Attention-driven Object Recognition

Motivation

Why do humans need attention? In computer science, our machine learning community experienced a revelation about attention through the paper "Attention is All You Need" (Vaswani, 2017). Our understanding of artificial intelligence shifted through comprehending languages and images via attention mechanisms. These mechanisms alone proved transformative for machine learning performance, and this breakthrough in artificial attention parallels our understanding of human biological attention systems. Considering in a driving situation, we seamlessly process multiple inputs - lights, roadside trees, cars, where we often has little memory on them and usually mindless of their presence. However, when a child suddenly runs into the street in front of us, our attention is snapped in. This automatic shift reveals the complex attention mechanisms that current AI systems haven't fully replicated. (Cvahte, 2019), and there are numerous of studies shows attention's influence on human's ability in recognition (Walker, 2018). For this project, my first goal is to determine how attention mechanisms influence our ability to recognize things in a biological setting.

My second goal began with an observation in movie theaters while choosing seats. Seat position fundamentally affects our viewing experience and always gives me different feelings. Sitting too close overwhelms the eyes with details, while sitting too far reduces the sense of immersion and loses nuance. The question arises in my mind why I am having these experiences? Does spatial location influence our ability to recognize things when we focus on them? Professional video game game players even measure the precise distance between the

screen and their eyes. This suggests to me that the spatial relation between observer and target actually matters, and by understanding how spatial relationships modulate attention in both artificial and biological systems, we can enhance our human computer interface and possibly develop more naturalistic AI attention mechanisms.

Model Structure

I started with building on the object recognition model with an attention mechanism.

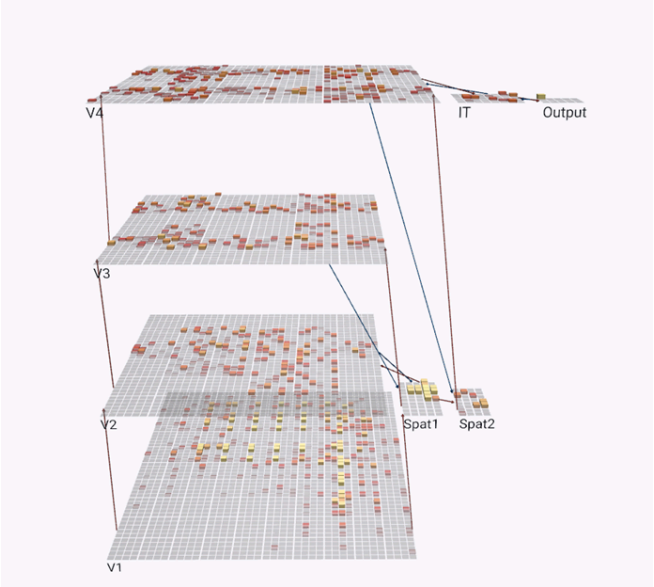

Figure 1. The model simulation in Cogent Core

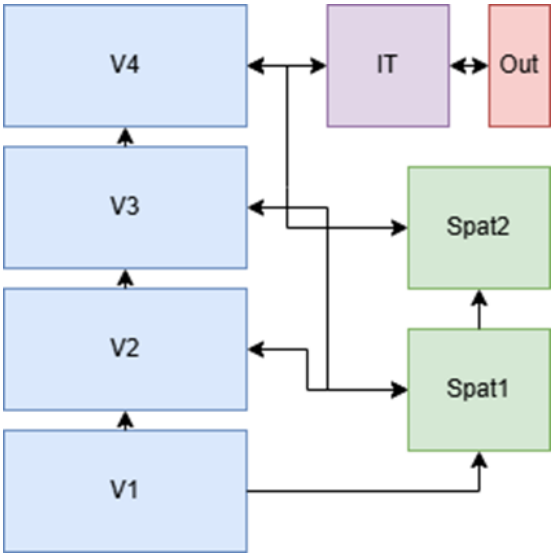

Figure 2. The model structure diagram

Figure 1 illustrates the actual model implementation, and Figure 2 refers to the diagram of model structure and connections. As visual signals enter the primary visual cortex V1, after feature extraction, it will pass the feature into deeper cortex layer V2, V3, and V4. The addition of V2 and V3 layers serves a critical theoretical role, reflecting the flexibility as we will increase the difficulty level of object recognition tasks and provide a more biological realistic structure. V2 neurons respond to more sophisticated combinations of orientations and textures than V1,

while V3 serves as an intermediate role that bridges the gap between simple and complex feature detections. Besides the biological visual cortex layers, the integration of spatial attention required some careful consideration of how information flows between visual and spatial processing streams, and I implemented them into two key layers Spat1 and Spat2. The early integration Spat1 has bidirectional connection with V2 and V4 layers, which intend to modulate intermediate feature processing, and later integration Spat2 is connected with V4 with bidirectional connection, which influences higher order object representations and facilitates object level selection with attention.

Model Setting

After the implementation of the expanded architecture with V2, V3, and spatial attention pathways. From this point, I expect the model should maintain a learnable state with error rates comparable to the original model, though requiring longer training time due to increased complexity. However, the network exhibited unexpected behaviors, struggling significantly to reduce training loss during object recognition training. I assume the additional processing layers are generating patterns that overwhelm the output layer in order to make clear decisions. This led me to question whether the spatial layers or the additional cortex layers were causing this learning failure. My first thought was to decrease the number of units in each layer, since we are adding more layers, there should be a balance between network depth and width. However, even with reduced unit, the model showed no signs of learning in the training epochs graph. I then considered that learning rate could be influencing the model's ability since there are more weights need to be adjusting during the training. I tested out if lowering the learning rates will

help, but the result shows it still shows not only no learning process and it takes much longer time in running the training, which is an unwanted approach.

After guidance on this I first found out that lowering the inhibition layer in the output will help the model restore the ability of object recognition in training, and the reason I am suggesting here is with more processing layers feeding into the output, the model needs to reduce the inhibition to allow a richer signals to influence the final classifications without being overly suppressed. I realized the inhibition level for each layer is crucial for preparing the function of the model. I started to test the inhibition parameters for each layer. For the initial V1 layer, I maintained the original parameters since modifying first-order feature extraction wouldn't be necessary. In the intermediate layers, inhibition parameters showed a significant impact. From my experiments, keeping default values led to a high peak percentage error compared to the lower inhibition parameters. While lower inhibition parameters reach a significantly reduced percentage error, testing revealed the network wasn't actually learning meaningful representations, and it was overfitting to training data. Where I realized the reduced inhibition allowed too many units to remain active, creating an easy learning situation that does not generalize to new inputs. After experimenting with more training, the training is dependent on the random seed but the direction of optimization is clear. I finalized the V1 inhibition parameter as default 2.5 to maintain strong initial feature selectivity while allowing the network to develop more distributed representations in higher layers V2 as 2.2, V3 as 2, and V4 as 1.8, the ending layers gradually increase back for clear decision IT as 1.9 and output as 2.0. From experience from previous experiments, the units of these layers are also modified by the logic of the training process, V2 and V3 layers use 4D layer (6,6,6,6). Though turning off weights between spatial and visual cortex pathways established our baseline, performance remained slightly worse than

the original model without V2 and V3. While they are certainly more to explore in optimizing, since I am doing a comparison test, and I am less care about the performance, comparing to the original model on this test is slightly worse. There are areas to explore for reasons, I hypothesize that this performance difference reflects a tension in which adding biological realism through intermediate stages may compromise a flexible ability in more complex tasks, but just like the human brain develops hierarchical processing over time, our model may require more complex tuning to fully leverage its ability, and this could be a trade-off between biological fidelity and performance.

As the baseline's setting is finalized, I am setting up the parameters for my comparison attention model. From analyzing the code from both original object recognition model and posner attention model, I plan to adjust weight scale in parameter settings with a Class to set up the paths. The path weights between spatial and object layers would need delicate turning in order to achieve the right balance of influence. Initially, I set stronger weights WtScale.Rel = 2 for spatial-to-object connections and WtScale.Rel = 0.5 for object-to-spatial connection, where the weights from posner attention model, my first thought is spatial attention need significant influence to modulate object processing, and test out what should be a relatively correct range for weight parameters. However, testing reveals this created overly dominant spatial effects that disrupted the normal object recognition task. Through iterative testing, I found that reducing these weights created more balanced interaction between attention and recognition processes. This process reveals a balancing state between attention and object recognition systems that need to work in harmony. I finalized the weights settings to WtScale.Rel = 0.2 for spatial-to-object connections and WtScale.Rel = 0.1 for object-to-spatial connection.

The next step of my setup is creating a way to test if the model can reflect object recognition faster, which would satisfy my goal 2 experiments. Taking inspiration from reaction time in the Posner attention model approach with CycleTheresholdStop, I implemented a similar approach in the object recognition model. I added it to the object recognition model. Initially, I was ambitious and tempted to track activation at multiple processing states; it needed to be fixed, and many buggy issues were produced by printing out the logs. But after all, I found out this wouldn't give us more insight into how recognition unfolds. Thinking deeper into understanding the recognition process, I realized something fundamental: the output layer represents the moment of conscious recognition and the point where the network has not just processed features but also reached a reportable decision. This mirrors how humans might internally process visual information through multiple stages but only report our recognition when it reaches conscious awareness. When we identify an object, we don't have access to our intermediate visual processing and test out when we recognize things. I then modified the code to log out the decisions at the output layer. By logging the output activation with each category, I can then measure and compare the reaction time. However, as I delved deeper into object recognition dynamics, I realized that a simple threshold setting, like the original poster task, is not sufficient. An object recognition task is not just about reaching a certain activation level and caring about only target detection; it is also about achieving a clear differentiation between the target and other alternatives; this insight led me to develop a more strict measure approach. The first solution I tried was to increase the threshold for target activation from 0.5 to 0.7, which would be a reasonable threshold, but this led to a problem in that most of my experiment runs reach and report a maximum number of reaction time cycles. I realized here increasing the threshold is not working well since, most of the time, the activation will not achieve that high. A more

sophisticated approach I then came out with is to stop the recording at the time that the target activation reaches the threshold and shows a compelling competitive inhibition with other alternatives, so I got the highest activation from the alternative unit and set the difference of the target activation to be higher than it by threshold of 0.2. Additional notices to this modification, the first is since we do not care about training, and I am assuming in a real situation we want to test out the ability of people who already understand the object, so this reaction time is only recorded in testing on later experiments, but during training, there are also shows decreasing of reaction time, which is intuitively prove human brain becoming faster in object recognition things during training, and how faster the rate is led to an area could explore more details on. There is a limitation, it is observed from later experiments that in a harder task situation like the scale factor is a range of 0.5-0.6, the random seed play is an unstable factor for the network to show its ability, and with chances the reaction time is not successfully recorded, also this happened with a overtrained network.

Experiments

The first experiment I want to run focuses on evaluating the differences in performance under a scenario where the object is straightforward to understand, and the features are well-defined. To achieve this, I modify the code to generate the LED tasks with random offsets on the X and Y axes are set to 0, the scale is fixed at 1, and the rotation is set to 0. A notable setting in this experiment is the higher inhibition layer at the intermediate level compared to the configuration demonstrated previously(the setting later tested out will make the results more unstable for the simple tasks). Under these conditions, the models generally achieve stable error rates close to 0 within fewer than 15 epochs. With higher inhibition remains stable results for

correcting all the categories besides cat 18 and cat 19. With attention enabled, reaction times range between 23 and 24, whereas without attention, they fall between 19 and 20. This slower processing with attention makes intuitive sense in a simple object recognition task, as the network needs to perform additional spatial processing and filtering.

After the simple experiment with fixed LED patterns gave us insights into attention's basic role, I realized we needed to explore more realistic scenarios that better reflect real-world visual challenges. My next experiment focused on introducing random offsets, which would make the recognition task significantly harder but also more relevant to actual visual processing, also the offsets are introduced in the range to ensure the features are enough to be learned by the network. The random offsets are set as X, Y translation from range -0.25 to 0.25, scale as 0.7 to 1, and rotation is positive and negative 360 degrees.

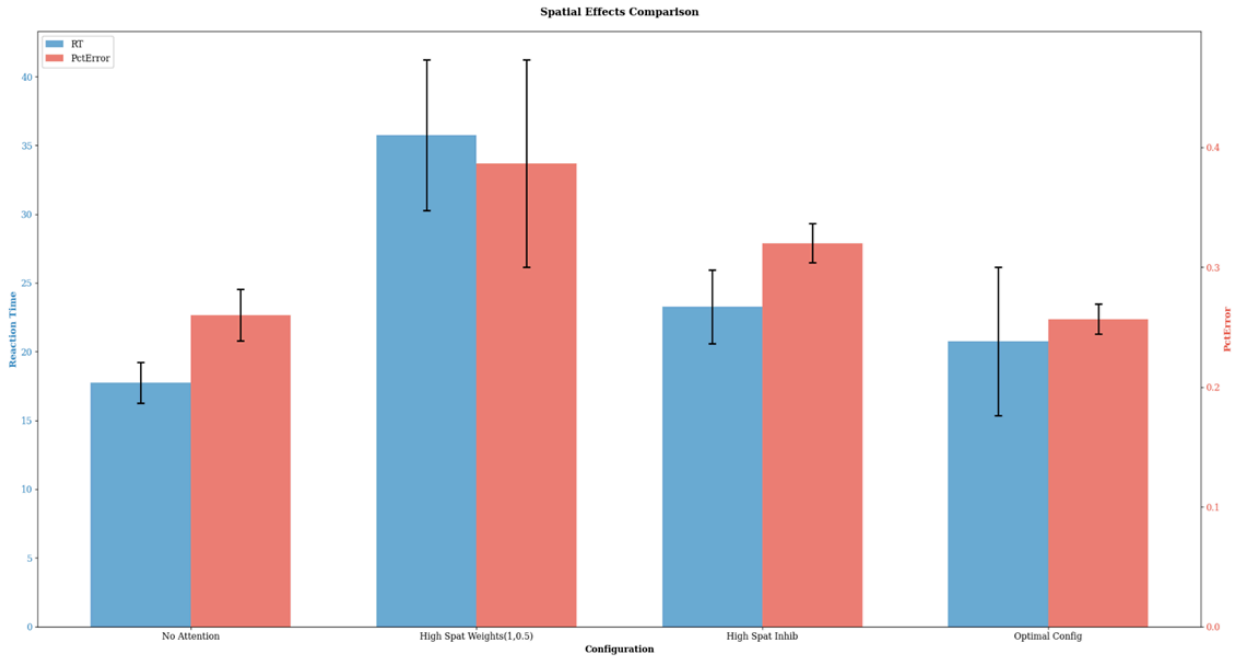

Figure 3: Spatial Effects graph compares the different settings of attention model compared to the baseline.

Figure 3 above presents a collection of experiment results from training and tested with random offsets. The parameters of visual processing cortex layers are discussed in my Model Setting section. The spatial layer parameters are modified here to assess their impact on performance. Typically, I run five epochs, discarding the highest and lowest results to mitigate potential negative effects caused by training randomness. Training stops once the percentage error reaches a local minimum. The No Attention column serves as the baseline where the weights of spatial pathways are set to 0. The second column High Spat weights column is the result are reported after setting spatial pathway weights higher than optimal, four settings of spatial-to-object and object-to-spatial weights are tested out; they are (2, 0.5), (1, 0.5), (0.5, 0.25), and (0.2, 0.1). Unsurprisingly, excessively high weights cause the network to fail entirely, leading to poor testing performance. The lowest weight setting (0.2, 0.1) shows the best performance across all and is identified as optimal. For comparison, (1, 0.5) is picked as the result slightly worse compared with (0.5, 0.25) to demonstrate typical effects of higher spatial weights. The High Spat Inhib column is the result of default inhibition of the spatial layers without reducing it to the adjusted value of 1.5. This adjustment, although small, represents the most significant change yielding distinct results. Initially, I used the default inhibition setting without question, but training revealed that the network struggled with spatial processing. Compared with the No Attention baseline, adding the spatial layer often degraded performance in terms of both reaction time and percentage error. The hypothesis I made behind it is that adding an attention mechanism to the simple task will cause the network to overthink since the features are straightforward. To test this hypothesis, I increased the offset ranges to make object recognition harder, and I hope the complex features would show the benefits of the spatial layer and make obvious performance improvements. However, repeated experiments disproved this

hypothesis. While examining the code, I suddenly realized the inhibition in the spatial layers needed theoretical adjustment. Since default inhibition is 2.5, it seems too restrictive and hurt the network's ability to allocate attention flexibly. After running with an ideal inhibition value of 1.5, it shows reliable performance improvement, as shown in the 4th column.

After the process of fixing and optimizing the parameters, now we can discuss the results from the graph. The baseline model without attention achieves an average percentage error of approximately 0.23 and an average reaction time of around 18. However, when the network is configured with high spatial weights, it exhibits significantly higher percentage errors compared to the baseline, it is not a small performance decline, and it reflects a fundamental disruption in the network's ability to process information efficiently, and suggests the network is focus too much on spatial location over visual features and destroy the harmony, much like how we overthinking the related attribute of something but ignores its core. When we set high inhibition, it shows better results compared to high spatial weights, it indicates rigid attention control impairs the network's ability to flexibly process visual information, but it still focuses excessively on certain parts of the spatial location, slightly degrading object recognition ability. The detailed effects between these two parameters should be adjustable and further compared outside the scope of this study. The breakthrough occurs with the optimal configuration, where both spatial weights and inhibition are properly balanced. Here, we see error rates comparable to the baseline while achieving a similar reaction time. Notably, the percentage error is slightly lower, and the mean is reduced, resulting in more stable accuracy when recognizing objects whose appearances constantly change. However, this stability comes with a trade-off, the average reaction times are slightly slower than the baseline and show greater variability. This balance reflects the complexity we see in biological systems, where improved accuracy

sometimes requires additional processing time. The wider range of reaction times might suggest that the attention system is more selective about when to commit the decision, much like how humans might take extra time to ensure accurate perception under challenging conditions, it also reminds us that in Posner cue tasks, the target appears in another direction different with the cue will make our brain take more time on processing, I think this could also be a reason why this situation happens and why the reaction time in the attention model has a higher deviation on reaction time.

For my second experimental goal, I used the scale range parameter to systematically explore how recognition performance varies across different simulated viewing distances. The range parameter provides a direct way to manipulate the apparent size of objects in the visual field, simulating how objects appear at varying distances from an observer. I designed 4 conditions, far range, mid range, close range, and extra close range. The Far Range has a range set from 0.1 to 0.2, which simulates maximum viewing distance, the Mid Range has a range set from 0.5 to 0.6 and 0.6 to 0.7, representing the intermediate distances and I intended to set it as a comfortable distance that we look at objects. Close Range 0.9 to 1.0 and 1.2 to 1.3 to approximate a nearby viewing, and extra close range 2.0 to 2.1, testing effects of extremely close viewing (The XClose setting may not completely reach 2.0 and 2.1, after examining from the LED image graph from training, it is demonstrating a limitation in maximizing the scale).

The Figure 4 below shows the performance are summarized results from 5 training sessions. Where the distance is far, the reaction time is consistently reaching the max of my limit. For the mid range 0.5-0.6 tests, there is a high chance that the network fails to achieve good activation states or reach a clear decision, depending on the random seed. Only runs with reasonable outputs are recorded for analysis. The network shows improved performance, with a

higher probability of valid runs, and these runs demonstrate lower percentage errors compared to the previous setting. For the results from two close ranges, the percentage errors keep low while being more stable, and the average of reaction time is showing an increase and more flexible in range.

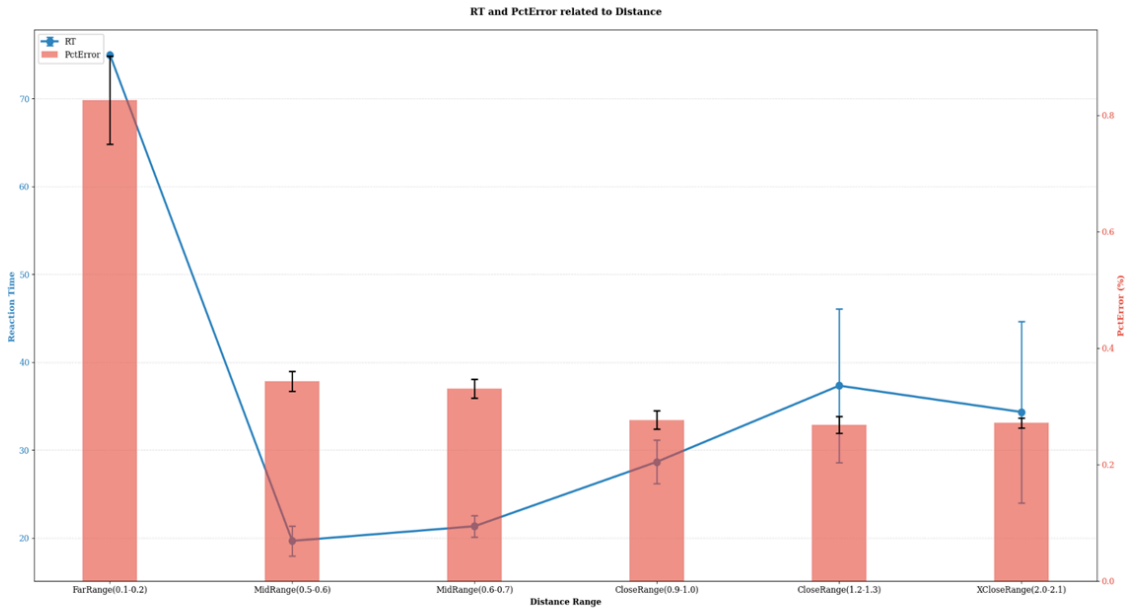

Figure 4: The RT and Percentage Error related to different Distance Range

Looking at the data, the results highlight the delicate balance between distance and recognition. The percentage error is showing a straight forward improvement as the scale factor increases. For reaction time, at far distances (as shown in the project slide), the network consistently struggles, pushing against the maximum cycle limits for reaction time. It reflects fundamental challenges in visual processing for objects at far distance that demonstrate with less first-order features. In the mid range distances, the 0.5- 0.6 range reveals high sensitivity to initial condition (random seed), marking a critical transition point in recognition(the range lower with 0.4-0.5 shows the same results as 0.1, 0.2), moving slightly further, the network finds more stable ground. This transition zone might represent something similar to our own visual system,

where we have enough detail for recognition while maintaining processing efficiency. The most surprising findings are in close range, 0.9-1.0 and 1.2-1.3, the network achieves more stable error rates in these ranges, but reaction times increase and become more variable. The sudden increase in the deviation of reaction times, particularly for the 1.2–1.3 range, was unexpected. This range shows a wider spread of reaction times compared to 0.9–1.0 and other previous settings. An interesting comparison arises with the optimal model from the previous experiment (scale range of 0.7- 1.0), which is not directly considered here, also demonstrating a wider range of reaction times while having the expected faster reaction times. My hypothesis is that as distance decreases beyond a certain threshold, spatial effects increasingly influence the network's processing. This shift may cause object recognition to rely more on location-specific features, which needs further exploration. For the extra far distance, the results are similar to those observed in the far range. However, the detailed relationships remain debatable due to limitations in the code's scaling capabilities, with the exact boundary values remaining undefined.
These findings tell about the relationship between distance, attention, and recognition. It is not simply that closer is better or that distance always impairs. Instead, there appears to be an optimal range where our visual system can balance detail and efficiency most effectively.

Future Studies

Multi-object recognition is a complex task that requires a more specialized model for thorough study and comparison. Taken from the optimal model from Experiment one, we can test out how attention works on multiple object recognition. Starting from LED environment code, I created a series of functions to create a situation that generates two different random LED objects, and these objects appear at distinct spatial locations with varying random offsets. This

setup allows us to examine how the model's spatial attention mechanisms handle multiple competing objects in the visual field. This experiment can be adapted for various purposes, such as focusing on both objects simultaneously or prioritizing one object over the other. Each purpose necessitates specific model modifications to test effectively. In the current setup, I configured the output to focus on the first LED object. Using this model, we can observe several key aspects of multi-object processing, although it remains limited in scope without additional adjustments. Figure 5 below shows how the activations shift during this process.

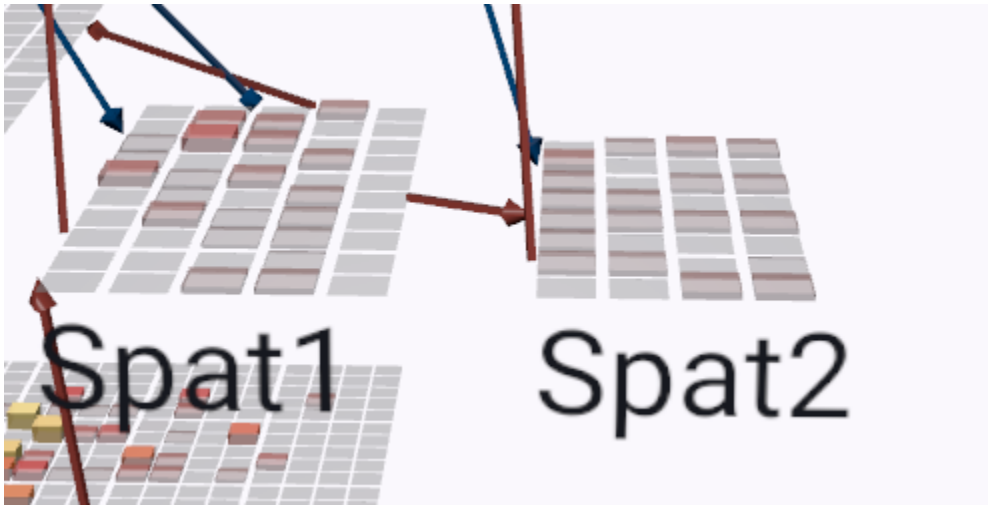

Figure 5. The activation of patterns during training on multi-object recognition

From the image above, in Spat1, we can observe two distinct regions of activity, likely corresponding to the spatial locations of the two LED objects. Between the initial layer and Spat1, there is strong activation concentrated in a small area at the top of the layer. After processing through Spat1, the activity in Spat2 becomes more spread out, demonstrating the network's capability to detect multiple objects simultaneously. This observation suggests that the network is actively processing the spatial layout of both objects, even without specific adjustments to the model ability during training.

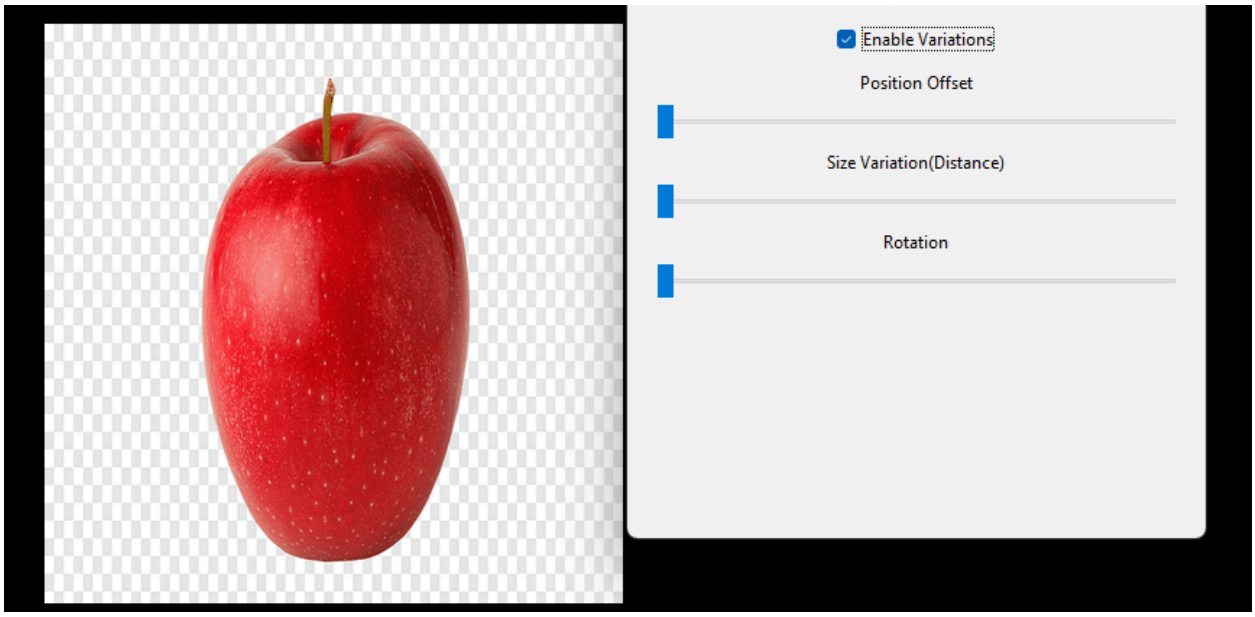

Figure 6. The current development of object recognition

Another area of study involves applying this knowledge to real world applications. Inspired by the Human Benchmark website, which collects and analyzes reaction time data from users, I am developing a program to test the speed of people's reactions to stimuli on their screens. A meaningful challenge lies in statistically analyzing the relationships between human attention, reaction time, actions, and object recognition within this program. These interactions present a complex problem that needs further exploration.

Conclusion

The project investigated the relation between attention, spatial processing, and object recognition. Through experiments demonstrate the critical role of balancing spatial weights and inhibition in optimizing our object recognition ability on both accuracy and reaction speed. More

insights into recognition across different distances and its potential influence on our perception of viewing. The future work focus on multi-object scenarios highlights the complexity of attention mechanisms that can be future explored and also a potential for real world applications. And hope these findings can contribute to our understanding and enhance attention-driven systems in domains like human computer interaction.

Reference

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I., Profile, A. V. B., Profile, N. S. B., Profile, N. P. R., Profile, J. U. R., Profile, L. J. R., Aidan N. GomezUniversity of TorontoView Profile, Profile, Ł. K. B., & Profile, I. P. R. (2017, December 4). *Attention is all you need: Proceedings of the 31st International Conference on Neural Information Processing Systems*. Guide Proceedings. https://dl.acm.org/doi/10.5555/3295222.3295349

Cvahte Ojsteršek, T., & Topolšek, D. (2019). Influence of drivers' visual and cognitive attention on their perception of changes in the traffic environment. *European transport research review*, *11*(1), 45.

Walker, H. E., & Trick, L. M. (2018). Mind-wandering while driving: The impact of fatigue, task length, and sustained attention abilities. *Transportation research part F: traffic psychology and behaviour*, *59*, 81-97.