Semantic Cognition Model from Images

Poornima Haridas, Kavya Sree Bhagavatula
New York University
New York City, NY 10012, USA
{ph1391, ksb456}@nyu.edu

Abstract

Humans are the best learners. They accumulate information in various forms of which visual input is the richest. The main purpose of this paper is to model a realistic semantic cognition model of human learning that incorporates both visual and factual components of the same. We also study the differentiation in development in the network as a result of learning.

1. Introduction

Semantic cognition is the ability of humans to use information acquired and stored in memory over their lifetime, to support innumerable verbal and non-verbal behaviours. Humans learn through a variety of input sources and visual aids are a primary component in human learning. We therefore try to model semantic cognition from images using neural networks.

McClelland [5] extensively studied differentiation and degradation in a neural network inspired semantic model. In this paper, we build on their idea to incorporate a visual front-end to their model that helps model the real world learning more accurately. We also study differentiation in our model given the visual input.

We tried to model the same in two different ways and compare it to the original McClelland[5] model as well. To formulate our model we first created a fact set for our world. A standard Convolutional Neural Network(CNN) model and a simple connectionist feedforward architecture as proposed by Rumelhart [6] [8] were used to learn our fact set. This hybrid network was then observed for differentiation.

1.1. Related Work

Very less work has been done in modelling vision based cognitive models using neural networks and this is our major motivation behind trying to model the same. On the other hand, a lot of extensive work has been done in the field of computer vision to model highly complex but accurate models for image recognition like ResNet[3], VGG[7] and AlexNet[4] (which we have used for our image recognition task). McClelland[5] modelled semantic cognition based on a set of facts and studied the differentiation patterns during learning and also the degradation of networks when noise is introduced to the network.

Therefore, we aim to explore this intersection between the advances in computer vision and semantic cognition models for visual perception.

2. Methodology

2.1. Data

The world that we have modeled contains 13 items and each item is associated with 5 relations. We gathered attributes for these 13 items with respect to each relation with many attributes being common to multiple items. There are a total of 74 attributes in our world. We modeled our world on 65 facts, each fact corresponding to an item and relation pair.
We extract images of our 13 items from Caltech-256 Object Category Dataset [2], publicly available dataset.

2.2. Convolutional Neural Network for Image Recognition

We use the well known CNN architecture AlexNet[4] (Figure 1) for our image recognition task. We used transfer learning to help us achieve high accuracy by first training AlexNet on the ImageNet dataset[1]. We then trained the network on our dataset sourced from Caltech-256[2]. We achieved a train accuracy of 97% and a validation accuracy of 91%.

2.3. Connectionist Feedforward Network To Learn Facts

This fully connected feedforward network is the crux of learning the facts of our world which describes each item and some of the item's selected properties. Figure 2 visualizes the model that is similar to the feedforward network

in [5]. The network consists of an item layer (13 neurons), relation layer (5 neurons), hidden layer (20 neurons) and the final attribute layer (74 neurons) that are trained with backpropagation algorithm.

3. Experiments/Models

3.1. Model 1

This model is a simple fully connected feedforward network which is an adaptation of the McClelland[5]'s network. This was created as a sanity check for our world of facts and to contrast with Models 2 3. This model assumes a clean input in the form of one-hot encoded arrays which cannot be assumed in the real world.

3.2. Model 2

This model incorporates the AlexNet and Feedforward networks to get our first new model. To understand the reasoning for this model, consider the following common real-world scenario of a child. During early stages of development, she takes in a wide range of visual input without much indication of what she is looking at. This is modelled as our AlexNet which is trained on our set of item images after taking advantage of transfer learning.

Next, as the the child grows, she is introduced to facts about the world where people around her point to things she has seen and augment the visual with facts about the scene("That is a dog!"). To model this, we then run through our set of images and obtain the predicted class, P, from AlexNet. The result of the AlexNet can be considered as the recognition that a child demonstrates on seeing an object they have seen before (act of bringing attention and pointing to the dog). This result is then fed as a one-hot encoded input into the feedforward architecture with known facts about the object ("Dog can bark"). This is repeated for all facts of all images in the set to complete training of our feedforward network.

3.3. Model 3

In the final model we take the probabilities of all classes produced by the AlexNet (in response to a given image) and directly take this as input to our feedforward network. This corresponds to another possible way a child might learn where, the very first time they see an object, we point to it and augment their visual input with facts about the same. This is again repeated for all facts for each image.

4. Discussion on Differentiation in Models

The class splits we are interested in are:- 'Flowers vs Fruits', 'Air vs Land vs Water animals', 'Motorized vs Non-motorized vehicles' higher classes like 'Non-Living vs Living' 'Plants vs Animals'.

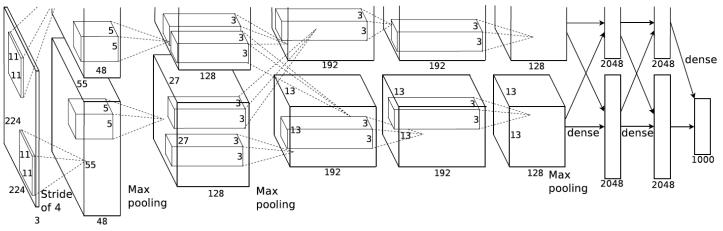

Figure 1. AlexNet

4.1. Model 1

Model 1 (Figure 3) is a slow learning model. It seems to have understood the difference between 'Non-Living vs Living' very soon (at about 1000 epochs), while it takes longer to come to the next distinction of 'Plants vs Animals' (around 3000 epochs). Finally, after 3600 epochs, our model has further levels of differentiation in our data. It clearly differentiates 'Flowers vs Fruits', 'Air vs Land vs Water animals', and 'Motorized vs Non-motorized vehicles' higher classes like 'Non-Living vs Living' 'Plants vs Animals'.

4.2. Model 2

Model 2 (Figure 4) is much faster and better than the previous model in terms of differentiation. We see that at epoch

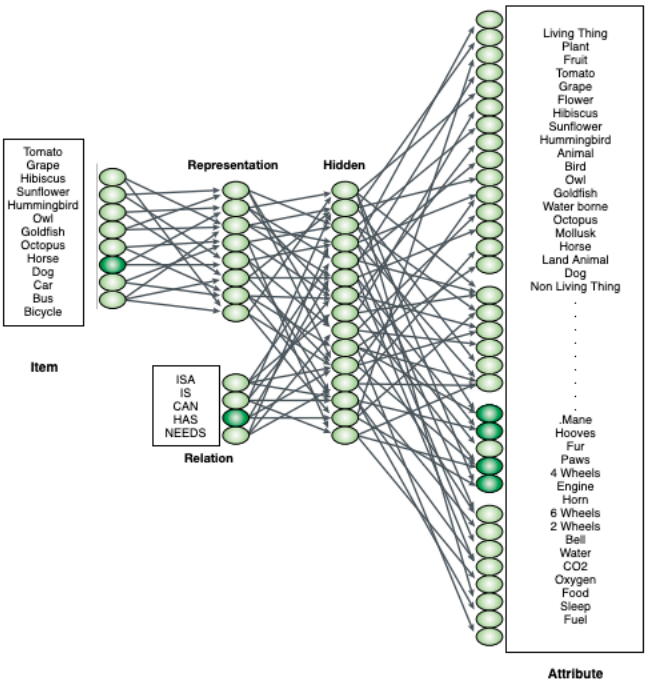

Figure 2. Connectionist Feedforward Network

20 itself we have the differentiation of 'Non-Living vs Living'. The differentiation of 'Plants vs Animals' starts to be distinguishable around epoch 50 and by epoch 75 we have complete differentiation of classes. The interesting thing to note here is the problematic separation of Horse and Dog (Land Animals). We studied the confusion matrix (Fig 6) of our AlexNet and realized that Horse (second to last row) is wrongly classified as Dog (last row) and vice versa. We also observed that Horse is often wrongly classified as the non-living objects Car and Bike we well. This could possibly be why the Land animals pair up and shift towards the non-living category.

4.3. Model 3

Model 3 differentiation, as shown in Figure 5 is by far the worst model. It was very slow to train and even after 100 epochs we had no significant differentiation. If we had to give the model a human analogy, it would be a very confused and dumb model that has failed to understand much about the classes in the data even after receiving rich inputs and being trained for much longer than Model 2.

5. Conclusion

We explored and compared three different semantic cognition models, two of which are novel models in semantic cognition for images. From the observations noted above, we can see that Model 2 is the best model in terms of differentiation, which makes it similar to how humans possibly interpret and differentiate data. The success of Model 2 leads us to the conclusion that humans (children) learn better when they are taught about facts (like "Dogs can bark") concerning a visual after they are familiar with the visual in advance (have seen multiple instances of a dog beforehand), while the failure of Model 3 makes us believe that loading a child with too much information (mixing new visual scenes

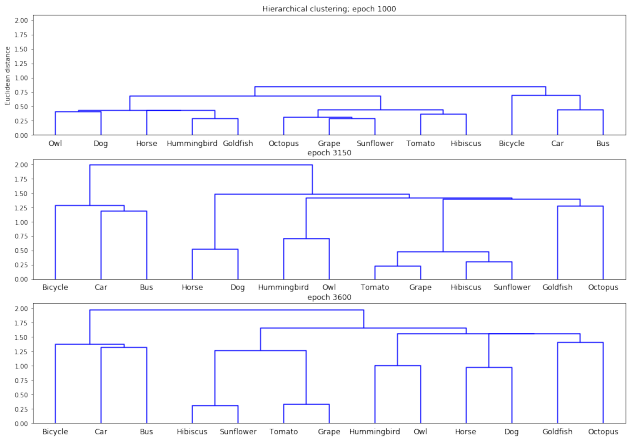

Figure 3. Differentiation in Model 1

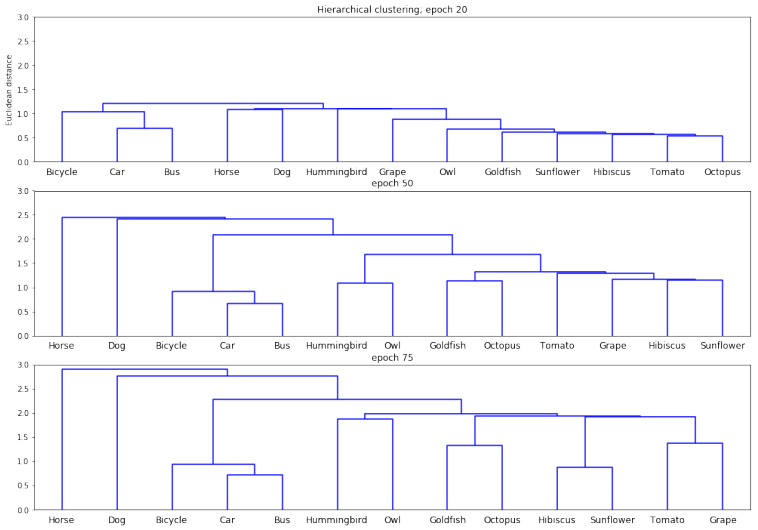

Figure 4. Differentiation in Model 2

with properties of the object) leads to slow and confused learning.

6. Future Work

The intersection of image recognition models and semantic cognition from images seems like a very interesting field that is relatively unexplored. We have various ideas to build on top of our work to study more properties in semantic cognition. Some of these are:

- Investigate the anomaly of Land Animals in an otherwise well formed Model 2. Possibly use different image recognition architectures to confirm or deny (and then look for other possibilities) whether the anomaly is due to the accuracy of the image recognition system.
- Study the effects of degradation in the above models as done in [5].
- This model can be combined with an LSTM to perform similar evaluations to a Visual Question Answering model.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[2] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. 2007.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

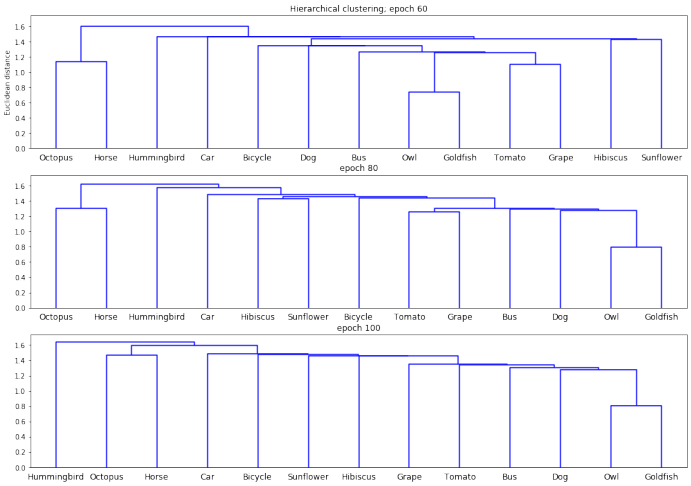

Figure 5. Differentiation in Model 3

```
tensor([[ 91.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   1.,   0.,   0.,  0.],
        [  0., 177.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   1.,   1.,   1.,  0.],
        [  0.,   0., 103.,   0.,   1.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,  0.],
        [  0.,   0.,   0.,  87.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   1.,  0.],
        [  0.,   0.,   0.,   0.,  97.,   0.,   0.,   0.,   0.,   0.,   1.,   1.,  0.],
        [  1.,   1.,   0.,   0.,   0.,  97.,   0.,   0.,   0.,   0.,   0.,   0.,  0.],
        [  0.,   0.,   0.,   0.,   0.,   0.,  72.,   0.,   0.,   0.,   0.,   0.,  0.],
        [  0.,   1.,   0.,   0.,   0.,   0.,   0., 101.,   0.,   1.,   0.,   1.,  0.],
        [  0.,   1.,   0.,   0.,   1.,   0.,   0.,   2., 100.,   2.,   0.,   0.,  2.],
        [  0.,   2.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,  81.,   0.,   0.,  0.],
        [  0.,   1.,   0.,   0.,   0.,   0.,   0.,   0.,   1.,   0.,  97.,   0.,  0.],
        [  0.,   2.,   1.,   0.,   2.,   0.,   0.,   0.,   2.,   0.,   0., 233.,  3.],
        [  0.,   1.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   3., 87.]])
```

Figure 6. Confusion Matrix for AlexNet (Horse is second to last row; Dog is last row

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[5] J. L. McClelland and T. T. Rogers. The parallel distributed processing approach to semantic cognition. *Nature reviews neuroscience*, 4(4):310, 2003.

[6] D. Rumelhart, P. Todd, D. Meyer, and S. Kornblum. Attention and performance xiv: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience. 1993.

[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[8] S. F. Zornetzer, C. Lau, J. L. Davis, and T. McKenna. *An introduction to neural and electronic networks*. Academic Press, Inc., 1994.