Human-Like Rule Learning from Images Using One-Shot Hypothesis Derivation

Dany Varghese[1(✉)], Roman Bauer[1], Daniel Baxter-Beard[2], Stephen Muggleton[3], and Alireza Tamaddoni-Nezhad[1]

[1] Department of Computer Science, University of Surrey, Guildford, UK
{dany.varghese,r.bauer,a.tamaddoni-nezhad}@surrey.ac.uk
[2] School of Computing, Newcastle University, Newcastle upon Tyne, UK
[3] Department of Computing, Imperial College London, London, UK
s.muggleton@imperial.ac.uk

Abstract. Unlike most computer vision approaches, which depend on hundreds or thousands of training images, humans can typically learn from a single visual example. Humans achieve this ability using background knowledge. Rule-based machine learning approaches such as Inductive Logic Programming (ILP) provide a framework for incorporating domain specific background knowledge. These approaches have the potential for human-like learning from small data or even one-shot learning, i.e. learning from a single positive example. By contrast, statistics based computer vision algorithms, including Deep Learning, have no general mechanisms for incorporating background knowledge. This paper presents an approach for one-shot rule learning called One-Shot Hypothesis Derivation (OSHD) based on using a logic program declarative bias. We apply this approach to two challenging human-like computer vision tasks: 1) Malayalam character recognition and 2) neurological diagnosis using retinal images. We compare our results with a state-of-the-art Deep Learning approach, called Siamese Network, developed for one-shot learning. The results suggest that our approach can generate human-understandable rules and outperforms the deep learning approach with a significantly higher average predictive accuracy.

1 Introduction

Deep Neural Networks (DNNs) [2,10,19] have demonstrated state-of-the-art results on many pattern recognition tasks, especially in image classification problems [6,12]. However, recent studies [27] revealed major differences between human visual cognition and DNNs. For example, it is easy to produce images that are completely unrecognizable to humans, though DNN visual learning algorithms believe them to be recognizable objects with over 99% confidence [27]. Another major difference is related to the number of required training examples. Humans can typically learn from a single visual example [15], unlike statistical learning which depends on hundreds or thousands of images. Humans achieve

N. Katzouris and A. Artikis (Eds.): ILP 2021, LNAI 13191, pp. 234–250, 2022.
https://doi.org/10.1007/978-3-030-97454-1_17

this ability using background knowledge, which plays a critical role. By contrast, statistics based computer vision algorithms have no general mechanisms for incorporating background knowledge.

The performance of DNNs in medical imaging analysis and clinical risk prediction has been exceptionally promising. For example, recent evidence suggests that certain neurological pathologies such as Alzheimer's and Parkinson's disease can be associated with retinal abnormalities [5,21]. Nevertheless, despite the many opportunities DNNs present for healthcare, their clinical use remains limited as they suffer from the black-box situation. This constitutes a significant problem for clinicians: a lack of understanding of the inner workings of such methods renders it problematic to explain the diagnosis and treatment process to their patients. Moreover, DNNs usually require large training data which are not always available, for example in the diagnosis of neurological conditions, which are highly heterogeneous and often involve comorbidities.

In this paper, we present an approach for one-shot rule learning called One-Shot Hypothesis Derivation (OSHD) based on logic program declarative bias which is first introduced in [31]. We apply this approach to two different the challenging tasks; Malayalam character recognition & neurological diagnosis using retinal images. We have created a dataset for Malayalam hand-written characters which includes high level properties of the language based on the 'Omniglot' dataset designed for developing human-level concept learning algorithms [16]. We present character recognition as an extended work of [31]. For neurological diagnosis, we have collected the images from the UK Biobank [28] and extracted retinal vascular features (RVFs). We compare our results with a state-of-the-art Deep Learning approach, called Siamese Network [14], which is popular for one-shot learning.

2 One-Shot Hypothesis Derivation (OSHD)

In this paper we adopt a form of ILP which is suitable for one-shot learning and is based on using a logic program declarative bias, i.e. using a logic program to represent the declarative bias over the hypothesis space. Using a logic program, declarative bias has several advantages. Firstly, a declarative bias logic program allows us to easily port bias from one problem to another similar problem (e.g. for transfer learning). Secondly, it is possible to reason about the bias at the meta-level. Declarative bias will also help to reduce the size of the search space for the target concept or hypothesis derivation [1,26]. We refer to this approach as One-Shot Hypothesis Derivation (OSHD) which is a special case of Top-Directed Hypothesis Derivation (TDHD) as described in [23].

Definition 1 (One-Shot Hypothesis Derivation). *The input to an OSHD system is the vector* $S_{TDHD} = \langle NT, \top, B, E, e\rangle$ *where* NT *is a set of "non-terminal" predicate symbols,* $\top$ *is a logic program representing the declarative bias over the hypothesis space,* B *is a logic program representing the background knowledge and* E *is a set of examples and* e *is a positive example in* E*. The following three conditions hold for clauses in* $\top$*: (a) each clause in* $\top$ *must contain*

at least one occurrence of an element of NT while clauses in B and E must not contain any occurrences of elements of NT, (b) any predicate appearing in the head of some clause in ⊤ must not occur in the body of any clause in B and (c) the head of the first clause in ⊤ is the target predicate and the head predicates for other clauses in ⊤ must be in NT. The aim of an OSHD learning system is to find a set of consistent hypothesised clauses H, containing no occurrence of NT, such that for each clause h ∈ H the following two conditions hold:

$$\top \models h \tag{1}$$

$$B, h \models e \tag{2}$$

The following theorem is a special case of Theorem 1 in [23].

Theorem 1. *Given* $S_{OSHD} = \langle NT, \top, B, E, e\rangle$ *assumptions (1) and (2) hold only if there exists an SLD refutation R of* $\neg e$ *from* $\top, B$*, such that R can be re-ordered to give* $R' = D_h R_e$ *where* D_h *is an SLD derivation of a hypothesis h for which (1) and (2) hold.*

According to Theorem 1, implicit hypotheses can be extracted from the refutations of e. Let us now consider a simple example on learning a rule describing the properties of a concept (alphabet).

Example 1. Let $S_{OSHD} = \langle NT, \top, B, E, e\rangle$ where NT, B , e and $\top$ are as follows:

$NT = \{\$body\}$
$B = b_1 = \text{property1(a)} \leftarrow$
$e = \text{alphabet(a)} \leftarrow$

$$\top = \begin{cases} \top_1 : \text{alphabet}(X) \leftarrow \$\text{body}(X) \\ \top_2 : \$\text{body}(X) \leftarrow \text{property1}(X) \\ \top_3 : \$\text{body}(X) \leftarrow \text{property2}(X) \end{cases}$$

Given the linear refutation, $R = \langle \neg e, \top_1, \top_2, b_1\rangle$, we now construct the re-ordered refutation $R' = D_h R_e$ where $D_h = \langle \top_1, \top_2\rangle$ derives the clause $h =$ alphabet(X) ← property1(X) for which (1) and (2) hold.

The user of OSHD can specify a declarative bias ⊤ in the form of a logic program. A general ⊤ theory can be also generated from user specified mode declarations. Figure 1 represents a simplified example of user specified mode declarations and the automatically constructed ⊤ theory.

The OSHD Learning algorithm can be described in 3 main steps:

1. Generate all hypotheses, H_e that are generalizations of e
2. Compute the coverage of each hypothesis in H_e
3. Build final theory, T, by choosing a subset of hypotheses in H_e that maximises a given score function (e.g. compression).

In step 1, H_e is generated using the OSHD hypothesis derivation described earlier in this section.

The second step of the algorithm, computing the coverage of each hypothesis, is not needed if the user program is a pure logic program (i.e. all relationships in the background knowledge are self contained and do not rely on Prolog built-in predicates). This is because, by construction, the OSHD hypothesis derivation

modeh(alphabet(+image)).
modeb(has_prop1(+image)).
modeb(has_prop2(+image)).

$$\top = \begin{cases} \top_1 : \text{alphabet}(X) \leftarrow \$\text{body}(X). \\ \top_2 : \$\text{body}(X) \leftarrow \\ \top_3 : \$\text{body}(X) \leftarrow \text{has\_prop1}(X), \$\text{body}(X). \\ \top_4 : \$\text{body}(X) \leftarrow \text{has\_prop2}(X), \$\text{body}(X). \end{cases}$$

Fig. 1. Mode declarations and a $\top$ theory automatically constructed from it

generates all hypotheses that entail a given example with respect to the user supplied mode declarations. This implies that the coverage of a hypothesis is exactly the set of examples that have it as their generalization. However, this coverage computation step is needed for the negative examples, as they were not used to build the hypothesis set.

For step 3, the compression-based evaluation function used for the experiments in this paper is:

$$\sum Covered\_Examples\_Weight - Total\_Literals \tag{3}$$

The weight associated to an example may be defined by the user but by default, positive examples have weight 1 and negative examples weight -1. In general, negative examples are defined with a weight smaller than 0 and positive examples with a weight greater than 0.

3 Siamese Neural Networks

In this paper, we use a state-of-the-art Deep Learning approach, called Siamese Network [14], which has been developed for one-shot learning. The original Siamese Networks were first introduced in the early 1990s by Bromley and LeCun to solve signature verification as an image matching problem [4]. A Siamese network is a Deep Learning architecture with two parallel neural networks with the same properties in terms of weight, layers etc. Each network takes a different input, and their outputs are combined using an energy function at the top to provide some prediction. The energy function computes a metric between the highest level feature representation on each side (Fig. 2). Weight tying guarantees that two extremely similar images could not possibly be mapped by their respective networks to very different locations in feature space because each network computes the same function. Also, the network is symmetric, so that whenever we present two distinct images to the twin networks, the top conjoining layer will compute the same metric as if we were to present the same two images but to the opposite twin.

A Siamese network model is defined using a convolutional neural network (CNN), mainly developed to work with image data. CNNs can be considered as regularized versions of multilayer perceptrons. Multilayer perceptrons usually mean fully connected networks; each neuron in one layer is connected to all neurons in the next layer. We stack three layers: the convolutional Layer, pooling Layer, and fully connected layer to form CNN.

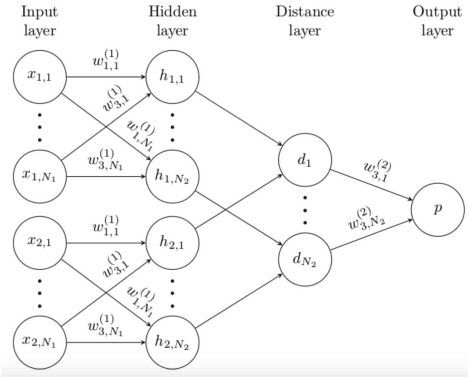


Fig. 2. A simple 2 hidden layer Siamese Neural Network [14]

- **Convolutional layers** apply a convolution operation to the input, passing the result to the next layer. A convolution converts all the pixels in its receptive field into a single value. For example, if one apply a convolution to an image, the image size is decreased as well as information in individual fields are brought together into a single pixel. The final output of the convolutional layer is a vector.
- **Pooling layers** reduce the dimensions of data by combining the outputs of neuron clusters at one layer into a single neuron in the next layer. Local pooling combines small clusters. There are two common types of pooling in popular use: max and average. Max pooling uses the maximum value of each local cluster of neurons in the feature map, while average pooling takes the average value.
- **Fully connected layers** connect every neuron in one layer to every neuron in another layer. It is the same as a traditional multi-layer perceptron neural network (MLP). The flattened matrix goes through a fully connected layer to classify the images.

Koch et al. [14] use a convolutional neural siamese architecture to classify pairs of omniglot images. In the experiments of this paper, we have adopted the same model as defined in [14]. Our standard model is a Siamese neural network with L fully-connected layers each with N_l units, where $h_{1,l}$ represents the hidden vector in layer l for the first twin, and $h_{2,l}$ denotes the same for the second twin. We use exclusively rectified linear (ReLU) units in the first $L-1$ layers. After the $(L-1)^{th}$ feed-forward layer, we compare the features computed by each twin via a fixed distance function.

4 One-Shot Learning for Malayalam Character Recognition

We apply OSHD as well as Deep Learning (i.e. Siamese Network) to the challenging task of one-shot Malayalam character recognition. This is a challenging task due to spherical and complex structure of Malayalam hand-written language.

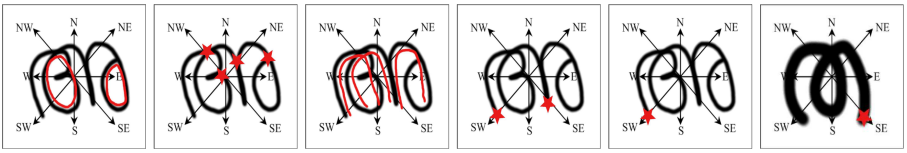
4.1 Character Recognition and Human-Like Background Knowledge

Malayalam is one of the four major languages of the Dravidian language family and originated from the ancient Brahmi script. Malayalam is the official language of Kerala, a state of India with roughly forty-five million people. Unlike for other languages, there is currently no efficient algorithm for Malayalam handwritten recognition. The handwriting recognition for Malayalam script is a major challenge compared to the recognition of other scripts because of the following reasons: presence of large number of alphabets, different writing styles, spherical features of alphabets, and similarity in character shapes. We selected the handwritten characters from the 'Omniglot' dataset [16]. Sample Malayalam alphabets from our dataset are shown in Fig. 3 (characters 'Aha' and 'Tha'). Feature extraction is conducted utilizing a set of advanced geometrical features [30] and directional features.

Geometrical Features. Every character may be identified by its geometric designations such as loops, junctions, arcs, and terminals. Geometrically, a loop means a closed path. Malayalam characters contain more intricate loops which may contain some up and downs within the loops itself. So we follow a concept as shown in Fig. 3(b). If the figure has a continuous closed curve then we will identify it as a loop. Junctions may be defined as a meeting point of two or more curves or line. It is easy for human to identify the junction from an image as shown in Fig. 3(c). As per dictionary definitions, an arc is a component of a curve. So in our case, a path with semi opening will be considered as an arc. Please refer to Fig. 3(d) for more details. Terminals may be classified as points where the character stroke ends, i.e. no more connection beyond that point. Figure 3(e) is a self-explanatory example for the definition.

We have included the visual explanation for the geometrical feature extraction in Fig. 3. We have selected two characters to explicate the features as shown in Fig. 3 and marked each geometrical features as we discussed. Table 1 will give an abstract conception about the dataset we have developed for the experiments from the 'Omniglot' dataset.

Directional Features. Every character may be identified by its directional specifications such as starting and ending points of the stroke. There are certain unwritten rules for Malayalam characters, e.g. the writing always commences from left and moves towards the right direction. Native Malayalam users can



(a) Loop (b) Junction (c) Arc (d) Terminals (e) Start Point (f) End Point

Fig. 3. Human-like feature extraction criteria

easily identify the starting and ending point. However, we will need to consider the starting and ending point as features so that these can be easily identified without semantic knowledge of a character. The starting and ending points are determined by standard direction properties. Figure 3(f) will give you an idea about developing the directional features from an alphabet. As we discussed, a user can identify both starting and ending point of the character displayed in Fig. 3(f) easily whereas the terminus point of Fig. 3(e) is arduous to determine.

4.2 Mode Declarations

In this section, we define the OSHD specific details of the declarative bias, defined by mode declaration and background knowledge representation used in our experiments. The first step was to develop and represent the background knowledge based on the concepts described in Sect. 4.1. Table 1 shows the geometrical and directional features of 5 different alphabets. Here, we use the same notations used in Progol [22] and Toplog [23].

For example, in our first experiment, *alphabet*(+*character*) is the head of the hypothesis, where +*character* defines the character identifier *character* as an input argument. We are using four predicates in the body part of the hypothesis as shown in the Listing 1.1. Note that +, −, # indicate input, output or a constant value arguments.

Listing 1.1. Mode declarations

```
:- modeh(1,alphabet(+character)).
:- modeb(*,has_gemproperties(+character,-properties)).
:- modeb(*,has_gemproperties_count(+properties,
                                   #geo_feature_name,#int)).
:- modeb(*,has_dirproperties(+character,-properties)).
:- modeb(*,has_dirproperties_feature(+properties,
                                #dir_feature_name,#featurevalue)).
```

The meaning of each *modeb* condition is defined as follows:

has_gemproperties/2 predicate was used to represent the geometrical features as defined in Table 1. The input argument *character* is the unique identifier for an alphabet, *properties* refers to the property names.

has_gemproperties_count/3 predicate outlines the count of the particular feature associated with the alphabet. The *properties* indicate unique identifiers for a

Table 1. Geometrical and directional properties

Character ID	Geometrical properties				Directional properties	
	No. loops	No. junctions	No. arcs	No. terminals	Starting point	Ending point
1	2	4	3	2	sw	null
2	3	4	3	2	sw	null
3	3	4	3	2	sw	null
4	1	2	3	2	null	se
5	1	3	3	2	nw	se

particular geometrical property of a particular alphabet, *geo_feature_name* refers to the property name and *int* stands for the feature count.

has_dirproperties/2 predicate used to represent the directional features mentioned in Table 1. The *character* is the unique identifier for the alphabet, *properties* refers to the property names.

has_dirproperties_count/3 predicate outlines the count of a particular directional feature associated with the alphabet. The *properties* is a unique identifier for a particular property of a particular alphabet, *dir_feature_name* refers to the property name and *featurevalue* stands for the feature vale.

5 One-Shot Learning for Neurological Diagnosis Using Retinal Images

Recent evidence suggests that certain neurological pathologies such as Alzheimer's and Parkinson's disease can be associated with retinal abnormalities [5,21]. Retinal fundus imaging and optical coherence tomography (OCT), powerful techniques for imaging the retina and highly informative from a clinical perspective [32], could therefore potentially be used for the early detection and diagnosis of these diseases. This is interesting also from a practical perspective, because retinal fundus photography and OCT are non-invasive, cheap, quick to perform, highly sensitive and specific. Moreover, these methods provide digital outputs that can be easily stored and analysed with modern AI tools. Indeed, a recent study shows relatively high accuracy in the usage of AI for the diagnosis of Alzheimer's disease from retinal fundus images [29]. However, numerous challenges remain to be addressed, such as regarding the capability for humans to understand the model, as well as the significant requirements for large amounts of data.

Here, we use our proposed OSHD approach for model construction and revision in retinal fundus image analysis. This approach has several benefits: it provides human-understandable reasoning that enable the medical practitioner to explain to the patient the computational diagnosis. Moreover, the OSHD approach allows for the incorporation of patient-specific information (e.g. age, weight, lifestyle, etc.) and so supports personalised treatments. Furthermore, the OSHD approach requires much fewer samples for training than state-of-the-art

DNN techniques. This is a highly valuable feature particularly for the diagnosis of neurological conditions, which are highly heterogeneous and often involve comorbidities. Hence, the most relevant and well-suited datasets are usually of limited size. The capability to make efficient use of limited datasets is therefore paramount for such computational methods. We here present a preliminary demonstration for this proposed approach.

5.1 Retinal Vasculature Features

Table 2. Retinal Vascular Features (RVFs) with the retinal zone of interest

Parameter	Description	Retinal zone
CRAE	Central Retinal Arteriolar Equivalent	B
CRVE	Central Retinal Venular Equivalent	B
AVR	Arteriole-Venular ratio	B
FDa	Fractal Dimension arteriole	C
FDv	Fractal Dimension venular	C
BSTDa	Zone B Standard Deviation arteriole	B
BSTDv	Zone B Standard Deviation venular	B
TORTa	Tortuosity arteriole	C
TORTv	Tortuosity venular	C

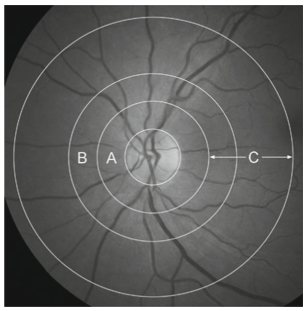


Fig. 4. Retinal zones considered in this study. Three concentric zones were introduced for the computation of retinal fundus image features, i.e. zones A, B and C [7].

Our selection of retinal vascular features (RVFs) for the analysis of the retinal images is based on the study by Frost et al. [7] where significant differences were found in specific RVFs between healthy controls and subjects with Alzheimer's disease. To simplify the derivation of a diagnostic algorithm, we decided to use only a subset of these features in our experiments. In particular, Frost et al. mentioned 13 RVF's in their experiment, while we selected 9 relevant and well-established features for our experiments. These selected RVFs are listed in Table 2. Briefly, the RVFs were measured based on the width and branching geometry of retinal vessels. In accordance with [7], we introduced concentric zones of interest A, B and C to compute the RVFs. Notably, the feature analysis did not incorporate information from zone A (region from 0 to 0.5 disc diameters away from the disc margin). The analysed two zones were 0.5-1.0 disc diameters away from the disc margin (zone B, Fig. 4) and 0.5-2.0 disc diameters away from the disc margin (zone C, Fig. 4). Restricting the measurements to these two zones ensured that the vessels had attained arteriolar status. The measured zone for each feature is listed in Table 2.

5.2 Retinal Fundus Image Dataset

The participants for this study were selected from the UK Biobank resource [28]. The UK Biobank recruited 500,000 people of ages between 40 and 69 to undergo a variety of tests and have their health followed throughout their lives. Notably, only a subset of these participants had their retinas imaged (in total 84,767). Retinal imaging was conducted using the TOPCON 3D OCT 1000 Mk2 which simultaneously performs OCT and takes a fundus photograph. The images produced are centred on the macula, have a 45° field of view and have the dimensions 2,048 by 1,536.

5.3 Extraction of Subject Data and Background Knowledge (BK) Preparation

Fig. 5. Demonstration of processing steps for vessel segmentation and artery *vs.* vein classification (Color figure online)

Extraction of Subject-Specific Clinical Information. The vast majority of the information about the participants was collated in the form of a large CSV (Comma Separated Values) file, where each row represents a participant and each column represents a data point. The UK Biobank online system provides explanations for the codes used for the column names and the associated data. Diagnoses in the dataset were encoded according to the International Classification of Diseases, Tenth Revision (ICD-10). A detailed analysis of the participant data file yielded 18 Alzheimer's, 133 Parkinson's and 54 vascular dementia sufferers who satisfied these two conditions: they had (1) fundus images taken, and (2) were diagnosed with exactly one of these three conditions. In addition to images from these diseased subjects, we used images of 528 subjects that were healthy with respect to these three conditions. Notably, we used only fundus images of the left eye.

Optic Disc Localisation. We followed an approach that uses pyramidal decomposition based on the Haar-discrete wavelet transform in order to localize the optic disc to certain regions. Template matching is then employed to further localize the optic disc. This particular approach is detailed by Lalonde et al. [17].

Artery/Vein Classification. The method we have used for artery/vein classification was developed by Galdran et al. [8]. In particular, we use a fully CNN to classify pixels into four separate categories: background, vein, artery or uncertain. Figure 5 shows the segmentation and artery/vein classification steps applied to a sample image. In the image, arteries are shown in red, veins in blue, and areas of uncertainty are coloured in green.

Vascular Feature Extraction. Retinal vascular features mentioned in Table 2 were extracted. To this end, vascular calibres were calculated for the six most extensive arterioles and six largest venules. The standard deviation of the width in zone B (BSTD) was calculated for the arteriolar and venular networks. Summary measures of vascular equivalent calibre were also calculated (central retinal arterial (CRAE) and venular (CRVE) equivalent calibre), based on the improved Knudston-Parr-Hubbard formula [11,13]. CRAE and CRVE represent the equivalent single-vessel parent calibre (width) for the six arterioles and venules, respectively. From these indices, the arteriole-to-venule ratio (AVR) was calculated (AVR = CRAE/CRVE). Natural patterns such as vessel networks often exhibit fractal properties, whereby they appear the same when viewed over a range of magnifications. The fractal dimension (FD) describes the range of scales over which this self-similarity is observed. In this study, the fractal dimension of the retinal vascular network was calculated using the box-counting method [20]. Larger values reflect a more complex branching pattern. Retinal vascular tortuosity is defined as the integral of the curvature squared along the path of the vessel, normalized by the total path length [9]. All vessels in the zone of interest with a width > 40 μm were measured. The estimates were summarized as the average tortuosity of the measured vessels. A smaller tortuosity value indicates straighter vessels.

Data Encoding and Background Knowledge Preparation. The RVF's from the measurement extraction step will be stored. During the data labelling step, the distribution of each RVF is analysed and classified into three categories: low, medium, and high. The next step is to prepare background knowledge from these labelled data and will be done according to the *mode* definitions. Sample background knowledge for a patient is displayed in Listing 1.2.

Listing 1.2. Sample BK for a retinal image using labelled RVF's

```
cr_arteriolar_equivalent(patientid_0,high).
cr_venular_equivalent(patientid_0,low).
av_ratio(patientid_0,high).
sd_arteriole(patientid_0,medium).
sd_venular(patientid_0,low).
fd_arteriole(patientid_0,low).
fd_venular(patientid_0,medium).
tortuosity_arteriole(patientid_0,low).
tortuosity_venular(patientid_0,low).
```

6 Experiments

In this section we evaluate the OSHD approach on two different applications. The first application is complex character recognition, which is the extension of an initial study presented in [31]. We have used 46 alphabets (46 classes) from the Malayalam language, and for each alphabet, 20 different handwritten images are selected and divided equally into training and test sets.

The second application is a novel diagnosis approach for three neurodegenerative diseases, namely vascular dementia, Parkinson's disease and Alzheimer's disease. We have used four different classes, as mentioned in Sect. 5.3, and each class contains 18 images, divided equally as training and test sets. Notably, this study is highly innovative as it investigates the usage of modern machine learning based analysis of retinal images for the classification of neurodegenerative diseases. In this section we test the following null hypotheses:

Null Hypothesis 1 OSHD cannot outperform Siamese Networks in one-shot learning for complex character recognition.

Null Hypothesis 2 OSHD cannot outperform Siamese Networks in one-shot learning for neurodegenerative disease identification.

Null Hypothesis 3 OSHD cannot learn human comprehensible rules for either of applications in Null Hypotheses 1 and 2.

6.1 Materials and Methods

The OSHD algorithm in this experiment is based on Top-Directed Hypothesis Derivation implemented in Toplog [23], and uses mode declarations and background knowledge which defined earlier in this paper. The Siamese Network used in the experiment is based on the implementation described in [14].

The data, codes and configuration input files used in the experiments in this section are available from: https://github.com/danyvarghese/One-Shot-ILP.

We endeavoured to reiterate the same concept of working with both architectures and repeated the experiments for different numbers of folds. Each fold consists of a single positive example and n negative examples. In our experiments, we use the term 'number of classes' to indicate the total number of negative examples (in addition to the one positive example) used for the cross-validation.

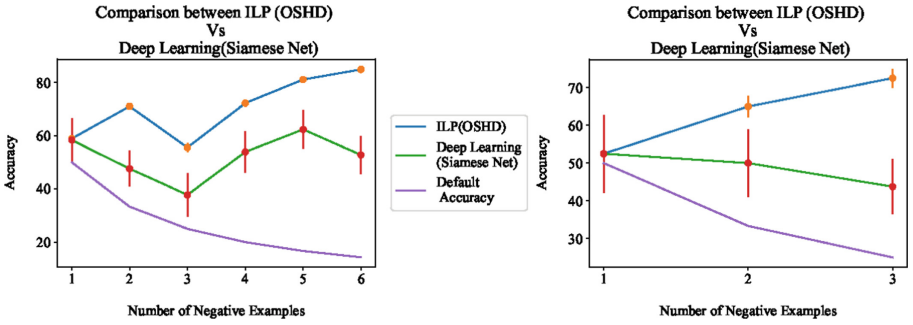
In order to reject the Null Hypotheses 1 and 2, we adapted an experimental setting used for one-shot learning by Siamese Networks as in [14]. We follow a '20-way N-class' experimental setting for each dataset. In this setting, 20 is the number of runs and N is the number of classes (N varies from 2 to 7 for character recognition and 2 to 4 for neurodegenerative diseases). As we are doing one-shot learning, we use one positive example at each run with negative examples which are from the other classes (the X axis in the learning curves of Fig. 6). In the following, we define specific parameter settings for each algorithm.

OSHD Parameter Settings. The following Toplog parameter settings were used in this experiment: *clause_length* (value = 15) defines the maximum number of literals (including the head) of a hypothesis, $evalfn$ (Value = compression) defines which function to use when scoring a clause. The default scoring function is compression. *positive_example_inflation* multiplies the weights of all positive examples by this factor. This parameter is set to 10 and 5 for character recognition and the neurodegenerative dataset, respectively, and negative example inflation multiplies the weights of all negative examples by this factor. This parameter is set to 5 for both datasets.

Siamese Networks Parameter Settings. For the implementation of the Siamese Network, we followed the same setups used by Koch et al. [14]. Koch et al. use a convolutional Siamese network to classify pairs of 'Omniglot' images, so the twin networks are both CNNs. The twins each have the following architecture: convolution with 64 (10×10) filters, 'max_pooling' convolution 128 (7×7), 'max_pooling' convolution 128 (4×4) filters and 'max_pooling' convolution 256 (4×4) filters, all with 'relu' activation functions. The twin networks reduce their inputs down to smaller and smaller 3D tensors. Finally, there is a fully connected layer with 4096 units.

In most implementations of Siamese Networks, the training model is developed using a large amount of data. Also, particularly in the case of character recognition, characters from one language are usually compared against the characters from other languages [3,18]. In contrast, in our experiments we only consider alphabets from a single language. This is advantageous because it renders the training process less data-demanding, as well as requires less expertise in choosing suitable comparison languages.

6.2 Results and Discussions



(a) Character Recognition (b) Neurological Diagnosis

Fig. 6. Average Predictive accuracy of ILP (OSHD) vs Deep Learning (Siamese Net)

Figure 6(a) shows the average predictive accuracy of ILP (OSHD) vs Deep Learning (Siamese Net) in One-shot character recognition with increasing number of negative examples (number of different classes). As shown in this figure, Siamese Net and OSHD have almost same accuracy for class 2. In all other cases, OSHD outperforms Siamese Net with significant difference. Figure 6(b) shows the average predictive accuracy of ILP (OSHD) vs Deep Learning (Siamese Net) in the neurodegenerative diseases with increasing numbers of negative examples. As in the character recognition dataset, Siamese net and OSHD have a similar accuracy for class 2. In all other cases, OSHD outperforms Siamese net with a significant difference. We have also used a T-test to check the validity of the null hypothesis. For the CR experiment, we obtained 2.813 as the t-value and 0.008 as the p-value, whereas we obtained 3.658 and 0.001 for the neurodegenerative dataset experiment, respectively. The T-test as well as the predictive accuracy reject the null hypotheses 1 and 2.

Due to the involved quantities and complexities, recording and displaying all the learned rules is beyond the scope of this work. However, we have shown some random learned rules generated during the experiments in Table 3, which allows to assess hypothesis 3. The rules from Table 3 suggest that OSHD can lean rules which are human comprehensible and can be easily communicated with domain experts, so null hypothesis 3 is also rejected.

Table 3. Example of learned rules

Character recognition	Neurological diagnosis
alphabet(A) :-	
has_gemproperties(A, B),	diagnosis(A,parkinson) :-
has_gemproperties(A, C),	fd_venular(A,medium),
has_gemproperties_count(B, junctions, 4),	cr_arteriolar_equivalent(A,medium),
has_gemproperties(A, D)	tortuosity_arteriole(A,low)

7 Conclusion

In this paper, we presented a novel approach for one-shot rule learning called One-Shot Hypothesis Derivation (OSHD) [31] that is based on using a logic program declarative bias. We applied this approach to two challenging computer vision tasks: 1) Malayalam character recognition and 2) neurological diagnosis from retinal images. The features used to express the background knowledge for character recognition were developed in such a way that it is well-suited for human visual cognition also. We demonstrated the learning of rules for each character which is more natural and in accordance with human visual understanding. For neurological diagnosis, we collected and quantified several retinal vascular features. We compared our results with a state-of-the-art Deep Learning approach, called Siamese Network, which has been developed for one-shot learning.

The results suggest that our approach can generate human-understandable rules and also outperforms the deep learning approach with a significantly higher average predictive accuracy. Its was clear from the results that deep learning paradigm needs more data and its efficiency is decreased when dealing with a small amount of data. As future work we would like to further extend the background knowledge for the nueurological diagnosis problem to include more semantic information and also explore the framework of Meta-Interpretive Learning (MIL) [24] for the purpose of one-shot learning from images [25].

Acknowledgements. Dany Varghese was supported by Vice Chancellor's PhD Scholarship Award at the University of Surrey. Roman Bauer was supported by the Engineering and Physical Sciences Research Council of the United Kingdom (EP/S001433/1). The authors also thankfully acknowledge that the retinal images were obtained using UK Biobank application number 1969. Alireza Tamaddoni-Nezhad and Stephen Muggleton were supported by the EPSRC Network Plus grant on Human-Like Computing (HLC).

References

1. Adé, H., Raedt, L.D., Bruynooghe, M.: Declarative bias for specific-to-general ILP systems. Mach. Learn. **20**, 119–154 (1995)
2. Bennett, C.H., et al.: Contrasting advantages of learning with random weights and backpropagation in non-volatile memory neural networks. IEEE Access **7**, 73938–73953 (2019)
3. Bouma, S.: One shot learning and Siamese networks in Keras (2017). https://sorenbouma.github.io/blog/oneshot/
4. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a "Siamese" time delay neural network. In: Proceedings of the 6th International Conference on Neural Information Processing Systems, pp. 737–744 (1993)
5. Cheung, C.Y.l., Ikram, M.K., Chen, C., Wong, T.Y.: Imaging retina to study dementia and stroke. Progr. Retinal Eye Res. **57**, 89–107 (2017)
6. Varghese, D., Shankar, V.: A novel approach for single image super resolution using statistical mathematical model. IJAER **10**(44) (2015)
7. Frost, S., Kanagasingam, Y., Sohrabi, H., Vignarajan, J., Bourgeat, P., et al.: Retinal vascular biomarkers for early detection and monitoring of Alzheimer's disease. Transl. Psychiatry **3**, e233 (2013)
8. Galdran, A., Meyer, M., Costa, P., MendonÇa, Campilho, A.: Uncertainty-aware artery/vein classification on retinal images. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 556–560 (2019)
9. Hart, W.E., Goldbaum, M., Côté, B., Kube, P., Nelson, M.R.: Measurement and classification of retinal vascular tortuosity. Int. J. Med. Inform. **53**(2), 239–252 (1999)
10. Hinton, G.: Learning multiple layers of representation. Trends Cogn. Sci. **11**, 428–434 (2007)
11. Hubbard, L.D., Brothers, R.J., King, W.N., et al.: Methods for evaluation of retinal microvascular abnormalities associated with hypertension/sclerosis in the atherosclerosis risk in communities study. Ophthalmology **106**(12), 2269–2280 (1999)

12. Neethu, K.S., Varghese, D.: An incremental semi-supervised approach for visual domain adaptation. In: 2017 International Conference on Communication and Signal Processing (ICCSP), pp. 1343–1346 (2017)
13. Knudtson, M., Lee, K.E., Hubbard, L., Wong, T., et al.: Revised formulas for summarizing retinal vessel diameters. Curr. Eye Res. **27**, 143–149 (2003)
14. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. In: Proeedings of International Conference on Machine Learning, vol. 37 (2015)
15. Lake, B., Salakhutdinov, R., Gross, J., Tenenbaum, J.: One shot learning of simple visual concepts. In: Proceedings of the 33rd Annual Conference of the Cognitive Science Society, pp. 2568–2573 (2011)
16. Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. Science **350**(6266), 1332–1338 (2015)
17. Lalonde, M., Beaulieu, M., Gagnon, L.: Fast and robust optic disc detection using pyramidal decomposition and Hausdorff-based template matching. IEEE Trans. Med. Imaging **20**, 1193–200 (2001)
18. Lamba, H.: One shot learning with Siamese networks using Keras (2019). https://towardsdatascience.com/one-shot-learning-with-siamese-networks-using-keras-17f34e75bb3d
19. Liu, X., He, P., Chen, W., Gao, J.: Multi-task deep neural networks for natural language understanding. CoRR 1901.11504 (2019)
20. Mainster, M.: The fractal properties of retinal vessels: embryological and clinical implications. Eye **4**, 235–241 (1990)
21. McGrory, S., Taylor, A.M., Kirin, M., et al.: Retinal microvascular network geometry and cognitive abilities in community-dwelling older people: the Lothian birth cohort 1936 study. Ophthalmology **101**(7), 993–998 (2017)
22. Muggleton, S.: Inverse entailment and Progol. N. Gener. Comput. **13**, 245–286 (1995)
23. Muggleton, S.H., Santos, J.C.A., Tamaddoni-Nezhad, A.: TopLog: ILP using a logic program declarative bias. In: Garcia de la Banda, M., Pontelli, E. (eds.) ICLP 2008. LNCS, vol. 5366, pp. 687–692. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89982-2_58
24. Muggleton, S., Lin, D., Tamaddoni-Nezhad, A.: Meta-interpretive learning of higher-order dyadic datalog: predicate invention revisited. Mach. Learn. **100**(1), 49–73 (2015)
25. Muggleton, S., Dai, W.Z., Sammut, C., Tamaddoni-Nezhad, A.: Meta-interpretive learning from noisy images. Mach. Learn. **107** (2018)
26. Nedellec, C.: Declarative bias in ILP (1996)
27. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 427–436 (2015)
28. Sudlow, C., et al.: UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PloS Med. (2015)
29. Tian, J., Smith, G., Guo, H., Liu, B., Pan, Z., et al.: Modular machine learning for Alzheimer's disease classification from retinal vasculature. Sci. Rep. **11**(1), 1–11 (2021)
30. Usman Akram, M., et al.: Geometric feature points based optical character recognition. In: 2013 IEEE Symposium on Industrial Electronics Applications, pp. 86–89 (2013)

31. Varghese, D., Tamaddoni-Nezhad, A.: One-shot rule learning for challenging character recognition. In: Proceedings of the 14th International Rule Challenge, CEUR, 2020, vol. 2644, pp. 10–27 (2020)
32. Zapata, M.A., Royo-Fibla, D., Font, O., Vela, J.I., et al.: Artificial intelligence to identify retinal fundus images, quality validation, laterality evaluation, macular degeneration, and suspected glaucoma. Clin. Ophthalmol. **14**, 419 (2020)