컨볼루션 신경망을 적용한 하이퍼네트워크 개념망 생성

남장군°,김경민,하정우,장병탁
서울대학교 컴퓨터공학부
{cjnan, kmkim, jwha, btzhang}@bi.snu.ac.kr

Hypernetwork-based concept network construction using convolutional neural networks

Chang-Jun Nan°, Kyung-Min Kim, Jung-Woo Ha, Byoung-Tak Zhang
School of Computer Science and Engineering, Seoul National University

요 약

기존 하이퍼네트워크 기반 멀티모달 개념계층모델은 MSER 알고리즘으로 이미지 패치를 추출하였지만 객체 인식율이 낮은 문제점이 있다. 본 논문에서는 최근 기계학습분야에서 높은 객체 인식율을 보이고 있는 컨볼루션 신경망으로 이미지 패치를 추출한 뒤 이를 멀티모달 개념계층모델에 적용하는 기법을 제안한다. 이 기법을 적용했을 때 이미지와 단어 벡터를 같은 공간상에서 더욱 의미있는 위치에 매핑시킴으로써 멀티모달 개념망의 정확도가 높아질 것을 기대한다.

1. 서 론

최근에 Deep Learning을 비롯한 다양한 멀티모달 학습기법에 대한 연구가 활발하게 진행되고 있다. 따라서 스마트폰과 구글그래스 등 웨어러블 센서의 발전을 통해 영상데이터가 엄청나게 증가하면서 멀티모달 데이터로부터 지식을 학습하는 컨텐츠 모델링 연구가 최근의 화제가 되고 있다. 그중 멀티모달 개념계층 모델[1]은 지속적으로 증가하는 데이터에 따라 효과적으로 학습할수 있는 계층적 구조의 개념학습 모델이다. 이 모델은 SPC(Sparse Population Coding)모델[2]과 다른, 계층구조로 구성되였다. 하위층은 이미지-테스트 조합의 고차 패턴을 표현하는 하이퍼그래프(Hypergragh)구조[3]로 구성되고 상위층은 개념변수들로 구성되였다. [1]에서 실험을 위해 유아용 만화 비디오 '뽀로로 시즌 3'을 사용하여 등장인물의 개념을 학습시켰다. 각 캐릭터와 인물특성을 고려하여 생성한 자막을 비교해본 결과 다른 모델보다 더 정확한 문장을 생성할수 있었다. 하지만 이미지와 관련이 없는 단어가 포함되는 경우도 존재하는 측면에서 개선의 여지가 있다고 판단되였다. 이러한 문제는 주로 데이터로부터 각 등장인물의 이미지 패치를 추출하는 과정에서 정확하게 추출하지 못하였기 때문이다.

본 논문에서는 이미지패치를 정확하게 추출하기 위해 컨볼루션 신경망 모델을 소개한다. 컨볼루션 신경망[4]은 다층 퍼셉트론의 한가지로서 최근에 음성신호처리와 이미지 인식 영역에서 훌륭한 성능을 보이고 있다. 그중 문자인식시스템 LeNet-5 모델[5]은 이미 은행 등 금융시스템에서 상용화되어 있고 Hinton은 컨볼루션 신경망 모델을 이용하여 ImageNet Challenge 2012에서 top5의 성능을 보였다[6]. 이외에도 [7]과 같은 성공적인 응용들이 많지만 본 논문에서는 [6]을 위주로 컨볼루션 신경망 모델을 소개하고자 한다.

2. Convolutional Neural Networks, CNNs

기존의 다층신경망은 이미지처리 영역에서 훌륭한 성능을 보여왔다. 하지만 알고리즘의 특성상 Fully connected network는 이미지의 사이즈에 비해 엄청나는 파라미터 개수가 포함되여 있어 Overfitting이 쉽게 생긴다. 이외에 Fully connected network는 입력한 이미지의 특성을 고려하지 않아 이미지의 작은 변화에 민감하지 않다. 이러한 결점으로 인해 다층신경망은 사용에 많은 극한성이 보였다. 이에 비해 컨볼루션 신경망은 Local receptive field, Share weights, Subsampling의 방식으로 데이터의 피처를 추출하여 모델의 성능을 높이고 있다.

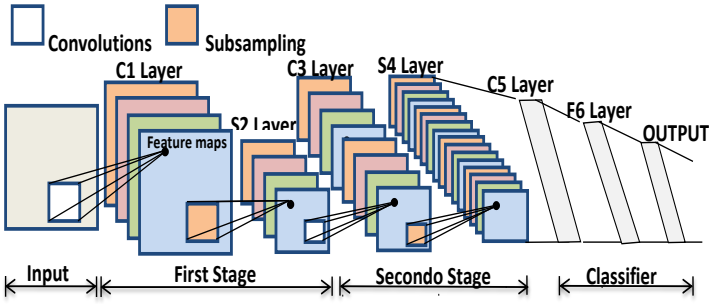

그림 1 컨볼루션 신경망

그림 1은 이미지 인식방면에서 가장 보편적으로 쓰이는 CNNs구조이다. 그중 Input은 이미지에 해당하고 각층은 상위층에서 추출한 피처맵으로 이루어 졌다. C1층은 Input에서 특정한 필러로 컨볼루션하여 추출한 피처맵이고 S2층는 C1에서 대응되는 피처맵을 Subsampling 하여 얻은 새로운 피처맵이다. S2층은 또 다시 컨볼루션하여 C3으로 전달되고 같은 방법으로 S4을 구성하게 된다. 마지막에 S4층에 생성된 피처맥은 Rasterization하여 Full connection 방식으로 5층과 6층으로 전달하여 결과를 출력한다.

그림 2에서 보다시피 피처의 추출과정은 크게 컨볼루션과 Subsampling의 두개 단계로 나누어져 있다.

(1)컨볼루션과정: Input부터 트레닝을 위한 필터 f_x와 바이어스 b_x를 통해 C_x층의 피처맵을 구성한다.

(2)Subsampling과정: C_x 의 피처맵의 각 근접하는 2×2 픽셀들을 합하고 다시 Weights sum W_{x+1} 와 바이어스 b_{x+1} 을 통과한다. 마지막으로 Sigmoid함수를 지나고 난 피처매핑S_{x+1}은 C_x의 1/4로 작아지게 된다.

위에서 정의한 C층은 피처추출층이다. 여기에 있는 각 하나의 뉴런은 상위계층의 Local receptive field와 연결되어 피처를 추출하게 된다. 따라서 각 뉴런사이의 위치가 정하게 된다. S층는 피처매핑층이다. 여기에 있는 각 피처매핑은 하나의 평면을 이루게 되며 각 평면상의 모든 뉴런의 Weights는 모두 같다. 피처매핑 스트럭처는 Sigmoid함수를 컨볼루션 네트웨크의 Activation 함수로사용하여 피처매핑이 Displacement invariance의 성질을 가지게 된다.

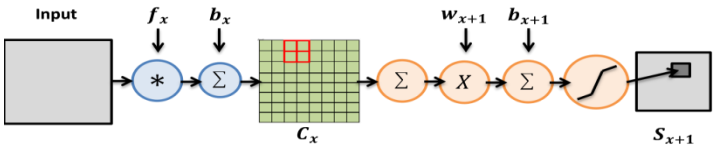

그림 2 컨볼루션과 Subsampling과정

하나의 매핑평면에 뉴련들이 가중치를 공유하고 있어 신경망에 필요되는 파라미터의 개수를 줄일수 있게 되여있다. C층과 S층이 조합되여 있어 피처를 2차추출하게 되는데 이러한 특성으로 Input이 변형된 상황에서도 일정한 성능을 보유할수 있다.

3. 멀티모달 개념계층 모델

멀티모달 개념계층 모델(Multimodal Concept Hierarchy, MuCH)는 계층구조 모델로 표현된다. 입력은 비디오 데이터에 해당되고 상위층변수는 비디오 각 캐릭터에 해당된다. 하위층은 하이퍼그래프구조를 이용하여 단어와 이미지 패치의 고차 패턴을 표현하는 하이퍼에지들의 집합이다. 상위층은 개념변수를 포함하여 있고 하위층은 이미지-테스트로 구성된 하이퍼에지 변수들이다. 상위층의 개념변수는 하위층의 연관성이 큰 하이퍼에지들의 부분집합과 연력되며 개념변수들은 하이퍼에지를 공유할수 있다. 모델의 파라미터 $\theta=(e,\alpha)$ 와 개념변수 $c=(c_1,\ldots c_k)$ 가 주어졌을 때 이미지 패치와 자막의 확률분포는 식 (1)로 표시된다.

$$P(\mathrm{r},\mathrm{w}|\mathrm{c})=\sum_{e,\alpha}P(r,w|e,\alpha,c)P(e,\alpha|c) \qquad (1)$$

그중 이진벡터와 $\mathrm{r}=(\mathrm{r}_1,\ldots\mathrm{r}_\mathrm{n})$ 와 $\mathrm{w}=(\mathrm{w}_1,\ldots\mathrm{w}_\mathrm{m})$ 는 이미지 패치와 단어이고 $\mathrm{c}=(\mathrm{c}_1,\ldots\mathrm{c}_\mathrm{k})$ 는 개념변수이다. 파라미터 e 는 하이퍼에지의 집합이고 α 는 하이퍼에지의가중치 집합이다. 모델의 학습은 순차적베이지안 가중치 집합이다. 모델의 학습은 순차적 베이지안 추론으로 이루어졌는데 수식으로 정의하면 아래와 같다

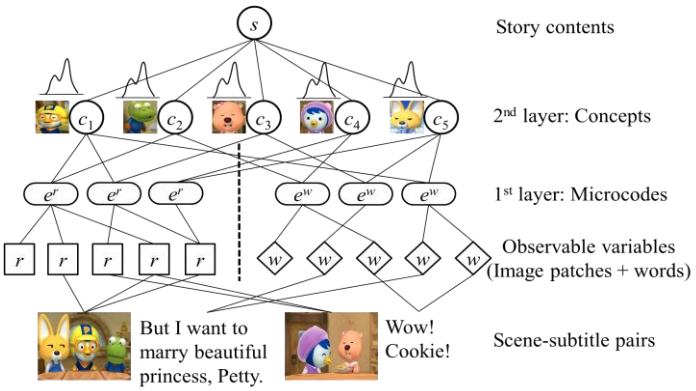

그림 3 멀티모달 개념계층 모델[8]. CNN을 이용하여 image patch를 캐릭터 단위로 보다 정교하게 뽑고자 함

$$P_t(\mathrm{e},\alpha|\mathrm{r},\mathrm{w},\mathrm{c}) = \frac{P(\mathrm{r},\mathrm{w}|\mathrm{c},\mathrm{e},\alpha)P(\mathrm{c}|\mathrm{c},\alpha)P_{t-1}(\mathrm{e},\alpha)}{P(\mathrm{r},\mathrm{w},\mathrm{c})} \quad (2)$$

P_t 는 t 번째 에피소드에서의 확률분포이고 t 번째 에피소드가 들어왔을 때 prior분포 $P_{t-1}(\theta)$ 는 Posterior분포를 계산하는데 사용 된다. 여기의 $P_t(\theta)$는 다음 단계의 Prior로 사용된다. 식 (2)을 아래와 같이 변형할수 있다.

$$\theta' = \underset{\theta}{\operatorname{argmax}} \left[\left\{ \sum_{d=1}^{D^t} \left(\log P\left(r^{(d)}, w^{(d)} | c^{(d)}, e, \alpha\right) + \log P\left(c^{(d)} | e, \alpha\right) \right) + D^t \log P_{t-1}(e, \alpha) \right\} \right] \quad (3)$$

여기서 D^t 는 t 번째 에피소드의 데이터 사이즈이고 학습은 log likelihood를 최대화하는 방식으로 이뤄진다.

4. 결론 및 향후 연구 계획

본 논문에서는 컨볼루션 신경망과 멀티모달 개념계층 모델에 대해 소개하였다. 기존의 멀티모달 개념계층 모델에서 사용된 유아용 만화 비디오는 이미지 프로세싱이 쉽게 구현되여 문제의 복잡도를 줄일 수 있었지만 실세계로부터 얻어지는 데이터에는 이미지 피처를 추출하는데 어려움이 보인다. 이러한 문제들로 인한 이미지와 테스트의 관련성이 낮는 등 문제부터 보면 아직 개선의 필요성이 보인다. 본 논문에서 컨볼루션 신경망의 특성을 분석한 결과 모델이 이미지 피처 추출 문제를 해결하기에 가장 적합하는 방법으로 보인다. 향후의 연구에서는 유아용 만화 비디오 '뽀로로' 데이터 베이스를 컨볼루션 신경망에 학습하여 이미지 피처 추출에 적용해 볼 것이다.

감사의 글

이 논문은 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구이며(NRF-2010-0017734-Videome), 정부(미래창조과학부 및 정보통신기술진흥센터)의 정보통신·방송 연구개발사업 지원 (10035348-mLife, 14-824-09-014, 10044009-HRI.MESSI)을 일부 받았음.

참고문헌

[1] 김경민, 하정우, 이범진, 장병탁, 멀티모달 개념망과 언어 모델을 이용한 이미지 설명문 생성, *2013 한국컴퓨터종합학술대회(KCC2013) 논문집*, pp. 1538-1540, 2013.

[2] B.-T. Zhang, J.-W. Ha, and M. Kang, Sparse population code models of word learning in concept drift, In *Proceedings of Annual Meeting of the Cognitive Science Society (CogSci 2012)*, pp. 1221-1226, 2012.

[3] B.-T. Zhang, Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory, *IEEE Computational Intelligence Magazine*, 3(3):49-63, 2008.

[4] D. C. Cireșan, U. Meier, J. Masci, L. M. Gambardella, J. Schmidhuber, Flexible, High performance convolutional neural networks for image classification, *the International Joint Conference on Artificial Intelligence (IJCAI-2011)*, pp. 1237-1242, 2011.

[5] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, pp. 2278-2324, 1998.

[6] A. Krizhevsky , I. Sutskever , G. E. Hinton, Imagenet classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems (NIPS 2012)*, 2012.

[7] Y Sun, X Wang, X Tang, Deep convolutional network cascade for facial point detection, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013)*, pp. 3476-3483, 2013.

[8] J.-W. Ha, K.-M. Kim, and B.-T. Zhang, Automated construction of visual-linguistic knowledge via concept learning from cartoon videos, In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2015)*, 2015.