인공신경망 은닉층 해석을 위한 토픽과의 비교

정영섭
순천향대학교 SCH 미디어랩스 빅데이터공학과
e-mail : bytecell@sch.ac.kr

Comparison Between Hidden Layers of Neural Networks and Topics for Hidden Layer Comprehension

Young-Seob Jeong
Dept. of Bigdata Engineering, Soonchunhyang University

요 약

데이터의 양이 증가하면서 인공신경망을 통한 데이터 분석 기술이 주목받고 있으며, 텍스트, 그림, 동영상 등에 이르기까지 다양한 종류의 데이터를 자동으로 분석하여, 번역기, 채팅봇, 그림 캡션 자동 생성 등에 대한 연구 및 서비스 개발에 활용되고 있다. 인공신경망 기반으로 수행된 많은 연구들이 공통적으로 가진 한계가 있는데, 그것은 은닉층에 대한 해석이 어렵다는 것이다. 가령, 입력층, 은닉층, 그리고 결과층으로 이루어진 인공신경망을 임의의 데이터로 학습시키면, 입력층과 은닉층 사이에 존재하는 행렬은 해당 데이터에 존재하는 패턴 정보를 내포하게 된다. 따라서, 행렬에 존재하는 패턴 정보를 직접 분석할 수 있다면, 인공신경망 결과물에 대한 해석이 가능할 뿐만 아니라 성능을 높이기 위해 어떤 조정이 필요한지에 대한 직관도 얻을 수 있을 것이다. 하지만, 이 행렬의 실체는 숫자로 이루어진 벡터이므로 사람이 직접 해석하는 것은 불가능하며, 지금까지 수행되어온 대부분의 인공신경망 연구들은 공통적으로 이러한 한계점을 가지고 있다. 본 연구는 데이터에 존재하는 패턴을 잡아내면서도 해석이 가능한 토픽 모델과 인공신경망의 결과물을 비교함으로써, 인공신경망 은닉층 해석에 대한 실마리를 찾기 위한 연구이다. 실험을 통해 토픽과 은닉층 패턴의 유사성을 검증하고, 향후 인공신경망 연구에서 은닉층 해석에 대한 가능성을 논한다.

1. 서론

데이터의 양이 폭발적으로 증가하면서, 최근 인공신경망을 통한 데이터 분석 연구가 활발히 수행되고 있다. 특히, 텍스트, 그림, 동영상 등의 다양한 종류의 데이터를 자동으로 분석함으로써, 번역기, 채팅봇, 인공지능 스피커, 그림캡션 자동생성, 자율주행 등에 대한 서비스 개발에 활용되고 있다 [1,2,3,4,5,6]. 인공신경망 모델은 사람의 뇌를 구성하는 신경망을 본떠서 제시되었으며, 충분한 양의 데이터가 주어질 경우, 사람이 미처 잡아내지 못하는 세부적인 패턴들도 자동으로 추출해준다. 인공신경망 모델들이 다양한 분야에 성공적으로 적용되고 있음에도 불구하고, 대부분의 모델들이 공통적으로 가진 한계가 있는데, 그것은 은닉층에 대한 해석이 어렵다는 점이다. 가령, 다수의 문장 데이터를 Convolutional Neural Networks (CNN) [7]에 입력하여 문장의 종류, 이를테면 명령문, 평서문 등을 맞추는 분류기를 만든다고 가정할 경우, CNN 모델의 은닉층은 문장 종류를 판가름하는 데 유용한 임의의 패턴들을 표현하게 된다. 따라서, 은닉층과 가시층을 연결하는 행렬 값들을 사람이 이해할 수 있게 된다면, 데이터로부터 인공신경망의 결과물에 이르기까지의 인과관계에 대한 직관을 얻을 수 있으며, 나아가 인공신경망 모델의 성능 향상을 보다 효과적으로 성취하게 될 것이다.

대부분의 인공신경망 모델들을 활용한 분류기들이 높은 성능을 보여줌에도 불구하고, 은닉층에 대한 해석이 어렵기 때문에 인공신경망 모델이 성공적으로 동작하는 정확한 이유를 파악하지 못하는 경우가 많으며, 성능 개선 방법을 구상하는 것도 매우 어렵다. 반면, 데이터에 내재된 패턴을 자동으로 추출 가능한 또 다른 기술인 토픽 모델은 그 결과물인 토픽을 사람이 직접 해석이 가능하다. 토픽 모델을 학습함으로써 얻는 토픽은 weight 가 매겨진 단어들의 군집으로 표현되는데, 높은 weight 를 가진 상위 단어들을 통해 각 군집(토픽)을 사람이 보고 해석할 수 있는 것이다.

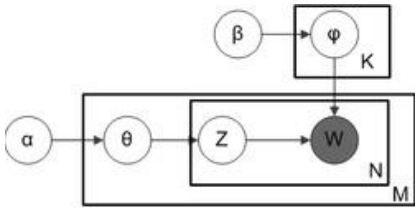
본 연구에서는, 같은 데이터에 대하여 학습된 대표적인 토픽모델인 Latent Dirichlet Allocation (LDA) [8] 모델의 결과물 토픽과 인공신경망 모델의 일종인 Restricted Boltzmann Machine (RBM) [9] 모델의 결과물인 가시층과 은닉층 사이의 행렬 값을 비교함으로써, 인공신경망 은닉층 패턴을 분석하는 것이 과연 가능한 것인지, 그리고 그것이 유의미할 것인지에 대한 가능성을 논한다.

본 논문의 2 장에서는 본 연구에 사용되는 토픽모델과 인공신경망 모델에 대한 배경지식을 소개하며, 3 장에서는 실험내용 및 결과를 분석하고, 4 장에서는 결론을 맺는다.

2. 배경

2.1 토픽 모델

토픽 모델은 데이터에 내재된 패턴을 자동으로 추출하는 모델로서, 주로 베이지안 모델로써 설계된다. 이 때 추출되는 패턴은 데이터들이 함께 등장하는 빈도수를 기반으로 데이터들을 군집화함으로써 표현된다. 예를 들어, 텍스트 데이터에서 '짜장면', '짬뽕', '우동' 등이 함께 등장하는 빈도가 컸을 경우, 이들을 임의의 군집으로 묶어줄 수 있으며 다른 단어들에 비해 높은 weight 를 가지는 상위 단어로 취급된다. 대표적인 토픽 모델로는 Probabilistic Latent Semantic Analysis (PLSA) [10] 와 Latent Dirichlet Allocation (LDA) [8] 등이 있으며, 이를 기반으로 다양한 목적을 가진 새로운 토픽 모델들도 설계되었다 [11,12].

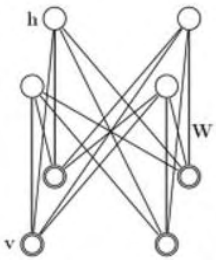


(그림 1) LDA 모델의 구조.

본 연구에서 실험에 사용될 토픽 모델은 LDA 모델이며, 그 구조는 그림 1 과 같다. K 는 토픽 개수, M 은 문서 개수, 그리고 N 은 문서 내의 단어 개수이며, 박스 내부는 반복됨을 의미한다. 예를 들어, 파라메터 Θ는 각 문서의 토픽 분포를 의미하며, 랜덤변수 z 는 각 단어 w 가 관련된 임의의 토픽을 의미한다. 단어를 의미하는 w 만 음영이 되어있는 이유는, 단어는 데이터에서 관측이 가능하다는 것을 표현하기 위함이다. 달리 말하면, LDA 모델은 단순한 텍스트 데이터만을 사용하여 학습이 가능한 모델이라는 의미가 된다. 파라메터 φ 는 토픽으로서, 함께 자주 등장했던 단어들의 군집으로써 표현된다. 하이퍼 파라메터 α 와 β 는 각각 Θ와 φ 에 대한 사전정보를 제공하는 용도로써 사용된다. 그림에서 눈여겨볼 점 한가지는, 하이퍼 파라메터 α 와 β 로부터 단어 w 에 이르기까지 방향선으로 연결되어있다는 것이다. 이 방향선을 통해, 데이터가 어떻게 생성되었을 것이라는 '가정'을 표현하게 되며, LDA 모델을 비롯한 모든 토픽 모델은 저마다의 고유한 '가정'을 가지고 있다.

LDA 모델을 학습할 때 사람이 정해줘야 하는 4 가지가 있다. 첫째, 토픽 개수 K 를 정해줘야 한다. 둘째, 하이퍼 파라메터 α 와 β 의 초기값을 설정해야 한다. 셋째, 각 단어 w 에 대한 랜덤 변수 z 의 초기값을 설정해야 한다. 넷째, 파라메터 Θ와 φ 를 학습시키기 위해 몇 번의 iteration 을 수행할지 결정해야 한다. 파라메터 학습 과정에는 보통 Variational Inference 또는 collapsed Gibbs sampling [8,13] 을 이용하는데, 두 방법 모두 iterative 방식으로 파라메터를 학습하므로, 적절한 iteration 횟수를 결정해주어야 한다.

LDA 모델을 학습시켜 얻은 파라메터 φ 는 각 토픽을 의미하며, 토픽은 weight 가 매겨진 단어들의 군집으로써 표현된다. 이 때, weight 값이 큰 상위 단어들을 열거함으로써 해당 토픽이 어떤 내용에 대한 것인지 이해하는 것이 가능하다. 예를 들어, '짜장면', '짱뽕', '우동' 등이 상위 단어로 존재하는 토픽은 '중국음식'이라는 주제라고 이해할 수 있을 것이다. 학습의 결과물로 얻은 또 다른 파라메터 Θ는 각 문서의 토픽 분포를 표현한다. 가령, 2 개의 토픽 '중국음식'과 '한국음식'이 존재할 경우, 각 문서는 이 두 개의 토픽이 각각 어느정도의 비율로 다뤄지고 있는지를 토픽 분포로써 표현하게 되는 것이다. LDA 모델은 위 두 개의 파라메터로써 데이터에 내재된 패턴을 표현하며, 문서 요약 [14], 사용자 생활패턴 분석 [15] 등에 활용될 수 있다.



(그림 2) RBM 모델의 구조.

2.2 인공신경망 모델

인공신경망 모델은 사람의 뇌를 구성하는 신경망을 본떠서 제시된 모델로서, 각 뉴런에 입력되는 신호부터 출력으로 내보내는 신호에 이르기까지의 과정을 그대로 모사하고 있다. 고전적으로 많이 사용된 인공신경망 모델은 Artificial Neural Networks (ANN) [16,17] 과 Auto Encoder [18] 이며, 최근에는 Deep Belief Networks (DBN) [19], Convolutional Neural Networks (CNN) [7], 그리고 Recurrent Neural Networks (RNN) [20] 등이 이미지 인식, 음성 인식 등의 분야에서 획기적인 성능 향상에 이바지하면서 많은 주목을 받고 있다. 하지만, 이러한 인공신경망 모델들은 공통적인 한 가지 한계점을 가지고 있는데, 그것은 은닉층에 대한 해석이 매우 어렵다는 것이다. 예를 들어, CNN 모델을 사용하여 문서 데이터 혹은 이미지 데이터를 분석할 경우, 은닉층과 다른 층을 연결하는 행렬에는 데이터에 내재된 임의의 패턴 정보가 표현될 것이다. 따라서, 행렬에 존재하는 패턴 정보를 직접 분석할 수 있다면, 인공신경망 결과물에 대한 해석이 가능할 뿐만 아니라 성능을 높이기 위해 어떤 조정이 필요한지에 대한 직관도 얻을 수 있을 것이다. 하지만, 이 행렬의 실체는 숫자로 이루어진 벡터이므로 사람이 직접 해석하는 것은 불가능하며, 지금까지 수행되어온 대부분의 인공신경망 연구들은 공통적으로 이러한 한계점을 가지고 있다. 따라서, 인공신경망 모델의 학습에서 얻는 행렬 값들을 직접 분석하고 해석할 수 있는 방법을 찾는 연구가 필요하다.

본 연구에서 실험에 사용될 인공신경망 모델은 Restricted Boltzmann Machine (RBM) [9] 이다. RBM 모델은 DBN 모델의 기초가 된 모델로서, 그림 2 와 같이 가시층과 은닉층 사이의 bipartite graph 구조를 지닌다. 그림에서 노드 v 를 포함한 아래쪽이 가시층이며, h 를 포함한 위쪽이 은닉층인데, 같은 층에 존재하는 노드들 사이에는 연결이 없고, 가시층과 은닉층 사이에만 연결 W 가 존재하는 이 구조로 인해 'restricted'라는 명칭이 붙게 되었다.

RBM 모델을 학습할 때 사람이 정해줘야 하는 3 가지가 있다. 첫째, 은닉층에 존재할 노드의 개수를 정해줘야 한다. 둘째, 연결 W 의 초기값을 설정해야 한다. 셋째, iteration 횟수 등의 제반사항을 결정해야 한다. RBM 모델의 학습에는 주로 Contrastive Divergence (CD) 기법 [21] 이 사용되는데, 이 기법은 가시층 노드들의 값을 바탕으로 은닉층 노드들의 값을 Gibbs sampling 한 후, 은닉층 노드들의 값을 바탕으로 가시층 노드들의 값을 다시 Gibbs sampling 하는 방식으로 동작하며, negative phase 의 횟수 n 에 따라 CD-n 이라고 표기한다.

RBM 모델의 학습에 문서 데이터를 사용할 경우, LDA 모델과 동일하게, 문서의 단어들만을 사용하여 파라메터 W 를 추론하게 된다. 학습의 결과물로 얻는 파라메터 W 값은 가시층과 은닉층 사이를 연결하는 weight 값이므로, weight 가 매겨진 단어들의 군집으로써 표현되는 LDA 모델의 토픽과 비교 가능하다. 달리 말하면, RBM 모델의 결과물인 행렬은 해석이 가능하며, LDA 모델의 결과물인 토픽들과 비교함으로써 인공신경망 모델의 행렬 값에 대한 해석이 가능할 것인지, 또한 그 해석이 유의미할 것인지 탐구할 수 있는 것이다.

3. 실험

본 연구는 LDA 모델의 결과물인 토픽과 RBM 모델의 결과물인 은닉층-가시층 사이의 행렬 값을 비교함으로써 은닉층에 내재된 패턴에 대한 사람의 이해가 가능한지를 살피고, 그 해석이 유의미할 것인지를 논하는 데 있다. 이를 위해, 같은 텍스트 데이터에 대하여 LDA 모델과 RBM 모델을 학습시킨 후, weight 가 매겨진 단어 리스트를 비교하도록 한다. 데이터는 소설 어린왕자 영문판을 사용하였으며, 소설에 존재하는 27 개의 챕터를 각각 문서라고 취급하여 LDA 모델을 학습하였다. 문서 단위로 단어들이 함께 등장하는 패턴을 추출하는 LDA 모델에 비교했을 때, RBM 모델은 각 문장 단위로 학습이 이루어지므로, 보다 공정한 실험을 위해 각 문장을 문서로 취급하여 LDA 모델 학습을 추가로 수행하여 비교하였다.

3.1 LDA 모델 결과물

LDA 모델 학습을 위해 토픽 개수 K 는 10 으로 설정하였으며, 하이퍼 파라메터들은 0.1, 0.001 로 설정하였다. 파라메터 추론을 위해 collapsed Gibbs sampling 기법을 이용하여 1000 번의 iteration 을 수행하였으며, 결과물로 얻은 토픽의 단어들은 TF-IDF 알고리즘을 토픽 단위로 적용하여 재정렬하였다. 10 개 토픽들 중의 일부 토픽 3 개에 대한 상위 5 개의 단어 리스트는 표 1 과 같다. 토픽의 해석은 일반적으로 사람이 직접 단어 리스트를 확인하여 이루어지며, 전반적으로 봤을 때 LDA 모델은 소설 어린왕자에서 다루는 내용 혹은 토픽들이 무엇인지 사람이 해석 가능한 형태로 결과물을 생성해내는 것을 확인할 수 있었다. 하지만, 소설 어린왕자의 각 챕터를 문서로 취급하여 LDA 모델을 학습한 반면, RBM 모델은 각 문장 단위로 학습이 이루어지므로 공정한 비교를 위해 LDA 모델을 각 문장 단위로 학습한 결과물도 확인하였으며, 일부 토픽 3 개에 대한 상위 단어 리스트는 표 2 와 같다.

<표 1> 어린왕자 챕터를 문서로 취급했을 때의 LDA 모델 결과물 예시.

토픽이름	명령하는 왕	그림 그리기	여우,장미와의 만남
Rank 1	King	Draw	Fox
Rank 2	Order	Sheep	Me
Rank 3	Lamp	Boa	Rose
Rank 4	Sunset	Grown-up	Tame
Rank 5	Lamplighter	Constrictor	Men

<표 2> 어린왕자 문장 단위로 학습한 LDA 모델 결과물 예시.

토픽이름	명령하는 왕	그림 그리기	친구맺기
Rank 1	King	Draw	Me
Rank 2	Order	Time	Make
Rank 3	Ah	Year	Great
Rank 4	Explorer	Boa	Friend
Rank 5	Geograph	Grown-up	Stars

챕터 단위로 학습한 LDA 모델 결과에 비교했을 때, 문장 단위로 학습한 LDA 모델의 결과물은 토픽들의 상위 단어가 하나의 토픽을 명확히 드러내지 못하여 coherence 가 높지 않았다. 예를 들어, 그림 그리기에 대한 토픽에서 time, year 등과 같은 단어들이 섞여있었다. 또한, 문서 단위로 패턴을 보지 못하므로, 여우, 장미와의 만남에 대한 토픽과 같이 일부 토픽은 아예 존재하지 않았다. 따라서, 전반적으로 LDA 모델을 문장 단위로 학습하면 챕터 단위로 학습한 경우에 비해 질적인 면에서 결과물이 부족했다.

3.2 RBM 모델 결과물

<표 3> 어린왕자 문서로 학습한 RBM 모델 결과물 예시.

토픽종류	토픽 1	토픽 2	토픽 3
Rank 1	Prince	Flower	Planet
Rank 2	Made	King	Prince
Rank 3	Me	Men	Man
Rank 4	Back	Conceit	King
Rank 5	Life	Prince	Time

RBM 모델은 소설 어린왕자에 등장했던 모든 단어들의 사전 V 를 구축한 후, 각 문장들에 등장한 단어들은 1, 그렇지 않은 단어들은 0 을 가지는 이진 벡터가 학습에 이용되었다. 학습 알고리즘은 CD-1 를 이

용하였으며, 1000 번의 iteration 수행의 결과물은 표 3 과 같다. RBM 모델의 결과물과 LDA 모델을 문장 단위로 학습시킨 결과물을 비교하면, LDA 모델의 결과물이 어린왕자 소설의 내용을 더 잘 표현하고 있음을 알 수 있었다. RBM 모델의 결과물과 LDA 모델의 결과물의 차이는 모델 구조적인 부분에서 온다고 해석할 수 있다. LDA 모델을 문장 단위로 학습한다고 하더라도, LDA 모델은 문장 단위의 토픽 분포를 모델링하는 반면, RBM 모델은 단순히 단어들의 빈도수만으로 weight 를 계산하기 때문이다. 즉, 모델 구조적인 차이가 결국 토픽 결과물에 대한 큰 영향을 끼쳤다고 해석할 수 있다. 따라서, 단순히 단어 빈도수에 기반하여 은닉층 행렬 값을 모델링하는 RBM 모델이 아닌, 순차적인 정보와 같은 부가적인 정보를 모델링하는 RNN 계열 모델을 추가실험해보는 것이 가치가 있을 것이다.

4. 결론

기존 인공신경망 관련 연구들이 공통적으로 가진 한계점이었던 은닉층 해석에 관한 부분에 대하여, 대표적인 토픽모델인 LDA 모델의 결과물인 토픽과 인공신경망 모델의 일종인 RBM 모델의 결과물을 비교함으로써 은닉층 해석에 대한 가능성을 모색하였다. 실험을 통해, RBM 모델의 결과물은 문장에서 함께 등장한 단어들의 의미적인 부분보다는, 더 많이 등장했던 단어들에 집중적인 결과물을 보여준다는 점을 알 수 있었다. 향후 연구에서는, 함께 등장한 단어, 혹은 순서에 의한 패턴까지 담을 수 있는 최신 인공신경망 모델들, 이를테면 RNN 계열 모델 등을 대상으로 분석할 필요가 있다.

감사의 글

본 연구는 2017 년도 정부(미래창조과학부)의 재원으로 한국연구재단 생애첫연구사업의 지원을 받아 수행된 연구임 (No. 2017017836), 본 연구는 순천향대학교 학술연구비 지원으로 수행하였음 (과제번호 20170265)

참고문헌

[1] Google translator, https://translate.google.co.kr/
[2] Slack, https://slack.com/
[3] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in Proceedings of the 32nd International Conference on Machine Learning (ICML), pp, 2048-2057, Lille, France, 6-11 July 2015.
[4] Tesla, https://www.tesla.com/
[5] Amazon Echo, https://www.amazon.com/Amazon-Echo-Bluetooth-Speaker-with-WiFi-Alexa/dp/B00X4WHP5E
[6] Apple Siri, http://www.apple.com/kr/ios/siri/
[7] Yann LeCun and Yoshua Bengio, "Convolutional Networks for Images, Speech, and Time-Series," The Handbook of Brain Theory and Neural Networks, MIT Press, 1995.
[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, 3, pp. 993-1022, January, 2003.
[9] Asja Fischer and Christian Igel, "An Introduction to Restricted Boltzmann Machines," LNCS 7441, pp. 14-36, 2012.
[10] Thomas Hofmann, "Probabilistic Latent Semantic Analysis," in Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI), pp. 289-296, Stockholm, Sweden, July 30-August 1, 1999.
[11] Young-Seob Jeong and Ho-Jin Choi, "Sequential Entity Group Topic Model for Getting Topic Flows of Entity Groups Within One Document," in Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) 2012, pp. 366-378, Kuala Lumpur, Malaysia, 29 May - 1 June, 2012.
[12] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth, "The Author-Topic Model for Authors and Documents," in Proceedings of the 20th conference on Uncertainty in artificial intelligence (UAI), pp. 487-494, Banff, Canada, July 07 - 11, 2004.
[13] Thomas L. Griffiths and Mark Steyvers, "Finding Scientific Topics," in Proceedings of the National Academy of Sciences of the United States of America, pp. 5228–5235, 2004.
[14] Guangbing Yang, Dunwei Wen, Kinshuk, Nian-Shing Chen, and Erkki Sutinen, "A Novel Contextual Topic Model for Multi-Document Summarization," Expert Systems with Applications, vol. 42, issue 3, pp. 1340-1352, February 2015.
[15] Seiter J., Derungs A., Schuster-Amft C., Amft O., and Troster G., "Daily Life Activity Routine Discovery in Hemiparetic Rehabilitation Patients Using Topic Models," Methods Inf Med, vol. 54, issue 3, pp. 248-255, 2015.
[16] Frank Rosenblatt, "The Perceptron: A Probabilistic Model For Information Storage And Organization In The Brain," Psychological Review, vol. 65, issue 6, pp. 386-408, 1958.
[17] Vidushi Sharma, Sachin Rai, and Anurag Dev, "A Comprehensive Study of Artificial Neural Networks," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2, issue 10, pp. 278-284, October 2012.
[18] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams, "Learning Representations by Back-Propagating Errors," Natur, vol. 323, issue 9, pp. 533-536, October 1986.
[19] Geoffrey Hinton, Simon Osindero, and Yee-Whye Teh, "A Fast Learning Algorithm for Deep Belief Nets," Neural Computation, vol. 18, issue 7, pp. 1527–1554, 2006.
[20] Jeffrey L. Elman, "Finding Structure in Time," Cognitive Science, 14, pp, 179–211, 1990.
[21] Geoffrey Hinton, "A Practical Guide to Training Restricted Boltzmann Machines," Technical report UTML TR 2010–003, 2010.