Artificial Intelligence

Moral Machine - Ethics

December 16, 2021

Diogo Rente - 2018294032
Francisco Pais - 2018288054
Rui Seabra - 2018274486

1 *Abstract*

Nowadays, we observe an exponential growth of technology that indirectly increases the influence in the daily life of human beings. This high growth has brought a set of facilities to human beings, for example, the use of robots in factories for a greater production that gives the ability to satisfy all people around the world, smartphones that allow communication with people on the other side of the world, the making of video calls and the sharing of images, etc. On the other hand, this sudden and abrupt evolution brings with it several consequences, one of which is the need to increasingly create or at least attempt to create an AI (artificial intelligence) and, consequently, the need to have it being regulated accordingly and to demand laws. Those who create them have to know how to transmit the value of ethics to AI. This is extremely important because the non-existence of these standards can very well lead to horrifying results like the ones we've all seen in the cinema, like the "Terminator", "I, Robot" or even the "Matrix", in reality. (Abstract)

2 Introduction

As Artificial Intelligence (AI) becomes more prevalent and its impact on human daily basis increases, ethical considerations are also being studied when it comes to implementing new algorithms. For instance, it is important to consider how data collected by AI systems will be processed in a way that is fair and not biased. This leads to new scenarios where the traditional ethics aren't enough to make decisions. One of the ethical concerns related to AI is the lack of transparency in its decisions and how the AI embodies our value system. They aren't always understandable to humans and being able to transmit the reasoning behind a decision is one of the core objectives to AI's improvement.

The idea that AI can inherit the positive characteristics of human morality is encouraging. It means that machines can be trained to make choices between wrong and right. In a world where AI is being used for tough decisions, we need it to be trustworthy, fair and properly moral.

In self-driving cars, where mistakes may be life-threatening, we enter gray areas in ethics. This is reflected in the moral machine project (http://moralmachine.mit.edu). This project has participants judge different scenarios facing autonomous vehicles which have malfunctioned (for example, the brakes not working) and have them select which of two outcomes they prefer. The information acquired from this project can be used as examples of the opinion on ethics of various people. Nonetheless, although the results may be interesting to discuss, the most chosen answer isn't the correct one. There isn't, in most cases, a right answer to these situations. To solve this problem we needed to create a priority list of ethic's laws like the rules that humans live by. This is impossible to accomplish since there is never going to be general consensus regarding these problems, because this topic has a lot of cultural influence.

3 Related Work

Of the various articles read, we found an interesting part of the investigation of the topic "Ethical Machine Learning and Artificial Intelligence" provided by the professor. In the article "Moral Choice Machine" he shows the application of machine learning to human texts to extract deontological ethical reasoning about "right" and "wrong" conduct. They start by creating a list of prompt and answer templates like "Should I [action]", "Is it okay to [action]?", etc. Their results indicated that MCM training in different temporal news and books from the year 1510 and 2008 demonstrated an evolution of moral and ethical choices in different periods of time. They also concluded that moral prejudices can be extracted, quantified, tracked and compared across cultures and over time.

After this reading, an enthusiasm for testing a MCM in real life contexts was revealed to us, in order to know how it would react to different scenarios, in which different variations in the environment were placed on it. We started by analyzing a simple game in which the user had to choose between two images. In these two images, regardless of the choice made, people would always end up dying, what would change would be the person's physiology, whether the direction of the car, whether the person was passing a red or green light, whether it was a child or adult between other aspects. After this variability of factors we found it interesting and managed to obtain the data set of this game where thousands of data from people's choices were stored. After obtaining the data set, we proceeded to its pre-processing since it was a huge file and then to the construction of several neural networks for the same data set. Each neural network was evaluated following these four criteria: accuracy, precision, recall and f1.

We think that the conclusion drawn by the previous group in their investigation can also be applied in a similar way here because from culture to culture the choices are different which can lead to identifying which culture we are present and above all knowing the type of people who leads to a certain choice, for example, a more aggressive person or a more passive person will have different choices so they are easier to identify through the way the MCM reacts to imposed situations.

4 Materials

We used the data set made available by the moral machine project. This data set contains 41 columns. We narrowed this number down to 23 features to work as input to our neural networks.

- Barrier: describes whether the potential casualties in this outcome are passengers or pedestrians
- CrossingSignal: represents whether there is a traffic light in this outcome, and light colour if yes
- DiffNumberOFCharacters: takes a value between 0 and 4; difference in number of characters between this outcome and the other outcome.
- NumberOfCharacters: takes a value between 1 and 5, the total number of characters in this outcome.
- Number of Character (by "type")
  - Man
  - Woman
  - Pregnant
  - Stroller
  - OldMan
  - OldWoman
  - Boy
  - Girl
  - Homeless
  - LargeWoman
  - LargeMan
  - Criminal
  - MaleExecutive
  - FemaleExecutive
  - FemaleAthlete
  - FemaleDoctor
  - MaleDoctor
  - Dog
  - Cat

We also used *ResponseID* (a unique, random set of characters that represents an identifier of the scenario. Since each scenario is represented by 2 rows, every row should share a *ResponseID* with another row) to group outcome of different scenarios and *Saved* (actual decision made by the user) as our Target for training/testing and validating the neural network.

The training data set contains 334013 examples. The test set and the validation set are both the same length with 50434 examples, so we have around a 70% percentage for training and (15%/15%) for validation and testing.

We worked with *Google Colab* to develop our neural networks using the *pandas*, *pytorch* and *sklearn* modules from *python*.

5 Methods

In the methods explained further in this section(starting in the next paragraph) we started by splitting the data set in two parts, one part represents the left image of the scenario and the other one represents the right image. The target is built through this images by defining only one output. In this case 0 represents the right image and 1 represents the left image as correct answer.

Initially, we explored the behaviour of the Decisions Trees. The division of the training data set was 70% for training and 30% for testing. For this structure we made two different tests, first we tested the tree with a combined input and the other just with one input, that is, the first one evolves two inputs at the same time referring to the same scenario and the other is by itself. For the next step of the decision tree's exploration we implemented a Random Forest architecture and Extra Trees as well. The last architecture we tested was Multi-layer Neural Network that always has as input the combined input of the two images from the same scenario. In this case we changed the number of layers and the number of neurons at each layer. The output layer has always the same neurons in order to respect the desired output. Furthermore, we used a sigmoidal activation function to all neurons and we used two different loss function: $Mean-Squared\ Error$ (MSELoss) and $Rectified\ Linear$ (ReLU). For the training we defined the batch size for each iteration and two data sets, one it's all for training and the other equally divided in two for testing and validation.

6 Experiments

As it was explained before in this document the experimentation of the structures built is made with the same data sets and what changes is the percentage for training, testing and validation. For all of the structures presented in the Method's section we studied the impact of the parameters in the performance and in the behavior of the system by changing them. To measure this impact we used four metrics: $accuracy$, $precision$, $recall$ and $f1\ score$.

Moving on to the training phase of the networks starting with **Decisions Trees**. For this model we used two different inputs mentioned and explained already in this document. Next we trained **Random Forest Tree** which is done the same way as the Binary Tree.

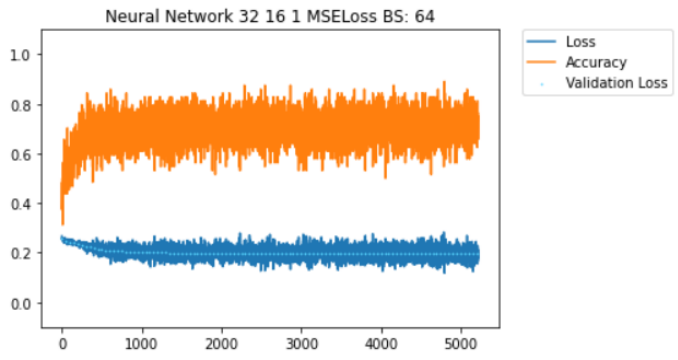
After training the decisions trees we decided to implement the **Multi layer Network**. All of the tested networks have the same 2 input layers (each describing one image of the same scenario) with 23 neurons (number of features retrieved from the dataset). They differ from each other in the next layers as in number of layers as in number of neurons in each layer. The first architecture trained had just one more layer besides the input layers that transforms the results provided by that layers and returns the desired output. This transformation from 32 results to just 1 may cause some performance issues. So we changed the architecture in order to solve that issue and tested two more networks. One with three more layers going from 24 neurons, 8 neurons and finally 1 neuron. The second was built with one layer with 16 neurons, other with 8 and the last one with 1 neuron.

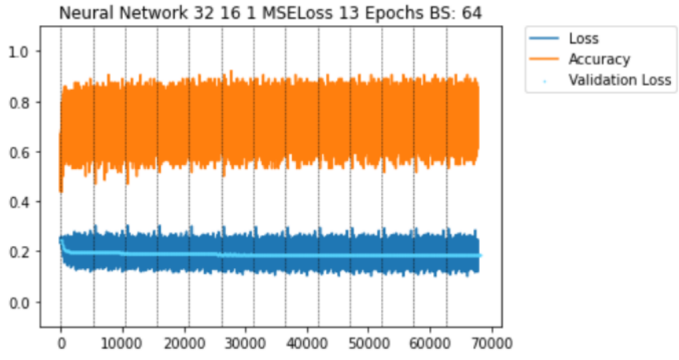
7 Results

These were the results obtained for each neural network constructed:

Test Results Table	Accuracy	Precision	Recall	F1 Score
Binary Tree Result(input not concatenated) Extra Trees	0.68	0.69	0.66	0.67
Binary Tree Result(input not concatenated) Decision Trees	0.67	0.68	0.63	0.66
Binary Tree Result(input concatenated) Extra Trees	0.70	0.73	0.71	0.72
Binary Tree Result(input concatenated) Decision Trees	0.66	0.69	0.67	0.68
Network MSELoss Results	0.72	0.74	0.75	0.75
Network Relu Results	0.53	0.54	1	0.7
Random Forest Results(input concatenated)	0.70	0.72	0.73	0.73
Random Forest Results(input not concatenated)	0.68	0.69	0.67	0.68

Train Results Table	Accuracy	Precision	Recall	F1 Score
Binary Tree Result(input not concatenated) Extra Trees	0.74	0.76	0.71	0.73
Binary Tree Result(input not concatenated) Decision Trees	0.74	0.76	0.71	0.73
Binary Tree Result(input concatenated) Extra Trees	0.84	0.85	0.84	0.85
Binary Tree Result(input concatenated) Decision Trees	0.84	0.85	0.84	0.85
Training graph below				
Training graph below				
Random Forest Results(input concatenated)	0.84	0.84	0.86	0.85
Random Forest Results(input not concatenated)	0.74	0.76	0.72	0.74





8 Discussion

As can be seen, we obtained an oscillation/variability of results. This fact may be due to the use of several neural networks that, through the network we are using, will influence the result obtained due to the way each one is built.

Another influence that may have existed and hence in some cases the results obtained are lower than expected, is due to the fact that we are using a part of the data set and not the entirety, which would lead to a larger sample of responses.

Other factor for the results not to be the most favorable is due to the fact that the answer itself depends from person to person, that is, it may happen that there is an equality of almost 50% of the choice of the image on the right side and others 50% of the choice of image on the left side, which makes neuronal networks learning difficult.

Taking a look at the graphs of training of the neural networks, its rather clear that the networks are over-fitting so, for that reason, we should've implemented a validation stop mechanism to prevent the network from over-fitting. Nonetheless, the training curve flattens at one third of the first epoch which is a bit odd.

However, taking into account the difficulty of the project theme, the results weren't that disappointing. The best overall result was in the **Multilayered Network** architecture with 72% accuracy, 74% precision, 75% recall and 75% f1 score.

9 Conclusions

In short, after analyzing the experiments and their results, one of the several conclusions we reached was the influence/importance that the data set will have in the construction of the neural network. In our case, although we are not using the data set in its entirety available, we can conclude that our neural network will be heavily influenced by that small part of the data set. For example: when selecting a part of the data set, we can select a field of people who have a certain ethical value different from other people.Now our neural network, if it is taught an ethics that for the majority of humanity is considered wrong but when it is trained with a data set by people whose ethics is not the most correct, what it will perform in this aspect is that the ethics it is applying is correct. In this case study of the car, knowing which life it should take, at first glance it seems like something very fictitious, perhaps a movie situation, but those who think so are mistaken because nowadays cars like those from the Tesla company already have a limited time for autonomous driving. So if these cars are already driving autonomously, it is to be expected that in a few years this limited time will change to an unlimited time and that there will no longer be a steering wheel and pedals for the driver to drive, leaving the departure and destination of the driver to the AI.If a failure in the brake system happens and the passenger will not be able to make a decision, then that's where the ethics of the AI that is driving the car comes in, which in a life-or-death situation will have to make a decision, may have one of those scenarios that we studied. Therefore, it should have been trained with the best data set, that is, that the people who have built it have a high standard of ethics. Once again, these issues of ethics, among others, are delicate and relative because it depends on the perspective of each person and their experiences in their daily lives. Therefore, as there are universal rules in justice, we believe that in ethics we should also start to create universal rules for them because the faster the better once the future passes through its construction, then the better it is to regulate this sector right from the ground and avoid greater harm.