

머신러닝 학습 부산물 추적을 위한 프레임워크

김은진^o, 이영섭*, 이성진*

^o경상대학교 기계항공정보융합공학부 항공우주및소프트웨어공학전공,

*경상대학교 기계항공정보융합공학부 항공우주및소프트웨어공학전공

e-mail: asd0755@naver.com, ys2lee.ac@gmail.com, insight@gnu.ac.kr

An Efficient Dynamic Workload Balancing Strategy

Eunjin Kim^o, Yeongseop Lee*, Seongjin Lee*

^oDepartment of Aerospace and Software Engineering ,Gyeongsang National University,

*Department of Aerospace and Software Engineering ,Gyeongsang National University

● 요약 ●

머신러닝이 보편화되면서 모델 학습을 돕기 위한 머신러닝 및 데이터과학 도구의 수요가 증가하고 있다. 머신 러닝을 사용하는 연구에서는 다양한 파라미터에 대한 실험이 진행되어 많은 학습 부산물이 생성 된다. 하지만 기존의 학습 부산물을 관리하는 프레임워크는 하나의 실험을 진행하는데 모든 경우의 수를 진행해 그 규모가 크다. 본 연구는 기존의 도구가 가지는 규모 문제를 개선하고, 주기적으로 메일을 사용자에게 전송해 실험과정을 보고하는 새로운 도구를 제안한다. 이러한 시스템은 학습과정에서 사용자가 의도한 파라미터의 학습이 진행되는지 추적가능하다.

키워드: Machine learning, Hyperparameter Optimization, Supporting Tool

I. Introduction

머신러닝이 점차 보편화 되어감에 따라 프레임워크에 독립적이며 유연하고 확장이 쉬운 머신러닝 및 데이터과학 도구에 대한 수요가 증가하고 있다. 머신러닝 실험에는 다양한 hyperparameter를 사용한 학습이 반복된다. 실험과정은 hyperparameter, 실험의 의도, 저장된 모델 가중치, 학습결과데이터 등 많은 학습부산물이 발생한다. 실험과정에서 만들어지는 방대한 학습부산물의 체계적인 관리가 필요하다. 불필요한 실험을 반복하는 문제를 해결하고 개발 과정에서 발생하는 최적의 값을 찾기 위해 많은 학습부산물을 관리할 체계적인 프레임워크가 필요하다.

기존 머신러닝 실험도구로는 TB (TensorBoard)[1], TW (TensorWatch)[2]등이 있다. 이중 TB에서 학습부산물을 관리하고 반복적인 실험을 도울 수 있는 ‘HParams Dashboard’을 도입하였다. 해당 기능은 실험에 사용할 값을 학습 전에 입력하고 가능한 경우의 수를 반복하기 때문에 최종 결과확인까지 시간이 오래 걸리고 실험 규모가 불필요하게 방대해지는 문제가 있다. 본 연구는 기존의 hyperparameter 최적화 도구의 규모 문제를 개선하고 학습에 사용할 값과 실험과정을 입력받아 사용자에게 보고하는 새로운 도구를 제안한다.

II. Preliminaries

1. Related works

기존 머신러닝 도구인 Google의 TensorBoard의 경우 TensorFlow와 같이 가장 많이 사용되는 머신러닝 디버깅, 시각화 등을 진행하는 종합 머신러닝 학습도구이다. 최근 추가된 ‘HParams Dashboard’는 실험을 진행할 hyperparameter를 각각의 리스트로 입력해

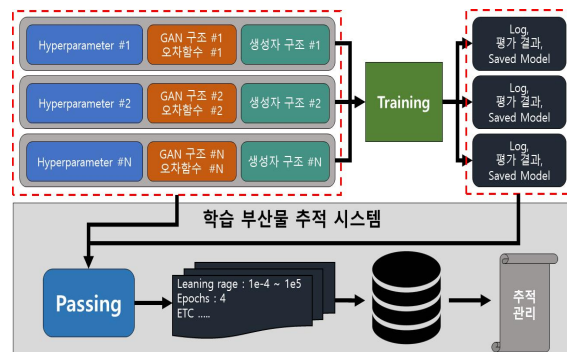


Fig. 1. 학습부산물 추적시스템 다이어그램

2019 December 27

Notification loss and accuracy of recent model.

History Table

Key	Value
GAN_loss	2
train_loss	2
val_loss	1
train_accuracy	0.78
val_accuracy	0.98

Fig. 2. Google Docs 문서

자동으로 반복 학습을 진행하여 TB 자체에서 테이블과 그래프로 시각화하는 기능을 가진다. 시각화 기능까지 포함되어 있어 간편하게 많은 실험을 하고 직관적으로 결과를 확인할 수 있는 장점이 있다. 하지만 최종 결과를 확인하기 위해서는 설정한 파라미터에 대한 모든 반복이 진행되어야 한다. 또한 작업 중간, 입력한 모든 경우의 수에 대해 기계적으로 반복을 진행한다. 따라서 사용자 의도를 반영할 수 없고 불필요하게 실험과정이 커지는 단점이 있다.

III. The Proposed Scheme

본 논문에서는 사용자가 학습과정에서 나타나는 부산물을 주기적으로 메일을 사용자에게 전송해 실험과정을 보고하는 메일 리포팅 시스템을 제안한다.

학습과정에서 발생하는 부산물은 hyperparameter, 모델구조, 오차함수 값, 모델평가결과, 모델 가중치가 있다. 이러한 데이터를 보고하는 테이블을 작성하는데는 Google Docs API를 사용했다. Google Docs API는 사용자가 편리하게 문서를 생성하고 수정할 수 있도록 해주는 플랫폼이다. 또한 문서를 작성한 뒤 원하는 내용을 추출하는 것이 용이하다.

본 논문에서 사용한 Google Docs 템플릿은 Fig.2처럼 메일이 전송된 날짜와 입력된 파라미터 테이블로 구성된다. 시스템이 파라미터를 받으면

Table. 1. 본 논문의 프레임워크와 TB 비교

특징	Proposed	TB
학습 부산물 기록	O	O
주기적인 학습 부산물 제공	O	X
학습 중 파라미터 추적 여부	O	X
특정 파라미터 제공	O	O

Google Docs 템플릿을 카피하고, 템플릿의 변수와 매핑되는 파라미터를 병합한다. 병합된 문서를 html과 pdf 형식으로 export하여 다운로드한다. html 파일을 인코딩하여 메일 본문에 파라미터 테이블을 출력하고, pdf 파일을 첨부하여 사용자에게 메일을 보낸다. 제안한 프레임워크의 특징은 Table. 1과 같다.

IV. Conclusions

기존의 머신러닝 도구는 반복적인 작업을 도와주지만 가능한 많은 경우의 수만큼 동작하므로 사용자가 불필요하다고 여기는 작업까지 수행한다. 본 논문에서 제시한 메일 리포팅 시스템은 사용자가 실험에 사용할 파라미터와 실험과정을 리포팅하여 사용자가 필요한 작업 위주로 모델이 학습되는지 확인가능하다.

ACKNOWLEDGEMENT

이 논문은 2018년도 정부(교육부)의 재원으로 한국과학창의재단(2019년도 학부생 연구프로그램)의 지원을 받아 수행된 연구이며 또한, 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2019R1G1A1100455).

REFERENCES

[1] TensorBoard, <https://www.tensorflow.org/tensorboard>