Australasian Transport Research Forum 2023 Proceedings
29 November – 1 December, Perth, Australia
Publication website: http://www.atrf.info

Evaluating deep neural networks using post-hoc analysis in discrete choice modelling

Niousha Bagheri Khoulenjani[1], Milad Ghasri[1], Michael Barlow[1]

[1]School of Engineering and Technology (SET), UNSW Canberra, ACT, Australia

Email for correspondence (presenting author): n.bagheri_khoulenjani@adfa.edu.au

1. Introduction

Discrete Choice Models (DCMs) have become a dominant theoretical framework for studying individual travel behavior. Econometric models, including DCMs, have been used to investigate individual decision-making over decades (Train, 2009). However, there is a growing inclination towards utilizing machine learning models, particularly Deep Neural Networks (DNNs), for analysing traveller's choices (Zhang et al., 2020). While DCMs have been the primary approach for travel behaviour, DNN models have been shown to offer superior prediction accuracy due to their advanced learning algorithms and flixable modelling structure (Thanh et al., 2019). However, a major drawback of DNNs is that they are often considered as black-box models, meaning that it can be challenging to understand why they make certain predictions. When it comes to critical decisions in transportation planning, such as congestion pricing and infrastructure investment, interpretability is of utmost importance. Therefore, the lack of interpretability of DNNs can be a significant limitation for applications where transparency is crucial, such as choice modeling.

Recent studies have utilized neural networks as a tool to learn more flexible behavior representations within DCMs, while still trying to maintain model interpretability (Sifringer et al., 2020, Wang et al., 2020a, Wang et al., 2020b, Wong and Farooq, 2021). The basic concept involves integrating a neural network into a DCM, which allows the neural network to learn a portion of the model specification, while the base model remains a DCM. This design is referred to as the Neural-Embedded Discrete Choice Model (NEDCM) (Han et al., 2020). In a recent study, Sifringer et al. (2020) proposed a NEDCM based on the Random Utility Maximization (RUM) theory, as the Learning Multinominal Logit model (L-MNL). This approach divides the systematic part of the utility specification into two components, a knowledge-driven and a data-driven one. The advantage of L-MNL is that it has a lighter architecture and sparser connectivity compared with a regular DNN model. However, these extended DNN models may not be considered transparent and reliable in choice modelling, as they may contain hundreds of layers and thousands of parameters. To be more precise, a reliable and transparent model must accurately represent the real relationships between the explanatory variables and the outcomes of the choices made, while providing dependable responses to questions about hypothetical scenarios at the disaggregated level.

As the significance of interpretability in advanced machine learning approaches such as DNNs becomes increasingly apparent, the Artificial Intelligence (AI) research community is dedicating more attention to the subject of explainability of DNNs (Arrieta et al., 2020). Post-hoc interpretation techniques have been introduced to solve the interpretation problem of DNNs by explaining the model's decisions with high-level insights (Lipton, 2018). These methods help to increase the reliability of DNNs and enable their deployment in critical applications with improved trustworthiness. It can also help developers find errors or biases in

their system and can even facilitate the development of more robust systems. In earlier research, there were attempts to extract information from the neural networks. For example, Hagenauer and Helbich (2017) and Golshani et al. (2018) conducted sensitivity analysis to measure the importance of different variables. Meanwhile, Wang et al. (2020b) demonstrated how to extract a comprehensive list of economic indicators from neural networks. Factors such as elasticities and Value of Time (VOT) have been calculated using the gradient of choice probabilities, and other information such as market share have been computed using choice probability in the DNN. However, Multi nominal Logit models (MNLs) are able to provide further explicit parametric form for each economic information while DNNs are not. Despite the increasing attention towards post-hoc analysis in the field of discrete choice modeling, only a limited number of studies have been dedicated to uncovering behavioral insights using these methods.

The purpose of this paper is to study the role of DNN architecture in the application of explainability methods in the context of choice modelling. To achieve this, we compare and evaluate the application of Integrated Gradient method, as a cutting-edge post-hoc explainability approach on NEDCMs and fully connected DNN model. Our main focus is to measure consistency between the statistical theories, such as RUM, and both NEDCM and DNN models.

2. Methodology

2.1. Deep neural networks

DNNs have emerged as a powerful tool for modeling complex datasets (Goodfellow et al., 2016). Unlike fully connected DNNs, which are typically trained to predict the probability of choosing each alternative based on the attributes of all alternatives, NEDCMs incorporate information about the decision-making process itself. NEDCMs embed the utility functions of each alternative in a high-dimensional space, where the distance between the embedded utilities reflects the similarity of the alternatives. This approach allows NEDCMs to capture the underlying decision-making process more realistically, and to account for the fact that individuals may have different preferences and decision-making strategies.

The Alternative-Specific Utility function DNN (ASU-DNN) represents a cutting-edge NEDCM proposed by Wang et al. (2020a), with impressive accuracy in modeling discrete choice data. The structure of ASU-DNN is presented in Figure 1. ASU-DNN is designed with a specific architecture composed of an input layer, two hidden layers, and an output layer. The input variables are partitioned into two separate vectors, individual-specific and alternative-specific variables. Every group of variables progresses through a fully connected network. Each neural network within the second layer represents a utility function, while the last layer computes the output probabilities. By leveraging the RUM theory, ASU-DNN estimates the utility of each alternative by analyzing individual-specific variables and their corresponding alternative-specific variables, thereby providing a more comprehensive and realistic understanding of the decision-making process.

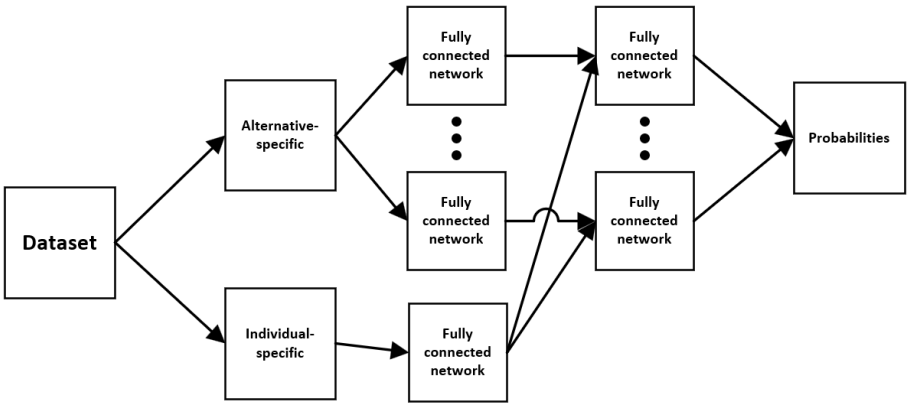

Figure 1: the structure of ASU-DNN model includes input layer, hidden layers, and the output layer.

The purpose of this paper is to compare two distinct DNN architectures, ASU-DNN and fully connected DNN, using real-world data in order to gain insight into how much the DNN's architecture impacts the significance of input variables.

2.2. Post-hoc analysis

The post-hoc explainability techniques have been proposed to shed light on the opaque behavior of DNNs. These methods are designed to provide a better understanding of how DNNs make predictions. Integrated Gradients is a post-hoc method for interpreting the predictions of DNNs by attributing importance scores to input features (Sundararajan et al., 2017). This method calculates the gradient of the output of a DNN with respect to the input features, integrating the gradients along a path from a baseline input to the actual input. Along this path, this technique calculates gradients at multiple points, which measure how sensitive the model's output is to changes in the input variable. By averaging these gradients over the entire path, the Integrate Gradient method provides information about how variation in input variables influence the model's prediction. The formulation of Integrated Gradient for ith feature of sample x is defined as:

$$IntegratedGrads_i(x) = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x')}{\partial x_i} d\alpha$$

x_i' refers to the ith dimension of the baseline x'. The resulting attribution scores reflect the contribution of each input feature x_i to the final output F, providing insights into the decision-making process of the DNN. Integrated Gradients has been shown to outperform other gradient-based attribution methods and has been applied to various domains, such as natural language processing, computer vision, and healthcare, to help improve model transparency and accountability (Lundstrom et al., 2022).

3. Experimental results

In this study, we evaluate the interpretability of ASU-DNN and the fully connected DNN using the Swissmetro dataset, which was collected in Switzerland in 1998 (Bierlaire et al., 2001). The dataset comprises responses from 1,192 individuals who were asked to choose their preferred mode of transportation among three alternatives: train, Swissmetro (SM), and car. For this analysis, we select Travel Time, Travel Cost, Age, and Income among the available variables for choice analysis. We use the Integrated Gradient method to provide insights into how these models make predictions. In this experiment, ASU-DNN and DNN include 2 layers

with 100 neurons in each layer. We also compare the performance of DNN models with a well-known choice model, the MNL, in terms of within-sample fit and out-of-sample prediction.

The dataset is divided into two parts: 70% for the training dataset and 30% for the test dataset. As presented in Table 1, loglikelihood and accuracy were calculated for fully connected DNN, ASU-DNN, and MNL models, using the test and train datasets. This result highlights the improvement in accuracy of ASU-DNN on the test dataset, reinforcing the ability of ASU-DNN in modeling unseen dataset. The number of parameters in ASU-DNN architecture is significantly less that the fully connected DNN. This means connections that are not supported by the theory are removed from the model. As a result, ASU-DNN avoids spurious correlations that could lead to overfitting, which is a common problem in deep learning models. Furthermore, as shown in the table, ASU-DNN demonstrates superiority over MNL in both training and test prediction accuracy and log-likelihood. Therefore, ASU-DNN has better generalization ability than fully connected DNN and MNL, which is a desirable property for a DNN model to have.

Table 1: the goodness of fit measurements of the fully connected DNN and ASU-DNN for the test and train datasets

Model	Number of parameters	Loglikelihood	Accuracy
Train			
ASU-DNN	**1,803**	-3519.97	74.92
Fully Connected DNN	21,403	**-1330.17**	**97.79**
MNL	10	-5095.18	65.11
Test			
ASU-DNN	**1,803**	**-1831.45**	**71.78**
Fully Connected DNN	21,403	-8861.69	67.20
MNL	10	-2203.35	66.21

Table 2 shows the average impact of input variables on the selection of each travel mode in the ASU-DNN model computed by the Integrated Gradient method. Travel Time and Travel Cost of each mode have negative effects on the selection of their corresponding mode, but positive effects on the selection of other modes. For example, travel time of train have negative impacts of -0.114 on the selection of train, but positive effects of 0.087 and 0.026 on Swissmetro and car selections respectively. This indicates that increasing the time and cost of a transportation mode would decrease its attractiveness, and meanwhile increase the chance of choosing other transportation modes. Moreover, the magnitude effects of Travel Time and Travel Cost are highest for their corresponding alternatives, which is also consistent with RUM theory. As an example, the magnitude impact value of train travel cost (0.26) is higher than its effect on Swissmetro and Car. Similarly, Table 2 shows Age and Income also have a significant impact on the choice of transportation mode, with Age having the highest impact on Swissmetro and Income having the highest impact on Train.

Table 2: Average impact of input variables on mode choice using Integrated Gradient for ASU-DNN

Input Variables	Train	SM	Car
Train travel time	**-0.113**	0.087	0.026

Train cost	**-0.260**	0.193	0.066
SM travel time	0.139	**-0.348**	0.208
SM cost	0.180	**-0.420**	0.239
Car travel time	0.022	0.100	**-0.123**
Car cost	0.042	0.243	**-0.285**
Age	0.270	-0.422	0.152
Income	-0.186	0.123	0.062

The results presented in Table 3 offer insight into the average impact of input variables on the transportation modes of Train, Swissmetro, and Car for the fully connected DNN model. Similar to ASU-DNN, the impact of time and cost on their corresponding alternative is negative for the fully connected DNN, however in some cases their impact on other alternatives is also negative. For instance, Train travel time has a negative impact on the selection of both Train and Car, even though it is expected to only have a negative effect on Train. Additionally, when considering the absolute values of impacts, some alternative-specific variables do not have the highest impact on their corresponding alternative. As an example, Car travel cost has the highest impact on Swissmetro, not Car. These findings suggest that there may be some inconsistencies between the fully connected DNN models and behavioral processes which could have implications for their usage in choice modeling.

Table 3: Average impact of input variables on mode choice using Integrated Gradient for DNN

Input Variables	Train	SM	Car
Train travel time	**-0.800**	**0.816**	**-0.016**
Train cost	**-0.428**	0.201	0.226
SM travel time	0.563	**-0.867**	0.303
SM cost	0.891	**-1.191**	0.299
Car travel time	**0.383**	**-0.189**	**-0.194**
Car cost	**-0.194**	**0.708**	**-0.513**
Age	-0.081	-0.082	0.164
Income	-0.254	0.160	0.093

The comparison between Table 2 and Table 3 reveals insights into the effects of RUM when applied within the DNN architecture. In Table 2, using Integrated Gradient to interpret the ASU-DNN framework, our observations align with the expectations of RUM theory, as evidenced by variables like travel time and travel cost, which yield expected effects on their respective modes and alternative modes, reflecting the inherent trade-offs in user decision-making. However, Table 3 presents a contrast within the fully connected DNN model, with unexpected negative impacts on unrelated alternatives, challenging RUM assumptions. This significant difference between the ASU-DNN and fully connected DNN models shows the relationship between the modeling structure and RUM principles. These findings highlight the limitation of the fully connected DNN models and the strength of ASU-DNN in capturing behavioral insights in the mode choice datasets.

4. Conclusion

In this paper, we evaluated the performance of ASU-DNN and fully connected DNN models using the Integrated Gradient method. This study showed the ability of the Integrated Gradient method in explaining the contribution of variables on the decision outcome is highly dependent on the adopted DNN architecture. Although the fully connected DNN had high prediction accuracy and low likelihood during training, ASU-DNN showed superior performance in modeling unseen data, as evidenced by its higher prediction accuracy and lowest log-likelihood during validation, surpassing both DNN and MNL models. Our interpretation analysis revealed that compared to the fully connected DNN model, ASU-DNN, with its theory-based architecture, is more consistent with RUM theory. The results indicated that in ASU-DNN, travel time and cost have a negative impact on their corresponding alternative, while their impacts on other alternatives are positive. In contrast, the fully connected DNN model showed that travel time and cost have a negative impact on other alternatives additional to their corresponding alternative. Furthermore, the absolute impact of travel time and travel cost on their corresponding alternative is the highest in ASU-DNN, while DNN shows not such connection between input variables and outputs.

The findings of this study demonstrate the potential of post-hoc analysis techniques for gaining insights into DNN models in the context of discrete choice modeling. With the highlighted interpretability offered by post-hoc methods, DNN models with theory-based architecture emerge as powerful tools for modeling travel mode choice datasets. The ability of these new models to capture the behavioral relationships from the dataset, along with post-hoc analysis, suggests informed decisions with a higher degree of accuracy for critical choices, such as transport investments. Future studies can explore additional post-hoc analysis techniques to extract further information from DNN models.

5. References

ARRIETA, A. B., DÍAZ-RODRÍGUEZ, N., DEL SER, J., BENNETOT, A., TABIK, S., BARBADO, A., GARCÍA, S., GIL-LÓPEZ, S., MOLINA, D. & BENJAMINS, R. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 58, 82-115.

BIERLAIRE, M., AXHAUSEN, K. & ABAY, G. The acceptance of modal innovation: The case of Swissmetro. Swiss Transport Research Conference, 2001.

GOLSHANI, N., SHABANPOUR, R., MAHMOUDIFARD, S. M., DERRIBLE, S. & MOHAMMADIAN, A. 2018. Modeling travel mode and timing decisions: Comparison of artificial neural networks and copula-based joint model. Travel Behaviour Society, 10, 21-32.

GOODFELLOW, I., BENGIO, Y. & COURVILLE, A. 2016. Deep learning, MIT press.

HAGENAUER, J. & HELBICH, M. 2017. A comparative study of machine learning classifiers for modeling travel mode choice. Expert Systems with Applications, 78, 273-282.

HAN, Y., PEREIRA, F. C., BEN-AKIVA, M. & ZEGRAS, C. 2020. A neural-embedded choice model: Tastenet-mnl modeling taste heterogeneity with flexibility and interpretability. arXiv preprint arXiv:.00922.

LIPTON, Z. C. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 16, 31-57.

LUNDSTROM, D. D., HUANG, T. & RAZAVIYAYN, M. A rigorous study of integrated gradients method and extensions to internal neuron attributions. International Conference on Machine Learning, 2022. PMLR, 14485-14508.

SIFRINGER, B., LURKIN, V. & ALAHI, A. 2020. Enhancing discrete choice models with representation learning. Transportation Research Part B: Methodological, 140, 236-261.

SUNDARARAJAN, M., TALY, A. & YAN, Q. Axiomatic attribution for deep networks. International conference on machine learning, 2017. PMLR, 3319-3328.

THANH, T. T. M., LY, H.-B. & PHAM, B. T. A possibility of AI application on mode-choice prediction of transport users in Hanoi. CIGOS 2019, Innovation for Sustainable Infrastructure: Proceedings of the 5th International Conference on Geotechnics, Civil Engineering Works and Structures, 2019. Springer, 1179-1184.

TRAIN, K. E. 2009. Discrete choice methods with simulation, Cambridge university press.

WANG, S., MO, B. & ZHAO, J. 2020a. Deep neural networks for choice analysis: Architecture design with alternative-specific utility functions. Transportation Research Part C: Emerging Technologies, 112, 234-251.

WANG, S., WANG, Q. & ZHAO, J. 2020b. Deep neural networks for choice analysis: Extracting complete economic information for interpretation. Transportation Research Part C: Emerging Technologies, 118, 102701.

WONG, M. & FAROOQ, B. 2021. ResLogit: A residual neural network logit model for data-driven choice modelling. Transportation Research Part C: Emerging Technologies, 126, 103050.

ZHANG, Z., JI, C., WANG, Y. & YANG, Y. 2020. A customized deep neural network approach to investigate travel mode choice with interpretable utility information. Journal of Advanced Transportation, 2020, 1-11.