엔트로피 기반 멀티모달 랜덤 하이퍼그래프 학습을 이용한 동영상에서의 이미지 태깅 방법

윤웅창○ 석호식 장병탁
서울대학교 컴퓨터공학부
wcyoon@bi.snu.ac.kr, hsseok@bi.snu.ac.kr, btzhang@bi.snu.ac.kr

Image annotation method in video corpus using Entropy based Multi-modal Random Hypergraph Learning

Woongchang Yoon○ Ho-Sik Seok Byoung-Tak Zhang
School of Computer Science and Engineering
Seoul National University Seoul, Korea

1. 서 론

대용량 데이터 처리 기술의 발달로 많은 멀티미디어 정보가 생산되고 있으며 이를 수정, 편집, 분류하는 작업의 필요성도 높아지고 있다. 그중에서도 영상데이터에 시맨틱한 정보를 추가하는 태깅 작업은 중요한 일로서 정지화상에서의 이미지 태깅은 주목한 만한 성과가 있지만 동영상에서는 작업은 쉽지않은 일이다. 본 연구에서는 자동으로 동영상의 이미지 정보에 태그를 부여하는 방법을 소개한다. 제안 방법은 동영상의 이미지 정보와 텍스트 정보를 모두 활용하여 이미지 정보의 변화에 상관없이 성공적으로 이미지에 태그를 부여할 수 있는 방법으로서 랜덤 하이퍼그래프 모델을 이용하여 이미지-텍스트 쌍을 생성한 후 정보량(엔트로피)를 조정하여 유용한 이미지-텍스트 쌍의 비율을 높이는 방법이다.

2. 본 론

Face-tag 쌍을 구성하는 개개의 image-text 쌍을 h_i(랜덤하이퍼그래프 모델에서 하이퍼에지)로 이루어진 집합을 H, 추정하고자 하는 목표 확률 분포를 π_o, 학습을 통해 추정된 근사 확률 분포를 P라고 표시할 경우 크로스 엔트로피 기반 인물 인식 방법의 목표는 주어진 훈련 데이터 D를 설명할 수 있는 H 중 π_0와 P간의 크로스 엔트로피를 최소화하는 H를 찾는 것이다. 즉 학습 목표를 다음과 같이 표현할 수 있다.

$$H = \arg min_H P_{CE}(\pi_0, P) \quad \text{여기서} \quad H = \arg max_H P(D|H)$$

$$\arg min_H P(\pi_o, P) = \int \pi_0 \ln \frac{\pi_o(x)}{P(x)} dx = \int \pi_o(x) \ln \pi_0(x) dx - \int \pi_0(x) \ln P(x) dx$$

여기서 $\int \pi_o(x) \ln \pi_0(x) dx$는 우리가 알고자 하는 목표 분포와 관련된 식으므로 어떤 값을 갖게 될 것인지 추정할 수 없다. 따라서 목표는 $\int \pi_0(x) \ln P(x) dx$를 최대화 할 수 있는 H를 찾는 것으로 변경된다. 본 논문에서는 P를 하이퍼에지 h_i와 h_i에 배정된 가중치 ω_i의 쌍 (h_i, ω_i)의 분포로 표현한다. 우리는 Maximum entropy 접근법을 이용하여 전체적인 학습 목표 및 학습 과정 정의하였다.

- 목표: $H^* = \arg max_H \int \pi_0(h) \ln P(D|H) dh$
  여기서 $H = \{h_i\}$, 만약 텍스트 x에 대하여 이미지 y가 얼굴인 경우 $h_i(x,y) = 1$, y가 얼굴이 아닌 경우 $h_i(x,y) = 0$으로 정의.
- H_Λ를 $\Lambda(p,w)$가 최대가 될 때의 H라고 표시하고 이 때의 $H(w_i$의 수치)를 $\Psi(\omega) \equiv \Lambda(P_\omega, \omega)$라고 하면

$$\Psi(\omega) = -\sum_x \tilde{P}(x) \log Z_\omega(x) + \sum_i \omega_i \tilde{P}(h_i), \; Z_\omega(x) = \sum_y \exp\left(\sum_i \omega_i h_i(x,y)\right)$$

· 각 하이퍼에지 h_i에 대하여 가중치 w_i를 부여하면 연산자 $\Lambda(p,w)$를 다음과 같이 정의할 수 있다.

$$\Lambda(P,w) \equiv H(P) + \sum_i w_i (P(h_i) - \tilde{P}(h_i)) \cdots (1)$$

(여기서 $\tilde{P}$는 h_i의 분포로 표현된 P, $H(P) \equiv -\sum_{x,y} \tilde{P}(x) P(y|x) \log P(y|x) = \sum_{x,y} \tilde{P}(x,y) h(x,y)$)

실험에서는 'Friends' 시트콤에서 대화가 이루어지는 시점의 400*300 픽셀 이미지로 1000장을 수집하여 이미지에 들장하는 인물들의 얼굴을 인식한 후 대본의 텍스트 정보를 이용하여 인물의 이름 단어를 추출해 낸다. 이렇게 만들어진 텍스트 이미지 데이터 쌍을 이용하여 학습을 수행한다.

실제로 제안한 방법이 이미지-텍스트 쌍 학습을 통해서 이미지 태깅이 잘 되는지는 여부는 하이퍼에지들의 가중치 변화와 이미지 태깅의 성공 비율로 알 수 있다. 하이퍼에지의 수는 1,000,000개로 이미지-텍스트 쌍은 1000개를 사용하여 실험을 진행하여 학습의 횟수가 증가함에 따라서 학습의 정확도와 하이퍼에지들의 가중치가 높아지는 것을 확인할 수 있었다.

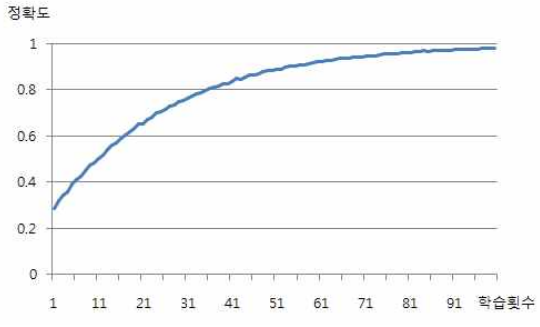

그림 1. 학습 데이터 중 이미지-태깅이 성공적으로 이루어지는 하이퍼에지의 비율

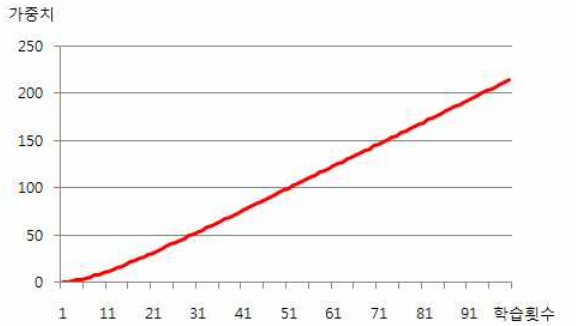

그림 2. 하이퍼에지들의 가중치의 변화

3. 결 론

본 논문에서는 기존의 정지화상이나 사진에서의 얼굴 인식 방법을 개선하여 멀티모달 데이터를 이용한 학습을 통해 동영상에 등장하는 이미지에 태그를 부여하는 방법을 제안하였다. 제안 방법에서는 이미지-텍스트 하이퍼에지를 생성한 후 하이퍼에지의 엔트로피를 조절하여 유용한 이미지-텍스트 하이퍼에지를 탐색한다.

감사의 글

본 연구는 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.

참고문헌

[1] Athanasios K. Noulas, Nikos Vlassis, and Ben J. A. Krose. Cross entropy for learning in multi-modal streams. In JointWorkshop on MultiModal Interaction and Related Machine Learning Algorithms, 2007.
[2] Pieter-Tjerk de Boer, A Tutorial on the Cross-Entropy Method, Annals of Operations Research Volume 134, Number 1, 19-67, 2005.
[3] Y. Tian, W. Liu, R. Xiao, F. Wen, and X. Tang, A face annotation framework with partial clustering and interactive labeling. In CVPR, 2007.
[4] Shin'ichi Satoh , Takeo Kanade, Name-It: Association of Face and Name in Video, Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97), p.368, June 17-19, 1997.
[5] Kienzle, W., G. Bakir, M. Franz and B. Schölkopf: Face Detection, Efficient and Rank Deficient, Advances in Neural Information Processing Systems 17, 673-680. (Eds.) Weiss, Y. MIT Press, Cambridge, MA, USA, 2005
[6] Berger, A. L., Pietra, S. A. D, and Pietra, V. J. D. A Maximum Entropy Approach to Natural Language Processing, Computational Linguistics, Vol. 22, No. 1, 39-71, 1996.
[7] http://www.kyb.mpg.de/bs/people/kienzle/facedemo/facedemo.htm
[8] Zhang, B.-T, Hypernetworks : A Molecular Evolutionary Architecture for Cognitive Learning and Memory, IEEE Computational Intelligence Magazine 3(3): 49-63. 2008.