Choose Research Area in Data Science

In the first phase we have to choose a research area from the given topic, as per the suggested topic ‘Training in noisy environments and incomplete data’ because I am new to this area I started doing research on each and every noisy environments in the real world. Before selecting the research area I should have an idea on

* What is noisy data
* How noisy data will add to the existing data
* How we have to train the noisy data.
* How to reduce the noisy data if we have in our data set.

And so on.

Noisy data is meaningless data. Noisy data can be stored, received or change the existing data that machine or program will run on it to provide unsatisfied or unexpected results if we did statistical analysis on it. Even though we have methods to reduce the noisy data but it will take huge memory for storage without any meaning.

Noisy data can be caused by hardware failures, programming errors and gibberish input from speech or optical character recognition (OCR) programs. Spelling errors, industry abbreviations and slang can also impede machine reading.

In some cases adding noise data to training dataset will improve the robustness of the network, resulting in better generalization and faster learning. For instance, we will get Training a neural network with a small dataset can cause the network to memorize all training examples, in turn leading to overfitting and poor performance on a holdout dataset. Small datasets may also represent a harder mapping problem for neural networks to learn, given the patchy or sparse sampling of points in the high-dimensional input space. One approach to making the input space smoother and easier to learn is to add noise to inputs during training.

The type of noise can be specialized to the types of data used as input to the model, for example, two-dimensional noise in the case of images and signal noise in the case of audio data. Noise is only added during the training of your model. Adding noise during training is a generic method that can be used regardless of the type of neural network that is being used. Make sure that any source of noise is not added during the evaluation of your model, or when your model is used to make predictions on new data. We cannot know how much noise will benefit your specific model on your training dataset. Experiment with different amounts, and even different types of noise, in order to discover what works best.

We have to use noise reduction techniques to minimize the noise level in the data. We have so many noise reduction techniques in present few techniques were discussed below most common technique we will use is to remove the unwanted fields from the input data, other method is to replace the unknown values with zeros or nan so that we can easily differentiate with other data for analysis. If it is readable data we can do above methods but what if it is audio, video or images it will be hard to replace or remove the noise data. For these kind of input noise data we have to use filters to remove the noise in it for example Kalman filter and Savitzky-Golay filter and so on.

**Background to select the Research Area:**

We have so many research areas to do on noise environments for example audio, video and image analysis etc., but these are common in now a days to do on these type of topics. So, I selected ***Health care Department*** to do my research. My goal is simply to inspire the world and open eyes to show how studying computer deep learning, AI, ML and then applying that knowledge to the medical field can make a big impact on the world. As per my research I see only few papers were published on health care when compare to other research areas. Most of the recent papers were published on ***EHR (Electronic health records)*** it might be diabetes, heart diseases, Allergy information, Emergency department visits, Hospital discharge summaries and reports, Drug and Pharmacy Service Information , Neurophysiology reports, Patient consultation reports etc., As per my knowledge we had few papers only worked on respiratory reports.

EHR Data is the data being collected when we see a doctor, pick up a prescription at the pharmacy, or even from a visit to the dentist. These are just a few of the examples where HER data is collected. This data is used for a variety of use-cases. From personalizing healthcare to discovering novel drugs and treatments to helping providers diagnose patients better and reduce medical errors.

EHR Data can come from many different sources in today's world. Data such as:

* Heart rate monitors
* X-Rays and other radiology scans
* Fitness trackers
* Other diagnostic tests

All this helps to provide a 360-degree view of a patient's health. Ensuring that this data is collected in an electronic health record is very valuable to the patient and their caregivers. This data becomes even more critical when you can get a longitudinal view of a patient's data and collect it in a way that allows healthcare professionals and data scientists to make meaningful and accurate predictions. To make this data even more impactful and create better insights and predictions, we need to be able to aggregate the data of many to help address some of the most significant opportunities in healthcare.

Since 2020, we see world suffering from dangerous virus called COVID-19. Once the patient effected by this virus it will act as a slow poison to him to die main reason for this is respiratory problem, this virus will infect all tissues inside the lungs so, slowly it will be very hard to the human to take breath and finally he will die. In mean time every doctor will do chest x-ray scan to measure the severity of the disease and infection inside the lungs.

Usually medical guys will know the status of the disease very easily by seeing the reports and X-ray images but the public who don’t have idea on medical terms they will not understand the infected area and infected severity.

**Research on Chest X-rays:**

After having above idea in my mind I just want to do research on available X-Rays to know the severity of the disease and level of infection inside the body so that public can alert and have a chance to consult or ask doctor for remedy for the damage and doctors can easily identify and give proper medication to the patient. I thought to explore X-ray images as doctors frequently use X-rays and CT scans to diagnose pneumonia, lung inflammation, abscesses, and enlarged lymph nodes. Since COVID-19 attacks the epithelial cells that line our respiratory tract, we can use X-rays to analyse the health of a patient’s lungs. Given that nearly all hospitals have X-ray imaging machines, it could be possible to use X-rays to test for COVID-19 without the dedicated test kits.

A drawback is that X-ray analysis requires a radiology expert and takes significant time which is precious when people are sick around the world. Therefore developing an automated analysis system is required to save medical professionals valuable time.