

# Introduction to Machine Learning (Spring 2020)

## Final Project (150 Pts, Due Date: June 21)

### Student ID 2015313754

Name 길태형

#### 1. noisy data를 구분하기 위한 방법

저는 CNN모델을 개선하거나 모델을 pretraining 하는 대신, noisy data를 구분하는 것에 초점을 뒀습니다. 그 이유는, 뛰어난 모델이라고 하더라도, Training data 제대로 Labeling 되어 있지 않다면, 학습에 저하가 될 수 있다고 생각했기 때문입니다. Noisy data를 구분하기 방법을 알기 위해, 논문 'Identifying Misabeled Data using the Area Under the Margin Ranking (Geoff Pleiss)' 를 참고하였습니다. 비교적 간단한 방법으로 noisy data를 구분하는 방법을 제시했기에, 본 논문을 참고하였습니다. 본 논문 내용에서는, noisy data를 구분하기 위해, softmax layer의 전 layer의 activation value인 logit value를 활용합니다.



그림 1. data들의 logit value(논문 발췌)

그림 1의 그래프는 각 이미지의 logit value 중, 두개의 class를 비교한 그림입니다. 초록색으로 표시된 선은 이미지가 라벨링 된 Class의 logit value 값이고, 빨간색으로 표시된 선은 라벨링 되지 않은 클래스 중, 가장 큰 logit value의 값입니다. 그림 1의 가장 왼쪽 이미지는 labeling이 정상적으로 되어 있으며, CNN 신경망이 학습하기에도 적합한 이미지의 logit value 그래프입니다. 가운데 이미지는 labeling은 정상적으로 되어 있으나, CNN 신경망이 학습하기엔 적합하지 않은 logit value 그래프입니다. 오른쪽 이미지는 labeling이 정상적으로 되어 있지 않은 이미지의 logit value 그래프입니다. 오른쪽 이미지의 logit value만이 빨간색 선이 초록색 선보다 위에 있는 것을 알 수 있습니다. 이는, 신경망은 이미지를 'Dog'가 아닌 'Bird'로 구분해서, 초록색의 'Dog'의 logit value보다, 빨간색의 'Bird'의 logit value가 더 큰 것일 겁니다.

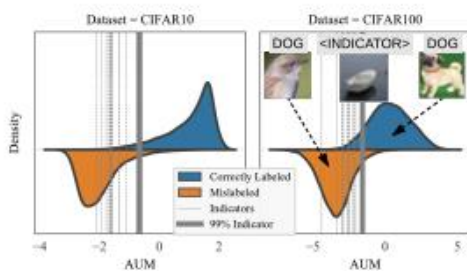


그림 2. Logit value 차이의 Histogram(논문 발췌)

각 이미지의 Logit value들의 차이를 기준으로 이미지들을 ascending sort한다면, 잘못 labeling이 되었을 확률이 큰 이미지들이 앞쪽에 정렬되고, 정상적으로 labeling이 되었을 확률이 큰 이미지들을 뒤쪽에 정렬될 것입니다. (그림2 참고)

논문에서는 새로운 가짜 Class를 생성하고, 기존의 이미지 데이터에 라벨링하여, noisy data를 나누는 기준이 되도록 구현하였지만, 제 프로젝트에서는 신경망 학습 시간이 부족하였기에, 기존의 Data들만을 이용하여, 어느정도 학습시킨 후에, Logit value들의 차이를 기준으로 Sorting하여 뒤쪽의 이미지 데이터들을 정상적으로 Labeling data로 간주하여, 학습을 처음부터 다시 진행하였습니다.

제가 Noisy data를 구분했던 과정은 다음과 같습니다.

1. CNN model을 제공해주신 Base CNN과 동일하게 설계하고, 전체 데이터에 대하여 40에폭 학습을 진행합니다.
2. 40 에폭 학습이 끝난 후, logit layer의 activation value를 사용하여 이미지 data들을 logit value의 차이를 기준으로 정렬합니다.
3. 차이가 큰 70%의 data들만을 남기고, 나머지 data들은 noisy data로 간주하여, 이후의 학습에는 참여하지 않습니다. 논문과 다르게 indicator sample을 구현하지 않아서, 기준을 정하는게 어려웠습니다. 학습을 여러 번 시도해보니, Data의 30%를 Noisy data로 간주했을 때, overfitting이 너무 일찍 일어나지 않을 정도로 Data의 양이 많았고, 정상적으로 Labeling된 data들의 비율이 많은 것 같았습니다.

## 2. noisy data를 구분한 이후의 학습



모든 data를 사용하여 학습한 Model을 삭제한 후, 다시 model을 초기화 하여, 정상적으로 labeling되었다고 가정하는 data만을 활용하여 학습을 진행합니다. Validation loss가 증가하기 전까지 학습을 진행하여, 80에폭 학습을 진행하였습니다.

## 3. overfitting 극복

Noisy data를 제외한 data는 많지 않아서, overfitting에 취약하다고 생각했습니다. Overfitting을 극복하기 위해 다음과 같은 시도를 하였습니다.

1. Dropout 계층 추가
2. Convolution Layer에서 L2 Regularization 적용
3. Convolution Layer 추가
4. Keras의 ImageDataGenerator를 활용하여 rotation, shift, zoom 등의 data augmentation 적용
5. 학습이 진행된 이후, Batch size와 Learning rate를 감소하여 학습이 세밀하게 진행될 수 있도록 조정하였습니다.

## 4. leader board 점수

30	xogud1231		0.50400	8	now
Your Best Entry 					
Your submission scored 0.50166, which is not an improvement of your best score. Keep trying!					