Computer Vision 연구자가 Deep Learning의 시대를 사는 법

Micheal S. Ryoo

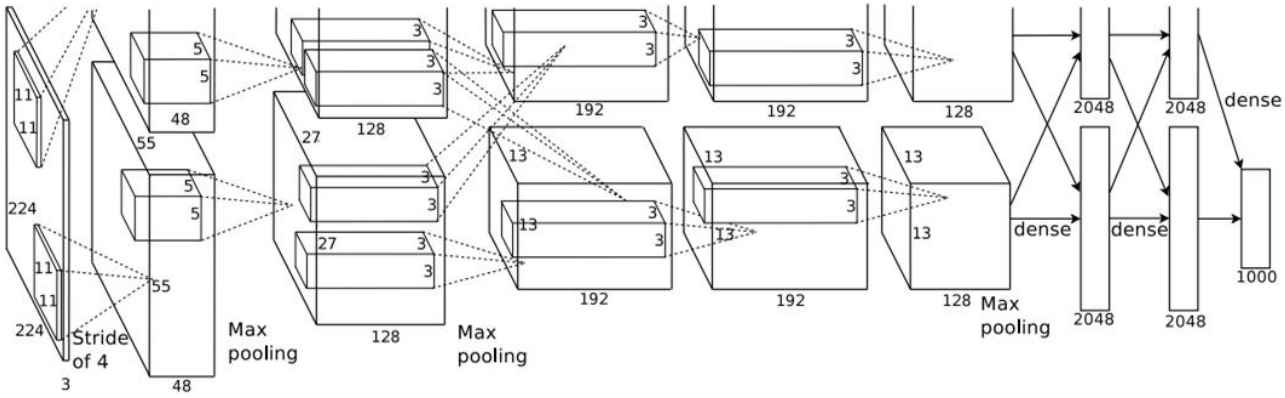
Deep learning의 시대이다. Big data라는 또 하나의 화려한 키워드와 결합되어 deep learning은 인공지능 분야 전체의 대세가 되었고, deep learning을 한다고 주장하지 않고서는 시대에 뒤떨어진 것으로 인식되는 상태에까지 이르렀다. Deep learning은 연구자가 과제제 안서를 쓰기위하여도 필수이며 거의 모든 (멀티미디어 관련) 스타트업 회사 또한 deep learning을 한다고 주장한다.

좀 더 technical하게는 convolutional neural network (CNN)이라고 불리는 이 deep learning은, 개념적으로는 1970~1980년대의 neural network의 현대판 부활이라고 볼 수 있다. 다수의 layer를 쌓고 학습 시킬수 있는 현대의 컴퓨터 처리 성능과 방대한 양의 데이터가 만나고, 거기에 몇가지 중요한 연구 (non-linear한 convolutional layer사용, layer의 Deep Boltzmann Machine으로서의 해석, stochastic gradient descent의 사용 등)가 더해져서 현재의 모습으로 탄생하게 되었다.[1,2]

Deep learning의 시대이다. Big data라는 또 하나의 화려한 키워드와 결합되어 deep learning은 인공지능 분야 전체의 대세가 되었다.

CNN은 Machine Learning framework의 하나로 볼 수 있는데, CNN과 종래의 framework과의 가장 큰 차이점은 raw data에서 feature를 자동으로 학습한다는 것이다. 종래의 Machine Learning framework은 (1) 데이터에서 사람의 전문가적 지식에 의해서 '설계된' feature (예: histogram of oriented gradients – HOG)를 추출한 후에 (2) SVM등의 classifier를 적용하는 단계로 이루어졌다. 이에 반하여 CNN은 다수의 convolutional layer를 통하여 raw input을 단계적으로 처리하여, 결과적으로 사람이 만든 것이 아니라 data에 최적화

Micheal S. Ryoo
Jet Propulsion Laboratory, NASA



〈그림 1〉 물체인식에 사용된 convolutional neural network의 예제[3].
5개의 convolutional layer를 가진다. 이 구조를 바탕으로 다수의 공개 프로그램 library가 개발되었다.[4, 5]

되어서 자동적으로 학습된 feature들을 convolutional layer들을 거치면서 추출하게 된다.

이러한 deep learning은 특히 speech recognition에서 현격한 성공을 보였으며, 최근 2~3년간 이미지를 기반으로한 Computer Vision 분야에서도 굉장한 성공을 보였다. 특히 백만개 이상의 이미지 데이터를 (학습에) 사용하는 object recognition (물체인식)에서 가능성을 보였는데, 일례로 ImageNet challenge에서 종래의 방법론 (앞서 언급된 사람이 설계한 HOG등 feature를 사용하는 방법)을 크게 뛰어넘는 성능을 보였다.[3] 〈그림 1〉은 그러한 ImageNet challenge의 물체인식을 위해서 2012년에 사용된 CNN의 구조를 나타낸다.

Deep learning의 성능을 결정짓는 가장 큰 요인은 (1) 데이터의 양과 (2) 컴퓨터의 연산 규모이다.

Approach 의 시대에서 Infrastructure의 시대로

Convolutional neural network (CNN)의 강점은 수백만/수천만 이상의 parameter를 내부에 포함하고 있는 high capacity model이라는 것에 있다. 따라서 만약 이를 모두 활용할 수 있는 방대한 양의 데이터 (big data!)가 제공되고 이를 바탕으로 학습을 수행할 수 있는 컴퓨터 연산 성능을 보유하고 있다면, CNN은 데이터에 최적화된 최상의 성능을 얻을 수 있다. 따라서, CNN 성능을 결정짓는 가장 큰 요인은 (1) 데이터의 양과 (2) 컴퓨터 연산 능력이라고 볼 수 있다.

생각해 보아야 할 것은 이러한 격차가 사람의 의해서 발생하는 것이 아니라 컴퓨터나 데이터에의 접근 가능 여부의 의해서 발생한다는 것이다. Deep learning의 시대에 성능의 관건은 연구자의 영감에 의해서 탄생한 새로운 알고리즘이 아니라 컴퓨터 연산과 데이터의 양이다. 아래에도 언급되겠지만, 이러한 연산 능력이나 데이터에 대한 준비 없이 deep learning으로 Google이나 facebook 같은 (모든 것을 보유한) 회사와 경쟁하겠다는 것은 굉장히 무모한 일이다. ImageNet 2014 object classification challenge에서 우승한 Google의 경우 22개의 convolutional layer를 가지고 있는 CNN을 학습시켜 사용하였다. 현재 학계에서 주로 사용하고 있는 CNN library (overfeat [4], caffe [5] 등)들이 고작 5개의 convolutional layer를 사용한다는 사실을 생각하여 보면 〈그림 1〉 학계와 기업 간의 격차가 발생하고 있다고 볼수 있다.

Deep learning을 '하는' 사람들

Deep learning을 '하는' 사람들은 CNN의 구조 그 자체나 그를 위하여 사용되는 학습 알고리즘 등을 연구하는 사람들이다. 이러한 연구자들과 그들의 성과는 주로 NIPS나 ICML, ICLR 등의 학회에서 찾아볼 수 있다. 특히 ICLR은 이러한 deep learning – CNN 연구를 출판하

기 위하여 탄생한 학회라고 봐도 좋다.

Deep learning 연구를 '하는' 가장 좋은 방법은 위에 언급된 컴퓨터 연산과 데이터를 제공해 줄 수 있는 기업들과 협력하거나 그러한 기업들의 일원이 되는 것이라고 할 수 있다. Deep learning의 선구자들이라 할 수 있는 토론토 대학의 Geoffrey Hinton은 Google에 부분적으로 몸을 담고 있고 NYU의 Yann LeCun은 아예 Facebook으로 주소속기관을 옮겼다. 예를 들어 Google 같은 경우는 내부적으로 수천만개 이상의 labeled image dataset을 가지고 있고 이러한 데이터의 활용은 다른 곳에서 불가능한 deep learning 연구를 가능케 한다. 기업들이 제공할 수 있는 컴퓨터 연산의 위력도 학계와는 비교를 불허한다.

물론 이러한 컴퓨터와 데이터에 대한 접근이 가능하지 않다고 하더라도 deep learning과 관련된 알고리즘 자체에 집중하는 연구를 수행하는 것은 가능하다. 이러한 경우는 (주로) 개발된 CNN 관련 방법론을 보다 작은 CNN (예: 앞서 언급된 5-layer CNN)에 적용하여 누구나 접근이 가능한 public dataset (예: ImageNet dataset) 등을 통하여 검증하는 방식으로 이루어진다. 물론, 이러한 연구가 실질적인 차원에서 빛을 보기 위해서는 후에 기업들이 이를 가져다가 자신들의 방대한 데이터 및 복잡한 CNN에 적용하는 것을 기대해야만 할 것이다.

Deep Learning을 '쓰는' 사람들

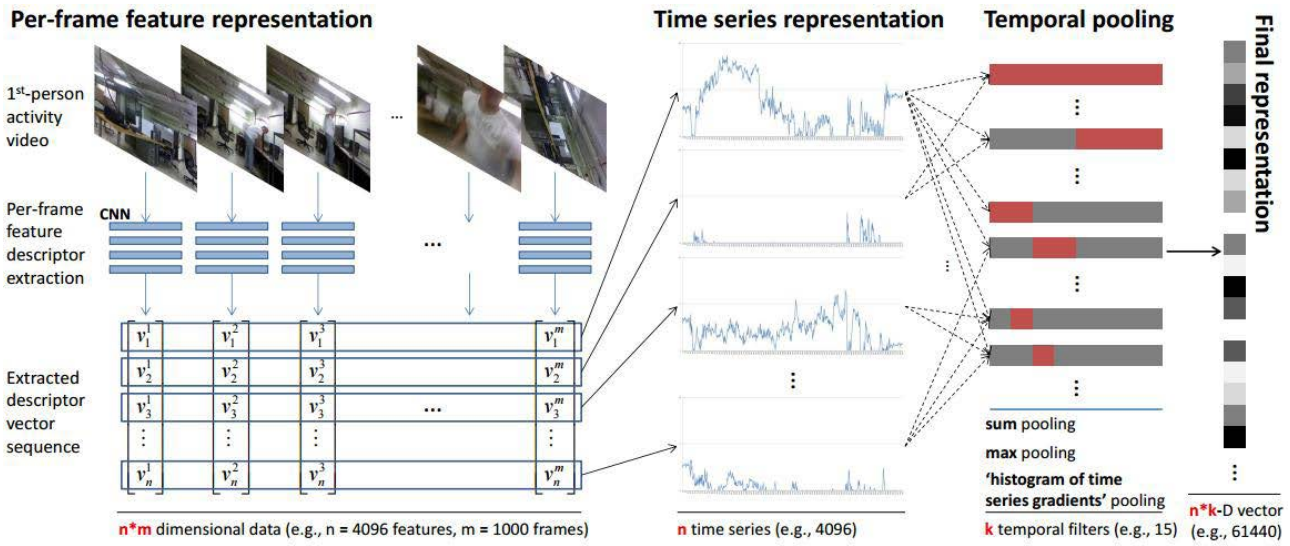
확실한 것은 deep learning은 자동 feature 학습에 대한 framework을 제공하여 주며, 이렇게 학습된 feature들은 성능적으로 종래의 feature에 비해서 우수하다. Feature로써의 deep learning은 대단히 유용하다. 이러한 deep learning의 '사용'을 보는 것은 Computer Vision분야에 있어서 어렵지 않다. Computer Vision 분야 전문 학회인 CVPR이나 ICCV의 학회지를 확인하여 보면 최근 매년 수백편의 논문들이 제목에 'deep' 또는 'convolutional network'등의 키워드를 포함한채 출판되는 것을 볼 수 있다. 이는 거의 대부분 deep learning의 사용에 관한 논문이다. CNN을 변경하지 않고 그대로 feature를 추출하는 것에 사용하는 경우도 있고, 사용되는 상황에 맞춰서 CNN 구조가 변형되거나 입력이 변형되는 케이스도 있다.

Deep learning을 위한 CNN은 convolutional layer들과 그 다음에 적용되는 classification을 위한 소수의 layer로 구성되었는데, deep learning을 '쓰는' 보편적인 방법은 이러한 CNN에서 중간 결과인 convolutional layer 결과값을 가져옴으로써 이루어진다. 실제로 새로운 데이터로 학습시킨 CNN을 사용하는 경우도 있으며, pre-trained CNN (이미 다른 dataset을 사용하여 학습된 CNN)을 사용하는 경우도 있다. 다음의 논문 [6]은 그러한 사용의 예이다: object classification을 위해서 ImageNet 데이터로 학습시킨 CNN을 이용하여, ImageNet이 아닌 다른 dataset에서 object detection을 수행하였다 (classification과 detection은 다르다). ImageNet dataset을 활용한 pre-trained CNN은 다양한 task에서 성공적인 결과를 얻었다.

> **Deep learning은 자동 feature 학습에 대한 framework을 제공하여 주며, 이렇게 학습된 feature들은 성능적으로 종래의 feature에 비해서 우수하다. Feature로써의 deep learning은 대단히 유용하다.**

비디오 인식을 위한 Deep Learning

현재까지 (몇몇 예외를 제외하고는) 대다수의 CNN의 사용은 이미지에 집중되어 왔다. 그러나 1장의 이미지 데이터만을 분석하는 것은 그 응용이 제한적이며, 보다 더 일반적인 상황에의 응용을 위해서는 비디오 분석에 대한 연구가 필수적이다. 이러한 비디오 분석의 예로는 행동인식 (activity recognition)이나 비디오 매칭 (비디오 searching이나 retrieval) 등이 있으며, 이미지에서 성공을 거둔 CNN을 비디오에 적용하여 행동인식 등에서 높은 성능을 거두는 것이 연구의 목표이다. 비디오는 이미지의 연속적인 sequence이고 이미지보다 더 고차원적인 (high-dimensional) 데이터이다. 당연히 이미지를 입력으로 하는 CNN 구조는 비디오에 변경없이 바로 적용이 불가능하다.



〈그림 2〉 CNN을 이용한 비디오 인식의 예. 비디오의 매 frame별로 CNN을 적용하여 추출되는 feature의 변화값을 추적하는 구조이다.[10] ImageNet pre-trained CNN이 적용되었다. IEEE©

최근 1~2년간 이미지 CNN을 기반으로 이를 비디오에 확장/적용하기 위하여 다음과 같은 노력들이 있었다. 이는 크게 세가지 종류로 구분가능하다.

첫번째는 비디오의 매 frame마다 이미지 CNN feature를 추출하여 (1) 그 값을 평균내어 feature로 사용하거나 또는 (2) 매 frame feature를 인식에 사용하고 그 결과값을 평균내는 방식이다. 다음의 논문이 (2)에 속한다.[7] 여기서는 이미지 데이터 뿐만 아니라 optical flow 데이터 또한 비디오에서 추출되어 사용되었다 (참고로 [7]의 1저자는 Google Deepmind로 소속을 옮겼다). 그러나 이는 '평균'이라는 과정을 통하여 frame간의 연속 정보가 희석되기 때문에 (즉, 시간적인 정보가 소실) 비디오 정보를 전부 활용하고 있다고 볼 수 없기에 한계가 있다.

두번째는 비디오를 3차원 XYT 데이터로 해석하여 3-D convolution을 통해서 CNN을 적용하는 방법이다.[8] 기본적으로 입력을 2차원 XY 이미지 데이터가 아닌 3차원 XYT데이터 (이미지가 시간축을 따라서 쌓인 것)으로 해석 하기 때문에 이를 처리하기 위하여 더 많은 parameter를 가지는 model이 필요하며, 이에 따라서 학습을 위하여 더 다수의 데이터가 필요하다. 이러한 방법은 가장 자연스러운 이미지 CNN의 비디오판 확장이라고 볼 수 있다.

Deep learning은 Computer Vision 연구자에게 있어서 대단히 유용한 도구가 될 수 있다. 다만 deep learning의 간편함은 연구자가 이러한 도구를 가져다 쓰는 것에 만족하게 만들 가능성이 있다.

세번째는 CNN을 매 frame마다 적용한 후 그 결과값의 연속을 pooling operator의 사용을 통하여 요약하는 방법이다. 이는 (1) CNN의 내부 구조 중 후반부 layer를 확장/변경하거나 또는 (2) CNN의 frame별 feature 값을 직접 pooling하는 형태로 이루어진다. Google의 최근 논문 [9]이 케이스 (1)의 단적인 예이다. CNN 내부의 temporal pooling이 강조되었다. 케이스 (2)의 예는 다음의 논문을 들수 있다.[10] 〈그림 2〉에서도 표현되었듯이, 매 frame 추출된 CNN feature를 time series 형태로 변환한 후에 time series pooling을 적용하여 비디오를 요약한다. 이러한 방향의 장점은 이미지 기반의 pre-trained CNN을 활용할 수 있다는 것에 있다. 비디오 데이터의 숫자가 부족한 상황에서 이 장점은 크게 다가온다.

마지막으로 한가지 짚고 넘어가야 할 것은 사람이 설계

한 비디오 feature (예: Improved Trajectory Features)에 비하여 비디오 CNN은 약간의 성능 향상만을 보이고 있다는 것이다 (UCF101 dataset에서의 인식 성능 87.9 대 88.6 [9]). 이미지보다 고차원인 비디오를 다루는 것은 생각처럼 간단한 일이 아니며, 이를 위한 연구는 계속 진행중이다.

마치며

Deep learning은 Computer Vision 연구자에게 있어서 (feature로써) 대단히 유용한 도구가 될 수 있다. 다만 deep learning의 높은 성능은 연구자의 연구를 이러한 도구를 가져다가 쓰는것에 그치게 만들 위험이 있다. 아울러서 만약 deep learning 자체에 대한 첨단의 연구를 하고 싶다면 Google, MS, facebook등의 회사에 구직 연락하는 것을 추천한다. 분야 선구자인 Geoffrey Hinton은 Google에 반쯤 소속되어 있으며 Yann LeCun은 facebook AI Lab의 director이다. 물론 Baidu도 나쁘지 않은 선택이다. Deep learning의 또 다른 선구자 중 하나인 스탠포드의 Andrew Ng의 경우 미국 Baidu 연구소의 director로서도 일하고 있다. 만약 Computer Vision의 미래가 deep learning이라면 이는 회사들(또는 회사와 긴밀히 협력 중인 학교들)에서 이뤄질 것이다.

References

[1] G. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets", Neural Computation, 2006.

[2] Y. Bengio, "Learning deep architectures for AI", Foundations and Trends® in Machine Learning, 2009.

[3] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks", NIPS, 2012.

[4] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun: "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks", International Conference on Learning Representations (ICLR 2014), April 2014.

[5] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional Architecture for Fast Feature Embedding", arXiv preprint arXiv:1408.5093, 2014.

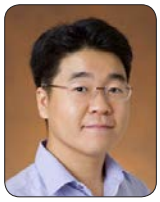
[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR, 2014.

[7] K. Simonyan, A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos", NIPS, 2014.

[8] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "C3D: Generic Features for Video Analysis", arXiv:1412.0767, 2014.

[9] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond Short Snippets: Deep Networks for Video Classification", CVPR, 2015.

[10] M. S. Ryoo, B. Rothrock, and L. Matthies, "Pooled Motion Features for First-Person Videos", CVPR, 2015.



Michael S. Ryoo

- 2004년 KAIST 학사
- 2006년 The University of Texas at Austin 석사
- 2008년 The University of Texas at Austin 박사
- 2008년~2011년 ETRI 전문연구요원
- 2011년~2015년 NASA-JPL, Research Technologist
- 2015년~ Indiana University, Assistant Professor