icassp 2022 Singapore

2022 IEEE International Conference on Acoustics, Speech & Signal Processing

Teaching CNNs to mimic Human Visual Cognitive Process & regularize Texture-Shape Bias

Satyam Mohla[1,2], Anshul Nasery[2], Biplab Banerjee[2]

[1]Value Creation Division, Digital Transformation Supervisory Unit, Honda Innovation Lab, Tokyo

[2]Indian Institute of Technology, Bombay, India

How CNNs do perception tasks like object recognition, has been a long-standing question in computer vision.

- Widely accepted intuition: CNNs combine low-level features to increasingly complex shapes(a) - also called **"shape hypothesis"**.
- Empirically, in agreement with visualization experiments of intermediate layer activation of CNNs(b), and shape bias observed in cognitive experiments with children(c).

Past belief indicated CNNs understand shape, but recent experiments demonstrate texture bias.

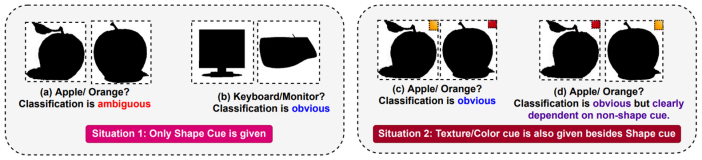
Figure 1.A & Figure 1.B. *from Geirhos et al.*

- Previous work demonstrated how the CNN classifies solely based on the texture. **[Figure 1.A]**
- Moreover, when processed cues of shape, edge etc. were tested on various CNNs, the accuracy decreased significantly for shape, and not at all for texture. **[Figure 1.B]**
- It seems thus, that CNNs, whilst being trained, may be fixating local texture with the object label, and failing to understand global shape, commonly referred to as **texture hypothesis**.

So, how to reduce texture bias?

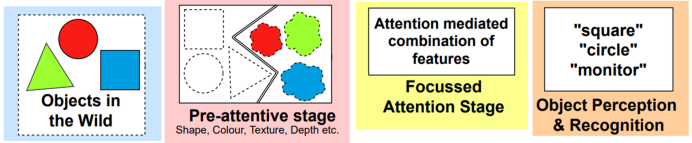
- Increasing shape bias makes CNNs more robust, and Geirhos et al. has successfully achieved this by fine-tuning on distorted texture image **[Figure 2.A]**
- While the fine-tuning does show some gain in accuracy and robustness both, its applicability is limited, and it does not really answer the greedy texture learning problem of CNNs

Gedankenexperiment



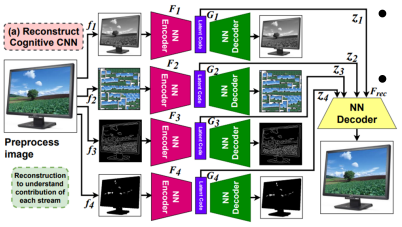
- Humans can give different relevance to different modalities (shape, texture, color etc.) depending on the recognition task. And perhaps CNNs should be able to do so too.

Feature Integration Theory (FIT)



- It theorizes, that humans perceive modalities like shape, color, size independently and integrate them in the visual field to perceive objects. **This provides an inspiration, perhaps, as to what might be missing from CNNs.**

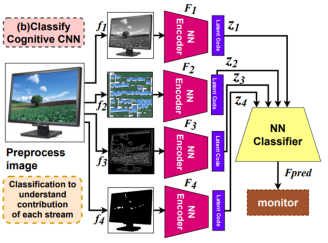
Training Setup Stage 1: Reconstruction



- We will classically process these modalities from Office31 dataset.
- We train modality encoders in a multitask setting, and demonstrate information completeness of the modalities by reconstructing original image.

Training Setup Stage 1: Classification

- We train a classifier using the generated latent code for these modalities from frozen encoders.
- The network by design is forced to evaluate these abstract representations of shape texture etc. explicitly to give its prediction.

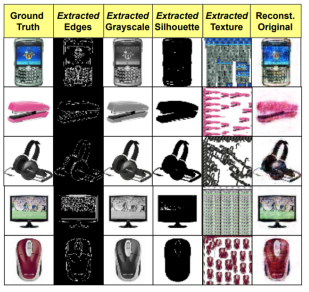


Quantifying bias using attention

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%

- To quantify the bias, we utilize self attention maps generated on concatenated latent code.
- The "relative relevance" of a modality for a given task empirically quantifies the relative weight given to that modality's latent code.
- When the weight given to a modality during reconstruction is different from the weight for a task, network is claimed biased.
- The bias can now be regularized by creating a loss term $[(RRMC - TRMC)]$^2

Results : Reconstruction & Accuracy



- The original image was successfully reconstructed using four latent codes.
- Just by virtue of separated modalities, structural to our CNN, we achieved an increase in accuracy. Regularization increased it further to 61.8%

Results : Robustness

- Our 4-stream regularized model performs very well, & achieves highest accuracy, & has a huge robustness gain!

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%

- Results show a huge increase in robustness for all CogCNN models, demonstrating the utility of the idea.