Fully Convolutional Network 기반 관심 영역 검출 기법의 속도 개선 연구

*, **황현수 *정진우 *김용환 **최윤식
*전자부품연구원, **연세대학교 전기전자공학부
hshwang@keti.re.kr

A Study on Improving Speed of Interesting Region Detection Based on Fully Convolutional Network

*, **Hwang, Hyun-Su *Jung, Jin-woo *Kim, Yong-Hwan **Choe, Yoon-Sik
*Korea Electronics Technology Institute
**School of Electrical and Electronic Engineering, Yonsei University

요약

영상의 관심 영역 검출은 영상처리 및 컴퓨터 비전 응용 분야에서 꾸준하게 사용되고 있는 기법이다. 특히, 근래 심층신경망 연구의 급격한 발전에 힘입어 심층신경망을 이용한 관심 영역 검출 기법에 대한 연구가 활발하게 진행되고 있다.

한편 Fully Convolutional Network(이하 FCN)은 본래 심층 예측(Dense Prediction)을 통한 의미론적 영상 분할(Semantic Segmentation)을 수행하기 위해 제안된 심층신경망 구조이다. FCN을 영상의 관심 영역 검출에 활용하여도 기존 관심 영역 검출 기법과 비교하여 충분히 좋은 성능을 발휘할 수 있다. 그러나 FCN에 사용되는 convolution 층의 수가 많고, 이에 따른 가중치(weight)의 개수도 기하급수적으로 늘어나 검출에 필요한 시간 복잡도가 매우 크다는 문제점이 있다.

따라서 본 논문에서는 기존 FCN이 가진 검출 시간 복잡도의 문제점을 convolution 층의 가중치 관점에서 해결하고자 이를 조절하여 FCN의 관심 영역 검출 속도를 향상시키는 방법을 제안한다. 적절한 convolution 층의 가중치를 조절함으로써, MSRA10K 데이터셋 환경에서 검출 정확도를 크게 저하시키지 않고도 최대 약 20.5%만큼 검출 속도를 향상시킬 수 있었다.

1. 서론

일반적으로 영상의 관심 영역 검출이란, 영상으로부터 고차원적인 정보를 얻기 위해 시각적으로 구분되는 영역을 추출하는 일련의 단계를 말한다. 관심 영역 검출의 핵심은 한정된 감각 자원을 관심 영역에 효율적으로 할당하고, 나머지는 무시할 수 있도록 인지를 억제하는 것이다. 관심 영역의 검출을 통해 장면을 대표하는 객체나 영역을 찾을 수 있고, 수많은 영상 데이터를 처리하는 과정에서 작지만 중요한 일부 객체나 영역을 찾을 수 있어 중요한 기법이라고 할 수 있다. 이러한 관심 영역의 검출은 영상 분할, 객체 검출, 동영상 요약, 동영상 압축 및 영상 합성 등 영상처리 및 컴퓨터 비전 분야에 널리 응용되고 있다.

기존의 관심 영역의 검출을 위한 다양한 기법들이 제안되었다. 초기에는 영상으로부터 색상이나 밝기, 질감(texture), 크기나 방향과 같이 저수준의 특징을 조합하여 관심 영역의 척도로 삼았다[1]. 그러나 이러한 저수준의 특징들은 영상의 배경이 복잡하거나 배경으로부터 관심 영역을 구분할 수 있을 만큼 특징간의 뚜렷한 차이가 적을 경우 관심 영역을 효과적으로 찾는 데 실패하는 경우가 많았다[2]. 근래에 심층신경망 연구가 활발히 이루어짐에 따라 이를 극복하기 위해 심층신경망을 이용한 관심 영역 검출 기법들이 연구되고 있다. 특히 convolution 신경망(이하 CNN)의 경우, 인간의 시각 시스템이 영상을 인식하는 과정과 유사한 구조를 가지고 영상의 특징을 추출한다[3]. 이러한 CNN 기반 특징들은 영상으로부터 복잡한 패턴을 스스로 학습할 수 있어 다양한 형태의 수많은 영상이 주어졌을 때 매우 효과적으로 관심 영역을 검출할 수 있다.

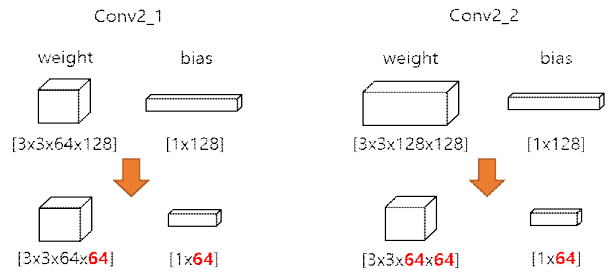

그림 1. 속도 개선을 위한 가중치 kernel 개수 감소

CNN의 종류 중 하나인 FCN은 본래 영상을 화소단위로 밀도 있게 예측하여 이를 각각 주어진 부류(class)별로 분류하는 의미론적 영상 분할을 위해 제안되었다[4]. FCN은 기존의 CNN구조와는 달리 망의 Fully Connected(이하 FC)층을 convolution 층으로 대체한 것이 특징으로, FC 층으로 인한 제약사항들을 개선할 수 있게 되었다. 관심 영역 검출 또한 영상 분할의 한 갈래이기 때문에, 이러한 목적에 맞게 FCN 구조를 변형하는 것은 어렵지 않다. 그러나 FCN은 다수의

convolution 층으로 구성되어 있어 검출 시에 convolution 연산량이 누적돼 많은 연산부하를 주고, 이는 검출의 시간복잡도 증가의 부작용으로 나타난다. 다시 말해, convolution 층에서 연산량 증가에 가장 중대한 역할을 하는 요인은 바로 가중치의 개수이다.

본 논문은 먼저 FCN 기반의 관심 영역 검출 구조에 대해서 간단하게 소개하고, 검출 속도를 개선하기 위해 가중치를 각 층에 순차적으로 조절하여 실험한 결과를 분석하였다. 그림 1은 제안한 구조의 일부 계층을 도시한 것이다.

2. 관심 영역 검출을 위한 FCN 구조

Long 등은 의미론적 영상 분할을 위해 FCN 구조를 제안하였다[4]. 이들의 FCN은 Simonyan 등이 영상 인식을 위해 제안한 VGGnet[5] 구조를 채용하였다. VGGnet과의 주요 차이점으로는 분류를 위한 마지막 층을 제거하고, FC 층을 모두 convolution 층으로 대체하였다. 또한 원본 영상과 동일한 크기의 결과를 얻기 위해, upsampling과정을 수행하는 deconvolution 층을 추가하였다. 마지막으로 convolution 과정을 지속적으로 거치면 영상의 세밀한 부분이 점차 사라지기 때문에, skip 층을 통해 이전 층의 결과와 결합하여 영상의 세밀한 부분을 보완하였다.

이러한 FCN 구조의 이점으로는 첫째, 기존 FC 층의 경우 특성상 입력영상과 가중치와의 행렬의 곱셈연산을 수행하기 때문에 입력영상의 크기가 고정되어야 하나, FCN은 이를 convolution 층으로 대체하고 sliding window 개념으로 간주하여 연산을 수행하므로 영상의 크기 제약에서 자유롭다. 둘째, 1×1 convolution 연산을 수행하기 때문에 입력 영상의 위치 정보가 보존된다. 마지막으로, skip과 upsampling 과정을 통해 영상의 세밀한 부분이 비교적 잘 보존되며, 원본 영상과 동일한 크기의 결과 영상을 얻을 수 있다. 그림 2는 본 논문에서 다루는 FCN 구조를 간략하게 도시한 것이다. 좌측 파란색 층들은 convolution 층을 나타내며, crop 층과 같이 부수적인 역할을 하는 층은 생략하였다.

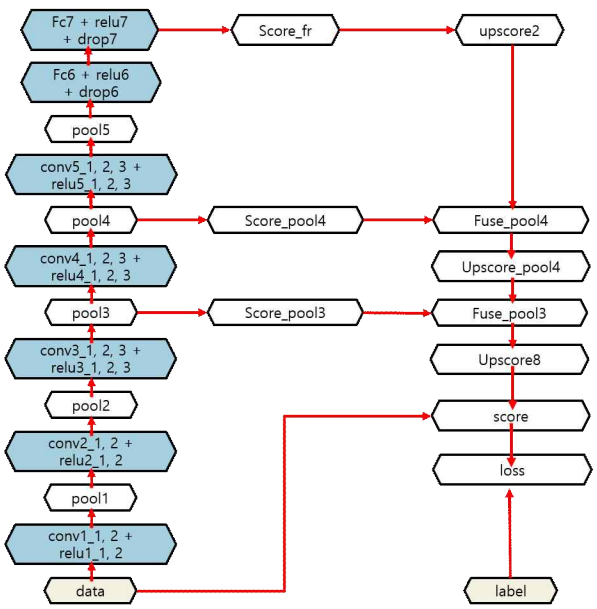

그림 2. 관심 영역 검출을 위한 FCN 구조도

기존 FCN 구조를 관심 영역 검출을 위해 변형하는 과정은 매우 간단하다. 관심 영역 검출의 결과는 관심 영역 지도(map)의 형태로 얻게 되며, 이는 관심 영역과 무관심 영역의 이진 영상형태로 나타낼 수 있다. 즉, 입력 영상의 레이블(label)로 관심 영역 지도를 사용하고, 가장 마지막 convolution 층에서 부류의 개수를 두 가지(관심 영역과 무관심 영역)로 축소하여 학습 및 검출을 수행할 수 있다. 그림 3은 관심 영역 검출을 위한 FCN의 입력 및 출력 영상을 나타낸 것이다.

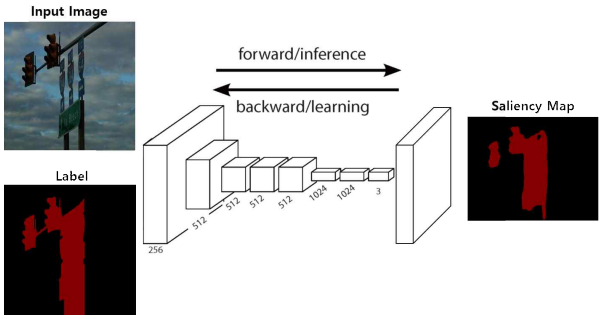

그림 3. 관심 영역 검출 과정의 입력 및 출력 영상

3. 검출 속도 개선을 위한 가중치 조절

그림 2에서 확인할 수 있듯이, FCN의 문제점은 다수의 convolution 층으로 인한 연산부하가 크다는 것이다. 즉 convolution 연산을 위해 필요한 가중치 및 편향(bias)의 개수가 매우 많다. 특히 convolution kernel의 배치(batch)수가 가장 크다. 표 1은 FCN의 각 convolution 층의 가중치와 편향의 차원을 각각 [너비×높이×채널(channel)×배치], [1×가중치의 배치]로 표기한 내역이다. 마지막 항목에 해당하는 배치의 수가 지속적으로 증가하고, 이는 바로 다음 층의 채널 수와 동일하기 때문에 배치수의 증가는 결국 전체 가중치의 큰 증가폭을 유도하게 된다.

층	가중치	편향
conv1_1	[3×3×3×64]	[1×64]
conv1_2	[3×3×3×64]	[1×64]
conv2_1	[3×3×64×128]	[1×128]
conv2_2	[3×3×128×128]	[1×128]
conv3_1	[3×3×128×256]	[1×256]
conv3_2	[3×3×256×256]	[1×256]
conv3_3	[3×3×256×256]	[1×256]
conv4_1	[3×3×256×512]	[1×512]
conv4_2	[3×3×512×512]	[1×512]
conv4_3	[3×3×512×512]	[1×512]
conv5_1	[3×3×512×512]	[1×512]
conv5_2	[3×3×512×512]	[1×512]
conv5_3	[3×3×512×512]	[1×512]
fc6	[7×7×512×4096]	[1×4096]
fc7	[1×1×4096×4096]	[1×4096]

표 1. FCN의 각 convolution 층의 가중치 및 편향의 차원

이와 같이 많은 연산량이 요구될 경우, 대용량의 데이터를 처리하는 심층신경망의 특성상 특히 검출해야 할 데이터의 개수가 늘어날수록 검출 속도의 현저한 저하를 불러올 수 있다. 따라서 본 논문에서는 각 convolution 층 가중치의 배치수를 기존의 1/2로 줄인 새로운 가중치가 검출 성능에 어떠한 영향을 주는지 확인하기 위한 실험을 진행하였다. 표 2는 실험별로 배치수를 조절하여 차원이 변경된 층의 내역이다. 이름이 '_half'로 변경된 층은 각 실험에서 실제로 배치수가 감소한 층이고, 그렇지 않은 층은 이전 층의 배치수가 감소하였기 때문에 단순히 입력 채널만 감소한 층이다. 각 실험별로 내역에 없는 나머지 층은 원래의 가중치 내역을 그대로 보존하여 실험을 진행하였다.

실험	층	가중치	편향
(a)	모든 conv 층 축소	-	-
(b)	conv2_1_half	[3×3×64×**64**]	[1×**64**]
	conv2_2_half	[3×3×**64**×**64**]	[1×**64**]
	conv3_1	[3×3×**64**×256]	[1×256]
(c)	conv3_1_half	[3×3×128×**128**]	[1×**128**]
	conv3_2_half	[3×3×**128**×**128**]	[1×**128**]
	conv3_3_half	[3×3×**128**×**128**]	[1×**128**]
	conv4_1	[3×3×**128**×512]	[1×512]
(d)	conv4_1_half	[3×3×256×**256**]	[1×**256**]
	conv4_2_half	[3×3×**256**×**256**]	[1×**256**]
	conv4_3_half	[3×3×**256**×**256**]	[1×**256**]
	conv5_1	[3×3×**256**×512]	[1×512]
(e)	conv5_1_half	[3×3×512×**256**]	[1×**256**]
	conv5_2_half	[3×3×**256**×**256**]	[1×**256**]
	conv5_3_half	[3×3×**256**×**256**]	[1×**256**]
	fc6	[7×7×**256**×4096]	[1×4096]
(f)	fc6_half	[7×7×512×**2048**]	[1×**2048**]
	fc7	[1×1×**2048**×4096]	[1×4096]
(g)	fc7_half	[1×1×4096×**2048**]	[1×**2048**]
(h)	fc6_half	[7×7×512×**2048**]	[1×**2048**]
	fc7_half	[1×1×**2048**×**2048**]	[1×**2048**]

표 2. 각 실험별로 변경한 가중치 및 편향의 차원

4. 실험 결과 및 분석

FCN 학습 및 검출을 위한 데이터셋은 관심 영역용으로 널리 사용되는 MSRA10K[6]와 DUT-OMRON[2]의 두 종류를 사용하였다. 심층 학습 프레임워크로 Caffe[6]를 사용하여 python 코드로 구현하였으며, GPU는 NVIDIA GTX 1080ti 한 개를 사용하였다. 학습의 경우 기존의 FCN-8s 모델을 표 2와 같이 조절하여 재학습을 수행하였으며, 최적화는 SGD 알고리즘을 이용하였다. 학습의 hyperparameter로 학습율(learning rate)은 1e-14로 고정하였고, momentum은 0.99, weight decay는 0.0005로 10만 번 반복하도록 하였다. 성능 평가의 척도는 검출 소요 시간과 식 (1)의 F-Score를 사용하였다[8].

$$F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 \times Precision + Recall} \quad (1)$$

표 3은 MSRA10K 데이터셋에 대한 실험 (a)-(h)의 관심 영역 검출 시간 및 F-Score 결과이다. 실험 (a)와 같이 무작정 모든 convolution 층의 가중치를 줄이는 것은 크게 성능을 저하시킨다. 실험 (f)의 경우, 검출 시간 측면에서 약 20.5%의 향상을 보인 반면, F-score의 저하는 약 1.8%에 불과해 나머지 실험에 비해 좋은 성능을 보임을 알 수 있다. 그림 4는 각 실험별 검출의 결과로 얻은 관심 영역 지도 영상이다. 상단의 경우 비교적 간단한 형태의 관심 영역으로 대부분의 실험이 비교적 유사하게 검출하는 반면, 하단의 경우 비교적 복잡한 형태를 가져 상단에 비해 검출이 쉽지 않음을 확인할 수 있다.

실험	검출 시간	F-Score
기존 FCN	84.44	0.942
(a)	54.409	0.5078
(b)	72.401	0.773
(c)	69.138	0.716
(d)	67.044	0.747
(e)	65.769	0.807
(f)	**67.139**	**0.9246**
(g)	70.742	0.92
(h)	66.452	0.9087

표 3. MSRA10K 데이터셋에 대한 검출 시간 및 F-score 결과

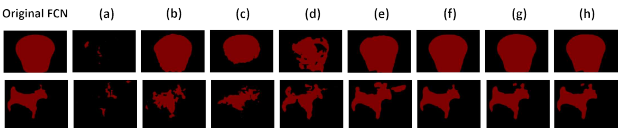

그림 4. MSRA10K 데이터셋에 대한 관심 영역 지도 영상

표 4는 DUT-OMRON 데이터셋에 대한 실험 (a)-(h)의 관심 영역 검출 시간 및 F-Score 결과이며, 그림 5는 각 실험별 검출의 결과로 얻은 관심 영역 지도 영상이다. DUT-OMRON에 대한 실험 또한 역시 실험 (f)의 결과가 검출 시간 측면에서 약 8.9%의 향상을 보인 반면, F-score는 4.6%의 저하를 보여 종합적으로 가장 우수한 성능을 보였다. 실험 (f)에서 변경한 fc6 층은 이후의 fc7 층과 더불어 후반부의 convolution 층에 위치해 있다. CNN의 특성상 초반부의 convolution 층은 영상의 세밀한 특징을 파악하고, 후반부의 convolution 층은 전체적인 특징을 파악하는 데 비중을 둔다. 따라서 앞선 층들에서 어느 정도 충분히 영상의 세밀한 특징들을 학습했기 때문에 후반부의 convolution 층의 가중치를 줄여도 검출 정확도가 비교적 큰 손실을 입지 않고, 검출 시간을 줄일 수 있다고 분석할 수 있다.

실험	검출 시간	F-Score
기존 FCN	38.242	0.6902
(a)	27.933	0.1731
(b)	36.742	0.5105
(c)	35.523	0.4648
(d)	34.926	0.4091
(e)	34.39	0.5547
(f)	**34.851**	**0.6583**
(g)	37.295	0.6768
(h)	34.473	0.6412

표 4. DUT-OMRON 데이터셋에 대한 검출 시간 및 F-score 결과

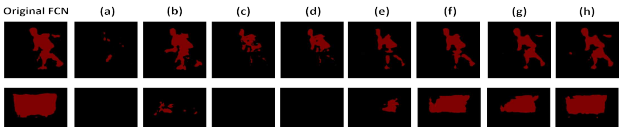

그림 5. DUT-OMRON 데이터셋에 대한 관심 영역 지도 영상

5. 결론

심층신경망을 이용한 관심 영역 검출 속도를 개선하기 위한 방법으로, 망의 대대적인 변형 없이 단순히 기존 모델의 가중치의 개수를 적절하게 줄임으로써 검출 정확도의 비교적 큰 손실이 없는 범위에서 검출 시간의 복잡도를 줄일 수 있다. 향후 연구에서는 좀 더 다양한 데이터셋에 대해서도 유사한 결과를 얻을 수 있는지를 검증할 예정이다. 또한 가중치의 차원을 감소시킨 후 가중치의 내역을 본 논문에서처럼 순차적으로 복제하는 대신에, 무작위나 감소한 부분의 가중치의 내역을 가지고도 검출 성능이 어떻게 변화하는 지에 대해서도 분석할 예정이다.

감사의 글

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2017년도 문화기술 연구개발 지원사업으로 수행되었음(R2017030018).

참고문헌

[1] A. Borji, M. M. Cheng, H. Jiang, and J. Li. (2014). “Salient object detection: A survey.” [Online]. Available: http://arxiv.org/abs/1411.5878

[2] G. Li and Y. Yu, “Visual saliency detection based on multiscale deep CNN features,” IEEE Trans. Image Process., vol. 25, no. 11, pp. 5012–5024, Nov. 2016.

[3] K. Fukushima, “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,” Biol. Cybern., vol. 36, no. 4, pp. 193–202, Apr. 1980.

[4] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vision Pattern Recog., 2015

[5] K. Simonyan, A. Zisserman, “Very deep convolutional networks for large-scale image recognition”, Proc. Int. Conf. Learn. Representations, 2015.

[6] M. M. Cheng, G. X. Zhang, N. J. Mitra, X. Huang, and S. M. Hu, “Global contrast based salient region detection,” in Proc. IEEE Conf. CVPR, Jun. 2011, pp. 409–416.

[7] Y. Jia, et al., “Caffe: Convolutional architecture for fast feature embedding,” arXiv:1408.5093, 2014.

[8] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, “Frequency-tuned salient region detection,” in Proc. IEEE Conf. CVPR, Jun. 2009, pp. 1597–1604.