Unveiling the Typicality Effect in Vision Models

Chaewon Park (CPARK382@Gatech.Edu)
College of Computing, Georgia Institute of Technology
Atlanta, GA, 30332 USA

Abstract

Over the past decades, neural networks have been in the spotlight in various fields of computer vision including image classification, and it has proved its performance through numerous tasks. However, beyond evaluating the model performance based on the prediction accuracy of the model, evaluating how similar the model perceives and interprets visual stimuli compared to the way humans do has been gaining great attention recently. In particular, the current study focuses on the typicality effect that is prevalent in human cognition. The typicality effect refers to a phenomenon where typical members of a category are likely to be used to categorize other instances than atypical members. That is, humans often recognize typical objects more quickly and accurately than atypical objects. In this work, we hypothesize that state-of-art computational models for image classification tasks also exhibit the typicality effect as observed in human cognition, and empirically prove the effect by exposing selected pre-trained Convolutional Neural Network (CNN) models of varying sizes (e.g., ResNet50) to a large dataset of images in various categories. We then evaluate whether the models' comprehension of images aligns with that of humans. The results from the above experiment show that typicality ratings reported by the CNN models are not highly correlated to human typicality ratings. This paper analyzes possible causes of demonstrated discrepancies and suggests future directions of the study, emphasizing the importance of continued exploration in this evolving intersection of human cognition and artificial cognition present in machine learning.

Keywords: Human Cognition; Typicality Effect; Image Classification; Vision Models; Convolutional Neural Networks (CNN)

Introduction

Over the past decades, neural networks have been in the spotlight in various fields of machine learning, ranging from image recognition to natural language generation. It has proved its outstanding performance exhibiting human-like behaviors in numerous tasks. While conventional metrics like accuracy have been widely used in evaluating the model performance, there has been a growing recognition of understanding how these computational artificial models perceive and interpret stimuli in ways that parallel human cognition.

The current study places a particular emphasis on investigating the intersection of human cognition and Convolutional Neural Networks, which have been in dominant use in various computer vision tasks.

Convolutional Neural Networks (CNN), characterized by their hierarchical feature extraction through multiple convolutional layers, have demonstrated a remarkable performance in image classification. State-of-art CNN models such as AlexNet, ResNet, and VGG, have exhibited a misclassification rate of less than 3% on a vast set of data. However, the depth of their understanding and alignment with human cognition processes remains an intriguing frontier yet to be fully explored.

One notable trait of human cognition is categorization. Despite the inability to completely separate objects in the real-world, we humans exhibit a natural tendency to group equivalent entities together (Rosch et al., 1976). The level of abstraction plays a pivotal role while determining the equivalency of two objects. The higher the level of abstraction, the greater inclusiveness of a category. The level of abstraction, thus, involves a sophisticated measure capable of successfully distinguishing in-group members from out-group members.

Such human categorization, inherently subjective and dependent on the level of abstraction, prompts us to explore the notion of *typicality effect.* The typicality effect refers to a phenomenon where typical members of a category are likely to be used to categorize other instances, rather than atypical members. By typical members, we mean objects that share a reasonable number of common attributes with the prototype image of a category. In practical terms, the typicality effect explains the human behavior, often recognizing typical objects more quickly and accurately than atypical objects.

In this work, we hypothesize that state-of-art CNN models for image recognition tasks also exhibit the typicality effect as observed in human cognition. To empirically prove the effect, we expose selected pre-trained Torch vision models of varying sizes (e.g., ResNet50) to a large dataset of images in various categories. We then implement the statistical measures to evaluate whether the models' comprehension of images aligns with that of humans. The results from the above experiment identified that the typicality ratings reported by the CNN models are not consistently correlated to human typicality ratings. This paper analyzes potential causes of demonstrated discrepancies, and offers valuable insights into complex

interplay between artificial and human cognition. Furthermore, the paper outlines the future directions of the study, emphasizing the importance of continued exploration in this evolving intersection of human cognition and artificial cognition present in machine learning.

Related Works

There have been an increasing number of studies on evaluating human cognition, and the typicality effect in particular, in various artificial computational models. One pioneering work in the field was by Lake et al. (2015), who evaluated the ability of CNNs to predict human typicality ratings from raw pixel data of category examples. Building on this, Saleh, Elgammal, and Feldman (2016) also contributed to this field by scrutinizing limitations of existing CNN models in generalizing to atypical instances. They proposed enhancements to the models to improve their generalization capacity, incorporating a typicality measure.

Our current study was largely inspired by a study titled "Evaluating Typicality in Combined Language and Vision Model Concept Representations" (Vemuri, Shah & Varma, 2023). The paper provides a comprehensive empirical analysis of the typicality effect in modern vision models and language models of varying sizes. While only an insignificant degree of correlation was observed in both models when experimented independently, the authors proposed an interesting approach – a combined language and vision model which, according to their findings, better emulates a human-like typicality effect.

These prior works lay the firm groundwork for our investigation into the typicality effect in Convolutional Neural Networks (CNNs) and serve as catalysts for the design and methodology of the present study. As we delve into the following sections, we aim to build upon these insights and contribute to the evolving study on the intersection of artificial and human cognition.

Approach

Data Preparation and Preprocessing

Human Data For human typicality ratings, category potency results from an experiment run by Castro, Curley, and Hertzog (2020) was used. Castro et al. conducted an experiment where participants in varying age groups were asked to provide as many exemplars as possible given a category, and reported mean potency scores. Potency score for each subcategory was calculated by dividing the number of responses by the total number of participants. In the original paper, scores were reported by age groups along with the total score, but only the total scores were taken into account for the purpose of the current study. The raw human typicality dataset initially had a total of 70 categories and ~2130 subcategories (exemplars). This paper followed the preprocessing of human typicality ratings outlined in "Evaluating Typicality in Combined Language and Vision Model Concept Representations" (Vemuri, Shah & Varma, 2023). This preprocessing includes renaming category names to match image data, removing NaN entries, and only leaving entries with sufficient image data. By such processing, the data was refactored to a total of 27 categories and 555 subcategories (Table 1).

Table 1: Categories of interest

Categories	Number of Subcategories
Green thing	25
Flying thing	19
Flower	20
Furniture	22
Tree	26
Four-legged animal	31
Instrument	25
Ship	25
Dwelling	21
Vehicle	21
Snake	19
Fish	27
Carpenters tool	17
Color	21
Gardeners tool	15
Fabric	17
Bird	33
Weapon	23
Toy	23
Kitchen utensil	16
Vegetable	25
Weather	21
Footwear	17
Earth formation	18
Fruit	24
Clothing	26
Insect	22

Images For images to feed into selected vision models, an image dataset preprocessed by Vemuri et al (2023) was used in this paper. The dataset was initially collected from the Google Image Search package (arrrlo as cited in Vermuri et al, 2023), which included exemplar images in various natural scenes. The background of images were removed and replaced with a plain white color by the authors (Figure 1). Removal of background helped minimize any potential noises that could interfere with the model analysis, allowing it to focus solely on analyzing the model performance on prototyping exemplars itself, independent of other trivial factors. These preprocessing steps aimed to provide a clean and standardized input for vision models, reducing potential distractions and variations posed by diverse backgrounds.

The resulting image dataset consisted of 6286 images distributed across the aforementioned 27 categories and 555 subcategories (Table 1). Each subcategory contained a minimum of 2 images and a maximum of 46 images. Notably, each image in the dataset was left in its original size during data preparation. This decision was made to accommodate the varying input size requirements of different CNN models considered in this study. The resizing of images was deferred until just before feeding the images into each specific model during the experiment phase.

Figure 1: Example images of exemplars

Vision Model Selection and Initialization

In the selection of vision models, we explored a range of Convolutional Neural Network (CNN) architectures taking two main factors into consideration.

First, our selection included pre-trained Torch vision models from various timelines. This decision was made to investigate whether the degree of the typicality effect exhibited by the models evolves or regresses as the model architecture undergoes advancements.

Second, we included models of varying sizes in our analysis. By comparing the results obtained from CNN models with different sizes, we aimed to evaluate the potential impact of model size on its ability to simulate human cognition.

A final set of Torch vision models, selected with a careful consideration of both temporal and size-related factors, is detailed in the following. Each selected model was originally designed for an image classification task. That is, given an image, the model outputs a classification label. Inspired by Vermuri et al.'s work (2023), the very last layer of each model was removed to obtain a raw feature vector of the input object returned by the model, aligning with the specific objectives of this study.

AlexNet Alexnet was first introduced in "ImageNet Classification with Deep Convolutional Neural Networks" by Krizhevsky, Sutskever and Hinton (2012). The initial architecture of the Alexnet consisted of five convolutional layers and three fully-connected layers, with a total of 60M parameters. Over time, the Torch implementation of AlexNet underwent a few refinements as outlined by Krizhevsky (2014), aiming to speed up the training by parallelization techniques, but the overall architecture of the model remains the same.

VGG VGG, introduced in "Very Deep Convolutional Networks for Large-scale Image Recognition" (Simonyan & Zisserman, 2015), represents a convolution network designed by stacking multiple small convolutional filters (3x3), coupled with subsequent max-pooling layers. Simonyan and Zisserman proposed two configurations for VGG architecture, both of which contain three fully connected layers, while one (VGG-16) has 16 weight layers and the other (VGG-19) has 19 weight layers. In this study, VGG-16 with 138M parameters was utilized which achieved a top-5 validation error of 7.5%.

ResNet ResNet is a convolutional neural network proposed in the paper "Deep Residual Learning for Image Recognition" by He et al (2016). It is designed to address challenges of training an extremely deep network, by introducing the concept of residual learning framework. Residual learning is implemented in the training phase to skip one or more layers utilizing *shortcut connections*, mitigating challenges associated with vanishing gradients. Such a ResNet architecture consists of a series of residual blocks each of which contain multiple convolutional layers. Diverse variants were also introduced by the authors with varying depths ranging from 34 layers to 152 layers. ResNet-18 and ResNet-50 are selected for the purpose of this study which involve approximately 12M and 26M parameters respectively.

SqueezeNet SqueezeNet is a compact CNN architecture that has achieved a comparable accuracy to AlexNet, but with 50 times fewer parameters (Iandola et al., 2017). SqueezeNet is structured with two convolutional layers and 8 Fire Modules. Fire Modules, introduced by Iandola et al.(2017), consists of a squeeze convolution layer, which is a set of 1x1 convolutional filters, and an expand layer, which is a combination of 1x1 and 3x3 convolutional filters, with the number of filters varying across layers. With suggested model architecture, a total of 1.2M parameters were required before pruning and 0.4M parameters after pruning. SqueezeNet v1.1 was specifically chosen for the current study, which demands 2.4 times less computation than v1.0 while retaining the accuracy.

DenseNet DenseNet was developed out of a recent finding that models with "shorter connections between layers close to the input and those close to the output" exhibit better accuracy and efficiency (Huang et al., 2018). In the proposed architecture, each layer has direct connections to its subsequent layers. That is, not only the feature map of its immediate predecessor, but also that of all preceding layers serve as its inputs of a layer, constructing a densely connected set of layers called a *dense block*. The DenseNet consists of an initial convolutional layer followed by a varying number of dense blocks, each of which is connected by a transition layer of one convolutional layer and a pooling layer. In this study, three variations of DenseNet with varying complexities were selected; DenseNet-121 (8M params), DenseNet-169 (15M params), and DenseNet-201(20M params).

ShuffleNet V2 ShuffleNet V2, presented in the paper "ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design" (Ma et al., 2018), stands out as an efficient as well as accurate CNN architecture. The *channel shuffle* operation is introduced to improve the accuracy, which facilitates information exchange between groups of channels. ShuffleNet V2 is composed of convolutional layers, a fully connected layer and so-called *stages*. Stage blocks consist of DenseNet blocks explained earlier. In this study, ShuffleNet V2 1.0x was utilized, leveraging its efficiency with a redacted number of parameters.

ConvNeXt ConvNeXt emerged as a response to the performance advancements made by Vision Transformers, such as Swin Transformers, which surpassed standard ConvNets in terms of its scalability, efficiency and accuracy in early 2020s (Liu et al., 2022). Proposed by Liu et al. (2022), ConvNeXt was constructed upon the architecture of standard ConvNets, with an intent to design an innovative architecture that competes state-of-art vision transformers while preserving the simplicity of the standard ConvNets. Four variants of ConvNet were introduced, with varying sizes; a tiny version with an unknown number of parameters, a small version with 22M parameters, a base version with 87M parameters, and a large version with 306M parameters. For the purpose of this study, the base model was selected taking computational costs into account ensuring the balance between efficiency and accuracy. The classification accuracy of the base ConvNeXt model on ImageNet-1K dataset was 82.0% which was not significantly lower than the larger model (82.6%).

A comprehensive list of models and their variations, along with the corresponding weight initialization and sizes, used for this study can be found in the following table (Table 2 and Table 3).

Table 2: Model selection and weights initialization

Model	Weights
AlexNet	AlexNet_Weights.DEFAULT
VGG-16	VGG16_Weights.DEFAULT
ResNet18	ResNet18_Weights.DEFAULT
ResNet50	ResNet50_Weights.DEFAULT
SqueezeNet	SqueezeNet1_1_Weights.DEFAULT
DenseNet-121	DenseNet121_Weights.DEFAULT
DenseNet-169	DenseNet169_Weights.DEFAULT
DenseNet-201	DenseNet201_Weights.DEFAULT
ShuffleNet v2	ShuffleNet_V2_X1_0_Weights.DEFAULT
ConvNeXt	ConvNeXt_Base_Weights.DEFAULT

Table 3: Year of model introduction and size

Model	Year	Size (# params)
AlexNet	2014	60M
VGG-16	2015	138M
ResNet18	2016	11.7M
ResNet50	2016	25.6M
SqueezeNet	2017	1.2M (before pruning)
DenseNet-121	2018	8M
DenseNet-169	2018	≦ 15M
DenseNet-201	2018	≦ 20M
ShuffleNet V2	2018	2.3M
ConvNeXt	2022	87M

Experiment Set-up

The experiment was carefully carried out in a series of steps in order to thoroughly examine the typicality effects exhibited by the selected CNN models. Throughout the experiment, input images were fed into the models, and the corresponding typicality ratings of each model were obtained. Typicality rating of each subcategory was computed by evaluating cosine similarity between its feature vector and the prototype vector of its parent category. Cosine similarities were employed to precisely quantify the extent to which the feature vector of each subcategory resembled the overall prototype vector derived from the model's outputs. The typicality ratings acquired through these steps align with the representation of human typicality ratings, falling within the range of 0.0 and 1.0. By utilizing cosine similarities, instances perceived as typical by the model exhibit ratings closer to 1.0, and instances perceived as atypical have ratings closer to 0.0. This pattern also aligns with the pattern observed in the human data, facilitating the comparison of the model data with the human data.

Various approaches were explored throughout the experiment while generating prototype vectors for each category, aiming to emulate the human cognitive process through which humans generate prototypes for distinct categories. In the initial attempt, the prototype vector was derived by averaging feature vectors of all subcategories falling under that category, as suggested in Vermuri et al.'s paper (2023). Subsequently, in the following attempt, a more nuanced approach was implemented, where outliers were identified and removed before calculating the average. The second approach aimed to construct a prototype vector that more accurately captures the essence of a category, minimizing the impact of outliers. We hypothesized the second approach would enhance the model's ability to replicate the human cognition, particularly in the context of generating prototypes, based on an intuitive understanding that humans tend to generate prototypes based on their typical perceptions of a category, rather than being heavily influenced by atypical instances.

For comparison of model typicality ratings with the human typicality ratings, Spearman correlation statistics were carefully reviewed. Spearman correlation serves as a

measure to quantitatively analyze the degree of association between the model's perception of typical instances and that of humans.

Results

Spearman Correlation between Typicality Ratings by Human and 10 Pre-trained CNN Models

The spearman correlation statistics obtained across all 10 models of our choice is detailed in Table 4. Every model demonstrated a positive mean correlation, indicating a certain degree of alignment between the model's typicality ratings and human typicality ratings. However, despite the positive trend of mean correlations, the overall degree of correlation between the model and human ratings was significantly low.

The maximum correlation was observed in ShuffleNet V2 with 2.3M parameters, while the minimum correlation was observed in ResNet50 with 25.6M parameters. However, no notable correlation was observed between the degree of correlation and the number of parameters. While models with fewer parameters (e.g., AlexNet, ShuffleNet) generally seemed to perform better than larger models, no direct, consistent correlation was evident, as exemplified by inconsistent trend of performance observed across three variations of DenseNets.

Furthermore, there was no correlation between the model's architectural advancements and its ability to replicate human cognition, especially in context of the typicality effect. In Table 4, models were listed in chronological order, yet no significant trend in mean correlation statistics was observed down the list. This lack of consistent trend in mean correlations suggests that the model complexities or advancements in its efficiency and accuracy does not play a significant role in determining a model's alignment with human cognition.

Table 4: Spearman correlation statistics

Model	Mean	StdDev
AlexNet	0.1351	0.2125
VGG16	0.0760	0.2464
ResNet18	0.0825	0.1861
ResNet50	0.0150	0.1837
SqueezeNet	0.1403	0.2300
DenseNet121	0.0948	0.2333
DenseNet169	0.0878	0.2295
DenseNet201	0.0958	0.2255
ShuffleNet V2	0.1433	0.2445
ConvNeXt	0.1118	0.2437

Extended Analysis on Correlation between Human and ShuffleNet V2 Typicality Ratings

While none of pre-trained CNN models of our interest demonstrated a remarkable alignment with human typicality effect, a closer examination was conducted to explore the disparities in how humans and models perceive typical and atypical instances. ShuffleNet V2, which exhibited the highest correlation with human typicality ratings, was chosen for this detailed analysis.

Table 5 presents the orders of the exemplars in the category Bird, sorted from highest to the lowest typicality ratings for both human and ShuffleNet V2 model. For simplicity, we considered the first half (17) of responses as typical and the second half (16) as atypical instances from the human and model perspectives.

Notably, only 11 exemplars (bolded) out of 33 fell into the same typical or atypical group as perceived by both humans and the model. This implies that humans' perception of typical instances mostly did not agree with the model perception.

Table 5: Order of typical-atypical bird instances of human(left) and a pre-trained ShuffleNet v2 model (right)

Human	ShuffleNet v2
bluejay	flamingo
robin	swan
eagle	oriole
cardinal	turkey
sparrow	penguin
hawk	**ostrich**
crow	**chicken**
parrot	canary
pigeon	goose
hummingbird	**duck**
dove	**bluejay**
chicken	owl
finch	black bird
duck	**parrot**
raven	seagull
woodpecker	vulture
ostrich	**robin**
owl	woodpecker
wren	**parakeet**
parakeet	hummingbird
seagull	**falcon**
flamingo	eagle
goose	cardinal
falcon	raven
mocking bird	finch
penguin	**swallow**
black bird	**mocking bird**
oriole	hawk
canary	**wren**
swan	crow
swallow	dove
turkey	sparrow
vulture	pigeon

We hypothesized that this misalignment may have stemmed from an inaccurate prototyping procedure that failed to capture the nuances of humans' category prototypes. Hence, an alternative approach in generating the

prototype vector was undertaken and investigated, as explained in the earlier section. All other steps of the experiment remained unchanged.

Surprisingly, even with the implementation of a new approach, the mean correlation across all categories exhibited only a modest improvement, reaching 0.1461 (compared to previous mean correlation of 0.1433). While this slight improvement might suggest a positive impact of the revised prototyping approach, it falls short of providing conclusive evidence for its effectiveness.

Conclusion

While the positive mean correlations between human typicality ratings and model typicality ratings (Table 4) indicate some level of agreement between human cognition and CNN model's behavior in context of typicality effect, the low overall degree of correlation with lack of a clear pattern with model features highlights the complexity of capturing human-like cognitive processes in existing pre-trained models.

Limitations of this study lie in the complex nature of human cognition. Human cognition involves an intense level of abstraction and complexity that can sometimes be affected by physical, environmental, or even mental factors. Therefore, it becomes clear that none of the existing models can perfectly simulate intricacies of human cognition.

Discussion & Future Work

The small improvement, while not definitive, observed with a modification in methodology used to generate prototype vectors emphasizes the need for continued refinement in such methodologies and the exploration of potential factors that affect the alignment between human and machine cognition. That said, future researchers should focus on addressing existing gaps discussed in this paper. Besides, the exploration of evaluation metrics beyond spearman correlation statistics may help.

References

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Krizhevsky, A. (2014). One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997*.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015, July). Deep Neural Networks Predict Category Typicality Ratings for Images. In *CogSci*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. *arXiv preprint arXiv:1602.07360*.

Saleh, B., Elgammal, A., & Feldman, J. (2016). The role of typicality in object classification: Improving the generalization capacity of convolutional neural networks. *arXiv preprint arXiv:1602.02865*.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Ma, N., Zhang, X., Zheng, H. T., & Sun, J. (2018). Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*.

van Dyck, L. E., Kwitt, R., Denzler, S. J., & Gruber, W. R. (2021). Comparing Object Recognition in Humans and Deep Convolutional Neural Networks-An Eye Tracking Study. *Frontiers in neuroscience*, *15*, 750639. https://doi.org/10.3389/fnins.2021.750639

Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Vemuri, S., Shah, R., & Varma, S. (2023). *Evaluating typicality in combined language and vision model concept representations.*