Exploring Hierarchical Graph Representation for Large-Scale Zero-Shot Image Classification

Kai Yi[1](✉), Xiaoqian Shen[1], Yunhao Gou[1,2], and Mohamed Elhoseiny[1]

[1] King Abdullah University of Science and Technology (KAUST),
Thuwal, Saudi Arabia
{kai.yi,xiaoqian.shen,yunhao.gou,mohamed.elhoseiny}@kaust.edu.sa
[2] University of Electronic Science and Technology of China (UESTC), Sichuan, China

Abstract. The main question we address in this paper is how to scale up visual recognition of unseen classes, also known as zero-shot learning, to tens of thousands of categories as in the ImageNet-21K benchmark. At this scale, especially with many fine-grained categories included in ImageNet-21K, it is critical to learn quality visual semantic representations that are discriminative enough to recognize unseen classes and distinguish them from seen ones. We propose a *H*ierarchical *G*raphical knowledge *R*epresentation framework for the confidence-based classification method, dubbed as HGR-Net. Our experimental results demonstrate that HGR-Net can grasp class inheritance relations by utilizing hierarchical conceptual knowledge. Our method significantly outperformed all existing techniques, boosting the performance by 7% compared to the runner-up approach on the ImageNet-21K benchmark. We show that HGR-Net is learning-efficient in few-shot scenarios. We also analyzed our method on smaller datasets like ImageNet-21K-P, 2-hops and 3-hops, demonstrating its generalization ability. Our benchmark and code are available at https://kaiyi.me/p/hgrnet.html.

Keywords: Zero-shot learning · Semantic hierarchical graph · Large-scale knowledge transfer · Vision and language

1 Introduction

Zero-Shot Learning (ZSL) is the task of recognizing images from unseen categories with the model trained only on seen classes. Nowadays, ZSL relies on semantic information to classify images of unseen categories and can be formulated as a visual semantic understanding problem. In other words, given candidate text descriptions of a class that has not been seen during training, the goal

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-20044-1_7.

S. Avidan et al. (Eds.): ECCV 2022, LNCS 13680, pp. 116–132, 2022.
https://doi.org/10.1007/978-3-031-20044-1_7

is to identify images of that unseen class and distinguish them from seen ones and other unseen classes based on their text descriptions.

In general, current datasets contain two commonly used semantic information including attribute descriptions (e.g., AWA2 [35], SUN [22], and CUB [34]), and more challenging unstructured text descriptions (e.g., CUB-wiki[5], NAB-wiki [6]). However, these datasets are all small or medium-size with up to a few hundred classes, leaving a significant gap to study generalization at a realistic scale. In this paper, we focus on large-scale zero-shot image classification. More specifically, we explore the learning limits of a model trained from 1K seen classes and transfer it to recognize more than 10 million images from 21K unseen candidate categories from ImageNet-21K [4], which is the largest available image classification dataset to the best of our knowledge.

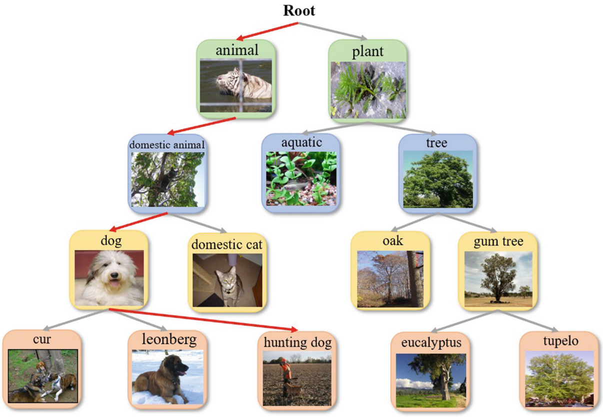


Fig. 1. Intuitive illustration of our proposed HGR-Net. Suppose the ground truth is `Hunting Dog`, then we can find the real-label path `Root` → `Animal` → `Domestic Animal` → `Dog` → `Hunting Dog`. Our goal is to efficiently leverage semantic hierarchical information to help better understand the visual-language pairs.

A few works of literature explored zero-shot image classification on ImageNet-21K. However, the performance has plateaued to a few percent Hit@1 performances on ImageNet-21K zero-shot classification benchmark ([7,13,20,32]). We believe the key challenge is distinguishing among 21K highly fine-grained classes. These methods represents class information by GloVe [23] or Skip-Gram [17] to align the vision-language relationships. However, these lower-dimensional features from GloVe or Skip-Gram are not representative enough to distinguish among 21K classes, especially since they may collapse for fine-grained classes. Besides, most existing works train a held-out classifier to categorize images of unseen classes with different initialization schemes. One used strategy is to initialize the classifier weights with semantic attributes [12,27,36,38], while another is to conduct fully-supervised training with generated unseen images. However, the trained MLP-like classifier is not representative enough to capture fine-grained differences to classify the image into a class with high confidence.

To resolve the challenge of large-scale zero-shot image classification, we proposed a novel *H*ierarchical *G*raph knowledge *R*epresentation network (denoted as HGR-Net). We explore the conceptual knowledge among classes to prompt the distinguishability. In Fig. 1, we state the intuition of our proposed method. Suppose the annotated image class label is `Hunting Dog`. The most straightforward way is to extract the semantic feature and train the classifier with cross-entropy. However, our experiments find that better leveraging hierarchical conceptual knowledge is important to learn discriminative text representation. We know the label as `Hunting Dog`, but all the labels from the root can also be regarded as the real label. We incorporate conceptual semantic knowledge to enhance the network representation.

Moreover, inspired by the recent success of pre-trained models from large vision-language pairs such as CLIP [24] and ALIGN [9], we adopt a dynamic confidence-based classification scheme, which means we multiply a particular image feature with candidate text features and then select the most confident one as the predicted label. Unlike traditional softmax-based classifier, this setting is dynamic, and no need to train a particular classifier for each task. Besides, the confidence-based scheme can help truly evaluate the vision-language relationship understanding ability. For better semantic representation, we adopt Transformer [29] as the feature extractor, and follow-up experiments show Transformer-based text encoder can significantly boost the classification performance.

Contributions. We consider the most challenging large-scale zero-shot image classification task on ImageNet-21K and proposed a novel hierarchical graph representation network, HGR-Net, to model the visual-semantic relationship between seen and unseen classes. Incorporated with a confidence-based learning scheme and a Transformers to represent class semantic information, we show that HGR-Net achieved new state-of-the-art performance with significantly better results than baselines. We also conducted few-shot evaluations of HGR, and we found our method can learn very efficiently by accessing only one example per class. We also conducted extensive experiments on the variants of ImageNet-21K, and the results demonstrate the effectiveness of our HGR-Net. To better align with our problem, we also proposed novel matrices to reflect the conceptual learning ability of different models.

2 Related Work

2.1 Zero-/Few-Shot Learning

Zero-Shot Learning (ZSL) is recognizing images of unseen categories. Our work is more related to semantic-based methods, which learn an alignment between different modalities (i.e., visual and semantic modalities) to facilitate classification [12,27,36,38]. CNZSL [27] proposed to map attributes into the visual space by normalization over classes. In contrast to [27], we map both the semantic text and the images into a common space and calculate the confidence. Experimental studies are conducted to show that mapping to a common space achieves higher accuracy. We also explore the Few-Shot Learning (FSL) task, which focuses on classification with only accessing a few testing examples during training [28,37,39]. Unlike [33] which defines the FSL task as extracting few training data from all classes, we took all images from seen classes and selected only a few samples from unseen classes during training. Our main goal here is to analyze how the performance differs from zero to one-shot.

2.2 Large-Scale Graphical Zero-Shot Learning

Graphical Neural Networks [11] are widely applied to formulate zero-shot learning, where each class is associated with a graph node, and a graph edge represents each inter-class relationship. For example, [32] trains a GNN based on the WordNet knowledge to generate classifiers for unseen classes. Similarly, [10] uses fewer convolutional layers but one additional dense connection layer to propagate features towards distant nodes for the same graph. More recently, [19] adopts a transformer graph convolutional network (TrGCN) for generating class representations. [31] leverages additional neighbor information in the graph with a contrastive objective. Unlike these methods, our method utilizes fruitful information of a hierarchical structure based on class confidence and thus grasps hierarchical relationships among classes to distinguish hard negatives. Besides, some works exploit graphical knowledge without explicitly training a GNN. For example, [15] employs semantic vectors of the class names using multidimensional scaling (MDS) [3] on the WordNet to learn a joint visual-semantic embedding for classification; [12] learns similarity between the image representation and the class representations in the hyperbolic space.

2.3 Visual Representation Learning from Semantic Supervision

Visual representation learning is a challenging task and has been widely studied with supervised or self-supervised methods. Considering semantic supervision from large-scale unlabeled data, learning visual representation from text representation [24] is a promising research topic with the benefit of large-scale visual and linguistic pairs collected from the Internet. These methods train a separate encoder for each modality (i.e., visual and language), allowing for extended to unseen classes for zero-shot learning. Upon these methods, [2] improves the data efficiency during training, [9] enables learning from larger-scale noisy image-text pairs, [40] optimizes the language prompts for better classifier generation. Our work adopts the pre-trained encoders of [24] but tackles the problem of large-scale zero-shot classification from a candidate set of 22K classes instead of at most 1K as in [24].

3 Method

3.1 Problem Definition

Zero-Shot Learning. Let $\mathcal{C}$ denote the set of all classes. $\mathcal{C}_s$ and $\mathcal{C}_u$ to be the unseen and seen classes, respectively, where $\mathcal{C}_s \cap \mathcal{C}_u = \emptyset$, and $\mathcal{C} = \mathcal{C}_s \cup \mathcal{C}_u$. For each class $c_i \in \mathcal{C}$, a d-dimensional semantic representation vector $t(c_i) \in \mathbb{R}^d$ is provided. We denote the training set $\mathcal{D}_{tr} = \{(\mathbf{x}_i, c_i, t(c_i))\}_{i=1}^N$, where $\mathbf{x}_i$ is the i-th training image. In ZSL setting, given testing images $\mathbf{x}_{te}$, we aim at learning a mapping function $\mathbf{x}_{te} \rightarrow \mathcal{C}_u$. In a more challenging setting, dubbed as generalized ZSL, we not only aim at classifying images from unseen categories but also seen categories, where we learn $\mathbf{x}_{te} \rightarrow \mathcal{C}_u \cup \mathcal{C}_s$ covering the entire prediction space.

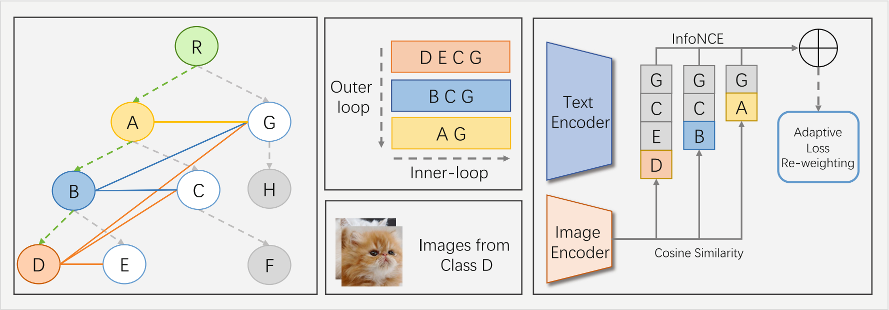


Fig. 2. HGR-Net: Suppose the annotated single label is D and we can find the tracked label path $\mathtt{R} \cdots \rightarrow \mathtt{A} \rightarrow \mathtt{B} \rightarrow \mathtt{D}$ from the semantic graph extended from WordNet. We first set D as the positive anchor and contrast with negatives which are sampled siblings of its ancestors (i.e., $\{\mathtt{E}, \mathtt{C}, \mathtt{G}\}$) layer by layer. Then we iterate to set the positive anchor to be controlled depth as $\mathtt{B}, \mathtt{A}$, which has layer-by-layer negatives $\{\mathtt{C}, \mathtt{G}\}$ and G, respectively. Finally, we use a memory-efficient adaptive re-weighting strategy to fuse knowledge from different conceptual level.

Semantic Hierarchical Structure. We assume access to a semantic Directed Acyclic Graph (DAG), $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \mathcal{C} \cup \{R\}$ and $\mathcal{E} \subseteq \big\{(x, y) \mid (x, y) \in \mathcal{C}^2,\ x \neq y\big\}$. Here the two-tuple (x, y) represents an parenting relationship between x and y, which means y is a more abstract concept than x. Here we manually add a root node R with a in-degree of 0 into $\mathcal{G}$. For simplicity, given any node c_i, we denote the ordered set of all its ancestors obtained by shortest path search from R to c_i as $\mathcal{A}^{c_i} = \{a(c_i)_j\}_{j=1}^{N_i^a} \subseteq \mathcal{C}$. Similarly, we denote the set of all siblings of c_i as $\mathcal{S}^{c_i} = \{s(c_i)_j\}_{j=1}^{N_i^s} \subseteq \mathcal{C}$. Finally, $d(c_i) \triangleq |\mathcal{A}^{c_i}|$ is defined as depth of node c_i.

3.2 HGR-Net: Large-Scale ZSL with Hierarchical Graph Representation Learning

We mainly focus on zero-shot learning on the variants of ImageNet-21K, the current largest image classification dataset to our knowledge. Previous strategies [7,13,20,32] adopt a N-way classification as the training task on all the N seen classes. However, we argue that this is problematic, especially in using a Transformer as the text encoder to obtain class semantic representations.

First, when conducting N-way classification, all the classes except the single ground-truth are regarded as negative ones. Even though this helps build a well-performing classifier in a fully-supervised learning scenario, we argue that this is harmful to knowledge transfer from seen to unseen classes in ZSL. Second, a batch of N samples is fed to the text Transformer [29] to obtain their corresponding text representations and to compute the class logits afterward. This strategy can be acceptable for datasets with a small number of classes. However, when the number of classes scales to tens of thousands, as in our case, it becomes formidable to implement the operations mentioned above. Therefore, we propose a memory-efficient hierarchical contrastive objective to learn transferable and discriminative representations for ZSL.

Intuitively, as illustrated in Fig. 2, suppose we have an image sample with annotated ground-truth label D according to ImageNet-21K. Then, we could find a shortest path $\mathtt{R} - \cdots \rightarrow \mathtt{A} \rightarrow \mathtt{B} \rightarrow \mathtt{D}$ to be the tracked true-label path $\mathcal{T}_{\mathtt{E}}$. With our definition of the hierarchical structure, the true labels for the image sample are defined by all the nodes along this path through different levels of conceptions, from abstract to concrete in our case. Therefore, to better leverage this hierarchical semantic knowledge, we propose a hierarchical contrastive loss that conducts two levels of pre-defined degrees of bottom-up contrasting.

Specifically, for node D with a depth of $d(\mathtt{D})$. In the outer-level loop, we iterate ground-truth labels of different levels, along the ancestor path $\mathcal{A}^{\mathtt{D}}$, we traverse from itself bottom-up $\mathtt{D} - \cdots \rightarrow \mathtt{B} \rightarrow \mathtt{A}$ until reaching one of its ancestors of a depth of $Kd(\mathtt{D})$, where K is the outer ratio. In the inner-level loop, fixing the ground-truth label, we conduct InfoNCE loss [21] layer by layer in a similar bottom-up strategy with an inner ratio M (e.g., when fixing current ground truth node as B in Fig. 2, for inner loop we consider $\langle \mathtt{B}, \mathtt{C} \rangle, \langle \mathtt{B}, \mathtt{G} \rangle$). We provide more details in Algorthim 1.

Formally, given an image of $\boldsymbol{x}$ from class c_i, we define its loss as:

$$\mathcal{L}_{\text{cont}} = \sum_{j=k_s}^{k_e} g(j, \mathcal{L}_j), \quad \mathcal{L}_j = \frac{1}{m_e - m_s + 1} \sum_{l=m_s}^{m_e} \mathcal{L}_{j,l}, \tag{1}$$

where $g(\cdot,\cdot)$ is an adaptive attention layer to dynamically re-weight the importance of labels given different levels j, $j \in [k_s, k_e]$ and $l \in [m_s, m_e]$ are the outer-level and inner-level loop respectively. k_s, k_e represents the start layer and the end layer for outer loop while m_s, m_e are the start layer and the end layer for the inner loop.

$$\mathcal{L}_{j,l} = -\log \frac{\text{pos}^j}{\text{pos}^j + \text{neg}^{j,l}}, \tag{2}$$

where

$$\text{pos}^{\text{j}} = \exp\left(\text{sim}\left(T(\boldsymbol{c}_j^+), V(\boldsymbol{x})\right)/\tau\right) \tag{3}$$

$$\text{neg}^{j,l} = \sum_{q=1}^{n_l} \exp\left(\text{sim}\left(T(\boldsymbol{c}_{j,l,q}^-), V(\boldsymbol{x})\right)/\tau\right) \tag{4}$$

where, sim$(\cdot)$ is the measure of similarity, τ is the temperature value. $V(\cdot)$ and $T(\cdot)$ are the visual and image encoders, $\boldsymbol{c}_j^+ = a(c_i)_j$ is the selected positive label on the tracked lable path at layer l. $\boldsymbol{c}_{j,l,q}^-$ is the *q-th* sibling of the *j-th* ground-truth at level l; see Agorithm 1.

Algorithm 1. Hierarchical Graph Representation Net (HGR-Net)

```
Require: Training set $\mathcal{D}_{\mathrm{tr}}$, text encoder $T$, visual encoder $V$, inner ratio $M$, outer
  ratio $K$, per layer sampling number threshold $\epsilon$, training label set $\mathcal{C}$, hierarchical
  graph $\mathcal{G}$
  Sample a batch of data $\mathbf{X}$ from class $c_i$
  Obtain its ancestor path $\mathcal{A}^{c_i}$
  Set the outer loop range $k_s = Kd(c_i), k_e = d(c_i)$
  for $j = k_s, k_s + 1, \dots, k_e$ do
      Set the current ground-truth label $\boldsymbol{c}_j^+ = a(c_i)_j$
      Prepare $\mathrm{pos}^{\mathrm{j}}$ according to Eq. 3
      Set the inner-loop ranges $m_s = Md(\boldsymbol{c}^+), m_e = d(\boldsymbol{c}^+)$
      for $l = m_s, m_s + 1, \dots, m_e$ do
          Prepare sibling set $\mathcal{S}^{c_j}$
          $n_l = \max(\epsilon, |\mathcal{S}^{c_j}|)$
          Sample $n_l$ negative sibling set $\left\{\boldsymbol{c}_{j,l,q}^-\right\}_{q=1}^{n_l}$
          Prepare $\mathrm{neg}^{\mathrm{j,l}}$ according to Eq. 4
          Compute $\mathcal{L}_{j,l}$ according to Eq. 2
      end for
      Compute $\mathcal{L}_j$ according to Eq. 1 right part
  end for
  Compute $\mathcal{L}_{\mathrm{cont}}$ according to Eq. 1 left part
```

4 Experiments

4.1 Datasets and the Hierarchical Structure

ImageNet [4] is a widely used large-scale benchmark for ZSL organized according to the WordNet hierarchy [18], which can lead our model to learn the hierarchical relationship among classes. However, the original hierarchical structure is not a DAG (Directed Acyclic Graph), thus not suitable when implementing our method. Therefore, to make all of the classes fit into an appropriate location in the hierarchical DAG, we reconstruct the hierarchical structure by removing some classes from the original dataset, which contains seen classes from the ImageNet-1K and unseen classes from the ImageNet-21K (winter-2021 release), resulting a modified dataset ImageNet-21K-D (D for Directed Acyclic Graph).

It is worth noticing that although there are 12 layers in the reconstructed hierarchical tree in total, most nodes reside between 2nd and 6th layers. Our class-wise dataset split is based on GBU [35], which provides new dataset splits for ImageNet-21K with 1K seen classes for training and the remaining 20, 841 classes as test split. Moreover, GBU [35] splits the unseen classes into three different levels, including "2-hop", "3-hops" and "All" based on WordNet hierarchy [18]. More specifically, the "2-hops" unseen concepts are within 2-hops from the known concepts. After the modification above, the training is then conducted on the processed ImageNet-1K with seen 983 classes, while 17,295 unseen classes from the processed ImageNet-21K are for ZSL testing, and 1533 and 6898 classes for "2-hops" and "3-hops" respectively. Please note that there is no overlap between the seen and unseen classes. The remaining 983 seen classes make our training setting more difficult because our model gets exposed to fewer images than the original 1k seen classes. Please refer to the supplementary materials for more detailed descriptions of the dataset split and reconstruction procedure.

4.2 Implementation Details

We use a modified ResNet-50 [8] from [24] as the image encoder, which replaces the global average pooling layer with an attention mechanism, to obtain visual representation with feature dimensions of 1024. Text descriptions are encoded into tokens and bracketed with start tokens and end tokens based on byte pair encoding (BPE) [26] with the max length of 77. For text embedding, we use CLIP [24] Transformer to extract semantic vectors with the same dimensions as feature representation. We obtain the logits with L2-normalized image and text features and calculate InfoNCE loss [21] layer by layer with an adaptive reweighting strategy. More specifically, a learnable parameter with a size equivalent to the depth of the hierarchical tree is used to adjust the weights adaptively.

Training Details. We use the AdamW optimizer [14] applied to all weights except the adaptive attention layer with a learning rate 3e-7. We use the SGD optimizer for the adaptive layer with a learning rate of 1e-4. When computing the matmul product of visual and text features, a learnable temperature parameter τ is initialized as 0.07 from [30] to scale the logits and clips gradient norm of the parameters to prevent training instability. To accelerate training and avoid additional memory, mixed-precision [16] is used, and the weights of the model are only transformed into float32 for optimization. Our proposed HGR model is implemented in PyTorch, and training and testing are conducted on a Tesla V100 GPU with a batch size of 256 and 512, respectively.

4.3 Large-Scale ZSL Performance

Comparison Approaches. We compare with the following approaches:

- **DeViSE** [7] linearly maps visual information to the semantic word-embedding space. The transformation is learned using a hinge ranking loss.
- **HZSL** [12] learns similarity between the image representation and the class representations in the hyperbolic space.
- **SGCN** [10] uses an asymmetrical normalized graph Laplacian to learn the class representations.

– **DGP** [10] separates adjacency matrix into ancestors and descendants and propagates knowledge in two phases with one additional dense connection layer based on the same graph as in GCNZ [32].
– **CNZSL** [27] utilizes a simple but effective class normalization strategy to preserve variance during a forward pass.
– **FREE** [1] incorporates semantic-visual mapping into a unified generative model to address cross-dataset bias.

Evaluation Protocols. We use the typical Top@K criterion, but we also introduce additional metrics. Since it could be more desirable to have a relatively general but correct prediction rather than a more specific but wrong prediction, the following three metrics evaluate a given model's ability to learn the hierarchical relationship between the ground truth and its general classes.

- **Top-Overlap Ratio (TOR).** In this metric, we take a further step to also cover all the ancestor nodes of the ground truth class. More concretely, for an image x_j from class c_i of depth q_{c_i}, TOR is defined as:

$$TOR(x_j) = \frac{|p_{x_j} \cap \{A_{c_i}, c_i\}|}{q_{c_i}} \tag{5}$$

  where c_i is the corresponding class to image x_j. A_{c_i} is the union of all the ancestors of class c_i and p_{x_j} is the predicted class of x_j. In other words, this metric consider the predicted class correct if it is an ancestor of the ground truth.
- **Point-Overlap Ratio (POR).** In this setting, we let the model predict labels layer by layer. POR is defined as:

$$POR(x_j) = \frac{|P_{x_j} \cap P_{c_i}|}{q_{c_i}}, \tag{6}$$

  where $P_{c_i} = \{c_{i_1}, c_{i_2}, c_{i_3}, \cdots, c_{i_{q_{c_i}-1}}, c_i\}$ is the union of classes from the root to the ground truth through all the ancestors, and P_{x_j} is the union of classes predicted by our model layer by layer. q_{c_i} is count of all the ancestors including the ground truth label, which is tantamount to the depth of node c_i. The intersection calculates the overlap between correct and predicted points for image x_j.

Results Analysis. Table 1 demonstrates the performance of different models on ImageNet-21K ZSL setting on Top@K and above-mentioned three hierarchical evaluation. Our proposed model outperforms SoTA methods in all metrics, including hierarchical measures, proving the ability to learn the hierarchical relationship between the ground truth and its ancestor classes. We also attach the performance on 2-hops and 3-hops in the supplementary.

4.4 Ablation Studies

Different Attributes. Conventional attribute-based ZSL methods use GloVe [23] or Skip-Gram [17] as text models, while CLIP [24] utilizes prompts (i.e., text description) template: "a photo of a `[CLASS]`", and take advantage of Transformer to extract text feature. Blindly adding Transformer to some attribute-based methods like HZSL [12] which utilizes unique techniques to improve their performance in the attribute setting result in unreliable results. Therefore, we conducted experiments comparing three selected methods with different attributes. The result in Table 2 shows that methods based on text embedding extracted by CLIP transformer outperform traditional attribute-based ones since the low dimension representations (500-D) from w2v [17] is not discriminative enough to distinguish unseen classes, while higher dimension (1024-D) text representations significantly boost classification performance. Our HGR-Net gained significant improvement by utilizing Transformer compared to the low dimension representation from w2v [17].

Table 1. Top@k accuracy, Top-Overlap Ratio (TOR), and Point-Overlap Ratio (POR) for different models on the ImageNet-21K-D only testing on unseen classes. Tr means text encoder is CLIP Transformer.

Method	Hit@ **k**(%)					TOR	POR
	1	2	5	10	20		
Devise [7]	1.0	1.8	3.0	15	23.8	–	–
HZSL [12]	3.7	5.9	10.3	13.0	16.4	–	–
SGCN(w2v) [10]	2.79	4.49	8.26	13.05	19.49	4.97	10.01
SGCN(Tr) [10]	4.83	8.17	14.61	21.23	29.42	8.33	14.69
DGP(w2v) [10]	3.00	5.12	9.49	14.28	20.55	7.07	11.71
DGP(Tr) [10]	5.78	9.57	16.89	24.09	32.62	12.39	15.50
CNZSL(w2v) [27]	1.94	3.17	5.88	9.12	13.73	3.93	4.03
CNZSL(Tr) [27]	5.77	9.48	16.49	23.25	31.00	8.32	7.22
FREE(w2v) [1]	2.87	4.91	9.54	13.28	20.36	4.89	5.37
FREE(Tr) [1]	5.76	9.54	16.71	23.65	31.81	8.59	9.68
CLIP [24]	15.22	22.54	33.43	42.13	50.93	18.55	14.68
HGR-Net(Ours)	**16.39**	**24.19**	**35.66**	**44.68**	**53.71**	**18.90**	**16.19**

Different Outer and Inner Ratio. Fig. 3 demonstrate the Top1, Top-Overlap Ratio (TOR) and Point-Overlap Ratio (POR) metrics of different `K` and `M`, where `K` and `M` $\in [0, 1]$. `K` and `M` are outer and inner ratio that determine how many samples is considered in the inner and outer loop respectively as earlier illustrated.

We explore different `K` and `M` in this setting and observe how performance differs under three evaluations. Please note that when `K` or `M` is 0.0, it means only the current node is involved in a loop. As `K` increases, the model is prone to

Table 2. Different attributes. DGP(w/o) means without separating adjacency matrix into ancestors and descendants, `CN` and `INIT` in CNZSL means class normalization and proper initialization respectively.

Attributes	Methods	Hit@ **k**(%)					TOR	POR
		1	2	5	10	20		
w2v	SGCN [10]	2.79	4.49	8.26	13.05	19.49	4.97	10.01
	DGP(w/o) [10]	2.90	4.86	8.91	13.67	20.18	3.96	11.49
	DGP [10]	3.00	5.12	9.49	14.28	20.55	7.07	11.71
	CNZSL(w/o CN) [27]	0.83	1.47	3.03	5.08	8.27	1.98	2.05
	CNZSL(w/o INIT) [27]	1.84	3.13	6.08	9.47	14.13	3.04	4.05
	CNZSL [27]	1.94	3.17	5.88	9.12	13.73	3.93	4.03
	FREE [1]	2.87	4.91	9.54	13.28	20.36	4.89	5.37
	HGR-Net(Ours)	2.35	3.69	7.03	11.46	18.27	4.38	5.76
Transformer(CLIP)	SGCN [10]	4.83	8.17	14.61	21.23	29.42	8.33	14.69
	DGP(w/o) [10]	5.42	9.16	16.01	22.92	31.20	7.80	15.29
	DGP [10]	5.78	9.57	16.89	24.09	32.62	12.39	15.50
	CNZSL(w/o CN) [27]	1.91	3.45	6.74	10.55	15.51	3.19	3.43
	CNZSL(w/o INIT) [27]	5.65	9.33	16.24	22.88	30.63	8.32	7.03
	CNZSL [27]	5.77	9.48	16.49	23.25	31.00	7.97	7.22
	FREE [1]	5.76	9.54	16.71	23.65	31.81	8.59	9.68
	HGR-Net(Ours)	**16.39**	**24.19**	**35.66**	**44.68**	**53.71**	**18.95**	**16.19**

obtain higher performance on hierarchical evaluation. An intuitive explanation is that more conceptual knowledge about ancestor nodes facilitates hierarchical learning relationships among classes.

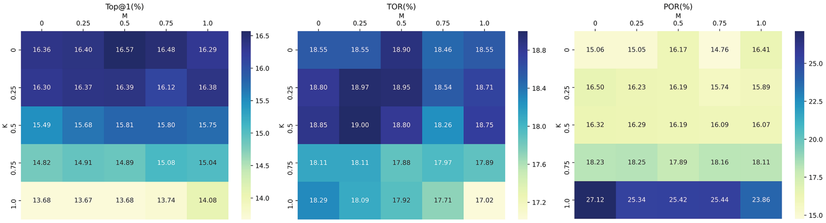


Fig. 3. Different outer ratio (K) and inner ratio (M)

Different Negative Sampling Strategies. We explore various sampling strategies for choosing negative samples and observe how they differ in performance. *Random* randomly samples classes from all the classes. *TopM* samples neighbour nodes from $(q_{c_i} - M)$ to q_{c_i} layers, where q_{c_i} is the depth of inner anchor c_i, and we set M as 1. *Similarity* calculates the similarity of text features and chooses the top similar samples with the positive sample as hard negatives. *Sibling* samples sibling nodes of the target class. Table 3 indicates that *TopM*

Table 3. Analysis of sampling strategies

Strategy	Hit@ **k**(%)					TOR	POR
	1	2	5	10	20		
Random	15.72	23.33	34.69	43.68	52.73	16.12	13.04
Sibling	16.25	23.95	35.29	44.16	53.09	17.91	13.46
Similarity	16.35	24.04	35.33	44.17	53.07	18.60	14.78
TopM(default)	**16.39**	**24.19**	**35.66**	**44.68**	**53.71**	**18.90**	**16.19**

outperforms other sampling strategies. Therefore, we adopt the *TopM* sampling strategy in the subsequent ablation studies.

Different Weighting Strategies. Orthogonal to negative sampling methods, we explore in this ablation the influence of different weighting strategies across the levels of the semantic hierarchy. The depth of the nodes in the hierarchical structure is not well-balanced, and the layers are not accessible for all objects. Therefore, it is necessary to focus on the importance of different layers. In this case, we experimented with 6 different weighting strategies and observed how they differ in multiple evaluations. As Table 4 shows, *Increasing* gives more weights to deeper layers in a linear way and ↑ *non-linear* is exponentially increasing weights to deeper layers. To balance the Top@K and hierarchical evaluations, the adaptive weighting method is proposed to obtain a comprehensive result. More specifically, *Adaptive* uses a learnable parameter with a size equivalent to the depth of the hierarchical tree to adjust the weights adaptively. We attached the exact formulation of different weighting strategies in the supplementary.

Table 4. Analysis of the weighting strategies when re-weighting in both inner and outer loop with K=0.25 and M=0.5.

Weighting	Hit@ **k**(%)					TOR	POR
	1	2	5	10	20		
Adaptive(default)	**16.39**	**24.19**	**35.66**	**44.68**	**53.71**	**18.90**	**16.19**
Equal	15.97	23.65	35.02	43.97	52.97	17.82	13.71
Increasing ↑	15.85	23.50	34.85	43.83	52.83	17.80	13.81
Decreasing ↓	16.08	23.77	35.16	44.10	53.09	17.84	13.59
↑ (non-linear)	15.58	23.13	34.43	43.44	52.46	17.79	14.12
↓ (non-linear)	16.19	23.89	35.26	44.18	53.13	17.87	13.47

Experiment on ImageNet-21K-P [25]. ImageNet-21K-P [25] is a pre-processed dataset from ImageNet21K by removing infrequent classes, reducing the number of total numbers by half but only removing only 13% of the original images, which contains 12,358,688 images from 11,221 classes. We select the intersection of this dataset with our modified ImageNet21K dataset to ensure DAG structure consistency. The spit details (class and sample wise) are demonstrated in the supplementary.

We show experimental results on ImageNet-21K-P comparing our method to different SoTA variants. Our model performs better in this smaller dataset compared to the original larger one in Table 1 and outstrips all the previous ZSL methods. We presented important results in Table 5 and we attached more results in the supplementary.

4.5 Qualitative Results

Figure 4 shows several retrieved images by implementing our model in the ZSL setting on ImageNet-21K-D. The task is to retrieve images from an unseen class with its semantic representation. Each row demonstrates three correct retrieved images and one incorrect image with its true label. Although our algorithm retrieves images from the wrong class, they are still visually similar to ground truth. For instance, the true label hurling and the wrong class American football belong to sports games, and images from both contain several athletes wearing helmets against a grass background. We also show some prediction examples in Fig. 5 to present Point-Overlap results.

Table 5. Result of ImageNet21K-P [25]. DGP(w/o) [10] means without separating adjacency matrix into ancestors and descendants, `CN` and `INIT` in CNZSL [27] means class normalization and proper initialization respectively, and Tr is Transformer of CLIP for short.

Models	Hit@ **k**(%)					TOR	POR
	1	2	5	10	20		
CNZSL(Tr w/o CN) [27]	3.27	5.59	10.69	16.17	23.33	5.32	7.68
CNZSL(Tr w/o INIT) [27]	7.90	12.77	21.40	29.50	38.63	11.23	12.56
CNZSL(Tr) [27]	7.97	12.81	21.75	29.92	38.97	11.50	12.62
FREE(Tr) [1]	8.15	12.90	21.37	30.29	40.62	11.82	13.34
CLIP [24]	19.33	28.07	41.66	53.77	61.23	20.08	20.27
HGR-Net (Ours)	**20.08**	**29.35**	**42.49**	**52.47**	**62.00**	**23.43**	**23.22**

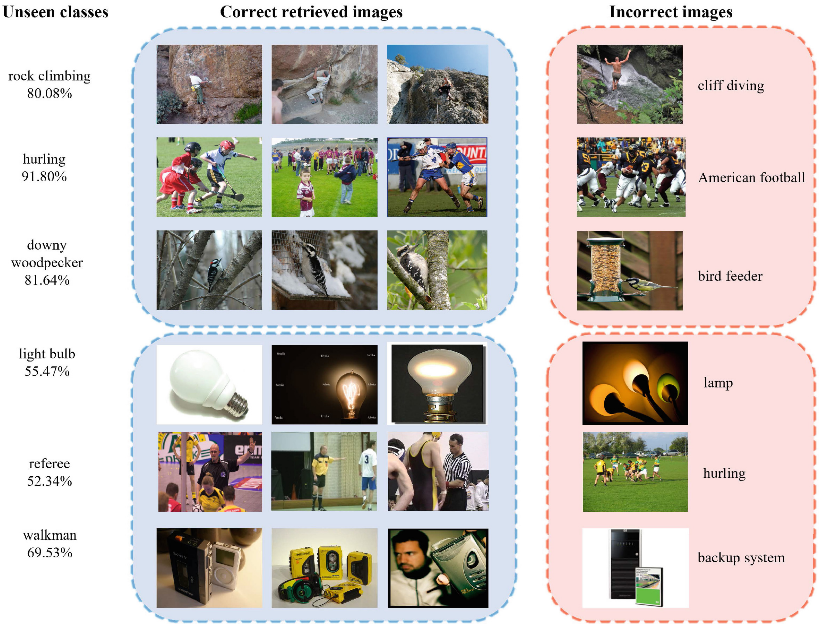
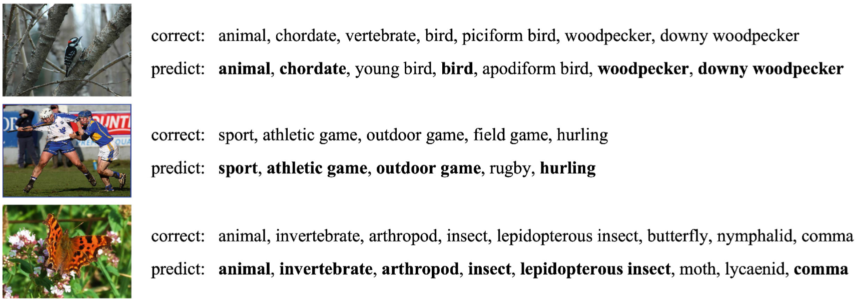


Fig. 4. Zero-shot retrieved images. The first column represents unseen class names and corresponding confidence, the middle shows correct retrieval, and the last demonstrates incorrect images and their true labels.

4.6 Low-shot Classification on Large-Scale Dataset

Apart from zero-shot experiments being our primary goal in this paper, we also explore the effectiveness of our method in the low-shot setting compared to several baselines. Unlike pure few-shot learning, our support set comprises two parts. To be consistent with ZSL experiments, all the training samples of 983 seen classes are for low-shot training. For the 17, 295 unseen classes used in the ZSL setting, k-shots (1,2,3,5,10) images are randomly sampled for training in the low-shot setting, and the remaining images are used for testing. The main goal of this experiment is to show how much models could improve from zero to one shot and whether our proposed hierarchical-based method could generalize well in the low-shot scenario. Figure 6 illustrated the few-shots results comparing our model to various SoTA methods. Although our approach gains trivial Top@k improvements from 1 to 10 shots, the jump from 0 to 1 shot is two times that from 1 to 10, proving that our model is an efficient learner.



correct: animal, chordate, vertebrate, bird, piciform bird, woodpecker, downy woodpecker

predict: **animal**, **chordate**, young bird, **bird**, apodiform bird, **woodpecker**, **downy woodpecker**

correct: sport, athletic game, outdoor game, field game, hurling

predict: **sport**, **athletic game**, **outdoor game**, rugby, **hurling**

correct: animal, invertebrate, arthropod, insect, lepidopterous insect, butterfly, nymphalid, comma

predict: **animal**, **invertebrate**, **arthropod**, **insect**, **lepidopterous insect**, moth, lycaenid, **comma**

Fig. 5. Predicted examples to show Point-Overlap. First row of each image is correct points from root to the ground truth and the second row show predicted points. The hit points are highlighted in bold.

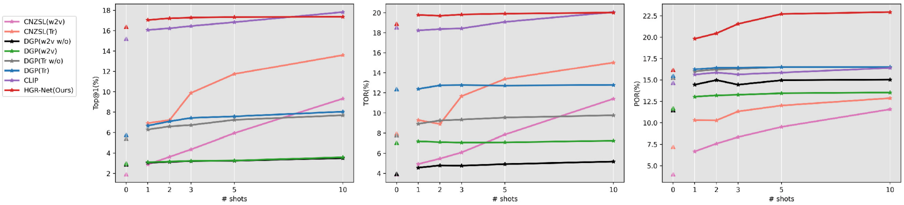


Fig. 6. Few shots comparison. DGP(w/o) [10] means without separating adjacency matrix into ancestors and descendants, `CN` and `INIT` in CNZSL [27] means class normalization and proper initialization respectively, and Tr is Transformer of CLIP [24] for short.

5 Conclusions

This paper focuses on scaling-up visual recognition of unseen classes to tens of thousands of categories. We proposed a novel hierarchical graphic knowledge representation framework for confidence-based classification and demonstrated significantly better performance than baselines over Image-Net-21K-D and Image-Net-21K-P benchmarks, achieving new SOTA. We hope our work help ease future research of zero-shot learning and pave a steady way to understand large-scale visual-language relationships with limited data.

Acknowledgments. Research reported in this paper was supported by King Abdullah University of Science and Technology (KAUST), BAS/1/1685-01-01.

References

1. Chen, S., et al.: Free: Feature refinement for generalized zero-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 122–131 (2021)

2. Cheng, R.: Data efficient language-supervised zero-shot recognition with optimal transport distillation (2021)
3. Cox, M.A., Cox, T.F.: Multidimensional scaling. In: Handbook of data visualization, pp. 315–347. Springer (2008). https://doi.org/10.1007/978-3-642-28753-4_101322
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
5. Elhoseiny, M., Saleh, B., Elgammal, A.: Write a classifier: Zero-shot learning using purely textual descriptions. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2584–2591 (2013)
6. Elhoseiny, M., Zhu, Y., Zhang, H., Elgammal, A.: Link the head to the" beak": Zero shot learning from noisy text description at part precision. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6288–6297. IEEE (2017)
7. Frome, A., et al.: Devise: A deep visual-semantic embedding model (2013)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
9. Jia, C., et al.: Scaling up visual and vision-language representation learning with noisy text supervision. arXiv preprint arXiv:2102.05918 (2021)
10. 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/CVF Conference on Computer Vision and Pattern Recognition, pp. 11487–11496 (2019)
11. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
12. Liu, S., Chen, J., Pan, L., Ngo, C.W., Chua, T.S., Jiang, Y.G.: Hyperbolic visual embedding learning for zero-shot recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9273–9281 (2020)
13. Long, Y., Shao, L.: Describing unseen classes by exemplars: Zero-shot learning using grouped simile ensemble. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 907–915. IEEE (2017)
14. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
15. Lu, Y.: Unsupervised learning on neural network outputs: with application in zero-shot learning. arXiv preprint arXiv:1506.00990 (2015)
16. Micikevicius., et al.: Mixed precision training. arXiv preprint arXiv:1710.03740 (2017)
17. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
18. Miller, G.A.: Wordnet: a lexical database for english. Commun. ACM **38**(11), 39–41 (1995)
19. Nayak, N.V., Bach, S.H.: Zero-shot learning with common sense knowledge graphs. arXiv preprint arXiv:2006.10713 (2020)
20. Norouzi, M., et al.: Zero-shot learning by convex combination of semantic embeddings. arXiv preprint arXiv:1312.5650 (2013)
21. Van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv e-prints pp. arXiv-1807 (2018)

22. Patterson, G., Hays, J.: Sun attribute database: Discovering, annotating, and recognizing scene attributes. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2751–2758. IEEE (2012)
23. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
24. Radford, A., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
25. Ridnik, T., Ben-Baruch, E., Noy, A., Zelnik-Manor, L.: Imagenet-21k pretraining for the masses. arXiv preprint arXiv:2104.10972 (2021)
26. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (2016)
27. Skorokhodov, I., Elhoseiny, M.: Class normalization for zero-shot learning. In: International Conference on Learning Representations (2021). https://openreview.net/forum?id=7pgFL2Dkyyy
28. Sun, Q., Liu, Y., Chen, Z., Chua, T.S., Schiele, B.: Meta-transfer learning through hard tasks. IEEE Trans. Pattern Anal. Mach. Intell. (2020)
29. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
30. Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M.: Rotation equivariant cnns for digital pathology. CoRR (2018)
31. Wang, J., Jiang, B.: Zero-shot learning via contrastive learning on dual knowledge graphs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 885–892 (2021)
32. 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, pp. 6857–6866 (2018)
33. Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M.: Generalizing from a few examples: a survey on few-shot learning. ACM Comput. Surv. (CSUR) **53**(3), 1–34 (2020)
34. Welinder, P., et al.: Caltech-ucsd birds 200 (2010)
35. Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. In: PAMI (2018)
36. Xie, G.S., et al.: Attentive region embedding network for zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9384–9393 (2019)
37. Ye, H.J., Hu, H., Zhan, D.C.: Learning adaptive classifiers synthesis for generalized few-shot learning. Int. J. Comput. Vision **129**(6), 1930–1953 (2021)
38. Yu, Y., Ji, Z., Fu, Y., Guo, J., Pang, Y., Zhang, Z.M.: Stacked semantics-guided attention model for fine-grained zero-shot learning. In: NeurIPS (2018)
39. Zhang, C., Cai, Y., Lin, G., Shen, C.: Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In 2020 IEEE CVF Conference on Computer Vision and Pattern Recognition, pp. 12200–12210 (2020)
40. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. arXiv preprint arXiv:2109.01134 (2021)