

Article

# Hierarchical Semantic Loss and Confidence Estimator for Visual-Semantic Embedding-Based Zero-Shot Learning

Sanghyun Seo  and Juntae Kim \*

Department of Computer Engineering, Dongguk University, 30, Pildong-ro 1-gil, Jung-gu, Seoul 04620, Korea  
\* Correspondence: jkim@dongguk.edu; Tel.: +82-2-2260-3712

Received: 11 June 2019; Accepted: 29 July 2019; Published: 2 August 2019



**Abstract:** Traditional supervised learning is dependent on the label of the training data, so there is a limitation that the class label which is not included in the training data cannot be recognized properly. Therefore, zero-shot learning, which can recognize unseen-classes that are not used in training, is gaining research interest. One approach to zero-shot learning is to embed visual data such as images and rich semantic data related to text labels of visual data into a common vector space to perform zero-shot cross-modal retrieval on newly input unseen-class data. This paper proposes a hierarchical semantic loss and confidence estimator to more efficiently perform zero-shot learning on visual data. Hierarchical semantic loss improves learning efficiency by using hierarchical knowledge in selecting a negative sample of triplet loss, and the confidence estimator estimates the confidence score to determine whether it is seen-class or unseen-class. These methodologies improve the performance of zero-shot learning by adjusting distances from a semantic vector to visual vector when performing zero-shot cross-modal retrieval. Experimental results show that the proposed method can improve the performance of zero-shot learning in terms of hit@k accuracy.

**Keywords:** zero-shot learning; hierarchical semantic loss; confidence estimator; zero-shot cross-modal retrieval; visual-semantic embedding

---

## 1. Introduction

Recently, various learning methodologies have been proposed in the field of machine learning, focusing on the deep learning model. Many of them belong to supervised learning based on training data with class labels. A higher performance of the trained model can be achieved if supervised learning is performed well with a large amount of labeled training data. However, supervised learning has some limitations on training and application in the real world. Firstly, obtaining labeled training data is costly and time-consuming. For example, in the object recognition area, a label needs to be attached to each image data [1]. In case of object detection, the image label and the location information must be collected as well [2,3]. Moreover, attaching labels to training data depends on each task, and if the problem domain is changed (even for the same task), new training data must be constructed. In addition, as supervised learning can only recognize the class labels given in training data, when the data of an unseen class which is not included in the labels of training data are input, the model cannot produce correct result. In other words, in the existing supervised learning paradigm, cost and time consumed in collecting training data are burdensome, and even if the training data are collected with difficulty, the unseen labels beyond the range of the training data cannot be recognized.

To address this issue, few-shot learning, one-shot learning, and zero-shot learning have been proposed [4–7]. In many practical machine-learning applications, there is often an imbalance in the classes of training data. Traditionally, over-sampling or under-sampling methodologies have been

applied to training data with imbalanced classes [8,9]. In this situation, few-shot learning or one-shot learning provides a way for a few classes in the training data to be trained more efficiently in a given class-imbalanced training dataset.

Meanwhile, zero-shot learning refers to a series of methodologies that perform learning on a class, even when there are no training data for a particular class [6,7,10–20]. Generally, the data and their class used in the training phase are referred to as *seen class data*, while the data and their class not used in the training phase are called *unseen class data*. Zero-shot learning can learn from the seen data and create an inference model to recognize the unseen class data properly. At this time, if we concentrate only on the processing of unseen class data, the performance of the seen data might degrade. Zero-shot learning that performs learning considering the performance of both seen and unseen class data is called *generalized zero-shot learning* [21].

There are two major approaches to zero-shot learning. The first approach is to train the model using visual data such as images and attributes in each visual data. In this method, when unseen class images are input, zero-shot inference is performed using their attributes to recognize the unseen class labels. For example, a bird has wings, beak, and other attributes. The zero-shot model can recognize the label of unseen class image by using attributes extracted from input unseen class image [10]. However, in such a case, an attribute for training visual data must be constructed in advance. The second approach for zero-shot learning is to embed semantic data that include a larger number of concepts at the same time as covering the classes of training data into a common vector space, so that it can be compared with the embedded visual data [6]. This approach also has the advantage of cross-modal retrieval because it embeds the heterogeneous data from different domains into a common space [22]. Recently, several embedding-based zero-shot learning methodologies that use deep neural networks as the embedding model have been proposed [7].

This paper proposes a zero-shot learning methodology based on the above second approach by using the embedding model with images as visual data and word vectors as semantic data. The proposed scheme has the following two main contributions.

- We propose a hierarchical semantic loss in the training phase. It uses hierarchical knowledge in semantic data to calculate loss function. It shows the more meaningful embedding results in performing zero-shot cross-modal retrieval.
- We propose an unseen confidence estimator in the inference phase. It estimates the confidence score for input query image and adjusts the distances between a query image vector and unseen class semantic vectors to improve the performance of generalized zero-shot learning.

Generally, in embedding-based zero-shot learning using word vectors as semantic data, the embedding model is trained such that the distance between the positive pair of visual data and the semantic data such as word vector of labels is closely embedded into common vector space by using mean squared error loss or triplet margin loss. However, as the existing triplet margin loss uses random sampling in the process of selecting a negative sample, unnecessary training occurs because the semantic data having no relation to it as a negative sample are used in training the embedding model [23]. In our way, hierarchical semantic loss selects a semantically more closely related data as a negative sample of the triplet margin loss and can create a better zero-shot learning model. The proposed methodology has a novelty in that it utilizes the common hierarchical knowledge of semantic data, which is easy to obtain, compared with the zero-shot learning methodologies using attributes that are relatively difficult to collect.

We also propose a generalized zero-shot learning methodology using a confidence estimator, which estimates whether the input visual data are seen class or unseen class. The proposed methodology increases the influence of the class label of unseen data if their confidence on the unseen class data is high. Experiments conducted using CIFAR-100 and CIFAR-10 datasets show that the proposed hierarchical semantic loss and confidence estimator can perform more efficient embedding-based zero-shot learning.

## 2. Related Work

### 2.1. Embedding-Based Zero-Shot Learning

As described above, zero-shot learning can be divided into two approaches. The first approach is to use attributes of visual data and the other is to embed visual data and semantic data into a common vector space. The second approach is to embed visual-semantic data into a common vector space to compare similarity and retrieve the most similar vectors from the cross-modal embedded vectors. There are three types of approaches to embedding-based zero-shot learning. The first is to fix the semantic space as an anchor and to embed the visual data into this space [7]. As a representative example of this approach, a methodology in which a pre-trained semantic vector is fixed and a visual model is used to approximate the image data to the semantic vector has been proposed. Although this approach can perform visual-semantic embedding efficiently, it has high dependency on the semantic representation of the language model because it uses the fixed semantic vector obtained through the pre-learned language model. This method uses ranking loss to decrease the distance between the positive and negative pairs by using the margin and similarity function, such as mean squared error, as the loss function [7,22].

The second approach of embedding-based zero-shot learning is to fix visual data as an anchor and embed semantic data into the visual space [19]. This approach aims at mitigating the hubness problem by finding the discriminative capacity of visual features. Recently, research using semantic inter-class relations has been conducted [24].

The final approach to embedding-based zero-shot learning is to embed data into latent intermediate spaces. This approach embeds the latent intermediate space using a compatibility function that distinguishes visual and semantic features from the class [13,15,25–27]. However, all these methodologies try to distinguish between classes, but there is a limitation on intra-class distribution.

The zero-shot learning of the various approaches discussed above can be roughly summarized by performing distance comparisons between embedded vectors in the common vector space. To improve the performance of embedding-based zero-shot learning, a method of reflecting the prior probabilities of embedded vectors to embedding has been proposed without further learning of the embedding model. This method is a semantic embedding technique that applies a simple convex combination by multiplying k-candidate semantic vectors obtained as a result of embedding by  $p(y|x)$ . This approach is more rigorous than the previous visual-semantic embedding model, so embedded vectors are more likely to stay on the manifold and better visual-semantic embedding can be expected [7,11].

### 2.2. Triplet Margin Loss

The simplest way to determine the similarity between images is to use the average of the pixels in each image. However, it is difficult to measure the degree of similarity with respect to the shape of an image by focusing on its brightness. To compensate for this problem, an image-embedding model can be constructed and the similarity can be measured by calculating the distance between the embedded image vectors [28]. For example, the distance function between images  $p$  and  $q$ , defined by the Euclidean distance in the image embedding space, can be expressed as

$$d(p, q) = \|f(p) - f(q)\|_2^2 \quad (1)$$

The function  $f(\cdot)$  is the image embedding model, and  $d(\cdot, \cdot)$  is the l2 norm distance, such as the Euclidean distance function calculated by the difference between  $f(p)$  and  $f(q)$ . If we use  $d(p, q)$  to find the most similar image  $q$  for a query image  $p$ , this is similar to the general  $k$ -nearest neighbor method. In other words, it is the same as solving a similarity ranking problem. That is,  $f(\cdot)$  is used in training the embedding model such that a similar image  $q$  for an anchor image  $p$  is embedded at distances closer than the other image.

There is a triplet margin loss as a loss function that is appropriate to solve this problem. Triplet margin loss uses triplet  $TL = (p, p^+, p^-)$ , where  $p$  is the query image,  $p^+$  is the positive image which is one of the images having the same semantic label for the query image  $p$ , and  $p^-$  is the negative image which is one of the images having different semantic label for the query image  $p$ . Thus, triplet margin loss can be expressed as

$$TL(p, p^+, p^-) = \max(0, \alpha + d(p, p^+) - d(p, p^-)) \quad (2)$$

Triplet margin loss implies that the embedded vector of the positive image  $p^+$  is embedded closer to the query image  $p$  than the negative image  $p^-$ . The triplet loss function is designed to learn that the distance of the positive pair is embedded as small as the margin  $\alpha$ . If the distance of the negative pair is larger than the margin of the positive pair, it is designed to be set to 0 so that no additional learning occurs.

The above triplet margin loss can be modified to measure the similarity of image-image, image-text, text-text, etc., beyond merely comparing the similarity between image and image [22]. For example, a pair is constructed on the basis of the semantic similarity between image and text, and image-text embedding model can be trained such that the positive-pairs of image-text are embedded closer [29]. Several embedding-based zero-shot learning methodologies have used triplet margin loss. However, when selecting a negative sample, it is pointed out that random sampling is very inefficient [23]. Nonetheless, as triplet loss can design a loss function that can control the ranking between embedded vectors, it is necessary to study a methodology that can compensate the current disadvantages.

### 2.3. Confidence Score

One of the various research areas on neural networks is to measure uncertainty in the output of neural networks [30]. This is an extension of traditional research that confirms the degree to which the model results can be relied upon by measuring the uncertainty of neural networks. One of the various approaches in measuring uncertainty is a method of designing a threshold-based detector that can calculate the confidence score [31]. This approach involves setting a certain threshold  $\varepsilon$ , setting the maximum value of the predictive distribution as the confidence score, and determining it as in-distribution when it is greater than  $\varepsilon$ . As this method is sensitive to the output value of softmax from neural networks, a method of using the temperature softmax method and the maximum value of scaled predictive distribution as the confidence score has been proposed [32]. Temperature softmax can relatively reduce the effect of extreme values in the existing methodology, because the difference in the distribution of softmax output is made smaller.

Confidence score measurement can be considered as a method of measuring the uncertainty of a neural network, but it is also possible to measure the in-distribution or out-of-distribution of input data. For example, the input data having a similar distribution to that of the data used during training a model can be judged as in-distribution by measuring a relatively high confidence score, and when the unseen data are input, the out-of-distribution will be more likely to be judged. A methodology for measuring out-of-distribution or performing novelty detection through such an approach has been proposed [33,34].

## 3. Hierarchical Semantic Loss and Confidence Estimator for Generalized Zero-Shot Learning

### 3.1. Zero-Shot Learning with Hierarchical Knowledge

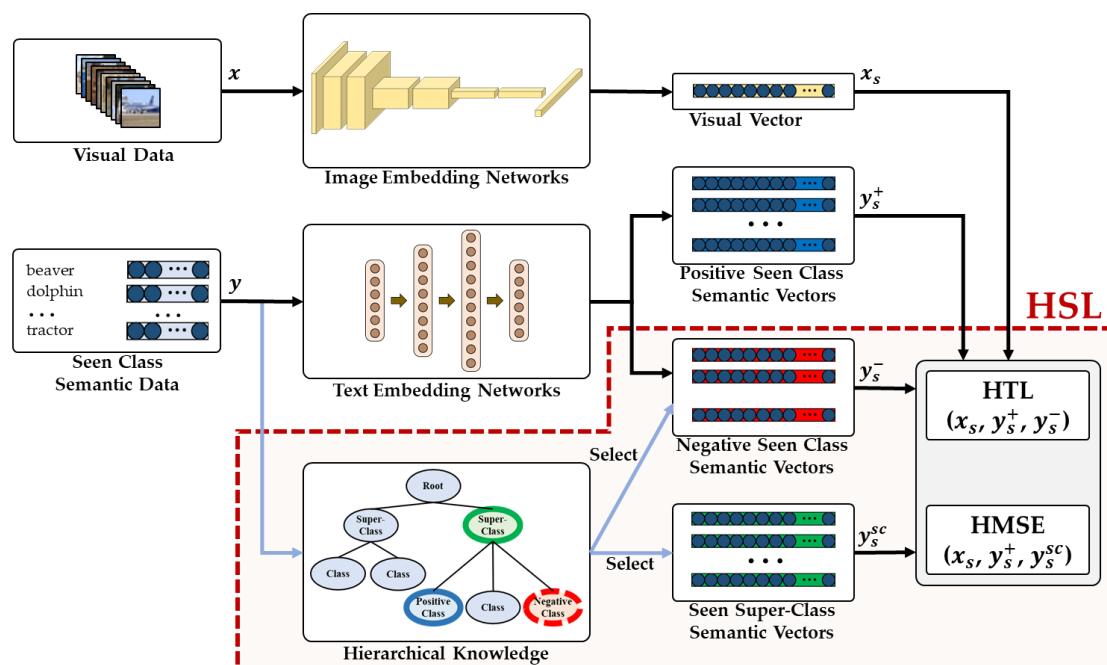
The proposed embedding-based generalized zero-shot learning model uses both visual data, e.g., images, and semantic data which is a class label of the visual data, as input data. Semantic data is a distributed representation vector obtained from a pre-trained language model. We can obtain visual and semantic vectors of the same dimension from the visual and semantic data through an embedding model based on deep neural networks. Unlike the general deep learning model which uses the same classes in both training and inference phase, the generalized zero-shot learning model distinguishes

seen and unseen classes. The seen-classes are used in both the training and inference phase, but the unseen-classes are used only in the inference phase. There are no overlapping classes in the seen and unseen classes.

The zero-shot architecture based on visual-semantic embedding classifies input visual data mainly through cross-modal retrieval, which involves finding a target class by retrieving a semantic vector of the closest distance by setting a visual vector as a query. At this time, zero-shot learning refers to the case where only the unseen-classes are used as the target semantic vectors for zero-shot cross-modal retrieval. In contrast, generalized zero-shot learning refers to the case where both seen and unseen classes are used as the target semantic vectors. That is, the difficulty of generalized zero-shot learning is higher than that of zero-shot learning.

This paper proposes the zero-shot learning methodology using hierarchical semantic loss (HSL) for improving the performance of visual-semantic embedding. The proposed loss function uses hierarchical knowledge such as pre-defined WordNet hierarchy, hierarchy of label structure, or manually constructed hierarchy. Hierarchical knowledge represents the relation of semantic data and is used to construct HSL. All semantic data consist of distributed vectors obtained through a pre-trained language model. Figure 1 shows the structure of the proposed generalized zero-shot learning process and the notations used in the proposed zero-shot learning process are as follows.

- $x$ : seen-class visual data such as an image in training dataset.
- $y$ : seen-class semantic data such as a text label of  $x$  in training dataset.
- $x_s$ : seen-class visual vector embedded from image embedding networks.
- $y_s^+$ : positive seen-class semantic vector embedded from text embedding networks.
- $y_s^-$ : negative seen-class semantic vector embedded from text embedding networks. It is selected by using hierarchical knowledge such that it has same super-class with  $y_s^+$ .
- $y_s^{sc}$ : seen super-class semantic vector obtained from pre-trained word vector and selected by using hierarchical knowledge. It is the super-class of  $y_s^+$  and  $y_s^-$ .



**Figure 1.** Zero-shot learning process with hierarchical semantic knowledge.

In the training phase, only seen-class semantic data are used with visual data. At this time, the proposed methodology uses a more efficient loss function by using hierarchical knowledge about the seen-class semantic data. The seen super-class semantic vector is obtained from hierarchical knowledge, which is a fixed vector from a pre-trained language model. There are three examples presented as hierarchical knowledge sources: WordNet, label structure of training dataset, and manually constructed hierarchy [35,36]. For example, a general ImageNet dataset consists of flat 1000 labels. In this situation, we can design the hierarchy by using WordNet, or manually construct the hierarchy directly based on the semantic similarity of the labels. In addition, if the hierarchical structure is already defined in datasets such as the CIFAR dataset, it can be used.

HSL is a loss function using hierarchical knowledge for training each embedding model. It consists of hierarchical mean squared error (HMSE) and hierarchical triplet margin loss (HTL). HMSE is used to minimize the distance between a visual vector  $x_s$  and a seen super-class vector  $y_s^{sc}$  and the distance between a seen class semantic vector  $y_s^+$  and  $y_s^{sc}$ . HTL allows the embedding networks to be trained such that the distance between the positive pairs becomes smaller than the distance between the negative pairs. Here, hierarchical knowledge is used in selecting the negative seen-class semantic vectors  $y_s^-$ .

HSL is based on the existing embedding-based zero-shot learning methodology that performs joint learning of image and text embedding networks. HSL is used in training each image and text embedding networks to embed semantically similar  $y_s^+$  in similar locations for  $x_s$ . In this case, HMSE is the sum of distances from  $x_s$  and  $y_s^+$  to fixed  $y_s^{sc}$ . However, using only HMSE makes it difficult to distinguish between visual vectors and other seen-class semantic vectors that have the same super-class, resulting in performance degradation of zero-shot inference. Therefore, HTL is applied to solve this problem. When selecting a negative sample of triplet loss, HTL selects  $y_s^-$  among the semantic vectors that have the same super-class as  $y_s^+$ . This enables embedding more efficient so that the distinction between similar semantic vectors become clearer.

The HMSE of the proposed HSL is given as:

$$HMSE(x_s, y_s^+, y_s^{sc}) = \|x_s - y_s^{sc}\|_2^2 + \|y_s^+ - y_s^{sc}\|_2^2 \quad (3)$$

In Equation (3), the distance is measured with l2 norm and the visual-semantic embedding networks are trained in the direction of minimizing it. In the above equation, because HMSE uses a fixed super class semantic vector  $y_s^{sc}$ , all embedding networks can be trained faster and stably.

The HTL of the proposed HSL is given as:

$$HTL(x_s, y_s^+, y_s^-, \alpha) = \max(0, \alpha + \|x_s - y_s^+\|_2^2 - \|x_s - y_s^-\|_2^2) \quad (4)$$

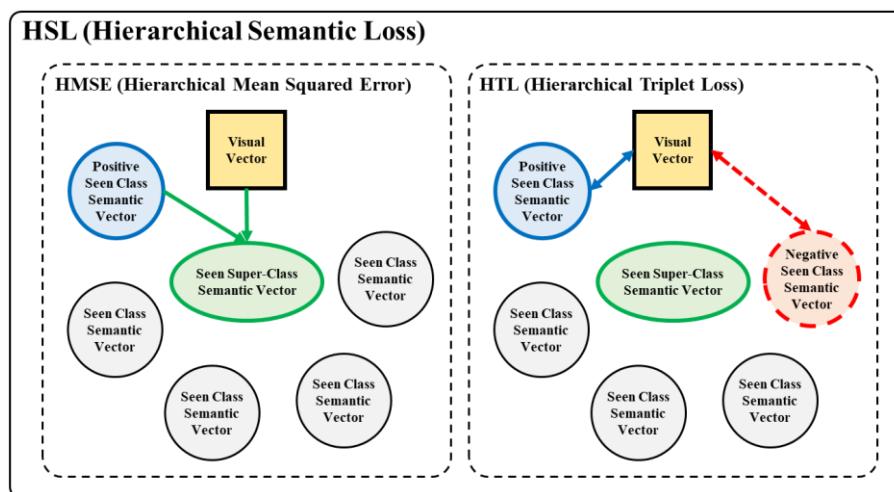
In Equation (4), The HTL is defined by differences between distance of anchor  $x_s$  to positive pair  $y_s^+$  and distance of anchor  $x_s$  to negative pair  $y_s^-$ . If the distance to the  $y_s^-$  from  $x_s$  is greater than the sum of margin  $\alpha$  and distance to  $y_s^+$  from  $x_s$ , the HTL is set to 0 so that learning does not proceed anymore.

In general, the triplet margin loss sets all seen-class semantic vectors as candidates of negative pair except the positive one. In contrast, the proposed method sets only the seen-class semantic vector that has the same super-class with positive visual vectors as candidates of negative pairs, reflecting the hierarchical knowledge of semantic class labels. Through this, the proposed HTL can train embedding networks more effectively for the input visual data.

$$HSL(x_s, y_s^+, y_s^-, y_s^{sc}, \alpha, \lambda) = \lambda * HMSE(x_s, y_s^+, y_s^{sc}) + (1 - \lambda) * HTL(x_s, y_s^+, y_s^-, \alpha, \lambda) \quad (5)$$

$$\underset{\theta_I, \theta_T \in \mathbb{R}}{\operatorname{argmin}} HSL(x_s, y_s^+, y_s^-, y_s^{sc}, \alpha, \lambda) \quad (6)$$

Equation (5) represents the proposed HSL adjusted by learning coefficient  $\lambda$ . For example, the embedding networks might be accelerated faster than the relative differences between the child semantic labels if we assign them a relatively large learning coefficient  $\lambda$ . On the contrary, when the learning coefficient  $\lambda$  is relatively small, we can expect detailed embedding results through hierarchical triplet margin loss. Finally, Equation (6) is a training objective function that minimizes HSL by adjusting weights of image embedding networks  $\theta_I$  and text embedding networks  $\theta_T$ . In this way, the proposed zero-shot learning methodology is different from the existing zero-shot learning methodology in that the loss function is designed using hierarchical knowledge [37]. Figure 2 shows an intuitive description of HSL using a hierarchical structure.



**Figure 2.** Description of hierarchical semantic loss.

The HMSE is defined such that the visual vector and the positive seen-class semantic vector are embedded close to the fixed seen super-class semantic vector. The green solid arrow in HMSE of Figure 2 indicates that the embedding models are trained to embed the visual data and seen-class semantic data to the corresponding direction. By training the embedding model to minimize the distance represented by the green solid arrow, we can embed the seen-class semantic data close to the corresponding visual data.

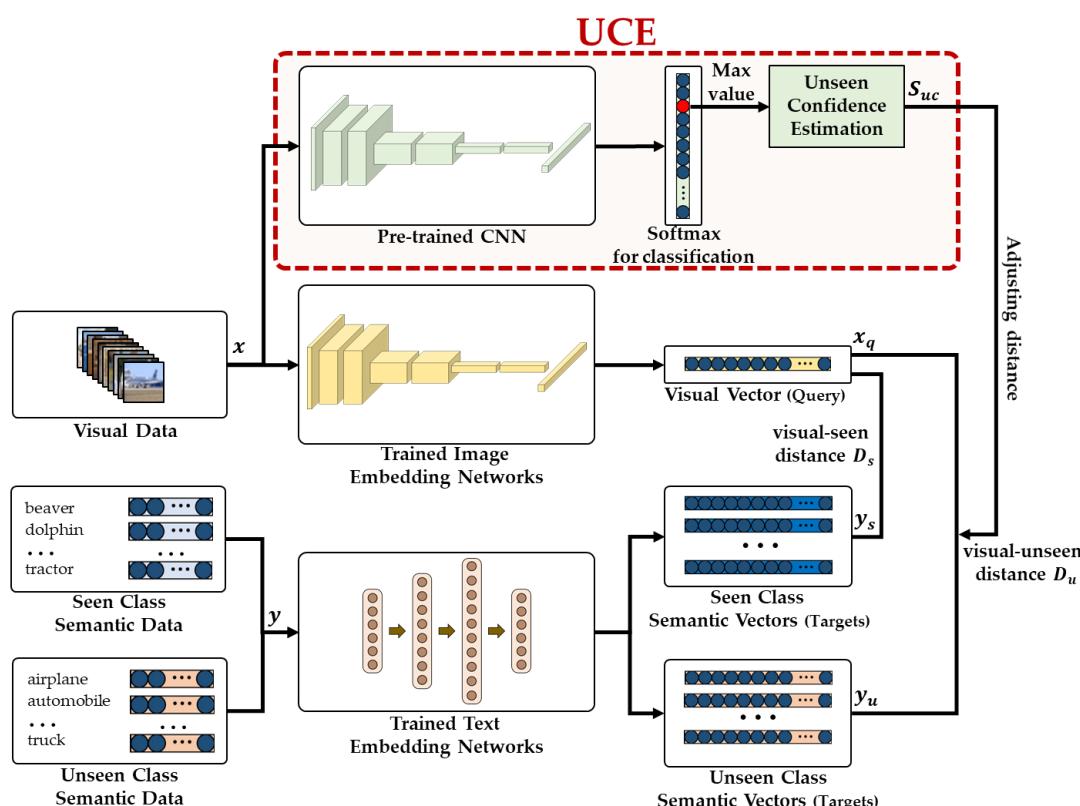
HTL is defined such that the visual vector and a positive seen-class semantic vector are embedded close, and the visual vector and a negative seen-class semantic vector are embedded far away, to distinguish those seen-class semantic vectors under the same super-class. The blue solid bidirectional arrow indicates the distance between a visual vector and a positive seen-class semantic vector become smaller, while the red dotted bidirectional arrow indicates the distance between a visual vector and a negative seen-class semantic vector become larger. Both detailed loss terms require hierarchical information, such as super-class relation.

### 3.2. Zero-Shot Inference with Unseen Confidence Estimator

The zero-shot architecture based on visual-semantic embedding classifies the input visual data mainly through cross-modal retrieval. Typical embedding-based zero-shot learning methods show performance degradation as the number of semantic candidates for inference increases. This is especially as most of the existing methodologies show higher performance degradation when the number of semantic classes for the observed data increases as compared to the case when the number of semantic classes for unseen data increases [7]. In other words, the existing zero-shot learning model tends to embed the unseen visual data into seen semantics. Therefore, a methodology is needed to mitigate the problem of embedding unseen visual data into the semantic of the seen data. A typical approach of this method is to design an estimator that estimates the confidence score to distinguish whether the input data are seen or unseen. Figure 3 shows the structure of the proposed generalized zero-shot

inference architecture and the notations used in the proposed generalized zero-shot inference phase are as follows.

- $x$ : visual data comprise unseen-classes not used in training phase.
- $y$ : semantic data comprise both seen-classes used in training phase and unseen-classes.
- $x_q$ : visual vector embedded from image embedding networks as query of cross-modal retrieval task.
- $y_s$ : seen-class semantic vector embedded from text embedding networks as targets of cross-modal retrieval task.
- $y_u$ : unseen-class semantic vector embedded from text embedding networks as targets of cross-modal retrieval task.
- $D_s$ : distances between visual vector query and seen-class semantic vectors.
- $D_u$ : distances between visual vector query and unseen-class semantic vectors.
- $s_{uc}$ : unseen confidence score from unseen confidence estimator adjusting visual-unseen distance  $D_u$ .



**Figure 3.** Generalized zero-shot inference architecture with unseen confidence estimator.

The proposed generalized zero-shot inference architecture comprises trained image embedding networks, trained text embedding networks, and an unseen confidence estimator (UCE) based on pre-trained convolutional neural networks for the classification of seen class visual data. The proposed architecture embeds both seen-class and unseen-class semantic data into a common vector space through text embedding networks for creating the target seen-class semantic vector  $y_s$  and unseen-class semantic vector  $y_u$ . We can perform zero-shot cross-modal retrieval using a visual vector  $x_q$  from image embedding networks as a query. However, as the embedding model is trained by the seen-classes, zero-shot cross-modal retrieval in this manner is not expected to achieve high performance.

In this paper, we calculate an unseen confidence score  $s_{uc}$ , which indicates whether the input visual data correspond to seen or unseen classes. We design  $s_{uc}$  to have a large value if the maximum value of softmax for the seen-class data is low. Then the  $s_{uc}$  adjusts the distance  $D_u$  between the query  $x_q$  and targets  $y_u$ .

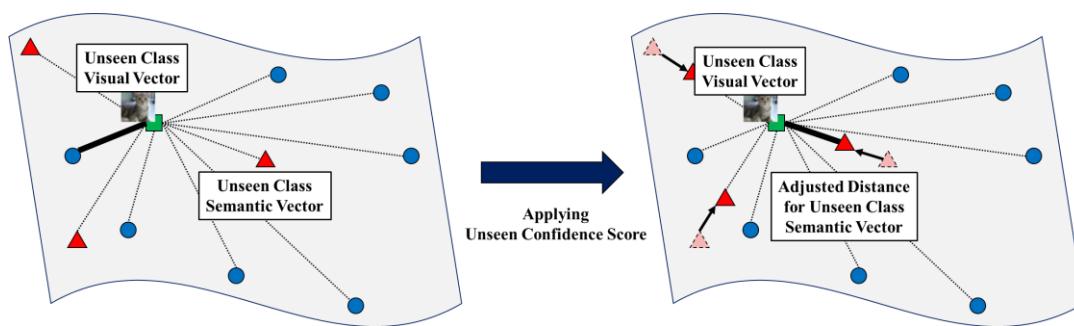
The confidence estimator includes pre-trained convolution neural networks, which perform classification on the seen-class training dataset. The classification result to the confidence estimator is expressed as a type of probability value for each class using the softmax function. The value of the largest softmax selected as the classification result is set as certainty for the seen-class. Using this, we calculate  $s_{uc}$  to apply it to  $D_u$  to make it smaller than  $D_s$  especially when the certainty for the seen class is very low. So, we can perform more efficient zero-shot inference using the unseen confidence score  $s_{uc}$  when unseen visual data are input.

$$s_{uc} = 2 - \max(\text{softmax}\left(\frac{f(x, \theta_c)}{T}\right)) \quad (7)$$

$$D_u(x_q, y_u) = \frac{\|x_q - y_u\|_2^2}{s_{uc}} \quad (8)$$

Equation (7) shows the UCE to calculate the unseen confidence score  $s_{uc}$ . Here,  $x$  denotes the visual data and  $f(\cdot, \theta_c)$  denotes a pre-trained CNN model that classifies all seen-classes in the training dataset. The  $s_{uc}$  represents the confidence of unseen classes. Dividing  $f(\cdot, \theta_c)$  by the temperature constant  $T$  in the Equation (7), the output distribution of softmax can be adjusted according to the purpose of specific tasks [36]. The largest value among the output values of the softmax function generally represents the probability of the classification result, and this paper considers this value as the confidence for the seen-class visual data. As the output value of softmax ranges between 0 and 1, we subtract this value from 2 for controlling the influence of  $s_{uc}$ . It becomes closer to 1 as the softmax value of the pre-trained CNN model approaches to 1 (high confidence of seen class), and closer to 2 as the softmax value approaches to 0 (low confidence of seen class). Therefore, the range of  $s_{uc}$  is designed to be between 1 and 2.

The generalized zero-shot inference using the proposed method is performed by reflecting the characteristics of the input data. Equation (8) shows how to calculate the distance between the query visual vector  $x_q$  and unseen-class semantic vector  $y_u$  by applying  $s_{uc}$ .  $D_u$  is calculated using l2 norm and divided by the  $s_{uc}$  obtained from Equation (7). So, if the  $s_{uc}$  become larger, the distances between  $x_q$  and all unseen semantic vectors become smaller. Figure 4 shows an example where  $s_{uc}$  is applied in the zero-shot inference process.



**Figure 4.** Example of zero-shot inference applying unseen confidence score.

In Figure 4, the blue circles indicate embedded seen-class semantic vectors, the red triangles indicate an embedded unseen-class semantic vectors, and the green square indicates the embedded visual vector. The simplest way to classify labels for visual data is to use the nearest neighbor method. For this purpose, we calculate the distances between the visual vector and all semantic vectors. The left side of Figure 4 shows that the seen-class semantic vector for the unseen-class image is classified as the closest one. If the distances between unseen-class visual vector and unseen-class semantic vectors become relatively small by applying the confidence score as shown on the right side of Figure 4, it helps to perform correct classification.

The proposed methodology adjusts the distances between the visual vector and the unseen-class semantic vector by using the  $s_{uc}$  obtained from UCE. An advantage of the proposed methodology is that the embedding networks and UCE are designed separately. In other words, the proposed methodology only needs to reflect the influence of  $s_{uc}$  through simple multiplication in the inference step without additional training to the embedding networks.

## 4. Experimental Setup

### 4.1. Dataset and Evaluation Metric

We conducted three main experiments. In the first one, we performed zero-shot inference with only unseen-class semantic vectors as targets for the visual vector query. In the second, we performed generalized zero-shot inference with both seen-class and unseen-class semantic vectors. Finally, we performed an intuitive analysis of the proposed methodology by embedding the visual data using the proposed embedding model, using dimension reduction on the visual vectors, and performing visualization.

In the first zero-shot inference experiment, a CIFAR-10 dataset comprising 50,000 images and 10 semantic classes was used [38]. Eight seen-classes were set as seen and the remaining two were set as unseen. The zero-shot classification performance for a total of five zero-shot category sets was measured. In the second generalized zero-shot inference experiment, generalized zero-shot retrieval performance was measured. In this experiment, a CIFAR-100 dataset with 50,000 images and 100 semantic classes was set as seen-classes and a CIFAR-10 dataset with 10,000 images and 10 semantic classes was set as unseen-classes. In the last experiment, we embedded various datasets using trained image embedding networks and text embedding networks and analyzed the visualization results of the embedded visual vectors using t-SNE [39].

In the first experiment, we used manually constructed hierarchical knowledge based on B-CNN [40]. In the second experiment, we used the existing hierarchical semantic label structure of CIFAR-100 dataset. This hierarchy consisted of 20 super-classes, with five classes sharing one super-class. All semantic vectors used in the experiments were 200-dimensional distributed representation vectors obtained from the pre-trained GloVe language model [41]. In the first experiment, zero-shot classification measured zero-shot accuracy to classify two zero-shot categories. In the second experiment, the performance of zero-shot cross-modal retrieval was measured by hit@k.

### 4.2. Network Structures and Training Details

We used three types of neural networks: image embedding networks, text-embedding networks, and pre-trained CNN. First, image embedding networks and confidence estimators were designed as custom convolution neural networks using a small  $3 \times 3$  filter by taking advantage of the structure of VGGNet [42]. Text embedding networks were designed as fully connected neural networks to reflect the characteristics of distributed representation semantic vectors.

Firstly, the pre-trained CNN for UCE was configured to classify 100 seen semantic labels by using the seen training dataset. In contrast, image-embedding networks use convolutional neural networks of the same structure, but the length of the output layer is 200 dimensions, which is designed to embed the visual data into the common space. Both CNNs comprise only global average pooling and one fully connected layer, minimizing the amount of computation required for learning. The text embedding networks consist of a general fully connected layer, which receives 200-dimension GloVe vectors and embeds them into the same common space. Table 1 summarizes the structure of the neural networks used in the experiments.

**Table 1.** Network architecture used in the proposed zero-shot learning model.

Unseen Confidence Estimator			Image Embedding Networks				Text Embedding Networks			
Layer	Weights Shape [W, H, IC, OC]	Feature Map [W, H, C]	Activation	Layer	Weights Shape [W, H, IC, OC]	Feature Map [W, H, C]	Activation	Layer	Weights Shape [IN, OH]	Activation
Input	-	[32, 32, 3]	-	Input	-	[32, 32, 3]	-	Input	[200]	-
Conv1	[3, 3, 3, 64]	[30, 30, 64]	ReLU	Conv1	[3, 3, 3, 64]	[30, 30, 64]	ReLU	FC1	[200, 256]	ReLU
Conv2	[3, 3, 64, 128]	[28, 28, 128]	ReLU	Conv2	[3, 3, 64, 128]	[28, 28, 128]	ReLU	FC2	[256, 512]	ReLU
Max pool	-	[14, 14, 128]	-	Max pool	-	[14, 14, 128]	-	FC3 (output)	[512, 200]	Identity
Conv3	[3, 3, 128, 256]	[12, 12, 256]	ReLU	Conv3	[3, 3, 128, 256]	[12, 12, 256]	ReLU			
Conv4	[3, 3, 256, 512]	[10, 10, 512]	ReLU	Conv4	[3, 3, 256, 512]	[10, 10, 512]	ReLU			
Conv5	[3, 3, 512, 1024]	[8, 8, 1024]	ReLU	Conv5	[3, 3, 512, 1024]	[8, 8, 1024]	ReLU			
GAP	-	[1, 1, 1024]	-	GAP	-	[1, 1, 1024]	-			
FC (output)	[1024, 100]	[100]	Identity	FC	[1024, 200]	[200]	Identity			

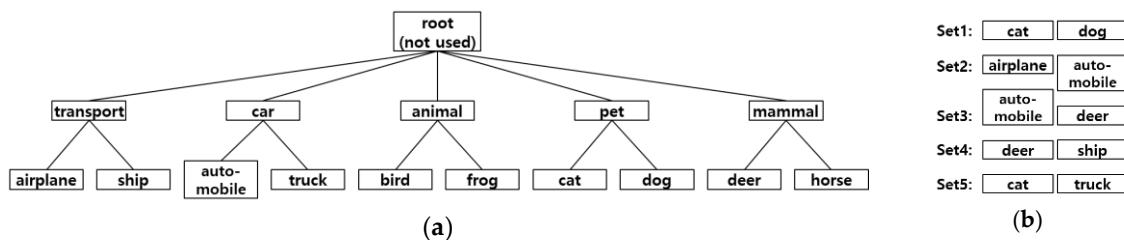
The training was performed for 100 epochs of all neural networks and the learning coefficient  $\lambda$  was set to 0.5, to create HSL in order to have an equal proportion of HMSE and HTL. The distances of all nearest neighbor methods used in the experiment were l2 norm. In the visualization experiment, a high-dimensional vector of 200 dimensions was reduced to two dimensions for visualization.

## 5. Experimental Results and Discussion

### 5.1. Zero-Shot Cross-Modal Retrieval Results

In the zero-shot learning experiment, we compared the performance of the proposed methodology with the baseline of the cross-modal transfer (CMT) methodology, which is one of the representative state-of-art methods using the CIFAR-10 dataset [6]. CMT selected two classes and created a zero-shot category of unseen-classes. In this process, the remaining eight labels were used for training as seen-classes, and an experiment was performed to classify the two selected unseen semantic labels. We used 38 k images in eight seen categories in the training model. In the test phase, 12 k images in two zero-shot (unseen) categories were used for zero-shot cross-modal retrieval.

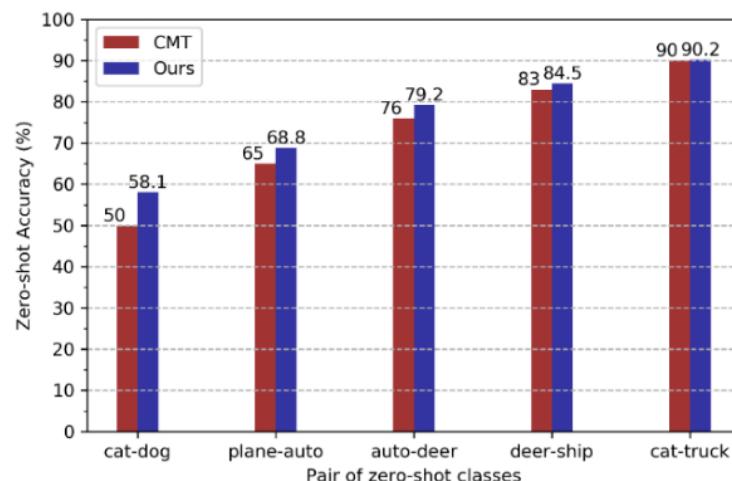
To conduct the zero-shot learning experiment, the hierarchical structure of the semantic classes was designed in the proposed methodology. The hierarchical structure was manually constructed based on B-CNN, which is one of the prior studies using hierarchical information of CIFAR-10 [40]. Figure 5a shows the CIFAR-10 hierarchical knowledge used in the experiment, which is constructed so that two child classes with high semantic similarity can share one super-class. Figure 5b shows the zero-shot category sets. Set1 is composed of a “cat-dog”, which has very high visual similarity and faces difficulty in getting help from other labels. In contrast, set5 consists of a “cat-truck”, which has different visual similarity, and might achieve relatively high zero-shot classification performance with the help of other semantic labels.



**Figure 5.** CIFAR-10 hierarchy knowledge and zero-shot category. (a) Cifar10 manually constructed hierarchy. (b) Zero-shot category sets.

Figure 6 compares the accuracy of classification of two zero-shot categories of CIFAR-10 using CMT, one of the existing state-of-arts methodologies, and the proposed methodology. The left dark red bar of each category represents the performance of CMT, and the right dark blue bar represents

the performance of the proposed methodology. It shows that the performance of our methodology is better than that of CMT for all zero-shot category sets. There is no large difference in sets such as “deer-ship” and “cat-truck,” which is relatively easy, but a relatively large difference in sets such as “cat-dog” and “airplane-automobile.” These experimental results confirm that the proposed HSL can make more efficient zero-shot cross-modal retrieval.



**Figure 6.** Hit@1 zero-shot cross-modal retrieval performance for five zero-shot category sets in CIFAR-10.

## 5.2. Generalized Zero-Shot Cross-Modal Retrieval Results

In the zero-shot learning experiment using CIFAR-10 dataset, we already showed how well the proposed model performs zero-shot cross-modal retrieval in Figure 6. However, to apply the proposed methodology in a more expandable manner, generalized zero-shot inference should be evaluated. The previous research has shown that the performance of zero-shot inference is significantly degraded in generalized case [7]. One way to improve the performance of generalized zero-shot learning is to determine if the input data is seen class or unseen class. In this experiment, we compare the performances of zero-shot model using HSL and zero-shot model using both HSL and UCE. In experiments, we use CIFAR-100 dataset as seen classes and CIFAR-10 dataset as unseen classes. The unseen semantic classes of CIFAR-10 used in the experiment are airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck.

Table 2 shows that the proposed HSL+UCE model outperforms the basic HSL model in recognizing unseen class visual data. The HSL model uses HTL and HMSE in training the embedding model by using seen class data only. Naturally, the generalized zero-shot accuracy is low. It is difficult to retrieve the correct label for unseen class visual data among 110 classes consist of 100 seen-classes and 10 unseen-classes. In contrast, the HSL+UCE model has an additional UCE module for adjusting the distances between query visual vector and unseen class semantic vectors when the confidence of unseen class is high, and it improves the performance of generalized zero-shot cross-modal retrieval.

**Table 2.** Hit@k performance on generalized zero-shot cross-modal retrieval for unseen class visual data in CIFAR-10.

Model (Semantic Candidates)	hit@1	hit@2	hit@3	hit@4	hit@5	hit@6	hit@7	hit@8	hit@9	hit@10
HSL (100 Seen Classes & 10 Unseen Classes)	0.03	0.09	0.15	0.15	0.20	0.25	0.33	0.37	0.41	0.45
HSL+UCE (100 Seen Classes & 10 Unseen Classes)	0.14	0.28	0.40	0.50	0.58	0.64	0.69	0.73	0.76	0.78

In the hit@1 of Table 2, the basic HSL shows a hit@1 performance of only 3.5%, while the HSL+UCE shows a performance of 13.8%. In hit@5, the basic HSL shows 19.9%, while the HSL+UCE shows 58% performance. The proposed methodology shows that the performance of generalized zero-shot cross-modal retrieval can be improved 3–4 times by applying UCE.

Table 3 shows the actual images and results of the retrieved semantic labels used in this experiment. The first column in Table 3 shows the input query image belonging to the unseen class, the second column shows retrieved labels using the HSL model only, and the third column shows retrieved labels using the HSL+UCE model. The retrieved labels are arranged in order from top-1 to top-5, with predicted true labels highlighted in bold and underlined.

**Table 3.** Retrieved semantic labels for unseen class image query.

Image Query	HSL	HSL+UCE
 (a) truck	<ul style="list-style-type: none"> <li>● train</li> <li>● streetcar</li> <li>● bus</li> <li>● house</li> <li>● tractor</li> </ul>	<ul style="list-style-type: none"> <li>✓ <b><u>truck (unseen)</u></b></li> <li>● train</li> <li>● airplane (unseen)</li> <li>● streetcar</li> <li>● automobile (unseen)</li> </ul>
 (b) horse	<ul style="list-style-type: none"> <li>● cattle</li> <li>● fox</li> <li>● lion</li> <li>● kangaroo</li> <li>● camel</li> </ul>	<ul style="list-style-type: none"> <li>● cattle</li> <li>● fox</li> <li>● lion</li> <li>✓ <b><u>horse (unseen)</u></b></li> <li>● kangaroo</li> </ul>
 (c) ship	<ul style="list-style-type: none"> <li>● castle</li> <li>✓ <b><u>ship (unseen)</u></b></li> <li>● bridge</li> <li>● skyscraper</li> <li>● rocket</li> </ul>	<ul style="list-style-type: none"> <li>✓ <b><u>ship (unseen)</u></b></li> <li>● airplane (unseen)</li> <li>● bird(unseen)</li> <li>● automobile (unseen)</li> <li>● horse (unseen)</li> </ul>

Each row with a different image query shows three different cases of generalized zero-shot learning using the proposed methodology. First, in the case of the truck image as input Table 3 (a), the HSL model did not retrieve the true label but retrieved the other labels in seen classes. In contrast, the HSL+UCE model retrieved true label in unseen classes as top-1. The other retrieved results including both seen and unseen classes are also semantically similar to each other in that they are person-made objects related to transportation. In this way, we can see a practical example of generalized zero-shot learning complemented by the UCE, which is difficult to solve with the HSL model only.

Next, in the second row with input Table 3 (b), while the HSL model could not properly predict the true label, the HSL+UCE model predicted the true label as top-4. This is an improvement over the results of the existing HSL model only, but the UCE is less influential than the example in Table 3 (a) and the true label is not adequately retrieved as top-1. The example of Table 3 (c) shows the result of correctly retrieving the true label of the unseen classes as top-1 through the HSL+UCE model. In contrast to the example of Table 3 (b), since all retrieved labels from top-1 to top-5 are in the unseen classes, we can surmise that the influence of UCE is over-applied, which shows the limitation that the seen classes cannot be retrieved properly. So, even though this is not the usual generalized zero-shot learning evaluation setting, we need to check the degradation of general cross-modal retrieval for seen class data.

### 5.3. Degradation of Cross-Modal Retrieval for Seen Class Visual Data

When UCE considers all input data as unseen classes, the performance of generalized zero-shot cross-modal retrieval is increasing but the performance of cross-modal retrieval for seen class input query may be decreasing. The performance evaluation on seen class data can measure the influence of UCE. So, the performance of cross-modal retrieval for seen class visual data is evaluated by using the CIFAR-100 dataset (seen classes) with the same model used in the previous generalized zero-shot cross-modal retrieval experiments. Table 4 shows the performance evaluation results for visual-semantic cross-modal retrieval for 10k CIFAR-100 test dataset (seen classes).

**Table 4.** Hit@k performance on visual-semantic cross-modal retrieval for seen class visual data in CIFAR-100.

Model (Semantic Candidates)	hit@1	hit@2	hit@3	hit@4	hit@5	hit@6	hit@7	hit@8	hit@9	hit@10
HSL (100 Seen Classes & 10 Unseen Classes)	0.53	0.65	0.70	0.75	0.77	0.79	0.81	0.82	0.84	0.85
HSL+UCE (100 Seen Classes & 10 Unseen Classes)	0.44	0.53	0.59	0.63	0.66	0.69	0.72	0.74	0.76	0.78

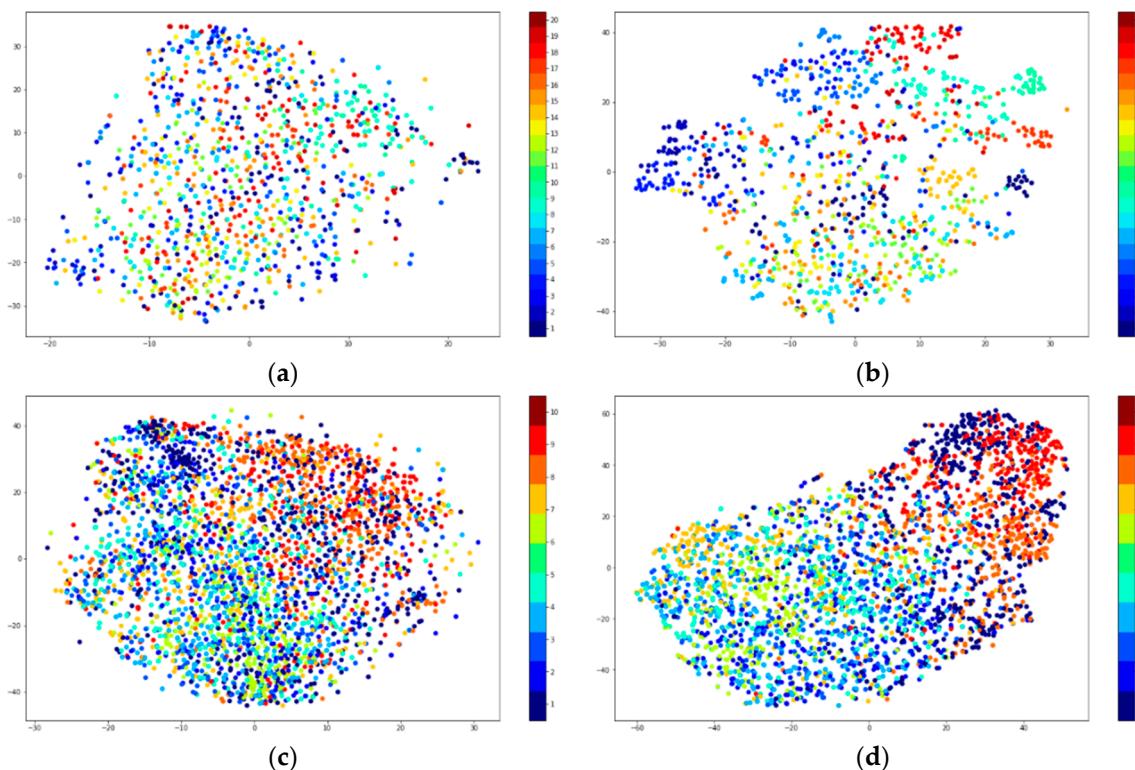
In this setting, the performance of HSL+UCE model in retrieving correct label for seen class visual data among 110 seen and unseen classes is somewhat lower than that of HSL model. The reason is that the unseen confidence score is applied so that it prefers the unseen semantic vectors. Performance improvements on the seen data are future research topics for generalized zero-shot learning. Nevertheless, considering the trade-off with the performance improvement of zero-shot learning obtained in Table 2, it can be considered an acceptable performance degradation. In terms of performance change ratio, the performance on unseen data increases by about 200–300% due to the addition of UCE, while the performance on seen data decreases by about 10–20%.

### 5.4. Visualization of Embedded Visual Vectors

A high-dimensional vector is not easy to compare intuitively. In the case of the visual or semantic vectors experimented here, it is difficult to understand how the vector is embedded as a 200-dimensional vector. Therefore, by using the t-SNE dimension reduction methodology, the vectors of 200 dimensions are reduced to two dimensions while maintaining their relative distances and are visualized as shown in Figure 7.

Figure 7 shows the visualization results using the CIFAR-100 test dataset (seen-classes) and CIFAR-10 test dataset (unseen-classes). For CIFAR-100, the color is expressed based on the 20 super-classes. For CIFAR-10, the color is expressed based on the classes. The left column of Figure 7 visualizes the raw pixel data of the images as it is, while the right column visualizes the embedded visual vectors through the proposed embedding model.

In Figure 7a,c which are visualization of raw images, it is difficult to find the relationship between visual vectors and each class. In contrast, Figure 7b,d which are visualization of embedded vectors from proposed embedding model, shows that visual vectors are embedded more cohesively according to their classes. As shown in Figure 7b, which is the visualization of the embedded visual vectors from CIFAR-100 test images, the embedding occurs in the meaningful manner per super-classes. This means that the embedding with proposed model performs relatively well on the seen-classes. In Figure 7d, the visualization of embedded visual vectors from CIFAR-10 test dataset is relatively complicated compared to Figure 7b, but still a little more meaningful than Figure 7c, the raw images. Through this visualization, we can get some intuition for the zero-shot learning using the proposed model.



**Figure 7.** Visualization of raw pixel image and embedded visual vectors. (a) Seen class images in CIFAR-100 test dataset (N: 1000). (b) Seen class embedded vectors of CIFAR-100 test dataset (N: 1000). (c) Unseen class images in CIFAR-10 test dataset (N: 3000). (d) Unseen class embedded vectors of CIFAR-10 test dataset (N: 3000).

## 6. Conclusions

While supervised learning can be used to classify data into labels in training data, zero-shot learning can classify unseen-classes that are not used in the training phase. One of the major approaches to zero-shot learning is embedding-based zero-shot learning. This paper proposes HSL and an unseen confidence estimator for more efficient embedding-based zero-shot learning. HSL uses a hierarchical mean squared error, which uses hierarchical knowledge of labels in the seen dataset for training visual-semantic embedding networks. In addition, when choosing a negative sample of triplet margin loss, we designed the loss function more efficiently using the hierarchical knowledge of training data. The confidence estimator estimates the degree to which the input data belongs to an unseen label that is not used during the training phase and improves the performance of the model by weighting it when performing zero-shot cross-modal retrieval. For the proposed methodology, quantitative experiments using CIFAR-100 and CIFAR-10 dataset are performed, and the proposed model shows better performance than the baseline model. In addition, visualization of the embedded visual vector is performed to confirm the effectiveness of the proposed methodology.

To improve the performance of zero-shot learning, it is important to use various side information related to visual data. In future research, we plan to study a zero-shot learning model which uses hierarchical knowledge more directly in training. Also, the research is needed to measure and apply unseen confidence scores more precisely to improve the performance of the generalized zero-shot learning.

**Author Contributions:** Writing—original draft preparation, S.S.; writing—review and editing, J.K.

**Funding:** This research was funded by the Ministry of Science, ICT, Republic of Korea, grant number (NRF-2017M3C4A7083279).

**Acknowledgments:** This research was supported by the Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT (NRF-2017M3C4A7083279).

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

## References

1. 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 Computer Vision and Pattern Recognition 2009, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
2. 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 2014, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
3. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [[CrossRef](#)]
4. Ravi, S.; Larochelle, H. Optimization as a Model for Few-Shot Learning. In Proceedings of the 5th the International Conference on Learning Representations, Toulon, France, 24–26 April 2017; pp. 1–11.
5. Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D. Matching Networks for One Shot Learning. In Proceedings of the Advances in Neural Information Processing Systems 2016, Barcelona, Spain, 5–10 December 2016; pp. 3630–3638.
6. 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 2013, Lake Tahoe, NV, USA, 5–10 December 2013; pp. 935–943.
7. Frome, A.; Corrado, G.S.; Shlens, J.; Bengio, S.; Dean, J.; Mikolov, T. Devise: A Deep Visual-Semantic Embedding Model. In Proceedings of the Advances in Neural Information Processing Systems 2013, Lake Tahoe, NV, USA, 5–10 December 2013; pp. 2121–2129.
8. Han, H.; Wang, W.Y.; Mao, B.H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In *International Conference on Intelligent Computing*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 878–887.
9. Chawla, N.V.; Japkowicz, N.; Kotcz, A. Special issue on learning from imbalanced data sets. *ACM Sigkdd Explor. Newsl.* **2004**, *6*, 1–6. [[CrossRef](#)]
10. Ampert, C.H.; Nickisch, H.; Harmeling, S. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 453–465. [[CrossRef](#)] [[PubMed](#)]
11. 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.
12. Zhang, Z.; Saligrama, V. Zero-shot Learning via Semantic Similarity Embedding. In Proceedings of the IEEE International Conference on Computer Vision 2015, Santiago, Chile, 7–13 December 2015; pp. 4166–4174.
13. 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 2015, Boston, MA, USA, 7–12 June 2015; pp. 2927–2936.
14. 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 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 69–77.
15. Romera-Paredes, B.; Torr, P. An Embarrassingly Simple Approach to Zero-Shot Learning. In Proceedings of the International Conference on Machine Learning 2015, Lille, France, 6–11 July 2015; pp. 2152–2161.
16. Akata, Z.; Perronnin, F.; Harchaoui, Z.; Schmid, C. Label-embedding for image classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1425–1438. [[CrossRef](#)] [[PubMed](#)]
17. 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 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 5327–5336.
18. Kodirov, E.; Xiang, T.; Gong, S. Semantic autoencoder for zero-shot learning. *arXiv* **2017**, arXiv:1704.08345.

19. 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 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 3010–3019.
20. 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**. [[CrossRef](#)] [[PubMed](#)]
21. Chao, W.L.; Changpinyo, S.; Gong, B.; Sha, F. An Empirical Study and Analysis of Generalized Zero-Shot Learning for Object Recognition in the Wild. In Proceedings of the European Conference on Computer Vision 2016, Amsterdam, The Netherland, 11–14 October 2016; pp. 52–68.
22. Aytar, Y.; Vondrick, C.; Torralba, A. See, hear, and read: Deep aligned representations. *arXiv* **2017**, arXiv:1706.00932.
23. Ge, W.; Huang, W.; Dong, D.; Scott, M.R. Deep Metric Learning with Hierarchical Triplet Loss. In Proceedings of the European Conference on Computer Vision 2018, Munich, Germany, 8–14 September 2018; pp. 272–288.
24. Annadani, Y.; Biswas, S. Preserving Semantic Relations for Zero-Shot Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7603–7612.
25. Lei Ba, J.; Swersky, K.; Fidler, S. Predicting Deep Zero-Shot Convolutional Neural Networks using Textual Descriptions. In Proceedings of the IEEE International Conference on Computer Vision 2015, Santiago, Chile, 7–13 December 2015; pp. 4247–4255.
26. Zhang, Z.; Saligrama, V. Zero-shot Learning via Joint Latent Similarity Embedding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 6034–6042.
27. Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P.H.; Hospedales, T.M. Learning to Compare: Relation Network for Few-Shot Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1199–1208.
28. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A Unified Embedding for Face Recognition and Clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
29. Chang, S.; Han, W.; Tang, J.; Qi, G.J.; Aggarwal, C.C.; Huang, T.S. Heterogeneous Network Embedding via Deep Architectures. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2015, Sydney, Australia, 10–13 August 2015; pp. 119–128.
30. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 1050–1059.
31. Hendrycks, D.; Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv* **2016**, arXiv:1610.02136.
32. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. *arXiv* **2017**, arXiv:1706.04599.
33. Lee, K.; Lee, H.; Lee, K.; Shin, J. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv* **2017**, arXiv:1711.09325.
34. Lee, K.; Lee, K.; Min, K.; Zhang, Y.; Shin, J.; Lee, H. Hierarchical Novelty Detection for Visual Object Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1034–1042.
35. Miller, G.A. WordNet: A lexical database for English. *Commun. ACM* **1995**, *38*, 39–41. [[CrossRef](#)]
36. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
37. Li, X.; Liao, S.; Lan, W.; Du, X.; Yang, G. Zero-shot image tagging by Hierarchical Semantic Embedding. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval 2018, Santiago, Chile, 9–13 August 2018; pp. 879–882.
38. Krizhevsky, A.; Hinton, G. *Learning Multiple Layers of Features from tiny Images*; Technical Report; University of Toronto: Toronto, ON, Canada, 2009; Volume 1, pp. 32–35.
39. Maaten, L.V.D.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
40. Zhu, X.; Bain, M. B-CNN: Branch convolutional neural network for hierarchical classification. *arXiv* **2017**, arXiv:1709.09890.

41. 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, Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
42. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).