A Neural Network Model for Taxonomic Responding with Realistic Visual Inputs

A Neural Network Model for Taxonomic Responding with Realistic Visual Inputs

Giorgia Fenoglio (giorgia.fenoglio@edu.unito.it)
Dipartimento di Informatica, Corso Svizzera 185,
10149, Torino, Italy

Roberto Esposito (roberto.esposito@unito.it)
Dipartimento di Informatica, Corso Svizzera 185,
10149, Torino, Italy

Valentina Gliozzi (valentina.gliozzi@unito.it)
Center for Logic, Language, Cognition
Dipartimento di Informatica, Corso Svizzera 185,
10149, Torino, Italy

Abstract

We propose a neural network model that accounts for the emergence of the taxonomic constraint in early word learning. Our proposal is based on Mayor and Plunkett (2010)'s neurocomputational model of the taxonomic constraint and overcomes one of its limitations, namely the fact that it considers artificially built, simplified stimuli. In fact, while in the original model the visual stimuli are random, sparse dot patterns, in our proposed solution they are photographic images from the ImageNet database. In our model the represented objects in the image can be of different size, color, location in the picture, point of view, etc.. We show that, notwithstanding the augmented complexity in the input, the proposed model compares favorably with respect to Mayor and Plunkett (2010)'s model.

Introduction

A central issue in the current understanding of early lexical acquisition concerns how infants learn the reference of words. Quine (1960) famously raised the point that for every word heard in a given circumstance, there are several possible references: in order to infer the appropriate one, infants have to rule out several possible alternatives. An influential solution to the issue has been proposed by Markman (1989), suggesting that infants rule out inappropriate references by means of three constraints. By the *whole object constraint* children assume that novel words refer to objects as a whole, rather than to their parts, substance, color, or other properties. By the *mutual exclusivity constraint* children assume that two labels usually do not refer to the same object. Last, but central to this paper, by the *taxonomic constraint* children extend words to taxonomically-related objects (at the level of basic categories): when a child hears the word "dog" pronounced by a caregiver while pointing at a specific dog, she generalizes the reference of "dog" to all dogs, not just to the one in front of her.

Here we propose a neural network model that accounts for the emergence of the taxonomic constraint in early word learning, and can process realistic *visual* stimuli[1]. This is the first step towards the development of a model able to cope with visual and auditory stimuli that are both realistic.

[1]For the time being we leave the question of realism of the acoustic part to future work.

Our starting point is Mayor and Plunkett (2010)'s neurocomputational model of the taxonomic constraint. The model consists of two self-organizing maps (a visual and an acoustic map) connected with Hebbian connections. The model successfully explains how it is possible to generalize a single word-object association to a whole class of objects. Essentially, this is the result of Hebbian learning creating word-object associations over a previous conceptual organization of the visual and acoustic space.

Here we want to go beyond one limitation of Mayor and Plunkett (2010)'s model, namely the fact that it considers artificially built, simplified stimuli: in their model the visual stimuli are random, sparse dot patterns, in the style of (Posner, Goldsmith, & Welton Jr, 1967), whereas the acoustic stimuli are manipulations of acoustic signatures extracted from sounds produced by a speaker, leading to a simplified acoustic input stimulus.

Would the model still work if we considered realistic visual inputs, instead? In order to address this question, we have expanded the original model's visual component making it able to process realistic visual stimuli, that in our case are images taken from the ImageNet dataset. More precisely, we have added to the visual component of Mayor and Plunkett (2010) an InceptionV3 deep network (Szegedy et al., 2015) which is at the state of the art in the image classification task. The deep network processes the visual scene in the image, builds a representation for it, and feeds the representation to the visual self-organizing map.

In order for the whole model to work, these representations need to contain a description of the main object of the visual scene, independent from the context. Understanding the nature of the image representations built at the various levels of the network is indeed one of the main points of debate in deep neural networks (Zeiler & Fergus, 2014; Zhou, Khosla, Lapedriza, Oliva, & Torralba, 2014; Agrawal, Girshick, & Malik, 2014). In order to assess whether the InceptionV3 deep network feeds into the visual self-organizing map meaningful object representations, we performed several clustering experiments. These experiments investigated whether representations deriving from images of objects can be clustered

together by off-the-shelf clustering algorithms. They showed that the representations provided by InceptionV3 are reasonably well organized. A further test investigated whether the visual self-organizing map could organize the representations received from the deep neural networks in a topologically satisfactory manner.

We then tested the whole model to see if it still exhibits a taxonomic responding, generalizing learned word-object associations to the whole category. Results show that our model, despite starting from more realistic visual stimuli, does replicate Mayor and Plunkett (2010)'s success on taxonomic responding when few joint word-object associations are considered.

Mayor and Plunkett (2010)'s model

Mayor and Plunkett (2010) neurocomputational model of taxonomic constraint (Figure 1) is based on two Self-Organizing Maps (SOMs): a visual map and an acoustic map, representing the primary visual cortex and the primary auditory cortex respectively.

The stimuli presented to the two maps are artificially built: the visual stimuli are random dot patterns, whereas the auditory stimuli are extracted from the acoustic signatures of uttered words; the acoustic signatures are manipulated in order to create simpler inputs.

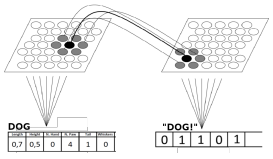

Figure 1: Mayor and Plunkett (2010) model

Learning is a two-phase process. First, the two maps are independently trained to learn to categorize the visual and the acoustic stimuli. This first learning phase is preliminary to word learning, and unsupervised. The two maps are trained using the standard learning algorithm for self-organizing maps. In short, a stimulus x is presented to each neuron of the map, and the *Best-Matching Unit* (BMU) is selected: this is the unit i whose weight vector w_i is closest to the stimulus x (i.e. $i = \arg\min_j \|x - w_j\|$).

The weights of the best matching unit and of its surrounding units are updated in order to maximize the chances that in the future the same unit (or the surrounding units) will be selected as the best matching unit for the same stimulus or for similar stimuli. At iteration $n+1$, the weights for neuron j are updated as follows:

$$w_j(n+1) = w_j(n) + \eta(n) h_{i,j}(n)(x - w_j(n)) \qquad (1)$$

where η is the *learning rate*, and $h_{i,j}$ is the neighborhood function between i and j $h_{i,j}(n)$ is defined as $h_{i,j}(n) = exp(-d_{i,j}^2/2\sigma(n)^2)$, where $d_{i,j}$ is the distance between i and j on the map's grid, and $\sigma(n)$ is the width of the gaussian.

After a while, the two maps learn to adequately represent the stimuli of their training set in a topologically significant way: close units respond similarly to similar stimuli. The *neural activation* a_j of a neuron j in response to a stimulus x is defined as: $a_j = e^{-\frac{q_j}{\tau}}$, where q_j is the *quantization error* (i.e., the distance between the input vector x and j's weight vector: $q_j = \|x - w_j(n)\|$)), and τ is a normalization constant.

Once the visual and acoustic maps have stabilized into a topological organization, proper word learning can start. This is the Hebbian Learning phase, in which the two kinds of stimuli are simultaneously presented to the model. For each joint presentation of a visual and acoustic stimulus, the synapses between the two maps are strengthened. In particular, for each neuron v on the visual map and neuron p on the acoustic map, the Hebbian connection $u_{v,p}$ is strengthened proportionally to the resulting neural activations a_v and a_p, as follows:

$$u_{v,p}(n+1) = u_{v,p}(n) + 1 - e^{-\lambda a_v a_p} \qquad (2)$$

where λ is the Hebbian training learning rate.

A single Hebbian learning event, combined with the previously acquired categorization capabilities of the visual and acoustic SOMs, allows the model to generalize the association to other stimuli belonging to the same category.

Once training is complete, the model is tested for its ability of comprehension and production. Comprehension is assessed by considering what visual category is retrieved when a word is presented to the auditory map and activation is propagated via Hebbian connections. Production is assessed by considering what word is produced by the auditory map when a visual stimulus is presented to the visual map, and activation is propagated through Hebbian connections.

The ability of the model to extend the learned word-object associations to other words and objects belonging to the same category is measured by the *Taxonomic Factor* which is the percentage of correct word-object associations generated by the model. Results show that when the SOMs are adequately trained the Taxonomic Factor reaches 80% after a single Hebbian learning trial.

One of the limitations of Mayor and Plunkett (2010)'s original model is that it uses artificially built input stimuli that are much simpler than what would derive from realistic contexts. Here we address this limitation, for what concerns the visual module, by introducing deep convolutional neural networks as shown in the next sections.

Deep Convolutional Neural Networks

In the last few years research on deep networks contributed to reach human (sometimes super-human) performances on several difficult tasks (Hinton et al., 2012; Li & Wu, 2015; Socher, Bauer, Manning, & Ng, 2013; Yue-Hei Ng et al., 2015). In particular, in 2011 a deep convolutional model achieved for the first time super-human performances in a visual pattern recognition task and, in the following year, the AlexNet Convolutional Neural Network (CNN) model won the ImageNet competition by a significant margin (Krizhevsky, Sutskever, & Hinton, 2012) over traditional competitors. These successes contributed to a growing interest in deep networks and today deep-network-based models are at the forefront of research in many different areas and are

setting performance records in tasks of interest for the cognitive sciences community such as image (e.g., (Russakovsky et al., 2015)) and speech recognition (e.g., (Xiong et al., 2016)).

Despite unheard performances achieved in many different tasks, deep models present important shortcomings that are far from being completely addressed. The most important problem from the point of view of the forthcoming discussion is the difficulty they present for what concerns the understanding of their internal working. Consequently, recent research investigated ways to make sense of the contents of the network providing interesting insights. For instance, Zeiler and Fergus (2014) use "deconvolution networks" to visualize the patterns that causes the activation of nodes in each layer; in Zhou et al. (2014) scenes are iteratively simplified or occluded to investigate which image patches and which objects contribute to the activation of nodes in a given layer; in (Agrawal et al., 2014) the authors investigate the presence of grand-mother-cells and of distributed representations in deep networks.

While the understanding of the representations built by these networks is still scattered and incomplete, some of the insights seem to be well supported. An important one concerns the hierarchical organization of the features: low-level (coarser) features are nearest to the network input, while higher-level (more abstract) features are nearest to the output (for an idea of the kind of features extracted at the different levels see for instance (Zeiler & Fergus, 2014)). Interestingly this organization mirrors a well known characteristic of the representations in the primate inferior temporal (IT) cortex, and hence it hints at a possible cognitive justification of this computational model. To this regard it is interesting to mention that recent research investigated the connection between the representations built by several computational models and the representations in the IT cortex and found that deep neural networks are among the best models (Serre, 2016; Kriegeskorte, 2015). For instance, in (Khaligh-Razavi & Kriegeskorte, 2014), the authors investigate a wide range of computational models and suggest that deep CNNs are, not only the best performing in term of accuracy, but also the best at explaining the IT representation (albeit still in an incomplete way).

Given the great accuracy they achieve and the possible cognitive plausibility of the CNNs, we have chosen to use these particular models as the visual component of the word-object association model we propose in the next section.

Proposed model

In order to solve the problem of the lack of realism in the visual stimuli in the Mayor and Plunkett (2010) model, we propose to replace the input of the visual SOM with a representation built by a CNN as shown in Figure 2.

The long term objective of this research is to find a cognitively plausible model able to reproduce the word-object association abilities observed in infants using realistic image *and* audio stimuli. In this paper we keep a simplified auditory input and focus instead on providing a visual module capable

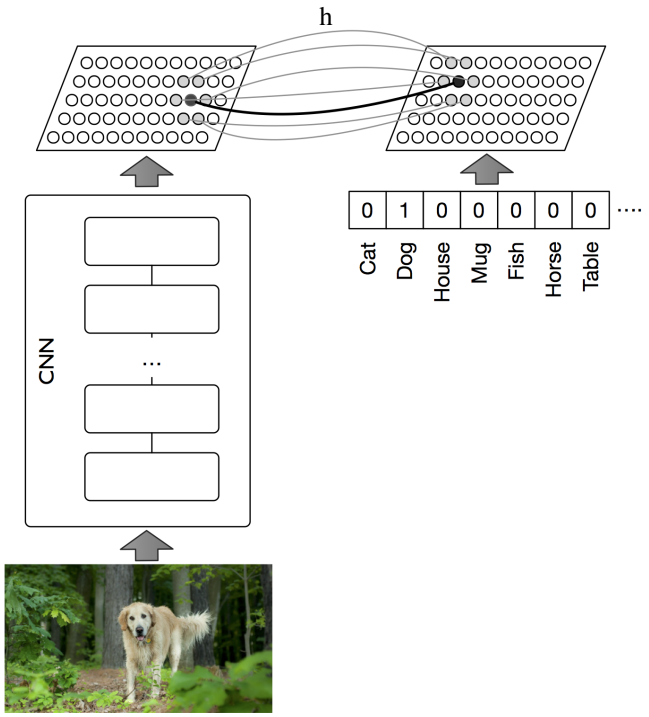

Figure 2: The proposed model: the visual component contains a deep convolutional network (InceptionV3) in order to process realistic images. The representation built by the deep neural network is then fed into the visual SOM. The acoustic component, on the contrary, only contains an acoustic SOM, as in Mayor & Plunkett (2010)'s model, and can only process simplified acoustic stimuli.

of handling realistic images. In fact, in this proposal the auditory input is a mere placeholder that does not provide any real processing ability.

In practice, we shall assume that an oracle provides the auditory SOM with a perfect representation of the auditory stimulus, or label, in the form of a binary vector. The vector contains a 1 in position i if the utterance provided to the auditory module corresponds to the i-th label; it contains a 0 otherwise.

The visual module shall, on the contrary, be able to cope with realistic images and, while we still assume that each image contains a main object corresponding to the concept to be learned, we pose no additional constraints. Images for a given concept can, for instance, be of different size, color, location in the picture, point of view, etc.. For instance, the "dogs" concept may be represented by images of dogs of different size, color, breed and be portrayed in different contexts, under different illuminations and poses.

The visual module is the concatenation of the InceptionV3 network and the SOM network we already introduced. InceptionV3 is a stack of Inception Modules, which parallelize and combine several convolution and pooling operations providing a richer output while still maintaining a small number of parameters. At the end of the stack of inception modules the

model contains a pooling layer of length 2048 which is fully connected with a shallow feed-forward neural network which is then used to classify the input image.

In the Inception architecture a representation of the input is propagated through the layers up to the top of the network where it is used to train the classifier. A question worth investigation is *if* and *where* a *good* representation of the concept in the input image is created by the deep network. In this paper we work under the assumption that such representation exists and argue that it has to be found in the last pooling layer (just before the fully connected neural classifier). Based on this assumption, we propose to use the vector containing the value of the 2048 neurons in that layer as the representation of the stimuli for the visual SOM. To verify the validity and the consequences of this assumption, we performed two sets of experiments: in the "Representation Quality" sub-section of the Experiments section we investigate the nature of the proposed representation, while in the subsequent section (Word Learning) we investigate the quality of the complete model. In our simulations we used the pre-trained Inception network provided by the TensorFlow library[2].

The complete system is trained as outlined in the "Mayor and Plunkett (2010)'s model" section. In summary, given the representation from the CNN and the simplified auditory input, the two SOMs (composed by 20×30 neurons each) are trained to cluster together similar representations using the standard SOM training algorithm. In our tests, the two SOMs attain their best topological organization of the objects in the training set after 60 epochs (the learning rate is set to 0.3 and decreases linearly at each epoch). Afterwards, the association between the visual and the auditory input is created using Hebbian connections between the two maps: two stimuli belonging to the same category are presented together to the model, the visual stimulus is processed to extract its representation and presented along with the auditory stimulus to the corresponding SOMs. Finally the SOMs activations are used to update the Hebbian connections using the update rule in Formula 2.

To better cope with the variability in the input representation, we introduce two variations to the Hebbian training (with respect to the procedure outlined in (Mayor & Plunkett, 2010)): *i)* we allow the network to learn from an increasing number of stimuli pairs (in the original paper a single pair of stimuli is presented to the network), this allows us to study how performances increase as the number of presentations grows; *ii)* we suppress the activation of a neuron in a SOM if its activation value is below 0.6.

Experiments

In the following two sub-sections we investigate two important facets of the proposed model. In particular, in the "Representation Quality" Section we show that the representation found in the last pooling layer of the InceptionV3 network allows one to cluster the input images into groups that correlate well with the classes assigned with the images themselves. This is arguably an evidence that such a representation can be usefully exploited as the input of the SOMs. In the "Word Learning" Section we focus on the complete model, replicate part of the experiments in (Mayor & Plunkett, 2010), and compare our results with those reported in that paper.

All the experiments have been performed on two datasets. A first dataset is composed by 10 classes associated with 100 stimuli each, for an overall 1.000 stimuli. A second dataset contains 100 classes associated with 100 stimuli each, for an overall of 10.000 stimuli. Since the results for the two datasets are very similar, for the sake of readability we focus on the smaller dataset and refer to (Fenoglio, 2016) for the details of the experiments on the larger dataset. The code for the complete model along with the datasets used can be found at `https://github.com/ml-unito/NNsTaxonomicResponding`.

Representation Quality

In order to assess the quality of the representation found in the last pooling layer of the InceptionV3 model, we investigate how well these representations can be clustered together. For each image we extract the representation found in the last pooling layer of the deep network, we then cluster the resulting representations using a K-means and an agglomerative algorithm. For both algorithms the number of clusters is set to 10. The clustering experiments have been conducted using the scikit-learn python library[3].

Figures 3 and 4 report results for K-means clustering. Analogous results hold for agglomerative clustering. In particular, Figure 3 reports, for each class, a bar showing how the class objects are partitioned among clusters; Figure 4 reports, for each cluster, a bar showing the distribution of the classes within it. We then investigate the topological organization provided by the visual SOM out of the representations created by the deep model. We report in Figure 5 a representation of the topology found by the visual SOM after 60 learning epochs.

Discussion The experiments show that the two clustering algorithms are able to find good, albeit not perfect, partitions for the representations. In particular, Figure 3 shows that the objects in 7 out of 10 classes are mostly assigned homogeneously to a single cluster: in two of the remaining cases the objects are almost all distributed among two classes, while in a single case (and only for the k-means clustering algorithm) the objects are distributed on three clusters. Figure 4 shows a similar picture, but from the point of view of the clusters: in almost all cases (8 out of 10) we have clusters which are almost pure. The remaining two clusters conglomerate objects from different classes acting almost as folders where all uncertain objects are put.

Overall, it seems that the clustering algorithms do find a way to partition the representations of the objects into co-

[2] `https://github.com/tensorflow/models/tree/master/inception`

[3] `http://scikit-learn.org/stable/modules/clustering`

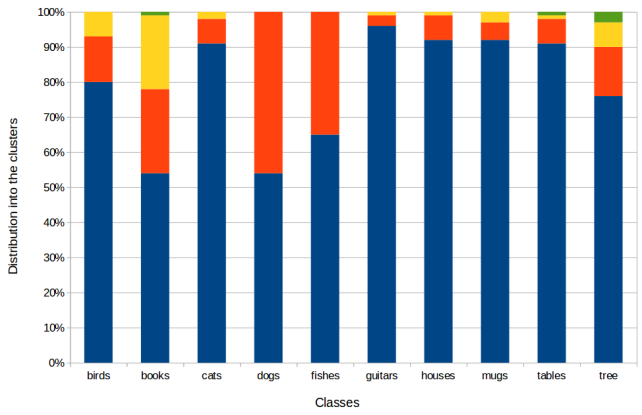

Figure 3: Per class distribution of objects into clusters (K-means clustering). Colors represents different clusters. Blue is used to represent the cluster containing the majority of the objects of a given class; orange, yellow, and green are used to represent the second, third, and fourth most represented clusters.

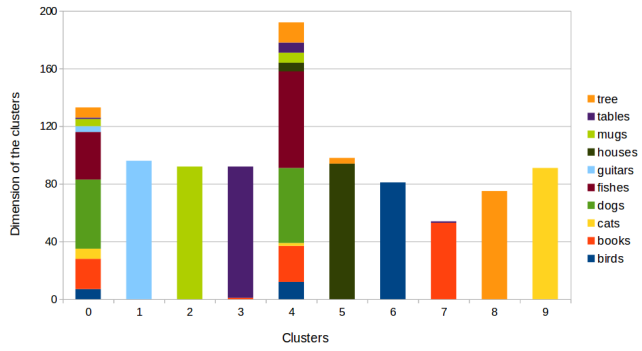

Figure 4: Distribution of classes among clusters (K-means clustering).

herent clusters. This is consistent with what happens for the topological organization that the SOMs create for the representations provided by the deep networks. Figure 5 shows that, with the exception of few cases, the visual SOM is able to group all related objects into nearby spaces.

Word Learning

In order to evaluate the performance of our model in the task of word learning, we calculate the *Taxonomic Factor* of the model as defined in (Mayor & Plunkett, 2010). We do so by testing the model for its production skills: for each class, 100 images are presented to the visual module, the activation is propagated through the deep neural network, then fed into the visual SOM. The activation of the visual SOM's best matching unit is propagated through the Hebbian connections up to the acoustic SOM. At the end of the process the resulting most active unit on the acoustic map is identified. It will be considered correct if it belongs to the area of the acoustic map

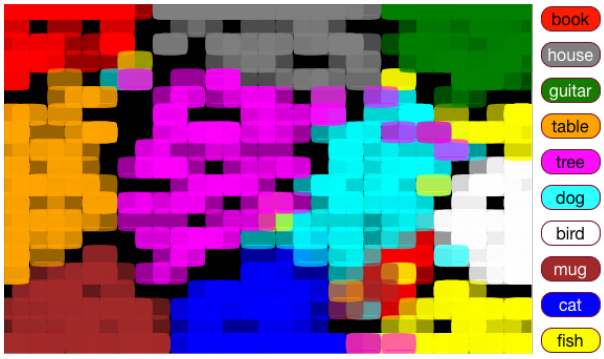

Figure 5: SOM clustering of the visual stimuli representations

associated to that word[4]. The percentage of correct words produced by the model when tested through all the classes is the Taxonomic Factor.

We have performed a number of experiments where we varied the number of presentations per class used to update the Hebbian connections. Specifically we let the number of presentations vary from 1 to 15. For each experiment we repeated the test over 1.000 different training sets (we kept fixed the SOM and let vary the images presented to the Hebbian learning module) and report the average taxonomic factor over an independent test set composed by additional 1000 images (100 images per class). Results are shown in Figure 6.

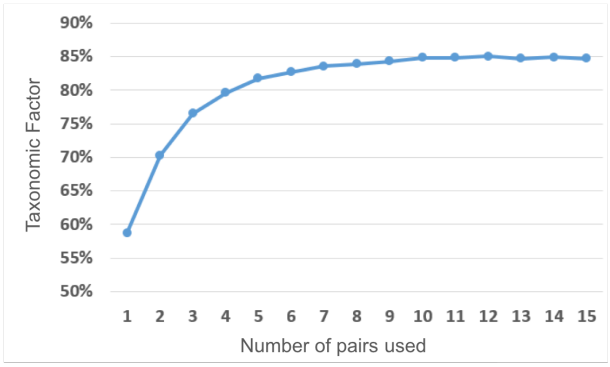

Figure 6: *Taxonomic factor* of the model, using an increasing number of pairs of stimuli per class during the training of the Hebbian Connection (on the x-axis).

Discussion The experiments show that the Taxonomic Factor steadily grows as more word-object associations are presented and reaches an accuracy above 80% (which is comparable with results in Mayor and Plunkett (2010)) at the fourth joint presentation.

[4] An area in the map is associated to a word if the activation of the neurons within it are at their peak when they respond to a stimulus of that particular word.

Conclusions

In this paper we have proposed an extension of the Mayor and Plunkett (2010) model for taxonomic responding. We have addressed the issue of adding realism to the visual stimuli. As a difference with respect to the original model in which these inputs were random dot patterns, the model can now deal with realistic images as those in the ImageNet dataset. This is possible thanks to the insertion of a deep convolutional neural network in the visual component of the model. Notwithstanding the higher complexity of the stimuli considered, our model exhibits taxonomic responding with performances comparable to the original one.

In our future work we will address the issue of making the acoustic module work with realistic stimuli. It can be interesting to explore whether a deep neural network for acoustic processing, as for instance the one proposed in (Xiong et al., 2016), could be nested into the acoustic part of the model in a way similar to what we already did for the visual component.

We will also explore whether the model proposed here can be used to provide a mechanistic account of the whole object constraint proposed by Markman (1989) by which a word is associated to the whole object instead of anyone of its properties. We conjecture that a model as the one proposed, with the deep component that extracts a representation of an object out of a more complex visual scene, can be adequate to the purpose: the whole object constraint may naturally emerge from the association of the word to the object's representation formed by the deep network. Important to this regard is the current discussion about the nature of the object representation built by deep networks (Ullman, Assif, Fetaya, & Harari, 2016; Tang & Kreiman, 2017).

References

Agrawal, P., Girshick, R., & Malik, J. (2014). Analyzing the performance of multilayer neural networks for object recognition. In *European conference on computer vision* (pp. 329–344).

Fenoglio, G. (2016). *A computational model for word learning on real world data through Deep Neural Networks*. Unpublished master's thesis, Turin University. (`https://github.com/gfbfenoglio/NNsTaxonomicResponding/blob/master/Documents/MasterThesis.pdf`, chapters 6,8)

Hinton, G., Deng, L., Yu, D., Dahl, G., rahman Mohamed, A., Jaitly, N., ... Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. *Signal Processing Magazine*.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014, November 6). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Computational Biology*, *10*(11).

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Li, X., & Wu, X. (2015). Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition. In *Acoustics, speech and signal processing (icassp), 2015 ieee international conference on* (pp. 4520–4524).

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. Mit Press.

Mayor, J., & Plunkett, K. (2010). A neurocomputational account of taxonomic responding and fast mapping in early word learning. *Psychological review*, *117 1*, 1–31.

Posner, M., Goldsmith, R., & Welton Jr, K. (1967). Perceived distance and the classification of distorted patterns. *Journal of Experimental Psychology*, *73*(1), 28-38.

Quine, W. V. O. (1960). Word and object. *Cambridge,Mass.: MIT Press*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Serre, T. (2016). Models of visual categorization. *Wiley Interdisciplinary Reviews: Cognitive Science*, *7*(3), 197–213.

Socher, R., Bauer, J., Manning, C. D., & Ng, A. Y. (2013). Parsing with compositional vector grammars. In *In proceedings of the ACL conference.*

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1–9).

Tang, H., & Kreiman, G. (2017). Recognition of occluded objects. *In Computational and Cognitive Neuroscience of Vision. (ed Zhao, Q). Singapore: Springer-Verlag*, *14*, 57–77.

Ullman, S., Assif, L., Fetaya, E., & Harari, D. (2016). Atoms of recognition in human and computer vision. *Proceedings of the National Academy of Sciences*, *113*(10), 2744–2749.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., ... Zweig, G. (2016). The microsoft 2016 conversational speech recognition system. *CoRR*, *abs/1609.03528*.

Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4694–4702).

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833).

Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., & Torralba, A. (2014). Object detectors emerge in deep scene cnns. *CoRR*, *abs/1412.6856*.