



Deep Hyperbolic Semantic Embedding Models for Zero-Shot Learning

Hermann Hälvää¹

MSc. Computational Statistics and Machine Learning

Supervisor: Prof. Bradley Love

Submission date: 7 September 2018

¹**Disclaimer:** This report is submitted as part requirement for the MSc CSML degree at UCL. It is substantially the result of my own work except where explicitly indicated in the text. The report may be freely copied and distributed provided the source is explicitly acknowledged

Abstract

Despite their impressive achievements on virtually all areas of computer vision, modern *deep learning* models remain constrained in their ability to generalize into novel environments. One of the most challenging problems is that of *zero-shot learning*, in which models attempt to label images from previously unseen classes. A common approach to this problem has been to project images into a semantic embedding space where zero-shot recognition can be performed by nearest neighbour search. A limitation of the embeddings spaces employed by previous works is that they only capture semantic similarity of labels, but fail to account for more complex hierarchical semantic relations that could further improve generalization performance.

This thesis aims is to endow deep learning models with a rich hierarchical semantic understanding and to investigate its benefits on zero-shot learning. To do this we introduce a *Deep Hyperbolic Semantic Embedding model* which use a deep learning model to map into a hyperbolic semantic embedding space. Hyperbolic embeddings have recently been shown to be especially adept at representing hierarchical graphs. The model was evaluated on a dataset of 20k previously unseen image classes. We found that even though the hyperbolic model's accuracy in predicting fine-grained image labels was lower than those of its Euclidean benchmarks, it made less predictions that were completely wrong. In particular, our models showed superior performance in classifying images into their higher level categories.

Acknowledgements

First, I would like to thank my supervisor Prof. Bradley Love, for all of his guidance, support and loads of great ideas. His genuine interest in the project and desire to help out really made the entire process very enjoyable. Massive thanks also to Olivia Guest, and Sebastian Bobadilla-Suarez for sharing their ideas and for pointing me to loads of relevant research I likely would have otherwise missed. Thank you to Adam Hornsby for helping me get my head around the different word embedding methods and for how I should think about presenting the project. Special thank you to Brett Roads for being incredibly helpful with setting up and helping to maintain the compute resources and data, which were required for this project; I was so impressed with the whole set up and how well it all functioned. For sure, nothing would have come of this project had I not received all this help, so for that - thank you!

Contents

1	Introduction	2
1.1	Project Motivation	2
1.2	Summary of Key Literature	5
1.3	Summary of Our Approach	7
1.4	Summary of Results	8
2	Background	11
2.1	Artificial Neural Networks & Deep Learning	11
2.1.1	Training ANNs	14
2.1.2	Convolutional Neural Networks (CNNs)	18
2.2	Exploiting Semantic Information in Computer Vision	23
2.2.1	Semantic Hierarchies and the Human Visual System	23
2.2.2	Semantic Hierarchies for Image Classification	24
2.2.3	Semantic Embeddings and Zero-shot Learning	26
2.3	Semantic Embeddings in Hyperbolic Space	33
2.3.1	Basics of Hyperbolic Geometry	33
2.3.2	Semantic Embeddings in Hyperbolic Spaces	40
2.4	Summary of Background	44

3 Our Approach	45
3.1 Approach Overview	45
3.2 Data & Model Details	46
3.2.1 Training Data	46
3.2.2 Poincare Embedding Construction	47
3.2.3 Deep Hyperbolic Semantic Embedding Model (Deep-HSEM)	47
3.2.4 Convex Hyperbolic Semantic Embedding Model (Convex- HSEM)	52
4 Experiments & Results	55
4.1 Overview	55
4.2 Baseline Models	55
4.3 Training Results	57
4.4 Experiment 1 - Zero-shot Classification	59
4.4.1 Experimental Setup	59
4.4.2 Results & Discussion	60
4.5 Experiment 2 - Robust Zero-Shot Classification	63
4.5.1 Experimental Setup	63
4.5.2 Results & Discussion	66
5 Conclusions & Ideas for Future Work	71
5.1 Summary	71
5.2 Limitations & Future Work	73

List of Figures

- | | |
|---|---|
| 1.1 Hypothetical example of zero-shot classification. Assume that we have used training images and their embedding vectors (blue dots) to learn a mapping from image features into the hyperbolic embedding space. Next an image from a previously unseen class is encountered, since the entity in the image looks a bit like a mix of golden retriever and Irish red setter, the predicted semantic embedding falls in between those two, which indeed happens to be close to the ground-truth embedding: Nova-Scotia Duck Tolling Retriever. | 3 |
| 1.2 Illustration of the Convex-HSEM model. In the first stage the model gives class predictions for any input image in terms of the 1000 training classes. Next the T -most likely classes and their corresponding ground-truth embedding vectors are taken. Their convex combination is then calculated in the hyperbolic space (orange dot) and nearest neighbour classification is performed subsequently. Note that a pre-trained deep learning model can be used; no training against embedding vectors is done | 8 |

1.3 An example of an image and model predictions on the <i>robuse zero-shot classification</i> task which measures how well models are able to predict semantic hierarchy of visual objects. This handpicked example illustrates the robustness of the Deep-HSEM model.	10
2.1 A simple Feed-forward ANN with two hidden layers. This figure illustrates a graphical model for Equation 2.1 - 3.3 and explicitly shows the matrix operations performed by the different neurons on a single vector input. The output dimension is set arbitrarily to be a 2×1 vector, which could for instance be used as class scores in binary classification; in practice, the dimension will be problem dependent.	13
2.2 A simple example of forward (top) and backward passes (bottom) for a simple two-layer ANN. This illustrates how gradients that flow to early layers are calculated by multiplying the local gradient of a given layer with the gradient that flows back from the upstream layer	17
2.3 Example of an ill conditioned Hessian. Momentum speeds up learning by accumulating the gradients of previous iterations such that velocity towards centre (lower loss) is established (shown in yellow). SGD without momentum keeps jumping across the loss surface as if in a downward sloping canyon, but fails to utilize the slope (blue trace). The contours depict different levels of loss that decreases inwards	18

2.4	A simple example of the computations performed by a single 2x2 convolution kernel on a 4x4 input. Assuming stride length of two squares, the kernel will see each of the four corners and performs a convolution operation on each of them. This is shown explicitly for the bottom left green square. The computation is hence essentially a dot product similarity metric between the kernel and the local image areas.	20
2.5	A graphical illustration of zero-shot learning. The left square contains all the input images. Red ones are training classes which are used to learn an approximation of the mapping $\mathcal{S} : X^d \rightarrow F^P$. The circle represents the semantic embedding space. Note that embedding vectors for zero-shot classes can be accessed as well. Upon seeing an image of unknown class (yellow square) the model outputs embedding vector prediction (yellow dot) and then the nearest plausible class (here fox) is found with a nearest neighbour function \mathcal{L} . Novel image is hence classified as a fox even though the model has never seen one before	28
2.6	Parallel lines in Hyperbolic space (Poincare disk model). The thick blue line is parallel to all the other lines since it never intersects any of them in the hyperbolic space (interior of the circle). All the lines are arcs of circles of different sizes that fall in the interior - these are known as geodesics. Image used under Creative Commons license, released into public domain by Trevorgoodchild of English Wikipedia	34
2.7	Saddle point in Euclidean space allows a representation of a hyperbolic triangle, for which the sum of its angles is less than 180 degrees. Image used under Creative Commons license from Wikimedia Commons.	35

2.8	Depiction of the Klein model of hyperbolic geometry. m , and n are parallel lines to l as they never intersect it.	36
2.9	Depiction of the Poincare Disk model of hyperbolic geometry. Line l represents an open chord passing through the centre of the circle, whilst line m is an open arc.	37
2.10	Upper sheet of the hyperboloid model in \mathbb{R}^3 and a projection of one of its geodesics on to the Poincare disk. While not possible to visually illustrate, the brim of the 'bowl' is at infinity and can never be reached. Image used under Creative Commons license from Wikimedia Commons.	39
2.11	Illustrations of Klein and Poincare Disks with hyperbolic tiling. All the tiles have equal area so this visually depicts how the area of the circles grows exponentially towards the boundary. Images used under Creative Commons license from Wikimedia Commons.	40
2.12	Naive example of embedding a hierarchical semantic taxonomy in a hyperbolic disk. All the concepts at each level of hierarchy are equidistant from the centre of the disk as portrayed by the dotted circles.	42
3.1	a.) Original VGG-16 structure except that we have taken out the final classification layer. The resulting model extracts a 4096 long feature vector from the input image. The dimensions of the layers are in the square brackets. b.) The VGG feature extractor is attached to a fully-connected layer which outputs a 10-dimensional embedding vector prediction. The distance of this prediction from the ground-truth embedding vector is used to compute a scalar loss. This is used to backpropagate gradients through the network.	49

3.2 Illustration of the Convex-HSEM model. The orange circle represents the Einstein midpoint of the semantic embeddings of the T most likely classes (yellow circles). The Einstein midpoint is mapped to the nearest ground-truth embedding in order to classify the input image	54
4.2 An example of top-3 & top-5 accuracy calculations in the robust zero-shot experiment. Predicting any of the five labels is allowed irrespective of which top- k is used, though only top- k predictions are counted. Image from the 2011 ImageNet database [1]	66
4.3 Four handpicked example test images from the robust zero-shot classification task. The first three illustrate successful high-level predictions by Deep-HSEM. Deep-Glove makes classifications at a very granular level and also makes more completely incorrect predictions. The fourth image is an example of a failure case for both models. Images from the ImageNet repository [1]	69

List of Tables

1.1	A small subset of results from our zero-shot classification experiment. Top- k classification accuracies for $k \in \{1, 2, 5, 10, 20\}$ on the three zero-shot datasets based on the 2011 20k ImageNet. None of the zero-shot classes were included in the training classes. The results for DeViSE and ConSE are directly from [2] and [3] respectively.	9
1.2	Small excerpt of Robust Zero-Shot Classification. In this task the models are rewarded for predicting the correct label or one of the correct higher-level categories of the image. Performance is measured with inverse distance-weighted top- k accuracies; predictions closer to the original label are credited more. The hierarchical 2011 20k datasets were derived from Experiment 1 by adding four levels of superordinate labels to each original label from the WordNet hierarchy.	10
4.1	Summary of all the models used in the evaluation experiments and their main differences.	56
4.2	Accuracies of all the models on the 2012 1k ImageNet validation data set.	58
4.3	Overview of the datasets in the zero-shot classification experiment	59

4.4	Top-k classification accuracies for $k \in \{1, 2, 5, 10, 20\}$ on the three zero-shot datasets based on the 2011 20k ImageNet, with increasing difficulty from top to bottom. None of the zero-shot classes were included in the training classes. Classification was always constrained to the labels of the dataset under consideration. The results for DeVISE and ConSE are directly from [2] and [3] respectively.	62
4.5	Robust Zero-Shot Classification results with inverse distance-weighted top- k accuracies. The hierarchical 2-hop, 3-hop and 2011 20k datasets were derived from Experiment 1 by adding four levels of superordinate labels to each original label from the WordNet hierarchy.	68

Chapter 1

Introduction¹

1.1 Project Motivation

Over the past decade deep learning has achieved state-of-the-art results on virtually all computer vision tasks [4]. The impressive performance of these models has, in turn, propelled paramount improvements in a wide range of application areas such as autonomous vehicles[5], face recognition [6] and robotics [7][8]. Despite the many successes of deep learning, most of the achievements rely on well defined training and test environments; the generalization performance of these models is still far from that of the human visual system. One particularly challenging task that illustrates this is zero-shot learning, that is, making predictions about previously unseen classes of images. The aim of this thesis is to improve the generalization performance of deep learning models in the context of zero-shot learning. We approach this by attempting to teach the models a more human-like hierarchical understanding of object categorization.

Some of the biggest advances in deep learning have come in the area of image classification. Inspired by the breakthrough performance of Krizhevsky

¹All the code developed for this project can be accessed at
<https://github.com/HHalva/msc-thesis>

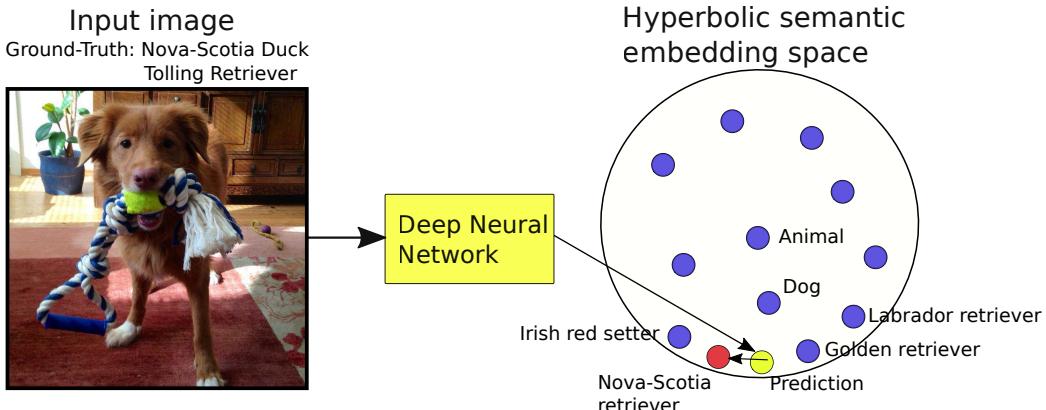


Figure 1.1: Hypothetical example of zero-shot classification. Assume that we have used training images and their embedding vectors (blue dots) to learn a mapping from image features into the hyperbolic embedding space. Next an image from a previously unseen class is encountered, since the entity in the image looks a bit like a mix of golden retriever and Irish red setter, the predicted semantic embedding falls in between those two, which indeed happens to be close to the ground-truth embedding: Nova-Scotia Duck Tolling Retriever.

et al. [9] on the ImageNet 2012 classification challenge [1], researchers have produced increasingly sophisticated deep learning architectures and as a result have recently surpassed human performance on this task [10]. Yet, the classifications these models are able to perform are inherently restricted; any new image can only be labelled as one of the training classes. In other words, it is not possible to make correct predictions about any new classes of images without further training data on them. Zero-shot learning [11] tries to address this limitation.

Whilst making predictions about previously unseen objects may sound impossible at first, humans can easily do this, under certain conditions. To see this, imagine for a moment that you have never seen a Zebra, but your zoologist friend aptly tells you that “a Zebra is like a horse with black and white stripes”. The next time you see an animal that fits this description, you could recognize the Zebra, even though you have never seen one before! This

represents the power of auxiliary semantic information in image classification. In typical zero-shot recognition models this information comes in the form of word embeddings. These are usually constructed so that words which occur in a similar semantic context are placed close to each other in a vector space. The classic zero-shot approach is to learn a mapping from training images into the word embedding. This mapping is then used to project any novel images into the embedding space where they can be classified using nearest neighbour search (see Figure 1.1 for an example).

A fundamental assumption that underpins this approach is that there is a close relationship between contextual semantic similarity and visual similarity. While this is often true [12], these similarity-based embeddings fail to account for more complex, hierarchical, semantic relations, which could better reflect visual characteristics of objects. For instance, biological organisms can be organized in a hierarchical taxonomy based on their physical traits [13]. In fact, the rich visual understanding that humans display is based on semantic and visual hierarchy of objects [14]. A related limitation is that similarity-based embeddings don't easily accommodate for subordinate relations: it might be possible to reason that the words 'cat' and 'pet' are grouped near each other but cannot deduce whether 'pet' is a 'cat' or 'cat' a 'pet'. In contrast to this, a human who sees a cat of an unknown breed, can traverse up the hierarchy and group it under the higher level category of cat, or even animals and mammals if necessary.

We propose a deep learning model with the ability to reason in semantic hierarchies by training the model to project input images into a hyperbolic embedding space. Hyperbolic geometry has recently been shown to be well suited for capturing hierarchical graph structures such as semantic taxonomies [15, 16]. We hypothesize that the resulting structure of the embedding space will better reflect the visual similarities of different objects, which should improve zero-shot classification. Our central research question can thus be stated as:

- *Does semantic hierarchical knowledge in the form of hyperbolic embeddings help improve the zero-shot prediction capability of deep learning models?*

Our research into zero-shot learning is also motivated by the fundamental scarcity of readily available labelled data. This is a problem for many real-world applications of deep learning. As an example, for autonomous vehicles to operate safely, they need to be able to perform inference in novel visual scenarios. Being able to identify an unknown object as a living entity can be incredibly valuable, even if more precise classification is not possible.

Overall, our results suggest that hyperbolic embeddings do in fact allow deep learning models make more robust predictions on previously unseen classes of images. In particular, our models perform better at predicting the higher level categories of novel visual objects than a benchmark that uses a similarity-based embedding. The cost of this robustness is, however, that the general zero-shot classification accuracy of our model is lower than for various baseline models from the literature. In conclusion, our models predict the exact class of an object correctly less often, but also makes less classifications that are completely wrong.

In the remainder of this chapter we provide a brief overview of the key literature that inspired this thesis, our methodology and display some of our key results more explicitly. In Chapter 2 we cover the necessary technical background for this project and review the literature in more detail. Chapter 3 presents our approach in detail, and the full results from our experiments are given in Chapter 4. Chapter 5 concludes our work with a discussion of the results and presents ideas for future research.

1.2 Summary of Key Literature

Modern deep neural networks take the form of Convolutional Neural Networks [17], which learn a hierarchical visual representation of objects over

multiple layers of convolution operations. Typically these models are trained against one-hot-encoded vectors of class labels: if there are N classes, the label of each image will be N -dimensional vector with a 1 in the dimension of the correct class and 0s in all other dimensions. Crucially, these vectors are orthogonal to each other. This is why standard deep learning models do not learn to capture any semantic relations between the visual objects, even though the hierarchical visual features that the model learn would be well suited for this. As a consequence, a standard deep learning model is not able to communicate, for example, that cats and dogs are both mammals if training labels are available only for the former two. Those few works that have attempted to use multi-level labels [18], [19], have only used up to three levels of hierarchy. Whilst their results are promising, such shallow hierarchies are unable to represent the full complexity of semantic relations. Additionally, the generalization ability of these models has only been tested on training labels but not on zero-shot prediction.

As discussed above, zero-shot learning has instead focused on improving generalization performance via word embeddings [11, 20, 21, 2, 3]. Note that this is possible because the embedding spaces are created so that they contain all the possible labels, including zero-shot ones. Hence, projecting images into this space can help with classification. Several methods have been used to create such all encompassing label-vector spaces. The most popular has been to utilize the word2vec family of models [22, 23]. These take in a large corpus of text, such the 5.2 billions words long Wikipedia data, and place words that occur in a similar context near to each other in the resulting, Euclidean, embedding space. For example, the model that is most similar to ours, DeViSE [2], uses a deep neural network to map into this space.

As alluded to earlier, similarity embeddings fail to capture the semantic hierarchy of objects, which is how humans categorize the world around us [24, 14]. One likely reason why hierarchical relations haven't been used by

previous works is that embedding them in Euclidean space is difficult. Only very recently have Nickel and Kiela [15] and Chamberlain *et al.* [16] shown that this can be done much more easily in hyperbolic geometry, and thus have introduced the idea of hyperbolic graph embeddings.

The main property than distinguishes hyperbolic spaces from Euclidean ones is that they have a constant negative curvature [25]. The consequence of this is that the area of a hyperbolic sub-space, such as a circle, expands exponentially as we move away from the centre, as opposed to the quadratic growth in Euclidean geometry. This exponential growth is also the rate at which hierarchical trees, with a constant branching factor, grow; hence the suitability of hyperbolic spaces to embed semantic hierarchies [15, 16].

1.3 Summary of Our Approach

We started by constructing a hyperbolic embedding that contains all the possible ground truth labels. For this we used the Poincare embedding model of Nickel and Kiela [15]. More specifically, this model was applied to the WordNet lexical data base which consists of 82,000 nouns and their semantic relations in the form of a graph. As a result, we acquired 10-dimensional Poincare embedding vectors for all training and zero-shot labels.

In order to learn a mapping from input images into the hyperbolic space, we trained two different deep learning models on the 1000 classes of the ImageNet 2012 database [1], approximately 1.3 million images. Our first model, called Deep Hyperbolic Semantic Embedding model (Deep-HSEM) employs a standard convolutional neural network in a similar vein as the DeViSE model of Frome *et al.* [2]. The general architecture of this model is depicted in Figure ???. Our second model, called Convex Hyperbolic Semantic Embedding model (Convex-HSEM) follows the approach of Norouzi *et al.* [3] and computes the predicted embeddings for zero-shot classes as convex combinations of known classes' embeddings. See Figure 1.2 for an illustration.

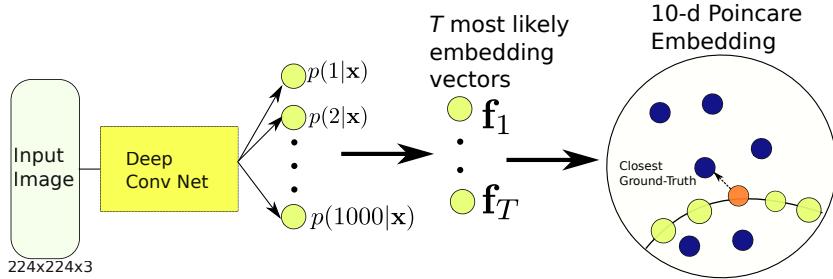


Figure 1.2: Illustration of the Convex-HSEM model. In the first stage the model gives class predictions for any input image in terms of the 1000 training classes. Next the T -most likely classes and their corresponding ground-truth embedding vectors are taken. Their convex combination is then calculated in the hyperbolic space (orange dot) and nearest neighbour classification is performed subsequently. Note that a pre-trained deep learning model can be used; no training against embedding vectors is done

All of our code is available at <https://github.com/HHalva/msc-thesis/>.

1.4 Summary of Results

After training, the Deep-HSEM and Convex-HSEM models were evaluated on 20k previously unseen zero-shot classes, obtained from the full ImageNet 2011 release [1]. This dataset contains a total of around 13 million images. In our first experiment we measured the standard classification accuracy of our models on this dataset and compared them to three baseline models that use similarity-based embeddings. These baselines were the DeViSE [2] and ConSE models [3] as well as Deep-Glove which is equivalent to the Deep-HSEM, except that it utilizes a Euclidean rather hyperbolic embedding. A small excerpts of our results is given in Table 1.1. To focus on our main results, compare our models to Deep-Glove: we can see that even though the top-1 accuracy of the hyperbolic models is much lower, the top-20 accuracy is higher. This provided the first indication that even though our models are worse at making precise classifications, the overall structure of the embed-

Test Dataset	Model	Top-k Accuracies(%)				
		1	2	5	10	20
2011 20k ImageNet	Deep-HSEM	0.2	0.4	1.0	2.0	3.5
	Convex-HSEM	0.2	0.4	1.1	2.2	3.9
	Deep-Glove	1.0	1.3	1.9	2.5	3.1
	DeViSE	0.8	1.4	2.5	3.9	6.0
	ConSE(10)	1.4	2.2	3.9	5.8	8.3

Table 1.1: A small subset of results from our zero-shot classification experiment. Top-k classification accuracies for $k \in \{1, 2, 5, 10, 20\}$ on the three zero-shot datasets based on the 2011 20k ImageNet. None of the zero-shot classes were included in the training classes. The results for DeViSE and ConSE are directly from [2] and [3] respectively.

dings may be better at relating visual and semantic categories.

One shortcoming of the above experiment is that it only gives credit for correct classifications. Yet, we hypothesized that the strength of the hyperbolic models would be in their hierarchical understanding of object categorization. It is thus possible that our models are not certain about the exact label of a new object but are still able to place them into appropriate higher-level categories. Thus we developed a second experiment, called *robust zero-shot classification* in which models were asked to predict both the correct label of an image as well as which super-categories it belongs under. Table 1.2 displays a snapshot of the results from this task. We can see that now Deep-HSEM and Convex-HSEM clearly outperform Deep-Glove; the ability of our models to classify images into their correct super-categories makes up for their lower accuracy in predicting the exact label. More detailed analysis of the results suggested that in general the hyperbolic models make less completely wrong predictions (see Figure 1.3 for an example) which warrants further research into this promising family of models.

Robust Zero-Shot Classification		Weighted Top-k Accuracies(%)				
Test Dataset	Model	1	2	3	4	5
Hierarchical 2011 20k ImageNet	Deep-HSEM	2.5	5.0	7.3	9.3	11.2
	Convex-HSEM	1.8	4.2	6.4	8.3	10.2
	Deep-Glove	1.1	1.4	1.7	1.8	2.0

Table 1.2: Small excerpt of Robust Zero-Shot Classification. In this task the models are rewarded for predicting the correct label or one of the correct higher-level categories of the image. Performance is measured with inverse distance-weighted top- k accuracies; predictions closer to the original label are credited more. The hierarchical 2011 20k datasets were derived from Experiment 1 by adding four levels of superordinate labels to each original label from the WordNet hierarchy.

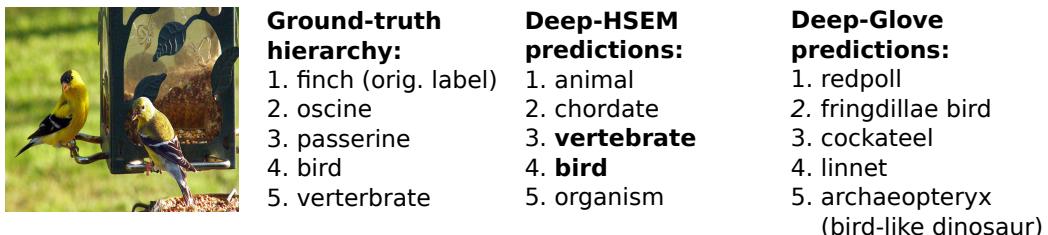


Figure 1.3: An example of an image and model predictions on the *robuse zero-shot classification* task which measures how well models are able to predict semantic hierarchy of visual objects. This handpicked example illustrates the robustness of the Deep-HSEM model.

Chapter 2

Background

2.1 Artificial Neural Networks & Deep Learning

Even though deep learning is often viewed as a new technique, in reality the recent breakthroughs are underpinned by decades of related research. For instance, earliest artificial neural networks, such as Rosenblatt's Perceptron [26] from the 1950s, are closely related to linear regressions; these models are all of the form $f(\mathbf{x}, \mathbf{w}) = \mathbf{x}^T \mathbf{w}$ where \mathbf{x} is a vector describing some input covariate and \mathbf{w} model weights. Much like in linear regression, the aim is to learn a mapping $f(\mathbf{w}, \mathbf{x}) = y$ that defines the relationship between the input \mathbf{x} and some category y . In a simple problem, y is binary and the weights \mathbf{w} are learned such that the resulting hyperplane can linearly separate the data into the two. These early models were largely restricted by their assuming linear separability and there was also no computationally feasible way of training the models at the time [4], [27]. Nevertheless, they form the basis for nearly all Artificial Neural Networks (ANNs).

Several steps were taken to increase to complexity of these models and to allow for non-linearities, many of them initially inspired to some degree

by biological neurons [?]. For instance, real neurons typically only fire once action potential exceeds a specific threshold [28]. In ANNs, similar behaviour is attained via activation functions on top of neuron outputs, most typically in the form of Rectified Linear Units (ReLU), which apply the following non-linearity: $f(\mathbf{x}, \mathbf{w}) = \max\{0, \mathbf{x}^\top \mathbf{w}\}$. ReLUs also improve the representational capacity of a network by imposing sparsity which can make it easier to disentangle the data [29] and hence lead to faster convergence of training [9].

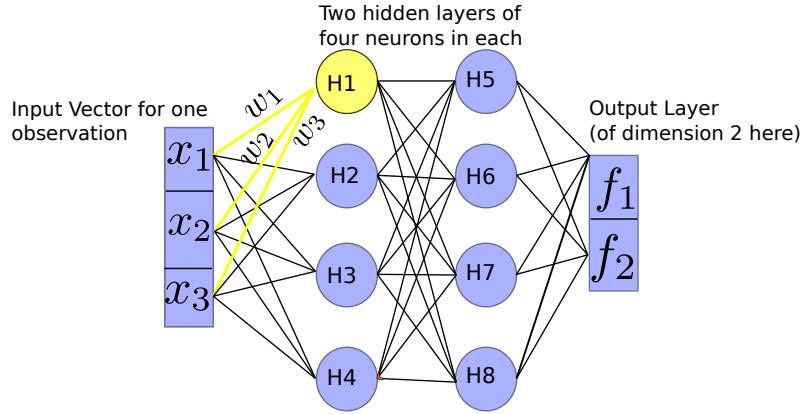
Another insight borrowed from neuroscience is that intelligence stems from groups of neurons acting together rather than from individual neurons [4]; this idea is behind deep ANNs where several layers of neurons are connected to each other and numerous neurons are present in each layer. Below equations and Figure 2.1 give a simplified example of an ANN with two hidden layers:

$$\mathbf{H}_1 = \max\{0, \mathbf{W}\mathbf{X} + \mathbf{B}\} \quad (2.1)$$

$$\mathbf{H}_2 = \max\{0, \mathbf{V}\mathbf{H}_1 + \mathbf{C}\} \quad (2.2)$$

$$\mathbf{F} = \mathbf{Z}\mathbf{H}_2 + \mathbf{D} \quad (2.3)$$

where \mathbf{X} is an $D \times N$ input data matrix of N observations in the columns and D is the dimension of a single observation. This matrix representation allows several datapoints to be fed through the network simultaneously, which is what is done in practice. \mathbf{W} , \mathbf{V} and \mathbf{Z} are the weights matrices of the two hidden layers and output layer respectively. The number of rows in each of the weight matrices is the number of neurons in that layer, whilst the number of columns corresponds to the input dimension. Notice also that constant bias matrices \mathbf{B} , \mathbf{C} , and \mathbf{D} are added to the neuron outputs. These bias matrices are akin to intercepts in regression analysis and increase the representation ability of the model. Mathematically these matrix operations



First Hidden Layer:

$$\begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \end{bmatrix} = \begin{bmatrix} w_1 & w_2 & w_3 \\ w_4 & w_5 & w_6 \\ w_7 & w_8 & w_9 \\ w_{10} & w_{11} & w_{12} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} \xrightarrow{\text{ReLU}} \begin{bmatrix} H1 \\ H2 \\ H3 \\ H4 \end{bmatrix} = \begin{bmatrix} \max(0, h_1) \\ \max(0, h_2) \\ \max(0, h_3) \\ \max(0, h_4) \end{bmatrix}$$

Second Hidden Layer:

$$\begin{bmatrix} h_5 \\ h_6 \\ h_7 \\ h_8 \end{bmatrix} = \begin{bmatrix} v_1 & v_2 & v_3 & v_4 \\ v_5 & v_6 & v_7 & v_8 \\ v_9 & v_{10} & v_{11} & v_{12} \\ v_{13} & v_{14} & v_{15} & v_{16} \end{bmatrix} \begin{bmatrix} H1 \\ H2 \\ H3 \\ H4 \end{bmatrix} + \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} \xrightarrow{\text{ReLU}} \begin{bmatrix} H5 \\ H6 \\ H7 \\ H8 \end{bmatrix} = \begin{bmatrix} \max(0, h_5) \\ \max(0, h_6) \\ \max(0, h_7) \\ \max(0, h_8) \end{bmatrix}$$

Output Layer:

$$\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \begin{bmatrix} z_1 & z_2 & z_3 & z_4 \\ z_5 & z_6 & z_7 & z_8 \end{bmatrix} \begin{bmatrix} H5 \\ H6 \\ H7 \\ H8 \end{bmatrix} + \begin{bmatrix} d_1 \\ d_2 \end{bmatrix}$$

Figure 2.1: A simple Feed-forward ANN with two hidden layers. This figure illustrates a graphical model for Equation 2.1 - 3.3 and explicitly shows the matrix operations performed by the different neurons on a single vector input. The output dimension is set arbitrarily to be a 2×1 vector, which could for instance be used as class scores in binary classification; in practice, the dimension will be problem dependent.

are just affine transformations on the input followed by ReLUs, which are applied elementwise. The output of each layer is thus a matrix, here \mathbf{H}_1 and \mathbf{H}_2 , where each row corresponds to an output from a specific neuron. Notice also that ReLUs are not added on the output, rather the output layer transforms the representations of the final hidden layer into a desired shape of output. For instance, for binary classification \mathbf{F} is $2 \times N$, and the output values for each of the N inputs reflects the relative likelihood of the two labels.

The ANN model we have described thus far is known as Feed-forward Network or a Fully-Connected Network, which refers to all neurons being connected to all other neurons in the preceding and succeeding layers. This is an important point as it facilitates a hierarchical structure in which neurons in the later layers combine features from earlier layers to create more complex features. Relatedly, this architecture enables distributed representation in which several types of features can be combined in different ways to represent an exponential number of inputs [?]. As an example, if we have 'squares', 'rectangles' and 'circles' and each of them could be either 'red', 'green' or 'blue', then we have 9 possible visual objects, yet all the possibilities could efficiently be represented by combining three colour neurons with three shape neurons [4].

2.1.1 Training ANNs

In a typical classification problem we have n possible labels for each input and wish to predict a probability distribution over them. In these applications the outputs of an ANN are transformed into 0 to 1 range. More formally, consider that there are n possible classes and hence the output from the above ANN for an input \mathbf{x} is the $n \times 1$ vector \mathbf{f} . We require, that each element of \mathbf{f} is between 0 and 1 and that $\sum_1^n f_i = 1$. This is most commonly done by assuming that the outputs of the ANN are unnormalized predictions of label log-probabilities: $f_i = \log \hat{P}(y = i|\mathbf{x})$ [4]. By taking exponents

and normalizing across possible labels, the predicted class probabilities are calculated as per below - this is known as the softmax function:

$$\text{softmax}(\mathbf{f})_i = \frac{\exp(f_i)}{\sum_{j=1}^n \exp(f_j)} = q_i \quad (2.4)$$

where $q_i \in [0, 1]$ captures the predicted probability that the input belongs to the class i . One reason for the softmax layer's popularity is that it is easily compatible with the cross entropy loss function defined as $L_i = -\sum_i p_i \log q_i$ [30]. In simple image classification tasks where the classes are mutually exclusive, we have $p_i = 0 \forall i \neq C$ and $p_i = 1$, with $i = C$ denoting the correct class. Plugging Equation 2.4 into this equation gives:

$$L_i = -\log \left(\frac{\exp(f_C)}{\sum_{j=1}^n \exp(f_j)} \right) = -f_C + \log \left(\sum_{j=1}^n \exp(f_j) \right) \quad (2.5)$$

which is the loss incurred from one observation, and is continuous and differentiable. The sequence of computations leading from model inputs all the way to scalar loss is known as a forward pass. Usually the forward pass is computed simultaneously for several inputs, known as a mini-batch, in which case the average loss across the mini-batch is typically used.

The aim of training an ANN is based on learning model weights that minimize some appropriate loss function, such as the cross-entropy loss above. Most supervised deep learning models, including the basic feed-forward-network described above, are nowadays trained using the back-propagation algorithm [31] [32]. The main idea of this algorithm is to use the chain rule to decompose the gradient of a loss function so that it can be efficiently passed back through the network. More formally, assume the loss of an ANN is produced by a sequence of m nested operations:

$$L(y_i, x_i) = f^{(m)}(y_i, f^{(m-1)}(\dots f^{(2)}(f^{(1)}(x_i)))) \quad (2.6)$$

where y_i is the correct label of the observation, x_i the input data, and the different $f^{(i)}$ may for example represent different types of layers. Employing the chain rule recursively, a simple decomposition gives:

$$\frac{\partial L}{\partial x_i} = \frac{\partial L}{\partial f^{(m)}} \frac{\partial f^{(m)}}{\partial f^{(m-1)}} \dots \frac{\partial f^{(1)}}{\partial x_i} \quad (2.7)$$

These computations are done in the opposite order of the forward-pass; hence back-propagation. Above example is simplistic since usually in addition to inputs from preceding layer, each layer has also its own parameters. The general idea of the chain rule still works, however now the gradient flow bifurcates at each layer: part of gradients flow into the earlier layers while the others flow back into the parameters. This is explained in more depth in Figure 2.2.

By representing ANNs as computational graphs, and using the chain rule, it becomes easy to see how backpropagation can be used to send appropriate gradients to right places even in complex network architectures. Importantly, backpropagation does this efficiently since at each operation all upstream gradients are collated and then passed one layer down. This is a lot more efficient than considering every single path through an ANN individually. To see this, consider Figure 2.1: with backpropagation, we can start from the output and calculate the derivative of the output layer with respect to $H5$ neuron, and pass this derivative to all neurons $H1 - H4$ simultaneously. This requires a lot less computation than considering all the paths that involve $H5$ separately.

After gradients have been calculated using back-propagation, they are used by an optimization algorithm to alter the model's parameters, with the

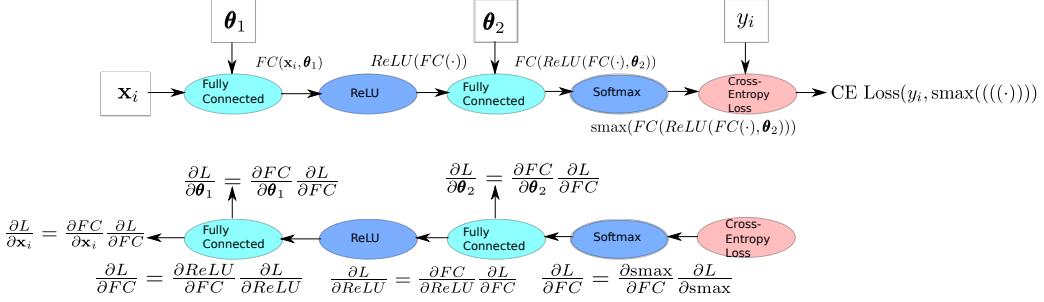


Figure 2.2: A simple example of forward (top) and backward passes (bottom) for a simple two-layer ANN. This illustrates how gradients that flow to early layers are calculated by multiplying the local gradient of a given layer with the gradient that flows back from the upstream layer

aim of minimizing the loss function. Here we consider stochastic gradient descent (SGD) [33] which is the most widely used optimization algorithm in deep learning. The algorithm is called stochastic because at each learning iteration only a mini-batch, of the data is used to calculate the gradients. It has been shown that SGD converges much faster than calculating gradients always on all available data; the time per update for the algorithm is independent of data size, as long as the batch size is held constant [4]. The parameter updates are based on moving 'downhill' i.e. in the direction of negative gradient:

$$\theta_{(t+1)} = \theta_{(t)} - \eta \nabla_{\theta_{(t)}} L(\mathbf{X}, \mathbf{y}) \quad (2.8)$$

where $\nabla_{\theta_{(t)}} L(\cdot)$ is the gradient of a loss function with respect to model parameters (multi-variable equivalent of derivative), and η is the learning rate.

Most deep learning models are very sensitive to the choice of learning rate. If it is too high, we are likely to miss minima, and conversely there is a risk of local minima and slow training when it is set too small. Usually learning rate is reduced linearly with training time, or in bigger steps at regular intervals.

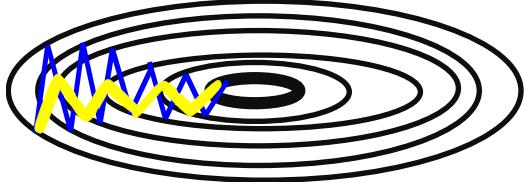


Figure 2.3: Example of an ill conditioned Hessian. Momentum speeds up learning by accumulating the gradients of previous iterations such that velocity towards centre (lower loss) is established (shown in yellow). SGD without momentum keeps jumping across the loss surface as if in a downward sloping canyon, but fails to utilize the slope (blue trace). The contours depict different levels of loss that decreases inwards

This reduces the impact of noise when we approach a minimum [4].

A common addition to the vanilla SGD is momentum [32], which can accelerate learning when there is a lot of noise or when the Hessian of the loss (matrix of 2nd order derivatives) is ill-conditioned, as shown in Figure 2.3. SGD with momentum amends the original SGD by introducing a velocity term that accumulates gradients from previous iterations with an exponential decay [4]. Rumelhart and Hinton [32] described momentum as if dropping a ball-bearing on loss surface and letting gravity accelerate the ball. Further, the loss landscape can be imagined to be immersed in a liquid with a specified level of viscosity that defines how quickly the ball's momentum fades. Algorithm 1 gives an example of full SGD momentum algorithm. In practice, backpropagation and SGD-based optimization can nowadays be done automatically in modern deep learning libraries such as PyTorch [34] and TensorFlow [35].

2.1.2 Convolutional Neural Networks (CNNs)

So far we have considered only simple feed-forward ANNs with fully connected layers. Most of the ground-breaking accomplishments over the past decade in deep learning have however been achieved with convolutional neu-

Algorithm 1 SGD with momentum (following [4])

Require: Learning rate η , Momentum decay parameter α

Require: Initial parameters θ , initial velocity ν

while Convergence not met **do**

 Sample a minibatch of m observations from training data $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$

 Get corresponding targets from training data $\{\mathbf{y}_1, \dots, \mathbf{y}_m\}$

 Compute gradient for the minibatch: $\mathbf{g} \leftarrow \frac{1}{m} \nabla_{\theta} \sum_i L(f(\mathbf{x}_i | \theta), \mathbf{y}_i)$

 Update velocity: $\nu \leftarrow \alpha \nu - \eta \mathbf{g}$

 Update parameters: $\theta \leftarrow \theta + \nu$

end while

ral networks (CNNs) [36] of some sort. This is particularly true for computer vision tasks such as image classification [27]. In fact, the term deep learning is often used synonymously with CNNs that have a large number of layers.

Unlike fully connected neurons in feed-forward networks, convolutional layer has neurons, usually called kernels, which are connected only to some parts of the input - together many such neurons span the entire input data. Furthermore, there is weight sharing between the kernels to allow for the recognition of a particular feature anywhere in the input space [17]. Convolution layers are therefore particularly suited for data with locally correlated structures such as images. Consider an input image of 225x225, a typical convolution filter may have size 3x3 and, after being trained on the image data, it could have learned to represent a particular visual feature such as a vertical edge. This filter is then replicated over the entire 225x225 input range so that the model can detect vertical edges anywhere in the image. Usually each layer has a multitude of kernels that are able to detect different features. Mathematically, this feature detection corresponds to the convolution operation between input data X and convolution kernels K . For discrete 2D data this is given as $S(i, j) = (X * K) = (i, j) \sum_m \sum_n X(i - m, j - n)K(m, n)$ [4] where i and j denotes a particular pixel location and m and n the location in the kernel layer. Figure 2.4 gives a brief toy example to illustrate this process.

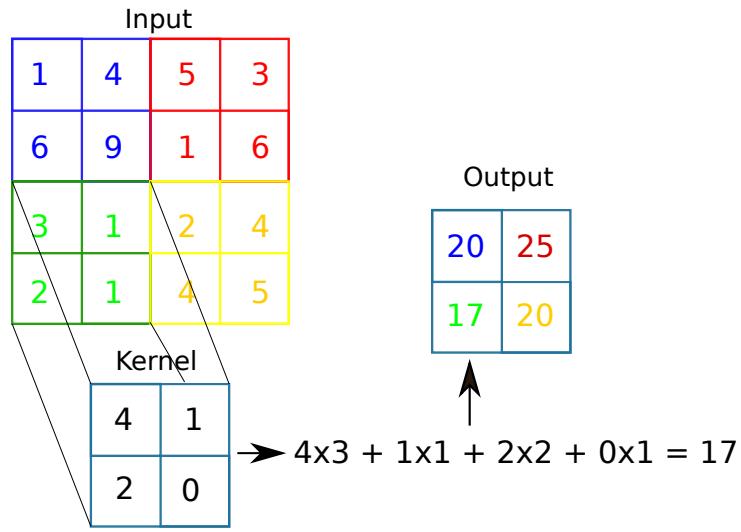


Figure 2.4: A simple example of the computations performed by a single 2×2 convolution kernel on a 4×4 input. Assuming stride length of two squares, the kernel will see each of the four corners and performs a convolution operation on each of them. This is shown explicitly for the bottom left green square. The computation is hence essentially a dot product similarity metric between the kernel and the local image areas.

The reason why convolutions have proven so effective for image data is that natural visual scenes present strong local correlations - the world we view is not just a collection of randomly ordered pixels. Additionally, visual features can appear anywhere in our visual field; thus the need for convolution layers to scan across the whole image [17]. Further, CNNs usually have several layers of convolutions on top of each other. When such deep CNNs are trained on image data, the kernels across the different layers usually learn a hierarchy of visual features: the early layers detect edges and corners, which are then used by middle layers to create contours and parts of objects, and finally later layers build up more complex representations, possibly of complete objects [37]. This shares some similarities with the mammalian visual cortex [38] in which earlier (V1) cells respond to edges and bars of different orientations [39] and later ones, like V4 and the IT cortex, to more complex shapes [40].

Convolution layers in CNNs are normally followed up by max-pooling functions. Usually these come in the form of the *max()* functions applied on top of convolution outputs. For example, in Figure 2.4 a max-pooling operation would only pass through the value of 25. The benefit of such down-sampling is local invariance: visual objects are not completely rigid (imagine a letter that's slightly rotated between different views). By passing on only the largest value, we are likely to produce the same output even if pixel values of the input moved around a bit [17].

One of the most fundamental properties of CNNs is that they can be trained end-to-end using back-propagation and SGD. A very important consequence is that the models automatically learn features that best fit the input data. This dispenses with the need to manually engineer visually features, which was the approach taken by earlier machine vision algorithms such as SIFT [41]. CNNs trained on backpropagation produced very large accuracy gains over these methods [42], and following the breakthrough performance of the seminal AlexNet CNN [9], deep CNNs of various designs

have become state-of-the-art in virtually all machine vision classification and detection tasks [27].

Different architectural designs can have a large impact on how well a CNN performs at a given task. For example, deep networks with a large number of convolution and pooling layers have been found to perform better [43] than shallow ones. This is exploited by several CNN models such as the ResNet architectures which can have more than 100 layers [44]. Theoretically, depth increases the representational capacity of the network [45]. Representational capacity is also increased by the use of ReLUs on top of convolution layers. In order to map feature representations into a vector of class predictions, usually the final layer of a CNN contains similar fully-connected layer as feed-forward networks.

Modern deep CNNs can easily have hundreds of millions of parameters and have recently started to match human performance on many image classification tasks [17]. Owing to their large representational capacity, these models can also easily overfit training data. The development of appropriate regularization techniques has thus played a large role in CNNs' success. Perhaps the most popular regularization technique, and the one applied in this work, is *dropout* [46]. At every training iteration, dropout deactivates each neuron with a user-specified probability. As a consequence, training is essentially done over exponentially many different sub-models, and the final model is their average [4]. This method regularizes the network as individual neurons cannot rely too much on other neurons. At test-time, dropout is turned off.

Many of the theoretical concepts of modern deep CNNs have been around for several decades [27]. The reason for their recent surge in popularity is to a large extent that modern computer resources have made them a lot easier and faster to train. In particular, the ability to train deep models on Graphical Processing Units (GPUs) has cut training times by at least an order of magnitude [17]. However, we have only been able reap the benefits

of faster compute and better models because of larger datasets. It has been suggested that CNNs with around 5000 labels per training category can on average start to equal human performance on image classification [4]. Over the past decade we have seen the advent of datasets that have up to tens of millions images on tens of thousands training classes [?], [1], [47].

2.2 Exploiting Semantic Information in Computer Vision

2.2.1 Semantic Hierarchies and the Human Visual System

Upon observing almost any visual scene, humans are able to effortlessly cut it up into segments of distinct objects [24], even if we have never seen the scene before. This is a remarkable ability considering that the world we live in can present us with an infinite amount of visual variation to our retinas. Yet we are able to easily recognize tens of thousands of distinct object categories [48] with invariance to various factors such as movement, lighting, shade, orientation and partial occlusion [49]. Rosch *et al.* [24] argue that this is because of the non-random, and hierarchical, structure of visual objects, for instance no breed of dogs has wings but they will often share many features such as four legs, snout and a tail among each other. In general entities that are closer to each other in a hierarchical semantic taxonomy will typically share more visual features: dogs and birds are still more alike than dogs and vehicles since both belong to a higher level category of animals which all share certain visual attributes. These observations are also corroborated by the importance of physical features in the organization of hierarchical biological taxa [13].

Standard machine learning and deep learning models used for image classification do not fully exploit this semantic taxonomy since they are trained

against one-hot encoded target vectors. For example, the 1K ImageNet data set [1] will have labels as 1000-dimensional vectors where there is a 1 for the dimension indicating the correct class and 0s elsewhere. It follows that label vectors for different classes are orthogonal, and a model trained on them will not account for any classes belonging to a common category at a higher level of a taxonomy. For example, in reality wolves and dogs fall under the same higher level categories of canines, animals, mammals and so forth - all of these share certain amount of visual features. The chosen level of hierarchy of a one-hot encoding is also usually arbitrary; should an image be labelled a dog or a Labrador retriever? Optimally both would be taken into account, we argue. The number of machine learning algorithms that do take this into account is limited, however; below we give an overview of them.

2.2.2 Semantic Hierarchies for Image Classification

The use of semantic hierarchies in computer vision dates back to early work on image retrieval [50], [51], [52] [53], where they have been used to expand potential query terms and to impose a more general association between image features and corresponding labels. These ideas were subsequently utilized to improve image annotation and classification [54], [55] [56]. For example, Marszalek and Schmid [55] used the WordNet database [57] to create a lexical hierarchy of their image labels and trained a separate binary SVM classifier [58] for each node of the hierarchy. The resulting model proves flexible and performs well under uncertainty - if the model is unsure about which dog breed is in an image, it can then move up one level and just predict a 'dog'. The dataset used by the authors is very simple by today's standards, however, and would be difficult to scale up. More recently, Redmon and Farhadi [59] employed a similar idea but in the context of deep learning. In particular, the label for each image was taken to be the full path from the root node of the WordNet tree to the original label. The model was then trained with multiple softmax functions, one for each level of hierarchy, in order produce

a chain of conditional probabilities from the root node to leaves. The authors' purpose for building this model was to learn simultaneously from two datasets of unequal hierarchies and to jointly optimize classification and detection on the two datasets. However, other possible benefits, like zero-shot learning, were not explored.

In their research on human categorization, Rosch *et al.* [24] defined the concept of 'basic level'. This is described as the most fundamental level in the semantic hierarchy of visual objects; it is the level that children usually learn first, and humans tend to identify objects most readily at this level. The level below the basic level is called 'subordinate level'. An example of this structure would be the basic level label of 'fox' and its subordinate 'arctic fox'. The importance of the basic level seems to stem from this level having the highest ratio of visual variation between categories to variation within categories [24], [14]. Therefore, it has been especially practical for humans to identify and label objects at this level [14].

Hillel and Weinshall [60] mimicked the idea of basic and subordinate levels by building a two-stage classifier where the first stage generates a vector representing the different parts of an objects and the second stage classifies instances based on this vector. On average, this strategy outperformed a traditional one-step algorithm. Wang and Cottrell [18] took these ideas to the deep learning-era, and trained a CNN on both, basic and subordinate, labels of the ImageNet 2012 data [61]. They found that this increased the standard top-5 classification accuracy. Another similar study [62] explored a dataset which mostly had coarse labelled images (akin to the basic level) and only some fine grained examples. The authors showed that training at both of the two levels helped to predict the fine-level classes. This suggests that the fine-grained model is able to borrow strength from the coarse labels and subsequently generalize better. Peterson *et al.* [19] extended this line of research by exploring the type of representations learned by a CNN trained on both 'basic' and 'sub-ordinate' labels. First, the authors found that including

basic-level labels in pre-training led to a much more clustered representation of the CNN features, e.g. the feature vectors of different breeds of dogs were now bundled up together. They also illustrated the generalization power of these representations by running a few-shot learning experiment in which the model was shown only a handful of examples of either sub- or basic-level objects. Zero-shot learning was not considered however. All of the above works have only considered hierarchies of few levels; we are not aware of any deep learning papers that would have explored the type of representations and the consequent generalization properties that would result from using a more complex hierarchy of labels.

2.2.3 Semantic Embeddings and Zero-shot Learning

Above studies illustrate the implicit connection between semantic and visual data processing in humans. Joliceur *et al.* [14] explored this in depth and illustrated these capabilities with several experiments. For instance, upon viewing a picture of a chair, we can use our semantic knowledge-base to generalize it to the category of furniture even though the image itself doesn't reveal that such a category even exists. Similarly, we can seamlessly access our visual memory to imagine concepts, even ones we may have never seen such as a 'cat wearing ice skates'. If we encounter an unknown dog breed, we are conscious of the uncertainty and will not try to claim that it is a Labrador retriever, instead we might say it's some sort of a dog that is similar to a Labrador. Or if we are completely unsure, we can just say that it's a dog. We are thus able to perform robust inference on previously unseen classes of objects.

Zero-shot learning (ZSL) [11] is concerned with making predictions about previously unseen classes of images. Formally, a classifier g is trained on features X_{TR} and training labels Y_{TR} such that $g : X_{TR} \rightarrow Y_{TR}$. ZSL requires that $Y_{TR} \cap Y_{ZSL} = \emptyset$ where Y_{ZSL} represents the test labels of any ZSL dataset; training and test labels are completely disjoint. A fundamental idea in this

area is to exploit the semantic relatedness of visually linked categories [63], [64], [11] by projecting input data into a lower dimensional semantic embedding space. This approach can be represented by the following equations [11] (see Figure 2.5 for a graphical explanation):

$$\mathcal{S} : X^d \rightarrow F^P \quad (2.9)$$

$$\mathcal{L} : F^P \rightarrow Y \quad (2.10)$$

where X^d is the d -dimensional input space, F^P a P -dimensional embedding space, Y the class-label space, and \mathcal{L} is a look-up table between embedding vectors and their class labels.

Further, assume that we have all the pairs $\{\mathbf{f}, y\}_{1:M} \in F^P \times Y$ for both the training and zero-shot classes, total M classes. In the training-phase of zero-shot learning, there are N training pairs $\{\mathbf{x}, y\}_{1:N}$, where $N \ll M$. Since each y has a corresponding embedding vector \mathbf{f} , a classifier $g(\cdot)$ can be trained on $\{\mathbf{x}, \mathbf{f}\}_{1:N}$ to approximate the mapping \mathcal{S} . After training, zero-shot prediction can be done by first mapping all the test images $\{\mathbf{x}\}_{(N+1):M}$ into the embedding space:

$$\hat{\mathbf{f}}_i = g(\mathbf{x}_i) \quad \forall i \in (M+1) : N \quad (2.11)$$

Next, a nearest neighbour classifier nn can be utilized to assign a ground-truth embedding to the predicted embeddings: $nn : \hat{\mathbf{f}} \rightarrow \mathbf{f}$ and subsequently \mathcal{L} is used to get the predicted class labels. To be clear, the term 'zero-shot' refers to the previously unseen $M - N$ classes of test images, however, in our semantic knowledge-base we do know about the existence of those classes and their names. The challenge is thus in how to construct the semantic embedding space and what type of classifier to use to approximate \mathcal{S} ; the following paragraphs will cover the existing approaches.

In an early work, Socher *et al.* [20] created word embedding vectors for all seen and unseen class labels from the Wikipedia text data using a

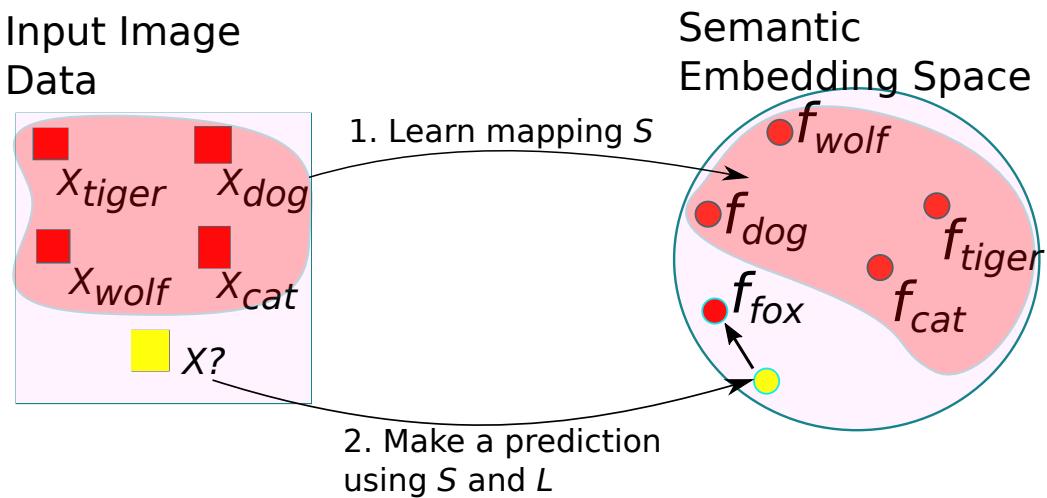


Figure 2.5: A graphical illustration of zero-shot learning. The left square contains all the input images. Red ones are training classes which are used to learn an approximation of the mapping $\mathcal{S} : X^d \rightarrow F^P$. The circle represents the semantic embedding space. Note that embedding vectors for zero-shot classes can be accessed as well. Upon seeing an image of unknown class (yellow square) the model outputs embedding vector prediction (yellow dot) and then the nearest plausible class (here fox) is found with a nearest neighbour function \mathcal{L} . Novel image is hence classified as a fox even though the model has never seen one before

methodology [21] that places words that occur in a similar context next to each other in the embedding space. A two-layer feed-forward network was used to learn a mapping from image features into the embedding space, and outlier detection was used to determine if the predicted embedding was one of the known or unknown classes. Only 8 known classes and 2 unknown classes were considered however. To address this and other limitations, Frome *et al.* [2] developed the Deep Visual-Semantic Embedding Model (DeViSE), which was trained on 1000 classes of the ImageNet dataset [1] and its zero-shot recognition abilities were tested on the remaining 21,000 ImageNet classes. The architecture of this model forms the foundation for the model we build in this paper: it takes a pre-trained deep learning model, AlexNet [9], removes the softmax layer and replaces it with a projection layer that attempts to predict the embedding vector of each image label. The word embeddings were extracted with another similarity-based model, word2vec, trained on 5.4 billion words of text from Wikipedia [23], [22].

The authors of DeViSE worried, however, that the model was overfitting the projection from the image space into the semantic embedding space. To overcome this they introduced another model, ConSE [3]. In this model, an image of unknown class is first passed into a pre-trained CNN. The CNN will try to classify the image as one of the known classes and outputs corresponding label probabilities, $p_0(y|\mathbf{x})$ such that $\sum_{y=1}^{n_{seen}} p_0(y|\mathbf{x}) = 1$. The predicted semantic embeddings for the zero-shot images are computed as a convex combination of the known classes' embedding vectors, weighted by the predicted probabilities from the previous step. More formally [3]:

$$f(\mathbf{x}) = \frac{1}{Z} \sum_{t=1}^T p(\hat{y}(\mathbf{x}, t)|\mathbf{x}) \cdot \mathcal{G}(\hat{y}(\mathbf{x}, t)) \quad (2.12)$$

where $\hat{y}(\mathbf{x}, t)$ is the t^{th} most likely class-label for image \mathbf{x} out of the known labels, with $p()$ giving the respective predicted probabilities. \mathcal{G} denotes a look-up table that maps labels to their embedding vectors. Z is a normalizing

constant, and T is user-defined. This model was shown to generalize slightly better to zero-shot classes than DeViSE. One contribution of this thesis is to show how the ConSE model can be adapted into hyperbolic space.

A potential problem with the above models, which use Wikipedia text data to create semantic embeddings, is that vectors for low occurrence labels may be imprecise. Li *et al.* [65] attempted to resolve this by exploiting the full WordNet semantic hierarchy [57]. To do this, they re-constructed the semantic embedding for each word by tracing its path to the root of the WordNet tree and calculated an inverse distance-weighted combination of the Wikipedia-based semantic embeddings along the path. For example, the embedding vector for 'arctic fox' was constructed as a combination of the word2vec semantic embeddings of the set ('arctic fox', 'fox', 'canine', 'carnivore' ... 'living being'). Despite the attempt to capture the semantic structure of the WordNet, the embeddings at each node are still based on the Wikipedia data, which does not account for this taxonomic structure. Further, each label only considers its own parent super-ordinate nodes. Consequently the resulting embedding manifold as a whole is sparse and hence unlikely to accurately reflect the full WordNet graph and its complex relations. In the next section we show how this has recently been achieved with hyperbolic embeddings.

Fu *et al.* [66] identified a central issue that maligns most ZSL approaches, namely the projection domain shift: the mapping that is learned from the training image space into the embedding space may not transfer well to the disjoint set of zero-shot classes. Transductive multi-view embedding was proposed as a solution by the authors: in addition to the semantic embeddings of the seen classes, the embeddings of unseen classes are also available and can hence be used at training time. In fact, multiple embeddings for each label were created and they were all jointly projected to common embedding space, which alleviated the domain shift problem. Kodirov *et al.* [67] proposed a similar approach: sparse coding was used to map from the unseen labels to

the semantic space. This helped to guide the domain projection from seen to unseen visual features. See also Rohrbach *et al.* for similar ideas [68]. The use of test labels during training may not be practical, however, and as an alternative Kodirov *et al.* [69] proposed a semantic autoencoder. The general idea is to learn one mapping (decoder) from input images to a latent semantic embedding space and another mapping (decoder) from there back to reconstruct the image data. The addition of the decoder constraints the model to solve the domain shift. Tsai *et al.* [?] used two autoencoders, one on the image data and another on text data, extracted the representation layer from each and minimized their mismatch.

Several other approaches have been suggested for learning the mapping from image data into the embedding space. For example, Fu and Sigal [70] used embeddings built on full semantic vocabularies, rather than just training labels, and showed that a better mapping was learned. Shigeto *et al.* [71], on the other hand, showed that instead of learning a mapping from image data into the label space, it can be beneficial to go in the opposite direction since nearest neighbour search is often easier in the lower dimensional image space. Changpinyo *et al.* [72] also based their models around this idea, with impressive results. Finding nearest neighbours in high dimensional spaces is a common challenge with Euclidean spaces, which we manage to avoid by working in low-dimensional hyperbolic geometry instead.

Another branch of works has focused on learning better embeddings. While most approaches learn a semantic embedding and the mapping function separately, Ba *et al.* [73] trained a feed-forward network from text input data to predict the visual features in the different layers of a CNN. Together these networks defined a joint-embedding, which was shown to capture fine-grained visual features more accurately. Ji *et al.* [74] showed how performance can be improved by ensuring that the embedding manifold preserves the local structure of the visual input space. Local similarity aware deep embeddings were explored by Huang *et al.* [75]. Zhang and Saligrama [76],

on the other hand, built semantic embedding labels for zero-shot classes as mixture distributions of the seen classes’ embedding vectors. In a follow up work the authors [77] provided a latent probabilistic framework; that is, zero-shot recognition is performed by estimating the posterior probability that an image matches one of the label vectors.

Graph-based approaches have also been popular in tackling ZSL recently [78, 79, 80]. Most notably, the current state-of-the-art [80] exploits both, explicit taxonomic relationships between categories from the WordNet graph, as well as similarity based semantic embeddings constructed from the Wikipedia text data [81]. More specifically, each node in the WordNet graph is replaced by its respective semantic embedding vector. A *graph convolutional network*, which performs convolutions between the nodes, is used to predict logistic regressors that are used for zero-shot recognition. The results of the this current state-of-the-art approach highlight the difficulty of the ZSL task: they achieve around 2 % top-1 accuracy and 12 % top-20 accuracy on predicting labels for 21k zero-shot classes. The approach taken by the authors is similar to ours in that they also try to explicitly represent semantic hierarchies, as well as similarities, however in our work both are captured by the embedding space. Further, our approach has the advantage that it can be combined with standard deep learning models, which is important given their ubiquity.

In addition to above, there is a vast array of approaches that focus on zero-shot learning via class attributes [82], [83], [84], [85], [86], [87], [88], [89]. A limitation of these approaches is that when the number of classes grows very big, it becomes harder to manage the space of attributes manually. The set of attributes is also hard to transfer into a new context [?] with different attributes. Therefore our work focuses solely on text and language based semantic embeddings. More importantly, attribute based ZSL often aims to predict fine-grained classes that are visually very similar [87] [90]. We on the other hand aim to build a framework that makes semantically robust zero-shot classifications, such as predicting a higher-level category when the

model is uncertain.

Finally, there exist various studies that combine language and attribute based semantic embeddings [91], [66] [92]. Akata *et al.* [91], for instance, combine three embeddings, one learned from the Wikipedia dataset, one from physical attributes of animals, and a hierarchical embedding based on the structure of the WordNet. The model performs well on zero-shot prediction of fine-grained classes on a relatively small data set; it's generalization properties on a large dataset such as the 21k label ImageNet is not explored, however. A more comprehensive review of further ZSL approaches and trends can be found in [93].

2.3 Semantic Embeddings in Hyperbolic Space

Section overview: As alluded to in the previous section, hyperbolic spaces provide an efficient way of embedding hierarchical graph structures. In this chapter we provide a brief overview of the required theoretical background and discuss the recent literature on building word embeddings in this space. Hyperbolic geometry and hyperbolic spaces are a vast fields containing thousands of years of work and formal statements and proofs are thus out of the scope of this work; we merely aim to provide the background necessary for the present work. For more thorough treatment of the topic, we refer the reader to Cannon *et al.* [] and Greenberg [25]; unless where otherwise stated, all the theories and intuitions provided below are based on these two books.

2.3.1 Basics of Hyperbolic Geometry

Hyperbolic geometry is a type of non-Euclidean geometry, derived by altering the fifth postulate of the classic Euclidean geometry [25]:

1. Any pair of points can be joined up by strictly one straight line segment
2. All lines can be extended indefinitely

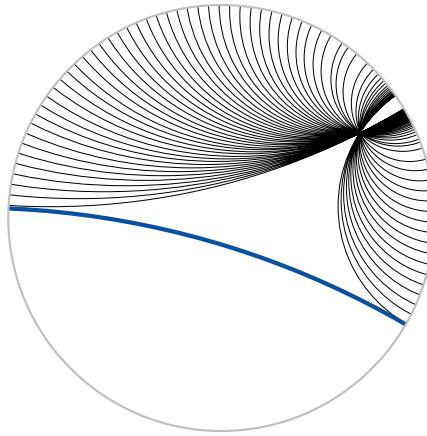


Figure 2.6: Parallel lines in Hyperbolic space (Poincaré disk model). The thick blue line is parallel to all the other lines since it never intersects any of them in the hyperbolic space (interior of the circle). All the lines are arcs of circles of different sizes that fall in the interior - these are known as geodesics. Image used under Creative Commons license, released into public domain by Trevorgoodchild of English Wikipedia

3. Exactly one circle of a specific circle and specific centre exists
4. All right angles are identical
5. Parallel postulate: Consider line L and a point P which does *not* lie on L. Then there exists only one other line M that passes through P and is parallel to L, no matter how much we extend the two lines.

In hyperbolic geometry the fifth postulate becomes its negation:

5. * Consider line L and a point P which does *not* lie on L. For any L, there are infinitely many lines that are both parallel to L and pass through P.

For us who are accustomed to thinking in Euclidean terms, this may seem impossible, at first. This becomes possible when we consider spaces

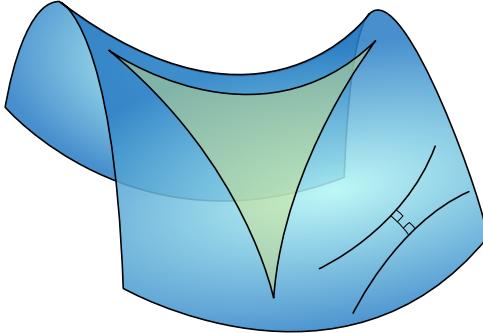


Figure 2.7: Saddle point in Euclidean space allows a representation of a hyperbolic triangle, for which the sum of its angles is less than 180 degrees. Image used under Creative Commons license from Wikimedia Commons.

of constant negative Gaussian curvature, which indeed is the fundamental property of hyperbolic spaces (Figure 2.6). For instance a hyperbolic plane is a 2-dimensional manifold with a constant negative curvature; similar logic holds in higher dimensions too [94]. As long as lines in hyperbolic space never intersect each other, they are parallel. Straight lines in hyperbolic space are more precisely known as geodesics - arcs of circle, and locally they define the shortest path between two points in the curved space [94]. It is illustrative to consider the shortest path between two locations on the globe drawn on a flat map, say new york and london; this too is an arc of the great circle of the earth's circumference, rather than a straight line (though in this instance due to a positive curvature). For an improved intuition of hyperbolic geometry, it should be noted that all points in a hyperbolic space can be viewed as saddle points of Euclidean geometry (Figure 2.7). There are several ways of constructing hyperbolic spaces from Euclidean geometry; the only requirement is that aforementioned necessary axioms are satisfied. Below we define the most important models, following [25].

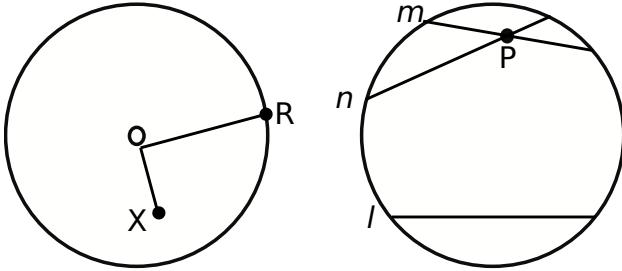


Figure 2.8: Depiction of the Klein model of hyperbolic geometry. m , and n are parallel lines to l as they never intersect it.

The Beltrami-Klein Model

Consider a circle γ of the Euclidean plane, its centre O , and radius OR (see Figure 2.8). The hyperbolic space is then defined as the interior of all points X for which $OX < OR$ [25]. Lines in the Klein model of hyperbolic space are open chords of γ , with open referring to a chord without its end points. This model satisfies the requirement of hyperbolic space since if there is a line l and a point *not* on it, P , then there are infinite number of parallel lines of l that go through P . An example is given in Figure 2.8. The reason why m and n are parallel to l in this figure is that they are open chords in γ and at no point of this space do they intersect l , which is the working definition of parallelism [25]. It is crucial to notice that only the interior of the circle forms the hyperbolic space and the boundary can never be reached, hence the *open* chords as lines. This, combined with the requirement that lines can be extended indefinitely, means that one can get arbitrarily close to the boundary of the circle but never reach it. As a consequence, the distance between two points on the line grows exponentially the further away from the centre we get and therefore we cannot use Euclidean distance (which is not visibly evident from the figures) [25]. Notice also that the appearance of lines as 'straight' is just a matter of mapping them from geodesics [25], such as those in Figure 2.6.

We can extend this model to any higher dimension n by re-defining the

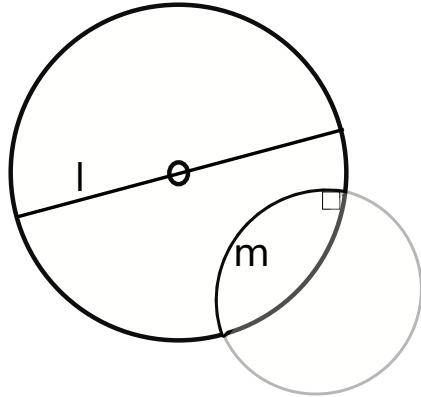


Figure 2.9: Depiction of the Poincare Disk model of hyperbolic geometry. Line l represents an open chord passing through the centre of the circle, whilst line m is an open arc.

disk as $\mathcal{K}^n = \{\mathbf{x} \in \mathbb{R}^n | \|\mathbf{x}\| < 1\}$, and taking all other axioms from above. Distance between two points \mathbf{p} and \mathbf{q} is defined as [95]:

$$\mathcal{D}_K(\mathbf{p}, \mathbf{q}) = \operatorname{arcosh} \left(\frac{1 - \mathbf{p} \cdot \mathbf{q}}{\sqrt{1 - \mathbf{p} \cdot \mathbf{p}} \sqrt{1 - \mathbf{q} \cdot \mathbf{q}}} \right) \quad (2.13)$$

The Poincare Model

In the Klein model arcs were mapped to straight lines. An alternative representation of hyperbolic space is the Poincare model. Here two types of lines exist: open chords that pass through the origin of the circle, which are like those in the Klein model, and arcs of all the circles orthogonal to γ i.e. *open* arcs (Figure 2.9) [25]. The other statements about Klein model transfer here; the boundary can never be reached and distances grow exponentially as we approach it. Again an infinite number of parallel lines exist. This is visualized in Figure 2.6.

To extend this formally into n higher dimensions we can define the hyperbolic space as an open unit ball \mathcal{B}^n as we did for the Klein model $\mathcal{B}^n = \{\mathbf{x} \in \mathbb{R}^n | \|\mathbf{x}\| < 1\}$, but as straight lines are arcs, we have a different distance

function [95]:

$$\mathcal{D}_P(\mathbf{p}, \mathbf{q}) = \operatorname{arccosh} \left(1 + 2 \frac{\|\mathbf{p} - \mathbf{q}\|^2}{(1 - \|\mathbf{p}\|^2)(1 - \|\mathbf{q}\|^2)} \right) \quad (2.14)$$

Whilst the Klein and Poincare models appear different, they both obey all the same axioms of hyperbolic spaces, in fact, the models are *isomorphic*: there is a one-to-one correspondence between all the points P and P' of the two models, as well as one-to-one correspondence between lines l and l' , such that ' P lies on l if and only if P' lies on l'' ' [25]. It can be proven that all models of hyperbolic geometry are isomorphic [25]. Therefore it is possible map from one model to another. We will employ the mapping from Poincare ball into the Klein model at a later stage and thus present the required equation here [25]:

$$\mathbf{s} = \frac{2\mathbf{u}}{1 + \|\mathbf{u}\|^2} \quad (2.15)$$

where \mathbf{s} is the point in the coordinates of the Klein model and \mathbf{u} in the Poincare model.

The Hyperboloid Model

This model stems from the theory of special relativity. If we denote x, y the 2-dimensional coordinates and t as time, then distances are measured by the Minkowski metric [25]:

$$ds^2 = dx^2 + dy^2 - dt^2$$

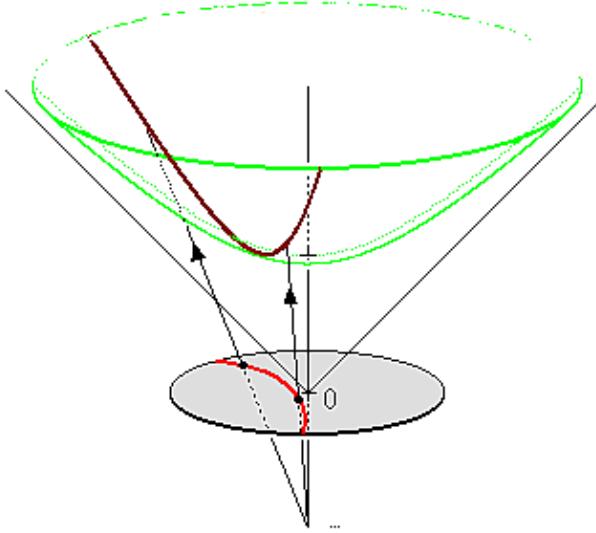


Figure 2.10: Upper sheet of the hyperboloid model in \mathbb{R}^3 and a projection of one of its geodesics on to the Poincare disk. While not possible to visually illustrate, the brim of the 'bowl' is at infinity and can never be reached. Image used under Creative Commons license from Wikimedia Commons.

The surface equation that corresponds to this metric is given by the equation below. See [25] for details.

$$x^2 + y^2 - t^2 = -1$$

In Euclidean space, this forms a hyperboloid object with two sheets [25], with the upper sheet shown in Figure 2.10. In that figure we also show how the model relates to the Poincare disk we saw above. The 'bowl' of the hyperboloid is infinite so the 'brim' can never be reached, thus satisfying the indefinite extensibility of lines, which are geodesics on the hyperboloid surface. In higher dimensions, the n dimensional hyperbolic space is defined as $\mathcal{H}^n = \{\mathbf{x} \in \mathbb{R}^{n+1} | \mathbf{x} * \mathbf{x} = -1\}$ where $\mathbf{x} * \mathbf{x} = \sum_{i=1}^n x_i x_i - x_0 x_0$ is the

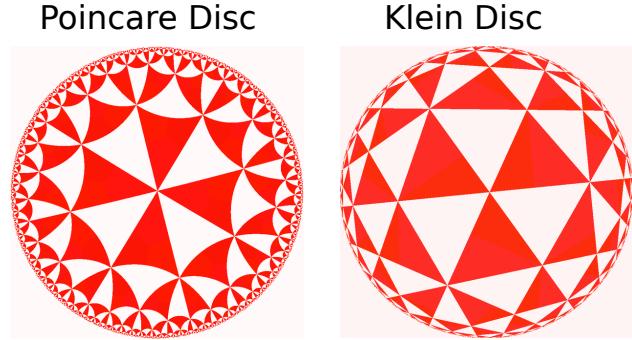


Figure 2.11: Illustrations of Klein and Poincare Disks with hyperbolic tiling. All the tiles have equal area so this visually depicts how the area of the circles grows exponentially towards the boundary. Images used under Creative Commons license from Wikimedia Commons.

Lorenzian inner product [96] [?]. The distance function is:

$$\mathcal{D}_{\mathcal{H}}(\mathbf{p}, \mathbf{q}) = \text{arcosh}(-\mathbf{p} * \mathbf{q}) \quad (2.16)$$

2.3.2 Semantic Embeddings in Hyperbolic Spaces

Most popular word embedding methods take into account the local context of words, e.g. the *word2vec* models [22, 23], or the global context like the *Glove* model [81], and produce vector representations so that semantically similar words are embedded near each other in the resulting vector space. The ability of these embeddings to capture complex relationships is constrained, however, as the required dimensionality becomes easily exceedingly large [15, 16, 96]. This limit on representational capacity is particularly relevant to our project as we aim to build an embedding space that accurately captures the hierarchical graph structure of word semantics. The idea of using hyperbolic spaces for representing semantic hierarchies more accurately was explored early on by [97], [98] and [99], though this was in the context of Self-Organizing Maps rather than word embeddings. More recently, Nickel

and Kiela [15], and Chamberlain *et al.* [16] have successfully demonstrated the advantages of hyperbolic word embeddings over Euclidean ones in terms of generalization performance and computational cost for representing hierarchical data. Additionally, Gulcehre *et al.* [100] have achieved success on various tasks, such as neural machine translation, by imposing hyperbolic geometry on the activation functions of their neural network model.

To understand the benefits of hyperbolic word embeddings, consider the Poincare and Klein disk models and contrast this to an Euclidean disk. The area of an Euclidean disk increases quadratically, since $A_E = \pi r^2$, whereas the negative curvature of hyperbolic space allows the hyperbolic disk area to grow exponentially with r as given by $A_H = 2\pi (\cosh(r) - 1)$ (assuming constant curvature of 1) where \cosh is the hyperbolic cosine function $\cosh(x) = \frac{e^{2x}+1}{2e^x}$ [25]. Intuitively the benefit of this is that in hyperbolic spaces there is much more room for embedding different concepts as we move outwards from the centre (see Figure 2.11). Further, Hyperbolic spaces can be interpreted as continuous space extensions of trees and are thus particularly efficient for embedding hierarchical structure in a way that preserves the associated graph structure, even for infinite trees [16]; like hyperbolic space, the number of leaves of a word-tree also grows exponentially as we increase the levels of hierarchy (assuming the number of splits per level remains constant) [15] (See Figure 2.12 for a naive example).

The hyperbolic embedding that we used is the Poincare Embedding model of Nickel and Kiela [15], mainly because they showed its effectiveness on the WordNet lexical hierarchy, which is the basis for the labels in the ImageNet dataset that we used. Here we give a brief overview of the way in which the Poincare embeddings are actually computed in this model; rest of this section is thus entirely based on [15]. To use the original notation of the paper, let $\mathcal{S} = \{x_i\}_{i=1}^n$ denote the set of words, and $\Theta = \{\boldsymbol{\theta}_i\}_{i=1}^n$ be their respective locations in the d dimensional Poincare embedding space \mathcal{P}^d i.e. $\boldsymbol{\theta}_i \in \mathcal{P}^d$. To

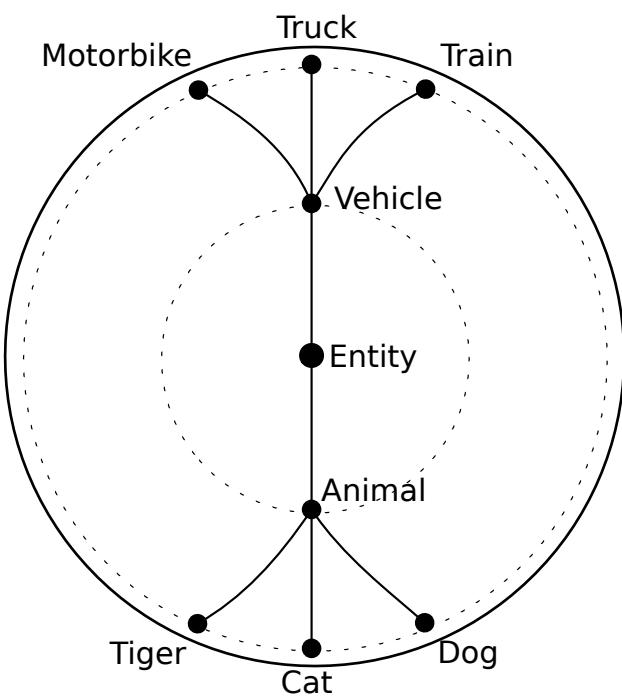


Figure 2.12: Naive example of embedding a hierarchical semantic taxonomy in a hyperbolic disk. All the concepts at each level of hierarchy are equidistant from the centre of the disk as portrayed by the dotted circles.

find embeddings, the authors set to solve the following optimization problem:

$$\Theta' \leftarrow \arg \min_{\Theta} \mathcal{L}(\Theta) \text{ s.t. } \forall \boldsymbol{\theta}_i \in \Theta : \|\boldsymbol{\theta}_i\| < 1$$

where the final inequality is needed to ensure the vectors fall inside the open unit ball. And their loss function is a ranking loss:

$$\mathcal{L}(\Theta) = \sum_{(u,v)} \log \frac{\exp\{-d(\mathbf{u}, \mathbf{v})\}}{\sum_{\mathbf{v}' \in N(u)} \exp\{-d(\mathbf{u}, \mathbf{v}')\}}$$

where the tuple (u, v) represents the nouns that are connected to each other in the WordNet graph, \mathbf{u} and \mathbf{v} are their estimated embedding vectors, and $N(u)$ those nouns v that are *not* connected to a given u . Finally, $d(\mathbf{u}, \mathbf{v})$ is the hyperbolic distance between the nouns. Since this embedding is based on the Poincare model, the relevant distance function is the one in Equation 2.14. A point of emphasis is that the model is not given any explicit hierarchical information but can learn it from data on which nouns are connected in the WordNet. It can therefore learn latent hierarchies.

The optimization problem above could be solved directly using full Riemannian optimization, but the authors show that the Riemannian gradients can also be obtained in terms of Euclidean gradients ∇_E , resulting in the following Riemannian SGD updates:

$$\boldsymbol{\theta}_{t+1} = \text{proj} \left(\boldsymbol{\theta}_t - \eta_t \frac{(1 - \|\boldsymbol{\theta}_t\|^2)^2}{4} \nabla_E \right)$$

where η_t is the learning rate, and the authors used

$$\text{proj}(\boldsymbol{\theta}) = \boldsymbol{\theta} / \|\boldsymbol{\theta}\| - 10^{-5} \text{ if } \boldsymbol{\theta} \geq 1; \text{ otherwise } \boldsymbol{\theta} \quad (2.17)$$

to ensure the updates remain with the open unit ball Poincare space. ∇_E can be derived using the backpropagation algorithm introduced in Section 2.1.1. This model achieved superior performance in comparison to an Euclidean

baseline in various experiments that tested for the models’ ability to preserve original graph structures [15].

2.4 Summary of Background

In this section we have covered the necessary technical background and the most relevant literature that inspired this thesis. Our literature review can be summarized as follow. First, deep learning has achieved unparalleled success in almost all computer vision tasks but in their standard form they are unable to generalize to previously unseen classes. Zero-shot learning tries to address this shortcoming, often by borrowing strength from different forms of semantic embeddings. This is motivated by the observation that semantic similarity often corresponds to visual similarity. Unlike similarity based embeddings, the human brain categorizes visual objects in a much richer representation of semantic *hierarchies* that allows robust inference on novel objects. We look to replicate this behaviour in the zero-shot setting. Whilst some hierarchical models have been introduced in this area previously, to our knowledge, no work has used the recent hyperbolic embeddings to explore their potential benefits over the standard Euclidean word embeddings.

Chapter 3

Our Approach

3.1 Approach Overview

We propose two model architectures: Deep Hyperbolic Semantic Embedding (Deep-HSEM) model and Convex Hyperbolic Semantic Embedding (Convex-HSEM) model. The general intuition is to replace the standard one-hot-encoded targets used in deep learning with Poincare embedding vectors, which should better reflect the location of each label in a hierarchical semantic taxonomy. By training against these embedding vectors, we look to exploit the fact that many objects of initially different categories share common visual attributes through shared parent categories. After learning a mapping from images into the hyperbolic embedding space, we perform zero-shot recognition by feeding previously unseen classes of images through the trained model and then search for the nearest label in the embedding space. We hypothesize that the hyperbolic embedding space, which is based on a semantic hierarchy, will reflect visual similarities better than the standard Euclidean similarity based word embeddings. To explore this, we also perform the zero-shot experiments with a version of our model that uses the Glove embedding [81] in place of the hyperbolic embedding. In particular, we test whether our hyperbolic model is able to make more robust predictions;

that is, even if our model is unable to predict the exact label of an image, it should be able to produce a better conjecture about what broader category the object might belong in.

3.2 Data & Model Details

3.2.1 Training Data

We used the lexical database WordNet [57] as our source of hierarchical semantic information. As a whole, the WordNet consists of 117,000 synsets (synonym sets; words in the same synset share a meaning), which are connected to each other in a graph based on their super- and subordinate relations. These relations are transitive so that if A is a subordinate of B and B is subordinate of C, then A is also subordinate of C. Several different *types* of words are encoded in the WordNet, such as verbs and nouns. In this work only nouns were considered. Overall, our final list, that was used to create the hyperbolic embedding, consisted of 82,000 of commonly used nouns. The total number of nouns far exceeds the number of image classes – this was done intentionally so that the embedding space is as all-encompassing as possible, and our model could easily be extended to other datasets in future.

Specifically, for training images, we used the ImageNet 2012 Large Scale Visual Recognition Challenge (ILSVRC) 1000 class dataset [61, 1]. From here on we refer to this as the '2012 1k ImageNet' data. The 1000 classes of the dataset provide a total of 1,281,167 images, with 730-1300 images per class. All of the images come in the RGB format and are 224×224 pixels. A large proportion of the image classes are fine-grained. For instance, there is no 'dog' category but instead around 90 classes of specific dog breeds [61]. Importantly, all the image labels are based on the WordNet hierarchy which made it easy to assign ground-truth embedding vectors; there is an exact mapping from each image class to a specific noun [61].

3.2.2 Poincare Embedding Construction

In order to learn a mapping between images and their locations in the hyperbolic embedding space, we first had to acquire ground-truth hyperbolic embedding vector for each image label. The specific hyperbolic embedding model we utilized is the Poincare embedding model of Nickel and Kiela [15], which was described in detail in Section 2.3.2. These authors also provided an open-source implementation of the model¹, which we used to construct Poincare embeddings for all the 82,000 nouns of the WordNet dataset. The dimensionality of the embedding vectors was chosen to be 10 for this project. This is much lower than the number of dimensions usually used by Euclidean word embeddings (50 to 500) and thus allowed us to demonstrate the relative efficacy of hyperbolic embeddings. To get our final label embeddings, we trained the Poincare model with the default hyperparameters for 1500 epochs (4 days run-time). An important detail is that the model was only provided with information about which nouns are connected to each other but no explicit hierarchical subordinate relations were given, rather the embedding self-organizes to learn the latent hierarchy.

3.2.3 Deep Hyperbolic Semantic Embedding Model (Deep-HSEM)

Model Architecture

A typically convolutional neural network takes in an image and outputs predicted probabilities for the different possible class labels. Our model is similar with the crucial exception, that for each image, the Deep-HSEM predicts a single vector that corresponds to a point in the Poincare embedding space. More precisely, we have taken a well known CNN architecture, VGG-16 [101], and replaced its final output layer with a 10-d fully connected layer that cor-

¹ Freely available under the Creative Commons license at <https://github.com/facebookresearch/poincare-embeddings>

responds to the dimensionality of the Poincare embedding space. We chose the VGG-16 model since it provides near state-of-the-art performance on the ImageNet 2012 1K data (top-1 accuracy 76.3%, top-5 accuracy 93.8%) but is simpler and less costly to train than many of the more recent CNNs. The architecture of the VGG-16 consists of 16 layers: 13 3x3 convolution layers followed by three fully-connected layers. Additionally, the network uses 2x2 max-pooling, ReLU non-linearities and dropout for regularization. We illustrate the Deep-HSEM model in Figure 3.1. The resulting architecture is similar to the DeViSE model of Frome *et al.* [2] (reviewed in Section 2.2.3); the most significant difference is that we used a hyperbolic embedding rather than the Euclidean word2vec embedding.

Formally, our aim is to learn to approximate the mapping \mathcal{S} from 224×224 RGB image space $X^{224 \times 224 \times 3}$ into the 10-dimensional Poincare embedding $F_{\mathcal{P}}^{10}$.

$$\mathcal{S} : X^{224 \times 224 \times 3} \rightarrow F_{\mathcal{P}}^{10}$$

In order to learn this mapping, the Deep-HSEM, denote it $h_{deep}(\mathbf{x}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes the model weights, is given M training pairs of input images and corresponding ground-truth embeddings $\{\mathbf{x}_i, \mathbf{f}_i\}_{i=1:M} \in X^{224 \times 224 \times 3} \times F_{\mathcal{P}}^{10}$ where \mathbf{f}_i is the embedding vector of its class $y_i \in Y_{train}$. The model takes an input image \mathbf{x} and predicts an embedding $\hat{\mathbf{f}}$:

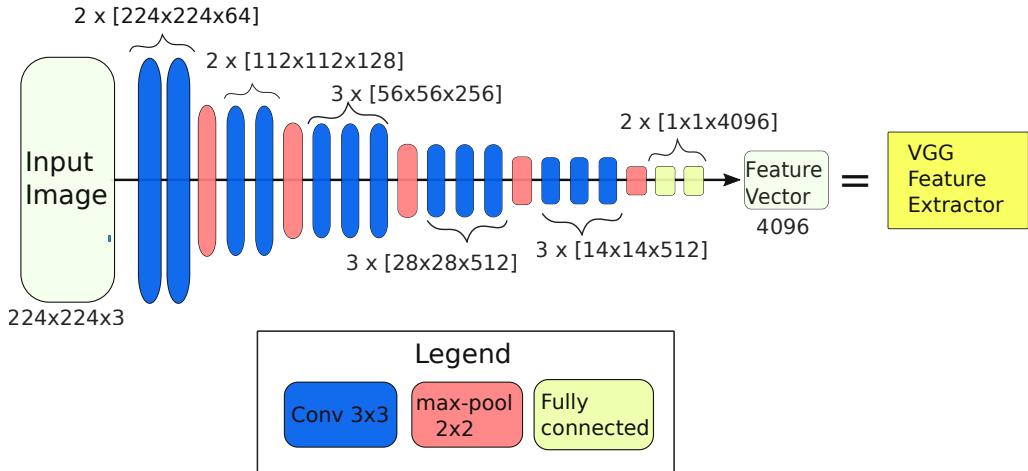
$$h_{deep}(\mathbf{x}_i, \boldsymbol{\theta}) = \hat{\mathbf{f}}_i$$

The Poincare hyperbolic space in this instance is defined as the 10-dimensional open unit ball:

$$F_{\mathcal{P}}^{10} = \{\mathbf{f} \in \mathbb{R}^{10} | \|\mathbf{f}\| < 1\}$$

and therefore we need to ensure that the Deep-HSEM outputs are constrained

a.) VGG as feature extractor



b.) Deep-HSEM

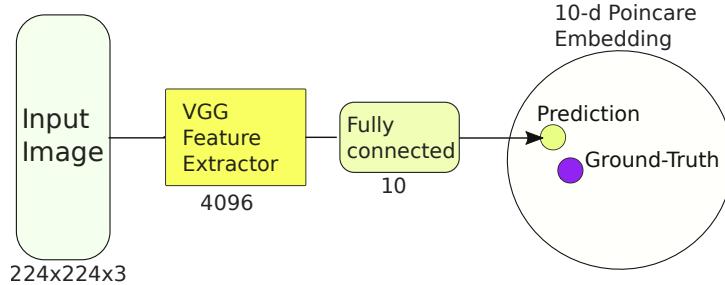


Figure 3.1: a.) Original VGG-16 structure except that we have taken out the final classification layer. The resulting model extracts a 4096 long feature vector from the input image. The dimensions of the layers are in the square brackets. b.) The VGG feature extractor is attached to a fully-connected layer which outputs a 10-dimensional embedding vector prediction. The distance of this prediction from the ground-truth embedding vector is used to compute a scalar loss. This is used to backpropagate gradients through the network.

into this space. This is ensured by the below operation, performed at the end of the network:

$$\hat{\mathbf{f}} = \begin{cases} \frac{\hat{\mathbf{f}}}{\|\hat{\mathbf{f}}\|}, & \text{if } \|\hat{\mathbf{f}}\| \geq 1 \\ \hat{\mathbf{f}}, & \text{otherwise} \end{cases} \quad (3.1)$$

In order to train the model parameters θ , a loss function $L(\hat{\mathbf{f}}_i, \mathbf{f}_i)$ is minimized. We considered various alternatives: squared Poincare distance, ranking hinge-loss on Poincare distance, and softmax-cross entropy loss on Poincare distances. The softmax-cross entropy loss was found to perform the best and was used for all the models presented here. In order to adapt the standard cross-entropy softmax (Equation 2.4) into this model, Poincare distances (Equation 2.14) $\mathcal{D}_{\mathcal{P}}(\hat{\mathbf{f}}_i, \mathbf{f}_j) \forall j \in \{1, \dots, i, \dots, M\}$ between the predicted embedding and all the possible ground-truth class embeddings were calculated. Since the aim is to minimize the distance to the correct ground-truth, inverses of these distances were used as class-scores i.e. logits. The final loss for a single prediction is hence:

$$\text{softmax}(\hat{\mathbf{f}})_i = \frac{\exp(-\mathcal{D}_{\mathcal{P}}(\hat{\mathbf{f}}_i, \mathbf{f}_i))}{\sum_{j=1}^M \exp(-\mathcal{D}_{\mathcal{P}}(\hat{\mathbf{f}}_i, \mathbf{f}_j))} := q_i \quad (3.2)$$

$$L(\hat{\mathbf{f}}_i, \mathbf{f}_i) = -\log(q_i) = \mathcal{D}_{\mathcal{P}}(\hat{\mathbf{f}}_i, \mathbf{f}_i) + \log \left(\sum_{j=1}^M \exp(-\mathcal{D}_{\mathcal{P}}(\hat{\mathbf{f}}_i, \mathbf{f}_j)) \right) \quad (3.3)$$

where $j \in \{1, \dots, i, \dots, M\}$.

During zero-shot evaluation, the Deep-HSEM takes all the new images $\{\mathbf{x}_z\}_{z=(M+1)}^N$ and predicts $\hat{\mathbf{f}}_z$ for each of them. After that, a nearest neighbour search is performed in the Poincare embedding using $\mathcal{D}_{\mathcal{P}}$ to map the prediction to the closest zero-shot class embedding vector: $nn_{\mathcal{D}_{\mathcal{P}}} : \hat{\mathbf{f}}_z \rightarrow \mathbf{f}'$ where $\mathbf{f}' \in F_{ZSL}^{10}$ and $F_{ZSL}^{10} \subseteq F_{\mathcal{P}}^{10}$. Finally, a look-up table $\mathcal{L} : \mathbf{f}' \rightarrow y_z$ assigns the class prediction (we can use a look-up table since we know the embedding vectors for all zero-shot classes). Note that $y_z \in Y_{ZSL}$ and $Y_{train} \cap Y_{ZSL} = \emptyset$.

Model Training

To implement the Deep-HSEM model, we started from a pre-trained open-source implementation of the VGG-16 network ² and adjusted the code to get the architecture described above. All of the implementations were done with PyTorch [34] which provides auto differentiation and, in general, is able to compute the gradients needed for backpropagation. However, since our loss function calculates hyperbolic distances, rather than Euclidean, we required the gradient output of the loss function to be Riemannian. Thus, we could not rely fully on the standard autodiff. Instead we amended the Poincare Distance function given in the open-source code of [15] to fit the rest of the Deep-HSEM architecture (see Footnote 1); this allows the Riemannian gradients to be computed appropriately.

We initially tested the model on a small 100 class subset of the full 2012 1k ImageNet dataset to ensure it was functioning as expected. We then proceeded to train the model on the full 1000 classes. Rather than training the model from scratch, we initialized the VGG portion of the Deep-HSEM with pre-trained VGG-16 weights available on PyTorch. The pre-trained model was trained on the same 2012 1k ImageNet dataset so gave better starting point for the model than random weights. Further, this reduced the necessary training time; training the model from scratch on such a big dataset would have taken over two weeks. With this set-up, two versions of the Deep-HSEM were trained. In the first, call it Deep-HSEM-FC, we only trained the final fully-connected layer that maps the 4096-long image feature vector into the embedding space. In the second, call it just Deep-HSEM, we trained all the layers.

Both Deep-HSEM-FC and Deep-HSEM were trained for a total of 90 epochs using SGD with momentum. The initial learning rate was set to 0.001 and it was reduced by factor of 10 every 30 epochs. We set momentum to 0.9. Batch-sizes of 32 images were used, with loss for each iteration averaged

²Available at <https://github.com/pytorch/examples/tree/master/imagenet>

over the batch images. The above hyperparameters were based on initial tests of 20 epochs of various settings. During the training-phase, the models were evaluated on additional 50,000 validation images provided with the 2012 ImageNet dataset. The overall training time for Deep-HSEM-FC was 4 days and 6 days for Deep-HSEM.

3.2.4 Convex Hyperbolic Semantic Embedding Model (Convex-HSEM)

The Convex-HSEM model is a hyperbolic embedding version of the ConSE model of Norouzi *et al.* [3]. To reiterate the discussion from Section 2.2.3, the ConSE approach can be described as two stages. In the first one, the model utilizes a standard pre-trained CNN to output predicted class probabilities in terms of the training classes. In our case, we use the original pre-trained VGG-16 CNN trained on the ImageNet 1k 2012 classes. Therefore, for any input image (including zero-shot ones), the CNN will output probabilities $p(y_1|\mathbf{x}_i) \dots p(y_{1000}|\mathbf{x}_i)$ where \mathbf{x}_i is the input image, $\sum_{i=1}^{1000} p(y|\mathbf{x}) = 1$, and $\{y_1, \dots, y_{1000}\} = Y_{train}$.

In the second stage of the model, the predictions from the previous step are sorted according to the probabilities and the corresponding T most likely class labels are extracted $\hat{y}(\mathbf{x}, 1), \hat{y}(\mathbf{x}, 2) \dots \hat{y}(\mathbf{x}, T)$. Next, their corresponding ground-truth embedding vectors are taken: $\mathbf{f}_1 = \mathcal{L}(\hat{y}(\mathbf{x}, 1)) \dots \mathbf{f}_T = \mathcal{L}(\hat{y}(\mathbf{x}, T))$ where \mathcal{L} is just a look-up table between class labels and embeddings. Finally, the predicted embedding vector is calculated as the weighted convex combination of the T most likely embeddings, where the weights are the respective softmax probabilities from the first stage of the model:

$$\hat{\mathbf{f}}(\mathbf{x}) = \frac{1}{Z} \sum_{t=1}^T p(\hat{y}(\mathbf{x}, t)|\mathbf{x}) \cdot \mathbf{f}_t \quad (3.4)$$

In the Convex-HSEM adaptation, the \mathbf{f}_t are Poincare embedding vectors of

the 2012 1K ImageNet classes. Calculating convex combinations in hyperbolic space is tricky, however. To see this, recall that due to constant negative curvature, straight lines in the Poincare model are arcs, and therefore, an Euclidean convex combination of two points will actually not fall on the shortest path arc between them. Further, distances cannot be treated as in the Euclidean case as they grow non-linearly away from the centre of the embedding space. Gulcehre *et al.* [100] solved a similar predicament, although in a different context, using the Einstein Gyromidpoints [102, 103]. Inspired by their approach, we used the Einstein midpoint to calculate convex combinations in our model. The Einstein midpoint in our setting is given as (see Ungar [103] Section 3 for derivation):

$$\hat{\mathbf{f}}(\mathbf{x})_{EM} = \sum_{t=1}^T \left[\frac{p(\hat{y}(\mathbf{x}, t) | \mathbf{x}) \gamma(\mathbf{v}_t)}{\sum_{l=1}^T p(\hat{y}(\mathbf{x}, l) | \mathbf{x}) \gamma(\mathbf{v}_l)} \right] \mathbf{v}_t$$

where \mathbf{v}_t are the embedding vectors expressed in Klein coordinates, and $\gamma(\mathbf{v}_t) = \frac{1}{\sqrt{1 - \|\mathbf{v}_t\|^2}}$, are known as Lorentz gamma factors. Since we used the Poincare model to construct our embeddings, we had to project all the embedding vectors into Klein coordinates, $\mathbf{f}_t \rightarrow \mathbf{v}_t$ (Equation 2.15). After calculating the Einstein midpoint, a nearest neighbour classifier, using the Klein distance function (see Equation 2.13), maps the midpoint to the closest label embedding \mathbf{v}_t :

$$nn_{\mathcal{K}} : \hat{\mathbf{f}}(\mathbf{x}_t)_{EM} \rightarrow \mathbf{v}_t$$

In zero-shot recognition we require that \mathbf{v}_t is a label embedding of one of the zero-shot classes expressed in Klein coordinates, denoted as \mathbf{v}_z . Finally, a look-up table is again used to get the predicted image label $\mathcal{L}(\mathbf{v}_z) \rightarrow y_z$ where $y_z \in Y_{ZSL}$.

The Convex-HSEM therefore consists of two step. First, the pre-trained VGG-16 model is used to produce predicted class probabilities in terms of

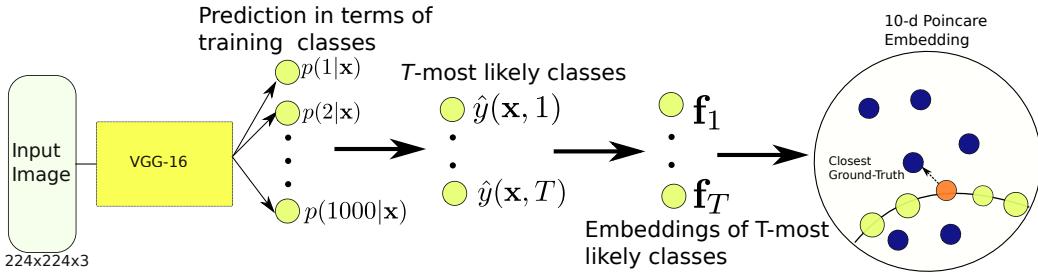


Figure 3.2: Illustration of the Convex-HSEM model. The orange circle represents the Einstein midpoint of the semantic embeddings of the T most likely classes (yellow circles). The Einstein midpoint is mapped to the nearest ground-truth embedding in order to classify the input image

training classes, out of which T most likely ones, and their corresponding embedding vectors, are chosen. The Einstein midpoint of these embeddings gives the predicted Poincare embedding for the input image, which allows it to be classified using nearest neighbour search over the possible label embeddings. Figure 3.2 illustrates these steps. Since we used the pre-trained VGG-16 model, Convex-HSEM did not require any additional training and was therefore much faster to implement than the Deep-HSEM. The only hyperparameter to set was the number of data points, T over which to calculate the Einstein midpoints. We ran our experiments for $T \in \{1, 2, 10, 100\}$. The motivation for this model architecture is to represent any unknown image as a combination of the most similar known image classes.

Chapter 4

Experiments & Results

4.1 Overview

This section starts by evaluating the Deep-HSEM and Convex-HSEM models on the training data, which provides several important observations for later discussion. We then move on to our two main zero-shot learning experiments. The first one measures the standard prediction accuracy of the models on approximately 21k previously unseen classes; this is the benchmark used most often in literature. The second experiment is a new task that we have developed in order to better study the robustness of the zero-shot classifications. Specifically, this tasks involves predicting not only the correct fine-grained labels, but also the superordinate labels of the images. In both experiments, we make comparisons to baseline models that use Euclidean similarity-based embeddings. The section below gives an overview of all the models that were evaluated in these tasks.

4.2 Baseline Models

All of the experiments were run on the following models from the previous section: Deep-HSEM-FC (only the fully connected layer trained), Deep-HSEM

Model	CNN model used	Embedding used	Embedding dimensions
Deep-HSEM-FC	VGG-16	Poincare	10
Deep-HSEM	VGG-16	Poincare	10
Convex-HSEM(1)	VGG-16	Poincare	10
Convex-HSEM(2)	VGG-16	Poincare	10
Convex-HSEM(10)	VGG-16	Poincare	10
Convex-HSEM(100)	VGG-16	Poincare	10
Deep-Glove	VGG-16	Glove [81]	50
DeViSE [2]	AlexNet[9]	word2vec [22]	500
ConSE(10) [3]	AlexNet	word2vec	500

Table 4.1: Summary of all the models used in the evaluation experiments and their main differences.

(all layers trained), Convex-HSEM(1), Convex-HSEM(2), Convex-HSEM(10) and Convex-HSEM(100). The T in Convex-HSEM(T) denotes that the T most likely classes were used in calculating the convex combination. The Euclidean embedding baseline models are the DeViSE [2] and ConSE [3] models upon which the Deep-HSEM and Convex-HSEM models are based respectively. Only ConSE(10) is considered out of the different ConSE(T) versions as it was the best performing one. Note that we were unable to obtain sufficient implementation details to train those two models ourselves and hence we present the results from the original papers. These models do not however provide a fair comparison for three major reasons. First, the CNN used by them is not as advanced as the VGG-16 model we utilized; our models might thus have an unfair advantage that's not based on the type of embedding used, which is our focus. Second, the Deep-HSEM model was only trained for 90 epochs, and while we don't know how long DeViSE was trained for, usually authors train their models up to thousands of epochs. Third, the dimensionality of the embeddings in the DeViSE and ConSE papers is much higher than ours (500-d vs. 10-d). Whilst the Poincare Embedding model

should do well even in low dimensions, this difference is so large that direct comparison is hard. For these reasons we trained a model called Deep-Glove which has exactly the same training details and architecture as Deep-HSEM except that it projects into a 50-dimensional ¹ Euclidean word embedding space. *Glove* [81] refers to the word embedding model used. Table 4.1 summarizes all the models and their key details.

4.3 Training Results

After the models were trained, their performance was evaluated on 50k validation images of the training classes. The results are displayed in Table 4.2. Since no zero-shot data was used, we have also included the original VGG-16 model as an additional baseline. This is indeed the one that performs by far the best. This is to be expected as the model is trained on one-hot-encoded labels - the benefit of orthogonal labels is that it makes the training data easier to separate (but as discussed, makes it impossible to do zero-shot classification).

Notice that the Deep-HSEM(1) model achieves the same Top-1 classification accuracy as the VGG-16. This too is expected as only the embedding of the most likely prediction is extracted, which is used to look-up the most likely class label, so the top-1 predictions of the two models are always the same. This no longer holds for top-5 as Convex-HSEM(1) will search for four other class embeddings closest to the embedding of the *top-1* most likely class, whilst VGG-16 outputs the five most likely classes. Additionally, note that the performance of the Convex-HSEM model deteriorates as we increase the size of the convex set. Indeed, there is no clear case for using convex combination models on training classes: if the fully trained VGG-16 predicts that an object is 60% cat and 20% dog, and these are the only two classes, taking

¹Lowest dimensionality readily available at <https://nlp.stanford.edu/projects/glove/>. We used the python open-source implementation of [80] to match the embedding to ImageNet classes: <https://github.com/JudyYe/zero-shot-gcn>

2012 1K Validation Data (training classes only)		
Model	Top-1 accuracy (%)	Top-5 accuracy (%)
VGG-16	71.4	90.2
Deep-HSEM-FC	52.9	62.6
Deep-HSEM	56.0	65.0
Convex-HSEM(1)	71.4	73.6
Convex-HSEM(2)	56.3	65.3
Convex-HSEM(10)	49.6	59.1
Convex-HSEM(100)	30.9	40.6
Deep-Glove	67.2	74.1
DeViSE [2]	53.2	76.7
ConSE(10) [3]	54.3	68.0

Table 4.2: Accuracies of all the models on the 2012 1k ImageNet validation data set.

an average of their embeddings will only reduce the prediction accuracy.

Another interesting result is that Deep-Glove has much better validation performance than the Deep-HSEM and Convex-HSEM models. As they all used the VGG-16 feature extractor, this suggests that it's either easier to learn the mapping into the Euclidean space from image features (with equal amounts of training time) or that it is structured so that the training classes are easier to separate. This will be discussed in more depth in analysing the zero-shot results. Finally, the Deep-Glove also has a significantly higher top-1 accuracy than the DeViSE model. This is likely the consequence of Deep-Glove using a more powerful CNN feature extractor. The top-5 accuracies of the two are similar, however. Since top-5 accuracy is a better judge of the embedding spaces, the most probable explanation is that the 500-dimensional word embedding helps DeViSE to make up for its less advanced CNN.

Dataset	# of classes
2-hop	1,549
3-hop	7,860
2011 20K ImageNet	20,841

Table 4.3: Overview of the datasets in the zero-shot classification experiment

4.4 Experiment 1 - Zero-shot Classification

4.4.1 Experimental Setup

This is the experiment used most widely to evaluate the prediction abilities of zero-shot models. The image and ground-truth labels for this task were extracted from the Full 2011 ImageNet dataset [1], which contains approximately 14 million images on 21,481 classes. Since the dataset subsumes also the training classes, those were excluded. Thus we ended up with previously unseen 20,841 classes; we will refer to this dataset as the '2011 20k ImageNet' data from hereon. Following previous works, we also constructed two additional subsets of this dataset, called *2-hop* and *3-hop* datasets. The 2-hop dataset contains all of the 2011 20k images which are within two tree-hops of the 1000 training classes in the WordNet tree hierarchy; the 3-hop labels are within three tree-hops. The labels for the three datasets were provided to us by [3]. Going from 2-hop, to 3-hop and finally to the full 2011 20k data we thus get an increasing level of difficult as the test labels are further from the training classes (note though that harder ones do contain the easier ones so the data are not completely disjoint). Table 4.3 summarizes this. Top-1, top-2, top-5, top-10, and top-20 prediction accuracies were the evaluation metrics. For example, top-20 accuracy means that the model outputs the 20 most-likely label predictions for each image and the classification is deemed to have 'hit' if any of the 20 match the ground-truth label.

4.4.2 Results & Discussion

The results for the zero-shot classification experiment for all the models and the three datasets are collated in Table 4.4. Overall, we can see that the baseline Euclidean similarity based embedding models DeViSE and ConSE(10) outperform our Deep-HSEM and Convex-HSEM models across all the test settings. ConSE(10) in particular is the best performing model. The relative underperformance of our hyperbolic models becomes more evident as we move further away from the training labels i.e. from the 2-hop data to the full ImageNet 20k data. Since our models actually used a more advanced CNN, the explanation must be related to the type of embeddings used. This hence seems to suggest that on average contextual semantic similarity may provide better visual guidance than semantic hierarchies in classifying novel visual objects; especially when the object is very different from the training labels. As noted earlier, however the embedding dimension of DeViSE and ConSE is much higher which makes comparisons less reliable. A better picture is hence given by comparing the hyperbolic models to Deep-Glove. If we focus on top-1 accuracy, a similar picture emerges; for example, Deep-Glove outperforms Deep-HSEM in the 2-hop data (7.5% vs. 2.5%) and even more clearly on the full data (1% vs. 0.2%). However, notice that as we move from top-1 accuracy to top-20 accuracy, the relative performance flips slowly. For instance, on the full data set Deep-HSEM has an accuracy of 3.5% vs. 3.1% of Deep-Glove. Easier the data set, the stronger this trend becomes e.g. contrast Deep-HSEM top-20 accuracy of 34.1% to 20.8% of the Glove equivalent on the 2-hop data. The underperformance of our models relative to DeViSE and ConSE is also much slighter for higher top- k values. It therefore seems that the Euclidean embeddings may get their predictions right more often, however when they don't, they are more likely to be in a wrong 'area' of the embedding space and thus benefit less from increasing the search radius (increasing k of top- k). Further, the poor performance of Deep-Glove relative to DeViSE for larger k implies that the Euclidean embedding models

are reliant on high dimensionality of the embedding.

Why would hyperbolic embeddings fare worse the further we get from the training labels? To explore this, we browsed through several concrete predictions. Figure 4.1a gives an example of a zero-shot class on which Deep-Glove does well and Deep-HSEM poorly, namely 'ski jumping'. The closest class in the training data is probably the word 'ski' which gives several images similar to those shown here, in that they include wintry scenes and people with skis. The contextual nature of the Euclidean embeddings places 'ski' close to many related terms and thus the model can extrapolate appropriately, though as a consequence some non-nonsensical predictions are made e.g. 'ski-lodge'. The reason why our hyperbolic model fails on such examples is likely due to the rigidity of the WordNet hierarchy, in which 'ski jumping' is under a super-category of 'sports', while the most relevant training label 'ski' is under 'devices'.

Figure 4.1b, on the other hand, displays an example of 'rattlesnake', which Deep-HSEM has labelled correctly, but Deep-Glove has not. These images are from the 2-hop dataset and the classification suggestions illustrate that Deep-HSEM has been able to exploit the semantic hierarchy of the WordNet to its advantage. The training data contained several snake species and thus the hyperbolic embedding has learned a good spatial representation of the relevant taxonomy: notice how one of the predictions is the super-category 'snake'. Given the abundance of different snakes in the training data, it is interesting that the Deep-Glove does so poorly. Most of the predictions are animal and nature related, and probably they occur relatively near in the embedding to 'snake' but the space is not structured enough for making correct predictions given the visual features of the images. This, in turn, is likely a reflection of the problems with low dimensional Euclidean embedding and helps to explain why Deep-Glove's accuracy is so far behind DeViSE's. As anticipated, these examples demonstrate that Hyperbolic embeddings do well when there is a clear hierarchical structure to be exploited, but suffer

		Top-k Accuracies(%)				
Test Dataset	Model	1	2	5	10	20
2-hop	Deep-HSEM-FC	2.4	5.4	14.6	23.2	33.2
	Deep-HSEM	2.5	5.6	15.0	23.7	34.1
	Convex-HSEM(1)	3.0	6.5	16.7	26.0	37.0
	Convex-HSEM(2)	1.4	3.6	10.9	18.5	27.2
	Convex-HSEM(10)	0.7	2.1	6.9	13.0	20.4
	Convex-HSEM(100)	0.6	1.8	6.1	11.5	18.3
	Deep-Glove	7.5	10.7	14.8	17.7	20.8
	DeViSE [2]	6.0	10.0	18.1	26.4	36.4
	ConSE(10) [3]	9.4	15.1	24.7	32.7	41.8
	Deep-HSEM-FC	0.4	0.8	2.3	4.7	8.3
3-hop	Deep-HSEM	0.4	0.8	2.3	4.7	8.4
	Convex-HSEM(1)	0.5	0.1	2.6	5.2	9.0
	Convex-HSEM(2)	0.2	0.5	1.6	3.4	6.3
	Convex-HSEM(10)	0.1	0.2	0.9	2.2	4.2
	Convex-HSEM(100)	0.1	0.2	0.8	1.9	3.7
	Deep-Glove	1.9	2.7	4.1	5.2	6.2
	DeViSE	1.7	2.9	5.3	8.2	12.5
	ConSE(10)	2.7	4.4	7.8	11.5	16.1
	Deep-HSEM-FC	0.2	0.3	1.0	1.9	3.4
	Deep-HSEM	0.2	0.4	1.0	2.0	3.5
2011 20k ImageNet	Convex-HSEM(1)	0.2	0.4	1.1	2.2	3.9
	Convex-HSEM(2)	0.1	0.2	0.7	1.5	2.7
	Convex-HSEM(10)	0.1	0.1	0.4	0.9	1.8
	Convex-HSEM(100)	0.0	0.1	0.3	0.8	1.6
	Deep-Glove	1.0	1.3	1.9	2.5	3.1
	DeViSE	0.8	1.4	2.5	3.9	6.0
	ConSE(10)	1.4	2.2	3.9	5.8	8.3

Table 4.4: Top-k classification accuracies for $k \in \{1, 2, 5, 10, 20\}$ on the three zero-shot datasets based on the 2011 20k ImageNet, with increasing difficulty from top to bottom. None of the zero-shot classes were included in the training classes. Classification was always constrained to the labels of the dataset under consideration. The results for DeViSE and ConSE are directly from [2] and [3] respectively.

from the lack of contextual information when this is not the case.

Another interesting result is that Convex-HSEM(1) performs the best out of our hyperbolic models. This implies that there is no benefit to calculating convex combinations which is at odds with ConSE(10) being the best ConSE model [3]. One potential explanation is that convex combinations are more beneficial with similarity based embeddings where the vast amount of contextual information creates noise which this method helps to iron out. Alternatively, it could be that calculating convex combinations in hyperbolic geometry does not provide similar impact as in Euclidean spaces; whilst the Einstein midpoint shares many similarities with the Euclidean convex combination, the two are not equivalent [103]. These are however just speculations and more work is needed to investigate them. The fact that Deep-HSEM performs worse than Convex-HSEM(1) could suggest that learning a mapping from the image features into the Poincare space is challenging. Finally, the overall low accuracies for all the models emphasizes the difficulty of zero-shot learning.

4.5 Experiment 2 - Robust Zero-Shot Classification

4.5.1 Experimental Setup

A limitation of the previous experiment is that classifications are considered correct only if the precise label of the image is predicted. For example, predicting 'arctic fox' as a 'fox' yields no points, even though in many situations that could be useful. The top-10 and top-20 accuracies partly alleviate this issue as they allow for a wider area to be searched, nevertheless they fail to provide a complete measure of how well the embedding space is structured. This is in particular the case for Poincare embeddings that place the top-level categories in the middle of the hyperbolic ball and lowest level ones near its



Deep-Glove:

1. **ski-jumping**
2. ski-lodge
3. ski-resort
4. ski-jacket
5. powder
6. compact car
7. ski-plane
8. ski jump
9. string instrument
10. water ski

Deep-HSEM:

1. artifact
2. device
3. being
4. person
5. instrument
6. electronic device
7. magnet
8. electrical device
9. equipment
10. implement

Deep-Glove:

1. ski jacket
2. pentathlete
3. ski-plane
4. throat protector
5. ski lodge
6. **ski jumping**
7. wild basil
8. bud brush
9. pear
10. bristle fern

Deep-HSEM:

1. artifact
2. device
3. being
4. person
5. instrument
6. electronic device
7. magnet
8. indicator
9. electrical device
10. heater

(a) Example Top-10 predictions on two images of 'ski jumping' from the full 2011 20k ImageNet data. Deep-Glove makes correct classifications while Deep-HSEM does not



Deep-Glove:

1. diamondback turtle
2. espadrille sandal
3. rotisserie
4. great ape
5. cinnabar mushroom
6. rhesus monkey
7. green monkey
8. jay bird
9. american water ouzel
10. suillus mushroom

Deep-HSEM:

1. snake
2. viper
3. **rattlesnake**
4. animal
5. cloubridsnake
6. boa
7. python
8. bird
9. iguanid lizard
10. black mamba

Deep-Glove:

1. gazella
2. ceratopsian dinosaur
3. alcazar
4. soapbox
5. milkcap
6. chihuahua lizard
7. goose
8. whiptail lizard
9. white sturgeon
10. espadrille sandal

Deep-HSEM:

1. snake
2. **rattlesnake**
3. viper
4. animal
5. cloubridsnake
6. python
7. boa
8. bird
9. black mamba
10. iguanid lizard

(b) Example Top-10 predictions on two images of 'rattlesnake' from the 2-hop data. Deep-HSEM makes correct classifications by exploiting the hierarchy of its embedding while Deep-Glove performs poorly. Images from the 2011 ImageNet dataset [1]

boundary. Given this structure, the model could end up predicting higher level categories of an object when it is uncertain about its exact class. This would be a desirable property as it would imply robustness of the model, but Experiment 1 does not account for correct higher level categorizations. We have therefore developed a 'robust zero-shot classification' task to investigate this, which involves predicting the correct label as well as its parent categories.

This experiment is similar to the *hierarchical precision* challenge in [2]. However, that task involved trying to predict labels within a certain radius around the correct label, and as a consequence rewarded incorrect predictions such as labelling an image of a brown bear as a polar bear. Moreover, the predictions in the hierarchical precision experiment are equally weighted irrespective of their distance from the correct label. Our robust zero-shot experiment, on the other hand, involves predicting only the correct label and its superordinate (higher level) categories. Additionally, we weigh predictions by the inverse of their distance (in edges) from the correct fine grained label. We decided to limit the levels of hierarchy in this task to five, that is, the ground truth label-set includes the original label and its superordinate category labels four levels up. These are extracted from the WordNet database [57]. This constraint was put into place to prevent very generic predictions like 'entity' being rewarded. The evaluation metric used is a weighted top- k accuracy, computed for all k from 1 to 5. As an example, weighted top-3 accuracy means that the three closest predictions are made and we calculate their (inverse) distance-weighted hit-rate over the 5 labels. Note that regardless of the k in the weighted top- k , predicting any of the five labels is allowed. Figure 4.2 gives a hypothetical example of this for one image.

The datasets for this experiment are based on the 2-hop, 3-hop and 2011 20k ImageNet data. However, now each of them contains also all the required superordinate nouns i.e. we now have hierarchical 2-hop, 3-hop and 2011 20k

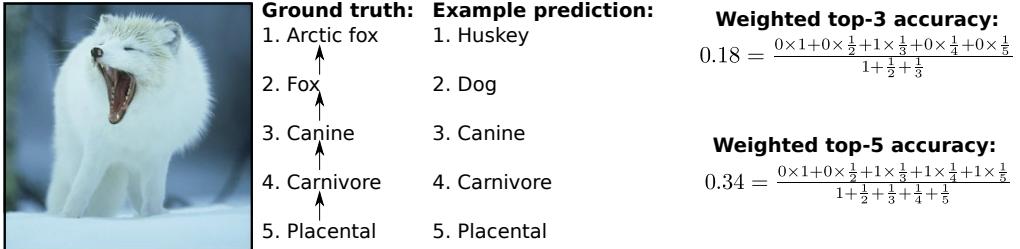


Figure 4.2: An example of top-3 & top-5 accuracy calculations in the robust zero-shot experiment. Predicting any of the five labels is allowed irrespective of which top- k is used, though only top- k predictions are counted. Image from the 2011 ImageNet database [1]

datasets. As there are now more labels in the embedding space, this might make it harder to find the correct nearest neighbour. On the other hand, being rewarded for predicting any of the five hierarchical labels should make the task easier if the model has a well structured embedding space. Since this a new experiment, we were only able to run it on our own models – DeVISE and ConSE were thus excluded from the baseline. Due to time constraints, we also chose to dispense with the Convex-HSEM(10) and Convex-HSEM(100) models as they were the two with worst performance in Experiment 1.

4.5.2 Results & Discussion

All the results for the robust zero-shot classification experiment are displayed in Table 4.5. Overall, we found that the Deep-HSEM and Convex-HSEM models perform better than the baseline Deep-Glove model across all the runs. In Experiment 1 we found that our hyperbolic models did progressively worse, relative to Euclidean models, as we moved further away from training labels. The reverse is true here: on the 2-hop data the weighted top-1 accuracy of Deep-HSEM is still worse than Convex-HSEM’s (6.2% vs. 7.6%), as was in Experiment 1, however on the 2011 20k data the corresponding numbers are 2.5% (Deep-HSEM) and 1.1% (Deep-Glove). The ability of

Deep-HSEM to find correct super-categories, even for labels that are very different from training data, makes up for its lower accuracy in predicting the fine-grained labels.

Together the results from the two experiments suggest that the contextual semantic information of Euclidean embeddings helps to find the correct label more easily, yet the overall structure of those models is less robust i.e. there are more instances where their predictions are completely off. The hyperbolic embeddings, in turn, may struggle to identify the exact class of an object but can still relatively often predict one of the superordinate categories correct. This is particularly evident from how our hyperbolic models' accuracies improve as we move from the weighted top-1 accuracy to top-5. This implies that the models are particularly adept at predicting the higher level categories of objects. By definition the higher level categories hold more objects so this may appear a trivial point, however, we can see that the accuracy of Deep-Glove only improves marginally for higher k as there is no hierarchical structure in the embedding to exploit. The result here is however only for the 50-dimensional Glove model, and likely the 500-dimensional one used by DeViSE and ConSE would do much better.

Overall, the findings here corroborate our hypothesis that the hyperbolic embedding models are able to exploit the relationships between visual and semantic hierarchies, which enables them to make more robust higher-level predictions. Figure 4.3 displays a few example images and predictions that illustrate this. Take as an example the first image of a 'puppy': even though Deep-HSEM is unable to predict this exact label, it successfully classifies the image as a dog and canine; even the 'wrong' predictions appear semantically reasonable. Deep-Glove, in turn, makes very granular classifications. Its prediction 'springer spaniel' looks appropriate (possibly the correct breed) however mozzarella and mammoth less so. Similar pattern is visible in the other examples.

Another interesting result is that the Deep-HSEM model performed bet-

Robust Zero-Shot Classification		Weighted Top-k Accuracies(%)				
Test Dataset	Model	1	2	3	4	5
Hierarchical 2-hop	Deep-HSEM-FC	5.9	10.9	15.5	19.7	23.6
	Deep-HSEM	6.2	11.3	15.7	20.1	24.2
	Convex-HSEM(1)	8.3	13.3	17.9	22.3	26.4
	Convex-HSEM(2)	4.5	8.6	12.5	16.1	19.4
	Deep-Glove	7.6	10.4	11.9	12.8	13.5
Hierarchical 3-hop	Deep-HSEM-FC	4.0	7.6	10.6	13.2	15.3
	Deep-HSEM	4.1	8.1	11.7	14.9	17.8
	Convex-HSEM(1)	3.6	7.5	11.1	14.3	17.3
	Convex-HSEM(2)	3.1	6.5	9.6	12.4	14.9
	Deep-Glove	2.9	3.3	3.9	4.3	4.6
Hierarchical 2011 20k ImageNet	Deep-HSEM-FC	2.0	5.0	7.3	9.3	11.2
	Deep-HSEM	2.5	5.0	7.3	9.3	11.2
	Convex-HSEM(1)	1.8	4.2	6.4	8.3	10.2
	Convex-HSEM(2)	2.0	4.2	6.3	8.0	9.8
	Deep-Glove	1.1	1.4	1.7	1.8	2.0

Table 4.5: Robust Zero-Shot Classification results with inverse distance-weighted top- k accuracies. The hierarchical 2-hop, 3-hop and 2011 20k datasets were derived from Experiment 1 by adding four levels of super-ordinate labels to each original label from the WordNet hierarchy.

	Ground-truth hierarchy: 1. puppy (orig. label) 2. pup 3. dog 4. canine 5. domestic animal	Deep-HSEM predictions: 1. domestic animal 2. dog 3. canine 4. hunting dog 5. carnivore	Deep-Glove predictions: 1. heckelphone 2. <i>springer spaniel</i> 3. mozzarella 4. spoonbill 5. mammoth
	Ground-truth hierarchy: 1. carp (orig. label) 2. cyprinid 3. cypriniform fish 4. soft-finned fish 5. teleost fish	Deep-HSEM predictions: 1. soft-finned fish 2. aquatic vertebrate 3. fish 4. teleost fish 5. bony fish	Deep-Glove predictions: 1. <i>crucian carp</i> 2. bethel 3. humming bird 4. coypu (mammal) 5. plasmodium
	Ground-truth hierarchy: 1. finch (orig. label) 2. oscine 3. passerine 4. bird 5. verterbrate	Deep-HSEM predictions: 1. animal 2. chordate 3. vertebrate 4. bird 5. organism	Deep-Glove predictions: 1. redpoll 2. fringillae bird 3. cockateel 4. linnet 5. archaeopteryx (bird-like dinosaur)
	Ground-truth hierarchy: 1. ameba (orig. label) 2. rhizopod 3. protozoan 4. prototist 5. micro-organism	Deep-HSEM predictions: 1. organism 2. person 3. artifact 4. vascular plant 5. plant	Deep-Glove predictions: 1. pantechnicon van 2. elasmobranch (class of fish) 3. tenterhook 4. phasianid bird 5. edentate (class of mammals)

Figure 4.3: Four handpicked example test images from the robust zero-shot classification task. The first three illustrate successful high-level predictions by Deep-HSEM. Deep-Glove makes classifications at a very granular level and also makes more completely incorrect predictions. The fourth image is an example of a failure case for both models. Images from the ImageNet repository [1]

ter at the robust zero-shot experiment than the Convex-HSEM, especially for the harder dataset. Even though Convex-HSEM was better at predicting the correct class in Experiment 1, its reliance on the embeddings of the closest training labels limits the model’s ability to generalize to higher level categories. In other words, the model can, by definition, only map those areas of the embedding manifold that are in the set of convex combination of training classes – this likely makes it harder to find superordinate categories for objects far away from the training examples. Deep-HSEM on the other hand is theoretically able to map into any area of the embedding space so is less tied to the local structure of training classes. Finally, the Deep-HSEM-FC and Deep-HSEM models perform very similarly to each other. It therefore seems that the pre-trained VGG-16 architecture we utilized comes with sufficiently good visual features that there is was no substantial benefit from training them further.

Chapter 5

Conclusions & Ideas for Future Work

5.1 Summary

The aim of this thesis has been to improve the generalization ability of deep learning models used for zero-shot learning. The precise research question we set out to explore was:

- *Does semantic hierarchical knowledge in the form of hyperbolic embeddings help to improve the zero-shot prediction capability of deep learning models*

Two zero-shot learning models were introduced to investigate this. The Deep-Hyperbolic Semantic Embedding Model (Deep-HSEM) applies a CNN to learn a mapping from training images into a hyperbolic Poincare embedding. The Convex-Hyperbolic Semantic Embedding Model (Convex-HSEM) predicts the Poincare embedding vectors of zero-shot images as a convex combination of training classes' embeddings. In both models, zero-shot classification is performed by finding the nearest neighbour of the predicted vector in the hyperbolic embedding space.

These models were evaluated on two related zero-shot experiments and their results were compared against baseline models that used similarity-based word embeddings. The first one measured the models’ classification accuracies on approximately 20k previously unseen image classes from the ImageNet database. In the second experiment, called robust zero-shot classification, the models were evaluated on how well they were able to predict both the correct label of new images as well as which higher level categories they belong in.

Overall, our results suggest that there are benefits to both similarity based embeddings and hyperbolic embeddings. The former have the advantage of being more flexible as they are ingrained with huge amounts of contextual information and are in this sense able to make associations between new classes and training images more easily, even if they are very different from each other. As a result they were better at predicting the fine-grained labels of zero-shot classes. On the other hand, the hyperbolic embedding models have the advantage of being able to exploit the relationship between visual and semantic hierarchies when those exist; for instance, our explorations suggest that this is often the case with biological categories, as expected. Further, we found that even though the Deep-HSEM and Convex-HSEM models predicted the correct class less often, they were also less often completely wrong. In particular, the robust zero-shot experiment results showed that the hyperbolic models were more adept at grouping image classes into their correct higher level categories than the similarity-based benchmark model. To answer the research question above, our results are promising and suggest several potential benefits, such as more robust classification of objects into their higher level categories. Nevertheless, many improvements and further tests need to be done before any strong conclusions can be drawn. These are discussed below.

5.2 Limitations & Future Work

Several important experiments were not completed due to the high compute cost of these models and the large datasets they needed to be trained on. One particularly important test that future studies should complete is to repeat above experiments with different dimensionalities of the Poincare embedding. As discussed in the previous chapter, our models did relatively well when compared to the Deep-Glove model that used only 50-dimensional Euclidean embedding, but could not match the performance of the DeViSE and ConSE models that used 500-d embedding vectors. It is likely that our decision to use only a 10-d Poincare embedding was too ambitious for such a large dataset. Alternate hyperbolic embedding models could also be investigated: Nickel and Kiela [96] demonstrated very recently the superiority of the Hyperboloid model over their Poincare model on various graph embedding tasks.

Several other architectural details should be experimented with. In particular, different loss functions should be explored in more detail. In the present study only cross-entropy loss was used. We did perform evaluations of other loss functions on the training data but not on the zero-shot tasks. For instance, the cross-entropy loss may be the optimal one for discriminating between different training classes, while some other function such as simple squared hyperbolic distance. This may help to learn a more precise mapping into the embedding space and generalizes better to zero-shot classes. Hyperparameter settings and different CNN architectures should also be experimented in more detail. Finally, on top of pre-training, our models were trained only for 90 epochs, which is much less than typically seen in deep learning. Longer training time could improve the results substantially.

Critique can also be levelled at our experimental design. One potential source of bias is that in the robust zero-shot classification task we used the WordNet database to construct the hierarchical labels. This is the very same database used to build our Poincare embeddings. One could argue that our models outperforming the Euclidean baseline, which used Wikipedia-based

embedding instead, is not surprising. Note though that the Wikipedia data also contained all the possible image labels so there were not restrictions on which labels could be predicted. There are two other counter-arguments, in support of our approach. First, the WordNet database is much larger and more general than the set of ImageNet labels. Subsequently, it should be possible to use the learned Poincare embedding on most other dataset as their labels too can be mapped into WordNet. Second, in the training phase the embedding is given no explicit hierarchical information, such as a dog is an animal, but only information about which nouns are directly linked to each other [15]. The model learns the hierarchical structure as part of the training, relying on the properties of the hyperbolic space. Nevertheless, these considerations need to be explored in more detail, preferably on a separate dataset. A baseline model that uses Euclidean WordNet embedding could also be included in the robust zero-shot experiment (though Nickel and Kiela [15] did precisely show their inferiority to hyperbolic embeddings in representing the WordNet graph). Finally, we should create our own implementations of the DeViSE and ConSE models so they could also be evaluated on the robust zero-shot classification task.

The standard zero-shot classification experiment should also be expanded. In our current version, the models perform nearest neighbour search only over the zero-shot classes. In the, so called, *generalized* zero-shot learning the models are allowed also to predict training classes (which will always be wrong). This is thus a step harder experiment but is also often used in literature. In future, we will look to include many more baseline models from recent research; many of these are much more powerful than the DeViSE and ConSE models. Nevertheless, our purpose was not to create the most powerful zero-shot model but rather explore the potential benefits of hyperbolic embeddings in this task and thus chose the models that allowed the relevant comparisons to be made.

Above paragraphs have covered the most immediate areas and shortcom-

ings of our methodology that need to be addressed. Future works should also concern themselves with more fundamental challenges that may be constraining the performance of the proposed deep hyperbolic zero-shot models. These can be grouped under two broader, related, categories. The first ones have to do with the quality and usefulness of the hyperbolic semantic embedding built on the WordNet hierarchy, and the second ones have to do with the quality of mapping we have learned from the image space into the embedding space.

Concerning the first set of limitations, one of our conclusions was that the vast amount of contextual information that is inherent in the datasets used by Euclidean embedding methods gives the model more flexibility than the rigid WordNet hierarchy, yet we also saw some benefits of hyperbolic embeddings. An interesting idea for future research would be to build a hyperbolic embedding based on both contextual data as well as explicit semantic word taxonomies. This is possible because the Poincare model of Nickel and Kiela [96] does not require explicit hierarchy as an input, instead, as discussed above, it only needs to be given data about which words are connected to each other. If such a dataset could be compiled from the Wikipedia text data, then we could use it to learn a latent hierarchy of those words and employ that in conjunction with the WordNet Poincare embedding, possibly for improved performance. Another, simpler, approach of combining the benefits of the two groups of models may be to take a similarity based embedding model and our hyperbolic embedding model and always pick the most likely prediction across the two.

One of the most fundamental challenges in zero-shot learning [74] is that the structure of the semantic embedding space should closely reflect the local smoothness of the visual feature-space. Certainly, the low overall accuracy of all of our models suggests that this might be an issue. A possible solution might be to utilize an approach similar to Kodirov *et al.* [69] who used an autoencoder to ensure that it was possible to reconstruct the input images from

the latent embedding. We could look to improve on this by initializing the latent vectors to be those of the hyperbolic embedding, but make them. The reconstruction constraint of the autoencoder could thus help the model to re-adjust the initial hyperbolic embedding space to be more visually smooth.

Another common issue encountered in zero-shot learning is that of the domain projection shift [66]; the training and test classes in zero-shot learning are fully disjoint and thus the learned mapping may not transfer well. We suspect that this issue might be relevant for hyperbolic embeddings. In particular, the hierarchical organization means that the embedding space is not populated evenly, and therefore the model learns a mapping to a more constrained area. This could explain why the hyperbolic models had lower accuracy on the standard zero-shot classification experiment than the Euclidean baselines, especially on classes that were far away from the training labels. Future research should investigate how this could be alleviated.

Finally, it would be worth exploring the possibility of imposing hyperbolic geometry on our models more broadly. For instance, Gulcehere *et al.* [100], showed that placing hyperbolic geometry on the activation functions of neural networks can increase their representational capacity and improve performance. Currently only our embedding is in hyperbolic space and in training Riemannian gradients from the Poincare distance loss function are combined with Euclidean gradients. If hyperbolic space was imposed on all model parameters, all the layers could be trained with Riemannian optimization. A more ambitious future research project would thus look at how this might be implemented and investigate its benefits on zero-shot learning.

Bibliography

- [1] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “ImageNet: A large-scale hierarchical image database,” *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.
- [2] A. Frome, G. Corrado, and J. Shlens, “Devise: A deep visual-semantic embedding model,” *Advances in Neural ...*, pp. 1–11, 2013.
- [3] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean, “Zero-Shot Learning by Convex Combination of Semantic Embeddings,” pp. 1–9, 2013.
- [4] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
- [5] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, “End to End Learning for Self-Driving Cars,” pp. 1–9, 2016.
- [6] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “DeepFace: Closing the gap to human-level performance in face verification,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1701–1708, 2014.

- [7] C. Finn, X. Y. Tan, Y. Duan, T. Darrell, S. Levine, and P. Abbeel, “Learning Visual Feature Spaces for Robotic Manipulation with Deep Spatial Autoencoders,” *arXiv preprint arXiv:1509.06113*, 2015.
- [8] T. Ganegedara, L. Ott, and F. Ramos, “Learning to navigate by growing deep networks,” *Australasian Conference on Robotics and Automation, ACRA*, vol. 2017-Decem, pp. 114–123, 2017.
- [9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Advances In Neural Information Processing Systems*, pp. 1–9, 2012.
- [10] S. Dodge and L. Karam, “A Study and Comparison of Human and Deep Learning Recognition Performance Under Visual Distortions,” 2017.
- [11] M. Palatucci, G. E. Hinton, D. Pomerleau, and T. M. Mitchell, “Zero-Shot Learning with Semantic Output Codes,” *Neural Information Processing Systems*, pp. 1–9, 2009.
- [12] T. Deselaers and V. Ferrari, “Visual and semantic similarity in ImageNet,” *Cvpr*, pp. 1777–1784, 2011.
- [13] M. Ohl, *Handbook of Paleoanthropology -Principles of Taxonomy and Classification*. No. July, Springer-Verlag Berlin, 2015.
- [14] P. Joliceur, M. A. Gluck, and S. M. Kosslyn, “Pictures and Naming: Making the Connection,” *Cognitive Psychology*, vol. 16, pp. 243–275, 1984.
- [15] M. Nickel and D. Kiela, “Poincaré Embeddings for Learning Hierarchical Representations,” 2017.
- [16] B. P. Chamberlain, J. Clough, and M. P. Deisenroth, “Neural Embeddings of Graphs in Hyperbolic Space,” 2017.

- [17] Y. Lecun, G. E. Hinton, and Y. Bengio, “Deep Learning,” *Nature*, vol. 521, pp. 436–444, 2015.
- [18] P. Wang and G. W. Cottrell, “Basic Level Categorization Facilitates Visual Object Recognition,” pp. 1–13, 2015.
- [19] J. C. Peterson, P. Soulou, A. Nematzadeh, and T. L. Griffiths, “Learning Hierarchical Visual Representations in Deep Neural Networks Using Hierarchical Linguistic Labels,” 2018.
- [20] R. Socher, M. Ganjoo, C. D. Manning, and A. Y. Ng, “Zero-Shot Learning Through Cross-Modal Transfer,”
- [21] E. H. Huang, R. Socher, C. D. Manning, and A. Y. Ng, “Improving-Word Representations via Global Context and MultipleWord Prototypes,” *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, no. July, pp. 873–882, 2012.
- [22] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “5021-Distributed-Representations-of-Words-and-Phrases-and-Their-Compositionality,” pp. 1–9.
- [23] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” pp. 1–12, 2013.
- [24] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem, “Basic objects in natural categories,” *Cognitive Psychology*, vol. 8, no. 3, pp. 382–439, 1976.
- [25] M. J. Greenberg, *Euclidean and Non-Euclidean Geometries - Development and History*. W. H. Freeman and Company, 3 ed., 1994.
- [26] F. Rosenblatt, “The perceptron: A probabilistic model for information storage and organization in . . . ,” *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958.

- [27] Jürgen Schmidhuber, “Deep learning in neural networks: An overview,” *Neural Networks*, vol. 61, pp. 85–117, 2015.
- [28] A. L. Hodgkin and A. F. Huxley, “A quantitative description of membrane current and its application to conduction and excitation in nerve,” *Bulletin of Mathematical Biology*, vol. 52, no. 1-2, pp. 25–71, 1990.
- [29] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” *AISTATS ’11: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, vol. 15, pp. 315–323, 2011.
- [30] C. E. Shannon, “A mathematical theory of communication,” *The Bell System Technical Journal*, vol. 27, no. July 1928, pp. 379–423, 1948.
- [31] S. Linnainmaa, “Taylor expansion of the accumulated rounding error,” *Bit*, vol. 16, no. 2, pp. 146–160, 1976.
- [32] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning Internal Representations by Error Propagation,” *Institute for Cognitive Science Report*, no. 8506, 1985.
- [33] H. Robbins and S. Monro, “A Stochastic Approximation Method,” *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951.
- [34] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in PyTorch,” in *NIPS*, 2017.
- [35] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, R. Jozefowicz, Y. Jia, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, M. Schuster, R. Monga, S. Moore,

- D. Murray, C. Olah, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, “TensorFlow: Large-scale machine learning on heterogeneous systems,” 2015.
- [36] Y. Lecun, B. Boser, J. Denker, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation Applied to Handwritten Zip Code Recognition,” 1989.
 - [37] M. Zeiler and R. Fergus, “Visualizing and Understanding Convolutional Networks,” in *European Conference on Computer Vision*, pp. 818–833, 2014.
 - [38] C. F. Cadieu, H. Hong, D. L. K. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, and J. J. DiCarlo, “Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition,” *PLoS Computational Biology*, vol. 10, no. 12, 2014.
 - [39] D. H. Hubel and T. Wiesel, “Receptive fields, binocular interaction, and functional architecture in the cat’s visual cortex,” *Journal of Physiology*, no. 160, pp. 106 – 154, 1962.
 - [40] E. Kobatake and K. Tanaka, “Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex,” *Journal of Neurophysiology*, vol. 71, pp. 856–867, 1994.
 - [41] D. Lowe, “Object recognition from local scale-invariant features,” *Proceedings of the Seventh IEEE International Conference on Computer Vision*, pp. 1150–1157 vol.2, 1999.
 - [42] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, “CNN features off-the-shelf: An astounding baseline for recognition,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 512–519, 2014.

- [43] R. K. Srivastava, K. Greff, and J. Schmidhuber, “Training Very Deep Networks,” pp. 1–11, 2015.
- [44] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” 2015.
- [45] S. Sun, W. Chen, L. Wang, X. Liu, and T.-Y. Liu, “On the Depth of Deep Neural Networks: A Theoretical View,” 2015.
- [46] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.
- [47] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, “Reading Digits in Natural Images with Unsupervised Feature Learning,” in *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.
- [48] I. Biederman, ““Recognition-by-components: A theory of human image understanding”: Clarification.,” *Psychological Review*, vol. 96, no. 1, pp. 2–2, 1989.
- [49] J. J. DiCarlo, D. Yoccoz, and N. C. Rust, “How does the brain solve visual object recognition?,” *Neuron*, vol. 73, no. 3, pp. 415–434, 2012.
- [50] Y. A. Aslandogan, C. Thier, C. T. Yu, J. Zou, and N. Rishe, “Using semantic contents and WordNet in image retrieval,” *ACM SIGIR Forum*, vol. 31, no. SI, pp. 286–295, 1997.
- [51] R. Zhao and W. I. Grosky, “Negotiating the semantic gap: From feature maps to semantic landscapes,” *Pattern Recognition*, vol. 35, no. 3, pp. 593–600, 2001.

- [52] W. Grosky, “Narrowing the semantic gap - improved text-based web document retrieval using visual features,” *IEEE Transactions on Multimedia*, vol. 4, no. 2, pp. 189–200, 2002.
- [53] K. Barnard and D. Forsyth, “Learning the semantics of words and pictures,” *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 2, pp. 408–415, 2001.
- [54] M. Srikanth, J. Varner, M. Bowden, and D. Moldovan, “Exploiting Ontologies for Automatic Image Annotation.,” *SIGIR Forum*, pp. 552–558, 2005.
- [55] M. Marszałek and C. Schmid, “Semantic hierarchies for visual object Recognition,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007.
- [56] G. Griffin, “Learning and Using Taxonomies for Visual and Olfactory Classification Thesis by,” vol. 2013, 2013.
- [57] G. A. Miller, “Wordnet: a lexical database for english.,” *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
- [58] B. Schölkopf and F. Bach, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.
- [59] J. Redmon and A. Farhadi, “Better , Faster , Stronger,” pp. 7263–7271.
- [60] A. B. Hillel and D. Weinshall, “Subordinate class recognition using relational object models,” *Nips*, vol. 19, pp. 73–80, 2007.
- [61] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

- [62] J. Lei, Z. Guo, and Y. Wang, “Weakly Supervised Image Classification with Coarse and Fine Labels,” *Proceedings - 2017 14th Conference on Computer and Robot Vision, CRV 2017*, vol. 2018-Janua, pp. 240–247, 2018.
- [63] F. Monay and D. Gatica-perez, “PLSA-based Image Auto-Annotation : Constraining the Latent Space,” *Proceedings of the 12th annual ACM international conference on Multimedia*, pp. 348–351, 2004.
- [64] J. Weston, S. Bengio, and N. Usunier, “Large scale image annotation: Learning to rank with joint word-image embeddings,” *Machine Learning*, vol. 81, no. 1, pp. 21–35, 2010.
- [65] X. Li, S. Liao, W. Lan, X. Du, and G. Yang, “Zero-shot Image Tagging by Hierarchical Semantic Embedding,” in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR 15*, pp. 879–882, 2015.
- [66] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, “Transductive Multi-View Zero-Shot Learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 11, pp. 2332–2345, 2015.
- [67] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, “Unsupervised Domain Adaptation for Zero-Shot Learning,” *IEEE International Conference on Computer Vision (ICCV)*, pp. 2452–2460, 2015.
- [68] M. Rohrbach, S. Ebert, and B. Schiele, “Transfer learning in a transductive setting,” *Advances in neural information . . .*, pp. 1–9, 2013.
- [69] E. Kodirov, T. Xiang, and S. Gong, “Semantic autoencoder for zero-shot learning,” *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 4447–4456, 2017.

- [70] Y. Fu and L. Sigal, “Semi-supervised Vocabulary-informed Learning,” 2016.
- [71] Y. Shigeto, I. Suzuki, K. Hara, M. Shimbo, and Y. Matsumoto, “Ridge regression, hubness, and zero-shot learning,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9284, pp. 135–151, 2015.
- [72] S. Changpinyo, W. L. Chao, and F. Sha, “Predicting Visual Exemplars of Unseen Classes for Zero-Shot Learning,” *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-Octob, pp. 3496–3505, 2017.
- [73] J. L. Ba, K. Swersky, S. Fidler, and R. Salakhutdinov, “Predicting deep zero-shot convolutional neural networks using textual descriptions,” *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 Inter, pp. 4247–4255, 2015.
- [74] Z. Ji, Y. Yu, Y. Pang, J. Guo, and Z. Zhang, “Manifold regularized cross-modal embedding for zero-shot learning,” *Information Sciences*, vol. 378, pp. 48–58, 2017.
- [75] C. Huang, C. C. Loy, and X. Tang, “Local Similarity-Aware Deep Feature Embedding,” vol. 1, no. Nips, pp. 1–9, 2016.
- [76] Z. Zhang and V. Saligrama, “Zero-Shot Learning via Semantic Similarity Embedding,” 2015.
- [77] Z. Zhang and V. Saligrama, “Zero-Shot Learning via Joint Latent Similarity Embedding,” 2015.
- [78] Z. Fu, T. A. Xiang, E. Kodirov, and S. Gong, “Zero-shot object recognition by semantic manifold distance,” *Proceedings of the IEEE Com-*

- puter Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June, pp. 2635–2644, 2015.
- [79] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, “Synthesized Classifiers for Zero-Shot Learning,” 2016.
 - [80] X. Wang, Y. Ye, and A. Gupta, “Zero-shot Recognition via Semantic Embeddings and Knowledge Graphs,” 2018.
 - [81] J. Pennington, R. Socher, and C. Manning, “Glove: Global Vectors for Word Representation,” *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.
 - [82] A. Farhadi, I. Endres, and D. Hoiem, “Attribute-centric recognition for cross-category generalization,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2352–2359, 2010.
 - [83] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele, “What helps whereand why? semantic relatedness for knowledge transfer,” *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, vol. 1, no. c, pp. 910–917, 2010.
 - [84] M. Rohrbach, M. Stark, and B. Schiele, “Evaluating knowledge transfer and zero-shot learning in a large-scale setting,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, no. March, pp. 1641–1648, 2011.
 - [85] S. J. Hwang and L. Sigal, “A Unified Semantic Embedding: Relating Taxonomies and Attributes,” pp. 71–74, 2014.
 - [86] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, “Label-Embedding for Image Classification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 7, pp. 1425–1438, 2016.

- [87] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, “Latent Embeddings for Zero-shot Classification,” 2016.
- [88] B. Romera-Paredes and P. H. S. Torr, “An Embarrassingly Simple Approach to Zero-Shot Learning,” vol. 37, pp. 11–30, 2017.
- [89] M. Bucher, S. Herbin, and F. Jurie, “Improving semantic embedding consistency by metric learning for zero-shot classification,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9909 LNCS, pp. 730–746, 2016.
- [90] S. Reed, Z. Akata, H. Lee, and B. Schiele, “Learning Deep Representations of Fine-Grained Visual Descriptions,” 2016.
- [91] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, “Evaluation of output embeddings for fine-grained image classification,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June, no. 1, pp. 2927–2936, 2015.
- [92] Z.-H. Zhou, D.-C. Zhan, and Q. Yang, “Semi-Supervised Learning with Very Few Labeled Training Examples,” *Artificial Intelligence*, vol. 22, no. 1, pp. 675–680, 2005.
- [93] Y. Xian, B. Schiele, and Z. Akata, “Zero-Shot Learning - The Good, the Bad and the Ugly,” 2017.
- [94] J. Stillwell, *Sources of hyperbolic geometry (History of mathematics v. 10)*. American Mathematical Society, 1991.
- [95] B. Iversen, *Hyperbolic geometry, volume 25*. Cambridge University Press, 1992.
- [96] M. Nickel and D. Kiela, “Learning Continuous Hierarchies in the Lorentz Model of Hyperbolic Geometry,” 2018.

- [97] H. Ritter, *Self-Organizing Maps on non-euclidean Spaces*. Elsevier, 1999.
- [98] J. Ontrup and H. Ritter, “Hyperbolic Self-Organizing Maps for Semantic Navigation,” *Advances in Neural Information Processing Systems 14*, pp. 1417–1424, 2002.
- [99] J. Walter, D. Webling, and K. Essig, “Interactive hyperbolic image browsing-towards an integrated multimedia navigator,” *Acm Sigkdd*, 2006.
- [100] C. Gulcehre, M. Denil, M. Malinowski, A. Razavi, R. Pascanu, K. M. Hermann, P. Battaglia, V. Bapst, D. Raposo, A. Santoro, and N. de Freitas, “Hyperbolic Attention Networks,” vol. i, pp. 1–15, 2018.
- [101] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *CoRR*, vol. abs/1409.1, 2014.
- [102] A. A. Ungar, *Analytic hyperbolic geometry: Mathematical foundations and applications*. World Scientific, 2005.
- [103] A. Ungar, *A Gyrovector Space Approach to Hyperbolic Geometry (Synthesis Lectures on Mathematics and Statistics)*. Morgan and Claypool Publishers, 2009.