

Visual Similarity: A Deep Neural Network and Human Subjects Approach

Dawood Abbas Abdul Malik

da2729@nyu.edu
New York University

Zixiao Chen

zc2194@nyu.edu
New York University

Audrey Chu

ac8839@nyu.edu
New York University

Elena Georgieva

etg259@nyu.edu
New York University

Abstract

Understanding similarity is a central question in the field of cognitive science. Similarity is used in industry to tackle topics such as semantic segmentation, recommendation systems, scene understanding, and even text mining. Deep neural networks have been the popular choice in solving these perception problems and have even reach or surpassed human-level accuracy. In this paper, we examine and extend on the methodologies implemented in *Adapting Deep Network Features to Capture Psychological Representations* (J. Peterson and T.Griffiths, 2016). In addition to developing three neural networks, we collect similarity ratings data from 57 human subjects. We examine the connection between representations learned by the networks and human judgements, and also manipulate the deep features to align more with human judgements. We further analyze human judgements by comparing similarity judgements based on survey participant behavior such as time to complete and rating by stimuli sequence. By understanding how image representations differ for the same person, we can apply this extended human behavior knowledge to areas such as image recommendation platforms.

Keywords: deep learning; computational cognitive modeling; neural networks; visual representation

1 Introduction

Neural networks are able to learn feature representations of high-dimensional inputs. They have been extremely successful in the field of semantic segmentation and computer visions for this reason. In recent work, Peterson et al. examined the relationship between neural network representations and human subject representations (J. Peterson and T.Griffiths, 2016). They developed a method for adapting neural network features to align with human similarity judgments.

Peterson’s work was done on a broader domain of images consisting of different animals like dogs, birds, and giraffes. In this paper, we are focused on examining a narrower domain: birds. In our narrow domain, similarity judgement is much harder than when the domain is

broader. When judging the similarity of birds, it can be difficult even for human subjects to make such judgements. We observed more doubtful judgements than certain judgements Figure 4. The goal of this paper was to understand how different pre-trained neural networks judge similarity in this scenario compared to human judgements.

Neural networks are able to learn feature representations of high-dimensional inputs, which are also a challenge for human perception (Austerweil, 2013). Some scientists have explored how representations learned by neural networks can be used in models of human behavior for tasks including predicting human judgements of category typicality (Lake, 2015) and predicting the memorability of objects in images (Dubey, 2015).

There have been many recent advances in the neural network representations of human perceptual tasks, but the problem is far from solved. In one study, a neural network was similarly able to classify images until it was systematically deceived by imperceptibly image transformations (Szegedy, 2013). Another group of authors attempt to explain adversarial examples in neural network training, where input results in the model outputting an incorrect answer with very high confidence (Goodfellow et al., 2015). Due to instances such as these, scientists wonder how similar neural network models truly are to real human perception, despite the fact that they can somewhat mimic human-like abilities.

In 2018, Peterson et al extended their own work. They observe that while the industry’s top object classification networks provide accurate predictions of human similarity judgments, they fail to capture some of the structure represented by humans. They use convex optimization to correct these discrepancies and enable the tractable use of large natural stimulus sets (Peterson et al., 2018).

In this paper, we examine the connection between representations learned by three state-of-the-art neural networks: GoogleNet, CaffeNet, and VGG166 to human judgements. We collected 57 human comparisons via an online survey and further analyze those results based on participants behavior. We understand how humans rate image similarity, and discuss how that can help with the development of neural networks for similarity.

Cognitive science researchers have recently explored the relationship between human classification and neural network design elements to provide a reference value for the future development of neural network design

(Xing Yang, 2021) and used human visual object shape similarity judgements to model data-trained networks (German and Jacobs, 2020). Clearly, combining human and neural network representations is a relevant topic in the field.

Understanding the relationship between the representations found by deep learning and those of humans is a central question in the field of cognitive science. Research in this area would likely benefit artificial intelligence including autonomous vehicles and film and video technology.

2 Deep Representations

Deep neural networks (DNN) are neural networks that have multiple hidden layers between input and output (Bengio, 2009). In recent years, there have been especially large advances in the field of computer vision, in part thanks to a particular DNN called a convolutional neural network (CNN) (LeCun, 1989). CNNs use convolutional layers, which allow for a great decrease in the number of parameters a network must learn in order to process high-dimensional inputs.

Notably, CNNs produce feature representations at each layer of the network, as opposed to just at their outputs. These representations have proven to be useful in predicting human behavior. Dubey et. al. used representations extracted from the last layer of a CNN to predict object memorability, and Lake et. al. used the same network and features to predict human typicality ratings (Dubey, 2015) (Lake, 2015). Lake’s work drew more cognitive scientists in to work on neural networks.

CNNs are also used in the neuroscience community. Yamins et al. used CNN activations to predict monkey IT cortex activity (Yamins, 2014), and Agrawal et al. used CNNs to predict low and high-level activity in human visual areas (YAgrawal, 2014). More recently, Yin et al. used a deep learning model called VIN-Net to classify images (Wei et al., 2020). This model, inspired by human visual interaction mechanisms, is successful on visual classification tasks.

In this paper, we investigate how well deep neural network features align with human perceptual recognition. We hope to advance the correspondence between the two by using human subjects data to influence our neural network design.

3 Evaluating Deep Representations

Before we investigate the relationship between human psychological representations and deep neural networks, we need to first find a way to measure human psychological representations quantitatively. Inspired by Peterson et al. (2016), we developed two $N \times N$ matrices of similarity ratings, one representing the similarity judgements by deep neural networks and another one representing human similarity judgements. However, since the input of deep neural networks is an image instead of an array, we performed feature extracting to form the matrix of deep neural networks similarity judgements. We

extracted features from each pair of images and took the dot product of those to generate paired similarity judgement vectors for both human subjects and each networks. Lastly, we calculated the correlation between these vectors and evaluated the relationship between human psychological representations and deep neural networks.

Stimuli. We selected 18 colored bird photos as our stimulus set as shown in Figure 1 from Kaggle (Kaggle, 2020). In order to have both inter-species and intra-species comparison, the stimulus set covers six different bird species, with three photos of each species. The six bird species were Bobolark, Shiny Cowbird, Indigo Bunting, Hooded Warbler, Painted Bunting, and Marsh Wren. Every photo was resized to 300x300 pixels to present the full body of birds while preventing the noise caused by different image qualities.



Figure 1: Bird Image Dataset

Behavioral Experiment. We organized 153 pairs of photos from the 18 pictures from the stimulus set and distributed them as a survey to 57 volunteers to get human similarity judgements. Volunteers were asked to give ratings on a likert scale from 0 to 6 on all 153 pairs of photos based on the similarity of paired photos. To get volunteers familiar with our survey, we started the survey with a short practice trial that included four pairs of photos. We collected 8721 human similarity judgement in total. Each pair of photos was rated by 57 different volunteers. We constructed the final result in to an 18×18 matrix, with each element representing the average rating of one pair of photos across 57 different individuals.

Feature Extraction. We conducted feature extraction by adopting three pre-trained CNNs: GoogLeNet, VGG16 and CaffeNet. The methods of feature extraction via three CNNs are quite similar, by passing flattened image vector into networks and generate feature feedback at each layer then looking at the last layer before the classification layer. The main difference between the three selected CNNs is that GoogLeNet has a 1000-dimensional average pooling layer as its last layer, while VGG16 and CaffeNet have a 4096-dimensional fully connected layer as the last layer.

3.1 Results: Deep Representations

We presented the correlation between human judgement and each deep neural network similarity in Table 1 showing the R^2 . From the result, we can see that all three models' results are quite similar and are medium correlated with human judgements, which indicates that the networks learned some part of human representations but not entirely.

Since the networks do not fully capture what human representations maintained, we adopted multidimensional scaling (MDS) and plotted hierarchical clusters to visualize the aspects that deep representations failed to capture. Using multidimensional scaling (MDS), we can find a low-dimensional representation of the data in which the distances scale with respect to our original high-dimensional space. MDS can be used to analyzing similarity data, and attempts to model similarity as distances in geometric space. In Figure 2, we can see how the concept of similarity has progressed based on human judgment data (left), GoogLeNet deep representations (middle). The human data view shows that our survey participants were able to distinguish images from the same species with little variation. The deep representations show that the networks are less successful in emulating the human behavior, evidenced by the more sparse image distribution within each cluster comparing to the one in human behavior panel. We also apply another classic psychological tool, hierarchical clustering, to analyze the psychological representations (Figure 3). This figure shows tree structures for human representations (left), deep representations (middle), and adapted representation (right). From the left panel, we observed that birds from the same species are under the same cluster, which suggests that human participants successfully categorized bird images based on their similarity. However, from the middle panel we drew the conclusion that deep representations is inconsistent to human judgements, since bird images from the same species are not always under the same cluster as it supposed to be.

Deep Representation Analysis. To further investigate why these three models perform differently, we looked at the differences between models based on different image pairs in Table 2. We have scaled our model similarity judgements to a scale of 0 to 6 to compare it against the human ratings. We picked out a subset of 10 paired images that represent various elements which influence human judgement and deep networks' prediction. These elements include but not limit to bird species, background color, bird poses, bird color, bird size, and object orientation, such as whether the bird faces camera or only shows a side of its body. We examined the correlation (R^2) between human judgement and each deep networks based on each pair of images independently and presented the result in Table 2. The smaller the difference between human judgement and deep representations, the closer the deep network representation is to the human.

From Table 2, we observed that GoogLeNet outperforms other networks for most of the time, except for

Table 1: Correlations between human and deep network similarities.

	CaffeNet	GoogLeNet	VGG
R^2	0.35	0.41	0.32

pair 3, 7, and 10, which is consistent to our correlation results in Table 1. However, for pair 3, 7, 10, where birds have similar orientations, CaffeNet is the best performed one. We concluded that CaffeNet is able to represent the orientation of the bird in the image very close to how humans perceive this structure. Similarly in image pairs having birds of different colors, GoogleNet seems to represent it much better compared to the other models. The human ratings in these images are not closer to GoogleNet, this maybe because of the other structures like the size of the bird, beak type etc. perceived by humans are not represented by GoogleNet. VGG also performs like GoogleNet in the aspect of colour but it is not performing well when the birds in the images are of different sizes. CaffeNet is affected more by the background of the image compared to the other models. To summarize CaffeNet represents the orientation of the birds better, GoogleNet represents the colours and VGG seems to have a balanced representation of different features.

4 Adapting Representations

The correlations between the deep neural network representations of the bird images and the corresponding human similarity judgements quantifies the discrepancy between both. The problem of classification in a neural network comes down to a linear transformation in the last layer, eg., Softmax function. Hence, the final layer representations we have can be considered the inputs to a transformation which gives the solution for the categorization problem. This idea can be represented as a linear transformation, whose solution would better capture human similarity judgements.

Similarity matrix in Deep Representation. In the earlier method we obtained a feature vector for each object from deep representations, the similarity judgment of two images given by a model is just FF^T . In other words

$$s_{ij} = F_i \cdot F_j$$

where s_{ij} is the similarity between i^{th} and j^{th} picture given by the model and F_i is the feature vector obtained for i^{th} image.

Similarity matrix in Adapting Representation. Applying the idea of linear transformation, any similarity matrix can be decomposed as a matrix product of a feature-by-object matrix F , F^T and a diagonal matrix with weights.

$$S = FW F^T$$

This formulation is similar to the one employed in *Adapting Deep Network Features to Capture Psychological Representations* (J. Peterson and T.Griffiths, 2016). With the

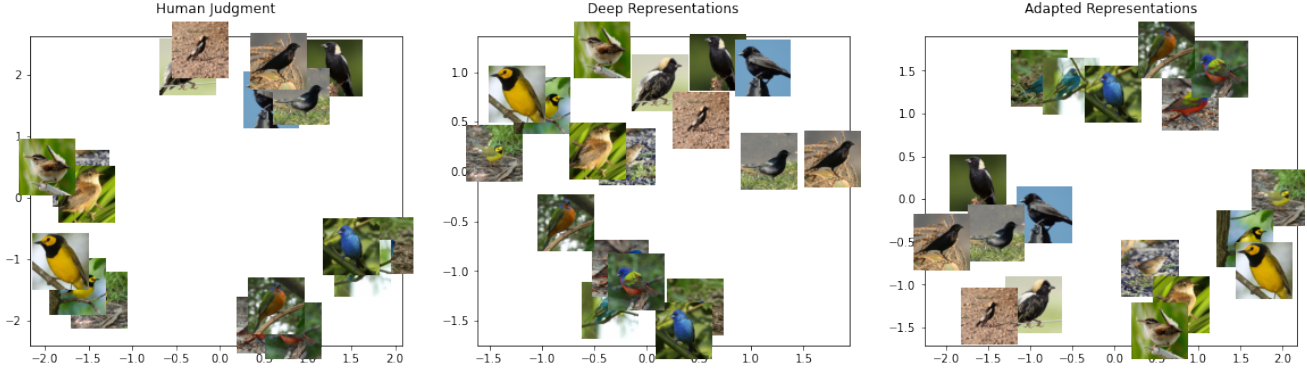


Figure 2: Multidimensional Scaling Solutions (MDS) for Similarity Matrices using GoogLeNet.

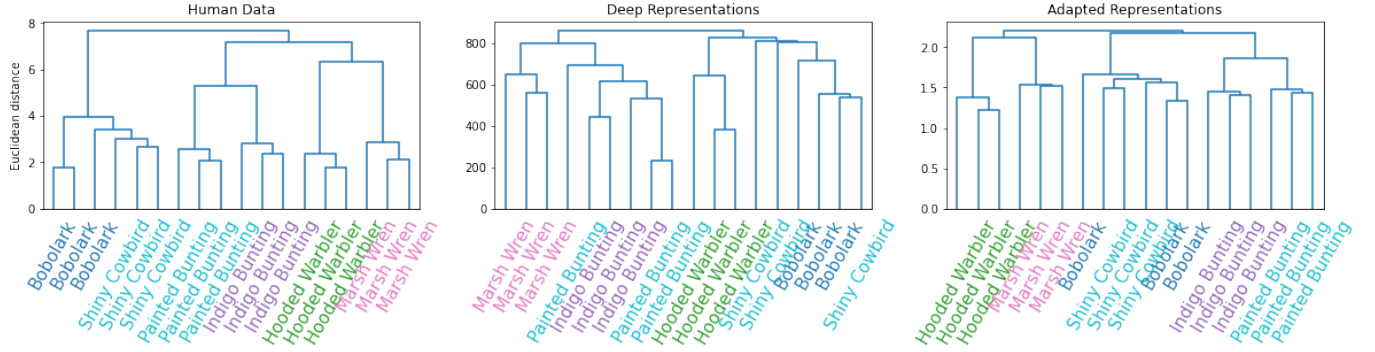


Figure 3: Hierarchical Clustering with GoogLeNet

feature-by-object matrix derived from deep representation, the weights W , which is a diagonal matrix can be solved using a regression mapping, where the predictors for each s_{ij} (similarity judgement between i^{th} and j^{th} image) is the dot product of the corresponding feature vector for images i and j .

Methodology. In this problem, since we have a large number of predictors there is a chance of overfitting in the regression mapping. Hence we have used ridge regression with L2 regularization. A Grid search was performed to find the best regularization parameter, using cross validation to validate performance. Now the similarity judgement is expressed as:

$$s_{ij} = \sum_{k=1}^n w_k f_{ik} f_{jk}$$

where f_{ik} is the k^{th} feature of image i and w_k is the weight. The squared error in adapting the model’s similarity judgements to the human similarity judgements is minimized by convex optimization from the ridge regression model. This new representation explains more variance of the dataset than the previous one.

4.1 Results: Adapted Representations

By incorporating human judgments into deep feature representations, we are able to achieve a higher correlation across all three deep networks (3). All models R^2 values either doubled or more than doubled in value. The

adapted VGG16 network performed the best with a R^2 of 0.87. In *Adapting Deep Network Features to Capture Psychological Representations* (J. Peterson and T.Griffiths, 2016), VGG16 also performed the best and achieved similar R^2 . The only control we have done on preventing overfitting in the regression mapping is the regularization. In our results from grid search for the best parameters, we observed a high value for the regularization, this maybe because of a smaller data set and the narrower domain here, with all feature vectors representing birds. This suggests that further randomness like, shuffling the features would help further in generalizing this model. The trends in the correlation between the human and different models is still going to be the same.

To further understand how the representations are being learned under the adapted method, we can extend on the classical psychological tools used in the Deep Representation results section. In Figure 2, comparing between the deep representation (middle) and adapted representation (right) views, birds from the same species are more tightly clustered than under the deep view. Additionally, in Figure 3, the dendrograms show that the branches of the adapted tree diverge at a much higher spatial distance than that of the deep representation tree. Comparing between the human judgment view (left) and adapted, we can see that GoogLeNet does indeed achieve the closest view to the human judgment view in both figures. This pattern was evident across all deep networks.











Image Pair	Human	CaffeNet	GoogLeNet	VGG16
	3.51	0.85	2.18	0.75
	1.02	0.38	0.65	0.15
	2.2	1.40	0.64	1.01
	4.95	1.33	1.80	0.64
	4.31	0.66	1.46	0.81
	4.51	1.43	2.87	2.79
	2.27	2.20	1.59	0.93
	3.73	1.09	1.98	0.58
	3.04	1.20	2.25	1.56
	1.69	2.52	0.58	0.43

Table 2: Similarity judgement for a subset of pairs for human judgement data and by model, (scaled to 0 - 6).

Table 3: Correlations between human and adapted network features.

	CaffeNet	GoogLeNet	VGG
R^2	0.70	0.84	0.87

5 Similarity Judgments by Human Behavior

To serve as somewhat of a “ground truth” for our Neural Networks, we collected pairwise similarity ratings for our bird stimulus set from 57 human participants. 28 participants identified as male, 26 identified as female, one identified as non-binary, and one preferred not to share their gender identity. 32 of our participants were in the age range 25-34, 22 were 18-24, two were 35-44, and none were over the age of 44. We created an approximately 15-minute long survey in Qualtrics. Each participant was asked to rate how similar a pair of images was on a likert scale ranging from 0 (not similar at all) to 6 (very similar). Following four practice trials, each participant rated all 153 possible pairings of the 18 photos, giving us 8721 total human similarity judgements.

The survey also included several demographic questions at the end, inquiring about the participant’s age, gender identity, race, as well as how long they have resided in the United States. The Qualtrics software also collected some information for us including how long it took each participant to complete the survey, the device they completed the survey on, and a timestamp.

5.1 Human Behavior Results

Following the survey, we put together an 18 x 18 similarity matrix, where each element is the inner product of the feature representations of each pair of images. This is a measure of similarity between the two vectors. All correlations were done with Cafenet. All similarity scores ranged between 0.153 and 0.677, with a mean value of 0.441. This is indicative of some correlation with human ratings.

Our behavioral experiment was somewhat based on the work done by Peterson et al. in 2016. However, our experiment was different in that our likert scale was between 0 to 6 (theirs was 0 to 10) and they had more participants rate some of the images, as opposed to us having a smaller number of participants who rated all of our image pairs.

Furthermore, our domain was narrower than Peterson’s (only birds vs. many animals). Similarity judgement is much more difficult when the domain is narrower, and we found it was be difficult even for humans to make the similarity judgements we asked. We observed more doubtful judgements (similarity scores 1-5) than certain judgements (similarity scores 0 or 6), Figure 4.

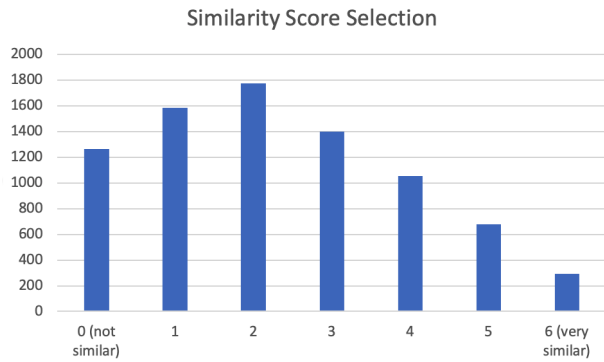


Figure 4: Distribution of the 8721 total Similarity Scores collected. Scores were significantly skewed right, with very few participants selecting a value of 6 (very similar) for an image pair.

5.2 Analysis

Unlike Peterson et al, we used a 0 to 6 likert scale. The most frequently selected similarity rating was a 2, and the mean similarity rating was 2.36 (Figure 4). Interestingly, the score 0 (not similar at all) was selected 1266 times total, and the score 6 (very similar) was selected just 298 times out of the 8721 total scores given. It seems participants rarely select the highest possible similarity rating, and we believe Peterson and his team had similar findings. Narrowing our likert scale to only 7 options, as opposed to 11 like Peterson, may have skewed our results to lower similarity scores.

We analyzed some of the human data collected for us by the Qualtrics software. We investigated the correlation between human and model similarity ratings on different human behaviours. On average, participants took around 11 minutes to complete the survey. Participants who took more than 11 minutes to complete the survey had a correlation value of 0.39 with the CaffeNet model, whereas those who took less than 11 minutes had a correlation of 0.32. This shows that some participants may have been rushed. In the future, it might be interesting to throw out the survey results of participants who went through the questions quickly. We would expect a stronger correlation with the output of the neural network.

Next, we looked at correlations in only the first half of the survey vs the second half of the survey. Correlation results were 0.39 and 0.31, respectively. This shows that participants may have experienced fatigue later in the survey. In the future, it may be interesting to shorten the survey so fatigue wouldn't be an issue. Alternatively, we could implement a mandatory break halfway through the exercise. Participants who completed the survey on a mobile screen vs a computer screen scored equivalent correlations at 0.34. This shows device did not make a difference in correlations with the neural network.

6 Discussion

In this work, we have analyzed the comparison between deep representation, adapting representation and human psychological representations, specifically in similarity judgements given on a narrow domain of images. We also observed that as in *Adapting Deep Network Features to Capture Psychological Representations* (J. Peterson and T.Griffiths, 2016), the high performing pre-trained neural networks were moderately correlated to the human representation, but failed to learn different structural characteristics like humans. The adapting representations derived from the similarity model overcome this drawback, but gave another problem in terms of generalizing the linear model. Though we see that we are able to adapt Peterson's work into a narrower subset domain like birds, the generalizability still cannot be judged until tested on a broader set of domains. These domains can vary based on size or even number of categories, and would further help produce a more generalized representation.

References

- [Austerweil2013] Griffiths T. L. Austerweil, J. L. 2013. A nonparametric bayesian framework for constructing flexible feature representations. *Psychological Review*, 120(4):817.
- [Bengio2009] Y Bengio. 2009. Learning deep architectures for ai. *Foundations and Trends in Machine Learning*, 2(1):1-127.
- [Dubey2015] Peterson J. Khosla A. Yang M.-H. Ghanem B. Dubey, R. 2015. What makes an object memorable? *International Conference on Computer Vision (ICCV)*.
- [German and Jacobs2020] Joseph Scott German and Robert A. Jacobs. 2020. Can machine learning account for human visual object shape similarity judgments? *Vision Research*, 167:87-99.
- [Goodfellow et al.2015] I. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572.
- [J. Peterson and T.Griffiths2016] J. Abbott J. Peterson and T.Griffiths. 2016. Adapting deep network features to capture psychological representations. <https://arxiv.org/abs/1608.02164>.
- [Kaggle2020] Kaggle. 2020. 200 bird species with 11788 images.
- [Lake2015] Zaremba W. Fergus R. Gureckis T. M. Lake, B. M. 2015. Deep neural networks predict category typicality ratings for images. *Proceedings of the 37th Annual Cognitive Science Society*.
- [LeCun1989] Boser B. Denker J. S. Henderson D. Howard R. E.-Hubbard W. Jackel L. D. LeCun, Y. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541-552.
- [Peterson et al.2018] Joshua C. Peterson, Joshua T. Abbott, and Thomas L. Griffiths. 2018. Evaluating (and improving) the correspondence between deep neural

networks and human representations. *Cognitive Science*, 42(8):2648–2669.

- [Szegedy2013] Zaremba W. Sutskever I. Bruna J. Erhan D. Goodfellow I.- Fergus R. Szegedy, C. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.
- [Wei et al.2020] Bing Wei, Haibo He, Kuangrong Hao, Lei Gao, and Xue song Tang. 2020. Visual interaction networks: A novel bio-inspired computational model for image classification. *Neural Networks*, 130:100–110.
- [Xing Yang2021] Meihua Li Wenying Wang Huichun Xie Jinping Du Xing Yang, Tingjun Yong. 2021. Relationship between cognitive learning psychological classification and neural network design elements. *Complexity*, 2021.
- [YAgrawal2014] Stansbury D. Malik J. Gallant J. L. YAgrawal, P. 2014. Pixels to voxels: Modeling visual representation in the human brain. *arXiv preprint arXiv:1407.5104*.
- [Yamins2014] Hong H. Cadieu C. F. Solomon E. A. Seib-ert D. DiCarlo J. J. Yamins, D. L. 2014. Performance-optimized hi- erarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23), 8619–8624.

7 Individual contributions

- Dawood Abbas Abdul Malik: CaffeNet for deep and adapted representations, writing discussion, adapted representation methodology, and report visuals.
- Zixiao Chen: VGG16 for deep and adapted representations, writing deep representations results and analysis
- Audrey Chu: GoogLeNet for deep and adapted representations, writing abstract, adapted representations results, and report visuals
- Elena Georgieva: Creation of Qualtrics human subjects survey, analysis of human subjects results, writing introduction, deep representations, and similarity judgments by human behavior

8 GitHub Repository

The code for this project can be found at https://github.com/audreychu/CCM_SimilarityRatings.