시선 정보를 이용한 만화 영상의 가속화된 선택적 학습 방법

김진화○[1], 장병탁[12]
서울대학교 인지과학 협동과정[1]
서울대학교 컴퓨터공학부[2]
{jhkim, btzhang}@bi.snu.ac.kr

Accelerated Selective Learning from Cartoon Videos with Eye-Gaze Information

요 약

이미지-텍스트 다중 모달(multimodal) 학습에 대한 많은 선행 연구들에서 입력 이미지에 대응하는 문자열을 생성하기 위한 다양한 모델을 찾을 수 있다. 본 연구에서는 콘볼루션(convolution) 시각 특징 벡터를 선택적으로 순환 학습하는 조건화 LSTM(Conditional Long Short-Term Memory)과 LRCN(Long-term Recurrent Convolutional Networks) 모델을 참고하여 학습하는 모델을 제안한다. 이 때 문자열을 구성하고 있는 한 단어는 주어진 이미지의 일부 영역과 대응될 수 있다는 가정을 전제한다. 본 연구에서는 만화 영상과 같이 입력 이미지들이 시간 축 상에서 문맥적 흐름을 가질 때 한 단어와 이미지의 한 영역의 대응 관계는 문맥에 따라 달라질 수 있다는 가정으로 조건을 완화하여 선행 연구를 확장한다. 문맥 기반 주의 모델을 학습하기 위하여 선행 학습으로 한 시간 분량의 비슷하지만 다른 영상 정보에 대하여 수집 된 다수 사람들의 시선 분포 정보를 이용함으로써 시각 주의에 대한 선택적 학습을 가속화 한다.

1 서 론

다중 모달 학습은 서로 다른 통계적 성질을 가진 모달 간의 공통된 정보를 학습한다. 그러나 이러한 각 모달의 통계적 차이가 입력된 데이터에 대하여 직접적인 공통된 정보를 얻기 어렵게 만든다. 따라서 많은 선행연구에서는 깊은 학습(deep learning)을 통해 각 모달의 상위 개념을 먼저 학습한 후 공통된 정보를 학습하는 방법을 이용하였다[1, 2, 3, 4]. 이미지-텍스트 다중 모달 학습에서 이미지에 포함된 여러 독립적인 객체들과 자연어 언어 모델 안에서 단어와의 관계를 문법적 정보나 예제 문장 틀과 같은 선행 지식 없이 자동적으로 학습하는 연구에서 성능적으로 많은 진보가 있었다[3, 5, 6].

Xu et al.[7]의 연구에서는 이미지 안의 여러 독립 객체들과 문장과의 관계를 자연어 언어 모델 안에서 동적으로 학습하는 모델을 제안하였다. 자연어 언어 모델 학습에 좋은 성능 보이고 있는 순환 네트워크의 일종인 LSTM 모델에서 은닉 상태에 따라 주어진 이미지의 콘볼루션(convolution) 시각 특징 벡터 일부를 공간 상에서 선택[8, 9], 순환 학습함으로써 주의 모델 기반 조건화 LSTM을 구현하였다. 변분 베이지안(variational Bayesian) 방법에서 변분 하한(variational lower bound)를 최대화하는 새로운 목적 함수를 정의하고, 이 새로운 목적 함수에 적용 가능한 몬테 카를로(Monte Carlo) 샘플링을 이용하여 비용 함수의 경사도를 계산하기 때문에 일반적인 역전파(back-propagation) 알고리즘을 통해 LSTM과 신경망 네트워크인 주의 모델을 일관된 방법으로 두 모듈을 동시에 학습할 수 있다.

하지만 기존 연구는 정적 이미지에 대한 기술적 문장을 생성하도록 학습하기 때문에, 만화 영상과 같이 시간의 흐름에 따라 문맥이 달라질 수 있는 데이터에는 적합하지 않다. 예를 들면 동일한 두 캐릭터가 대화를 나누는 같은 장면도 앞에서 무슨 행동이나 대화를 나누었냐에 따라 문맥이 달라지며 해당 장면에 대한 대사가 달라질 것이다. 따라서 우리는 기술적으로 기존 연구에서 두 가지 요소를 변경하여 문제를 해결하고자 한다. 첫 번째는 LSTM의 메모리 벡터 $\mathbf{c}_0$와 은닉 벡터 $\mathbf{h}_0$를 새로운 이미지에 따라 초기화 하지 않고 보존한다. 두 번째는 주의 모델 $\mathbf{f}_{att}$를 랜덤 파라미터로 초기화하는 대신 한 시간 분량의 비슷하지만 서로 다른 영상을 본 다수 사람들의 시선 분포 정보를 이용하여 파라미터 값을 초기화 함으로써 선택적 학습을 가속화 하고자 한다.

따라서 본 연구는 입력 이미지들이 시간 축 상에서 문맥적 흐름을 가질 때 한 단어와 이미지의 한 영역의 대응 관계는 문맥에 따라 달라질 수 있다는 가정으로 선행 연구를 확장한다. 이 때 입력 이미지 내 독립 객체들의 상호 작용이 문맥 조건에 따라 달라지고 문장 내의 단어들과의 상호 작용도 더욱 동적으로 일어난다.

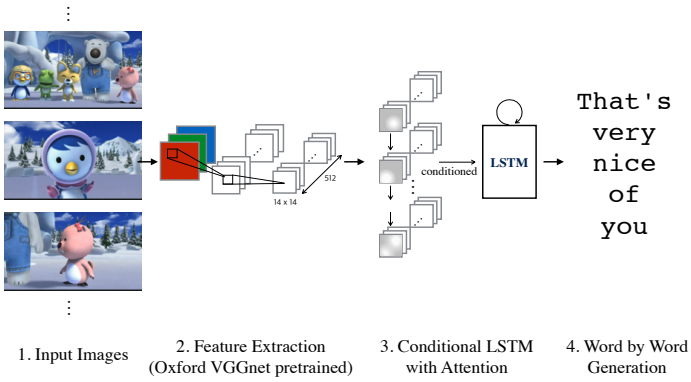

그림 1: 시선 정보를 활용하는 방법을 도식화. A는 원본 이미지, B는 A를 응시한 서로 다른 13명의 응시 좌표들을 십자가 기호로 표시. C는 응시 좌표들을 가우시안 혼합 모델(Gaussian Mixture Model)로 나타냈을 때의 확률 분포를 등고선 표현. D는 몬테 카를로 방법으로 샘플링할 확률 가중치를 도식적으로 표현.

2 Attentional LSTM

LSTM 모델은 입력 벡터의 일부가 출력되는 재귀적인 구조를 가진다. $(t-1)$ 시간에서의 출력 벡터 $\mathbf{y}_{t-1} \in \mathbb{R}^K$의 선형 사영값 $\mathbf{E}\mathbf{y}_{t-1} \in \mathbb{R}^m$와 은닉 벡터 $\mathbf{h}_{t-1} \in \mathbb{R}^n$, 그리고 t 시간에서의 문맥 벡터 $\hat{\mathbf{z}}_t \in \mathbb{R}^D$를 입력하면 t 시간에서의 은닉 벡터 $\mathbf{h}_t$를 출력한다. 여기에서 $\mathbf{y}$는 단어를 나타내는 비트 지시 벡터(해당 단어 인덱스가 1이고 나머지는 0인 벡터)이고 $\mathbf{h}$는 LSTM의 학습 파라미터 벡터, $\hat{\mathbf{z}}$은 주의 모델 $\mathbf{f}_{\text{att}}$의 출력 값이다. LSTM 모델에서 정규화 방법을 제안한 Zaremba et al.[10]의 방법에 따라 LSTM을 구현하면 다음과 같이 표현할 수 있다.

$$\begin{pmatrix} \mathbf{i_t} \\ \mathbf{f_t} \\ \mathbf{o_t} \\ \mathbf{g_t} \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \mathbf{T} \begin{pmatrix} \mathbf{E}\mathbf{y_{t-1}} \\ \mathbf{h_{t-1}} \\ \hat{\mathbf{z}}_\mathbf{t} \end{pmatrix} \tag{1}$$

$$\mathbf{c_t} = \mathbf{f_t} \odot \mathbf{c_{t-1}} + \mathbf{i_t} \odot \mathbf{g_t} \tag{2}$$

$$\mathbf{h_t} = \mathbf{o_t} \odot \tanh(\mathbf{c_t}) \tag{3}$$

$\mathbf{E} \in \mathbb{R}^{m \times K}$는 단어를 나타내는 비트 지시 벡터를 실수값으로 선형 사영한다. $\mathbf{T} \in \mathbb{R}^{(D+m+n) \times n}$는 LSTM의 내부 파라미터 행렬이다. $\mathbf{E}$와 $\mathbf{T}$는 연쇄 법칙에 따라 목적 함수에 의해 역전파 알고리즘으로 학습되는 파라미터 행렬이다. 두 개의 초기 값 파라미터, $\mathbf{c}_0$, $\mathbf{h}_0$는 각각 별도의 MLP(Multi-Layer Perceptron)으로 학습하며, 두 MLP는 채널에 대한 $\mathbf{a}$의 평균 값($\in \mathbb{R}^D$)을 같은 입력으로 가진다.

수식3에 따라 $\mathbf{h}_t$를 구한다면 다음 식을 이용하여 $\mathbf{y}_t$를 구할 수 있게 된다.

$$p(\mathbf{y}_t | \mathbf{y}_1^{t-1}, \mathbf{a}) \propto \exp \mathbf{L}_o(\mathbf{E}\mathbf{y}_{t-1} + \mathbf{L}_h \mathbf{h}_t + \mathbf{L}_z \hat{\mathbf{z}}_t)) \tag{4}$$

여기에서 $\mathbf{L}_o \in \mathbb{R}^{K \times m}$, $\mathbf{L}_h \in \mathbb{R}^{m \times n}$, $\mathbf{L}_z \in \mathbb{R}^{m \times D}$는 모두 목적 함수에 의해 학습되는 파라미터 행렬이다. $\mathbf{a} \in \mathbb{R}^{L \times D}$는 미리 학습된 콘볼루션 네트워크의 반응 값으로 $\mathbf{L}$은 반응 크기(예, $\mathbf{L} = 14 \times 14$), $\mathbf{D}$는 채널 수이다.

$\mathbf{f}_{\text{att}}$가 반응 벡터 $\mathbf{a}$와 은닉 벡터 $\mathbf{h}_{t-1}$를 입력으로 하는 MLP일 때, $\hat{\mathbf{z}}_\mathbf{t}$는 다음과 같이 정의한다.

$$e_{ti} = \mathbf{f}_{\text{att}}(\mathbf{a_i}, \mathbf{h_{t-1}}) \tag{5}$$

$$\alpha_{\mathbf{ti}} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})} \tag{6}$$

$$\hat{\mathbf{z}}_\mathbf{t} \sim \text{Multinoulli}_L(\{\alpha_{ti}\}) \tag{7}$$

문맥 벡터 $\hat{\mathbf{z}}_\mathbf{t}$는 역전파 알고리즘으로 학습된 $\mathbf{f}_{\text{att}}$의 결과에 따라 샘플링 된다. 이 때, 목적 함수 L을 로그 합 부등식을 이용, marginal log-likelihood $\log p(\mathbf{y}|\mathbf{a})$의 변분 하한으로 정의하고 모델의 파라미터 집합 W로 편미분 하면 다음과 같이 정리할 수 있다[7]. 이 때 $p(\hat{z}|a)$에 대하여 몬테-카를로 샘플링(Monte Carlo sampling) 방법을 사용하여 근사한다.

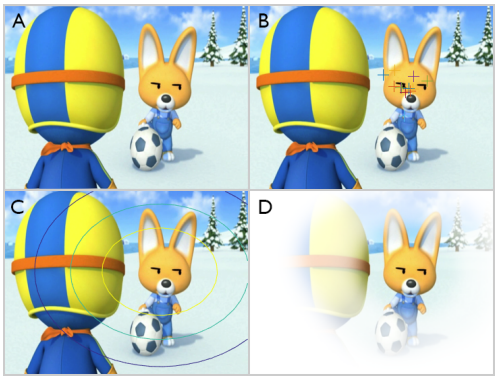

그림 2: 시선 정보를 활용하는 방법을 도식화. A는 원본 이미지, B는 A를 응시한 서로 다른 13명의 응시 좌표들을 십자가 기호로 표시. C는 응시 좌표들을 가우시안 혼합 모델(Gaussian Mixture Model)로 나타냈을 때의 확률 분포를 등고선 표현. D는 몬테 카를로 방법으로 샘플링할 확률 가중치를 도식적으로 표현.

$$\frac{\partial L}{\partial W} \approx \frac{1}{N} \sum_{n=1}^{N} \left[\frac{\partial \log p(\mathbf{y} \mid \hat{z}, \mathbf{a})}{\partial W} + \log p(\mathbf{y} \mid \hat{z}, \mathbf{a}) \frac{\partial \log p(\hat{z} \mid \mathbf{a})}{\partial W} \right] \tag{8}$$

3 Story-Aware Attentional LSTM

만화 영상과 같이 입력 이미지들이 시간 축 상에서 문맥적 흐름을 가질 때 이러한 문맥을 보전하기 위해 이전 기억 상태와 은닉 상태를 유지할 필요가 있다. 따라서, $\mathbf{c_0}$와 $\mathbf{h_0}$를 이전 이미지 $\mathcal{I}_{\text{prev}}$의 마지막 상태 값 $\mathbf{c_T}$와 $\mathbf{h_T}$로 초기화 한다.

$\mathbf{f}_{\text{att}}$는 콘볼루션 반응 벡터 a_i를 은닉 벡터 $\mathbf{h}_{t-1}$를 통해 샘플링 하는 기제를 가진다. 은닉 벡터 $\mathbf{h}_{t-1}$에 따라 주의 성능에 영향을 미치게 되지만 이는 $\hat{\mathbf{z}}_{t-1}$에 의존적이므로 만화 영상 데이터에 대한 실험에서 성능을 크게 저하시킬 원인이 된다. 따라서 랜덤 파라미터로 $\mathbf{f}_{\text{att}}$를 초기화하는 대신 다수 사람들의 시선 분포 정보를 이용하여 파라미터 값을 초기화함으로써 $\mathbf{f}_{\text{att}}$에 대한 학습을 가속화하고 전체 모델에 대한 효율을 제고 하고자 한다.

관측된 시선 정보에 가우시안 혼합 모델(Gaussian Mixture Model)을 통하여 $\{\tilde{\alpha}_{ti}\}$를 얻은 후, $\mathbf{h}_{t-1}$는 랜덤 파라미터로 초기화 한 후 $\mathbf{a}_i$에 대한 $\mathbf{f}_{\text{att}}$를 학습한다. $\mathbf{h}_{t-1}$는 $\mathbf{f}_{\text{att}}$의 상위 층에서 $\mathbf{a}_i$의 추상 표현과 결합하게 정의하였기 때문에 전이 학습의 효과를 상대적으로 크게 얻을 수 있다[11].

4 실험 계획

콘볼루션 반응 값 $\mathbf{a}$를 얻기 위해 Oxford VGGnet[12]의 네번째 콘볼루션 층의 max-pooling 바로 전의 반응 값을 사용한다. 반응 크기는 14×14, 채널 수는 512 이다. 경험적으로, 일반 이미지에 대한 학습 후에는 학습된 특징 탐지기(feature detectors)들이 충분하기 때문에 추가적인 세밀 조정 학습은 실시하지 않았다. 관찰된 시선 정보에 대한 설명은 그림 2를 참고한다.

5 토 론

시간적 문맥과 시각-언어 개념 학습 조건에서는 유연한 모델과 학습 방법을 요구한다 [13]. 몬테 카를로 방법을 이용한 주의 모델은 한 이미지 내에서 필요한 정보를 능동적으로 탐색하므로 적은 파라미터 수로도 유연한 데이터 적용의 접점을 제공한다.

본 연구에서는 몬테 카를로 방법을 사용하고 있는 주의 모델에서 관찰된 사람의 시선 정보를 이용한다. 다수 사람들의 시선 분포는 시간에 따른 문맥을 고려한 정보이므로 깊은 순환 네트워크 모델의 빠른 수렴을 위한 도움을 준다.

6 결 론

기존 연구에서는 LSTM과 MLP에 기반한 주의 모델을 단일한 정보 흐름으로 역전파 알고리즘을 통해 학습할 수 있는 방법을 제안하였다. 그러나 이미지-문장 변환만 고려되어 있어 만화 영상과 같이 시간의 흐름에 따라 문맥이 달라질 수 있는 데이터에는 적합하지 않았다. 본 연구에서는 메타 정보인 사람의 시선 분포 정보를 이용하여 시간에 따른 문맥 정보를 고려할 때 발생할 수 있는 계산 비용 문제를 풀고자 하였다.

7 감사의 글

이 논문은 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원(NRF-2010-0017734-Videome)과 정보통신기술진흥센터의 지원(R0126-15-1072-SW스타랩, 10035348-mLife, 10044009-HRI.MESSI)을 받아 수행된 연구임.

참고 문헌

[1] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal Deep Learning," *Proceedings of The 28th International Conference on Machine Learning (ICML)*, pp. 689–696, 2011.

[2] N. Srivastava and R. R. Salakhutdinov, "Multimodal Learning with Deep Boltzmann Machines," in *Advances in Neural Information Processing Systems 25* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), pp. 2222–2230, Curran Associates, Inc., 2012.

[3] R. Kiros, R. Zemel, and R. Salakhutdinov, "Multimodal Neural Language Models," *Proc NIPS Deep Learning . . .* , 2013.

[4] K. Sohn, W. Shang, and H. Lee, "Improved Multimodal Deep Learning with Variation of Information," in *Advances in Neural Information Processing Systems 27* (Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds.), pp. 2141–2149, Curran Associates, Inc., 2014.

[5] H. Yu and J. M. Siskind, "Grounded Language Learning from Video Described with Sentences," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL2013)*, pp. 53–63, 2013.

[6] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," in *28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, 2015.

[7] K. Xu, A. Courville, R. S. Zemel, and Y. Bengio, "Show , Attend and Tell : Neural Image Caption Generation with Visual Attention," in *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, 2015.

[8] S. Oh, M. S. Lee, and B. T. Zhang, "Ensemble Learning With Active Example Selection For Imbalanced Biomedical Data Classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 2, pp. 316–325, 2011.

[9] B.-T. Zhang, "Accelerated Learning By Active Example Selection," *International Journal of Neural Systems*, vol. 5, no. 01, pp. 67–75, 1994.

[10] W. Zaremba, I. Sutskever, O. Vinyals, and G. Brain, "Recurrent Neural Network Regularization," no. 2013, pp. 1–8, 2015.

[11] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Advances in Neural Information Processing Systems 27* (Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds.), pp. 3320–3328, Curran Associates, Inc., 2014.

[12] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Sept. 2014.

[13] B.-T. Zhang, "An Incremental Learning Algorithm That Optimizes Network Size And Sample Size In One Trial," in *Neural Networks, 1994. IEEE World Congress on Computational Intelligence., 1994 IEEE International Conference on*, vol. 1, pp. 215–220, IEEE.