A comparison of deep learning models and human vision

1st Prithviraj Pawar
College of Computing (of Aff.)
Georgia Institute of Technology (of Aff.)
Atlanta, USA
prithviraj.pawar@gatech.edu

***Abstract*—Deep learning models have found to be highly successful in achieving state-of-the-art results for image recognition tasks. Since AI started as a field to represent human's cognitive behavior, it would be interesting to learn how cutting edge models perform on certain tasks that are mundane to humans. This paper explores the connections between certain deep learning models and human cognition especially in the area of vision. It starts by giving an overview about AI, deep learning and cognitive science and describes vision transformers in brief. This project aims to compare these models with humans by building a computational model using vision transformers and convolutional neural network trained on CIFAR10 dataset. The results of these models are compared to humans using various metrics. Finally it also explores similarities in architecture between transformers and human brains. The learning of the paper has implications on building more robust models that are closer to humans not only in a laboratory environment but also in real life complex tasks.**

***Index Terms*—cognitive science, deep learning, artificial intelligence, neural networks, biological vision**

I. Introduction

This document tests deep learning architectures namely CNN and Vision Transformers in terms of neuro-visual cognition. Deep learning belongs to an even broader topic of Artificial Intelligence that has strong ties to cognitive science. Moreover I believe that the field of AI provides some explanations and tools to test cognitive science hypotheses. It provides insights of the brain function and neurocognition using deep learning architectures. This document focuses on deep learning architectures in vision and image recognition and their relation to human vision and tries to answer following questions:

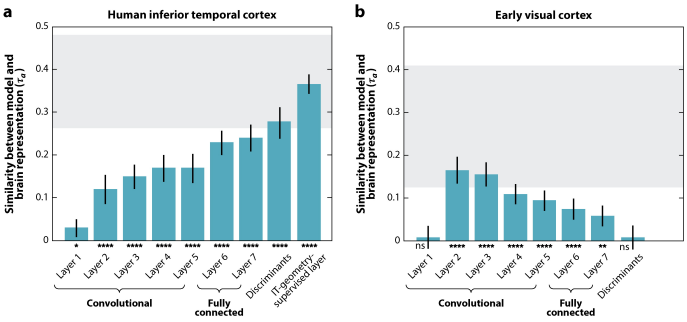
- How does a vision transformer, a deep learning model compare to human vision?
- Which of the deep learning algorithms produce human-like robustness in vision after image transformations like blurring, texture manipulation, shape deformation, etc?

Since AI is an emerging field in recent years specifically in the image processing domain, we must check the robustness of its models and how they can produce human-like results. I feel we must analyze how the state-of-the-art models in deep learning are closer to human cognition. I want to continue on this path and explore additional robustness in these architectures and how they relate to neurocognition. In cognitive science vision is considered a complex task involving inference, object detection and past knowledge. Using a computer for the same tasks that performs similar to humans is considered a great achievement and hence it excites me to explore this field. As per Langley; AI studies high-level cognition and many ideas in intelligent systems like knowledge representation, planning, natural language, learning are inspired from cognitive psychology. In the early days of AI, researchers focused on modeling human intellectual behavior which led to the fields such as HCI however more recently the focus is shifted on designing models that are more suitable for machines to solve like neural networks. Initially there was much focus on knowledge representation, reasoning and building symbolic structures that can form comprehensive theories of mind but now statistical approaches are prevalent which give great performance on narrowly designed tasks but no insight into human intelligence. We must use the study of cognitive systems as heuristics to search for better theories of mind and AI systems can play a key role when we understand more about their performance. AI can help in understanding what type of data needs to be collected in cognitive science experiments and it can give a lot of insight in representation and organization of knowledge in memory. AI can improve accuracy for natural language processing and common sense reasoning by training agents with relevant knowledge and then programming them as per intelligent algorithms. Deep learning models are dominated by neural networks which are based on the neural synapses in the brain. They can now identify objects in images, translate text and extract context from the text translate languages and also defeat humans in various complex games. These systems are based on biological brains and use only computations that are identified in human brains. They also provide tools to test cognitive theories [1]. We not only need to build networks that can explain brains behavior and human responses but also analyze the representations and parameters in the model. DNNs have received great attention in visual object and face recognition and are able to reach immaculate results. There are still questions on how can we explain results of a network to study brain behavior and come with a theory of mind.

II. Related Work

Deep learning and cognitive science are highly correlated and many studies have compared deep learning with human brains. Most importantly the state of the art Convolutional neural networks (CNNs) are inspired by research in biological

vision [2]. They are considered successful models in computer vision and state-of-the-art models related to neural vision. Experiments applying neuro-modularity effects of visual attention on CNN blocks have shown to increase their performance and its observed that these models are able to incorporate behavioral and neural effects of perceptual learning. Some studies [3] have also considered deep learning as a model of biological vision. These studies compared recurrent neural networks and CNNs with human brains based on their internal representation and performance levels. Its observed that these models follow the connectionist paradigm and use representational spaces that are more similar to those of the inferior temporal cortex of the brain. This correlation confirms that computer vision can learn from biological vision. The layers of neural networks explain visual and inferior temporal representations in images. Its also considered that the brain might rely on a combination of neural networks mainly feedforward and recurrent. There



AR Kriegeskorte N. 2015.
Annu. Rev. Vis. Sci. 1:417–46

Fig. 1. Similarity between model and brain representation

has also been research around how can we improve robustness of CNNs by including complex blurring techniques [4]. In human vision most of the light cast into retina is degraded and blurring is a natural process of human vision. Hence training CNNs on blurred inputs can bring them closer to human vision and it has seen to improve accuracy compared to vanilla CNNs. This also enriches our understanding of biological visual system and asserts the importance of blurry visual experiences. More recently vision transformers [5] are considered state-of-the-art in image recognition and much of the computer vision which are based on the transformer architecture [6] originally proposed for NLP problems. Transformers excel in modelling longer dependencies between input and output sequences and also support parallelism which makes training easier [7]. Transformers have been used in a myriad of image tasks namely image recognition, image and scene generation using generative modeling and learning unsupervised representations. They are also used in image restoration algorithms to preserve texture and edges information. Not these but transformers have also been used in multi-modal tasks like visual question answering, visual commonsense reasoning and image captioning. Since images are widely available using datasets like ImageNet [8], we can exploit transformer ability by training on large scale data. The effective features of vision transformers are due to their receptive fields which is in turn a result of self attention [7]. The self attention layer can flexibly attend to multiple sequence of pixels. Some studies have shown that ViT are naturally robust to occlusion and natural perturbations. If we properly train vision transformers on shape-based features then they can achieve image recognition capabilities to near-human accuracy. There have been similar attempts to this paper's work to compare ViTs to human vision [9]. There is a natural comparison of results or performance of model to human's performance but some have also compared using error metrics like error overlap, class-wise JS distance and Inter-class JS distance. Transformers are shown to be very close to human vision. So in this paper I would like to extend the work and analyze how they relate to cognitive science concepts.

III. Experiment Design

A. Transformers

Transformers are deep learning models based on neural networks consisting of multiple layers. It uses attention to improve its performance in NLP tasks and is used more universally in recent times. In NLP the model uses Encoder Decoder architecture [6] which accepts an input sentence, encoding its context and maps it to the output labels as per training and fine tuning of the weights. The different layers of transformer contribute to producing accuracy. The input embedding layers are of two types positional and one hot or word embedding. The word embedding is used to reduce dimensions of input vectors which consist of individual pixels or input words based on the task at hand. The positional embedding maintain position or context of the input vectors. This layer is important since transformers have no convolution unlike its predecessor CNN. Hence in order to maintain sequence of inputs we must add about relative position of the tokens in input. These layers have learn-able parameters so during neural network training the weight matrices get optimized and the model learns the positions after huge amount of training. The next layer is the Encoder Block which consists of normalization, multi-headed attention and multi layer perceptron (MLP) sub-layers. The normalization layer prevents mean and STD of the embedding layers from changing and hence prevents unstable and slow learning. Multi-headed attention is the most important layer in a transformer which calculates a dot-product attention and then self attention. The dot product is important to understand the context of the inputs and it maps how it relates to the output labels. It carries the relevant information around an input ahead in the network so that its not lost. At the end of the encoder block we have position wise feed forward network or a multi-layer perceptron. Its used to introduce non-linearity which is a standard practice in neural network training.

B. Computational Model Design

For this research I designed a computational model that compares the results of a neural network on image recognition with the results of image recognition by humans. It will also compare error metrics which would indicate which type

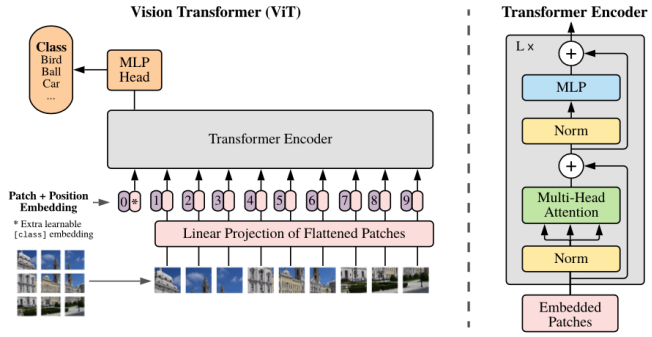


Fig. 2. ViT High level Architecture

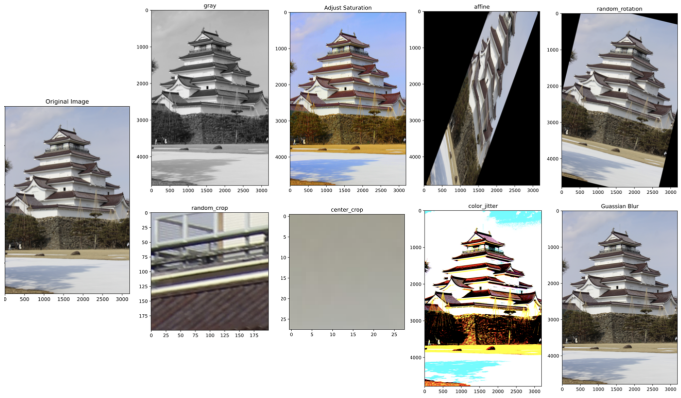


Fig. 4. Pytorch Image transformations wrt to original image

of images are difficult to recognize. Figure 1 is the rough design of the model. The image transformations block will transform the input images before training the neural network. The vision transformer is state of the art model which is based on a neural network equipped with attention [10]. This model trains CIFAR10 which is a dataset of 32x32 small images. When comparing with humans the model compares robustness, error metrics and accuracy results. The image

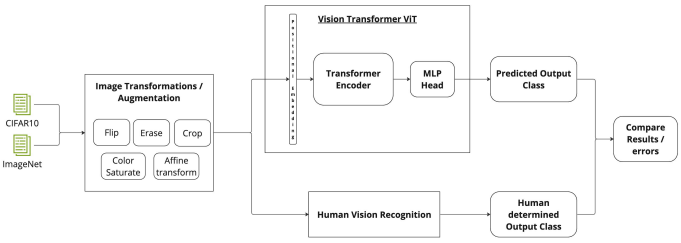


Fig. 3. A rough design of the Computational Model

transformation block of the model uses techniques from pytorch [11] such as random rotation, center crop, random crop, color jitter, grayscale, Gaussian blur, color saturation and affine transformations. Using these techniques, the images are randomly transformed as shown in Figure 3. When we use these image transformation techniques CNNs and ViTs acquire increased sensitivity towards shape information and show greater robustness. Its good to add these techniques to compare with humans and also to avoid overfitting.

C. Computational Model Implementation

1) Neural Network with Image transformation: The computational model was implemented using Jupyter notebook in python and it uses pytorch as the main library for deep learning algorithms. The models were mainly trained on CIFAR10 with image size 32, number of output classes as 10, batch size as 32 and number of epochs as 200. The learning rate for training was 1e-3 with a decay of 1e-1. The training loss was calculated using Cross Entropy Loss function and the optimizer was Adam [12]. The transformer has an embedding layer with shape 32x32 and 256 parameters. It contains positional embedding to mark relative positions of the pixels in an image. The layers have residual connections of size 256 which are added after the output of each sub layer. Its a Sequential layer that does simple operations. The Self attention layer consists of 8 head with input channels as 256 and head channels as 32. This is as per the original transformer design proposed. I wanted to use vanilla transformer to compare the performance with humans. The training time for CNNs was 11 hours and for transformers it was 48 hours.

IV. Results

A. Results of Neural Network training

Following are the results after training both the models. I observed that ViTs are more accurate and CNN suffer from over-fitting. CNNs have a natural bias towards texture and shape hence a vanilla algorithm would suffer from over-fitting while ViTs have no such bias. The test accuracy for CNN was 0.71 and for ViT it was 0.75. Also after some time CNNs accuracy had plateaued and no matter the number of epochs, it gave the same performance. Vision Transformers results are more accurate than CNNs and it has also been observed in many papers. ViT shows a clear gap between train and validation losses. Hence from this section I used ViTs as the only model to compare with humans' performance.

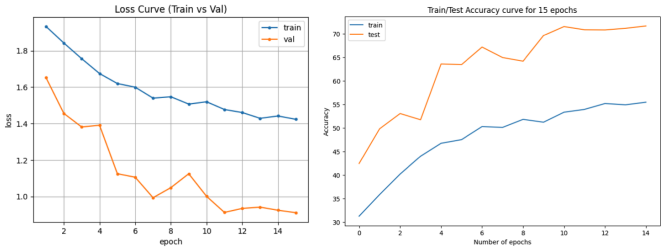


Fig. 5. CNN Accuracy and Loss curves for CIFAR10

B. Comparing Neural Network Results with Humans

This part of comparison involved result based comparison which is comparing the results produced by a vision transformer trained in previous steps on a sample dataset with the

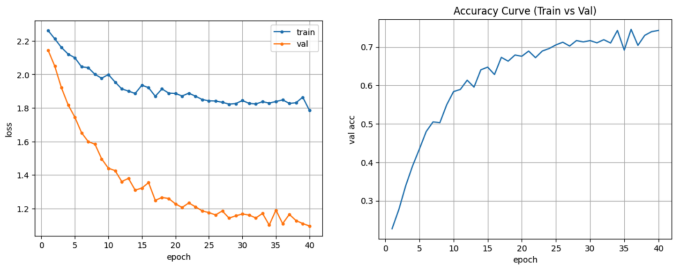


Fig. 6. Vision Transformer Accuracy and Loss curves for CIFAR10

results by humans. The task at hand for comparison ins recognizing images. In this section ideally we must ask machine and the model to identify same set of images to make a good comparison. However since I couldn't reach out to humans in real life I used similar research done by model-vs-human project [13]. The project provides a good tool to compare the results from any neural network model with humans for cognitive vision tasks. It uses pytorch and Tensorflow models on 17 datasets with high-quality human comparison data. The human comparison datra is collected under highly controlled lab conditions at Wichmannlab [14] . All these datasets have some image transformation / augmentation techniques like paramteric or binary image distortions, grayscale, contrast, high-pass, blurring, uniform noise, etc. which makes it perfect to test with my model. To use this tool I added my ViT

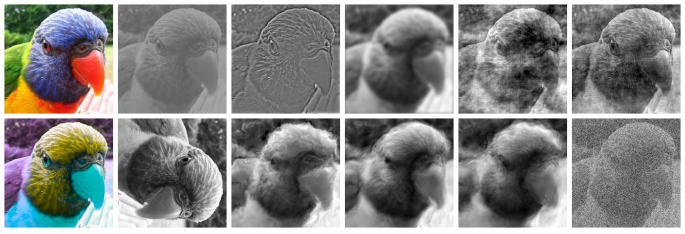


Fig. 7. Source : https://github.com/bethgelab/model-vs-human/tree/master

model using its custom APIs and then generated a report about its performance on sample datasets. It uses various metrics to compare the performance of a given model with the performance of humans. The metrics used are:

- Accuracy difference: It compares the raw accuracy of the model with human's accuracy by calculating the accuracy difference across all 17 datasets. This measures how far is our model with human's accuracy. Lower the value, closer the model to humans
- Observed consistency: It measures the percentage of samples where both human and machine get right or wrong. Higher the value, closer the model to humans. A standard value would be around 0.4 to 0.5.
- Error Consistency: It tracks if the errors are consistent on certain images which were classified incorrectly. Higher the value, closer the model to humans. A standard value would be around 0.2 to 0.3.

This comparison results generated an accuracy difference of 20.56, observed consistency of 0.0567 and error consistency of 0.0345. This indicates that my model was poorly performing compared to humans. I believe that are many reason why my model performed poorly. Some of the main reasons are:

- Not enough parameters: Since the state-of-the-art vision transformers scale to billions of parameters they are able to generate a good accuracy. My model had 6,482,138 parameters and still took 2 days to train on a regular 6-core CPU. To achieve the same state-of-the-art accuracy I would have to train for days if not months which was not feasible.
- Not enough training data: Some of the benchmark models mentioned in the project were trained on raw ImageNet dataset which takes multiple days on a highly powerful GPU. Since I had limitation of time and resources I couldn't achieve the same performance.

To circumvent the above problems I used a pre-trained ViT which is available as ViT-L [5]. This model has 14 million parameters and is trained in ImageNet and it resulted to near-human performance in some cases. It has a better OOD accuracy than humans in fact. It is almost consistent with humans except for error metrics.

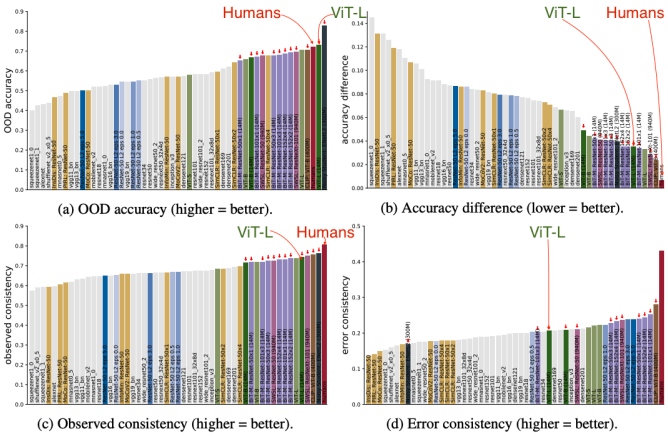


Fig. 8. The ViT-L is one of the best performing models as per this project

C. Result-based Comparison Implications

The computational model developed in this project indicates that Vision Transformers are highly accurate given the amount of training. Of course I couldn't't build a state of the art model but there are many references attached which state that Vision transformers flourish in computer vision tasks among deep learning models. Also as indicated in previous sections, that the gap is closing very fast between humans and models for basic computer vision tasks. ViTs show promising robustness and it could be improved if we train it with real life situations that human vision solves everyday. From the result-based comparison, I understood that in terms of accuracy the gap between humans and models is closing quickly. However we should consider that human vision is capable of analysing deep

patterns in an image once exposed for longer time. The state-of-the-art models are not robust enough in terms of errors and consistency. However this could be improved by training the models on huge amounts of data which involves robustness. But attaining human level robustness is not so far it seems.

D. Comparing attention of transformers with attention in human brains

Transformer architectures are highly powerful due to their attention heads. Hence to compare between human brains I considered origins, reasons and behavior of attention in human brains with transformers.

1) Attention in human brains: Our brain uses its processing power to focus on certain details in a visual scene to identify a predefined concept or understanding an intriguing thing. This is roughly the job of attention. For example, is there a person on a road and is it my friend? Certain neurophysical studies have provided insights that when we pay attention to an object using our eyes, the neural responses from that sense is enhanced while sound, taste or touch is decreased. Also, in terms of neurons firing in our brains, attention increases synchronization between firing of neurons within a particular area. However much of this work is observed in an isolated situation thereby reducing complexity of the real world tasks. Some studies have shown that a neuron's receptive field plays a key role in filtering out the stimuli. If two stimuli from different sensors are presented in a receptive field, they suppress each and the final neural response is a weighted average of the individual responses. There is a bias that attends to one of the two stimuli thereby filtering out distractions and resolving a race.Such biases are present in neurons and originate based on their evolution. However the biases are also present on a spatial level [15].

2) Attention in transformers: Attention in transformers is a result of a scaled dot product between two vectors. The purpose of this layer is to understand the context of the input vectors, in NLP it is used to predict the output words in a different language.

$$ScaledDotProduct = ||a||.||b||cos(\theta)$$

Here theta is the angle between them. It is the maximum when theta=0 and minimum when theta=180. Basically it tries to retain context of input vectors in subsequent layers and matches the same with the output context. In terms of Vision transformers, the context is present in different areas of the input image. It tries to direct the model to focus on relevant features and captures more information. Transformers come with multiple attention heads and each attention heads calculate a scaled dot product between the input image and features. Following is the representation of a single attention layer

$$Attention(Q, K, V) = AV = softmax(QK^T \div (\sqrt{dK}))V$$

Here Q is the query matrix which indicates output features multiplied by weight matrices and K is the key matrix with input features multiplied by weight matrices. V is the weighted sum of different pixels in the image that represents the context. Thus, we are projecting queries from output vectors with input features onto the context in the image. Hence

$$Q = YW^Q, K = XW^K, V = XW^V$$

Here all weight matrices W indicate how much the model should pay attention to each pixel in the image and we call them "attention weights".

E. Attention Comparison Implications

I believe that attention in human brains is quite similar to self attention in transformers. The biases present in human brains to attend to certain stimuli are similar to attention weights. The biases are learned through years of learning and evolution of our brains similar to training the transformer weights and optimizing them to yield best results. Also in human brains when there is a race among stimuli at the receptive fields, biological neurons calculate a weighted average along with neuron bias to determine the winner. This is quite similar to the weighted average calculated for V vector in self attention. As we have not yet fully uncovered the reasons behind human attention and the origin of neuron's bias, we can only hypothesize that it comes from learning. Also how does our brain improves attention by spending more time on the input image is still unclear. This is a question for future research.

V. Conclusion

Deep learning's implications in image processing, image recognition are well known and this study tried to compare these implications based on human's cognitive abilities for the same tasks. The computational model proved that vision transformers are very close to achieving near-human accuracy for image recognition and surpass many models in doing so. Also the study compared transformer's attention mechanism with human brain's attention. There are certainly similarities between the two but we need more insight about human brains to conclusively say so.

VI. Limitation

This project verified that transformers a state-of-the-art deep learning model performs similar to humans in terms of certain metrics in image recognition tasks. However Model-vs-human exposed humans to an image for around 200 ms and hence the results are closer to humans. Normally human's attention result in a visual scene increases if we spend more time looking at it. Also in day to day life our vision encounters multiple visuals at once and we are still able to recognize relevant information, backtrack in our memory and analyze certain patterns. This type of complex scenarios are yet to be explored by transformers. Currently the tasks where such models are exposed are simple compared to humans day to day tasks hence we cannot compare abilities of human visions with deep learning models. As we have not yet fully uncovered the reasons behind brain cognition like the origin of neuron's bias and behavior of receptive fields we can only hypothesize about

them and compare them with AI models. We need to study our brain more to prove the hypotheses and it can also make way for a different approach in AI. Then we need to build the deep learning models closer to human's cognitive behavior so that we can make fair comparison and also test our theories of mind with such models.

REFERENCES

[1] K. R. Storrs1 and N. K. A. A. P. J.-L. U. G. G. M. B. Z. M. B. B. I. D. of Psychology Department of Neuroscience Department of Electrical Engineering Columbia University USA, "Deep learning for cognitive neuro-science," *https://arxiv.org/pdf/1903.01458.pdf*, 2019.

[2] G. W. Lindsay, "Convolutional neural networks as a model of the visual system: Past, present, and future," *Gatsby Computational Unit/Sainsbury Wellcome Centre, University College London*.

[3] N. Kriegeskorte, "Deep neural networks: A new framework for modeling biological vision and brain information processing," *Annual Review of Vision Science*, 2015.

[4] H. ang and F. Tong, "Improved modeling of human vision by incorporating robustness to blur in convolutional neural networks." *bioRxiv*, 2023.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[7] S. H. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *CoRR*, vol. abs/2101.01169, 2021. [Online]. Available: https://arxiv.org/abs/2101.01169

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[9] S. Tuli, I. Dasgupta, E. Grant, and T. L. Griffiths, "Are convolutional neural networks or transformers more like human vision?" *CoRR*, 2021.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[11] "https://pytorch.org/vision/main/transforms.html," *Pytorch Image Augmentation*.

[12] J. B. Diederik P. Kingma, "Padam: A method for stochastic optimization," *conference paper at the 3rd International Conference for Learning Representations, San Diego*, 2015.

[13] R. Geirhos, K. Narayanappa, B. Mitzkus, T. Thieringer, M. Bethge, F. A. Wichmann, and W. Brendel, "Partial success in closing the gap between human and machine vision," 2021.

[14] Http://www.wichmannlab.org/.

[15] M. V. Peelen and S. Kastner, "Attention in the real world: toward understanding its neural basis," *Trends in Cognitive Sciences*, vol. 18, no. 5, pp. 242–250, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364661314000473