25th International Conference on Pattern Recognition
(ICPR-2020)

Collaborative Human Machine Attention Module for Character Recognition

Chetan Ralekar[1], Tapan Kumar Gandhi[1], Santanu Chaudhury[1,2]

[1]Department of Electrical Engineering, IIT Delhi, India

[2]Department of Computer Science and Engineering, IIT Jodhpur, India

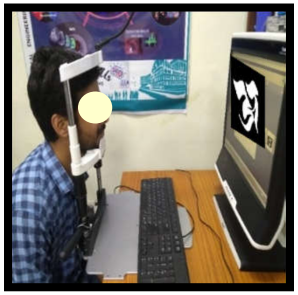
Presented by: Chetan Ralekar (chetan.ralekar@gmail.com)

Introduction

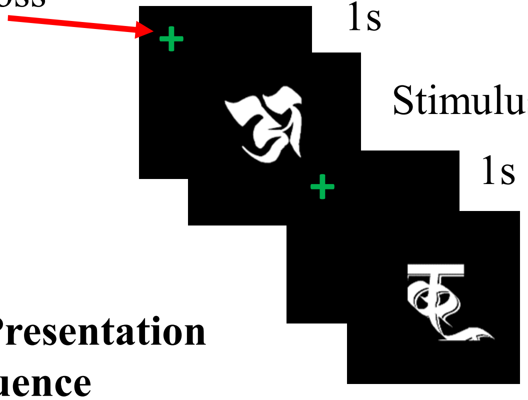
- Convolutional neural networks (CNNs) have shown impressive performance on various vision tasks by learning rich representations
- Attention mechanism tells ‘where to focus’ and improves representations*
- Most of the attention models considered attention mechanism to be a pure machine vision optimization problem
- Visual attention remains a neglected aspect

*Woo et al. [2018]

Visual Attention



Participant sitting in front of Eye-tracker



Fixation cross

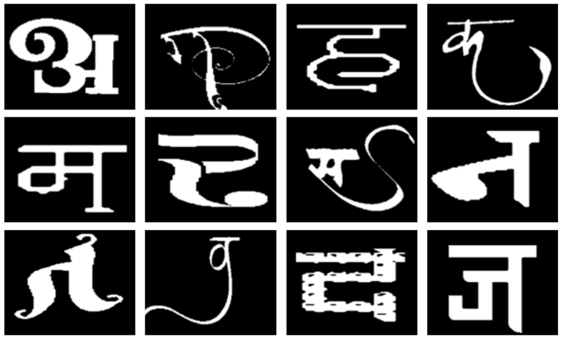
1s

Stimulus

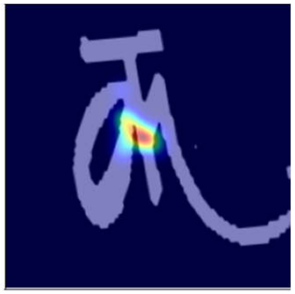
1s

Stimulus

Stimulus Presentation Sequence

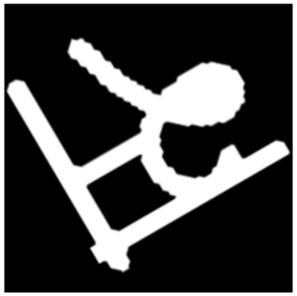


Stimuli

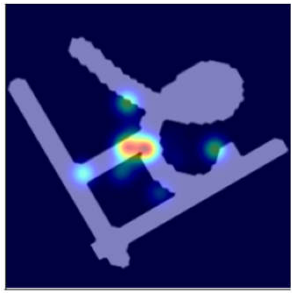


Heat Map of Visual Fixations

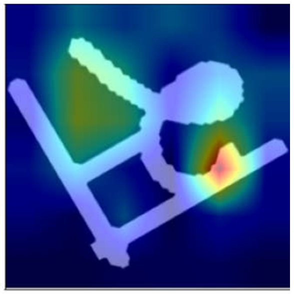
Motivation



Input character



Visual attention map

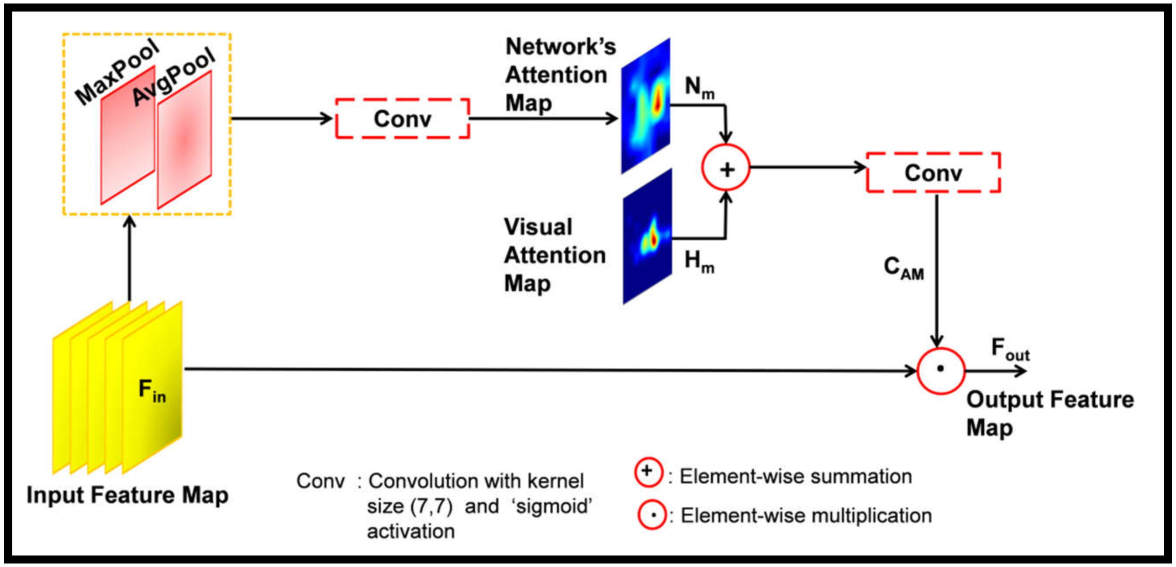


Visualization map for AlexNet Using Grad-CAM* technique

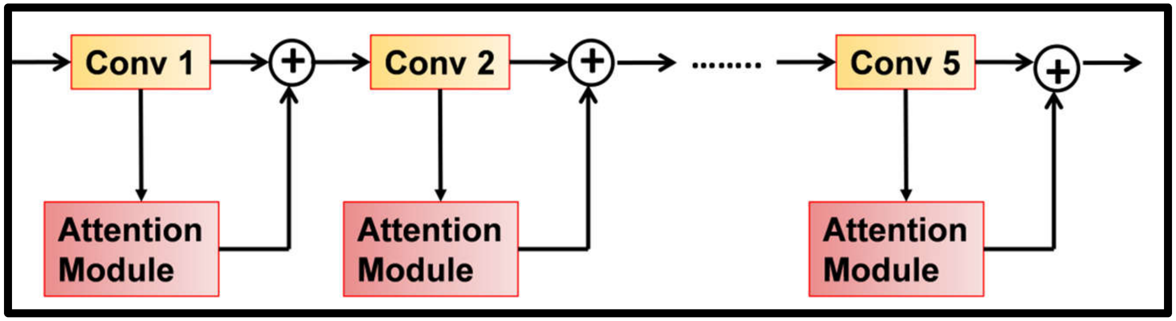
- Human attention is highly focused
- Eye-tracker captures foveal regions
- Para-foveal regions important for recognition

*Selvaraju et al. [2017]

Proposed Attention Module



Schematic of Collaborative Human-Machine Attention Module



Placement of module in the network

Ablation Experiments

Network's Spatial Attention

Sr. No.	Network	Test Accuracy
1	Baseline (AlexNet)	80.08
2	Spatial attention using Average Pool	82.77
3	Spatial attention using Max Pool	82.05
4	**Concatenation of Max and Average Pool**	**83.61**

Combining Human and Network Attention map

Sr. No.	Network	Test Accuracy
1	Baseline (AlexNet)	80.08
2	**Element wise summation**	**83.61**
3	Element wise Multiplication	82.91

Comparative Analysis

Sr. No.	Network	Test Accuracy
1	Baseline (AlexNet)	80.08
2	DeepSupervision (ICDARW-19*)	82.05
3	Fusion all layers (inspired by SonoEyeNet**)	68.71
4	Late fusion (SonoEyeNet**)	78.15
5	**Proposed Module**	**83.61**

* Chetan et al. [2019], ** Cai, Yifan, et al [2018]

Conclusions

- Collaborative Human-Machine attention module decides ‘where’ to focus
- The visual attention map covers image regions focused by humans and spatial attention maps spans other relevant regions
- The combination of visual and spatial attention maps brings finer refinement in feature maps
- The proposed module can be integrated with any CNN architecture resulting in better performance

Acknowledgements

- We would like to thank all the participants participating in the study
- Thanks to Visvesvaraya PhD scheme for providing financial assistance to first author (MEITY-PHD-589) to carry out research