CAP 6778

Advance Data Mining and Machine Learning

Noisy Datasets and Their Effect in Deep Neural Networks

December 12, 2021

Carlos Gross-Martinez

Z23226341

Dr. Taghi Khoshgoftaar

Abstract

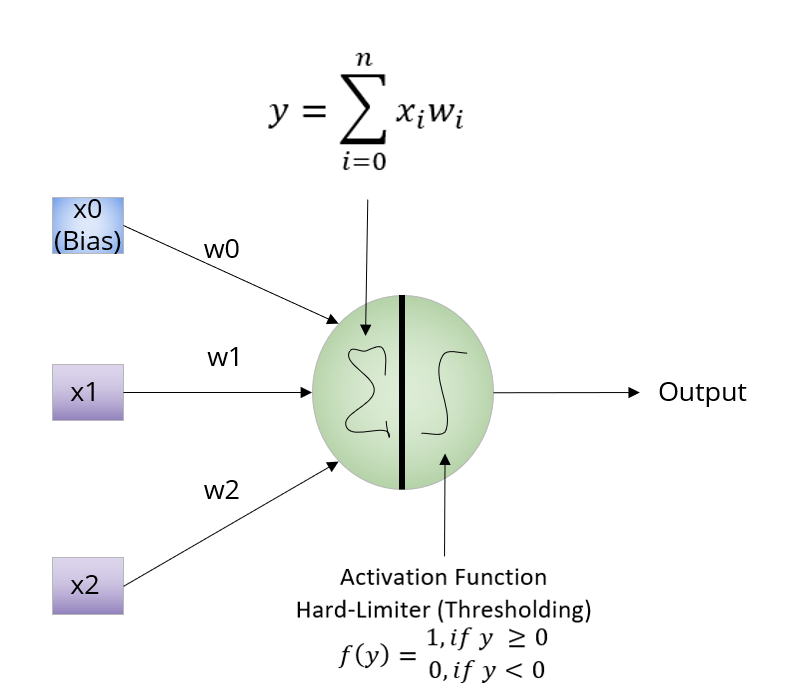
Artificial intelligence is one of the most aggressive growing fields in the technology industry today. The notion to fuse mathematical concepts, or other techniques with software, in order to design programs with the ability to inherit some sort of human rationale to complete tasks that only a person can perform, makes our time one of the most exciting times in the history of human kind. One such artificial intelligence technique which has taken some of the most monumental leaps in its development and implementation in the past decade is known as deep learning or deep neural networks. Today, deep neural networks are one of the most important and fast-growing fields within the artificial intelligence and machine learning domain. Nevertheless, as this technology continues to evolve, the complexity and challenges of its training process has increased as well. One big challenge of training a deep neural network to properly conduct a defined task, is the existence of noise in the dataset used to create and build the model. This paper focuses on introducing the notion of deep neural networks and the challenges it faces in the training process due to the utilization of noisy datasets. Moreover, it covers some of the different mechanisms used in order to try to mitigate this problem, as well as the flaws of the mechanisms.

Introduction

The Merriam-Webster dictionary defines Artificial Intelligence as “an area of computer science that deals with giving machines the ability to seem like they have human intelligence”. This idea has revolutionized today’s industry as AI continues to be at the fore front of the vast majority of topics covered in the technology field. In our modern society, companies understand that in order to compete and succeed in such a crowded and demanding market, the implementation of AI technology is not only necessary but indispensable in order to enhance their functions and operations. One of the most notable Artificial Intelligence techniques that company seek to implement and integrate is known as deep learning or deep neural networks. This method falls within the machine learning domain which is in itself a subcategory of the artificial intelligence field. In it most simple definition, a deep neural network is the interconnections of multiple layers of artificial neurons which tries to replicate the functionality and connectivity of a human neurons in order to identify and learn different patterns and characteristics of data so that it can make statistical predictions of new information. Their implementation in our society is huge and more continue to be added by the day. These deep neural networks integrate into cars to help them drive themselves, they combine into cameras in cellphones so that better pictures can be taken, they are found in software development tools such as IDEs so that developers can code faster and more efficiently, and thousands of other implementations which help us increase our productivity and quality of life. Nevertheless, it is necessary to mention that although deep neural networks continue to evolve and make our lives easier, the training mechanisms used in order to create and build these models continue to face their fare share of challenges as well. One of such big challenges is found in the ability to obtain clean and reliable datasets to build and train and deep neural network. The rapid expansion of the internet in past two decades, have provide the ability to gather huge amounts data that can be used in order build and train deep neural networks. Nonetheless, there is so much data out there that it makes it extremely difficult to properly sort and categorize this data. Since data cannot be properly categorized, then the information within it becomes corrupt and misleading. Hence, when the data used to build and train deep neural network models is misleading and corrupt, then a built model from this data is not reliable.

Neuron Basics in Deep Neural Networks

The concept of deep neural networks was inspired through the understanding of neurons in our body and how these neurons transmit information, and it uses the concept of a network of neurons which can be tuned in harmony in order to perform different tasks such as object classification. In order to grasp a better understanding of the impact of training a deep neural network with datasets containing misleading or corrupt information, it is necessary to have a fundamental understanding the pieces of this technique and how it works in order to create and build a model.

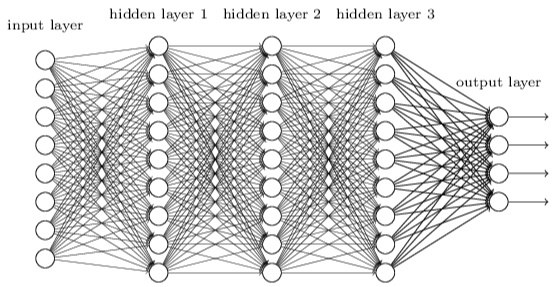


Picture 1: Single Neuron Structure for Deep Learning Architecture

In the same manner, in order to understand the overall technique of deep neural networks, it is essential to have a notion in the strategy behind its most fundamental concept which is the neuron. The idea of Deep Learning is born from the implementation of various interconnected neurons in a network to build a deep neural network model with the ability to perform certain tasks. Picture 1 above, shows a single neuron classification strategy once it has been trained. On the left, the picture shows the parameters of the neurons which consists on the features of the input to the neuron alongside the bias. Moreover, the picture shows that each feature in the neuron contains an individual weight. During training and validation, each weight for each parameter is altered constantly to reduce error and obtain the greatest probability of proper classification of the data sample. Once the model has been trained and validated, the classification strategy is very simple consisting of computing the summation of the products of all the weights times their respective parameters. With the results from the summation, an activation function is executed in the results the prior step to output the final prediction or classification of the model. Moreover, picture one above displays the Hard Limiter activation function which is mainly used for perceptron modeling. Nevertheless, for Deep Learning neural network design other activation functions such as ReLU, SoftMax, and sigmoid are used in order for the model to be able to adapt to more complex architectures.

Deep Neural Networks Basics

With a strong understanding of the underlying piece which build a deep neural network, it is now possible to expand in this concept to its implementation into deep neural networks. A deep neural network is comprised by stacking neurons on top of each other. Each neuron in the stack has a specific purpose and its parameter can be individually tuned to perform the overall assignment. A deep neural network is developed when layers of stacked neurons are group and connected together in order to conduct a specific job such as classifying an object to a specific class.



Picture 2: Sample Structure for Deep Neural Network Architecture

Picture 2 above displays a sample architecture of a deep neural network. In there it can be observed how stacked neurons are stacked in layers in order to form the full architecture of the model. It is essential to note that deep neural networks can have even more layers and stacked neurons per layer in its design depending on the purpose of the model. Additionally, it is essential note that other deep neural networks designs can have a its own independent number of neurons in each layer and they do not have to be equal in each layer as shown above.

Datasets

With a strong understanding of deep neural networks’ architecture and operational concept, we can now move to understanding what a dataset is. A dataset according to the Oxford dictionary can be defined as “a collection of data that is treated as a single unit by a computer”. Furthermore, a dataset can be considered as a collection of items which have similar attributes. E.g., a dataset containing pictures of dos and cats. Each picture in the dataset is a sample for each category, either a picture of a cat or dog. Hence, each picture is labeled either as cat or a dog in the dataset, and each animal has its own set of attributes which distinguishes it from one another. A normal person is able to just pick a random picture from this dataset and quickly determine if the picture is a dog or a cat. Therefore, it can be concluded that if there are 100 pictures in the dataset, a person should be able to determine by looking at the pictures if the picture belongs to the cat or a dog class. With this in mind, it can then be determined that we are able to classify all the pictures in the dataset with two labels. Either the picture contains a cat or a dog. Moreover, any individual which knows the difference between these two animals has the ability to easily classify each picture in the dataset, but how is this classification even possible? In this case if a person were to reach into the dataset and pull a picture from it, the person will look for specific characteristics or attributes in the picture in order to classify the picture as a cat or a dog. For example, if the picture the animal has an elongated face, it is big, and possesses long ears, then the animal in the picture is more likely a dog. On the other hand, if the animal’s face in the picture is short, and it is small with short ears, then the animal is more likely a cat.

Datasets and Deep Neural Networks

In order for a deep neural network to conduct the same classification as any person, it needs to be trained with a dataset to help them learn and distinguish the differences between the classes or labels in the dataset. In the same manner as a human, a deep neural network model will receive the attributes or characteristics of a sample in a dataset as its input to the neurons and it will adjust the weights of the model in order to increase the probabilities of a proper classification of the picture in each iteration. The deep neural network will have to iterate as many times as necessary in order to fine tune its weights in order to make the proper classifications of the dataset. In order to accomplish this feat, the dataset needs to be divided into two separate subsets of the main datasets. One subset will contain all the samples related to one specific class or label, while the second subset will contain the remining sample of the second class. By conducting this segmentation for each class, we are able to feed the samples to the deep neural network in order to train it to distinguish between the two separate classes or labels.

Noisy Datasets and Deep Neural Networks

Now that there exists a solid understanding of what a deep neural network is, how it works, and how a dataset is used to train these models, we can now discuss the challenges being faced by engineers that are trying to train and build models with datasets. The expansion of the internet has allowed for the gathering of data to increase monumentally in the past two decades. Today, data is being gathered through cellphones, social media, research, and many other forms to build massive knowledge bases which allows for information to be available to the masses for mining and exploitation purposes. Nevertheless, these massive amounts of information are very hard to categorized and maintain and they continue to grow with the passage of each day. With this in mind, the lack of proper classification of the information has allowed for the data to be misleading in various senses causing big problems. Referring back to the example of the dataset with samples from cats and dogs, this misleading information can be thought of pictures in the dataset which are for cats but are mistakenly labeled or classified as dogs. In the same manner, a picture of a dog in the dataset mistakenly labeled or classified as a cat. Therefore, it can be concluded that these samples are misleading since they belong to a class which is not properly represented in their true label. This misrepresentation in the label to the sample can be considered as noise in the dataset. In order words, noise in a dataset can be defined as the improper labeling of samples in a dataset in regards to the proper label or classification. Now imagine that in a dataset of 100 pictures with labels of cats and dogs, where 10 pictures in each subset of the dataset are mislabeled or erroneously categorized. If regular people were to go through these pictures, they will be able to quickly determine that some pictures were improperly classified. Nevertheless, if the same noisy dataset was to be used to train a deep neural network with the capability to distinguish between these two classes, cats and dogs, then the models will not be properly trained because some of the samples in the dataset are wrongly categorized and labeled which prevent the classifier from properly learning the distinct and unique characteristics and features for each category or label in the dataset. Instead, it will cause confusion as the model learns from the samples that dogs and cats are virtually the same in specific instances based on the information contained in the noisy dataset. This will cause the classifier to not properly learn the differences between the two categories or classes in the dataset and therefore, perform the improper classification of new samples after it has been trained. This misclassification not only reduces the classification performance of the deep neural network, but depending on the implementation of the technology, it can be dangerous since the misclassification of an object can have catastrophic consequences in real world scenarios.

Techniques to Tackle Noisy Datasets and Deep Neural Networks

In order to mitigate the problem with noisy datasets, the industry has come up with different approaches in order to address this problem. One of the most rudimentary solutions is to have a group of people to go through each sample in a dataset in order to ensure that the samples are categorized properly. Nevertheless, this approach is not the best way to confront the problem since if a dataset contains 1 million samples, then a lot of people will have to perform this tedious task in order to completely remove the noise of the dataset, and even then, factors like fatigue and human error can also have an impact in properly reducing all the noise form the classes in the dataset. Another approach proposed by Jinchi Huang et al, consist in implementing a novel technique call O2U. Which consist on building deep neural networks with the ability to learn with noisy samples and still build robust models. In their research, they argue that having noisy samples in a dataset causes a model to overfit the classifier. Hence, they would need to underfit the model in order to compensate for the overfitting in order to create a balance which will provide the highest level of performance of the model with the available data. The underfitting of the model follows an iterative algorithm to increase the underfitting of the model and achieve the necessary balance of the model. The issue with this approach is that as the noise in the dataset increases, the classification performance of the classifier decreases. Therefore, this technique is only beneficial with datasets which contain a low percentage of noise in the dataset. It is essential to note as well that the number of cycles in the O2U approach also have a great impact in the classification performance of the classifier. As the number of iterations increase, the model gains better accuracy. Nevertheless, it still does not perform properly when the noise in the dataset increases. Others researches in the industry such as Taghi et al, studies and researches the impact of noise in a dataset based on the model by implementing different techniques to select better features from the dataset to build more robust models even if noise is present in the dataset. Additionally, Taghi et al, also studies the impact of using different sampling techniques in noisy datasets in order to produce more robust deep neural networks classifiers that are able to learn with noisy data. Nevertheless, the impact of noisy samples in a dataset base on the volume of the noise can render the best technique or approach in the industry incapable of building a strong model. Another approach in the process is to use loss functions in the model which are capable of creating better classifiers even though noise is present in a dataset. Nonetheless, it remains a fact that no matter the technique or approach, noise is an issue which greatly decreases the performance of a classifier when compared to models which use a training dataset which is free of noise.

Conclusion

Although there exists a great number of clean datasets which allows for the proper training of a deep neural network model, the vast majority of the data available is mostly noisy and the resources to conduct the cleansing of the dataset is just not feasible. With this information in mind, it is essential develop models which can be trained with noisy datasets and still the same level of performance as in models trained with clean datasets. Additionally, it is essential to understand that although there are techniques in the industry in order to build robust models with noisy datasets, these techniques are far from perfect and still require a lot of research in order to ensure that models can still be trained and performed as desired with mislabeled information. Although it is believed that having an extremely low volume of noise in a dataset is good in order to increase the robustness in the deep neural network model. The existence of high volumes of noise greatly decreases the classification performance of the same. Some researches in the industry continue to study how noise in dataset impacts the different architectures in deep learning models and how to build better models which are capable of learning with different levels of noise in the dataset. Nevertheless, these studies still require a lot of research in order to finally determine the best methodology in order to ensure that a deep neural network can be trained with a high volume of noise in a dataset.

References

“Adding Noise for Robust Deep Neural Network Models.” *DebuggerCafe*, 3 Feb. 2020, [debuggercafe.com/adding-noise-for-robust-deep-neural-network-models/](file:///C:\Users\Carlos\Downloads\debuggercafe.com\adding-noise-for-robust-deep-neural-network-models\).

Borodinov, Nikolay, et al. “Deep Neural Networks for Understanding Noisy Data Applied to Physical Property Extraction in Scanning Probe Microscopy.” *Npj Computational Materials*, vol. 5, no. 1, 22 Feb. 2019, pp. 1–8, [www.nature.com/articles/s41524-019-0148-5](file:///C:\Users\Carlos\Downloads\www.nature.com\articles\s41524-019-0148-5), [10.1038/s41524-019-0148-5](file:///C:\Users\Carlos\Downloads\10.1038\s41524-019-0148-5). Accessed 12 Dec. 2021.

“Data-Set Noun - Definition, Pictures, Pronunciation and Usage Notes | Oxford Advanced Learner’s Dictionary at OxfordLearnersDictionaries.com.” *Oxfordlearnersdictionaries.com*, 2021, [www.oxfordlearnersdictionaries.com/us/definition/english/data-set?q=data+set](file:///C:\Users\Carlos\Downloads\www.oxfordlearnersdictionaries.com\us\definition\english\data-set?q=data+set).

Hulse, Jason, et al. *Evaluating the Impact of Data Quality on Sampling*.

Kennedy, Robert, et al. *The Effects of Class Label Noise on Highly-Imbalanced Big Data*.

Merriam-Webster. “Definition of Artificial Intelligence.” *Merriam-Webster.com*, 2019, [www.merriam-webster.com/dictionary/artificial%20intelligence](file:///C:\Users\Carlos\Downloads\www.merriam-webster.com\dictionary\artificial%20intelligence).

Seiffert, Chris, et al. “An Empirical Study of the Classification Performance of Learners on Imbalanced and Noisy Software Quality Data.” *Information Sciences*, vol. 259, Feb. 2014, pp. 571–595, [10.1016/j.ins.2010.12.016](file:///C:\Users\Carlos\Downloads\10.1016\j.ins.2010.12.016). Accessed 12 Dec. 2021.

Wald, Randall, et al. *The Effect of Measurement Approach and Noise Level on Gene Selection Stability*.