applied sciences

MDPI

Article

Zero-Shot Recognition Enhancement by Distance-Weighted Contextual Inference

Doo Soo Chang, Gun Hee Cho and Yong Suk Choi *

Artificial Intelligence Laboratory, Hanyang University, Seoul 04763, Korea; tim0225@hanyang.ac.kr (D.S.C.); gunneng@hanyang.ac.kr (G.H.C.)

* Correspondence: cys@hanyang.ac.kr

Received: 22 July 2020; Accepted: 14 October 2020; Published: 16 October 2020

Abstract: Zero-shot recognition (ZSR) aims to perform visual classification by category in the absence of training samples. The focus in most traditional ZSR models is using semantic knowledge about familiar categories to represent unfamiliar categories with only the visual appearance of an unseen object. In this research, we consider not only visual information but context to enhance the classifier's cognitive ability in a multi-object scene. We propose a novel method, *contextual inference*, that uses external resources such as knowledge graphs and semantic embedding spaces to obtain similarity measures between an unseen object and its surrounding objects. Using the intuition that close contexts involve more related associations than distant ones, distance weighting is applied to each piece of surrounding information with a newly defined distance calculation formula. We integrated contextual inference into traditional ZSR models to calibrate their visual predictions, and performed extensive experiments on two different datasets for comparative evaluations. The experimental results demonstrate the effectiveness of our method through significant enhancements in performance.

Keywords: zero-shot recognition; similarity measures; distance-weighting; knowledge graph; semantic embedding

1. Introduction

Demands to expand the scale of categories available for object recognition have been aroused by a rapid increase in the sizes and types of image data and the recent success of large-scale recognition systems [1]. However, manually constructing additional annotations and retraining existing image classifiers through supervised learning is impractical and costly, which limits the scalability of existing systems. To alleviate that limitation, a variant of transfer learning, zero-shot learning (ZSL) has drawn the attention of the computer vision community [2–5].

ZSL is inspired by the human capacity to recognize objects without any visual samples using background knowledge already read or heard. The key is to transfer semantic knowledge about familiar (seen) objects to imagine unfamiliar (unseen) objects. For instance, a human can easily recognize an unseen zebra based on visual experiences with a horse and a watermelon, if it is known that a zebra looks like a horse with stripes on its body.In the same way, the objective of ZSL methods is to increase the cognitive capability of a visual classifier by using annotated training sets of seen class labels and external knowledge about the semantic relations between seen and unseen categories to allow the classifier to infer the class labels of novel objects. In this context, external knowledge is generally represented as non-visually using attributes [6,7], semantic embedding [8], and knowledge graphs [9,10]. To transfer knowledge, zero-shot recognition (ZSR) assumes that visually similar objects tend to also be semantically similar, which implies that the vector representations of their class labels are close. Most existing ZSL methods thus focus on learning to recognize inherent visual features (e.g., color, shape, and texture) and providing a map between the visual and semantic representations.

However, the correlation between visual and semantic information is not always assured. For example, the unseen object, monitor, may be confused with a frame or a window because it has a square, edged shape with inside contents. Thus, existing ZSR models have critical limitations because they rely on visual information about a single unseen object. In contrast, when people try to identify an unseen object, they naturally refer to the circumstances surrounding the object, such as other objects and their relative positions, as well as the visual characteristics of the target object. In addition, it is common to infer more from relatively close objects than from more distant objects (e.g., the toothpaste beside a toothbrush and not the mirror in the scene). When modeling a zero-shot classifier, the correct label monitor could be inferred as appropriate when the surroundings suggest an office environment through the appearance of seen objects such as a keyboard and a notebook. That is, the classifier should use surrounding information to determine the type of object appropriate in a given environment.

In this paper, we use those intuitions to propose a novel ZSR method that leverages context based on similarity measurements and distance-weighting between a target unseen object and surrounding objects. We aim to enhance the performance of existing instance-level ZSR that relies only on individual visual information about the target. To identify the context, our method uses cognitive information about the surrounding objects and obtains their similarity information for the target object using three different measures on a knowledge graph, a semantic embedding space, and both together. Moreover, distance weighting is applied to each piece of similarity information to focus on nearby surrounding objects by defining a distance calculation formula. To evaluate the effectiveness of our method, we adopted several existing ZSL models as baselines and performed extensive experimental evaluations on two different datasets containing small and large amounts of target categories.

The main contributions of this work are as follows. (1) We propose an advanced ZSR method that references to the similarity-based contextual information in a multi-object scene to alleviate the dependence of traditional methods on visual information about an unseen object. (2) With the intuition that nearby objects have more reliable relationships with the target object than distant objects, we newly formulate a distance calculation between the objects' bounding boxes to enable distance-weighted contextual inference. (3) Our method maintains modularity and can be integrated with any instance-level zero-shot classifier because it does not require an additional training process for the contextual inference. (4) Extensive experiments on two datasets with different target category scales show that our system offers performance enhancements compared with existing instance-level and context-aware ZSL models, often by a large margin.

This paper is structured as follows. Section 2 outlines related works and sets baselines for the evaluation. We detail our method for contextual inference and distance-weighting in Section 3. Section 4 presents our experiments and results, and Section 5 gives our conclusions.

2. Related Work

2.1. Instance-Level ZSL

Existing ZSL models differ in their use of semantic embedding spaces or other external knowledge sources. Early models used manually constructed attribute spaces [11–17] to represent categories as binary vectors that implied the presence of attributes. In general, the use of attributes has seemed promising [18–21] in various research fields, including ZSL, but it has limited scalability due to domain dependency and the cost of manual construction. To relax those limitations, semantic word-vector spaces that are automatically trained on a textual dataset have been used used in more recent ZSL works [6,22–25]. The word-vector spaces from text corpora, such as word2vec [26] and GloVe [27], motivated the use of large-scale ZSR with many unseen categories because they are unrestricted and less costly than manually annotated attributes. Some ZSL works has used knowledge graphs instead [9,28–30]. In particular, some recent works [9,30] based on a graph convolutional network

(GCN) [31] have used the WordNet [32] taxonomy to propagate classifier weights from seen to unseen categories.

Most existing ZSL works make instance-level inferences about an individual unseen object by transferring external knowledge from seen categories based on visual similarities, and they have shown promising prediction results in test sets with a specific type of categories such as Caltech-UCSD Birds (CUB) [33], Stanford Dogs (Dogs) [34], and Animals With Attributes (AWA) [35]. However, despite many attempts of ZSR with large-scale datasets such as ImageNet [36], which contain various types of categories for generic objects, instance-level ZSL works have produced comparatively poor performance, which implies that different types of categories have irrelevant semantic characteristics despite their visual similarities. Therefore, we consider not only visual prediction about an unseen object but also its surrounding information to determine the most likely category.

2.2. Contextual Recognition

Many works [37–44] have emphasized the importance of context and tried to enhance recognition or detection using context as an additional resource. However, they are inappropriate for ZSL, in which supervised learning cannot be applied. Exploiting context in a ZSR task is still challenging and not well standardized. Only a few recent works have proposed that a ZSR task be aware of context. For instance, the authors of [45] leveraged visual context and the geometric relationships between multiple objects using a conditional random field. The authors of [46] presented a method based on conditional likelihoods that combined three independent models of contextual, visual, and prior information. Both of those works complement our research, but we consider visual candidates of surrounding objects as potential contexts based on similarity measurements and apply distance-weighting according to the positional differences of the objects. In other words, we propose a method that effectively exploits contextual information, and we validated our proposed model by comparing it with several instance-level ZSL models [6,9,22,23] and one context-aware model [45] through the experiments described in Section 4.

3. Method

3.1. Problem Definition

Let C be a set of class labels c that is split into two disjoint subsets, $S = \{s_m\}$ and $U = \{u_n\}$ that, respectively, denote a set of M seen class labels and a set of N unseen class labels. Under the setting of the ZSR task, the two subsets of labels satisfy $S \cap U = \oslash$ and $S \cup U = C$. Each class label has its representative semantic embedding vector e in a semantic embedding space $E \in \mathbb{R}^{d_e}$.

For training, a labeled dataset of images $I_s = \{(i_m, s_m)\}$ is given, in which each image is represented by a d_i-dimensional feature vector, $i_m \in \mathbb{R}^{d_i}$, and a class label, $s_m \in S$. A test dataset $I_u = \{(i_n, u_n)\}$ is provided for testing in which $i_n \in \mathbb{R}^{d_i}$ and $u_n \in U$. In general, the goal of ZSL is to learn a classifier f to produce the correct class label for an unseen individual image.

3.2. Model Overview

Traditional ZSR methods classify an unseen object using only its individual visual information. On the contrary, our proposed model aims to classify an unseen object with the help of contextual information obtained from its surroundings in a multi-object scene. We define potentially related surrounding objects under the following assumptions; (1) they represent non-unseen classes and are not the targets of our recognition task; (2) they are detected by a pre-trained object detector or classifier before the ZSR of the target object; and (3) each one includes predicted candidates and corresponding prediction probabilities. Consequently, the surrounding information (SI) of an unseen image feature i with multiple surrounding objects is specified by the following equations,

$$SI(i) = \{si_j\}_{j=1}^{N^{sobj}}, \quad (1)$$

$$si = (\{(c_k, p_k)\}_{k=1}^{N^{cand}}, d), \tag{2}$$

where si denotes the surrounding information for a single surrounding object and SI consists of multiple sis with the number, N^{sobj}. c_k and p_k are the class label of the k-th predicted candidate of a surrounding object and the prediction probability of the corresponding candidate c_k, respectively. In particular, $c_k \in S$, and p_k is a soft-max value of the prediction score among N^{cand} candidates.

Our proposed model takes an unseen image feature i and its surrounding information $SI(i)$ as its inputs and predicts the most likely class label c as its output. Specifically, let the aforementioned classifying function be $f : I \rightarrow E$ between a visual feature space $I \in \mathbb{R}^{d_i}$ and a semantic embedding space $E \in \mathbb{R}^{d_e}$ for class labels. The classifying function f outputs the semantic embedding vector e that maximizes the scoring function F as follows and our model finally produces a prediction for the corresponding class label c of the output semantic embedding vector.

$$f(i, SI(i)) = \arg\max_{e \in E} F(i, e, SI(i)). \tag{3}$$

For the process of ZSR, as shown in Figure 1, our model acts on two collaborating branches, the instance-level visual inference of an unseen object and the contextual inference with surrounding information. The visual inference basically follows the process of traditional ZSL models: the extraction of a feature vector from a target unseen image and the prediction of a visual score for each target class label using a zero-shot classifier. The contextual inference, a novel process, performs similarity measurements between each of target class labels and visual candidates of the surrounding objects to obtain a contextual score.Similarities are measured using a knowledge graph and/or a semantic embedding space. The prediction of the visual inference is calibrated and advanced by the result of the contextual inference. More detailed processes and formulations are described in Sections 3.3 and 3.4. Consequently, the scoring function F for a specific class label can be specified by combining the instance-level visual inference function G and the contextual inference function H as follows, where α is a balancing factor for an usage ratio between the visual and contextual scores:

$$F(i, e, SI(i)) = \alpha \cdot G(i, e) + (1 - \alpha) \cdot H(e, SI(i)). \tag{4}$$

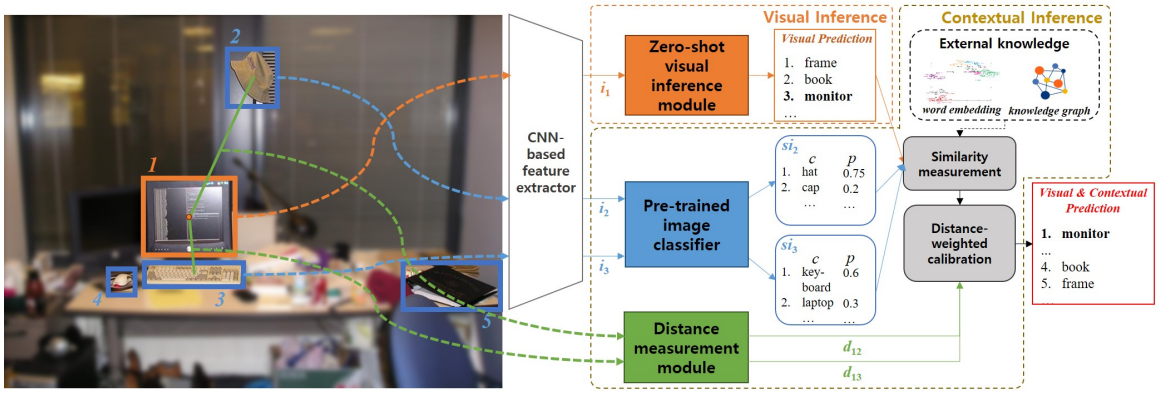

Figure 1. The architecture of our proposed model with an example in an office environment. The model performs Zero-Shot Recognition (ZSR) on two main branches: the instance-level visual inference and the contextual inference.The visual inference first infers the class label of the extracted feature vector i_1 of the target object, which is a process used by existing methods. A distance-weighted, similarity-based calibration is then performed between the target and its surrounding objects to refer to the contextual information, which is the novel process we propose.

As shown in the intuitive and assumptive example in Figure 1, an office environment contains several objects including the target unseen object (orange bounding box), a monitor, and its surrounding

objects (blue bounding boxes). The visual inference first predicts the orange-boxed image as a frame, which has the highest rank due to its square appearance and raised edge, with a monitor, which is the correct label, ranked third. Using only visual information about a target object can thus lead to failed prediction. Therefore, we exploit additional context information from the surrounding objects to calibrate the visual prediction result. Most of the surrounding objects are ranked as likely to be computer devices, such as a keyboard and a mouse, by the pre-trained image classifier for seen class labels. However, an unwanted object in the context, such as a hat, could disturb the contextual inference.Under the assumption that closer objects are more relevant than more distant objects, the contextual classifier gathers more information from the keyboard and mouse than from the hat. That proper use of surrounding information enables the classifier to correctly rank the target as a monitor that is "computer-related", square-shaped, and edged. We experimentally validate this overall intuition in Section 4.

3.3. Visual Inference

The visual inference process aims to classify an unseen object using only its individual visual feature as input. As a result of that process, the classifier predicts a visual score for each unseen class label as output. In general, classifiers use a model trained with fully supervised learning for seen class labels or an indirectly trained model based on semantic relations between seen and unseen class labels. Because the main concept of this paper is the additional use of context, we adopt several existing ZSR models [6,9,22,23] for the visual inference function $G : I \rightarrow E$ and use them as baselines in the comparison of models with and without the contextual inference in Section 4. The G functions used by the baseline methods to measure the visual scores are specified in Table 1.

Table 1. Visual inference functions of baseline zero-shot recognition (ZSR) models.

Model	Function $G(i, e)$
SJE	$i^T W e$
LatEm	$\max_{1 \le j \le K} i^T W_j e$
ConSE	$\cos\left(\frac{1}{Z}\sum_{c \in S} p(c\|i) \cdot e_c,\ e\right)$
GCN	$\hat{w}_e \cdot i$

Notations: SJE (Structured Joint Embeddings), LatEm (Latent Embedding), ConSE (Convex Combination of Semantic Embedding), GCN (Graph Convolutional Network), W (image-semantic embedding matrix), K (latent variable, $K \geq 2$), Z (normalization factor, $Z = \sum_{c \in S} p(c|i)$), $\hat{w}$ (predicted classifier weight).

Structured Joint Embeddings (SJE) [6] trains an image-semantic embedding matrix $W \in \mathbb{R}^{d_i \times d_e}$ and infers a class label of an unseen object by measuring the distance between an embedded image feature of the object and unseen semantic embeddings. Latent Embedding (LatEm) [22] tries to relax the limitation of linearity in SJE by using multiple image-semantic embedding matrices. LatEm proposes a nonlinear compatibility function with K indexes over the latent choices.

The convex combination of semantic embeddings (ConSE) [23] uses a pretrained image classifier trained on seen class labels with full supervision. For the inference, softmax-output values p of the classifier for an input unseen image are used to create its representation vector by weighting the semantic vectors of seen class labels e_c. An unseen class label for the semantic vector nearest to the representation vector is then predicted as an appropriate class label.

A multi-layer GCN model [9] begins by training an image classifier in the same manner, but it uses classifier weights from the image classifier as ground-truth to learn predicted classifier weights with semantic embeddings for class labels and their adjacency matrix as input. At test time, the visual inference for the GCN conducts a dot-product estimation between an image feature vector i and a predicted classifier weight $\hat{w}$ for unseen class labels.

3.4. Contextual Inference

When encountering an unseen object, humans unconsciously refer to not only its visual appearance but also the surrounding environment and the types and relationships among nearby objects to infer its identity. We propose a novel approach that derives a contextual score based on the associations between an unseen object and its surrounding objects, and calibrates the prediction results from the visual score of existing ZSR methods. In particular, the associations are obtained through similarity measures that use external knowledge sources. Furthermore, a distance calculation formula ensures that nearby surrounding objects are more important to the contextual score than distant objects.

3.4.1. Similarity-Based Association Measurements

To grasp the context of an unseen object, we determine associations through similarities between the class label of the target object and those of its surrounding objects. Note that we assume that the surrounding objects are recognized before beginning ZSR for the unseen object, as explained in Section 3.2. When measuring similarities for contextual inference, we consider not only the top-1 classified label of the surrounding objects, but also all potentially-ranked labels of their candidates. This helps to alleviate the problem of recognizing misclassified surrounding objects as the representative context and allows our system to consider multiple potential contexts. External knowledge sources—a knowledge graph, a semantic embedding space, and both together—are used for three different similarity measurements: semantic similarity (SM), cosine similarity (CS), and the harmony of both (HM), respectively.

- Semantic Similarity Measurement

The SM metric is defined over documents and is based on the likeness of conceptual meanings [47]. In this paper, we use a hierarchical knowledge graph, ontology, representing the hierarchical concepts of objects for the SM measurement. Various measures [48] can be used with a knowledge graph, such as path and depth measures, information content-based measures, feature measures, and hybrid measures. We adopt three typical path and depth-based measures for our experiments, as presented in Section 4.2.2. A SM-based association, S^{SM}, between one unseen class label and a single surrounding object, is given by

$$S^{SM}(e, si) = \max_{1 \le k \le N^{cand}} p_k \cdot sem(c_e, c_k), \tag{5}$$

where the inputs are a semantic embedding e of an unseen class label c_e and the surrounding information si for a single surrounding object. c_k denotes a class label of the k-th predicted candidate for the surrounding object. Each SM value between c_e and c_k is weighted by p_k, a prediction probability for the candidate that implies the reliability of the measured similarity. An association's output is the maximization of the weighted similarities for candidates because finding the best combination of an unseen class label and all candidates for all surrounding objects is equivalent to finding and combining the best candidate for each surrounding object.

- Cosine Similarity Measurement

To measure associations, we next use CS, which is available in semantic embedding spaces. Various types of semantic embedding spaces are used in ZSR, such as manually-annotated attributes [11,12], text descriptions of images [49], word embeddings [26,27], and rdf graph embeddings [50]. Among those, the attribute space represents fine-grained concepts of each class label with a binary value depicting the presence/absence of an attribute, based on human annotated description (e.g., for attributes horse-like, stripe, and green, $e_{zebra} = [1, 1, 0]$, $e_{tiger} = [0, 1, 0]$, and $e_{watermelon} = [0, 1, 1]$, respectively). The word embedding space contains vectors learned by neural net with a certain feature dimension according to the mutual frequency of words in the context in the

text corpus (e.g., due to simultaneous occurrence of words, "monkey" and "banana", the magnitude and direction of e_{monkey} and e_{banana} can be similar). In terms of ZSR performance, embedding spaces based on manual construction such as attributes and text descriptions of images are generally more effective than a word embedding space [6,12,51,52]. However, a word embedding space constructed in an unsupervised manner still has higher versatility and utility than the other options because it costs much less and enables large-scale recognition with many class labels [53]. We thus use word embedding spaces as the semantic embedding space E for the evaluation, although our proposed model works independently on the type of semantic space. A CS-based association, S^{CS}, is measured by calculating two vectors, as follows,

$$S^{CS}(e, si) = \max_{1 \le k \le N^{cand}} p_k \cdot cos(e, e_{c_k}). \tag{6}$$

The overall metric is similar to Equation (5) except for the use of external knowledge and the similarity measurement. e_{c_k} denotes a semantic embedding vector for c_k in the semantic embedding space E.

- Harmonic Similarity Measurement

As a harmonic approach, we combine the association results of the SM and CS, which implies the use of both a knowledge graph and a semantic embedding space when referring to the surrounding information. As a harmonic association, S^{HM} is specified simply with a balancing factor on the two measurements as follows,

$$S^{HM} = \beta \cdot S^{SM} + (1 - \beta) \cdot S^{CS}, \tag{7}$$

3.4.2. Distance-Weighted Calibration for Multiple Surrounding Objects

Associations for all surrounding objects are eventually synthesized to derive the contextual inference to calibrate the results of the visual inference. Using the intuition that objects near a target object offer better context than objects farther away, the context inference applies a distance-weighted average of associations rather than a simple average. We assume that the bounding boxes, $b = (x, y, w, h)$ of the objects needed to calculate distance are given, where (x, y) is the center point of a bounding box and w and h denotes its width and height, respectively.

However, we do not simply use the Euclidean distance between center points. In Figure 2, the Euclidean distances between the two objects in panels (a,b) are the same, but the actual relative distances in panel (b) are much smaller. By generalizing all the related cases, we define a distance calculation equation that relaxes the Euclidean distance with the size of two objects.

$$d = \begin{cases} \dfrac{\sqrt{\tilde{x}^2 + \tilde{y}^2}}{\dfrac{\tilde{x}}{\tilde{x} + \tilde{y}} \cdot \bar{w} + \dfrac{\tilde{y}}{\tilde{x} + \tilde{y}} \cdot \bar{h}}, & \text{if } \tilde{x} \neq 0 \text{ or } \tilde{y} \neq 0 \\ 0, & \text{else} \end{cases} \tag{8}$$

where $\tilde{x} = |x_1 - x_2|$ and $\bar{w} = \frac{w_1 + w_2}{2}$. A distance is calculated when center points are not exactly the same, and it is fixed to 0 otherwise. When calculating the distance, the application of the Euclidean distance is adjusted to the average width and height of two objects, where the width and height are referenced using the ratio of the x-axis and y-axis intervals between the objects.

As previously explained, because the context of close objects should be exploited more, the reciprocal of the distance is considered as a weight. By weighted-averaging all obtained associations,

the set of $S(e, si)$, and the corresponding weights are combined to derive the contextual score for an unseen class label as follows,

$$H(e, SI(i)) = \frac{\sum_{j}^{N^{sobj}} r_j \cdot S(e, si_j)}{\sum_{j}^{N^{sobj}} r_j}, \quad (9)$$

where $r = 1/(d + \epsilon)$ and ϵ is a smoothing factor whose value is fixed at 0.001 in the experiments.

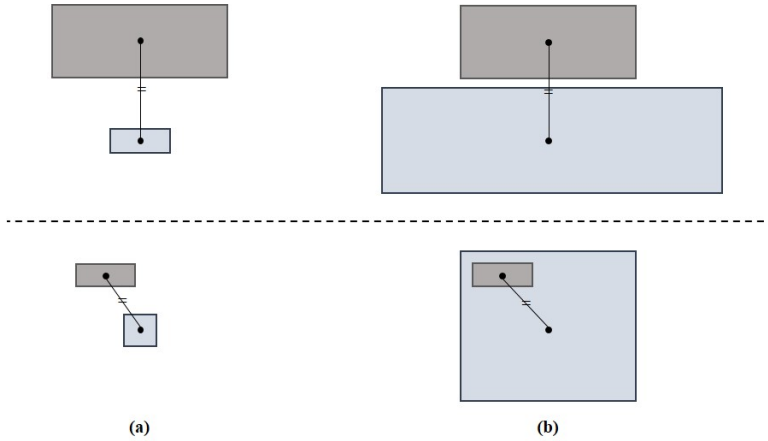

Figure 2. The Euclidean distances are the same in (**a**,**b**) for each case, but the actual distances seem remarkably different. The upper and lower case of (**b**) are likely to show an "on" and "attached" relation, respectively, which represents the definite nearness of the objects

4. Experiments

4.1. Overall Experimental Scenario

The main task in the experiments is to predict an appropriate class label for an unseen object from among target class labels by using its visual information and the contextual information of surrounding objects in a multi-object scene. Recall that we assume that the surrounding objects are detected and recognized by the existing object detector, which is pretrained only on non-unseen class labels and never on unseen ones. YOLOv2 [54] is mainly used as the object detector, the pretrained image classifier in Figure 1 and is trained for 80 categories on the COCO dataset [55]. Among those categories, the ones that share concepts with the unseen class labels are excluded for detection. The namespace of the rest is matched and anchored to that of a knowledge graph and semantic embedding space (presented in the following subsections). Moreover, confidence scores provided by YOLOv2 are exploited as the reliability values p in the aforementioned equations for those whose values are higher than the fixed threshold of 0.3.

We conduct two experiments with different dataset scales and types of class labels: (1) experiments on ImageNet categories with less unseen class labels, and (2) experiments on Visual Genome categories based on the split in [45,56], with relatively more unseen class labels and a larger scale test set.

4.2. Experiments on Imagenet Categories

4.2.1. Dataset

In the first experiment, we use the ImageNet dataset [36] and Visual Genome (VG) dataset [57] for training and testing, respectively. Each image in ImageNet contains an annotation for only one category, whereas an image in VG represents a multi-object situation and has multiple annotations.

We use ImageNet (ILSVRC) 2012 1K, which is composed of 1K class labels and more than 1.2 million images for training, but we consider only 944 class labels with available semantic embedding as seen classes, with 1,211,266 training images that are used to train some of aforementioned baseline models for the visual inference. The weight matrix in SJE and 6 weight matrices (i.e., $K = 6$) in LatEm are trained with 150 epochs and a learning rate of 0.001. For the GCN, we use the same training settings as [9], which are 6 convolutional layers with an output channel D of 1024. The classifier weights and image features are L2-normalized for the GCN and the ADAM [58] optimizer is used with 300 epochs, a learning rate of 0.001 and a weight decay of 0.0005. The adjacency matrix in the GCN is constructed based on the WordNet knowledge graph and considers sibling classes as adjacent classes as well. More details about the knowledge graph are provided in the next subsection.

To adopt 34 unseen class labels for testing, we define and observe four criteria: (1) they are selected among 360 categories in ImageNet 2010 1K that are disjoint from ImageNet 2012 1K, which is similar to the settings in [53], (2) they are categories for generic objects, (3) they have available semantic embeddings, and (4) they are also annotated in VG with more than 20 image instances. Additionally, we randomly sample a maximum of 200 image instances per class label, which indicates that some might have fewer than 200. Thus, we have 4720 test images in VG for 34 unseen class labels.

4.2.2. Visual/Semantic Embeddings and Knowledge Graphs

For the visual inference in this experiment, we use the entire model of Inception-V1 [57,58] on the ConSE baseline model and 1024-dimensional outputs from the top-layer pooling units of Inception-V1 on the other baselines to extract visual features of image instances, which is the same technique used in other recent ZSL researches [49,59]. We consider Inception-V1 to be reasonably appropriate for our experimental setting because it is pre-trained on the ImageNet 2012 1K dataset, so that all of our seen class labels are involved but our unseen class labels are not. Z-score normalization on the dimensions is applied to extracted image features, and the mean and standard deviation values of the training visual features are used on testing visual features for normalization.

Our approach uses a semantic embedding space to measure the CS between the class labels of a surrounding object and a target object. Recall that word vectors are considered to be a semantic embedding space E in this evaluation. In this experiment, we use a skip-gram model [26] trained on Wikipedia English in February 2015 with a window size of 10. The dimension of the word vectors is set to 1000, which is a column dimension of embedding matrices for baselines. All the word vectors are L2 unit-normed, and all class labels are anchored to their own semantic embedding. Specifically, each class label that consists of multiple words has a representative semantic embedding that is the averaged word vector for all the words.

For the SM measurement, a sub-graph of WordNet [32] is used in the proposed model. WordNet is a large-scale hierarchical database with more than 100K English words. The concepts in WordNet are represented as synset IDs, which is the same as in the ImageNet dataset and allows us to integrate the namespace of the class labels into the form of synset IDs. We adopt three SM metrics provided by the NLTK library in WordNet, *path*, *lch*, and *wup*, from the work in [48]. In particular, the distance relationships for *path* are scaled to the similarity measurement in NLTK to range the value from 0 to 1. We conduct a performance evaluation separately for each metric, as presented in the next subsection.

4.2.3. Results

The experimental evaluation is performed under two ZSL settings: classic and generalized. Only unseen class labels and both unseen and seen class labels are considered as target categories for prediction in the classic setting and the generalized setting, respectively. Note that the ground-truths of the test image instances represent only unseen class labels. The performance of each model is evaluated in terms of the average per-class accuracy, "*per-class*" and an overall accuracy for all instances, "*per-instance*". Note that the accuracy values are expressed in percentage in all our experiments.

We apply the proposed contextual calibrations to the visual inferences of four baseline models for validation, giving seven types of evaluation per baseline model—three SMs (*path*, *lch*, and *wup*), CS, and HMs—for each of the three semantic similarities.

- Comparative Evaluation to Baselines

Table 2 shows the results of our comparison with the baselines. All the prediction models using any type of the contextual inferences significantly outperform the baseline models in both settings. In most cases, the contextual inferences with harmonic measures are more effective than the others.

Table 2. Comparative results to the baseline models. Numerical values represent top-1 accuracy in percentage. SM, CS and HM imply the semantic, cosine and harmonic similarity measurements, respectively.

Method			Classic		Generalized	
			Per-Class	**Per-Instance**	**Per-Class**	**Per-Instance**
SJE (**baseline**)			11.64	13.2	1.7	2.03
SJE + contextual inference (**proposed**)	SM	path	15.86	16.93	3.2	4.43
		lch	15.54	16.97	3.3	4.51
		wup	15.39	16.67	3.31	4.56
	CS		17.1	17.52	1.7	2.06
	HM	path	18.56	18.77	3.2	4.43
		lch	**19.49**	**19.68**	3.3	4.51
		wup	19.25	19.2	**3.31**	**4.56**
LatEm (**baseline**)			11.99	13.22	0.78	0.83
LatEm + contextual inference (**proposed**)	SM	path	15.7	16.25	3.2	4.43
		lch	15.69	16.46	3.3	4.51
		wup	15.92	16.7	3.31	4.56
	CS		16.83	16.95	1.43	2.06
	HM	path	17.36	17.65	3.2	4.43
		lch	17.85	18.24	3.3	4.51
		wup	**18.27**	**18.37**	**3.31**	**4.56**
ConSE (**baseline**)			11.72	12.03	0.43	0.38
ConSE + contextual inference (**proposed**)	SM	path	15.15	14.85	3.2	4.43
		lch	15.26	15.51	3.3	4.51
		wup	15.07	15.11	3.31	4.56
	CS		14.49	14.34	1.43	2.06
	HM	path	16.09	15.59	3.2	4.43
		lch	**16.36**	**16.91**	3.3	4.51
		wup	16.35	16.29	**3.31**	**4.56**
GCN (**baseline**)			13.37	11.8	2.96	2.37
GCN + contextual inference (**proposed**)	SM	path	16.34	15.15	4.54	5.11
		lch	16.39	15.13	4.34	5.4
		wup	15.93	14.28	3.99	5.3
	CS		18.44	16.91	2.96	2.37
	HM	path	19.77	17.69	**4.54**	5.15
		lch	20.43	**18.52**	4.34	**5.4**
		wup	**20.55**	17.92	3.99	5.3

SJE, LatEm, ConSE, and GCN baselines result in *per-class* accuracies of 11.64, 11.99, 11.72, and 13.37 in the classic setting, respectively, and *per-class* accuracies of 1.7, 0.78, 0.43, and 2.96 in the generalized setting, respectively. Note that the bold values in all tables represent the maximum accuracies. Compared to those performances, when applied to the baselines, SJE, LatEm, ConSE, and GCN, our model deduces *per-class* maximum accuracies (bold values) of 19.49 (increase rate of 67%), 18.27 (52%),

16.36 (40%), and 20.55 (54%) in the classic setting, respectively and 3.31 (95%), 3.31 (324%), 3.31 (670%), and 4.54 (53%) in the generalized setting, respectively. The enhancements in the generalized setting are mostly higher than those in the classic setting, which indicates that the influence of contextual inference increases as the number of target categories increases. That in turn implies that the surrounding objects do have a common relationship, and thus, the context induces predictions of relevant types of categories among the various types of the entire categories.

CS shows better performance than SM in the classic setting but poor performance in the generalized setting. In several evaluations, it has no effect at all. We attribute that difference to the properties of the external knowledge sources. A semantic embedding space is trained from a text corpus in a sub-symbolic manner; thus, it contains less intuitive information, such as concept relations or category types, than a symbolic knowledge graph, which explains the relatively poor effectiveness of CS when the number of categories is large.

Moreover, for SJE, LatEm, and ConSE, the calibrated prediction performances are mostly the same in the generalized setting, which indicates that the visual inference results are not referred in the predictions in those cases. In other words, our experiments have verified the importance of considering contextual information.

- Top-n Result

The top-N evaluation is conducted on the HM_{path} method which shows generally flat performance in the previous hit@1 evaluation (Table 3). Our proposed method enhances performance compared with the baselines even in the top-n. As n increases, the rate of performance improvement tends to decrease slightly. In particular, according the performance of the ConSE, baseline at top-n is relatively higher than top-1, the improvement rate becomes lower. In other words, it implies that context plays a particularly important role in giving the correct category the highest ranking. Moreover, the significant effectiveness of similarity-based contextual inference is validated in the top-n prediction results.

Table 3. Top-n prediction results of HM_{path} by applying contextual inference (CI) to the baselines.

Method	Classic				Generalized			
	hit@3		hit@5		hit@5		hit@10	
	per-cls	per-ins	per-cls	per-ins	per-cls	per-ins	per-cls	per-ins
SJE	25.26	27.42	35.36	37.56	6.14	7.39	10.83	11.89
SJE + CI	**35.95**	**36**	**47.06**	**46.04**	**8.06**	**10.28**	**12.79**	**14.7**
LatEm	24.64	26	32.85	34.32	4.72	5.28	7.36	7.67
LatEm + CI	**30.15**	**30.59**	**41.05**	**39.66**	**6.19**	**7.92**	**9.19**	**10.45**
ConSE	25.91	25.7	36.16	34.36	5.51	5.76	8.18	8.09
ConSE + CI	**33.74**	**31.65**	**45.27**	**42.23**	**6.84**	**7.78**	**9.4**	**9.89**
GCN	29.24	26.34	40.01	37.59	7.34	6.38	12.29	10.3
GCN + CI	**36.43**	**34.13**	**47.15**	**44.94**	**10.11**	**9.85**	**16.4**	**15.02**

- Ablation Study

We conduct an additional experiment to validate whether a nearby object is highly relevant to the target object. The ablation study is performed by excluding the module that applies the defined distance calculation from the entire model. For the contextual inference, the HM_{path} calibration is applied with a normal average instead of the distance-weighted average in Equation (9). As seen in Table 4, even with only the similarity measurement, our proposed approach outperforms the existing models by a large margin. In most cases, however, the ablation study confirms that distance weighting boosts performance, verifying the validity of our defined distance calculation.

Table 4. An ablation study on HM_{path}.

Method	Classic/U				Generalized/U			
	hit@1		hit@5		hit@1		hit@5	
	per-cls	per-ins	per-cls	per-ins	per-cls	per-ins	per-cls	per-ins
SJE	11.64	13.2	35.36	37.56	1.7	2.03	6.14	7.39
SJE + sim	17.32	17.75	46.17	45.42	2.98	4.17	7.94	10.09
SJE + sim + dis	**18.56**	**18.77**	**47.06**	**46.04**	**3.2**	**4.43**	**8.06**	**10.28**
LatEm	11.99	13.22	32.85	34.32	0.78	0.83	4.72	5.28
LatEm + sim	16.46	17.2	39.67	39.09	2.4	3.67	6.05	7.75
LatEm + sim + dis	**17.36**	**17.65**	**41.05**	**39.66**	**3.2**	**4.43**	**6.19**	**7.92**
ConSE	11.72	12.03	36.16	34.36	0.43	0.38	5.51	5.76
ConSE + sim	14.31	14.85	44.74	41.34	2.59	3.67	6.73	7.67
CONSE + sim + dis	**16.09**	**15.59**	**45.27**	**42.23**	**3.2**	**4.43**	**6.84**	**7.78**
GCN	13.37	11.8	40.01	37.59	2.96	2.37	7.34	6.38
GCN + sim	19.34	17.06	**47.57**	**44.94**	4.12	4.64	**10.23**	9.77
GCN + sim + dis	**19.77**	**17.69**	47.15	**44.94**	**4.54**	**5.15**	10.11	**9.85**

4.3. Experiments on Visual Genome Categories

4.3.1. Dataset

VG contains more than 100K images, each of which has 35 object categories on average, and it is separated into two subsets: part-1, with about 60K images, and part-2, with about 40K images. In this experiment, we use categories and images in VG for both training and testing.

Our main goal in this experiment is to validate our method against both the baseline model and a recently proposed context-aware ZSR model [45] that is already evaluated on the VG dataset using the same setting as in [56]. We thus adopt the same split of seen and unseen class labels used in [56], 478 seen class labels and 130 unseen class labels, for a larger number of unseen class labels than in the first experiment. Similar to the work in [45], we use 55,038 images with 621,770 instances from part-1 of the VG dataset for training and 7818 images with 33,921 instances from part-2 for testing. The exact number of images differs slightly, but it is still considered within tolerance.

4.3.2. Visual Classifier and Semantic Embedding Space

Following the same experimental setting and using ConSE [23] as the baseline model for the comparative evaluation, prediction results from a visual classifier are needed to obtain a visual score for the target objects. We fine-tune ResNet-50 (without freezing the conv. layers) pretrained on ImageNet 2012 1K by building the dimension of the output layer to be the same as the number of seen class labels. The SGD optimizer is used for fine-tuning with approximately 240,000 iterations, and the learning rate, momentum, weight decay, and batch size are set to 0.001, 0.9, 0.0001, and 8, respectively.

For the sementic embedding space in this experiment, we have applied GloVe [27] with 300 dimensions pretrained on Wikipedia 2014 and Gigaword 5 [60]. All seen and unseen class labels individually have individually corresponding semantic embeddings or representative semantic embedding, just as in the first experiment. In addition, note that we use the same knowledge graph, WordNet, and SM measurements as well.

4.3.3. Results

- Evaluation Result

This experiment is conducted by matching its experimental configuration to Context-Aware (CA) ZSR [45] to the hilt, but the performance of the ConSE baseline is reproduced slightly differently due to a minute difference in the datasets, fine-tuning of the visual classifier, and the word embedding space.

We thus evaluate performance in terms of the improvement rate using the results given in Table 5. Recall that the numerical values represent accuracy in percentage. In the top-1 of the classic setting, CA presents its performance as 19.6 (increase rate of −1.51%) and 30.2 (9.02%) for a baseline performance of 19.9 and 27.7 on per-class and per-instance, respectively, whereas our proposed method has a performance of 15.34 (2.33%) and 32.62 (3.16%), for 14.99 per-class and 31.62 per-instance. Compared with CA, our method shows enhanced per-class performance, but the per-instance influence is slightly deficient. In the top-1 in the generalized setting, our method produces relatively low enhancement (0.05–0.27 and 0.29–0.81, for per-class and per-instance, respectively) compared with that of the CA (0.1–5.8 and 0.6–20.7). In the top-5 evaluations, our absolute performance of 33.58 on per-instance in the generalized setting outperforms CA's 29.4, although the performance of the ConSE baseline is lower than that of CA.

Consequently, the proposed method offers fair performance improvement in classifying an appropriate category compared with the existing method, and it more often gives categories related to the category type of the target object a high ranking, as shown by the first experimental results. This is furthermore validated through qualitative evaluations by actual exemplary analysis, which is detailed in the following subsection.

Table 5. Evaluation results for the proposed methods with the YOLOv2-80 detector on Visual Genome categories.

Method			Classic				Generalized			
			hit@1		hit@5		hit@1		hit@5	
			per-cls	per-ins	per-cls	per-ins	per-cls	per-ins	per-cls	per-ins
ConSE (**baseline**)			14.99	31.62	34.45	57.64	0.05	0.29	13.97	32.28
ConSE + CI (**proposed**)	SM	path	15.29	32	34.64	57.86	0.2	0.29	13.97	32.31
		lch	15.17	32.02	34.55	57.76	0.23	0.29	13.97	32.31
		wup	15.06	31.81	34.45	57.64	**0.27**	0.74	13.97	32.3
	CS		15.11	32.09	34.77	58.31	0.15	**0.81**	**14.39**	**33.58**
	HM	path	**15.34**	32.26	**34.95**	**58.45**	0.2	**0.81**	**14.39**	**33.58**
		lch	15.2	32.22	34.92	58.4	0.23	**0.81**	**14.39**	**33.58**
		wup	15.13	**32.62**	34.77	58.33	**0.27**	**0.81**	**14.39**	**33.58**

- Detector Comparison

In the previous experiment, surrounding information is obtained by detecting and recognizing surrounding objects using YOLOv2 on 80 categories. However, of those 80 categories, we use the results from only 65 that are disjoint from VG's unseen categories. That limited reference to surrounding objects for a somewhat small amount of classification categories could lead to poor performance improvement. For ascertainment, we conduct an additional experiment by applying another pretrained detector with a large scale category, YOLOv2-9K [54], and applying the other source, ground-truths for seen class labels.

YOLOv2-9K produces detected results above the confidence threshold of 0.1 for 8955 categories out of 9K that have available semantic embedding and disjoint from unseen categories. In both YOLOv2-80 and YOLOv2-9K, we exclude detections whose intersection-over-union with the bounding box of the target object is 0.8 or higher.

Ground-truths for seen class labels, GT_S, are applied as surrounding information with a fixed confidence by exploiting bounding boxes from annotations on the seen class labels in the test images. The value of the prediction probability p in Equation (2) is fixed to 1.0 in GT_S. To alleviate the problem of noise caused by high-frequency objects, such as a category, window in the ground-truths, we set the system to reference only one randomly selected object per class label.

The evaluation result of the variation in detection of surrounding objects is presented in Table 6. The performance of YOLOv2-9K is generally poor because of its tendency to return detection results for

higher-level categories or abstract concepts such as *whole*, *instrumentality*, and *creation*. That negative tendency causes confusion by misreferencing the surrounding information. GT_S outperforms the models to which other detectors are applied in most cases, which means that the contextual inference is well harmonized with the visual inference in ZSR.

Table 6. A comparative evaluation in the differentiation of sources for surrounding information with the HM_{path} method.

Method	Classic						Generalized					
	hit@1		hit@5		hit@10		hit@1		hit@5		hit@10	
	per-cls (Δ(%))	per-ins (Δ(%))	per-cls (Δ(%))	per-ins (Δ(%))	per-cls (Δ(%))	per-ins (Δ(%))	per-cls (Δ(%))	per-ins (Δ(%))	per-cls (Δ(%))	per-ins (Δ(%))	per-cls (Δ(%))	per-ins (Δ(%))
ConSE (baseline)	14.99	31.62	34.45	57.64	45.42	68.18	0.05	0.29	13.97	32.28	23.63	46.21
ConSE + CI (YOLOv2- 80)	**15.34 (2.33↑)**	32.26 (2.02↑)	**34.95 (1.45↑)**	58.45 (1.41↑)	46.28 (1.89↑)	69.06 (1.29↑)	0.2 (300↑)	0.81 (179↑)	14.39 (3.01↑)	33.58 (4.03↑)	24.15 (2.2↑)	46.94 (1.58↑)
ConSE + CI (YOLOv2- 9K)	15.22 (1.53↑)	32.32 (2.21↑)	34.55 (0.29↑)	58.15 (0.88↑)	45.63 (0.46↑)	68.79 (0.89↑)	**0.69 (1280↑)**	1.11 (282↑)	14.14 (1.22↑)	33.8 (4.71↑)	23.91 (1.18↑)	47.4 (2.58↑)
ConSE + CI (GT_S)	15.14 (1↑)	**32.94 (4.17↑)**	34.67 (0.64↑)	**59.12 (2.57↑)**	**46.81 (3.06↑)**	**69.65 (2.16↑)**	0.17 (240↑)	**1.3 (348↑)**	**14.4 (3.08↑)**	**33.81 (4.74↑)**	**24.34 (3↑)**	**48.22 (4.35↑)**

However, there is not a significant gap between the performance of YOLOv2-80 and that of GT_S. In other words, our method is easy to apply practically to existing detectors without needs to provide accurate surrounding information because it can reference the potentials of classification candidates. Furthermore, it is promising to show optimized performance with a detector that stably recognizes a wide variety of categories.

- Qualitative Analysis

We also use GT_S-based surrounding information for qualitative evaluations to clearly analyze the effects of contextual inference. The HM_{path} method on ConSE is applied to the contextual inference, and its optimized values of balancing parameters, α and β on GT_S, are used for the evaluation.

Figure 3 depicts several qualitative experimental results. The upper three and lower two results show the positive and negative effects of our method, respectively. In the first example, the ground-truth class label of the target object (orange box) is *chair*. The visual inference predicts *chair*-like *bathhub* and *toilet* in the first and second rank, respectively, with *chair* in the third. By referring to the nearby objects, *table* and *sofa*, which are related to a living room, the contextual inference calibrates the prediction results to rank *chair* first. Another surrounding object, *fan*, far from the target, does not adversely affect the positive calibration because the degree of reference is reduced by distance-weighting. This verifies our assumption that related objects are likely to be located nearby. In fact, the CSs of *sofa* and *table* with *chair* in GloVe are both above 0.4, whereas that of fan is approximately 0.11. Similar positive phenomena are confirmed from the second and third examples, as well. Our methods infers the correct class label, *flower*, which is not even in the top-5 of the visual-only prediction, into the top-3 by using the contexts close to the target, *leaf* and *vase*, in the second example. In the third example, which has little useful surrounding information, the rank of *collar* is slightly increased by the closest object, *dog*, with our distance formula.

Although the overall performance is enhanced by the positive calibration of contextual inference, it still has some negative effects, particularly in the cases of general categories irrelevant to specific objects. For example, the visually fourth-ranked ground-truths, *writing* and *sky* are excluded from the top-5 and re-ranked to ninth and sixth, respectively. The tableware objects surrounding *writing* negatively affect the prediction and produces high rankings for tableware-related categories such as *coffee* and *chair*. *Sky*, which appears in most outdoor images regardless of the particular environment, generally does not have a semantic relation with a specific object. In the last example, the ground-truth category is undervalued by the contextual inference and rather structure-related *window* and *building* are overvalued due to the low similarities between *sky* and its surrounding *roof* and *train*. However,

our results generally validate that our methods based on the contextual inference positively affect ZSR in an advanced way, as shown by the previous evaluation results.

Figure 3. Qualitative examples evaluated on HM_{path} in the classic setting using the ground-truths of seen categories instead of the pretrained detector. The orange and blue boxes indicate the target unseen objects and ground-truth surrounding objects, respectively.

5. Conclusions

We have proposed a novel approach to ZSR that enhances the performance of existing ZSL methods. Our method uses surrounding information as context by measuring the similarities of each surrounding object and applying distance-weighted averaging with a defined distance calculation formula to calibrate the visually predicted results. We performed experimental evaluations with various combinations of similarity measures to validate the comparative performance of our proposed method on two different datasets with ImageNet and Visual Genome categories. Our experimental results demonstrate that our method enhances performance by a large margin compared with existing methods. The ablation and differentiation in detectors studies verified the effectiveness of distance-weighting and the potential practicality of our method, respectively. Future research could consider the topological relationships among objects in an image and optimized semantic embedding for ZSR using annotations of train images as sources.

Author Contributions: Conceptualization, D.S.C.; data curation, D.S.C. and G.H.C.; formal analysis, D.S.C.; funding acquisition, Y.S.C.; investigation, D.S.C. and Y.S.C.; methodology, D.S.C. and G.H.C.; resources, G.H.C.; software, D.S.C. and G.H.C.; supervision, Y.S.C.; validation, D.S.C. and G.H.C.; writing—original draft, D.S.C.; writing—review and editing, D.S.C., G.H.C., and Y.S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Technology Innovation Program (10077553, Development of Social Robot Intelligence for Social Human-Robot Interaction of Service Robots) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea), supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2020-0-01373, Artificial Intelligence Graduate School Program (Hanyang University)), and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1A2C1014037).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sun, C.; Shrivastava, A.; Singh, S.; Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the IEEE INTERNATIONAL Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 843–852.
2. Farhadi, A.; Endres, I.; Hoiem, D.; Forsyth, D. Describing objects by their attributes. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1778–1785.
3. Akata, Z.; Perronnin, F.; Harchaoui, Z.; Schmid, C. Label-embedding for attribute-based classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–27 June 2013; pp. 819–826.
4. Kodirov, E.; Xiang, T.; Gong, S. Semantic autoencoder for zero-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3174–3183.
5. Xian, Y.; Lampert, C.H.; Schiele, B.; Akata, Z. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2251–2265. [CrossRef] [PubMed]
6. Akata, Z.; Reed, S.; Walter, D.; Lee, H.; Schiele, B. Evaluation of output embeddings for fine-grained image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–12 June 2015; pp. 2927–2936.
7. Changpinyo, S.; Chao, W.L.; Gong, B.; Sha, F. Synthesized classifiers for zero-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 25 June–1 July 2016; pp. 5327–5336.
8. Frome, A.; Corrado, G.S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; Mikolov, T. Devise: A deep visual-semantic embedding model. In Proceedings of the Advances in Neural Information Processing Systems, Douglas County, NV, USA, 5–10 December 2013; pp. 2121–2129.
9. Wang, X.; Ye, Y.; Gupta, A. Zero-shot recognition via semantic embeddings and knowledge graphs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 19–22 June 2018; pp. 6857–6866.
10. Rohrbach, M.; Stark, M.; Schiele, B. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In Proceedings of the IEEE CVPR 2011, Colorado Springs, CO, USA, 21–23 June 2011; pp. 1641–1648.
11. Ferrari, V.; Zisserman, A. Learning visual attributes. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–13 December 2008; pp. 433–440.
12. Lampert, C.H.; Nickisch, H.; Harmeling, S. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 453–465. [CrossRef] [PubMed]
13. Farhadi, A.; Endres, I.; Hoiem, D. Attribute-centric recognition for cross-category generalization. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2352–2359.
14. Duan, K.; Parikh, D.; Crandall, D.; Grauman, K. Discovering localized attributes for fine-grained recognition. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3474–3481.
15. Changpinyo, S.; Chao, W.L.; Gong, B.; Sha, F. Classifier and exemplar synthesis for zero-shot learning. *Int. J. Comput. Vis.* **2020**, *128*, 166–201. [CrossRef]

16. Jayaraman, D.; Grauman, K. Zero-shot recognition with unreliable attributes. In Proceedings of the Advances in Neural Information Processing Systems, Montréal, QC, Canada, 8–13 December 2014; pp. 3464–3472.
17. Misra, I.; Gupta, A.; Hebert, M. From red wine to red tomato: Composition with context. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1792–1801.
18. Chen, H.; Gallagher, A.C.; Girod, B. What's in a name? first names as facial attributes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–27 June 2013; pp. 3366–3373.
19. Douze, M.; Ramisa, A.; Schmid, C. Combining attributes and fisher vectors for efficient image retrieval. In Proceedings of the IEEE CVPR 2011, Colorado Springs, CO, USA, 21–23 June 2011; pp. 745–752.
20. Liu, J.; Kuipers, B.; Savarese, S. Recognizing human actions by attributes. In Proceedings of the IEEE CVPR 2011, Colorado Springs, CO, USA, 21–23 June 2011; pp. 3337–3344.
21. Scheirer, W.J.; Kumar, N.; Belhumeur, P.N.; Boult, T.E. Multi-attribute spaces: Calibration for attribute fusion and similarity search. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2933–2940.
22. Xian, Y.; Akata, Z.; Sharma, G.; Nguyen, Q.; Hein, M.; Schiele, B. Latent embeddings for zero-shot classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 25 June–1 July 2016; pp. 69–77.
23. Norouzi, M.; Mikolov, T.; Bengio, S.; Singer, Y.; Shlens, J.; Frome, A.; Corrado, G.S.; Dean, J. Zero-shot learning by convex combination of semantic embeddings. *arXiv* **2013**, arXiv:1312.5650.
24. Socher, R.; Ganjoo, M.; Manning, C.D.; Ng, A. Zero-shot learning through cross-modal transfer. In Proceedings of the Advances in Neural Information Processing Systems, Douglas County, NV, USA, 5–10 December 2013; pp. 935–943.
25. Fu, Y.; Hospedales, T.M.; Xiang, T.; Fu, Z.; Gong, S. Transductive multi-view embedding for zero-shot recognition and annotation. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 584–599.
26. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the Advances in Neural Information Processing Systems, Douglas County, NV, USA, 5–10 December 2013; pp. 3111–3119.
27. Pennington, J.; Socher, R.; Manning, C. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
28. Mensink, T.; Verbeek, J.; Perronnin, F.; Csurka, G. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 488–501.
29. Palatucci, M.; Pomerleau, D.; Hinton, G.E.; Mitchell, T.M. Zero-shot learning with semantic output codes. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 6–9 December 2009; pp. 1410–1418.
30. Kampffmeyer, M.; Chen, Y.; Liang, X.; Wang, H.; Zhang, Y.; Xing, E.P. Rethinking knowledge graph propagation for zero-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 11487–11496.
31. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
32. Miller, G.A. WordNet: A lexical database for English. *Commun. ACM* **1995**, *38*, 39–41. [CrossRef]
33. Wah, C.; Branson, S.; Perona, P.; Belongie, S. Multiclass recognition and part localization with humans in the loop. In Proceedings of the 2011 IEEE International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2524–2531.
34. Khosla, A.; Jayadevaprakash, N.; Yao, B.; Li, F.F. Novel dataset for fine-grained image categorization: Stanford dogs. In Proceedings of the CVPR Workshop on Fine-Grained Visual Categorization (FGVC), Colorado Springs, CO, USA, 25 June 2011; Volume 2.
35. Lampert, C.H.; Nickisch, H.; Harmeling, S. Learning to detect unseen object classes by between-class attribute transfer. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 951–958.

36. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
37. Song, Z.; Chen, Q.; Huang, Z.; Hua, Y.; Yan, S. Contextualizing object detection and classification. In Proceedings of the IEEE CVPR 2011, Colorado Springs, CO, USA, 21–23 June 2011; pp. 1585–1592.
38. Desai, C.; Ramanan, D.; Fowlkes, C.C. Discriminative models for multi-class object layout. *Int. J. Comput. Vis.* **2011**, *95*, 1–12. [CrossRef]
39. Torralba, A.; Murphy, K.P.; Freeman, W.T. Using the forest to see the trees: Exploiting context for visual object detection and localization. *Commun. ACM* **2010**, *53*, 107–114. [CrossRef]
40. Divvala, S.K.; Hoiem, D.; Hays, J.H.; Efros, A.A.; Hebert, M. An empirical study of context in object detection. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1271–1278.
41. Rabinovich, A.; Vedaldi, A.; Galleguillos, C.; Wiewiora, E.; Belongie, S. Objects in context. In Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–20 October 2007; pp. 1–8.
42. Yu, R.; Chen, X.; Morariu, V.I.; Davis, L.S. The role of context selection in object detection. *arXiv* **2016**, arXiv:1609.02948.
43. Chen, X.; Li, L.J.; Fei-Fei, L.; Gupta, A. Iterative visual reasoning beyond convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 19–22 June 2018; pp. 7239–7248.
44. Galleguillos, C.; Belongie, S. Context based object categorization: A critical survey. *Comput. Vis. Image Underst.* **2010**, *114*, 712–722. [CrossRef]
45. Luo, R.; Zhang, N.; Han, B.; Yang, L. *Context-Aware Zero-Shot Recognition*; AAAI: Menlo Park, CA, USA, 2020; pp. 11709–11716.
46. Zablocki, E.; Bordes, P.; Soulier, L.; Piwowarski, B.; Gallinari, P. Context-aware zero-shot learning for object recognition. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 7292–7303.
47. Jiang, J.J.; Conrath, D.W. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv* **1997**, cmp-lg/9709008.
48. Taieb, M.A.H.; Aouicha, M.B.; Hamadou, A.B. Ontology-based approach for measuring semantic similarity. *Eng. Appl. Artif. Intell.* **2014**, *36*, 238–261. [CrossRef]
49. Reed, S.; Akata, Z.; Lee, H.; Schiele, B. Learning deep representations of fine-grained visual descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 25 June–1 July 2016; pp. 49–58.
50. Ristoski, P.; Paulheim, H. Rdf2vec: Rdf graph embeddings for data mining. In Proceedings of the International Semantic Web Conference; Springer: Cham, Switzerland, 2016; pp. 498–514.
51. Zhang, Z.; Saligrama, V. Zero-shot learning via semantic similarity embedding. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015; pp. 4166–4174.
52. Romera-Paredes, B.; Torr, P. An embarrassingly simple approach to zero-shot learning. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2152–2161.
53. Fu, Y.; Sigal, L. Semi-supervised vocabulary-informed learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 25 June–1 July 2016; pp. 5337–5346.
54. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
55. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision;* Springer: Cham, Switzerland, 2014; pp. 740–755.
56. Bansal, A.; Sikka, K.; Sharma, G.; Chellappa, R.; Divakaran, A. Zero-shot object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 384–400.
57. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–12 June 2015; pp. 1–9.
58. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.

59. Zhang, L.; Xiang, T.; Gong, S. Learning a deep embedding model for zero-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2021–2030.
60. Parker, R.; Graff, D.; Kong, J.; Chen, K.; Maeda, K. *English Gigaword Fifth Edition, Linguistic Data Consortium*; Linguistic Data Consortium: Philadelphia, Pennsylvania, 2011.