합성곱 신경망에서의 신뢰도 보정

*심재훈[1] 김세윤[2] 조남익[1]

[1]서울대학교 전기정보공학부 뉴미디어통신공동연구소

[2]VUNO Inc.

*joinjhs@ispl.snu.ac.kr light4u@gmail.com nicho@snu.ac.kr

Confidence Calibration in Convolutional Neural Network

*Shim, Jae Hoon[1] Kim, Seyun[2] Cho, Nam Ik[1]

[1]Department of ECE, INMC, Seoul National University

요약

본 논문에서는 합성곱 신경망을 이용한 이미지 분류에서 신뢰도와 실제 예측 정확도가 다른 문제점을 해결하기 위하여 변형된 두 가지 목적 함수를 제안하였다. 첫 번째는 기존 교차 엔트로피 함수에 새로이 신뢰도와 정확도의 차이를 더해준 것이고, 두 번째는 예측값의 최댓값을 0.5로 제한한 것이다. 새로운 목적 함수를 통해 학습해본 결과 정확도의 차이는 거의 나지 않았고, 신뢰도와 실제 정확도는 매우 근접하게 되는 결과를 얻을 수 있었다.

1. 서론

최근 딥 러닝을 활용한 연구들이 활발하게 이루어지고 있고, 그 적용 분야는 보행자 검출(pedestrian detection), 분류 문제(classification problem), 초해상도 영상(image super-resolution), 영상 분할(image segmentation), 압축 등 컴퓨터를 활용할 수 있는 거의 모든 분야로 확장되었다. 분류 문제의 경우, 가장 흔히 쓰이는 합성곱 신경망 구조는 모델의 출력에 softmax라는 함수를 사용하여 그 결과를 신뢰도(confidence)로 사용한다. 신뢰도는 결과를 얼마나 신용할 수 있는지를 알려주는 지표로 사용할 수 있다.

예를 들어, 강아지와 고양이의 사진을 분류하는 신경망을 설계했다고 가정하자. 어떤 사진을 입력으로 넣었을 때, 이 신경망은 이 사진이 신뢰도 0.9(90%)로 이 사진을 강아지로 예측한다. 그렇다면 우리는 이 사진이 정말 90% 확률로 강아지라고 말할 수 있을까? 이에 대한 대답은 결과부터 말하자면 그렇지 않다고 할 수 있다. softmax란 함수는 단순히 신경망의 출력에 지수함수를 취해 그 비율을 내는 함수에 지나지 않기 때문에 실제 정확도와 정확히 같기는 힘들다.

따라서 본 논문에서는 신뢰도의 값을 유의미하게 사용하기 위하여 이를 예측의 실제 정확도에 근접하게 만들고자 했다. 목적 함수에 신뢰도와 정확도의 차이를 넣음으로써 신경망이 신뢰도를 실제 정확도에 근접하게 예측할 수 있도록 하였다. 또한 분류 문제에 목적 함수로 자주 사용하는 교차 엔트로피 함수는 높은 신뢰도를 선호하기 때문에 과하게 높은 신뢰도 값이 나오는 것을 방지하기 위하여 예측값을 제한해서 교차 엔트로피 함수에 넣는 방법을 사용하였고, 그 결과를 Reliability Diagram이라는 도표를 사용하여 나타내었다.

2. Reliability Diagram과 ECE, MCE

2.1 Reliability Diagram

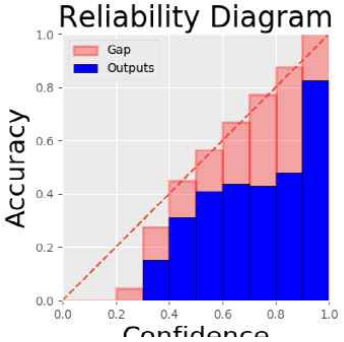

그림 1 CIFAR-10 분류 신경망의 Reliability Diagram

Reliability Diagram[1]이란, 모델의 출력의 신뢰도를 M개의 구간으로 나누고, 각 구간에서의 실제 예측 정확도를 막대로 표현한 도표이다. 그림 1은 한 예시인데, 푸른 막대는 실제 예측 정확도를 나타내고 붉은 막대는 실제 예측 정확도와 신뢰도 사이의 차이를 나타낸다.

2.2 ECE와 MCE

ECE(Expected Calibration Error)[1]란 신뢰도 보정의 결과를 수치적으로 나타내기 위한 하나의 통계값으로, 각 신뢰도 구간에서 실제 예측 정확도와 신뢰도 사이의 차이의 가중 평균으로 정의된다.

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)| \quad (1)$$

여기서 B_m은 각 신뢰도 구간에서의 출력 집합이고, n은 총 샘플 개수이다. $acc(B_m)$은 해당 신뢰도 구간에서의 정확도이고 $conf(B_m)$는 신뢰도를 나타낸다.

신뢰도 보정의 결과를 MCE(Maximum Calibration Error)[1]로 표현할 수도 있다. 이 통계값은 전체적인 신뢰도 보정의 결과가 아닌 가장 오차가 큰 신뢰도 구간에서의 신뢰도와 정확도의 차이로 정의된다.

$$MCE = \max_{m \in 1,\ldots,M} |acc(B_m) - conf(B_m)| \quad (2)$$

그림 1을 예시로 들어 설명하면, 붉은 막대 중 가장 긴 막대의 길이가 MCE에 해당한다. 그림 1에서의 MCE는 [0.8, 0.9] 구간에서 나타나고, 0.5372의 값을 갖는다.

3. 방법 제안

3.1 목적 함수

일반적으로 분류 문제에 사용하는 목적 함수는 교차 엔트로피인데, 아래 식 (3)과 같이 나타난다.

$$L_{CE} = -\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{C} t_{ij}\log(s_{ij}) \quad (3)$$

여기서 N은 샘플의 개수이고, C는 분류하고자 하는 클래스의 개수이다. t_{ij}는 실제값이며 해당 클래스의 이미지일 경우 1의 값을, 나머지 경우에는 0의 값을 갖는다. s_{ij}는 예측값으로 [0,1] 구간의 값을 가진다.

본 연구에서는 기존의 교차 엔트로피에 예측 정확도와 신뢰도 사이의 차이에 절댓값을 취한 값을 추가로 사용하였다. 해당 목적 함수는 아래와 같다.

$$L_{conf} = |\frac{1}{N}\sum_{j=1}^{N}\max_{i}(s_{ij}) - acc_N| \quad (4)$$

여기서 $\max_i(s_{ij})$는 C개의 클래스의 예측값 중 가장 큰 값을 가지는 값으로, 신뢰도를 의미한다. acc_N은 샘플 N개에 대한 예측을 실행했을 때의 정확도이다. 맞힌 샘플의 수가 k개라면, $acc_N = \frac{k}{N}$의 값을 가진다. 따라서 식 (4)는 샘플의 신뢰도 평균과 정확도의 차이를 의미하는 항이 된다.

3.2 예측값 제한

목적 함수로 사용하는 교차 엔트로피는 신뢰도가 1에 가까운 값일수록 낮은 값을 갖게 된다. 따라서 이 목적 함수를 사용하면 실제 정확도와는 별개로 신경망은 신뢰도에 높은 값을 부여하게 되고, 확실하지 않은 결과에도 강한 확신을 주장하게 된다. 이런 현상의 완화를 위해 교차 엔트로피 함수에 들어가는 예측값을 제한하는 방법을 제안한다.

구체적으로, 어떤 클래스에 대한 입력의 예측값이 0.5 이상이면 신경망은 입력을 해당 클래스로 예측하게 되기 때문에 해당 예측값을 0.5로 대체하여 넣어도 예측 결과에는 아무런 지장이 없게 된다. 따라서 교차 엔트로피에 예측값인 s_{ij} 대신 $\min(s_{ij}, 0.5)$를 대입한다면 신경망의 클래스 예측 성능에는 영향을 주지 않으면서 신경망이 과한 신뢰도를 부여하지 않도록 할 수 있다. 따라서 수정된 교차 엔트로피 목적 함수는 아래와 같으며,

$$L_{CE,clip} = -\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{C} t_{ij}\log(\min(s_{ij}, 0.5)) \quad (5)$$

학습에 사용된 총 목적함수는 앞의 목적함수와 더하여 다음과 같이 표현할 수 있다.

$$L = L_{CE,clip} + \lambda L_{conf} = -\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{C} t_{ij}\log(\min(s_{ij}, 0.5)) + |\frac{1}{N}\sum_{j=1}^{N}\max_{i}(s_{ij}) - acc_N| \quad (6)$$

4. 실험 결과

4.1 실험 세부 사항

실험에는 CIFAR-10 데이터셋을 사용하였고, 신경망은 ResNet[2] 구조를 사용하였다. 입력이 들어오면 3×3 conv 레이어를 거쳐 32개의 피쳐맵을 사용하는 9개의 residual block을 지난다. 각 residual block은 3×3 conv 레이어 2개와 skip connection으로 이루어져 있다. 이후에는 64개의 피쳐맵을 사용하는 18개의 residual block을 지난 후 average pool을 거쳐 1000차원의 fully connected 레이어를 통해 10개의 클래스를 가지는 레이블을 출력한다. 해당 구조는 그림 2에 나타내었다. 목적 함수의 λ는 10을 사용하였다.

4.2 실험 결과

실험 결과는 reliability diagram을 통해 나타내었고, 그림 3에 세 가지 reliability diagram을 도시하였다. 가장 왼쪽의 첫 번째 도표는 목적 함수로 교차 엔트로피만 사용한 경우의 reliability diagram이고,

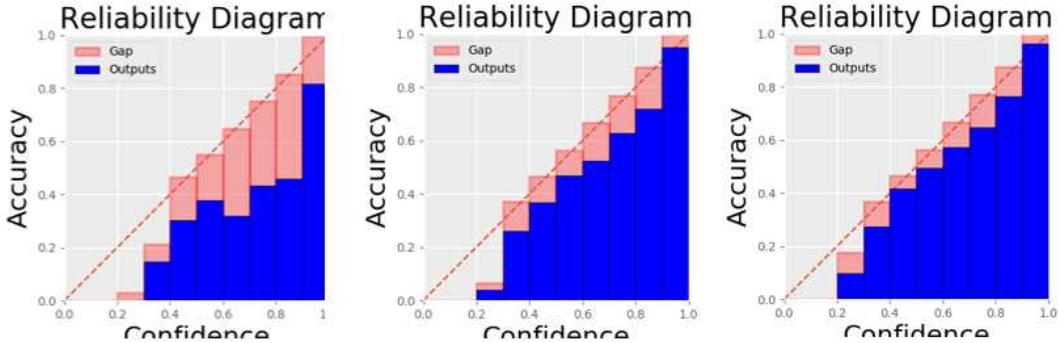

그림 2 실험 결과 reliability diagram. 각각 목적 함수로 교차 엔트로피만 사용한 경우(왼쪽), 예측 정확도와 신뢰도 사이의 차를 목적 함수에 더한 경우(가운데), 예측값을 제한하는 식을 추가했을 때(오른쪽)를 나타낸다.

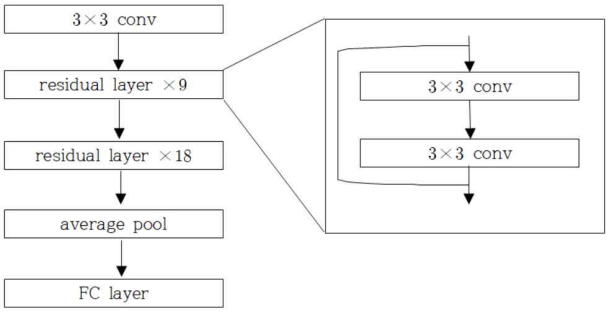

그림 3 사용한 합성곱 신경망 구조

가운데의 도표는 예측 정확도와 신뢰도 사이의 차를 목적 함수에 더한 경우의 결과, 오른쪽의 도표는 여기에 예측값을 제한했을 때의 결과이다. 붉은색 막대의 길이는 신뢰도와 정확도의 오차를 나타내는데, 오른쪽으로 갈수록 붉은색 막대들이 짧아지는 것을 볼 수 있다.

목적 함수	L_{CE}	$L_{CE}+\lambda L_{conf}$	$L_{CE,dip}+\lambda L_{conf}$
정확도	81%	81%	81%
ECE	0.1620	0.0936	0.0762
MCE	0.5728	0.3417	0.3202

표 1 각 목적 함수에 따른 실험 결과

표 1은 세 경우에서 ECE, MCE, 분류 정확도를 나타낸 결과이다. 표에서 볼 수 있듯 제안된 방법들을 적용했을 때 결과가 유의미하게 개선된 것을 볼 수 있고, 동시에 예측 정확도에는 영향을 거의 주지 않는 것을 볼 수 있다.

5. 결론

본 논문에서는 합성곱 신경망에서의 신뢰도를 실제 정확도와 근접시키기 위해 목적 함수에 정확도와 신뢰도의 차이를 추가하고 예측값을 제한하는 방법을 통해 성능을 더욱 개선시켰다. 해당 실험 결과를 reliability diagram과 ECE, MCE를 통해 분석하여 신뢰도와 실제 정확도가 유의미한 수준으로 근접함을 확인할 수 있었다. 해당 목적 함수는 다른 분류 신경망에도 그대로 사용할 수 있기 때문에 활용 다양성이 높을 것으로 예상된다.

감사의 글

이 논문은 삼성전자 및 2020년도 BK21 플러스 창의정보기술 인재양성 사업단에 의하여 지원되었음.

참고문헌

[1] Guo, Pleiss, Sun, Weinburger, On Calibration of Modern Neural Networks, International Conference on Machine Learning, 2017.
[2] He, Zhang, Ren and Sun, Deep Residual Learning for Image Recognition, arXiv preprint arXiv:1512.03385, 2015.
[3] Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. The elements of statistical learning, volume 1. Springer series in statistics Springer, Berlin, 2001
[4] Krizhevsky, Alex and Hinton, Geoffrey. Learning multiple layers of features from tiny images, 2009.
[5] Niculescu-Mizil, Alexandru and Caruana, Rich. Predicting good probabilities with supervised learning. International Conference on Machine Learning, pp. 625-632, 2005.