Attention과 LIME기법을 활용한 순환신경망의 의사결정 요인 분석

윤주성† · 박종철† · 하종수† · 안진현† · 김현철†
† 고려대학교 컴퓨터학과

Attention/LIME method to analyze decision process of RNN

Joo-Sung Yoon† · Jong-Cheol Park† · Jong-Su Ha† · Jin-Hyeon An† · Hyeon-Cheol Kim†
† Dept. of Computer Science and Engineering, Korea University

요 약

딥러닝으로 만들어진 모델의 내부는 black box와 같은 특성을 가져 동작 규칙을 알기 어렵다. 최근 기계학습의 발전으로 인공지능이 전보다 더 복잡한 문제를 해결할 수 있으나 위와 같은 이유로, 모델이 내린 판단의 근거를 알기 어렵다. 그러므로 딥러닝의 동작 규칙을 사람이 이해할 수 있는 형식으로 나타내려는 노력이 필요하다. 본 연구에서는 Attention과 LIME 기법을 활용하여 IMDB 데이터를 감성 분석한 순환신경망의 의사결정 요인을 분석하였다. 각 기법을 활용했을 때의 장단점과 실제 구현에 있어 등장하는 문제에 대해 알아보고자 한다.

1. 서 론

오늘날 인공지능 분야에서 다양한 접근 방법이 시도되고 있으며 그 기술력이 급속도로 발전하여 일부 영역에서는 인공지능이 사람의 능력을 뛰어넘는 단계에까지 이르렀다. 인공지능의 발전으로 더 복잡한 문제들을 해결할 수 있게 되면서, 이미지 인식과 자연어처리와 같은 분야들은 지금도 계속 발전하고 있다. 하지만 딥러닝과 같은 기계학습으로 학습된 모델의 성능이 사람보다 우수해도 결과가 도출된 이유나 근거를 사람이 이해할 수 없다는 black box 문제가 존재한다. Rasmussen과 Williams에 의하면 기계학습 커뮤니티에서 가우시안 과정이 black box 안에서 이루어지는 것은 좋은 예측 결과만 제공한다면 중요하지 않다고 보는 경향이 생겼다고 주장한다[1]. 하지만 모델에 입력이 들어가서 어떠한 과정을 통하여 출력 값이 나오는지에 대한 설명을 제공하지 않고, 사람이 이해할 수 없다면 모델의 결과에 대해 신뢰하기가 어렵다. 본 논문은 이러한 black box 문제를 해결하기 위해 기존의 Attention 기법과 최근에 나온 LIME 기법을 활용하여 순환신경망 (RNN)이 어떠한 방식으로 의사결정 하는지를 이해하는 것을 목표로 한다[2].

본 논문의 구성은 다음과 같다. 2장에서는 이론적 배경을 소개하고, 3장에서는 Attention과 LIME 기법으로 의사결정 요인 분석하는 과정을 설명한다. 4장에서는 실험 및 결과를 분석하고 5장에서는 결론을 맺는다.

2. 이론적 배경

2.1 Attention mechanism

2.1.1 Soft attention

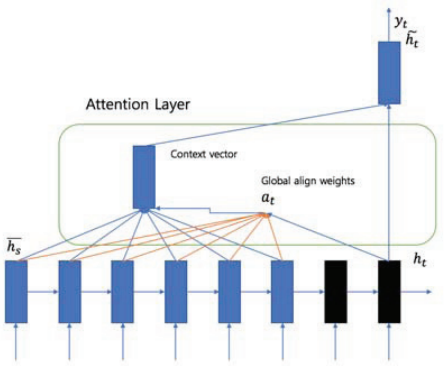
Attention mechanism은 모든 포지션에 대해 attention을 하는 것과 특정 포지션에 대해 attention을 하는 것에 따라 soft attention과 hard attention으로 나뉜다[3]. 본 연구에서는 soft attention을 사용함으로써 전체적인 context를 고려하였다. context를 고려한 hidden state를 h_t라 하고 입력데이터의 문맥 벡터를 c_t라 하자. attentional hidden state는 수식 (1), (2)와 같이 구할 수 있다. 그림1은 이전 context를 고려하는 attention model의 예시다.

$$\tilde{h}_t = \tanh(W_c[c_t;h_t]) \quad (1)$$

$$p(y_t|y_{<t},x) = softmax(W_s\tilde{h}_t) \quad (2)$$

2.1.2 Hierarchical Attention Network (HAN)

우리는 Hierarchical Attention Network를 사용하였다[4]. 이 네트워크는 크게 sentence attention, sentence encoder, word attention, word encoder 네 부분으로 구성되어있다. bidirectional GRU를 사용하였다. GRU는 기존의 순환신경망에 gated unit을 추가한 형태를 말한다[5][6]. 이 유닛의 특징은 LSTM과 다르게 출력 게이트를 사용하지 않는다[7].



[그림 1] Attention mechanism

대신 r_t게이트와 z_t게이트만을 사용한다. 시간 t에서 정보가 어떻게 상태에 갱신되는지 다음과 같이 계산한다.

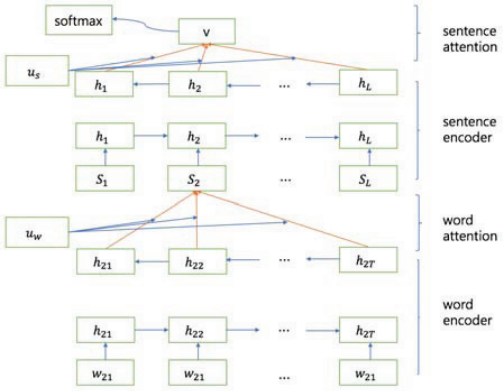
$$h_t = (1-z_t)\odot h_{t-1} + z_t \odot \widetilde{h_t} \quad (3)$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad (4)$$

여기서 x_t 는 time step 당 입력 벡터를 의미한다. $\widetilde{h_t}$는 아래의 과정을 거쳐 갱신된다.

$$\widetilde{h_t} = \tanh(W_h x_t + r_t \odot (U_h h_{t-1}) + b_h) \quad (5)$$

HAN에서는 한 문서를 문장 단위와 단어 단위로 나누어서 처리한다. 한 문서는 L개의 문장을 포함하고 있고 각 문장s_i은 T_i개의 단어를 갖는다. 이때 각 단어는 w_{it}로 나타낸다.



[그림 2] Hierarchical Attention Network

[그림 2]에서 보듯이 encoder 상의 동작은 아래와 같이 표현될 수 있다. (6)번 수식은 word embedding matrix로부터 word embedding을 구하는 과정이며, (7),(8)번 수식은 bidirectional GRU의 hidden state를 계산하는 과정이다[8].

$$x = W_e w, \ t \in [1, T] \quad (6)$$

$$\overrightarrow{h} = \overrightarrow{GRU}(x), \ t \in [1, T] \quad (7)$$

$$\overrightarrow{h_i} = \overrightarrow{GRU}(s_i), \ i \in [1, L] \quad (8)$$

본 모델의 attention은 앞서 설명한 soft attention과 같은 방식으로 동작한다. 또한 word attention 부분과 sentence attention 부분은 단어끼리 attention을 잡고 문장끼리 attention을 잡는 방식으로 동작한다. (9)번 수식은 hidden layer에서 통과된 값을 1-layer MLP를 통과시켜 u벡터와 연산할 벡터를 구하는 과정이며 (10)번 수식은 u벡터와 context vector u_w의 similarity를 구하고 그 값들에 대해서 정규화된 중요도 weight α를 구한 것이다. 이때 학습된 α를 통해 어떤 요소가 attention을 받는지 파악할 수 있고 이를 통해 모델이 의사결정을 내릴 때 중요하게 여기는 요소가 무엇인지 유추할 수 있다. (11)번 수식에서는 sentence vector를 계산하며, 출력값은 sentence encoder의 입력 값으로 사용된다. sentence attention도 이와 같은 과정을 거친다.

$$u_{it} = \tanh(W_w h_{it} + b_w) \quad (9)$$

$$\alpha = \exp(u^T u_w) / \sum_t \exp(u^T u_w) \quad (10)$$

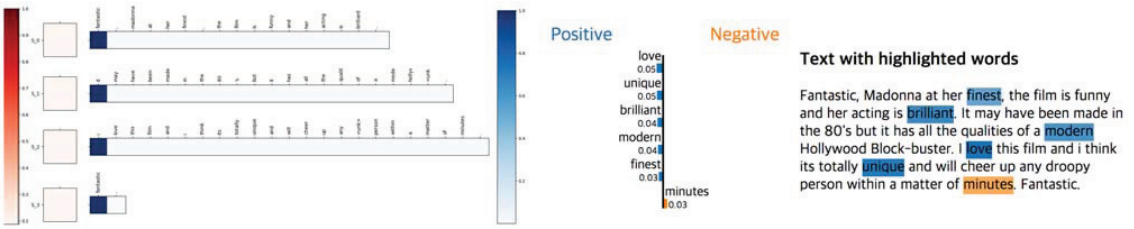
$$s_i = \sum_t \alpha_{it} h_{it} \quad (11)$$

sentence attention에 대한 결과 벡터 v는 document classification을 하기 위해 softmax layer에서 확률값으로 변환되어 class에 대한 예측값을 도출해낸다.

$$p = softmax(W_c v + b_c) \quad (7)$$

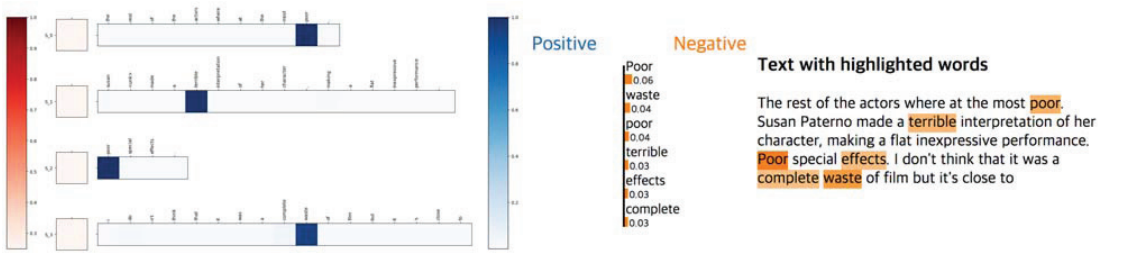
2.2 LIME

현재까지 딥러닝 모델의 설명과 시각화에 집중한 많은 연구가 있었고, 모델을 global understanding 또는 individual understanding을 목표로 한 접근 방식 2가지로 나눌 수 있다. Global understanding을 목표로 한 접근 방식은 주로 인공 신경망의 가중치 행렬과 같은 특성들을 분석하고, individual understanding을 목표로 한 접근 방식은 입력된 데이터를 중심으로 모델의 예측을 분석한다. LIME(Local Interpretable



[그림 3] Attention과 LIME기법의 Heatmap 비교1 - IMDB Positive 리뷰 데이터

Model-Agnostic Explanations)은 후자의 방법으로 데이터 하나를 여러 컴포넌트로 나눠서 각각 컴포넌트의 영향을 linear regression을 통해서 분석하는 기법이다[1]. 컴포넌트들을 조합해서 만든 세그먼테이션 된 인스턴스들의 classification 결과를 y값으로, 컴포넌트들의 유무를 x값으로 만든 linear regression에서 계수를 통해 각 클래스에 끼치는 영향의 정도를 측정한다.



[그림 4] Attention과 LIME기법의 Heatmap 비교2 - IMDB Negative 리뷰 데이터

3. Attention과 LIME기법 실험

본 연구에서는 순환신경망의 의사결정 요인을 분석하기 위해 IMDB 데이터에 대해서 감성 분석을 진행하였다. IMDB 데이터는 영화리뷰와 0부터 10까지의 Rating이 되어있는 라벨 데이터로 구성되어 있다. 본 실험에서는 리뷰데이터 중 0~4까지의 Rating을 갖는 리뷰는 Negative class로, 7~10까지의 Rating을 리뷰는 Positive class로 지정하여 위의 2가지 class에 대해 감성 분석을 하였다. 데이터는 요약하면 <표 1>과 같다[9].

<표 1> IMDB 데이터

Corpus	Positive	Negative
Train	12,500	12,500
Test	12,500	12,500
Total	25,000	25,000

감성 분석에 사용된 모델은 순환신경망의 종류중 하나인 Hierarchical Attention Network를 사용하였다. Hyper parameter 및 모델의 기본 구조는 Zichao가 제안한 구조와 동일하게 구성하였다. 전처리 및 토크나이징은 NLTK를 사용하였으며 본 모델은 PyTorch를 통해 구현되었다. 본 모델을 훈련하기 위해 GTX 1080 8GB GPU를 CUDA와 cuDNN 라이브러리와 함께 사용하였다. IMDB의 테스트 데이터 세트에 대해서 HAN 모델은 81.3%의 Accuracy를 나타냈다.

HAN에서 어떠한 요인을 통해 의사결정을 내리는지 파악하고 비교하기 위해 HAN이 나타내는 Attention을 시각화하고 LIME 기법을 통해 어떤 단어가 Classification probability에 영향을 가장 크게 미치는지에 대해서도 [그림 3][그림 4]를 통해 시각화하였다.

실험 결과 Attention은 Negative class에 대해서 'boring', 'worst' 등 부정적인 단어를 제대로 추출하였지만 Positive class에 대해서는 문장의 맨 앞에 있는 단어에 집중하거나 'i', 'you', 'a'등 다소 중의적인 단어들에 대해서 집중하는 것을 확인할 수 있었다. 이는 Negative class를 결정하는 단어들에만 집중해도 2가지 class에 대해서는 잘 분류할 수 있기 때문으로 보인다. 반면, LIME의 경우 Positive class에서 'love', 'brilliant' 등 긍정적인 단어들을 제대로 추출한 것을 확인할 수 있었고 Negative class에서는 Attention과 비슷한 단어들을 추출한 것을 확인할 수 있었다. 이러한 결과로 미루어볼 때, HAN이 Positive class를 갖는 문장의 경우 긍정적인 단어에 Attention이 많이 주어지지 않지만, 긍정적인 단어가 classification probability에 여전히 영향을 주고 있다고 여겨진다. 또한, 각각의 class에 대해서 순환신경망이 어떤 요인에 의해서 의사결정을 하는지를 더 정확하게 파악하기 위해서는 Attention보다는 LIME 기법이 더 정확한 성능을 보일 수 있음을 유추할 수 있다.

4. 결론 및 논의

본 연구는 Neural Network의 black box 문제를 해결하기 위해 순환신경망이 의사결정을 할 때 중요하게 여기는 요소에 대해서 분석을 진행하였다. IMDB 데이터에 대해서 감성 분석을 하였으며, 순환신경망이 Attention과 LIME기법을 통해 어떤 단어를 중요하게 여기는지에 대해서 분석함으로써 순환신경망의 의사결정 요인을 분석하였다. Negative class에 대해서는 Attention과 LIME 모두 비슷한 단어에 초점을 맞췄지만, Positive class에 대해서는 LIME이 더 정확한 결과를 나타냈다. 하지만 실험을 진행하면서 LIME은 분석시간이 Attention에 비해 오래 걸리는 문제가 있음을 발견했다. Attention과 LIME의 결과가 비슷한 특정 class에 대해서 분석하는 경우에는 Attention기반의 기법도 활용가능 할 것으로 보인다. 향후 연구에서는 토픽 카테고리 분류 문제, 소스 코드에 대한 기능성 분류 문제에 본 연구를 적용해보고자 하며, LIME의 분석 속도와 Attention을 개선하기 위한 연구 또한 진행해보고자 한다.

감사의 글

" 이 논문은 2017년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2017R1A2B4003558)."

참 고 문 헌

[1] Rasmussen, C. E., &Williams, C. K. (2006). *Gaussian processes for machine learning* (Vol. 1). Cambridge: MIT press.

[2] Ribeiro, M. T., Singh, S., &Guestrin, C. (2016, August). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144). ACM.

[3] Luong, M. T., Pham, H., &Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

[4] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. J., &Hovy, E. H. (2016). Hierarchical Attention Networks for Document Classification. In *HLT-NAACL* (pp. 1480-1489).

[5] Bahdanau, D., Cho, K., &Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[6] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

[7] Sak, H., Senior, A., &Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth Annual Conference of the International Speech Communication Association*.

[8] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

[9] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 142-150). Association for Computational Linguistics.