TEACHING CNNS TO MIMIC HUMAN VISUAL COGNITIVE PROCESS & REGULARISE TEXTURE-SHAPE BIAS

Satyam Mohla[†], *Anshul Nasery, Biplab Banerjee*

Indian Institute of Technology, Bombay, India

✉ satyammohla@iitb.ac.in

ABSTRACT

Recent experiments in computer vision demonstrate texture bias as the primary reason for supreme results in models employing Convolutional Neural Networks (CNNs), conflicting with early works claiming that these networks identify objects using shape. It is believed that the cost function forces the CNN to take a greedy approach and develop a proclivity for local information like texture to increase accuracy, thus failing to explore any global statistics. We propose CognitiveCNN, a new intuitive architecture, inspired from feature integration theory in psychology to utilise human-interpretable feature like shape, texture, edges etc. to reconstruct, and classify the image. We define novel metrics to quantify the "relevance" of "abstract information" present in these modalities using attention maps. We further introduce a regularisation method which ensures that each modality like shape, texture etc. gets proportionate influence in a given task, as it does for reconstruction; and perform experiments to show the resulting boost in accuracy and robustness, besides imparting explainability to these CNNs for achieving superior performance in object recognition.

Index Terms— attention, bias, texture-shape, cognitive, explainable.

1. INTRODUCTION

CNNs, considered as the computational model for primate visual system [1, 2], have been shown to exhibit representation hierarchy in terms of feature selectivities of edges, shapes and objects in early, mid and deep level units. The fact that complex objects and shapes appear after edges in intermediate layer activation visualisations of CNNs seem to support empirically a theoretical understanding of interpretable selectivities [3, 4], also in agreement with the shape bias observed in experiments with children [5].

However, recent works demonstrate texture bias as the reason for the superior performance of CNNs [6]. Similar conclusions were drawn in [7], where texturised images of dogs were correctly classified, even when global statistics were highly distorted. It seems that CNNs, in order to maximise accuracy, greedily learned to use texture to solve the problem and thus failed to learn the global features relevant for the task. [6] attempts to reduce this bias by training an ImageNet pretrained CNN with a stylised texture image dataset. The method although novel, is ad-hoc and does not address the underlying problem of greedy learning in CNNs. Moreover, such techniques are difficult to apply in tasks where any non-domain training data can lead to possible distortion and thus to loss in accuracy & robustness, or where the image data is inherently of low quality.

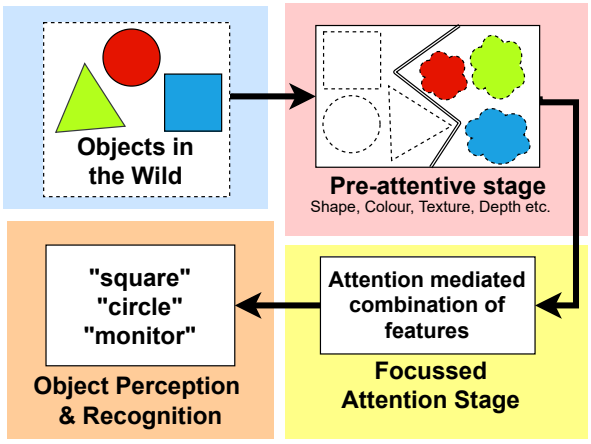

Fig. 1: Stages of FIT based object recognition.

In this paper,we propose methods to utilise self-attention for quantifying the texture-shape bias in CNNs in an intuitive & interpretable manner. Furthermore, we introduce new metrics to regulate bias & demonstrate resulting gains in performance & robustness in object recognition.

2. FEATURE INTEGRATION THEORY (FIT):

In cognitive psychology, FIT refers to an attention model which suggests that when perceiving objects, we first synthesise and separate features in an automatic & parallel way, directing attention serially to each item in turn afterwards [8], as shown in Figure 1. This has been supported by many experiments [9–11].The features isolated in pre-attentive stage include shape, colour, size, curvature, lines etc. [12].

FIT provides a novel inspiration to combat our problem: we provide different feature selectivities as input to the network, emulating the pre-attentive stage. Our model can

[†]Corresponding Author. Currently at Digital Transformation Supervisory Unit, Honda Innovation Lab Tokyo.

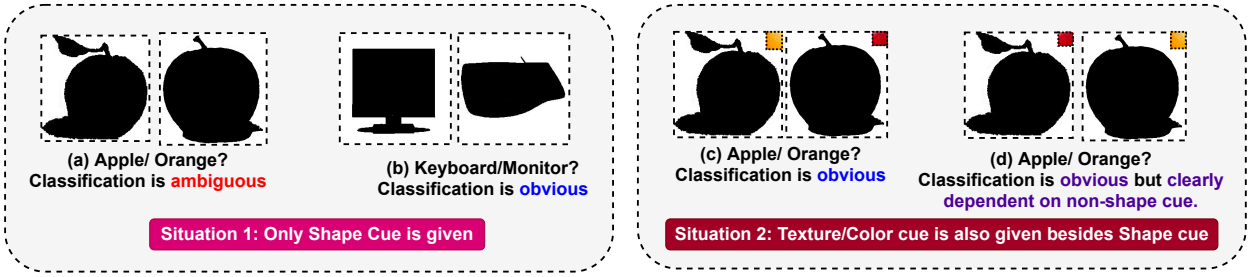

Fig. 2: Gedankenexperiment: Importance of Shape Cue vs Shape + Texture cue for varying classes

thus train with the knowledge of various features like texture, shape and edges and explore more avenues to maximise performance.

2.1. Gedankenexperiment

Let us begin with a thought experiment. Consider humans tasked with classifying a pair of silhouette of objects, for example, orange/apple & keyboard/monitor as in Fig 2. In situation 1, where only shape cue is given, the classes are somewhat ambiguous in 2(a), but obvious in 2(b). In situation 2 however, where additional information of texture/colour is given, a human can easily classify orange/apple.

Note that in case of classifying orange/apple silhouette, no information was imparted by the silhouette (shape) to human. It was the texture/colour cue that assisted us make the decision. To confirm, consider 2(d), which has opposite texture cue as to 2(c). Does the result of your classification reverses when the texture cue reverses too?

Essentially the CNN, just like humans, should learn to give different relevance to different modalities of information depending on the task. Feature Integration Theory (FIT) has been experimentally observed in various studies & thus provides the inspiration for our investigation.

3. EXPERIMENTAL SETUP

3.1. Preprocessing

Let the original dataset be $\{x, y\}_{i=1}^{m}$ where x is an image and y is the associated label. Further, let $f_1, f_2, ..., f_n$ be a set of *feature transform based classical image processing algorithm* which can be applied in preprocessing stage to each x to extract a new modality. In our case, f_1, f_2, f_3, and f_4 are instantiated to extract shape, texture, greyscale image and edges respectively, & each resulting $f_i(x)$ is a new modality. Additional modalities are generated similarly as in [6]:

Greyscale: Images are processed in Matlab and converted to greyscale using `skimage.color.rgb2gray`

Silhouette: These are generated by thresholding colour images on white background. As such the outermost contour is interpreted as the "perceived" shape, which is how human perception would also interpret the object when looked.

Edges The edge representation for the image is generated in Matlab using `edge(I, 'Canny')`

Texture We utilise the interpretation of [6] to define texture as repetition: Many repeated 'things' become 'stuff' [6, 13]. We utilise [14], to generate texture classically (to ensure deterministic & reproducible outputs) as shown in Fig4.

3.2. Architecture

Next, we describe our model's architecture and training method. We represent our model by the tuple $(F_i(x)_{i=1}^{i=4}, F_{rec}, F_{pred})$ where each $F_i(x)_{i=1}^{i=4}$ acts as a modality encoder for corresponding $f_i(x)$, converting it to a latent vector z_i. F_{rec} represents the network tasked with reconstructing the original image, and F_{pred} represents the network for predicting the labels for a given set of input modality tuple. G_i are decoders tasked to assist encoders F_i to learn the latent representation of modality in autoencoder setting, as in Fig 3.

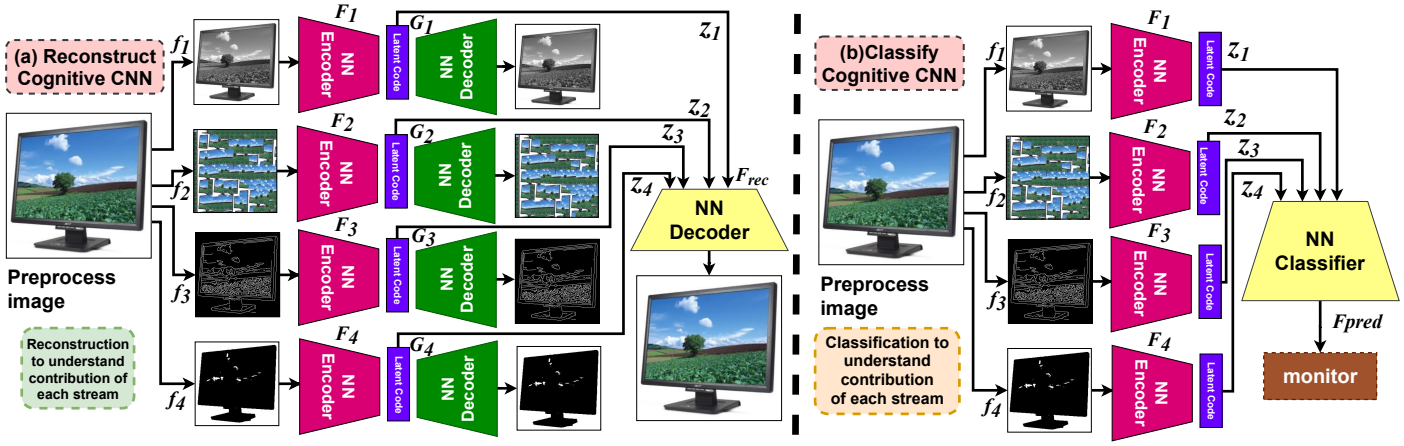

Fig. 3: Experimental Setup: FIT model adapted to CNN for quantification and regularisation

For the modality encoder $(F_i(\theta_i))$, we used a SegNet backbone [15], and for generating attention maps (described in Section 4), we used a 2-layer Conv Network (with 128 filters, kernel size of 2) taking the entire concatenated latent vector(outputs of block 5 of SegNet) as input & generating attended vectors as output. Lastly, to predict the label, we feed these attended vectors through a BatchNorm & Dense layer with softmax activation.

3.3. Training

In the first stage of training, we train our modality encoders $F_i(\theta_i)$ with learnable parameters θ_i and reconstruction network F_{rec} to reconstruct the original image from the given modality encoders. The input to F_{rec} are the concatenated latent vectors $z_1, z_2, ..., z_n$ (Figure 3 (a)). The purpose of this part of training is to tune the modality encoders, and to gauge the "relevance" of each modality in the reconstruction for the image. The reconstruction also confirms the assumption that all the information of the image is captured in these four modalities. Formally, this stage of training can be summarized as in equation (1):

$$\underset{\theta_1,\theta_2,...,\theta_n,\theta_{rec}}{\arg\min} \mathcal{D}\Big(F_{rec}\Big(F_1\big(f_1(x)\big), F_2\big(f_2(x)\big), ... \\ ..., F_n\big(f_n(x)\big)\Big), x\Big) + \lambda \sum_{i=1}^{n} \mathcal{D}\Big(G_i\big(F_i\big(f_i(x)\big)\big), f_i(x)\Big) \tag{1}$$

where $\mathcal{D}$ represents the pixel-wise Euclidean distance between original and reconstructed images.
Once the networks have converged, we train the prediction network F_{pred} to predict the label of each input given its latent vectors $z_1, z_2, ..., z_n$. (Figure 3 (b)). Formally, this stage aims to find $F_{pred}(\theta_{pred})$ as in equation (2):

$$F_{pred}(\theta_{pred}) \leftarrow \underset{(\theta_{pred})}{\arg\min} \mathcal{L}_{ce}\Big(F_{pred}\Big(F_1\big(f_1(x)\big), \\ F_2\big(f_2(x)\big), F_3\big(f_3(x)\big), ..., F_n\big(f_n(x)\big)\Big), y\Big) \tag{2}$$

where $\mathcal{L}_{ce}$ is the cross-entropy loss between the predicted labels and true labels. Our setup is summarized in Fig.3.
Now that we have classified or generated the image from the human interpretable features, we want to quantify the "relative relevance" among them for the different tasks, for which we use attention maps.

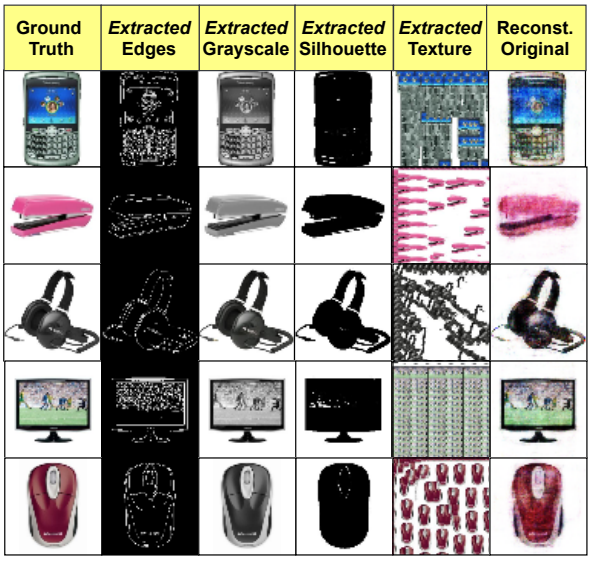

Fig. 4: Some examples of reconstructions produced by our Reconstruct Cognitive Network

4. QUANTIFYING BIAS USING ATTENTION

In this section, we introduce a self-attention based mechanism to quantify the bias in the dataset as well as our prediction network. There have been previous attempts to use attention as a tool for neural feature selection ([16], [17]). We extend this technique to utilize attention as a means to weigh the relative relevance of each modality in prediction and reconstruction, and to further regulate information flow to prediction network in an unbiased manner to make it more robust.

Let $A_{j=1}^{j=4}$ be the attention layers corresponding to the modalities. We add self-attention layers A_{pred} and A_{rec} to the network which act on the concatenated latent vectors $z_1, ..., z_n$ to give weighted vectors $\hat{z}_1, ..., \hat{z}_n$. These are then passed to F_{pred} and F_{rec} to classify and reconstruct respectively. Formally,

$$\hat{z} = \sigma(A_j(z)) \odot z, \qquad z = (z_1 \parallel z_2\|, ..., \|z_n)$$

where $\odot$ represents element wise product, z is the concatenated latent vector.

4.1. Measuring Shape & Texture Cue

It is these attention maps we use to quantify the biases. We define the measures *Reconstruction Relevant Modality Coefficient (RRMC)* and *Task Relevant Modality Coefficient (TRMC)* for each modality i for a particular example as

$$\textit{RRMC}_i(z) = \frac{\mathbb{E}(\sigma(A_{rec}(z))_i)}{\sum_{j=1}^{n} \mathbb{E}(\sigma(A_{rec}(z))_j)}$$

$$\textit{TRMC}_i(z) = \frac{\mathbb{E}(\sigma(A_{pred}(z))_i)}{\sum_{j=1}^{n} \mathbb{E}(\sigma(A_{pred}(z))_j)}$$

where $\mathbb{E}$ represents the mean of a vector over its dimensions. $\textit{RRMC}_i$ and $\textit{TRMC}_i$ represent the abstract measure of amount of "relative relevance" of a given modality i for reconstruction and prediction networks respectively. This "relative relevance" is reflected in the attention maps that the network generates, assigning maximal importance to that modality that

assists it the most; to reconstruct the image in $RRMC_i$, & towards classifying the given image in $TRMC_i$. Finally, prediction network is defined as biased if there is a mismatch in this "relative relevance" of the modality for the two tasks, namely reconstruction and prediction. i.e. when $TRMC_i$ is not equal to $RRMC_i$

4.2. Regularising Shape-Texture Bias

To control the shape-texture bias, we add a regularizer $\sum_{i=1}^{n=4} ||TRMC_i - RRMC_i||^2$ to the loss function, which forces the prediction network to give as much importance ($TRMC_i$) to a modality for a given task (herein the task is prediction), as much as it was important ($RRMC_i$) for reconstruction.

5. RESULTS

Since our tasks involve pre-processing using classical techniques, we utilised a dataset with a white background, like Amazon Office-31 dataset [18]. We perform experiments & show the efficacy of our measures & regularizer, and demonstrate gains in performance and robustness due to our method.

5.1. Reconstructions

The first stage of training involves reconstructing the image from the four modalities. The reconstructions (Figure 4) look very close to the original image, which demonstrates that (i) the chosen modalities contained all information of the original image and (ii) the current setup was able to extract all the information from these modalities into the latent code, essentially *demonstrating an empirical scheme to extract numerical representation of abstract modalities like shape, texture, edge cues etc.*, basically what we set out to achieve.

5.2. Accuracy

We trained a network using our method (Figure 3) to classify the Amazon Office-31 dataset and recorded the value of $RRMC_i$ and $TRMC_i$ for each feature, besides the accuracy. Unregularised network (4UC-CogCNN) reported an accuracy of 58.7%, while the accuracy in network (4RC-CogCNN) resulting due to regularisation (Section 4.2) increased to 61.8%, shown in Table 1.

We incorporate these ideas into our baseline CNN, and propose Cue Augmented CNN (CueAugCNN) which takes all 4 features as additional input channels alongside the image itself. We compared all the methods with a baseline CNN having the same architecture. All our 4 stream networks (4UC, 4RC, CueAugCNN) perform superior to the baseline network, with the highest accuracy being achieved by CueAugCNN at 62.5%. The results are shown in Table 2.

Table 1: Comparing relative relevance of different modalities

Stream	$RRMC_i$	$TRMC_i$ **4UC-CogCNN**	$TRMC_i$ **4RC-CogCNN**
Shape	23.7%	21.0%	24.0%
Texture	22.3%	22.8%	22.2%
Greyscale	30.4%	31.4%	30.7%
Edges	23.4%	24.6%	23.0%
Accuracy		58.7%	61.8%

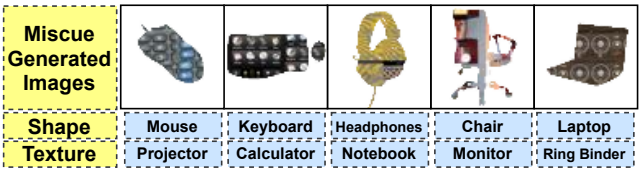

Fig. 5: Examples of Miscue Conflict

5.3. Robustness

To test for robustness, we performed a texture-shape miscue experiment as done in [6]. In order to demonstrate miscue conflict, we classically generated images by overlaying the texture of one object on the shape of an object from another class, as shown in Fig 5.

All CogCNN approaches performed consistently better in robustness than the baseline by a large margin. CueAugCNN however, based on conventional CNN architecture performed poorly, at the cost of increase in accuracy. In an ablation of CogCNN, we considered only 2 streams (shape-texture). The network still performed comparable to baseline (only 0.7% decrease in accuracy) for a huge gain in robustness. Our results are tabulated in Table 2.

Table 2: Accuracy & robustness for different models

Method	**Accuracy**	
	Original	**Miscue**
Conventional CNN (Baseline)	58.3%	14.5%
2 Stream Reg (2RC-CogCNN)	57.6%	49.3%
4 Stream Unreg (4UC-CogCNN)	58.7%	52.0%
4 Stream Reg (4RC-CogCNN)	**61.8%**	**56.9%**
CueAugmented (CueAugCNN)	**62.5%**	11.1%

6. CONCLUSION

We demonstrated an empirical scheme to extract numerical representation of abstract modalities like shape, texture, edge cues etc., imparting explainability to the model in tasks like reconstruction & classification. We developed novel metrics & regulariser to control bias between different modalities, like texture-shape bias in the network. We showed that training a CNN with human-interpretable modalities like shape/texture/edge cues, as inspired from FIT, lead to increase in accuracy & robustness against cue conflicts. Lastly, we adapted the ideas to conventional CNNs, and achieved highest accuracy. Our future work includes preprocessing the input image in-situ to present an end-to-end network so that it can be readily used on any dataset.

7. REFERENCES

[1] Charles F Cadieu, Ha Hong, Daniel LK Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo, "Deep neural networks rival the representation of primate it cortex for core visual object recognition," *PLoS Comput Biol*, vol. 10, no. 12, pp. e1003963, 2014.

[2] Jonas Kubilius, Stefania Bracci, and Hans P Op de Beeck, "Deep neural networks as a computational model for human shape sensitivity," *PLoS computational biology*, vol. 12, no. 4, pp. e1004896, 2016.

[3] Nikolaus Kriegeskorte, "Deep neural networks: a new framework for modeling biological vision and brain information processing," *Annual review of vision science*, vol. 1, pp. 417–446, 2015.

[4] Umut Güçlü and Marcel AJ van Gerven, "Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream," *Journal of Neuroscience*, vol. 35, no. 27, pp. 10005–10014, 2015.

[5] Samuel Ritter, David GT Barrett, Adam Santoro, and Matt M Botvinick, "Cognitive psychology for deep neural networks: A shape bias case study," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 2940–2949.

[6] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel, "Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.," in *International Conference on Learning Representations*, 2019.

[7] Wieland Brendel and Matthias Bethge, "Approximating cnns with bag-of-local-features models works surprisingly well on imagenet," *arXiv preprint arXiv:1904.00760*, 2019.

[8] Anne Treisman, "A feature integration theory of attention," *Cognitive Psychology*, vol. 12, pp. 97–136, 1980.

[9] A Treisman and H Schmidt, "Illusory conjunctions in the perception of objects.," *Cognitive psychology*, vol. 14, no. 1, pp. 107–141, 1982.

[10] Stacia R Friedman-Hill, Lynn C Robertson, and Anne Treisman, "Parietal contributions to visual feature binding: Evidence from a patient with bilateral lesions.," *Science*, 1995.

[11] Ian H Robertson, Tom Manly, Jackie Andrade, Bart T Baddeley, and Jenny Yiend, "Oops!': performance correlates of everyday attentional failures in traumatic brain injured and normal subjects," *Neuropsychologia*, vol. 35, no. 6, pp. 747–758, 1997.

[12] Anne Treisman, "Features and objects in visual processing," *Scientific American*, vol. 255, no. 5, pp. 114B–125, 1986.

[13] Leon A Gatys, Alexander S Ecker, and Matthias Bethge, "Texture and art with deep neural networks," *Current opinion in neurobiology*, vol. 46, pp. 178–186, 2017.

[14] Alexei A Efros and William T Freeman, "Image quilting for texture synthesis and transfer," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001, pp. 341–346.

[15] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[16] Qian Wang, Jiaxing Zhang, Sen Song, and Zheng Zhang, "Attentional neural network: Feature selection using cognitive feedback," in *Advances in Neural Information Processing Systems*, 2014, pp. 2033–2041.

[17] Ning Gui, Danni Ge, and Ziyin Hu, "Afs: An attention-based mechanism for supervised feature selection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 3705–3713.

[18] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell, "Adapting visual category models to new domains," in *European conference on computer vision*. Springer, 2010, pp. 213–226.