”USING DEEP CONVOLUTIONAL NEURAL NETWORKS TO TEST WHY HUMAN FACE RECOGNITION WORKS THE WAY IT DOES”

REPLICATION AND EXTENSION OF RESEARCH PAPER
TEAM: CHADGPT

Srujana Vanka
2020102005

Hariharan Kalimuthu
2020115015

Kushal Jain
2019111001

***Abstract*—The project aims to replicate and extend the inferences of the paper titled ”Using deep convolutional neural networks to understand why human face recognition works the way it does.” In the original study, Convolutional Neural Networks (CNNs) trained on specialising in certain tasks such as face recognition, object recognition etc. were evaluated on tasks such as target matching, similarity matching, multi-arrangement etc. and their performances were compared to that of human subject who had undertaken the same tasks.**

The performances of these CNNs and their comparisons to human beings’ performances lead to a lot of inferences that helped gain insights on the mechanism of face recognition in humans. We saw how certain behavioral signatures of human face recognition such as high accuracy in face identification tasks, face inversion effect, other race effect could be seen in CNNs as well, and how these observations could be explained on the grounds of specialization by means of optimization for tasks in the human brain. The paper also concluded how there was inherently nothing ’special’ about the face stimuli, the specialization which caused such dramatic effects to be shown.

***Index Terms*—behavioral signatures, convolutional neural networks (CNNs), face inversion effects, other race effect, target matching, similarity matching, multi arrangement, perception**

I. INTRODUCTION

It has always been known that the human face recognition system exhibits distinct properties such as identification of face with very high accuracy, face inversion effect(drop in accuracy when tasked to identify face stimuli presented upside down), other race effect (Drop in accuracy when asked to identify people of race unfamiliar to the participant), etc.

While cognitive psychologists have documented these characteristics for decades, the underlying computations and optimizations that give rise to these phenomena remain unclear. With the availability of task-optimized deep neural networks in recent years, the original paper makes use of the same to arrive at interesting inferences and deductions. By testing whether the behavioral signatures of human face perception emerge in any of the CNNs trained on specific tasks, we gain a better understanding of the mechanisms underlying human face recognition. In this project, we aim to replicate these signature behaviors of human face perception in CNNs and explore the possibilities of adding on to our limited knowledge of the underlying mechanisms behind human face perception and recognition.

II. METHODOLOGY

We aimed to replicate the inferences arrived at in the paper based on the datasets that were available to us. We contacted the author to procure as much data that was used in the original paper as possible. Wherever data was unavailable, we substituted the same with datasets available online to the public domain.

The methodology involved training the specialized CNNs required to carry out certain experiments that were performed, wherever pre-trained models were unavailable, and then coding up the tasks that were performed in order to arrive at inferences after evaluating human and CNN performances across these tasks. The next section discusses in detail the experiments involved and the observations we arrived at, and whether they were in accordance with the paper. The code to ranging from cleaning datasets, to replicating experiments to training CNNs were all written by us from scratch and are available for anybody who wishes to replicate this work.

We divided the experiments conducted into 6 tasks. These tasks are segregated based on the datasets the CNN model is trained and tested on. We use a variety of CNN models trained specifically for particular kinds of tasks to replicate such behavioral signatures and shape our understanding of the human face recognition system through comparison and contrasts.

To see the tasks performed and the results obtain, you may click here.

III. EXPERIMENTS AND RESULTS

A. Do CNNs achieve human-level accuracy in face recognition?

The first task aimed to measure human face recognition performance in a target-matching task, and to compare this

performance to that of four convolutional neural networks (CNNs) that were trained to specialise for different tasks. All four were based on the VGG16 architecture: one trained to discriminate face identities (Face-ID CNN), one trained on object categorization (Obj-Cat CNN), one trained on object categorisation with face as a category (including all the faces and object training images, but assigning all face images to a single category; Obj-Face-Cat CNN), and one untrained, randomly initialized CNN (Untrained CNN).

The same stimuli that were presented to the human participants were presented to the CNNs. The stimuli contained 40 unique identities with 5 images per identity. The the activation of the penultimate layer was used to compute the correlation distance between the target image and two images presented, one of which was identical to the target image. This was used to determine the accuracy of the CNNs in this target matching task.

Observations

We see that the InceptionResnetV1 CNN trained on VGGFace2 dataset did an excellent task of identifying which amongst the two images presented was identical to the target image. The performance of the object classification CNN which was exposed to faces has been slightly better at this task, despite being trained on a much smaller dataset. (We took a pre-trained VGG16 object classification CNN, and the object classification model with face as a class, was trained by us).

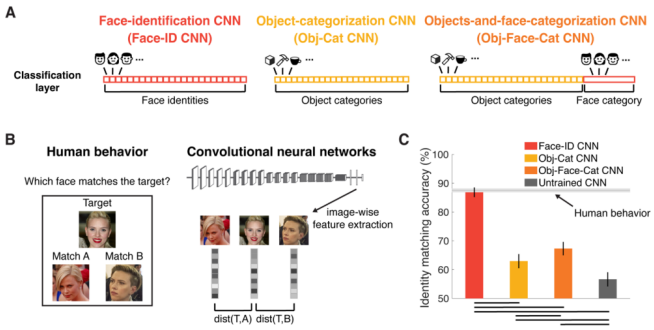

Fig. 1. Results of Target matching task performed on humans and CNNs from the paper

Inferences

Based on the results, it can be inferred that the CNN trained specifically for face recognition was able to perform close to the level of human face recognition in the target-matching task, in fact, the InceptionResnetV1 model performed even better than Humans at this task, in our runs.

Hence, CNNs specialised for the task of face recognition can achieve human like performances in face identification, and the exposure to faces during training does make the CNNs do better in this task as well.

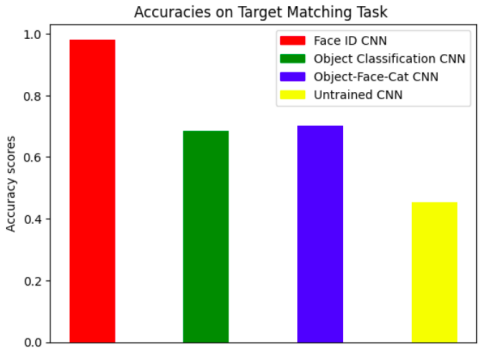

Fig. 2. Our Results of Target matching task performed task performed by the CNNs

B. Do CNNs represent faces in a similar fashion to humans?

CNNs trained on face recognition achieve similar levels of accuracy as humans. However, it's unclear if they achieve this in the same way as humans. To investigate this, the authors of the original paper used representational similarity analysis (RSA) to compare the perceived similarity of face images in humans to the similarity of face representations in CNNs. The goal was to determine if the CNNs' face representations resemble those found in humans.

1) Multi-Arrangement Task: In the second task, participants were given a multi-arrangement task, where they were asked to arrange faces similar in identity close to each other and dissimilar ones far apart. From this task, a representational dissimilarity matrix (RDM) was obtained for each participant. However, we do not conduct this human experiment, nor was the data made available to us.

We constructed four RDMs by computing pairwise dissimilarity analysis (by calculating the correlation distance (1 - Pearson's r) between the activations in the penultimate fully-connected layer for each pair of stimuli used in the multi-arrangement task) of all the stimuli that was used in this experiment. Comparison of this RDM using Spearman Rank correlation to that obtained for Human Beings would have given us valuable insights about the representational similarity in Humans and the CNNs. This analysis would have provided insights into whether the CNNs were using similar mechanisms to humans in recognizing and representing facial features, if the data of the experiment was made available.

Observations

Following were the RDMs generated from these tasks. We do see that they have very high similarity rating for different images of an identical face. (The blue boxes along the diagonal indicate this).

Inferences

If human RDMs were provided to us, we would have been able to conduct Representational Similarity Analysis(RSA) and see if results of the paper could be replicated. However, owing to the unavailability of the data regarding human

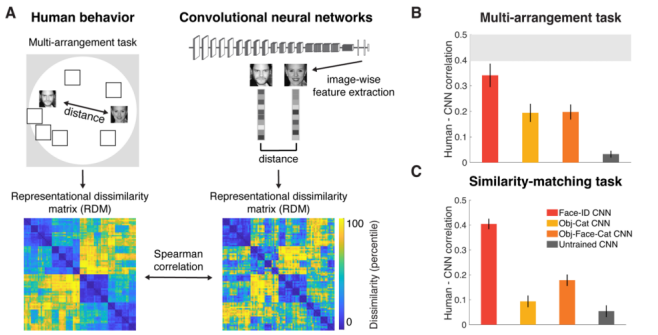

Fig. 3. Results of Multi-Arrangement and Similarity matching task performed on humans and CNNs from the paper

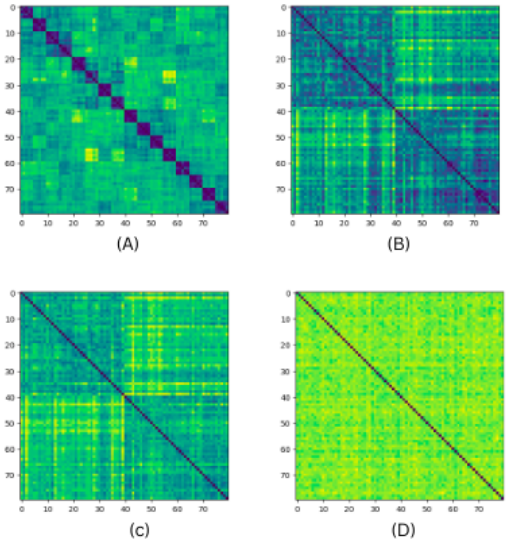

Fig. 4. RDMs generated from Multi-Arrangement task performed on CNNs (A) - Face ID CNN (B) - Obj-Face CNN (c) - Obj-Cat CNN (D) - Untrained CNN

performances, we could not proceed further with this experiment.

2) Similarity Matching Task: In order to gain further insight into the similarity of representational spaces between humans and CNNs, a similarity-matching task was designed for participants to perform.

However, due to the lack of data and information on how the stimuli were presented in this study, we were unable to replicate the results of the CNNs. It is unfortunate that this task could not be replicated, as it would have allowed us to draw more comprehensive conclusions about the similarities and differences between the representational spaces of humans and CNNs.

C. Do CNNs show classic signatures of human face processing?

The above tasks suggest that human face recognition performance can be attained by CNNs and are similar in fashion to human face recognition when trained for that specific purpose, rather than for generic object categorization or face detection.

This section explores whether classic behavioral signatures of human face processing, such as the other-race effect and the face inversion effect, are seen in CNNs in any fashion.We will now use a variety of CNN models trained specifically for particular kinds of tasks in an attempt to find if any of them show such behavioral signatures.

1) The Inversion Effect: The target-matching task was performed by presenting images upside-down to both human beings and CNNs. it is known that amongst humans, a performance drop in accuracy occurs in identification when the face stimuli is presented upside down.

Observations

We see that there was a significant drop in performance on the CNN trained for face recognition, although all the other CNNs virtually remained unaffected by the stimuli being presented upside down.

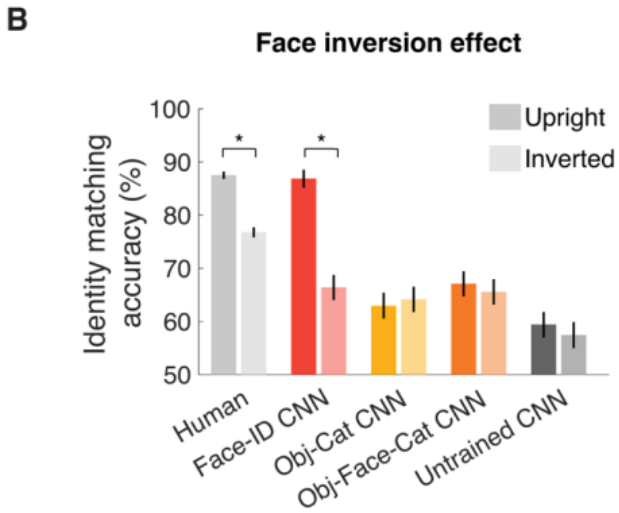

Fig. 5. Results of Face Inversion Effect from the paper

Inferences

Human beings performed worse when they were asked to identify the target image, given two images if the faces were inverted. The results of the drop in accuracy of the CNN trained for Face recognition in our run of the task concurred with that of the study. CNN trained on face recognition showed a drop in performance for inverted images. There was no significant difference in the performance of the other CNNs for upside-down images.

We see that such drops in performance of human beings can be explained in terms of the optimisation of human

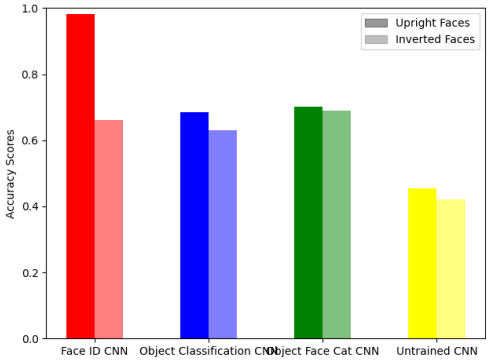

Fig. 6. Our results of Face Inversion Effect

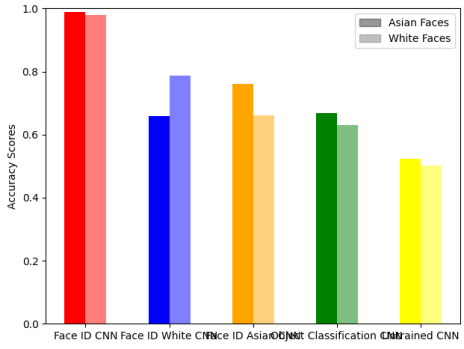

Fig. 8. Our results of Other-Race Effect

beings' face recognition system to upright faces, which takes a hit when stimuli is presented upside down.

2) The Other-Race Effect: To investigate the other-race effect in both humans and CNNs, the study conducted was a target-matching task with White and Asian participants being tested on white and Asian face images. It was observed that asian participants performed worse when tasked to identify white faces while the white participants did a worse job when presented with asian faces as stimuli as compared to white faces. This indicated that people are good at recognising faces of races they are familiar with.

We trained two CNNs: one on face recognition using only Asian identities, and another on a dataset with only White identities. Asian Faces and White faces were presented to both these CNNs and the results were compared. This allowed us to test the other-race effect in CNNs.

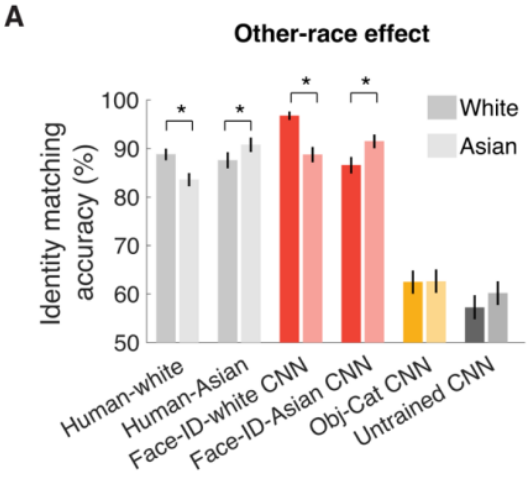

Fig. 7. Results of Other-Race Effect from the paper

Observations

We see that the face recognition CNN which was trained on datasets with contain significant representations of all races did no suffer drop in accuracy when presented with stimuli of different races. However, the CNN trained on asian faces did a poorer job when tested on white faces and vice-verse for the case of CNN trained only on white faces. This did not affect any CNNs that were not trained for face recognition.

Inferences

Humans showed decreased recognition performance when presented with faces of races they were unfamiliar with, and CNNs trained on both white and Asian identities exhibited similar behavior as their human counterparts.

The results of the drop in accuracies of our CNNs coincided with those of the study. The fact that the other CNNs did not show such a drop in performance indicates how optimizations work with respect to the training data in the case of CNNs and the familiarity in the case of human beings.

D. Is the 'Face' stimuli special in any sense?

The question of whether any stimulus category can produce inversion effects given sufficient training has been long debated in the psychology literature. Since human beings are not nearly exposed to any other stimuli as much as human faces, we could not proceed to address this question before CNNs came into the picture.

We check the performance of the Face ID CNN on upright and inverted images and compare how it fares when a CNN trained on inverted faces is tested on upright faces as well as inverted faces. For this, we trained the CNN ourselves, not the pre-trained one we've been using so far, and both were trained on the same dataset, with one of them being fed inverted images for training(i.e. rotated by 180 degrees).

Observations

We see that there is a drop in performance when the CNN trained on inverted faces is given to identify the image of an upright face, and vice-versa for the CNN trained on upright images.

Inferences

While the CNN trained on upright faces showed higher accuracy for upright than inverted faces, the CNN trained on inverted faces showed the opposite. The performance

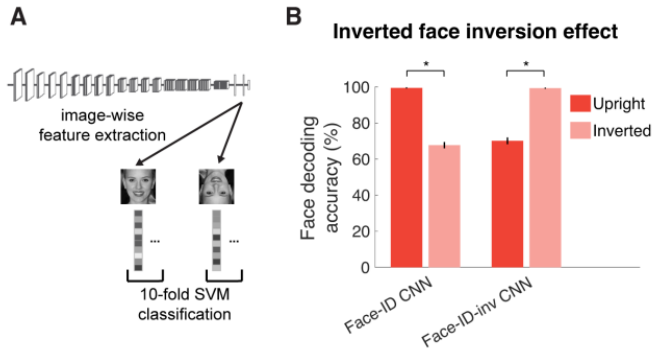

Fig. 9. Results of Inverted Face Inversion Effect from the paper

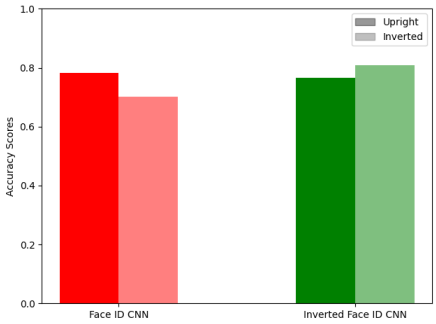

Fig. 10. Our results of Inverted Face Inversion Effect

of the CNN which specialized in face recognition dropped significantly, as compared to that of the other CNNs(as shown in the previous section). Now we see whether this effect is restricted only to face stimuli or can occur for other CNNs (or systems) which are optimised to specialise in identifying some other entity.

The Car Inversion Effect

To investigate if the inversion effect is unique to faces, we conducted experiments on fine-grained car decoding. For this purpose, we trained a VGG16 architecture on a car model using the CompCars dataset. Additionally, we wanted to see if other CNNs show an inversion effect, and thus, we contrasted the performances of various CNNs when presented with upright and inverted car images.

Observations

These results are similar to the results from the paper. Our findings indicate that inversion effects are not specific to faces but can arise naturally for other stimulus classes when trained on only upright stimuli.

Inferences

Only the car-trained CNN showed an inversion effect for cars, i.e. lower performance for inverted than upright cars. However, we did not observe this effect in networks trained

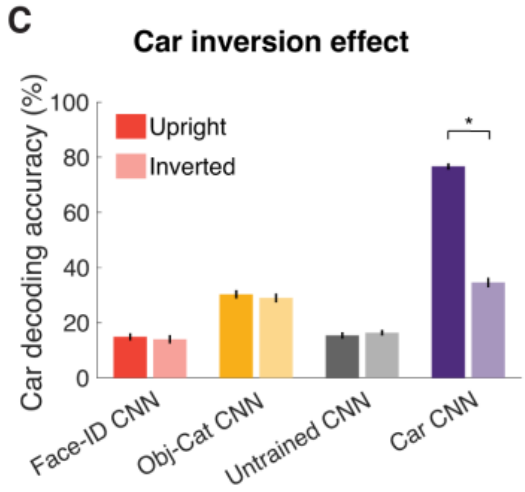

Fig. 11. Results of Car Inversion Effect from the paper

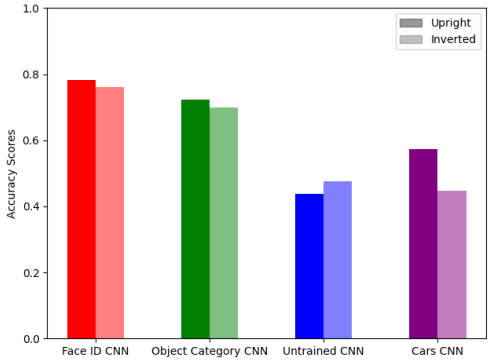

Fig. 12. Our results of Car Inversion Effect

on object categorization or even face-identity recognition, or untrained CNNs.

This tells us that the effects such as face inversion effects and the other race effects observed in human face recognition is not due to the stimuli being special, but due to underlying computational mechanisms and optimisation towards the task which involves faces.

IV. Conclusion

We successfully replicated all the tasks that were presented in the paper and hence, seconded the inferences and conclusions drawn by the authors in their work. We gained a lot of experience while working with CNNs, right from training them to getting distances between various activation layers and using such information to develop an understanding of the principles behind Human Face Recognition.

We realised how optimisation and specialisation towards a particular task gives rise to the phenomenon observed in mechanisms such as face recognition in human beings. This paper made us realise how the use of tools such as neural networks can help gain an understanding of the Human Brain.

V. Future Scope and Extension

We realised that there could be very interesting insights gained when we work with the question 'Upto what extent of specialisation and optimisation in the task of face recognition can we start observing the human face recognition characteristics?'

In the spectrum of CNNs designed to detect and classify a presented stimuli as a face to identifying a particular face, we can design CNNs which lie somewhere in between these two ends of the spectrum in terms of feature extraction given a face image. If we design several of such CNNs with varying degrees of specialisations towards face recognition and then see where effects such as the other race effects and face inversion effects show up, we can arrive at an answer to the research question posed above. This is our idea of extending the knowledge gained from the original research paper and helping us move a minuscule step closer to understanding human face perception.

We will try to implement this extension of paper and are currently learning about MTCNNs, while exploring several other architectures that could be pit against the CNNs used in the project. Their performances in various tasks such as these may end up providing us with valuable information.

VI. Code and Datasets

Here is the link to all the code used in our project. To access all the training data and stimuli data, refer to the README.md file.

References

[1] 'Dobs, Katharina Yuan, Joanne Martinez, Julio Kanwisher, Nancy. (2022). Using deep convolutional neural networks to test why human face recognition works the way it does. 10.1101/2022.11.23.517478.'