Learning Graph Embeddings for Compositional Zero-shot Learning

Muhammad Ferjad Naeem[1,3], Yongqin Xian[2], Federico Tombari[3,4], Zeynep Akata[1,2,5]
[1]University of Tübingen, [2]MPI for Informatics, [3]TUM, [4]Google [5]MPI for Intelligent Systems

Abstract

In compositional zero-shot learning, the goal is to recognize unseen compositions (e.g. old dog) of observed visual primitives states (e.g. old, cute) and objects (e.g. car, dog) in the training set. This is challenging because the same state can for example alter the visual appearance of a dog drastically differently from a car. As a solution, we propose a novel graph formulation called Compositional Graph Embedding (CGE) that learns image features, compositional classifiers and latent representations of visual primitives in an end-to-end manner. The key to our approach is exploiting the dependency between states, objects and their compositions within a graph structure to enforce the relevant knowledge transfer from seen to unseen compositions. By learning a joint compatibility that encodes semantics between concepts, our model allows for generalization to unseen compositions without relying on an external knowledge base like WordNet. We show that in the challenging generalized compositional zero-shot setting our CGE significantly outperforms the state of the art on MIT-States and UT-Zappos. We also propose a new benchmark for this task based on the recent GQA dataset. Code is available at: https://github.com/ExplainableML/czsl

1. Introduction

A "black swan" was ironically used as a metaphor in the 16th century for an unlikely event because the western world had only seen white swans. Yet when the European settlers observed a black swan for the first time in Australia in 1697, they immediately knew what it was. This is because humans posses the ability to compose their knowledge of known entities to generalize to novel concepts. Since visual concepts follow a long tailed distribution [43, 48], it is not possible to gather supervision for all concepts. Therefore, recognizing shared and discriminative properties of objects and reasoning about their various states has evolved as an essential part of human intelligence. Once familiar with the semantic meaning of these concepts, we can recognize unseen compositions of them without any supervision. While there is a certain degree of compositionality in modern vision systems, e.g. feature sharing, most models are not compositional in the classifier space and treat every class as an independent entity requiring training for any new concept.

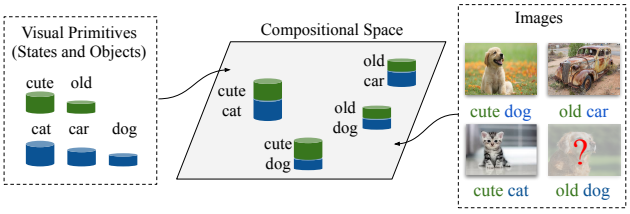

Figure 1: We aim to build a classifier for a novel state of a known object (e.g. `old dog`) given the knowledge of the shared primitives state and object in the training set.

In this work, we study the state-object compositionality problem also known as Compositional Zero-Shot Learning (CZSL)[34]. The goal is to learn the compositionality of observed objects and their states as visual primitives to generalize to novel compositions of them as shown in figure 1. Some notable existing works in this field include learning a transformation network on top of individual classifiers [34], treating states as linear transformations of object vectors [35], learning modular networks conditioned on compositional classes [39] and learning object embeddings that are symmetric under different states [28]. However, these works treat each state-object composition independently, ignoring the rich dependency structure of different states, objects and their compositions. For example, learning the composition `old dog` is not only dependent on the state `old` and object `dog`, but also can be supported by other compositions like `cute dog`, `old car`, etc. We argue that such dependency structure provides a strong regularization which allows the network to better generalize to novel compositions. We therefore propose to exploit this dependency relationship by constructing a compositional graph to learn embeddings that are globally consistent.

Our contributions are as follows: (1) We introduce a novel graph formulation named Compositional Graph Embedding (CGE) to model the dependency relationship of visual primitives and compositional classes. This graph can

be created independently of an external knowledge base like WordNet [32]. (2) Observing that visual primitives are dependent on each other and their compositional classes (figure 1), we propose a multimodal compatibility learning framework that learns to embed related states, objects and their compositions close to each other and far away from the unrelated ones. (3) We propose a new benchmark called C-GQA for the task of CZSL. This dataset is curated from the recent GQA[15] dataset with diverse compositional classes and clean annotations compared to datasets used in the community. (4) Our model significantly improves the state of the art on all the metrics on MIT-States, UT-Zappos and C-GQA datasets.

2. Related work

Compositionality can loosely be defined as the ability to decompose an observation into its primitives. These primitives can then be used for complex reasoning. One of the earliest attempts in computer vision in this direction can be traced to Hoffman [14] and Biederman [4] who theorized that visual systems can mimic compositionality by decomposing objects to their parts. Compositionality at a fundamental level is already included in modern vision systems. Convolutional Neural Networks (CNN) have been shown to exploit compositionality by learning a hierarchy of features[57, 25]. Transfer learning[6, 8, 10, 38] and few-shot learning[12, 40, 30] exploit the compositionality of pretrained features to generalize to data constraint environments. Visual scene understanding[18, 9, 17, 29] aims to understand the compositionality of concepts in a scene. Nevertheless, these approaches still requires collecting data for new classes.

Zero-Shot Learning aims at recognizing novel classes that are not observed during training [24]. This is accomplished by using side information that describes novel classes e.g. attributes [24], text descriptions [41] or word embeddings [44]. Some notable approaches include learning a compatibility function between image and class embeddings [1, 58] and learning to generate image features for novel classes [52, 59]. Graph convolutional networks (GCN) [21, 46, 19] have shown to be promising for zero-shot learning. Wang et al. [46] propose to directly regress the classifier weights of novel classes with a GCN operated on an external knowledge graph (WordNet [32]). Kampffmeyer et al.[19] improve this formulation by introducing a dense graph to learn a shallow GCN as a remedy for the laplacian smoothing problem [27].

Graph Convolutional Networks are a special type of neural networks that exploit the dependency structure of data (nodes) defined in a graph. Current methods [21] are limited by the network depth due to over smoothing at deeper layers of the network. The extreme case of this can cause all nodes to converge to the same value [27]. Several works have tried to remedy this by dense skip connections [53, 26], randomly dropping edges [42] and applying a linear combination of neighbor features [49, 23, 22]. A recent work in this direction from Chen et al.[33] combines residual connections with identity mapping.

Compositional zero-shot learning stands at the intersection of compositionality and zero-shot learning and focuses on state and object relations. We aim to learn the compositionality of objects and their states from the training set and are tasked with generalizing to unseen combination of these primitives. Approaches in this direction can be divided into two groups. The first group is directly inspired by [14, 4]. Some notable methods including learning a transformation upon individual classifiers of states and objects [34], modeling each state as a linear transformation of objects [35], learning a hierarchical decomposition and composition of visual primitives[54] and modeling objects to be symmetric under attribute transformations[28]. An alternate line of works argues that compositionality requires learning a joint compatibility function with respect to the image, the state and the object[2, 39, 47]. This is achieved by learning a modular networks conditioned on each composition [39, 47] that can be "rewired" for a new compositions. Finally a recent work from Atzmon et al. [2] argue that achieving generalization in CZSL requires learning the causality of visual transformation through a causal graph where the latent representation of primitives are independent of each other.

Our proposed method lies at the intersection of several discussed approaches. We learn a joint compatibility function similar to [2, 39, 47] and utilize a GCN similar to [46, 19]. However, our approach exploits the dependency structure between states, objects and compositons which has been overlooked by previous CZSL approaches [2, 39, 47]. Instead of using a predefined knowledge graph like WordNet [32] to regress pretrained classifiers of the seen classes [46, 19], we propose a novel way to build a compositional graph and learn classifiers for all classes in an end-to-end manner. In contrast to Atzmon et al.[2] we explicitly promote the dependency between all primitives and their compositions in our graph. This allows us to learn embeddings that are consistent with the whole graph. Finally, unlike all existing methods [34, 35, 2, 39, 47, 54], we do not rely on a fixed image feature extractor and train our pipeline end-to-end.

3. Approach

We consider the image classification task where each image is associated with a label that is composed of a state (e.g. `cute`) and an object (e.g. `dog`). The goal of compositional zero-shot learning (CZSL) [34] is to recognize the compositional labels that are not observed during

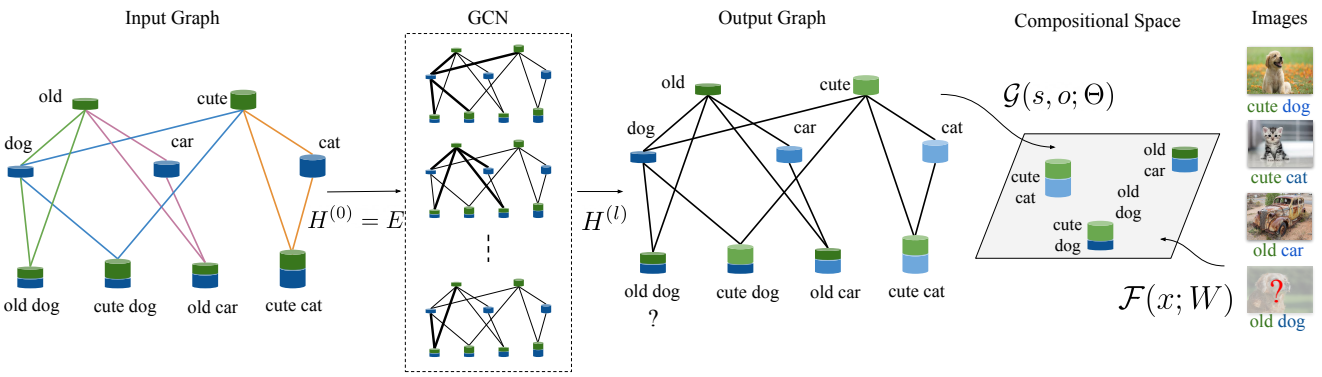

Figure 2: Compositional Graph Embed (CGE) learns a globally consistent joint embedding space between image features and classes of seen and unseen compositions from a graph. In our novel graph formulation, nodes are connected if a dependency exists in form of a compositional label e.g. old, car and old car. We backpropagate the classification loss through the seen compositional nodes to the GCN $\mathcal{G}$ and the feature extractor $\mathcal{F}$. Hence, the representation of e.g. the dog is compatible with its different states and the representation of old dog aggregates the knowledge from old, dog, cute dog, old car etc.

training. This is particularly challenging as the states significantly change the visual appearance of an object hindering the performance of the classifiers.

We propose a novel formulation to the problem, namely Compositional Graph Embedding (CGE), which constructs a compositional graph and adopts a graph convolutional network to learn the dependency structure between labels. An overview of our approach is shown in Figure 2. It builds on the compatibility learning framework that learns a class-agnostic scoring function between an image and a compositional label. The input image is encoded with an image feature extractor $\mathcal{F}$, while the classifier weights for the compositional label are learned by a composition function $\mathcal{G}$. The key insight of our approach is that leveraging the dependency relationship between states, objects and their compositions is beneficial for recognizing unseen compositions.

3.1. Compatibility Learning Framework for CZSL

Task formulation. We formalize the CZSL task as follows. Let $\mathcal{T} = \{(x, y)|x \in \mathcal{X}, y \in \mathcal{Y}_s\}$ where $\mathcal{T}$ stands for the training set, x denotes an image in the RGB image space $\mathcal{X}$ and y is its label belonging to one of the seen labels $\mathcal{Y}_s$. Each label is a tuple $y = (s, o)$ of a state $s \in \mathcal{S}$ and an object $o \in \mathcal{O}$ with $\mathcal{S}$ and $\mathcal{O}$ being the set of states and objects respectively. The task of CZSL is to predict a set of novel labels $\mathcal{Y}_n$ that consists of novel compositions of states $\mathcal{S}$ and objects $\mathcal{O}$ i.e., $\mathcal{Y}_s \cap \mathcal{Y}_n = \emptyset$. Following [39, 51], we study this problem in the generalized compositional zero-shot setting where the test set includes images from both seen and novel compositional labels $\mathcal{Y} = \mathcal{Y}_s \cup \mathcal{Y}_n$.

Compatibility function. Learning state and object classifiers separately is prone to overfit to labels observed during training because states and objects are not independent e.g. the appearance of the state `sliced` varies significantly with the object (e.g. `apple` or `bread`). Therefore, we chose to model them jointly by learning a compatibility function $f : \mathcal{X} \times \mathcal{S} \times \mathcal{O} \longrightarrow \mathbb{R}$ that captures the compatibility score between an image, a state and an object. Given a specific input image x, we predict its label $y = (s, o)$ by searching the state and object composition that yields the highest compatibility score:

$$f(x, s, o) = \mathcal{F}(x; W) \cdot \mathcal{G}(s, o; \Theta) \tag{1}$$

where $\mathcal{F}(x; W) \in \mathbb{R}^d$ is the image feature extracted from a pretrained feature extractor, $\mathcal{G}(s, o; \Theta) \in \mathbb{R}^d$ is a function that outputs the label embedding of the state-object pair (s, o), (W, Θ) are respectively the learnable parameters of $\mathcal{F}$ and $\mathcal{G}$, and $(\cdot)$ is the dot product. The compatibility function assigns high scores to the correct triplets i.e., image x and its label (s, o), and low scores to the incorrect ones. The label embedding can be also interpreted as the classifier weights for the label (s, o) and we use the two terms interchangeably.

Our compatibility learning framework is closely related to [34, 39]. LabelEmbed [34] parameterizes the compositional embedding function with a multi-layer perceptron and computes the compositions from the word embeddings (e.g. word2vec [31]) of states and objects, while TMN [39] adopts a modular network as the image feature extractor and a gating network as the compositional embedding function. We argue that there exists a complex dependency structure between states, objects and their compositions and learning this dependency structure is crucial. To this end, we propose to integrate the compositional embedding function $\mathcal{G}$ as a graph convolutional neural network (GCN) which adds an inductive bias to the inherent structure between states, objects, and their combination defined by our compositional graph introduced next.

3.2. Compositional Graph Embedding (CGE)

We propose the Compositional Graph Embedding (CGE) framework integrating the Graph Convolutional Networks (GCN) [21] to the compositional embedding function $G(s, o)$ that learns the label embedding for each compositional label $y = (s, o) \in \mathcal{Y}$ in an end to end manner. The GCN network exploits the dependency structure in a predefined compositional graph from states, objects and their compositions (including both seen and unseen labels). In the following, we first define the compositional graph, then introduce the node features and finally explain how to learn a GCN for the CZSL task.

Compositional graph. Our graph consists of $K = |\mathcal{S}| + |\mathcal{O}| + |\mathcal{Y}|$ nodes that represent states $\mathcal{S}$, objects $\mathcal{O}$ and compositional labels $\mathcal{Y}$. Two nodes are connected if they are related. The key insight of our graph is that each compositional label $y = (s, o) \in \mathcal{Y}$ defines a dependency relationship between the state s, object o and their composition y. To this end, we build the edges of the graph by connecting (s, o), (s, y) and (o, y) for every $y = (s, o) \in \mathcal{Y}$. In addition, each node is also connected to itself. Note that the edges in our graph are unweighted and undirected, leading to a symmetric adjacency matrix $L \in \mathbb{R}^{K\times K}$ where element $L_{ij} = 1$ if there is a connection between nodes i and j otherwise $L_{ij} = 0$. Despite its simplicity, we find that our compositional graph provides the accurate dependency structure to recognize unseen compositional labels.

Node features. GCN [21, 33] operates on node features in a neighborhood defined by the graph. Therefore, after obtaining the compositional graph, we need to represent each node with a proper feature embedding. We chose to use the word embeddings [31, 5] pretrained on a large text corpus e.g. Wikipedia, because they capture rich semantic similarities among words i.e., `dog` is closer to `cat` than to `car` in the word embedding space. Specifically, every state or object node in the compositional graph is represented by the word embedding associated to its corresponding state or object name. We compute the node features of the the compositional label (e.g. `cute dog`) by averaging the word embeddings of the corresponding state (e.g. `cute`) and object (e.g. `dog`) names. As indicated in [31], by adding word embeddings we achieve compositionality in the semantic space. We represent the input node features with a matrix $E \in \mathbb{R}^{K\times P}$ where K is the total number of nodes and each row denotes the P-dim feature of a graph node.

Graph convolutional network for CZSL. GCN [21] is an efficient multi-layer network to learn new feature representation of nodes for a downstream task that are consistent with the graph structure. Here, we apply the GCN to tackle the CZSL task by directly predicting the compositional label embeddings. The input of our GCN consists of the compositional graph, represented by the adjencency matrix $L \in \mathbb{R}^{K\times K}$ and the node feature matrix $E \in \mathbb{R}^{K\times P}$. Specifically, each GCN layer computes the hidden representation of each node by convolving over neighbor nodes using a simple propagation rule [21] also known as a spectral convolution,

$$H^{(l+1)} = \sigma(D^{-1}LH^{(l)}\Theta^{(l)}) \tag{2}$$

where σ represents the non-linearity activation function ReLU, $H^{(l)} \in \mathbb{R}^{K\times U}$ denotes the hidden representations in the l^{th} layer with $H^{(0)} = E$ and $\Theta \in \mathbb{R}^{U\times V}$ is the trainable weight matrix with V learnable filters operating over U features of $H^{(l)}$. $D \in \mathbb{R}^{K\times K}$ is a diagonal node degree matrix which normalizes rows in L to preserve the scale of the feature vectors. By stacking multiple such layers, the GCN propagates the information through the graph to obtain better node embeddings for both the seen and unseen compositional labels. For example, our GCN allows an unseen compositional label e.g. `old dog` to aggregate information from its neighbor nodes e.g. `old`, `dog`, `cute dog`, and `old car` that are observed (see Figure 2).

Objective. As the objective of the GCN is to predict the classifier weights of the compositional labels, the node embedding of the output layer in the GCN has the same dimentionality as the image feature $\mathcal{F}(x)$. This indicates that our compositional embedding function becomes $\mathcal{G}(s, o) = H^{(N)}_y$ where $H^{(N)}$ is the output node embedding matrix and $H^{(N)}_y$ denotes the row corresponding to the compositional label $y = (s, o)$. We then optimize the following cross-entropy loss to jointly learn the image feature extractor and GCN in an end-to-end manner,

$$\min_{W,\Theta} \frac{1}{|\mathcal{T}|} \sum_{i=1}^{|\mathcal{T}|} -log(\frac{\exp f(x_i, s_i, o_i)}{\sum_{j\in\mathcal{Y}_s} \exp f(x_i, s_j, o_j)}) \tag{3}$$

where f is the compatibility function defined in Equation 1, $y = (s_i, o_i)$ is the ground truth label of image x_i, label $y' = (s_j, o_j)$ denotes any seen compositional class, W and Θ are the learnable parameters of the feature extractor and the GCN respectively. Intuitively, the cross-entropy loss enables the compatibility function to assign the high scores for correct input triplets.

Inference. At test time, given an input image x, we derive a prediction by searching the compositional label that yields the highest compatibility score,

$$\arg\max_{y=(s,o)\in\mathcal{Y}} f(x, s, o). \tag{4}$$

It is worth noting that our model works in the challenging generalized CZSL setting [39], where both seen and unseen compositional classes (i.e. $\mathcal{Y} = \mathcal{Y}_s \cup \mathcal{Y}_n$) are predicted.

Discussion. To the best of our knowledge, our Compositional Graph Embedding (CGE) is the first end-to-end learning method that jointly optimizes the feature extractor $\mathcal{F}$

			Training		Validation			Test		
Dataset	s	o	sp	i	sp	up	i	sp	up	i
MIT-States[16]	115	245	1262	30k	300	300	10k	400	400	13k
UT-Zappos[55]	16	12	83	23k	15	15	3k	18	18	3k
C-GQA (Ours)	453	870	6963	26k	1173	1368	7k	1022	1047	5k

Table 1: **Dataset statistics for CZSL**: We use three datasets to benchmark our method against the baselines. C-GQA (ours): our proposed dataset splits from Stanford GQA dataset [15]. (s: # states, o: # objects, sp: # seen compositions, up: # unseen compositions, i: # images)

and the compositional embedding function $\mathcal{G}$ for the task of compositional zero-shot learning.

Compared to prior CZSL works [39, 28, 34, 35] our CGE does not overfit while optimizing the CNN backbone of $\mathcal{F}$ (see supplementary) as it is regularized by the compositional graph that defines the dependency relationship between classes making the end-to-end training beneficial. Compared to previous GCN work [46, 19] that utilizes GCNs to regress the fixed classifier weights to learn classifiers of novel classes, we directly use image information to learn classifiers for both seen and novel classes. Moreover, while [46, 19] rely on a known knowledge graph like WordNet[32] describing the relation of novel classes to seen classes, our CGE cannot rely on existing knowledge graphs like WordNet[32] because they do not cover compositional labels. Therefore, we propose to construct the graph by exploiting the dependency relationship defined in the compositional classes. We find that propagating information from seen to unseen labels through this graph is crucial for boosting the CZSL performance.

4. Experiments

After introducing our experimental setup, we compare our results with the state of the art, ablate over our design choices and present some qualitative results.

Datasets. We perform our experiments on three datasets (see detailed statistics in Table 1). MIT-States[16] consists of natural objects in different states collected using an older search engine with limited human annotation leading to significant label noise [2]. UT-Zappos[55, 56] consists of images of a shoes catalogue which is arguably not entirely compositional as states like *Faux leather* vs *Leather* are material differences not always observable as visual transformations. We use the GCZSL splits from [39].

To address the limitations of these two datasets,we propose a split built on top of Stanford GQA dataset [15] originally proposed for VQA and name it Compositional GQA (C-GQA) dataset (see supplementary for the details). C-GQA contains over 9.5k compositional labels making it the most extensive dataset for CZSL. With cleaner labels and a larger label space, our hope is that this dataset will inspire further research on the topic. Figure 4 shows some samples from the three datasets.

Metrics. As the models in zero-shot learning problems are trained only on seen $\mathcal{Y}_s$ labels (compositions), there is an inherent bias against the unseen $\mathcal{Y}_n$ labels. As pointed out by [7, 39], the model thus needs to be calibrated by adding a scalar bias to the activations of the novel compositions to find the best operating point and evaluate the generalized CZSL performance [39] for a more realistic setting.

We adopt the evaluation protocol of [39] and report the Area Under the Curve (AUC) (in $\%$) between the accuracy on seen and unseen compositions at different operating points with respect to the bias. The best unseen accuracy is calculated when the bias term is large leading to predicting only the unseen labels, also known as zero-shot performance. In addition, the best seen (base class) performance is calculated when the bias term is negative leading to predicting only the seen labels. As a balance between the two, we also report the best harmonic mean. To emphasize that this is different from the traditional zero-shot learning evaluation, we add the term "best" in our results. Finally, we report the state and object accuracy on the novel labels to show the improvement in classifying the visual primitives. We emphasize that the dataset splits we propose for C-GQA and use from [39] for MIT-States and UT-Zappos do no not violate the zero-shot assumption as results are ablated on the validation set. Some works in CZSL use older splits that lack a validation set and thus use indirect full label supervision[51] by ablating over the test set. We therefore advice future works to rely on the new splits.

Training details. To be consistent with the state of the art, we use a ResNet18 [13] backbone pretrained on ImageNet as the image feature extractor $\mathcal{F}$. For a fair comparison with the models that use a fixed feature extractor, we introduce a simplification of our method named CGE_{ff}. We learn a 3 layer fully-connected (FC) network with ReLU[36], LayerNorm[3] and Dropout[45] while keeping the feature extractor fixed for this baseline. We use a shallow 2-layer GCN with a hidden dimension of 4096 as $\mathcal{G}$ (detailed ablation on this is presented in section 4.2). On MIT-States, we initialize our word embeddings with a concatenation of pretrained fasttext[5] and word2vec models[31] similar to [50]. On UT-Zappos and C-GQA, we initialize the word embeddings with word2vec(ablation reported in supplementary).

We use Adam Optimizer[20] with a learning rate of $5e^{-6}$ for $\mathcal{F}$ and a learning rate of $5e^{-5}$ for $\mathcal{G}$. We implement our method in PyTorch[37] and train on a Nvidia V100 GPU. For state-of-the-art comparisons, we use the authors' implementations where available. The code for our method and the new dataset C-GQA will be released upon acceptance.

Method	MIT-States							UT-Zap50K							C-GQA						
	AUC		Best					AUC		Best					AUC		Best				
	Val	Test	HM	Seen	Unseen	s	o	Val	Test	HM	Seen	Unseen	s	o	Val	Test	HM	Seen	Unseen	s	o
AttOp[35]	2.5	1.6	9.9	14.3	17.4	21.1	23.6	21.5	25.9	40.8	59.8	54.2	38.9	69.6	0.9	0.3	2.9	11.8	3.9	8.3	12.5
LE+[34]	3.0	2.0	10.7	15.0	20.1	23.5	26.3	26.4	25.7	41.0	53.0	61.9	41.2	69.2	1.2	0.6	5.3	16.1	5.0	7.4	15.6
TMN[39]	3.5	2.9	13.0	20.2	20.1	23.3	26.5	36.8	29.3	45.0	58.7	60.0	40.8	69.9	2.2	1.1	7.7	21.6	6.3	9.7	20.5
SymNet[28]	4.3	3.0	16.1	24.4	25.2	26.3	28.3	25.9	23.9	39.2	53.3	57.9	40.5	71.2	3.3	1.8	9.8	25.2	9.2	14.5	20.2
CGE_{ff} (ours)	6.8	5.1	17.2	28.7	25.3	27.9	32.0	38.7	26.4	41.2	56.8	63.6	45.0	73.9	3.6	2.5	11.9	27.5	11.7	12.7	26.9
CGE (ours)	**8.6**	**6.5**	**21.4**	**32.8**	**28.0**	**30.1**	**34.7**	**43.2**	**33.5**	**60.5**	**64.5**	**71.5**	**48.7**	**76.2**	**5.0**	**3.6**	**14.5**	**31.4**	**14.0**	**15.2**	**30.4**

Table 2: **Comparison with the state of the art:** We compare our Compositional Graph Embed (CGE) with the state of the art on Validation and Test AUC (in %); best unseen, seen and harmomic mean (HM) accuracies (in %) as well as state (s) and object (o) prediction accuracies (in %) on widely used MIT-States and UT-Zappos datasets as well as our proposed C-GQA dataset.

4.1. Comparing with the State of the Art

We compare our results with the state of the art in Table 2 and show that our Compositional Graph Embed(CGE) outperforms all previous methods by a large margin and establishes a new state of the art for Compositional Zero-shot Learning. Our detailed observations are as follows.

Generalized CZSL performance. Our framework demonstrates robustness against the label noise on MIT-States noted previously in [2]. For the generalized CZSL task, our CGE achieves a test AUC of 6.5% which is an improvement of over 2× compared to the last best 3.0% from SymNet. Similarly, as our method does not only improve results on seen labels but also unseen labels, it significantly boosts the state of the art harmonic mean, i.e. 16.1% to 21.4%. When it comes to state and object prediction accuracy, we observe an improvement from 26.3% to 30.1% for states and 28.3% to 34.7% for objects. Although our results significantly improve the state of the art on all metrics, the state and object accuracies are quite low, partially due to the label noise for this dataset.

Similar observations are confirmed on UT-Zappos, where we achieve a significant improvement on the state of the art with an AUC of 33.5% compared to 29.3% from TMN. An interesting observation is that SymNet, i.e. the state of the art on MIT States, with an AUC of 23.9% does not achieve the best performance in the generalized CZSL setting on UT Zappos. We conjecture that this is because the state labels in this dataset are not entirely representing visual transformations, something this method was designed to exploit. In this dataset, our fully compositional model improves the best harmonic mean wrt the state of the art significantly (45.0% with TMN vs 60.5% ours). Note that, this is due to a significant accuracy boost achieved on unseen compositions (60.0% vs 71.5%).

Finally on the proposed splits of the GQA dataset [15], i.e. C-GQA dataset, we achieve a test AUC of 3.6% outperforming the closest baseline by a 2×. Note that, since C-GQA has a compositional space of over 9.3k concepts, it is significantly harder than MIT-States and UT-Zappos while being truly compositional and containing cleaner labels. The state and the object accuracies of our method are 15.2% and 30.4%, *i.e.* significantly higher than the state of the art. However these results also indicate the general difficulty of the task. Similarly, our best seen and best unseen accuracies (31.4% and 14.0%) indicate a large room for improvement on this dataset, which may encourage further research with our C-GQA dataset on the CZSL task.

We also make an interesting observation on all three datasets. While SymNet uses an object classifier that is trained independently from the compositional pipeline, our method consistently outperforms it on object accuracy. We conjecture that this is because a compositional network sensitive to the information about the states is also a better object classifier, since it disentangles what it means to be an object from the state it is observed in, preventing biases to properties like textures [11]. This insight can be an avenue for future improvement in object classification.

Impact of feature representations. To quantify the improvement of our graph formulation on the same feature representations as the state of the art, we also present results of our CGE with a fixed feature extractor (Resnet18), i.e. denoted by CGE_{ff}, in Table 2. We see that this version of our model also consistently outperforms the state of the art by a large margin on MIT-States and C-GQA while matching the performance on UT-Zappos. Especially on MIT-States, the improvement over the state of the art is remarkable, i.e. 5.1% test AUC vs 3.0% test AUC with SymNet. In summary, this shows that our method benefits from both the knowledge propagation in the compositional graph and from learning better image representations.

For a fair comparison, we also allowed the previous base-

Connections in Graph	AUC	Best HM
a) Direct Word Embedding	5.9	19.4
b) (s,y) and (o,y), no self-loop on y	7.6	18.6
c) (s,y) and (o,y)	8.1	22.7
d) ***CGE*:** (s,y), (o,y) and (s,o)	**8.6**	**23.3**
e) Extra WordNet hierarchy on $\mathcal{O}$	7.9	22.0

Table 3: **Ablation over the graph connections** validates the structure of our proposed graph on the validation set of MIT-States dataset. We start from directly using the word embeddings as classifier weights to learning a globally consistent embedding from a GCN as the classifier weights (s: states, o: objects, y: compositional labels).

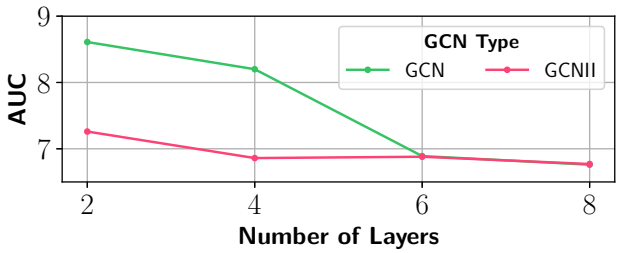

Figure 3: **Graph convolution and depth:** We compare the spectral convolution GCN[21] with the recent GCNII[33] that aims to address the over smoothing issue at increasing depth. We perform the comparison at various depths of the GCN network on the validation set of MIT-States.

lines to train end-to-end with $\mathcal{F}$. However, this results in a significant performance drop indicating they are unable to jointly learn the feature extractor against the RGB space. To address this limitation, some works[52, 47] have proposed to use a generative network to learn the distribution of image features in zero-shot problems. We, on the other hand, don't need to rely on an expensive generative network and jointly learn the image representations and the compositional classifiers in an end-to-end manner.

4.2. Ablation study

In this section we ablate our CGE model with respect to the graph connections, the graph depth and graph convolution variants.

Graph connections. We perform an ablation study with respect to the various connections in our compositional graph on the validation set of MIT-States and report results in Table 3. In the Direct Word Embedding variant, i.e. row (a) our label embedding function $\mathcal{G}$ is an average operation of state and object word embeddings. We see that, directly using word embedding of compositional labels as the classifier weights leads to an AUC of 5.9. In row (b) we represent a graph with connections between states (s) to compositional labels (y) and objects to compositional labels (y) but remove the self connection for the compositional label. In this case, the final representation of compositional labels from the GCN only combines the hidden representations of states and objects leading to an AUC of 7.6.

Row (c) represents the graph that has self connections from each compositional label in addition to the connections between states and compositional labels as well as objects and compositional labels as in row (b). We see that this variant achieves an AUC of 8.1 indicating that the hidden representation of compositional classes is beneficial.

Row (d) is our final model where we additionally incorporate the connections between states and objects in a pair to model the dependency between them. We observe that learning a representation that is consistent with states, objects and the compositional labels increases the AUC from 8.1 to 8.6 validating our choice of connections in the graph. We again emphasize that in the absence of an existing knowledge graph for compositional relations, our simple but well designed graph structure is able to capture the dependencies between various concepts.

While our final CGE does not employ an external knowledge graph, we can utilize an existing graph like WordNet [32] to get the hierarchy of the object classes similar to some baselines in zero-shot learning [46, 19]. Row (e) represents a model that exploits object hierarchy in addition to our compositional graph discussed earlier. This leads to additional 418 nodes to model the parent child relation of the objects. We see that this results in a slight performance drop with an AUC of 7.9 because this graph may not be compatible with the compositional relations.

Graph architecture. We ablate over the architecture of the graph at various depths from 2-8 layers to quantify the degree of knowledge propagation needed to achieve best performance. From Figure 3 we observe that a shallow architecture at 2 layers achieves the best AUC of 8.6 outperforming the deeper configuration. This is an established problem for the spectral graph convolution and is caused by laplacian smoothing across deeper layers[27]. To study if we are limited by a shallow representation, we use a more recent formulation of graph convolution named GCNII[33]. This method introduces a few key improvements like skip connections that remedy the laplacian smoothing problem. We see that while GCNII suffers less from the smoothing problem and maintains performance at deeper architectures, It only achieves an AUC of 7.2 for the best performing model. Since our graph is exploiting close relations between the states, objects and the compositions introduced by the dense connections for visual primitives, we are not held back by the shallow architecture. We advice future works to explore richer graphs that can facilitate deeper models.

(a) Retrieving compositional labels

(b) Retrieving images

Figure 4: **Qualitative results.** Left: We show the top-3 predictions of our model for some examples. We observe from the first four columns that all the predictions of the model are meaningful, but the model is only incentivized when it matches the label. The task of CZSL is a multi label one and future datasets need to account for this. The last column shows some examples of suboptimal labels and wrong predictions. Right: We show good candidates for retrival on all three dataset and then we perform cross-dataset retrieval for a unseen composition across C-GQA and MIT-States.

4.3. Qualitative results

We show some qualitative results for the novel compositions with their top-3 predictions in Figure 4 (left). The first three columns present some instances where the top prediction matches the label. For MIT-States and C-GQA, we notice that the remaining two answers of the model contain information visually present in the image but not in the label highlighting a limitation of current CZSL benchmarks. Different groups of states like color, age, material etc. can represent different information for the same object and are thus not mutually exclusive.

For example in column 4, row 1 the image of the cat consist of a size, surface texture and age all present in the label space of the dataset and the output of the model. However the label for this image only contains its surface texture. This puts an upper limit on compositional class accuracy dependent on the number of groups associated with an object in the label space. Specifically, column 4 of Figure 4 (left) counts as a wrong prediction but all of the top 3 predictions represent correct visual information for MIT-States and C-GQA. Unless the model learns the annotator bias, it can not achieve a perfect accuracy. Finally in column 5, we show some instances of sub-optimal and wrong labels. Specifically, the image in row 1 is entirely missing the thawed meat represented in the label, the image in row 2 can not sufficiently communicate the material information while the label in row 3 does not contain the dominant information in the image.

In Figure 4 (right) we first show image retrieval results from seen and unseen compositions. We can see that for all three datasets our method returns the correct top images for the query. We then perform cross-dataset retrieval between MIT-States and C-GQA for an unseen composition. We show a representative image from the original dataset and the top-3 retrievals from the cross dataset. While the datasets have a distribution shift between them, we see that retrievals are still meaningful. On MIT-States 2/3 retrieved images match a Mossy pond while the 3rd image is a grass field confused with the query. Similar trend is observed for the model trained on C-GQA for retrieval of a puffy pizza. The model confuses the top retrieval with a casserole followed by two images of pizzas. Nevertheless, the cross dataset retrieval shows promise towards further generalization for future works.

5. Conclusion

We propose a novel graph formulation for Compositional Zero-shot learning in the challenging generalized zero-shot setting. Since our graph does not depend on external knowledge bases, it can readily be applied to a wide variety of compositional problems. By propagating knowledge in the graph against training images of seen compositions, we learn classifiers for all compositions end-to-end. Our graph also acts like a regularizer and allows us to learn image representations consistent with the compositional nature of the task. We benchmark our method against various baselines on three datasets to establish a new state of the art in CZSL in all settings. We also highlight the limitations of existing methods and knowledge bases. We encourage future works to explore datasets with structured compositional relations and richer graphs that will allow for deeper graph models.

Acknowledgments Partially funded by ERC (853489 - DEXIM) and DFG (2064/1 – Project number 390727645).

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *CVPR*, 2013. 2

[2] Yuval Atzmon, Felix Kreuk, Uri Shalit, and Gal Chechik. A causal view of compositional zero-shot recognition. In *NeurIPS*, 2020. 2, 5, 6

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[4] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987. 2

[5] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. 4, 5

[6] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997. 2

[7] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV*, 2016. 5

[8] Jonghyun Choi, Mohammad Rastegari, Ali Farhadi, and Larry S Davis. Adding unlabeled samples to categories by learned attributes. In *CVPR*, 2013. 2

[9] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *CVPR*, 2017. 2

[10] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *ECCV*, 2014. 2

[11] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019. 6

[12] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *CVPR*, 2017. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[14] Donald D Hoffman and Whitman A Richards. Parts of recognition. *Cognition*, 18(1-3):65–96, 1984. 2

[15] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 2, 5, 6

[16] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *CVPR*, 2015. 5

[17] Seong Jae Hwang, Sathya N Ravi, Zirui Tao, Hyunwoo J Kim, Maxwell D Collins, and Vikas Singh. Tensorize, factorize and regularize: Robust visual relationship learning. In *CVPR*, 2018. 2

[18] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *CVPR*, 2015. 2

[19] Michael Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, and Eric P Xing. Rethinking knowledge graph propagation for zero-shot learning. In *CVPR*, 2019. 2, 5, 7

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[21] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 2, 4, 7

[22] Johannes Klicpera, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized pagerank. *arXiv preprint arXiv:1810.05997*, 2018. 2

[23] Johannes Klicpera, Stefan Weißenberger, and Stephan Günnemann. Diffusion improves graph learning. In *NeurIPS*, 2019. 2

[24] Christoph Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. In *TPAMI*, 2013. 2

[25] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989. 2

[26] Guohao Li, Matthias Muller, Ali Thabet, and Bernard Ghanem. Deepgcns: Can gcns go as deep as cnns? In *CVPR*, 2019. 2

[27] Q. Li, Z. Han, and X.-M. Wu. Deeper Insights into Graph Convolutional Networks for Semi-Supervised Learning. In *AAAI*, 2018. 2, 7

[28] Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. Symmetry and group in attribute-object compositions. In *CVPR*, 2020. 1, 2, 5, 6

[29] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, 2016. 2

[30] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In *ECCV*, 2012. 2

[31] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013. 3, 4, 5

[32] George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244, 1990. 2, 5, 7

[33] Zhewei Wei Ming Chen, Bolin Ding Zengfeng Huang, and Yaliang Li. Simple and deep graph convolutional networks. In *ICML*, 2020. 2, 4, 7

[34] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *CVPR*, 2017. 1, 2, 3, 5, 6

[35] Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *ECCV*, 2018. 1, 2, 5, 6

[36] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010. 5

[37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 5

[38] Novi Patricia and Barbara Caputo. Learning to learn, from transfer learning to domain adaptation: A unifying perspective. In *CVPR*, 2014. 2

[39] Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. In *ICCV*, 2019. 1, 2, 3, 4, 5, 6

[40] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017. 2

[41] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *CVPR*, 2016. 2

[42] Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. Dropedge: Towards deep graph convolutional networks on node classification. In *ICLR*, 2019. 2

[43] Ruslan Salakhutdinov, Antonio Torralba, and Josh Tenenbaum. Learning to share visual appearance for multiclass object detection. In *CVPR*, 2011. 1

[44] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, 2013. 2

[45] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 5

[46] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *CVPR*, 2018. 2, 5, 7

[47] Xin Wang, Fisher Yu, Trevor Darrell, and Joseph E Gonzalez. Task-aware feature generation for zero-shot compositional learning. *arXiv preprint arXiv:1906.04854*, 2019. 2, 7

[48] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *NIPS*, 2017. 1

[49] Felix Wu, Tianyi Zhang, Amauri Holanda de Souza Jr, Christopher Fifty, Tao Yu, and Kilian Q Weinberger. Simplifying graph convolutional networks. *arXiv preprint arXiv:1902.07153*, 2019. 2

[50] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *CVPR*, 2019. 5

[51] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE TPAMI*, 41(9):2251–2265, 2018. 3, 5

[52] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, 2018. 2, 7

[53] Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. Representation learning on graphs with jumping knowledge networks. *arXiv preprint arXiv:1806.03536*, 2018. 2

[54] Muli Yang, Cheng Deng, Junchi Yan, Xianglong Liu, and Dacheng Tao. Learning unseen concepts via hierarchical decomposition and composition. In *CVPR*, 2020. 2

[55] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *CVPR*, 2014. 5

[56] Aron Yu and Kristen Grauman. Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *ICCV*, 2017. 5

[57] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 2

[58] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *CVPR*, 2017. 2

[59] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *CVPR*, 2018. 2