COGNITIVECNN: MIMICKING HUMAN COGNITIVE MODELS TO RESOLVE TEXTURE-SHAPE BIAS

Satyam Mohla *, **Anshul Nasery***, **Biplab Banerjee, Subhasis Chaudhari**
Indian Institute of Technology Bombay, Mumbai, India
satyammohla@gmail.com

ABSTRACT

Recent works demonstrate the texture bias in Convolutional Neural Networks (CNNs), conflicting with early works claiming that networks identify objects using shape. It is commonly believed that the cost function forces the network to take a greedy route to increase accuracy using texture, failing to explore any global statistics. We propose a novel intuitive architecture, namely CognitiveCNN, 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 two metrics, namely *TIC* and *RIC* to quantify the importance of each stream using attention maps. We introduce a regulariser which ensures that the contribution of each feature is same for any task, as it is for reconstruction; and perform experiments to show the resulting boost in accuracy and robustness besides imparting explainability. Lastly, we adapt these ideas to conventional CNNs and propose *Augmented Cognitive CNN* to achieve superior performance in object recognition.

1 INTRODUCTION

CNNs, considered as the computational model for primate visual system (Cadieu et al. (2014), Kubilius et al. (2016)), 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 seem to support a theoretical understanding of interpretable selectivities (Kriegeskorte (2015), Güçlü & van Gerven (2015)), and also agrees with the shape bias observed in experiments with children(Ritter et al. (2017)). However, recent experiments demonstrate texture bias as the reason for the superior performance of CNNs (Geirhos et al. (2019)). Similar conclusions were drawn in Brendel & Bethge (2019), 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 learn to use texture to solve the problem and thus fail to learn any global features relevant for the task. Geirhos et al. (2019) 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. Moreover, such techniques are difficult to apply in tasks where high accuracy and robustness is important or where the image data is inherently of low quality.

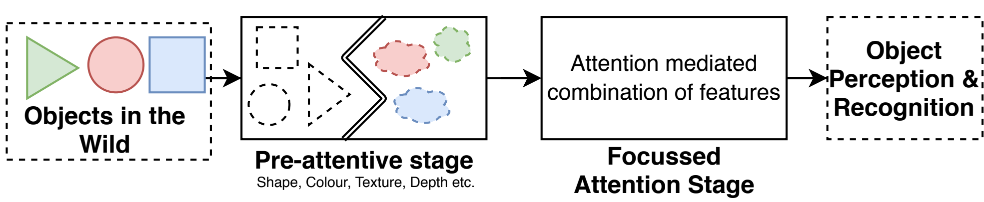

Figure 1: The stages of Feature Integration Theory(FIT) *Adapted from Treisman (1980)*

Feature Integration Theory (FIT) In cognitive psychology, feature integration theory (FIT) refers to an attention model which suggests that when perceiving objects, we synthesise and separate features initially, automatically and in parallel, directing attention serially afterwards to each

*Equal Contribution

item in turn (Treisman (1980)). This has been supported by many experiments (Treisman & Schmidt (1982), Friedman-Hill et al. (1995), Robertson et al. (1997)).The features isolated in pre-attentive stage include shape, colour, size curvature, lines etc. (Treisman (1986)) as summarised in Figure 1.

FIT provides a novel inspiration to combat our problem: we provide different feature selectivities as input to the network, emulating the pre-attentive stage. It can now explore more avenues and benefit from training with the knowledge of various features like texture, shape and edges.

This paper proposes *CognitiveCNN*, which attempts to utilise human-interpretable features like shape, texture, edges etc. to reconstruct, and classify the image. We define two metrics, namely *Reconstruction Information Coefficient (RIC)* and *Task Information Coefficient (TIC)* to quantify the importance of each stream using attention maps. We also introduce *Information Content Regulariser (ICR)* which ensures that the contribution of each feature is same for any task as it is for reconstruction and perform experiments to show the boost in accuracy and robustness besides imparting explainability. Lastly, we adapt these ideas to conventional CNNs and propose *Augmented Cognitive CNN* to achieve superior performance in object recognition.

2 Methodology

2.1 Preprocessing and training

Preprocessing Let the original dataset be $\{x, y\}_{i=1}^{n}$ where x is an image and y is the associated label. Further, let $f_1, f_2, ..., f_n$ be a set of transforms which can be applied in preprocessing stage to each x to obtain a feature. In our case, f_1, f_2, f_3, and f_4 are instantiated to extract shape, texture, greyscale image and edges respectively, and we refer to each $f_i(x)$ as a stream of information. Notably, each f_i is a *feature transform based classical image processing method.*

Architecture Next, we describe our model's architecture and training method. We represent our model by the tuple $(F_1, F_2, F_3, F_4, F_{rec}, F_{pred})$ where each F_i acts as a feature extractor for the corresponding $f_i(x)$, converting it to a latent vector z_i. F_{rec} represents a neural network for reconstructing the original image, and F_{pred} represents a neural network for predicting the labels for a given set of feature streams. Note that each F_i are one half of an autoencoder with D_i as the decoder, as shown in Fig 2

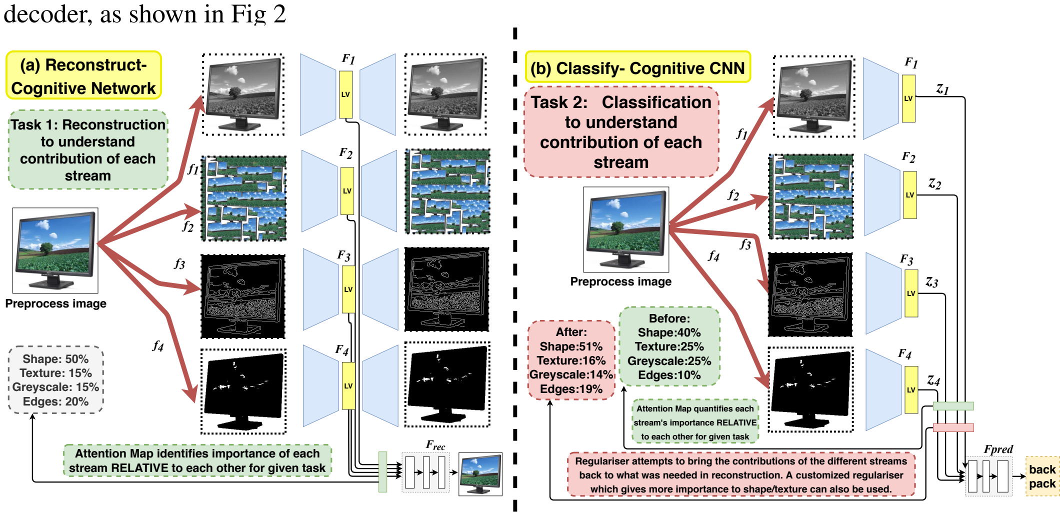

Figure 2: Experimental Setup: FIT model adapted to CNN for quantification and regularisation

Traning In the first stage of training, we train our feature extractors $F_i(\theta_i)$ with learnable parameters θ_i and reconstruction network F_{rec} to reconstruct the original image from the given feature vectors. The input to F_{rec} is the concatenated latent vectors $z_1, z_2, ..., z_n$ (Figure 2 (a)). The purpose of this part of training is to tune the feature extractors, and to gauge the importance of each stream in the reconstruction for the image. This also formally ensures that all the information of the image is captured in these four features. Formally, this stage of training can be summarized as

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

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 2 (b)) Formally, this stage aims to find

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

where $\mathcal{L}_{ce}$ is the cross-entropy loss between the predicted labels and true labels. Our model and method is summarized in Figure 2.
Now that we have classified or generated the image from the human interpretable features, we want to quantify the importance among them for the different tasks, for which we use attention maps.

2.2 Measuring Bias using Attention Maps

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 (Wang et al. (2014), Gui et al. (2019)). We extend this technique to utilize attention as a means to weigh the relative importance of each stream in prediction and reconstruction, and further to regulate the flow of information to the prediction network to make it more robust.

Let A_j be the attention layer corresponding to the task. 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.
In order to quantify the biases, based on these attention maps, we define the measures *Reconstruction Information Coefficient (RIC)* and *Task Information Coefficient (TIC)* for each stream for a particular example as

$$\mathit{RIC}_i(z) = \frac{\mathbb{E}(\sigma(A_{rec}(z))_i)}{\sum_{j=1}^{n} \mathbb{E}(\sigma(A_{rec}(z))_j)} \qquad \mathit{TIC}_i(z) = \frac{\mathbb{E}(\sigma(A_{pred}(z))_i)}{\sum_{j=1}^{n} \mathbb{E}(\sigma(A_{pred}(z))_j)}$$

where $\mathbb{E}$ represents the mean of a vector over its dimensions. *RIC* and *TIC* represent the measure of importance of each feature for reconstruction and prediction networks respectively. RIC_i indicates a measure of the amount of content present for a given feature in an image, since it is the importance that the network assigns to it while reconstructing the image, whereas TIC_i corresponds to the importance of i^{th} stream towards classifying the given image. Finally, we define the prediction network to be biased if there is a mismatch in the relative importance of the streams, i.e. $\mathit{TIC}(= \mathit{TIC}_1, \mathit{TIC}_2, ..., \mathit{TIC}_n)$ is not equal to $\mathit{RIC}(= \mathit{RIC}_1, \mathit{RIC}_2, ..., \mathit{RIC}_n)$.

2.3 Information Content Regulariser (ICR): Controlling Shape-Texture Bias

Inspired from the measures defined in the previous section, we propose Information Content Regulariser, a self-supervised regularizer to control the shape-texture bias in CNNs by adding $\sum_{i=1}^{n} ||\mathit{TIC}_i - \mathit{RIC}_i||^2$ to the loss function. This forces the prediction network to give as much importance to a feature for a given task TIC_i, as much as it was important for reconstruction RIC_i. We introduce different attention maps for classification and reconstruction tasks, as the exact same components of the latent vectors might not be useful for both tasks, but we expect the general importance of the streams to be the same.

3 Experiments and Results

We performed experiments to show the efficacy of our measures and regularizer, and the performance and robustness of our method. Since our tasks involve preprocessing using classical techniques (refer Appendix for details), our method requires a dataset that has a white background, thus we use the Amazon Domain of the Office-31(Saenko et al. (2010)) dataset for our experiments.

Accuracy We trained a network using our method (Figure 2) to classify the Amazon Office-31 dataset and recorded the value of RIC_i and TIC_i for each feature, besides the accuracy. 4UC-CogCNN reported an accuracy of 58.7%. When the network was regularized (4RC-CogCNN) using ICR, TIC_i became similar to RIC_i and accuracy increased to 61.8%, as mentioned in Table 1. We

Table 1: Comparison of importance of different streams

Stream	$\boldsymbol{RIC_i}$		$\boldsymbol{TIC_i}$
		4UC-CogCNN	**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%

incorporate these ideas into our baseline CNN, and propose Augmented CogCNN (AugCogCNN) which takes all 4 features as input alongside the image itself (Fig 3). We compared all the methods with a baseline CNN having the same architecture. All our 4 stream networks (4UC, 4RC, AugCog CNN) perform superior to the baseline network, with the highest accuracy being achieved by AugCogCNN at 62.5%. The results are mentioned in Table 2.Our reconstructions are exact and thus the extracted features carry complete information as the real image and so comparable accuracies, as observed in Table 2, are expected.

Table 2: Comparison of accuracy and robustness of different streams

Method	**Accuracy**	
	Original	**Under Miscue**
Conventional CNN (Baseline)	58.3%	14.5%
2 Stream Regularised CogCNN (2RC-CogCNN)	57.6%	49.3%
4 Stream Unregularised CogCNN (4UC-CogCNN)	58.7%	52.0%
4 Stream Regularised CogCNN (4RC-CogCNN)	**61.8%**	**56.9%**
Augmented Cognitive CNN (AugCogCNN)	**62.5%**	11.1%

Robustness In order to test for robustness, we performed a texture-shape miscue experiment as done in Geirhos et al. (2019), by *classically* generating images which had the shape from one class and texture from another .Our CogCNN approach performed consistently better than the baseline by a large margin. AugCogCNN however, based on conventional CNN architecture performed poorly, at the cost of increase in accuracy.
We also performed an ablation of CogCNN by considering 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

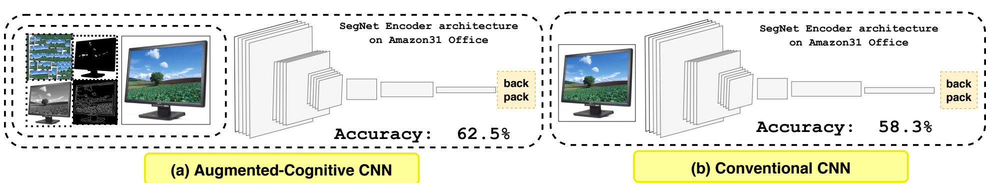

Figure 3: CognitiveCNN: FIT Inspired modification to conventional CNNs

4 Conclusion and Future Work

We showed that training a CNN with pre-processed images, as inspired from FIT, leads to an increase in the accuracy and robustness against cue conflicts. We highlighted how the method imparts explainability regarding contribution of various human-interpretable features like shape, texture, edges etc in tasks like reconstruction and classification. We also developed a novel regulariser to control bias between different features in the network. Our regularised and unregularised CogCNN performed better than the baseline in terms of accuracy, besides being robust to cue conflicts, supporting this proposed work. Lastly, we adapted the ideas to conventional CNNs for easy utilisation and achieved the highest accuracy. Our future work includes processing the input image in-situ to present an end-to-end CognitiveCNN so that it can be readily used on any dataset.

References

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

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*, 10(12):e1003963, 2014.

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.

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. URL `https://openreview.net/forum?id=Bygh9j09KX`.

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*, 35(27):10005–10014, 2015.

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*, volume 33, pp. 3705–3713, 2019.

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

Jonas Kubilius, Stefania Bracci, and Hans P Op de Beeck. Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology*, 12(4):e1004896, 2016.

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*, pp. 2940–2949. JMLR. org, 2017.

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*, 35(6):747–758, 1997.

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

A Treisman and H Schmidt. Illusory conjunctions in the perception of objects. *Cognitive psychology*, 14(1):107–141, 1982.

Anne Treisman. A feature integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

Anne Treisman. Features and objects in visual processing. *Scientific American*, 255(5):114B–125, 1986.

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