Few-shot Learning with Contextual Cueing for Object Recognition in Complex Scenes

Mathieu Pagé Fortin
Laval University
Québec, Canada
mathieu.page-fortin.1@ulaval.ca

Brahim Chaib-draa
Laval University
Québec, Canada
brahim.chaib-draa@ift.ulaval.ca

Abstract

Few-shot Learning aims to recognize new concepts from a small number of training examples. Recent work mainly tackle this problem by improving visual features, feature transfer and meta-training algorithms. In this work, we propose to explore a complementary direction by using scene context semantics to learn and recognize new concepts more easily. Whereas a few visual examples cannot cover all intra-class variations, contextual cueing offers a complementary signal to classify instances with unseen features or ambiguous objects. More specifically, we propose a Class-conditioned Context Attention Module (CCAM) that learns to weight the most important context elements while learning a particular concept. We additionally propose a flexible gating mechanism to ground visual class representations in context semantics.

We conduct extensive experiments on Visual Genome dataset, and we show that compared to a visual-only baseline, our model improves top-1 accuracy by 20.47% and 9.13% in 5-way 1-shot and 5-way 5-shot, respectively; and by 20.42% and 12.45% in 20-way 1-shot and 20-way 5-shot, respectively.

1. Introduction

The need for large quantity of data to train Convolutional Neural Networks (CNN) is one the most important drawback of these networks. This severely reduces their applicability in real-world scenarios where the data may be too expensive to annotate or unavailable in sufficient quantity. For this reason, in the last few years an increasing amount of effort has been done in Zero-shot Learning (ZSL) and Few-shot Learning (FSL) to develop approaches that reduce the number of examples required to train efficient models. Progress in this direction will enable new applications such as robots that actively learn new concepts on the fly from their environment [34].

Recent ZSL and FSL approaches generally focus on visual features and tend to consider object instances in isolation [30, 25, 27, 5, 39, 32, 36, 14, 1]. These methods are therefore primarily evaluated on datasets composed of images with only one centered object (e.g. miniImagenet [30], Omniglot [12], CUB-200 [33]).

However, such clean datasets can differ from those encountered in real applications where the target objects would be located in complex scenes. For example, robots are more likely to encounter scenes with several entities such as Fig. 1 than close-ups of isolated objects. This scenario has been neglected so far, and while it can be more relevant for real applications, we argue that complex scenes also offer an interesting opportunity to leverage additional semantic information, thereby facilitating ZSL and FSL. Instead of considering other objects as noise that must be
eliminated (e.g. by only focusing on the image region containing the focal object), we propose a method that exploits the scene context, determined by the presence of other objects, as additional semantic information for learning and recognizing concepts.

Recently, Zablocki et al. [37] introduced the use of contextual information in ZSL. They showed that leveraging visual information and classes of surrounding entities in a complex scene can help recognize unseen objects. Their approach combines the probabilities obtained from three independent models that use contextual, visual and prior information. Our approach differs firstly in that we learn end-to-end a joint representation by grounding class embeddings in their context. This idea is motivated by the first functional feature of context defined by Dohn et al. [6]:

**Supplementary role of context.** [Context] is brought in, or added to, the understanding of a phenomenon on the focal object that would not have been adequately understood had it been considered in isolation. A context thus completes the conditions for understanding the focal object. [6]

Similarly, psychological studies showed that humans learn new concepts by integrating them in their existing conceptual network, rather than only memorizing visual appearance [21]. We apply this principle to FSL by building on Prototypical Networks [25]: our model learns to modify class representations according to the context in which they appear by refining the class prototypes. By doing so, the prototypes do not only rely on visual data but also on complementary semantic information from context. This improves class representations and facilitates the recognition of unseen examples thanks to contextual cueing [3]. The latter is closely related to the second functional feature of context described by Dohn et al. [6]:

**Relative role of context.** The context is centered around the object. The context is not a neutral layout of things or properties near the focal object, nor is it a set of circumstances or an indefinite background. It is ordered and organized by its relations to the focal object, which codetermines what properties of the surroundings are relevant and thus part of the context. [6]

We integrate this principle in our approach by proposing a Class-conditioned Context Attention Module (CCAM) such that our model can learn to attend to context elements that are relevant to the focal object in the training examples. Psychological studies showed that contextual cueing in humans improves object recognition in scenes, by capitalizing on the fact that most objects co-occur more often with certain objects and not others [3]. But additionally, the relative role of context specifies that not all co-occurrences are equal [6]. Fig. 1 shows an example of what is considered important by our model for learning the concept ‘headphones’. CCAM rightly focuses on concepts related to computers, while eliminating irrelevant co-occurrences such as ‘sticker’ (1.04%), ‘switch’ (0.33%), or ‘floor’ (0.06%).

In summary, our contributions are as follows. First, based on the supplementary role of context [6], we propose a Few-shot model that learns class representations grounded in contextual semantics. To this end, we propose a gated visuo-semantic unit, a flexible module to combine visual prototypes with semantic information. Unlike recent FSL work that improve visual features [5, 39, 15, 16], feature transfer [14, 26, 23], or the training procedure [7, 10, 8], our context-aware method exploits an orthogonal direction by leveraging complementary semantics from scene context.

Second, based on the relative role of context [6], we propose CCAM, a context module that automatically learns to attend to the most important elements in scenes relatively to the focal object.

Third, we conduct extensive experiments on Visual Genome [11], which is made of complex scenes images. Our experiments show that, compared to a visual-only baseline [25], our model improves the accuracy by 20.47% and 9.13% in 5-way 1-shot and 5-way 5-shot, respectively; and by 20.42% and 12.45% in 20-way 1-shot and 20-way 5-shot settings, respectively.

2. Related Work

**FSL approaches.** Several FSL work, including ours, build on Prototypical Networks [25]. This approach learns a metric space by computing class centroids from the examples in the support (train) set. It then compares query (test) image embeddings with these prototypes and assigns a class by performing nearest neighbor search. Other approaches consider different ways to compare support and query embeddings, such as Relation Networks [27] that automatically learn the distance function with a neural network, or the approach of Li et al. [15] that compares support and query images based on several descriptors.

Data augmentation and feature transfer methods seek to improve visual features [5, 39, 14]. For instance, Zhang et al. [39] use Generative Adversarial Networks to “hallucinate” new samples, thus virtually augmenting the training set. Other approaches leverage relations between classes to transfer visual features to new ones. Wang et al. [31] proposed to use a Graph Convolutional Network to transfer features between classes based on a knowledge base that encodes relations between these categories. Li et al. [14] developed an approach that learns from predicting class hierarchies, which facilitates feature transfer. A similar idea has been proposed to transfer explicit attributes between classes [1].

Another research avenue in learning from few examples explores improvements in the training procedure, which MAML [7] is a typical example. MAML is a meta-learning
algorithm that aims to generalize such that new tasks can be learned with few update steps. Several work build on MAML, e.g. \cite{10,8}.

Auxiliary semantics in FSL. Recently, additional cues that were only considered in ZSL have proved to be also useful in FSL, especially when the quantity of examples for each class is very low. Xing et al. \cite{36} built on Prototypical Networks \cite{25} by adding word embeddings in the formation of class prototypes in their approach called AM3. This improves the accuracy in 1-shot by almost 10\% on miniImagenet \cite{30} and by 5\% on CUB-200 \cite{33}. Schwartz et al. \cite{24} built on AM3 by additionally using text descriptions of classes extracted from WordNet, and thereby improved 1-shot accuracy by an additional 2\% on miniImagenet. These authors also proposed a general framework for learning prototypes from multiple semantics.

Context semantics. All the FSL work cited above focus on visual information. Attributes, word embeddings and text descriptions need to encode features that can be detected visually from the appearance of a new concept. Semantic-based approaches that leverage relations between classes with knowledge bases are also restricted to the transfer of visual features. However, this is not the only form of semantic information that can be available in images. The presence of other objects in a scene can also inform which classes are more or less likely to appear \cite{21,3,37}. The context has been used recently to improve object detection within deep learning models \cite{17,35,4,20}. For example, Liu et al. \cite{17} have developed an approach that uses scene context and object-object relationships to infer the presence of undetected objects. In a similar vein, Woo et al. \cite{35} established a new state-of-the-art in scene graph prediction by modeling global context and spatial relations between objects.

Recently Zablocki et al. \cite{37} introduced the use of scene context in ZSL. They showed that their model, with the use of Word2vec \cite{19} embeddings, could learn to rank unseen classes according to their likelihood of appearing in an image given the presence of other objects. This suggests that word embeddings implicitly encode co-occurrences of other classes in real visual scenes, even if they have been trained on text corpora \cite{19}. This is closely related to the distributional hypothesis \cite{9}, which states that words that appear in similar contexts often have similar meanings. This is exploited by skip-gram and continuous bag-of-words (CBOW) models such as Word2vec \cite{19}, and the results from Zablocki et al. \cite{37} suggest that it also generalizes to visual scenes: items denoted by words that have similar meanings tend to occur in similar scenes. This idea has been explored by Lddecke et al. \cite{18}, where they proposed a method to learn semantic word embeddings explicitly from images showing objects in context.

To the best of our knowledge, scene context has not been used in previous FSL work. In this paper, we argue that FSL would benefit from considering objects with their context, because it places new concepts in relation with others by learning from their co-occurrences in real scenes \cite{18}. We hypothesize that in this way object classification should increase in robustness and accuracy, because even if few examples cannot cover all intra-class variations, contextual cueing offers a complementary signal that can help disambiguate instances with unseen features.

We introduce this idea of using scene context in FSL to perform object recognition from few examples in complex images by building on Prototypical Networks \cite{25} and we propose to learn class prototypes grounded in context. Unlike Zablocki et al. \cite{37}, our model jointly learns to embed visual information with context semantics and class word embeddings. Whereas some recent semantic-based approaches use relations between classes to share common visual features, our use of scene context with class word embeddings exploit a different form of semantic relation which is complementary and orthogonal to visual features.

3. Our model

3.1. Preliminaries

FSL aims to solve the task of $M$-way $K$-shot classification, where $M$ is the number of classes in a given task and $K$ is a small number of examples for each class. Generally, Few-shot models are trained on a large dataset $D_{train}$, with a set of classes $C_{train}$ that is disjoint from the categories $C_{test}$ in $D_{test}$. The goal is to learn a representation model $f_{0}$ on $D_{train}$ such that it can learn to recognize new categories from $C_{test}$ only with $K$ examples. This is generally done by simulating the episodic test scenario of $M$-way $K$-shot classification during training. That is, even if a large number of examples are available for each class at train time, $f_{0}$ is trained by sampling at each episode $e$ 1) a support set $S_{e} = \{(x_{i}, y_{i})\}_{i=1}^{M \times K}$ that contains $K$ examples for each $M$ class and 2) a query set $Q_{e} = \{(q_{j}, y_{j})\}_{j=1}^{n_{q}}$ containing $n_{q}$ images from the same set of classes sampled in $C_{train}$. The model is then trained according to the cross-entropy loss:

$$\mathcal{L}(\theta) = -\frac{1}{n_{q}} \sum_{t=1}^{n_{q}} \log p_{0}(y_{t}|q_{t}, S_{e}) \quad \text{(1)}$$

Prototypical networks \cite{25} offer a simple and efficient way to model $p_{0}(y|q, S_{e})$. Each of the images in the support set are embedded by a CNN denoted by $f_{0} : \mathbb{R}^{D} \rightarrow \mathbb{R}^{d_{e}}$. Then a prototype is built for each class by averaging the $K$ vector embeddings from the same class:

$$c_{k} = \frac{1}{|S_{k}|} \sum_{(x_{i}, y_{i}) \in S_{k}} f_{0}(x_{i}) \quad \text{(2)}$$
Finally, the class distribution of a query image \( q \) is assigned by computing the softmax over the euclidean distances \( d \) of its embedding and all class prototypes:

\[
p_\theta(y = k | q, S_e) = \frac{\exp(-d(f_\theta(q), c_k))}{\sum_{k'} \exp(-d(f_\theta(q), c_{k'}))}
\]  

(3)

### 3.2. Context-Aware prototypes learning

Following previous work on the supplementary role of context in learning [6] and recognition [3, 21], we propose to learn class prototypes that embed knowledge about their context. To achieve that, we augment the support and query sets with scene context \( S_e \), where

\[
S_e = \{(x_i, S_i, y_i)\}_{i=1}^{M} \times \{ (q_j, y_j)\}_{j=1}^{n_q}.
\]

We adapt the formulation of each prototype \( c_k \) as:

\[
\hat{c}_k = \frac{1}{|S_k|} \sum_{(x_i, y_i) \in S_k} \phi(f(x_i), g(c_i)),
\]  

(4)

where \( c_i \) is the context representation of object \( i \) obtained from \( S_i \) (see section 3.2.1 for more details), \( g(c_i) \) is a small neural network projecting \( c_i \) in the same space than the image embedding, and \( \phi(\cdot, \cdot) \) is a function that adapts the image embedding according to the scene context (see section 3.2.2). An overview of our approach is shown in Figure 2. We now describe these components in more details.

#### 3.2.1 Context representation

**Scene context.** We model the scene context by using the class name annotations of the surrounding objects. This is done by leveraging word embeddings learned from a semantic model such as Word2vec [19]. Therefore, the scene context of an object is represented by the matrix \( S \in \mathbb{R}^{d_w \times n_s} \), where \( d_w \) is the word embeddings dimension and \( n_s \) is the number of surrounding objects.

**Class-conditioned Context Attention.** The relative role of context [6] suggests that some elements are more important than others when understanding a particular object. For instance, a switch might be important with respect to a lamp, but irrelevant while learning the concept of headphones (see Fig. 1). Therefore, we propose a Class-conditioned Context Attention Module (CCAM) that enables our model to weight the importance of each context elements in \( S \) while learning a particular concept \( w \) (see CCAM in Fig. 2). This is done by computing a scaled dot-product attention score \( A [29] \) between the class word embedding \( w \) and each element in \( S \) after linear transformations:

\[
K = W_K S \\
Q = W_Q w \\
A = \text{softmax} \left( \frac{K^T Q}{\sqrt{d_c}} \right)
\]  

(5)
where $W_K, W_Q \in \mathbb{R}^{d_x \times d_w}$ are projection matrices, and $\frac{1}{\sqrt{d_v}}$ is a scaling factor proposed in [29] to obtain smoother scores. $A$ reflects the relative role of each object in $S$, which is used to weight the contribution of context entities with respect to the focal object.

Context averaging $C_{avg}$. Note that the attention mechanism in CCAM is exclusively applied on contexts from the support set since it depends on the class category $w$. For query instances, the context representation $c_q$ is obtained by averaging all class embeddings $w_q \in S_q$:

$$c_q = \frac{1}{n_w} \sum_{w_q \in S_q} w_q$$   \hspace{1cm} (6)

3.2.2 Gated visuo-semantic unit

To combine visual embeddings with context semantic according to the supplementary role of context [6], we first experimented the convex combination as proposed by Xing et al. [36] to adapt visual prototypes with class word embeddings. We found however that this was too restrictive to consider the complex relations between visual and contextual information as the weighting factor always converged to zero during training, thus ignoring image embeddings. To solve this issue, we propose a gated visuo-semantic unit, a more expressive module to adaptively combine each feature individually from both representations based on a gating mechanism:

$$\phi(f(x), g(c)) := z \cdot f(x) + (1 - z) \cdot g(c)$$

$$h_v = \tanh(W_v \cdot f(x))$$

$$h_c = \tanh(W_c \cdot g(c))$$

$$z = \sigma(W_z \cdot [h_v, h_c])$$   \hspace{1cm} (7)

where $W_v \in \mathbb{R}^{d_x \times d_z}, W_c \in \mathbb{R}^{d_x \times d_w}, W_z \in \mathbb{R}^{d_x \times 2d_z}$ are projection matrices, and $\sigma$ is the sigmoid function.

Our fusion mechanism differs from the convex combination proposed by Xing et al. [36] in two ways. First, it computes the weighting factors $z$ based on both inputs, whereas in [36] it is only function of the class semantics. Second, we perform an element-wise convex combination, which enables the resulting point to lie anywhere inside the minimal surrounding box of $f(x)$ and $g(c)$, that is the hyper-rectangle in which $f(x)$ and $g(c)$ are the two most opposite corners (represented by a dashed rectangle in 2D in Fig. 2). This lets more degrees of freedom than the usual convex combination that places the resulting point on the segment between $f(x)$ and $g(c)$.

Note that our module is in the same vein than the Gated Multimodal Unit (GMU) [2], with a slight modification. In the original GMU, the output representation is $h = z \cdot h_v + (1 - z) \cdot h_c$, whereas in our formulation the intermediate representations $h_v$ and $h_c$ are used to compute the weighting factors to apply on each dimension of $f(x)$ and $g(c)$.

To sum up, in our approach each image embedding is combined with contextual information, and a context-aware class prototype $\hat{c}$ is obtained by averaging representations from the same class as described in Eq. 4. However, while these context-aware prototypes encode visual and context information, at this point they still ignore the class semantic of the focal object. In the next section, we show how we add the class word embedding to further refine each prototype.

3.3. Multi-semantics prototypes

We adopt a similar mechanism that Xing et al. [36] proposed to combine visual prototypes with their class word embeddings, since it proved to be particularly useful in settings with less data. Our context-aware prototypes are thus refined by:

$$\hat{c}_k' = \lambda \cdot \hat{c}_k + (1 - \lambda) \cdot \hat{w}_k$$   \hspace{1cm} (8)

where $\hat{w}_k$ is a transformation of the word embedding $w_k$ and $\lambda$ is a coefficient between 0 and 1. Both $\hat{w}_k$ and $\lambda$ are obtained with a two-layer neural network that uses $w_k$ as input.

Finally, the class distribution of a query image $q$ is computed as:

$$p(y = k|q, S_q, S_c, w) \propto \exp(-d[\phi(f(q), g(c_q)), \hat{c}_k'])$$   \hspace{1cm} (9)

4. Experiments

4.1. Dataset and settings

Visual Genome [11]. Traditional FSL datasets such as miniImageNet [30] and CUB-200 [33] mainly contain images with only one object and little context except background. Therefore, we rather experiment on Visual Genome, which is a large dataset of scenes with several objects in each image.

We randomly split the images in 70%/10%/20% train, validation and test sets, respectively. We start by using the public splits by Zablocki et al. [37] that keep 50% of classes for train and 50% for test. However, a closer look at those sets showed that some test classes are very similar to train classes, which could bias generalization evaluations. For instance, “bottle” and “television” are in the train set, but “bottles” and “TV” are test classes. To solve this issue, we filter the test classes whose Word2vec [19] embeddings have a cosine similarity higher than 0.75 with any of the
train classes. It effectively removes singular/plural nouns and closely related concepts such as “police” (train class) and “policeman” (test class). This will prevent our model from picking on those biases that would overestimate FSL performance.

We use the bounding box annotations to crop image parts that correspond to objects, and we remove examples whose smallest side is less than 25 pixels. Following this, we remove the classes that appear in less than 10 images. This finally results in 1211 train classes $C_{\text{train}}$ and 829 test classes $C_{\text{test}}$.

### 4.2. Implementation details

We employ a ResNet-12 CNN backbone as described in [22]. It is made of 4 blocks with 3 layers of 3x3 convolutions and a 2x2 maxpooling operation at the end of each block. The first block has 64 filters in each layer, and this number is doubled after each block. The output image embedding is a 512-dimension vector. Image crops from Visual Genome bounding box annotations are rescaled to $84 \times 84 \times 3$.

Each model is trained for 30,000 episodes with Adam optimizer initialized with a learning rate of $10^{-3}$ and is divided by a factor of 10 every 10,000 episodes.

### 5. Results

Since we want to study the contribution of using context information and class semantics in addition to the appearance of objects, we compare the following versions of models that all build on Prototypical Networks (ProtoNet) backbone [25]. **ProtoNet** is our reimplementation of Prototypical Networks [25]. **AM3-Proto** is our reimplementation of AM3 [36]. **Proto-$C_{\text{avg}}$** learns context-aware prototypes by averaging context elements from the support set. **Proto-CCAM** learns context-aware prototypes by applying the relative role of context [6] with our CCAM. **Proto-$C_{\text{avg}}$-W2V** adds class word embeddings to context-aware embeddings, but processes the support contexts by averaging the elements.

Table 1 shows the main results on Visual Genome dataset (Top-1 accuracy). Results are averaged over 4000 test episodes. The models are explained in Section 5. $C_S$ and $C_T$ refer to whether the context elements are picked from $C_{\text{train}}$ or $C_{\text{test}}$, respectively. Bold and underline show the best and second best result, respectively.

| Model                  | 5-way 1-shot | 5-way 5-shot | 20-way 1-shot | 20-way 5-shot |
|------------------------|-------------|-------------|---------------|---------------|
| ProtoNet [25]          | 52.23 ± 0.76% | 69.37 ± 0.63% | 25.71 ± 0.29% | 42.61 ± 0.70% |
| AM3-Proto [36]         | 62.50 ± 0.66% | 72.07 ± 0.74% | 34.36 ± 0.32% | 44.84 ± 0.61% |
| Proto-$C_{\text{avg}}$ ($C_S$) | 61.20 ± 0.65% | 76.66 ± 0.52% | 35.61 ± 0.46% | 53.42 ± 0.52% |
| Proto-$C_{\text{avg}}$ ($C_T$) | 57.49 ± 0.62% | 74.23 ± 0.58% | 36.58 ± 0.47% | 48.15 ± 0.52% |
| Proto-$C_{\text{avg}}$ ($C_S \cup C_T$) | 62.89 ± 0.64% | 77.74 ± 0.56% | 36.23 ± 0.50% | 53.19 ± 0.53% |
| Proto-CCAM ($C_S$)     | 63.56 ± 0.68% | 77.16 ± 0.55% | 35.60 ± 0.41% | 54.05 ± 0.54% |
| Proto-CCAM ($C_T$)     | 61.48 ± 0.72% | 76.82 ± 0.56% | 33.10 ± 0.44% | 50.66 ± 0.56% |
| Proto-CCAM ($C_S \cup C_T$) | 63.71 ± 0.71% | 77.98 ± 0.55% | 35.84 ± 0.42% | 54.37 ± 0.56% |
| Proto-$C_{\text{avg}}$-W2V ($C_S$) | 69.10 ± 0.59% | 76.63 ± 0.58% | 42.58 ± 0.44% | 53.40 ± 0.53% |
| Proto-$C_{\text{avg}}$-W2V ($C_T$) | 64.54 ± 0.58% | 75.86 ± 0.56% | 37.29 ± 0.44% | 50.03 ± 0.50% |
| Proto-$C_{\text{avg}}$-W2V ($C_S \cup C_T$) | 68.87 ± 0.59% | 77.76 ± 0.59% | 43.21 ± 0.41% | 54.80 ± 0.52% |
| Ours ($C_S$)           | 71.54 ± 0.57% | 78.50 ± 0.55% | 46.13 ± 0.47% | 54.72 ± 0.49% |
| Ours ($C_T$)           | 68.48 ± 0.61% | 76.20 ± 0.58% | 41.57 ± 0.50% | 51.23 ± 0.49% |
| Ours ($C_S \cup C_T$)  | 72.70 ± 0.51% | 77.34 ± 0.56% | 46.01 ± 0.48% | 55.06 ± 0.51% |

Table 2: Top-5 accuracy for 50-way and 100-way classification (%).

| Model                  | 50-way 1-shot | 50-way 5-shot | 100-way 1-shot | 100-way 5-shot |
|------------------------|-------------|-------------|---------------|---------------|
| ProtoNet [25]          | 39.68 ± 0.48% | 59.52 ± 0.50% | 27.98 ± 0.47% | 46.20 ± 0.52% |
| AM3-Proto [36]         | 51.78 ± 0.44% | 64.14 ± 0.44% | 38.22 ± 0.44% | 51.19 ± 0.44% |
| Ours ($C_S$)           | 64.10 ± 0.47% | 72.86 ± 0.50% | 50.49 ± 0.50% | 61.19 ± 0.49% |
| Ours ($C_T$)           | 68.47 ± 0.44% | 76.20 ± 0.50% | 51.23 ± 0.49% | 56.61 ± 0.49% |
| Ours ($C_S \cup C_T$)  | 66.39 ± 0.48% | 73.25 ± 0.50% | 52.20 ± 0.50% | 60.25 ± 0.50% |

**Supplementary role of context [6].** We can observe that the use of context information outperforms the visual-only ProtoNet [25] by a large margin, which confirms the supplementary role of context. Even Proto-$C_{\text{avg}}$ in 1-shot, which represents the most basic use of context, is better than using only the appearance of objects. Additionally, the use of class semantics in Proto-$C_{\text{avg}}$-W2V, AM3 [36], and our model, confirms its importance in 1-shot classification. Indeed, compared to ProtoNet [25] that only uses visual information, AM3-Proto [36] increases the accuracy by 10.27%.
and 8.65% in 5-way 1-shot and 20-way 1-shot, respectively, and the use of context with CCAM in our best model gives an additional boost of 10.20% and 11.77%.

Our model also performs reasonably well on larger-scale experiments shown in Table 2. With only one example per class in 100-way classification, our model almost doubles the top-5 accuracy of ProtoNet [25], with 52.20% compared to 27.98%.

Relative role of context [6]. The relative role of context states that the importance of context elements is function of the focal object, which is supported by our results. We can see from Table 1 that Proto-CCAM is better than Proto-C_{avg} in most scenarios, especially when context elements are chosen from C_{test}. The evaluations made with context elements picked only from C_{test} is a more challenging setting since the context module needs to assign the relative importance of classes that have never been encountered during training. Interestingly, the fact that Proto-CCAM obtains better results than Proto-C_{avg} shows that CCAM is still able to attend to relevant contextual elements.

Figure 3 shows examples of CCAM outputs when our model needed to learn the concepts “carrot”, “fish”, “paw” and “spectator”, respectively. Our model correctly gives more weight to semantically relevant co-occurring concepts and ignores background elements such as “tiles” in Fig. 3c and “grass” in Fig. 3d.

To further study the contribution of context and CCAM, we show in Figure 5 a t-SNE visualization [28] of embeddings produced by our model using different amount of information. Visual embeddings (Fig. 5a) seem to produce ambiguous clusters, similar to context averaging (Fig. 5b). On the opposite, our CCAM produces surprisingly good clusters (Fig. 5c), which shows its ability, and the importance, to attend to discriminative elements in the context.

Some visually different objects seem to share similar contexts, as shown by the mixed cluster in the center of Fig. 5c. This is mostly solved by our gated visuo-semantic unit that combines visual and contextual information (Fig. 5d).

Robustness to small objects and noise. Real scenes datasets contain additional challenges, one of which is to recognize ambiguous low-resolution objects. We experimented our model on subsets of test images based on bounding box sizes. Figure 4 shows the results for our model and ProtoNet [25] on 20-way 5-shot classification. Small objects (area in the interval of [0, 25^2] pixels) are very problematic for ProtoNet as it obtains 25.08% in accuracy, whereas our model still obtains 42.02%. These results are consistent with the fact that context semantics offer a complementary cue to recognize objects.

Moreover, we evaluated the robustness of our model while adding noise in support and query context labels.
With a probability $p_{\text{noise}}$, we randomly swap context elements by another label from $C_{\text{train}}$. The results for 20-way 5-shot classification with noise are shown in Figure 6. Interestingly, our model seems reasonably tolerant to false objects. With $p_{\text{noise}} = 0.5$, our model still outperforms AM3 [36] and ProtoNet [8].

Learning semantic word embeddings through visual scenes. Our approach also offers an auxiliary result that could be further investigated: CCAM implicitly learns and enriches semantic word embeddings by acting similarly to a CBOW model. To evaluate this aspect, we inputted to CCAM a matrix $S$ that contains all classes in $C_{\text{train}}$ and we conditioned its attention on a few words $w$ to see how CCAM would weight $S$. We show in Table 3 a few examples. Interestingly, the contextual concepts defined by CCAM strongly differ from those obtained with Word2vec [19] embeddings, which shows that CCAM captures different semantic relations between concepts.

6. Conclusion and Future work

In this work, we proposed a Few-shot learning model that uses scene context semantics to improve learning and recognition. Our approach integrates the supplementary role of context [6] by building context-aware class prototypes, and we apply the relative role of context [6] with our CCAM, a module that proved to be able to focus on discriminative elements in the scene with respect to the class semantics (see Fig. 3, Fig. 5c and Table 3).

Our experiments on Visual Genome [11] showed promising results of using context information, by increasing the accuracy of Prototypical Networks [25] by a large margin (e.g. by 20.42% in 20-way 1-shot and by 13.02% in 100-way 5-shot top-5 accuracy, see Table 1, 2).

More generally, our multi-semantics model is a step towards holistic approaches of few-shot object recognition that can be applied in challenging real scenarios.

As future work, we plan to reduce the amount of supervision by replacing ground-truth class annotations of context by automatic object detection and classification. Our experiment on robustness to noise suggests that even with many labeling errors, our method could still outperform models that ignore the context.

We also plan to explore the ability of our model to learn semantic word embeddings through visual scenes. This could be further investigated in line with work on learning multimodal word embeddings [13, 38, 18].
References

[1] Ziad Al-Halah, Makarand Tapaswi, and Rainer Stiefelhagen. Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5975–5984, 2016. 1, 2

[2] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. arXiv preprint arXiv:1702.01992, 2017. 5

[3] Moshe Bar. Visual objects in context. Nature Reviews Neuroscience, 5(8):617, 2004. 1, 2, 3, 4

[4] Xinlei Chen and Abhinav Gupta. Spatial memory for context reasoning in object detection. In Proceedings of the IEEE International Conference on Computer Vision, pages 4086–4096, 2017. 3

[5] Zitian Chen, Yanwei Fu, Yu-Xiong Wang, Lin Ma, Wei Liu, and Martial Hebert. Image deformation meta-networks for one-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8680–8689, 2019. 1, 2

[6] Nina Dohn, Stig Hansen, and Søren Klausen. On the concept of context. Education Sciences, 8(3):111, 2018. 2, 4, 5, 6, 7, 8

[7] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 1126–1135. JMLR. org, 2017. 2

[8] Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. In Advances in Neural Information Processing Systems, pages 9516–9527, 2018. 2, 3, 4, 8

[9] Zellig S Harris. Distributional structure. Word, 10(2-3):146–162, 1954. 3

[10] Muhammad Abdullah Jamal and Gao-Jun Qi. Task agnostic meta-learning for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 11719–11727, 2019. 2, 3

[11] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, and Andriy Bursuc. Dense classification and implanting for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 9258–9267, 2019. 2

[12] Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. Combining language and vision with a multimodal skipgram model. arXiv preprint arXiv:1501.02598, 2015. 8

[13] Wenbin Li, Lei Wang, Jinglin Xu, Jing Huo, Yang Gao, and Jiebo Luo. Revisiting local descriptor based image-to-class measure for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7260–7268, 2019. 2

[14] Yann Lecun, Yannis Avrithis, Sylvaine Picard, and Androi Bursuc. Dense classification and implanting for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 9258–9267, 2019. 2

[15] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Structure inference net: Object detection using scene-level context and instance-level relationships. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6985–6994, 2018. 3

[16] Timo Lüddeke, Alejandro Agostini, Michael Fauth, Minija Tomasianaitait, and Florentin Wörgötter. Distributional semantics of objects in visual scenes in comparison to text. Artificial Intelligence, 274:44–65, 2019. 3, 8

[17] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119, 2013. 3, 4, 5, 8

[18] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 891–898, 2014. 3

[19] Aude Oliva and Antonio Torralba. The role of context in object recognition. Trends in cognitive sciences, 11(12):520–527, 2007. 2, 3, 4

[20] Boris Oreshkin, Pau Rodriguez 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. 6

[21] Zhimao Peng, Zechao Li, Junge Zhang, Yan Li, Guo-Jun Qi, and Jinhu Tang. Few-shot image recognition with knowledge transfer. In Proceedings of the IEEE International Conference on Computer Vision, pages 441–449, 2019. 2

[22] Eli Schwartz, Leonid Karlinsky, Rogerio Feris, Raja Giryes, and Alex M Bronstein. Baby steps towards few-shot learning with multiple semantics. arXiv preprint arXiv:1906.01905, 2019. 3

[23] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In Advances in Neural Information Processing Systems, pages 4077–4087, 2017. 1, 2, 3, 6, 7, 8

[24] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 403–412, 2019. 2

[25] 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 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1199–1208, 2018. 1, 2
[38] Hongguang Zhang, Jing Zhang, and Piotr Koniusz. Few-shot learning via saliency-guided hallucination of samples. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2770–2779, 2019. 1, 2