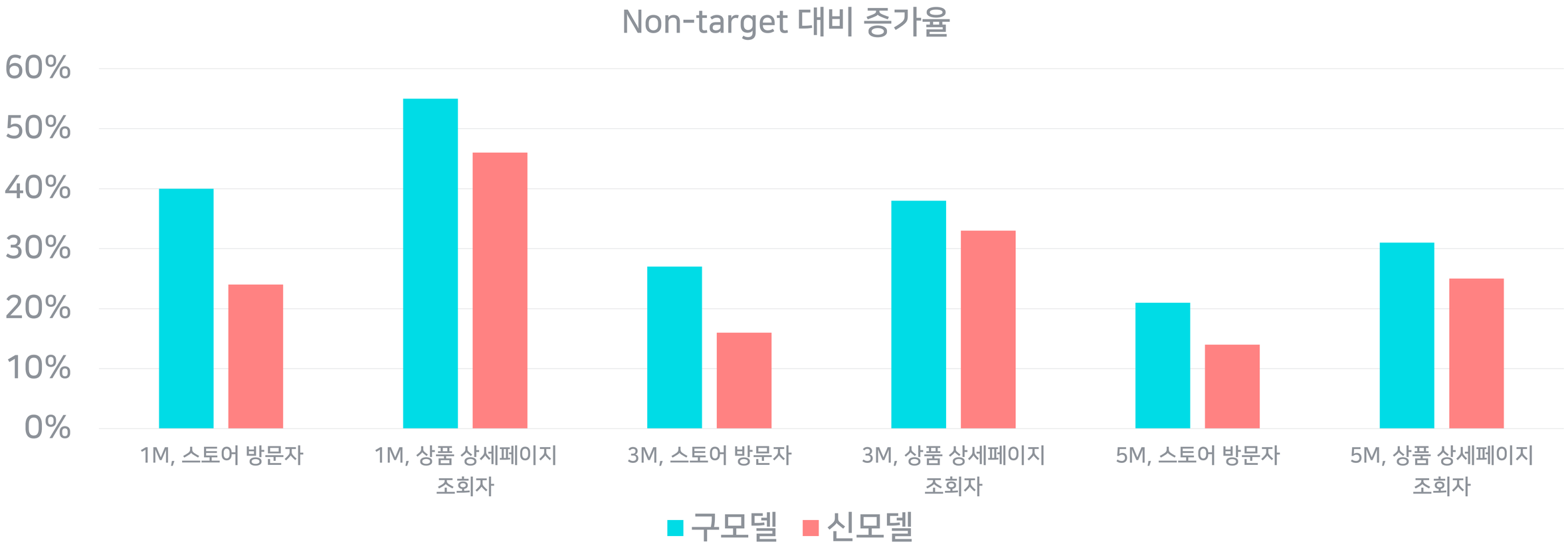
6.1. 오프라인 테스트

CVR 결과



6.1. 오프라인 테스트

CTR 결과



6.2. 실 서비스 성능 테스트

DEVIEW 2021

광고 선정 조건

모델 변경 전/후 2주일 광고 집행

유의미한 노출 수 보유 및 전환 수 존재

Look-alike 추천 사용자를 사용한 광고

지표	전	후	증감률
IMP(광고 노출)	580,908	483,931	-16.69%
CLICK	1,887	1,506	-20.19%
CONVERSION	10	50	400.00%
CTR	0.32%	0.31%	-4.20%
CVR	0.53%	3.32%	526.49%
CTCVR	0.002%	0.010%	500.20%

일반화 할 수 없으나 두 케이스에서 매우 큰 (CT)CVR의 상승이 존재

6.3. 성능 개선 실험

CTR 개선 실험

CVR은 기존 모델(18’) 대비 크게 향상, CTR은 광고에 따라 상승, 하락 존재

실험: User embedding learning task의 target (Y) 변경

- 실험군: 상품 구매 + 검색광고 클릭
- 대조군: 상품 구매

결과: CTR, CVR 소폭 하락

하락 이유 분석

- 상품 구매가 검색광고 클릭보다 CTR, CVR에 좋은 데이터일 수 있음
- 성질이 다른 두 target을 함께 학습해서 학습 난이도 상승, fitting이 잘 안됨

6.3. 성능 개선 실험

DEVIEW 2021

노출 수 개선 실험

타게팅 기능을 사용하기 위해, 일정 광고 노출량 확보 필요

실험: 유저 filtering

네이버 서비스 활동성 지표 생성

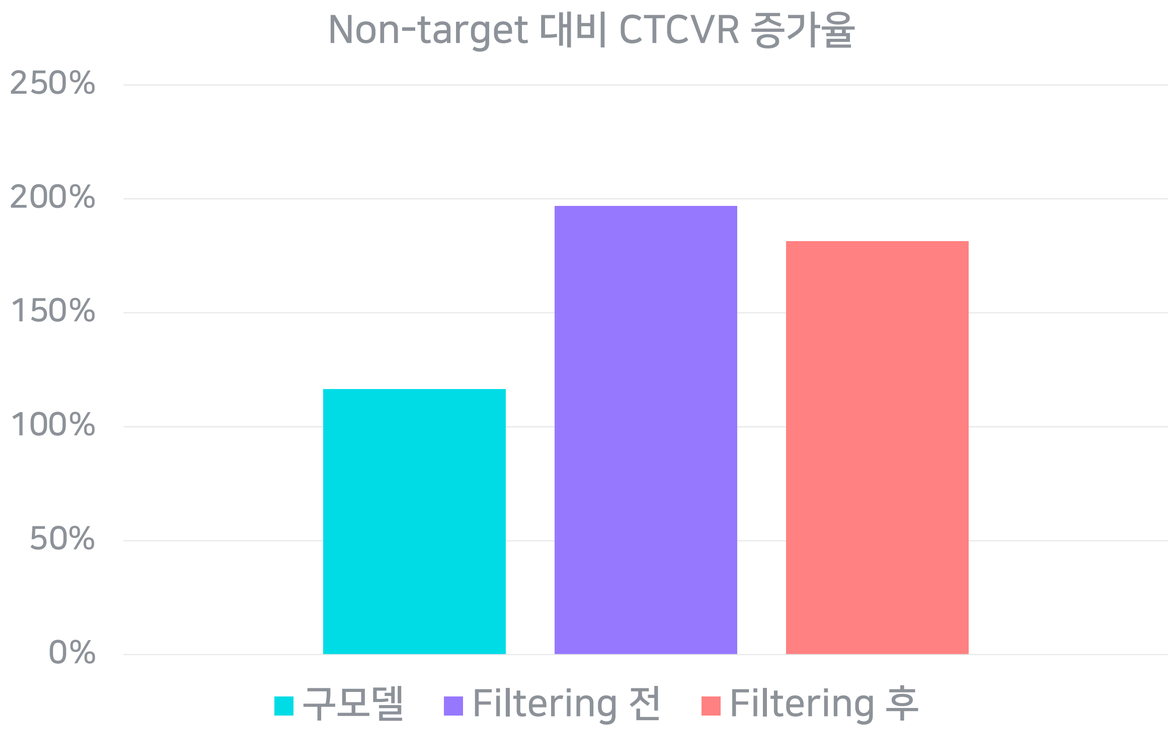
지표가 낮은 유저는 추천 후보에서 제거

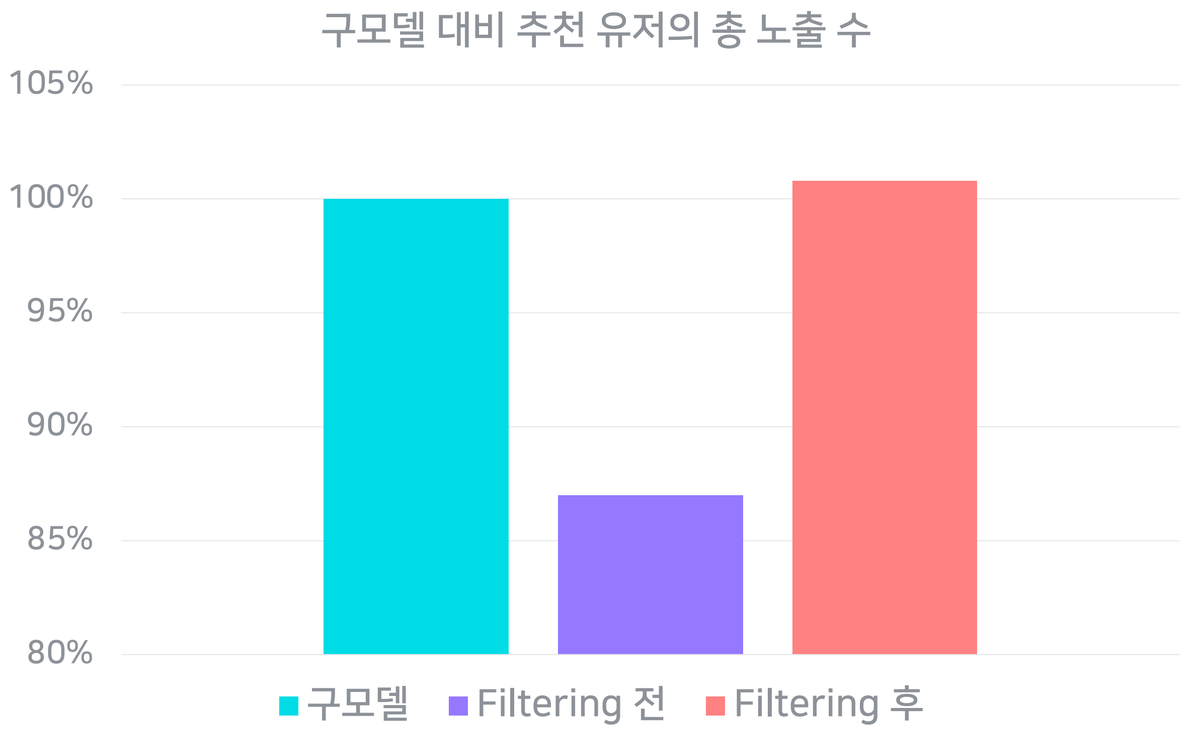
확대 규모: 3M

6.3. 성능 개선 실험

노출 수 개선 결과

(CT)CVR 소폭 하락, 구 모델(18')과 비슷한 수준으로 노출 수 확보





7.Conclusion

7.1. 방법 및 성과 정리

방법

딥 러닝을 통해 대량의 유저 행동 데이터를 embedding
Embedding 데이터를 사용해 실시간 learning 및 서비스

성과

기존 기법 대비 CTR 소폭 감소하나, 매우 높은 CVR을 얻음

향후 발전 방향

광고주 seed 외 네이버 내부 seed를 사용, 자유롭게 특화 사용자 셋 생성 가능

Reference

[1] Real-time Attention Based Look-alike Model for Recommender System, KDD 2019
[2] https://iabtechlab.com/standards/audience-taxonomy/, IAB Audience Taxonomy
[3] Epileptic Focus Localization Based on iEEG by Using Positive Unlabeled (PU) Learning, APSIPA 2018

Thank You

Appendix

IAB(Interactive Advertising Bureau)

온라인 광고 산업의 표준 개발, 연구 및 법률 지원을 제공하는 조직
세계 여러 나라의 650여개 이상의 회사들이 가입

Audience Taxonomy 제공 - Demographics, Purchas Intent, Interest

Purchase Intent

상품 또는 서비스를 당장 구매하려 할 사용자를 분류하는 경우에 사용하는 체계