랜덤 하이퍼그래프 모델을 이용한 순차적 멀티모달 데이터에서의 문장 생성

윤웅창°, 장병탁
서울대학교 컴퓨터공학부
{wcyoon, btzhang}@bi.snu.ac.kr

Sentence generation on sequential multi-modal data using random hypergraph model

Woongchang Yoon°, Byoung-Tak Zhang
Seoul National University
School of Computer Science & Engineering

요 약

인간의 학습과 기억현상에 있어서 멀티모달 데이터를 사용하는 것은 단순 모달리티 데이터를 사용하는 것에 비해서 향상된 효과를 보인다는 여러 연구 결과가 있어왔다. 이 논문에서는 인간의 순차적인 정보처리와 생성현상을 기계에서의 시뮬레이션을 통해서 기계학습에 있어서도 동일한 현상이 나타나는지에 대해서 알아보고자 하였다. 이를 위해서 가중치를 가진 랜덤 하이퍼그래프 모델을 통해서 순차적인 멀티모달 데이터의 상호작용을 하이퍼에지들의 조합으로 나타내는 것을 제안 하였다. 이러한 제안의 타당성을 알아보기 위해서 비디오 데이터를 이용한 문장생성을 시도하여 보았다. 이전 장면의 사진과 문장을 주고 다음 문장의 생성을 시도하였으며, 단순 암기학습이나 주어진 룰을 통하지 않고 의미 있는 실험 결과를 얻을 수 있었다. 단순 텍스트와 텍스트-이미지 쌍의 단서를 통한 실험을 통해서 멀티 모달리티가 단순 모달리티에 비해서 미치는 영향을 보였으며, 한 단계 이전의 멀티모달 단서와 두 단계 및 한 단계 이전의 멀티모달 단서를 통한 실험을 통해서 순차적 데이터의 단계별 단서의 차이에 따른 영향을 알아볼 수 있었다. 이를 통하여 멀티 모달리티가 시공간적으로 미치는 기계학습에 미치는 영향과 순차적 데이터의 시간적 누적에 따른 효과가 어떻게 나타날 수 있는지에 대한 실마리를 제공할 수 있었다고 생각된다.

1. 서 론

인간에 있어서 멀티모달 데이터를 이용한 학습이 단순 모달리티 데이터를 이용한 학습 보다 더 좋은 효과를 보인다는 결과의 연구를 많이 찾아볼 수 있다[1]. 이러한 '멀티미디어 원칙'을 통해서 사진과 단어를 통한 기억이 단순한 단어를 이용한 기억보다 오래간다는 결과를 보이고 있다. 인간의 정보처리 및 생성을 기계에서도 모사해 보려는 시도는 많이 있어 왔다[2]. 그중에서 기억에 관련된 접근은 쉽지 않은 시도이며 만족할 만한 결과를 아직 얻지는 못하였다. 이 논문에서는 인간에게서 일어나는 멀티모달 데이터를 이용할 때의 기억의 상승효과를 기계를 통해서 알아보려고 하였으며, 특히 인간의 순차적 회상 기억 현상[3]을 대한 시뮬레이션을 시도 하였다. 이전 연구에서는 멀티모달 데이터를 이용한 정보처리[4], 이미지 검색[5] 등은 많이 시도되었지만 문장생성[6]이나 자연어 처리[7]의 시도는 많지 않았다. 이 실험에서는 시계열 데이터인 비디오 데이터를 이용한 기계학습을 통해서 문장생성을 시도해 봄으로써 순차적 회상 기억 현상을 모사해 보고자 하였다.

이를 위해서 멀티모달 데이터의 상호작용을 랜덤 하이퍼그래프 모델로 나타낼 수 있다고 생각하였다. 가정에 따라서 가능한 모델이 많이 존재할 수 있지만, 이 실험에서는 진화적 하이퍼네트워크 모델[8]을 사용하였다. 이 모델은 기본 구성은 가중치를 가지는 여러 개의 하이퍼에지들의 집합으로 이루어진다. 하나의 하이퍼에지는 랜덤하게 전체 데이터의 조각 정보를 가지게 되고, 이러한 하이퍼에지들의 조합을 통해서 전체 데이터를 나타낼 수 있다.

이 실험에서는 멀티모달 데이터가 미치는 영향을 문장생성을 통해서 알아보고자 하였다. 상호 모달리티를 통한 의미 있는 데이터의 생성 및 조합은 여러 실험을 통해서 시도되었지만 문장의 회상이나 생성은 제한적으로 시도 되었다. 문장 생성은 문법의 정확성, 문맥의 흐름과의 일치성, 의미적인 타당성 등 여러 가지 판단 요소를 가질 수 있으며, 이를 객관적으로 수치화 하는 것은 쉽지 않은 문제일 수 있다. 본 실험에서는 대화문에 사용된 문장을 만족할 만한 문법과 의미성을 가진다고 생각하여 실제 생성된 문장과의 일치성을 문장 생성의 정확성이라고 판단하여 측정하였다.

2. 데이터 처리

멀티모달 데이터로 사용할 수 있는 데이터의 형태는 여러 가지가 존재할 수 있다. 하지만 비디오 데이터는 화상, 소리, 문자 등 복합적인 모달리티가 결합되어 있으며 시계열 데이터, 수집/ 가공의 용이성이라는 특성을 가지고 있다.

더욱이 실생활을 잘 반영하고 표현하고 있다는 점에서 인간의 행동과 지적현상을 시뮬레이션 하고 있는 본 실험에서 볼 때, 실제 인간의 행동을 분석, 연구한 실험과의 비교를 위한 공통적인 실험데이터로서 쓰일 수 있다는 장점이 있다.

이 실험에서는 '프렌즈(Friends)'라는 잘 알려진 미국의 TV 시트콤 시리즈를 사용하였으며 실험에 사용한 사진과 텍스트의 양은 비디오에서 추출한 343쌍을 사용하였다.

실험을 위한 노이즈 제거를 위해서 스크립트와 대화문을 비교분석하여 의성어 등 불필요한 낱말을 삭제하였으며, 구어체적 표현이나 문법이 틀린 문장 등을 알맞게 고치는 작업을 하였다. 이후, 문장을 단어 수준으로 나눈 후에 단어사전을 만들어서 각 문장이 어떠한 단어로 이루어져 있는지를 암호화 하였다.

표1. 순차적 텍스트-이미지 쌍의 예

시간 / 매체	$(t-2)^{th}$ Text-Image Pair
Image	
Text	He would be okay with Ethan.
시간 / 매체	$(t-1)^{th}$ Text-Image Pair
Image	
Text	There is an Ethan?
시간 / 매체	t^{th} Text
Text	Ethan is my boy friend.

이미지의 경우에는 자막이 등장하는 모든 화상을 200*150 픽셀 크기로 캡처 하였다. 다음에는 각 이미지를 균등한 92개의 그물망 형태로 자른 후 K-군집화 기법(K-means clustering)을 사용하여 2000개의 군집으로 나누었다. 각 군집은 비슷한 이미지들의 집합을 대표하는 의미를 가지게 되며 이를 바탕으로 이미지 사전을 작성 하였다. 텍스트와 마찬가지로 각 이미지의 각각의 조각들은 어떠한 군집에 속하는지를 암호화하여 표현하였다.

3. 실험 방법

우리는 이 실험에 있어서 크게 2가지 현상을 관찰하고자 하였다. 첫째는 멀티 모달리티가 단순 모달리티에 비하여 대비되는 효과와 두 번째는 순차적인 단서의 차이가 기억 현상에 미치는 효과를 알아보려 하였다.

실험 1에서는 t-1의 텍스트 단서를 주고 생성되는 문장과 t-1의 이미지와 텍스트 쌍을 주고 생성되는 문장 결과를 비교하는 실험을 하였다.

실험 2에서는 t-1의 이미지, 텍스트 쌍을 주고 생성되는 결과와 t-2, t-1의 이미지, 텍스트 쌍을 주고 생성되는 결과를 서로 비교하였다.

실험 1과 실험 2의 개괄적인 알고리즘은 표2와 같다.

표2. 하이퍼네트워크 학습과 문장 생성 알고리즘

```
H : 하이퍼에지들의 집합, h ∈ H
T : 텍스트 데이터 집합, t ∈ T
I : 이미지 데이터 집합, i ∈ I
C : 후보군들의 집합, c ∈ C
q : 특정 i에 대한 이전 시간의 단서
r : 특정 t에 대한 이전 시간의 단서
S : 생성된 문장, s ∈ S
For 1 to EpochTime do
    - H ← T, I에 대해서 랜덤 샘플링
    - 모든 i, t에 대해서 h가 일치/ 불일치에 따라
        하이퍼에지의 가중치를 증가/ 감소
    - 특정 하이퍼에지의 가중치가 역치 이하이면
        하이퍼에지를 제거하고 새로 랜덤 샘플링
End For
End For
For 1 to size(H) do
    If q와 r에 대해서 특정 h가 일치하면 then
        C ← h
    End If
End For
For 1 to size(q) do
    If q, r에 대한 위치 데이터가 c와 일치 then
        If 복수개의 c가 존재하면 then
        S ← 가장 큰 가중치를 가지는 c
        End If
    End If
End For
```

알고리즘에 관련된 개략적인 설명을 하자면, 모든 이미지와 텍스트 데이터에 대해서 랜덤 샘플링을 통해서 정해진 수의 하이퍼에지를 생성한다. 이후에 이미지와 텍스트 데이터와 비교하여 일치하는지 여부에 따라서 가중치를 증가하거나 감소시키며 일정 역치 값 이하로 떨어지는 가중치를 가지는 하이퍼에지를 제거하고 새로 생성하여 해집단의 품질을 높이는 작업을 일정시간 동안 수행한다. 이렇게 하여 생성된 하이퍼네트워크에서 생성될 문장의 이전 문장 또는 이미지 단서를 주고 이와 일치하는 하이퍼에지를 모두 후보군집에 모은다. 모아진 후보군집과 단서의 위치 데이터를 비교하여 문장을 생성하며, 충돌이 일어나면 가장 큰 가중치를 가지는 후보를 채택하여 문장을 생성한다.

랜덤 샘플링 시에는 이미지 또는 텍스트의 시작 위치는 랜덤하게 선정하였지만 시작 위치 다음의 샘플링은 $n-gram$ 기법을 사용하여 샘플링을 하였다.

위치정보를 사용하는 이유는 위치정보 없이 문장을 생성하게 되면 문법적으로 말이 되지 않는 전혀 의미 없는 단어의 나열에 불과한 결과를 얻었기 때문이며 각 위치에 대한 후보가 여럿일 경우에는 가중치가 가장 큰 후보가 가장 문법적으로 그럴 듯한 구문일 것이라는 가정이 있었기 때문이다. 따라서 생성된 문자의 정확도는 다음과 같이 나타낼 수 있다.

$$\text{생성의 정확도} = \frac{\text{정확하게 생성된 문장의 집합}}{\text{전체 생성된 문장의 집합}}$$

순차적 멀티모달 데이터간의 상호작용을 랜덤 하이퍼그래프인 하이퍼네트워크 모델[8]을 사용하였는데, 랜덤 하이퍼그래프 모델은 H = (X, E, W)로 표현될 수 있다. 각각 X, E, W는 정점, 연결선, 가중치를 나타낸다. 데이터의 집합 $D = \{d^{(i)}\}_{i=1}^{m}$가 주어졌을 때, h의 집합을 H라고 보면 구하고자 h에 관한 식은

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

$$h_{MAP} = \arg\max_{h \in H} P(D|h)P(h) \approx \arg\max_{h \in H} P(D|h)$$ 으로

나타내어 질 수 있다. 이를 위한 필수적인 가정은

$$P(D|h) = \prod_{i=1}^{m} p(d^{(i)}|h)$$

$$p(d^{(i)}) = \frac{\sum_{j=1}^{J} w^{(j)}}{Z(W)}$$ 이다.

j번째 하이퍼에지가 가지는 가중치가 시간 t에서 $w_t^{(j)}$로 나타내어지고, 하이퍼에지의 일치와 불일치에 따른 가중치를 표현하면,

$$w_t^{(j)} = \begin{cases} w_t^{(j)} = w_{t-1}^{(j)} + \gamma \cdot \triangle w \\ w_t^{(j)} = w_{t-1}^{(j)} - \gamma \cdot \triangle w \end{cases}$$ 이고,

여기서 γ는 학습의 비율을 나타낸다.

4. 실험 결과 및 분석

실험 1과 실험 2는 343쌍의 텍스트 또는 텍스트-이미지 쌍과 데이터에 따라 미치는 영향을 보기 위한 절반 수준인 150쌍을 사용하여 진행 하였다.

하이퍼에지의 크기(해집단)는 통상 데이터의 10배 정도로 유지하였으며, 각 제약조건에 따른 10000번의 문장 생성을 실행하여 평균값을 결과로 사용하였다. 텍스트 정보를 담는 하이퍼에지의 크기는 3으로, 이미지를 정보를 담는 하이퍼에지의 크기는 7로 일정하게 유지하였다.

멀티 모달리티와 단순 모달리티 간의 차이와 이를 이용한 문장생성에 미치는 효과를 본 실험 1의 결과는 그림 1과 같이 나타났다. 이를 통해서 멀티 모달리티 데이터를 이용한 문장 생성(TI2T)이 단순 모달리티를 통한 학습(T2T) 보다 좀 더 나은 성능을 보여주는 것을 알 수 있다.

순차적 단서의 차이에 따른 기억성능의 차이를 알아보려한 실험 2에서는 t-2(두 단계 이전 시점)과 t-1(한 단계 이전 시점)의 단서를 준 문장 생성(TI2T)이 t-1(한 단계 이전 시점)의 단서를 준 문장생성((TI-TI)2T) 보다 좀 더 잘 되는 것을 그림 2에서 볼 수 있다.

실험 1, 2 모두 데이터의 크기가 커질수록 문장 생성의 정확도가 줄어드는 것을 확인 할 수 있다.

위의 실험 결과를 분석하여 볼 때, 단순 모달리티 보다 멀티 모달리티 데이터가 기계 학습에 도움을 주며, 좀 더 많은 시점의 과거의 단서가 정보의 회상과 생성에 도움을 주는 실험 결과라고 볼 수 있다.

그림 1. T2T와 TI2T의 문장 생성 결과 비교

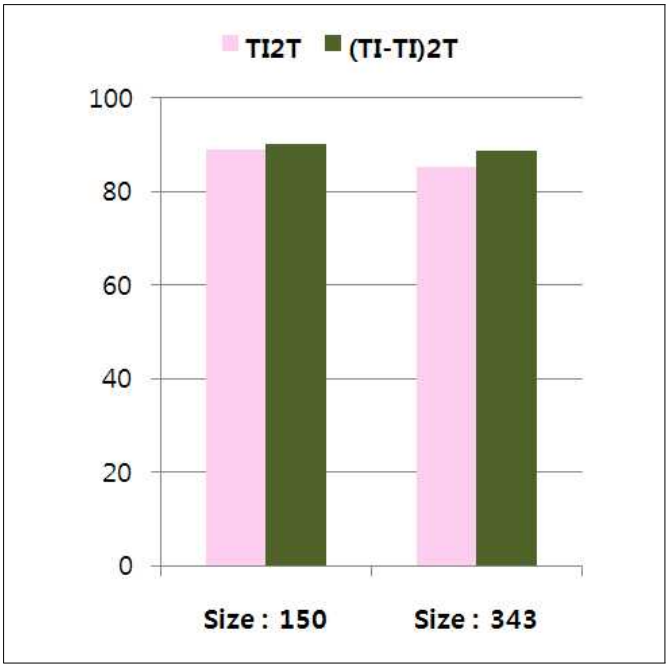

그림 2. TI2T와 (TI-TI)2T의 문장 생성 결과 비교

5. 결론 및 향후 연구

이 논문에서는 멀티 모달리티 데이터가 단순 모달리티 데이터 보다, 더 많은 시점의 과거의 단서가 정보의 회상 및 생성에 있어서 도움을 준다는 가정을 문장 생성을 통해서 확인하는 실험을 하였다.

다만, 실험 데이터의 양이 많다고 보기 어려우며 실험의 제약조건상 또는 여러 학습 중에서도 문장생성에서만 이러한 현상이 나타난다고 생각될 수 있다. 따라서 다른 매체나 미디어를 활용한 멀티 모달리티 데이터에 대한 실험(이미지 생성, 이미지 검색, 사운드 처리)이 필수적으로 뒤따라야 할 것이며, 다른 특성을 가진 좀 더 많은 데이터를 활용한 실험도 이루어져야 할 것이다.

감사의 글

이 논문은 교육과학기술부 재원에 의한 국가연구재단(314-2008-1-D00377, Xtran/ No. 2010-0017734), 지식경제부 및 한국산업기술평가관리원의 IT산업원천기술개발사업(K1002138, 차세대 맞춤형 서비스를 위한 기계학습 기반 멀티모달 복합 정보 추출 및 추천 기술 개발, MARS) 및 교육과학기술부의 BK21-IT 사업, 서울대학교 컴퓨터연구소에 의해 지원되었음.

[참고 문헌]

[1] Cisco Systems, Inc, Multimodal Learning Through Media: What the Research Says. (2008)

[2] Verena Rieser, Oliver Lemon, Using Machine Learning to Explore Human Multimodal Clarification Strategies, Annual Meeting of the ACL archive Proceedings of the COLING/ACL, p659-666,(2006)

[3] Ghalib, H. and Huyck, C.: A Cell Assembly Model of Sequential Memory. Proceedings of International Joint Conference on Neural Networks. (2005)

[4] Zhang, R., Zhang, Z., Li, M., Ma,W.-Y., and Zhang, H.-J.: A probabilistic semantic model for image annotation and multi-modal image retrieval. Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV''05) Vol. 1, 846-851 (2005)

[5] Sivic, J. and Zisserman, A.: Video Google: A Text Retrieval Approach to Object Matching in Videos. Proceedings of the Ninth IEEE International Conference on Computer Vision, Vol. 2, 1470 (2003)

[6] Oh, A. H., Rudnicky, A. I.: Stochastic natural language generation for spoken dialog systems. Computer Speech and Language. Vol. 16, 387-407 (2002).

[7] Rieser, V. and Lemon, O.: Learning human multimodal dialogue strategies. Natural Language Enginnering, Vol. 16, 3-23 (2009)

[8] Zhang, B.-T.: A Molecular Evolutionary Architecture for Cognitive Learning and Memory. IEEE Computational Intelligence Magazine. Vol. 3, No. 3, 49-63 (2008)