SAMPLE-EFFICIENT LEARNING OF NOVEL VISUAL CONCEPTS

Sarthak Bhagat*, **Simon Stepputtis***, **Joseph Campbell, Katia Sycara**
The Robotics Institute, Carnegie Mellon University
{sarthakb, sstepput, jacampbe, sycara}@andrew.cmu.edu

ABSTRACT

Despite the advances made in visual object recognition, state-of-the-art deep learning models struggle to effectively recognize novel objects in a few-shot setting where only a limited number of examples are provided. Unlike humans who excel at such tasks, these models often fail to leverage known relationships between entities in order to draw conclusions about such objects. In this work, we show that incorporating a symbolic knowledge graph into a state-of-the-art recognition model enables a new approach for effective few-shot classification. In our proposed neuro-symbolic architecture and training methodology, the knowledge graph is augmented with additional relationships extracted from a small set of examples, improving its ability to recognize novel objects by considering the presence of interconnected entities. Unlike existing few-shot classifiers, we show that this enables our model to incorporate not only objects but also abstract concepts and affordances. The existence of the knowledge graph also makes this approach amenable to interpretability through analysis of the relationships contained within it. We empirically show that our approach outperforms current state-of-the-art few-shot multi-label classification methods on the COCO dataset and evaluate the addition of abstract concepts and affordances on the Visual Genome dataset.

1 INTRODUCTION

The ability to recognize objects from visual inputs (Zhao et al., 2018) is crucial for the success of agents that interact in real or simulated environments (Bai et al., 2020; Stepputtis et al., 2020; Shahria et al., 2022). Beyond applications in agent development, object recognition is also vital for image captioning (Stefanini et al., 2022), scene understanding (Xie et al., 2020), vision-language understanding (Uppal et al., 2022), and many other domains. Recent contributions to foundational vision models (Dosovitskiy et al., 2021; He et al., 2022) and a wider availability of computational resources has enabled many of these applications. One benefit of such models is their ability to drastically reduce the amount of training data needed when utilizing them as priors to train new visual tasks, e.g. in the domain of object recognition. However, while very capable, these pre-trained models often fail to perform well in few-shot learning settings that require them to recognize novel objects from a small set of sample images (Wang et al., 2023). Beyond object recognition, assigning abstract concepts and affordances is an even more challenging task as concepts such as *wearable* are only indirectly related to a visual representation. Inspired by how humans learn to utilize few-shot learning by connecting novel concepts to their prior domain knowledge and experience, neuro-symbolic architectures (Hassan et al., 2022) can address some of these shortcomings by imbuing neural networks with symbolic knowledge graphs (KG) (Abu-Salih, 2020). Utilizing the interconnected domain knowledge of the graph, novel concepts can be added in a few-shot manner by augmenting the graph with the new nodes and thus, also limiting the need to re-train large parts of the neural architecture. Depending on the representation of novel nodes, such approaches are largely invariant to the topology of the graph, only requiring the final neural outputs to be expanded and trained while intermediate components may be able to only require little fine-tuning. The availability of interconnected domain knowledge also allows easy integration of non-visual abstract concepts and affordances as relationships can be formed between such concepts and existing entities.

In the spirit of Marino et al. (2016), our approach utilizes a neural network approach in conjunction with an optimized KG constructed from the Visual Genome Multi-Label (VGML) (Marino et al., 2016) dataset. In this work, we improve and extend this setup to few-shot multi-label classification (FS-MLC) by proposing a pipeline that adds new information to existing domain knowledge via `RelaTe` : a multimodal relationship prediction transformer that, given a small set of images and a latent representation of the linguistic concept, automatically connects novel objects, abstract concepts, and affordances to existing domain knowledge. In particular, `RelaTe` will evaluate the information propagated through the KG that is relevant to these images in the context of a latent concept representation from GloVe (Pennington et al., 2014) and determine which nodes are applicable to be connected to the novel target concepts. Subsequently,

*These authors contributed equally to this work

we propose to extend the capacity of the final multi-label classifier by adding an output neuron associated with the new concept. The related weights of this extra neuron as well as the graph neural network are then trained and fine-tuned, respectively, to learn how to incorporate the new information. Thus, our approach utilizes a dynamically changing neuro-symbolic architecture that efficiently incorporates additional concepts in a sample-efficient manner.

Our proposed method provides improvements over existing work in two categories: 1) with the help of our proposed `RelaTe` module and training methodology, we show that it outperforms current state-of-the-art FS-MLC models on COCO and 2) our approach goes beyond object recognition by being able to also learn how abstract concepts and affordances can be incorporated in a few-shot manner, thus, continuously updating the neuro-symbolic architecture to accommodate the new concepts.

The availability of a KG also makes this approach amenable to interpretability as the propagations throughout the graph can be reviewed after the classification is made (Tiddi & Schlobach, 2022). The example in Figure 1 indicates that the *motorcycle* concept caused the attribute *two-wheeled*, *red*, and the affordance *transport*. In summary, our main contributions are as follows: (a) We propose a sample-efficient few-shot methodology to recognize novel concepts from a small set of images by utilizing a knowledge graph that is amenable to interpretability, showing that it outperforms current state-of-the-art few-shot methods on the COCO dataset. (b) We introduce `RelaTe` – a multimodal approach that predicts the existence of edges in the knowledge graph between novel concepts and existing nodes, allowing efficient integration of domain knowledge. (c) We also show the utility of having access to interconnected domain knowledge to effectively add abstract concepts, affordances, and scene summaries.

2 Related Works

Few-shot multi-label classification (FS-MLC) remains a challenging problem despite some recent advances (Alfassy et al., 2019; Chen et al., 2020; Yan et al., 2022). On its own, few-shot classification is difficult due to various factors like catastrophic forgetting (Goodfellow et al., 2013) and limited data sets; however, these problems are amplified in the multi-label case when novel target classes occur in conjunction with already existing concepts, making their identification and training even more challenging. One avenue of addressing this issue is the utilization of domain knowledge, which can reduce the complexity of this problem by reducing the reliance on labeled data (Wang et al., 2020; Chen et al., 2021) and instead, drawing from the encoded knowledge. Such domain knowledge can be acquired in multiple ways, either by explicitly formulating and utilizing a data structure or by utilizing a foundational neural network that is "large enough" to encode the general knowledge. Examples of such large models are GPT (OpenAI, 2023), particularly MiniGPT-4 (Zhu et al., 2023), CLIP (Radford et al., 2021), and Flamingo (Alayrac et al., 2022). However, in this work, we focus on imbuing neural networks with symbolic knowledge in the form of a Knowledge Graph as such data structure is amenable to human interpretation (Guo et al., 2022) and quick augmentation in order to address the FS-MLC problem. Nevertheless, we will compare our approach to publicly available large-language models (LLMs).

Few-Shot Multi-Label Classification Utilizing concepts has shown to be an efficient approach to learning interpretable policies (Zabounidis et al., 2023). One approach to learning the FS-MLC task is to define novel objects as the sum of their parts, allowing such approaches to learning how to recombine known, simpler concepts that represent the target class (Lake et al., 2011; Jia et al., 2013). However, the addition of novel fundamental concepts remains an active field of research. Several approaches have addressed the problem of adding new concepts from a small number of samples by utilizing additional modalities (Mao et al., 2015; Mei et al., 2022), structured primitives (Qian et al., 2019), generative modeling (Rostami et al., 2019; Bhagat et al., 2020), and meta-learning methods (Cao et al., 2021). However, these approaches are usually limited to simulated (Mordatch (2018); Qian et al. (2019)) or less demanding datasets (Rostami et al. (2019); Mao et al. (2019); Cao et al. (2021); Mei et al. (2022)) that do not reflect the richness and intricacy of real-world concepts that we encounter in our daily lives. One of the first papers addressing the problem of FS-MLC in great detail is Alfassy et al. (2019), which tackled the problem of limited data by representing sample images and their labels in a latent space and defining various set operations over these representations to synthesize additional samples through the combination of latent image features. Similarly, the work presented in Yan et al. (2022) proposed a multimodal approach that utilized word embeddings to align verbal and visual representations in a latent feature space, allowing the creation of a mechanism that obtains visual prototypes for unseen labels by sampling an image from the latent space pinpointed by a language description of the novel entity. In our work, we propose a framework for extracting abstract concepts from complex real-world images and demonstrate enhanced performance over current few-shot learners by utilizing the connection between linguistic and visual concept representations.

Neuro-Symbolic Few-Shot Learning In addition to the techniques discussed above, incorporating domain-specific knowledge shows great potential as it can assist in recognizing and adding new concepts in a more sample-efficient

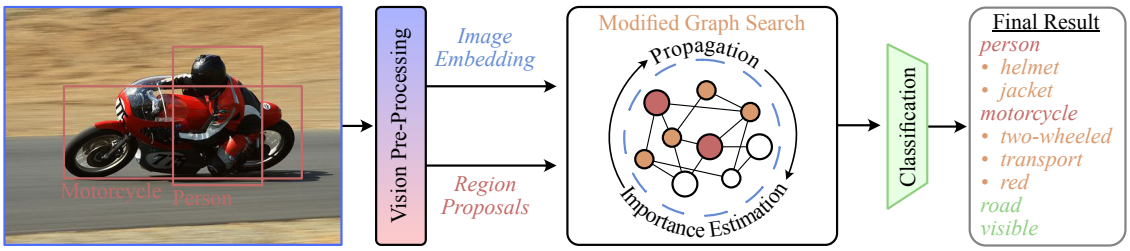


Figure 1: Example inference explaining our general inference pipeline, inspired by Marino et al. (2016). Given a novel image, we utilize ViT and Faster R-CNN to extract an image embedding (blue) and a set of initial object proposals (red). The initial proposals initialize our knowledge graph (red nodes) while the Modified Graph-Search identifies additional nodes (orange) conditioned on the overall image embedding (blue). Finally, our classifier evaluates the active nodes and produces a list of detected objects (green) in addition to the already detected nodes from the graph.

manner, especially in scenarios with limited data. A commonly used approach to utilize symbolic KGs in deep learning is graph neural networks (Xie et al., 2019; Wan et al., 2019; Lamb et al., 2020), providing a multitude of benefits from interpretability (Tiddi & Schlobach, 2022) to the utilization of interconnected information. The hierarchy and structure present in KGs have resulted in their use as priors for neuro-symbolic vision systems (Jiao et al., 2023) for a number of applications ranging from transfer learning (Alam et al., 2022) to vision-language pre-training (Alberts et al., 2021). Chen et al. (2020) introduced a static knowledge-guided graph routing framework consisting of two graph propagation frameworks to transfer both visual and semantic features, enabling information transfer between correlated features to train a better classifier with limited samples. Marino et al. (2016) and Fang et al. (2017) utilized this structured knowledge to identify the underlying concepts present in the image. With a comprehensive graph, the structured knowledge embedded in it can even be used to extract critical information about previously unseen classes in either a few- (Chen et al., 2019; Peng et al., 2019) or zero-shot (Kampffmeyer et al., 2018; Wang et al., 2018; Huang et al., 2020; Wang & Jiang, 2021; Wei et al., 2022; Lee et al., 2017) manner. However, a limitation of these approaches is the use of a static KG. The work presented in Wang et al. (2021) and Kim et al. (2020) partially addressed this problem by dynamically changing edge weights and re-computing latent node representations respectively, but the graph's structure and encoded knowledge fundamentally remain the same. In this work, we propose a mechanism to update both aspects of the neuro-symbolic architecture by dynamically extending the KG with novel nodes, computing representations that are conditioned on the target images, and updating the neural components of our classification approach both structurally and in regards to its trained weights. This also allows our approach to incorporate novel objects that go beyond the visual domain, including abstract concepts and affordances while alleviating the assumption, as in prior work (Zhu et al., 2014; Szedmak et al., 2014; Chuang et al., 2017; Ardón et al., 2019; Gretkowski et al., 2022) that an exhaustive KG has to exist.

3 Few-Shot Object Recognition with Neuro-Symbolic Architectures

In this work, we propose a twofold approach. Firstly, we employ a neuro-symbolic object recognition approach called Graph Search Neural Networks (GSNN), as originally introduced by Marino et al. (2016). To enhance the performance of this pipeline, we propose multiple modifications, namely adding image conditioning and incorporating node types (see Section 3.1). Secondly, we introduce a novel approach called `RelaTe` , which automatically extends the KG to integrate new concepts while, simultaneously, extending the neural components of the system to incorporate them (see Section 3.2). In the following sections, we provide detailed explanations of each component.

3.1 Neuro-Symbolic Object Recognition

At its core, our work considers the problem of detecting a set $\mathbb{C}$ of concepts in a given image $\boldsymbol{I}$, while affording the ability for a human Subject Matter Expert (SME) to extend the system's capability by detecting additional, novel concepts in a sample efficient manner from a small set of images. In this section, we first describe our inference pipeline, inspired by Marino et al. (2016) before discussing novel concept addition in Section 3.2. Figure 1 describes the three main steps of the inference pipeline: First, we extract a set of candidate objects $\mathbb{F}_{\boldsymbol{I}}$ that initialize our KG $\mathcal{G}$ and extract a global image embedding $\boldsymbol{e_I}$ (see Section 3.1.1); Second, we utilize the GSNN to propagate information through the graph while extending the prior work to also condition on the global image embedding $\boldsymbol{e_I}$, alleviating the need for edge types, and utilizing semantic node types; Third, the final classifier evaluates all active nodes of graph

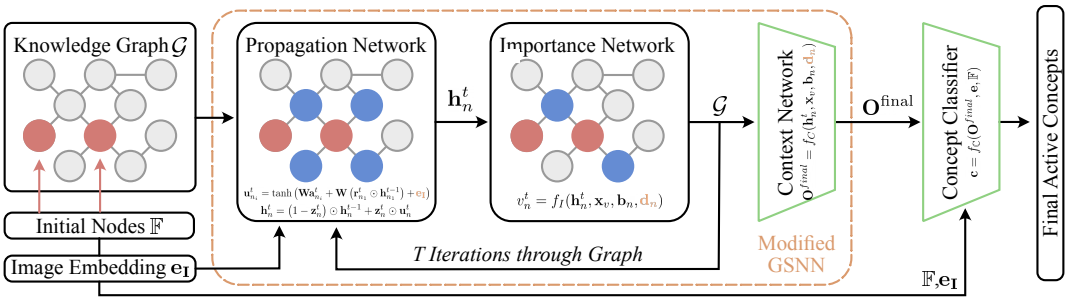


Figure 2: Overview of the Modified Graph Search Neural Network (GSNN). In contrast to prior work, we condition the propagation network on the global image embedding $\boldsymbol{e_I}$ and introduce a node type d_n while dropping edge types.

$\mathcal{G}$ (where $\mathcal{P}_{\boldsymbol{I}}$ is the sub-graph of $\mathcal{G}$ containing the active nodes for a particular image $\boldsymbol{I}$) in order to provide a holistic view of the input image and predict the final set of concepts $\mathbb{C}_{\boldsymbol{I}}$.

3.1.1 Extracting Candidate Objects

In the first step, we employ a pre-trained object detection pipeline, namely Faster R-CNN (Ren et al., 2015) to extract the initial set $\mathbb{F}_{\boldsymbol{I}}$ of candidate objects from image $\boldsymbol{I}$. Faster R-CNN is pre-trained on the COCO (Lin et al., 2014) dataset to predict the 80 concepts of COCO, but omit the 16 classes designated for our FS-MLC experiments as defined in Alfassy et al. (2019) for a total of 64 trained concepts $\mathbb{C}_{\text{COCO}}$. For this approach, we did not conduct any further fine-tuning on other datasets, nor did we change the outputs of Faster R-CNN. The initial set of objects $\mathbb{F} \subset \mathbb{C}_{\text{COCO}}$ is then utilized to activate the initial set of nodes $\mathbb{N}_{\mathbb{F}}$ in graph $\mathcal{G}$. In contrast to the prior work that uses VGG (Simonyan & Zisserman, 2014), we use a pre-trained ViT (Dosovitskiy et al., 2021) model to calculate an overall image embedding $\boldsymbol{e_I} \in \mathbb{R}^v$ with feature size v that is utilized to provide a global context for our modified graph-search approach as well as the final classifier. ViT is pre-trained on the ImageNet-21k (Deng et al., 2009) dataset and then fine-tuned on the ImageNet-10k dataset without any further modifications.

3.1.2 Modified Graph Search Neural Network

In this section, we provide a detailed explanation of the different components of the GSNN inspired by Marino et al. (2016), as well as the proposed modification of conditioning its components on the input image, removing edge types, and introducing node types. Figure 2 shows the modified GSNN over graph $\mathcal{G}$ which contains three core components: a) the propagation network which computes an embedding for each node given its neighbors in the context of the current image $\boldsymbol{I}$, b) the importance network which decides which nodes are relevant and should be kept for potential future expansion given the current image $\boldsymbol{I}$, and c) the context network which generates final node embeddings. The context network is dependent on both the current image and the associations derived from the KG via multiple rounds of applying the propagation and importance network. The goal of the GSNN is to, in an iterative manner, propagate and prepare the information encoded in the KG that is relevant to image $\boldsymbol{I}$ by alternating the propagation and importance network over T steps. After T rounds, the context network provides the representation needed for the final concept classifier. The selection of T is a crucial hyper-parameter of our method and Section A.6 evaluates this choice in detail. Each of these components and the final classifier are described in the following sections along with our proposed modifications.

Propagation Network: Given the initial set of nodes $\mathbb{N}_{\mathbb{F}}$, the propagation network is designed to produce an output $\boldsymbol{O} \in \mathbb{R}^{N \times F}$, where N is the number of nodes and F the feature size for the latent node embedding, encoding the information of each node's neighborhood. Each row of $\boldsymbol{O}$ represents a feature vector $\boldsymbol{h}$ for the respective node which is initialized with all zeros outside of the first element, which contains the node ID, x_v. We utilize the graph structure, encoded in an adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ to retrieve the hidden states $\boldsymbol{h}$ of active nodes based on their neighborhood in the graph. In contrast to prior work, we also provide the propagation network with the global image encoding $\boldsymbol{e_I}$ in order to ensure that information is propagated according to the image context (Ablations are provided in Section A.7)

Initially, we calculate a vector $\boldsymbol{a}_n$ representing the neighbourhood of each active node at iteration t given that $\boldsymbol{A}_{n_i}$ is the relative adjacency matrix for node n_i:

$$\boldsymbol{a}_{n_i}^t = \boldsymbol{A}_{n_i}^{\mathsf{T}} \left[\boldsymbol{h}_1^{t-1}, \boldsymbol{h}_2^{t-1}, \ldots, \boldsymbol{h}_N^{t-1}\right]^{\mathsf{T}} + b \tag{1}$$

Given this neighbourhood vector $\boldsymbol{a}_{n_i}^t$ for each node, we calculate $\boldsymbol{z}_{n_i}^t$ and $\boldsymbol{r}_{n_i}^t = \sigma\left(\boldsymbol{W}\boldsymbol{a}_{n_i}^t + \boldsymbol{W}\boldsymbol{h}_{n_i}^{t-1}\right)$, where all $\boldsymbol{W}$ are different and trainable weights of the neural network. Subsequently, we calculate the update $\boldsymbol{u}_{n_i}^t$ for each node's hidden state as follows, where each $\boldsymbol{W}$ is a separately trainable weight matrix:

$$\boldsymbol{u}_{n_i}^t = \tanh\left(\boldsymbol{W}\boldsymbol{a}_{n_i}^t + \boldsymbol{W}\left(\boldsymbol{r}_{n_1}^t \odot \boldsymbol{h}_{n_1}^{t-1}\right) + \boldsymbol{e_I}\right) \tag{2}$$

In contrast to the work presented in Marino et al. (2016), we calculate this update conditioned on the global image context $\boldsymbol{e_I}$, allowing the modified GSNN to incorporate image-specific information. Section 4.2.1 evaluates this benefit in further detail. The final hidden state $\boldsymbol{h}_{n_i}^t$ for each node in $\mathcal{P}_{\boldsymbol{I}}$ is subsequently calculated as a weighted sum of the previous hidden state $\boldsymbol{h}_{n_i}^{t-1}$ and the previously computed update vector $\boldsymbol{u}_{n_i}^t$:

$$\boldsymbol{h}_{n_i}^t = \left(1 - \boldsymbol{z}_{n_i}^t\right) \odot \boldsymbol{h}_{n_i}^{t-1} + \boldsymbol{z}_{n_i}^t \odot \boldsymbol{u}_{n_i}^t \tag{3}$$

Together with the importance network detailed in the next section, the propagation throughout the graph is done over T iterations, thus, allowing the utilization of the interconnected knowledge provided in the knowledge graph $\mathcal{G}$. Learning to utilize the symbolic knowledge of the graph efficiently is of utmost importance for our few-shot learning goal described in Section 3.2.

Importance Network. The importance network alternates with the propagation network over T cycles and decides whether or not an adjacent node to a currently active node should be made active. This is an important step as purely expanding nodes at every step has the potential to become computationally impractical if $\mathcal{G}$ is large. The importance v_n^t of each node at timestep t is calculated as follows:

$$v_n^t = f_I(\boldsymbol{h}_n^t, \boldsymbol{x}_v, \boldsymbol{b}_n, \boldsymbol{d}_n) \tag{4}$$

where, in contrast to the original GSNN, we propose the addition of $\boldsymbol{d}_n$ which represents a one-hot vector describing the node type ("object", "affordance", or "attribute") instead of using an edge type and $f_I(\dots)$ is a multi-layer perceptron (MLP). Nodes above a certain threshold γ are maintained for the next propagation cycle. Additionally, we also learn a node bias term $\boldsymbol{b}_n$ for each node in the knowledge graph that intuitively captures a global meaning of the respective node. Note that this bias is not depending on a particular image $\boldsymbol{I}$.

Context Network. After T iterations, the final node embeddings are created via the context network. Similar to the importance network, it is formulated as:

$$\boldsymbol{O}^{\text{final}} = f_C(\boldsymbol{h}_n^t, \boldsymbol{x}_v, \boldsymbol{b}_n, \boldsymbol{d}_n) \tag{5}$$

However, instead of predicting a scalar value indicating a node's importance, it generates the final state representation of the expanded nodes in $\mathcal{G}$, where $f_C(\dots)$ is another MLP.

3.1.3 Final Concept Classifier

The third and final step is the classification of the active concepts $\mathbb{C}_{\boldsymbol{I}} \subset \mathbb{C}$ in the input image $\boldsymbol{I}$ where $\mathbb{C}$ is a set of all possible concepts. These concepts are computed from the state representations $\boldsymbol{O}^{\text{final}}$ of all the expanded nodes in the active graph $\mathcal{P}$ along with the global image embedding $\boldsymbol{e_I}$ and the originally detected classes $\mathbb{F}_{\boldsymbol{I}}$ from Faster R-CNN. Utilizing a single fully connected layer, a probability distribution over all the concepts is predicted $\boldsymbol{c} = f_{\mathbb{C}}(\boldsymbol{O}^{final}, \boldsymbol{e_I}, \mathbb{F}_{\boldsymbol{I}})$. In order to make the result amenable to interpretation by a human user, we also provide the graph of active nodes $\mathcal{P}$, thus providing insights into why certain classifications may have been made.

3.2 Novel Concept Learning in a Dynamic Neuro-Symbolic Architecture

In addition to improving the neuro-symbolic architecture of GSNN our remaining two main contributions are as follows: a) a multi-modal Relation Prediction Transformer – `RelaTe` , that aids a human SME when adding novel concepts to the symbolic knowledge graph (Section 3.2.1) and b) introducing a framework to also dynamically updating neural parts of the inference pipeline described in Section 3.1.

3.2.1 Extending the Knowledge Graph with Relation Prediction Transformer

Figure 3 introduces our proposed approach – `RelaTe` . Given a small set of SME-provided images $\mathbb{I}_{\text{SME}}$ showing a novel concept as well as the partial graphs $\mathcal{P}$ for each image, `RelaTe` predicts how this novel concept can be incorporated into the existing knowledge graph. Thereby, `RelaTe` provides an efficient and intuitive way of quickly adding new symbolic knowledge to the graph.

Multi-modal Cross-Attention Framework. While we use the image processing pipeline of Dosovitskiy et al. (2021), we introduce a multi-modal approach to relating linguistic concept representations to images that contain a novel

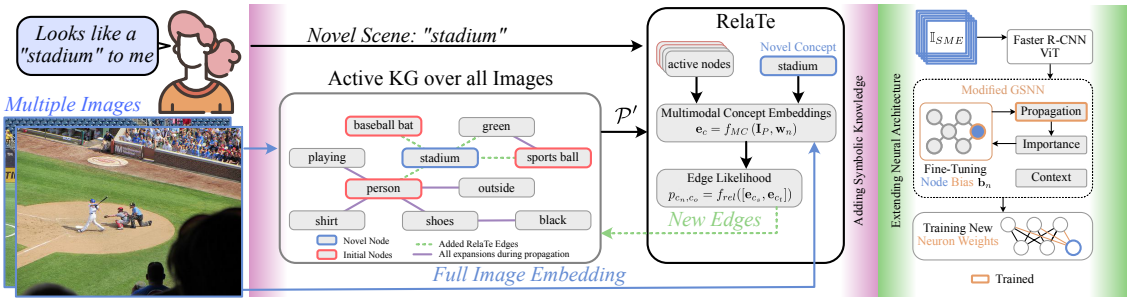


Figure 3: Given a novel concept of a *stadium* by a human expert along with one or more images for it, `RelaTe` estimates the optimal connectivity between the novel concept and existing nodes in the KG (i.e. *person*, *baseball bat*, *sports ball*, *green*). Subsequent inferences on similar images will yield the novel concept and allow for generalization through the domain knowledge encoded in the KG.

concept. In particular, we utilize GloVe (Pennington et al., 2014) word embeddings in order to retrieve a context-invariant representation $\boldsymbol{w}_c = f_{GloVe}(c) \in \mathbb{R}^{F_w}$ for any given concept $c \in \mathbb{C}$. These representations are particularly useful when generalizing to novel concepts due to the potential similarity of new concept embeddings to a semantically similar word that may be known to the KG already. In order to combine the linguistic and visual representations, we utilize a cross-attention framework in which the image is represented as a sequence of patches $\boldsymbol{I}_P$ (Dosovitskiy et al., 2021; Gan et al., 2022). We then combine both modalities as follows:

$$\boldsymbol{e}_{c_e} = f_{MC}\left(\boldsymbol{I}_P, \boldsymbol{w}_n\right) \tag{6}$$

where, $\boldsymbol{w}_c$ is the word embedding of any given concept c, and f_{MC} is explained in detail in Section A.1.

Post-Attention Fusion. Given the embeddings $\boldsymbol{e}_{c_e}$ for each node in each $\mathcal{P}_{\boldsymbol{I}}$ across all images in $\mathbb{I}_{\text{SME}}$ and the novel concept $\boldsymbol{e}_{c_n}$, we create pairs between the novel concept embedding $\boldsymbol{e}_{c_n}$ and the existing nodes' embeddings $\boldsymbol{e}_{c_e}$. We calculate the likelihood of an edge being present between a source and target nodes by concatenating the embeddings of each node pair as follows:

$$p_{c_s,c_t} = f_{rel}([\boldsymbol{e}_{c_s}, \boldsymbol{e}_{c_t}]) \tag{7}$$

Here, $f_{rel}(\dots)$ is an MLP predicting a scalar likelihood that an edge is present between the pair of nodes in the direction from source to target. $\boldsymbol{e}_{c_s}$ and $\boldsymbol{e}_{c_t}$ are populated by every combination of the novel concept node and existing graph nodes for novel objects; however, for abstract concepts and affordances, we only calculate incoming connections (see Section 4.1 for details). Finally, at most k nodes that are above a specified threshold $p_{c_n,c_o} > \gamma$ are added to the KG. We empirically choose a suitable k depending on the concept type while γ is a global hyper-parameter.

3.2.2 Updating the Neural Architecture

Adding the novel concept to the KG alone does not directly yield improved classification performance as the new node does not have a trained node bias $\boldsymbol{b}_n$ yet, nor does the propagation network know how to generate node embeddings $\boldsymbol{h}_n$ for the novel concept. Additionally, the final classifier needs to be extended to enable the prediction of the novel class. In this section, we detail the process of training the node bias, fine-tuning the propagation module, and extending the classifier in further detail (also see Figure 3, describing how we extend the neural architecture).

Fine-tuning the Node Bias and Graph Propagation. To fine-tune the propagation network and train the node bias $\boldsymbol{b}_n$ of the novel concept, we utilize a small dataset $\mathcal{D}_{\text{SME}}$ generated from the images $\mathbb{I}_{\text{SME}}$ given by the human SME that demonstrate the novel concept. $\mathcal{D}_{\text{SME}}$ is subsequently expanded by applying transformations to all images in $\mathbb{I}_{\text{SME}}$. Further, we define a small curated dataset $\mathcal{D}_C$ with the intention of preventing catastrophic forgetting that contains $\sim 2\%$ of the original VGML training data. The dataset $\mathcal{D}_C$ is selected through Maximally Diverse Expansion Sampling (MDES) which selects a representative set of inputs from the original VGML dataset that activates a diverse set of nodes in the graph $\mathcal{G}$ (see Section A.9 for details). Prior to training the propagation network and node bias, the novel node's bias $\boldsymbol{b}_n$ is initialized as the average of its adjacent nodes' bias, and the corresponding novel node is forcefully activated in each training sample in $\mathcal{D}_{\text{SME}} \cup \mathcal{D}_C$ that contains an image from $\mathcal{D}_{\text{SME}}$. Forcing the activation of the novel node includes it in the downstream classification task and thus enables the fine-tuning of the propagation network and node bias.

Extending the Classification Module After training the propagation network and node bias for a limited number of epochs, the classification module with the novel neuron for concept c_n is unfrozen and added to the fine-tuning

process. Furthermore, we reduce the learning rate of the propagation module and freeze the node bias $\boldsymbol{b}_n$ in this step of fine-tuning. As the classifier is depending on a valid node bias and the propagation network produces meaningful node embeddings for c_n, we delay the training of the classifier; however, we allow for continuous training of the propagation network to better capture the image-conditioned representation learning of the novel node. When training the classifier, its training objective is reduced to a binary classification problem that predicts whether or not the novel concept is active in $\mathcal{D}_{\text{SME}} \cup \mathcal{D}_C$ while only calculating gradients for the added neuron so as to not alter the prediction capabilities of the existing concepts. This approach drastically reduces the number of parameters that need to be optimized, allowing the dataset to be comparatively small.

4 Evaluation

We evaluate the effectiveness of our proposed approach for novel concept recognition in two separate settings. First, we compare it against current state-of-the-art FS-MLC baselines on the COCO (Lin et al., 2014) dataset, and second, we perform a qualitative analysis of adding abstract concepts, affordances, and scene summaries to the underlying neuro-symbolic architecture. Particularly, we demonstrate that our method, when trained on Visual Genome, outperforms FS-MLC baselines on COCO and even improves performance further when trained on the COCO training data. Further, we conduct an extensive ablation study analyzing the impact of the various components of the neuro-symbolic architecture as well as our proposed `RelaTe` approach. Finally, we further investigate the implications of different node addition strategies, while additional experiments regarding the number of iteration step T, curated dataset $\mathcal{D}_C$, the addition of node types, and further qualitative analysis including deeper analysis in regards to failure cases and large language models are available in the Appendix. The source code can be found at: `https://github.com/sarthak268/sample-efficient-visual-concept-learning`.

4.1 Data and Metrics

As our method depends on the existence of an initial knowledge graph, we initialize the graph $\mathcal{G}$ from the Visual Genome Multi-Label (VGML) (Marino et al., 2016) dataset. While based on Visual Genome (Krishna et al., 2016), VGML improves VG by drastically simplifying the graph, only using the 200 most common objects and 100 most common attributes, plus an additional 16 nodes to completely cover all COCO classes. We subsequently modify the graph for our FS-MLC task by removing 16 nodes that are defined as test nodes in Alfassy et al. (2019), resulting in a total of 300 nodes while also removing all images related to these 16 FS-MLC target nodes from the training dataset of VGML. Furthermore, we impose the requirement that all nodes representing affordances and attributes must be leaf nodes in the graph to simplify graph structure further. Additionally, we remove edge labels and introduce a one-hot vector indicating whether a node is an object, attribute, or affordance, and discovered that the edge types did not impact the performance of the approach. We evaluate these changes in comparison to the original knowledge graph from Marino et al. (2016) in Section A.8.

The coverage of all COCO classes is allowing us to compare novel object recognition performance between multiple COCO baselines and our approach. Particularly, the 16 FS-MLC test classes include *bicycle*, *boat*, *stop sign*, *bird*, *backpack*, *frisbee*, *snowboard*, *surfboard*, *cup*, *fork*, *spoon*, *broccoli*, *chair*, *keyboard*, *microwave*, and *vase*. In our additional evaluation of novel abstract concepts, affordance, and scenes, we utilize the full knowledge graph with all 316 nodes. We utilize our modified VGML dataset to train the GSNN, classification head of ViT, and final concept classifier in an end-to-end fashion. Further, `RelaTe` is trained on the entire Visual Genome dataset after removing the 16 test classes from it. The training for `RelaTe` includes concepts that are not present in VGML.

Evaluation Metric. In order to compare the efficacy of our approach in the FC-MLC task, we utilize mean average precision (mAP), macro average precision (Macro AP), and the top-K score. mAP is computed by taking the mean of the AP scores computed for each label, where AP is the area under the precision-recall curve plotted for each label. Similarly, Macro AP is computed by averaging the AP scores for each label across all instances and then averaging the results across all classes. To compute the top-K score, we compute the percentage of K most confident predictions of our model that are correctly predicted, i.e. precision of K most confident predictions.

4.2 Novel Concept Recognition

In this section, we detail our comparison of utilizing `RelaTe` with our updated neuro-symbolic architecture to add novel concepts. Particularly, we compare against multiple state-of-the-art baselines on FS-MLC tasks over the COCO dataset while training our model on VGML and later fine-tuning on COCO. Further, we demonstrate the utility of adding scene summaries, e.g. *kitchen* from kitchen appliances, abstract concepts, and affordances in a comprehensive ablation study on the VGML dataset.

4.2.1 COCO NOVEL MULTI-OBJECT RECOGNITION

We evaluate the efficacy of adding novel visual objects in a sample-efficient manner using our approach by comparing it against current state-of-the-art baselines in FC-MLC applications. As defined in Alfassy et al. (2019), we use a set $\mathbb{I}_{SME}$ of five images per novel class and train all 16 novel classes one by one. Table 1 shows the results comparing our method to three state-of-the-art baselines as well as four additional ablations. For each method, "Source" indicates the training dataset for the respective model while $\mathcal{D}_{SME}$ and $\mathcal{D}_C$ indicate which dataset was used for the few-shot learning. If both datasets are used, they are randomly interleaved. Lines 1 to 3 in Table 1 demonstrate the performance of our three baselines. Using a naive approach to adding novel concepts to Marino et al. (2016), line 4 trains only the final classifier (by adding a novel neuron) on the same training dataset as used in lines 4 to 7 without adding novel information to the knowledge graph altogether. In addition to training the classifier, adding the novel node to the knowledge graph, but not training the propagation network and node bias is shown in line 5, indicating that our modified GSNN is mostly invariant of the knowledge graph despite the lack of fine-tuning, underlining the strength of having domain knowledge (compare line 4 and 5), yielding a 17% performance increase. However, further improvements can be done when fine-tuning the propagation network and node bias. Compared to Yan et al. (2022) in line 3, which achieves 68.12% (the best baseline), utilizing the neuro-symbolic architecture and our proposed `RelaTe` architecture in line 6, we achieve a Macro AP score of 70.26, despite training on VGML, which is statistically significant with a standard deviation of $\sigma = 0.45$ at p-value $1.252e^{-3}$ trained over four seeds. Further, we also fine-tuned our method from line 6 on the training set of COCO and report the results of 70.30% with $\sigma = 0.19$, with a p-value of $9.1e^{-5}$ over four seeds in line 7 of Table 1. In each case, we parameterize `RelaTe` with an unlimited k value to add as many relations as possible. Given that our results in line 6 are resulting from a model trained on an entirely different dataset, i.e. VGML, yet performs very similarly to being trained on COCO allows the conclusion that our approach has the ability to transfer knowledge between datasets through the utilization of a knowledge graph.

4.2.2 RECOGNIZING AFFORDANCES, ATTRIBUTES, AND SCENES

Unlike other approaches to few-shot novel concept detection that rely on novel objects being visible in the input image, our approach can go beyond such limitations through the utilization of interconnected information in the knowledge graph. In addition to adding visual concepts as shown in Section 4.2.1, we demonstrate how non-visible concepts like abstract concepts, attributes, and scene summaries can be added. While the borders between what is visual and what is not are sometimes blurry, particularly in the case of scenes, utilizing the knowledge graph highlight the ability to draw conclusions from a set of partial observations. E.g., given that *refrigerator*, *oven*, and *microwave* were detected, we can conclude that the input image likely shows the *kitchen* concept, which can subsequently be added to the knowledge graph as a novel concept. In the following two sections, we discuss the addition of abstract concepts and scene summaries.

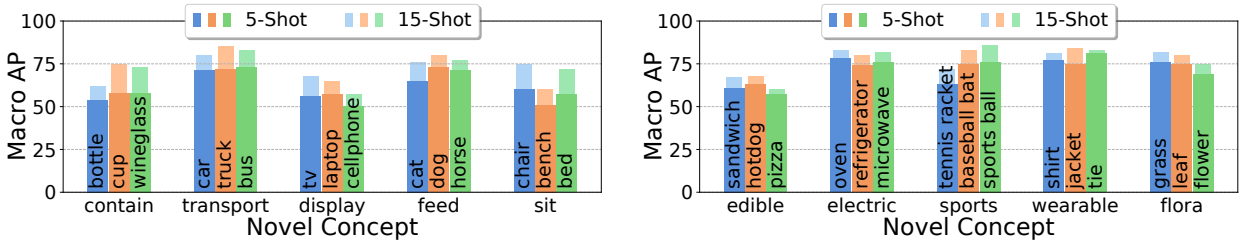
	Method	Source	$\mathcal{D}_{SME}$	$\mathcal{D}_C$	Macro AP
1	Alfassy et al. (2019)	COCO	✓	—	58.10
2	Chen et al. (2020)	COCO	✓	—	63.50
3	Yan et al. (2022)	COCO	✓	—	68.12
4	Fine-tuning (classifier)	VGML	✓	✓	52.22
5	Fine-tuning + `RelaTe`	VGML	✓	✓	69.26
6	Ours	VGML	✓	✓	70.26
7	Ours	COCO	✓	✓	**70.30**

Table 1: Experimental results on COCO dataset for five-shot multi-label classification of previously unseen concepts.

Adding Non-Visual Concepts. We conduct further experiments to assess the ability of `RelaTe` to incorporate non-visual concepts into the knowledge graph by relating it to relevant existing domain knowledge. In contrast to novel object recognition, we parameterize `RelaTe` with a threshold $k = 3$ in order to enforce a sparser connection of affordances and attributes to existing nodes. Figure 4 shows our experiments on adding novel affordances (Figure 4a) and attributes (Figure 4b). In each case, we selected a set $\mathbb{I}_{SME}$ with five and fifteen sample images containing three separate concepts that should be assigned to the novel concept. Subsequently, we evaluate the performance of the resulting model on 50 test images from within the same concept classes as well as 50 test images that do not show any of the trained targets. Our results show that added non-visual concepts have an average Macro AP of 66.7% given five sample images and 75.1% given fifteen images. Recall that for novel object detection as shown in Table 1, the Macro AP score is 70.26%. We hypothesize that the slightly lower performance on abstract concepts roots from the difficulty of not having a clear visual representation for such concepts. However, when increasing the training samples to fifteen, we outperform the object detection reported in Table 1, which, in the context of deep learning, is still a relatively small sample size.

	Model	Scene Concept *stadium*	*kitchen*	*zoo*	*school*	*bedroom*	Avg.
8	CLIP (0-shot)	16	100	100	56	72	68.8
9	Flamingo (0-shot)	20	4	40	24	28	23.2
10	Mini-GPT (0-shot)	24	96	64	24	96	60.8
11	Flamingo (5-shot)	68	36	40	72	80	59.2
12	Ours (5-shot)	90	84	84	72	92	**84.4**

Table 2: Novel scene prediction in comparison to free-form text generation models.



(a) Novel affordances learning given a set of five images.

(b) Novel attribute learning, given a set of five images.

Figure 4: Analysis of the performance when adding novel affordances and attributes to the knowledge graph. We evaluate the performance on five-shot (dark colors) and fifteen-shot (bright colors) learning.

Novel Scene Recognition: In addition to adding novel objects and abstract concepts, `RelaTe` can also assist in the addition of compound concepts. For example, the existence of a *oven*, *microwave* and *refrigerator* implies a scene that can be defined as a *kitchen* that is the sum of the underlying parts. Encoding such knowledge poses a slightly different problem as compound concepts require reasoning over multiple adjacent concepts. This ability can be imbued by our KG, but can also be found in large foundational neural networks, particularly LLMs. Table 2 demonstrates the ability of three LLM baselines, MiniGPT-4 (Zhu et al., 2023), CLIP (Radford et al., 2021), and Flamingo (Alayrac et al., 2022), to draw higher-level conclusions about the general scene shown in an image to our method, attempting the same task. We evaluate each scene on 25 test images of previously unseen samples and report the existence of the compound concept within the estimated concepts. For the LLMs, particularly the free-form response models, we queried the models if the image shows any of the five target scenes. With an average accuracy of 84.4%, this experiment underlines the utility of having interconnected knowledge that augments our few-shot detection pipeline, allowing the GSNN to successfully draw high-level conclusions from a set of basic concepts. While LLMs demonstrate partial success in identifying the high-level scenes, explicitly modeling the symbolic knowledge provides significant improvements despite the LLM's large general knowledge encoded within their trained model architecture. In additional experiments for the *kitchen* example, we showed that the likelihood of classifying the *kitchen* scene from a *refrigerator* or *microwave* alone is 24% and 28% respectively while the likelihood to identify it from an image containing both base concepts is 88%, showing that our model accurately learned that a *kitchen* is the sum of its parts.

4.2.3 Ablations

	Components			Fine-tuned		KG Configuration		All Classes			Novel Classes		
	ViT	FRCNN	KG	GSNN	CLF	RelaTe	MDES	T-1	T-5	mAP	T-1	T-5	mAP
1	✓	-	-	-	-	-	-	84.2	63.4	31.8	85.7	65.5	34.6
2	✓	✓	-	-	✓	-	-	84.7	64.2	33.0	86.1	67.2	37.3
3	✓	✓	✓	-	✓	-	-	87.4	68.8	36.5	91.8	72.8	68.0
4	✓	✓	✓	-	✓	✓	-	89.8	69.8	38.5	91.6	72.8	68.2
5	✓	✓	✓	✓	✓	✓	-	**90.4**	72.4	41.7	92.2	73.6	69.3
6	✓	✓	✓	-	✓	✓	✓	90.0	70.1	39.5	91.8	73.0	68.8
7	✓	✓	✓	✓	✓	✓	✓	90.3	**72.9**	**42.0**	**92.4**	**73.9**	**69.5**

Table 3: Experimental results on Visual Genome dataset, ablating the components of our method

In this section, we ablate the different components of our FS-MLC pipeline on the VGML dataset, recognizing the novel objects defined in Alfassy et al. (2019). Table 3 summarizes these results where lines 1 and 2 show the performance on the test set across all classes and our 16 novel few-shot classes given their Top-1 (T-1) and Top-5 (T-5) performance when using pure neural end-to-end architectures. In each case, novel classes are trained with five demonstration images and evaluated on the test set of VGML. Line 3 adds a KG with the GSNN approach proposed in Marino et al. (2016) and fine-tunes the final classifier (CLF-column) on the novel classes with an $\sim 3\%$ improvement in Top-K score. From this, we conclude that novel classes may also need to be added to the knowledge graph. Line 4 uses our proposed `RelaTe` approach to add the novel classes to the graph; however, does not tune the GSNN with respect to the propagation network and node bias (GSNN-column). Adding nodes to the graph yields another $3-5\%$ improvement over line 3. Line 5 fine-tunes the propagation network and node bias with our methodology described in Section 3.2.2, improving performance by another $\sim 2\%$. Finally, lines 6 and 7 show the impact of our curated fine-tuning dataset $\mathcal{D}_C$ as compared to an equally sized random dataset over the original VGML dataset. This demonstrates the importance of MDES to prevent catastrophic forgetting. In summary, Table 3 highlights our approach's ability to effectively expand its understanding of novel concepts with limited samples by effectively utilizing the knowledge graph. Further

experiments on the original KG are available in Section A.3 while a qualitative comparison of the ground-truth graph connections in comparison to the ones `RelaTe` adds is available in Section A.4.

4.2.4 Evaluating Node Addition Procedure

While `RelaTe` allows the addition of novel concepts individually, or as a group, we hypothesize that the node addition strategy has an impact on the overall performance of the model. Figure 5 shows the performance on the VGML dataset when adding the 16 nodes one-by-one (blue) or all at once (orange), omitting intermediate nodes 6-9 and 11-15 for simplicity. The trends show that adding one concept at a time and fine-tuning the classifier as well as the GSNN for each of them before adding the next concept yields a higher performing model. This not only facilitates the extraction and comprehension of new concepts but also prevents the model from getting overwhelmed with multiple concepts simultaneously, thus, minimizing the risk of forgetting previously acquired knowledge.

4.2.5 Interpretability of Results

Our approach provides interpretability through the explicit propagation of the initially detected concepts $\mathbb{F}_{\boldsymbol{I}}$ through the graph $\mathcal{G}$, providing insights as to why certain final concepts have been classified. However, while these propagations are not a direct output of the model, they provide an auxiliary insight into the internal workings of the FS-MLC pipeline. Figure 1 shows how these propagations can be useful to interpret the concept classifier's result.

4.3 Limitations and Future Work

While our method is capable of learning to recognize various objects, abstract concepts, and affordances in a sample-efficient manner, it is dependent on the comprehensiveness of the underlying knowledge graph Additionally, the reachability of the desired target class from the initially detected concept $\mathbb{F}_{\boldsymbol{I}}$ depends on the number of propagation steps T. Further, the accuracy of the model depends on the initial object detections of Faster R-CNN (see Appendix A.10 for a brief analysis). Another factor adding to this is that `RelaTe` requires any potentially related knowledge with respect to a novel concept to be expanded by $\mathbb{I}_{SME}$ due to the prohibitive computational complexity of checking against every node in $\mathcal{G}$. In future work, we plan to address these issues by choosing the number of propagation steps dynamically, allowing for further expansions, while also exploring the options of allowing SMEs to review proposed connections that `RelaTe` introduces.

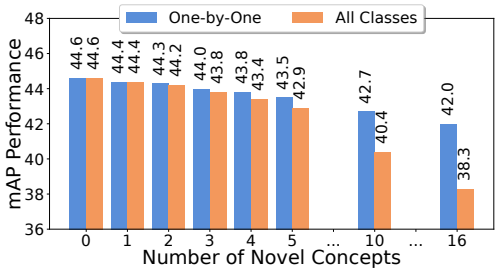


Figure 5: Strategy for adding the 16 novel concepts: one-by-one vs. all

5 Conclusion

In this work, we present a novel few-shot multi-label classification approach that uses domain knowledge in the form of a knowledge graph to recognize novel objects, abstract concepts, and affordances. We show that this neuro-symbolic architecture is particularly well suited for few-shot learning as novel concepts can be added to the knowledge graph, and thus alleviates the need to re-train large neural networks, but rather, utilizes a small dataset to perform targeted fine-tuning. As part of this methodology, we propose `RelaTe` , a novel approach of automatically connecting novel concepts to the existing domain knowledge and show its utility of not only adding novel objects but also adding non-visual concepts. Finally, we show that this approach outperforms current state-of-the-art approaches in few-shot multi-label classification on the COCO dataset. We also show that this approach performs well when on the COCO test classes despite being trained on the VGML dataset, demonstrating the transferability of knowledge through the utilization of a knowledge graph.

Acknowledgements

We gratefully acknowledge the financial support received from multiple funding sources, which made this research possible. Specifically, we would like to acknowledge the support from the Defense Advanced Research Projects Agency (DARPA) under the ASIST grant HR001120C0036, the Air Force Office of Scientific Research (AFOSR) under grants FA9550-18-1-0251 and FA9550-18-1-0097, and the Army Research Laboratory (ARL) under grant W911NF-19-2-0146 and W911NF-2320007.

REFERENCES

Bilal Abu-Salih. Domain-specific knowledge graphs: A survey. *J. Netw. Comput. Appl.*, 185:103076, 2020.

Mehwish Alam, Davide Buscaldi, Michael Cochez, Francesco Osborne, Diego Reforgiato Recupero, Harald Sack, Sebastian Monka, Lavdim Halilaj, Achim Rettinger, Mehwish Alam, Davide Buscaldi, Michael Cochez, Francesco Osborne, Diego Refogiato Recupero, and Harald Sack. A survey on visual transfer learning using knowledge graphs. *Semant. Web*, 13(3):477–510, jan 2022. ISSN 1570-0844. doi: 10.3233/SW-212959. URL https://doi.org/10.3233/SW-212959.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198, 2022.

Houda Alberts, Ningyuan Huang, Yash Deshpande, Yibo Liu, Kyunghyun Cho, Clara Vania, and Iacer Calixto. VisualSem: a high-quality knowledge graph for vision and language. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pp. 138–152, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.mrl-1.13. URL https://aclanthology.org/2021.mrl-1.13.

Amit Alfassy, Leonid Karlinsky, Amit Aides, Joseph Shtok, Sivan Harary, Rogério Schmidt Feris, Raja Giryes, and Alexander M. Bronstein. Laso: Label-set operations networks for multi-label few-shot learning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6541–6550, 2019.

Paola Ardón, Éric Pairet, Ronald P. A. Petrick, Subramanian Ramamoorthy, and Katrin Solveig Lohan. Learning grasp affordance reasoning through semantic relations. *IEEE Robotics and Automation Letters*, 4:4571–4578, 2019.

Qiang Bai, Shaobo Li, Jing Yang, Qisong Song, Zhiang Li, and Xingxing Zhang. Object detection recognition and robot grasping based on machine learning: A survey. *IEEE Access*, 8:181855–181879, 2020. doi: 10.1109/ACCESS.2020.3028740.

Sarthak Bhagat, Shagun Uppal, Zhuyun Yin, and Nengli Lim. Disentangling multiple features in video sequences using gaussian processes in variational autoencoders. In *European Conference on Computer Vision (ECCV)*, 2020.

Kaidi Cao, Maria Brbić, and Jure Leskovec. Concept learners for few-shot learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Jiaoyan Chen, Yuxia Geng, Zhuo Chen, Jeff Z. Pan, Yuan He, Wen Zhang, Ian Horrocks, and Hua zeng Chen. Zero-shot and few-shot learning with knowledge graphs: A comprehensive survey. 2021.

Riquan Chen, Tianshui Chen, Xiaolu Hui, Hefeng Wu, Guanbin Li, and Liang Lin. Knowledge graph transfer network for few-shot recognition. In *AAAI Conference on Artificial Intelligence*, 2019.

Tianshui Chen, Liang Lin, Riquan Chen, Xiaolu Hui, and Hefeng Wu. Knowledge-guided multi-label few-shot learning for general image recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:1371–1384, 2020.

Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 975–983, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. 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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Yuan Fang, Kingsley Kuan, Jie Lin, Cheston Tan, and Vijay Ramaseshan Chandrasekhar. Object detection meets knowledge graphs. In *International Joint Conference on Artificial Intelligence*, 2017.

Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. Vision-language pre-training: Basics, recent advances, and future trends, 2022.

Ian J. Goodfellow, Mehdi Mirza, Xia Da, Aaron C. Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgeting in gradient-based neural networks. *CoRR*, abs/1312.6211, 2013.

Andrzej Gretkowski, Dawid Wiśniewski, and Agnieszka Ławrynowicz. Should we afford affordances? injecting conceptnet knowledge into bert-based models to improve commonsense reasoning ability. In Oscar Corcho, Laura Hollink, Oliver Kutz, Nicolas Troquard, and Fajar J. Ekaputra (eds.), *Knowledge Engineering and Knowledge Management*, pp. 97–104, Cham, 2022. Springer International Publishing. ISBN 978-3-031-17105-5.

Yue Guo, Joseph Campbell, Simon Stepputtis, Ruiyu Li, Dana Hughes, Fei Fang, and Katia Sycara. Explainable action advising for multi-agent reinforcement learning. *arXiv preprint arXiv:2211.07882*, 2022.

Muhammad Hassan, Haifei Guan, Aikaterini Melliou, Yuqi Wang, Qianhui Sun, Sen Zeng, Wenqing Liang, Yi wei Zhang, Ziheng Zhang, Qiuyue Hu, Yang Liu, Shun-Dong Shi, Lin An, Shuyue Ma, Ijaz Gul, Muhammad Akmal Rahee, Zhou You, Canyang Zhang, Vijay Pandey, Yuxing Han, Yongbing Zhang, Ming Xu, Qi Huang, Jiefu Tan, Qinwang Xing, Peiwu Qin, and Dongmei Yu. Neuro-symbolic learning: Principles and applications in ophthalmology. *ArXiv*, abs/2208.00374, 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

He Huang, Yuan-Wei Chen, Wei Tang, Wenhao Zheng, Qingguo Chen, Yao Hu, and Philip S. Yu. Multi-label zero-shot classification by learning to transfer from external knowledge. *ArXiv*, abs/2007.15610, 2020.

Yangqing Jia, Joshua T. Abbott, Joseph L. Austerweil, Thomas L. Griffiths, and Trevor Darrell. Visual concept learning: Combining machine vision and bayesian generalization on concept hierarchies. In *NIPS*, 2013.

Licheng Jiao, Jing Chen, F. Liu, Shuyuan Yang, Chao You, Xu Liu, Lingling Li, and Biao Hou. Graph representation learning meets computer vision: A survey. *IEEE Transactions on Artificial Intelligence*, 4:2–22, 2023.

Michael C. Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, and Eric P. Xing. Rethinking knowledge graph propagation for zero-shot learning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11479–11488, 2018.

Taesup Kim, Sungwoong Kim, and Yoshua Bengio. Visual concept reasoning networks. In *AAAI Conference on Artificial Intelligence*, 2020.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. URL https://arxiv.org/abs/1602.07332.

Brenden M. Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua B. Tenenbaum. One shot learning of simple visual concepts. *Cognitive Science*, 33, 2011.

L. Lamb, Artur S. d'Avila Garcez, Marco Gori, Marcelo O. R. Prates, Pedro H. C. Avelar, and Moshe Y. Vardi. Graph neural networks meet neural-symbolic computing: A survey and perspective. In *International Joint Conference on Artificial Intelligence*, 2020.

Chung-Wei Lee, Wei Fang, Chih-Kuan Yeh, and Y. Wang. Multi-label zero-shot learning with structured knowledge graphs. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1576–1585, 2017.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes words and sentences from natural supervision. *ArXiv*, abs/1904.12584, 2019.

Junhua Mao, Xu Wei, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Loddon Yuille. Learning like a child: Fast novel visual concept learning from sentence descriptions of images. *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2533–2541, 2015.

Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Kumar Gupta. The more you know: Using knowledge graphs for image classification. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20–28, 2016.

Lingjie Mei, Jiayuan Mao, Ziqi Wang, Chuang Gan, and Joshua B. Tenenbaum. FALCON: Fast visual concept learning by integrating images, linguistic descriptions, and conceptual relations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=htWIlvDcY8.

Igor Mordatch. Concept learning with energy-based models. *ArXiv*, abs/1811.02486, 2018.

OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

Zhimao Peng, Zechao Li, Junge Zhang, Yan Li, Guo-Jun Qi, and Jinhui Tang. Few-shot image recognition with knowledge transfer. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 441–449, 2019.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

Peng Qian, Luke B. Hewitt, Joshua B. Tenenbaum, and R. Levy. Inferring structured visual concepts from minimal data. In *Annual Meeting of the Cognitive Science Society*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015.

Mohammad Rostami, Soheil Kolouri, James L Mcclelland, and Praveen K. Pilly. Generative continual concept learning. In *AAAI Conference on Artificial Intelligence*, 2019.

Md Tanzil Shahria, Md Samiul Haque Sunny, Md Ishrak Islam Zarif, Jawhar Ghommam, Sheikh Iqbal Ahamed, and Mohammad H Rahman. A comprehensive review of vision-based robotic applications: Current state, components, approaches, barriers, and potential solutions. *Robotics*, 11(6), 2022. ISSN 2218-6581. doi: 10.3390/robotics11060139. URL https://www.mdpi.com/2218-6581/11/6/139.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: a survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):539–559, 2022.

Simon Stepputtis, Joseph Campbell, Mariano Phielipp, Stefan Lee, Chitta Baral, and Heni Ben Amor. Language-conditioned imitation learning for robot manipulation tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 13139–13150. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/9909794d52985cbc5d95c26e31125d1a-Paper.pdf.

Sandor Szedmak, Emre Ugur, and Justus Piater. Knowledge propagation and relation learning for predicting action effects. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 623–629, 2014. doi: 10.1109/IROS.2014.6942624.

Ilaria Tiddi and Stefan Schlobach. Knowledge graphs as tools for explainable machine learning: A survey. *Artificial Intelligence*, 302:103627, 2022. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2021.103627. URL https://www.sciencedirect.com/science/article/pii/S0004370221001788.

Shagun Uppal, Sarthak Bhagat, Devamanyu Hazarika, Navonil Majumder, Soujanya Poria, Roger Zimmermann, and Amir Zadeh. Multimodal research in vision and language: A review of current and emerging trends. *Information Fusion*, 77:149–171, 2022. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2021.07.009. URL https://www.sciencedirect.com/science/article/pii/S1566253521001512.

Hai Wan, Jialing Ou, Baoyi Wang, Jianfeng Du, Jeff Z. Pan, and Juan Zeng. Iterative visual relationship detection via commonsense knowledge graph. In *Joint International Conference of Semantic Technology*, 2019.

Hang Wang, Youtian Du, Guangxun Zhang, Zhongmin Cai, and Chang Su. Learning fundamental visual concepts based on evolved multi-edge concept graph. *IEEE Transactions on Multimedia*, 23:4400–4413, 2021. doi: 10.1109/TMM.2020.3042072.

Jin Wang and Bo Jiang. Zero-shot learning via contrastive learning on dual knowledge graphs. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 885–892, 2021. doi: 10.1109/ICCVW54120.2021.00104.

X. Wang, Yufei Ye, and Abhinav Kumar Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6857–6866, 2018.

Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.*, 53(3), jun 2020. ISSN 0360-0300. doi: 10.1145/3386252. URL https://doi.org/10.1145/3386252.

Yudong Wang, Ma Chang, Qingxiu Dong, Lingpeng Kong, Zhifang Sui, and Jingjing Xu. Worst-case few-shot evaluation: Are neural networks robust few-shot learners?, 2023. URL https://openreview.net/forum?id=53yQBJNQVJu.

Jiwei Wei, Yang Yang, Zeyu Ma, Jingjing Li, Xing Xu, and Heng Tao Shen. Semantic enhanced knowledge graph for large-scale zero-shot learning. *ArXiv*, abs/2212.13151, 2022.

Lin Xie, Feifei Lee, Li Liu, Koji Kotani, and Qiu Chen. Scene recognition: A comprehensive survey. *Pattern Recognition*, 102:107205, 2020. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2020.107205. URL https://www.sciencedirect.com/science/article/pii/S003132032030011X.

Yaqi Xie, Ziwei Xu, M. Kankanhalli, Kuldeep S. Meel, and Harold Soh. Embedding symbolic knowledge into deep networks. In *Neural Information Processing Systems*, 2019.

Kun Yan, Chenbin Zhang, Jun Hou, Ping Wang, Zied Bouraoui, Shoaib Jameel, and Steven Schockaert. Inferring prototypes for multi-label few-shot image classification with word vector guided attention. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):2991–2999, Jun. 2022. doi: 10.1609/aaai.v36i3.20205. URL https://ojs.aaai.org/index.php/AAAI/article/view/20205.

Renos Zabounidis, Joseph Campbell, Simon Stepputtis, Dana Hughes, and Katia P Sycara. Concept learning for interpretable multi-agent reinforcement learning. In *Conference on Robot Learning*, pp. 1828–1837. PMLR, 2023.

Zhong-Qiu Zhao, Peng Zheng, Shou tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 30:3212–3232, 2018.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023.

Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *European Conference on Computer Vision*, 2014.

Approach	Affordances	Attributes
CLIP (0-shot)	28.4	35.0
Flamingo (0-shot)	0	0
MiniGPT (0-shot)	18.4	24.4
Flamingo (5-shot)	30.8	47.8
Ours (5-shot)	**61.2**	**69.8**

Table 4: Novel non-visual concept prediction in comparison to free-form text generation models.

A Appendix

A.1 Cross-Modal Attention Mechanism in ReLaTe

In this section, we elaborate further on the cross-modal attention mechanism to fuse the linguistic concept representation $\boldsymbol{w}_c \in \mathbb{R}^{F_w}$ with the image representation $\boldsymbol{I}_P \in \mathbb{R}^{P \times F_P^2 C}$. Fundamentally, this is a standard cross-attention approach in which the word embedding is considered as query and image patch embedding as key and value. However, for completeness, we outline the process as follows. Particularly, we define

$$\boldsymbol{e}_c = f_{MC}\left(\boldsymbol{I}_P, \boldsymbol{w}_n\right) \tag{8}$$

The transformer encoder architecture is built as L sequential layers each composed of a multi-head cross-attention and multi-layer perceptron block where each block is preceded by layer normalization and followed by a residual connection.

The initial input to the encoder is a sequence $\mathbf{Z}^0$ of length P where each element $\boldsymbol{z}_p \in \mathbb{R}^{F_l}$ of size F_l is the computed as follows for each patch in $\boldsymbol{I}_{P[i,:]}$:

$$\boldsymbol{z}_i^0 = \boldsymbol{I}_{P[i,:]}\boldsymbol{E}_{[i,:]} + \boldsymbol{P}_{[i+1,:]} \tag{9}$$

where $\boldsymbol{E} \in \mathbb{R}^{(F_p^2 C) \times F_l}$ is a learnable projection matrix and $\boldsymbol{P} \in \mathbb{R}^{(P+1) \times F_l}$ is a learnable positional embedding for each patch in $\boldsymbol{I}_P$. Further, we insert a `CLS` token at the beginning of the list $\mathbf{Z}_1^0 = \boldsymbol{P}_{[0,:]}$. Provided that the word embedding $\boldsymbol{w}_c$ for the concept c, for each layer $l \in [1, \ldots, L]$, the embedding $\boldsymbol{z}_i$ is given by the following equations:

$$\boldsymbol{z}_i^{l'} = f_{CA}\big(\text{layernorm}(\boldsymbol{z}_i^{l-1}), \boldsymbol{w}_c\big) + \boldsymbol{z}_i^{l-1} \tag{10}$$

$$\boldsymbol{z}_i^{l} = \text{MLP}\big(\text{layernorm}(\boldsymbol{z}_i^{l'})\big) + \boldsymbol{z}_i^{l'} \tag{11}$$

In the above equation, the cross-attention is computed by querying the concept embedding $\boldsymbol{w}_c$ against the patchwise encoding of the previous layer, initialized by the patches from image $\boldsymbol{I}_P$. The cross-attention module, $f_{CA}(\ldots)$, is a multi-head approach encompassing h heads. Following the standard transformer architecture, we compute $f_{CA}(\ldots)$ as follows:

$$f_{CA}(\boldsymbol{k}, \boldsymbol{v}, \boldsymbol{q}) = \text{softmax}\Big(\frac{\boldsymbol{q}\boldsymbol{k}^\mathsf{T}}{\sqrt{D_A}}\Big)\boldsymbol{v} \tag{12}$$

The key $\boldsymbol{k}$, query $\boldsymbol{q}$, and value $\boldsymbol{v}$ for the individual cross-attention heads are given by:

$$[\boldsymbol{k} \quad \boldsymbol{v}] = \text{layernorm}(\boldsymbol{z}_i^{l-1})\boldsymbol{W}_{kv}, \quad \boldsymbol{q} = \boldsymbol{w}_c\boldsymbol{W}_q \tag{13}$$

where $\boldsymbol{W}_{kv}$ and $\boldsymbol{W}_q$ are trainable weights and $D_A = \frac{F_l}{\beta}$, where β is a hyper-parameter. The final embedding for concept $\boldsymbol{e}_{c_n}$ is obtained by extracting the representation corresponding to the 1^{st} element in sequence $\mathbf{Z}$ after all L layers.

$$\boldsymbol{e}_c = f_{MC}\left(\boldsymbol{I}_P, \boldsymbol{w}_c\right) = \text{layernorm}(\mathbf{Z}_0^L) \tag{14}$$

A.2 Evaluation of Novel Non-Visual Concept Extraction

Our model was assessed for its ability to predict non-visual concepts such as affordances and attributes, in comparison to the free-form language generation baselines, namely MiniGPT-4 (Zhu et al., 2023), Flamingo (Alayrac et al., 2022), and CLIP (Radford et al., 2021), explained in Section 4.2.2. The methodology we utilize to obtain the predictions for each baseline is mentioned below:

- CLIP (0-shot) (Radford et al., 2021): We evaluated a standard CLIP model by tasking it with a multi-label classification task over our concepts. The language prompt for CLIP is the list of all 316 nodes plus the novel concept node and we considered the detection to be successful if the targeted concept was part of the N most confident classes, where N is the number of the respective image's ground-truth classes plus one. As CLIP does not provide an easy few-shot learning opportunity, we only evaluated the zero-shot case.
- Open-Flaming (0-shot) (Alayrac et al., 2022): We utilize Open-Flamingo in lieu of the official Flamingo, as official models are not publicly available. However, in the zero-shot case, we provided our test images and prompted Flamingo with the following query: "Does the image show an item that can contain, display, feed, sit, or transport?" for the affordances and "Does the image show an item that is edible, electric, flora, sports, or wearable?" for the attributes. We then evaluated the generated text manually to determine whether or not Flamingo detected the concept correctly. For example, we counted a response like "Yes, the image shows hot dogs with cheese on them, which are edible." as successful identification of the concept *edible*.
- Open-Flamingo (5-shot) (Alayrac et al., 2022): In the five-shot use-case, we provided further context to Flamingo by providing all 25 sample images (five for each class) with their respective label to Flamingo and then prompting for a single label for each of the novel test images.
- MiniGPT (0-shot) (Zhu et al., 2023): Finally, we also employed MiniGPT-4 in order to also utilize a multi-modal GPT baseline. Here, we provided the image as context and asked the same question as in Open-Flamingo while evaluating the generated response manually.

The results for the same are presented in Table 4. Our approach outperforms the best baseline by an average score of 26.2% on the prediction of non-visual concepts. The superior performance of our model can be attributed to its capacity to deduce non-visual concepts by connecting them with visual concepts derived from the visual inputs.

A.3 Quantitative Evaluation of RelaTe

	Fine-tuned		KG Configuration		All Classes			Novel Classes		
	GSNN	CLF	RelaTe	O-KG	T-1	T-5	mAP	T-1	T-5	mAP
1	-	✓	-	✓	90.6	70.2	38.8	91.4	72.2	67.6
2	✓	✓	-	✓	90.2	**72.8**	41.6	91.2	**73.7**	69.0
3	-	✓	✓	-	89.8	69.8	38.5	91.6	72.8	68.2
4	✓	✓	✓	-	**90.4**	72.4	**41.7**	**92.2**	73.6	**69.3**

Table 5: Experimental results on Visual Genome dataset.

In addition to the ablations of Table 3, we provide a quantitative evaluation regarding the ability of RelaTe to restoring the ground-truth KG of the VGML dataset for the 16 novel classes. Recall that we intentionally removed the test classes from the KG used in our few-shot experiments. Ideally, RelaTe would restore or create an even better KG through the proposed edge-addition framework. Table 5 presents a quantitative comparison between our proposed edge addition methodology, RelaTe , and using the original knowledge graph without removing nodes corresponding to the 16 novel classes. Rows 1 and 2 demonstrate the use of the original KG (O-KG) while rows 3 and 4 denote the models that use the KG populated by our RelaTe framework. All four of these models are trained without the use of MDES on a random selection of images from the original dataset. The results demonstrate that our approach effectively incorporates novel concepts into the KG. In fact, our method outperforms the model that utilized the original KG for some metrics. This is because our approach not only restores the previously removed edges but also introduces additional connections that are observed in the SME-provided images, thus, improving performance.

A.4 Qualitative Evaluation of RelaTe

In this section, we provide examples of connections recommended by our RelaTe framework. For each novel concept, we pick 4 images and pass them through our edge addition framework to demonstrate the edges that it populates into the graph $\mathcal{G}$. In Figure 6, each example starts with the concept that was removed from the knowledge graph (red) and its initial connections (purple). The suggested connections by our RelaTe framework are shown in the green box, which was generated when the system was given a set of 4 images. These results demonstrate the effectiveness of our relation prediction approach in adding back relevant connections that are prominent in the provided images. Moreover, our model suggests some additional relations that may not have been present in the original graph but are relevant and can provide significant information about the scene content.

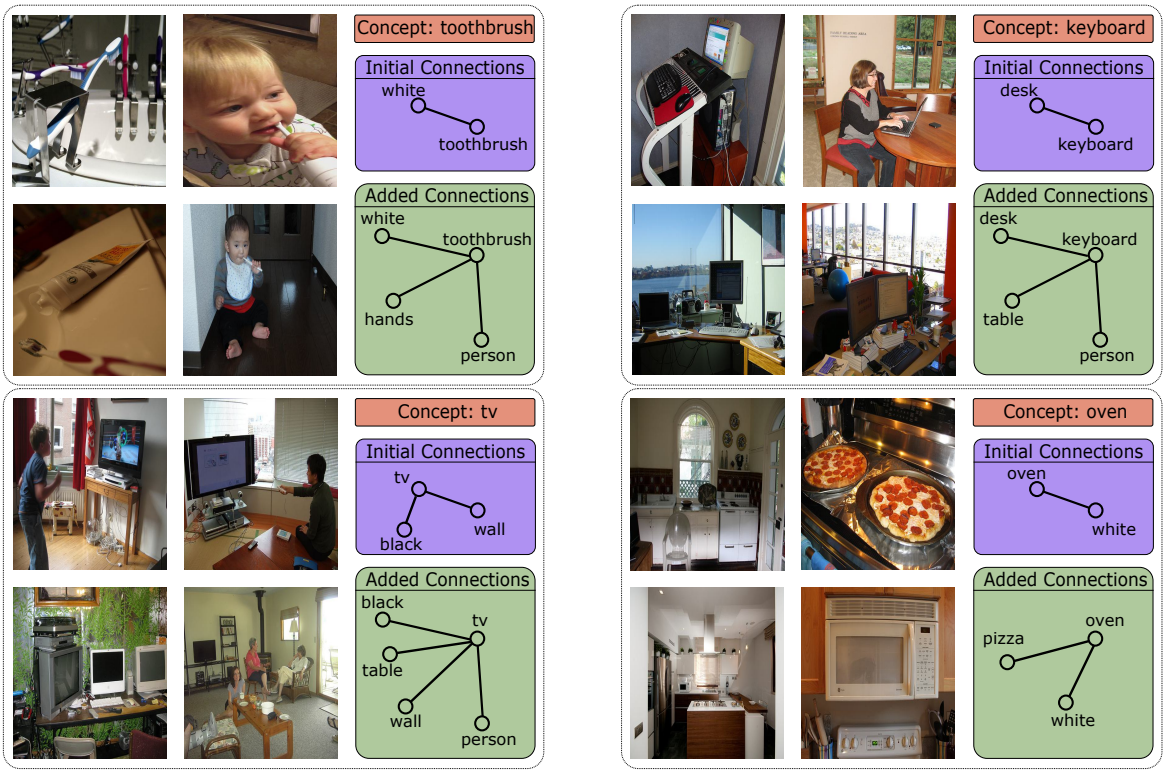


Figure 6: Qualitative Evaluation of Edges added by the `RelaTe` approach. In each example, we include the concept being added, the edges that were present in the knowledge graph originally, and the nodes that were suggested by `RelaTe` for a set of images.

Steps of Expansion, T	Expansion %	mAP
2	39.5	39.1
3	93.4	**42.0**
4	100.0	41.3

Table 6: Percentage of samples that required T steps of expansion and the corresponding mAP performance of our model with that T.

A.5 Countering the Classifier Bottleneck

We demonstrate the crucial role of fine-tuning the propagation network and the node bias when adding novel concepts to the graph. In Figure 7, we plot the mAP performance on the entire VGML dataset as a function of the number of novel classes added to the system for the model where we fine-tune either only the classification or both classification and the propagation module including the node biases. During training with five images per concept, we utilize the one-by-one node addition strategy which showed improved performance (see Figure 5). Initially, for just a few nodes, not training the GSNN does not have a huge influence; however, the plot shows that the model in which we only fine-tune the classifier experiences a substantial performance drop which is proportional to the number of concepts added compared to the model in which both the modules are fine-tuned.

A.6 Ablation on Number of Propagation Steps

We experiment with different values for T that define the number of iterations between the propagation and importance network during inference of the GSNN module as described in Section 3.1.2. We aim to select the minimum possible value of T that ensures the complete expansion of most of the samples in our test dataset within the first T steps. In Table 6, we report the performance of our model on all the classes of the VGML test dataset along with the percentage of samples that were expanded to full capacity by varying the number of expansion steps. The results we obtained in Table 6 highlight that 3 is the optimal value of T since we start to obtain diminishing returns following steps greater than 3. Expanding less than two steps doesn't allow the model to experience many relevant connections while expanding beyond the third level makes it challenging for the model to identify concepts that are related to the original concepts $\mathbb{F}_{\boldsymbol{I}}$.

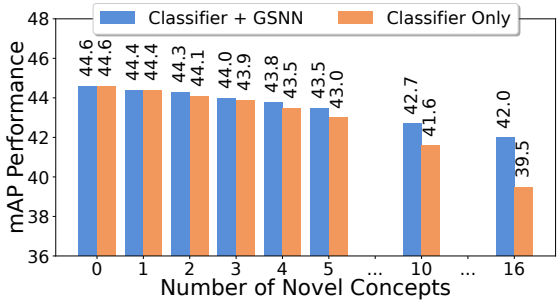


Figure 7: Fine-tuning GSNN + Classifier vs Classifier Only

A.7 Significance of Image Conditioning on Node Embeddings

We explicitly enforce conditioning of the image content on the embeddings generated for each of the nodes during the graph propagation. These embeddings are utilized by both the importance and the context network and form the backbone of the entire graph expansion procedure. Unlike Marino et al. (2016), which does not enforce this constraint, our model prioritizes expanding nodes that are relevant to the image content rather than simply expanding nodes that are only dependent on the initial class detection that would result in the same propagated nodes even for dissimilar images. We demonstrate the significance of our proposed conditioning by selecting three vastly distinct images from the dataset in Figure 8. The original GSNN fails to distinguish between these images in terms of expanded nodes in the KG, whereas our approach expands a unique set of nodes for each image. The following are the final classifications with and without image conditioning on the propagation network:

- Image 1:
  - w/o Conditioning: *person*, *bench*, *shirt*, *black*, *white*, *gray*
  - w Conditioning: *person*, *bench*, *shirt*, *wooden*, *brown*, *black*, *sunglasses*
- Image 2:
  - w/o Conditioning: *person*, *bench*, *green*, *sitting*, *shirt*, *white*

Figure 8: Samples from the dataset where the initial propagation begins with the concepts of *person* and *bench*.

  - w Conditioning: *person*, *bench*, *jacket*, *green*, *visible*, *sitting*

- Image 3:
  - w/o Conditioning: *person*, *bench*, *shirt*, *sitting*, *pink*, *wooden*, *black*
  - w Conditioning: *person*, *bench*, *shirt*, *black*, *jacket*, *head*, *wooden*

While the model without image conditioning expands a generic list of nodes, our approach identifies image-specific concepts such as *sunglasses* for the first image and *jacket* for the second image, demonstrating the improvements imposed by this additional conditioning.

A.8 Ablation for Node types and Edge types

	Edge types	Node types	mAP
1	✓	-	42.3
2	-	-	42.1
3	-	✓	**44.6**

Table 7: Experimental results with ablations of edge and node types.

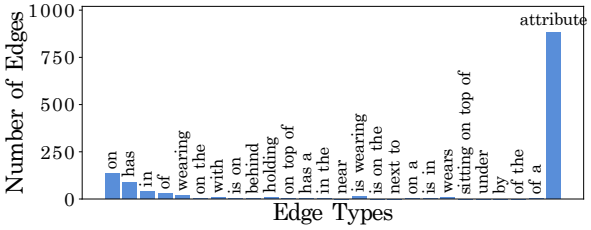
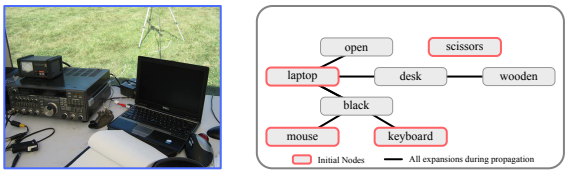


Figure 9: Edge type distribution of the KG used by Marino et al. (2016).

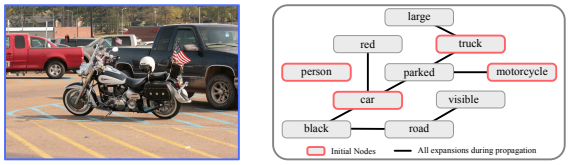
In Section 4.1, we introduced the changes to the KG as described in Marino et al. (2016). Here, we ablate these choices in greater detail. Table 7 shows the performance of the algorithm with the original 26 edge types in line 1, no edge types (i.e. just a single unlabeled edge) in line 2, and our modified KG without edge types, but an additional one-hot indicating the node-type in line 3. The results show that edge types hinder the performance of the inference pipeline and indicating the node type improves performance. We hypothesize that this is due to the strong imbalance of the encoded edge types, as shown in Figure 9, where the *has attribute*, comprising almost two-thirds of all edge types.

A.9 Maximally Diverse Expansion Sampling

To select a small subset of the original dataset that allows us to maximize the diversity of expanded nodes in our KG, we adopt a binning-based approach. We begin with a single bin spanning all nodes and traverse the dataset to identify the image that can expand a node in the largest bin. Upon finding such an image, we use each expanded node in that image as the dividing line between the new bins. If an image does not expand a node that would divide the largest bin, the image is not added to our curated dataset $\mathcal{D}_C$. We only process the dataset once until either we have a set of images that expand all the possible nodes or all images have been either added to $\mathcal{D}_C$ or have been discarded. Under the assumption that rare classes are randomly distributed in the dataset, we ensure that at least some images containing that class are added to $\mathcal{D}_C$. As a result, we create a dataset $\mathcal{D}_C$ containing approximately 2% of the original VGML dataset.

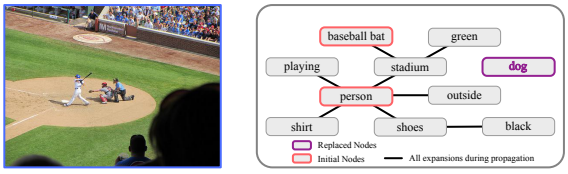


(a) Mitigated failure case: While Faster R-CNN detected a *scissors*, our propagation and importance network did not incorporate this node any further.

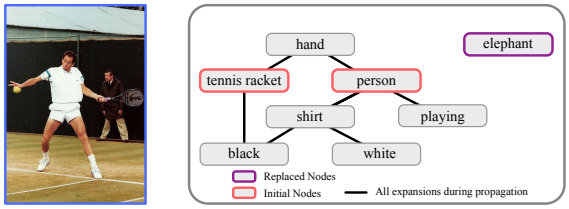


(b) Mitigated failure case: While Faster R-CNN detected a *person*, our propagation and importance network did not incorporate this node any further.

Figure 10: Robustness to wrong graph initialization by Faster R-CNN detections.



(a) Node Replacement: Here, we removed the *sports ball* and replaced it with a *dog*, demonstrating how our approach does not incorporate the wrong node.



(b) Node Replacement: Here, we replaced the *sports ball* with an *elephant*. An interesting result here is that not only was the wrong node not expanded, but it was also removed from the final classification.

Figure 11: Robustness to wrong graph initializations that are manually enforced.

A.10 Analysis of Dependence on Object Detectors

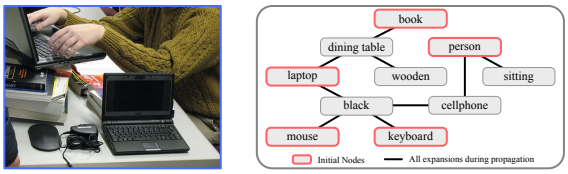
To test the resilience of our model against inaccuracies in the object detection module, we conducted an evaluation by replacing the detected objects with random concepts (that were not originally present in the respective example), and observing whether our model expands upon them. We conducted a small-scale experiment on 30 test images, where we introduced an additional random node that is unrelated to the actual image. We observed that the propagation and importance networks ignore these wrong nodes in 63% of the cases by not expanding them any further. Further, in 16.7% of the cases, the final classifier removes these nodes altogether. In the current work, the importance network can not remove previously added nodes; however, this capability could be explored in future work. Figure 10 demonstrates two instances where Faster R-CNN mistakenly detects a non-existent object class in the image. Furthermore, we provide a few instances in Figure 11 where we substitute one of the original object detections in the image with an entirely unrelated object, and our model refrains from further propagating the modified node, demonstrating its resistance to such potential issues.

Additionally, we also analyzed a potential failure case in which wrong edges exist in the graph. The only potential source for such edges is if `RelaTe` predicts wrong edges during novel concept addition. When testing the performance of `RelaTe` by removing a known node for which the desired edges are known, we observe that 84% of these edges are restored when re-adding the target node using our approach. However, while the remaining 16% of edges are not necessarily wrong, we analyzed the impact of potentially wrong edges by manually introducing them between the initially detected nodes and an arbitrary, unrelated node. This was evaluated on 30 images, as for the prior experiments. We observe that in 76.6% of the cases, the propagation and importance network ignore this wrong connection.

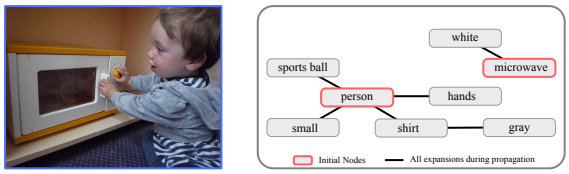
This highlights the robustness of our model to erroneous initialization. Moreover, we empirically observed that Faster R-CNN rarely introduces wrong nodes, thus further mitigating this potential error source.

A.11 Failure Analysis

As part of our failure analysis, we highlight some examples where our model hallucinates non-existent concepts in the image. Although such misclassifications are not common, they offer valuable insights into how our approach functions and where it may be prone to errors.
In the given scenario depicted in Figure 12a, the model has incorrectly identified the concept of a *cellphone*. This



(a) Failure case: The model identified an erroneously identified and further integrated a *cellphone*.



(b) Failure case: The model identified an erroneously identified and further integrated a *sports ball*.

Figure 12: Failure cases of our model in which wrong nodes are integrated into the graph.

error can be attributed to the model's tendency to associate objects with certain visual characteristics, which can result in confusion between objects that share common properties. For instance, in this case, both the *laptop* and *cellphone* have a screen, and therefore the affordance of being able to display something, leading to the misidentification of the object as a *cellphone*. The second example in Figure 12b demonstrates another instance where our model has made an incorrect prediction by identifying the object in the image as a *sports ball*. This error can be attributed to the model's tendency to rely on the way people interact with objects in the scene when identifying them. In this case, the child's hand gripping the knob of the *microwave* may resemble the way one would grip a ball, leading the model to mistakenly classify it as a *sports ball*.

Finally, we evaluate potential failure cases in which `RelaTe` may be tasked to add edges between contextually unrelated nodes. It is a key feature of `RelaTe` to automatically determine the nodes that are relevant for a novel concept while not adding edges to nodes that are contextually different. To evaluate this, we attempt to add edges between nodes from the *bedroom* context and nodes from the *stadium* context. In this case, we observe that `RelaTe` only adds an edge in 20% of the queried connections. However, it is important to note that some connections are in fact reasonable, as connections between the *person* node in the stadium context have a valid connection to *bed* in the bedroom context.

A.12 Runtime Complexity

We trained our model on a single RTX 6000 GPU for ≈ 100 hours of total training time. When adding a novel concept, the two-staged tuning of our model takes approximately 45 minutes. Finally, during inference, it takes approximately 30 seconds per image to obtain predictions using our approach.