MOONEY TASK: MODELING HUMAN ONE-SHOT PERCEPTUAL LEARNING THROUGH DISAMBIGUATION

A PREPRINT

Xujin "Chris" Liu[*†] **Ayaka Hachisuka**[‡§] **Jonathan Shor**[‡§] **Lavender Yao Jiang**[*¶]
Yao Wang[†] **Biyu J. He**[‡‖**††] **Eric K. Oermann**[*††**]
{chris.liu, lavender.jiang, yaowang}@nyu.edu
{ayaka.hachisuka, jonathan.shor, biyu.he, eric.oermann}@nyulangone.org

September 30, 2024

ABSTRACT

Recent advances in deep neural networks (DNNs) suggest that they can function as powerful models of human and animal behavior and neural function. Thus, many studies have harnessed DNNs to model visual perception in the human brain in terms of task performance and internal representations. In this study, we propose the Mooney task, adapted from a one-shot perceptual learning task in humans. We develop a series of DNN models that apply temporal attention to the features generated by different backbone models that operate on individual images only. Our results are threefold. Firstly, our DNN models can replicate the behaviors observed in human volunteers without the need for task-specific backbone model finetuning. Secondly, we find ShuffleNet V2-based models to have the most brain-like representations overall. Thirdly, we find that representations in the temporal layers of our model, rather than those in the spatial layers, are more similar to low level and high level visual regions after one-shot learning. This suggests that recurrent neural mechanisms are involved in Mooney effect. Taken together, our study demonstrates the potential of using DNN's to model human perceptual learning, and improve our understanding of its mechanisms. [9]

1 Introduction

Using models to test hypotheses about the brain is an important part of computational neuroscience. Recently, deep neural network (DNN) models have shown promising results at both achieving human-like performance on various tasks [23] as well as modeling complex brain representations in humans [28]. Recent modeling studies of primate visual cortex [12] have shown that recurrent computation is required to efficiently capture the representations of the visual cortex. Indeed, studies have found that the visual cortex performs recurrent computations even when visual stimuli is static [13]. All of these studies to date, however, use static images and simple image classification, and thus the role of DNNs in modeling the recurrent human perceptual learning effects remain open to be explored. Here, we propose to use DNNs to model one-shot perceptual learning using Mooney images. We propose to model one such effect, the Mooney effect [9], using DNN models. The Mooney effect is when subjects learn to recognize ambiguous images in a one-shot

[*]Department of Neurosurgery, NYU Langone Health, New York, NY
[†]Electrical and Computer Engineering Department, NYU Tandon School of Engineering, New York, NY
[‡]Neuroscience Institute, NYU Grossman School of Medicine, New York, NY
[§]Vilcek Institute of Graduate Biomedical Sciences and Neuroscience Insitute, NYU Grossman School of Medicine, New York, NY
[¶]Center for Data Science, NYU, New York, NY
[‖]Departments of Neurology, NYU Grossman School of Medicine, New York, NY
[**]Department of Neuroscience & Physiology, NYU Grossman School of Medicine, New York, NY
[††]Department of Radiology, NYU Grossman School of Medicine, New York, NY
[9]Our training and model code is available at `https://anonymous.4open.science/r/MinimalMooneyNet-8D81/`, visualization code available at `https://anonymous.4open.science/r/MooneyPlotting-80CC/`.

learning manner. This effect is intriguing due to the persistence of learned perceptual gains, and a unique independence from episodic memory [24]. We propose the Mooney task, an adaptation of the human psychophysics experiment for studying the Mooney effect, for DNN modeling as a test bed to model human one-shot perceptual learning.

One-shot perceptual learning refers to the improvement in an organism's ability to interpret sensory inputs after being exposed to a single example or stimulus. For instance, being able to recognize a tiger after encountering it just once is an example of this effect. The Mooney Faces Test [19] is a widely recognized experiment that demonstrates one-shot perceptual learning. In this experiment, an ambiguous (binary) image is presented, followed by several clear images, with one of them being the clear version of the ambiguous image. The participant is then asked to identify the original ambiguous image, to determine if the brief exposure to the clear image can disambiguate the original one. Figure 1 illustrates an example sequence of the stimuli used in the human experiment described in [9]. It used 33 gray scale images (and their binarized versions), selected from Pascal VOC dataset [4] and Caltech 101 dataset [5].

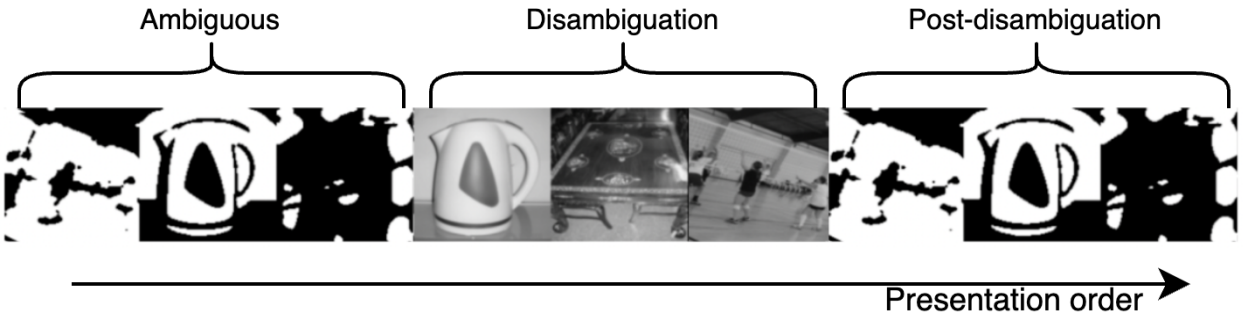


Figure 1: An example Mooney task sequence

One would expect the Mooney effect to be mediated by episodic memory, however the underlying mechanism is interestingly not episodic memory. In a key study, subjects with medial temporal lobe lesion and deficits forming episodic memories still display the Mooney effect [24], which is shown to persist for up to months or even a lifetime after their exposure to the clear image [17]. Prior works postulated that this effect is mediated by prior knowledge ("priors") that are learned from exposure to the original images. An important question then is: what kinds of mechanisms are involved in the task? Recent studies suggest that content-specific representations [6] across the entire cortical hierarchy contribute to this effect. To further understand the neural mechanism behind this effect, we employ DNN modeling.

We create models for this task by combining a spatial vision backbone with temporal transformers. The vision backbone processes images at each timestep independently, and the temporal transformer takes the visual features at each time step up to the current time and updates the current features. This simplistic 2-stage architecture divides the work of spatial information processing and temporal information integration. We construct multiple models by using many vision backbones. We exploit this separation of function to answer following questions about the Mooney effect:

1. Are task-specific spatial representations necessary for Mooney task?
2. What kinds of visual backbones generate the most brain-like representations?
3. Which brain regions are similar to which layers?

We formulate the first question about task-specific representation as varying the training process of our models. We use pretrained, then finetuned vision backbones to model task-specific spatial representations in the brain; we use pretrained, then frozen vision backbones to model general purpose spatial representations that are learned from non-Mooney experiences. Our experiments show that task-specific spatial representation is not required to achieve good Mooney task performance or produce high brain similarity, although finetuning still affects the distribution of representation similarities across brain regions.

The second question can be answered by inspecting which models perform the best in terms of representation similarity. We define measures of representational similarity in later sections. Understanding which type of visual models produce the most brain-like representation can help us test hypotheses about the computation of interest in Mooney effect. We show that models based on ShuffleNet V2 achieve highest overall brain similarity among all the models that we investigated.

We ask the third question to understand better which regions of the brain our model is similar to. Since our models are divided into specialized temporal and backbone layers, they might help us understand the function of regions they are most similar to. Our similarity analysis shows that overall, the temporal layers are more similar to high level regions of

the brain; before and after the presentation of grayscale images, this higher similarity extends to low and high level visual regions as well. This shows that widespread recurrence might underlie the Mooney effect. Based on these observed similarity patterns, we provide a possible theory of the neural mechanism underlying the Mooney effect.

2 Related works

Brain-score [22] proposes a similarity measure that consists of representational and behavioral similarities to monkey subjects with implanted electrode arrays in the late visual cortex. The representation similarity is measured by how well the DNN representations can be used to predict the brain representations; the behavioral similarity is measured by subject's classification and the model's classification. The stimuli presented to the subjects are a large quantity of static grayscale images of naturalistic objects. An important finding of this work is that brain score seems to correlate with model performance on ImageNet ([21]).

Sensorium [27] is a similar benchmark that explores DNN model's representational and behavioral similarity to large scale electrode recordings of rats, in response to static images. It also uses regression error to estimate representation similarity between models and rat brain, but the rat behavior is based on the motions of the rat instead of classification. A similar benchmark, Algonaut [8], evaluates representational similarity between DNN models and human brain as measured from fMRI. The stimuli shown is static, colored scene images taken from COCO dataset [14].

Our work shares a similar goal as existing benchmarks that aim to model the visual perception of human brain. The main difference from them is that we try to model a narrowly defined phenomenon, the Mooney effect, that involves perceptual learning, instead of image classification tasks.

3 Methods

3.1 Mooney task

We adapted and simplified the experiment setup in [9]. We present a sequence of black and white images divided into three phases: pre-phase, grayscale-phase, and post-phase. The images shown in pre-phase and post-phase are Mooney images (binarized), while the grayscale phase presents grayscale images. We display 6 images per phase, and thus 18 images in total for each trial. The grayscale images can be presented in a different order than the pre and post images, and there is a 10% chance that it is replaced by an unrelated, randomly sampled image. A shortened version with 3 images per phase is shown in Fig. 1. The images shown in each phase was randomly chosen from 33 preselected images belonging to 20 classes. The images in these datasets are similar in that they are all medium size images with a clearly defined core object. The classes covered by these datasets include animals, artifacts, and everyday objects.

For each input image sequence, our models are trained to classify the current image conditioned on all previously shown images in the sequence. We train DNN models using images from the ImageNet1k dataset [21]. To generate the Mooney images, we first convert color image to grayscale by setting RGB intensities to the average of all colors. Second, we resize image to 128x128. Third, we apply a Gaussian blurring to make the edge smoother. Finally, we apply a global binarization at 50% percentile intensity level. The pre-phase images serve as a control before the display of the grayscale image and set a baseline for model performance. If the model properly associates the pre-phase and post-phase image with its grayscale-phase counterpart, then the classification accuracy in post-phase (post accuracy) will be higher than the pre-phase accuracy (pre accuracy).

Since the vision backbone models all have different abilities of extracting visual features, the grayscale accuracy for a particular model is the upper-bound of post accuracy for that model. Then, if the post accuracy becomes identical to grayscale accuracy, we can assert that the model has integrated all the priors possible. We formally define the metric *normalized disambiguation strength* (NDS) in equation 1, to measure the strength of disambiguation. As it measures the difference of accuracy between phases, it is a task performance that measures the holistic behavior of the network's overall ability to integrate priors, instead of focusing on the absolute performance in each individual phase alone. We use it as the measurement of behavioral similarity to the humans, since the humans display very strong Mooney effect.

$$\text{NDS} = 100\% \times \frac{\text{Post accuracy} - \text{Pre accuracy}}{\text{Gray accuracy} - \text{Pre accuracy}} \tag{1}$$

3.2 DNN models and training

Our general architecture is divided into two parts: the vision backbone and the temporal attention, shown in Fig. 2. The vision backbone processes images presented at each time step independently; the temporal attention processes a

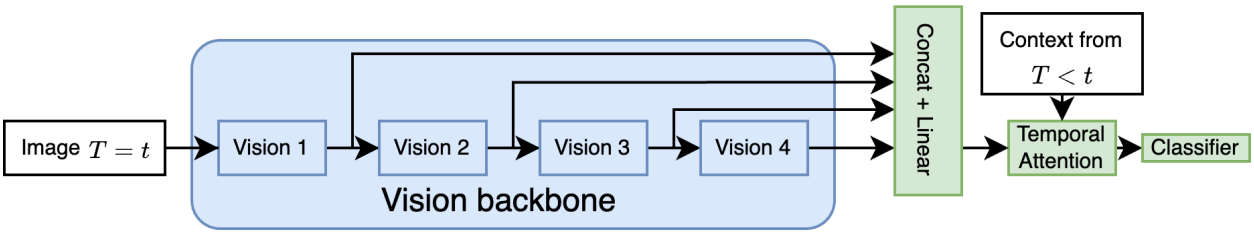


Figure 2: General model architecture

Table 1: Vision backbone used

Model type	Models
CNN	ResNet [10], WideResNet [29], ConvNext [16]
Mobile CNN	MobileNetV3 [11], ShuffleNetV2 [18], EfficientNet V2 [25]
Visual attention	BEiT-V2 [20], CoATNet [2], Swin Transformer [15]

sequence of visual features and is responsible for integrating the prior information. We deliberately choose this structure so we can separate the portion of the model that processes visual information (the backbone) and sequential information (the temporal attention).

Vision backbones are common pretrained CV models that take in images and generate spatial visual features. We list the models we used in table 1. We can divide any vision backbone into 4 parts to extract visual features of different levels. For example, ResNet's layer 1-4 would correspond to vision 1-4 in our formulation. The output of these 4 parts are then spatially pooled and concatenated together to form the final extracted visual features.

The temporal attention block is a standard pre-norm variant of transformer encoder. The visual features from a current time step t attends to the features from all previous time steps by treating the current feature vector as the query, and the feature vector from the current and each previous step as the key. The output feature at the current step is followed by a linear layer to compute the logits for each class. We use a hidden size of 768, 8 heads, 0.1 dropout, and 10 layers for the temporal attention. Since the model shouldn't see inputs from the future, we apply a causal mask when calculating attention weights. We have experimented with RNN style models such as GRU [1] and LSTM [7], but these models ended up learning identity functions despite our hyperparameter tuning. We thus only report transformer-based results.

We use pytorch, and train on 2 seeds for all the models on a single A100 GPU on our in-house cluster with varying batch sizes to minimize the total negative log likelihood loss. Note that when we use pretrained vision backbones, the different random seeds will not affect the parameter of the backbone itself. We warm up the learning rate from 0 to 1e-4, and then use a cosine cycle throughout the training with a maximum learning rate of 1e-4 with 4 cycles. We train for a maximum of 700 epochs, and evaluate the model on the checkpoint that has lowest total validation loss. When generating training sequences, we randomly sample images from ImageNet1K dataset and construct the pre, grayscale, and post-phase sequences as descried in 3.1. For each training epoch, we sample 50,000 such sequences.

The pre, grayscale, and post phase accuracies are evaluated on ImageNet1K, which has 1000 classes. For the human experiments, there was a total of 20 classes. It might seem that the DNN models are performing a harder task; however, it is important to note that the humans do not know what classes of images would be presented before the trials begin, and so their response space is any possible object. However, we assume the chance of a successful random guess from our models or the human subjects to be equivalent.

3.3 Measuring model's brain similarity

To measure the similarity between different layers of a model and a brain region, we use a method called representation similarity analysis (RSA) [3] in keeping consistent with the existing literature. Since the representation from the brain and the DNN layers are different, we cannot directly compare them. Instead, RSA compares systems with different internal representations based on the (dis)similarities of their representations to a common set of stimuli.

We calculate the so-called representation dissimilarity matrix (RDM) as a second-order descriptor of a system's representational behavior. For a function (either a DNN layer or a region of the brain) f, we denote its response to input x as $f(x)$. Then given a set of N inputs $X = \{x_0, x_1, ...x_{N-1}\}$, we find a $N \times N$ matrix D^f that collects the pairwise

Table 2: Rough cortical hierarchy and functions of brain regions. Please note that the hierarchy is not a well-defined measure, but rather an approximate idea of computation precedence. Their commonly known functions are also highly context-dependent.

Region name	Commonly known functions
V1-V4	Visual feature extraction. Low level visual region.
Fusiform Gyrus (FG)	Face, landmarks, familiar objects. High level visual region.
Lateral Occipital Cortex (LOC)	Object Detection & Recognition. High level visual region.
Frontal Parietal Network (FPN)	Task-relevant attention, spatial processing, working memory
Default Mode Network (DMN)	Self-referential, autobiographical memory, social learning

correlation distance between $f(x_i)$ and $f(x_j)$:

$$D^f_{i,j} = 1 - \rho(f(x_i), f(x_j)),$$

where ρ is Pearson's correlation. Since D^f is a symmetric matrix, we can simply take the flattened lower triangle of it, $d^f \in [0, 2]^{(N^2-N)/2}$, as a description of system f's response to X. We can calculate such a representation descriptor for any system. For example, d^f can represent the descriptor of a ResNet's 3rd layer f to a set of images, or the measured neural activity in a brain region f. Having a pair of representation descriptors enables us to apply correlation again to compare the representation similarity between two systems. We use ρ_a, a version of Spearman's correlation with random tie breaking. For example, given d^f and d^{V1}, their correlation $\rho_a(d^f, d^{\mathrm{V1}})$ tells us how similar the functions f's and V1's representations are. [10]

Since we are studying the change of representation before, during, and after the exposure to grayscale images, we treat the same "origin" image in the 3 different phases as 3 different images. Because the Mooney experiment used 33 different grayscale images (common between the human experiment and DNN inference), we have $N = 99$ in the RDM we generated for the brain and DNN model. The first 33 images correspond to the pre-phase images, the next 33 the post-phase images, and the last 33 the grayscale-phase images. Because the order of images in each phase might affect the representation, both in the human trials and the DNN model evaluations we repeat each "origin" image for multiple times with different permutations and average the resultant representation over these repetitions. In our experiments, we compute all the representation descriptors $d^l, \forall l \in \{\mathrm{V1}, \mathrm{V2}, ...\}$, across a range of brain regions of interest as investigated in [9] for each subject. We include a short description of each major region investigated in table 2. To determine the similarity of a model layer with a brain region, we calculate the correlations of the model layer's RDM with the RDM of each subject in this region, and then average the correlations.

Let $L = \{1, 2, 3, ...\}$ indicate the set of layers in the DNN model M. If we want to evaluate which layer of a model M has the best representational correspondence with a certain brain region, e.g. V1, we simply find $\mathrm{argmax}_{l \in L}\, \rho(d^l, d^{\mathrm{V1}})$. We call this the best fit layer. In order to evaluate an entire model M's similarity to a brain region such as V1, we use the best fit layer's similarity to that brain region: $S(M, \mathrm{V1}) = \max_{l \in L} \rho(d^l, d^{V1})$. When we evaluate the whole brain similarity of a model, we simply take the average S over all regions of the brain. Although our DNN models are trained on ImageNet1k, we compute DNN models' similarity to human brains using the same set of images as the human experiments.

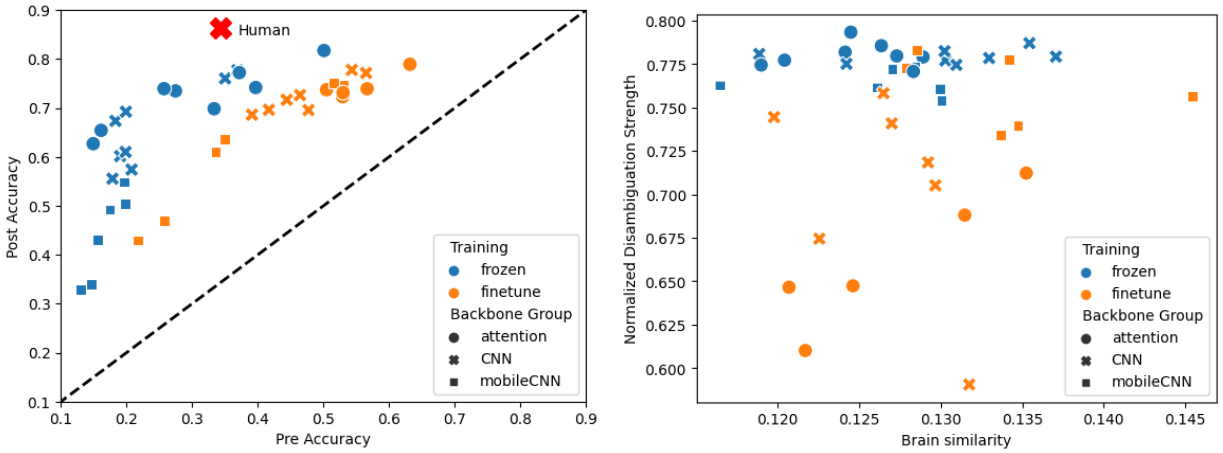
When evaluating the brain similarity of a subset of a model's layers, we perform similar steps as above, but we take the max operation over the desired subset. For example, if we want to find a model's temporal layers' similarity to a brain region, instead of taking the max similarity over all the layers, we take the max similarity over only the temporal layers.

3.4 Experiments

We train DNN models (described in 3.2) using the Mooney task (described in section 3.1) and evaluate the brain similarity (described in section 3.3). For each model, we either freeze the backbone or finetune the backbone. After these experiments are done, we evaluate the similarity between each pair of model layer and brain region. We further bin constituent regions of the brain together for easier interpretation: V1-V4 are summarized as low level visual areas, FG and LOC are summarized as higher visual areas, and we keep the same binning for FPN and DMN, as shown in Table 2.

[10] We use `https://github.com/rsagroup/rsatoolbox` python toolkit to compare similarities.

4 Results



(a) Task performance of models

(b) Model task performance (NDS) vs. brain similarity

Figure 3: Brain similarity and task performance results of all trained models. (a) All models have a higher post accuracy than pre accuracy, indicating that they have integrated the priors in grayscale images. (b) Frozen models have higher task performance in terms of NDS, but have similar range of brain similarity as the fine tuned models.

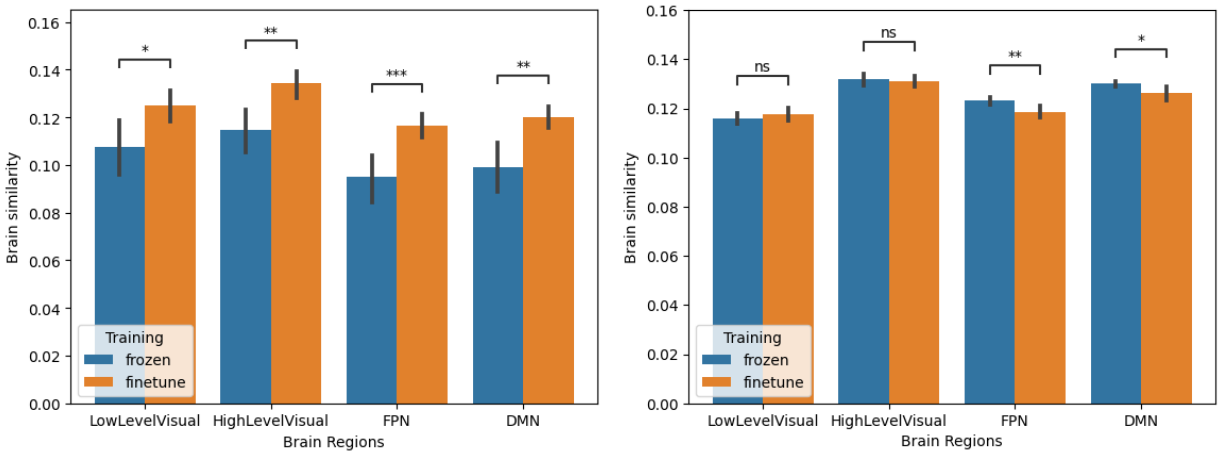
All DNN models are able to incorporate priors from grayscale images. Figure 3a shows the performance of DNN models before and after the exposure to the grayscale images. All models have a higher post accuracy than pre, suggesting that the models are able to extract and integrate the priors in the grayscale phase. In general, the pre accuracy of the model correlates well with the post accuracy. The human performance (red X in the figure) has a higher post accuracy than all of the models we train, but a lower pre accuracy than some of the best DNN models, especially those where the vision backbone was refined. These models likely adapted to the binary Mooney images during finetuning. Some models from the frozen category, however, can outperform humans in terms of pre accuracy without adapting the backbone to the binary images. Interestingly, the strongest performing models in pre accuracy are also the strongest in post accuracy. Figure 3b shows the task performance (in terms of the NDS) and brain similarity of all the models that we trained. Although the brain similarity correlation magnitude is not high, all models display statistically significant whole brain similarity (i.e. correlation greater than 0 with small p-values). In general, there is no obvious correlation between models with high brain similarity and models with high task performance.

Task-specific spatial representations are not necessary for the Mooney task, but it alters the distribution of brain similarity. Generally, finetuned models do not perform better than frozen models in terms of NDS ($77.54\% \pm 0.056$ for frozen models, $71.11\% \pm 0.010$ for finetuned models) or brain similarity (0.13 ± 0.0064 for frozen models, 0.13 ± 0.0052 for finetuned models). This shows that task-specific spatial representations are indeed not necessary for achieving a higher Mooney task performance or brain similarity. Through a more granular similarity analysis, Fig. 4a shows that finetuning the backbone layers increases their correlation to all brain regions significantly. However, Fig. 4b shows that finetuning reduces the temporal layers' similarity to FPN and DMN significantly. This redistribution of similarity results in a similar overall brain similarity between finetuned models and frozen models.

Table 3: Top 5 backbone models in terms of brain similarity

Backbone Model	Task performance	Brain similarity
Finetuned Base ShuffleNet V2	75.65%	$\mathbf{0.145 \pm 0.0065}$
Frozen Large WideResNet	77.92%	0.137 ± 0.011
Frozen Large ResNext	$\mathbf{78.69\%}$	0.135 ± 0.010
Finetuned Base CoAtNet	71.22%	0.135 ± 0.0077
Finetuned Base EfficientNet V2	73.94%	0.134 ± 0.0052

ShuffleNet V2-based models produce the most brain-like representations. Table 3 shows that the top model in terms of the similarity to the brain is based on ShuffleNet V2. Besides this, we find the top 5 models covers both frozen and finetuned models, and include different types of architecture from CNN, attention and mobile CNN categories. It is



(a) Similarity of backbone layers to each brain region

(b) Similarity of temporal layers to each brain region

Figure 4: Brain region similarities (*denotes significant difference, n.s. denotes no significant difference. Mann-Whitney U test with two-sided alternative.). **(a)** Finetuned models' backbone layers have higher similarity to all brain regions compared to frozen models. **(b)** Finetuned models' temporal layers have lower similarity to FPN and DMN.

worth noting that although shufflenet V2-based models produce the most similar representations, their task performance is not the highest.

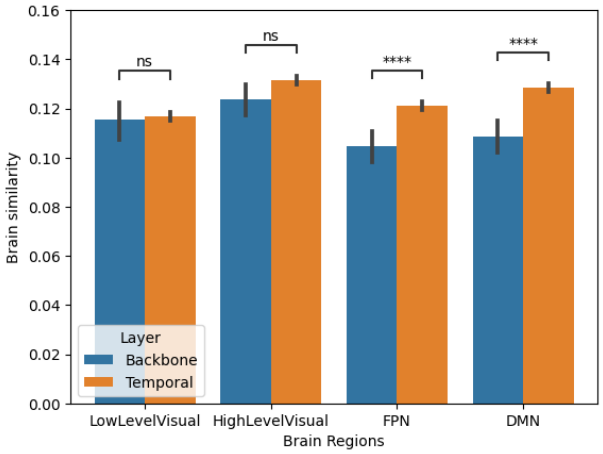
Overall, temporal layers are more similar to the deeper regions of the brain and have comparable similarity to visual regions, compared to the backbone layers. Seen in Fig. 5a, evaluating the models' brain similarity using only either the temporal layers or the backbone layers, we find that the temporal layers are significantly more similar to FPN and DMN compared to the backbone layers. For the low and high level visual regions, we observe that the two groups of layers have negligible difference in terms of similarity. There is a slight increase in similarity in high level visual region, although not statistically significant.

Backbone layers are more similar to the brain during grayscale phase, and temporal layers are more similar to the brain during pre and post phase. Figure 5b shows that backbone layers are significantly more similar than temporal layers to each region in the brain when evaluated on grayscale phase only. There is a gradient of decreasing similarity as the region becomes more high level. For the temporal layers, it has the highest brain similarity with the high level visual region, but still significantly less similar than the backbone layers. However, if we evaluate on pre phase images only, as shown in 5c, we see that the temporal layers are significantly more similar to low and high level visual region and DMN. A similar trend exists when we evaluate similarity on post phase. Figure 5d shows that temporal layers' similarity to low level visual region and FPN are significantly higher than backbone layers.

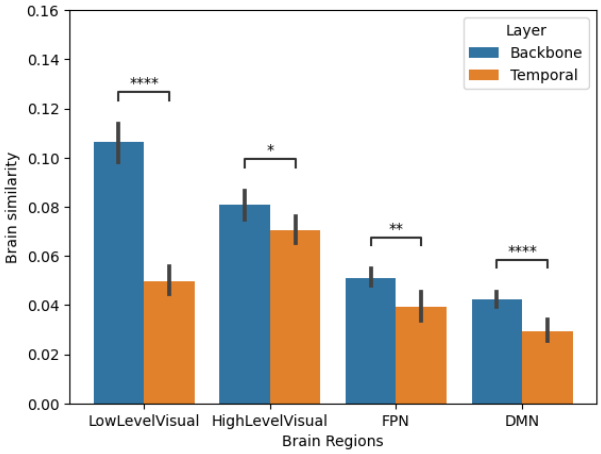
5 Discussion

Our results suggest that there could be recurrent computation underlying Mooney effect. In our brain similarity analysis, we found that the temporal layers are more similar than or as similar as the backbone layers in general (Fig. 5a,5c,5d). Further, while we do not expect low-level visual regions (V1-V4) to show similarity to temporal layers because they are understood as convolution filters, Fig. 5d shows that V1-V4 are more similar to the temporal layers than the backbone layers before and after one-shot learning. This is consistent with current evidence of widespread recurrent computations across the visual cortex [12, 13], as well as the previous works studying Mooney effect [9]. This consistency between our findings and previous evidence implies that there might also be widespread recurrent computation underlying Mooney effect.

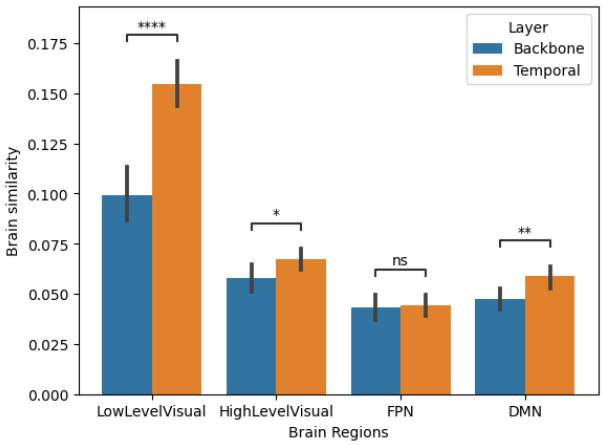
Based on the different similarity patterns during grayscale phase and post phase, we have a working hypothesis that the recognition of grayscale image is largely a process driven by feedforward visual processing, with high level visual regions accumulating information which are then used in the post phase to disambiguate. During the grayscale phase (Fig. 5b), backbone layers of our model is more similar to the entire brain, while temporal layers are mainly similar to the high level visual region; in the post phase (Fig. 5d), the low level visual region similarity of temporal layer is higher than that of the backbone layers. We form the following hypothesis to explain these observations. The recognition of grayscale image is largely a process driven by feedforward visual processing (causing the gradient of higher backbone



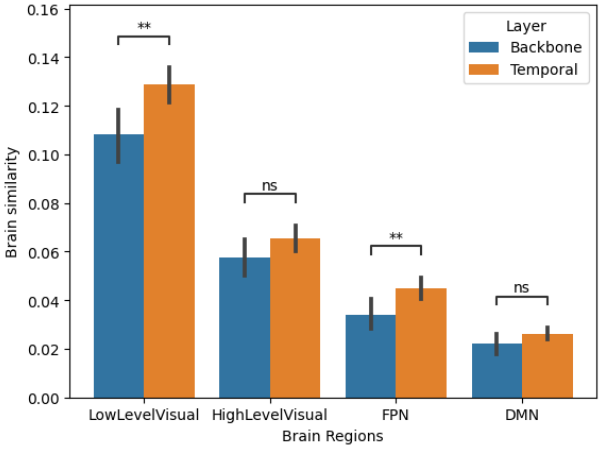
(a) Similarity to each brain region in all phases



(b) Model similarity breakdown in grayscale phase



(c) Model similarity breakdown in pre phase



(d) Model similarity breakdown in post phase

Figure 5: Granular breakdown of brain region similarities of finetune & frozen models (*denotes significant difference, n.s. denotes no significant difference. Mann-Whitney U test with two-sided alternative.). **(a)** Overall, temporal layers are significantly more similar to FPN and DMN compared to backbone layers. No significant different in low and high level visual regions. **(b)** In grayscale phase, backbone layers are significantly more similar to the brain. **(c)** In pre phase, temporal layers are more similar to the low level visual regions, high level visual regions, and the DMN. **(d)** In post phase, temporal layers are significantly more similar to FPN and low level visual regions.

similarity), with high level visual regions accumulating information (causing the high temporal layer similarity to high level visual regions). In the post phase, the information stored during the grayscale phase is used to modulate the representations from low level visual regions (causing the high temporal layer similarity to low level visual region), making one-shot learning possible. However, this theory does not directly explain why temporal layers are also more similar to low level visual region in the pre phase, as no information should have been accumulated in the brain or the model at this point. It's possible that the lack of expectation is also encoded in a similar way in the lower visual region and the temporal layers.

Certain architectural choices in ShuffleNet V2 might be important in producing brain-like representations. ShuffleNet V2 [18] is a convolution neural network developed to optimize inference and parameter efficiency in trade-off with the classification performance. This optimization goal coincides with the evolutionary necessity of the human visual cortex to have fast processing speed, and thus it's possible that this common goal drives a convergent representation. However, we note that no other models from the mobile CNN category produce this level of similarity. Thus, the general practice of optimizing inference speed itself is not likely to be the cause for high brain similarity. Looking into its network architecture, one unique operation introduced by the ShuffleNet family of model is the channel shuffle operation. Instead of applying convolutions on all channels at the same time, the convolutions are divided into smaller

groups, and then feature maps are shuffled among different groups. Although not explicitly designed to do so, this operation could be modeling structures in the brain such as hypercolumns, where a pattern of dense small clusters coupled with inter-cluster communication is common [26].

6 Limitations and future works

We caution the use of DNN models in computational neuroscience. Although we have wide knowledge about optimizing their performance, we don't fully understand their inner workings. Due to the ease of producing satisfying metrics, too much emphasis on modeling without physical validation might easily send us chasing after phantoms. Before physical validation is warranted, one might alleviate this by surveying comprehensive architectures to rigorously evaluate claims. We attempt to do this in our study, but due to the constraints of compute, scale is still limited. A larger scale studying across a larger space of model might reveal more robust results.

Trained DNNs do not explicitly model the real-time synaptic weight changes in human brain. The Mooney effect is likely encoded in synaptic weight changes in the brain due to its temporal persistence. The analogy of this in DNN models would be that the model weights are updated on- or offline upon the integration of these priors. However, in our computational models, the priors integrated do not alter the weights in the model, only the activations generated during the inference of one sample sequence. It is unclear how this affects the brain similarity of DNN models. One way to bridge this gap is to develop online weight updates to model the synaptic changes. Another consequence of this difference is that the priors accumulated in human brain is persistent through time, whereas the priors accumulated by our models are only available throughout the inference of one sample sequence. Again, an online update of model weight might be able to bridge this gap.

7 Conclusion

In this work, we proposed to use DNN to model Mooney effect, a perceptual learning effect seen in human that involves one-shot learning of ambiguous images through disambiguation. We showed that our baseline models consisting of vision backbones and temporal attention displays Mooney effect and have significant brain similarities. We found that task-specific spatial representations are not necessary for modeling the Mooney effect. We found that temporal layers have a stronger similarity to high level brain regions throughout the process of visual learning, and comparable similarity to low level regions. We also discovered that ShuffleNet V2 backbone produces the most brain-like representations. Our work provides promising results to improve our understanding of human one-shot perceptual learning.

References

[1] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-Decoder approaches. Sept. 2014.

[2] Z. Dai, H. Liu, Q. V. Le, and M. Tan. CoAtNet: Marrying convolution and attention for all data sizes. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3965–3977. Curran Associates, Inc., 2021.

[3] J. Diedrichsen and N. Kriegeskorte. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Comput. Biol.*, 13(4):e1005508, Apr. 2017.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.*, 88(2):303–338, June 2010.

[5] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.*, 106(1):59–70, Apr. 2007.

[6] M. W. Flounders, C. González-García, R. Hardstone, and B. J. He. Neural dynamics of visual ambiguity resolution by perceptual prior. *Elife*, 8:e41861, Mar. 2019.

[7] F. A. Gers, J. Schmidhuber, and F. Cummins. Learning to forget: continual prediction with LSTM. *Neural Comput.*, 12(10):2451–2471, Oct. 2000.

[8] A. T. Gifford, B. Lahner, S. Saba-Sadiya, M. G. Vilas, A. Lascelles, A. Oliva, K. Kay, G. Roig, and R. M. Cichy. The algonauts project 2023 challenge: How the human brain makes sense of natural scenes. Jan. 2023.

[9] C. González-García, M. W. Flounders, R. Chang, A. T. Baria, and B. J. He. Content-specific activity in frontoparietal and default-mode networks during prior-guided visual perception. *Elife*, 7:e36068, July 2018.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[11] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le. Searching for MobileNetV3. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1314–1324, Oct. 2019.

[12] K. Kar, J. Kubilius, K. Schmidt, E. B. Issa, and J. J. DiCarlo. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat. Neurosci.*, 22(6):974–983, June 2019.

[13] T. C. Kietzmann, C. J. Spoerer, L. K. A. Sörensen, R. M. Cichy, O. Hauk, and N. Kriegeskorte. Recurrence is required to capture the representational dynamics of the human visual system. *Proc. Natl. Acad. Sci. U. S. A.*, 116(43):21854–21863, Oct. 2019.

[14] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. Lawrence Zitnick, and P. Dollár. Microsoft COCO: Common objects in context. May 2014.

[15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, Oct. 2021.

[16] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, June 2022.

[17] R. Ludmer, Y. Dudai, and N. Rubin. Uncovering camouflage: amygdala activation predicts long-term memory of induced perceptual insight. *Neuron*, 69(5):1002–1014, Mar. 2011.

[18] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun. ShuffleNet v2: Practical guidelines for efficient CNN architecture design. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part XIV*, pages 122–138, Berlin, Heidelberg, Sept. 2018. Springer-Verlag.

[19] C. M. Mooney. Age in the development of closure ability in children. *Can. J. Psychol.*, 11(4):219–226, Dec. 1957.

[20] Z. Peng, L. Dong, H. Bao, Q. Ye, and F. Wei. BEiT v2: Masked image modeling with Vector-Quantized visual tokenizers. Aug. 2022.

[21] 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. Sept. 2014.

[22] M. Schrimpf, J. Kubilius, H. Hong, N. J. Majaj, R. Rajalingham, E. B. Issa, K. Kar, P. Bashivan, J. Prescott-Roy, F. Geiger, K. Schmidt, D. L. K. Yamins, and J. J. DiCarlo. Brain-Score: Which artificial neural network for object recognition is most Brain-Like? Jan. 2020.

[23] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, Oct. 2017.

[24] L. R. Squire, J. C. Frascino, C. S. Rivera, N. C. Heyworth, and B. J. He. One-trial perceptual learning in the absence of conscious remembering and independent of the medial temporal lobe. *Proc. Natl. Acad. Sci. U. S. A.*, 118(19), May 2021.

[25] M. Tan and Q. Le. EfficientNetV2: Smaller models and faster training. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10096–10106. PMLR, 2021.

[26] D. Y. Ts'o, M. Zarella, and G. Burkitt. Whither the hypercolumn? *J. Physiol.*, 587(Pt 12):2791–2805, June 2009.

[27] K. F. Willeke, P. G. Fahey, M. Bashiri, L. Pede, M. F. Burg, C. Blessing, S. A. Cadena, Z. Ding, K.-K. Lurz, K. Ponder, T. Muhammad, S. S. Patel, A. S. Ecker, A. S. Tolias, and F. H. Sinz. The sensorium competition on predicting large-scale mouse primary visual cortex activity. June 2022.

[28] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U. S. A.*, 111(23):8619–8624, June 2014.

[29] S. Zagoruyko and N. Komodakis. Wide residual networks. In E. R. H. Richard C. Wilson and W. A. P. Smith, editors, *Procedings of the British Machine Vision Conference 2016*, number 87. British Machine Vision Association, 2016.