DKT

알아서 잘 딱 깔끔하고 센스있게.

김건우, 김동우, 박기정, 심유정, 이성범

프로젝트의 방향성

EDA

> 모델을 활용한 데이터의 패턴 분석

Model Architecture

> 데이터의 패턴을 제대로 반영할 수 있는 Custom model 개발

Feature Engineering & Training Method

> 일반화된 모델을 만들기 위한 학습 방법

모델을 활용한 데이터의 패턴 분석

GMF

- 유저 임베딩의 영향으로 모델이 쉽게 과적합
- 문항에 대한 정보를 추가할수록 모델의 성능 향상

(유저마다 존재하는 데이터의 수가 적어 적절한 유저 임베딩 학습 불가)

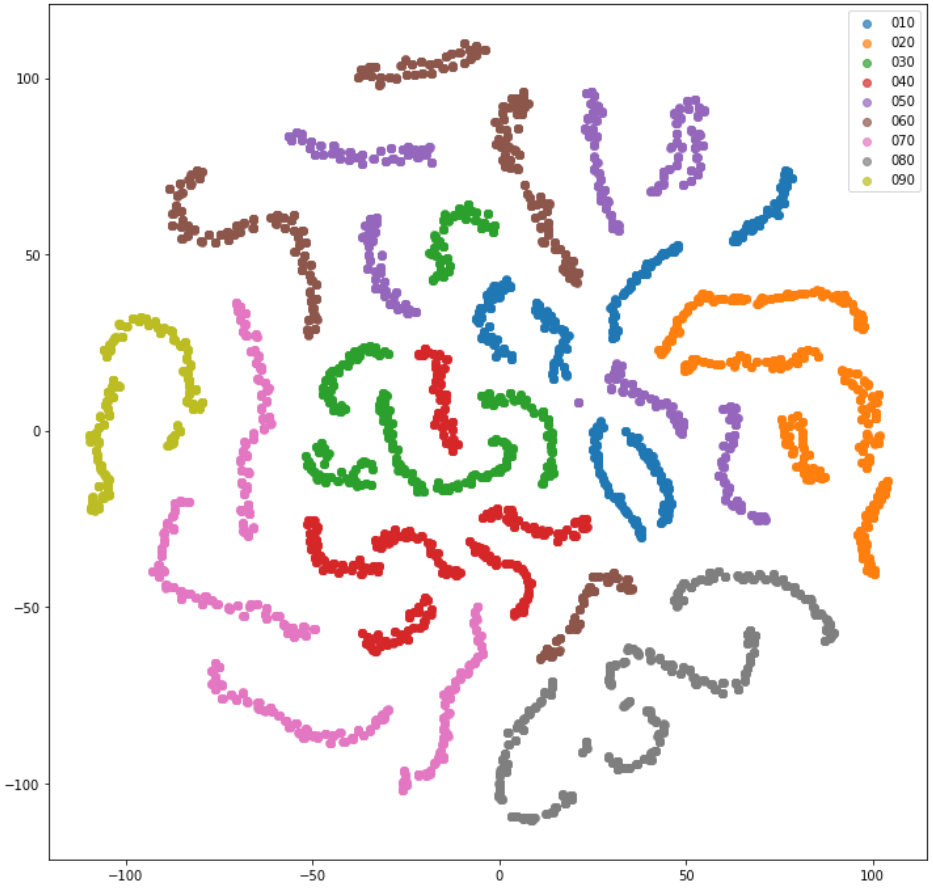
LightGCN

- 유저-문항을 그래프 형태로 표현하여, 유저-유저 연관성을 기반으로 데이터를 표현하면 모델 성능이 좋지 않다는 것을 확인

Item2Vec

- 유저의 seq를 이용한 문항 정보를 임베딩 했을 때, 유저의 seq 내에 특정한 문항 패턴이 존재

(문제 풀이 내역 마다 공통된 특정한 문항 풀이 패턴이 존재)



Word2Vec Model 학습 결과

모델을 활용한 데이터의 패턴 분석

BERT(양방향) VS Transformer(단방향)

- 문제 풀이 내역을 양방향으로 학습하는 것보다 **단반향으로 학습을 하는 것이 더 좋은 성능**을 보임
- 즉, **유저의 문제 풀이 순서는 매우 중요한 패턴이 내재되어 있다**는 것을 알 수 있음
- Transformer를 이용하는 경우, **단순히 문항을 이용해 다음에 등장할 문제를 예측하더라도 좋은 정확도**를 보임

정리)

1. **유저를 표현할 수 있는 데이터의 수는 매우 적음**

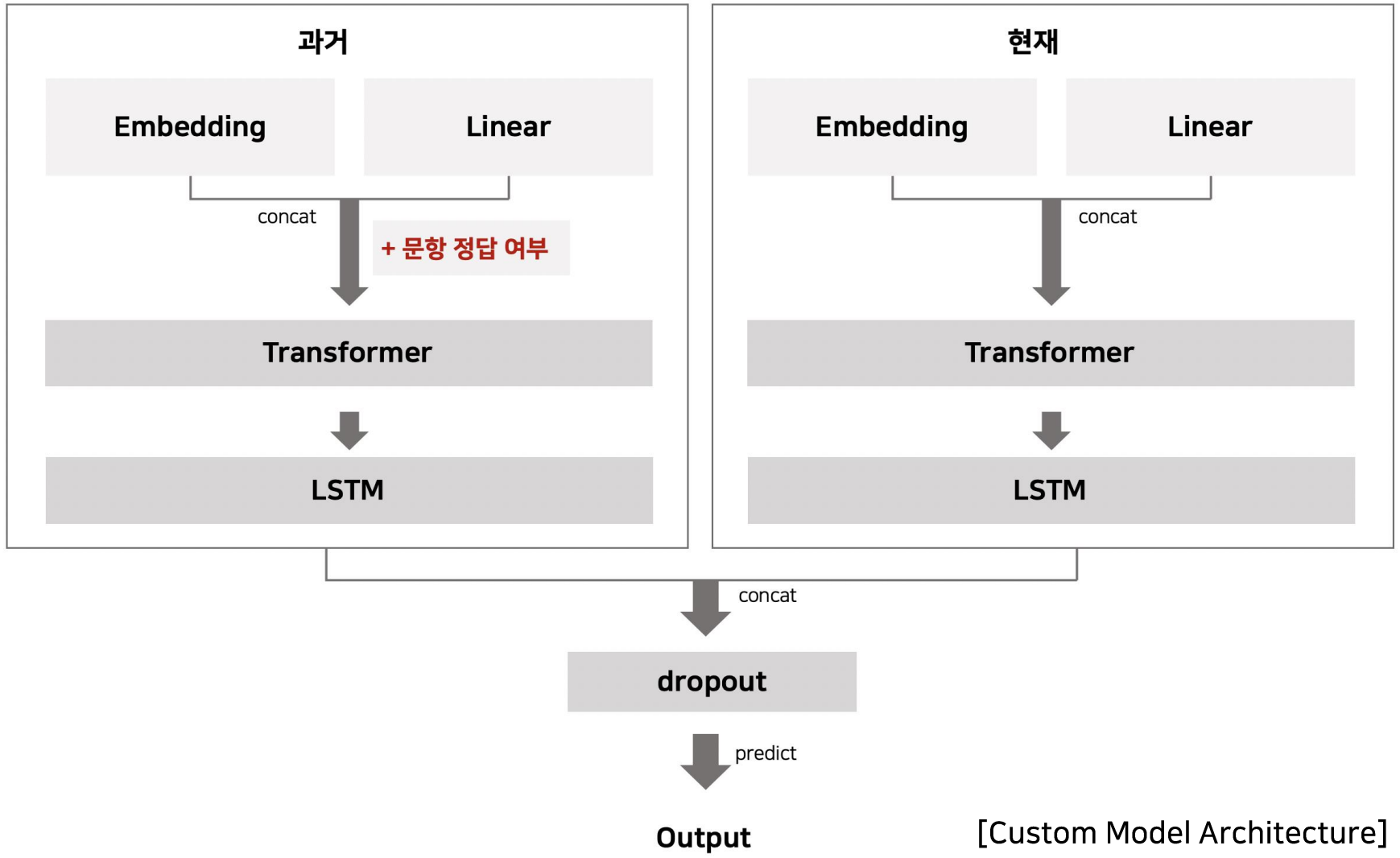
   → 유저를 임베딩 하는 것보다 **최대한 문항 정보를 활용하여 parameterized function을 만드는 것**이 중요

2. **유저의 문제 풀이 순서는 매우 중요한 패턴**

   → **시간적 순서를 효과적으로 표현할 수 있는 Model Architecture 설계**가 중요

데이터의 패턴을 제대로 반영할 수 있는 Custom model 개발

알잘딱깔센



[Custom Model Architecture]

- 과거 풀이 정보와
  현재 풀이 정보 간의 연관성을 표현
- 서로 다른 Embedding Layer를 두어
  non-convex 한 목적 함수를
  convex하게 만듦

모델의 일반화 성능을 높이기 위한 학습 방법

알잘딱깔센

1. Representation Learning

- 데이터의 표현을 효과적으로 학습할 수 있는 Model Architecture
- 최소한의 변수로 최대의 성능

2. Feature Selection

- EDA를 통한 Feature Selection
- 범주형 - 문항, 시험지, 태그, 시험지 대분류, 시간, 요일 등
- 수치형 - 정답률의 평균 / 표준편차, 풀이 시간의 평균 / 표준 편차 등

3. Loss

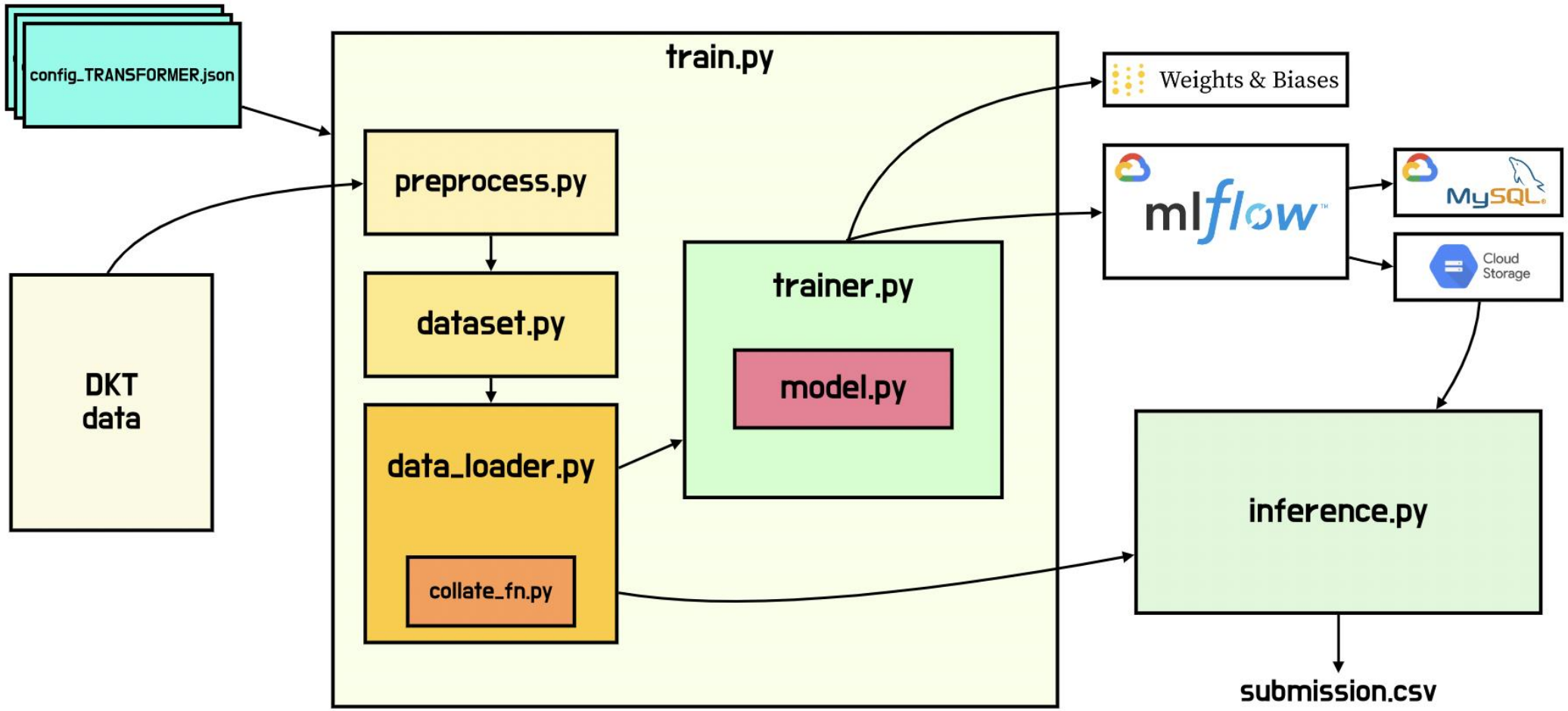
- 전체 Time-step에 대하여 Loss를 계산하여 유저 데이터 부족 문제를 해결

4. padding

- Max-len과 mean-len의 차이가 크기 때문에, 배치 마다 서로 다른 크기의 padding을 두어 모델을 학습 시킴

전체 학습 프로세스

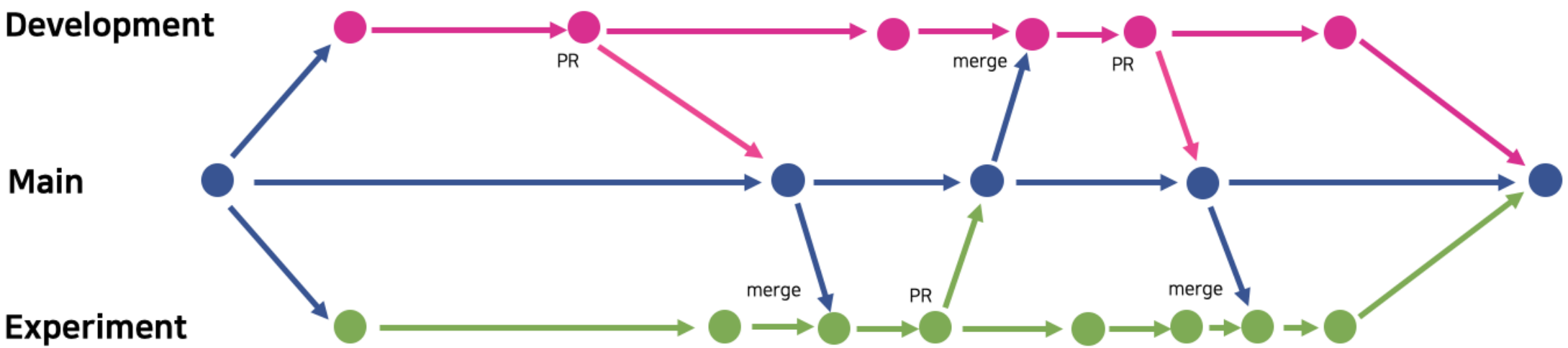
알잘딱깔센



https://github.com/victoresque/pytorch-template 기반으로 제작

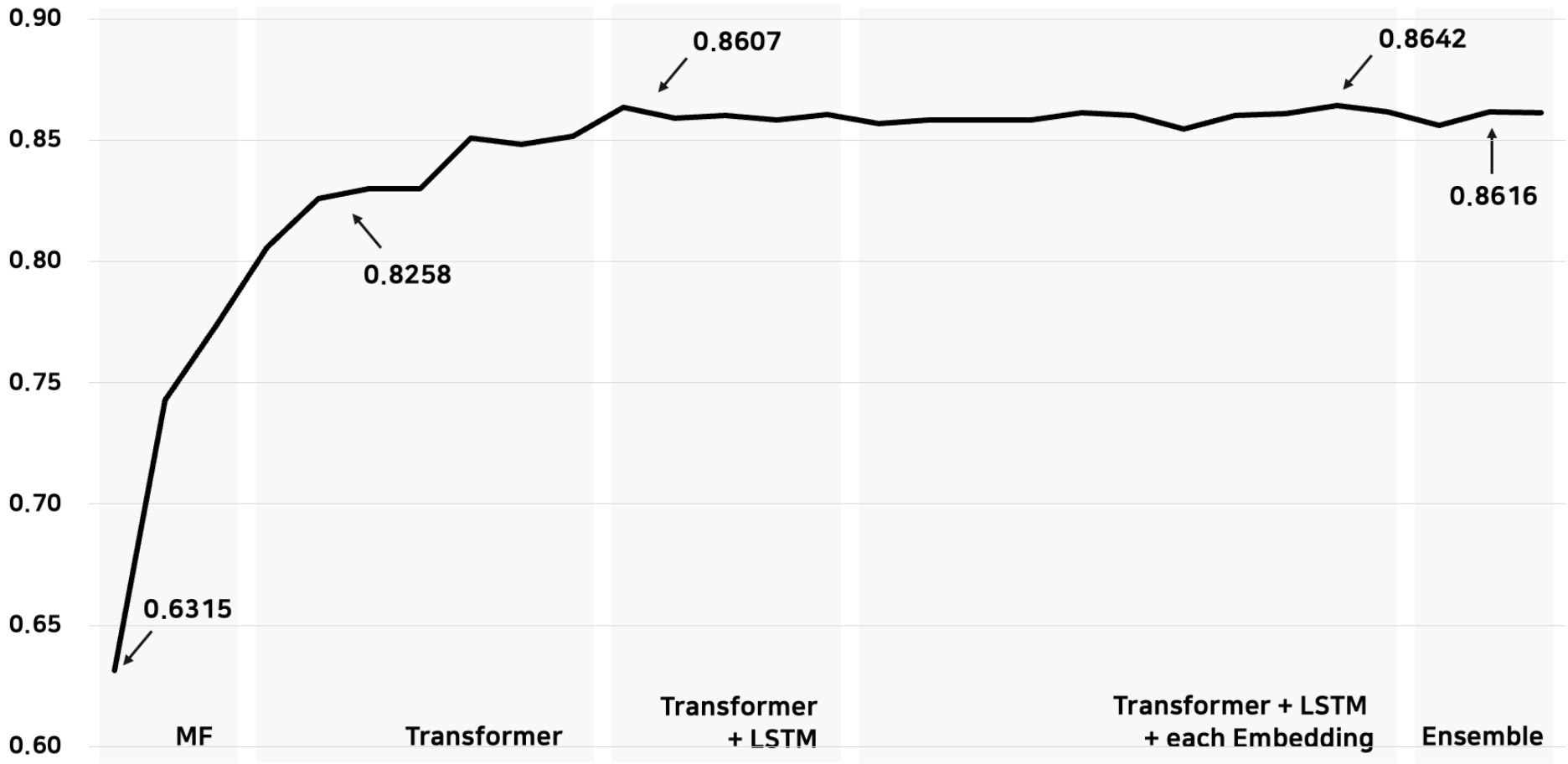
깃허브 브랜치 전략

알잘딱깔센



[Git Branch Strategy]

리더보드



*최종 점수 기준

End of Document
Thank You.