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

INTRODUCTION:

- Convolutional neural networks (CNNs) have achieved impressive performance on various vision tasks.
- Performance improvement through depth, width, cardinality and attention mechanism
- Attention mechanism tells 'where to focus'
- Most of the works consider 'Attention a pure machine vision optimization problem'
- Visual attention remains a neglected aspect

EYE TRACKING:

Cognitive Link

Eye Tracker

Fixation   Saccade

Basic Eye Tracking Features

- Eye fixation a measure of information processing by Brain
- Fixations indicate selective attention on different image regions

EXPERIMENTAL DETAILS:

Participant Performing the Experiment

Stimuli: अ क ग ज त द र व ह स म न

Fixation Cross

1s

1s

Experiment Protocol

PROPOSED MODULE:

Input character

Need for foveal and para-foveal regions for recognition

Visual attention map   **Visualization map**

Conv : Convolution with kernel size (7,7) and 'sigmoid' activation

⊕ : Element-wise summation
⊗ : Element-wise multiplication

Schematic of Collaborative Human-Machine Attention Module
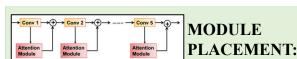
ABLATION EXPERIMENTS:

Table I: 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**

Table II: Combining spatial and visual attention map

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

MODULE PLACEMENT:
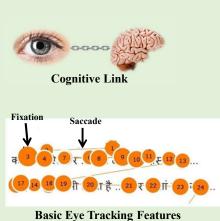
COMPARATIVE ANALYSIS:

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

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 bring finer refinement in feature maps
- The proposed module can be integrated with any CNN architecture

References:

1. Ralekar, Chetan, et al. "Intelligent Identification of Ornamental Devanagari Characters Inspired by Visual Fixations." *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. Vol. 5. IEEE, 2019.
2. Cai, Yifan, et al. "SonoEyeNet: Standardized fetal ultrasound plane detection informed by eye tracking." *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018.
3. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Communications of the ACM* 60.6 (2017): 84-90.

*Chetan Ralekar (chetan.ralekar@gmail.com)