Explaining predictions of artificial neural networks for travel behaviour analysis using layer-wise relevance propagation

Ahmad Alwosheel, Sander van Cranenburgh[1], Caspar G. Chorus
Transport and Logistics Group, Department of Engineering Systems and Services, Delft University of Technology

1. Introduction

Artificial Neural Networks (ANNs) are emerging as a new tool to analyse travel behaviour. Recent examples include modelling lane-changing behaviour of drivers (Xie et al. 2019), predicting mode choice behaviour (Sun et al. 2018), and investigating travellers' decision rules (Alwosheel et al. 2017; Van Cranenburgh and Alwosheel 2019). This increase in ANNs' use in transportation is inspired by impressive achievements of ANNs in applications outside of transportation, such as in computer vision and natural language processing (Karlaftis and Vlahogianni 2011; Goodfellow et al. 2016; Mckinney et al. 2020), and further fuelled by the increase of data becoming available to transportation researchers (Chen et al. 2016).

Despite ANN's often superior prediction performance compared to their traditionally more widely-used theory-driven counterparts, such as Random Utility Theory (RUT) based discrete choice models, their use is hampered by their black box nature. That is, it is impossible to interpret or diagnose ANNs by looking at the weights of the network. The weights will tell the analyst nothing about the importance of attributes, or whether the ANN has learned intuitively correct relationships. In other words, after having trained an ANN the analyst is in the dark about whether and when he can trust the ANN's predictions.

Recently, the development of techniques for opening up and explaining ANN's black-box has been the subject of many research efforts in a variety of research fields (Lipton 2016). Notably, in the computer vision field much progress has been made to shed light on the inner workings of trained ANNs (Simonyan et al. 2013; Samek et al. 2017; Montavon et al. 2018). One technique, called Layer-wise Relevance Propagation (LRP), has emerged as one of the most popular techniques to inspect the rationale behind ANNs' predictions (Adebayo et al. 2018). LRP generates a so-called heat map. For example, the heat map of an ANN trained to discriminate between dogs and cats based on pictures, highlights which parts of an image (e.g., pixels representing cat whiskers) were most relevant for the produced prediction (in casu: cat). The generated heat map reveals the rationale of a trained ANN and as such allows for intuitive investigation of what let the ANN to produce a particular prediction. In case the rationale aligns with the mental map of the analyst, the analyst gains trust in that prediction, and by extension in the trained ANN as a whole. In case the exhibited rationale does not align with the mental map of the analyst, it could point the analyst to biases or imbalances in the training data, or causes the analyst to update his own mental map.

This paper reconceptualises the use of LRP-based heat map generation and pioneers its use in a transportation context. In particular, we show that by properly reconceptualising the notion of heat maps, they can provide meaningful explanations for predictions made by ANNs which were trained for predicting travel mode choices. As such, our paper presents a method to help analysts gain trust in ANNs' predictions in transportation contexts. For the empirical part of our study, we use a recently collected Revealed Preference (RP) mode choice data dataset (Hillel et al. 2018a).

[1] Corresponding author: s.vancranenburgh@tudelft.nl

2. Reconceptualising Layer-wise Relevance Propagation

2.1. Layer-wise Relevance Propagation

LRP operates by propagating the activation strength of the node of interest backward, through hidden layers, to the input layer. In this study, we limit our focus on understanding the ANN prediction; hence, we are mainly concerned with back propagating the activation at the *output* nodes backwards through the hidden layers, using local propagation rules, until it allocates a relevance score R_i to each *input* variable x_i (Samek et al. 2017). Each R_i can be interpreted as the contribution an input x_i has made to a prediction (see Fig. 1).

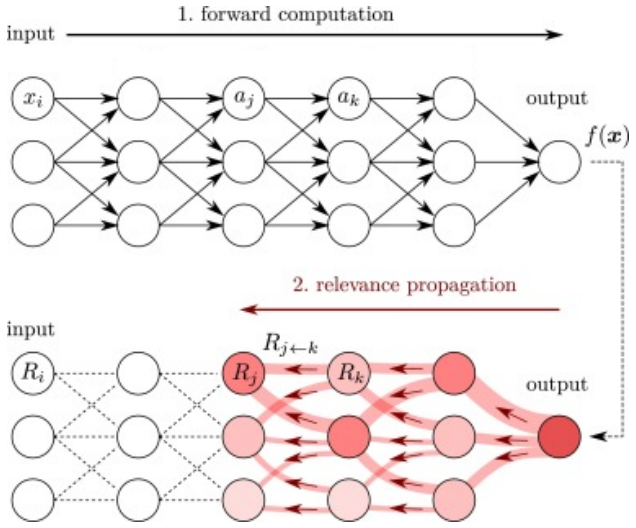

Fig. 1 : Diagram of the LRP procedure (Montavon et al. 2018). Red arrows indicate the relevance propagation flow.

The key property of the relevance redistribution process used in LRP is that the total relevance at every layer of the ANN (from the output layer to the input) is maintained; this property is known as relevance conservation and can be described as follows:

$$\sum_i R_i = \cdots = \sum_j R_j = \sum_k R_k = \cdots = f(\mathbf{x}) \qquad (1)$$

where i, j and k are the indices for nodes on the layers, and R_k is the relevance of node k for the relevance $f(\mathbf{x})$. This equation highlights that the method computes the decomposition of $f(\mathbf{x})$ (most right) in terms of the input variables (most left). To ensure Equation (1) holds, two rules are imposed:

$$\sum_j R_{j\leftarrow k} = R_k \qquad (2)$$

$$R_j = \sum_k R_{j\leftarrow k} \qquad (3)$$

where $R_{j\leftarrow k}$ is defined as the share of R_k that is redistributed to node j in the lower layer (see Fig. 1). The redistribution of the relevance resembles the process of forward propagation (used to produce predictions). In forward propagation, the activation function $z(.)$ of the node k generates one output a_k that is fanned out to other neurons and can be described as follows:

$$a_k = z\left(\sum_j w_{jk}a_j + w_k\right) \quad (4)$$

where w_{jk}, w_k are the weight and bias parameter of the neuron. The main principle used by LRP to back propagate the relevance is that what has been received by a node should be redistributed to the nodes at the lower layer proportionally. In the literature, different ways in which relevance is back propagated have been proposed. Empirical studies have shown that some of these rules yield better relevance redistribution depending on many factors such as the used activation function and position of the hidden layer (i.e., the layer deepness). In this study, we use the ϵ-rule (as described in (Samek et al. 2019)), which back propagates the relevance to each neuron as follows:

$$R_j = \sum_k \left(\frac{w_{jk} * a_j}{\sum_j (w_{jk} * a_j) + \epsilon}\right) R_k \quad (5)$$

where ϵ is a fixed constant of small value ($\epsilon = 10^{-7}$) which is added to the denominator to prevent division by zero (not to be confused with the error in discrete choice models). Doing so avoids the relevance values to become too large. This equation shows that the relevance is propagated proportionally depending on: 1) the neuron activation a_j (i.e., more activated neurons receive larger share of relevance), and 2) the strength of the connection w_{jk} (more relevance flows through more strong connection). In this study, we focus only on the rule shown in Equation (5), and for more detailed description of LRP and comprehensive discussion on alternative relevance redistribution rules, interested readers are referred to Samek et al. (2019) and Lapuschkin et al. (2016).

2.2. A reconceptualization of LRP using Monte Carlo experiments

This subsection conducts a series of Monte Carlo experiments to get a feeling for how heat maps can be re-conceptualised and used in the context of discrete choice data. Table 1 shows the parametrisations of the three synthetic data sets that we generated. Each data set consists of three alternatives and two generic attributes: X_1 and X_2. Parameters have different values across data sets (we use negative, positive and neutral parameter values). Each data set consists of 10,000 hypothetical respondents, each making a single choice. Attribute levels are generated using a random number generator between zero and one. To create the synthetic choices, the total utility of each alternative is computed and the highest utility alternative is assumed to be chosen, following a Logit (RUM-MNL) model where the random part of utility is distributed Extreme Value type I with variance $\pi^2/6$.

Table 1: Synthetic data specification and parametrisation

Dataset no.	Model specification	Parametrisation	Cross-entropy (RUM-MNL)	ρ^2 (RUM-MNL)	Cross-entropy (ANN)	ρ^2 (ANN)
A1	$V_{un} = \sum_m \beta_m x_{umn}$	$\beta_1 = -6$ $\beta_2 = -4$	-0.53	0.51	-0.54	0.50
A2		$\beta_1 = +6$ $\beta_2 = +4$	-0.53	0.51	-0.54	0.50
A3		$\beta_1 = -6$ $\beta_2 = 0$	-0.59	0.45	-0.61	0.44

For each data set, a three-layers ANN with 4 hidden nodes on the hidden layers is trained. As has also been found in previous studies (e.g., Alwosheel et al. (2018)), the ANNs are able to learn the a RUM-MNL data generating process with high accuracy, in the sense that the prediction performance of the ANN almost matches that of the true underlying data generating process encoded in a corresponding discrete choice model.

For the first data set (A1), the negative sign of the parameters imposes a dislike for higher attribute values (i.e., the lower the attribute values, the more attractive the alternative becomes). Hence, the attribute values of the *chosen* alternative are expected to contribute negatively to the choice probability prediction for that alternative (as reducing the attribute values would increase the attractiveness of the chosen alternative). In contrast, we expect that high attribute values of the *non-chosen* alternatives contribute positively to the prediction, implying that the attractiveness of these non-chosen alternatives increases as these attribute values increase. These expectations are confirmed in Table 2, where we see the relevance of the attribute values that are computed using the LRP method[2], alongside the choice probabilities predicted by the ANN, for three randomly selected observations from the synthetic data.

Table 2: Results of observations randomly selected from A1 data set

	Attribute Values						Relevance						True chosen alternative			ANN prob		
	X_1			X_2			X_1			X_2								
	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt
Obs. 1	0.127	0.8887	0.916	0.038	0.871	0.742							1	0	0	0.99	0	0.0
Obs. 2	0.95	0.004	0.97	0.725	0.133	0.75							0	1	0	0.01	0.99	0
Obs. 3	0.936	0.957	0.045	0.882	0.866	0.157							0	0	1	0.01	0	0.9

Consider the three observations shown in Table 2, where alternative 1 is chosen in the first observation, alternative 2 is chosen in the second observation, and alternative 3 is chosen in the third observation. The blue diagonal values show that the attribute values of the chosen alterative have contributed negatively toward the predicted probability of the alternative being chosen. In contrast, the off-diagonal cells, which here are associated with the non-chosen alternatives, are coloured red. This means that the attribute values of these unattractive alternatives[3], which are comparatively high, positively contribute to the prediction that alternative 1 is chosen in observation 1, alternative 2 in observation 2, etc. Hence, increasing the attribute values of the non-chosen alternatives would in this situation further increase the probabilities of the chosen alternatives, which is exactly as expected.

Compared to the first data set, in the second data set (A2) the parameters have flipped signs. Hence, lower attribute values are more attractive than higher ones. Table 4 shows the results for three randomly selected observations. We use the same colour map and intensity as in Table 2. As can be seen, Table 3 reveals the same patterns as shown in Table 3, but colours are flipped, i.e., cells on the diagonal are red, and cells off the diagonal are blue. This is fully in line with expectations, as here an increase (decrease) in the attribute levels of the chosen (non-chosen) alternative positively contributes to the choice probability that is predicted for the chosen alternative.

Table 3: Results of observations randomly selected from A2 data set

	Attribute Values						Relevance						True chosen alternative			ANN prob		
	X_1			X_2			X_1			X_2								
	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3	Alt 1	Alt2	A
Obs. 1	0.97	0.19	0.06	0.95	0.07	0.107							1	0	0	0.99	0.01	
Obs. 2	0.108	0.98	0.0594	0.063	0.865	0.35							0	1	0	0.01	0.99	
Obs. 3	0.025	0.05	0.83	0.32	0.305	0.99							0	0	1	0.01	0	0.

Lastly, Table 4 presents the results for data set A3. Again, three randomly selected observations are shown. In this data set, β_2 is zero. This means that the attribute X_2 does not impact the decision makers' choices. As such, we expect the relevancies for these attribute values to have values that are close to zero. In line with expectation, Table 4 shows that all cells for X_2 are (almost) white – meaning that the values of this attribute do neither positively or negatively contribute to the predicted choice probabilities.

[2] To generate heat maps, the LRP method is used and implemented in the Python environment using the open source library iNNvestigate (Alber et al. 2019).

[3] Note that the ANN predictions are correct with very high confidence as shown by predicted choice probabilities of 0.99 for the chosen alternative in each of the three observations.

Table 4: Results of observations randomly selected from A3 data set

	Attribute Values						Relevance						True chosen alternative			ANN prob		
	X_1			X_2			$X1$			$X2$								
	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3	Alt 1	Alt2	Alt3	Alt 1	Alt2	A
Obs. 1	0.05	0.665	0.979	0.99	0.001	0.757							1	0	0	0.98	0.1	0
Obs. 2	0.97	0.05	0.99	0.323	0.385	0.329							0	1	0	0.01	0.99	
Obs. 3	0.98	0.9	0.00038	0.017	0.154	0.94							0	0	1	0	0.01	0

In sum, this application on synthetic data shows how the LRP method can be used to inspect the rationale based on which an ANN makes it predictions in a travel mode choice context, and it provides a first sign of face validity of the method. The next section presents an application of the method on a real empirical data set.

3. Applying the LRP method to empirical data

For this study, we use revealed preference (RP) data from a study conducted for travel mode choice analysis in London city (Hillel et al. 2018b). Three processing steps have been executed to prepare the data for this study: First, features that were considered redundant are removed, or merged with others. Second, we noticed that the dataset is highly imbalanced in terms of the chosen mode: walking (17.6%), cycling (3.0%), public transport (35.3%) and driving (44.2%). Such imbalances could affect the reliability of the trained ANNs (Haykin 2009). Therefore, the data imbalance is 'repaired' by eliminating the cycling alternative from the dataset. Third, we excluded very short trips (i.e., less than two minutes), as these were deemed not to contain a mode trade-off. The resulting dataset that is used for this study consists of 77,638 mode choice observations.

3.1. Results

Tables 5 to 7 show the back-propagated relevance extracted for three randomly selected observations.[4] It can be seen, that predictions are made with different levels of confidence. In the context of our analyses, a high confidence level means the network assigns a choice probability of more than 0.80 to one the modes, and a low confidence level means that the highest (across travel mode alternatives in the context of a particular observation) predicted choice probability is still below 0.40. As such, for diversification purposes and to build trust in the model as a whole, the three observations are randomly selected as follows: two predictions with high confidence levels and one prediction with low confidence level. Tables 5 to 7 show the ANN probabilities, the attributes' values, and relevancies obtained using LRP for the selected observations. A heat map is employed to visualise the relevancies.

Table 5 shows an observation of a middle-aged female, who holds a driver license and owns two cars, who chose the driving alternative, which indeed seems to be the most attractive travel alternative in this case as is the fasted and cheapest alternative. In line with intuition, the ANN predicts a choice for the driving alternative with a very high level of confidence (assigning a 0.99 choice probability to that alternative). The relevance values show that car travel time receives a strong negative relevance, as expected (given that lower travel times are preferred). The relatively long travel times offered by the non-chosen alternatives receive a strong positive relevance, as expected (given that the high travel times of these alternatives makes driving alternative more appealing). Furthermore, the number of owned cars (two) and the driving license availability have a positive relevance. Together, these analyses reveal on what basis the ANN model has predicted that this traveller would choose the driving alternative. From a travel behaviour perspective, all these points are in line with expectations. As such, the analyst equipped with the proper domain knowledge can safely trust this prediction.

Moving forward to Table 6, for this observation, a young female traveller chose the walking alternative. As before, the travel time and cost of the non-chosen alternatives and the relatively high traffic on the driving route have high positive relevance for the predicted choice probability for walking. Furthermore, we see that the travel time of the chosen alternative, and the travelled distance have

[4] From the subset of observations that are correctly assigned by the ANN

negative relevance values. All these relations are expected from a travel behaviour perspective; hence, this prediction too can be safely trusted.

Lastly, in the Table 7 the alternative with highest predicted probability is walking; however, this mode receives a predicted probability which is only one percentage point higher than that of the other mode, implying that the ANN has low confidence in this prediction. The relevance values show that attribute values with negative relevance for the predicted choice probability for the walking alternative, are the relatively long distance and walking travel time, suggesting that shorter distance and less walking time would have made the walking alternative more attractive. This is expected from a travel behaviour perspective. Further, it can be noticed that the red and blue colours associated in this heat map are less bright, meaning that the ANN is less outspoken about what determined its prediction; this too is to be expected, given that the ANN assigns almost equal choice probabilities to each of the three alternatives.

Table 5: Results of observation 1: true chosen alternative is Drive

	Alternatives' Characteristics						Other Characteristics		
	Alt.1: Drive	Alt.2: PubTr	Alt.3: Walk	Alt.1: Drive	Alt.2: PubTr	Alt.3: Walk			
	Attribute value			Relevance			Attribute	Value	Relevance
TT (min)	23.48	137	160				AG	47	
TC (£)	1.58	7.50					FEM	0	
TRAF	0.03						DL	1	
INTER		4					CO	2	
ANN probs	0.99	0	0.01				DIS	9,556	

Table 6: Results of observation 2: true chosen alternative is Walk

	Alternatives' Characteristics						Other Characteristics		
	Alt.1: Drive	Alt.2: PubTr	Alt.3: Walk	Alt.1: Drive	Alt.2: PubTr	Alt.3: Walk			
	Attribute value			Relevance			Attribute	Value	Releva
TT (min)	7.38	4	6				AG	26	
TC (£)	10.70	2.40					FEM	1	
TRAF	0.31						DL	1	
INTER		0					CO	0	
ANN probs	0	0.01	0.99				DIS	368	

Table 7: Results of observation 3: true chosen alternative is Walk.

	Alternatives' Characteristics						Other Characteristics		
	Alt.1: Drive	Alt.2: PubTr	Alt.3: Walk	Alt.1: Drive	Alt.2: PubTr	Alt.3: Walk			
	Attribute value			Relevance			Attribute	Value	Relevar
TT (min)	7.23	22	24				AG	15	
TC (£)	0.30	0.00					FEM	1	
TRAF	0.44						DL	0	
INTER		0					CO	1	
ANN probs	0.33	0.33	0.34				DIS	1,271	

4. Conclusions

This study re-conceptualised the use of heat maps, generated using a Layer-wise Relevance Propagation method, to explain predictions of Artificial Neural Networks in the context of travel behaviour analysis. Our analysis suggests that the proposed LRP-based heat maps provide a valuable tool to understand the rationale behind ANN predictions in the context of travel choice behaviour, however it is important to acknowledge that the proposed method does not completely solve the ANNs' black-box puzzle as it will never completely explain the inner workings of the network.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M. & Kim, B. (2018). Sanity checks for saliency maps. Advances in Neural Information Processing Systems.

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., Samek, W., Müller, K.-R., Dähne, S. & Kindermans, P.-J. (2019). iNNvestigate neural networks! *Journal of Machine Learning Research*, *20(93)*, 1-8.

Alwosheel, A., van Cranenburgh, S. & Chorus, C. (2017). Artificial neural networks as a means to accommodate decision rules in choice models. International Choice Modelling Conference 2017.

Alwosheel, A., van Cranenburgh, S. & Chorus, C. G. (2018). Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *Journal of Choice Modelling*, *28*, 167-182.

Chen, C., Ma, J., Susilo, Y., Liu, Y. & Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies*, *68*, 285-299.

Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep learning*: MIT press Cambridge).

Haykin, S. S. (2009). *Neural networks and learning machines*: Pearson Upper Saddle River, NJ, USA:).

Hillel, T., Elshafie, M. & Ying, J. (2018a). Recreating passenger mode choice-sets for transport simulation: A case study of London, UK. *Proceedings of the Institution of Civil Engineers - Smart Infrastructure and Construction*, *171(1)*, 29-42.

Hillel, T., Elshafie, M. Z. & Jin, Y. (2018b). Recreating Passenger Mode Choice-Sets for Transport Simulation. *Proceedings of the Institution of Civil Engineers-Smart Infrastructure and Construction*, 1-49.

Karlaftis, M. G. & Vlahogianni, E. I. (2011). Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies*, *19(3)*, 387-399.

Lapuschkin, S., Binder, A., Montavon, G., Müller, K.-R. & Samek, W. (2016). The LRP toolbox for artificial neural networks. *The Journal of Machine Learning Research*, *17(1)*, 3938-3942.

Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

Mckinney, S. M., Sieniek, M., Gilbert, F., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M. & Corrado, G. C. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, *577*, 89-94.

Montavon, G., Samek, W. & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1-15.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K. & Müller, K.-R. (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*: Springer Nature).

Samek, W., Wiegand, T. & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Simonyan, K., Vedaldi, A. & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Sun, Y., Jiang, Z., Gu, J., Zhou, M., Li, Y. & Zhang, L. (2018). Analyzing high speed rail passengers' train choices based on new online booking data in China. *Transportation Research Part C: Emerging Technologies*, *97*, 96-113.

Van Cranenburgh, S. & Alwosheel, A. (2019). An artificial neural network based approach to investigate travellers' decision rules. *Transportation Research Part C: Emerging Technologies*, *98*, 152-166.

Xie, D.-F., Fang, Z.-Z., Jia, B. & He, Z. (2019). A data-driven lane-changing model based on deep learning. *Transportation research part C: emerging technologies*, *106*, 41-60.