Introduction to Deep Learning

Course Instructor: Dr. M. Umair
Topic: Datasets for Deep Learning Models

Agenda

1. Introduction
2. Dataset Development Recommendations

Introduction

Introduction

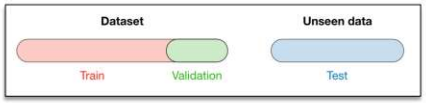
What is a Dataset

- Dataset is a *collection* of *single* or *various types of data* stored in a *digital* form.
- Datasets primarily consist of *images*, *text*, *audio*, *videos*, and *numerical data points*, etc.
- Finding a *quality dataset* is a fundamental requirement to build a real-world AI application.

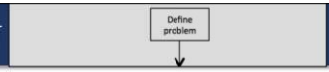
Introduction

Importance of a Datasets

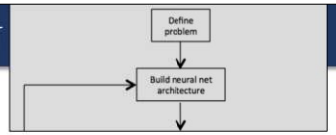
- Carefully *curated* and *annotated datasets* are the foundation of machine learning.
- Deep learning requires *large-scale datasets* to reach the desired *accuracy* and *generalization* performance.
- Once the model has been chosen, it is *trained* on the *entire dataset* and *tested* on the *unseen test set*.



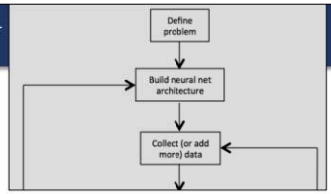
Introduction



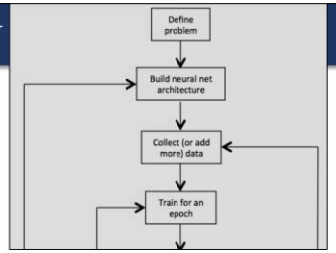
Introduction



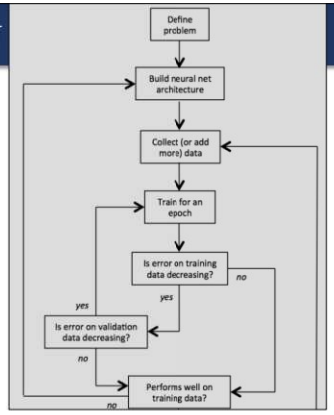
Introduction



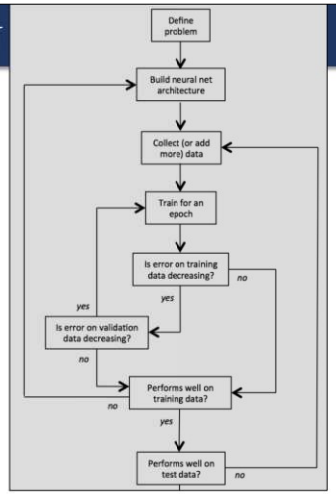
Introduction



Introduction



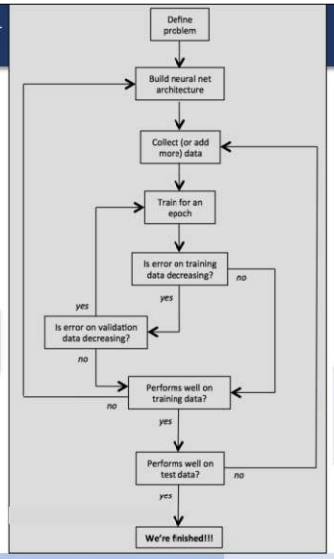
Introduction



Introduction

Hyperparameters tuning (learning rate, dropout rate)

The *validation* set will tell us how the model does on data it has yet to see.



Adjust internal parameters, for example weights and bias

If the *accuracy* on the *training set* continues to *increase* while the accuracy on the validation set stays the *same (or decreases)*, it's a good sign that it's time to stop training because we're *over-fitting*.

Introduction

Should I Develop a Dataset?

- For a *highly specific problem* statement, you have to *create a dataset* for a domain, *clean* it, *visualize* it, and understand the *relevance* to get the result.
- However, if the *problem* statement is *common*, one can use a few *dataset search engines*.

Introduction

Dataset Search Engines

- UC Irvine Machine Learning Repository
- Hugging Face
- Dataset Search – Google
- Kaggle

Dataset Development Recommendations

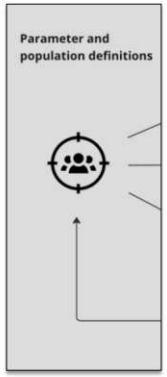
Dataset Development Recommendations

Introduction

- The *creation of datasets* is a *crucial* aspect of machine learning research.
- **FACT:** No dataset is a neutral, complete, or apolitical representation.
- Practitioners must make a *series of decisions* throughout the dataset *creation process*.

Dataset Development Recommendations

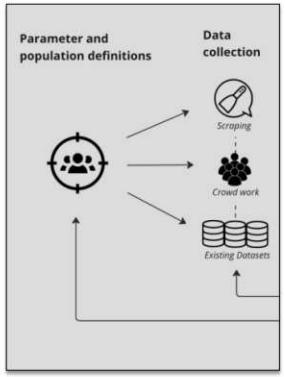
Dataset Creation Process



- This phase involves *identifying* the *specific variables of interest* within the broader group from which data will be collected.
- This phase is *crucial* for *defining the phenomena the dataset intends to capture*, what is within *scope*, and what might be considered *inappropriate*, *invalid*, or *irrelevant* data.

Dataset Development Recommendations

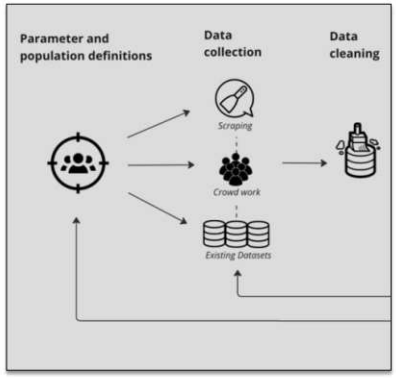
Dataset Creation Process



- *Data* is gathered from *various sources* according to the defined parameters and population.
- This can involve *different methodologies* such as *scraping* content from the web, employing various types of *sensors*, or utilizing and repurposing *existing datasets*.
- *Methods* used are often *shaped* by the *intended objectives* of the dataset, as well as *operational constraints*, such as institutional *requirements* or resource *limitations*.

Dataset Development Recommendations

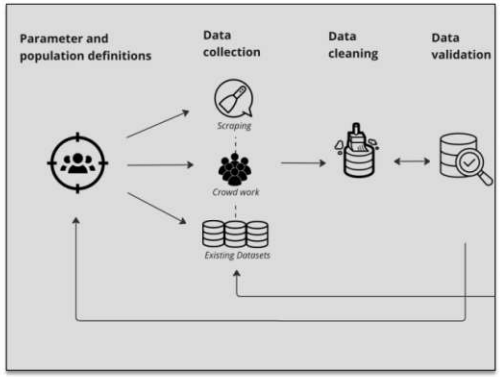
Dataset Creation Process



- Once data is collected, it often contains *errors*, *duplicates*, or *missing values* that need to be addressed.
- Data cleaning involves processing the data to correct inaccuracies, remove irrelevant information, and handle missing data.

Dataset Development Recommendations

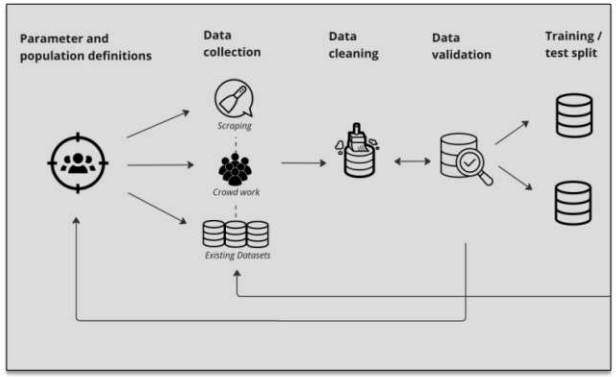
Dataset Creation Process



- This phase *checks* the dataset for *accuracy* and *consistency* with real-world phenomena.
- Data validation ensures that the dataset *meets the necessary standards and assumptions* for its intended use.

Dataset Development Recommendations

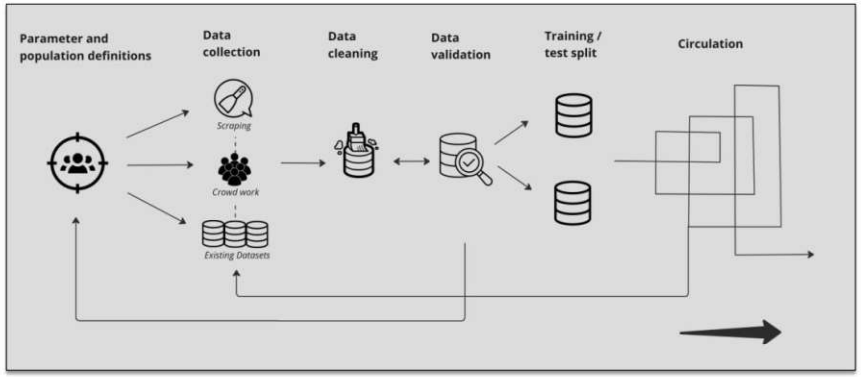
Dataset Creation Process



- The dataset is divided into *training* and *test* sets.
- The *training set* is used to build and *train the model*, while the *test set* is used to *evaluate* its performance and generalizability to unseen data.

Dataset Development Recommendations

Dataset Creation Process



References

References

1. There Is No Data Like More Data – Datasets for Deep Learning in Earth Observation, Michael Schmitt et al., 2023
2. Importance of datasets in machine learning and AI research, DataToBix
3. Fundamental of Deep Learning,
4. Building Better Datasets: Seven Recommendations for Responsible Design from Dataset Creators, Will Orr, 2024