이미지 기반 사물 인식을 위한 개선된 딥러닝 네트워크

최승도, 장윤석, 김효진, 김태현, 허준
고려대학교
{skstmdeh1, joe9210, gywls1743}@naver.com, {samecorey, junheo}@korea.ac.kr

An Enhanced Deep Learning Network for Image-based Object Recognition

Choi Seung Do, Kim Hyo JIn, Jang Yoon Suk, Kim Tae Hyun, Heo Jun
Korea Univ.

요 약

본 논문은 딥러닝 학습에 이용하는 기본적인 데이터베이스 중 하나인 Mixed National Institute of Standards and Technology database(MNIST)를 이용하여 완전연결망(fully connected network) 하나와 서로 다른 두 구조의 또한 그 결과를 두 가지 다른 특성의 테스트 집단(test set)에 대해 확인해 보려 한다. 두 테스트 집단 중 하나는 아무 가공도 하지 않은 MNIST 데이터베이스 그대로의 것이고 다른 Convolution Neural Network(CNN) 구조를 학습시켜 보려고 한다. 하나는 원래의 테스트 이미지 내의 물체의 위치를 변화시킨 것이다. 이 두 가지 데이터 집합의 비교를 통해 어떠한 구성의 딥러닝 네트워크가 이미지 안에서 물체의 위치 변화에 대해 더 둔감하게 되는지 제시하려 한다.

Ⅰ. 서 론

딥러닝 이론에서 Convolution Neural Network(CNN)[1]은 동물의 시각 인지과정에서 영감을 받아 고안된 것으로, 필터를 통해 전체 이미지를 부분적으로 인식하여 학습하는 네트워크 구조이다. CNN의 큰 장점중 하나는 학습 정확도가 입력 데이터의 위치 변동에 둔감하다는 것이라고 알려져 있다. 여기서 둔감하다는 것은 사진의 어느 부분에 물체가 있든 인지할 가능성이 높다는 뜻이다. 사람은 당연하게도 사진의 어느 부분에 물체가 있든 인식할 수 있다. 따라서 이미지 상에서의 물체의 위치가 딥러닝 네트워크로 물체를 인식하는 데에 영향을 주어서는 안 된다. 하지만 CNN은 그 내부 구조에 따라서 학습 성능이 매우 달라지기 때문에, 구조에 따라서 입력 데이터의 위치 변동을 인지하지 못하게 되는 경우도 있다.

본 논문에서는 몇 가지 간단한 딥러닝 네트워크[3]를 이용하여, 딥러닝에서 보편적으로 쓰이는 데이터 셋인 Mixed National Institute of Standards and Technology database(MNIST)[1]를 이용한 필기체 인식 학습을 진행할 것이다. 구체적으로 각 네트워크를 필기체 숫자가 중앙에 위치하는 이미지들로 학습시킨 뒤 이를 위치를 이동하기 전의 이미지와 이동 후의 이미지들로 테스트하여, 이미지 위치 변동에 둔감한 딥러닝 네트워크 설계 방향을 제시하려 한다.

Ⅱ. 본론

각 네트워크가 어떤 특성을 가질 것인지 예측해 보기 위해 먼저 각 네트워크의 구조도를 간단히 살펴보려 한다. 사용한 완전연결망(fully connected network)는 3개의 히든 레이어, 4개의 학습 가능한 가중치(weight) 행렬을 가지는 네트워크를 이용하여 학습하였다. 아래의 그림은 앞으로 비교할 두 종류의 서로 다른 CNN 네트워크의 구조이다. 두 네트워크는 그 기본적인 구조가 같다. 먼저 그림 1에서 보이고 있는 공통된 구조 부분을 보면, 3개의 convolution layer가 있으며, 이것들에는 각각 학습할 수 있는 필터가 존재한다. convolution layer마다 하나씩 pooling layer가 존재하는데 이 레이어에는 학습에 관여하는 가중치는 존재하지 않는다. 이 과정을 거친 후 얻어진 데이터는 128*4*4의 크기를 가지는 행렬인데, 이를 1차원으로 재배열한 후 완전연결망을 사용하여 최종 결과를 도출한다. 그림 2에서 조금 변화를 준 네트워크의 일부를 살펴 볼 수 있는데, 재배열하기 전에 average pooling을 하는 레이어를 하나 추가하였다는 점이다. 재배열하기 전의 행렬은 학습된 필터들이 적용된 특성지도(feature map)의 형태이다. 다시 말해 필터를 통해 인식할 수 있는 각각의 특성들에 대한 위치 정보 값을 가진다는 뜻이다. 이러한 행렬을 average pooling 하여 각 특성지도에 있는 16개의 정보를 하나로 합쳤으며, 이를 통해 극단적으로 위치정보에 대한 의존도가 없는 결과가 나오기를 기대하였다.

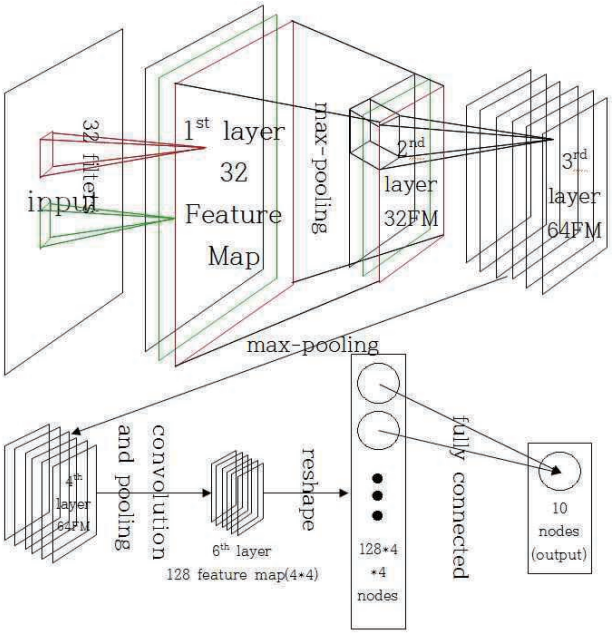

그림 1 두 CNN에서 기본적으로 사용하는 구조

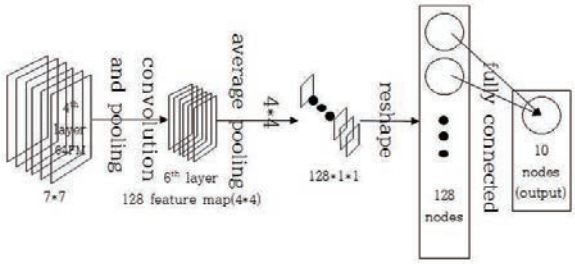

그림 2 Average-pooling을 추가하여 수정한 CNN 구조의 일부

앞에서 설명한 네트워크들을 Tensorflow라는 Python 기반의 오픈소스 라이브러리[2]를 이용하여 구현하였고, 이 네트워크들에 MNIST를 학습시켜 보았다. 학습 집단(Training set)에는 아무런 조작도 하지 않고 학습을 진행하였다. 그리고 이렇게 학습된 네트워크에 특별한 조작을 하지 않은 테스트 집단(Test set)과 이미지 내에서 숫자의 위치를 이동시킨 테스트 집단 각각에 대해 인식률을 확인하였다. 숫자의 위치는 구체적으로 왼쪽 방향과 위 방향으로 각각 4칸 씩 이동하였으며, 이미지의 크기가 28*28인 것을 고려하면 14%정도 이동한 것임을 알 수 있다. 이러한 이동은 행렬 연산을 통해 구현하였다. 아래에 첨부한 그림 3에서 데이터에 어떤 조작을 했는지 시각적으로 확인할 수 있다.

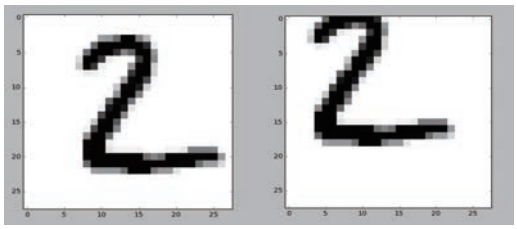

그림 3 두 가지 Test set의 이미지 사진 예시

아래에 첨부된 그래프들은 Tensorflow에서 지원하는 기능인 tensorboard를 통해 확인한 결과이며 가로축은 학습 집단의 반복학습 횟수, 세로축은 테스트 집단에 대한 정확도(인식률)를 나타낸다. 그림 4, 5, 6은 순서대로 완전연결망, average pooling을 하지 않는 CNN, average pooling을 하는 CNN의 결과를 나타낸다. 먼저 그림 4를 살펴보면 숫자를 이동시키지 않은 테스트 데이터에 대해 정확도가 약 98%정도에 수렴하는 것을 볼 수 있다. 그러나 숫자의 위치가 이동된 데이터에 대해 정확도를 확인하면 정확도가 7.5퍼센트 정도로 수렴하며, 거의 인식하지 못함을 확인할 수 있다.

그림 4 완전연결망의 반복 횟수에 대한 정확도 그래프

그림 5에서는 원래의 테스트 데이터 셋에 대해서 정확도가 99%이상으로 수렴하며, 매우 좋은 성능을 나타냄을 알 수 있다. 그러나 흔히 설명하는 CNN의 특징과는 반대로 숫자의 위치가 변경된 이미지에 대해서는 정확도가 상당히 떨어짐을 확인할 수 있다. 완전연결망의 정확도보다는 개선되었지만, 아직도 이미지의 위치에 대해 상당히 과적합(overfitting) 되어 있다고 말할 수 있다.

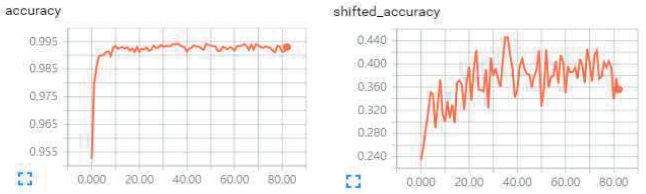

그림 5 Average pooling을 사용하지 않은 CNN의 반복 횟수에 대한 정확도 그래프

Average pooling을 추가하면 그림 5의 문제가 상당히 해결되는 것을 그림 6에서 확인할 수 있다. 먼저 기본적인 테스트 집단에 대한 정확도는 추가하기 전과 같은 값으로 수렴한다. 다만 수렴속도에 조금 차이가 있다. 전자는 학습 집단을 한번 학습하고 나면 정확도가 95% 가까이 되는 반면, 후자는 약 30%에 머문다. 정확도가 99%를 넘어가는 시점도 후자가 약간 늦다. 같은 수의 반복학습을 하는 데에 걸리는 시간 또한 계산량이 늘어났기 때문에 느리다. 그러나 이 네트워크는 기대했던 것처럼 숫자의 위치가 이동된 데이터에 대해 상대적으로 정확도가 높다. 그래프를 보면 알 수 있듯이 세 개의 네트워크 중 가장 높은 최대 90%의 정확도를 보인다. 최고점에서 점점 정확도가 감소하는 것은 중앙에 위치한 숫자 이미지에 대해 네트워크가 과적합 되어 가는 것이라고 할 수 있겠다.

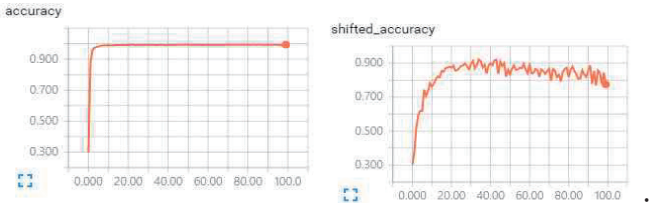

그림 6 Average pooling을 사용한 CNN의 반복 횟수에 대한 정확도 그래프

Ⅲ. 결론

본 논문에서는 몇 가지 딥러닝 네트워크의 성능을 두 종류의 다른 특징을 가지는 테스트 집단에 대해 비교해 보고 이미지 내의 물체의 위치와 정확도 관계를 네트워크 구조와 연관 지어 살펴보았다. average pooling을 이용한 네트워크 구조가 수렴 속도는 약간 느리지만 위치가 이동된 이미지에 대한 높은 학습 정확도를 얻었다. 물론 물체의 위치가 이동된 이미지에 대한 인식률을 높이는 방법으로 이 방법이 최선인 것은 아니며, 위치에 대한 정보를 완전히 없앰으로서 생기는 문제도 있을 것임을 생각해 볼 수 있다. 본 논문의 결과는 극단적으로 위치정보를 제거했을 때의 장점을 보여주는 것이며, 위치 정보도 무의미 하지 않을 것이라 생각한다면 average pooling을 한 데이터와 하지 않은 데이터를 병렬적으로 처리한 네트워크를 사용해야 할 것이라고 예상한다.

ACKNOWLEDGMENT

"본 연구는 미래창조과학부 및 정보통신기술진흥센터의 대학ICT연구센터육성 지원사업의 연구결과로 수행되었음" (IITP-2016-R0992-16-1017)

참 고 문 헌

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition." Proceedings of the IEEE, 86(11):2278-2324, November 1998.

[2] M. Abadi, et al, "TensorFlow: Large-scale machine learning on heterogeneous systems", 2015. (www.tensorflow.org)

[3] 딥러닝 네트워크 예제 : https://github.com/nlintz/