*Journal of the Korean Data & Information Science Society*
*2017,* **28**(6), 1217–1227

http://dx.doi.org/10.7465/jkdi.2017.28.6.1217
한국데이터정보과학회지

# 딥러닝 모형의 복잡도에 관한 연구[†]

김동하[1] · 백규승[2] · 김용대[3]

[123]서울대학교 통계학과

접수 2017년 10월 31일, 수정 2017년 11월 22일, 게재확정 2017년 11월 23일

**요 약**

딥러닝은 영상 인식, 음성 인식 등 기존의 머신 러닝 기법들로 해결이 어려웠던 분야에서 매우 우수한 성능을 보였고, 그로 인해 딥러닝의 폭발적인 연구의 증가가 있었다. 좋은 성능을 보이는 모형 및 모수 추정 방법에 대한 연구들이 주를 이루고 있는 현 흐름 속에서 딥러닝의 이론적인 연구 또한 조심스럽게 진행되고 있다. 본 논문에서는 딥러닝의 성공을 딥러닝 함수가 복잡한 함수를 효율적으로 잘 표현할 수 있음에서 해답을 찾고, 이에 관련된 이론적인 연구들을 조사하여 분석하고자 한다.

주요용어: 딥러닝, 복잡도, 선형 영역, 심층 신경망 함수, 함수 궤적, 함수 전이.

## 1. 서론

딥러닝은 동물과 인간의 뉴런 구조를 모사한 심층 인공 신경망 모형 (deep neural network model; Larochelle 등, 2007)을 응용하여 만든 모형 및 이를 학습하기 위한 알고리즘을 총칭한다. 딥러닝은 특히 예전까지 좋은 성능을 내지 못했던 분야들 (사진 인식, 음성 인식, 비디오 인식 및 자연어 처리)에서 압도적인 성능을 보이고 있다 (Hochreiter와 Schmidhuber, 1997; Krizhevsky 등, 2012; Sutskever 등, 2014; Chung 등, 2008; Ioffe와 Szegedy, 2015; He 등, 2016). 다양하고 복잡한 모형들(Krizhevsky 등, 2012)과 다양한 학습 알고리즘의 개발 (Kingma와 Ba, 2014), 데이터 규모의 증가, 그리고 다중 GPU (graphics processing units)의 사용 (Krizhevsky 등, 2012) 등으로 인해 딥러닝은 급격한 성장을 이루게 되었고, 그 결과 딥러닝을 이용한 인공지능 프로그램인 AlphaGo (Silver 등, 2016; Silver 등, 2017)는 지금까지 인공지능에게 금단의 구역이라 여겨지던 바둑에서 인간 프로기사를 압도적으로 이기는 상황에 오게 되었다. 또한 의료 보건 서비스, 금융 시장 등 사회 전반적인 분야에 딥러닝 모형을 적용하여 좋은 성과를 거두고 있다 (Lee 등, 2015; Lee와 Chun, 2016; Lee, 2017; Miotto 등, 2017).

대부분의 딥러닝에 관한 연구들은 뛰어난 성능을 보이도록 하는 모형 또는 기법, 그리고 좋은 추정량을 제공해주는 학습 방법들에 대한 제안이 주를 이루고 있다. 딥러닝 연구의 시발점은 2006년 G. E. Hinton이 제한된 볼츠만 기계 (restricted Boltzmann machine; Smolensky, 1986)를 이용하여 심층 신경망 모수의 초기값을 설정해주는 pre-training 방법이며, 이를 통해 과거에 비해 좋은 추정량을 얻을 수 있게 되었다 (Hinton 등, 2006; Hinton과 Salakhutdinov, 2006). 딥러닝의 한 요소인 활성 함수에 대한 연구도 활발히 진행되었는데 2010년에 제안된 ReLU (rectified linear unit; Nair와 Hinton, 2010)가

[†] 이 논문은 삼성미래기술육성재단의 지원을 받아 수행된 연구임. (과제번호 SSTF-BA1601-02).
[1] (08826) 서울시 관악구 관악로 1, 서울대학교 통계학과, 박사과정.
[2] (08826) 서울시 관악구 관악로 1, 서울대학교 통계학과, 박사과정.
[3] 교신저자: (08826) 서울시 관악구 관악로 1, 서울대학교 통계학과, 교수.
E-mail: ydkim0903@gmail.com

대표적이다. 이후에는 ReLU를 응용하여 LeakyReLU (Maas 등, 2013), PReLU (He 등, 2015), ELU (Clevert 등, 2015) 등 많은 활성 함수가 제안되었다.

딥러닝은 많은 모수를 사용하기 때문에 과적합될 가능성이 높다. 따라서 과적합을 방지하기 위한 다양한 정규화 (regularization) 방법 또한 연구되었는데 2012년에 발표된 은닉 노드들의 상관관계를 줄여주는 drop-out (Hinton 등, 2012)과 각 은닉 노드들의 분포를 일관성있게 바꾸어 좋은 추정값을 얻을 수 있도록 해주는 batch-normalization algorithm (Ioffe와 Szegedy, 2015)이 대표적이다. 또한 모수를 효율적으로 빠르게 추정하는 그래디언트 기반의 알고리즘 또한 다양하게 개발되었는데, 대표적인 알고리즘으로는 RMSProp (Tieleman과 Hinton, 2012), Adadelta (Zeiler, 2012), Adam (Kingma와 Ba, 2014) 등이 있다. 위에서 언급한 여러 연구들을 이용하여 GoogLeNet (Szegedy 등, 2015), ResNet (He 등, 2016), WaveNet (Oord 등, 2016) 등 다양한 모형들이 개발되었고, 사진 인식, 음성 인식 등 인공지능 분야에서 월등한 성능을 보이고 있다.

이러한 딥러닝 모형의 성공과 더불어 딥러닝 모형이 어떻게 잘 작동하는지, 다른 방법론들에 비해 왜 뛰어난 성능을 가지는지에 대한 이론적 연구 또한 조심스럽게 진행되고 있다. 최초로 인공 신경망 모형의 이론적인 특성을 밝힌 논문은 Hornik 등 (1989)과 Cybenko (1989)의 연구로, 단층 인공 신경망 모형이 임의의 연속 함수를 원하는 정밀도로 근사할 수 있음을 수학적으로 증명하였다. 최근에 발표된 대부분의 연구들은 한 개의 은닉층을 갖는 단층 인공 신경망 모형 (shallow neural network)과 여러개의 은닉층을 갖는 심층 인공 신경망 모형 (deep neural network)를 비교하는 것을 목표로 하였다. 한 예로 Eldan과 Shamir (2016)는 단층 인공 신경망으로 심층 인공 신경망을 근사하기 어렵다는 것을 예제를 통해 보였다.

또한, 다양한 측도를 이용하여 심층 인공 신경망 모형이 단층 인공 신경망 모형에 비해 높은 복잡도 (complexity)를 갖는다는 사실을 이론적으로 밝히기도 하였는데 Pascanu 등 (2013)과 Montufar (2014)는 모형의 복잡도를 나타내는 측도로써 함수가 가지는 선형 영역의 수를 이용하였다. 이 복잡도를 이용하여 같은 개수의 노드를 갖고 있더라도 은닉층의 개수가 커질수록 함수의 복잡도가 증가한다는 사실을 수학적으로 증명하였다. 또한 Raghu 등 (2016)은 선형 영역의 수 뿐만 아니라 더 나아가서 입력 변수가 움직이면서 만들어내는 함수의 궤적의 길이를 복잡도의 측도로 제안하였다. 이 복잡도를 이용해서 다양한 모수에서 함수의 복잡도가 은닉층의 개수에 따라 지수적으로 증가한다는 사실을 보였다. 본 논문에서는 선형 영역의 수와 함수 궤적의 길이를 이용하여 단층 인공 신경망 모형과 심층 인공 신경망 모형의 복잡도를 비교하는 이론 연구들에 대해 자세히 살펴보고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 인공 신경망 모형에 대해 간단한 언급을 한다. 3장에서는 인공 신경망 모형의 복잡도를 나타내는 두가지의 측도에 대해 설명하고, 이 측도를 사용했을 때 알 수 있는 이론적인 결과에 대해 설명한다. 4장에서는 3장에서 소개한 다양한 복잡도를 실제로 측정하여 이론적인 결과와 비교해보고, 마지막으로 5장에서 결론을 맺는다.

## 2. 신경망 모형

인공 신경망 모형 (artificial neural network)은 기계 학습에서 연구하는 분야 중 하나로, 동물의 뇌 구조를 모방하여 만든 수학적 모형이다. 사용하고자 하는 목적에 따라 컨볼루션 신경망 모형 (convolutional neural network; LeCun 등, 1998), 순환 신경망 모형 (recurrent neural network; Mikolov 등, 2010) 등이 있지만 본 논문에서는 가장 간단한 형태인 완전 연결된 인공 신경망 모형 (fully connected neural network; Larochelle 등, 2007)만을 고려하도록 한다.

$m$차원의 벡터 $x$를 입력 변수로 하는 임의의 인공 신경망 모형을 $A$라 하고, $A$ 모형의 구조를 갖는 인공 신경망 예측 함수를 $F_A(x; W, b)$라 하자. 이 때 $(W, b)$는 함수에 필요한 모수를 의미한다. 인공

신경망 모형 중에서 $k$차원의 은닉층 $n$개를 갖는 완전 연결된 인공 신경망 모형을 $A_{n,k}$라 하자. 이 때 $n = 1$인 인공 신경망 모형을 단층 인공 신경망 모형이라 하고, $n > 1$인 인공 신경망 모형을 심층 인공 신경망 모형이라 한다. $A_{n,k}$ 모형의 구조를 갖는 인공 신경망 예측 함수는 $F_{A_{n,k}}(x; W, b)$라 표현할 수 있으며, 이를 수식으로 나타내면 다음과 같다:

$$\begin{aligned} h^{(0)} &= x \\ z^{(l+1)} &= W^{(l)} h^{(l)} + b^{(l)}, \quad l = 0, ..., n-1 \\ h^{(l+1)} &= \sigma\left(z^{(l+1)}\right), \quad l = 0, ..., n-1 \\ F_{A_{n,k}}(x; W, b) &= W^{(n)} h^{(n)} + b^{(n)}. \end{aligned}$$

이 때 $\sigma(\cdot)$는 활성 함수 (activation function)라 불리우며, 선형 변환을 통해 계산된 값에 비선형성을 갖도록 하는 역할을 한다. 많이 사용하는 함수로는 시그모이드 함수, 탄젠트의 역함수, ReLU 함수, hard tanh 함수 등이 있고, 앞으로 전개될 내용에 대해서는 특별한 언급이 없으면 ReLU 함수 ($\mathrm{ReLU}(x) = \max(0, x)$)를 활성 함수로 사용하는 것으로 가정한다. 또한 위 식에서 나타나는 $h^{(l)}$의 원소 하나하나를 노드라고 부르도록 한다. 앞으로 전개될 이론에 대해서 예측 모형의 출력 변수의 차원의 수는 중요하지 않으므로 편의상 1차원이라 정한다 (i.e. $F_A(\cdot; W, b) : \mathbb{R}^m \to \mathbb{R}$).

## 3. 신경망 모형의 복잡도

### 3.1. 선형 영역의 수를 이용한 복잡도

인공 신경망 모형 $A$ 구조를 갖는 예측 함수 $F_A(x; W, b)$의 형태를 생각해보자. 많은 모수들로 이루어져있는 복잡한 함수이지만 결국 $F_A(x; W, b)$는 입력 변수 $x$에 대한 부분 선형 (piecewise linear) 함수라는 것을 파악할 수 있다 (아래의 Figure 3.1에서 인공 신경망 예측 함수가 어떤 형태를 띠고 있는지 확인할 수 있다.). 즉, 주어진 모형 $A$와 모수 $W, b$에 대해서 예측 함수 $F_A(x; W, b)$를 다음과 같이 쓸 수 있다:

$$F_A(x; W, b) = \sum_{p=1}^{\mathcal{R}(A;W,b)} (\alpha_p + x^T \beta_p) \cdot I(x \in R_p).$$

위의 식에서 $R_p$는 같은 선형식 $\alpha_p + x^T \beta_p$을 같는 정의역 내의 가장 큰 연결 집합으로, 선형 영역 (linear region)이라고 불린다. 그리고 $\mathcal{R}(A; W, b)$는 예측 함수 $F_A(x; W, b)$가 가지는 선형 영역의 개수를 의미한다. 이 식을 통해 인공 신경망 함수는 연속적으로 이어져 있는 선형 함수들로 구성되어 있음을 알 수 있으며, 따라서 선형 영역의 개수가 많으면 많을수록 주어진 인공 신경망 함수가 복잡한 형태를 띠고 있다고 생각할 수 있다. 또한 선형 함수의 개수는 선형 영역의 개수와 일치한다는 사실을 알 수 있다. Pascanu 등 (2013)는 주어진 인공 신경망 모형 $A_{n,k}$가 가질 수 있는 선형 영역의 개수를 처음으로 연구하였으며 신경망 구조를 이루고 있는 노드의 총 숫자가 같더라도 은닉층의 개수가 많아질수록 심층 인공 신경망 모형이 단층 인공 신경망 모형보다 훨씬 더 많은 수의 선형 영역을 가질 수 있음을 수학적으로 증명하였다. 구조 $A_{n,k}$를 갖는 인공 신경망 함수가 가질 수 있는 선형 영역의 개수의 최대값을 $\mathcal{R}(A_{n,k})$라 정의하자 (i.e. $\mathcal{R}(A_{n,k}) = \max_{W,b} \mathcal{R}(A_{n,k}; W, b)$). 이 때 단층 인공 신경망 모형과 심층 인공 신경망 모형에 대해서 아래의 정리들이 성립한다.

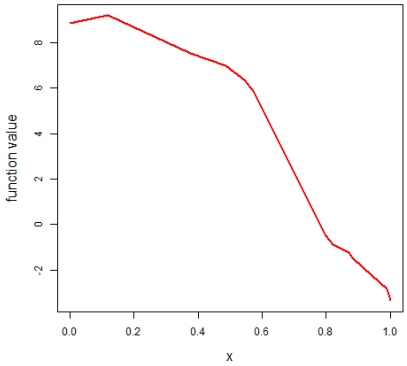



**Figure 3.1** Example of an artificial neural network prediction function

**정리 3.1** (Proposition 1 of Pascanu 등 (2013)) $n \times k$개의 노드를 갖는 단층 인공 신경망 모형 $A_{1,nk}$를 생각하자. 이 때 모형 $A_{1,nk}$가 가질 수 있는 선형 영역의 개수에 대해서 다음의 식이 성립한다:

$$\mathcal{R}(A_{1,nk}) = \sum_{j=1}^{m} \binom{nk}{j} = O\left(n^m k^m\right).$$

(여기서 $f(x) = O(g(x))$는 모든 $x \geq x_0$에 대해서 $f(x) \leq C \cdot |g(x)|$가 성립하는 실수 $x_0$와 양수 $C$가 존재한다는 의미이다.)

**정리 3.2** (Theorem 1 of Pascanu 등 (2013)) $n$개의 은닉층을 가지며, 각 층마다 $k$개의 노드로 이루어져 있는 심층 인공 신경망 모형 $A_{n,k}$을 생각하자. 이 때 모형 $A_{n,k}$가 가질 수 있는 선형 영역의 개수에 대해서 다음과 같은 식이 성립한다 ($n \geq m$이라 가정한다.):

$$\begin{aligned} \mathcal{R}(A_{n,k}) &\geq \left(\left\lfloor \frac{k}{m} \right\rfloor^{n-1}\right) \sum_{j=0}^{m} \binom{k}{j} \\ &= \Omega\left(\left\lfloor \frac{k}{m} \right\rfloor^{n-1} k^m\right). \end{aligned}$$

(여기서 $f(x) = \Omega(g(x))$는 모든 $x \geq x_0$에 대해서 $f(x) \geq c \cdot |g(x)|$가 성립하는 실수 $x_0$와 양수 $c$가 존재한다는 의미이다.)

Montufar 등 (2014)는 위에서 언급한 Pascanu 등 (2013)의 정리 3.2를 발전시켜서 심층 신경망 모형에서의 더욱 큰 $\mathcal{R}(A_{n,k})$를 만들 수 있음을 증명하였다.

**정리 3.3** (Theorem 4 of Montufar 등 (2014)) Theorem 2와 같은 가정을 만족할 때 다음과 같은 식이 성립한다:

$$\begin{aligned} \mathcal{R}(A_{n,k}) &\geq \left(\left\lfloor \frac{k}{m} \right\rfloor^{m(n-1)}\right) \sum_{j=0}^{m} \binom{k}{j} \\ &= \Omega\left(\left\lfloor \frac{k}{m} \right\rfloor^{m(n-1)} k^m\right). \end{aligned}$$

입력 변수의 차원 $m$과 각 은닉층마다의 노드의 개수 $k$가 고정되어 있다고 가정하자. 정리 3.1에 의해서 단층 인공 신경망 모형이 가질 수 있는 선형 영역의 최대 개수는 은닉층의 개수 $n$에 대해서 다항식의 증가세를 가지고 있음을 알 수 있다. 이와 대비하여 정리 3.2와 정리 3.3에 의해서 알 수 있듯이 심층 인공 신경망 모형이 가질 수 있는 선형 영역의 최대 개수는 은닉층의 개수 $n$에 대한 지수함수의 증가세를 가진다. 특히 $k \geq 2m$을 만족한다면, 은닉층의 개수 $n$이 커질수록 단층 인공 신경망에 비해서 심층 인공 신경망이 더욱 많은 선형 영역을 가질 수 있을 것이다. 이 사실은 주어진 인공 신경망 모형이 가질 수 있는 선형 영역의 수의 최대값을 모형의 복잡도의 측도로 사용할 때, 충분히 큰 $n$에 대해서 심층 인공 신경망 모형이 단층 인공 신경망 모형보다 높은 복잡도를 가짐을 의미한다.

위에서 언급한 3개의 정리들로부터 주어진 인공 신경망 모형 $A$에 대해서 $\mathcal{R}(A)$를 인공 신경망 모형 $A$에 대한 복잡도를 나타내는 측도로 사용할 때, 구조 전체의 노드의 수가 동일하더라도 은닉층의 개수가 커질수록 심층 인공 신경망 모형이 단층 인공 신경망 모형에 비해 더 높은 복잡도를 가짐을 알 수 있었다. 하지만 $\mathcal{R}(A)$는 정의 그대로 선형 영역의 개수의 최대값이기 때문에 훈련 자료를 통해 학습한 모수 $\hat{W}, \hat{b}$로부터 얻을 수 있는 예측 함수 $F_A(x; \hat{W}, \hat{b})$의 선형 영역의 수 $\mathcal{R}(A; \hat{W}, \hat{b})$와는 큰 차이가 있을 수 있다 (i.e. $\mathcal{R}(A) >> \mathcal{R}(A; \hat{W}, \hat{b})$). 따라서 $\mathcal{R}(A)$를 이용한 복잡도는 주어진 인공 신경망 모형 $A$의 일반적인 함수들을 대표하는 복잡도라고 보기는 어렵다는 한계가 있다.

### 3.2. 함수의 궤적의 길이를 이용한 복잡도

함수의 복잡도를 측정하는 또 다른 방법으로 정의역 내에서 곡선 (trajectory)을 따라 움직일 때 함수가 어떻게 변화하는지를 살펴볼 수 있다. $x$가 동일한 곡선을 따라 움직일 때 지나가는 선형 영역의 수가 많을수록 더 복잡한 함수라고 말할 수 있다. 이는 3.1절에서 살펴본 선형 영역의 개수를 세는 방법과 유사한 방식으로 복잡도를 측정하는 것이라고 볼 수 있으나, 탐색 영역을 정의역 전체 ($\mathbb{R}^m$)에서 일차원으로 매개화된 곡선으로 축소해서 측정하는 방식으로 받아들일 수 있다. Raghu 등 (2016)은 은닉층의 개수가 많을 수록 한 곡선이 지나는 선형 영역의 개수가 지수적으로 증가한다는 사실을 보였으며, 매 은닉층을 지날 때 마다 변화하는 곡선의 길이를 이용해서 곡선이 지나는 선형 영역의 개수를 측정할 수 있음을 보였다. 또한 이러한 신경망 모형의 복잡도가 모수 $(W, b)$의 선택과는 무관하다는 것을 밝혔다.

본 내용에 들어가기에 앞서 필요한 개념들에 대한 정의를 하겠다. 정의역 내의 곡선은 일차원 변수로 매개화된 정의역의 부분집합으로, $x(\cdot) : [0, 1] \to \mathbb{R}^m$ 으로 표기하겠다 (예: 정의역 내의 임의의 두 점 $x_0, x_1 \in \mathbb{R}^m$을 잇는 선분 $x(t) = x_0 + t(x_1 - x_0), t \in [0, 1]$). 가능한 곡선의 종류는 가장 단순한 형태인 선분부터 원호 등 다양한 형태가 있지만, 여기에서는 일반적인 곡선에 대해서 다룰 것이다. 주어진 신경망 모형 함수에 대해서, 정의역 내의 곡선 $x(\cdot)$은 매개변수에 따라 서로 다른 선형 영역에 포함될 수 있다. 매개변수가 변화함에 따라 곡선위의 점이 속하는 선형 영역이 변화하는 순간을 전이 (transition)가 일어났다고 정의한다. 신경망 모형 $A$와 모수 $(W, b)$에 대해서 곡선 $x(\cdot)$가 전이를 일으키는 횟수를 $\mathcal{T}(F_A(x(\cdot); W, b))$로 표현하겠다. 전이 횟수 $\mathcal{T}$는 곡선 $x(\cdot)$의 복잡도에 영향을 받을 수 있으나, 고정된 곡선에 대해서는 모형 $A$와 모수 $(W, b)$가 예측 함수 $F$를 복잡하게 만들수록 큰 값을 갖게 될 것이다.

전이 횟수를 이용해서 너비가 $k$인 $n$개의 은닉층을 갖고 있는 완전 연결된 인공 신경망 모형 $A_{n,k}$의 복잡도에 대해서 알아보고자 한다. 특정 모수가 아닌 다양한 모수 $(W, b)$에 대해서 모형 $A_{n,k}$의 복잡성을 알아보기 위해서, 모든 가중치 $W$는 $\mathcal{N}(0, \sigma_w^2/k)$에서, 모든 절편항 $b$는 $\mathcal{N}(0, \sigma_b^2)$에서 임의로 추출되었다고 가정하였다. 이 때 다음과 같은 정리가 성립한다.

**정리 3.4** (Theorem 4 of Raghu 등 (2016)) 활성함수로 hard tanh 함수 (hard $\tanh(x) = \min\{1, \max\{x, -1\}\}$)를 사용할 때 임의의 곡선 $x(\cdot)$에 대해서 인공 신경망 모형 $A_{n,k}$의 전이 횟수 $\mathcal{T}(F_{A_{n,k}}(x(\cdot); W, b))$는 다음과 같은 식을 만족한다.

$$\mathbb{E}_{W,b}\left[\mathcal{T}(F_{A_{n,k}}(x(\cdot); W, b))\right] = O\left(\frac{\sqrt{k}}{1+\sigma_b^2/\sigma_w^2}\right)^n$$

$\mathcal{T}$가 함수의 복잡도를 나타낸다는 사실로 미루어 볼 때, 위 정리는 신경망 모형의 복잡도가 은닉층의 너비 $k$보다는 은닉층의 개수 $n$에 많은 영향을 받는다는 것을 시사한다. 은닉층의 너비 $k$에 대해서는 다항함수의 증가세를 보여주지만 은닉층의 개수 $n$에 대해서는 $\mathcal{T}$가 지수적으로 증가하는 것을 알 수 있다.

흥미롭게도, 정리 3.4의 결과는 곡선의 길이와도 밀접한 연관을 갖는 것을 확인할 수 있었다. 임의의 곡선의 길이 $l(x(\cdot))$는 다음과 같이 계산할 수 있다.

$$l(x(\cdot)) = \int_0^1 \left|\left|\frac{dx(t)}{dt}\right|\right| dt \tag{3.1}$$

곡선 $x(\cdot)$과 신경망 예측 함수 $F_A(\cdot; W, b)$의 합성함수 $F_A(x(\cdot); W, b) : [0,1] \to \mathbb{R}$은 일차원 변수로 매개화된 치역의 부분집합으로 바라볼 수 있으므로 치역 내의 곡선이라고 할 수 있다. 이를 이미지 곡선이라고 하자. 마찬가지로 신경망 모형 $A_{n,k}$에 대해서 $d$ $(d \le n)$번째 은닉층의 사전활성값 $z^{(d)}(x(\cdot))$ 또한 $\mathbb{R}^k$상의 곡선으로 생각할 수 있으며, 식 3.1를 이용해서 길이를 측정할 수 있다.

**정리 3.5** (Theorem 3 of Raghu 등 (2016)) 신경망 모형 $A_{n,k}$에 대해 예측 함수 $F_{A_{n,k}}(\cdot; W, b)$에 대해 모든 가중치 $W$는 $\mathcal{N}(0, \sigma_w^2/k)$에서, 모든 절편항 $b$는 $\mathcal{N}(0, \sigma_b^2)$에서 임의로 추출되었다고 가정하자. 임의의 곡선 $x(\cdot)$에 대해 신경망 모형 $A_{n,k}$의 $d(d \le n)$번째 은닉층의 사전활성값 $z^{(d)}(x(\cdot))$의 길이는 다음 부등식을 만족한다.

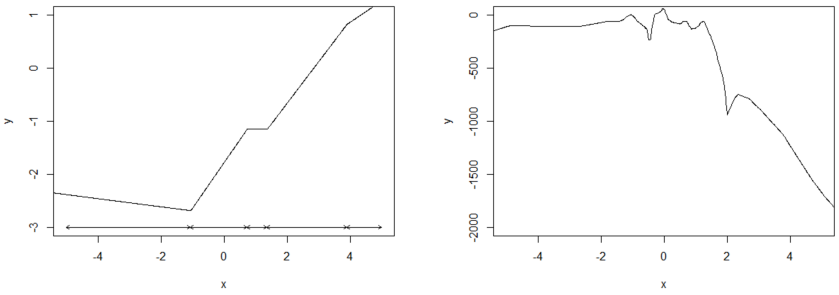
$$\mathbb{E}_{(W,b)}\left[l(z^{(d)}(\cdot))\right] \ge \begin{cases} \Omega\left(\dfrac{\sigma_w\sqrt{k}}{\sqrt{k+1}}\right)^d l(x(\cdot)) & \text{(ReLU)} \\ \Omega\left(\dfrac{\sigma_w\sqrt{k}}{\sqrt{\sigma_w^2+\sigma_b^2+k\sqrt{\sigma_w^2+\sigma_b^2}}}\right)^d l(x(\cdot)) & \text{(hard tanh)} \end{cases} \tag{3.2}$$

$d = n$인 경우에는 $F_{A_{n,k}}(x(\cdot); W, b) = z^{(d)}(x(\cdot))$가 성립하게 된다. $d = n$인 경우에 집중해서 정리 3.5를 살펴본다면 이 정리는 정의역에서의 곡선 $x(\cdot)$이 예측 함수 $F_{A_{n,k}}(\cdot; W, b)$를 거치면서 길이가 증가하는 정도를 식으로 표현했다고 할 수 있다. 이러한 관점에서 $\mathbb{E}_{(W,b)}\left[l(z^{(d)}(x(\cdot)))\right]/l(x(\cdot))$을 이미지 곡선의 길이의 증가율이라고 할 수 있다. 정리 3.5의 결과는 정리 3.4의 결과와 매우 밀접하다. 이미지 곡선의 길이의 증가율이 은닉층의 개수에 지수적으로 비례한다는 점은 전이 횟수가 은닉층의 개수에 지수적으로 비례한다는 정리 3.4의 결과와 유사하다. 무엇보다 중요한 것은 식 3.2에서 활성화 함수로 hard tanh함수를 사용했을 때의 결과이다. 만일 가중치의 분산 $\sigma_w$가 절편항의 분산 $\sigma_b$보다 매우 크다면 정리 3.5에서 밝힌 이미지 곡선의 길이의 증가율의 하계와 전이 횟수의 차수 (order)가 일치하게 된다. 따라서 이미지 곡선의 길이의 증가율을 인공 신경망 모형의 복잡성을 설명하는 측도로써 사용할 수 있는 것이다.

정리 3.4와 정리 3.5를 통해 인공 신경망 모형의 복잡도에는 은닉층의 개수가 중요한 영향을 끼친다는 사실을 알 수 있었다. 주어진 곡선을 따라 움직일 때 인공 신경망 모형의 예측 함수가 어떻게 변하는지 살펴보는 것은 선형 영역의 개수를 세는 것과 동일한 철학을 공유하는 방법이지만 문제를 단순화함으로써 좀 더 수학적인 접근을 가능케 하였다. 정리 3.4의 결과 인공 신경망 모형의 전이 횟수를 복잡도의 측도로 이용할 경우에 복잡도는 은닉층의 개수에 대해 지수적으로 증가한다는 사실을 확인할 수 있었으며, 이 결과는 모수 $(W, b)$의 선택과는 큰 관련이 없다는 점 또한 확인할 수 있었다. 다른 복잡도와 관련된 연구들의 경우 특정한 모수의 선택 하에서의 복잡도에 대해 다루었던 점을 상기하면 (Pascanu 등, 2013; Montufar 등, 2014), 정리 3.4는 다양한 모수에서 신경망 모형의 복잡도에 대해서 탐구하였다고 할 수 있다. 또한 정리 3.5를 통해 전이 횟수는 이미지 곡선의 길이의 증가율과 밀접한 연관이 있음을 밝혀냈으며, 이를 통해 이미지 곡선의 길이의 증가율을 인공 신경망 모형의 복잡도를 나타내는 측도로 사용할 수 있는 이론적인 근거를 마련하였다.

## 4. 실험

앞에서 소개한 인공 신경망 모형의 복잡도를 실제 인공 신경망 함수에 대해서 측정해보았다. 앞에서의 모든 결과들이 은닉층의 개수가 증가할수록 완전 연결된 신경망 모형의 복잡도가 지수적으로 증가한다는 사실을 보였으므로, 은닉층의 너비는 고정시키고 은닉층의 개수만을 변화시켜가며 어떠한 결과가 나오는지 확인해 보았다. 편의를 위해 정의역과 공역은 1차원으로 고정시켰으며, 은닉층의 너비는 2로 고정하였고 은닉층의 개수는 1부터 10까지 변화시켜보았다. 활성화함수로는 ReLU 함수를 사용하였으며, 각 모수들은 표준정규분포에서 임의로 생성하였고 추가적인 학습을 진행하지는 않았다.




(a) Depth=2 (b) Depth=10

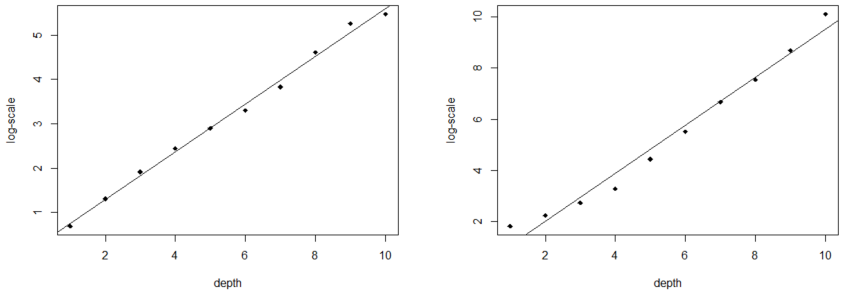
**Figure 4.1** Two graphs of deep neural network

Figure 4.1은 각각 은닉층의 개수가 2, 10인 경우의 신경망 모형 함수의 그래프이다. 활성화 함수로 부분 선형함수인 ReLU함수를 사용하였기 때문에 최종 신경망 모형 함수 또한 부분 선형 함수의 형태를 띄는 것을 확인할 수 있었다. 또한 두 그래프를 비교하면 은닉층의 개수가 많을수록 좀 더 복잡한 형태의 함수가 만들어지는 것을 확인할 수 있다.

3.1절에서 소개한 선형 영역은 이 경우에는 같은 선형 함수를 공유하고 있는 구간을 의미한다. 은닉층의 개수가 2개인 경우 선형 영역을 Figure 4.1 (a)의 하단에 화살표모양으로 나타내었다. 정의역이 일차원인 경우 선형 영역의 개수는 함수가 꺾이는 지점의 개수에 1을 더한 값과 동일한 것을 알 수 있다.

따라서 선형 영역의 개수가 많다는 것은 함수가 자주 꺾인다는 사실과 일치한다고 볼 수 있으며, 꺾이는 횟수가 많을수록 복잡한 함수라고 생각한다면 선형 영역의 개수를 함수의 복잡도를 나타내는 측도로 사용하는 것이 타당하다고 볼 수 있다.

이를 실험적으로 입증하기 위해, 은닉층의 개수를 1부터 10까지 변화시킬 때 선형 영역의 개수가 어떻게 변화하는지를 실험을 통해 알아보았다. 모수의 선택에 의한 효과를 무시하기 위해 각 은닉층의 개수마다 함수를 100개씩 생성하였으며, 각각의 함수의 선형 영역의 개수를 계산하였다. 각 은닉층의 개수 별로 계산된 선형 영역의 개수의 평균에 로그를 씌운 값과 은닉층의 개수 간의 그래프를 Figure 4.2 (a)에 나타내었다.




(a) The number of linear regions (b) The length of the image trajectory

**Figure 4.2** Complexity measures of deep neural networks

실험 결과 정리 3.2와 정리 3.3의 결과를 뒷받침하는 결과를 얻을 수 있었다. 은닉층의 개수가 증가할수록 선형 영역의 개수 또한 증가하는 것을 확인해 볼 수 있었으며, 이러한 증가가 지수적으로 이루어진다는 사실 또한 확인할 수 있었다.

Raghu 등 (2016)가 제시한 측도인 함수의 궤적의 길이 또한 측정해 보았다. 정의역 상에서의 곡선 [-10,10]이 신경망 모형을 거친 후 길이변화가 어떻게 되는지를 실험해 보았다. 앞의 실험과 마찬가지로 은닉층의 개수는 1부터 10까지 변화시켜보았으며, 각 은닉층의 개수마다 함수를 100개씩 생성한 후 곡선 [-10,10]이 신경망 모형을 거친 후 변화한 자취의 길이의 평균을 측정하였다. Figure 4.2 (b)를 통해 자취의 길이의 평균에 로그를 씌운 값과 은닉층간의 개수 간의 관계를 확인할 수 있다. 실험 결과, 자취의 길이 또한 선형 영역의 개수처럼 은닉층의 개수가 증가함에 따라 지수적으로 증가한다는 사실을 확인할 수 있었다. 이는 정리 3.5의 결과를 실험적으로 입증한다고 할 수 있으며, ReLU를 선형함수로 사용하였을 때 선형 영역의 개수와 자취의 길이 간에 밀접한 연관이 있다는 점 또한 시사하는 결과이다.

## 5. 결론

본 논문에서는 딥러닝의 성공의 이유를 딥러닝 함수의 높은 표현력에서 찾고, 함수의 복잡도를 나타낼 수 있는 여러 측도들을 이용하여 은닉층의 수가 증가할 수록 복잡도가 지수적으로 커짐을 연구한 다양한 논문들을 정리하였다. 구체적으로 인공 신경망이 나타낼 수 있는 선형 영역의 최대값을 복잡도로 정의한 논문들 (Pascanu 등, 2013; Montufar 등, 2014)에 대해서 조사하였고, 이 복잡도의 단점을 개선하여

전이 횟수 및 함수 궤적의 길이 등 특정 모수에 구애받지 않는 새로운 복잡도를 제안한 논문 (Raghu 등, 2016)에 대해서도 살펴보았다. 또한 복잡도에 관한 이론적인 결과를 모의 실험을 통해 확인하였다.

향후 연구로는 이 논문에서 조사한 연구들을 포함하여 지금까지 제안된 여러 복잡도가 갖는 통계적인 성질을 규명하고, 이를 바탕으로 새로운 복잡도 측도를 개발하여 이 측도에 대한 다양한 이론적 성질들을 밝혀낼 예정이다.

## References


Chung, J., Gulcehre, C., Cho, K. and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555.*

Clevert, D., Unterthiner, T. and Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289.*

Cybenko G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)*, **2**, 303-314.

Eldan, R. and Shamir, O. (2016). The power of depth for feedforward neural networks. *Conference on Learning Theory*, 907-940.

He, K., Zhang, X., Ren, S. and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE International Conference on Computer Vision*, 1026-1034.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770-778.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, **313**, 504-507.

Hinton, G. E., Osindero, S. and Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, **18**, 1527-1554.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580.*

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, **9**, 1735-1780.

Hornik, K., Stinchcombe, Ma. and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, **2**, 359-366.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167.*

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 1097-1105.

Larochelle, H., Erhan, D., Courville, A., Bergstra, J. and Bengio, Y. (2007). An empirical evaluation of deep architectures on problems with many factors of variation. *Proceedings of the 24th International Conference on Machine Learning*, 473-480.

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, **86**, 2278-2324.

Lee, K. J., Lee, H. J. and Oh, K. J. (2015). Using fuzzy-neural network to predict hedge fund survival. *Journal of the Korean Data & Information Science Society*, **26**, 1189-1198.

Lee, W. (2017). A deep learning analysis of the KOSPI's directions. *Journal of the Korean Data & Information Science Society*, **28**, 287-295.

Lee, W. and Chun, H. (2016). A deep learning analysis of the Chinese Yuan's volatility in the onshore and offshore markets. *Journal of the Korean Data & Information Science Society*, **27**, 327-335.

Maas, A. L., Hannun, A. Y. and Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. *Proceedings of the 30th International Conference on Machine Learning*, **30**.

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J. and Khudanpur, S. (2010). Recurrent neural network based language model. *Interspeech*, **2**.

Miotto, R., Wang, F., Wang, S., Jiang, X. and Dudley, J. T. (2017). Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics.*

Montufar, G. F., Pascanu, R., Cho, K. and Bengio, Y. (2014). On the number of linear regions of deep neural networks. *Advances in Neural Information Processing Systems*, 2924-2932.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. *Proceedings of the 27th International Conference on Machine Learning*, 807-814.

Oord, A., and Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Pascanu, R., Montufar, G. and Bengio, Y. (2013). On the number of response regions of deep feed forward networks with piece-wise linear activations. *arXiv preprint arXiv:1312.6098*.

Raghu, M., Poole, B., Kleinberg, J., Ganguli, S. and Sohl-Dickstein, J. (2016). On the expressive power of deep neural networks. *arXiv preprint arXiv:1606.05336*.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. and others. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, **529**, 484-489.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A. and others. (2017). Mastering the game of go without human knowledge. *Nature*, **550**, 354-359.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. *Colorado University at Boulder Department of Computer Science*.

Sutskever, I., Vinyals, O and Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 3104-3112.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1-9.

Tieleman, T. and Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *Coursera: Neural Networks for Machine Learning*, **4**.

Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

*Journal of the Korean Data & Information Science Society*
*2017,* **28**(6), 1217–1227

http://dx.doi.org/10.7465/jkdi.2017.28.6.1217
한국데이터정보과학회지

# A study on complexity of deep learning model[†]

Dongha Kim[1] · Gyuseung Baek[2] · Yongdai Kim[3]

[123]Department of Statistics, Seoul National University

Received 31 October 2017, revised 22 November 2017, accepted 23 November 2017

**Abstract**

Deep learning has been studied explosively and has achieved excellent performance in areas like image and speech recognition, the application areas in which computations have been challenges with ordinary machine learning techniques. The theoretical study of deep learning has also been researched toward improving the performance. In this paper, we try to find a key of the success of the deep learning in rich and efficient expressiveness of the deep learning function, and analyze the theoretical studies related to it.

*Keywords*: Complexity, deep learning, deep neural network, linear regions, trajectory of a function, transition of a function.

[†] This work was supported by Samsung Science and Technology Foundation under Project Number SSTF-BA1601-02.

[1] Ph.D.Candidate, Department of Statistics, Seoul National University, 56-1 Mountain, Sillim-dong, Gwanak-gu, Seoul metropolis, 151-742, Korea.

[2] Ph.D.Candidate, Department of Statistics, Seoul National University, 56-1 Mountain, Sillim-dong, Gwanak-gu, Seoul metropolis, 151-742, Korea.

[3] Corresponding author: Professor, Department of Statistics, Seoul National University, 56-1 Mountain, Sillim-dong, Gwanak-gu, Seoul metropolis, 151-742, Korea. E-mail: ydkim0903@gmail.com