VERIFYING PERCEPTION AS A SENSORY-MOTOR LOOP PROCESS THROUGH DEEP LEARNING

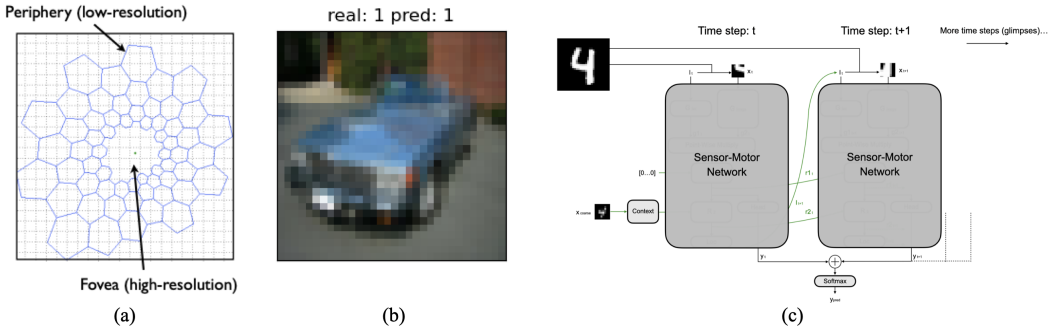
Yingsi Qin, Gary Liu

1 ABSTRACT

We computationally verified the Neuroscience hypothesis of the hierarchical structure of our visual cortex being a nested sensory-motor loop. We built a neural network framework that explores the impact of a sensory-motor loop structure of a network on the efficacy of the representation of a scene. The network architecture, specifically, is of a recurrent structure that mimics the hypothesis of V1, V2, V3. etc., being in a recurrent loop. We have the network attend itself to one small region of interest (a glimpse) in the image at a time, and learn to decide where to gimpse next based on the location and feature information of the current glimpse. Since our sensory-motor network (SMN) can predict the next sequential locations to attend to, it is thus able to, in between glimpses, develop a semantic understanding about the global object in the scene.

To train the SMN, we adopt representation learning techniques to ensure the construction of the latent space, while performing a downstream image classification task. We trained separately the "teacher" models which developed effective latent representation and use them to teach our "student" SMNs. The loss is a combination of the contrastive representation loss and the class prediction loss. Since our work is essentially a proof of concept, we trained on multiple datasets using the same SMN structure without increased complications. With the baseline SMN, we experimented on the 10-class MNIST, 1000-class Triple MNIST, and 10-class CIFAR-10 dataset, with the same architecture and achieved respective validation accuracy of 97.43%, 65.58%, and 52.26%.

If the performance is good, this can imply that the sensory-motor loop structure does work for visual tasks. Our results suggest that the sensory-motor loop structure does work on visual tasks. Compared to previous relevant works which used reinforcement learning instead of representation learning to train the networks, the main contributions of our project are twofold: (1) we use contrastive representation learning method and showed that the latent scene representations generated from limited input information (a glimpse) can be effective as well, (2) we increased the training efficiency by not using reinforcement learning techniques which update the gradients over every glimpse. We introduced a non-reinforcement learning method to train a sensory-motor loop network for visual attention.



(a) (b) (c)

Figure 1: Overview of motivation and model structure

Figure 1 shows a high-level overview of our motivation. (a) indicates the retinal transformation that is much like the fovea of the human retina, which extracts high-resolution information only in the neighborhood of the attention pixel, and extracts lower-resolution information at the periphery. This inspired us to write an attention mechanism of a small region. (b) is an image from the CIFAR-10 dataset. In a sensory-motor loop, the agent would attend to the essential small regions, one at a time, and from them, eventually develop a perception of the scene. Our network, shown in (c) on a high-level, is an analogy to this process.

2 RELEVANT WORKS

The vision system builds on the perception from a continuous motion of the eyes. It is commonly assumed that the saccadic exchange of regions of interest is included in sensorimotor loops. However, since actions purely based on sensor-feedback are slow and unstable, the brain developed a strategy generate a predictive feedback that estimate the current state of the effector, compensating the delays of real sensory feedback(Hagai Lalazar (2008)). It is proved to be necessary to use a forward model when implementing the sensorimotor system's feature of maximizing optimal performance(Todorov (2004)). Then the prediction from the model is compared with the real sensor's perception to compensate for the possible noises and inaccuracy of the prediction.(Wolpert DM (1995)).
This structure of learning is similar to the concept of reinforcement learning with recurrent models. The study in (Denil et al. (2011)) performed classification using restricted boltzmann machine with attention and decaying resolution toward the periphery of the gaze. Other works also include detecting the salient image regions, motivated by the idea that human perception are saliency detectors(Torralba A (2006)). But this approach may lose important information since they are mainly based on low-level image properties.
Our work is most similar to other attempts to develop a recurrent model of visual attention. Such approach will extract information from an image by selecting small patches of the image and learn through processing such small patches with reinforcement learning technique to update weights(Mnih et al. (2014)). The main difference is that our model will also learn from a teacher network. The teacher enables our model to improve predict accuracy not only by the accuracy of each glimpse, but also by the representation of the whole image.

3 METHODS

3.1 THE FRAMEWORK DESIGN

In order to verify learning a representational space through a computational sensory-motor loop, our network uses a recurrent sensory-motor neural network (discussed in section 3.1.1) while applying the contrastive learning method (discussed in section 3.1.2) to develop a localized, oriented, and sparse representation space, capable of spanning across the image space of the dataset being trained on. It has been previously shown (Olshausen (1996)) that such a representation space is closely similar to those found in the human primary visual cortex.

3.1.1 THE RECURRENT NETWORK: SENSORY-MOTOR NETWORK (SMN)

Recurrent neural networks (RNN) are repeated instances of the same network with a previous instance passing information to the next instance. It adopts the chain like structure with the output of one single cell being passed onto the next. The weights are shared across time, so the weights of the network are the same for each time instance.

Since our goal is to have the sensory-motor loop like network learn which region of the image to attend to in order to extract necessary information to calculate an effective representation, we adopt the recurrent architecture so that the network can adjust itself and orient to the next area in the image based on information in the current area.

Figure 2 shows the structure of our Sensory-Motor Network (SMN). The network learns to attend itself to a small region of interest (ROI) at a time, which we call a glimpse. The network is recurrent, as it is looped over how ever many glimpses we choose to take.

At the top of each time step, we see that a small glimpse x_t from the large image and the location l_t of the glimpse are fed into the network. Based on the location and the local features in the glimpse, the network outputs the next glimpse location l_{t+1} to look at, for time step $t+1$. At each time step, hidden vectors r_1 and r_2 from recurrent layers are passed as additional inputs into the next time step as well. For a classification task, the prediction y_t from each time step t also contributes to the final prediction y. We can then see that this structure allows the network to be able to decide where to look for the next glimpse based on the current glimpse, and that since it can guess the next location, the network develops semantic understanding about the global object between glimpses.

Essentially, the network develops a location-based feature understanding of the current glimpse, g_t. Then the first recurrent block R_1 uses the current g_t and its own previous output r_{t-1} to develop the current r_t. After that, the second recurrent block also takes its own previous outputs and the output from R_1 to develop vector r_2, which is used to generate the next glimpse location. In the mean time, the output of the first recurrent block

is used to make a glimpse-based prediction of the class. In Figure 2 when there are two time steps, the final prediction vector y is a Softmax of the sum of all y_t's obtained.

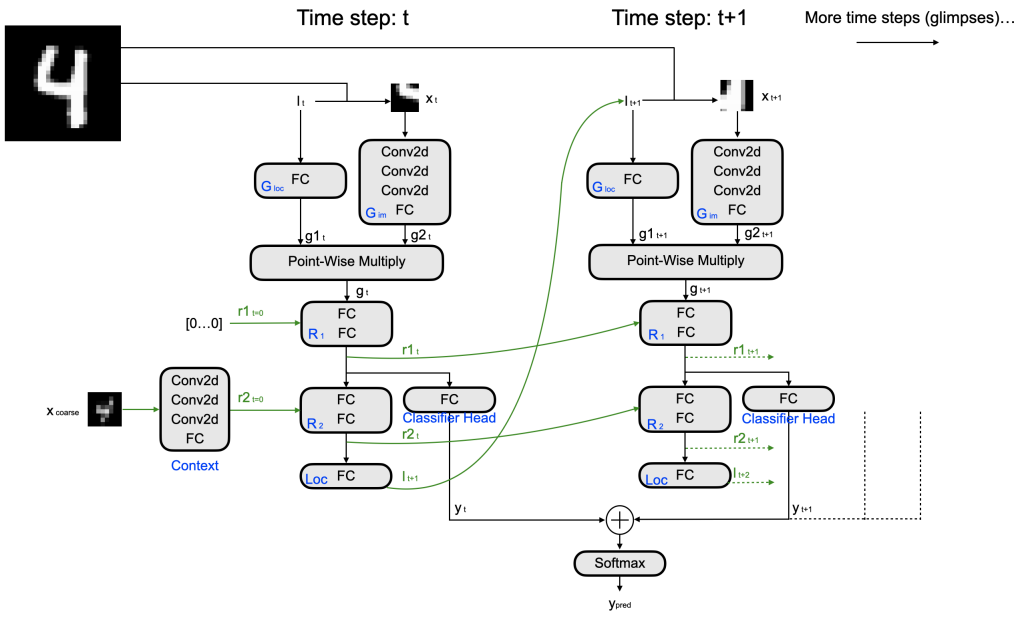


Figure 2: Our Sensory-Motor Network at 2 time steps

3.1.2 The Training Pipeline: Contrastive Representation Learning

Learning representations allows machine perception on natural images and thus enables different downstream vision tasks using the same representation. Contrastive representation learning, in particular, applies data augmentation methods such as jigsaw patching, perspective transformations, and rotations, etc, on the image inputs into the pipeline. Previous works by Chen et al. (2020) and Le-Khac et al. (2020) have shown that the composition of data augmentations plays a critical role in defining effective predictive tasks.

In our case, we adopt the contrastive representation learning method to allow our SMN to learn a localized, oriented, and sparse latent representation space. We do this by introducing a **Student-Teacher** model, where the student is our Sensor-Motor network and the teacher is a a well-trained Convolutional Neural Network that takes the entire image as the input. The teacher teaches the student using its latent representation via contrastive loss and the student learns via contrastive loss and a downstream prediction loss. The SMN needs latent representation learning because it needs to understand the image at the global level, instead of each glimpse. Our workflow is shown in Figure 3.

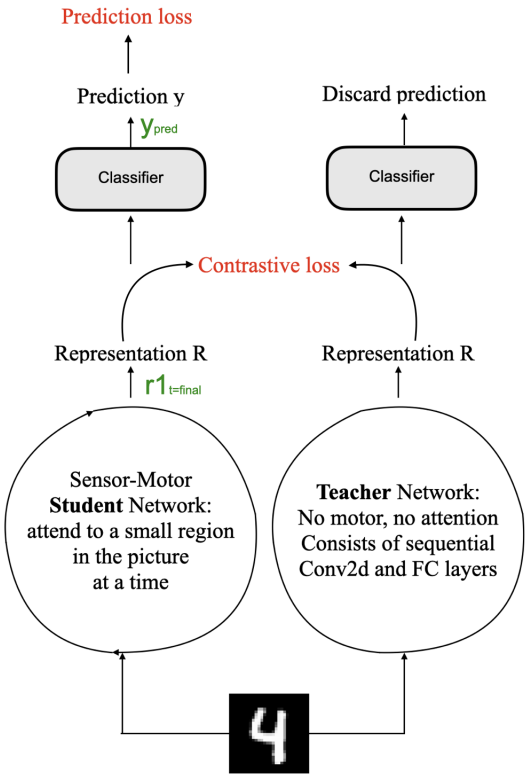


Figure 3: Student-Teacher Model for the Training Pipeline

4 Experimental Results

4.1 Training Pipeline

In training, we implemented a student-teacher model. The teacher model is a well-trained classifier trained with the whole image. It is used to teach the student with its latent representation via contrastive loss. Then the student will use both the contrastive loss and the prediction loss to update the weights accordingly. This structure is adapted with hope that the sensory-motor-student-network will learn through understanding the image at the global level, and use that information to improve it's ability to predict locations of glimpses.

4.2 Results

Data	MNIST	Triple-MNIST	CIFAR-10
# of Glimpses	4	9	4
Dataset Image Size	28x28	84x84	32x32
Teacher Model	3-layer Convnet	VGG-16	VGG-16
Training Accuracy(Student)(%)	99.34	65.94	99.58
Validation Accuracy(Student)(%)	97.43	64.73	54.46
Validation Accuracy(Teacher)(%)	97.99	77.90	85.98

Table 1: Training results on different datasets

4.3 Learning to Classifiy Digits

The SMN model is first tested on the traditional MNIST handwritten digit dataset. SMN is allowed to take 4 glimpses in total since this is a single-object-classification task. The size of the glimpsed patch is around 30 percent of the image size. A prediction vector is generated after each glimpse and the results are summed element-wise in the end to generate final prediction. This task evaluates the performance of SMN to learn single object with minimal noise. Overall, the model performs well in this baseline experiment.

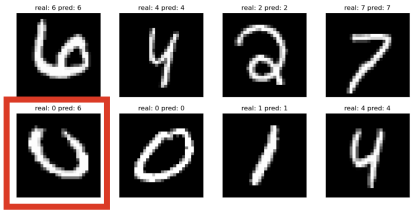


Figure 4: Validation results on the MNIST dataset

4.4 Learning to Classify Multiple Digits

Next, we test the performance of SMN with a more challenging task. The triple-mnist dataset contains 1000 classes and each class represents a digit from 0 to 999. There are 3 digits in the image and the label of the image is the number read from left to right. This task is more challenging because the 3 digits are sparsely distributed in the image–it is hard to predict the location of next digit after reading the first one. The accuracy is relatively low comparing to the first task. One reason is, as stated above, due to the digits being separate without global objects features that connect them. It becomes hard for the SMN to tell where to look next, based on the current glimpse.

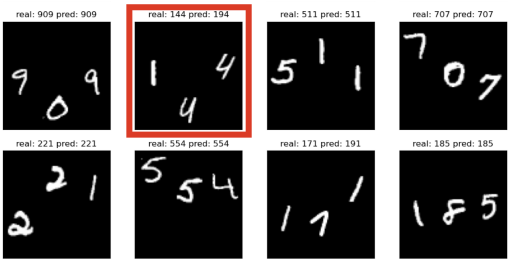


Figure 5: Validation results on the Triple MNIST dataset

4.5 Learning to Classify Different Objects

Finally, instead of digits,we evaluate the model based on it's performance on object classification. We choose the CIFAR-10 dataset for the object recognition task. The CIFAR-10 dataset contains 10 classes, each being a diverse object. We can observe that there is still a huge gap between our model's accuracy and the state-of-the-art model. CIFAR-10 is a hard dataset by nature. It is low resolution and many classes have common local aspects(for example, frogs and cats), which means that it is challenging especially for our model since it builds it's prediction purely on local aspects learned through glimpses. Our model does not perform well particularly for images of animals probably because of many common features shared among the animals. The training accuracy is higher than validating accuracy by 0.45, suggesting that the model is overfitting on the training data. This is most likely due to the huge amount of noise in the dataset. We will use more techniques like grey-scaling and data augmentation to prevent overfitting in further studies. In general, we can conclude that our feature is great at learning local aspects but not that good at classifying objects that shares common local feature. Our data is also sensitive to noises in the image and needs cautious preprocessing on the dataset before training. Furthermore, the teacher network does not perform well on the dataset as well, which may pass false information into the student network and confuse it. But in real-world-scenario(in the brain), we will have a very confident representation of the object that we are viewing and use that to improve our ability to predict labels with glimpses in future. This might also be part of the reason for this accuracy.

Figure 6: Validation results on the CIFAR-10 dataset

5 Future Work

Future work to explore the SMN on multi-object classification or 3D scenes would help understand the efficacy of the SMN model without reinforcement learning. It would be also helpful if future studies look into how localized, oriented, and sparse the latent space is on a dataset and how it changes with the number of different training hyper-parameters for the SMN. Another possible direction is, if given enough computational power, to expand the traning of SMN on larger datasets that have more diverse features and more classes that may share common local features, such as the ImageNet.

6 Appendix

The code, additional training details and params, and model checkpoints for this project are available via the google drive link here and the datasets are available on our Google Cloud Storage bucket here.

References

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

Misha Denil, Loris Bazzani, Hugo Larochelle, and Nando de Freitas. Learning where to attend with deep architectures for image tracking, 2011.

Eilon Vaadia Hagai Lalazar. Neural basis of sensorimotor learning: modifying internal models. *Current Opinion in Neurobiology*, 18(6):573–581, 2008. URL https://doi.org/10.1016/j.conb.2008.11.003.

Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934, 2020. ISSN 2169-3536. doi: 10.1109/access.2020.3031549. URL http://dx.doi.org/10.1109/ACCESS.2020.3031549.

Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention, 2014.

Field D. Olshausen, B. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(3):607–609, 1996. URL https://doi.org/10.1038/381607a0.

E. Todorov. Optimality principles in sensorimotor control. *Nat Neurosci*, 2004. URL https://doi.org/10.1038/nn1309.

Castelhano MS Henderson JM Torralba A, Oliva A. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychol Rev*, 2006. URL https://doi.org/10.1037/0033-295X.113.4.766.

Jordan MI. Wolpert DM, Ghahramani Z. An internal model for sensorimotor integration. *Science*, 1995. URL https://doi.org/10.1126/science.7569931.