Department of Computer Science and Information Systems
Birkbeck College, University of London

The Role of Depth in Neural Network's Multimodal Word-Learning Assumption

Akira Charoensit

A project proposal submitted for the degree of
Master of Science in Data Science

Spring Term 2019

Abstract

This work studies the correlation between depth information, whole-object assumption, and language learning ability in artificial neural network (ANN). I attempt to answer three research questions: (1) does depth information improve word-learning rate in ANN; (2) does depth information induce whole-object bias; and (3) is there a correlation between whole-object assumption and word-learning rate. To answer questions (1) and (2), I plan to implement an agent with two convolutional neural networks (CNN), one to process RGB information and the other to process depth information. The output of the two CNNs are then combined with a recurrent neural network (RNN) which processes linguistic information. Then I intend to train the model using the opensource DeepMind Lab environment and evaluate the agent on word-learning task. To answer (3) I apply the genetic algorithm (GA) to the initial agent and examine whether the later generations of models show higher whole-object bias as the learning rate increases.

Academic Declaration

I have read and understood the sections of plagiarism in the College Policy on assessment offences and confirm that the work is my own, with the work of others clearly acknowledged. I give my permission to submit my report to the plagiarism testing database that the College is using and test it using plagiarism detection software, search engines or meta-searching software.

Akira Charoensit

Table of Contents

Chapter 1 – Introduction

This project examines the analogy between word learning in humans and artificial neural network. More specifically, I examine the relationship between depth information and whole-object assumption in neural network language learner by implementing an agent which encompasses a convolutional neural network (CNN) module [1, 2, 3] for processing depth, and evaluating its performance and language learning behaviours on a language learning task.

On being exposed to a new word, children show a tendency to assume that the word refers to a 'whole object' and not its properties or parts. This tendency is known as 'object bias' and is a well-established observation in linguistics [4, 5, 6, 7, 8, 9]. The bias is universal and is argued to be an *innate* property of human cognition as it helps children to learn words quickly [4]: [9, p. 100] argues that a whole object is a more salient referent for a linguistic form because children 'parse the world into discrete physical objects'. By contrast, a recent study shows that artificial neural network agent [10] prioritises colour over shape of objects. This is the direct opposite of human subjects who favours shape over colours, as found in an experiment by [11] amongst others. Specifically, in a setting where the agent is exposed to an equal number of colour and shape words, it is predisposed to assume that a new, unseen word refers to the colour of the object, rather than the form.

My initial hypothesis is that the convolutional neural network used in this model only processes the RGB colours and not depth, hence the model's colour bias. Therefore, in the scope of this project, I will implement a multimodal language learner which takes as its input both linguistic instruction and RGB-D visual information. Then I apply the genetic algorithm to the implemented agent and conduct a series of experiments to answer the three following research questions:

1) Can depth information improve the rate of language learning?
2) Does the agent which can take depth information into consideration show whole-object assumption like human children in their first language acquisition?
3) Is there a relationship between whole-object assumption and the language learning rate?

In the field of cognitive linguistics, the use of language facilities has been examined to shed new lights on the way in which humans perceive and process the world. By the same tokens, the aim of this work is to achieve a better understanding in the field of robotic and computer visions as well as to improve language learning ability in artificial intelligence.

In Chapter 2, I briefly explore relevant topics in cognitive linguistics and examine other related works on multimodal language learning. In Chapter 3, I discuss architecture of the agent and the data used to train the model. Chapter 4 discusses the design of experiments and methods of evaluation. In Chapter 5 I present a GANNT chart on the planned timeline of this study.

Chapter 2 – Background & Literature Review

The relationship between cognitive linguistics and deep neural networks has been explored by works such as [12]. This chapter is divided into three sections. In the first, I briefly outline relevant points in cognitive linguistics and how children acquire first language. In the second, I motivate the choice of multimodal learning and review the earlier works that have been done in field. In the third section, I briefly outline the genetic algorithm.

2.1 Cognitive Science

The notion of object bias has been introduced in the last chapter. This section explores the discussion on the subject with a particular focus on empirical findings.

How do children learn words? According to [13], there are three stages of human language acquisition:

1. Acquisition of linguistics signals
2. Acquisition of non-linguistic sensory (e.g. visual)
3. Mapping between word forms and non-linguistic sensory

Before the first word is achieved, therefore, many prelinguistic skills must be obtained. One of the most important tasks is to distinguish objects from its background. In an experiment by [14], children are asked to count a set of familiar objects, such as a set of forks with one broken as in figure 1:

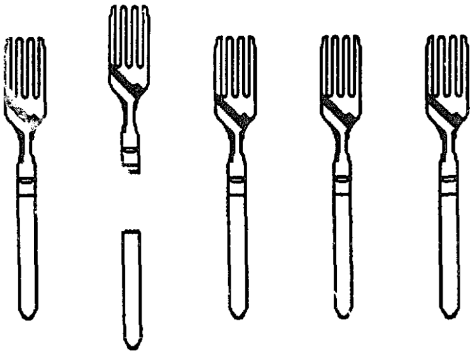

Figure 1: Picture of forks shown to children in (Shipley & Shepperson, 1990) [14]

They found that pre-school children tend to answer six rather than five or four. That is, they already parse the world into a set of discrete objects.

[6, p. 4] writes:

> "It is obvious that an infant has the capacity to distinguish from the rest of the physical environment an object which his mother draw to his attention and names. It seems clear too that in such circumstances he adopts the strategy of taking the he hear as a name for the object as a whole rather than as a subject of its properties, or for its position, or weight, or worth, or anything else."

When mapping linguistic form to the physical world, however, there are infinite candidate referents. The form can refer to property (colour, size, weight, value, etc.) or parts of the object. To be able to learn quickly, children have the arguably innate ‘object bias’ whereby they assume that the new word refers to the whole objects, unless there is a clear reason not to (e.g. mutual exclusivity) [15]. There are many empirical evidences to support this claim. For instance, in [11], 40 children (2 to 3 years old) are shown differently coloured objects (‘see this *zom*?’)and asked to find a matching object (‘can you show me another *zom*?’). The subjects almost exclusively interpret ‘zom’ as the object name (rather than the colour name) and choose an object of matching form, not colour. Children are also more predisposed to think that a word refers to the whole objects rather than part of object.

An important point to consider is what counts as an object. Spelke proposes four main criteria:

i) Cohesion: an object must be connected and discrete (they must move as one)
ii) Continuity: an object must exist continually in space and time
iii) Solidity: an object must be solid (i.e. they cannot pass through a grate, or each other)
iv) Contact: an inanimate object only moves when touched

Of these four principles, cohesion is the most important criterion [16, 17].

2.2 Multimodal Learning

As discussed in the previous section, the external sensory information such as visual is clearly central to grounding linguistic forms to the physical world. This section explores three models which have explored the field of multimodal learning and discusses their advantages and disadvantages.

As early as in 1972, there has been a study in grounding natural language in physical world [18]. The SHRDLU system introduced in [18] can answer questions or execute certain actions based on linguistic instruction and visual cue.

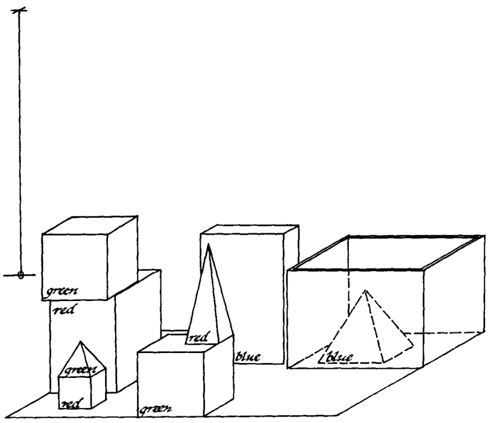

Figure 2: A simulated 3D-scene from the SHRDLU system (Winograd, 1972)

For instance, given a simulated 3D-scene in Figure 2, it can move objects and answer questions about where objects are (e.g. the red pyramid in on top of the green box). While impressive, the model can only work on a predefined grammar and finite set of words. This type of models clearly is not suitable for our purpose.

Since [18], multimodal learner agent has been studied in in many other works. In 2005, [19] proposes a more robust model with two modules: the visual module and the linguistic module, each producing motor and linguistic outputs, respectively. The input space of the network contains 480 objects which belong to 4 categories of 120 objects each, and 4 words in the dictionary corresponds to different categories of objects. Each of the four words has 120 variations, corresponding to the variation in the way human speaks. The goal is to map the words to the objects.

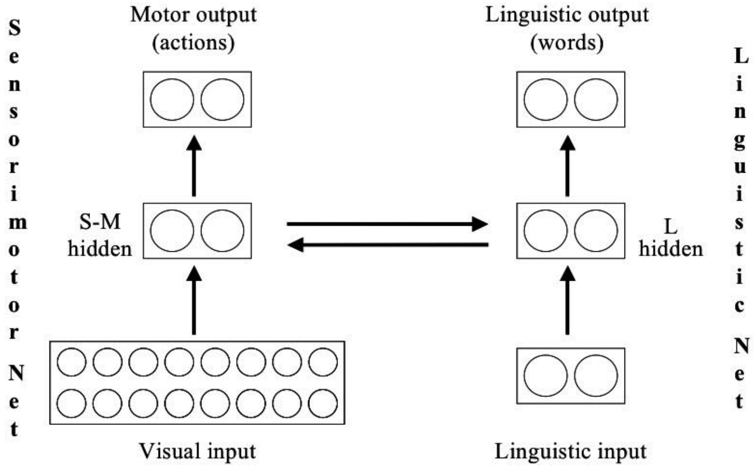

Figure 3: the design of multimodal language learning from [19]

To integrate visual and linguistic inputs, the hidden layers of the two networks are connected by two-way weight matrices so that the outputs of each also influence the other. The training process has two phases. In the first phase, the two connecting matrices are not activated, and the two model are first trained separately. The sensory motor net is trained to classify visual inputs and linguistic net to imitate words. This is to simulate a child's language learning. The first stage has 1,000 training episodes. In the second stage, the connecting weight matrices are activated. A word and an object which belongs to the corresponding class is given simultaneously. This stage also has 1,000 training episodes. It is reported that at the end of the second stage the agent achieves 100% accuracy in both naming (producing the right word for an object) and comprehension (outputting the right category for the right word). The main disadvantage of this model is its lack of scalability.

In 2017, [10] proposes a more robust model. The agent takes as its inputs a natural language instruction and a continuous pixel generated from a 3D-simulated world using Deepmind Lab [20], and produces as output a series of motor movement actions required to follow the instruction. As a result, in addition to learning relative position from static images, the agent must learn them from a continuous stream of pixels. The model has four main modules:

1. the language module $\mathbb{L}$
2. the visual module $\mathbb{V}$
3. the mixing module $\mathbb{M}$
4. the core module $\mathbb{C}$

The language module $\mathbb{L}$ is a recurrent neural network (RNN) [21, 22] which processes the instruction as a string. The visual module $\mathbb{V}$ is a three-layer convolutional neural network, which processes the visual input, and recurrent neural network (Long Short-Term Memory -- LSTM), which processes the linguistic inputs. The outputs of the two networks is fed to the mixing module in which they are combined and fed forward to the core of the model, which is another LSTM and outputs one of the actions in the action space (consisting of eight-directional movements, jumping, and turning the viewpoint). As each movement action is taken, the model's visual field is moved and the new pixel is fed to $\mathbb{V}$ [23]. The system employs reinforcement learning strategy; in each training episode, the agent receives a specific task and gets a score when the goal of the task is accomplished.

In [10], it is reported that this design cannot successfully learn words even after millions of training episodes, so four additional networks are implemented. The first is the temporal autoencoding which combines actions from the action space with the current embedding representation of visual input at timestep *t*, v_t, from $\mathbb{V}$ and returns a predicted vision. The second is a language prediction module which guesses the instruction associate with the current vision. By adding these two extra modules, along with reward prediction and value replay modules, the agent begins to learn their first words after around 250,000 training episodes (after having acquired pre-linguistic knowledge e.g. figure-ground disambiguation).

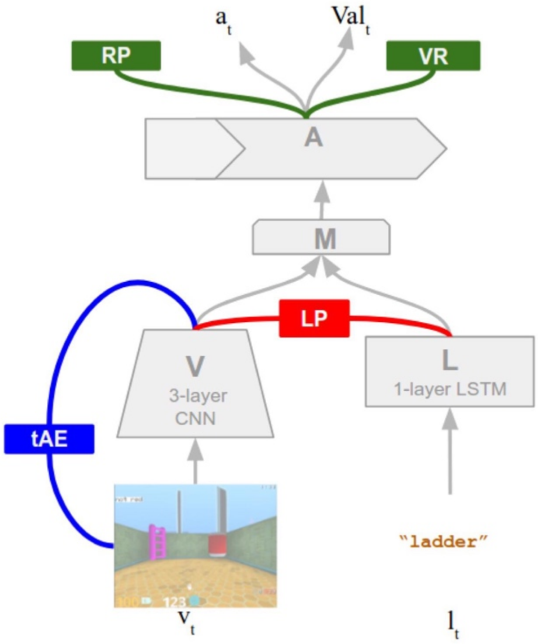

Figure 4: The design of multimodal language learner in [23]

This approach is ideal for our purpose because it has been shown to successfully processs 3D inputs. Moreover, DeepMind Lab [20] provides a useful

In addition to these, I have also looked for a model which focuses on depth information. [24] proposes an object recognition model which focuses on depth recognition as well as RGB recognition.

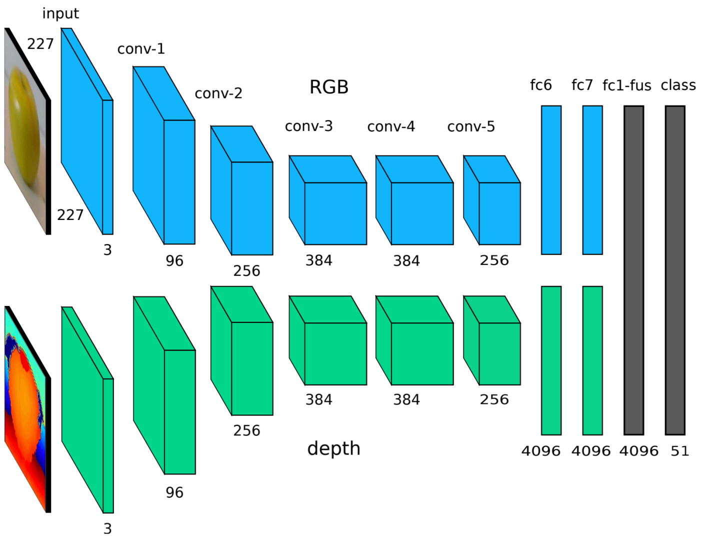

Figure 5: The design of two convultional networks in [24]

Their model has two deep CNNs. The first CNN is trained on an RGB labelled (ImageNet) input, while the other is trained on a depth labelled dataset. However, as large depth-labelled dataset is not available, they encode depth information as RGB (i.e. the closeness of the object is represented by a colour on RGB map) instead. The outputs from the two CNNs. The network is evaluated on recognising objects in the Washington RGB-D Object Dataset [25] with added noise and achieve 91% accuracy.

2.3 Genetic Algorithm/Genetic Programming

Genetic algorithm (GA) is a type of optimisation algorithms pioneered by John Holland in 1972 [26]. The algorithm is inspired by the process of evolution and knowledge transfer between individuals. A GA has two main phases: encoding and evolution.

2.3.1 Encoding

In the encoding phase, one must first design the artificial neural network (ANN). And then one can select any parameters such as variables the number of hidden units, initial learning rate, and learning algorithm from the ANN. These parameters are encoded into a string of bits or ‘genomes, which will be manipulated in the evolution phase. For instance, if one only chooses two parameters: the number of hidden units and learning rate, and the length of each parameter is 10 bits, then each model is represented using a 20-bit binary string.

2.3.2 Evolution

The evolution phase starts with generating a random population of binary strings or genomes. Each of which represent a certain configuration of the ANN as discussed in 2.3.1. Then, the encoded ANNs are trained and evaluated on a certain set of tasks. The genomes with the best performance on the tasks are selected for reproduction, while others perish.

Consider an 8-bit genome with 4 genes, each representing a parameter. Let's consider two genomes selected from the group of population with the highest score: e.g. the father has the genome 10 01 00 01 and the mother has the genome 11 10 11 00. The reproduction process goes as follows:

1. The genome is split into halves or 'chromosomes'. The first member in each pair of the gene form one half. The father's genome is split into 1000 0101 and the mother's into 1110 1010.
2. Sperms and eggs are generated from the resulting chromosome by applying single crossover point. At this point we have six sperms (1101, 1001, 1001, 0000, 0100, 0100) and six eggs (1010, 1110, 1110, 1110, 1010, 1010).
3. Then the sperms and eggs are combined using positional combination. For instance, the first sperm and the first egg combined to get 11 10 01 10. The offspring is always of the same length as the parents.

Apart from mating, the genetic algorithm also allows random mutation, where at random point one of the bit is randomly flipped to ensure that the evolution does not get stuck.

Then, the process is repeated. The offspring is trained and evaluated again and the best are selected for reproduction (to ensure that the size of population is always the same). As the algorithm converges, the new generations will perform better on the evaluated tasks as their parameters are an amalgamation of the last best configuration. That is, the knowledge of the previous generations is transferred to the current one and becomes an 'innate' part of the architecture. The main advantage of this algorithm is that it can explores vast probability space in parallel.

Chapter 3 – Analysis, Requirements, and Design

This chapter is divided into two parts. Section 3.1 describes the architecture of the agent. Section 3.2 discusses the data used to train the model.

3.1 Multimodal Learning Agent and Implementation Methods

The initial model here is largely inspired by the model described in Figure 4 [23]. However, to be able to process depth information, I am planning to start by adding an extra CNN module as shown in Figure 6:

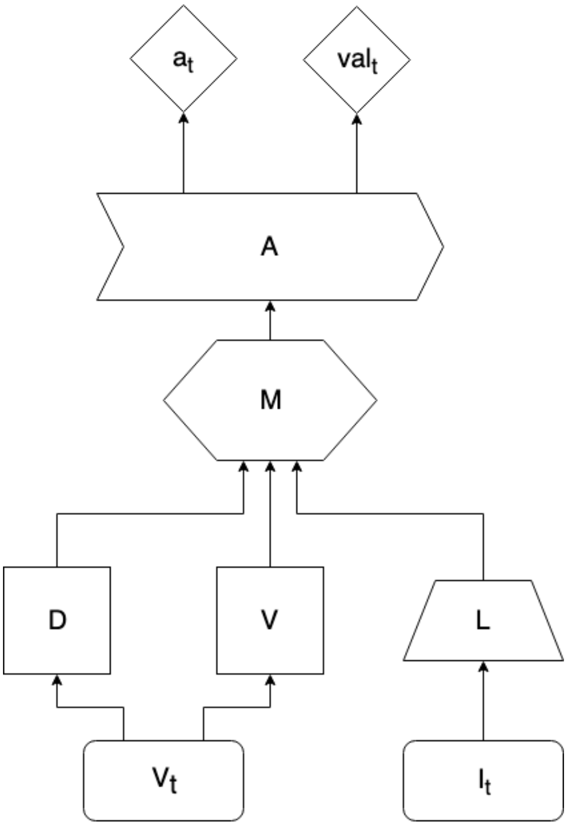

Figure 6: The Architecture of the Agent

Like in the earlier model, the system takes as its inputs the linguistic instruction (I_t) and the visual information V_t. However, there are five main modules instead of four. The main modules are:

1. The visual module V, comparable to the visual module 𝕍 in [10, 23] (described in Section 2.2). This module processes RGB information.
2. The visual module D which processes depth information. This is module will be a three-layered CNN model inspired by the work of [24]
3. The language module L, comparable to the language module in [10, 23]
4. The mixing module M, comparable to module 𝕄 in [10, 23] which combines the result from modules M, D, and L

5. The action module A, comparable to module $\mathbb{A}$ in [10, 23]

The visual input V_t is an RGB-D image with depth plane of the vision field at time step *t*, instead of an RGB image. V_t is simultaneously fed to two modules: V and D, which is the main implementation of this project. Linguistic instructions I_t is processed through module L, which remains the same as the original $\mathbb{L}$ module described in the previous section. Then the outputs of modules D, V, and L are processed through module M and A which also remain the same as the original model, to give two outputs: the action A_t from the action space and the state [21]value, Val_t.

I plan to implement the model in Python, using TensorFlow [27]. I will also use the DEAP (Distributed Evolutionary Algorithm in Python) [28] or similar library to perform the Genetic Algorithm needed in Phase 3 of the experiment described in Chapter 4.

3.2 Data

This section describes the two datasets I intend to use to train the network: Deepmind Lab [20] and RGB-D Dataset [25].

3.2.1 RGB-D Dataset

The RGB-D dataset [25] contains video sequence of 300 physical objects. The objects are arranged into a hierarchy according to the WordNet word relation [7] as shown in Figure 6:

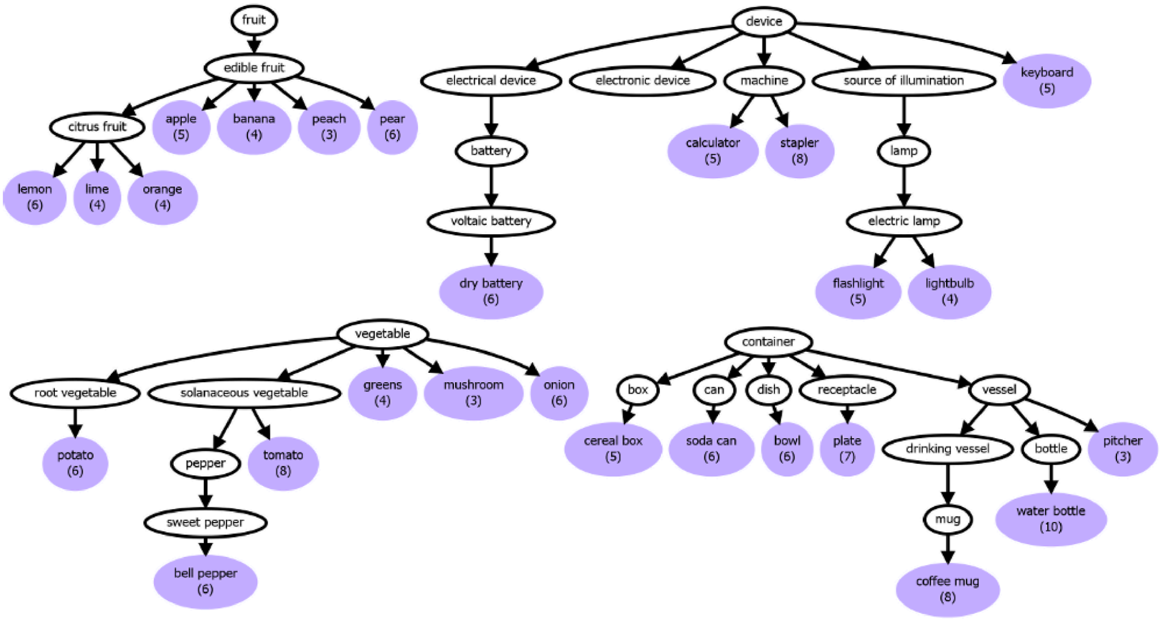

Figure 7: Hierarchy of objects in the RGB-D Dataset [25]

There are four main categories: fruit, vegetable, device, and container. The objects are instances of one of the 51 terminal nodes e.g. potatoes, apples, calculators, cereal boxes. Each of the physical item is placed on a turning table and is recorded using RGB-D camera from 30 different height angle (30°, 45°, and 60°). The camera uses an invisible infrared light projecting to actively record depth. The video recording has approximately 250 frames, resulting in 250,000 images in total. Each frame of the resulting video contains 640×480 pixels and four planes: red, green, blue and depth.

3.2.2 DeepMind Lab

DeepMind lab [20] will be used as both training and test environment. The environment is a simulated 3D-world. The agent is placed inside the simulated world and explore the world from first-person point of view using an action from the action space (e.g. move in 3D direction, look up, jump, etc.). After each action is taken, three observations are available: the current viewpoint of the agent, the scalar reward signal, and the velocity signal. Deepmind Lab is ideal for our experiment because of its breadth and reliable quality. Most importantly, it also provides depth information of the viewpoint at each time step. This study makes use of the first two observations: the current viewpoint of the agent in RGB-D format and the scalar reward signal.

Chapter 4 – Experiment Design

The experiment is divided in three phases to answer each of three questions stated in Chapter 1.

Phase 1 examines the role of depth plane in language learning rate in general, compared to a model which only use RGB information. The agent is trained using the reinforcement learning method. The agent learn a word by being placed in a 3D-room generated using the DeepMind lab engine. As seen in Figure 8, there are two objects in the room which the agent can choose by touching (walking into) the object.

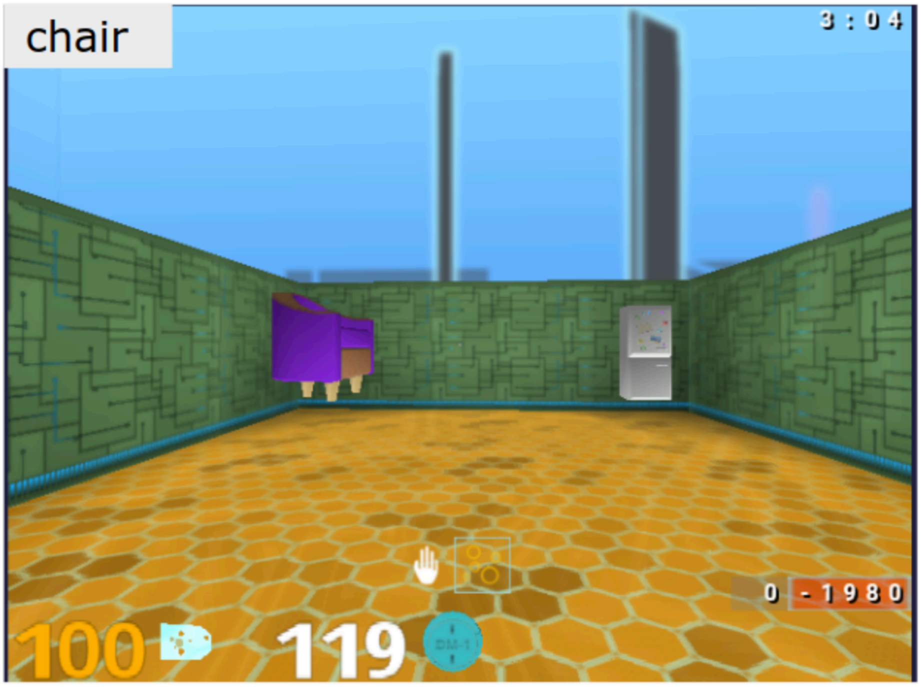

Figure 8: the training episode in DeepMind lab [10]

If the agent chooses the correct object, then it receives a positive reward. If it does not choose the correct object within a predefined number of steps or chooses the wrong object, then it receives a negative reward. There are two groups of words to be learnt: shape words (e.g. ‘apple’) and colour words (e.g. ‘green’). The two groups contain the same number of words. The learning rate of the depth-encompassing model described in Section 3.1 will be compared to that of the original model which only take RGB visual inputs [23]. If my hypothesis is correct, then the depth-encompassing agent should be able to learn words faster than RGB agent.

Phase 2 is a replication of the experiment described in [5] and [10]. After the agent has successfully learnt the words in the lexicon in the training episodes, in the test episode, the agent is exposed to a new word and must choose between two objects, (a) an object with an unseen shape and a familiar colour, and (b) an object with a familiar shape and an unseen colour. If the agent chooses object (a) then, like humans, it shows whole-object assumption, i.e. it assumes that a new word refers to a novel shape. Thus, my hypothesis that incorporating depth will induce a whole-object bias is confirmed. This also carries the implication that the depth plane allows the model to parse its environment in a manner which resembles more closely that of human. Otherwise, it shows a colour bias, i.e. it assumes that a new word refers to a new colour.

Finally, in the third phase of the experiment, I will apply the genetic algorithm to the model and explore whether there is a correlation between whole-object assumption and word learning rate. A population of genomes will be generated. Each of the genome string represents parameters of all the modules of the agent described in Section 3.1. They will all be trained and evaluated on the same language learning task in Phase 1 of the experiment. Then the agents with the best performance are selected to reproduce using the algorithm described in section 2.3.2. It is predicted that the learning rate should improve after a number of generation via knowledge transfer. If as learning rate improves, the whole-object bias also increases, this would suggest that whole-object assumption could improve machine's language learning ability as it does children's.

Chapter 5 – Timescale

The timeline in Figure 9 below shows the planned timescale for each phase in the project. This is an AGILE chart where there are overlaps in tasks. The general research started in November 2018 and the main implementation phase will start in June.

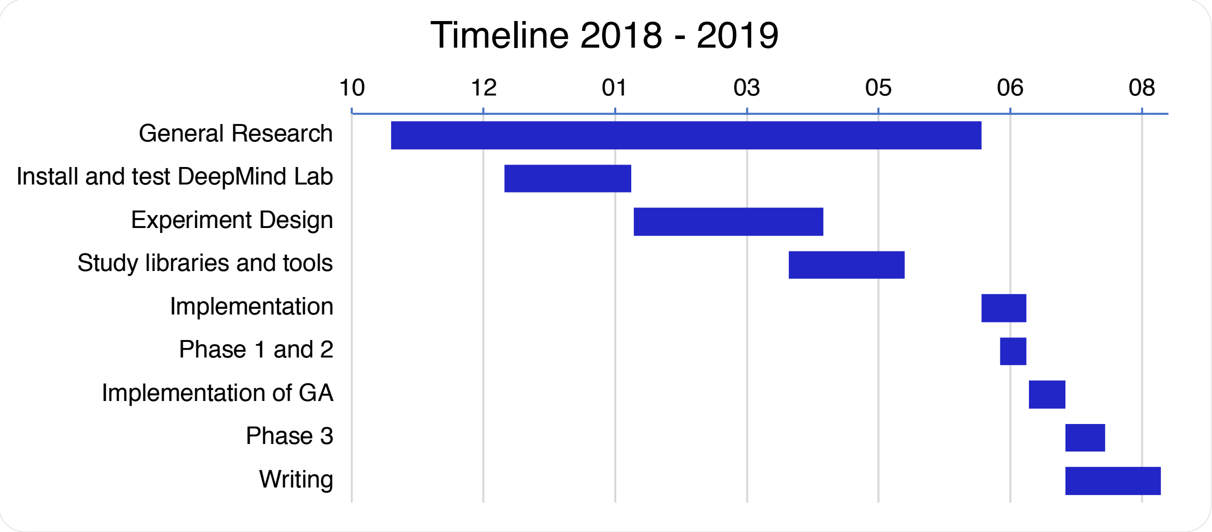

Figure 9: Work Plan

Bibliography

[1] Y. LeCun, K. Kavukcuoglu and C. Farabet, “Convolutional Networks and Applications in Vision,” *Circuits and Sysmtems (ISCAS), Proceedings of 2010 IEEE International Symposium on,* pp. 253-256, 2010.

[2] M. D. Zeiler and R. Fergus, “Visualizing and Understanding Convolutional Networks,” *European Conference on Computer Vision,* pp. 818-833, 2014.

[3] K. Simonyan, A. Vedaldi and A. Zisserman, “Deep inside Convolutional Networks: VIsualising image classification models and saliency maps,” *Workshop at International Conference on Learning Representations,* 2014.

[4] B. Landau, L. B. Smith and S. S. Jones, “The Importance of Shape in Early Lexical Learning,” *Cognitive Development,* pp. 233-321, 1988.

[5] B. Landau, L. Smith and S. Jones, “Object Perception and Object Naming in Early Development,” *Trends in Cognitive Sciences,* pp. 19-24, 1998.

[6] J. Macnamara, “Cognitive basis of language learning in infants,” *Psychological Review,* pp. 1-13, 1972.

[7] G. A. Miller, “WordNet: a lexical database for English,” *Communications of the ACM,* pp. 39-41, 1995.

[8] D. Roy and E. Reiter, “Connecting language to the world,” *Aritificial Intelligence,* vol. 167, pp. 1-12, 2005.

[9] P. Bloom, How Children Learn Meanings of Words, London: The MIT Press, 2000.

[10] F. Hill, K. M. Hermann, P. Blunsom and S. Clark, “Understanding Grounded Language Learning Agents,” *CoRR,* 2017.

[11] D. A. Baldwin, “Priorities in Children's Expectations about Object Label Reference: Form over Color,” *Child Development,* pp. 1291 - 1306, 1989.

[12] S. Ritter, D. G. Barrett, A. Santoro and M. M. Botvinick, “Cognitive Psychology for Deep Neural Networks: A shape bias case study,” *Proceedings of ICML,* 2017.

[13] C. Kit, “How does lexical acqusition begin? A cognitive perspective,” *Cognitive Science,* pp. 1-50, 2002.

[14] E. F. Shipley and B. Shepperson, “Countable entities: Developmental changes,” *Cognition,* pp. 109-136, 1990.

[15] E. M. Markman and G. F. Wachtel, “Children's use of mutual exclusivity to constrain the meanings of words,” *Cognitive Psychology,* pp. 121-157, 1988.

[16] E. S. Spelke, “Initial Knowledge: Six suggestions,” *Cognitive,* pp. 443-447, 1994.

[17] E. S. Spelke, K. Breinlinger, K. Jacobson and A. Phillips, “Gestalt relations and object perception: A developmental study,” *Perception,* pp. 1483-1501, 1993.

[18] T. Winograd, “Understanding Natural Language,” *Cognitive Psychology,* pp. 1-191, 1972.

[19] M. Mirolli and D. Parisi, “Languge as an Aid to Categorisation: A neural network model of early language acquisition,” *Progress in Neural Processing. Modeling language, cognition and action: Proceedings of the Ninth Neural Computation and Psychology Workshop,* pp. 97-106, 2005.

[20] C. Beattie, J. Z. Leibo, D. Teplyashin, T. Ward, M. Wainwright, H. Küttler, A. Lefrancq, S. Green, V. Valdés, A. Sadik, J. Schrittwieser, K. Anderson, S. York, M. Cant, A. Cain and A. Bolton, “Deepmind Lab,” *CoRR,* 2016.

[21] T. Mikolov, M. Karafiát, L. Burget, J. Černocký and S. Khudanpur, “Recurrent Neural Network Based Language Model,” *INTERSPEECH-2010,* pp. 1045-1048, 2010.

[22] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Compuation,* vol. 9, no. 8, pp. 1735-1780, 1997.

[23] K. M. Hermann, F. Hill, S. Green, F. Wang, R. Faulkner, H. Soyer, D. Szepesvari, W. M. Czarnecki, M. Jaderberg, D. Teplyashin, M. Wainwright, C. Apps, D. Hassabis and P. Blunsom, “Grounded Language Learning in a Simulated 3D World,” *CoRR,* 2017.

[24] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller and W. Burgard, “Multimodal Deep Learning for Robust RGB-D Object Recognition,” *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),* 2015.

[25] K. Lai, L. Bo, X. Ren and D. Fox, “A large-scale hierarchical multi-view rgb-d object dataset,” *2011 IEEE international conference on robotics and automation,* pp. 1817-1824, 2011.

[26] J. Holland, “Genetic Algorithms,” *Scientific American,* pp. 66-73, 1992.

[27] Google Brain, “TensorFlow: A system for large-scale machine learning,” *12th USENIX Symposium on Operating Systems Design and Implementation ,* 2016.

[28] F.-A. Fortin, F.-M. De Rainville, M.-A. Gardner, M. Parizeau and C. Gagné, “DEAP: Evolutionary Algorithms Made Easy,” *Journal of Machine Learning Research,* pp. 2171-2175, 2012.

[29] T. Ramalho, T. Kočisky, F. Besse, S. M. A. Eskami, G. Melis, F. Viola, P. Blunsom and K. M. Hermann, “Encoding Spatial Relations from Natural Language,” *CoRR,* 2018.

[30] R. Rojas, Neural Networks, Berlin: Springer-Verlag, 1996.

[31] B. L. Whorf, Language, Thought, and Reality, London: THe MIT Press, 2012.

[32] E. M. Markman, “Constraints on Word Meaning in Early Langauge Acquisition,” *Lingua,* vol. 92, pp. 199-227, 1994.

[33] C. Silberer and M. Lapata, “Learning Grounded Meaning Representations with Autoencoders,” *Proceedings of 52nd Annual Meeting of the ACL,* pp. 721-732, 2014.

[34] N. Krishnaswamy, S. Friedman and J. Putejovsky, “Combining Deep Learning and Qualitative Spatial Reasoning to Learn Complex Structures from Sparse Examples with Noise,” *CoRR,* 2018.

[35] B. Landau, “Where's what and what's where: the language of objects in space,” *Lingua,* vol. 92, pp. 250 - 96, 1994.

[36] N. Giralt and P. Bloom, “How Special are Objects? Children's Reasoning About Objects, Parts, and Holes,” *Psychological Science,* pp. 497-501, 2000.

[37] R. Casati and A. Varzi, “Holes and Other Superficialities,” *Philosophy and Phenomenological Research,* pp. 734-736, 1997.

[38] J. Deng, W. Dong, R. Socher, L. Li and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” *Computer Vision and Pattern Recognition 2009,* pp. 248 - 255, 2009.