

# SQuAD 를 활용한 MRC 알고리즘 성능 분석 연구

임중혁

고려대학교 컴퓨터정보통신대학원 빅데이터융합학과  
e-mail : eastjade@korea.ac.kr

## A Study on Performance Analysis of MRC Algorithm Using SQuAD

Jong-Hyuk Lim

Dept. of Big Data Convergence, Korea University Graduate School of Computer & Information Technology

### 요 약

MRC(기계독해)는 Passage, Question, Answer 로 이루어진 Dataset 으로 학습된 모델을 사용하여 요청한 Question 의 Answer 를 같이 주어진 Passage 내에서 찾아내는 것을 목적으로 한다. 최근 MRC 시스템의 성능 측정 지표로 활용되는 SQuAD Dataset 을 활용하여 RNN 의 한 분류인 match-LSTM 과 R-NET 알고리즘의 성능을 비교 분석하고자 한다.

### 1. 서론

최근 Question Answering 시스템에 MRC 엔진의 적용이 활발해짐에 따라 내부 알고리즘 대한 연구 및 개선활동이 활발히 이루어지고 있다. 본 논문에서는 최근 적용되고 있는 알고리즘 중 Match-LSTM 과 R-NET 을 연구하고 SQuAD Dataset 을 활용하여 Training 하고 그 성능을 비교해 보고자 한다.

### 2. Match-LSTM

기존 FFnets(Feed-forward neural networks) 와 같은 전통적 neural network 가 가지는 주요 단점은 시간과 순서를 고려하지 않음에 있었다. (그림 1)

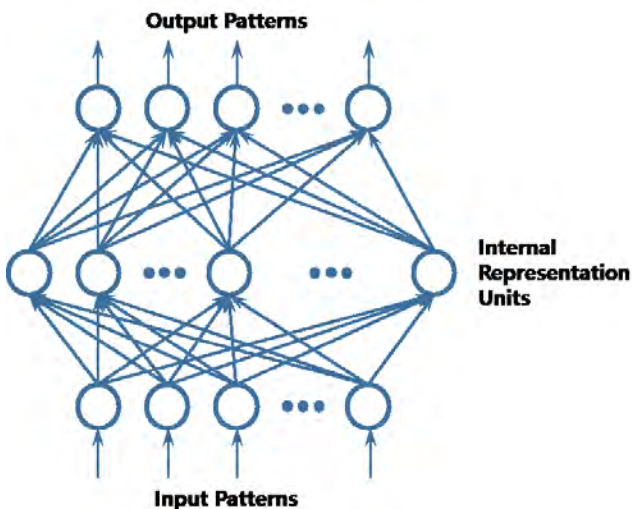


그림 1. FFNETs Architecture

과거의 출력을 다시 입력할 수 있는 소위 피드백구조를 가진 RNNs 이 나타나면서 그러한 단점이 해결되었으나 또 다시 과거의 출력이 거슬러 갈수록 소실되는 Vanishing Gradients 문제에 봉착하게 되었다. (그림 3)

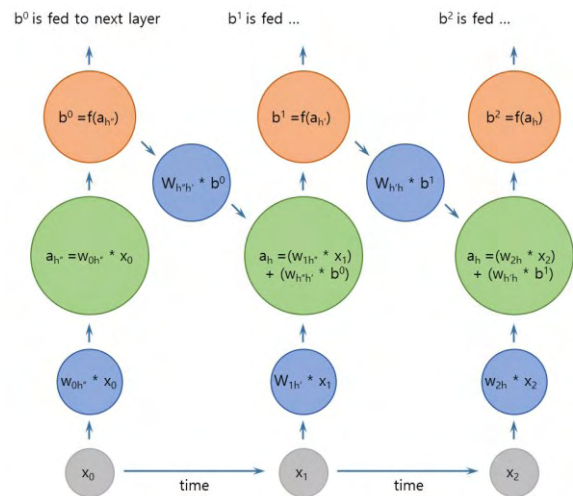


그림 2. RNN Architecture

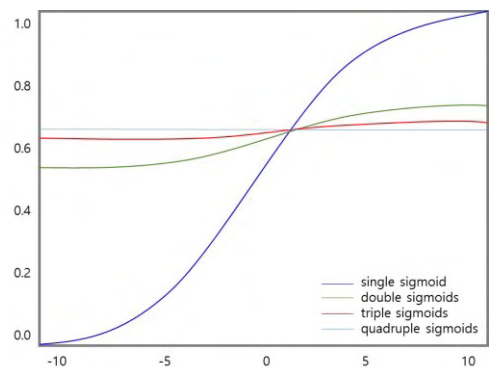


그림 3. Vanishing Gradients

다시 이를 해결하기 위하여 Backpropagation 을 몇단계로 제한한 LSTM(Long Short-Term Memory)으로 발전되었다. 이는 더 나아가 자연어 추론을 위하여 Match-LSTM 으로 발전하였다.

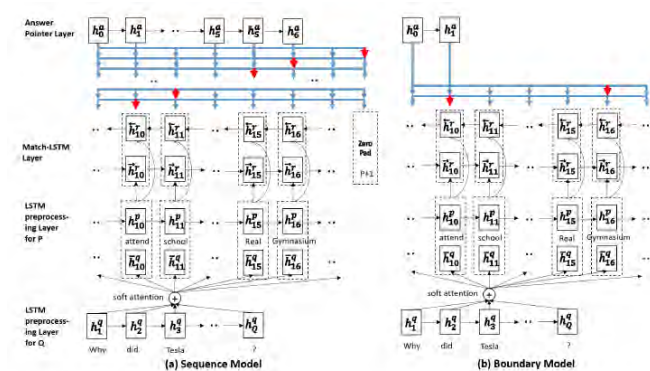


그림 4. Match-LSTM Architecture

### 3. R-NET

Match-LSTM 에서 다시 그림 5 와 같이 1) Passage 와 Question 을 별도로 표현하기 위한 반복적인 Network Encoder, 2) Question 과 Passage 를 일치시키는 Gated matching layer, 3) 전체 Passage 에서 정보를 모으는 Self-matching layer, 4) Network 기반의 응답 경계 예측 layer 구조를 가진 R-NET 알고리즘을 발전하였다..

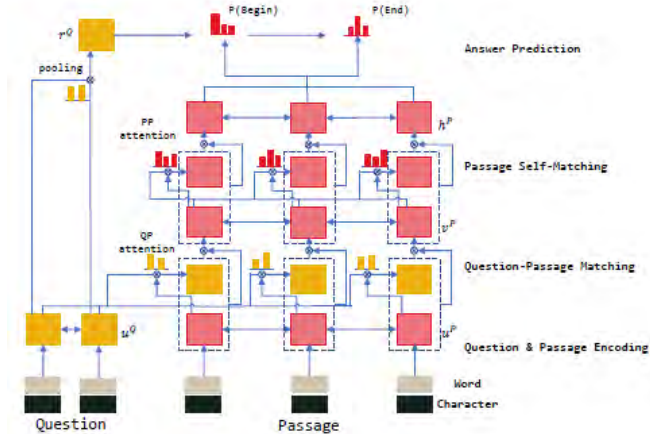


그림 5. R-NET structure overview

### 4. 구현

Match-LSTM 과 R-NET 알고리즘을 구현한 Open Source 를 기반으로 학습, 평가 및 실행 프로그램을 구현하였다. 언어는 Phthon 버전 3.5 로 작성되었고, 프로그램의 구조는 크게 1) SQuAD 데이터 전처리, 2) Tensorflow 를 활용한 반복학습, 3) 학습 모델 Evaluation, 4) 실행으로 구성되어있다.

### 5. 학습 및 평가

각각은 SQuAD Training Set v1.1 으로 학습하고 Dev Set v1.1 으로 Evaluation 을 진행하며 각각 F1 과 EM 을 모두 계산한다.

### 6. 결론

본 논문에서는 질의-답변을 목적으로 하는 MRC 시스템에 활용되는 최신의 알고리즘을 연구하고 그 성능을 비교하였다. 가능 나중에 개선된 알고리즘의 성능이 가장 높게 평가됨이 어찌면 당연하게 판단되겠지만 실용화의 관점에서 학습 데이터의 전처리 용이성과 학습에 소모되는 비용의 차이 또한 크게 고려되어야 할 문제일 것이다.

### 참고문헌

- [1]Hershed Tilak, Yangyang Yu, Michael Lowney, "Match LSTM based Question Answering Codalab Username: yyu10," 2017.
- [2]Natural Language Computing Group, Microsoft Research. Asia, R-NET: Machine Reading Comprehension with Self-matching Networks," ACL 2017.
- [3]Shuohang Wang, Jing Jiang, "Learning Natural Language Inference with LSTM," ArXiv 2016.
- [4]Shuohang Wang, Jing Jiang, "Machine comprehension using match-LSTM and answer pointer," ArXiv 2016.