Concept Learners for Generalizable Few-Shot Learning

Kaidi Cao* **Maria Brbić*** **Jure Leskovec**
Department of Computer Science
Stanford University
{kaidicao, mbrbic, jure}@cs.stanford.edu

Abstract

Developing algorithms that are able to generalize to a novel task given only a few labeled examples represents a fundamental challenge in closing the gap between machine- and human-level performance. The core of human cognition lies in the structured, reusable concepts that help us to rapidly adapt to new tasks and provide reasoning behind our decisions. However, existing meta-learning methods learn complex representations across prior labeled tasks without imposing any structure on the learned representations. Here we propose COMET, a meta-learning method that improves generalization ability by learning to learn along human-interpretable concept dimensions. Instead of learning a joint unstructured metric space, COMET learns mappings of high-level concepts into semi-structured metric spaces, and effectively combines the outputs of independent concept learners. We evaluate our model on few-shot tasks from diverse domains, including a benchmark image classification dataset and a novel single-cell dataset from a biological domain developed in our work. COMET significantly outperforms strong meta-learning baselines, achieving 9–12% average improvement on the most challenging 1-shot learning tasks, while unlike existing methods also providing interpretations behind the model's predictions.

1 Introduction

Deep learning has reached human-level performance on domains with the abundance of large-scale labeled training data. However, learning on tasks with a small number of annotated examples is still an open challenge. Due to the lack of training data, models often overfit or are too simplistic to provide good generalization. On the contrary, humans can learn new concepts very quickly by drawing upon prior knowledge and experience. This ability to rapidly learn and adapt to new environments is a hallmark of human intelligence.

Few-shot learning [1–3] aims at addressing this fundamental challenge by designing algorithms that are able to generalize to new tasks given only a few labeled training examples. Meta-learning [4, 5] has recently made major advances in the field by explicitly optimizing the model's ability to generalize, or learning how to learn, from many related tasks [6–9]. Motivated by the way humans effectively use prior knowledge, meta-learning algorithms acquire prior knowledge over previous tasks so that new tasks can be efficiently learned from a small amount of data. However, recent works [10, 11] show that simple baseline methods perform comparably to existing meta-learning methods, opening the question about which components are crucial for rapid adaptation and generalization.

Here, we argue that there is an important missing piece in this puzzle. Human knowledge is structured in the form of reusable concepts. For instance, when we learn to recognize new bird species we

*The two first authors made equal contributions.

Preprint. Under review.

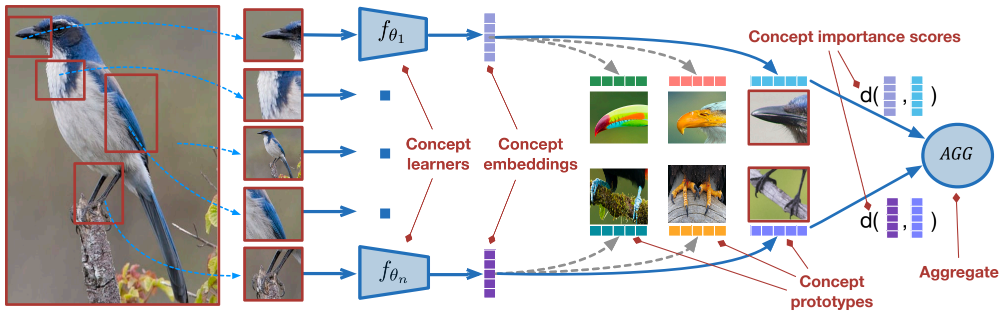

Figure 1: Along each concept dimension, COMET learns concept embeddings using independent concept learners and compares them to concept prototypes. COMET then effectively aggregates information across concept dimensions, assigning concept importance scores to each dimension.

are already equipped with the critical concepts, such as wing, beak, and feather. We then focus on these specific concepts and combine them to identify a new species. While learning to recognize new species is challenging in the complex bird space, it becomes remarkably simpler once the reasoning is structured into familiar concepts. Moreover, such a structured way of cognition gives us the ability to provide reasoning behind our decisions, such as "ravens have thicker beaks than crows, with more of a curve to the end". We argue that this lack of structure is limiting the generalization ability of the current meta-learners. The importance of compositionality for few-shot learning was emphasized in [12, 13] where hand-designed features of strokes were combined using Bayesian program learning.

Motivated by the structured form of human cognition, we propose COMET, a meta-learning method that discovers generalizable representations along human-interpretable concept dimensions. COMET learns a unique metric space for each *concept dimension* using concept-specific embedding functions, named *concept learners*, that are parameterized by deep neural networks. Along each high-level dimension, COMET defines *concept prototypes* that reflect class-level differences in the metric space of the underlying concept. To obtain final predictions, COMET effectively aggregates information from diverse concept learners and concept prototypes. Three key aspects lead to a strong generalization ability of our approach: (i) semi-structured representation learning, (ii) concept-specific metric spaces described with concept prototypes, and (iii) ensembling of many models. The latter assures that the combination of diverse and accurate concept learners improves the generalization ability of the base learner [14, 15]. Remarkably, the high-level universe of concepts that are used to guide our algorithm can be discovered in a fully unsupervised way, or we can use external knowledge bases to define concepts. In particular, we can get a large universe of noisy, incomplete and redundant concepts and COMET learns which subsets of those are important by assigning local and global concept importance scores. Unlike existing methods [6, 7, 16, 17], COMET's predictions are interpretable—an advantage especially important in the few-shot learning setting, where predictions are based only on a handful of labeled examples making it hard to trust the model. As such, COMET is the first domain-agnostic interpretable meta-learning approach.

We demonstrate the effectiveness of our approach in two extremely diverse domains, computer vision and biology. To define concepts on the benchmark bird image classification dataset, we use already available body part-based annotations. Additionally, we consider the scenario in which concepts are not given in advance, and test COMET's performance with automatically extracted visual concepts. Going beyond computer vision applications, we conduct the first systematic comparison of meta-learning algorithms in the biological domain. We develop a new meta-learning dataset and define a novel benchmark task to characterize single-cell transcriptome of all mouse organs [18, 19], publicly available at `https://snap.stanford.edu/comet`. Our experimental results show that on both domains COMET significantly improves generalization ability, achieving 9–12% relative improvement over state-of-the-art methods in the most challenging 1-shot task. Furthermore, we demonstrate the ability of COMET to provide interpretations behind the model's predictions, and support our claim with quantitative and qualitative evaluations of the generated explanations.

2 Proposed method

Problem formulation. In few-shot classification, we assume that we are given a labeled training set $\mathcal{D}_{tr}$, an unlabeled query set $\mathcal{D}_{qr}$, and a support set $\mathcal{S}$ consisting of a few labeled data points.

Each labeled data point $(\mathbf{x}, y)$ consists of a D-dimensional feature vector $\mathbf{x} \in \mathbb{R}^D$ and a class label $y \in \{1, ..., K\}$. The label space between training and test set is disjoint, *i.e.*, $\{Y_{tr}\} \cap \{Y_{qr}\} = \emptyset$, whereas support set and query set share label space, *i.e.*, $\{Y_S\} = \{Y_{qr}\}$. Given a training set of previously labeled tasks $\mathcal{D}_{tr}$ and the support set $\mathcal{S}$ of a few labeled data points on a novel task, the goal is to train a model that can generalize to the novel task and label the query set $\mathcal{D}_{qr}$.

2.1 Preliminaries

Episodic training. To achieve successful generalization to a new task, training of meta-learning methods is usually performed using sampled mini-batches called episodes [7]. Each episode is formed by first sampling classes from the training set, and then sampling data points labeled with these classes. The sampled data points are divided into disjoint sets of: (i) a support set consisting of a few labeled data points, and (ii) a query set consisting of data points whose labels are used to calculate a prediction error. Given the sampled support set, the model minimizes the loss on the sampled query set in each episode. The main idea behind this meta-learning training scheme is to improve generalization of the model by trying to mimic the low-data regime encountered during testing. Episodes with balanced training sets are usually referred to as "N-way, k-shot" episodes where N indicates number of classes per episode ("way"), and k indicates number of support points (labeled training examples) per class ("shot").

Prototypical networks. Our work is inspired by prototypical networks [6], a simple but highly effective metric-based meta-learning method. Prototypical networks learn a non-linear embedding function $f_{\boldsymbol{\theta}} : \mathbb{R}^D \to \mathbb{R}^M$ parameterized by a convolutional neural network. The main idea is to learn a function $f_{\boldsymbol{\theta}}$ such that in the M-dimensional embedding space data points cluster around a single prototype representation $\mathbf{p}_k \in \mathbb{R}^M$ for each class k. Class prototype $\mathbf{p}_k$ is computed as the mean vector of the support set labeled with the class k:

$$\mathbf{p}_k = \frac{1}{|\mathcal{S}_k|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}_k} f_{\boldsymbol{\theta}}(\mathbf{x}_i), \tag{1}$$

where $\mathcal{S}_k$ denotes the subset of the support set $\mathcal{S}$ belonging to the class k. Given a query data point $\mathbf{x}_q$, prototypical networks output distribution over classes using the softmax function:

$$p_{\boldsymbol{\theta}}(y = k|\mathbf{x}_q) = \frac{\exp(-d(f_{\boldsymbol{\theta}}(\mathbf{x}_q), \mathbf{p}_k))}{\sum_{k'} \exp(-d(f_{\boldsymbol{\theta}}(\mathbf{x}_q), \mathbf{p}_{k'}))}, \tag{2}$$

where $d : \mathbb{R}^M \to \mathbb{R}$ denotes the distance function. Query data point $\mathbf{x}_q$ is assigned to the class with the minimal distance between the class prototype and embedded query point.

2.2 Meta-learning via concept learners

Our main assumption is that input dimensions can be separated into subsets of related dimensions corresponding to high-level, human-interpretable concepts that guide the training. Such sets of potentially overlapping, noisy and incomplete human-interpretable dimensions exists in many real-world scenarios. For instance, in computer vision concepts can be assigned to image segments; in natural language processing to semantically related categories; whereas in biology we can use external knowledge bases and ontologies. In many problems, concepts are already available as a prior domain knowledge [20–23], or can be automatically generated using existing techniques [24–26]. Intuitively, concepts can be seen as part-based representations of the input and reflect the way humans reason about the world. Importantly, we do not assume these concepts are clean or complete. On the contrary, we show that even if there are thousands of concepts, which are noisy, incomplete, overlapping, or redundant, they still provide useful guidance to the meta-learning algorithm.

Formally, let $\mathcal{C} = \{\mathbf{c}^{(j)}\}_{j=1}^N$ denote a set of N concepts, where each concept $\mathbf{c}^{(j)} \in \{0, 1\}^D$ is given as a binary vector such that $c_i^{(j)} = 1$ if i-th dimension should be used to describe the j-th concept. We do not impose any constraints on $\mathcal{C}$, meaning that the concepts can be disjoint or overlap.

Instead of learning single mapping function $f_{\boldsymbol{\theta}} : \mathbb{R}^D \to \mathbb{R}^M$ across all dimensions, COMET separates original space into subspaces of predefined concepts and learns individual embedding functions $f_{\boldsymbol{\theta}}^{(j)} : \mathbb{R}^{D^{(j)}} \to \mathbb{R}^M$ for each concept j, where $D^{(j)}$ are dimensions of j-th concept (Figure 1).

Concept embedding functions $f_{\boldsymbol{\theta}}^{(j)}$, named *concept learners*, are non-linear functions parametrized by a deep neural network. Each concept learner j produces its own *concept prototypes* $\mathbf{p}_k^{(j)}$ for class k computed as the average of concept embeddings of data points in the support set:

$$\mathbf{p}_k^{(j)} = \frac{1}{|\mathcal{S}_k|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}_k} f_{\boldsymbol{\theta}}^{(j)}(\mathbf{x}_i \circ \mathbf{c}^{(j)}), \tag{3}$$

where $\circ$ denotes Hadamard product. As a result, each class k is represented with a set of N concept prototypes $\{\mathbf{p}_k^{(j)}\}_{j=1}^N$.

Given a query data point $\mathbf{x}_q$, we compute its concept embeddings and estimate their distances to the concept prototypes of each class. We then aggregate the information across all concepts by taking sum over distances between concept embeddings and concept prototypes. Specifically, for each concept embedding $f_{\boldsymbol{\theta}}^{(j)}(\mathbf{x}_i \circ \mathbf{c}^{(j)})$ we compute its distance to concept prototype $\mathbf{p}_k^{(j)}$ of a given class k, and sum distances across all concepts to obtain a distribution over support classes. The probability of assigning query point $\mathbf{x}_q$ to k-th class is then given by:

$$p_{\boldsymbol{\theta}}(y = k|\mathbf{x}_q) = \frac{\exp(-\sum_j d(f_{\boldsymbol{\theta}}^{(j)}(\mathbf{x}_q \circ \mathbf{c}^{(j)}), \mathbf{p}_k^{(j)}))}{\sum_{k'} \exp(-\sum_j d(f_{\boldsymbol{\theta}}^{(j)}(\mathbf{x}_q \circ \mathbf{c}^{(j)}), \mathbf{p}_{k'}^{(j)}))}. \tag{4}$$

The loss is computed as the negative log-likelihood $L_{\boldsymbol{\theta}} = -\log p_{\boldsymbol{\theta}}(y = k|\mathbf{x}_q)$ of the true class, and COMET is trained by minimizing the loss on the query samples of training set in the episodic fashion [6, 7]. In equation (5), we use euclidean distance as the distance function. Experimentally, we find that it outperforms cosine distance (Appendix A), which agrees with the theory and experimental findings in [6]. We note that in order for distances to be comparable, it is crucial to normalize neural network layers using batch normalization [27].

2.3 Interpretability

Local and global concept importance scores. In COMET, each class is represented with N concept prototypes. Given a query data point $\mathbf{x}_q$, COMET assigns local concept importance scores by comparing concept embbeddings of the query to concept prototypes. Specifically, for a concept j in a class k the local importance score is obtained by inverted distance $d(f_{\boldsymbol{\theta}}^{(j)}(\mathbf{x}_q \circ \mathbf{c}^{(j)}), \mathbf{p}_k^{(j)})$. Higher importance score indicates higher contribution in classifying query point to the class k. Therefore, explanations for the query point $\mathbf{x}_q$ are given by local concept importance scores, and directly provide reasoning behind each prediction. To provide global explanations that can reveal important concepts for a set of query points of interest or an entire class, COMET computes average distance between concept prototype and concept embeddings of all query points of interest. Inverted average distance reflects global concept importance score and can be used to rank concepts, providing insights on important concepts across a set of examples.

Discovering locally similar examples. Given a fixed concept j, COMET can be used to rank data points based on the distance of their concept embeddings to the concept prototype $\mathbf{p}_k^{(j)}$ of class k. By ranking data points according to their similarity to the concept of interest, COMET can find examples that locally share similar patterns within the same class, or even across different classes. For instance, COMET can reveal examples that well reflect a prototypical concept, or examples that are very distant from the prototypical concept.

3 Experiments

3.1 Experimental setup

Datasets. We apply COMET to datasets from two extremely diverse domains: computer vision and biology. In the computer vision domain, we consider the fine-grained image classification task. We use CUB-200-2011 (referred as CUB hereafter) [22] dataset, which contains $11,788$ images of 200 bird species. CUB is a standard and commonly used meta-learning benchmark dataset. To define concepts, CUB provides part-based annotations, such as beak, wing, and tail of a bird. Parts were annotated by pixel location and visibility in each image. The total number of 15 parts/concepts is available; however concepts are incomplete and only a subset of them is present in an image. In case concept is not present, we rely on the prototypical concept to substitute for a missing concept. Based on the part coordinates, we create a surrounding bounding box with a fixed length to serve as the concept mask $\mathbf{c}^{(j)}$. We followed the evaluation protocol in [10] and split the dataset into 100 base, 50 validation, and 50 test classes in the exactly same split.

In the biology domain, we introduce a new cross-organ cell type classification task together with a new dataset. We develop a novel single-cell transcriptomic dataset based on the Tabula Muris dataset [18, 19] that comprises $105,960$ cells of 124 cell types collected across 23 organs of the mouse model organism. The features correspond to the gene expression profiles of cells. Out of the $23,341$ genes, we select $2,866$ genes with high standardized log dispersion given their mean. We define concepts using Gene Ontology [20, 28], a resource which characterizes gene functional roles in a hierarchically structured vocabulary. We select gene ontology terms at level 3 that have at least 64 assigned genes, resulting in the total number of 190 terms that define our concepts. Note that the assignment is not exclusive so the same gene might be assigned to multiple gene ontology terms. In both datasets, we include a concept that captures the whole input, which corresponds to a binary mask of all ones. We propose the evaluation protocol in which different organs are used for training, validation, and test splits. Therefore, a meta-learner needs to learn to generalize to unseen cell types across organs. In particular, we use 15 organs for training, 4 organs for validation, and 4 organs for test, resulting into 59 base, 47 validation, and 37 test classes corresponding to cell types. This novel dataset will be made publicly available at the time of the publication along with the cross-organ evaluation splits. To our knowledge, this is the first meta-learning dataset from the biology domain.

Implementation details. For CUB, we use the widely adopted four-layer and six-layer convolutional backbones (Conv-4, Conv-6). We perform standard data augmentation, including random crop, rotation, horizontal flipping and color jittering. For the Tabula Muris dataset, we use a simple backbone network containing two fully-connected layers with batch normalization, ReLu activation and dropout. More details are provided in the Appendix C. Code is publicly available at `https://github.com/snap-stanford/comet`.

Evaluation. We test all methods on the most broadly used 5-way classification setting. In each episode, we randomly sample 5 classes where each class contains k examples as the support set in the k-shot classification task. We construct the query set to have 16 examples, where each unlabeled sample in the query set belongs to one of the classes in the support set. We choose the best model according to the validation accuracy, and then evaluate it on the test set with novel classes. We report the mean accuracy by randomly sampling 600 episodes in the fine-tuning or meta-testing stage.

3.2 Results

Performance comparison. We compare COMET's performance to five baselines, including FineTune/Baseline++ [10], Matching Networks (MatchingNet) [7], Model Agnostic Meta-Learning (MAML) [9], Relation Networks [16], and Prototypical Networks (ProtoNet) [6]. COMET outperforms all baselines by a remarkably large margin on both CUB and Tabula Muris datasets (Table 1). Specifically, on the CUB dataset COMET achieves 9% relative improvement over the best performing baseline in 1-shot task, and 6–7% improvement in the 5-shot task depending on the backbone network used. Note that the best performing baseline is changing depending on the classification task and the backbone network, but the performance improvements attained by COMET remain consistent across architectures and tasks. On the single-cell Tabula Muris dataset, COMET achieves 11–12% improvement over the best performing baseline on both classification tasks. Strikingly, COMET improves the result of the ProtoNet baseline by 19% and 23% in the 1-shot task on the CUB and Tabula Muris datasets, respectively. These improvements clearly demonstrate the power of introducing concept learners. As an additional baseline, we compare our concept aggregation approach to the majority voting approach in which each concept learner votes on the final prediction. Whereas majority voting outperforms all other baselines supporting the importance of introducing concept learners, our aggregation approach significantly outperforms majority voting on both domains (Appendix B).

Effect of number of concepts. We further systematically evaluate the effect of the number of concepts on COMET's performance (Figure 2). In particular, we start from ProtoNet's result that can be seen as using a single concept in COMET that covers all dimensions of the input. We then gradually increase number of concepts and train and evaluate COMET with the selected number of concepts. For the CUB dataset, we add concepts based on their visibility frequency, whereas on the

Table 1: Results on 1-shot and 5-shot classification on the CUB and Tabula Muris datasets. We report average accuracy and standard deviation over 600 randomly sampled episodes.

Method	CUB: Conv-4 1-shot	CUB: Conv-4 5-shot	CUB: Conv-6 1-shot	CUB: Conv-6 5-shot	Tabula Muris 1-shot	Tabula Muris 5-shot
Finetune	61.4 ± 1.0	80.2 ± 0.6	66.0 ± 0.9	82.0 ± 0.6	65.3 ± 1.0	82.1 ± 0.7
MatchingNet	61.0 ± 0.9	75.9 ± 0.6	66.5 ± 0.9	77.9 ± 0.7	71.0 ± 0.9	82.4 ± 0.7
MAML	52.8 ± 1.0	74.4 ± 0.8	66.3 ± 1.1	78.8 ± 0.7	50.4 ± 1.1	57.4 ± 1.1
RelationNet	62.1 ± 1.0	78.6 ± 0.7	64.4 ± 0.9	80.2 ± 0.6	69.3 ± 1.0	80.1 ± 0.8
ProtoNet	57.1 ± 1.0	76.1 ± 0.7	66.4 ± 1.0	82.0 ± 0.6	64.5 ± 1.0	82.5 ± 0.7
COMET	$\mathbf{67.9 \pm 0.9}$	$\mathbf{85.3 \pm 0.5}$	$\mathbf{72.2 \pm 0.9}$	$\mathbf{87.6 \pm 0.5}$	$\mathbf{79.4 \pm 0.9}$	$\mathbf{91.7 \pm 0.5}$

Tabula Muris we are not limited in the coverage of concepts so we randomly select them. The results demonstrate that on both domains COMET consistently improves performance when increasing the number of concepts. Strikingly, by adding just one most frequent concept corresponding to a bird's beak on top of the whole image concept, we improve ProtoNet's performance on CUB by 10% and 5% in 1-shot and 5-shot tasks, respectively. Moreover, with only the 'beak' concept COMET achieves state-of-the-art performance on CUB, while with 6 concepts it improves the score of the best baseline by 5% on both tasks. On the Tabula Muris, with just 8 concepts COMET significantly outperforms all baselines and achieves 7% and 17% improvement over ProtoNet in 1-shot and 5-shot tasks, respectively. To demonstrate the robustness of our method to a huge set of overlapping concepts, we extend the number of concepts to 1500 by capturing all levels of the Gene Ontology hierarchy, therefore allowing many redundant relationships. Even in this scenario, COMET slightly improves the results compared to 190 concepts obtained from a single level. These results demonstrate that COMET outperforms other methods even when the number of concepts is small and annotations are incomplete, as well as with many overlapping and redundant concepts.

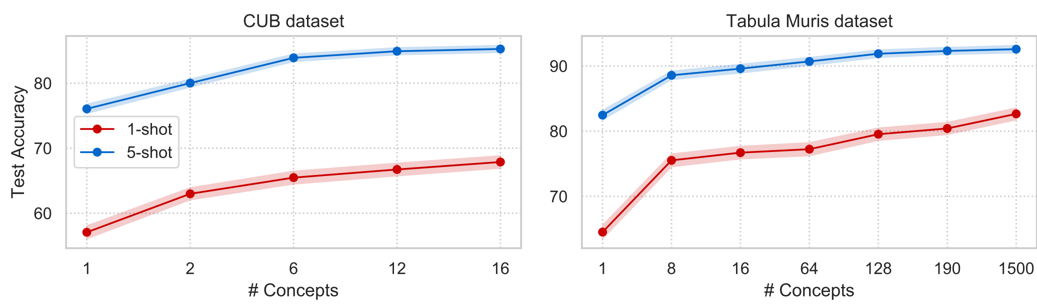

Figure 2: The effect of number of concepts on COMET's performance. COMET consistently improves performance when we gradually increase number of concept terms.

Unsupervised concept annotation. While COMET achieves remarkable results with human-validated concepts given as external knowledge, we next investigate COMET's performance on automatically inferred concepts. To automatically extract visual concepts from the CUB dataset, we train the autoencoding framework for landmarks discovery proposed in [25]. The encoding module outputs landmark coordinates that we use instead of part-based coordinates. We set the number of concepts to 30. We visualize resulting landmarks in Appendix D. Although extracted coordinates are often noisy and capture background, we find that COMET significantly outperforms all other methods (Table 2). This analysis shows that the benefits of our method are expected even with noisy concepts extracted in a fully automated and unsupervised way.

Table 2: Results on 1-shot and 5-shot classification on the CUB dataset with automatically extracted concepts. We report average accuracy and standard deviation over 600 randomly sampled episodes. We show the average relative improvement of COMET over the best and ProtoNet baselines.

Accuracy	**Conv-4: 1-shot**	**Conv-4: 5-shot**	**Conv-6: 1-shot**	**Conv-6: 5-shot**
COMET	63.7 ± 1.0	81.9 ± 0.6	69.8 ± 1.0	84.2 ± 0.6
Improvement of COMET...				
over best baseline	2.5%	2.1%	4.9%	2.6%
over ProtoNet	11.6%	7.6%	5.1%	2.6%

3.3 Interpretability

We analyze the reasoning part of COMET by designing case studies aiming to answer the following questions: (i) Which concepts are the most important for a given query point (*i.e.*, local explanation)? Which concepts are the most important for a given class (*i.e.*, global explanation)?; (iii) Which examples share locally similar patterns?; (iv) Which examples well reflect a prototypical concept? We perform all analyses exclusively on classes from the novel task that are not seen during training.

Concept importance. Given a query point, COMET ranks concepts based on their importance scores, therefore identifying concepts highly relevant for the prediction of a single query point. We demonstrate examples of local explanations in the Appendix E. To quantitatively evaluate global explanations that assign concept importance scores to the entire class, we derive ground truth explanations on the Tabula Muris dataset. Specifically, using the ground truth labels on the test set, we obtain a set of genes that are differentially expressed for each class (*i.e.*, cell type). We then find gene ontology terms that are significantly enriched (false discovery rate corrected p-value< 0.1) in the set of differentially expressed genes of a given class, and use those terms as ground-truth concepts. We consider only cell types that have at least 2 assigned gene ontology terms. To obtain COMET's explanations, we rank global concept importance scores for each class and report number of relevant terms that are successfully retrieved in top 20 concepts with the highest scores in the 5-shot setting (Figure 3a). We find that COMET's importance scores agree remarkably well with the ground truth annotations, achieving 0.71 average recall@20 across all cell types. We further investigate global explanations on the CUB dataset by computing the frequency of the most relevant concepts across the species (Figure 3b). Beak, belly and forehead turn out to be the most relevant features, supporting common-sense intuition. For instance, 'beak' is selected as the most relevant concept for 'parakeet auklet' known for its nearly circular beak; 'belly' for 'cape may warbler' known for its tiger stripes on the belly; while 'belted kingfisher' indeed has characteristic 'forehead' with its shaggy crest on the top of the head. This confirms that COMET correctly identifies important class-level concepts.

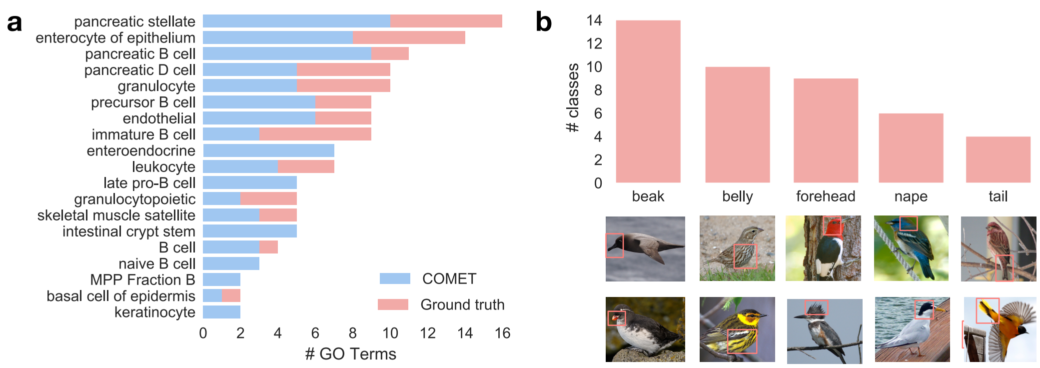

Figure 3: (a) Quantitatively, on the Tabula Muris dataset COMET's global importance scores agree well with the ground truth important gene ontology terms estimated using differentially expressed genes. (b) Qualitatively, on the CUB dataset importance scores correctly reflect the most relevant bird features.

Locally similar patterns. Given a fixed concept of interest, we apply COMET to sort images with respect to the distance of their concept embedding to the concept prototype (Figure 4). COMET finds images that locally resemble the prototypical image and well express prototypical concept, correctly reflecting the underlying concept of interest. On the contrary, images sorted using the whole image as a concept often reflect background similarity and can not provide intuitive explanations. Furthermore, by finding most distant examples COMET can aid in identifying misannotated or non-visible concepts (Appendix F) which can be particularly useful when the concepts are automatically extracted. These analyses suggest that COMET can be used to discover, sort and visualize locally similar patterns, revealing insights on concept-based similarity across examples.

4 Related work

Our work draws motivation from a rich line of research on meta-learning, compositional representations, and concept-based interpretability.

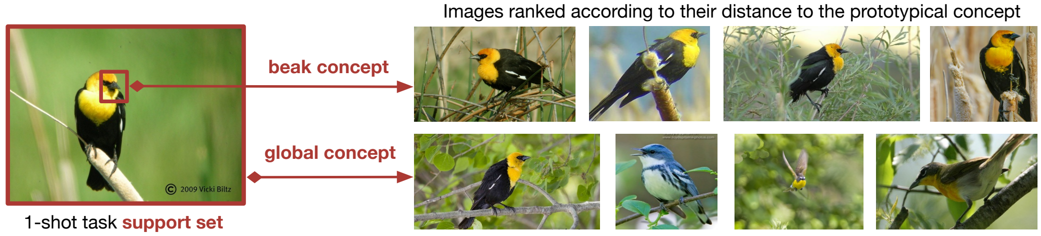

Figure 4: Top row shows images with beak concept embeddings most similar to the prototypical beak. Bottom row shows images ranked according the global concept that captures whole image. COMET correctly reflects local similarity in the underlying concept of interest, while global concept often reflects environmental similarity.

Meta-learning. Recent meta-learning methods fall broadly into two categories. Optimization-based methods [9, 29–32] aim to learn a good initialization such that network can be fine-tuned to a target task within a few gradient steps. Prominent examples include MAML [9], and its extensions Reptile [30] and LEO [29], that formulate the problem as a two level optimization procedure. On the other hand, metric-based methods [6, 7, 16, 17] learn a metric space shared across tasks such that in the new space target task can be solved using nearest neighbour or simple linear classifier. Prototypical networks [6] learn a metric space such that data points cluster around a prototypical representation computed for each category as the mean of embedded labeled examples. It has remained one of the most competitive few-shot learning methods [33], resulting in many follow-up works [16, 34–37]. Our work builds upon prototypical networks.

Compositionality. The idea behind learning from a few examples using compositional representations originates from work on Bayesian probabilistic programs in which individual strokes were combined for the handwritten character recognition task [12, 13]. This approach was extended in [38] by replacing hand designed features with symmetry axis as object descriptors. Although these early works effectively demonstrated that compositionality is a key ingredient for adaptation in a low-data regime, it is unclear how to extend them to generalize beyond simple visual concepts. Recent work [39] revived the idea and showed that deep compositional representations generalize better in few-shot image classification. However, this approach requires category-level attribute annotations that are impossible to get in domains not intuitive to humans, such as biology. Moreover, even in domains in which annotations can be collected, they require tedious manual effort. On the contrary, our approach is domain-agnostic and generates human-understandable interpretations in any domain.

Interpretability. There has been much progress on designing interpretable methods that estimate the importance of individual features [40–45]. However, individual features are often not intuitive, or can even be misleading when interpreted by humans [46]. To overcome this limitation, recent advances have been focused on designing methods that explain predictions using high-level human understandable concepts [46, 47]. TCAV [46] defines concepts based on user-annotated set of examples in which the concept of interest appears. On the contrary, high-level concepts in our work are defined with a set of related dimensions. As such, they are already available in many domains, or can be obtained in an unsupervised manner. Once defined, they are transferable across problems that share feature space. As opposed to the methods that base their predictions on the posthoc analysis [43–46], COMET is designed as an inherently interpretable model and explains predictions by gaining insights from the reasoning process of the network. The closest to our work are prototypes-based explanation models [48, 49]. However, they require specialized convolutional architecture for feature extraction and are not applicable beyond image classification, or to a few-shot setting.

5 Conclusion

We introduced COMET, a novel metric-based meta-learning algorithm that learns to generalize along human-interpretable concept dimensions. We showed that COMET learns generalizable representations with incomplete, noisy, redundant, very few or a huge set of concept dimensions, selecting only important concepts for classification and providing reasoning behind the decisions. Our experimental results showed that COMET does not make a trade-off between interpretability and accuracy and significantly outperforms existing methods on tasks from diverse domains, including a novel benchmark dataset from the biology domain developed in our work.

Acknowledgments and Disclosure of Funding

The authors thank Yusuf Roohani, Michihiro Yasunaga and Marinka Zitnik for their helpful comments. We gratefully acknowledge the support of DARPA under Nos. FA865018C7880 (ASED), N660011924033 (MCS); ARO under Nos. W911NF-16-1-0342 (MURI), W911NF-16-1-0171 (DURIP); NSF under Nos. OAC-1835598 (CINES), OAC-1934578 (HDR), CCF-1918940 (Expeditions), IIS-2030477 (RAPID); Stanford Data Science Initiative, Wu Tsai Neurosciences Institute, Chan Zuckerberg Biohub, Amazon, Boeing, Chase, Docomo, Hitachi, Huawei, JD.com, NVIDIA, Dell. J. L. is a Chan Zuckerberg Biohub investigator. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views, policies, or endorsements, either expressed or implied, of DARPA, NIH, ARO, or the U.S. Government.

References

[1] Erik G Miller, Nicholas E Matsakis, and Paul A Viola. Learning from one example through shared densities on transforms. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 464–471, 2000.

[2] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

[3] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2, 2015.

[4] Jürgen Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. PhD thesis, Technische Universität München, 1987.

[5] Samy Bengio, Yoshua Bengio, Jocelyn Cloutier, and Jan Gecsei. On the optimization of a synaptic learning rule. In *Preprints Conf. Optimality in Artificial and Biological Neural Networks*, volume 2, 1992.

[6] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017.

[7] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pages 3630–3638, 2016.

[8] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International Conference on Learning Representations*, 2017.

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135, 2017.

[10] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations*, 2019.

[11] Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? Towards understanding the effectiveness of MAML. In *International Conference on Learning Representations*, 2020.

[12] Brenden Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua Tenenbaum. One shot learning of simple visual concepts. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33, 2011.

[13] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[14] Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990.

[15] Nikita Dvornik, Cordelia Schmid, and Julien Mairal. Diversity with cooperation: Ensemble methods for few-shot classification. In *IEEE International Conference on Computer Vision*, pages 3723–3731, 2019.

[16] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018.

[17] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4367–4375, 2018.

[18] Tabula Muris Consortium et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, 562(7727):367, 2018.

[19] Angela Oliveira Pisco, Nicholas Schaum, Aaron McGeever, Jim Karkanias, Norma F Neff, Spyros Darmanis, Tony Wyss-Coray, Stephen R Quake, et al. A single cell transcriptomic atlas characterizes aging tissues in the mouse. *bioRxiv*, page 661728, 2019.

[20] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[21] Alexey G Murzin, Steven E Brenner, Tim Hubbard, and Cyrus Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536–540, 1995.

[22] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 dataset. 2011.

[23] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 909–918, 2019.

[24] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[25] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2694–2703, 2018.

[26] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *Advances in Neural Information Processing Systems*, pages 4016–4027, 2018.

[27] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[28] Gene Ontology Consortium. The Gene Ontology resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1):D330–D338, 2019.

[29] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations*, 2019.

[30] Alex Nichol and John Schulman. Reptile: A scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2:2, 2018.

[31] Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical Bayes. In *International Conference on Learning Representations*, 2018.

[32] Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your MAML. In *International Conference on Learning Representations*, 2019.

[33] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Evci, et al. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *International Conference on Learning Representations*, 2020.

[34] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. TADAM: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, pages 721–731, 2018.

[35] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018.

[36] Lu Liu, Tianyi Zhou, Guodong Long, Jing Jiang, and Chengqi Zhang. Learning to propagate for graph meta-learning. In *Advances in Neural Information Processing Systems*, pages 1037–1048, 2019.

[37] Chen Xing, Negar Rostamzadeh, Boris Oreshkin, and Pedro O Pinheiro. Adaptive cross-modal few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4848–4858, 2019.

[38] Alex Wong and Alan L Yuille. One shot learning via compositions of meaningful patches. In *IEEE International Conference on Computer Vision*, pages 1197–1205, 2015.

[39] Pavel Tokmakov, Yu-Xiong Wang, and Martial Hebert. Learning compositional representations for few-shot recognition. In *IEEE International Conference on Computer Vision*, pages 6372–6381, 2019.

[40] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-CAM: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016.

[41] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328, 2017.

[42] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smooth-Grad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[43] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[44] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

[45] David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems*, pages 7775–7784, 2018.

[46] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning*, pages 2668–2677, 2018.

[47] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, pages 9273–9282, 2019.

[48] Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[49] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems*, pages 8928–8939, 2019.

A Ablation study on distance function

We compare the effect of distance metric on the COMET's performance. We find that Euclidean distance consistently outperforms cosine distance in both domains.

Table 3: The effect of distance metric on COMET's performance.

	CUB: Conv-4		**Tabula Muris**	
Distance	**1-shot**	**5-shot**	**1-shot**	**5-shot**
Cosine	65.7 ± 1.0	82.2 ± 0.6	77.1 ± 0.9	90.1 ± 0.6
Euclidean	**67.9 $\pm$ 0.9**	**85.3 $\pm$ 0.5**	**79.4 $\pm$ 0.9**	**91.7 $\pm$ 0.5**

B Ablation study on aggregation

In COMET, we combine concept learners by aggregating distances across all concept prototypes in the softmax function. In this section, we compare our aggregation approach with the majority voting in which each concept learner votes on the final prediction. In particular, we design a following baseline method. Given a query point $\mathbf{x}_q$, concept prototypes $\{\mathbf{p}_k^{(j)}\}_k$ and concept embeddings $f_{\boldsymbol{\theta}}^{(j)}(\mathbf{x}_q \circ \mathbf{c}^{(j)})$ obtained using the COMET, the method outputs probability distribution for each concept j:

$$p_{\boldsymbol{\theta}}^{(j)}(y = k|\mathbf{x}_q) = \frac{\exp(-d(f_{\boldsymbol{\theta}}^{(j)}(\mathbf{x}_q \circ \mathbf{c}^{(j)}), \mathbf{p}_k^{(j)}))}{\sum_{k'} \exp(-d(f_{\boldsymbol{\theta}}^{(j)}(\mathbf{x}_q \circ \mathbf{c}^{(j)}), \mathbf{p}_{k'}^{(j)}))}. \tag{5}$$

Each concept then outputs its own class prediction, and we combine predictions across all concepts with the majority voting. The results shown in Table 4 demonstrate that the aggregation used in COMET significantly outperforms majority voting on both domains.

Table 4: Comparison between aggregation approach used in COMET and majority voting approach in which each concept votes on the final prediction.

	CUB: Conv-4		**Tabula Muris**	
Method	**1-shot**	**5-shot**	**1-shot**	**5-shot**
Concept majority voting	64.0 ± 0.9	82.1 ± 0.6	75.6 ± 0.8	88.2 ± 0.5
COMET's aggregation	**67.9 $\pm$ 0.9**	**85.3 $\pm$ 0.5**	**79.4 $\pm$ 0.9**	**91.7 $\pm$ 0.5**

C Implementation details

For the CUB dataset, we use the widely adopted four-layer and six-layer convolutional backbones (Conv-4, Conv-6) with an input size of 84×84 [2]. We perform standard data augmentation, including random crop, rotation, horizontal flipping and color jittering. We use the Adam optimizer [3] with an initial learning rate of 10^{-3} and weight decay 0. We train the 5-shot tasks for $40,000$ episodes and 1-shot tasks for $60,000$ episodes [4]. To speed up training of COMET, we share the network parameters between concept learners. In particular, we first forward the entire image $\mathbf{x}_i$ into the convolutional network and get a spatial feature embedding $f_{\boldsymbol{\theta}}(\mathbf{x}_i)$, and then get the j-th concept embedding as $f_{\boldsymbol{\theta}}(\mathbf{x}_i) \circ \mathbf{c}^{(j)}$. Since convolutional filters only operate on pixels locally, in practice we get similar performance if we apply the mask at the beginning or at the end while significantly speeding up training time. In case the part is not annotatated (i.e. visible), we use the prototypical concept to replace the missing concept. For the Tabula Muris dataset, we use a simple backbone network structure containing two fully-connected layers with batch normalization, ReLu activation and dropout. We use Adam optimizer [3] with an initial learning rate of 10^{-3} and weight decay 0. We train the network for $1,000$ episodes.

D Unsupervised concept annotation: landmarks examples

To asses the performance of COMET using automatically extracted visual concepts on the CUB dataset, we applied autoencoding framework for landmarks discovery proposed in [1]. We use default parameters and implementation provided by the authors, and set the number of landmarks to 30. The encoding module provides coordinates of the estimated landmarks. To train the autoencoder, we use the same parameters as [1], and set the number of concepts to 30. Examples of extracted landmarks for 20 images from the CUB dataset are visualized in Figure 1.

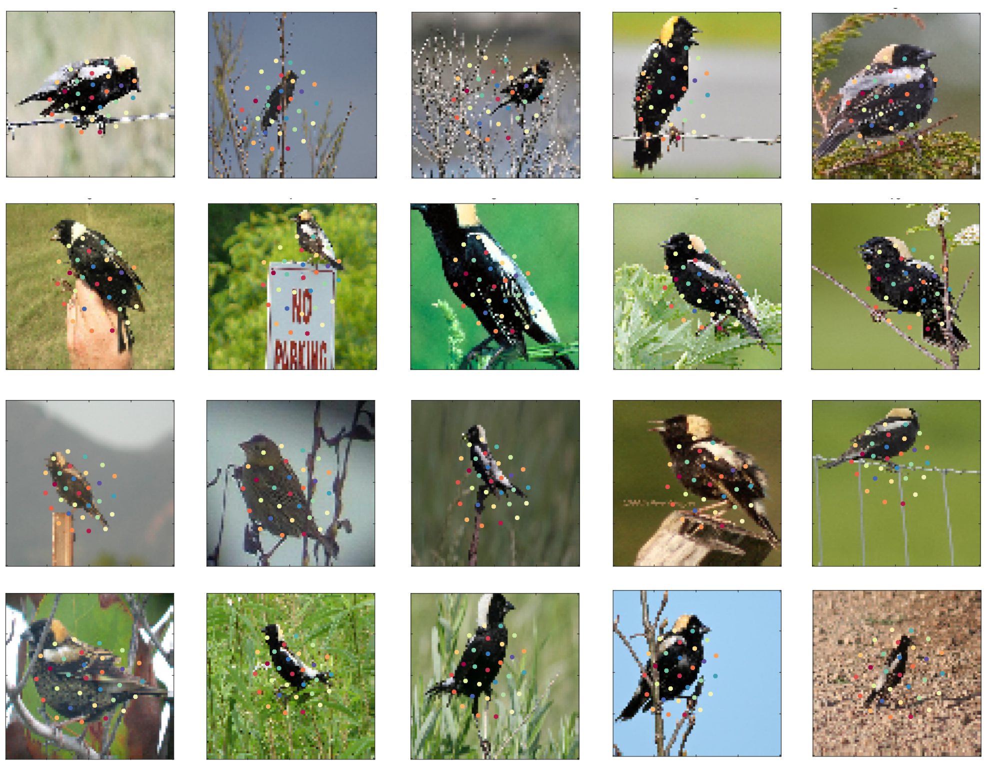

Figure 1: Examples of automatically extracted landmarks using [1] on the CUB dataset.

E Interpretability: local explanations

In this section, we demonstrate COMET's local explanations on the CUB dataset. Given a single query data point, COMET assigns local concept importance scores to each concept based on the distance between concept embedding of the query data point to the prototypical concept. We then rank concepts according to their local concept importance scores. Figure 2 shows examples of ranked concepts. Importance scores assigned by COMET visually reflect well the most relevant bird features.

F Interpretability: local similarity

Given fixed concept of interest, we apply COMET to sort images with respect to the distance of their concept embedding to the concept prototype. Figure 4 shows example of chipping sparrow images with the belly concept embedding most similar to the prototypical belly, and images with the belly concept embedding most distant to the prototypical belly. Most similar images indeed have clearly visible belly part and reflect prototypical belly well. On the contrary, most distant images have only small part of belly visible, indicating that COMET can be used to detect misannotated or non-visible concepts.

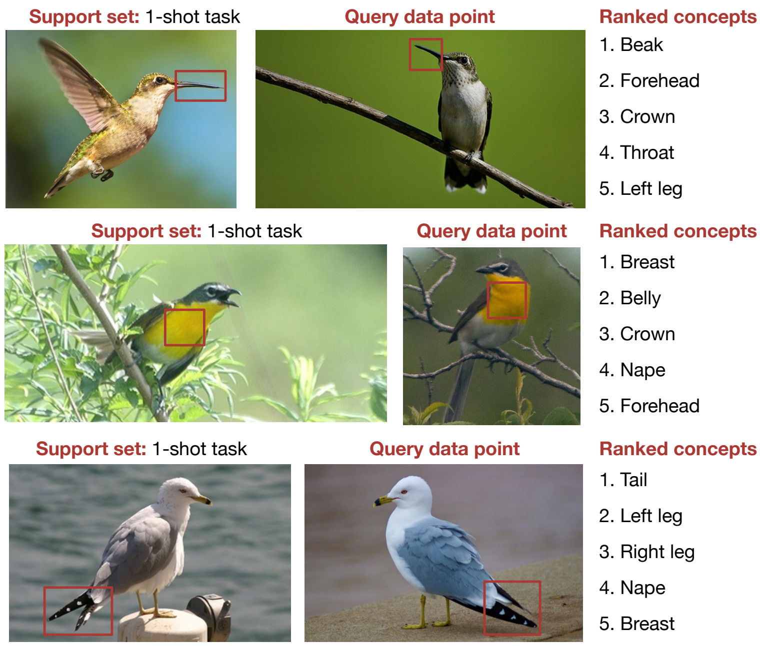

Figure 2: Examples of COMET's local explanations on the CUB dataset. Concepts are ranked according to the highest local concept similarity scores. Qualitatively, local importance scores correctly reflect the most relevant bird features.

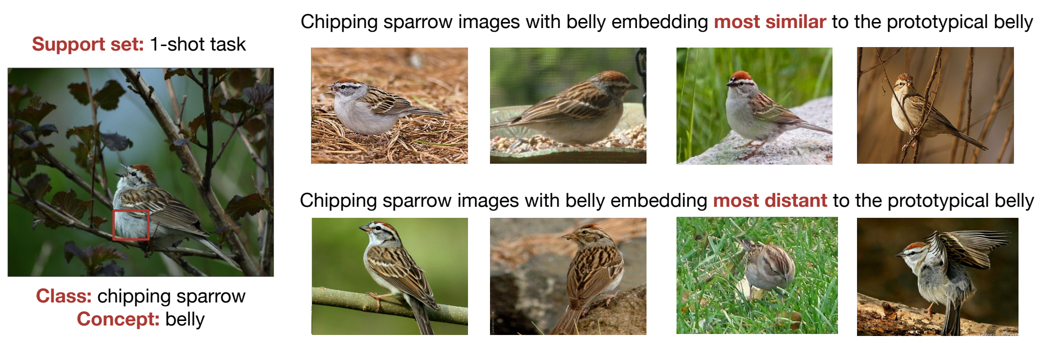

Figure 3: Images ranked according to the distance of their belly concept embedding to the belly concept prototype. Most similar images (top) and most distant images (bottom). Images closest to the prototype have clearly visible belly part that visually looks like prototypical belly of a chipping sparrow, whereas most distant images do not have belly part clearly visible.

References

[1] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* , pages 2694–2703, 2018.

[2] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems* , pages 4077–4087, 2017.

[3] Diederik Kingma, Jimmy Ba, and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[4] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *International Conference on Learning Representations*, 2019.