Predicting First Interaction Decision on Mobile Interfaces Using a Deep Convolutional Network

Sami Nieminen
sami.nieminen@aalto.fi
Aalto University
Espoo, Finland

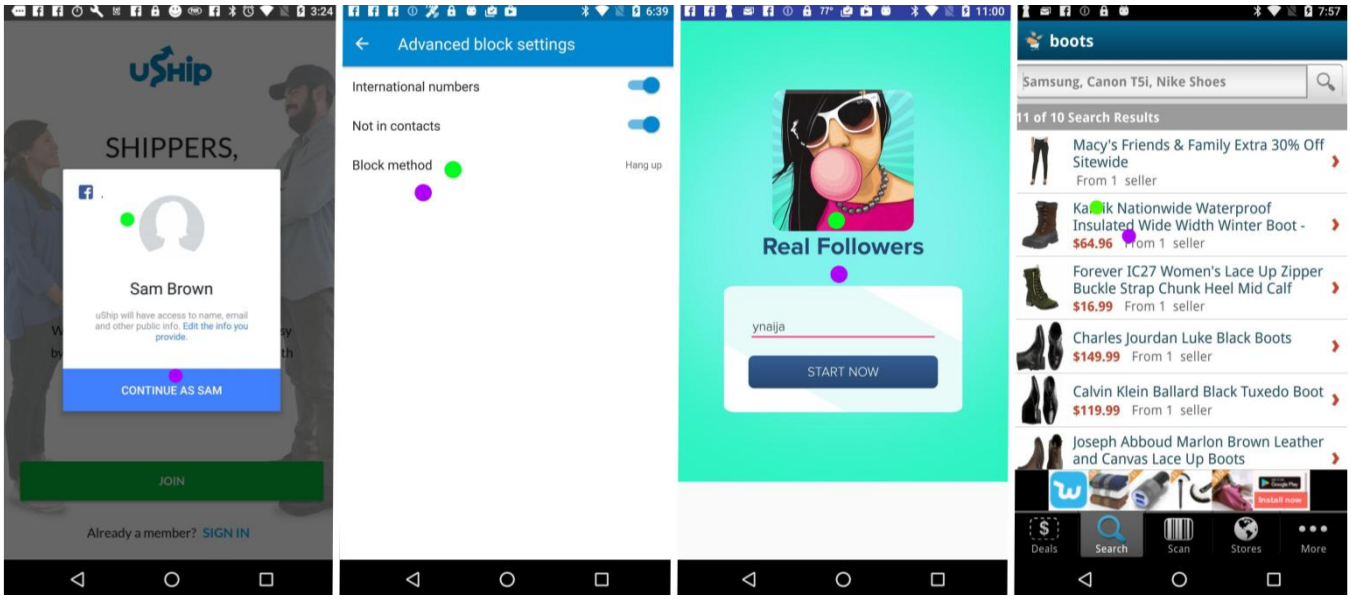

Figure 1: Four randomly chosen example UIs containing touch interaction prediction as purple dot and actual touch interaction in green. From left to right the screenshots have Rico ids 5956, 39293, 7627 and 41937.

ABSTRACT

An interaction decision prediction can be generated from a model trained to learn factors affecting user decisions to interact with an interface. The predictive models can be trained for various interfaces where interactions using various hand gestures such as swipes, taps and clicks may be performed. The decision to perform an interaction with the interface can be affected by factors related to human decision making process such as factors affecting attention like saliency and expectation of a reward. Data-driven models have been developed to predict component based mobile user interface saliency and click-sequence predictions. With components we refer to items visible on the screen such as buttons and menu boxes. However, the overall process of first interaction decision prediction especially in touch interaction settings for mobile interfaces has not been studied. A deep convolutional neural network model trained to minimize euclidean distance between a known touch point on the screen and the prediction converges towards the correct touch area highlighting the effect of attention and interface structure on human decision making process. This work expands on the possible study area of decision making factors like attention on mobile interfaces as previously saliency models have been trained to study attention on mobile interfaces. The results help form a link between the human decision making process and mobile interfaces. The work provides a starting point for studying decision making factors on mobile interfaces from both bottom-up and top-down perspectives. Additionally the work contributes towards improving user experience and enhancing application business performance.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ELEC-E7861 '21, June 03–05, 2021, Espoo, Finland
© 2021 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

CCS CONCEPTS

• **Human-computer interaction**; • **Computational modeling**; • **Mobile interfaces**;

KEYWORDS

convolutional network, data-driven modeling, decision making, deep learning, interaction modeling, mobile interface

ACM Reference Format:
Sami Nieminen. 2021. Predicting First Interaction Decision on Mobile Interfaces Using a Deep Convolutional Network. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

When a human is starting to interact with a computer, interface or software it is not immediately clear how the human is going to choose to interact with it. This is especially unclear during the first interaction decision when previous interaction data may not be available, or the user may not be known yet. We define the first interaction decision as the outcome of when a human has made the decision to interact with an interface and performed the intended interaction such as a tap or a swipe. What affects possible modes of interaction are the available types of pointing devices and interfaces such as a touchscreen on a mobile device. On a touchscreen it is possible to perform various hand gestures such as taps, pinches and swipes [4, 21] to interact with a selected component visible on the interface. With components we refer to areas on the interface that contain multiple elements such as a pop-up window or a slider [1]. Components form another constraint on possible human interaction choices as interacting with a button for example is primarily done via taps and similar gestures on a touchscreen. Using data collected from possible interaction choices on interfaces it is possible to predict where the user will perform their next interaction or where they choose to point gaze as they direct visual attention on the screen [6, 8, 9, 11, 12, 15, 21].

During human interaction with interfaces such as mobile applications it is helpful to know how the user will experience the interface, what they find interesting about it or how it links to their goals. How a user directs attention through selectivity in perception [13] links to how the interface is experienced and what goals they have in mind. The targeting process of visual attention is controlled by directing the eyes to make a particular scene in the environment visible with fovea of the retina displaying a high-resolution central area for detailed processing. When considering attention it affects how humans consider making choices during the decision making process [13]. As visual attention is directed through pointing eyes to display a particular area on an interface, investigating linkages between first interaction decision, interface structure, visual attention process and decision making is possible. Orquin et al. performed a review of studies investigating eye movements in decision making. The identified categories for attention included stimulus-driven attention, saliency, goal-oriented attention and lastly an intersection between working memory and attention [13].

Attention has been studied for mobile interfaces from the perspectives of visual saliency [6, 11] and touch saliency [12, 21]. Saliency characterizes parts of a scene that stand out relative to the surrounding area [2, 13]. Xu et al. [21] had introduced touch saliency as an alternative to visual saliency for natural images with a follow-up by [12] proposing a model using both touch saliency and visual saliency to improve saliency prediction performance. Gupta et al. demonstrated that specifically training a deep learning autoencoder for predicting saliency on mobile user interfaces performs better than models trained on natural images [6]. Leiva et al. demonstrated that data-driven models to perform better than classic parameter-free saliency models on an annotated mobile UI data set of 193 mobile UIs. In addition, the results highlight the role of expectations when users choose where to look especially towards the top left corner. [11] It has been shown for mobile UIs that placing task targets in high saliency areas on mobile UIs can improve task completion times [19]. Lastly interaction choices on mobile interfaces have been studied based on historical usage data and recent choice sequences to gain insight into what the user might do next so a marketing intervention may be performed [8, 15]. Previous works have not studied the linkage between what actually led to the decision to physically interact with a particular area of an interface. Additionally, to the best of the authors' knowledge no attempt has been made at trying to predict the outcome of the first interaction decision and what affected the decision.

We propose a deep convolutional neural network (CNN) model to predict the first interaction decision for mobile UIs based on touch data containing taps and swipes. The touch data has been obtained from Rico dataset [4] while the UIS are from Enrico [10] which is a collection of curated Rico interfaces. The CNN model input data is essentially the same as the data Leiva et al. [10] used for classifying interfaces into 20 UI design topics. However, the prediction generated by our model is markedly different as we are predicting two-dimensional Euclidean distances measured in pixels on the interfaces. The Euclidean distance is measured as the difference between the actual tap or first point of swipe and the predicted point of first interaction. Consequentially, our model is a regression problem rather than a classification problem. The predicted touch point and actual touch point are also classified into components on the UI in order to study if the actual and predicted touch points are within the same component.

This work shows that it is possible to predict interaction decisions on mobile interfaces in the form of touches as the regression problem learned by the CNN model converges towards the actual touch points. Additionally, the model had a better classification performance than randomly guessing a component on the interface. This work opens up possibilities to study and understand human interaction choices on interfaces. Understanding human choices to interact with interfaces contributes to possibilities for improving user experience and enhancing the business performance of various interfaces. From research perspective we now know that it should be possible to form a linkage between human interaction decisions, attention and the final interaction outcomes.

2 RELATED WORK

2.1 Rico Dataset

Rico is a dataset comprising of over 72,000 mobile UI screenshots and related touch interaction traces. The screenshots comprise approximately 9,700 mobile applications available from Android store. The interaction traces can be subdivided into touch interaction and swipe interaction traces. Both types of interaction traces measure the first interaction and then cut off. Rico also contains view hierarchies and Android app store metadata for example. The data has been obtained through both human exploration and automated exploration. The work also included training a deep autoencoder to learn a 64 dimensional representation of each UI layout. [4] Given

the large amount of interaction traces and UI layouts the data can be used to study user interaction behaviors and predicting UI topics based on annotations for example.

2.2 Enrico, Rico and Modeling the Data

Leiva, Hota and Oulasvirta [10] investigated topic modeling based on Rico dataset and formed enhanced Rico (Enrico) dataset. The dataset is generated from investigating 10k UIs selected randomly resulting to a 1460 high quality UI dataset comprising 20 design topics such as settings, login and gallery. The originally selected 10k UIs were revised and bad examples discarded using a web-based interface displaying a screenshot of the UI side-by-side with a semantic wireframe. The screenshots were also annotated using an annotation interface with topic categories available for the screenshots. As a demonstration of the dataset applications a deep convolutional autoencoder was trained to perform topic classification. The best classification accuracy was achieved with UI screenshots as opposed to wireframes and embeddings. [10] In the work an open question was discovered regarding the best algorithm to discover the latent space for Enrico in 2D. When UMAP algorithm was used the categories in the latent space were quite mixed up.

2.3 Touch Saliency and Finger Touch Data

Xu et al. [21] followed up and expanded by Ni et al. [12] developed and researched the concept of touch saliency for mobile interfaces. The data set was formed from NUSEF dataset containing natural images. The data comprised two parts where eye-tracking data was collected and additionally finger touch data for zooming in on parts of the images was collected. From the data both touch saliency maps and visual saliency maps were obtained. Xu et al. tested algorithms such as IT, AIM and ICL with AUC and CC metrics for evaluative capability on both visual saliency and touch saliency. The saliences were found to have comparable evaluative capabilities but touch saliency was more sensitive to object shape. Ni et al. tested multiple saliency prediction models on the data such as context-aware based saliency detection (CSD) and multi-task sparsity pursuit (MTSP). The metrics used included ROC, AUC and CC. In particular MTSP and MTSP-Mid were found to performed well and especially on touch ground truth.

2.4 Decision Making and Attention

Attention and decision making are affected by saliency, value and reward. In particular value is increased positively or negatively through increased magnitude and probability of a reward. Saliency can be increased by both positive and negative value when considering learned saliency through association. Additionally saliency can mean visual saliency which is affected by physical properties like color. [7]

2.5 Decision Making on Interfaces

Decision making processes on interfaces can be modeled as an emergent property of partially observable Markov decision processes (POMDP). In POMDP approach a problem space is defined that represents the interaction between an agent and a partially observable stochastic environment. A POMDP definition by itself is not sufficient for understanding how a human chooses to interact with for example a side-menu or a full interface. Rather a specific solution to the POMDP model has to be learned. [14] Additionally Todi et al. demonstrated the application of a reinforcement learning algorithm with MDP problem space definition for adaptive side-menus [20]. Interestingly an alternative approach to study the problem of the paper is to model it from MDP perspective and apply a reinforcement learning model to the MDP problem space representation.

Table 1: Screen complexities

Metric	all	train	test
Screens	1338	1168	170
Max	52	52	26
Min	1	1	1
Mean	9.41	9.48	8.96
Std	6.49	5.93	6.57

3 DATA ANALYSIS

Analysis of Enrico interface components for understanding model predictive power compared to random choices. Table 1 contains measures of screen complexities for the whole dataset, train set and test set. Max and min values mean the minimum number of components found for some screen. In particular the mean value helps understand how well on average the model predicts the correct component compared to randomly choosing one component on the screen as the correct one. The number of components on the screen is defined as the number of first level components which may have child components within their bounds.

Figure 2 depicts the most complicated screens found on both the training and test sets by maximum first level components measure. The dataset contains positive skew with a longer tail to the right as seen in figure 3. The skew does affect interpreting how much better the model is at choosing correct component compared to random guesses as right tail brings up the average first level component count. Due to the positive skew it is interesting to also ascertain separately that the model performs better than randomly guessing for screens with first component counts in the 2 to 6 range for example.

4 COMPUTATIONAL MODELING APPROACH

We approach the touch decision prediction problem using data-driven computational modeling.

With mapping from Enrico to Rico touch traces we are able to select interaction traces that will more likely match to elements on screen annotations more accurately as Enrico annotation scheme has already been validated by humans. Additionally the dataset is large enough to attempt predicting touch decision making with a deep learning approach. This approach also allows us to test if touch decisions are more biased towards some of the interactable interface elements which we call element interaction bias. Element interaction bias can be measured by comparing the ratios of touched interactable elements to total count of the interactable element. If there is no bias then it would be expected that interactable elements

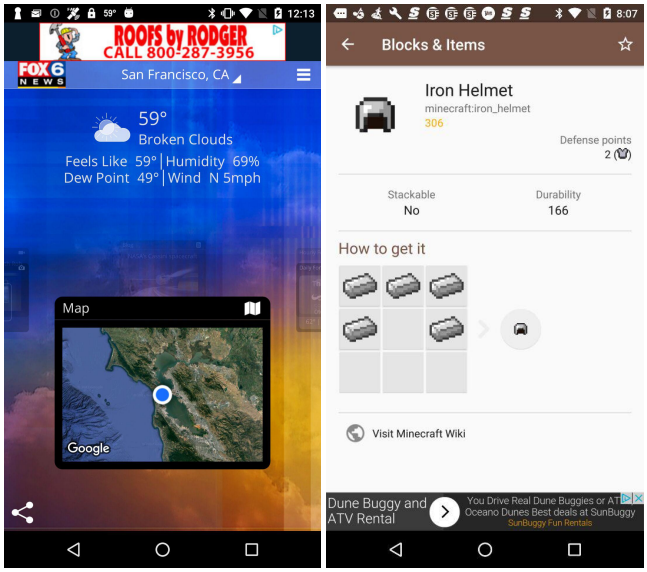

Figure 2: Screenshot id 49581 from training set and screenshot id 69867 from test sets with the highest numbers of first level components

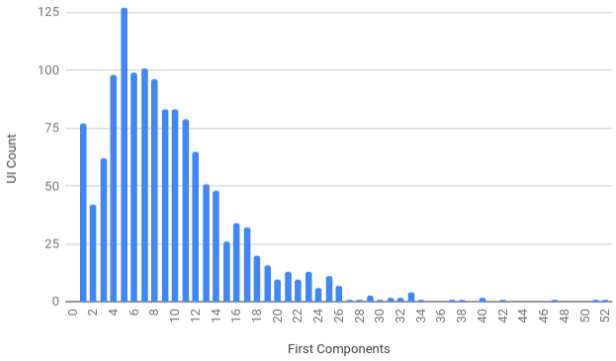

Figure 3: First level component distribution of Enrico dataset.

are touched with amounts corresponding to their count of the total interactable element population.

Convolutional neural networks (CNNs) are suitable for learning from data with an already known grid-like topology such as images [5] and also demonstrated by [10, 18].

4.1 Model Implementation

We implement a series of simple convolutional networks to study where we can obtain good regression convergence for tap prediction as the problem is different from other studies in the area as typical standard is based on classification approach. The input layer will be kept constant with dimension sizes 360, 640 and 3. For expanding the CNN architecture we will be expanding the architecture using

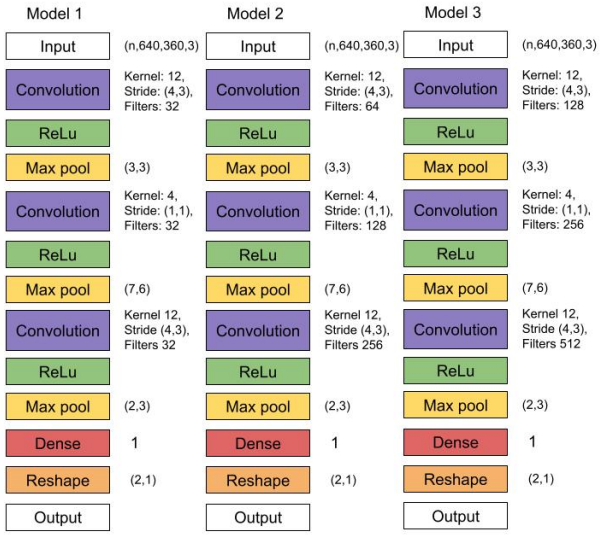

Figure 4: Designs of implemented CNN network architectures with convolution and max pooling for dimensionality reduction. The models are the same but employ a different number of filters.

approaches presented by [5, 18]. Figure 3 displays CNN architectures to be tested first. As an optimizer we will be using Adam due to its robustness [5].

A typical layer in a CNN consists of three parts, namely convolution, activation and pooling functions which we discuss briefly below. Convolution operation applies a linear operation using a kernel with weights over an input grid such as a 2D image. In machine learning discrete convolution is typically performed without kernel flipping which is similar to cross-correlation and offers a measure of similarity when one function, such as the kernel, is moved across the other function over a distance [16]. Discrete cross-correlation is defined as

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i+m, j+n) K(m,n) \quad (1)$$

Analysis of Enrico interface components for understanding model

where I is a 2D tensor with dimensions i and j. The kernel K has dimensions m and n. [5] In addition the convolution operation can have stride which offers a way to downsample the dimensions by skipping some of the column and row combinations in the input. A commonly used activation function is Rectified linear unit (ReLU) defined as

$$g(z) = max(0, z), \quad (2)$$

where any input values below zero are converted to zero [5].

Pooling is an operation where an output value is obtained from a rectangular neighborhood. In particular, max pooling is common in convolutional layers where the maximum value from the neighborhood is selected [5].

Figure 4 depicts the currently applied network architectures. A special feature of the architectures is the input being non-square

which requires down-sampling the y-dimension faster for easier convergence towards (2,1,1) output shape after dense layer.

- model 1 employs a very sparse model by adding strides, larger kernel sizes and few filters.
- model 2 is the same as model 1 but employs a larger number of filters.
- model 3 has a yet larger amount of filters but is otherwise the same as models 1 and 2.

4.2 Training and testing

We use a 80 percent training to 20 percent test split with random selection. Between the reported model results the training and test data is kept the same for better comparability. As we are predicting human touch choice, in particular taps on discrete space bounded screen elements we want to loss optimize based on a normalized distance metric. The norm

$$L^1 = ||x||_1 = \sum_i |x_i| \tag{3}$$

is commonly used in machine learning to help distinguish values close to zero. However, as our purpose is to ensure fitting predicted tap points inside correct components, weighting predictions most likely outside the correct component is desirable. As a consequence our optimization metric will be the norm

$$L^2 = ||x||_2 = \sqrt{\sum_i |x_i|^2} \tag{4}$$

as the norm increases slowly near the ground-truth tap point. [5] The final loss function is

$$L(\hat{y}, y) = \sqrt{\sum_i |\hat{y_i} - y_i|^2}, \tag{5}$$

where $\hat{y}$ is the predicted point and y is the actual point.

4.3 Evaluation metrics

In order to evaluate if the predicted coordinates were within the right elements the true tap coordinates are mapped to element bounds extracted from view hierarchy. The predicted taps are binary classified according to the classification function

$$C(x_1, x_2, X, Y) = \tag{6}$$

$$X_{min} <= x <= X_{max} \cap Y_{min} <= x_2 <= Y_{max}, C \in \{0, 1\}, \tag{7}$$

where Y and X contain minimum and maximum boundaries for an UI element. x_i are the predicted touch coordinates. In equation 6. 0 indicates false positive so the element is wrong and 1 indicating true positive so the tap exists within the desired element. From here we obtain precision metric

$$Pr = \frac{TP}{TP + FP}, \tag{8}$$

where TP is number of true positives, and FP is number of false positives [17].

Additionally, we are interested in understanding how close the predicted taps go to the actual taps in, we are also employing mean squared error [5] modified for a two-dimensional grid

$$MSE = \frac{1}{m} \sum_m L(\hat{y_m}, y_m)^2, \tag{9}$$

where $\hat{y_m}$ is the predicted value, y_m is the actual value and m is the number of predictions. We may additionally consider estimating separate errors for x and y dimensions to understand if one dimension is overrepresented which may for example highlight issues with model implementation.

5 RESULTS

We were able to identify a clickable component ratio of 0.34 for the correct component. This could be either due to current component selection algorithm design or due to inconsistencies with the way a touchable element is classified in Rico and Enrico datasets. Table 2 depicts prediction accuracy for finding the correct touch component for the rudimentary model. Figure 5 shows two randomly chosen interaction predictions.

Table 2: Prediction results by model

Model	Accuracy	Loss
Model 1	0.268	0.0796
Model 2	0.299	0.670
Model 3	0.268	0.012

Based on the initial results from first three models and some additional ones that were tried ad-hoc it seems like developing the model further will require an alternative approach than adding more filters. From algorithmic perspective developing a model with one to several more deep layers and going up to 512 filters as in model 3 could yield a better prediction. Model 3 had the best loss which did not result to more classification accuracy which could be an issue with the depth compared to filter count. The convergence for the loss measure indicates that the models have the ability to learn where the touch might occur and it can be expected that it is possible to obtain a better convergence without overfitting. In particular it seems like model 3 overfitted on the training data as the better loss measure did not lead to a better accuracy for prediction.

For classification performance from pure probabilistic perspective the model performs with similar accuracy to random guessing for interfaces with 3 first-level UI componenents and worse for those with less than 3 components. However, we expect that in reality the models perform well for first component counts in the range of 1 to 7 and worse for larger first-level component counts at the right-end skew of figure 3 distribution. This needs to be verified during future work which will be discussed below. Figure 6 depicts an example of an UI with 10 first-level components where we suspect the predictive power may be worse.

6 DISCUSSION

This work presented the first study on decision making for first touches on mobile interfaces using data-driven modeling with deep convolutional networks. What the results have shown is that human choice to make an interaction decision on a mobile interface

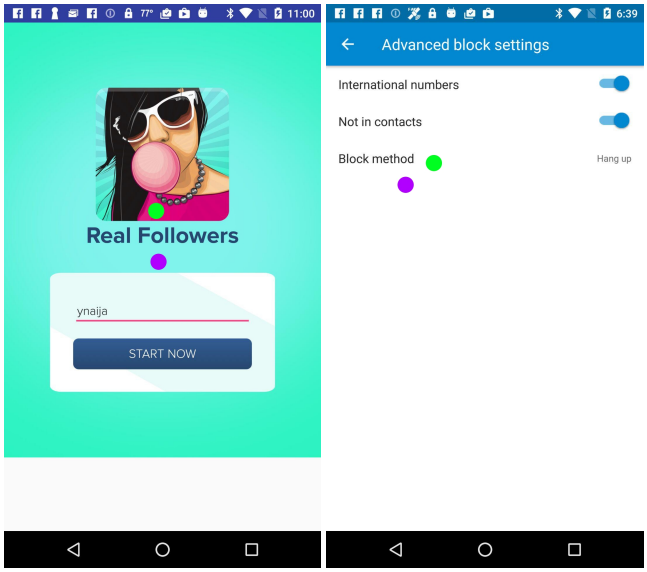

Figure 5: Two randomly chosen predictions with purple indicating prediction spot and green indicating actual touch point. The Rico ids for the screenshots are 9293 and 7627.

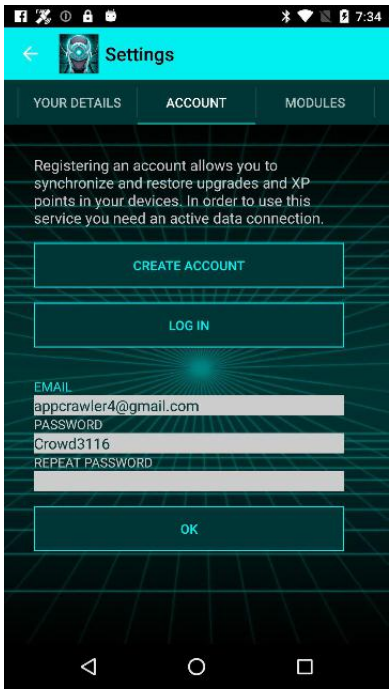

Figure 6: Screenshot id 514 as an example of an interface with a more complicated component structure.

can be partially explained by the visual composition of the interface . When we consider the number of touchable items on the screenshots, then the models might have already captured a fundamental section or sections of human decision making process. This work also contributes a new perspective to study interfaces and interaction from HCI perspective as most comparable work with CNN models has focused on classification problems rather than regression problems.

6.1 Limitations and Future Work

The most pressing issue with the model accuracies can be inferred from observing figure 5 prediction results. What we see is that the predictions made by model 2 fell outside where the wireframe boxes would be for the correct clicks. Furthermore, it can be noted that the predictions were converging towards the correct location from Euclidean distance perspective. This means that the model did not learn to infer wireframe-like shapes and assign weights to them in a manner that would emphasize placing the prediction points within the boundaries of the inferred wireframes. This indicates that the convolution kernel sizes should be optimized to obtain a better predictive capability. Especially the convolution filters from figure 4 with kernel sizes of 12 might be problematic. An competitive alternative to changing the kernel sizes could be changing the input to contain both screenshots and wireframes at the same time. This requires very little work and does not change complexity considerably as Enrico and Rico already contain the wireframes [4, 10] This automatically ensures that the wireframe structures will be present. This seems sensible especially when we consider the right screenshot in figure 5 where the CNN learning edges approximating the wireframe may be challenging.

The classification and regression accuracy could also be improved by using a well-known CNN approach such as VGG [18] and replacing the VGG classifier with an optimizer for mean squared error instead. Another factor that was missing from the models that could have had a positive effect on predictive performance is using dropout [5] which helps prevent overfitting. In particular overfitting was observed with model 3 results. An interesting alternative approach to the work includes applying a reinforcement learning algorithm to learn a solution to the problem space represented by a POMDP where [3], [20] and [14] may offer a promising starting point. As saliency is a known factor affecting decision making process [7] a data-driven approach could be to generate visual saliency maps from eye-tracking data or predictions with an existing algorithm and study how saliency and eye-tracking results affect predictive performance. As an example the model generated by Leiva et al. [11] could be applied.

When it comes to studying the linkage between decision making, attention and final interaction result we are currently limited by the available datasets although the interaction choices measured in Rico already function as a proxy for the end result of the decision choice. This does not however account for what happens during the time when a human is observing the interface and and considering which interaction choice to make and rather measures only the abstracted end result of that process. Factors identified by Orguin et al. [13] such as goals for example are not well visible from current research. With the current data and model there is however a good basis for studying how interface structure affects the interaction decisions and if biases identified by Leiva et al. [11] can be linked to the interaction results.

7 CONCLUSION

In this work we have shown that it is possible to predict first interaction decisions made by humans when they are using mobile applications. Secondly, we have demonstrated a methodology for studying HCI problems with deep convolutional neural networks applied to regression problems as studying classification problems has been a more common theme. We also suspect that other problems and themes in HCI that can be studied with machine learning could be approached as CNN regression problems to obtain new perspectives. Generally speaking any problem where a human must perform an operation over some distance could be similarly studied as a regression problem that is still convertible to a classification type result as well. Additionally, it is interesting to consider how interactions with an interface such as a mobile UI will link to neural activity when a person is considering which decision to make as they direct attention to the interface.

REFERENCES

[1] 2021. *User Interface Elements.* Retrieved May 15, 2021 from https://www.usability.gov/how-to-and-tools/methods/user-interface-elements.html

[2] A. Borji and L. Itti. 2013. State-of-the-Art in Visual Attention Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 1 (2013), 185–207. https://doi.org/10.1109/TPAMI.2012.89

[3] Xiuli Chen, Gilles Bailly, Duncan P. Brumby, Antti Oulasvirta, and Andrew Howes. 2015. The Emergence of Interactive Behavior: A Model of Rational Menu Search. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 4217–4226. https://doi.org/10.1145/2702123.2702483

[4] Biplab Deka, Zifeng Huang, Chad Franzen, Joshua Hibschman, Daniel Afergan, Yang Li, Jeffrey Nichols, and Ranjitha Kumar. 2017. Rico: A Mobile App Dataset for Building Data-Driven Design Applications. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada) *(UIST '17)*. Association for Computing Machinery, New York, NY, USA, 845–854. https://doi.org/10.1145/3126594.3126651

[5] I. Goodfellow, Y. Bengio, and A. Courville. 2016. *Deep Learning.* MIT Press. http://www.deeplearningbook.org

[6] P. Gupta, S. Gupta, A. Jayagopal, S. Pal, and R. Sinha. 2018. Saliency Prediction for Mobile User Interfaces. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1529–1538. https://doi.org/10.1109/WACV.2018.00171

[7] T. Kahnt and P.N. Tobler. 2017. Chapter 9 - Reward, Value, and Salience. In *Decision Neuroscience*, Jean-Claude Dreher and Léon Tremblay (Eds.). Academic Press, San Diego, 109–120. https://doi.org/10.1016/B978-0-12-805308-9.00009-9

[8] Dennis Koehn, Stefan Lessmann, and Markus Schaal. 2020. Predicting online shopping behaviour from clickstream data using deep learning. *Expert Systems with Applications* 150 (2020), 113342. https://doi.org/10.1016/j.eswa.2020.113342

[9] S. Lee, R. Ha, and H. Cha. 2019. Click Sequence Prediction in Android Mobile Applications. *IEEE Transactions on Human-Machine Systems* 49, 3 (2019), 278–289. https://doi.org/10.1109/THMS.2018.2868806

[10] Luis A. Leiva, Asutosh Hota, and Antti Oulasvirta. 2020. Enrico: A High-quality Dataset for Topic Modeling of Mobile UI Designs. In *Proceedings of the 22nd International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct (MobileHCI'20)*. https://doi.org/10.1145/3406324.3410710

[11] Luis A. Leiva, Yunfei Xue, Avya Bansal, Hamed R. Tavakoli, Tuğçe Köroğlu, Jingzhou Du, Niraj R. Dayama, and Antti Oulasvirta. 2020. Understanding Visual Saliency in Mobile User Interfaces. In *Proc. MobileHCI*. https://doi.org/10.1145/3379503.3403557

[12] B. Ni, M. Xu, T. V. Nguyen, M. Wang, C. Lang, Z. Huang, and S. Yan. 2014. Touch Saliency: Characteristics and Prediction. *IEEE Transactions on Multimedia* 16, 6 (2014), 1779–1791. https://doi.org/10.1109/TMM.2014.2329275

[13] Jacob L. Orquin and Simone Mueller Loose. 2013. Attention and choice: a review on eye movements in decision making. *Acta psychologica* 144 1 (2013), 190–206.

[14] Antti Oulasvirta, Per Ola Kristensson, Xiaojun Bi, and Andrew Howes. 2018. *Computational Interaction.* Oxford University Press. http://www.oup.com/academic/product/9780198799610

[15] C. Pu, Z. Wu, H. Chen, K. Xu, and J. Cao. 2018. A Sequential Recommendation for Mobile Apps: What Will User Click Next App?. In *2018 IEEE International Conference on Web Services (ICWS)*. 243–248. https://doi.org/10.1109/ICWS.2018.00038

[16] K. F. Riley, M. P. Hobson, and S. J. Bence. 2006. *Mathematical Methods for Physics and Engineering: A Comprehensive Guide* (3 ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511810763

[17] Simon Rogers and Mark Girolami. 2016. *A First Course in Machine Learning, Second Edition* (2nd ed.). Chapman Hall/CRC.

[18] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs.CV]

[19] Jeremiah Still, John Hicks, and Ashley Cain. 2020. Examining the Influence of Saliency in Mobile Interface Displays. *AIS Transactions on Human-Computer Interaction* (03 2020), 28–44. https://doi.org/10.17705/1thci.00127

[20] Kashyap Todi, Gilles Bailly, Luis Leiva, and Antti Oulasvirta. 2021. Adapting User Interfaces with Model-based Reinforcement Learning. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM. https://kashyaptodi.com/adaptive/

[21] Mengdi Xu, Bingbing Ni, Jian Dong, Zhongyang Huang, Meng Wang, and Shuicheng Yan. 2012. Touch Saliency. In *Proceedings of the 20th ACM International Conference on Multimedia* (Nara, Japan) *(MM '12)*. Association for Computing Machinery, New York, NY, USA, 1041–1044. https://doi.org/10.1145/2393347.2396378