Enriching discrete choice models with computer vision for understanding choice behaviour in the presence of visual stimuli

Sander van Cranenburgh*[1]

[1] Department of Engineering Systems and Services,

Delft University of Technology, The Netherlands

SHORT SUMMARY

Discrete Choice Models (DCMs) are a key methodology in the transportation. However, current DCMs literally suffer from a blind spot: they cannot handle visual information. This blind spot hampers (1) a deeper understanding of human choice behaviour in the presence of visual stimuli and (2) using DCMs to deduce economic outputs for policies that involve changes to the visual environment. This study aims to bring visual information, in the form of images, to the realm of choice modelling. Specifically, it develops a series of discrete choice models –with computer vision parts embedded in different ways– to model the behaviour of decision makers when confronted with alternatives comprising both visual stimuli and conventional numeric attributes.

Keywords: Computer vision, Discrete choice modelling, Machine Learning.

1. INTRODUCTION

Since its inception in the 1970s, Discrete Choice Models (DCMs) have become a key methodology in the transportation, as well as in various other adjacent fields such as environmental economics, marketing and health economics (Hess and Daly 2014). However, current DCMs literally suffer from a blind spot: they cannot handle visual information. This blind spot hampers (1) a deeper understanding of human choice behaviour in the presence of visual stimuli and (2) using DCMs to deduce economic outputs for policies that involve changes to the visual environment. For instance, not seldom billions of euros are spend extra to integrate new transport infrastructure into a landscape in such a way that its visual character is preserved (e.g. using tunnels). Yet, such decision are often contested and postponed due to a lack of underpinning by reliable WTP estimates for visual improvements.

In the last decade major developments have taken place in Computer Vision (CV). State-of-the-art CV models are able to accurately detect scenes and objects in images (Gu et al. 2018). Nowadays, CV is used in a vast range of applications from detection faults in production processes to detecting tumours in MRI scans. Moreover, many pre-trained CV models are nowadays available. Such models have been trained on large amounts of data and can be further trained, using limited amounts of data, to conduct new related tasks. Thereby, the computational burden and the need for large amounts of data is substantially reduced. Very recently, applications of (pre-trained) CV models start to emerge in the travel behaviour and planning fields (e.g. Ramírez et al. 2021).

This research aims to further the use of computer vision models for choice behaviour modelling. Specifically, we aim to bring visual information, in the form of images, to the realm of choice modelling. We develop a series of discrete choice models with CV parts embedded in DCMs in different ways to model the behaviour of decision makers whom are confronted with alternatives comprising both visual stimuli and conventional numeric attributes. The developed models range from relatively straightforward to implement and apply to fairly challenging. In the relatively more simple models first features from the images are extracted using a pre-trained Convolution

Neural Network, after which a standard MNL model is estimated using these features plus the numeric attributes. In the more sophisticated models, we simultaneously train the CV part and the choice model. We compare the performance and model outcomes across the proposed models. We show that by embedding CV parts into DCMs relevant information for explaining choice behaviour can be extracted from images. In the absence of an existing data set consisting of choice tasks that comprise alternatives with images and numeric attributes, we use a synthetic data set to test our models. In this synthetic data set we emulate choice behaviour in which trade-offs are made between visual stimuli (aesthetic value) embedded in images and price level (a numeric attribute). We use images from the AVA data set (Murray et al. 2012). This data set consists of a plethora of images, ranging from artefacts to landscapes. All images in this data set are rated on their aesthetic value. As such, the AVA data set allow us to synthetically create choice data involving a trade-off between aesthetic value and a hypothetical price level associated with the alternative.

2. METHODOLOGY

2.1. Modelling framework

Most traditional CV models conceptually consists of two parts: a Convolution Neural Network (CNN), which extracts features from the image, and a (object) classifier (see Figure 1). The feature extractor typically comprises of a series of convolution layers which ultimately map the image onto a lower dimensional space, called the feature map. The object classifier, typically an Artificial Neural Network, (ANN) in turn classifies the image based on the extracted features.

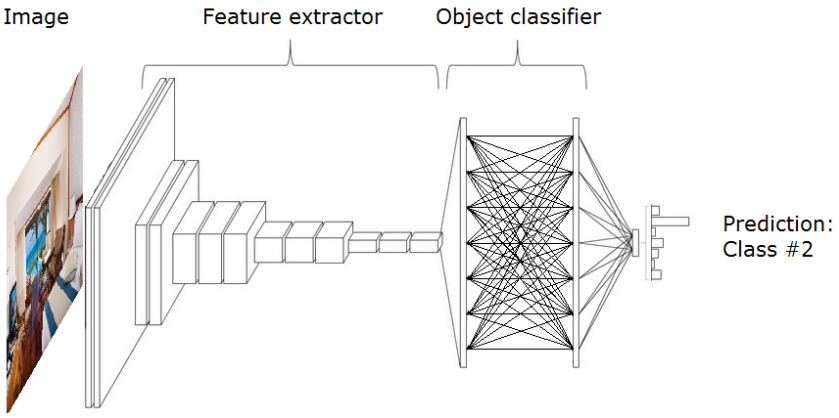


Figure 1: Typical CV model

In our modelling framework we treat the feature map extracted from a CNN as explanatory variables that enter the utility function of our DCMs, just like e.g. travel cost would. The observed part of utility V_i is assumed to take a linear-additive form, both for the utility associated with numeric attributes m as well as for the utility associated with the image's features k, see equation 1, where x_{ikn} denotes the kth feature for the image of alternative i, and w_k denotes the weight associated with feature k. Another possibility is to encode the feature space extracted from the CNN into another (lower dimensional space) feature space, using an e.g. an Artificial Neural Network, see equation 2. The rationale to do this is that the aesthetic value of an image might not only be a linear-additive function of the features, but could also be caused by particular interactions between features.

Figure 2 visualises of both approaches conceptually. The left-had side plot depicts the situation in which the feature space directly enters the utility function; the right-hand side plot depicts the

situation in which the feature space is first fed to an ANN, before it enters the utility function. Note that the utilities of the left and right alternatives are depicted in Figure 2 by the upper and lower rectangular blocks on the right-hand sides, respectively. Importantly, the upper and lower parts of the network are fully identical: both in architecture as well as in the networks' weights. In some ways, the proposed model architecture is very similar to Siamese architectures (Bromley et al. 1994), which are typically used for tasks like determination of whether, or not, two images belong to the same class (e.g. depict the Eiffel tower). In these networks typically a distance measure is computed between the features spaces of the images, where a low distance implies a large probability that both images belong to the same class. In contrast, in our models we do not compute the distance between features, rather we compute the total utilities.

$$U_{in} = \sum_m \beta_m x_{imn} + \sum_k w_k x_{ikn} + \varepsilon_{in} \tag{1}$$

$$U_{in} = \sum_m \beta_m x_{imn} + \sum_k w_k \tilde{x}_{ikn} + \varepsilon_{in} \tag{2}$$
$$where\ \tilde{x}_{ikn} = f\left(x_{ikn}\right)$$

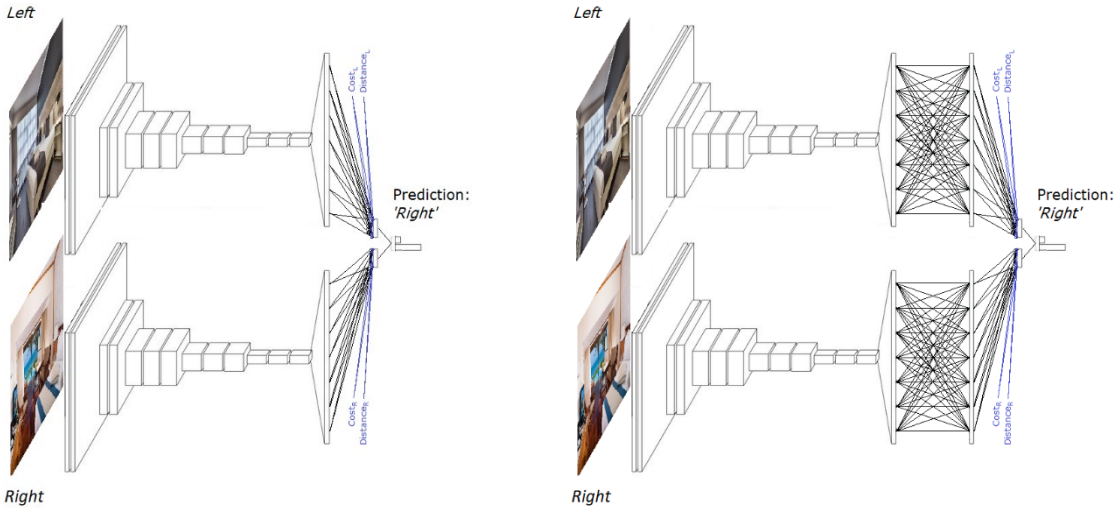


Figure 2: Visualisation of modelling framework (numeric attributes depicted in blue).

Importantly, upon recovering the model parameters (β,w) the weights w are not as interpretable as in fully theory-driven DCMs. Although the weights can then still be conceived as marginal utilities –after all they reflect the marginal effect on utility– their interpretation is hampered because the features x_{ikn} themselves do not carry a meaningful behavioural interpretation.

2.2. Computer vision enriched discrete choice models

Table 1 provides an overview of the models we estimate and train. Models 1 to 3 are conventional DCMs. Model 1 uses numeric attributes, but ignores information of the images. This model serves as a benchmark to assess the increase in explanatory power when accounting for the visual information. Model 2 uses the feature map, extracted from the images using a ResNet50 CNN, but ignores the numeric attributes. This model could provide a lower bound of the explanatory power embedded in the feature maps. As both models 1 and 2 are 'half-blind', in the sense that they do

not use all explanatory variables that matter for the choice behaviour, we expect these model to attain a low prediction performance. In model 3 the numeric and visual information are both used. The set-up of this model corresponds to the left-hand side plot of Figure 2. As this model has access to both the numeric and visual information we expect this model to outperform models 1 and 2.

Model 4 to 6 use a ANN classifier. That is, in these models the extracted features are first encoded onto another (lower dimensional) feature space using an ANN, before they enter the utility function (final layer). The set-up of this model corresponds to the right-hand side plot of Figure 2. The model variables used in models 4 to 6 are similar to those of models 1 to 3. In case nonlinearities and complex interactions between the features result in higher (or lower) aesthetic values, then we expect to see that models 5 and 6 outperform models 2 and 3. Conversely, when the aesthetic value is predominantly a linear-additive function of the feature map with no particularly strong interaction effects, then we will see that models 5 and 6 will perform on par with models 2 and 3.

In models 7 and 8 we jointly train the computer vision model and DCM. Model 7 uses a classic ResNet50 CNN architecture, while models 8 uses a so-called vision transformer architecture, specifically DieT base (Touvron et al. 2021). Nowadays, vision transformers are an increasingly popular approach in computer vision, outperforming CNNs models on various data sets. The ResNet50 of model 7 consumes 23.5m weights; the DieT_base of model 8 consumes 86m weights. For training these models we use the state-of-the-art AdamW algorithm (Loshchilov and Hutter 2017) with the following hyperparameter settings: a minibatch size of 16 for model 7 and of 8 for model 8, with a learning rate of 1e-5 and 3e-5. Comparing the performance of models 7 and 8 with the performance of the models in which only the DCMs and encoder are trained (i.e. models 1 to 6) will shed light on the extent to which the considerable extra efforts and computational complexity of jointly trained models are worth it.

Table 1: Model overview

Model	Model variables		Features	Computer vision	Training algorithm
	Numeric	Feature map	Extracted or Trained	architecture	
1	X		N/A	N/A	Quasi-Newton
2		X	Extracted	ResNet50	Quasi-Newton
3	X	X	Extracted	ResNet50	Quasi-Newton
4	X		N/A	N/A	SGD
5		X	Extracted	ResNet50	SGD
6	X	X	Extracted	ResNet50	SGD
7	X	X	Trained	ResNet50	AdamW
8	X	X	Trained	DieT_base	AdamW

3. Data

In the absence of a suitable existing data set – i.e. a data set that is large enough and consisting of choice tasks that comprise alternatives involving trade-offs between visual stimuli and numeric attributes, we create one ourselves. For this, we make use of the AVA data set (Murray et al. 2012). This data set consists of approximately 172k images. Each image is rated on a scale from one to ten, by on average 222 people (mostly amateur photographers) on its aesthetic value. We use these images to create binary choice tasks, in which each alternative consists of an image and a price tag.

However, not all images in the AVA data set are deemed suitable for use in this study. Specifically, we exclude images with a high standard deviation of their rating (std. dev >2.5). We do this to avoid overcomplicating training of the model, by including images with highly ambiguous ratings. To acquire a sense of the images in our data set and their ratings, Figure 3 shows a two

montages of images that are used in this study. The left-hand side plot shows a random selection of low rated images; the right-hand side plot shows a random selection of high rated images. Below the plots, the mean ratings are shown in heatmaps.

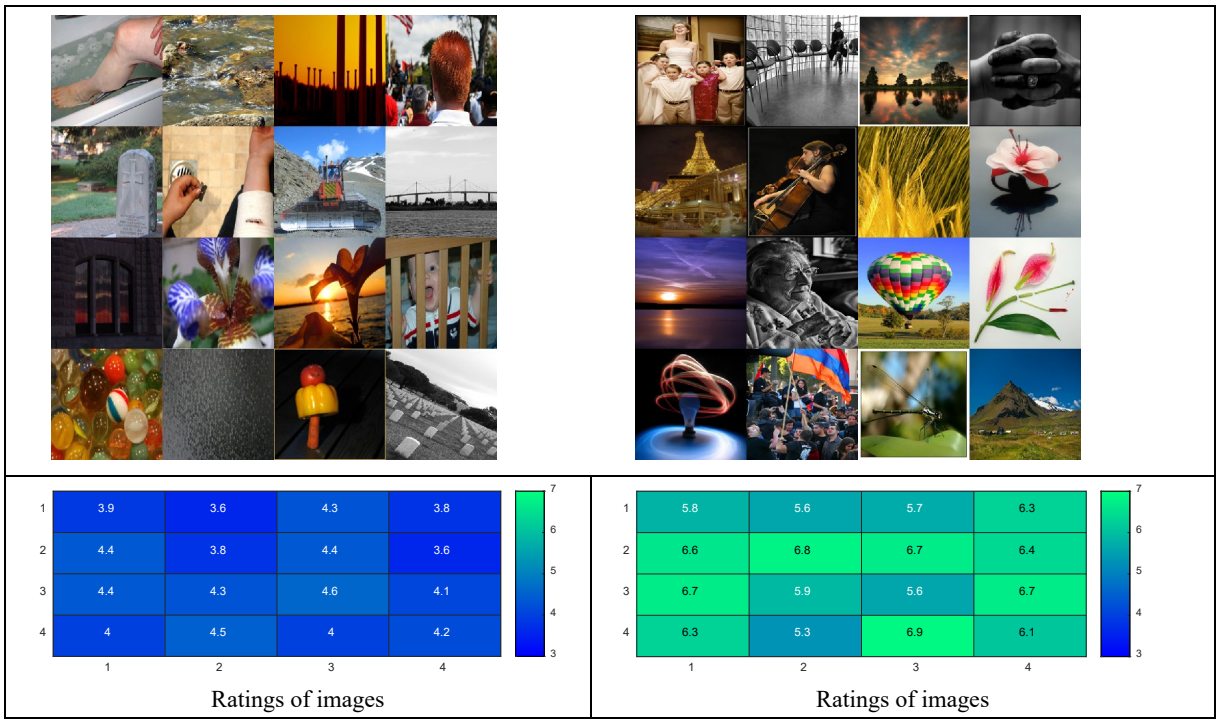


Figure 3: Illustration of a random set of low (left) and high (right) rated images

To construct our training and test data sets, first we split the available images in two subsets: one for creating the training data set, and one for creating the test data set. Next, we created choice tasks –for both the training and test data set– by randomly sampling two images. Each image is assigned a price tag which is randomly drawn between 0 and 10 euros. The complete training data set comprises 50k choice tasks, created using 26k unique admissible images. The test data set comprises of 10k choice tasks, created using the remaining 18k unique admissible images.

Figure 4 shows statistics on the distribution of the created training data set. Because the ratings are (approximately) normally distributed (left-hand side plot), the distribution of the difference in the rating between two alternatives is bimodal distributed (middle plot). Furthermore, as the price levels are drawn from a uniform distribution their differences are triangular distributed. The distribution of these differences is important as in discrete choice models only differences in utility matter.

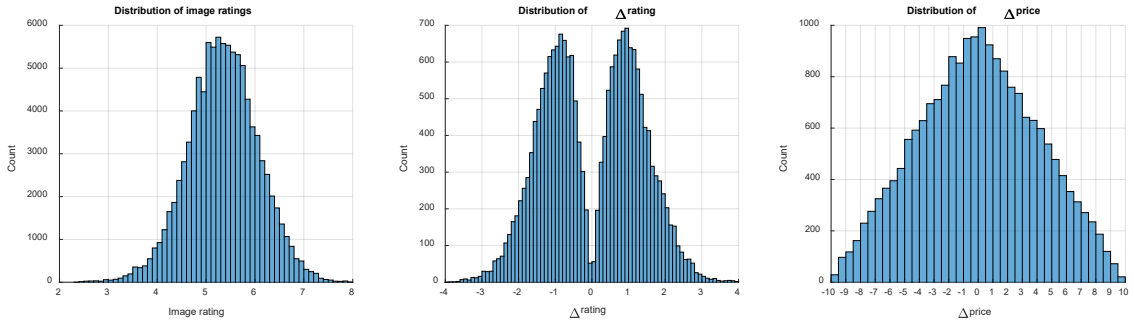


Figure 4: Statistic on the training data set

Our synthetic decision makers are assumed experience a positive utility from the aesthetic value of an image and a negative utility from its price. Equation 3 shows the linear-additive utility function:

$$U_{in} = \beta_{rating} \cdot R_{in} + \beta_{price} \cdot P_{in} \quad (3)$$
$$where \quad \beta_{rating} = 0.6, \ \beta_{price} = -0.2$$

where R_{in} denotes the rating of the image of alternative i and P_{in} denotes the price of alternative i. Hence, the image rating is taken as its aesthetic value. In line with our conceptual model, decision-makers are assumed to maximise utility (equation 4), y_{in} denotes the chosen alternative.

$$y_{in} = \begin{cases} 1 & if \quad U_{in} > U_{jn} \ \forall \ i \neq j \\ 0 & otherwise \end{cases} \quad (4)$$

Note that the choices are deterministically determined, as the utility function does not involve the usual error term ϵ (equation 3). It was deemed that the data generating process was sufficiently stochastic by itself due to the fact that the rating R is a stochastic variable. Figure 5 shows a randomly selected choice observation from the test data set. The aesthetic ratings of the left and right images are respectively, 7.36 and 3.99. Applying equation 3 yields the highest utility for alternative 1 (2.72 for alternative 1 against 1.79 for alternative 2). Hence, alternative 1 is the chosen alternative in this choice observation.

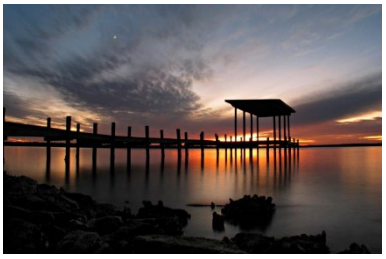
	Alternative 1	*Alternative 2*
		
Price	€8.5,-	€3
Choice	X	

Figure 5: Randomly selected observation from the training data set

4. PRELIMINARY RESULTS

Table 2 reports the estimation and training results. To assess the performance, we look at the rho square and its machine learning counterpart: the cross entropy. These performance metrics are computed by evaluating the models' performance on the test data set. Based on Table 2 a number of observations can be made. Firstly, the feature space holds explanatory power to explain choice behaviour. This can be inferred from the non-zero rho squares of model 2 and model 5. Secondly, the improved model performance the models that use both sources of information: numeric plus visual (models 3, 6, 7 and 8) over the models that do not (models 1, 2, 4 and 5) suggests visual information can embedded in choice models. These models seem to have been able to capture the trade-off between the numeric attributes and the visual stimuli. Thirdly, counter to our prior expectations, encoding the feature map using ANNs, as done in model 4 to 6, does not seem to meaningfully improve the model performance, compared to models 1 to 3. Fourthly, comparing

the computational time of model 3 and 6, we see that –in line with the findings by Lederrey et al. (2021)– for optimisation problems with a large number of parameters SGD is orders of magnitudes faster than the commonly used quasi-Newton based optimisation algorithms for estimating DCMs. Fifthly, counter to our expectations, we see that the models that jointly train the CV model and the DCM (model 7 and 8) hardly improve the model performance, as compared the sequential models. In case this result holds up after more extensive testing, it would lower the burden for choice models to use visual information into account in their choice models. After all, it does not require full-fledge training of computationally expensive CV models. However, it also raises questions regarding how much relevant information can be extracted from images for predicting choice behaviour. Finally, we do not find a meaningful difference between the performance of the CNN and Transformer based models.

Table 2: Estimation / training results

MODEL	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Model variables								
Cost	X		X	X		X	X	X
Latent features		X	X		X	X	X	X
Features extracted or trained		Extracted	Extracted		Extracted	Extracted	Jointly trained	Jointly trained
Computer vision model architecture		ResNet50	ResNet50		ResNet50	ResNet50	ResNet50	DieT_tiny
Training algorithm	quasi-newton	quasi-newton	quasi-newton	SGD	SGD	SGD	Adam	Adam
No. parameters/weights trained	1	2048	2049	53	33k	33k	23.5m	86m
Cross-entropy (out-of-sample)	0.495	0.649	0.398	0.489	0.645	0.417	0.407	0.408
rho^2 (out-of-sample)	0.29	0.06	0.43	0.29	0.07	0.40	0.41	0.41
Computation time	8 sec[I]	6 h[I]	6 h[I]	2 sec[II]	25 sec[II]	25 sec[II]	4 h[II]	4 h[II]

[I] Using 4 CPUs (Xeon @ 3.60 GHz)

[II] Using GPUs (GeForce RTX 2080Ti)

5. CONCLUSIONS

This research has proposed a method to bring visual information into the realm of discrete choice modelling. In our view, it is a first step towards choice models that would allow choice models to work with visual data. There are plenty avenues for further research. The next step is to collect empirical data involving trade-offs between numeric attributes and visual stimuli (images). This can be done in a variety of contexts. We plan to administer a Stated Choice (SC) experiments in which respondents are presented a city park decision, and need to make a trade-off between travel time to the park and the park's aesthetics – as captured in photos of the parks. Additionally, being able to build CV enriched discrete choice model is only the first step. The next steps is to find ways to extract meaningful behavioural insights from such models, such as Willingness-to-Pay estimates for e.g. city park improvements. This could be done by building forth on the swath of explainable AI techniques currently being developed in the machine learning field.

ACKNOWLEDGMENT

This work is supported by the TU Delft AI Labs programme.

REFERENCES

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E. & Shah, R. (1994). Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J. & Chen, T. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, 77, 354-377.

Hess, S. & Daly, A. (2014). *Handbook of choice modelling*: Edward Elgar Publishing).

Lederrey, G., Lurkin, V., Hillel, T. & Bierlaire, M. (2021). Estimation of discrete choice models with hybrid stochastic adaptive batch size algorithms. *Journal of Choice Modelling*, 100226.

Loshchilov, I. & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Murray, N., Marchesotti, L. & Perronnin, F. (2012). AVA: A large-scale database for aesthetic visual analysis. 2012 *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE.

Ramírez, T., Hurtubia, R., Lobel, H. & Rossetti, T. (2021). Measuring heterogeneous perception of urban space with massive data and machine learning: An application to safety. *Landscape and Urban Planning*, 208, 104002.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. *International Conference on Machine Learning*, PMLR.