Representation Learning to Aid Human Decision Making

Ben Wolfson bw916@nyu.edu
Rutvik Shah rss638@nyu.edu
Samantha Lee skl384@nyu.edu
Enyi Lian el2986@nyu.edu

Abstract

Rosenfeld and Hilgard are among the first to tackle machine inspired human-decisions, demonstrating that machines can be trained to manipulate output visualizations to direct human decision making to be more optimal [1]. In this paper, we develop an extension of their work, extending their framework to selecting between different types of representation. We find that machines can learn both representation parameters and type of representation. We see applications of our extension when choosing between representations that are suited to different types of humans and how they interpret the output.

1 Introduction

In the last 10 years, machine learning (ML) methods have made incredible inroads into the business world, and the predictive capabilities of such methods have proven both accurate and consistent. However, as humans are the ultimate decision makers in many ML applications, the usage of ML algorithms in practice remains largely dependent on human response to their output. To date, humans have struggled to interpret and effectively use ML outputs when making decisions, as the outputs are often numeric and lack interpretability [2]. In *Learning Representations by Humans, for Humans* Rosenfeld and Hilgard address the lack of model transparency by demonstrating that machines can be trained to manipulate the visualizations of the ML-output, thus optimizing the decisions made by the human users of said visualizations. For example, instead of a numeric output (of an X-Ray analysis) given to a radiologist with a probability of cancer, the ML algorithm learns the best way to present the information to the radiologist such that it induces the confidence of the radiologist to make the best decision (e.g. pointing out the suspicious area and including excerpts from the patient's medical records). Rosenfeld and Hilgard are among the first to advocate for shifting from the "learning to predict" framework to the "learning to represent" framework. Their goal was to optimize, not a "machine-generated output", but a "MoM" – "man composed with machine"[1].

In the "learning to represent" framework, ML is meant to integrate into the human reasoning process. Under ideal conditions; the algorithm not only recommends a course of action, but is also able to explain why that course is chosen. The explanation is essential to fostering high human confidence in the model.

Rosenfeld and Hilgard's "learning to represent" framework focuses on three different aspects of decision making, all important to learning in the cognitive realm:

1. **Interpretability** is essential in evaluating criteria that are difficult to quantify
2. **Incorporating human feedback** can provide labels that guarantee certainty and preference in the evaluation methods.
3. **Expertise, trust and agency** in the model ensures that the outputs are accepted as accurate predictions that the humans are satisfied with. Alternatively, if the model appears incorrect, they have the domain knowledge and understanding to intervene.

Rosenfeld and Hilgard's experiments focus on a single representation of the decision process that is modified to optimize the human decision. In this paper, we extend their analysis to select from *multiple* possible representations while optimizing the representation's role in the human decision making process.

1.1 Rosenfeld and Hilgard's Experiments

The original paper presents a series of experiments based on tasks that increase in complexity.

1.1.1 Learning the right 2D projection

The first experiment made high-dimensional data more easily manageable by projecting it into a lower dimension. In each instance, a human is given a 2D projection of either an "X" or and "O" that has noise added in the third dimension. The human is tasked with determining its label. The ML algorithm learns the correct projection under which humans can immediately and accurately guess the original label - the progress is driven completely by the user guesses and achieves 100% human accuracy in 5 epochs.

1.1.2 Learning a human understandable summary

A second experiment revolved around comparing the MoM framework with the LIME method [3]. While they both can provide an explanation on how a black box prediction can work, LIME is primarily successful when the human intuition is correct, whereas MoM is focused on adjusting itself according to how humans make decisions. When text summaries outputed by LIME (composed of the top and bottom three words with highest coefficients) were given as input to humans, human performance was only 65%. If given summaries generated by MoM, human performance reached 76%. Furthermore, Rosenfeld and Hilgard found that MoM summaries were better understood by humans, while the ones produced by LIME were not as helpful.

1.1.3 Optimizing loan decisions

The third experiment "focused on the problem of approving loans using the Lending Club dataset. Given details of a loan application, the task of a decision maker is to decide whether to approve the loan." [1]. The information was augmented with an avatar (picture of a face) making a specific facial expression. The ML algorithm learned to convey algorithmic advice through modifying the facial expressions to the loan-approving agent.

This taps into human cognition and empathy; by reading facial signals we can create a judgment. The results of this experiment showed more accuracy as the user feedback accumulated, but did not always reach the level of accuracy of the machine level benchmark at the high dimension. This variance is due to how the experimental subjects followed predictive advice, and proved there is a trade-off between accuracy and the ability to reason.

1.2 Conceptual Extension

Rosenfeld and Hilgard combine human decision making with the choice of the best form of a representation. Our project aims to further extend into the selection of the best format of representation. Based on picture superiority effect, our brain processes graphical data in a different way to a single number, and images are more likely to be retained in memory [4]. Visual stimulation over text translation allows the brain to consume the material with more consummate ease. In our experiment we posit a "trust factor" as a proxy for how much a human trusts a numeric output versus a more intuitive graphical representation. We allow the our MoM to learn the best representation (visual vs numeric) to show to the human as well as the optimal parameters for the chosen implementation.

Rosenfeld and Hilgard found that the way machine output is represented is important to inspiring the human to give the "correct" output. While their experiment focused on optimizing a singular representation $\arg\min_{\phi_W} L(h(\phi_W(x)), y)$, we extend this notion to not only optimizing over ϕ_W, but also selecting among different types of representations (ϕ_i's (see Figure 2)).

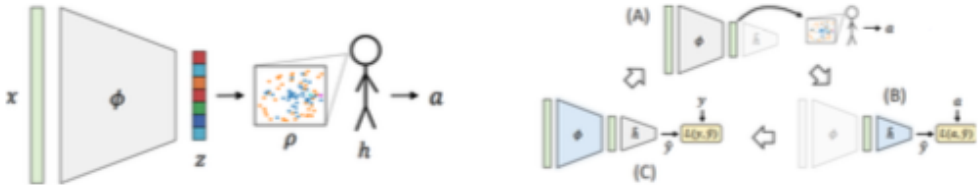

Figure 1: **Left**: The MoM framework. **Right**: The learning process.

2 Method

2.1 The Rosenfeld and Hilgard Set Up

We began with the same setting as Rosenfeld, where $x \in X$ sampled from some distribution D, for which humans are tasked with deciding on an action $a \in A$ [1]. Our assumption is that users are seeking to choose action a, which we model as $a = h(\phi_i(x))$. The machine generates visualizations, $\phi(x)$, that are human-centric representations of the input with the intention of helping a human make as accurate of a guess as possible. Given a training set S, we wanted to minimize the empirical loss using mean squared error (MSE):

$$MSE = L(y, \hat{y}) = \frac{1}{N}\sum_{i=0}^{N}(y - \hat{y}_i)^2 \quad \text{where } \hat{y}_i = h(\phi(x_i))$$

Similar to how humans apply what they have learned previously to new tasks, the architecture of a multi-layered neural network **N(x)** is a useful tool for learning representations. We split the network at some partition into a parameterized representation mapping ϕ and a predictor $f(\phi)$ (see Fig. 1). By re-training on new human inputs, the network improves over time. Rosenfeld and Hilgard chose to remove $f(\phi)$ and plug in the human decision function $h(\phi)$, to create a *Man Composed with Machine* ("**MoM**") framework illustrated in Figure 1. The intention is to use the human decision function to optimize ϕ. Instead of a human input layer, $h(\phi)$, we chose to develop a function $g_\tau(\phi)$ that generates a guess based on ϕ_i.

2.2 Preliminaries

We model two human-centric representations of the underlying data as ϕ_m and ϕ_s, where $\phi_m(x) : x \in R^{200} \rightarrow R^1$ and $\phi_s : x \in R^{200} \rightarrow R^{200}$, representing the "numeric" output and the scatter plot respectively, as seen in Figure 3. Furthermore we use $g_\tau(\phi)$ to simulate the $h(\phi)$, defined as follows:

$$\begin{cases} \phi_m(x) + \dfrac{1}{\tau} \cdot N(0,1), & \text{if } \phi = \phi_m \\ \mu(\phi_s(x)) + \sigma(\phi_s(x)) \cdot N(0,1), & \text{if } \phi = \phi_s \end{cases}$$

We use $g_\tau(\phi)$ to model a human guess with the following reasoning: if the human sees a single numeric output s/he is likely to guess the output with some added or subtracted noise depending on how much s/he trusts the machine (herein modeled by the τ factor). Similarly, if the human sees a scatter plot, we think that s/he will guess the perceived average of the scatter plot. However, if there is a large variance in the distribution, we assume that the guess will be less precise and therefore scale it by the standard deviation of the distribution.

Our model chooses which output to present to the human based on a learned parameter ϵ, used to balance exploring and exploiting the different solutions. We update epsilon at every time step using the following formula:

$$\epsilon_{t+1} = \min(\max(\epsilon_t + \frac{e_m - e_s}{e_m + e_s}, 0.05), 0.95)$$

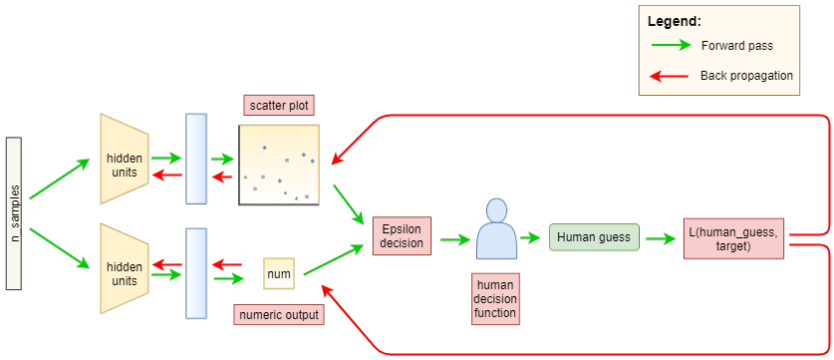

Figure 2: Model Architecture

where e_i, is an array of the losses generated by the representation $\phi_i \in \{m, s\}$ and where ϵ_{t+1} is bound to ensure exploration in the early stages. Finally, the model uses L to backpropagate loss through the relevant neural network representing ϕ_i. To implement, we build a neural network to output the two representation, ϕ_i's. At each time step t, we backpropagate the loss and update ϵ_t.

2.3 Experiments

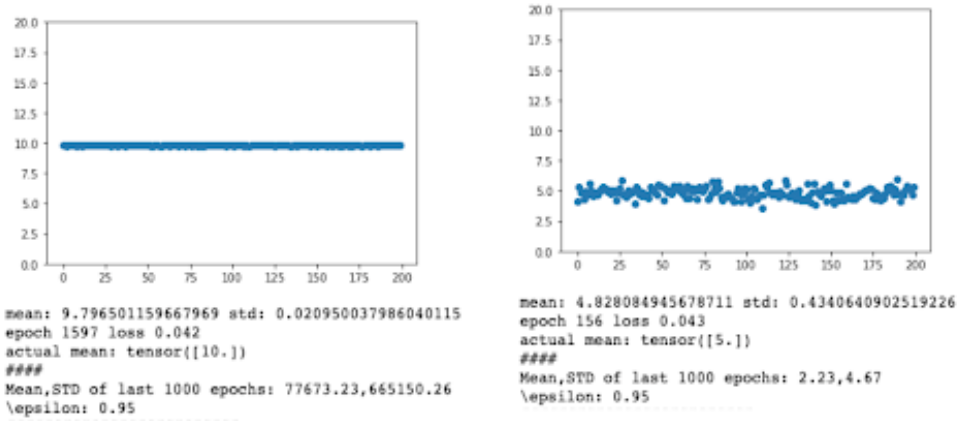

Figure 3: Example of model returns at one iteration

We want to model a human's guess of the mean of a distribution as a function of the representation that s/he gets and his/her trust factor τ (i.e., how likely the human trust the numeric output of a machine). Specifically, we look at ϕ_s and ϕ_m and see how the model learns which one is best suited to the h_τ.

We modeled the above behavior by drawing a random sample of 200 numbers from $x_i = N(k, l), k, l \in \{1, 4\}, \{0, 20\}$ and running our model for as many epochs as it took to generate less than < 0.05 loss for 5 consecutive epochs (the point at which we believe the model "found" the best representation).

We run this experiment for different values of τ to show that as the human trust in the machine changes, our model adapts to the preference and learns both the best representation and representation type to present to the human (ϕ_m or ϕ_s, with their best parametrization).

2.4 Results

The goal was for our model to learn the parameters from the representation and decide which visualization was the ideal one to return, as shown in Figure 4. The model learned the corresponding representation depending on the value of τ - with high values of τ the model chose ϕ_m and for low values it chose ϕ_s more often. Furthermore, we can see the parameters of the relevant $\phi_i(x)$, are also learned.

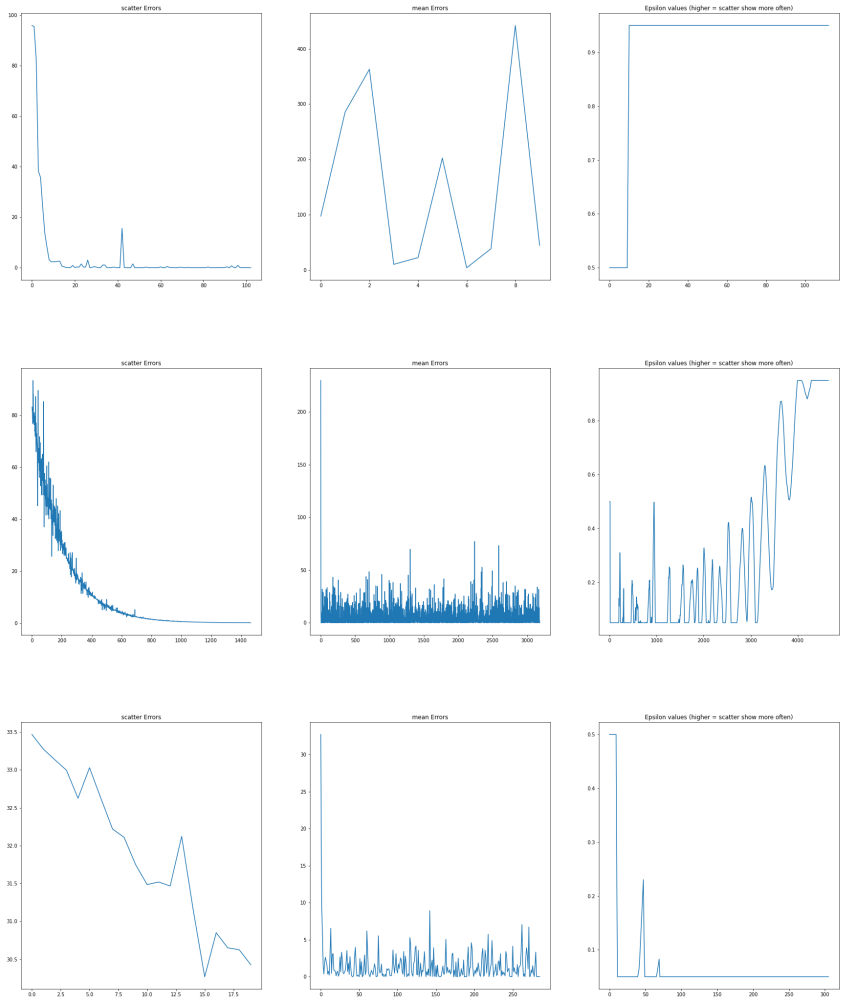

Figure 4: Results with $\tau \in [0.1, 0.5, 1]$, showing the error when showing the numeric output ϕ_m, the error when showing the scatter ϕ_s, and the value of ϵ in every epoch

3 Discussion

Our results demonstrated that we can extend the Rosenfeld and Hilgard framework to choosing the specific type of representation of data. Since in many situations the ultimate output of a ML algorithm is used to inform human decisions, tailoring the output *representation* becomes drastically paramount. But different humans respond differently depending on the data representation. Some are more likely to believe numbers, while others need a more holistic view of the data. We showed that we can train ML models to modify their output to these variations in human interpretation.

There are a few challenges that we faced while conducting the experiments. One of our primary difficulties was finding relevant representations of data that could be tested. There is an abundance of data available and many ways to depict them, but being able to tap into human cognition and create an explainable, impartial model that could be learned easily required heavy brainstorming and discussion. If we had more time, we would ideally experiment with other representations of data in different dimensions and more complex trust parameterization.

3.1 Further Research

We note that human decision making can hardly be modeled by a simple function, be it rule or neural-network based. Nevertheless, Rosenfeld and Hilgard have show that using a functional approximation of $h(x)$ can lead to efficient training in practice and positive results in real world scenarios. We extend Rosenfeld and Hilgard to show that the functional approximation can apply to different *types* of representations as well.

The field of representation learning is still growing, and the success of algorithmic modeling is reliant on how the output of ML algorithms is interpreted and subsequently used [5]. Ideally, the output of models should be interpretable, and by augmenting human participation in the calculation of the prediction, we can induce trust in the model. However, the large number of hyper-parameters in the deep neural network limits the methods and types of the exploration. Oftentimes, by having a human input layer such as h or g_τ, the composed ML algorithm may sacrifice high accuracy while optimizing for the "correct" human decision. We believe that the trade-off is worthwhile, because the end model is easier to understand (as it is built for humans) and more applicable in various professional settings.

Acknowledgement

Thank you to Sophie Hilgard and Nir Rosenfeld for jump starting us on this project and for spear-heading this new and exciting area of research. We hope to encourage further discussion and experimentation on the possibilities around representation learning and how it relates to human decision making.

References

[1] Sophie Hilgard, Nir Rosenfeld, Mahzarin R Banaji, Jack Cao, and David C Parkes. Learning representations by humans, for humans. *arXiv preprint arXiv:1905.12686*, 2019.

[2] Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, 2019.

[3] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[4] Margaret Anne Defeyter, Riccardo Russo, and Pamela Louise McPartlin. The picture superiority effect in recognition memory: A developmental study using the response signal procedure. *Cognitive Development*, 24(3):265–273, 2009.

[5] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.