Using Eye-Gaze to Evaluate Neural Attention

Dissertation Submitted In Partial Fulfillment Of The Requirements For The Degree Of

Master of Technology
in
Computer Science

by

Shahansha Salim
[Roll No: CS1803]

Under the Guidance of

Dr. Utpal Garain
Professor
Computer Vision and Pattern Recognition Unit(CVPR)

Indian Statistical Institute
Kolkata-700108, India

CERTIFICATE

This is to certify that the dissertation entitled **"Using Eye-Gaze to Evaluate Neural Attention"** submitted by **Shahansha Salim** to Indian Statistical Institute, Kolkata, in partial fulfilment for the award of the degree of **Master of Technology in Computer Science** is a *bona fide* record of work carried out by him under my supervision and guidance. The dissertation has fulfilled all the requirements as per the regulations of this institute and, in my opinion, has reached the standard needed for submission.

Dr. Utpal Garain
Professor,
Computer Vision and Pattern Recognition Unit,
Indian Statistical Institute,
Kolkata-700108, India.

Acknowledgment

It was a great privilege and learning experience, to work with Dr. Utpal Garain. He had been a constant source of support, starting from my first year. I want to thank all the research scholars in NLP Lab and would like to extend my sincere gratitude to Akshay Chaturvedi and Joy Mahapatra. Thanks to all the friends of MTech-CS 2018 who had been there with me for the past two years.

Contents

Abstract

The ability to selectively concentrate on areas of interest while ignoring the rest is termed as attention in human beings. This ability has played a key role in survival as well as information processing. Neural Attention is said to be an effort to bring similar action of selectively concentrating areas of relevance in deep neural networks. This simple yet powerful concept has attracted a lot of research in recent years, yielding breakthrough results in Natural Language Processing (NLP) problems and main stream Computer Vision problems such as Image Caption Generation, Neural Machine Translation (NMT), Visual Question Answering (VQA), Action Recognition, Image Segmentation, etc. Only few works has been there articulating the relation between human attention and machine attention. Some recent efforts suggest that automatically learned attention maps can capture informative parts of an input signal and highlight human sensible regions of interest. Also, as the neural attention gets better, so is the performance of the network. However, there are no formal way of bench marking how good the learned attention is in a network. This seems necessary since visualizing attention as a means for logical correctness of the network is common. Since eye-gaze can better capture human visual reasoning, With this work, we are investigating how well neural attention compares with the visual grounding given by human cognitive modality on VQA tasks.

Introduction

Humans rely on multiple sensory modalities when examining and reasoning over images. In recent years, eye-tracking has become widespread ranging from Virtual Reality to assisted communication (Padmanaban et al.,2017 [1]; Holmqvist et al., 2017 [2]). Gaze data, such as fixation duration and how fixations are changed during gazing can reveal crucial information about the cognitive process of human understanding and what is happening inside our brain. One such area in which researchers have used gaze measurements is to understand where drivers look and to identify differences in experts' and novices' viewing behaviors in domain-specific tasks (Underwood et al., 2003 [3]; Eivazi et al.,2012 [4]). There has been numerous studies which highlight the potential importance of analyzing gaze data to shed light on how humans cognitively process information, make decisions and how this behaviour varies among humans (Fiedler and Glockner, 2012 [5]; Hayes and Henderson, 2017 [6];Brunye and Gardony, 2017 [7]).

Our work is based on the eye-mind hypothesis (Just and Carpenter 1980 [8]) which states that when a person views a word/object, they also processes it cognitively for approximately the same amount of time the person fixates on it. With the advent of ubiquitous eye-tracking hardware, we are now equipped to make use of this hypothesis into various fields such as education(Lima Sanches et al., 2018 [9]), Natural language understanding, Image processing, etc. Effectiveness of online education materials (Abhijit Mishra et al., 2017 [10]) can now be measured and adjusted more effectively. The effort reader has to invest in reading an article along with the eye-gaze and movement of eye-gaze (termed scanpath) can indicate the quality of material as well as the understanding ability of the student without cross-examining post the activity. Eye tracking is also useful in human computer interaction, real time game scene tracking and many more.

Another area of research which witnessed a surge in popularity is the attention mechanisms encoded within deep neural networks. This is inspired from the selective attention of the human visual cortex, and artificial attention is designed to focus more on the part of the input which is most relevant for making a prediction. This offers great help in post-hoc analysis and explanation of results from neural networks. In simpler terms, attention mechanisms highlight that part of the input where the neural network looks to make predictions, just like humans do. Introduction of attention in deep neural networks has helped to create new state of the art techniques in Neural Machine Translation (NMT, Jacob Devlin et al., [11]) Visual Question Answering (VQA ,Yu et al., [12]), etc. VQA is particularly interesting since it combines multiple modalities from domains of Computer vision and Natural language processing to produce question answering models.

Even though attention helped to model various state of the arts for many known problems, we have little ways to see the correctness of attention. Sergey Zagoruyko et al., [13] suggests that by

properly defining attention for convolutional neural networks, we can actually significantly improve the performance of a CNN. Many recent research works showcase attention maps as a means of the correctness and explainability of a network. This suffers from the disadvantage of being empirical and not able to capture the performance of the network as a whole or key inputs where network might suffer. Also, sometimes networks predict true answers while 'attending' areas where no significant information is conveyed. This indicates the necessity for a systematic understanding and evaluations on the neural attention mechanism.

Naturally, since attention is an effort to focus the regions of significance in input signals, by intuition this can correlate to human visual attention facilitated by eye-gaze. There has been little studies before studying eye-gaze as a means for bench marking ground truth in input data. Recently, there has been a number of publications claiming being successful in improving performance of Visual Question Answering models by giving additional grounding on train data by means of human like visual and linguistic modalities (Selvaraju et al., 2019 [14]). However, it has been proven very recently that the visual grounding techniques does not impart any significant performance improvement, rather the improvement was a result of a regularization effect which prevents over fitting to linguistic priors (Srestha et al.,2020 [15]). This demands a proper analysis on how attention would behave with actual human data.

We use a non-intrusive comparatively cheaper eye-tracker to record eye-gaze data for the evaluation purpose. Since Visual Question Answering combines both visual and language modalities, VQA models are naturally a candidate which suits the evaluation. With this work, we are trying to study how much human and machine attention correlate on a selective subset of VQA dataset. Also, a critical evaluation of how good the machine attention compared to human attention is also done.

In the first part, some preliminaries are introduced which would be necessary to understand this work. Later the implementation, Dataset and Evaluation of the data are discussed.

1.1 Preliminaries

Before diving deeper into the problem and the solutions, definition of some important terms needs to be understood as they will be extensively used throughout this report.

1.1.1 Human visual attention

Human visual attention has been extensively studied for decades in the cognitive science[16] and computer vision domains dating back to 1998 (Itti et al., [17]). Visual attention falls under two main categories: bottom-up(exogenous) and top-down(endogenous). Similar analogical models can be found in neural(machine) attention as well

Bottom-up attention: This is driven by noticeable external stimuli due to color, motion, etc. of a subject. This corresponds to looking, then processing much like our day to day vision. Looking at a flower because of it's popping colours or looking at a moving object then correcting our travel route are examples of bottom-up attention (Katsuki and Constantinidis,[18]). Most early computational attention models were of bottom-up nature.

Top-down attention: instead of being stimulus-inspired, it is an internally induced process based on prior knowledge or goals. For instance, when inspecting surveillance videos, guards are more

likely to allocate their attention to moving people for detecting suspicious behaviors. Endogenous attention is accompanied by longer-term cognitive factors and is very common in our daily-life.

1.1.2 Eye-Gaze

Placed on or within the machine interface and using the eye's reflection of near infrared light beams, eye tracking technology calculates data about the user; detecting presence, attention, and focus as well as the position of a person's eye and pupil size. This is done using cheap and non-intrusive, screen mountable devices from Tobii in our case. The working principle is rather straight forward - Using cameras to detect the near infrared light reflected off of user's eye and then using filtering and triangulation to estimate accurate *(x,y)* coordinates where the pupil is focused.

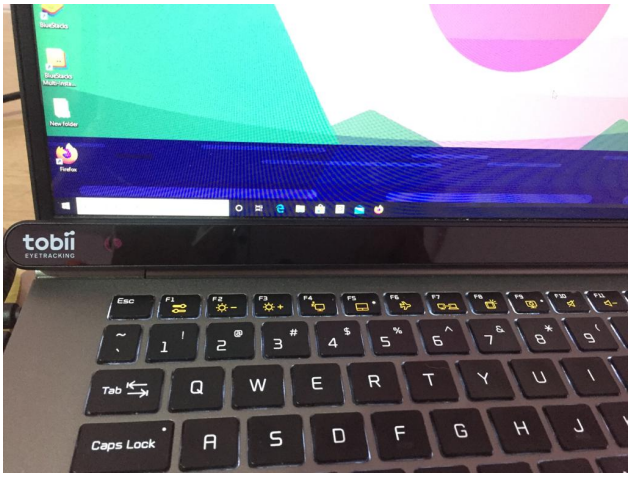

Figure 1.1: Tobii 4C mounted on a laptop computer.

The **Interest Area (IA)** is an area of the screen which is under interest. We mainly look at objects in the image for reasoning the answer for the question. A **fixation** takes place when gaze is focused on a point in the screen. A **saccade** is the movement of gaze between two fixations. At any point in time, the eye-tracker gives an *(x,y)* coordinate. The eye-tracker we used for this project has a frequency of 90Hz.

1.1.3 Recurrent Neural Networks

Classic Neural Networks when used for inputs in the sequential form, there is a chance for the number of parameters to explode or it may not even be possible to decide the size of parameters beforehand. This is where we make use of Recurrent Neural Networks(RNN). An RNN can incrementally process inputs at each time steps separately and the output is an encoding for the whole sequence. Since the parameters used at each time step are shared, the explosion of the number of parameters can be avoided. Each input x of length n is a sequence represented as $(x_1, x_2, ..., x_{n-1}, x_n)$. The RNN at each time step t calculates the hidden representation called hidden state h_t as

$$h_t = F(x_t, h_{t-1}) \tag{1.1}$$

The output o_t for each time step is defined in terms of h_i as

$$o_t = G(h_t) \tag{1.2}$$

In the above equations, $F()$ is an RNN Cell and $G()$ is a Feed-Forward neural network.

1.1.4 Long Short Term Memory Networks

RNN inherently has issues with exploding gradients and vanishing gradients. Long Short Term Memory is one way to address these issues. LSTM was introduced by Hochreiter et al. [19] in 1997. Along with the hidden state defined for RNN, LSTM keeps track of a cell state, which acts as a bypass for the gradients to pass easily. Thus, the cell state helps the LSTM to keep track of long term dependencies. LSTM consists of modules called gates which are: *Input gate*, *Forget gate* and *Output gate*.

$$\begin{aligned} f_t &= F(x_t, h_{t-1}) \\ i_t &= I(x_t, h_{t-1}) \\ o_t &= O(x_t, h_{t-1}) \\ c_t &= f_t \circ c_{t-1} + i_t \circ C(x_t, h_{t-1}) \\ h_t &= o_t \circ H(c_t) \end{aligned} \tag{1.3}$$

where f_t, i_t and o_t are outputs from forget, input and output gates respectively, c_t is the cell state and h_t is the hidden state.

1.1.5 Transformers

Transformers are a simple network architecture based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. This was introduced by Vaswani et al. in the paper "Attention is All You Need" [20]. They show that transformer give results superior in quality while being more parallelizable and requiring significantly less time to train. Multiple kinds of attention mechanisms are used in transformer. Self attention will be used independently by encoder and decoder sides to do attention on the input sentence or the generated sentences respectively. Encoder-Decoder attention is a combined attention mechanism which takes care of the input-output sentence alignments. Figure 1.2 shows the architecture for a Transformer model. Bidirectional Encoder Representations from Transformers(BERT) [11] which is derived from Transformers is one of the most hot-topics in NLP in recent times. It is a pre-trained language model, which can be fine-tuned solve other tasks.

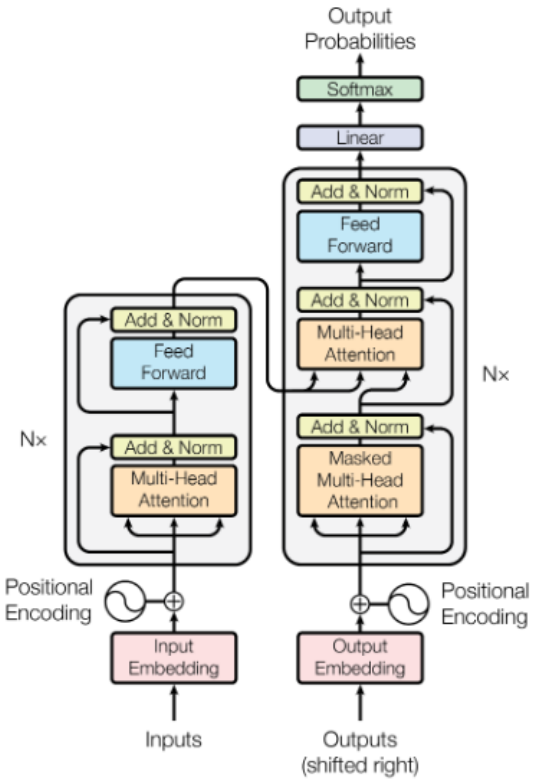

Figure 1.2: Transformer model as given in the paper: Attention is all you need [20].

1.1.6 Region proposal networks : The Faster R-CNN

Faster R-CNN is an object detection model designed to identify instances of objects belonging to certain classes and localize them with bounding boxes [21]. Faster R-CNN detects objects in two stages. The first stage, described as a Region Proposal Network (RPN), predicts object proposals. A small network is slid over features at an intermediate level of a CNN. At each spatial location the network predicts a class-agnostic objectness score and a bounding box refinement for anchor boxes of multiple scales and aspect ratios. Using greedy non-maximum suppression with an intersection-over-union (IoU) threshold, the top box proposals are selected as input to the second stage. In the second stage, region of interest (RoI) pooling is used to extract a small feature map (e.g.14×14) for each box proposal. These feature maps are then batched together as input to the final layers of the CNN. The final output of the model consists of a softmax distribution over class labels and class-specific bounding box refinements for each box proposal [22].

1.1.7 Visual Question Answering

The task of Visual Question Answering as an yearly challenge and the VQA dataset was introduced in 2015 [23]. Given an image and an open-ended, natural language question about the image, the task is to provide an accurate natural language answer. The present iteration of VQA dataset is more unbiased and balanced, known as the *VQA V2.0*. There are nearly 200K images, nearly 1 Million questions and 11 Million ground truth answers [24]. Every year, the VQA workshop hosts

talks with best performing networks which are able to solve the VQA problem. Figure 1.3 shows typical examples from *VQA V2.0* dataset. The questions are categorized based on the answer type. There are *Number*, *Yes/No* and *Other(textual)* type answers and questions. The ground truth answers are recorded using Amazon Mechanical Turk(AMT) over multiple observers, and any prediction is a true prediction if 3 or more ground truth answers are same as the prediction. The state of the art VQA network from 2019 is used for this work and it is explained in section 1.1.8 below.

Figure 1.3: Examples of questions and images in VQA [24].

1.1.8 Modular Co-Attention Network (MCAN)

MCAN consists of a cascade of modular co-attention (MCA) layers, each of which consists of the self-attention and guided-attention units to model the intra and inter modal interactions synergistically. MCAN achieves state of the art performance in VQA by stacking MCA layers in deep using the encoder-decoder strategy. Fig 1.4 shows the building blocks and overall processing pipeline of the network.

The image features used here, denoted by X is the set of vectors each corresponding to the objects detected by a faster R-CNN [21]. Y is the question input which is obtained from the GloVe + LSTM module. These are passed on though a deep co-attention model consisting of L MCA layers cascaded in depth (denoted by $\text{MCA}^{(1)}$, $\text{MCA}^{(2)}$... $\text{MCA}^{(L)}$). The output from one layer is fed as the input to the next layer, denoted as,

$$[X^{(l)}, Y^{(l)}] = MCA^{(l)}([X^{l-1}, Y^{l-1}]). \tag{1.4}$$

After the deep co-attention learning, the output image features $X^{(L)} = [x_1^{(L)}; ...; x_m^{(L)}] \in \mathbb{R}^{m \times d}$ and question features $Y^{(L)} = [y_1^{(L)}; ...; y_n^{(L)}] \in \mathbb{R}^{m \times d}$ already contain rich information about the attention weights over the question words and image regions. This is passed through a 2 layer MLP. for example, the attended feature $\bar{x}$ is obtained as follows.

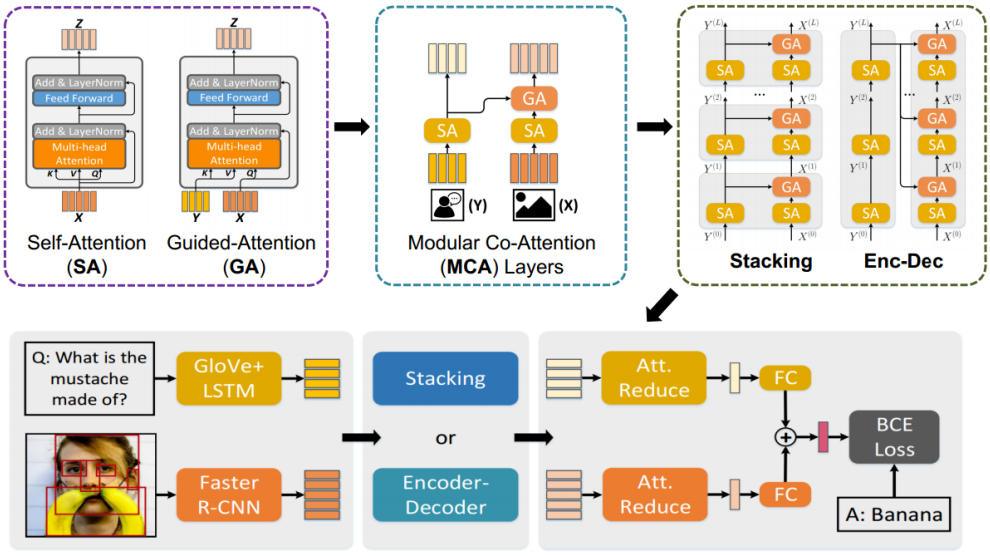

Figure 1.4: MCAN VQA network overall building blocks [12].

$$\alpha = softmax(MLP(X^{(L)}))$$
$$\bar{x} = \sum_{i=1}^{m} \alpha_i x_i^{(L)} \tag{1.5}$$

where $\alpha = [\alpha_1, \alpha_2, ..., \alpha_m] \in \mathbb{R}^m$ are the learned attention weights for the m objects detected initially using the faster R-CNN. It should be noted that the objects that are most relevant to the task gets the most attention. also, $\sum \alpha = 1$. Using the computed $\bar{x}$ and $\bar{y}$, the final prediction is obtained using the following linear multimodal function.

$$z = LayerNorm(W_x^T \bar{x} + W_y^T \bar{y}) \tag{1.6}$$

where W is the weight matrix learned through training.

1.2 Previous Works

Attention mechanisms were first introduced in the text processing domain. Ever since the inception, it is treated monumentally and has attracted a great amount of research. The major milestones in attention research is highlighted below.

1. Sequence-to-sequence(seq2seq) model (Cho et al. 2014 [25], Sutskever et al. 2014 [26])
2. Alignment models (Bahdanau et al. 2015 [27], Luong et al. 2015)
3. Visual attention (Xu et al. 2015 [28])
4. Hierarchical attention (Yang et al. 2016 [29])

5. Transformers (Vaswani et al. 2017 [20])

Above mentioned works are some of the most cited publications in the attention prowess. Xu et al. mentions that they intend to produce attention framework for caption generation based on how human would attend to informative areas of a scene.
Visual Question Answering (VQA) was introduced by Anton et al. in 2015 [23]. Since then, It has received a lot of research interest, some of them being quite contradictory in the inferences. Das et al. [30] in 2016 pointed out that for VQA tasks, machine attention need not necessarily fall in line with human attention. They generated mouse click based human annotations data known as VQA-HAT for a small number of questions using Amazon Mechanical Turk for this task. The annotator examines the question and the image and point out the areas which might be used by a human observer to reason the answer for the question. Later, Qiao et al. [31] used the same VQA-HAT applied to VQA to generate human like attentions for image-question pairs. They claim that adding human-like supervision yields a more accurate attention together with a better performance, showing a promising future for human-like attention supervision in VQA. However, the problem we see with VQA-HAT annotations is that this need not be capturing the human cognitive attention directly. Eye gaze on the other hand is intrinsic and is a direct reflection of human decision making. Shrestha et al. in 2020 [15] suggested that the performance improvement that is obtained while incorporating visual cues like VQA-HAT is due to a regularization effect which prevents over-fitting to linguistic priors. With so many conflicting results present, we need to have a proper understanding of how good the present neural attention when compared to human attention to see if networks are reliable with the attention map based explanations they give.

1.3 Motivation

Understanding how cognition works in humans has always been a topic of interest. Eye-gaze is an excellent choice when it comes to non-intrusive ways of gathering brain information. Based on this, characterizing human reading behaviour with mathematical models were a major research of past. SWIFT [32] and E-Z reader [33] are noteworthy. Even though the research based on eye-gaze started as early as 1990's [34] [17] [8] , researchers started using it extensively in text and language processing domain fairly recently. So far, using eye-gaze has proven successful in rating quality of article [35], Rating quality of reader [36] [37] [38], Sentiment and sarcasm understanding [39], etc. Eye-gaze analysis on Japanese Manga comics has helped to identify conceptually difficult parts, then later use this result in their learning system namely 'Tadoku' [40].
Even though the possibilities are rich, Eye-gaze is far less utilized in the image processing domain. Vaidyanathan et al. [41] showed that using eye-gaze and spoken narratives, successful image segmentation is possible. Attention is one area which is particularly interesting. Partly due to the recent breakthroughs and immense research participation it gained, and also that eye-gaze naturally corresponds to human attention, Characterizing if neural networks are looking at the correct place is necessary. Lai et al. [42] suggests that the better eye-gaze and neural attention coincide, the better performance of the network will be. Also, the methodology usually adopted by researchers while explaining the correctness of a model is to empirically show for few examples that, attention is in the logically relevant place. This is flawed since few examples does not reflect the flaws in the system. Therefore, a proper comparison with human attention is necessary for neural attention to understand the reliability of inferences that machine attention gives.

Methodology

This section explains in more detail, the approaches taken to solve the problem. We designed some preliminary experiments to get insights on how to formally design the solution. Following sections describe the experiment, the evaluation metric and the dataset in detail.

2.1 Preliminary Exploration

With our hypothesis, the first try was on public eye-gaze datasets which are already available. Most existing visual attention datasets collected gaze data during free viewing. i.e., subjects were asked to view scenes without any particular task in mind. This deemed not useful, since our task was to capture the interest areas as per human subject and then compare it to that of machine. Other openly available datasets such as PASCAL-S, DUT-O, Hollywood-2, etc are mainly used for salient object segmentation and video action recognition. Lai et al.[42] suggests that machine attention may not be able to extract meaningful areas in case of complex tasks like action recognition. To start with something intuitive and explicit neural attention based, VQA seemed a good candidate for the task. As no such study where eye-gaze has been recorded for VQA, A preliminary study of recording eye-gaze for a subset of questions was conducted to see if machine and human looks at the same place for answers, and also, if there are any questions for which the machine attention is reasonably far from what should have been the 'ground truth'.

2.1.1 Dataset Selection

VQA V2.0 [24] is a recent dataset used for vqa task, which is more unbiased and balanced than the original VQA. The experiment was chosen to conduct on validation set consisting of nearly 200,000 questions on nearly 40,000 images. The annotated training dataset of VQA consists of 443,753 questions over 82,783 images. The data was collected using Amazon Mechanical Turk and each validation annotation consists of 9 answers, for which if machine predicts an answer with atleast 3 agreements, it is regarded as a true answer. The image features are extracted using the bottom-up-attention strategy [22], with each image being represented as an dynamic number (from 10 to 100) of 2048-D features. In addition to this, the dataset contains VQA samples from visual genome [43] dataset to expand the training samples. The bottom up attention feature extractor model also gives the set of bounding boxes (limited to 100) for all the images.

2.1.2 Model Selection

There are 5 state of the art VQA networks over the years, out of that, one particularly suited for this task. The Modular Co-Attention Network or MCAN as explained in section 1.1.8 not only is the latest state of the art of 2019, But has an explicit attention component per detected objects (see equation 1.5). Also, the attention is a modified transformer and with the stacked or encoder-decoder co-attention learning, Image features will contain rich question based and self attention data.

2.1.3 Experiment

We designed and tested simple interfaces for conducting VQA test on humans and for recording data. A Tobii SDK based eye-gaze recording software was also implemented. The gaze recorder software runs in the background recording data from the subject while they are interacting with the VQA interface. We experimented with multiple variants of the data collection interface. Analysis of the interfaces as well as details of the human evaluation studies conducted to converge on the final interface used for results in this main document are explained in the implementation section. We collected human attention data from 3 subjects who went through qualification screening to make sure that they are qualified to comprehend the questions, perceive the image and well versed for overall interaction with the interface using a computer. The data was collected on carefully selected 300 questions which can reveal the flaws of underlying model. The confidence of prediction from the model was taken as a means for choosing the subset of questions. Present day neural networks may have problems with calibration, which is a well studied problem [44]. To rely on confidences, calibrating the neural network is necessary in a normal scenario. But in our case, it was decided to proceed without calibration as calibrating the network may improve it by reducing confidence on wrong answers for which it was overconfident initially [44]. The data from eye-tracker is recorded and processed for noises, removed the noise and mapped on to the corresponding VQA question-image pair. Fig 2.1 - 2.3 shows neural attention, raw eye-gaze when plotted onto the image and eye-gaze quantized to objects when plotted back to the image for the question *What is the bus number?*. The neural network gave a correct answer prediction even though it *attends* an area which is more than what is necessary for human subject.

Figure 2.1: Neural attention

Figure 2.2: Eye-gaze plotted

Figure 2.3: Quantized Eye-Gaze

Implementation and Dataset

3.1 Implementation

In this section, the implementation and system level information regarding the user interface and eye-gaze tracker software are discussed.

3.1.1 Data Processing Pipeline

A Python and Kivy based GUI consisting of multiple windows is developed for the purpose of conducting the experiment. The user is allowed to have a username, and continue where they left off. A text csv file is kept for each user to store the collected data. The VQA GUI collects the answer given by the subject in text, The start and end times of interaction for each question-image pair, the bottom left coordinates of the image and the dimensions of the image displayed. This is combined on the preprocessing stage with eye-tracker data to generate (question, image, gaze) triple. Fig 3.1 and 3.2 shows screen shots of the UI for collecting vqa data.

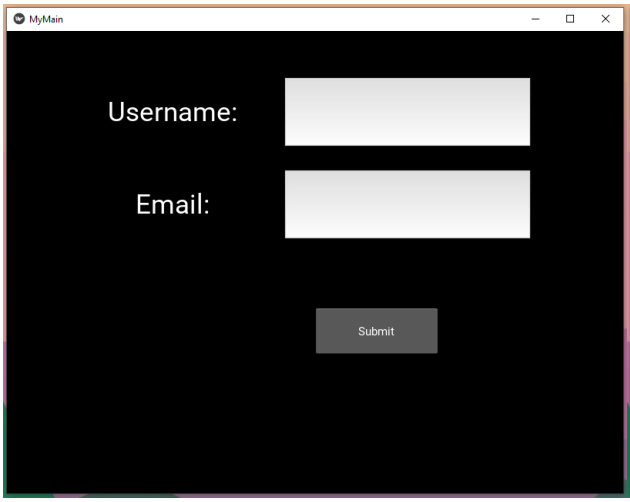

Figure 3.1: Login window of the vqa interface

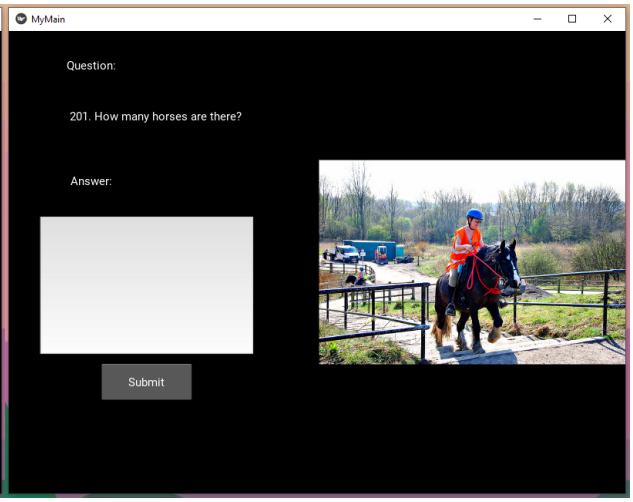

Figure 3.2: Main window of the vqa interface

Eye-gaze tracker is calibrated before each recording session per user. Calibration is the process whereby the geometric characteristics of a subject's eyes are estimated as the basis for a fully-customized and accurate gaze point calculation. Before an eye tracking recording is started, the

user is taken through a calibration procedure. During this procedure, the eye tracker measures characteristics of the user's eyes and uses them together with an internal, anatomycal 3D eye model to calculate the gaze data. This model includes information about shapes, light refraction and reflection properties of the different parts of the eyes (e.g. cornea, placement of the fovea, etc.). During the calibration the user is asked to look at specific points on the screen, also known as calibration dots. During this period several images of the eyes are collected and analyzed. The resulting information is then integrated in the eye model and the gaze point for each image sample is calculated. This process is handled by the eye-tracker accompanied software provided by Tobii Technology. Fig 3.3 shows the collected calibration data before each user session.

Figure 3.3: Eye-Tracker calibration data

The eye-tracker runs in the background recording the triple $(time_g, x, y)$ at $90Hz$ where (x, y) is a fixation. Data is recorded in sessions of 50 questions to make sure the annotator is well rested. For each new session, the eye-tracker is adjusted for calibration. For the data $(t_1, t_2, question, image)$ from the GUI, all the gaze points for which $t_1 \leq time_g \leq t_2$ is mapped onto the question-image pair. All the data is conveniently stored in csv format, and manipulated using the Pandas data analytics library of Python.

3.1.2 Extracting Neural Attention

The *PyTorch* implementation of Modular Co-Attention Network (MCAN)[12] for VQA is taken from their Github repository[1]. The input image is represented as a set of regional visualfeatures in a bottom-up manner[22]. These features are the intermediate features extracted from a Faster R-CNN model(with ResNet-101 as its backbone)[45] pre-trained on the visual genome dataset [43]. By mean pooling the convolutional features from the detected region, a feature set of size $m, m \in \mathbb{R}$ is extracted. $m \in [10, 100]$ and the image will be represented as a feature matrix $X \in \mathbb{R}^{m \times d_x}$. The features are stored as npz files in the file system. Similarly, the question features Y is generated using an LSTM[19] network with d_y hidden units. Prior to this, each word is embedded using a 300-D GloVe word embedding[46] pretrained on a large-scale text corpus. Finally the question feature matrix Y is formed with $Y \in \mathbb{R}^{n \times d_y}$.

[1] https://github.com/MILVLG/mcan-vqa

The dimensionality of input image features d_x, input question features d_y and fused multimodal features d_z are 2,048, 512, and 1,024, respectively. Number of heads h in multihead attention is 8 and latent dimensionality is 512. The number of Co-Attention layers is $L \in \{1, 2, 4, 6, 8\}$. To train the MCAN model, Adam Solver [47] is used with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. The model is trained for 13 epochs with a batch size of 32. A validation is run using the *Val* set of VQA, and the attention from equation 1.5 is stored into Python Object Serialized files (Pickle) which is of dimension $d_{att} \in \mathbb{R}^{m \times d_x}$

3.1.3 Dataset

A subset of 300 questions are chosen from the *Val* set for the experiment. There are three answer classes in VQA dataset namely, 'Yes/No', 'Number' and 'Other' depending on the type of answer they indicate. To ensure fairness and to provide critical performance evaluation of the model, following classes of questions are chosen

1. Questions with 50 least confident correct answers from each class
2. Questions with 50 most confident wrong answers from each class.

Upon filtering eye-gaze and combining vqa data obtained with gaze data, it is combined from multiple annotators into a ground truth based on a majority voting of at least 2 out of 3. The ground truth is verified to conform to individual annotators. On the gaze-vqa data along with neural attention obtained from section 3.1.2, Evaluations are performed.

3.1.4 System Level Information

All data recording experiments are run on a portable notebook computer with *Intel(R) Core(TM) i5-8250U CPU @ 1.60GHz*. The machine has 8 GB of RAM. The MCAN [12] model was run on *GeForce GTX 1080 Ti* GPU with 8118MiB RAM. CPU used is *Intel(R) Core(TM) i7-6700 CPU @ 3.70GHz*. The machine has 31.4 GB of RAM.

Experiments and Results

In this section, we discuss the evaluations performed on the data in section 4.1 and the comparison metric and results from eye-gaze with attention data in section 4.2. The combined eye-gaze and vqa data obtained from each of the annotators is stored in a csv file as explained in section 3.1.1. The neural attention from the model MCAN-VQA is also obtained as explained in section 3.1.2.

4.1 Evaluations on Data

4.1.1 Inferences from Recorded Eye-Gaze

There are on average 80K gaze instances per 100 question data. The gaze data is filtered to extract only those which falls in the image displayed. Once filtered, we filter the gazes to eliminate possible saccades. Tables below indicates the Euclidian distance metric between two adjacent gazes extracted from a subset of 100 vqa question instances. Table 4.1 indicates the statistics before performing the filtering operation and table 4.2 refers to gaze data once the saccades are eliminated. Clearly, it can be seen that the mean distance between successive fixations are reduced drastically. Also, the percentiles at 0.25, 0.50 and 0.75 parts when the data is sorted is also given to indicate the manner of distribution and balance in the overall data.

Count	39042
mean	4.886
std	20.600
min	0
least 25%	0.572
least 50%	1.146
least 75%	2.477
max	561.662

Table 4.1: Distance between successive gaze instances before filtering.

Count	33844
mean	1.277
std	1.032
min	0
least 25%	0.510
least 50%	0.966
least 75%	1.756
max	4.885

Table 4.2: Distance between successive gaze instances after filtering for saccades.

Once the eye-gaze and vqa are combined per session per annotator, table 4.3 shows the combined average statistics on number of gaze instances for 100 question-gaze pairs. Percentiles are also given as a representative of distribution of data.

The ground truth gaze data is obtained from a majority voting basis of at least 2 out of 3 from

Count	100
mean	338.440
std	142.414
min	19
least 25%	249.500
least 50%	307.000
least 75%	395.500
max	909.000

Table 4.3: Gaze count statistics on combined gaze-vqa data for 100 instances

the individual annotators. Once all 900 instances of questions from 3 annotators are combined, we have on average 150 gaze instances per question.

4.2 Reliability of Attention

4.2.1 Intersection Over Union(IOU)

To further check the overlap between objects indicated by machine attention and objects annotated by the ground truth, we use a rather simple metric, inspired by Segmentation tasks, called Jaccard Index. formally,

$$IOU = \frac{n_{same}}{n_{total}} \tag{4.1}$$

Where n_{same} is the number of objects match and n_{total} is the union of ground truth and observation. Based on IOU, following parameters are defined.

True Positive(TP): A correct attention. Matching with $IOU \geq$ threshold

False Positive(FP): A wrong attention. Matching with $IOU <$ threshold

False Negative(FN): A ground truth not detected

True Negative(TN): Does not apply. It would represent a corrected misdetection. In the object detection task there are many possible bounding boxes that should not be detected within an image. Thus, TN would be all possible bounding boxes that were correctly not detected (so many possible boxes within an image). That's why it is not used by the metrics.
threshold : depending on the metric, it is usually set to $50\%, 75\%$ *or* 95%.

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

We consider overall and answer type based precisions. Table 4.4 - 4.5 shows the results of evaluation.

Key Inferences:

1. Overall nearly 50% of the time when IOU=50%, the neural attention is in line with human attention.

IOU	Precision
50%	0.48
75%	0.103
95%	0.06

Table 4.4: Precision Overall.

IOU	Precision
50%	0.406
75%	0.08
95%	0

Table 4.5: Precision on the correctly predicted class.

IOU	Precision
50%	0.553
75%	0.126
95%	0.013

Table 4.6: Precision on the wrongly predicted classes.

2. True answers and False answers don't show any difference in trends of attention-gaze overlap and they are significantly low once IOU falls above 75%

4.2.2 Correctly Predicted but Improperly Grounded (CPIG)

Shrestha et al. in 'A Negative Case Analysis of Visual Grounding Methods for VQA' [15] introduced a new metric in order to qualitatively assess visual grounding. $CPIG$ is defined as,

$$\%CPIG = \frac{N_{correct\ pred, improper\ grounding}}{N_{correct\ pred}} \times 100\% \tag{4.3}$$

which is the number of instances for which the most sensitive visual region is not within top-3 most relevant ground truth regions, normalized by the number of correct predictions. Here also, the metric is evaluated over different values of IOU. Table 4.7 indicate CPIG evaluations.

IOU	CPIG
50%	0.593
75%	0.92
95%	1

Table 4.7: Correctly Predicted but Improperly Grounded $CPIG$

Key Inferences:

1. 59% of the times, even if the answer is correct, attention is in the wrong place.
2. as IOU goes up, the fall in attention correctness is evident.

4.3 Consistency of Ground Truth

We see if the ground truth conforms to data from individual annotators. We have 3 annotators denoted as P_1, P_2 and P_3 with the merged ground truth being denoted as G. A simple averag-

ing based comparison as well as chance agnostic Kohen's Kappa coefficient is also calculated as explained in [48].

4.3.1 Agreement without chance correction

Percentage of instances on which both G and P_i agree. The agreement value agr_i for all items $i \in I$ is 1, if both G and P_i have the same answer class when evaluated using IOU.
Observed agreement over the values agr_i for all items $i \in I$ is then:

$$A_0 = \frac{1}{i} \sum_{i \in I} agr_i \tag{4.4}$$

The evaluation is performed using an IOU threshold of 50%. Table 4.8 shows % of agreement between each of P_i and G. We can infer that the ground truth conforms with each annotator by over 90% in call cases.

P_1	0.95
P_2	0.95
P_3	0.93

Table 4.8: Percentage of agreement between ground truth G and each of annotators P_i

4.3.2 Chance corrected coefficient

Cohen's kappa score [49] express the level of agreement between two annotators. Here we consider the level of agreement between each of annotators and the ground truth. It is defined as,

$$\kappa = \frac{(p_o - p_e)}{(1 - p_e)} \tag{4.5}$$

where p_o is the empirical probability of agreement on the label assigned to any sample (the observed agreement ratio), and p_e is the expected agreement when both annotators assign answer class labels randomly. κ is estimated using a per-annotator empirical prior over the answer class labels.

P_1	0.859
P_2	0.866
P_3	0.833

Table 4.9: Cohen's Kappa score between ground truth G and each of annotators P_i

Table 4.9 indicates Cohen's kappa score calculated between each of annotators P_i and ground truth G. near 1 values of the score indicates almost perfect agreement.

Figures 4.1 - 4.3 represent some examples of neural attention, eye-gaze and quantized eye-gaze plotted on to the image. The question and the inference are provided in the caption.

(a) Attention (b) Gaze (c) Gaze after quantization

Figure 4.1: Question:*How many horses are there?* , Network gave wrong answer, attention fails to highlight all the relevant regions

(a) Attention (b) Gaze (c) Gaze after quantization

Figure 4.2: Question: *What street number is this?* , Network gave correct answer, but wrong attention

(a) Attention (b) Gaze (c) Gaze after quantization

Figure 4.3: Question: *What is in the bowl?* , Both network attention and eye-gaze highlight same region

Conclusion and Future Works

Human cognitive reasoning process and how eye helps shape it has been a topic of study for decades [16]. The vision process takes up about 30% of the brain cortex compared to that of touch which takes about 8% and hearing merely 3% shows the importance of eyes as information gatherers in human body. Eye-gaze as a human attention modality has been successful in providing insightful results in text processing domain such as, for tasks which involve both understanding nature of the reader as well as understanding the nature of the text. This has an advantage of being intrinsic representative of human cognition, which when combined with carefully designed experiments can have a pivotal role in Artificial Intelligence tasks. In future, availability of eye-trackers and the sensor data would be much common, facilitated by high definition mobile phone cameras and sensors, we can hope some action in this field in the future. Right now, phone eye sensors are of limited use up to tasks like automatic scrolling, screen wake and shutdown according to eye movement of user, etc.

We also observed the rising popularity of neural attention ever since the inception and the wide variety of problems it solves. Attention is usually interpreted as the counter part of human information sensitive reasoning in neural networks [28]. Since neural networks are evolving in the direction to achieve human level performance or more, knowing if neural attention delivers what it promises is a matter of interest.

Recently, there has been a number of research in VQA highlighting how well human attention can provide visual grounding for the neural attention, and how to incorporate human attention using a loss function to improve the network [14]. Also, there are results that shows the improvements are not an effect of additional grounding provided via human attention [15]. However, with the motivation these results give, we can expect more research in this matter in the future. This also demands an analysis of whether human and machine looks at the correct place currently.

The intention of this project was to see how reliable neural attention is while providing visual grounding for the decision basis. i.e., how well they point out human sensible regions in images during predictions. Empirically showing neural attention visualized on some of the input examples as a means of logical correctness of the experiment is seen commonly in publications today. This property may appear trivial in high confidence straight forward predictions yet it may not reflect the trend in lower confidence prediction inputs. This demands the need for analyzing how well present day networks capture human sensitive regions in input signals. The natural candidate to capture human cognition in figures is eye-gaze.

We designed experiments to capture both human and machine attention in the process. Upon

evaluation, taking the Intersection Over Union(IOU) as a metric, we see that for an IOU of 50%, we see that nearly 60% of the times, the attention does not exactly highlight human sensible regions. We also observed that there is no significant difference in how neural attention highlights the area in case of true and false predictions. For this evaluation, we took a small subset of carefully chosen input data that can capture the trend in those low confidence answers where the inference might matter. At present, we see that neural attention may not capture the intended trend in input signals. Clearly, the illustrations made from certain high confidence answers are not to be trusted to capture the whole trend. The attention mechanisms still need refining to match human level performance, which would be an interesting research to see in the future.

Bibliography

[1] Nitish Padmanaban, Robert Konrad, Tal Stramer, Emily A. Cooper, and Gordon Wetzstein. Optimizing virtual reality for all users through gaze-contingent and adaptive focus displays. *Proceedings of the National Academy of Sciences*, 2017.

[2] Holmqvist K Orquin, J.L. Threats to the validity of eye-movement research in psychology. *Behav Res*, 50:1645–1656, 08 2018.

[3] Geoffrey Underwood, Peter Chapman, Neil Brocklehurst, Jean Underwood, and David Crundall. Visual attention while driving: Sequences of eye fixations made by experienced and novice drivers. *Ergonomics*, 46:629–46, 06 2003.

[4] Shahram Eivazi, Roman Bednarik, Markku Tukiainen, Mikael Fraunberg, Ville Leinonen, and Juha Jääskeläinen. Gaze behaviour of expert and novice microneurosurgeons differs during observations of tumor removal recordings. *Eye Tracking Research and Applications Symposium (ETRA)*, 03 2012.

[5] Susann Fiedler and Andreas Glöckner. The dynamics of decision making in risky choice: An eye-tracking analysis. *Frontiers in psychology*, 3:335, 10 2012.

[6] Taylor Hayes and John Henderson. Scan patterns during real-world scene viewing predict individual differences in cognitive capacity. *Journal of Vision*, 17:23, 05 2017.

[7] Thora Tenbrink, Holly Taylor, Tad Brunyé, Stephanie Gagnon, and Aaron Gardony. Cognitive focus affects spatial decisions under conditions of uncertainty. *Cognitive Processing*, 21, 01 2020.

[8] M. A. Just and P. A Carpenter. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87, 1980.

[9] Charles Lima Sanches, Olivier Augereau, and Koichi Kise. Estimation of reading subjective understanding based on eye gaze analysis. *PLoS ONE*, 13, 2018.

[10] Abhijit Mishra and Pushpak Bhattacharyya. *Scanpath Complexity: Modeling Reading/Annotation Effort Using Gaze Information: An Investigation Based on Eye-tracking*, pages 77–98. ACL, 08 2018.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[12] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6281–6290, 2019.

[13] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *Proc, ICLR*, 2017.

[14] Ramprasaath R. Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Dhruv Batra, and Devi Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. pages 2591–2600, 2019.

[15] Robik Shrestha, Kushal Kafle, and Christopher Kanan. A negative case analysis of visual grounding methods for vqa. 2020.

[16] Marisa Carrasco. Visual attention: The past 25 years. *pubmed*, 51, 2011.

[17] L. Itti, C. Koch, , and E. Niebur. A model of saliency-based visualattention for rapid scene analysis. *IEEE Trans. on Pattern Analysisand Machine Intelligence*, 20:254–1259, 1998.

[18] Katsuki and C. Constantinidis. Bottom-up and top-down attention:different processes and overlapping neural systems. *The Neuroscientist*, 20:509–521, 2014.

[19] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, USA, 2017. Curran Associates Inc.

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detectionwith region proposal networks. In *Proc, NIPS*, 2015.

[22] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proc, CVPR*, 2018.

[23] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

[24] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[25] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.

[26] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA, 2014. MIT Press.

[27] Dzmitry Bahdanau, Kyunghyun Cho, and Y Bengio. Neural machine translation by jointly learning to align and translate. *ArXiv*, 1409, 09 2014.

[28] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention, 2015.

[29] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California, June 2016. Association for Computational Linguistics.

[30] Abhishek Das, Harsh Agrawal, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[31] Tingting Qiao, Jianfeng Dong, and Duanqing Xu. Exploring human-like attention supervision in visual question answering, 2017.

[32] R Engbert, A Nuthmann, E M Richter, and R Kliegl. Swift: A dynamical model of saccade generation during reading. *Psychological Review*, 112(4):777–813, October 2005.

[33] E.D. Reichle and Heather Sheridan. E-z reader: An overview of the model and two recent applications. *Oxford handbook of reading*, pages 277–292, 01 2015.

[34] Keith Reyner. Eye movements in reading and information processing: 20 years of research. *Psychol Bull*, 124(3):372–422, 11 1998.

[35] Sandeep Mathias, Diptesh Kanojia, Kevin Patel, Samarth Agrawal, Abhijit Mishra, and Pushpak Bhattacharyya. Eyes are the windows to the soul: Predicting the rating of text quality using gaze behaviour. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2352–2362, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[36] Joachim Bingel, Maria Barrett, and Sigrid Klerke. Predicting misreadings from gaze in children with reading difficulties. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 24–34, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[37] Abhijit Mishra and Pushpak Bhattacharyya. *Scanpath Complexity: Modeling Reading/Annotation Effort Using Gaze Information: An Investigation Based on Eye-tracking*, pages 77–98. 08 2018.

[38] Utpal Garain, Onkar Pandit, Olivier Augereau, Ayano Okoso, and Koichi Kise. Identification of reader specific difficult words by analyzing eye gaze and document content. pages 1346–

1351. 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 11 2017.

[39] Abhijit Mishra, Kuntal Dey, and Pushpak Bhattacharyya. Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 377–387, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[40] Charles Sanches, Olivier Augereau, and Koichi Kise. Manga content analysis using physiological signals. pages 1–6. proc. MANPU '16, 12 2016.

[41] Preethi Vaidyanathan, Emily T. Prud'hommeaux, Jeff B. Pelz, and Cecilia O. Alm. SNAG: Spoken narratives and gaze dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 132–137, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[42] Qiuxia Lai, Wenguan Wang, Salman Khan, Jianbing Shen, Hanqiu Sun, and Ling Shao. Human vs machine attention in neural networks: A comparative study. *ArXiv*, abs/1906.08764, 2019.

[43] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016.

[44] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1321–1330. JMLR.org, 2017.

[45] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2015.

[46] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[47] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

[48] Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596, December 2008.

[49] Jacob Willem Cohen. A coefficient of agreement for nominal scales. volume XX. Educational and Psychological Measurement, 1960.