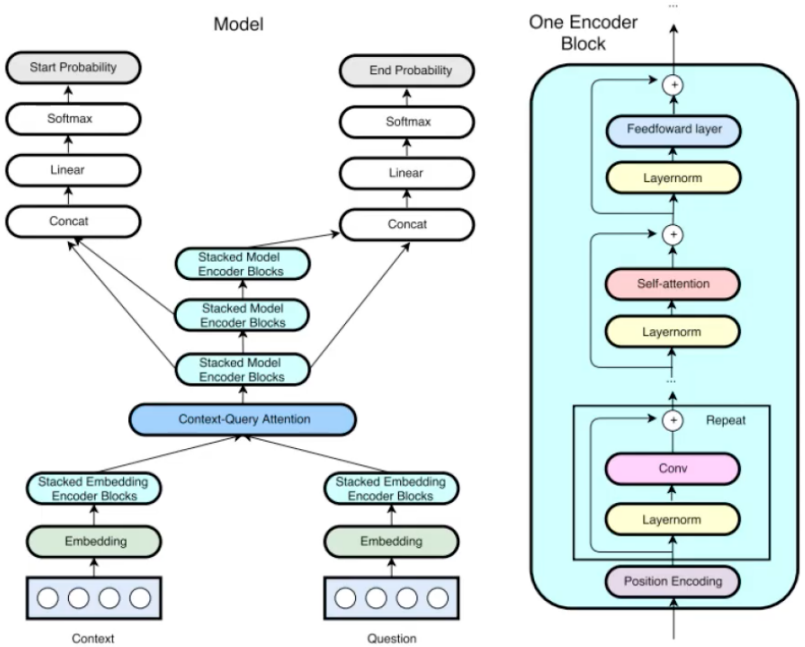
모델 설명

3가지 모델을 구현했다. f1 score는 아래 3 모델 중 3. modified DrQA 에서 구하였다.

1. QANet: **2019_17577_**정지후**_Final_Assignment_1. ipynb** 모든 모듈에 대한 설명이 다 적혀 있는 파일이다**. (**다른 파일들에는 설명 많이 생략함**)** 다른 모델들에서 공통적으로 사용 vocab, train data 등을 저장했다.

   QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension 논문에 나온 모델을 사용했다.
   https://github.com/kushalj001/pytorch-question-answering 여기의 implementation을 변형하여 이용했다.



2. DrQA : 2019_17577_정지후_Final_Assignment_2. ipynb 에는 vocab, train data 등을 직접 구현하지 않고, 이전 파일에서 저장한 데이터셋**, vocab** 등을 그대로 사용하였다**.**
   Reading Wikipedia to Answer Open-Domain Questions 이 논문의 모델이다.
   https://github.com/kushalj001/pytorch-question-answering 여기의 implementation을 변형하여 이용했다.

3. modified DrQA : 2019_17577_정지후_Final_Assignment_3. ipynb 에는 vocab, train data 등을 직접 구현하지 않고, 이전 파일에서 저장한 코드를 그대로 사용하였다.
   앞선 DrQA는 bi directional lstm을 기반으로 만든 모델인 반면, modified DrQA에서는 RNN을 사용하였다. f1 score를 구했다.

결과 분석

전반적으로 inference 성능은 좋지 않았다. 그 이유는 첫째로, 메모리 이슈로 데이터의 극히 일부만 사용했었기 때문이며 둘째로, colab gpu 성능 이슈로 많은 epoch을 training하며 성능을 올리는 것에 실패하였기 때문이다. 그 외의 tokenize가 불완전하다든지 등의 이슈가 있을 수 있다.

F1 score

앞선 결과분석에서처럼 여러 가지 이유로 모델 성능이 좋지 못하여, f1 score가 전반적으로 매우 낮으나, 높은 그룹 (즉 중복도가 높은 그룹) 에서 f1 score가 더 높게 나온 것으로 보아 예상한 결과와 같게 나왔음을 확인할 수 있다.