A Comparative Study of Human and Machine Attention in Visual Question Answering

Ashish Rai, Aman Gupta, Nishtha Agrawal, Chandralekha Chinnam

(arr8134@nyu.edu, ag9900@nyu.edu, na3533@nyu.edu, cc8445@nyu.edu)
New York University, Courant Institute of Mathematical Sciences

Abstract

Vision has been a long-researched domain in Human Cognitive Modeling. Human vision is a complex process that involves attention, perception, and cognition. Over the years, extensive research has been dedicated to understanding the intricacies of human visual processing, resulting in various computational models attempting to mimic these processes. Concurrently, recent advancements in deep learning have led to the emergence of visual language foundation models, such as SAN (Stacked Attention Network) and Contrastive Language-Image Pre-training (CLIP) and their variants, which have showcased remarkable success in understanding and generating text based on visual inputs. The aim of this study is to compare attention mechanisms in human vision with those in visual language foundation models, focusing on the specific task of visual question answering (VQA). Our primary dataset for analysis is VQA-HAT, a comprehensive dataset designed explicitly for studying human attention in VQA tasks. By contrasting attention patterns in humans with those predicted by state-of-the-art visual language models (via saliency maps), we seek to gain insights into the similarities, differences, and potential advancements in attention modeling. In this study, analysis involves examining the similarities and differences in attentional focus, the regions of interest, and the overall attentional strategies.

Keywords: VQA, Attention Map, CLIP, SAN

1. Introduction

A key component of human cognitive processing is attention, which allows for quick interpretation of visual images by focusing attention on particular details rather than observing things in their whole. Because of the human capacity for prioritizing visual stimuli, deep learning and cognitive science have seen a rise in interest, which has resulted in the creation of computational models of attention[1].

The purpose of our research is to examine human attention and develop a model of attention tailored to the Visual Question Answering (VQA) problem. When allocating attention to visuals, humans concentrate on pertinent areas or objects according to the question posed. They focus on the most important features, such as objects, situations, or details that are essential for understanding, as they scan the image to find significant elements that help answer the question. This procedure is comparable to how models instinctively focus on various aspects of images in order to comprehend their meaning[2].

By emphasizing regions of interest within the image or question that contribute to answer inference, attention maps—a commonly used technique based on heatmaps for clarifying Visual Question Answering (VQA) systems—help users understand the model's decision-making process. An essential component of cognitive processing is attention. The human mind is capable of quickly comprehending an image by focusing attention on individual elements of the picture instead of digesting the image as a whole. Computational models of attention have been developed as a result of the human ability to prioritize particular parts of visual information. This has prompted research in the fields of computer vision and deep learning.

Through meticulous evaluation and analysis, we aim to elucidate the strengths and limitations of each model, thereby facilitating informed decision-making and advancing the state-of-the-art in multimodal learning and representation.

Visual quality assurance algorithms also have to handle a wide range of questions regarding the image. These questions can be directly related to the image itself, like identifying objects like books placed under televisions or figuring out the color of a boat, or they can be more general, like figuring out why a baby is upset or which chair is the most expensive. Creating robust algorithms for visual quality assurance that can perform at human-levels is a critical step in the development of artificial intelligence technologies.

Moving away from the field of picture captioning, where a basic understanding is sufficient to produce broad descriptions, visual queries tend to interact with different aspects of an image, including background details and contextual foundations. This study emphasizes on how useful explicit or implicit attentional frameworks may be in VQA models to help with precise question answering. This research is focused on answering the following questions: 1) When asked to provide an answer, what particular areas of a picture do people usually focus on? 2) How well can deep

VQA models with attention mechanisms match human attentional patterns when choosing areas of images?

Within the particular field of Visual Question Answering (VQA), attention map visualization over pictures has become a popular means of explanation, clearly indicating the relevant regions that the system uses to answer questions.
In this study, we examine two models: SAN and CLIP. SAN uses a unique multi-layer attention technique to gradually concentrate on the pertinent image regions needed to provide an answer to a particular inquiry. The SAN is able to identify the precise image regions that are strongly suggestive of the solution by progressively filtering out unnecessary visual information thanks to the multi-step reasoning made possible by the stacked attention layers[5].

The most advanced model for comprehending text and images together in a multimodal environment is called CLIP. Through extensive training on an extensive collection of photos and text from the internet, CLIP learns to correlate images with written descriptions. This makes it possible for CLIP to comprehend picture content in a manner that is comparable to that of humans[3].

In this work, saliency maps were generated and compared with human attention maps using models trained with CLIP and SAN. We seek to comprehend the interpretability and human perception correspondence between model-generated and human attention, offering insights into the visual reasoning skills of the model.

2. Related Work

Attention mechanisms in Visual Question Answering (VQA) have garnered significant attention within the research community. A prevalent approach involves utilizing convolutional neural networks (CNNs) to identify pertinent regions within an image in response to a posed question. One notable model, the Stacked Attention Networks (SAN) as introduced by Yang et al. (2015), leverages Long Short-Term Memory (LSTM) encodings of question words to establish a spatial attention distribution across the features extracted by the convolutional layers of the image.

Building upon this framework, the Hierarchical Co-Attention Network proposed by Lu et al. (2016) introduces a multi-layered approach to image attention, considering words, phrases, and entire questions. This model has achieved prominence, emerging as the leading solution in the VQA Challenge.[2]

In recent studies, an intriguing method involves employing question parsing for the construction of neural networks using modular components, with attention being one of the tasks handled by these modules. Notably, these endeavors primarily revolve around unsupervised attention models, wherein "attention" serves merely as an intermediate variable, specifically a spatial distribution generated by the model to optimize downstream loss, typically measured through VQA cross-entropy. It should be emphasized that the interpretability of some (the exact proportion remains uncertain) of these spatial distributions arises serendipitously. Conversely, the investigation focuses on the gaze patterns of humans when addressing visual inquiries. These human-generated attention maps present a means to assess unsupervised attention maps.[2]
Human studies have extensively explored the acquisition of eye-tracking data from human participants as a means to comprehend image saliency and visual perception. The utilization of eye-tracking data for investigating natural visual exploration has been acknowledged for its utility, yet its acquisition on a large scale is challenging and cost-prohibitive. Jiang et al. (2015) introduced mouse tracking as a viable method for acquiring precise attention maps.[2]
Alternative models incorporate visual attention mechanisms, thereby facilitating the elucidation of the network's learning process in selectively attending to pertinent local regions within the image correlating to the semantic content of the posed question. It is noteworthy to observe that within an initial study on Visual Question Answering (VQA) , the rudimentary Bag-of-Words (BOW) technique combined with image features demonstrates superior performance compared to Long Short-Term Memory (LSTM)-based models. This assessment was conducted on a synthetic VQA dataset derived from image captions within the COCO dataset.[1]
Recent efforts in the field have shown an increasing interest in the exploration of visual question answering (VQA). However, it is noteworthy that these endeavors primarily operate within constrained environments, often utilizing synthetic datasets of limited scale. For instance, the works confine its analysis to questions pertaining only to a predefined closed domain comprising 16 fundamental colors or 894 distinct object categories. Similarly, the other model adopts a template-based approach, restricting inquiries to a fixed lexicon encompassing objects, attributes, and spatial relationships.[6]

3. Methodology

3.1 Dataset And Features

A customized dataset called VQA-HAT (Visual Question Answering Human Attention) was created to make it easier to analyze how human attention matches the attention patterns produced by artificial visual language models when doing Visual Question Answering (VQA) activities.
Two foundational datasets serve as the basis for the VQA-HAT dataset:

- Microsoft Common Objects in Context, or MSCOCO: MSCOCO is a comprehensive dataset that is extensively utilized within the computer vision field. It offers a vast amount of image data that is richly annotated with details on object segmentation, object detection, and captioning.
- VQA version 1: A benchmark dataset for assessing visual question answering models is the VQA dataset version 1. It asks questions concerning the substance of the photographs in pairs and provides multiple-choice solutions.

Human attention maps are added to the VQA-HAT dataset, which improves upon the original VQA dataset. These are the maps:

- Produced from Eye-Tracking Information: The VQA-HAT dataset's human attention maps are created using eye-tracking information gathered from participants while they respond to questions on the pictures. This method records the areas of a picture where people typically glance when coming up with a response to a particular query.
- Designed to Compare Models: The dataset allows researchers to directly compare the attention mechanisms of AI models (such as those in SAN and CLIP) with human visual attention by making these attention maps available. This comparison can provide insights into a model's interpretability and decision-making processes by demonstrating how closely a model's "attention" resembles that of humans.

3.2 Models

3.2.1 Stacked Attention Network

The Stacked Attention Network (SAN) architecture consists of an image model using a CNN to extract feature vectors for different image regions, a question model like an LSTM or CNN to encode the question into a vector representation, and one or more stacked attention modules. SAN takes the question vector and combines it with the image region vectors through a feedforward network to produce an initial attention distribution over the image regions. It calculates a weighted sum of the image region vectors based on this attention and combines it with the question vector to produce a refined query vector. This refined query vector is then fed into the next stacked attention layer to compute a new attention distribution over the image regions. This process repeats over multiple stacked attention layers, allowing each layer to focus more sharply on the relevant visual regions for answering the question using the previous layers' representations. Finally, the attended visual features from the last stacked attention layer are combined with the final query vector to predict the answer [5].

3.2.2 Contrastive Language-Image Pre-training (CLIP)

Vision-language models (VLMs) like CLIP [3] have shown significant improvements over traditional attention-based models like Stacked Attention Networks (SANs) for multimodal tasks involving vision and language. VLM models like CLIP consist of 3 key elements: an image encoder, a text encoder, and a strategy to fuse representations from the two encoders. These key elements are tightly coupled together as the loss functions are designed around both the model architecture and the learning strategy [4].
Subsequently, the encoded outputs from both encoders are projected in a joint-embedding space. A contrastive loss maximizes cosine similarity between matched image-text embeddings while minimizing similarity with other pairings in a training batch. Notably, CLIP diverges from conventional vision-centric training paradigms by capitalizing on the wealth of language supervisions derived from a corpus of 400 million web-crawled image-text pairs. This enables CLIP to undertake diverse image classification tasks without the need for task-specific optimization.

CLIP (Contrastive Language-Image Pre-training) models have shown promising results for visual question answering (VQA) tasks by leveraging their strong multimodal representations learned from large-scale image-text data. The pre-trained model can be fine-tuned on a VQA dataset of interest to increase the accuracy of the model.

3.3 Training

3.3.1 SAN Training

For SAN, the VGG19 backbone is used as the visual encoder and a LSTM network to extract the text embeddings. Two attention models were used to produce a spatial attention distribution over the visual features of the image. The model is then trained to reduce Cross entropy loss on the vQA v1 dataset. This model is referred to as SAN-2 hereafter.
The attention map generated by SAN is the actual activation of the second attention layer. This is stored during the forward pass of the model and laid over the input image to

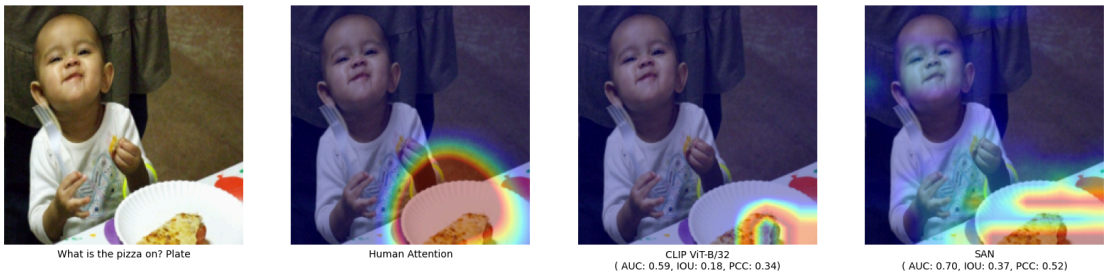

Figure 1. (a)

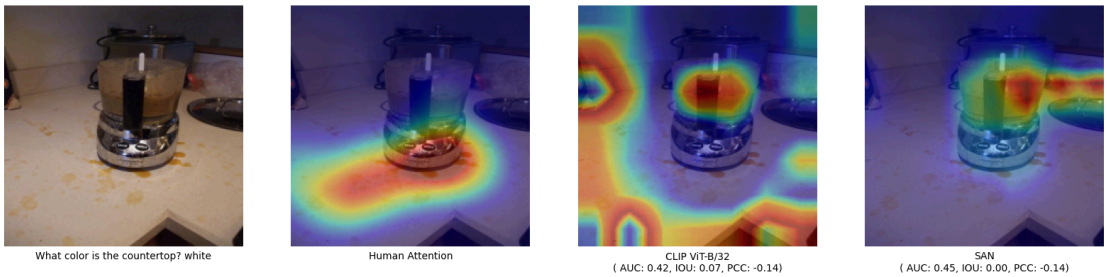

Figure 1. (b)

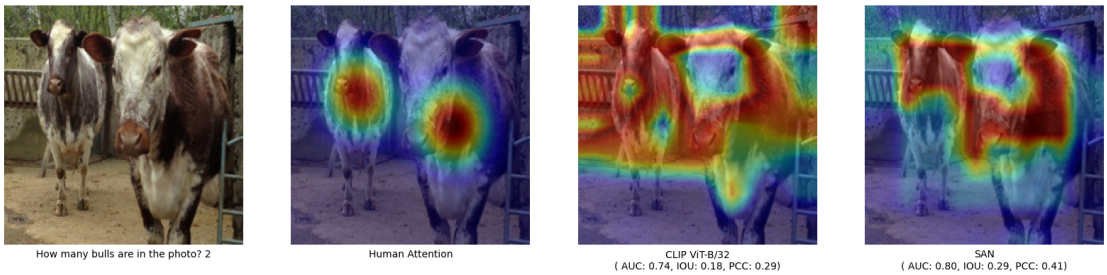

Figure 1. (c)

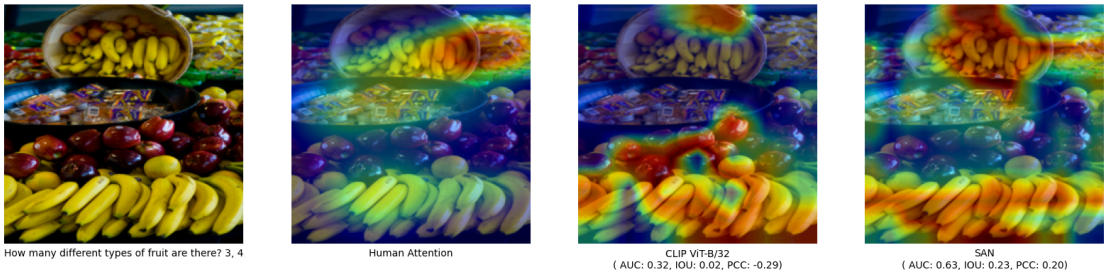

Figure 1. (d)

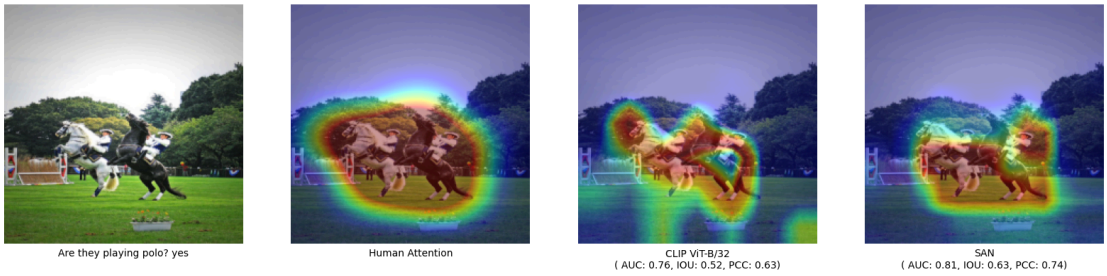

Figure 1.(e)

Figure 1. : Random samples of human attention (column 2) v/s attention in VQA models (columns 3 and 4). The comparison scores with respect to the human attention map is shown below model generated attention maps.

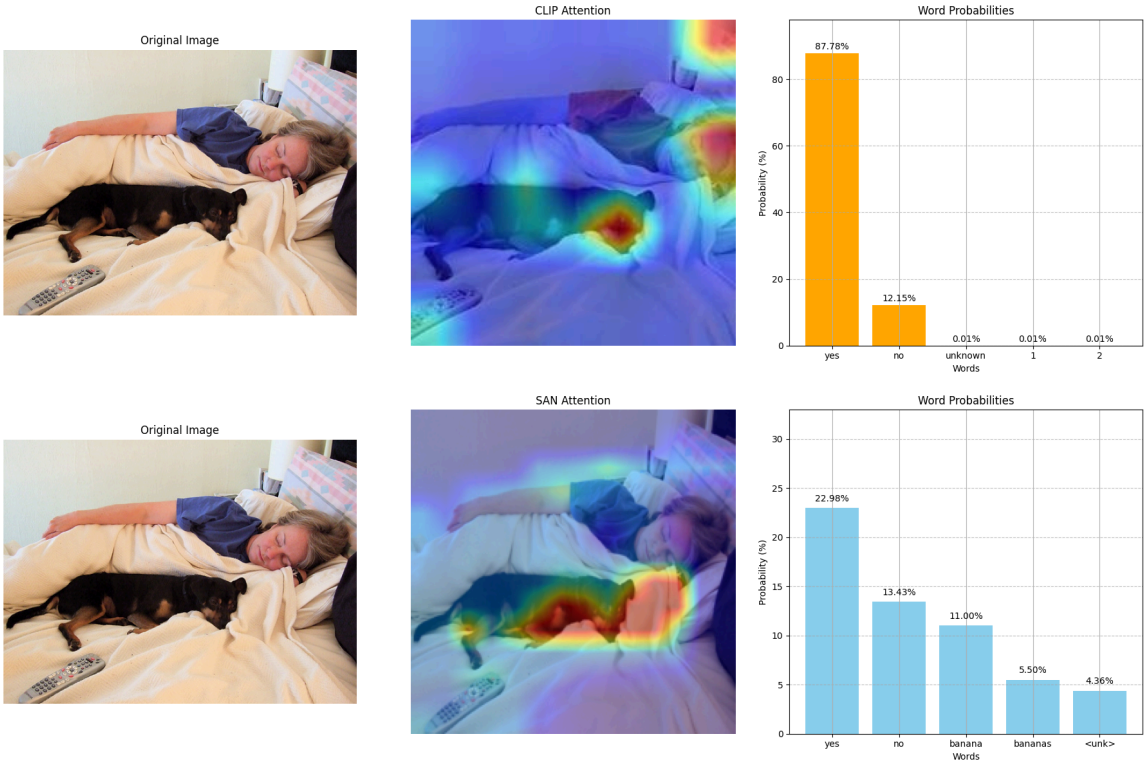

Figure 2. : Correlation of attention with model output logits

visualize the regions of the image the model attends to for answering a question.

3.3.1 CLIP Training We use a cross-modal attention layer to combine the image and text embeddings and then pass it through linear layers to predict the output tokens.
The output is 3129 class logits, similar to other contemporary VLM models. Cross Entropy is used to train the model on the dataset. The initial weights are loaded from the CLIP-ViT-B/32 pre-trained weights released with the original paper [3].
The attention map for CLIP is generated using the GRADCAM [7] algorithm. The activations from the last layer of the visual encoder are weighted with the gradients of that layer with respect to the class index with the highest likelihood.

4. Results and Analysis

4.1 Attention in Humans vs VQA Models

We compare the attention maps from the human dataset (VQA-HAT) to the saliency maps generated by the two VQA models:

- SAN-2: Stacked Attention Network with 2 attention layers
- CLIP-ViT-B/32: CLIP model with ViT-B/32 backbone

Our project has been linked to human cognition and behavior. We have contrasted the output of our model with the specific dataset of ratings or human behavior.

We present both statistical and qualitative comparison of human attention maps to model generated saliency maps in the following sections.

4.1.1 Statistical Comparison

Table 1 gives the accuracy scores of the two different models based on their evaluation on the test set of the VQA v1 dataset. CLIP-ViT-B/32 achieves better accuracy owing to a better way of fusing image and text embeddings. The pre-training done in CLIP also helps it acquire important bias required to answer generic questions.

Table 1: Accuracy

Model	Accuracy
SAN-2	50.34%
CLIP-ViT-B/32	65.32%

Table 2: Mean similarity scores on the validation set

Metrics	AUC	IOU	PCC
SAN-2	0.67	0.3	0.35
CLIP	0.58	0.208	0.18

The average similarity scores [8] for the three important metrics—the Pearson Correlation Coefficient (PCC), the Area Under the Curve (AUC), and the Intersection over Union (IoU)—computed for both models are shown in Table 2. These metrics function as numerical gauges of the models' efficacy and performance in identifying key areas within the photos.

Area Under the Curve (AUC): It is calculated from the receiver operating characteristic (ROC) curve, AUC measures the trade-off between true positive rate and false positive rate. Since AUC does not take into account the negative rate, models are not penalized for attending to larger regions than required. It does not account for the magnitude or location of the salient regions.

Intersection over Union (IoU): IoU measures the overlap between the predicted saliency map and the ground truth map. It is commonly used in object detection tasks and can indicate the accuracy of the saliency map in capturing relevant regions. IOU does not account for true negatives either.

Pearson Correlation Coefficient (PCC): PCC measures the linear correlation between two sets of data. In saliency map evaluation, it can assess the similarity between predicted and ground truth maps.

From both the tables, it is seen that the similarity scores are inversely related to the accuracy when the two models are compared.

4.1.2 Qualitative Comparison

The attention and saliency maps are blended with the original image with an alpha value of 0.4 to visualize for qualitative assessment of saliency maps.

Figure 1 shows examples of human attention and machine-generated attention maps with corresponding similarity metrics where the predicted output matches the ground truth answer for both the models. We see three major takeaways from this study.

Firstly, VQA models are not “attending” to the same regions in an image for answering a question.

Compared to CLIP-ViT-B/32, SAN-2 attends conservatively and presents sharper attention (seen with darker red regions and less fuzzy yellow regions in the saliency maps). CLIP is able to attend to larger portions of an image and results in a better accuracy in generating the answer.

In contrast to lesser similarity with human attention, CLIP-ViT-B/32 presents a better separation of predicted class logits. This can be seen in Figure 2.

In this study we are comparing different models’ attentions with the human attention.
As seen in the above figures there is major overlap between the attentions acquired by both.
Table 1 shows the mean value of metrics that was found over the validation dataset.

4.2 Analysis and Discussion

4.2.1 Role Language bias in VQA

The task of Visual Question Answering (VQA) was chosen to model and compare language conditioned (via questions) attention in humans. However, answering a question also requires one to use linguistic abilities. While humans excel at interpreting nuanced language and context, VQA systems often struggle due to limited training data and biased datasets.

In Figure 1 (b), SAN-2 attends to a region where the blender is kept in the image for a question pertaining to the kitchen counter which is a prevalent example of language bias introduced in the model due to dataset bias. CLIP-ViT-B/32 is better in this regard as it generates lesser irrelevant answers.

Cultural biases and the difficulty in aligning visual and linguistic cues further complicate the task for VQA models. Unlike humans, these models lack adaptability and may fail to generalize to diverse scenarios as well.

4.2.2 Outside Knowledge in VQA

Humans utilize vast external knowledge to answer questions, whereas VQA systems primarily rely on visual and textual features, facing challenges in reasoning with external context. Humans also possess the ability to better guess an answer with the help of inherent biases of

grammatical knowledge whereas a VQA model is entirely dependent on approximating a distribution over the image and textual features. The pre-training of CLIP helps to incorporate outside knowledge to better answer difficult questions.

4.2.3 On Recording Human Attention

VQA-HAT does serve as a good proxy for human attention for two major reasons:

- There is no upper bound on the size of the attention representation unlike that in the human visual cortex.
- The methodology in VQA-HAT generates a reward-driven conservative attention map which might not directly correlate with how visual attention is performed in human vision.

Humans do not attend conservatively in a bottom-up manner. Better approximations of human attention modeling would help bridge the gap between human and machine attention.

4.2.4 Grounding visual concepts in language

Humans tend to ground visual concepts in language and vice versa. There is a strong correlation between the image and textual interpretation in humans. However, most VQA models are trained to approximate the distribution of image and text modalities.

This can be seen in Figure 3 for the CLIP-ViT-B/32 where it is asked an irrelevant question and it attends to random parts of the image to generate an irrelevant answer. VQA models can be trained with contrastive image-question pairs where the questions pertain to unrelated entities.

Drawing from developmental psychology, this research area aims to develop AI systems that understand and communicate about the visual world like humans.

5. Conclusion

VQA models have come a long way in incorporating image and text modalities to answer a question in natural language. We see that even with low correlation to human attention, the models are able to predict the answer with high confidence. This leads to a discussion of the importance of attention for visual question answering. Better ways of capturing human attention can help create better computational models of human attention. Human attention maps can be fed to models as error correction maps to better attend to the salient regions in an image.

Figure 3: CLIP Attention and response for contrastive entity questions

6. References

[1] Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., & Fergus, R. (2015). *Simple Baseline for Visual Question Answering.* arX-iv Preprint. arXiv preprint arXiv:1512.02167.

[2] Das, A., Agrawal, H., Zitnick, L., Parikh, D., & Batra, D. (2017). *Human attention in visual question answering: Do humans and deep networks look at the same regions?*. Computer Vision and Image Understanding, 163, 90-100.

[3] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). *Learning transferable visual models from natural language supervision.* In International conference on machine learning (pp. 8748-8763). PMLR.

[4] Song, H., Dong, L., Zhang, W. N., Liu, T., & Wei, F. (2022). *Clip models are few-shot learners: Empirical studies on vqa and visual entailment. arXiv preprint arXiv:2203.07190*.

[5] Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). *Stacked attention networks for image question answering.* In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 21-29).

[6] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). *Vqa: Visual question answering.* In Proceedings of the IEEE international conference on computer vision (pp. 2425-2433).

[7] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). *Grad-cam: Visual explanations from deep networks via gradient-based localization.* In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

[8] Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., & Durand, F. (2018). *What do different evaluation metrics tell us about saliency models?*. IEEE transactions on pattern analysis and machine intelligence, 41(3), 740-757.