장단기 기억 신경망과 공간적 순환 신경망을 이용한 배경차분

*추성권, **조남익
서울대학교 전기정보공학부 뉴미디어통신공동연구소
*chewry@ispl.snu.ac.kr

Background subtraction using LSTM and spatial recurrent neural network

*Sungkwon Choo, **Nam Ik Cho
INMC, School of Electrical and Computer Engineering, Seoul National University

요 약

본 논문에서는 순환 신경망을 이용하여 동영상에서의 배경과 전경을 구분하는 알고리즘을 제안한다. 순환 신경망은 일련의 순차적인 입력에 대해서 내부의 루프(loop)를 통해 이전 입력에 의한 정보를 지속할 수 있도록 구성되는 신경망을 말한다. 순환 신경망의 여러 구조들 가운데, 우리는 장기적인 관계에도 반응할 수 있도록 장단기 기억 신경망(Long short-term memory networks, LSTM)을 사용했다. 그리고 동영상에서의 시간적인 연결 뿐 아니라 공간적인 연관성도 배경과 전경을 판단하는 것에 영향을 미치기 때문에, 공간적 순환 신경망을 적용하여 내부 신경망(hidden layer)들의 정보가 공간적으로 전달될 수 있도록 신경망을 구성하였다. 제안하는 알고리즘은 기본적인 배경차분 동영상에 대해 기존 알고리즘들과 비교할만한 결과를 보인다.

1. 서론

최근 깊은 신경망 구조를 사용하여 빅데이터에 대한 학습을 통해 정보를 이해, 분석하는 연구가 활발히 이루어지고 있다. 그 중에서도 정지 영상에 대한 성공적인 연구[1,2]들을 기반으로 깊은 신경망 구조를 사용하는 연구들이 영상처리의 다양한 분야에서 진행되는 중이다. 이는 정지영상과 밀접한 관련이 있는 동영상에서도 이루어지고 있는데, 동영상 설명(video captioning), 동영상 분류(video classification) 의 분야에서 가장 활발하게 진행되고 있다. 하지만 이러한 연구들은 정지영상에서의 시각적 학습에 의존적으로 진행되고 있으며 알고리즘의 출력 역시 정지영상에 대한 알고리즘과 비슷하게 하나의 동영상에 대한 하나의 문장, 분류의 형태를 갖는다. 동영상이 정지영상과 구분되는 가장 큰 특징은 정지영상의 순차적인 나열을 통하여 시간적 정보를 포함하고 있는 것이라고 할 수 있다. 동영상에 대한 기존의 영상처리 알고리즘들은 확률 모델을 생성하거나 다른 시간에서의 같은 위치를 찾는 방법 등으로 쌓여있는 정지영상들에서 시간적 정보를 추출하였다. 동영상 설명이나 분류 알고리즘들도 시간적 정보를 추출하기는 하지만, 기존 정지영상에 대해 학습한 깊은 신경망 구조를 사용하고, 동영상 전체에서 하나의 특징을 추출한기 때문에 시간적 정보를 충분히 사용하는 것을 확인하기 어렵다.

본 논문에서 제안하는 깊은 신경망을 이용한 배경차분 기법은 기존의 정지영상에서 학습한 깊은 신경망 구조를 사용하지 않으면서 동영상과 같은 크기, 길이의 출력을 생성한다. 각 프레임마다 결과가 출력되기 때문에 시간적 정보가 추출되는 것을 확인할 수 있다. 시간적 정보에 대한 학습을 위해 순환 신경망의 여러 구조들 가운데, 우리는 장기적인 상관관계에도 반응할 수 있도록 장단기 기억 신경망(Long short-term memory networks, LSTM)[3]을 사용했다. 여기에 더하여 공간적인 연관성을 사용하기위해 공간적 순환 신경망을 내부 상태(hidden state)에 적용하는 구조를 설계하였다.

깊은 신경망 구조인 deep auto-encoder 나 convolutional 신경망 구조를 사용하여 배경차분을 하는 알고리즘들이 최근 제안되었지만[4,5], 우리가 아는 선에서 순환 신경망을 사용하여 배경차분 알고리즘을 수행하는 것은 제안하는 방법이 처음이다.

본 논문의 구성은 다음과 같다. 2 절에서는 장단기 기억 신경망과 공간적 순환 신경망의 구조에 대해 살펴본 후, 3 절에서는 실험 환경과 결과에 대해서 서술한다. 마지막으로 4 절에서는 본 논문에 대한 결론을 맺으며 향후 연구방향을 정리한다.

2. 배경차분

이 절에서는 배경차분을 위해 본 논문이 제안하는 알고리즘의 구조를 설명한다.

2.1 장단기 기억 신경망(Long short-term memory)

장단기 기억 신경망(Long short-term memory, 이하 LSTM)은 역전달(back propagation)과정에서 입력 순서 상 멀리 있는 입력까지의 미분 값이 사라지지 않도록 구현된 신경망이다. LSTM 에도 여러가지 종류가 있지만 제안하는 알고리즘에서는 peephole 연결이 있는 것을 사용하였다. 전체 gate(input, forget, output)와 cell, hidden state 의 식은 아래와 같다.

$$
\begin{aligned}
i_t &= \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right)\\
f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right)\\
c_t &= f_t c_{t-1} + i_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right)\\
o_t &= \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right)\\
h_t &= o_t \tanh(c_t)
\end{aligned}
$$

동영상 전체를 하나의 단일한 입력으로 본다면 동영상의 크기는 시간 T × 높이 H × 너비 W × 깊이 C 의 크기가 될 것이고, 이것은 하나의 LSTM 신경망으로 처리하기에는 물리적 메모리의 한계가 있다. 때문에 제안하는 알고리즘에서는 픽셀별로 LSTM 신경망을 만들어 가중치 행렬(weight matrix)은 공유하면서 내부 상태(hidden state)는 각 픽셀별로 다른 값을 갖도록 구조를 설계하였다. 이를 통해 각 LSTM 신경망은 배경을 학습하는 같은 작업을 수행하지만, 입력 값에 따라 내부의 학습된 배경 모델은 서로 다른 값을 학습하게 된다. 이러한 구조는 기존의 배경차분 알고리즘에서 픽셀 별로 똑같은 작업을 수행하지만 입력 값에 따라 서로 다른 확률 모델을 만드는 과정[6,7]과 유사하다. 이와 같이 제안하는 방법은 LSTM 이 배경모델을 생성하는 과정을 자동적으로 학습할 수 있도록 모든 픽셀이 같은 구조를 갖는 신경망을 설계하였다.

2.2 공간적 순환 신경망(spatial recurrent neural network)

배경모델을 학습하기 위해 2.1 의 내용과 같이 각 픽셀에서 동일한 작업을 수행하는 LSTM 을 설계하였다. 하지만 배경차분에서는 시간적 정보뿐만 아니라 공간적으로 가까운 영역들간의 연관성(spatial correlation) 역시 배경여부를 판단하는 데에 중요한 역할을 한다. 공간적 연관성을 위해 최근에 순환 신경망을 공간적 시퀀스에 적용하는 연구들[8,9]이 좋은 성능을 보이고 있는데, 우리는 간단한 4 방향 순환 신경망을 사용하여 이를 구현하였다. 순환 신경망으로는 GRU(Gated Recurrent Unit)[10]를 사용하였는데 이는 속도면에서 LSTM 보다 빠르면서 길이가 긴 입력에 대해서도 대응할 수 있도록 설계된 신경망이기 때문이다.

2.1 의 과정을 통해 생성된 출력을 2 차원 형태로 변환한 뒤 4 방향(좌-우, 우-좌, 상-하, 하-상) 으로 GRU 에 통과시킨 다음, 이것을 쌓아서 convolutional 신경망을 통해 4 방향의 정보를 결합시킨다. 이를 통하여 공간적 연관성을 결합한 특징을 학습하게 된다. 학습된 특징들은 모두 연결된(dense, fully connected) 신경망을 통하여 하나의 이진 결과값을 출력하게 된다.

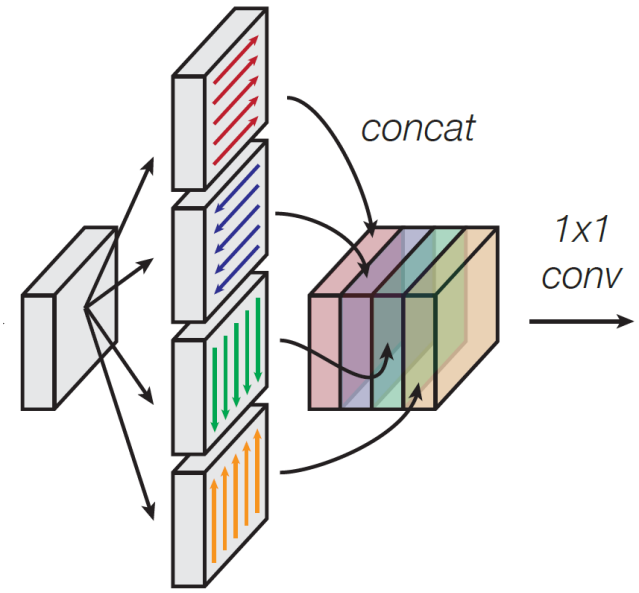

그림 1. 공간적 순환 신경망 구조

3. 실험 결과

실험은 changedetection2012 데이터[11]를 사용하였다. Over-fitting 현상을 막기 위해, 같은 dropout 네트워크를 공유하는 Bayesian dropout[12] 을 사용하였다. 입력과 내부 네트워크에 대한 dropout 모두 0.25 로 같은 값을 사용하였다.

깊은 신경망을 사용하기 위해서는 학습이 필요한데, 우리는 데이터에서 제공하는 groundtruth 에서 각 동영상의 전반부 절반을 학습에 사용하였고 이를 통해 배경차분 결과를 추출하였다. 동영상의 공간적인 영역은 너무 가까운 경우에는 정보량에 차이가 없기 때문에 간격(stride)를 2 로 설정하여 모든 픽셀을 학습하는 대신, 빠르게 반복학습할 수 있도록 하였다. LSTM-GRU-CNN 을 하나의 세트로 하여 3 개의 층(layer)를 쌓은 구조로 설계하여 학습하였다. 시간적인 정보를 추출하는 LSTM 은 30 개의 노드를 갖고, 마지막 LSTM 에서만 50 개의 노드를 갖도록 했다. 공간적 정보를 추출하는 GRU 는 모든 층에서 20 개의 노드로 설정하였다. 4 방향의 순환 신경망은 LSTM 신경망이 모든 픽셀에서 같은 가중치 행렬을 공유하는 것과 마찬가지로 모두 같은 가중치 행렬을 공유하도록 했는데, 이는 방향이 다를 뿐 같은 목적을 갖는 신경망이기 때문이다. 4 방향의 순환 신경망 결과를 결합하기 위한 convolutional 신경망은 1×1 크기의 필터를 30 개 사용하였고, 마지막 층에서만 10 개의 필터를 사용하도록 하였다. 이후의 모두 연결된 신경망에서는 하나의 출력으로 전경 배경을 구분하도록 연결하였으며 sigmoid 함수를 출력에 사용하였다. Loss 함수는 전경 배경 분류 문제이기 때문에 binary cross entropy 함수를 사용했고, 최적화 방법은 순환 신경망 구조에서 널리 사용되는 Adam[13] 최적화 방법을 적용하였다. 동영상의 3 채널 정보를 모두 사용하는 것과 1 채널 정보만을 사용하는 것이 큰 차이가 없어서, grayscale 영상으로 변환한 영상을 입력으로 사용하였다. Baseline 카테고리에 대하여 실험한 결과는 아래 표와 같다.

algorithm	video	recall	precision	F
SuBSENSE [14]	pedestrians	0.9615	0.9311	0.9461
	PETS2006	0.9446	0.9189	**0.9315**
	highway	0.9518	0.9355	0.9436
	office	0.9053	0.9721	0.9375
FTSG[15]	pedestrians	0.9786	0.8902	0.9323
	PETS2006	0.9630	0.8830	0.9212
	highway	0.9555	0.9338	**0.9446**
	office	0.9081	0.9611	0.9338
proposed	pedestrians	0.9500	0.9582	**0.9541**
	PETS2006	0.8400	0.7418	0.7878
	highway	0.9762	0.8819	0.9267
	office	0.9566	0.9471	**0.9518**

표 1. 알고리즘 별 성능비교

비교 알고리즘은 단일 알고리즘 중에서 가장 성능이 좋은 알고리즘[14,15]의 결과를 표시한 것이다. 제안하는 방법이 하나의 동영상을 제외하고는 비교할만한 성능을 보이는 것을 확인할 수 있었다.

그림 2. 제안하는 방법의 결과

4. 결론

본 논문에서는 순환 신경망을 이용하여 시간적, 공간적 정보를 추출하여 배경차분을 수행하였고 가장 기본적인 카테고리의 동영상에 대해 좋은 성능을 보였다. 하지만 다른 카테고리의 경우 성능이 하락하는 결과를 보였는데, 이는 신경망이 표현할 수 있는 크기가 작고, 충분한 학습을 하기 위한 데이터가 부족하기 때문인 것으로 보인다. 이러한 문제를 해결하기 위한 새로운 신경망 구조와 전처리 방법을 구현하는 것이 후속 연구로 요구된다.

감사의 글

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 대학 ICT 연구센터육성지원사업의 연구결과로 수행되었음(IITP-2016-R2718-16-0014)

참고문헌

[1] A. Krizhevsky, S. Ilya, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[2] K. Simonyan, A. Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[3] S. Hochreiter, J. Schmidhuber. "Long short-term

memory." Neural computation 9.8 (1997): 1735-1780.

[4] M. Braham, Marc Van Droogenbroeck. "Deep Background Subtraction with Scene-Specific Convolutional Neural Networks." International Conference on Systems, Signals and Image Processing, Bratislava 23-25 May 2016. IEEE, 2016.

[5] Zhang, Yaqing, et al. "Deep learning driven blockwise moving object detection with binary scene modeling." Neurocomputing 168 (2015): 454-463.

[6] Z. Zivkovic. "Improved adaptive Gaussian mixture model for background subtraction." Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. Vol. 2. IEEE, 2004.

[7] A. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background subtraction," in Proc. Eur. Conf. on Computer Vision, Lect. Notes Comput. Sci. 1843, 751-767 2000.

[8] Byeon, Wonmin, et al. "Scene labeling with lstm recurrent neural networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[9] Bell, Sean, et al. "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks." arXiv preprint arXiv:1512.04143 (2015).

[10] K. Cho et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078 (2014).

[11] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, changedetection.net: A new change detection benchmark dataset, in Proc. IEEE Workshop on Change Detection (CDW-2012) at CVPR-2012, Providence, RI, 16-21 Jun., 2012.

[12] Y. Gal, Z. Ghahramani "A theoretically grounded application of dropout in recurrent neural networks." Advances in neural information processing systems. 2016.

[13] D.P. Kingma, J. Ba, "Adam: A Method for Stochastic Optimization". in Proc. The International Conference on Learning Representations (ICLR), San Diego, 2015

[14] St-Charles, P.-L., Bilodeau, G.-A., Bergevin, R., "SuBSENSE : A Universal Change Detection Method with Local Adaptive Sensitivity". IEEE Transactions on Image Processing 24.1 (2015): 359-373.

[15] R. Wang, F. Bunyak, G. Seetharaman and K. Palaniappan "Static and Moving Object Detection Using Flux Tensor with Split Gaussian Models", in proc of IEEE Workshop on Change Detection, 2014