순차적 비디오 데이터를 위한 구성 구조 학습

온경운[O1], 김은솔[2], 허유정[1], 장병탁[1,3,4]
[1]서울대학교 컴퓨터공학부, [2]카카오브레인, [3]인지과학 협동과정, [4]뇌과학 협동과정
kwon@bi.snu.ac.kr, epsilon.kim@kakaobrain.com, {yjheo, btzhang}@bi.snu.ac.kr

Compositional Structure Learning for Sequential Video Data

Kyoung-Woon On[O1], Eun-Sol Kim, Yu-Jung Heo, Byoung-Tak Zhang[1,2,3]
[1]Department of Computer Science and Engineering, [2]Kakao Brain, [3]Brain Science Program, [4]Cognitive Science Program, Seoul National University

요 약

RNNs (Recurrent Neural Networks)와 같은 기존의 순차 학습 기법은 연속되는 입력 간의 상호 작용, 즉 1차 마르코프 의존성에 초점을 맞춘다. 그러나 비디오 데이터와 같은 대부분의 순차적 데이터는 가변 길이의 의미적 흐름과 그들의 구성으로 이루어진 복잡한 시간 종속성을 가지며, 이는 기존 방법으로 포착하기 어렵다. 본 논문에서는 이러한 복잡한 비디오 구조를 스스로 발견하여 비디오 데이터를 학습하는 Temporal Dependency Network (TDN)을 제안한다. TDN은 비디오의 프레임을 정점으로, 프레임 사이의 종속성을 간선으로 갖는 그래프로 비디오 입력을 표현한다. TDN은 그래프 컷 및 그래프 컨볼루션 기법을 이용하여 다단계 그래프(multilevel graph) 형식으로 데이터의 구성 종속성을 찾는다. 실험으로, 대규모 비디오 데이터 세트 Youtube-8M에서 제안된 방법을 평가한다. 실험 결과는 제안하는 모델이 비디오 데이터의 복잡한 의미 구조를 효율적으로 학습한다는 것을 보여준다.

1. Introduction

순차적 데이터 학습의 근본적인 문제는 더 나은 표현 학습을 위한 시퀀스의 의미적 구조를 학습하는 것이다. 특히 가장 어려운 문제는 여러 개의 의미 단위(semantic unit)으로 긴 길이의 전체 시퀀스를 분할하고 의미 단위 간의 구성 구조(compositional structure)를 찾는 것이다.
뉴럴 네트워크 관점에서 순차적 데이터의 학습은 순차적 입력을 자연스럽게 받아들이는 RNNs (Recurrent Neural Networks) 계열의 모델을 주로 사용한다. 그러나 RNN 기반 모델은 연속적인 입력 사이의 변화 패턴을 학습하기 위한 방법으로서 전체 입력의 장거리의 시간 종속성(long-term temporal dependency)를 학습하기 어렵다. 이를 해결하기 위해 Long Short-Term Memory (LSTM)[1], Gated Recurrent Units (GRU)[2]와 같은 변형들이 제안되었다. 그러나 이러한 변형들을 통해서도 여러 의미적 흐름을 유지하고 계층적 및 구성적 관계를 학습하는 것은 어렵다.

본 논문에서는 비디오 입력의 구성 종속 구조(compositional dependency structure)를 찾아가며 이를 비디오의 표현 학습에 활용할 수 있는 Temporal Dependency Network (TDN)을 제안한다. 제안하는 모델은 구성 구조를 다중 레벨 그래프(multilevel graph) 형식으로 정의하고, 긴 길이 입력에 대한 종속성과 계층적 관계를 효과적으로 찾아가며 학습을 한다.

실험으로, 비디오 이해 문제를 다루는 대용량 실세계 비디오 데이터셋인 YouTube-8M[3]을 사용하여 평가한다. 정성적 분석으로 실제 비디오 입력에 대해 학습하여 얻어지는 구성 종속 구조를 시각화하고, 정량적 분석으로 다른 베이스라인 모델과 비교하여 비디오 분류 성능의 향상을 보인다.

2. Problem Statement

입력되는 비디오는 각 프레임을 정점(node)로 갖고, 각 프레임 사이의 종속성의 강도를 간선(edge)의 가중치(weight)로 갖는 그래프 G로 표현할 수 있다.

N개의 연속적인 프레임을 갖는 입력 비디오 데이터의 각 프레임은 그래프 G의 정점 $v \in V$이고, 두 프레임 v_i, v_j 사이의 종속성 강도는 가중 간선 $e_{ij} \in E$이다. 또한 비디오의 각 프레임 v는 m-차원의 feature vector $x \in R^m$을 갖는다. 그래프 $G \in (V,E)$로부터, 비디오 프레임 사이의 종속 구조는 $A_{ij} = e_{ij}$인 가중인접행렬 A로 표현될 수 있다. 위의 표기법과 정의를 기반으로, 비디오 표현 학습 문제를 다음과 같이 정의할 수 있다.

*주어진 비디오 프레임 **feature vector matrix** $X \in R^{N \times m}$으로부터 프레임 간의 종속 구조를 표현하는 가중인접행렬 $A \in R^{N \times N}$을 찾는다.*

$$f : X \rightarrow A \quad (1)$$

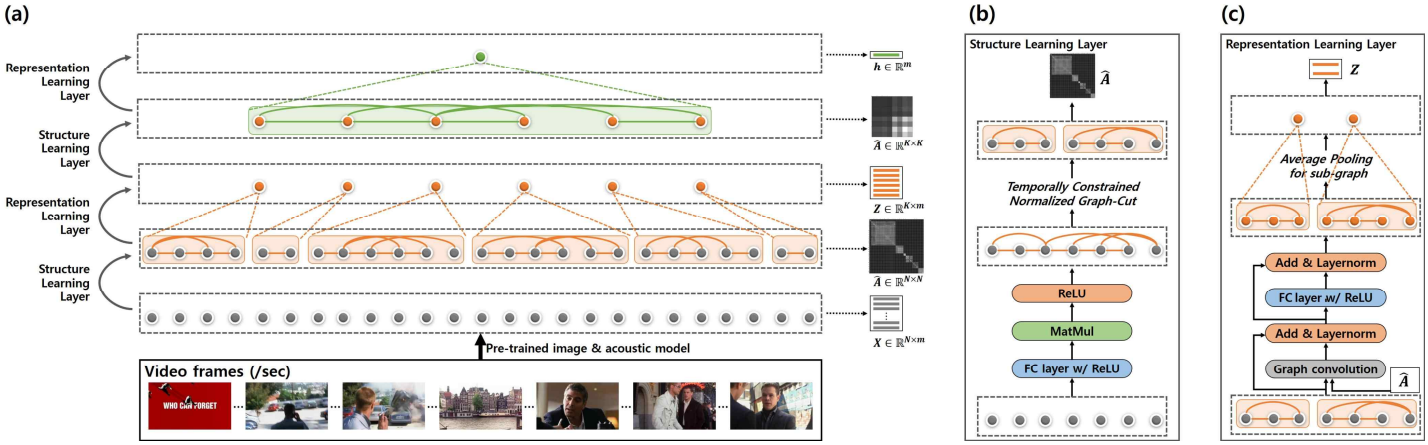
X와 A로부터 함수 g를 통해 최종 비디오 표현 $h \in R^l$ 얻는다.

$$g : \{X, A\} \rightarrow h \quad (2)$$

얻어진 비디오 표현 h는 다양한 비디오 학습 문제를 다룰 수 있다. 본 논문에서는 multi-label 분류 문제를 다룬다.

3 Temporal Dependency network

제안하는 모델은 그래프 분할 기반의 Structure learning layer와 그래프 컨볼루션 기반의 Representation learning layer 두 가지로 구성되어있다. Structure learning layer에서는 학습 가능한 커널과 시간적 제약



[그림 1] (a) 비디오 분류 문제를 위한 Temporal Dependency Networks의 전체 구조. (b),(c): Structure learning layer 및 Representation learning layer 세부 명세.

기반 정규화 그래프 분할 기법(temporally constrained normalized graph cut)을 이용하여 프레임간 종속 구조를 나타내는 인접행렬 $\hat{A}$를 추정한다. Representation learning layer에서는 추정된 $\hat{A}$를 기반으로 그래프 컨볼루션 및 풀링 기법을 통해 표현을 학습한다. 또한, 이 모듈들을 여러 층 쌓음으로써, 전체 비디오 프레임의 구성 구조가 다단계 그래프(multilevel graph)형식으로 찾아진다. [그림 1]은 TDN의 전체 구조를 나타낸다. 다음절에서는 각 모듈의 구조를 상세하게 설명한다.

2.1 Structure Learning Layer

Structure learning layer는 두 단계로 이루어진다. 첫 번째 단계에서는 모든 비디오 프레임 사이의 종속성을 커널 K를 통해 학습한다:

$$\hat{A}_{ij} = K(x_i, x_j) = ReLU(f(x_i)^\top f(x_j)) \quad (3)$$

이 때, $f(x)$는 비선형 활성함수가 없는 단층 신경망이다:

$$f(x) = W^f x + b^f, \ where \ W^f \in R^{m \times m} \text{ and } b^f \in R^m \quad (4)$$

그 후, $\hat{A}$은 Normalized 그래프 분할 알고리즘[4]을 통해 정제된다. 그래프 분할의 목표 함수는 식(5)와 같다.

$$Ncut(V_1, V_2) = \frac{\sum_{v_i \in V_1, v_j \in V_2} \hat{A}_{ij}}{\sum_{v_i \in V_1} \hat{A}_{i\cdot}} + \frac{\sum_{v_i \in V_1, v_j \in V_2} \hat{A}_{ij}}{\sum_{v_j \in V_2} \hat{A}_{\cdot j}} \quad (5)$$

식 (5)는 이산최적화 문제로 정리할 수 있고, 이산 문제를 연속 문제로 이완(relaxation)함으로써 eigen-value 문제로 변환할 수 있고 이는 $O(n^2)$의 시간복잡도를 필요로 한다 [4]. 비디오 데이터의 경우 시간에 따른 연속적인 서브시퀀스들로 구성되어 있으며 따라서 두 부분으로 분할된 하위 그래프들은 물리적 시간 내에서 겹침이 없다 [5, 6]. 그리하여 식 (6)과 같은 시간적 제약(Temporal constraint)을 둘 수 있다:

$$(i < j \text{ or } j < i) \text{ for } all \ v_i \in V_1, v_j \in V_2 \quad (6)$$

시간적 제약 조건 하에서 그래프 분할은 시간 축에 따라서만 분할이 될 수 있고 이에 따라 그래프 분할은 선형 시간에($O(n)$) 최적해를 구할 수 있다. Structure learning layer에서 시간적 제약 기반 그래프 분할 기법은 재귀적으로 여러번 적용되어 정제된 $\hat{A}$와 다수의 분할된 sub-graph들을 얻는다. sub-graph의 개수는 식(7)과 같이 비디오 길이 N에 의해 결정된다.

$$K = 2^{\lfloor \log_2 \sqrt{N} \rfloor - 1} \quad (7)$$

[그림 1(b)]는 Structure learning layer의 세부 구조를 보여준다.

2.2 Representation Learning Module

얻어진 가중인접행렬 $\hat{A}$로부터, 표현 학습 모듈은 그래프 컨볼루션 연산[7]과 position-wise 단층 신경망으로 각 프레임의 표현을 학습한다. 또한 residual connection[8]과 layer normalization[9]을 적용한다.

$$Z' = LN(\sigma(\hat{D}^{-1}\hat{A}XW^Z) + X) \quad (8)$$

$$Z = LN(\sigma(Z'W^Z + b^Z) + Z') \quad (9)$$

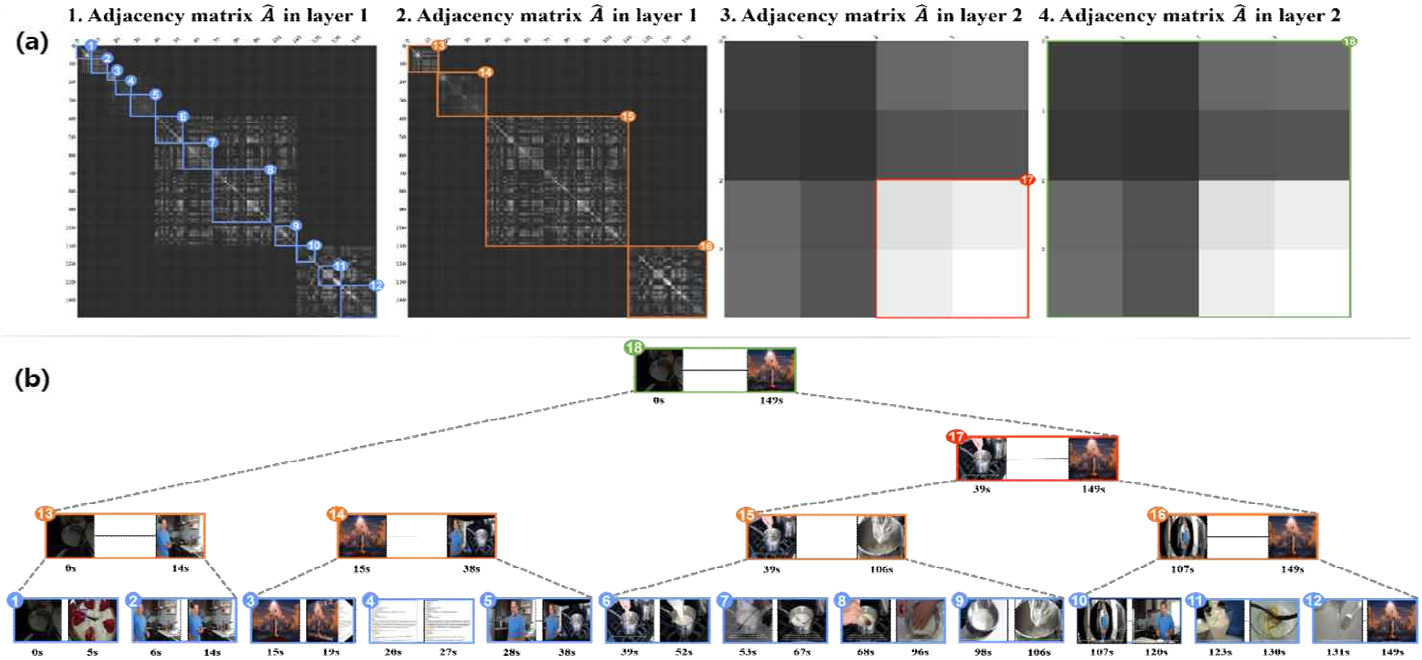
각 $W^f \in R^{m \times m}$, $b^f \in R^m$은 학습가능한 파라미터이다. 얻어진 표현 Z에 대해 sub-graph기반 average pooling을 적용하여 k개의 sub-graph에 대한 상위 레벨 표현 $Z \in R^{K \times m}$을 얻을 수 있다([그림 1(c)]. 이러한 방식으로, 얻어진 Z를 다시 새로운 Structure learning layer 및 Representation learning layer에 입력으로 넣어 최종 비디오 레벨의 표현 $h \in R^m$을 얻을 수 있다.

4. Experimental Results

실험으로 비디오의 주제를 분류하는 벤치마크 데이터셋인 YouTube-8M 데이터셋을 사용하였다. YouTube-8M 데이터셋은 YouTube로부터 약 6백만개의 비디오 클립을 수집하여 초당 1개의 이미지 피쳐[] 및 오디오 피쳐[]를 제공한다. 평가방법으로는 YouTube-8M competition과 동일하게 20개의 예측된 레이블의 confidence score를 계산하는 Global Averge Precision (GAP)를 사용한다.

4.1 Quantitative result

먼저, 다른 순차데이터 베이스라인 모델과 비교하여 제안하는 모델의 분류 성능을 평가하였다. 각 베이스라인 모델과 TDN의 성능은 [표 1]에 정리되어있다. TDN은 다른 순차 데이터 학습 모델보다 더 높은 GAP score를 달성하였다.



[그림 2] 실제 입력 비디오 “Rice Pudding”(https://youtu.be/cD3enxnS-JY)으로부터 구축된 구성 종속 구조(compositional dependency structure). 비디오의 주제(labels)는 {Food, Recipe, Cooking, Dish, Dessert, Cake, baking, Cream, Milk Pudding and Risotto}. (a) 학습된 1, 2 layer의 가중인접행렬로 1에 가까울수록 흰색, 0에 가까울수록 검정색. (b) 학습된 구성 종속 구조에 대한 개념적 예시.

4.2 Qualitative result

정성적 분석으로 TDN이 다단계 그래프 형태로 구성 구조를 학습하는 능력에 대해 검증하였다. [그림 2]는 실제 예시중 하나인 “ Rice Pudding” 비디오를 통하여 시각화한 모습이다. [그림 2(a)]의 1, 2 layer의 가중인접행렬로 밝을수록 강한 연결을 나타낸다. 가장 왼쪽에 파란박스는 프레임간 연결이 강한 부분을 표시하였으며 (b)의 최하단과 매칭된다. 실제 비디오에서 파란박스 구간은 화면전환시기와 상당히 유사함을 확인할 수 있었다. (a)의 두 번째의 주황박스는 실제 그래프 분할이 이루어진 구간으로((b) 주황색과 일치), 각각 요리채널 및 라이스푸딩 소개, 라이스푸딩 재료 및 레시피 설명, 라이스 푸딩 만드는 절차 자세히 설명, 요리 완성 및 outro 구간으로 분할 됨을 확인할 수 있었다. (a)의 빨간색, 초록색 박스는 2 layer에서의 가중인접행렬로 주제를 분류하는데 중요한 구간이 강한 연결로 표시됨을 확인할 수 있었다.

5. Conclusion

본 논문에서는 순차적 비디오 데이터의 구성 종속 구조를 학습함으로서 비디오의 표현을 학습할 수 있는 TDN을 제안하였다. 정량적, 정성적 분석을 통해 제안하는 모델이 효율적으로 내재적인 종속 구조를 잘 학습할 뿐아니라, 분류문제를 위한 더 좋은 표현을 학습할 수 있음을 보였다.

[표 1] Validation 데이터셋에 대한 GAP 성능 비교표.

Frame-level model	GAP
Average pooling	0.7824
DeepBoF (4096 custer)	0.8079
NetVLAD (256 cluster)	0.8396
LSTM (2 layers)	0.8446
GRU (2 layers)	0.8160
BiLSTM (2 layers)	0.8410
BiGRU (2 layers)	0.8079
Self-Attention (4 head, 2 layers)	0.8499
TDNs(2 layers)	**0.8557**

References

[1] Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735-1780, 1997.

[2] Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. Empiricalevaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[3] Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici,G., Varadarajan, B., and Vijayanarasimhan, S. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[4] Shi, J. and Malik, J. Normalized cuts and image segmenta-tion. *IEEE Transactions on pattern analysis and machineintelligence*, 22(8):888-905, 2000.

[5] Rasheed, Z. and Shah, M. Detection and representation ofscenes in videos. *IEEE transactions on Multimedia*, 7(6):1097-1105, 2005.

[6] Sakarya, U. and Telatar, Z. Graph-based multilevel temporal video segmentation. *Multimedia systems*, 14(5):277-290,2008.

[7] Kipf, T. N. and Welling, M. Semi-supervised classifica-tion with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[8] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learn-ing for image recognition. *InProceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778, 2016

[9] Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.