# Ontology-based $n$-ball Concept Embeddings Informing Few-shot Image Classification

Mirantha Jayathilaka[1], Tingting Mu[1] and Uli Sattler[1]

[1]Departmeent of Computer Science, The University of Manchester, UK

**Abstract**

We propose a novel framework named ViOCE that integrates ontology-based background knowledge in the form of $n$-ball concept embeddings into a neural network based vision architecture. The approach consists of two main components - converting symbolic knowledge of an ontology into continuous space by learning $n$-ball embeddings that capture properties of subsumption and disjointness, and guiding the training and inference of a vision model using the learnt embeddings. We evaluate ViOCE using the task of few-shot image classification, where it demonstrates superior performance on two standard benchmarks.

**Keywords**

Background Knowledge, Ontology, Machine Learning, Few-shot Learning.

## 1. Introduction

This study sheds light on the use of ontologies in a machine learning context. In the proposed ViOCE framework, we adopt a technique to embed ontology-based knowledge as $n$-balls inspired by the work done by Kulmanov et al. [2]. This embedding can represent specialisations (e.g., Dog SubclassOf Animal) using the property of one $n$-ball enclosing another and partonomies (e.g., Dog hasPart Tail) using translations of $n$-ball positions. In this study, we directly utilise two loss design components of [2] to capture subsumption and disjointness axioms, while extending their approach with more regularisation components in order to embed large hierarchies in a favourable manner for a downstream vision task. Additionally, we propose the use of the inferred class hierarchy of the input ontology during the embedding learning process. The learnt $n$-ball embeddings can be seen as definitions of space for each concept in consideration, that preserves the inferred class hierarchy entailed by the ontology. Next, we introduce a method to use a vision model [3, 4] to map input images to the space defined by the concept embeddings, informing the vision task with the knowledge captured from the ontology during a few-shot image classification task [5]. Few-shot learning in an image classification context focuses on effectively learning the visual features of a class with very few examples. Figure 1 shows a snapshot of a few predictions for some miniImageNet[1] classes 'miniature poodle', 'hotdog' and 'street sign' made by a model trained using the ViOCE framework.

*15th International Workshop on Neural-Symbolic Learning and Reasoning (NeSy'2021)*

mirantha.jayathilaka@manchester.ac.uk (M. Jayathilaka); tingting.mu@manchester.ac.uk (T. Mu); uli.sattler@manchester.ac.uk (U. Sattler)

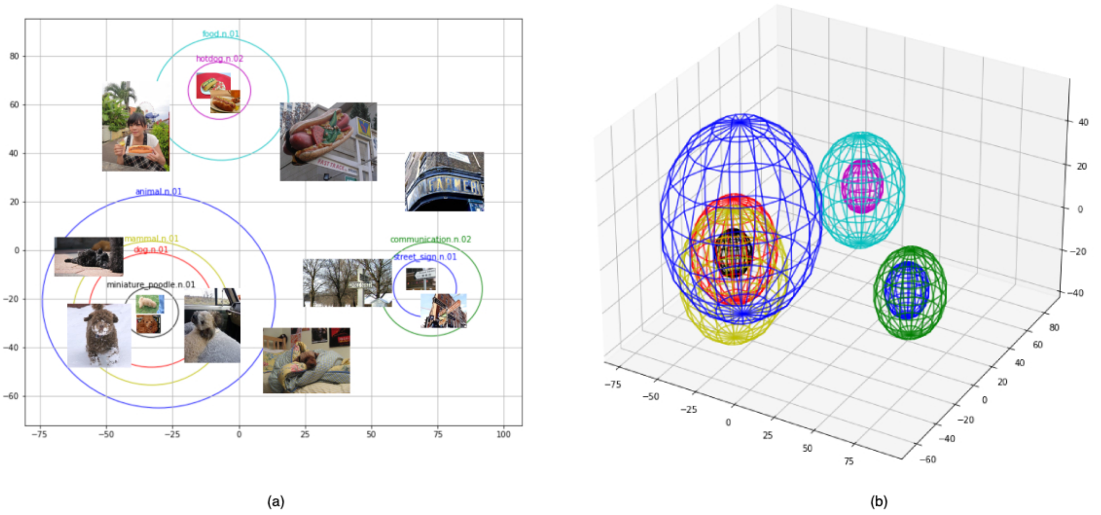


**Figure 1:** The proposed approach classifies images by projecting them towards concept $n$-balls defined in high-dimensional space according to ontology-based background knowledge. (a) shows a snapshot of a few predictions for three miniImageNet[1] classes 'miniature poodle', 'hotdog' and 'street sign' made by a model trained using the ViOCE framework during few-shot image classification. The dimensionality of the $n$-balls is reduced to 2 for visualisation purposes. A correct prediction is an image projected to be inside of the $n$-ball of its ground truth label. (b) is a visualisation of the same set of $n$-balls in (a) reduced to a 3-dimensional space in order to provide a clearer idea on the nature of $n$-ball shape and placing.

Overall, we extend an approach to capture knowledge from an ontology in the form of $n$-ball embeddings and show that they are favourable for the downstream vision task of few-shot image classification. We propose a technique to utilise the $n$-balls to guide a vision model during its training and inference stages.

## 2. Related Work

An area that inspires the investigation of background knowledge integration in vision is knowledge-based vision systems [6]. The choice of knowledge form can be very much based on the considered vision application, as pointed out in [6], where the authors curate a number of vision tasks along with the forms of knowledge used to inform the learning process. Out of these, the use of scene graphs, probabilistic ontologies and first-order logic rules grab the attention as promising paths to explore. Investigation into the use of background knowledge in the form of first-Order Logic (FOL) is found as well [7]. As shown in [7], adaptation of logical knowledge as constraints during the learning process has generated promising results, that reinforces the attempts to use ontologies as background knowledge. The area of neuro-symbolic approaches also provides insights into the use of logical knowledge during the training of artificial neural networks [8].

Studies such as [3, 4, 9] show mapping of image features to a vector space defined by language embeddings [10]. In the case of [3], the knowledge from an unstructured text corpus is captured in the form of word embeddings to be integrated to the vision architecture. These approaches were mostly evaluated on zero-shot image classification, making use of the distance between points in the vector space defined. In terms of few-shot learning [11, 12], our study is motivated by metric learning methods [13] due of their ability to extend standard vision architectures [14]. These approaches exploit image feature similarities [15] when learning and predicting a vision task.

## 3. *n*-balls and EL Embeddings

The mathematical concept of ball refers to the volume space bounded by a sphere and is also called a solid sphere. An $n$-ball usually refers to a ball in an $n$-dimensional Euclidean space. The EL embeddings study [2] attempts to encode logical axioms by positioning $n$-balls. We explain how it works for encoding subsumption and disjointness as they are the most relevant to our work. Each concept $P$ is embedded as an $n$-ball with its centre denoted by $c_P \in \mathbb{R}^n$ and the radius by $r_P \in \mathbb{R}$. The basic idea is to move one ball inside the other for subsumption and to push two $n$-balls to stay away for disjointness. The following loss is minimised to encode $\mathcal{O} \models P \sqsubseteq Q$:

$$\begin{aligned} l_{P\sqsubseteq Q}(c_P, c_Q, r_P, r_Q) = &\max(0, \|c_P - c_Q\|_2 + r_P - r_Q - \gamma) \\ &+ \big|\|c_P\|_2 - 1\big| + \big|\|c_Q\|_2 - 1\big|, \end{aligned} \tag{1}$$

where $\|\cdot\|_2$ denotes the $l_2$ norm and $\gamma \in \mathbb{R}$ is a user-set hyperparameter. It enforces the inequality $\|c_P - c_Q\|_2 \leq r_Q - r_P + \gamma$, meanwhile regulates the ball centres to be close to a unit sphere. Through controlling the sign of $\gamma$, the user can adjust whether to push the $P$ ball completely inside the $Q$ ball. In a similar fashion, the loss for encoding $\mathcal{O} \models P \sqcap Q \sqsubseteq \bot$ is given as

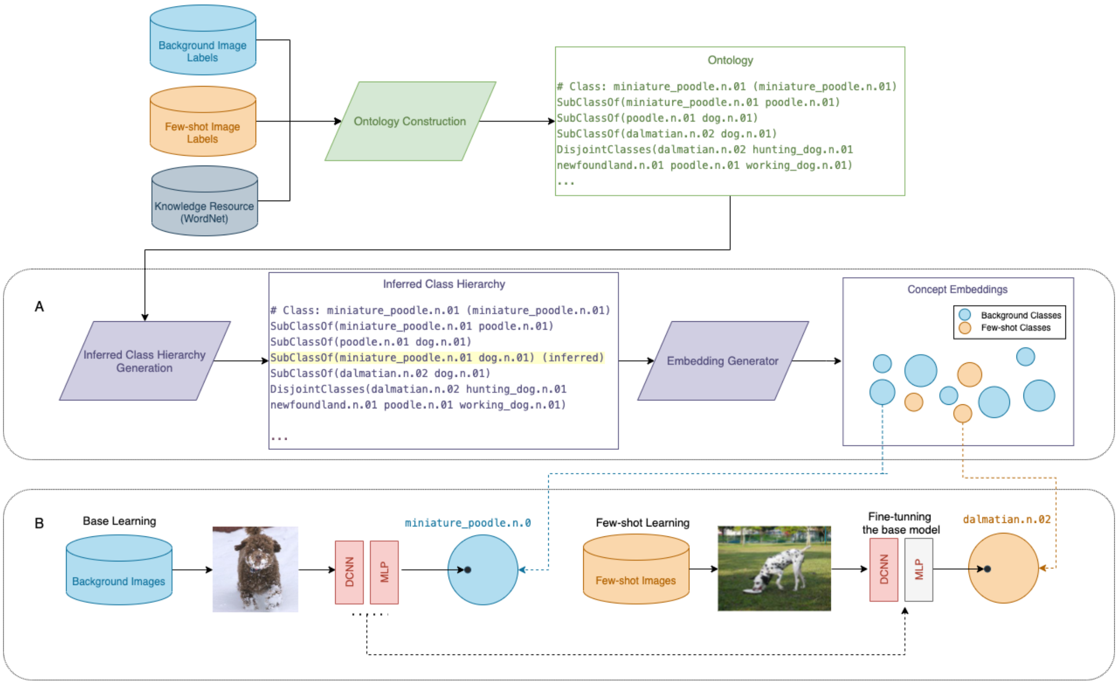
$$\begin{aligned} l_{P\sqcap Q\sqsubseteq\bot}(c_P, c_Q, r_P, r_Q) = &\max(0, -\|c_P - c_Q\|_2 + r_P + r_Q + \gamma) \\ &+ \big|\|c_P\|_2 - 1\big| + \big|\|c_Q\|_2 - 1\big|. \end{aligned} \tag{2}$$

It enforces the inequality $\|c_P - c_Q\|_2 \geq r_Q + r_P + \gamma$. According to the setting of $\gamma$, the user can decide how far the two $n$-balls are pushed away.

## 4. Proposed Method: ViOCE

Adopting few-shot image classification as our benchmark [16], we train a neural vision model using a set of background images $BI = \{(I_i, y_i)\}_{i=1}^m$ (base set) from $\mathcal{K}$ classes with $y_i \in C_B = \{c_1, c_2, \ldots c_{\mathcal{K}}\}$ and a set of few-shot images $FI = \{(I_i, y_i)\}_{i=1}^s$ (novel set) from $w$ classes with $y_i \in C_F = \{\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_w\}$, where $C_B \cap C_F = \emptyset$, and $I_i$ denotes the raw image vectors containing pixel values. The few-shot success is usually assessed by how accurate a model can select a correct class from the candidate class set $C_F$ for a new image from the few-shot classes. This is often referred to as the $w$-way $s$-shot few-shot image classification. We construct an ontology $\mathcal{O}$ by using the class label information $C_B$ and $C_F$, and also WordNet. It provides information on relationships that can exist among the class labels, containing knowledge regarding to "SubClassOf" and "DisjointClasses".

We propose ViOCE as a framework to improve few-shot image classification by integrating information provided by $\mathcal{O}$, $BI$ and $FI$. It is composed of two main components: (1) to embed classes in $C_B$ and $C_F$ as $n$-balls based on the constructed $\mathcal{O}$, (2) to embed images in the same Euclidean space as the $n$-balls with a suitable arrangement, and to infer the class for a query image based on its image embedding and the $n$-ball embeddings of the candidate classes. Figure 2 shows the general framework flow with an overview of all processes and data inputs.



```
# Class: miniature_poodle.n.01 (miniature_poodle.n.01)
SubClassOf(miniature_poodle.n.01 poodle.n.01)
SubClassOf(poodle.n.01 dog.n.01)
SubClassOf(dalmatian.n.02 dog.n.01)
DisjointClasses(dalmatian.n.02 hunting_dog.n.01
newfoundland.n.01 poodle.n.01 working_dog.n.01)
...
```

```
# Class: miniature_poodle.n.01 (miniature_poodle.n.01)
SubClassOf(miniature_poodle.n.01 poodle.n.01)
SubClassOf(poodle.n.01 dog.n.01)
SubClassOf(miniature_poodle.n.01 dog.n.01) (inferred)
SubClassOf(dalmatian.n.02 dog.n.01)
DisjointClasses(dalmatian.n.02 hunting_dog.n.01
newfoundland.n.01 poodle.n.01 working_dog.n.01)

...
```

**Figure 2:** The overview of the proposed ViOCE framework. If a suitable ontology for the task does not exist, the approach starts from constructing an ontology for the image labels capturing relationships between them based on an external knowledge resource (in our case WordNet). Subsequently the approach follows the two main components of the framework - A) Concept embedding learning process that starts with computing the inferred class hierarchy ($ICH$) of the input ontology and then generates $n$-ball embeddings for all the concepts found in the ontology. B) Visual model (DCNN+MLP) training where, first the background images are used to train a base model which gets fine-tuned (only MLP) using the few-shot images to produce the final model. During both base learning and few-shot learning processes, the concept embeddings guide the learning process by setting the objective of the model to project the image feature points inside the correct $n$-ball representing the ground truth label of an input image.

## 4.1. *n*-Ball Concept Embeddings

We build upon the EL embedding technique [2] to learn a set of $n$-balls for all concepts $\widetilde{\mathcal{O}}$ in the ontology $\mathcal{O}$, which is referred to as a *concept embedding*. We extract subsumption and disjointness axioms to define the class hierarchy of the ontology $\mathcal{O}$. It has been noticed that the entailed transitive relations such as if *Poodle SubclassOf Dog* and *Dog SubclassOf Animal*, then *Poodle SubclassOf Animal* are usually not well reflected by the learned $n$-balls. To overcome this, we use the inferred class hierarchy (ICH). Assuming all the concepts are satisfiable with $\mathcal{O}$, the ICH is computed such that $\text{ICH}(\mathcal{O}) = \{P \sqsubseteq Q | P \neq Q, P, Q \in \widetilde{\mathcal{O}}, \mathcal{O} \models P \sqsubseteq Q\}$. ICH contains all possible subsumption relations according to the definition of $\mathcal{O}$.

If simply to follow Eqs. (1) and (2), the radius of the learned $n$-ball for a leaf concept, which corresponds to an image class in $C_B$ or $C_F$, can end up being very small, in order to fit into the balls of its ancestor concepts. Since in the image embedding learning, we will map each image as a data point inside the $n$-ball corresponding to its ground truth class, an overly small radius can affect the learning accuracy. To tackle this, we introduce a regularisation term in Eq. (4) to prevent radius shrinkage. Moreover, the embedding quality can deteriorate as the class hierarchy of the input ontology becomes larger. To improve the embedding quality, we introduce an extra hyperparameter $\phi$ in Eq. (5) to explore potentially more expressive design spaces, which is supported by an additional parameter tuning process. Finally, we minimise the following loss function:

$$l_c\left(\{c_P\}_{P\in\widetilde{\mathcal{O}}}, \{r_P\}_{P\in\widetilde{\mathcal{O}}}\right) = \sum_{\text{ICH}(\mathcal{O})\models P\sqsubseteq Q} \max(0, \|c_P - c_Q\|_2 + r_P - r_Q - \gamma) \tag{3}$$

$$+ \sum_{\mathcal{O}\models C\sqcap D\sqsubseteq\bot} \max(0, -\|c_P - c_Q\|_2 + r_P + r_Q + \gamma)$$

$$+ \sum_{P\in\widetilde{\mathcal{O}}} \max(0, \psi\sqrt{N_h - L(P)} - r_P) \tag{4}$$

$$+ \sum_{P\in\widetilde{\mathcal{O}}} N(P)\big|\|c_P\|_2 - \phi\big| \tag{5}$$

Here, $N_h$ denotes the total level number contained by the class hierarchy, and $L(P)$ denotes the level of the concept $P$ in the hierarchy, e.g.„ the top-most concept has level 1. $N(P)$ denotes the number of times the concept $P$ appears in the extracted axioms. Both $\psi, \phi > 0$ are hyperparameters. Eq. (4) restrict the radius of the concept $P$ 's $n$-ball to be no less than $\psi\sqrt{N_h - L(P)}$. The top-level concepts are allowed to have larger $n$-balls than the bottom ones.

### 4.2. Image Embedding Learning

Our vision model is composed of a base DCNN architecture coupled with a multi-layer perceptron (MLP). The DCNN computes the visual features for an image by taking its raw pixel representation vector as the input: $f_i = \phi_{\text{D}}(I_i, \theta_{\text{D}})$ where $f_i \in \mathbb{R}^d$. The MLP is responsible for mapping the visual features $f_i$ to the $n$-dimensional Euclidean space where the $n$-ball concept embeddings sit: $h_i = \phi_{\text{M}}(f_i, \theta_{\text{M}})$ where $h_i \in \mathbb{R}^n$. We use $\theta_{\text{D}}$ and $\theta_{\text{M}}$ to denote the neural network parameters to be trained for the DCNN and MLP, respectively. The idea is to identify visual features of an image (using a DCNN) so that they can be mapped (by an MLP) as a data point inside the $n$-ball of its ground truth class. For example, an image containing the visual features of a "poodle" should be mapped inside the $n$-ball of the "poodle" concept learnt from the ontology.

To achieve this, the following a pairwise ranking loss is used to optimise the network parameters:

$$l_I(\theta_{\mathrm{D}}, \theta_{\mathrm{M}}) = \sum_{i=1}^{m} \left[ \max\left(0, \|c_P - h_i\|_2 - \mu r_P\right) + \sum_{Q \in C_i^{(-)}} \max(0, \nu r_Q - \left\|c_Q - h_i\right\|_2) \right], \tag{6}$$

where $\mu, \nu > 0$ are hyperparameters. The set $C_i^{(-)}$ contains the negative classes defined for each image $I_i$ of the positive class with its embedding computed by $h_i = \phi_{\mathrm{M}}(\phi_{\mathrm{D}}(I_i, \theta_{\mathrm{D}}), \theta_{\mathrm{M}})$. When setting $\mu = \nu = 1$, the loss enforces $\|c_P - h_i\|_2 \leq r_P$, pushing the embedded image point to stay inside the $n$-ball of the correct concept class $P$, while $\left\|c_Q - h_i\right\|_2 \geq r_Q$, to stay outside the $n$-ball of the incorrect concept class $Q$. The hyperparameters $\mu$ and $\nu$ are placed to control the intensity of this effect, e.g.,, $\mu < 1$ requiring to lie closer to the center which makes the task harder. The negatives classes $C_i^{(-)}$ are selected from the most similar classes to the positive class.

In practice, we first train the DCNN and MLP from scratch by minimising Eq. (6) using the background images $BI$. This is called base learning (BL). Then, we fine tune the MLP by using the few-shot images $FI$ by minimising the same loss, but keep the weights of DCNN fixed. This is called the few-shot learning (FSL). We test the vision model using the testing images of $FI$ ($FI_{te}$) after the fine-tuning of MLP in the $FSL$ stage. During inference, a prediction is made by finding the $n$-ball which an image feature projection lies in. Let $\mathcal{U} = \|c_P - h\| - r_P$, where $h$ is an output feature for a query image from the vision model and $c_P$ and $r_P$ are the centre and radius of a selected $n$-ball of $P$ respectively. If $\mathcal{U} \leq 0$, we find that $h$ lies inside the $n$-ball of $P$. Hence the classification of $h$ will be class $P$. In case some $h$ does not lie inside any of the $n$-balls of the $w$ classes in the few-shot task, we choose the closest lying $n$-ball centre $c_i$ out of the classes to $h$, where $\mathrm{argmin}_{c_i(i=1,2,..,w)} \left(\|c_i - h\|\right)$, as the prediction. The proportion of the correct predictions out of all images in $FI_{te}$ is recorded as the accuracy of the vision model in this study.

## 5. Experimental Results

MiniImageNet dataset consists of 60,000 images of 100 classes from ImageNet, where each class carries 600 example images [1]. Following the same splitting as in [17], 80 and 20 classes were allocated for training and testing respectively. TieredImageNet dataset is larger in size than miniImageNet, containing 608 classes from ImageNet [18]. Its classes are acquired based on 34 higher-level categories. We use a training set of 26 higher-level categories with 448 classes, and testing set of 8 higher-level categories with 160 classes.

We construct two new ontologies based on the image labels of the datasets. All selected datasets are subsets of ImageNet [19], where WordNet [20] synsets are used to annotate all images. This offered the opportunity to use the information from WordNet to formulate more knowledge about the image labels. To obtain a class hierarchy, we chose the hypernym tree of WordNet where given a label, the corresponding synset name together with all other synsets above it until the root (*entity.n.01*) was extracted. All these concepts were included in the ontology. The dimensionality of the concept embeddings was chosen to be 300. During all experiments, ResNet50 [21] architecture was chosen to be the base network and the MLP was composed of 5 layers with sizes of 2048, 1024, 512, 512 and 300.

ViOCE is evaluated by comparing with the performance of several existing approaches according to [22] under the same configuration. We conduct experiments for $w = 5$ and $s = \{1, 5\}$. Table 1 reports the 5-way 1-shot and 5-shot performance comparisons. It can be seen that ViOCE surpasses the the performance of all other approaches in every 5-way task in both datasets, while achieving >90% accuracy in miniImageNet 5-shot task. We argue that this is the result of the *n*-ball embeddings learnt using background knowledge, guiding a more meaningful distribution of image feature points.

**Table 1**
5-way 1-shot and 5-shot accuracy comparison with existing approaches using miniImageNet and tieredImageNet benchmarks. All accuracies are reported with 95% confidence intervals.

| **Model** | **miniImageNet 5-way** | | **tieredImageNet 5-way** | |
|---|---|---|---|---|
| | **1-shot (%)** | **5-shot (%)** | **1-shot (%)** | **5-shot (%)** |
| MAML (Finn et al.) | 48.70 ± 1.84 | 63.11 ± 0.92 | 51.67 ± 1.81 | 70.30 ± 1.75 |
| Matching Networks (Vinyals et al.) | 43.56 ± 0.84 | 55.31 ± 0.73 | - | - |
| Prototypical Networks (Snell et al.) | 49.42 ± 0.78 | 68.20 ± 0.66 | 53.31 ± 0.89 | 72.69 ± 0.74 |
| Relational Networks (Sung et al.) | 50.44 ± 0.82 | 65.32 ± 0.70 | 54.48 ± 0.93 | 71.32 ± 0.78 |
| AdaResNet (Munkhdalai et al.) | 56.88 ± 0.62 | 71.94 ± 0.57 | - | - |
| TADAM (Oreshkin et al.) | 58.50 ± 0.30 | 76.70 ± 0.30 | - | - |
| Shot-Free (Ravichandran et al.) | 59.04 ± n/a | 77.64 ± n/a | 63.52 ± n/a | 82.59 ± n/a |
| MetaOptNet (Lee et al.) | 62.64 ± 0.61 | 78.63 ± 0.46 | 65.99 ± 0.72 | 81.56 ± 0.53 |
| Fine-tuning (Dhillon et al.) | 57.73 ± 0.62 | 78.17 ± 0.49 | 66.58 ± 0.70 | 85.55 ± 0.48 |
| LEO-trainval (Rusu et al.) | 61.76 ± 0.08 | 77.59 ± 0.12 | 66.33 ± 0.05 | 81.44 ± 0.09 |
| Embedding-distill (Tian et al.) | 64.82 ± 0.60 | 82.14 ± 0.43 | 71.52 ± 0.69 | 86.03 ± 0.49 |
| ViOCE | **65.71 ± 0.13** | **93.65 ± 0.07** | **73.4 ± 0.13** | **88.95 ± 0.09** |

## 6. Conclusion

We show that the introduction of ontology-based background knowledge in the form of learnt concept embeddings to a visual model can improve its performance in the task of few-shot image classification. The proposed embedding learning method is capable of representing symbolic knowledge as *n*-ball embeddings, by capturing the subsumption and disjointness relations among concepts in a large ontology. The proposed ViOCE framework is capable of utilising the concept embeddings in an effective way to inform the training and inference procedures of a vision model, and produces superior performance in two benchmarks in few-shot image classification. We plan to extend this study to evaluate the semantically meaningful errors in classification as future work.

## References

[1] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, D. Wierstra, Matching networks for one shot learning, arXiv preprint arXiv:1606.04080 (2016).

[2] M. Kulmanov, W. Liu-Wei, Y. Yan, R. Hoehndorf, El embeddings: Geometric construction of models for the description logic el++, arXiv preprint arXiv:1902.10499 (2019).

[3] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, T. Mikolov, Devise: A deep visual-semantic embedding model (2013).

[4] M. Jayathilaka, T. Mu, U. Sattler, Visual-semantic embedding model informed by structured knowledge, arXiv preprint arXiv:2009.10026 (2020).

[5] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, J.-B. Huang, A closer look at few-shot classification, arXiv preprint arXiv:1904.04232 (2019).

[6] T. de Souza Alves, C. S. de Oliveira, C. Sanin, E. Szczerbicki, From knowledge based vision systems to cognitive vision systems: a review, Procedia Computer Science 126 (2018) 1855–1864.

[7] Z. Hu, X. Ma, Z. Liu, E. Hovy, E. Xing, Harnessing deep neural networks with logic rules, arXiv preprint arXiv:1603.06318 (2016).

[8] L. Serafini, A. d. Garcez, Logic tensor networks: Deep learning and logical reasoning from data and knowledge, arXiv preprint arXiv:1606.04422 (2016).

[9] X. Wang, Y. Ye, A. Gupta, Zero-shot recognition via semantic embeddings and knowledge graphs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6857–6866.

[10] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).

[11] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 1126–1135.

[12] M. Jayathilaka, Enhancing generalization of first-order meta-learning (2019).

[13] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al., Matching networks for one shot learning, in: Advances in neural information processing systems, 2016, pp. 3630–3638.

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

[15] J. Snell, K. Swersky, R. S. Zemel, Prototypical networks for few-shot learning, arXiv preprint arXiv:1703.05175 (2017).

[16] Y. Hu, V. Gripon, S. Pateux, Leveraging the feature distribution in transfer-based few-shot learning, arXiv preprint arXiv:2006.03806 (2020).

[17] X. He, P. Qiao, Y. Dou, X. Niu, Spatial attention network for few-shot learning, in: International Conference on Artificial Neural Networks, Springer, 2019, pp. 567–578.

[18] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, R. S. Zemel, Meta-learning for semi-supervised few-shot classification, arXiv preprint arXiv:1803.00676 (2018).

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database, in: CVPR09, 2009.

[20] G. A. Miller, Wordnet: a lexical database for english, Communications of the ACM 38 (1995) 39–41.

[21] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016). URL: http://dx.doi.org/10.1109/cvpr.2016.90. doi:10.1109/cvpr.2016.90.

[22] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, P. Isola, Rethinking few-shot image classification: a good embedding is all you need?, arXiv preprint arXiv:2003.11539 (2020).