P Stage 3 : Machine Reading Comprehension

팀이쿄! 너도 할 수 있어! *

TEAM-IKYO 권태양 류재희 박종헌 오수지 이현규 정익효

boostcamp

| 목차 |

01 History

History

40 ~ 50

Retriever

- Wiki data 전처리

Reader

- Question random masking (token 기준)
- KoELECTRA 모델 사용

Inference

- Elastic Search k <= 5

History

40 ~ 50

Retriever

- Wiki data 전처리

Reader

- Question random masking (token 기준)
- KoELECTRA 모델 사용

Inference

- Elastic Search k <= 5

50 ~ 60

Retriever

- Elastic Search
- Wiki data 불용어 제거

Reader

- Question random masking (단어 기준)
- Pretrained KorQUAD
- xlm-roberta-large 모델 사용
- Custom model conv1d layer 추가

Inference

- Elastic Search k <= 10

History

40 ~ 50

Retriever

- Wiki data 전처리

Reader

- Question random masking (token 기준)
- KoELECTRA 모델 사용

Inference

- Elastic Search k <= 5

50 ~ 60

Retriever

- Elastic Search
- Wiki data 불용어 제거

Reader

- Question random masking (단어 기준)
- Pretrained KorQUAD
- xlm-roberta-large 모델 사용
- Custom model conv1d layer 추가

Inference

- Elastic Search k <= 10

60 ~ 70

Reader

- Context masking (span 기준 중요도에 따라)
- Pretrained AIHUB data
- Custom model로 deep한 모델 사용

Inference

- wiki data split
- Elastic Search k > 10
- Sentence Transformer
- Hard Voting Ensemble

02 Retriever

Search Engine

Search engine	precision@5	precision@10	precision@15
Baseline retrieval using morpheme tokenizer	59%	67%	77%
Baseline retrieval using word tokenizer	65%	75%	83%
BM25	83%	88%	92%
Dense Retrieval (elastic search top 100)	86%	90%	96%
Elastic Search	88%	92%	96%

Elastic Search

Elastic Search Settings

- tokenizer : nori_tokenzier
- decompound mode (복합명사 처리 방식) : mixed (복합명사로 분리 + 원본 데이터도 유지)

Wiki Dataset Pre-processing

- “₩n”, "₩₩n", "₩s+", “#” 등 answer에 존재하지 않는 특수문자 제거
- 불용어 사전 : 중요도가 낮은 한국어 조사/어미를 불용어로 사용

03 Reader

Additional Data

- **korQuAD** : 1,560 개의 Wikipedia article에 대해 10,645 건의 문단과 66,181 개의 질의응답 쌍
  - https://korquad.github.io/KorQuad%201.0/
- **AIHUB Data** : 뉴스 본문 기반 학습 데이터셋 45만 건 중 표준 데이터셋인 질문과 답(25만 건) 사용
  - https://aihub.or.kr/aidata/86

Masking Techniques

1. Question token 기준으로 Random masking

   “부스트캠프 AI Tech 2기 지원은 어디서 해?” → “부스트캠프 <MASK> Tech 2기 지원<MASK> 어디서 해?”

2. Question 단어 기준으로 Random masking

   “부스트캠프 AI Tech 2기 지원은 어디서 해?” → “부스트캠프 AI Tech 2기 <MASK>은 어디서 해?”

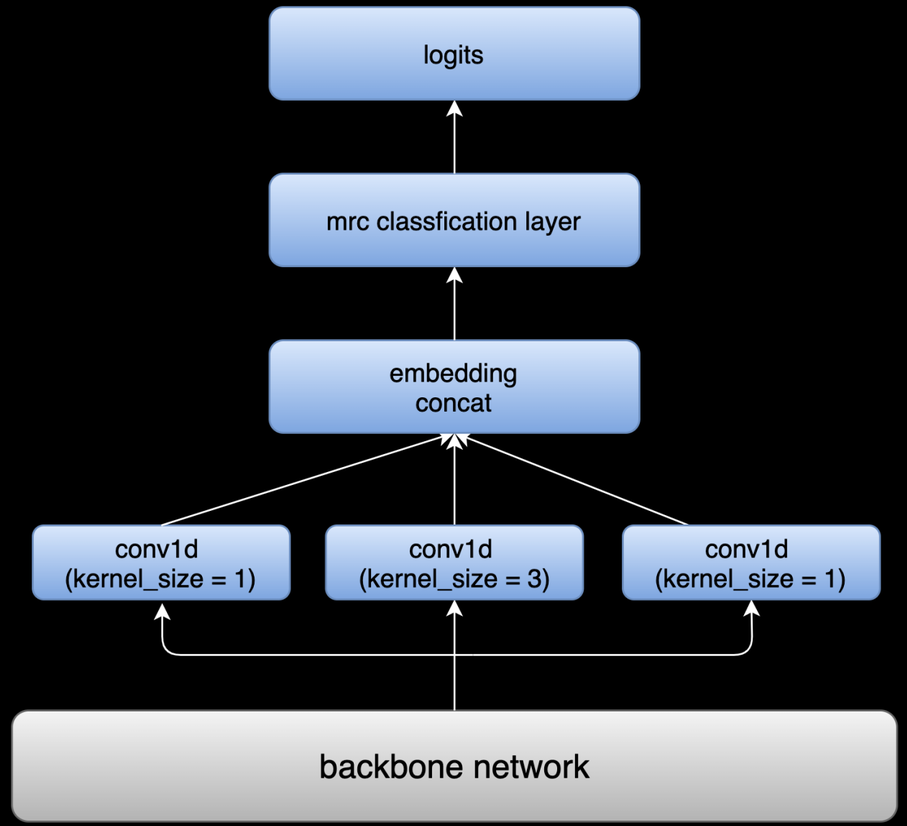
3. Context 중요 단어 기준으로 Random masking

   - SentenceTransformer를 이용해 Question과 유사도 높은 단어 마스킹

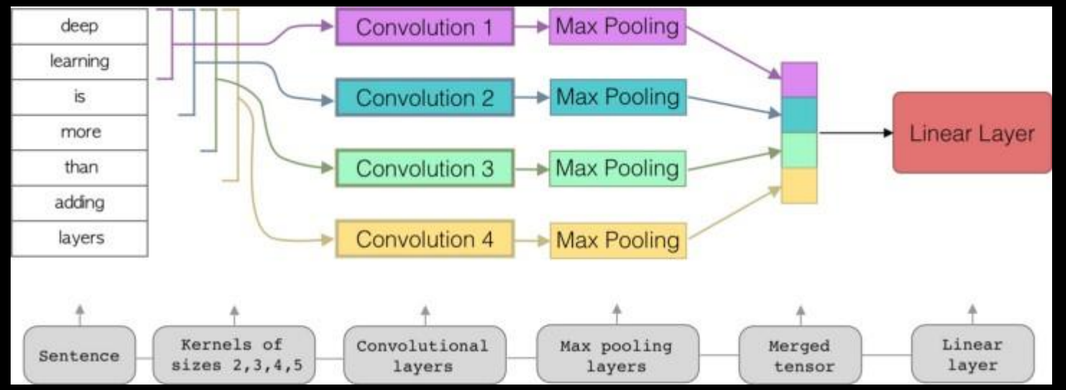
   Q: “이순신의 출생지는?” A: “한성”

   “본관은 덕수, 자는 여해, 시호는 충무였으며, 한성 출신이었다.” → “본관은 덕수, 자는 여해, 시호는 충무였으며, 한성 <MASK>이었다.”

Conv Model

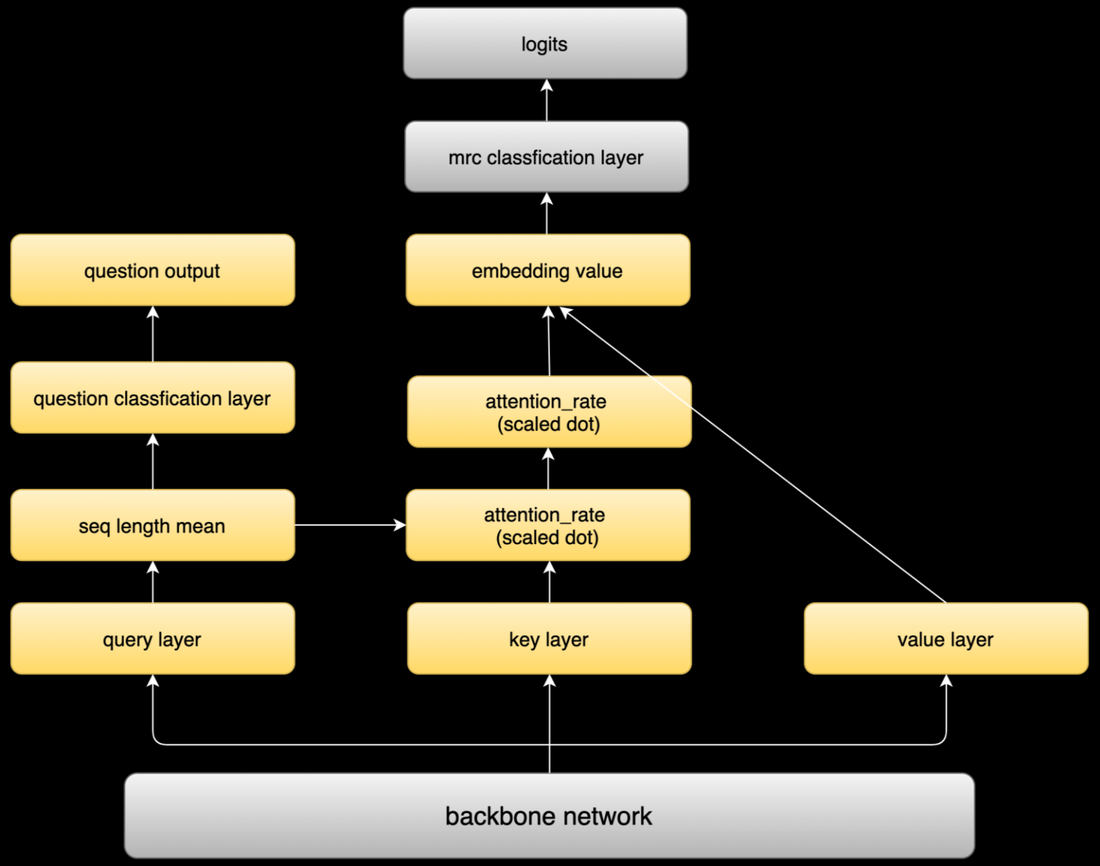


- Convolution 연산을 이용한 **n-gram 기반의 아이디어** 적용



conv model : https://ichi.pro/ko/pytorcheseo-cnneul-sayonghan-tegseuteu-bunlyu-18777046640543

Question Attention Model



- Question에 해당하는 문장의 embedding과 각 토큰의 **attention**을 활용하는 아이디어 적용
- A data를 이용하여 question type classification을 동시에 학습하는 **multi-task learning**

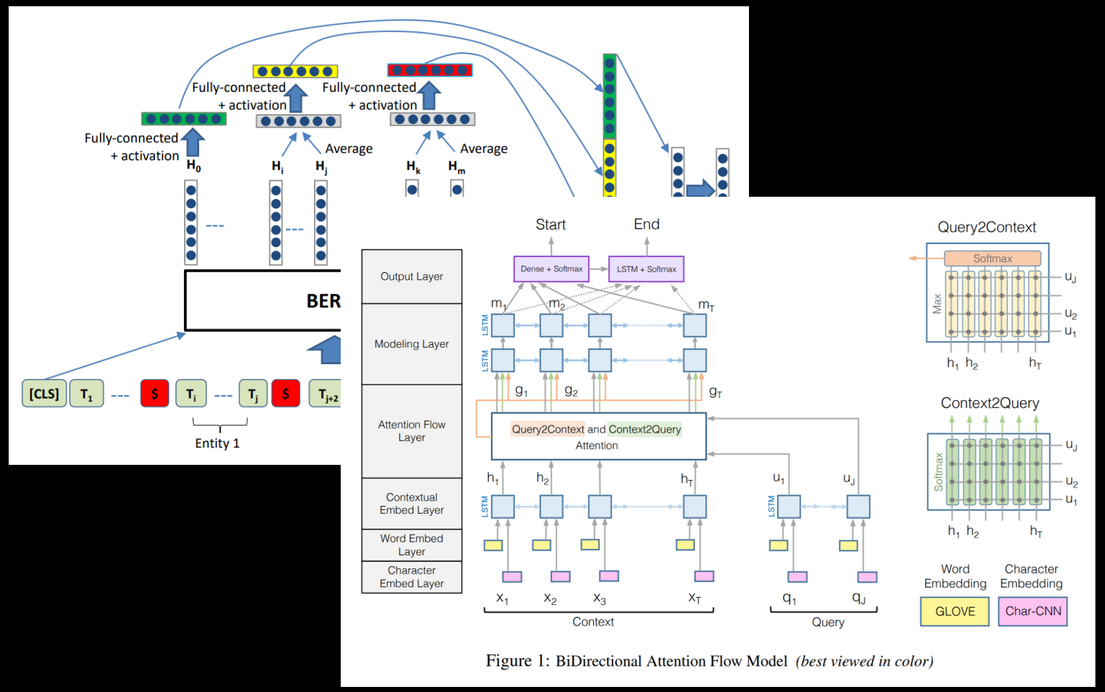
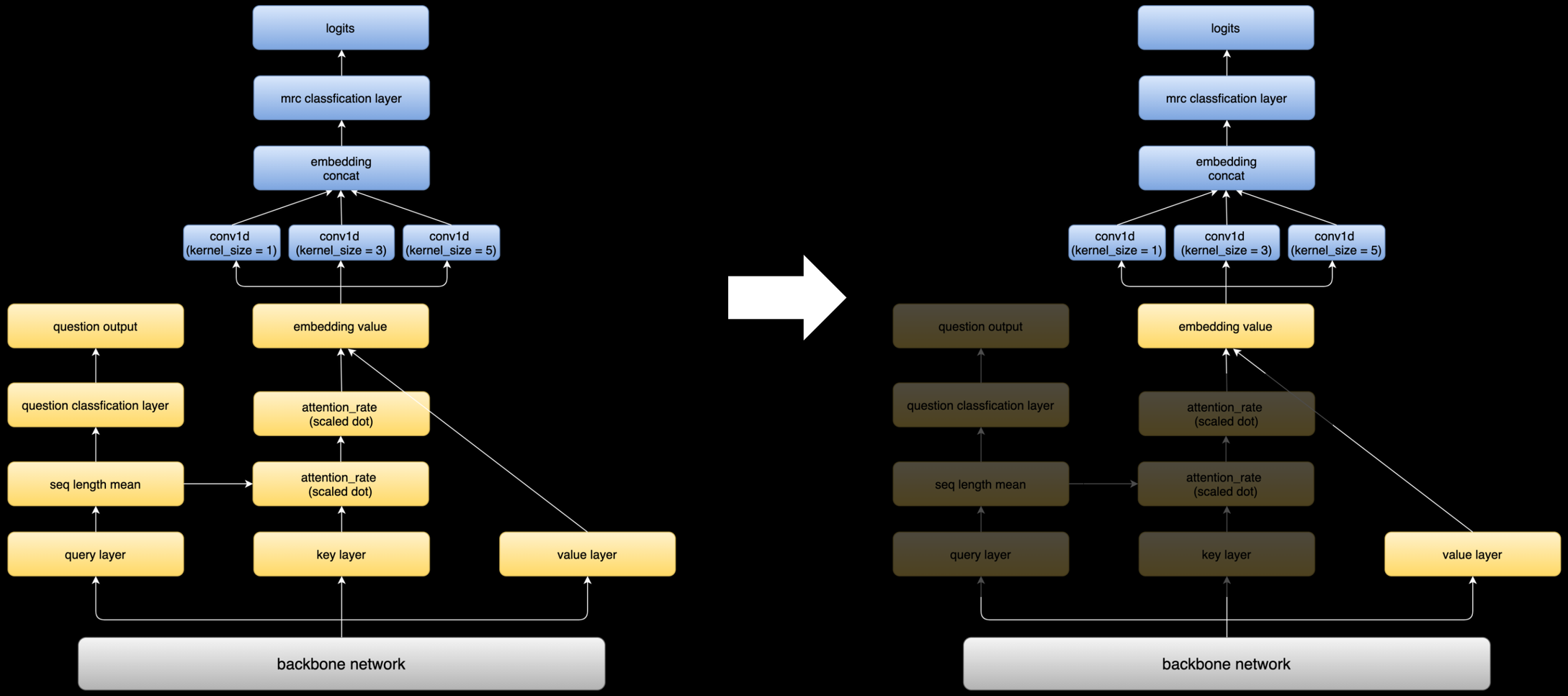


Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

Wu. S. Enriching Pre-trained Language Model with Entity Information for Relation Classification. arXiv:1905.08284

Minjoon Seo. BI-DIRECTIONAL ATTENTION FLOW FOR MACHINE COMPREHENSION. ICLR 2017.

QA-Conv Model



04 Inference

Pre-processing

Wiki Data Split

기준 이상의 길이로 되어 있는 context를 sentence 단위로 나누어 50% 크기를 가진 2개의 context로 나눈 후 elastic search 사용

wiki data split	precision@5	precision@10	precision@15
1000	88%	91%	95%
800	87%	91%	95%
400	86%	90%	94%

Concatenate

elastic search 기준 K = 35개의 context를 concat하여 사용

Sentence Transformer

- Context를 kss 라이브러리로 문장 단위로 분리
- hugging face의 pretrained model을 사용하여 Question과 Context의 문장을 각각 embedding
  (Model name : sentence-transformers/xlm-r-100langs-bert-base-nli-stsb-mean-tokens)
- Question과 Context의 embedding된 값을 코사인 유사도 비교 후 -0.2 보다 낮은 문장은 Question과 관련이 없는 문장으로 판단 후 제거

Post-processing

조사 버리기

```
mrc-0-005407: "숙의 정씨는"
mrc-0-004445: " 2019년 8월 1일에"
mrc-0-000540: " 빌바인의"
mrc-1-000387: " 정보통신윤리회의"
```

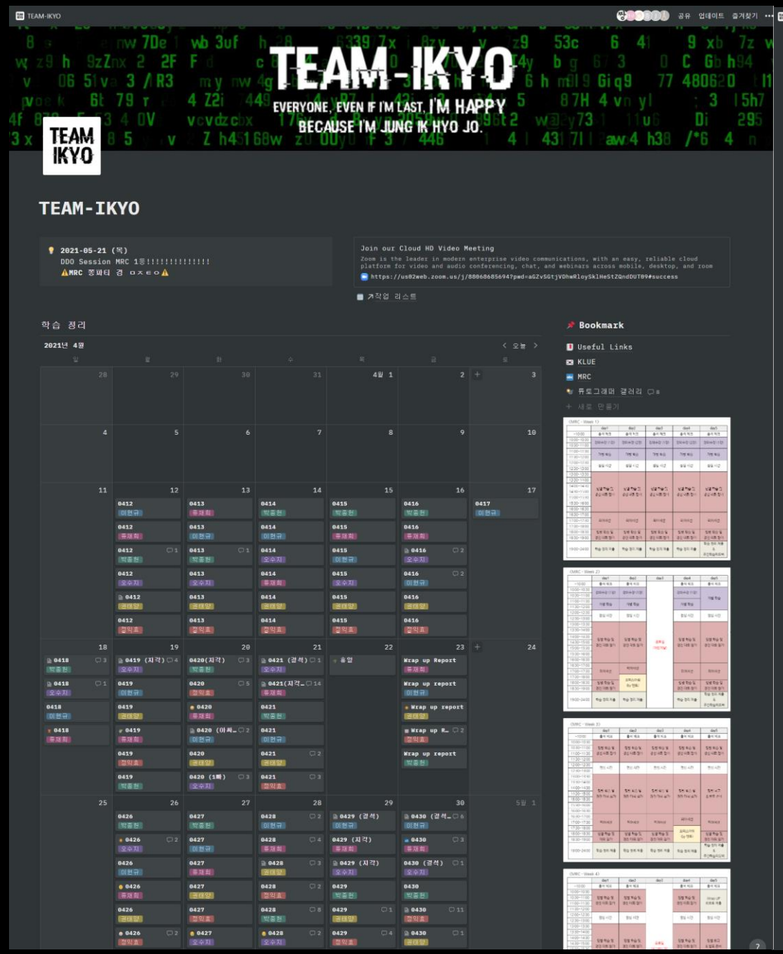
```
mrc-0-005407: "숙의 정씨"
mrc-0-004445: " 2019년 8월 1일"
mrc-0-000540: " 빌바인"
mrc-1-000387: " 정보통신윤리회의"
```

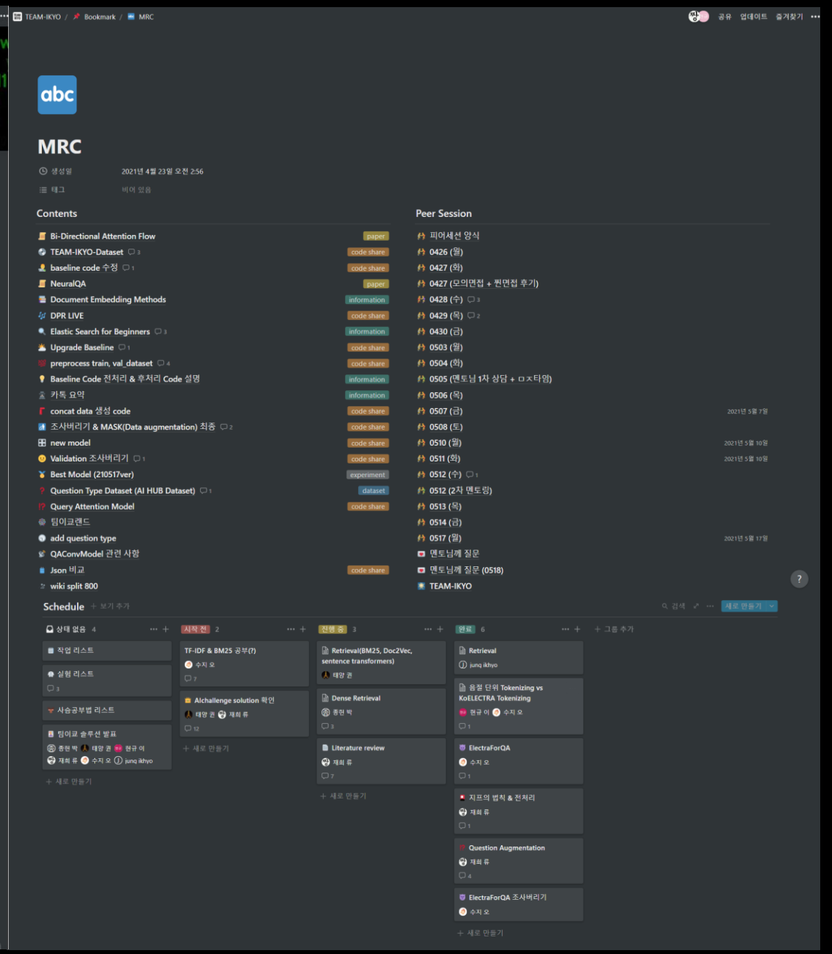
Hard Voting Ensemble

- 단일 모델로 EM score 65% 이상인 13개 모델들에 대해 앙상블
- 사용한 backbone 모델
  - deepset/xlm-roberta-large-squad2
  - a-ware/xlmroberta-squadv2
  - xlm-roberta-large

05 팀이쿄 자랑 

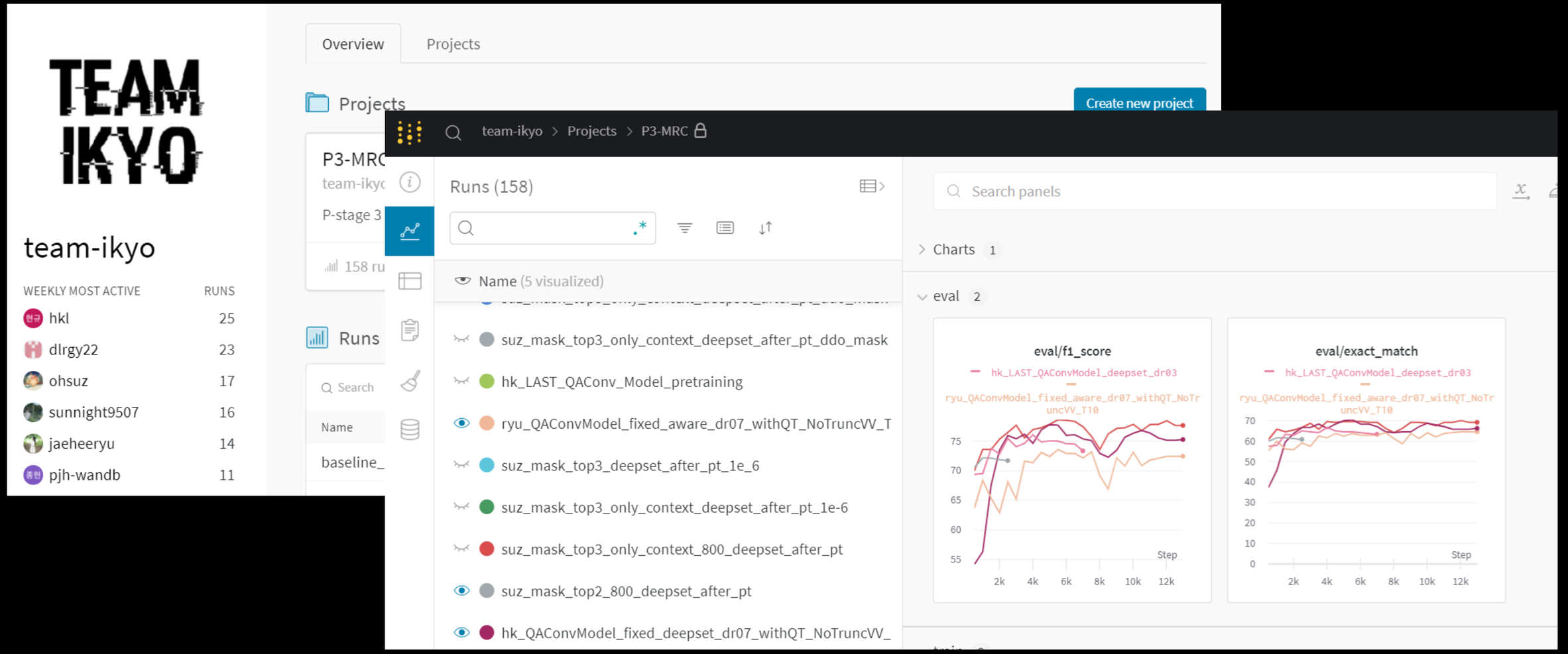
Notion



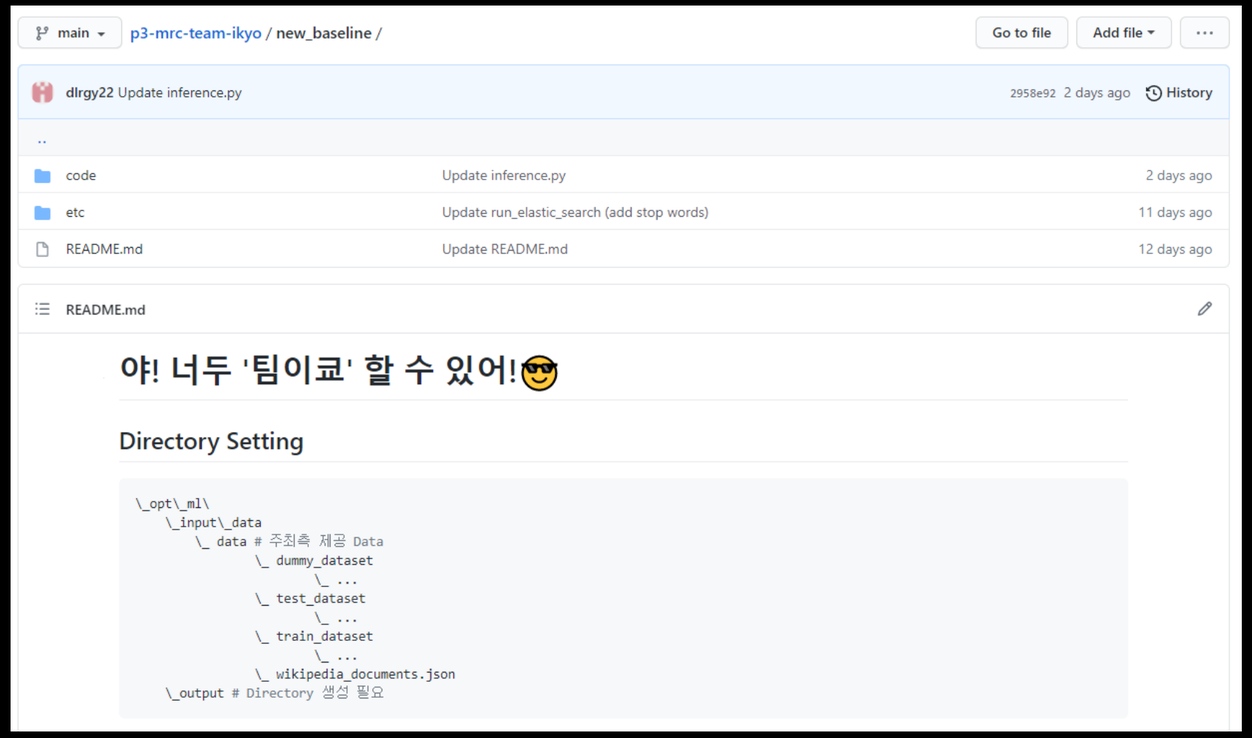


- 학습 정리 공유
- 피어세션 기록
- 각자의 진행 상황 공유 및 기록
- 대회 관련 정보 및 코드 공유
- 실험 리스트 및 결과 정리

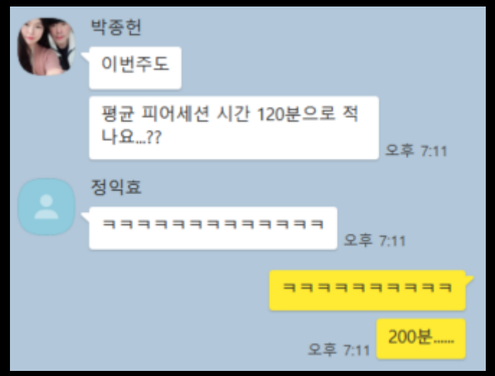
Wandb

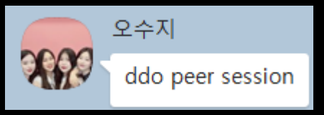

TEAM IKYO
team-ikyo
WEEKLY MOST ACTIVE
RUNS
hkl 25
dlrgy22 23
ohsuz 17
sunnight9507 16
jaeheeryu 14
pjh-wandb 11
Overview
Projects
Projects
Create new project
P3-MRC
team-ikyo
P-stage 3
158 ru
Runs
Search
Name
baseline_
team-ikyo > Projects > P3-MRC
Runs (158)
Name (5 visualized)
suz_mask_top3_only_context_deepset_after_pt_ddo_mask
hk_LAST_QAConv_Model_pretraining
ryu_QAConvModel_fixed_aware_dr07_withQT_NoTruncVV_T
suz_mask_top3_deepset_after_pt_1e_6
suz_mask_top3_only_context_deepset_after_pt_1e-6
suz_mask_top3_only_context_800_deepset_after_pt
suz_mask_top2_800_deepset_after_pt
hk_QAConvModel_fixed_deepset_dr07_withQT_NoTruncVV_
Search panels
Charts 1
eval 2
eval/f1_score
hk_LAST_QAConvModel_deepset_dr03
ryu_QAConvModel_fixed_aware_dr07_withQT_NoTruncVV_T10
75
70
65
60
55
Step
2k
4k
6k
8k
10k
12k
eval/exact_match
hk_LAST_QAConvModel_deepset_dr03
ryu_QAConvModel_fixed_aware_dr07_withQT_NoTruncVV_T10
70
60
50
40
30
20
10
0
Step
2k
4k
6k
8k
10k
12k

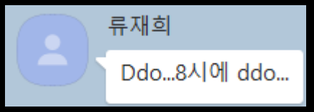
Team Baseline



- 개인의 모델을 구현하는 것이 아니라 하나의 **TEAM 모델** 구현
- 아이디어, 접근 기법 당 1~2명의 담당을 정하고 **파트별로 운영**
- 팀원 모두가 각자의 파트에 대해 비슷한 이해도를 가질 수 있도록 **공유**에 많은 시간 투자 **(feat. ddo 피어세션…)**







Mentoring

멘토님께 질문💌

❤️ Question

1) 최고 성능 모델 공유 및 개선점 도출

Ranking	User Name	EM	F1	Entries	Last
1	정익호_T1189	63.75%	73.36%	61	9 hours ago

⁉️ 현재 팀이코 최고 성능 모델(1등😎)의 구현 디테일은 아래와 같습니다.
혹, 멘토님의 시선에서 조언해주실 만한 개선점이 있다면 조언 부탁드립니다!🙇
(최대한 자세히 작성한다고 써봤는데..
아무래도 더 자세한 디테일은 만나뵙고 설명드려야 할 듯합니다.)

Retrieval Model

- ***Train*** : 정답(GT) context + *Elastic search* 상위 *Top4 concat*
- ***Inference*** : *Elastic search* 상위 *Top20 concat*
- ***Elastic search settings***
  - *Nori tokenizer*(*similarity measure* : *BM25* + 불용어 사전(조사 + 어미))
  - ***Elastic search 최적화 시도***
    - ***BM25*** : 어떤 것을 보고(기준) 어떤 것을 수정해야 하는지 모르겠습니다
      BM25를 활용해서, 추가적인 성능 향상을 위해 어떤 형태의 실험이 가능할까요?
    - ***Elastic search***를 더 잘쓰기 위한 방법은 또 어떤 것이 있을까요?

MRC Model

- ***KorQuAD pretrain*** 후 ***KLUE data train***

  (* 두 훈련 모두 같은 모델을 활용했으며 모델의 디테일은 아래와 같습니다)
  - **backbone network** : *xlm_roberta_large* + ***conv1d + dense layer***
  - ***conv1d + dense layer***

    conv1d(kernel size : 3, padding 1, output_dim : 256) + dense layer(256, 2)

    ⇒ 현재 ***multi_conv1d_layer + output concat*** 하여 성능 검토 진행 중입니다.

    ⇒ ***vanilla convolution layer*** 외에 `residual`, `efficientnet`, `depthwise separable convolution` 등 ***Vision***에서 핫한 개념을 활용해 볼 수도 있을까요?
    관련 사례가 있는지도 궁금합니다.
- ***Post processing***
  - *조사* 버리기(정답 마지막에 Macab 기준의 *조사* 포함 시 제거) + *후처리*
  - *후처리 디테일*(정상 처리되지 않는 조사 제처리)
    1. answer 마지막 글자가 "의"인지 확인 후 조사일 경우 제거
    2. answer의 마지막 token이 "에서", "는", "은", "에" 등 조사일 경우 제거

TEAM git repository

- 팀이코 Team baseline을 정리한 Github입니다!(`join` 환영입니다!💌)

github.com
https://github.com/boostcampaitech1/p3-mrc-team-ikyo

를 9→20으로 늘리니 EM 기준 2% 상
지 의심됩니다. 이렇게 되면 처음 보
의 최고 성능을 신뢰해도 될지 모르

드립니다!

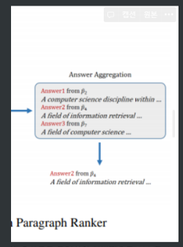
을 찾을 때의 성능 증가의 폭이 훨씬 클

나올 것이다.

인이다.
으니까)
것이다.

board에서는 너무 떨어졌습니다(EM

가 생각해본 해당 접근의 개선 방법입



를 통해 GT context를 찾는 방식의 retrieval을 만들고 있
입니다. 하지만 retrieval 자체의 성능은 조금씩 개선되
하는 것이 맞는 건지 의문입니다. (소요 시간 : 약 2주)
ieval : 56%

시면 감사하겠습니다! ㅎㅎ

어느 정도 성능이 올라와서 안정화 되었다고(감히) 생각
~~upper bound라고 언급하셨지만 그 가목을 개속 깨고~~

향성에 대해 고민이 듭니다.
개선하는 게 맞는지
하는 게 맞을지

능하긴 하지만, 멘토님께서 저희의 입장이라면 지금의 상
어떤 식으로 역할을 분담하는 것이 좋다고 생각하시는지

이 다를 것 같은데, 어떤 부분이 다르고, 현재의 저희가 어

질문을 드려도 될까요?😊

1. passage를 자르는 기준은 *Korean Sentence splitter(KSS)* 사용, overlap은 활용하지 않음
2. context length가 1,000이 넘는 context에 대해 문장 단위로 나누고 전체 길이의 절반에 가깝도록 문장들을 part 1, 2로 나눔
3. 여기서 train 및 validation의 GT answer index가 약간 변형됨
   (***KSS의 특성에 의해 문장 사이 공백이 사라지거나 생김***)

   ⇒ raw context의 index와 위치상 가장 가까운 answer을 GT로 처리

- **팀이코 막내의 개별 요청❣️)** 부스트캠프에서 만난 캠퍼분들과 함께 SKT AI Fellowship 대외활동에 ***"KoBERT/KoGPT/KoBART 기반 언어처리 Application 개발"***이란 주제로 지원을 할 예정입니다. 5월 16일이 마감이라 현재까지 간략하게 연구 계획서를 작성해보았는데 아무래도 학부생끼리 모여 작성하다보니 세부적인 측면에서 아직 많이 부족한 것 같습니다. 혹시 괜찮으시다면 프로젝트의 방향성이나, 연구 계획서에서 어느 점을 보충하면 좋을지 등에 관련해서 간단히라도 조언해주실 수 있을까요?

  혹, 부탁드려도 된다면 빠른 시일 내 연구 계획서 송부 드리겠습니다!🙇
  (아래는 지원 내용 참고용 링크입니다.)

감사합니다*

TEAM-IKYO 권태양 류재희 박종헌 오수지 이현규 정익효

boostcamp