

팀 프로젝트 지도 활동 보고서

교과목명	AI및데이터분석의 기초		
강사	한상호 교수		
활동일자	2020년 09월 19일 (토요일)		
학과	성명	연락처	서명
스마트시스템SW학과	정학제	010-3895-1306	
스마트시스템SW학과	정재윤	010-9008-4847	
스마트시스템SW학과	정해문	010-7176-4289	
스마트시스템SW학과	조재홍	010-4062-9523	
활동내용	<p>(미팅시 팀원들 사진 붙이기 - 선택사항)</p> <p>- 미팅 결과물에 대한 URL: https://github.com/Hott-J/AI-BigData-Basic 팀장 github 계정 아래 project repo를 하나 만들고 (repo 이름은 project 취지에 맞게 작명), 코드 파일들, PPT 파일 등을 commit 하고 해당 파일 접근을 위한 URL을 여기에 작성하기</p> <p>- 2페이지에 작성했습니다. 데이터 수집에 관해 조사하고 의견을 나누었습니다.</p>		

< 정학제 팀장 >

-9주차

1. 데이터 불균형

<https://wikidocs.net/73827>

2. mobilenetV2?

<https://minimin2.tistory.com/43>

전이학습? : 잘 된 모델의 가중치를 가져와서 만들 모델에 적용

과제제시 및 전체적인 핸들링

< 조재홍 팀원 >

오버피팅(과적합)은 특정 데이터만 과하게 학습된 상태를 말합니다. 오버피팅이 발생하면 해당 데이터에서는 높은 정확도를 보여주지만 다른 데이터가 나오면 제대로 동작하지 못하는 것입니다.

반대로 언더피팅(과소적합)은 데이터가 너무 적어서 이전에 학습했던 데이터에서 불필요한 패턴이나 노이즈까지 학습하게 됩니다.

이러한 문제를 해결하기 위한 것이 데이터 불균형(데이터 증강), 모델의 복잡도 줄이기, 배치 정규화, 드롭아웃이 있습니다.

데이터 불균형은 기존에 있는 데이터를 약간 변형하여 사용하는 것입니다. 예를 들어 이미지를 돌리거나, 반전시키거나, 특정방향으로 밀어내는 등의 방법이 있습니다.

모델의 복잡도를 줄이는 방법은 신경망의 레이어의 수를 줄이거나 노드의 수를 줄이는 방법이 있습니다.

배치 정규화는 가중치를 0~1 사이로 제한을 하는 방법입니다. 활성화 함수의 출력 값을 정규화 하며 초기화가 자유롭고 학습 속도가 빠르며, 기울기 소실을 해결할 수 있습니다.

드롭아웃은 신경망 학습을 할 때, 모든 계층에 대해 학습을 수행하는 것이 아니라 일부를 생략하고 학습을 수행하는 방식입니다. 이를 통해 과적합된 모델이 아니라 일반화된 모델을 학습할 수 있습니다.

< 정해문 팀원 >

오버피팅이란, 샘플 데이터만 가지고 학습을 한 결과 너무 정확하게 학습이 된 상태를 말한다.

오버피팅(과적합)이 발생할 경우 기존에 가지고 있던 샘플 데이터에 대해선 100%에 가까운 정확도를 선보이지만 만일 우리가 예상치 못했던 다른 데이터가 개입할 경우 정확도가 급격하게 떨어지는 문제를 야기하게 된다. 이런 오버피팅을 막기 위한 방법들이 있다.

첫번째로 드롭아웃이 있다.

드롭아웃은 우리가 설정한 노드를 이용하여 예측할 때 노드 중 일부분을 랜덤으로 활성화하지 않는다는 것이다. 예를 들어 30%의 노드를 랜덤으로 선택하여 이번 예측에 활용하지 않고 나머지 70%의 노드만 사용하여 예측값을 내기 때문에 너무 지엽적인 특징점에 의해 결괏값이 나오는 것을 방지, 즉 오버피팅을 방지 하게 된다.

두번째로 데이터를 증강 시켜주는 것이다.

하나의 인물에 대하여 사진 하나하나 중복 없이 구하는 건 상당히 힘들고 시간도 많이 든다.

하지만 keras의 이미지 증강해 주는 기능을 활용하여 하나의 이미지를 가지고 회전, 반전, 이동 등을

통해 마치 다수의 이미지인 것처럼 만들어 주는 것이다.

세번째로 배치 정규화가 있다.

배치 정규화는 활성화 함수의 출력값(활성화값)을 정규화하는 작업을 뜻한다. 분포를 강제로 바꿔주기 때문에 학습은 더 어려워지지만, 모델이 과하게 학습 데이터에 맞춰지는 것을 방지 할 수 있다. 학습 속도도 빨라진다.

< 정재윤 팀원 >

모델을 생성할 때 오버피팅이 날 경우 보완해주며 모델의 정확도를 높여주는 방법에 대해서 배치 노멀라이제이션, 드롭아웃 등의 방법들을 조사해보았고, 이러한 방법들은 학습 중에 적용하며 사용해 보았다.

이번에는 오버피팅이 아닌, 데이터가 부족한 상황에서의 학습에는 어떠한 해결책이 있을까? 에 대해서 조사해보았다. 첫번째로는 데이터 프로세싱 및 Augmentation과정을 통한 방법이 있고, 두번째로는 전이학습을 통한 방법이 있다.

우선 Data Preprocessing과 augmentation을 하면 거의 성능이 좋아지고, 원본에 추가되는 개념이기 때문에 모델의 성능이 떨어지지 않고, 쉽고 패턴이 정해져 있다. 기본적인 예로 좌우반전, 이미지 랜덤으로 자르기, 밝기 조절 등의 응용방법이 있다. 이것 외에도 회전시키고, 크기조절하고, 짜그러트리는 등의 방식들이 있다. 이를 통해 데이터의 질과 양을 동시에 늘릴 수 있다. 데이터가 부족하다면, 이러한 데이터 프로세싱을 통해서 데이터의 질과 양을 늘릴 수 있을 것이다.

더욱 나아가, 사진데이터에서 가장 많이 사용하는 Preprocessing & Augmentation 방식은 VGGNet 가 했던 방식으로 RGB값을 각각 빼서, RGB값의 평균을 0으로 만들어 loss의 수렴을 빠르게 하고, 같은 이미지를 3가지 크기로 만든 뒤 224*224px로 랜덤 crop해서 일부만 보더라도 인식할 수 있도록 해서 성능을 향상시킬 수 있다고 한다.

두번째로는 데이터가 부족한 상황에서 사용할 수 있는 전이학습이라는 방법이 있다고 한다. 전이학습이란 학습 데이터가 부족한 분야의 모델 구축을 위해 데이터가 풍부한 분야에서 훈련된 모델을 재사용하는 머신러닝 학습 기법으로 미리 학습된 모델의 가중치를 새로운 모델의 적용하는 Pre-trained Model 기법이나, 기존 모델의 일부 레이어를 재사용하는 Layer Re-use 등의 기법들이 있다고 한다.

팀프로젝트 지도 수행과 관련하여 상기와 같이 활동하였음을 확인합니다.

2020 년 11월 26 일

담당 교수 : 한 상 호 (인)